

الجمهورية الجزائرية الديمقراطية الشعبية

People's Democratic Republic of Algeria

وزارة التعليم العالي والبحث العلمي

Ministry of Higher Education and Scientific Research

جامعة زيان ماحور بالجلية
Ziane Achour University of Djelfa



كلية العلوم والتكنولوجيا
Faculty of Science and Technology

Civil Engineering Department

Order N° : / 2026

Defense authorization N° : / 2026

DOCTORAL THESIS

3rd Cycle Doctoral (D-LMD)

Presented by :

BOUGUEDRA Billel

With a view to obtaining the doctoral diploma in 3rd Cycle (D-LMD)

Branch : Civil Engineering

Specialty : Structures

Topic

**Application of machine learning techniques to predict
the subgrades-soil strength for pavement structures**

Supported, on 07/07/2026, before the jury composed of :

Last and first name :	Grade :	Institution of affiliation :	Designation :
Dr. ATTIA Ahmed	Lecturer A	University of Djelfa	President
Dr. SANDJAK Khaled	Lecturer A	University of Boumerdas	Supervisor
Pr. OUANANI Mouloud	Professor	University of Djelfa	Co-Supervisor
Pr. MESSAFER Tahar	Professor	University of Boumerdas	Examiner
Dr. CHERIET Fayssal	Lecturer A	University of Djelfa	Examiner

Djelfa University, FST - 2026



THESE DE DOCTORAT

Doctorat 3^{ème} Cycle (D-LMD)

Présentée par :

BOUGUEDRA Billel

En vue de l'obtention du diplôme de Doctorat en 3^{ème} Cycle (D-LMD)

Filière : Génie Civil

Spécialité : Structures

Thème

**Application des techniques d'apprentissage automatique pour la prédiction
de la rigidité des sols-soutiens des structures de chaussées**

Soutenue publiquement, le 07/07/2026, devant le jury composé de :

Nom et prénom :	Grade :	Etablissement de rattachement :	Désignation :
Dr. ATTIA Ahmed	M. C. A	Université de Djelfa	Président
Dr. SANDJAK Khaled	M. C. A	Université de Boumerdès	Rapporteur
Pr. OUANANI Mouloud	Professeur	Université de Djelfa	Co-Rapporteur
Pr. MESSAFER Tahar	Professeur	Université de Boumerdès	Examineur
Dr. CHERIET Fayssal	M. C. A	Université de Djelfa	Examineur

Remerciements

La gratitude commence par la louange à ALLAH qui nous a permis de réussir, nous a guidés vers le succès et nous a conduits à ces rangs et honneurs.

Je tiens à exprimer ma profonde reconnaissance et ma gratitude aux : Les superviseurs de ma thèse Dr. SANDJAK Khaled et Pr. OUANANI Mouloud pour leurs conseils, leur suivi, leur assistance et leur soutien tout au long de ma formation doctorale.

J'ai également l'honneur de remercier les membres du jury : du président du comité Dr. ATTIA Ahmed aux examinateurs Pr. MESSAFER Tahar et Dr. CHERJET Fayssal.

Mes remerciements également vont à :

Pr. Hegazy Rezk de l'Université du Prince Sattam Bin-Abdulaziz et Dr. Backar Ahmed de l'Université d'Umm Al-Qura d'Arabie Saoudite pour leur contribution et leur aide dans la publication de l'article.

Ing. MAIZA Hamza directeur du laboratoire de sol et des matériaux du CTTP d'Alger, pour sa supervision et ses conseils pendant mon séjour au laboratoire.

Dr. Allout Naas le vice-doyen et Pr. Guettala Salim Chef de CFD, pour leur soutien et leurs facilités administratives.

Pr. BOUCHAIR Abdelhamid responsable de génie civil à l'université de Clermont Auvergne INP de France pour son accueil et ses conseils pendant mon stage en France.

À toute l'administration de la faculté ST, de la doyenette à la direction.

Dédicace

La dédicace est une confirmation du don et un cadeau pour les proches. Je dédie ce travail à :

*Mes chers parents pour leur encouragement, leur soutien et leur fierté de mes réussites.
Merci à vous deux pour tout.*

Mes chers frères Bachir, Mohamed, Oussama et Abdelhakj, ainsi qu'à toute ma famille.

Mes chers amis Riadh et Younes Missoum de Paris.

À tous ceux qui ont contribué à mon succès et m'ont soutenu, même par un simple mot d'encouragement.

Billel.

Résumé :

Le CBR est un indicateur clé de la rigidité des structures de chaussées. Cependant, sa détermination nécessite des efforts sur le terrain et en laboratoire, ainsi qu'un temps considérable. C'est pourquoi l'AI a été utilisée pour prédire les valeurs du CBR. Grâce à l'observation et à le questionnaire au niveau des archives du laboratoire de sol et des matériaux du Organisme CTTP, une base de données a été constituée, comprenant 1123 échantillons de sols fins provenant de divers régions du nord-est de l'Algérie. Cette base de données comprend 12 paramètres du sol qui représentent la granulométrie, la plasticité et les caractéristiques de compactage. Cinq modèles de ML ont été développés à savoir : MLR, MLPNN, RF, GPR et SVR. La technique de validation croisée en 10-fold a été intégrée aux les cinq modèles pour une division optimale des données entre l'entraînement et le test et pour éviter le sur-apprentissage. De plus, les deux modèles GPR et SVR ont été hybride avec l'Algorithme d'Optimisation Bayésienne (TPE) pour ajuster les hyper-paramètres de ces deux modèles afin d'obtenir une meilleure efficacité. Les performances des modèles sont évaluées selon les valeurs de coefficient de détermination (R^2) ainsi que les erreurs (MSE, RMSE, MAE et MAPE(%)). D'autre part l'utilisation des techniques avancées de sélection des données à travers les deux méthodes RFE et EFS a permis de réduire la taille des données de 12 entrées à seulement 7, ce qui a également amélioré la performance de modèles. Les résultats de l'étude ont donné des valeurs satisfaisantes avec les deux modèles GPR et SVR, avec des valeurs de R^2 dépassant le seuil de 95% et de faibles taux d'erreur, ce qui prouve la bonne réponse des deux modèles aux techniques d'optimisation dans cette étude. Cependant, à travers la comparaison, il a été démontré que le GPR est plus efficace et surpasse les modèles conventionnels, En plus de sa supériorité sur les modèles hybrides (SVR, H-DNN, H-XGB, H-LGB et H-Ada). Les résultats de performance de modèle GPR hybridé sont : ($R^2 = 0.9982/0.9935$, $MSE = 0.0768/0.2974$, $RMSE = 0.2771/0.5453$, $MAE = 0.2107/0.4364$ et $MAPE(\%) = 5.32/13.79$). En fin l'application de la méthode SHAP sur le modèle de GPR et Sobol sur le modèle de SVR a montré que la densité sèche optimale (MDD) des échantillons est l'entrée ayant le plus d'impact sur la sortie CBR.

Mots-clés : sols, prédiction, rigidité, apprentissage automatique, structures de chaussées, supports.

Abstract :

The CBR is a key indicator of the strength of pavement structures. However, its determination requires efforts in the field and in the laboratory, as well as considerable time. This is why AI has been used to predict CBR values. Through observation and a questionnaire at the archives of the soil and materials laboratory of the CTTP organization, a database was created, including 1123 samples of fine soils from various regions of north-eastern Algeria. This database includes 12 soil parameters that represent grain size, plasticity, and compaction characteristics. Five ML models were developed, namely: MLR, MLPNN, RF, GPR, and SVR. The 10-fold cross validation technique was integrated into the five models for optimal data partitioning between training and testing and to avoid overfitting. Additionally, the GPR and SVR models were hybridized with the Bayesian Optimization Algorithm (TPE) to fine-tune the hyperparameters of these two models for better efficiency. The performance of the models is evaluated according to the coefficient of determination (R^2) as well as errors (MSE, RMSE, MAE, and MAPE(%)). Furthermore, the use of advanced data selection techniques through the two methods RFE and EFS allowed for reducing the size of the data from 12 inputs to only 7, which also improved the performance of the models. The results of the study provided satisfactory values for both GPR and SVR models, with R^2 values exceeding the threshold of 95% and low error rates, demonstrating the good responsiveness of both models to optimization techniques in this study. However, through comparison, it has been shown that GPR is more effective and surpasses conventional models, in addition to its superiority over hybrid models (SVR, H-DNN, H-XGB, H-LGB, and H-Ada). The performance results of the hybrid GPR model are : ($R^2 = 0.9982/0.9935$, $MSE = 0.0768/0.2974$, $RMSE = 0.2771/0.5453$, $MAE = 0.2107/0.4364$ et $MAPE(\%) = 5.32/13.79$). Finally, the application of the SHAP method on the GPR model and Sobol on the SVR model showed that the optimal dry density (MDD) of the samples is the input having the most impact on the CBR output.

Key-words : soil, prediction, strength, machine learning, pavement structures, subgrades.

ملخص :

يُعتبر CBR مؤشراً رئيسياً لصلابة هياكل الرصف. ومع ذلك، يتطلب تحديد قيمه جهوداً في الميدان والمختبر، فضلاً عن الوقت المستغرق. لهذا السبب، تم استخدام الذكاء الاصطناعي للتنبؤ بقيم CBR. من خلال الملاحظة والاستبيان في أرشيفات مختبر التربة والمواد لهيئة CTPP، تم إنشاء قاعدة بيانات تتضمن 1123 عينة من التربة الدقيقة من مختلف مناطق الشمال الشرقي للجزائر. تحتوي هذه القاعدة على 12 معلمة للتربة تمثل حجم الحبوب، واللدونة، وخصائص الدمك. تم تطوير خمسة نماذج من التعلم الآلي، وهي : MLR، MLPNN، RF، GPR و SVR. تم دمج تقنية التحقق المتقاطع بنظام 10-طيات في النماذج الخمسة لتحقيق تقسيم مثالي للبيانات بين التدريب والاختبار ولتجنب التكيف الزائد. بالإضافة إلى ذلك، تم دمج نموذجي GPR و SVR مع خوارزمية تحسين بايزي (TPE) لضبط المعلمات الفائقة لهذين النموذجين من أجل تحقيق كفاءة أفضل. يُقِيم أداء النماذج وفقاً لقيم معامل التحديد (R^2) وكذلك الأخطاء (MAE، RMSE، MSE) و (MAPE(%). علاوة على ذلك، سمحت تقنيات اختيار البيانات المتقدمة من خلال الطريقتين RFE و EFS بتقليل حجم البيانات من 12 مدخلاً إلى 7 فقط، مما حسن أيضاً من أداء النماذج. قدمت نتائج الدراسة قيماً مرضية لكل من نموذجي GPR و SVR، حيث تجاوزت قيم R^2 عتبة 95٪ ومعدلات خطأ منخفضة، مما يثبت الاستجابة الجيدة لكلا النموذجين لتقنيات التحسين في هذه الدراسة. ثم ومن خلال المقارنة، تم التوصل إلى أن GPR الهجين أكثر فعالية ويتفوق على النماذج التقليدية، بالإضافة إلى تفوقه على النماذج الهجينة (SVR، H-DNN، H-XGB، H-LGB و H-Ada). كانت نتائج أداء نموذج GPR الهجين كما يلي : $R^2 = 0.9982/0.9935$, $MSE = 0.0768/0.2974$, $RMSE = 0.2771/0.5453$, $MAE = 0.2107/0.4364$ والنسبة $(MAPE(\%)) = 5.32/13.7$. وأخيراً، أظهر تطبيق تقنية SHAP على نموذج GPR و Sobol على نموذج SVR أن الكثافة الجافة المثلى (MDD) للعينات هي المدخل الأكثر تأثيراً على المخرج CBR.

الكلمات المفتاحية : التربة، التنبؤ، الصلابة، التعليم الآلي، بنيات الأرصفة، الداعمة.

Des travaux et publications issus de cette thèse :

❖ Articles internationales :

B, Bouguedra; K, Sandjak; M, Ouanani; A.H, Backar; H, Rezk. A Hybrid GPR Framework with Feature Selection for Enhanced Prediction and Explainability of Subgrade Soil Strength. *Scientific Reports*, 16(1), 18603, pp 01-27, ISSN: 2045-2322, Springer Nature, UK, 2026. <https://doi.org/10.1038/s41598-026-48770-2> [Revue A selon DGRSDT, Q1 selon Scopus et PubMed].

B, Bouguedra; K, Sandjak; M, Ouanani. Evaluating the strength of subgrade soil for pavement design: an analysis using support vector regression model optimized by bayesian algorithm. *Studies in Engineering and Exact Sciences*, v5 (2), pp 01-23, ISSN: 2764-0981, Curitiba, Brazil, 2024. <https://doi.org/10.54021/seesv5n2-159> [Revue non catégorisé a l'Algérie].

❖ Conférences internationales :

B, Bouguedra; K, Sandjak; M, Ouanani; H, Rezk. Leveraging Noise Injection to Mitigate Samples Limitations and Enhance Predictive Model Performance. *The 18th International Conference on Optimization and Information Systems COSI'26*, Springer, pp. 01-11, Batna. Algeria, 12th-14th May, 2026.

B, Bouguedra; K, Sandjak; M, Ouanani; H, Rezk. Exhaustive Features Selection with Randomized Decision Tree Model to Forecast the Soil Types According to AASHTO Classification. *Second Edition of the International Seminar on Materials and Engineering Structures ISMES'25*, pp. 01-10, Relizane. Algeria, 7th-8th February, 2026.

B, Bouguedra; K, Sandjak; M, Ouanani. Machine Learning techniques to estimate the strength value of clayey-silty soil from Algerian highlands. *The First International Conference on Recent Challenges in Civil Engineering and Environmental Sustainability ICRCCEES'25*, pp. 01-08, Djelfa. Algeria, 10th-11th December, 2025.

B, Bouguedra; K, Sandjak; M, Ouanani. Bayesian Optimization Algorithm based Support Vector Regression analysis for Predicting the Strength of Subgrade soil for Pavement Design. *18th African Regional Conference on Soil Mechanics and Geotechnical Engineering ARCSMGE'24*, ISSMGE proceeding, pp. 1035-1045, ISBN: 978-9969-9836-0-9, Algiers. Algeria, 6th-9th October, 2024. <https://www.issmge.org/publications/author/billel-bouguedra>

❖ Communications nationales :

B, Bouguedra; K, Sandjak; M, Ouanani; H, Rezk. Accurate Prediction of Optimum Moisture Content for Fine-Grained Soils Using XGBoost-RFE Model. *1st National Seminar on Innovation and Artificial Intelligence in Computational Process and Chemical Engineering for Green Applications I-AICPCGE'26*. pp. 01-10, Médéa. Algérie, 20 Mai, 2026.

B, Bouguedra; K, Sandjak; M, Ouanani; H, Rezk. A Trees-Based Approach to Estimate the Bearing Capacity of Soaked Soil. *The 1st National Congress on Materials Energy and Environment NCMEE'26*. pp. 01-10, Tipaza. Algérie, 25 Avril, 2026.

B, Bouguedra; K, Sandjak; M, Ouanani. Hybridation de model SVR avec l'algorithme d'optimisation Bayésienne pour la prédiction de la portance des sols-supports de chaussées. *1^{er} Séminaire National de Génie Civil SNGC'25*. pp.01-04, ISBN: 978-9969-9733-6-5, Boumerdès. Algérie, 8-9 Décembre, 2025. <https://dspace.univ-boumerdes.dz/handle/123456789/15929>

B, Bouguedra; K, Sandjak; M, Ouanani. Différentes techniques d'optimisation des modèles de ML pour prédire des valeurs expérimentales en physique. *Atelier sur les applications de l'intelligence artificielle en physique*, "Poster, pp. 01, Médéa. Algérie, 6 Novembre, 2024.

B, Bouguedra; K, Sandjak; M, Ouanani. BOA based SVR analysis for Predicting the Strength of Subgrade soil for Pavement Design. *Congrès National de Génie Civil et d'Hydraulique CNGCH'23*. pp. 01-13, Guelma. Algérie, 12-13 Novembre, 2023.

❖ Prix internationaux :

Bright Spark Lecture Award. Attribué par le président d'organisation internationale ISSMGE comme meilleur papier Algérien lors la conférence ARCSMGE'24.

Sommaire de matières :

Matière :	Page :
Résumé	I
Des travaux issus de cette thèse	IV
Sommaire de matières	VI
Liste des tableaux	XII
Liste des figures	XIII
Liste des abréviations	XVI
Introduction générale	2
Chapitre 1 : Analyse des recherches antérieurs	
I. Introduction	
II. Intelligence artificielle et la prédiction de de l'indice de portance Californien CBR	6
III. Les recherches précédentes	
III.1. Les travaux de Semachew et Betelhem 2023	
III.1.1. Application d'étude sur les sols fins	7
III.1.2. L'étude statistique de base des données	8
III.1.3. Développement de modèles de forêts aléatoires RF, les arbres de décision DT et multi-linaire perceptron MLP pour prédire les valeurs de CBR	
III.1.3.1.Le modèle d'arbre aléatoire RF	9
III.1.3.2.Le modèle Arbre de décision DT	
III.1.3.3.Le modèle de régression linéaire MLP	10
III.1.4. Evaluation des résultats d'étude de Semachew et Betelhem 2023	
III.2. Les travaux de Yusuf Erzin et autres 2022	12
III.2.1 Application d'étude sur les sols argileux traités à la chaux avec déchet de pneus	
III.2.2. Choix de modèles ANN et MLR, en évaluant par la variance VAF pour prédire le CBR	13
III.2.3. Evaluation des résultats d'étude de Yusuf Erzin et autres 2022	14
III.3. Les travaux de Van et Hai Quan 2021	15
III.3.1. Mélange des cendres volantes avec les sols cohérents pour prédire le CBR	
III.3.2. L'étude statistique de base des données	16
III.3.3. Evaluation de performance de la bibliothèque LightGBM pour prédire le CBR	17
III.4. SK Kamrul et autres 2020	18
III.4.1. Application sur les sols fins de bassin Bengale a l'Inde	
III.4.2. Développement de modèles GEP, ANN et la méthode de Kriggen pour prédire le CBR	19
III.4.2.1.La programmation génétique d'expression (GEP)	
III.4.2.2.Les réseaux de neurones artificiels (ANN)	20
III.4.2.3.La méthode de Kriggen	
III.4.3. Résultats et discussion	21
III.5. Taskiran 2010	22
III.5.1. La base des données et la région d'échantillonnage	
III.5.2. Comparaison de performance de modèle ANN et de GEP pour prédire le CBR	23
III.5.2.1.Les réseaux de neurones artificiels (ANN)	
III.5.2.2.La programmation de l'expression génétique (GEP)	24
III.5.3. Evaluation des résultats d'étude de Taskiran 2010	25

IV.	Conclusion et suggestions	27
Chapitre 2 : Collection et préparation de base des données		
I.	Introduction	29
II.	L'étude de terrain	
II.1.	Carte d'identification par laboratoire d'étude	
III.	La base de données	
III.1.	La structure des chaussées (La couche sélectionnée)	30
III.1.1.	Sol-support	
III.1.2.	Couche de forme	
III.1.3.	Couches d'assise	
III.1.4.	Couches de surfaces	31
III.2.	Les zones d'échantillonnage	
III.3.	Les protocoles expérimentaux pour les tests de laboratoire CTTP	
III.3.1.	L'analyse granulométrique selon les normes NF P 94-056	
III.3.1.1.	Objectif de l'essai	32
III.3.1.2.	Principe de l'essai	
III.3.1.3.	Matériel requis	
III.3.1.4.	Exécution de l'essai	
III.3.2.	L'essai sédimentométrique selon la norme NF P94-057	34
III.3.2.1.	Objectif de l'essai	
III.3.2.2.	Principe de l'essai	
III.3.2.3.	Matériel requis	
III.3.2.4.	Exécution de l'essai	35
III.3.3.	L'essai de limites d'Atterberg selon les normes NF EN ISO 17892-12: 2018	
III.3.3.1.	Objectif de l'essai	
III.3.3.2.	Principe de l'essai	
III.3.3.3.	Matériel requis	36
III.3.3.4.	Exécution de l'essai	
III.3.4.	L'essai de Proctor selon les normes NF P 94-093	
III.3.4.1.	Objectif de l'essai	
III.3.4.2.	Principe de l'essai	37
III.3.4.3.	Matériel requis	
III.3.4.4.	Exécution de l'essai	
III.3.5.	L'essai de CBR (California Bearing Ratio) selon les normes NF P 94-078	
III.3.5.1.	Objectif de l'essai	38
III.3.5.2.	Principe de l'essai	
III.3.5.3.	Matériel requis	
III.3.5.4.	Exécution de l'essai	
III.4.	Classification des sols	39
III.4.1.	Classification selon USCS	
III.4.2.	Classification selon AASHTO	
III.4.2.1.	Groupe des sols fins	
III.5.	Exemple de base de données	40
III.6.	Etude statistique de base de données	43
III.6.1.	Un aperçu statistique des valeurs de variables	
III.6.2.	Le coefficient de corrélation (Pearson) r	
III.6.3.	La matrice de corrélation (Pearson)	
III.6.4.	L'histogramme d'erreur	44
		45

IV.	Conclusion	47
Chapitre 3 : Architecture des modèles		
I.	Introduction	49
II.	La prédiction	
III.	L'environnement de la modélisation des modèles	
III.1.	Programme PYTHON	50
IV.	Choix des modèles et de ses architectures	
IV.1.	Division de la base des données	
IV.2.	La méthode de validation croisée K-fold	
IV.3.	L'apprentissage automatique (Machine Learning ML) et ces types	51
IV.3.1.	L'apprentissage supervisé	
IV.3.2.	L'apprentissage non supervisé	
IV.3.3.	L'apprentissage par renforcement	
IV.4.	Choix de la méthode d'apprentissage automatique	52
IV.4.1.	Identification du format de la variable de sortie	
IV.4.2.	Identifier le type de problème d'apprentissage machine	
IV.4.3.	Choix de la méthode	
IV.5.	La régression	
IV.5.1.	Régression linéaire	53
IV.5.1.1.	Régression simple SLR	
IV.5.1.2.	Régression multiple MLR	54
IV.5.1.3.	L'hyper-paramètre de la régression linéaire	
IV.5.2.	Régression non linéaire	
IV.6.	Théories de modèles	55
IV.6.1.	Réseaux de Neurones Artificiel ANN (Artificial Neural Network)	
IV.6.1.1.	Lecture biologique sur le modèle d'ANN	
IV.6.1.2.	Bref historique sur le développement de modèle ANN	56
IV.6.1.3.	Neurone formel	
IV.6.1.3.1.	Perceptron	57
IV.6.1.3.2.	Perceptron multicouches MLP (multi-layer perceptron)	58
IV.6.1.3.3.	Correction des erreurs	59
IV.6.1.4.	Les hyper-paramètres de modèle de MLP d'ANN	60
IV.6.2.	Les forêts aléatoires (RF)	
IV.6.2.1.	Bref historique sur le développement de modèle de RF	
IV.6.2.2.	Arbre de décision DT	61
IV.6.2.3.	Bref historique sur le développement de modèle de DT	
IV.6.2.4.	La structure de modèle de RF et (DT)	
IV.6.2.5.	Processus de division des données	
IV.6.2.6.	Les paramètres clés de modèle de RF et (DT)	63
IV.6.3.	Régression par Processus Gaussien GPR	
IV.6.3.1.	Bref historique sur le développement de modèle GPR	64
IV.6.3.2.	Processus Gaussien GP	
IV.6.3.3.	Présentation théorique de l'algorithme de GPR	66
IV.6.3.4.	Les noyaux de modèle de GPR	
IV.6.3.4.1.	Noyau Square Exponentiel (SE) / (RBF)	67
IV.6.3.4.2.	Noyau Matern	68
IV.6.3.4.3.	Noyau Exponentiel	
IV.6.3.4.4.	Noyau Exp-Square-Sine (Périodique)	69

IV.6.3.4.5. Noyau de covariance Rational Quadratique (RQ)	70
IV.6.3.4.6. Noyau Linéaire / Dot-Product (Quadratique)	71
IV.6.3.5. Résoudre du bruit dans les données (Le noyau de bruit blanc)	72
IV.6.3.6. Les hyper-paramètres de modèle de GPR	74
IV.6.4. Régression à Vecteurs de Support SVR	
IV.6.4.1. Bref historique sur le développement de modèle de SVR	
IV.6.4.2. Présentation théorique de l'algorithme de SVR	75
IV.6.4.3. Les noyaux de modèle de SVR	76
IV.6.4.3.1. Noyau linéaire	
IV.6.4.3.2. Noyau polynomial	77
IV.6.4.3.3. Noyau Gaussien (ou RBF)	
IV.6.4.3.4. Noyau sigmoïde	78
IV.6.4.4. Les hyper-paramètres de modèle de SVR	79
V. Sélection de caractéristiques d'entrées (Features selection technique)	81
V.1. Techniques de sélection de caractéristiques supervisées	
V.1.1. Approche basée sur les filtres	
V.1.1.1. Gain d'information	82
V.1.1.2. Test du Chi-carré	
V.1.1.3. Score de Fisher	
V.1.1.4. Ratio de valeurs manquantes	
V.1.2. Approche basée sur des Wrappers	
V.1.2.1. Sélection avant	83
V.1.2.2. Sélection arrière	
V.1.2.3. Sélection exhaustive des caractéristiques (EFS)	84
V.1.2.4. Élimination récursive des caractéristiques (RFE)	
V.1.3. Approche intégrée	
V.1.3.1. Régularisation	
V.1.3.2. Importance des forêts aléatoires	
V.2. Techniques de sélection de caractéristiques non supervisées	
V.2.1. Analyse en Composantes Principales (ACP)	
V.2.2. Analyse en Composantes Indépendantes (ACI)	85
V.2.3. Factorisation de Matrice Non Négative (NMF)	
V.2.4. Intégration Stochastique à Proximité t-distribuée (t-SNE)	
V.2.5. Auto-encodeur	86
VI. Méthode d'optimisation	
VI.1. Hyper-paramètres	87
VI.1.1. Identification des hyper-paramètres	
VI.1.2. Réglage des hyper-paramètres	
VI.1.3. Fonctionnement de réglage des hyper-paramètres	88
VI.1.4. Les techniques d'ajustement des hyper-paramètres	
VI.1.4.1. Optimisation Bayésienne	
VI.1.4.2. Recherche dans la grille	
VI.1.4.3. Recherche aléatoire	
VI.2. L'Optimisation Bayésienne BOA	89
VI.2.1. Principaux composants (stratégies) de l'Optimisation Bayésienne BOA	
VI.2.1.1. Le modèle surrogate	
VI.2.1.2. La fonction d'acquisition	
VI.2.1.3. La stratégie d'équilibre entre exploration et exploitation	

VI.2.1.4.	L'estimateur de Parzen structuré arborescent (TPE)	
VI.2.2.	Le Processus Itératif de l'Optimisation Bayésienne BOA	
VI.2.3.	Bonnes pratiques pour le réglage des hyper-paramètres	90
VI.2.4.	Présentation théorique d'Algorithme d'Optimisation Bayésien BOA et TPE	
VI.2.5.	Les étapes d'exécution d'un Algorithme d'Optimisation Bayésien BOA / TPE	92
VI.2.5.1.	BO avec (Processus Gaussien - GP)	
VI.2.5.2.	TPE (Tree-structured Parzen Estimator)	93
VI.2.6.	Comparaison entre les deux stratégies de BOA (avec GP) et TPE	94
VI.3.	L'hybridation des modèles	95
VII.	Critères d'évaluation des modèles	
VII.1.	Erreur Quadratique Moyenne (MSE)	
VII.2.	Racine Carrée de l'Erreur Quadratique Moyenne (RMSE)	
VII.3.	Erreur Absolue Moyenne (MAE)	
VII.4.	Erreur Absolue Moyenne en Pourcentage (MAPE)	
VII.5.	Coefficient de Détermination (R^2)	96
VIII.	Analyse de sensibilité (Impact des entrées sur la sortie CBR)	97
VIII.1.	Types d'Analyse de Sensibilité	
VIII.1.1.	Analyse de sensibilité locale	
VIII.1.2.	Analyse de sensibilité globale	
VIII.2.	Effet d'ordre 1 et Effet d'interaction	
VIII.3.	Méthodologies et les techniques principales d'analyse de sensibilité	98
VIII.3.1.	One-At-a-Time (OAT)	
VIII.3.2.	Méthodes basées sur les dérivées	
VIII.3.3.	Méthodes de screening (ex. méthode de Morris)	
VIII.3.4.	Méthodes basées sur la variance (ex. indices de Sobol)	
VIII.3.5.	Méthodes modernes liées au ML	
VIII.3.5.1.	SHAP (SHapley Additive exPlanations)	
VIII.3.5.2.	Importance par permutation	
VIII.3.5.3.	LIME (Local Interpretable Model-agnostic Explanations)	
VIII.4.	Application à la prédiction du CBR	99
VIII.5.	Présentation théorique de la méthode de SHAP	
VIII.6.	Présentation théorique de la méthode d'analyse de Sobol	100
VIII.6.1.	Les indices de Sobol	
VIII.7.	Comparaison entre les deux méthodes (SHAP et Sobol)	101
IX.	Organigrammes de modélisation	
IX.1.	Modélisation des modèles conventionnels (par défaut)	102
IX.2.	Modélisation des modèles hybrides	
IX.2.1.	Algorithme d'exécution des techniques RFE et EFS	103
IX.2.2.	Algorithme d'exécution de technique d'optimisation TPE	
IX.2.3.	Algorithme global de l'étude	104
X.	Conclusion	105
Chapitre 4 : Résultats et optimisation		
I.	Introduction	
II.	Développement de modèles conventionnels (CML)	107
II.1.	Développement de modèle MLR	
II.2.	Développement de modèle MLPNN d'ANN	108
II.3.	Développement de modèle RF	
III.	Développement des modèles hybrides	109

III.1.	Cadre méthodologique pour le développement des modèles hybrides	
III.2.	Développement du modèle hybride TPE-GPR-RFE	110
III.3.	Développement du modèle hybride TPE-SVR-EFS	112
III.4.	Intégration méthodologique des deux modèles	
IV.	Évaluation des résultats de performance	114
IV.1.	Comparaison du modèle GPR hybride avec le modèle GPR par défaut	
IV.2.	Comparaison de performance du modèle hybride TPE-GPR-RFE avec des modèles conventionnels (CML)	116
IV.2.1.	Performance du modèle de MLR	
IV.2.2.	Performance du modèle de MLPNN	
IV.2.3.	Performance du modèle de RF	118
IV.2.4.	Comparaison de performance des modèles par le diagramme de Taylor	
IV.2.5.	Résumé et conclusion de la comparaison	119
IV.3.	Comparaison du modèle SVR hybride avec le modèle SVR par défaut	120
IV.4.	Comparaison entre les modèles hybrides	121
V.	Analyse de sensibilité (Sensitivity Analysis)	
V.1.	Application de la méthode de SHAP sur le modèle hybride TPE-GPR-RFE	124
V.2.	Application de la méthode de Sobol sur le modèle hybride TPE-SVR-EFS	131
V.2.1.	Les indices de premier ordre (S_i)	
V.2.2.	Les indices totaux (ST_i)	
V.2.3.	Les indices d'interaction ou bien deuxième ordre (S_{ij})	132
V.2.4.	Analyse interprétative	133
V.3.	Comparaison entre les résultats de méthode de SHAP sur TPE-GPR-RFE et de Sobol sur TPE-SVR-EFS	134
VI.	Conclusion	135
Conclusion générale et recommandations		137
Références bibliographiques		XVIII

Liste des tableaux :

Tableau :	Page :
Chapitre 1 : Recherches antérieures	
Tableau 1.1. Résumé de recherches précédentes	6
Tableau 1.2. Les résultats d'étude (Semachew et Betelhem 2023)	10
Tableau 1.3. Les résultats d'étude (Yusuf Erzin et autres 2022)	14
Tableau 1.4. Les résultats d'étude (Van et Hai Quan 2021)	17
Tableau 1.5. Exemple sur la base des données de (SK Kamrul 2020)	19
Tableau 1.6. La performance statistique de modèles ANN/GEP	26
Chapitre 2 : Collection et préparation de base des données	
Tableau 2.1. Résumé des symboles d'entrée et de sortie et de leurs significations et expériences	32
Tableau 2.2. Regroupe des sols d'après la granulométrie selon AASHTO système	40
Tableau 2.3. Identification entre les deux systèmes de classification de sol AASHTO et USCS	41
Tableau 2.4. Exemple de base de données	42
Tableau 2.5. Les valeurs statistiques de base de données à la partie (d'entraînement et de test)	43
Chapitre 3 : Architecture des modèles	
Tableau 3.1. Résumé de modèles sélectionnés, les noyaux, les hyper-paramètres et les (+ avantages, – inconvénients) de chacun noyau	80
Tableau 3.2. Comparaison entre les deux stratégies de BOA (GP et TPE)	94
Chapitre 4 : Résultats et optimisation	
Tableau 4.1. Les hyper-paramètres de modèle MLPNN d'ANN utilisé	108
Tableau 4.2. Les hyper-paramètres optimisés pour le modèle GPR utilisé	110
Tableau 4.3. Résumé statistique des valeurs mesurées et des prédictions du modèle TPE-SVR-EFS	112
Tableau 4.4. Les hyper-paramètres optimisés pour le modèle SVR utilisé	113
Tableau 4.5. Les indices de performance comparée entre le modèle de GPR par défaut et l'hybride TPE-GPR-RFE à la partie d'entraînement et de test	115
Tableau 4.6. Les indices de performance comparée entre le modèle TPE-GPR-RFE et les autres modèles conventionnels à la partie d'entraînement et de test	117
Tableau 4.7. Les indices de performance comparée entre le modèle de SVR par défaut et l'hybride TPE-SVR-EFS à la partie d'entraînement et de test	120
Tableau 4.8. Les indices de performance comparent les modèles hybrides à la partie d'entraînement et de test	122
Tableau 4.9. Les indices de Sobol de premier ordre (S_i) avec le modèle TPE-SVR-EFS	132
Tableau 4.10. Les indices de Sobol totaux (ST_i) avec le modèle hybride TPE-SVR-EFS	132
Tableau 4.11. Les indices de Sobol d'interaction entre MDD et les autres entées avec le modèle hybride TPE-SVR-EFS	133

Liste des figures :

Figure :	Page :
Chapitre 1 : Recherches antérieurs	
Figure 1.1. La matrice de corrélation (recherche de Semachew et Betelhem 2023)	8
Figure 1.2. Distribution des attributs à l'aide d'un diagramme de densité	9
Figure 1.3. Comparaison de MAE, MSE et RMSE pour les algorithmes	10
Figure 1.4. Les valeurs réelles et prédites des ensembles d'entraînement et de test à l'aide de RF model	11
Figure 1.5. Test de sensibilité (recherche de Semachew et Betelhem 2023)	
Figure 1.6. L'architecture de modèle d'ANN	13
Figure 1.7. Comparaison entre les valeurs de CBR obtenus d'après les essais expérimentaux et les valeurs prédictives dans les deux étapes d'entraînement et de test pour le modèle d'ANN	
Figure 1.8. Comparaison entre les valeurs de CBR obtenus d'après les essais expérimentaux et les valeurs prédictives dans toutes les étapes pour le modèle de MLR	14
Figure 1.9. La relation entre l'erreur de pourcentage mise à l'échelle et la fréquence cumulative pour le modèle d'ANN à gauche et de MLR à droite consécutivement	
Figure 1.10. La matrice de corrélation (recherche de Van et Hai 2021)	16
Figure 1.11. Distribution et ligne de corrélation de chaque variable d'entrée et de sortie	
Figure 1.12. Analyse de corrélation entre le CBR expérimental et prédit des sols expansifs stabilisés	17
Figure 1.13. Test de sensibilité (recherche de Van et Hai 2021)	
Figure 1.14. Le meilleur réseau pour prédire le CBR imbibé et non imbibé	20
Figure 1.15. Comparaison du CBR non-imbibé expérimental et prédit	21
Figure 1.16. Comparaison du CBR imbibé expérimental et prédit	
Figure 1.17. L'architecture des réseaux de neurones artificiels à propagation-avant avec sept entrées	24
Figure 1.18. Algorithme de GEP	25
Figure 1.19. Comparaison entre les valeurs estimées (par ANN ou GEP) et les valeurs mesurées de CBR	26
Figure 1.20. Analyse de sensibilité pour les résultats de CBR	
Chapitre 2 : Collection et préparation de base des données	
Figure 2.1. Coupe transversale dans une structure des chaussées	30
Figure 2.2. Répartition des zones de collecte d'échantillons selon les wilayas Algérie	31
Figure 2.3. Appareillage de l'essai d'analyse granulométrique	33
Figure 2.4. Appareillage de l'essai de sédimentométrie	34
Figure 2.5. Appareillage de l'essai de limites d'Atterberg	35
Figure 2.6. Appareillage de l'essai de Proctor	37
Figure 2.7. Appareillage de l'essai de CBR	38
Figure 2.8. Plan de classification de sols d'après 'limite de liquidité LL et l'indice de plasticité PI' selon USCS système	39
Figure 2.9. Plan de classification de sols d'après 'limite de liquidité LL et l'indice de plasticité PI' selon AASHTO système	41

Figure 2.10.	Les types de relation de corrélation linéaire	44
Figure 2.11.	Matrice de corrélation (Pearson)	45
Figure 2.12.	Histogramme avec une distribution probable des erreurs a) pour les 12 entrées; b) pour le sorte CBR.	46-47
Chapitre 3 : Architecture des modèles		
Figure 3.1.	La méthode de K-fold cross validation	51
Figure 3.2.	Étapes d'identification de la méthode d'apprentissage automatique	52
Figure 3.3.	Parcours de choix de méthode d'apprentissage automatique	53
Figure 3.4.	Régression linéaire	54
Figure 3.5.	Exemple de régression linéaire simple SLR	54
Figure 3.6.	Exemple de régression linéaire multiple MLR	55
Figure 3.7.	Régression non linéaire	56
Figure 3.8.	Simulation entre la structure biologique d'un neurone et ANN	56
Figure 3.9.	Bref historique sur les développements scientifiques de modèle ANN	57
Figure 3.10.	Exemple de trois entrées de neurone formel	57
Figure 3.11.	Exemple d'une représentation d'un MLP	58
Figure 3.12.	Forme explicative de rétro-propagation de l'erreur dans MLP	60
Figure 3.13.	Forme exemplaire et explicative de la manière d'exécution de la régression par un modèle de forêts aléatoires	62
Figure 3.14.	Exemple de changement entre le priori et le posteriori de Processus Gaussien en raison de change des paramètres entre deux phases	65
Figure 3.15.	Le changement entre le priori et le posteriori du noyau SE (RBF)	68
Figure 3.16.	Le changement entre le priori et le posteriori du noyau Matern	69
Figure 3.17.	Le changement entre le priori et le posteriori du noyau Exponentiel	70
Figure 3.18.	Le changement entre le priori et le posteriori du noyau Exp-Square-Sine (périodique)	71
Figure 3.19.	Le changement entre le priori et le posteriori du noyau Rational Quadratique (RQ)	72
Figure 3.20.	Le changement entre le priori et le posteriori du noyau Noyau Linéaire / Dot-Product	73
Figure 3.21.	L'effet de white kernel sur les noyaux de GPR entre le priori et le posteriori	76
Figure 3.22.	Représentation schématique de la régression à vecteurs de support SVR	77
Figure 3.23.	Représentation de noyau linéaire de modèle SVR	78
Figure 3.24.	Représentation de noyau polynomial de modèle SVR	83
Figure 3.25.	Représentation de noyau RBF de modèle SVR	86
Figure 3.26.	Représentation de noyau sigmoïde de modèle SVR	88
Figure 3.27.	Exemple sur la performance de modèle avec la méthode de sélection avant	94
Figure 3.28.	Organigramme explicatif pour les types de méthode de features selection en montrant les techniques adoptées dans notre étude	102
Figure 3.29.	Exemple d'utilisation de l'optimisation bayésienne BOA	103
Figure 3.30.	Exemple d'évolution temporelle de l'optimisation bayésienne BOA : GP vs TPE	
Figure 3.31.	Organigramme de la modélisation des modèles conventionnels, CML dans notre étude	
Figure 3.32.	Organigramme d'exécution des techniques de RFE et EFS	

Figure 3.33.	Organigramme d'exécution de technique d'optimisation TPE	104
Figure 3.34.	Organigramme de la méthodologie de cette étude	
Chapitre 4 : Résultats et optimisation		
Figure 4.1.	L'architecture de modèle MLPNN 12-8-1 d'ANN	109
Figure 4.2.	Résultats de la recherche du sous-ensemble optimal de caractéristiques a) Résultats de MSE ; b) Résultats de R^2	111
Figure 4.3.	Test des prédictions avec estimations d'incertitude pour les valeurs CBR	
Figure 4.4.	Le progrès de l'optimisation bayésienne des hyper-paramètres pour le modèle SVR	113
Figure 4.5.	Diagramme en boîte des valeurs mesurées et prédites du CBR pour le modèle SVR	114
Figure 4.6.	Le graphique de la ligne ajustée des valeurs prédites et mesurées du CBR pour le modèle hybride TPE-GPR-RFE	115
Figure 4.7.	Nuage de points des résidus par rapport aux valeurs prédites pour le modèle hybride TPE-GPR-RFE	116
Figure 4.8.	Comparaison des performances des modèles hybrides et conventionnels : valeurs CBR prédites vs valeurs CBR réelles	117
Figure 4.9.	Analyse des résidus des modèles : résidus de prédiction à travers les cinq modèles de Machine Learning (hybride et CML)	118
Figure 4.10.	Diagramme de Taylor des 5 modèles de ML utilisant l'ensemble de test	119
Figure 4.11.	Diagramme de Taylor des 5 modèles de ML utilisant l'ensemble de test	
Figure 4.12.	Nuage de points des résidus par rapport aux valeurs prédites pour le modèle hybride TPE-SVR-EFS	121
Figure 4.13.	Diagramme de Taylor compare les cinq modèles hybrides	123
Figure 4.14.	Résultats des explications globales SHAP utilisant le TPE-GPR-RFE	125
Figure 4.15.	Explications globales SHAP moyennes	
Figure 4.16.	Graphique de dépendance de MDD	126
Figure 4.17.	Graphique de dépendance de MDD avec interaction OMC	
Figure 4.18.	Graphique de dépendance de MDD avec interaction PI	
Figure 4.19.	Graphique d'interaction 3D de la valeur SHAP de la MDD avec G, OMC et MDD	127
Figure 4.20.	Graphique d'interaction 3D de la valeur SHAP de l'OMC avec G, OMC et MDD	128
Figure 4.21.	Graphique d'interaction 3D de la valeur SHAP de la MDD avec PI, OMC et MDD	
Figure 4.22.	Graphique d'interaction 3D de la valeur SHAP de l'OMC avec PI, OMC et MDD	129
Figure 4.23.	Graphique en cascade pour l'explication locale	130
Figure 4.24.	Application de l'analyse de sensibilité de Sobol sur modèle TPE-SVR-EFS	131
Figure 4.25.	Graphique d'interaction 3D des indices de Sobol (MDD avec sept autres variables d'entrées)	134

Liste des abréviations :

Symboles :	Description en français :	Description en anglais :
CC	Coup de Compactage = 10, 25 ou 55	Compaction pressure = 10, 25 or 55
G	Teneur en Gravier dans les échantillons	Gravel content in samples
S	Teneur en Sable dans les échantillons	Sand content in samples
FC	Teneur en Fins dans les échantillons	Fines content in samples
Cu	Coefficient d'Uniformité	Uniformity Coefficient
Cc	Coefficient de Courbure	Curvature Coefficient
LL	Limite de Liquidité	Liquid Limit
LP	Limite de Plasticité	Plastic Limit
PI	Indice de Plasticité	Plasticity Index
MC	Teneur en Eau dans chaque CC	Water Content in each CC
DD	Densité Sèche dans chaque CC	Dry Density in each CC
OMC	Teneur en Eau Optimale	Optimal Water Content
MDD	Densité Sèche Optimale	Optimal Dry Density
CBR	Coefficient de Portance Californie	Californian Bearing Ratio
AI	Intelligence Artificielle	Artificial Intelligence
ML	Apprentissage automatique	Machine Learning
CML	Modèles d'Apprentissage Automatique Conventionnels	Conventional Machine Learning Models
R²	Coefficient de Régression ou de Détermination	Coefficient of Regression or Determination
MSE	Erreur Quadratique Moyenne	Mean Square Error
RMSE	Racine Carrée de l'Erreur Quadratique Moyenne	Root Mean Square Error
MAE	Erreur Absolue Moyenne	Mean Absolute Error
MAPE(%)	Erreur Absolue Moyenne en Pourcentage	Mean Absolute Percentage Error
MLR	Régression Multi-Linéaire	Multiple Linear Regression
SLR	Régression Linéaire Simple	Simple Linear Regression
ANN	Réseaux de Neurones Artificiels	Artificial Neural Network
DNN	Modèle de Réseaux de Neurones Profonds	Deep or Dense Neural Network Model
MLPNN	Modèle de Réseaux de Neurones Multi-Percéptron	Multi-Layer Perceptron Neural Network Model
DT	Modèle d'Arbre de Décision	Decision Trees Model
RF	Modèle de Forêt Aléatoire	Random Forest Model
GPR	Modèle de Régression de Processus Gaussien	Gaussian Process Regression Model
SVR	Modèle de Régression par Vecteurs de Support	Support Vector Regression Model
H-XGB	Modèle Hybride de XGBoost	Hybrid XGBoost Model
H-LGB	Modèle Hybride de LightGBM	Hybrid LightGBM Model

H-Ada	Modèle Hybride d'Adaboost	Hybrid Adaboost Model
H-DNN	Modèle Hybride de Réseaux de Neurones Denses	Hybrid Dense Neural Network
BOA	Algorithme d'Optimisation Bayésienne	Bayesian Optimization Algorithm
TPE	Estimation d'Erreur de Parzen Structurée par Arbre	Tree-Structured Parzen Estimator Error
GP	Processus Gaussien	Gaussian Process
XAI	Méthodes d'AI explicable	eXplainable AI methods
SHAP	Explications Additives de Shapley	SHapley Additive exPlanations
Sobol	Analyse de sensibilité de Sobol	Sensitivity analysis of Sobol
S_i	Indices de Sobol de premier ordre	First order Sobol indices
S_T	Indices de Sobol Totaux	Total Sobol indices
S_{ij}	Indices de Sobol d'interaction	Interaction Sobol indices

Introduction générale

Introduction générale :

Les routes sont parmi les infrastructures les plus importantes dans le monde et constituent un indicateur de la capacité économique d'un pays. Elles sont l'épine dorsale des réseaux de transport de marchandises, du commerce et des déplacements de la population en général.

Les routes sont une structure civile dont la surface est constituée de la couche de roulement en bitume. Leur structure sous-jacente est constituée de plusieurs couches superposées pour le revêtement, telles que la couche de fondation, la couche de base, etc. Cette étude met en lumière la couche de sol de support ayant une granulométrie fine (limon, argile).

La rigidité des couches reflète le bon comportement de la structure routière, sa capacité de charge et sa durabilité. Le coefficient de portance CBR est un bon indicateur de cette rigidité. Dans la pratique habituelle, la valeur du CBR est déterminée par des essais en laboratoire. Des échantillons de sol sont prélevés sur le terrain, nettoyés et immergés (imbibée) dans l'eau pendant quatre jours pour reproduire les conditions environnementales sur le terrain. Ils sont ensuite placés dans une machine d'essai Proctor pour obtenir les valeurs du CBR à chaque énergie de compactage. Cet essai nécessite des déplacements, des efforts, des travaux en laboratoire, des compétences techniques, du temps et même des dépenses. Par conséquent, l'intégration de la technologie, telle que la prédiction des valeurs du CBR par l'utilisation de l'intelligence artificielle, pourrait être une solution appropriée car elle économise les efforts, le temps et l'argent.

L'intelligence artificielle est actuellement l'un des axes développements les plus importants de la science moderne. Elle est largement utilisée dans plusieurs domaines tels que la numérisation, l'apprentissage automatique, la prédiction, la classification, la sélection génétique, la reconnaissance des empreintes de mains et des yeux, etc. D'autre part, l'apprentissage automatique est l'une des plus grandes branches de l'intelligence artificielle, il englobe de nombreux algorithmes tels que les réseaux neuronaux artificiels (ANN), la régression à vecteurs de support (SVR), les arbres de décision (DT) et bien d'autres.

L'efficacité de la recherche scientifique menée par de nombreux chercheurs à travers le monde avec l'intégration de l'intelligence artificielle dans le domaine du génie civil (en particulier la géotechnique) dans le but de prédire les valeurs du CBR, nous a incités à nous lancer dans l'expérimentation sur ce sujet.

Avant de commencer la modélisation, il est nécessaire d'avoir une base de données adaptée à la nature de la recherche. À cette fin, nous avons collecté une base de données comprenant des échantillons de sol à granulométrie fine (limon, argile) imbibées pendant quatre jours pour la simulation de l'environnement naturelle, ces échantillons sont déduire de la couche de sol-support provenant de plusieurs régions du nord de l'Algérie, comprend 12 entrées de paramètres relatives à la granulométrie, la plasticité et aux paramètres de compactage du sol, et une seule sortie qui est la valeur CBR.

Avant de passer à l'étape de modélisation, une analyse statistique de la base de données a été effectuée, ce qui nous a permis d'éliminer les échantillons à valeurs aberrantes ce qui a réduit le nombre d'échantillons dans la base de données (de 1680 à 1123).

Les objectifs des modèles d'apprentissage automatique sont variés : certains sont classificatoires ou régressifs (comme la prédiction), quantitatifs ou qualitatifs, ou encore supervisés ou non supervisés. Dans notre étude, l'objectif est de prédire la valeur du CBR, ce qui nous place dans le

cadre de la régression quantitatif et surveillé. Après plusieurs tests, nous avons choisi cinq modèles parmi tous les modèles disponibles d'apprentissage automatique qui sont : régression multilinéaire (MLR), les réseaux de neurones artificiels (ANN) de type multicouches perceptron (MLPNN), les forêts aléatoires (RF), régression par processus Gaussien (GPR) et la régression à vecteur de support (SVR).

L'évaluation de la performance des modèles se fait en suivant les valeurs du coefficient de régression (détermination) R^2 pour les étapes d'entraînement et de test, qui doivent s'inscrire dans l'intervalle $[0.8, 1]$. Nous suivons également les valeurs des coefficients d'erreurs (MSE, RMSE, MAE et MAPE) qui doivent être faibles les plus faibles possible.

Eviter le cas de sur-ajustement (over-fitting) est le souci de beaucoup de chercheurs car le modèle devient alors mémorisant les mêmes valeurs d'entrée sans être capable de produire de nouveaux résultats pour de nouvelles données. Pour éviter cela, nous avons utilisé la technique K-fold cross-validation qui divise la base de données en k ensembles pour l'entraînement et le test de manière cyclique. Dans notre étude, la division qui a donné les meilleurs résultats est celle à 10 plis.

D'autre part l'utilisation de techniques avancées de sélection d'entrées, en choisissant les techniques de basée sur le Wrapper, de type de l'élimination Récursive (RFE) avec le modèle de GPR et de type de la sélection Exhaustive (EFS) avec le modèle de SVR, a permis de réduire la taille de données d'entrée de 12 à seulement 7, cette réduction des données d'entrée n'a pas seulement réduit le temps de calcul mais a aussi permis l'amélioration de la performance de ces deux modèles. Finalement, les modèles hybrides développés dans ce cas sont: GPR-RFE et SVR-EFS.

L'amélioration des modèles vise à passer d'un modèle ordinaire à un modèle ayant une meilleure performance. Dans notre travail, nous avons hybridé les deux modèles GPR et SVR avec l'Algorithme d'Optimisation Bayésienne (BOA), en utilisant le type de l'estimateur de Parzen structuré arborescent (TPE) qui sélectionne le noyau optimal du modèle en réglant ses hyper-paramètres aux valeurs optimales. Nous avons ainsi obtenu deux modèles optimisés sont: TPE-GPR-RFE et TPE-SVR-EFS.

Les deux modèles GPR et SVR ont répondu de manière significative à l'amélioration (l'optimisation) et à l'apprentissage dans les deux étapes (entraînement et test) et ont donné des résultats très satisfaisants caractérisés par des valeurs d'erreurs très minimales et des valeurs du coefficient de détermination R^2 dépassant le seuil de 80%.

En comparant les performances des cinq modèles développés, les résultats ont montré la supériorité du modèle TPE-GPR-RFE hybride par rapport aux trois modèles conventionnels (modélisés par défaut) de manière claire, ainsi qu'une meilleure performance par rapport au modèle hybride TPE-SVR-EFS et mêmes des quatre autres modèles modernes et hybridés (H-DNN, H-XGB, H-LGB et H-Ada) avec des valeurs dans les paires (Entraînement/Test) suivants :

$(R^2 = 0.9982/0.9935, \text{ MSE} = 0.0768/0.2974, \text{ RMSE} = 0.2771/0.5453, \text{ MAE} = 0.2107/0.4364 \text{ et } \text{MAPE}(\%) = 5.32/13.79).$

Enfin, pour déterminer la sensibilité de la sortie de calcul (CBR) aux variations des différentes entrées, en appliquant la méthode des valeurs de SHAP sur le modèle de TPE-GPR-RFE et le test de sensibilité de Sobol sur le modèle de TPE-SVR-EFS, et les résultats de simulation ont montré que la densité sèche optimale (MDD) des échantillons est l'entrée ayant le plus d'impact sur la sortie CBR.

Ce manuscrit de thèse comporte quatre chapitres :

- Dans le premier chapitre de cette thèse, nous présentons un aperçu général des recherches antérieures relatif à l'utilisation des modèles d'intelligence artificielle pour la prédiction de la valeur CBR.
- Le deuxième chapitre, se penche sur la collecte et la préparation d'une base de données locale utilisée pour la prédiction des valeurs CBR de sols-supports des chaussées.
- Le troisième chapitre traite des aspects liés à l'architecture des modèles de prédiction utilisés dans le présent travail de thèse.
- Dans le quatrième chapitre nous présentons les principaux résultats des simulations réalisées, en utilisant plusieurs modèles prédictifs qui seront comparés au modèle de TPE-GPR-RFE.
- Enfin, cette thèse de doctorat sera clôturée par des conclusions générales et des recommandations utiles sur le plan pratique, qui retraceront les grandes lignes du travail effectuée et souligneront les différents résultats obtenus.

Chapitre 1

***A**analyse des recherches antérieures*

I. Introduction : L'intelligence artificielle est un domaine de plus en plus présent dans notre vie quotidienne, et elle est également en train de révolutionner le monde du génie civil.

L'intelligence artificielle (IA) est un domaine de l'informatique qui vise à développer des systèmes capables de simuler l'intelligence humaine pour résoudre des problèmes complexes. Le machine learning ML et réseaux de neurones sont des sous-catégories de l'IA qui utilise des algorithmes pour apprendre à partir de données et améliorer ses performances au fil du temps.

En ce qui concerne la prédiction, l'IA et le machine learning peuvent être utilisés pour prédire les performances des structures et des matériaux, pour optimiser la conception des bâtiments et des ponts, et pour prédire les risques de défaillance ou de dommages...ect.

L'intelligence artificielle et le machine learning peuvent également être utilisés dans le domaine de la géotechnique pour prédire les paramètres de sols tels que la résistance à la compression, la perméabilité, la densité,...etc.

Dans ce chapitre, nous présenterons l'une des applications de ce travail et ses succès en présentant une série de résultats de recherches précédentes concernant la prédiction des valeurs CBR.

II. Intelligence artificielle et la prédiction de l'indice de portance Californien (CBR) : Le CBR (California Bearing Ratio) est un indicateur de la capacité portante des sols utilisé dans la conception des routes et des pistes d'atterrissage.

La prédiction précise du CBR est essentielle pour garantir la sécurité et la durabilité de ces infrastructures.

L'IA et le machine Learning peuvent être utilisées pour prédire les valeurs de CBR à partir de données géotechniques telles que la densité, l'humidité, la granulométrie et la teneur en argile des sols...etc.

Les algorithmes d'apprentissage automatique peuvent analyser ces données pour détecter des modèles et des relations complexes entre les différentes variables, ce qui permet de prédire avec précision les valeurs de CBR pour différents types de sols.

III. Les recherches précédentes : Les recherches précédentes ont montré ces dernières années des avancées remarquables dans le domaine de l'ingénierie civile et géotechnique spécifiquement, en particulier en ce qui concerne l'utilisation de l'intelligence artificielle pour prédire les valeurs du coefficient CBR, ce qui est très intéressant et prometteur et les résultats obtenus sont très encourageants.

Le tableau 1.1 présente les résultats des principales études de prédiction du CBR lors de la dernière décennie.

Tableau 1.1. Résumé de recherches précédentes.

No	Type de sol	Entrées	Sortes	Modèles	R ² ou R	MSE ou RMSE	Année	Référence
1	fin	AASHTO M 145 – MDD - LL - PL - PI - MDD - OMC - gonflement	CBR	RF - DT - LR - ANN	R ² 0.8400	MSE 66.2400	2023	(Semachew et Betelhem 2023)
2	Argile	γ_{dry} - W% - TB% - L%	CBR	ANN - MLR	R ² 0.9861	RMSE 3.3200	2022	(Yusuf Erzin et al 2022)
3	fin	Ash type - Ash% - LL% - LP% - OMC% - MDD	CBR	Light GBM	R ² 0.9473	RMSE 0.3303	2021	(Van and Hai 2021)
4	fin	G - Cu - Cc - LL - LP - PI - OMC -MDD	CBR	ANN - GEP - méthode de Kriggen	R 0.9980	MSE /	2020	(SK Kamrul et al 2020)
5	fin	LL - PI - γ_d - Wopt - C+M - S - G	CBR	ANN - GEP	R ² 0.9100	MSE 2.2070	2010	(Taskiran 2010)

III.1. Les travaux de Semachew et Betelhem 2023 : Les chercheurs ont présenté la définition du CBR et son rôle dans les études géotechniques, ainsi que la méthode pour déterminer sa valeur. Ils ont souligné que le CBR est un paramètre important en génie civil, en particulier en génie des matériaux et en géotechnique, car il mesure la résistance des matériaux de base dans les structures routières, les pistes aériennes et les chemins de fer. Les tests CBR sur le terrain impliquent la poussée du piston dans le sol et les matériaux de base sur le site de test à l'aide d'un vérin de charge pour déterminer la force du sol sur site et des matériaux de base pour la conception des routes. Les équipements de test CBR sur le terrain sont coûteux et peu pratiques à transporter dans différentes régions. Les valeurs CBR des sols et des matériaux de base sont généralement déterminées en laboratoire. Le CBR est mesuré en laboratoire en plaçant un piston de diamètre standard dans un échantillon de sol compacté préparé à l'humidité optimale à un taux de 1,3 mm / minute. Les valeurs CBR peuvent être calculées pour tout échantillon de sol humide ou sec. Les valeurs CBR des échantillons de sol immergés sont généralement inférieures à celles des échantillons non immergés, par conséquent, les valeurs CBR des échantillons immergés sont généralement utilisées pour estimer la qualité des matériaux faibles. Étant donné que la détermination du CBR prend beaucoup de temps, le processus a un impact important sur le retard de la construction. Cependant, les échantillons de sol préparés à l'humidité optimale doivent être conservés dans des conditions saturées d'eau pendant 4 jours, ce qui est considéré comme la pire situation si on s'attend à des précipitations pendant 4 jours.

Le calcul du CBR pour de nombreux échantillons est généralement nécessaire, ce qui nécessite du temps et des coûts, Le test CBR nécessite des techniciens de laboratoire formés, ce qui retarde la fin du projet. Une solution à ce problème est de connaître ou de prédire la valeur CBR pour éviter de perdre du temps en utilisant des techniques d'apprentissage automatique.

Les chercheurs ont estimé que la régression est une méthode pour déterminer comment les caractéristiques ou variables indépendantes sont liées à la caractéristique ou résultat dépendant. Au cours des dernières années, des exemples réels de géotechnique ont été intégrés à des techniques d'apprentissage automatique telles que le réseau neuronal artificiel (ANN), le réseau de neurones à couches multiples (MLP), la programmation génétique (GEP), la machine à vecteurs de support (SVM) et la méthode de traitement de données multiples pour prédire les données souhaitées. Une liste des travaux précédemment réalisés a été incluse.

III.1.1. Application d'étude sur des sols fins :

a) La région d'étude : La zone d'étude se situe dans le projet de route de la région d'Amhara en Éthiopie, le long d'une route de 48,92 km entre les villes de Mekane Eyesus et Semada. Des échantillons de sol ont été prélevés le long de la route pour effectuer des tests en laboratoire.

b) La base des données : La base de données examinée par les chercheurs peut être divisée en deux parties principales :

Les entrées, qui comprennent sept éléments suivant :

1) La classification des sols : Les échantillons de sol ont été classés en fonction du système américain AASHTO, donnant les groupes suivants : **A-7-5, A-2-4, A-2-7, A-2-7, A-2-6, A-7-6.**

2) Limite de liquidité LL : Ces valeurs ont été déterminées par les essais de limites d'Atterberg, où la valeur maximale a atteint 83.1% et la valeur minimale de 37%, avec une moyenne de 61.1%.

3) Limite de plasticité LP : Ces valeurs ont été déterminées aussi par les essais de limites d'Atterberg, où la valeur maximale a atteint 46.29% et la valeur minimale de 21.78%, et avec une moyenne de 31.55%.

4) **L'indice de plasticité IP** : Qu'est égal à **LL-LP** et il est varié entre valeur maximale de 45.67% et autre minimale de 9.63% avec une moyenne de 29.54%.

5) **La densité sec optimale MDD** : Ces résultats sont obtenus par l'essai de Proctor, qui détermine la valeur optimale de la densité sec du sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur varie entre une valeur maximale de 1.80 et une valeur minimale de 1.23 avec moyenne de 1.60.

6) **Le teneur en eau optimal OMC** : Ces résultats sont obtenus aussi par l'essai de Proctor, qui détermine la valeur optimale de teneur en eau dans le sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur est donnée en % et varie entre une valeur maximale de 39.50% et une valeur minimale de 2.60% et avec moyenne de 22.03%.

7) **Gonflement**: Sol gonflant contenant des minéraux d'argile de montmorillonite ou d'autres, donné en pourcentages et il est déterminé d'après les essais de sédiment-métrique et l'analyse chimique d'échantillons ses valeurs sont variées entre 35 et 36%.

Les sortes, on a un seul sort **CBR** : Ou bien CBR trempé pour simuler les pires conditions: le matériau de la couche de fond gagne de l'humidité. Cette valeur est déterminée en laboratoire en se basant sur l'expérience CBR mentionnée ci-dessus, voir [les travaux de Semachew et Betelhem 2023], et ses valeurs sont variées entre 124 et 0.24 avec une moyenne de 15.17.

III.1.2. L'étude statistique de base des données : Les deux chercheurs ont présenté un tableau montrant les résultats statistiques des variables d'entrée-sortie dans la base de données qui ont été discutés lors de la présentation de la base de données étudiée. Ils ont également présenté une étude d'image montrant les corrélations entre les paramètres du sol, comme indiqué dans la figure 1.

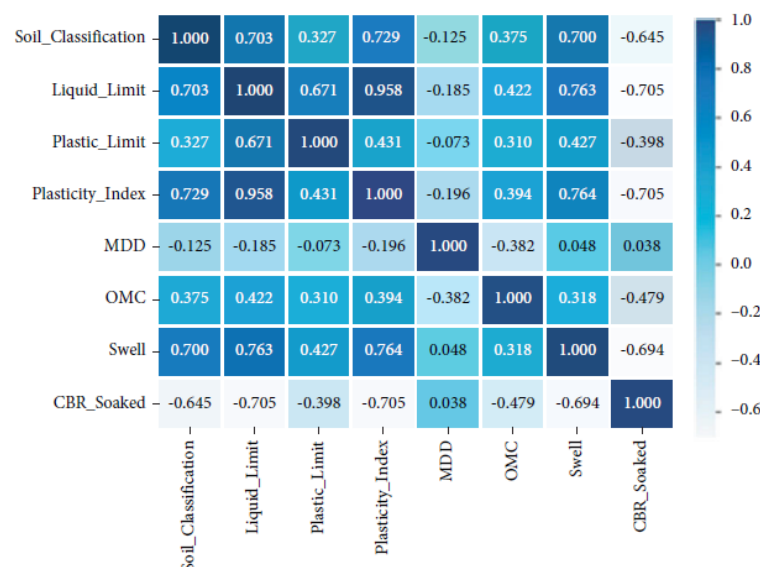


Figure 1.1. La matrice de corrélation (recherche de Semachew et Betelhem 2023).

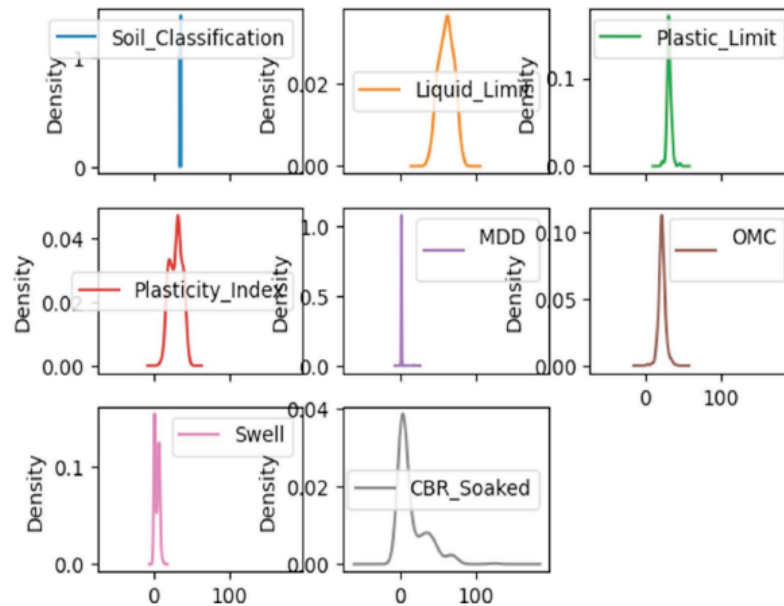


Figure 1.2. Distribution des attributs à l'aide d'un diagramme de densité.

III.1.3. Développement de modèles de forêts aléatoires RF, les arbres de décision DT et multi-linéaire perceptron MLP pour prédire les valeurs de CBR : Dans cette étude expérimentale, le langage de programmation Python a été utilisé pour développer un modèle de prédiction. 80% de l'ensemble de données ont été utilisés pour l'entraînement, tandis que le reste de 20% a été utilisé pour la validation. En fonction de la linéarité des données, de la taille de l'échantillon et du nombre de paramètres, des algorithmes d'apprentissage supervisé tels que les forêts aléatoires RF, les arbres de décision DT et la régression linéaire RL ont été choisis, ainsi que des techniques d'apprentissage en profondeur telles que les réseaux neuronaux artificiels (ANN) pour prédire le coefficient CBR. Cette technique utilise des échantillons de sol collectés dans la zone d'étude. En comparant la précision de ces techniques, celles ayant une bonne précision de prédiction sont sélectionnées. Enfin, la moyenne quadratique de l'erreur MSE, la racine carrée de la moyenne quadratique de l'erreur RMSE, l'erreur absolue moyenne MAE et le coefficient de détermination R^2 sont utilisés pour évaluer les performances de la méthode.

III.1.3.1. Le modèle d'arbre aléatoire RF : Les chercheurs estiment que 'Cook et Carter' ont été les premiers à présenter un ensemble d'arbres de décision, et 'Breiman' a été l'un des premiers à démontrer, à la fois mathématiquement et pratiquement, que l'agrégation de différentes répétitions du prédicteur dans un ensemble pouvait conduire à des gains significatifs de fiabilité.

Les chercheurs ont mentionné l'une des méthodes pour améliorer les performances des modèles, qui est la technique de l'agrégation (Bagging). Cette technique est utile dans une variété de cas, car elle a l'avantage de pouvoir être utilisée pour améliorer n'importe quel type de modèle, et pas seulement les arbres de décision. Dans l'étude des forêts aléatoires (RF) réalisée par Breiman, chaque nœud inclut la technique d'agrégation en plus de la sélection de variables aléatoires. En unifiant ces deux méthodes et en ajustant l'aléatoire, on obtient l'un des algorithmes d'apprentissage automatique les plus efficaces et performants pour presque toutes les tâches.

III.1.3.2. Le modèle Arbre de décision DT : L'arbre de décision est un modèle hiérarchique d'apprentissage supervisé, où la région locale est déterminée par une série de divisions répétées en moins d'étapes. L'arbre de décision se compose de nœuds de décision internes et de feuilles finales.

Cette technique est utilisée pour la classification et les transformations. La construction de l'arbre de transformation est similaire à la construction de l'arbre de classification, à l'exception que la mesure d'impureté utilisée dans la classification est remplacée par une mesure utilisée dans la transformation.

III.1.3.3. Le modèle de régression linéaire MLP : Est une méthode permettant de déterminer la relation entre les prédictors (variables indépendantes) et les variables de réponse (variables dépendantes). L'analyse de régression se présente sous trois formes différentes. La première est l'analyse de régression linéaire simple, utilisée pour modéliser la relation linéaire entre deux variables, à savoir la variable indépendante x et la variable dépendante y . Le deuxième type d'analyse est appelé analyse de régression linéaire multiple, qui est un modèle de régression linéaire avec une variable dépendante et plusieurs variables indépendantes. L'analyse de régression linéaire multiple suppose que la variable dépendante est une fonction linéaire des coefficients du modèle et qu'il y a plusieurs variables indépendantes. Les coefficients de régression pour le troisième type d'analyse supposent que la relation entre la variable dépendante et la variable indépendante n'est pas linéaire.

III.1.4. Evaluation des résultats d'étude de Semachew et Betelhem 2023 : La performance du modèle de forêt aléatoire a été évaluée en utilisant trois indicateurs d'évaluation standard, y compris le coefficient de détermination (R^2), l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (MSE) et la racine carrée de l'erreur quadratique moyenne (RMSE).

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y}_i)^2} \quad \dots (1.1)$$

(Mauvaise valeur $-\infty$, meilleure valeur $+1$)

$$MSE = \frac{1}{n} \sum_{i=1}^n |Y - Y'|^2 \quad \dots (1.3)$$

(Mauvaise valeur $+\infty$, meilleure valeur 0)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |Y - Y'|^2} \quad \dots (1.2)$$

(Mauvaise valeur $+\infty$, meilleure valeur 0)

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i| \quad \dots (1.4)$$

(Mauvaise valeur $+\infty$, meilleure valeur 0)

Les résultats qui obtenus sont résumées dans le tableau et les figures suivants et elles sont représenté les valeurs d'évaluation statistiques.

Tableau 1.2. Les résultats d'étude (Semachew et Betelhem 2023).

Algorithm	MAE	RMSE	MSE	R^2
RF	4.80	8.13	66.24	0.86
DT	5.30	11.92	143.30	0.66
LR	8.90	11.08	122.80	0.53
ANN	5.69	10.36	107.46	0.67

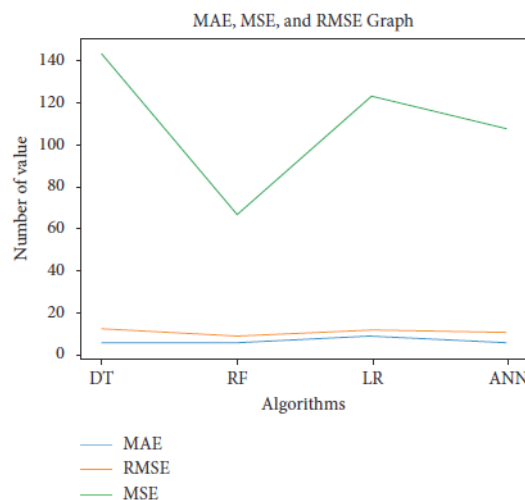


Figure 1.3. Comparaison de MAE, MSE et RMSE pour les algorithmes.

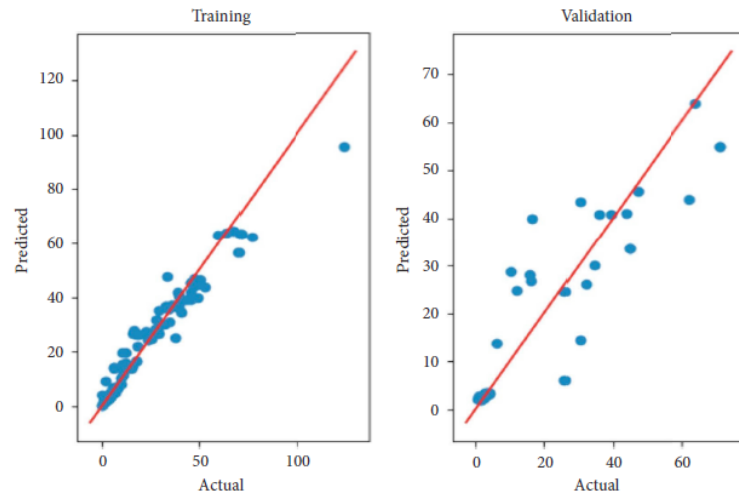
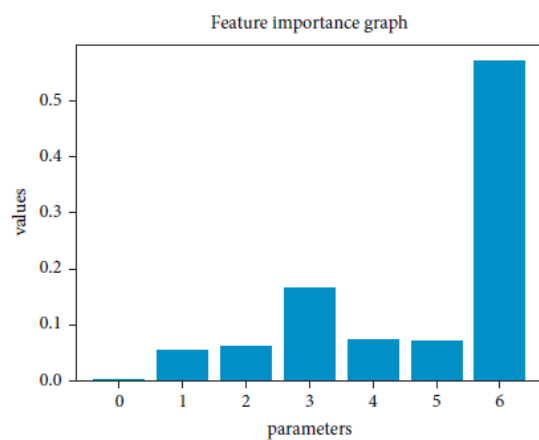


Figure 1.4. Les valeurs réelles et prédites des ensembles d'entraînement et de test à l'aide de RF model.



0 classification de sol, 1 limite de liquidité, 2 limite de plasticité, 3 indice de plasticité, 4 densité sèche optimale, 5 teneur en eau optimal, et 6 pourcentage de gonflement.

Figure 1.5. Test de sensibilité (recherche de Semachew et Betelhem 2023).

Dans cette étude, les chercheurs ont utilisé les algorithmes des forêts aléatoires, des arbres de décision, de la régression linéaire et du réseau de neurones artificiels pour entraîner les données à 80% et les tester à 20%. De cela, le modèle de forêts aléatoire est prévu pour avoir les meilleures prédictions, avec une plus faible erreur lorsqu'il est comparé aux autres en termes de MAE, MSE et RMSE. De plus, le modèle de forêts aléatoire obtient une bonne valeur du coefficient de détermination par rapport aux autres. En comparant, les chercheurs ont trouvé que le modèle RF est le plus optimal parmi les autres modèles étudiés, car il a donné la valeur R^2 la plus élevée.

De plus, il a donné des valeurs plus petites pour MAE, RMSE et MSE. Le test de sensibilité qu'ils ont effectué a montré que le gonflement a le plus grand impact sur la prédiction de CBR.

Conclusion : Dans cette étude, le CBR du sol a été prédit en utilisant les algorithmes des forêts aléatoires, des arbres de décision, de la régression linéaire et du réseau de neurones artificiels, qui ont été formés et construits sur un ensemble d'entraînement de 80% et testés sur un ensemble de test de 20%. Les variables des modèles incluent SC, LL, PL, PI, OMC, MDD, gonflement et CBR. L'auteur a essayé d'utiliser ces algorithmes et d'étudier pour prédire les valeurs du CBR dans un sol spécifique.

Les résultats de l'étude évaluent les méthodes sur cet ensemble de données et identifient l'algorithme performant qui prédit bien dans cet ensemble de données. Par conséquent, l'algorithme RF a obtenu une valeur de coefficient de détermination (R^2) de 0,84 et des valeurs plus petites pour MAE, MSE, et RMSE. [1]

III.2. Les travaux de Yusuf Erzin et autres 2022 : Chaque année, de grandes quantités de pneus usagés sont stockées, ce qui représente des risques environnementaux et sanitaires importants. Les pneus usagés peuvent être utilisés comme matériau de remblai léger dans des projets routiers et de terrassement tels que des barrières et des ponts, ou comme matériaux de drainage dans les décharges.

De nombreux chercheurs ont étudié le comportement des pneus en caoutchouc et des mélanges de sable en utilisant des tests triaxiaux et de cisaillement direct. Il a été déterminé que la résistance au cisaillement augmente avec l'augmentation du pourcentage de pneus, atteignant une valeur maximale de 35% avant de diminuer. De nombreux chercheurs ont également étudié l'effet de l'ajout de matériaux de formation et de matériaux usagés sur le sol. Les réseaux de neurones artificiels peuvent être utilisés pour identifier des modèles et détecter des tendances qui sont très complexes.

Les valeurs CBR ont été déterminées en utilisant un modèle ANN et un modèle MLR (Régression Linéaire Multiple). Le poids volumique sec γ_d , la teneur en eau W, la teneur en chaux L et la teneur en pneus en caoutchouc TB ont été utilisés comme variables d'entrée dans la création des modèles ANN et MLR, tandis que la valeur CBR observée a été utilisée comme paramètre de sortie. Le modèle ANN a montré une capacité supérieure à prédire les valeurs observées par rapport au modèle MLR.

III.2.1. Application d'étude sur les sols argileux traités à la chaux avec déchet de pneus :

a) La région d'étude et lieu de collecte de matériau : Une terre argileuse provenant du campus de l'Université de Gaziantep a été utilisée dans les études expérimentales. Les remblais de pneus ont été achetés dans une usine à Gaziantep, en Turquie. Il s'est avéré que les remblais de pneus avaient une gravité spécifique de 1,08, La chaux vive (CaO) a été achetée auprès d'une entreprise à Gaziantep. L'activité de la chaux, qui dépend de la quantité de CaO présente, a été déterminée à 88%.

b) La base des données : La base de données examinée par les chercheurs peut être divisée en deux parties principales :

Les entrées, qui comprennent quatre éléments comme suivant :

1) La densité sec optimale γ_d : Ces résultats sont obtenus par l'essai de Proctor, qui détermine la valeur optimale de la densité sec du sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur varie entre une valeur maximale de 17.70 et une valeur minimale de 11.30 avec moyenne de 14.67.

2) Le teneur en eau optimal W : Ces résultats sont obtenus aussi par l'essai de Proctor, qui détermine la valeur optimale de teneur en eau dans le sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur est donnée en % et varie entre une valeur maximale de 27.50% et une valeur minimale de 13.00% et avec moyenne de 20.12%.

3) Le teneur en chaux L : Il représente le pourcentage de chaux ajoutée de 88% et le bien déterminé, où la valeur ajoutée maximale est de 6% et la valeur minimale est de 0%, tandis que la moyenne de cette valeur est égale à 3%.

4) Le teneur en pneus en caoutchouc TB : Il représente le pourcentage de pneus de voiture jetés, où la valeur ajoutée maximale est de 15% et la valeur minimale est de 0%, tandis que la moyenne de cette valeur est égale à 7,5%.

Les sorties, on a un seul sort **CBR** : Les expériences de cette étude ont été menées en utilisant l'équipement CBR d'une capacité de 28 kilo newtons (ASTMD1883). Quatre concentrations d'eau

différentes et des quantités variables de déchets de pneus et de chaux ont été utilisées pour créer des échantillons de sol argileux. Des ratios différents de déchets de pneus et de chaux ont été choisis, à savoir 0, 5, 10 et 15 pour cent et 0, 2, 4 et 6 pour cent respectivement, en fonction du poids sec de l'échantillon. La quantité d'eau requise a été ajoutée après le séchage du sol argileux et son mélange avec de la chaux ou des déchets de pneus. Les mélanges ont été stockés dans un endroit fermé en utilisant un sac en plastique pendant seulement 24 heures pour sécher les échantillons d'argile et les déchets de pneus, et seulement une heure pour mélanger le mélange de sol argileux, de chaux et / ou de déchets de pneus. Selon ASTM D1883, le mélange a été écrasé dans un moule CBR en cinq couches. Un poids supplémentaire de 4,5 kilogrammes a été appliqué sur l'échantillon à l'intérieur du moule. Le taux de gonflement a été mesuré pendant la période d'immersion en utilisant un indicateur de pompage et un triptyque métallique. Après les mesures initiales de gonflement, les échantillons ont été immergés dans un bassin d'eau pendant quatre jours. Les échantillons contenant de la chaux et / ou des déchets de pneus ont été conservés dans un sac en plastique pendant trois jours pour sécher à température ambiante après compression. Ensuite, les échantillons ont été immergés pendant quatre jours. Après une période d'immersion de quatre jours, le taux de gonflement a été mesuré à nouveau.

Les échantillons ont été testés et les valeurs CBR ont été calculées pour eux à une profondeur de pénétration de 2,54 mm et 5,08 mm en utilisant les données collectées toutes les 0,25 mm (Cabalar et al. 2014), et ses valeurs sont variées entre 97.80 et 0.50 avec une moyenne de 21.74.

III.2.2. Choix de modèles ANN et MLR, en évaluant par la variance VAF pour prédire le CBR :

Les auteurs de cette étude ont utilisé le modèle de **réseau de neurones ANN** et le modèle de **régression multilinéaire MLR** pour prédire le CBR. Ils ont expliqué que cinq neurones ont été utilisés dans la couche cachée du modèle d'ANN.

La modélisation des deux modèles a été effectuée à l'aide du logiciel MATLAB et les échantillons ont été divisés en 80% pour l'entraînement et 20% pour la validation.

L'efficacité de la performance des modèles a été évaluée en utilisant la valeur de régression R^2 et la valeur de l'erreur absolue moyenne MAE, de l'erreur quadratique moyenne RMSE et la variance expliquée VAF.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad \dots (1.5) \quad MAE = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i| \quad \dots (1.6)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |Y - Y'|^2} \quad \dots (1.7) \quad VAF = \left[1 - \frac{\text{var}(Y - \hat{Y}_i)}{\text{var}(Y)} \right] * 100 \quad \dots (1.8)$$

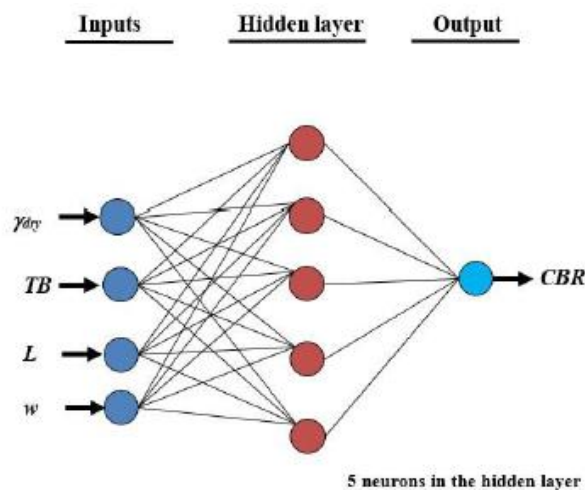


Figure 1.6. L'architecture de modèle d'ANN.

III.2.3. Evaluation des résultats d'étude de Yusuf Erzin et autres 2022 : Les résultats qui obtenus sont résumées dans le tableau 1.3 et les figures (1.7 et 1.8) suivants et elles sont représenté les valeurs d'évaluation statistiques.

Tableau 1.3. Les résultats d'étude (Yusuf Erzin et autres 2022).

Modèle	Data	R ²	MAE	RMSE	VAF(%)
ANN	Entrainement	0.9721	2.41	3.32	96.93
	Test	0.9861	2.32	2.84	98.28
MLR	Tous	0.4215	10.77	14.59	42.15

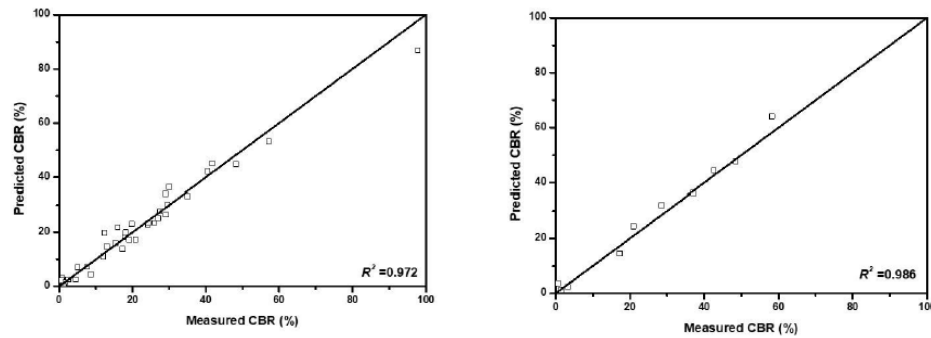


Figure 1.7. Comparaison entre les valeurs de CBR obtenus d'après les essais expérimentaux et les valeurs prédictives dans les deux étapes d'entraînement et de test pour le modèle d'ANN.

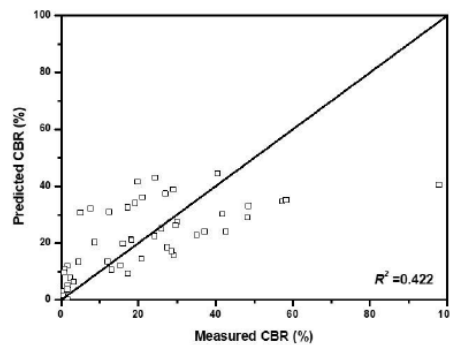


Figure 1.8. Comparaison entre les valeurs de CBR obtenus d'après les essais expérimentaux et les valeurs prédictives dans toutes les étapes pour le modèle de MLR.

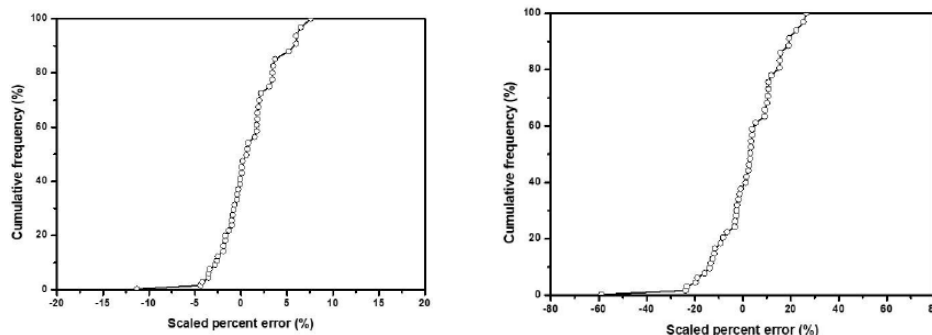


Figure 1.9. La relation entre l'erreur de pourcentage mise à l'échelle et la fréquence cumulative pour le modèle d'ANN à gauche et de MLR à droite consécutivement.

Les chercheurs ont clarifié que le modèle avec une valeur de R² proche de l'unité, une valeur de RMSE proche de zéro et une valeur de VAF proche de 100% est le meilleur modèle pour prédire les valeurs de CBR pour les sols fins stabilisés par des pneus usagés et traités à la chaux.

En comparant les résultats, il a été démontré que le modèle ANN est le meilleur, tandis que le modèle MLR a donné des résultats relativement faibles avec un R^2 de 0,42.

Conclusion : Cette étude a examiné l'application de modèles de réseau de neurones artificiels (ANN) et de régression multiple (MLR) pour prédire la valeur CBR des sols fins stabilisés par des pneus usagés et de la chaux. À cette fin, une série d'expériences CBR sur un sol argileux mélangé à des pneus usagés et de la chaux a été réalisée. Les résultats des tests ont ensuite été utilisés pour développer les modèles ANN et MLR. Le poids unitaire sec (γ_{sec}), la teneur en eau (w), la teneur en chaux (L) et la teneur en pneus usagés (TB) ont été utilisés comme paramètres d'entrée pour les modèles ANN et MLR, avec la valeur CBR mesurée agissant comme paramètre de sortie. Lorsque les sorties des modèles ANN et MLR ont été comparées aux résultats expérimentaux, il a été observé que le premier modèle produisait des prévisions plus fiables que le second. Pour évaluer les capacités de prédiction des modèles ANN et MLR proposés, de nombreux indices de performance ont été construits, notamment R^2 , VAF, MAE et RMSE. [2]

III.3. Les travaux de Van et Hai Quan 2021 : Le coût des chercheurs est d'utiliser des matériaux cimentaires dérivés de déchets agricoles tels que la cendre de bagasse et la cendre de coque d'arachide, etc. pour aider à stabiliser les sols expansifs qui sont sujets au gonflement et à la contraction, ainsi que pour protéger l'environnement et favoriser le développement durable. Les chercheurs ont également mentionné le coefficient de dureté CBR des sols porteurs de routes et d'aéroports en tant qu'indicateur de laboratoire. Les chercheurs ont mentionné le protocole pour déterminer ce coefficient, comme mentionné dans les recherches précédentes [comme mentionné a les travaux de Semachew et Betelhem 2023]. Les chercheurs ont salué le rôle de l'intelligence artificielle et de l'apprentissage automatique dans la prédiction de la valeur de ce coefficient, en particulier ce qui a été présenté par le chercheur Taskiran en 2010. La performance des modèles développés est évaluée en suivant la valeur de la corrélation R et les valeurs MAE et RMSE.

III.3.1. Mélange des cendres volantes avec les sols cohérents pour prédire le CBR :

a) La région d'étude : De cette étude, une base de données a été construite en collectant des données de Rajakumar et Reddy, 207 échantillons de sols cohérents ont été conçus en utilisant différents types de cendres (trois types).

b) La base des données : La base de données examinée par les chercheurs peut être divisée en deux parties principales :

Les entrées, qui comprennent six éléments comme suivant :

1) Les types de cendres volantes Ash : Ils sont trois types cendres de charbon-type 1, cendres de bagasse-type 2 et cendres de cosse de cacahuète type-3.

2) Le pourcentage de cendres volantes Ash% : Il présente le pourcentage de cendres ajouté de trois type mentionnés précédemment et ses valeurs comme suivant, une valeur maximale de 60% et minimale de 0% et de moyenne de 14.02%.

3) Limite de liquidité LL : Ces valeurs ont été déterminées par les essais de limites d'Atterberg, où la valeur maximale a atteint 64% et la valeur minimale de 17%, avec une moyenne de 39.38%.

4) Limite de plasticité LP : Ont été déterminées aussi par les essais de limites d'Atterberg, où la valeur maximale a atteint 30.10% et la valeur minimale de 12.80%, et avec une moyenne de 20.64%.

5) Le teneur en eau optimal OMC : Ces résultats sont obtenus aussi par l'essai de Proctor, qui détermine la valeur optimale de teneur en eau dans le sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur est donnée en % et varie entre une valeur maximale de 32.50% et une valeur minimale de 8.91% et avec moyenne de 17.77%.

6) La densité sec optimale MDD : Ces résultats sont obtenus par l'essai de Proctor, qui détermine la valeur optimale de la densité sec du sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur varie entre une valeur maximale de 1.88 et une valeur minimale de 1.37 avec moyenne de 1.61.

Les sortes, CBR : Ou bien CBR trempé pour simuler les pires conditions: le matériau de la couche de fond gagne de l'humidité. Cette valeur est déterminée en laboratoire en se basant sur l'expérience CBR mentionnée ci-dessus (comme mentionné par Semachaw) en respect l'addition des cendres par un pourcentage étudié, et ses valeurs sont variées entre 6.46 et 1.86 avec une moyenne de 3.77.

III.3.2. L'étude statistique de base des données : Les figures ci-dessus montrant la corrélation et la distribution de paramètres dans cette étude :

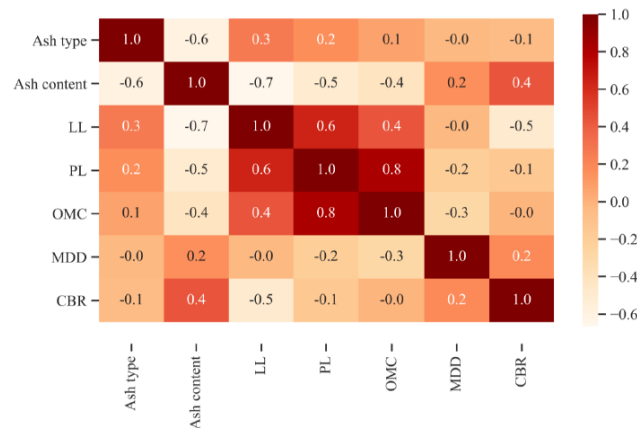


Figure 1.10. La matrice de corrélation (recherche de Van et Hai 2021).

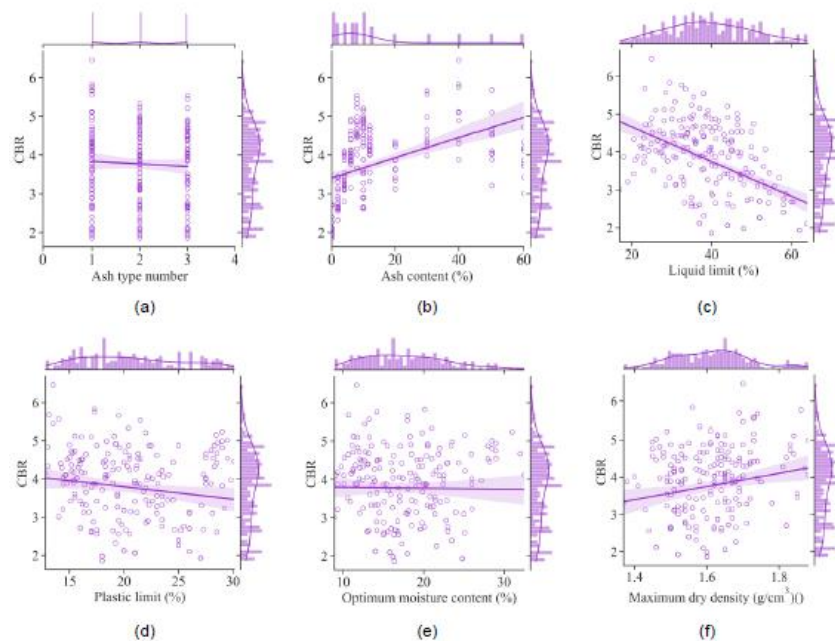


Figure 1.11. Distribution et ligne de corrélation de chaque variable d'entrée et de la sortie CBR.

III.3.3. Evaluation de performance de la bibliothèque LightGBM pour prédire le CBR :

LightGBM est une bibliothèque open source qui offre une mise en œuvre efficace d'un cadre de renforcement graduel basé sur des algorithmes d'apprentissage basés sur des arbres. Cet algorithme a été conçu par Microsoft depuis 2016.

L'algorithme présente certains avantages tels que la vitesse d'apprentissage, la haute précision et la fiabilité avec une faible utilisation de la mémoire. En ce qui concerne l'évaluation des performances, comme pour les autres recherches, les valeurs de coefficient de corrélation R et MSE et RMSE ont été suivies.

Les résultats qui obtenus sont résumées dans le tableau et les figures suivants et elles sont représenté les valeurs d'évaluation statistiques.

Tableau 1.4. Les résultats d'étude (Van et Hai Quan 2021).

Modèle	Data	R	MAE	RMSE
LightGBM	Entraînement	0.9473	0.2530	0.3303
	Test	0.9385	0.2506	0.3037

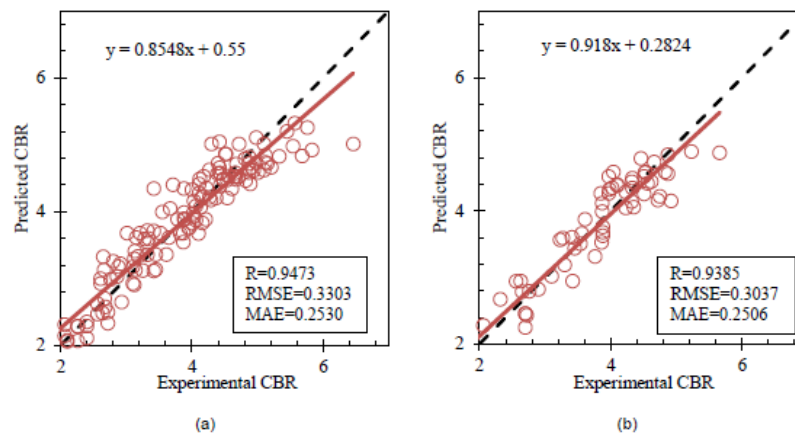


Figure 1.12. Analyse de corrélation entre le CBR expérimental et prédit des sols expansifs stabilisés. (a) ensemble de données d'entraînement, (b) ensemble de données de test.

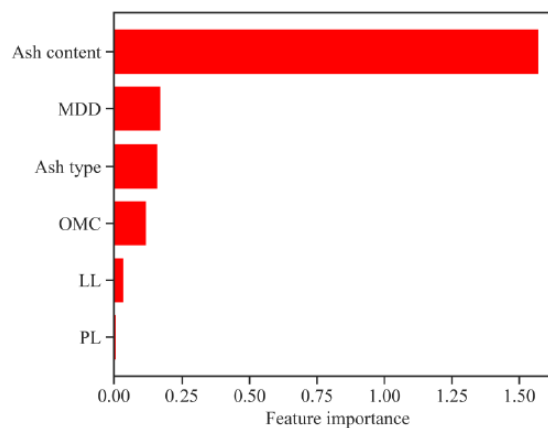


Figure 1.13. Test de sensibilité (recherche de Van et Hai 2021).

L'analyse de corrélation entre le CBR expérimental et prédit des sols expansifs stabilisés montre que le modèle de stabilisation est performant dans la prédiction de la valeur du CBR pour les sols traités avec des cendres résiduelles agricoles.

Nous pouvons ignorer le paramètre LP parmi les entrées pendant l'entraînement pour la force du coefficient de corrélation entre-ils et le coefficient LL, ainsi que leur effet convergent sur le sortie CBR.

Les valeurs de corrélation R et d'erreur sont respectivement de RMSE et MAE pour l'ensemble de données d'entraînement et de test. Ces résultats indiquent que le modèle de LightGBM est capable de prédire avec précision la valeur du CBR pour les sols expansifs stabilisés, ce qui est confirmé par les résultats satisfaisants obtenus dans les deux phases de l'étude, à savoir l'entraînement et la vérification.

Le test de sensibilité qu'ils ont effectué a montré que le type de cendre ajouté Ash a le plus grand impact sur la prédiction de CBR.

Conclusion : Cette étude présente la capacité des techniques d'intelligence artificielle à prédire le CBR des sols gonflants stables. La base de données utilisée pour construire le modèle a été obtenue à partir de la littérature. Un modèle Light Gradient Boosting Machine (LightGBM) a été développé pour réduire le temps et les coûts des expériences. En utilisant les paramètres par défaut, le CBR des sols gonflants stables peut être prédit avec succès en utilisant le modèle LightGBM avec une grande précision, les valeurs de performance étant: R (0,9452), RMSE (0,3225) et MAE (0,2522).

Par conséquent, cet algorithme est une bonne technique pouvant être appliquée pour estimer le CBR des sols gonflants stables. Les caractéristiques qui affectent le CBR par ordre d'importance sont: la teneur en cendres, MDD, le type de cendres, OMC, LL et PL. Le paramètre d'entrée PL peut être ignoré lors de l'entraînement du modèle LightGBM pour prédire le CBR des sols gonflants stables à l'avenir. La précision de prédiction du CBR pour le modèle LightGBM peut être améliorée en utilisant des algorithmes d'intelligence artificielle pour identifier les paramètres importants du modèle. [3]

III.4. SK Kamrul et autres 2020 : Dans cette étude, les chercheurs soulignent l'importance de la prédiction et de l'intelligence artificielle dans le développement et la construction des routes, étant donné que l'Inde dispose du deuxième plus grand réseau routier au monde après les États-Unis. La détermination de la résistance et de la préparation des routes à supporter les structures de chaussées dépend de la connaissance de la valeur CBR, qui nécessite des fonds et des efforts en laboratoire. Les chercheurs reconnaissent le travail de Yildiz et Guneydin en 2011 dans le développement d'une équation de corrélation entre CBR et l'analyse des tamis, les limites d'Atterberg, etc., en utilisant l'analyse de régression linéaire multiple et des coefficients statistiques qui ont montré la fiabilité du modèle. Ils saluent également le travail de Taskiran en 2010, qui a développé des modèles ANN et GEP pour prédire les valeurs CBR sur des sols humides et fins, en utilisant des coefficients statistiques tels que le coefficient de détermination (R^2) et l'erreur quadratique moyenne (MSE) pour évaluer la performance du modèle. Dans cette étude, trois modèles prédictifs pour les valeurs CBR ont été adoptés, à savoir les techniques de calcul doux telles que GEP, ANN et la méthode Krigging. Des relations ont été établies entre CBR imbibé et non imbibé et différentes propriétés du sol pour le sol de vallée dans l'État du Bengale occidental, en Inde, en utilisant ces méthodes. Ces relations ont été comparées aux corrélations disponibles précédemment et améliorées pour obtenir la valeur CBR maximale avec des propriétés de sol spécifiées.

III.4.1. Application sur les sols fins de bassin Bengale a l'Inde :

a) La région d'étude : De cette étude, les échantillons de sol ont été prélevés d'après le district de Murshidabad et la région de North 24 Parganas dans l'État du Bengale occidental en Inde. Ces échantillons ont été collectés et testés en laboratoire selon les spécifications définies par les normes indiennes concernées. Il convient de noter que le bassin du Bengale est formé de sédiments fluviaux et marins pour les systèmes fluviaux du Gange et du Brahmapoutre, constituant ainsi la plus grande île deltaïque du monde.

b) La base des données : La base de données examinée par les chercheurs peut être divisée en deux parties principales :

Les entrées, ils ont adopté 8 entrées présentés dans les lignes et le tableau suivants:

1) La densité spécifique du sol G : Elle a déterminée selon le test de pycnomètre qui a été effectué conformément aux spécifications (IS-2720 (partie 3) 1980).

2) Limite de liquidité LL et 3) Limite de plasticité LP ainsi que 4) Indice de plasticité IP : Ces valeurs ont été déterminées d'après le test de limite d'Atterberg qui a été effectué selon (IS-2720 (partie 5) 1985).

5) Le teneur en eau optimal OMC et 6) la densité sèche optimale MDD : Ces résultats sont obtenus aussi par l'essai de Proctor, qui détermine la valeur optimale de teneur en eau dans le sol lors d'un coup de compactage à 55 coups. Selon les chercheurs, cette valeur est donnée en % selon (IS-2720 (partie 7) 1980).

7) Le coefficient de uniforme Cu et 8) de courbure Cc : Une analyse granulométrique a été réalisée en tamisant des échantillons de sol à travers une série de tamis pour déterminer les caractéristiques de la distribution de la taille des particules dans le sol (IS-2720 (partie 4) 1985) et après on peut calculer les deux coefficients Cu et Cc.

Les sorties, on a deux sortes de **CBR :** Les sorties du test CBR sont effectuées conformément à (IS-2720- (partie 16) 1987) sur un sol compacté pour les deux cas imbibé et non imbibé en eau.

Classification des échantillons de sol, Les résultats des tests montrent que les limites de liquéfaction varient entre 14,85% et 70%. Cela indique que le sol sélectionné pour l'étude se situe entre des limites liquides basses et élevées. Les données de la limite plastique indiquent que la plupart des échantillons de sol étaient moyennement plastiques, à l'exception des échantillons M2, M4, M11, N5 et N9. L'indice de plasticité des échantillons de sol varie de 1,62% à 37,23%, ce qui indique une forte proportion d'argile dans le sol. En ce qui concerne les propriétés de plasticité, le sol a été classé comme **CL, ML, CL-ML, CI-MI, CI, CH-MH et MH** selon classification unifiée des sols.

Tableau 1.5. Exemple sur la base des données de (SK Kamrul 2020).

N échant	G	Cu	Cc	LL	PL	IP	Class-Sol	OMC	MDD	CBR Imbibé	CBR Non Imbibé
M1	2.5	6.0	0.3	35.0	17.7	17.2	CL-ML	13.5	1.7	14.0	10.0
M2	2.5	10.0	0.8	31.0	15.9	15.1	CL-ML	13.2	1.7	15.8	10.6
M3	2.2	6.6	0.5	28.7	13.3	15.4	CI-MI	14.9	1.7	16.4	8.9
N1	2.1	18.1	0.5	33.0	18.8	14.1	CL-ML	12.5	1.8	15.8	9.1
N2	2.2	11.2	1.0	30.0	19.0	10.9	CL	13.0	1.7	14.5	8.4
N3	2.1	3.5	2.1	63.4	32.7	30.6	MH	21.0	1.5	8.3	3.0

III.4.2. Développement de modèles GEP, ANN et la méthode de Kriggen pour prédire le CBR :

III.4.2.1. La programmation génétique d'expression (GEP) : Suit l'algorithme des gènes. Les étapes de GEP consistent à créer aléatoirement des chromosomes dans la population initiale en utilisant ce programme. Ensuite, les chromosomes sont transférés et la forme physique de chaque individu dans la population est évaluée. Ensuite, en fonction de la forme physique, l'individu approprié avec une mutation est sélectionné, ce qui conduit à la naissance d'une nouvelle descendance avec de nouvelles caractéristiques.

Dans la nouvelle génération, l'individu subit le même processus - affichage des génomes et compétition environnementale de sélection et reproduction avec mutation.

Ce processus se poursuit pendant un certain nombre de générations ou jusqu'à ce qu'une solution spécifique soit atteinte. Pour exprimer le CBR saturé et non saturé séparément en tant que fonction de G, Cu, Cc, LL, PL, PI, OMC et MDD, GeneXpro Tools 4.0 (www.genexprotools.soft112.com/) a été utilisé dans cette étude. Dans les deux cas, trente chromosomes ont été utilisés, chacun contenant trois gènes. "Addition" a été utilisée comme fonction de liaison. La valeur maximale de R² a été atteinte dans les générations 1025 et 124 respectivement. Les équations de CBR imbibé et non imbibé sont représentées comme suivant :

$$CBR_{imbibé} = ATAN((((d6/d5)3) - (d5)1/2) * SIN(d4)1/2))) + EXP((d1 - (COS((d8 - EXP(G2C1)))) 1/2)) + COS(((d2 + ATAN((d2/- G3C1))) + ((d1)2)1/2)) \dots (1.9)$$

$$CBR_{non-imbibé} = (sin(sin(sin((d[3])1/3))) + d[7]) + ((d[0] + (d[2] + G2C0))1/3) + (log(sin((d[5] + d[7])/d[6]))) + (d[7] + d[0])) \dots (1.10)$$

III.4.2.2. Les réseaux de neurones artificiels (ANN) : Sont une technique mathématique dérivée de la structure des cellules nerveuses du corps, et sont utilisés pour traiter des signaux non linéaires composés d'unités connectées appelées neurones. Ces réseaux sont entraînés en expérimentant et en présentant des données pour trouver des relations fonctionnelles entre elles, mais ces relations sont souvent inconnues ou difficiles à interpréter. L'algorithme de rétro-propagation est utilisé comme partie du modèle de réseau neuronal, et est particulièrement utilisé pour la prédiction et la modélisation de données.

Le processus de réseau neuronal artificiel consiste à ajuster les poids lorsqu'un ensemble de données d'entraînement apparaît et à utiliser une règle d'apprentissage pour prédire qu'un ensemble de poids peut être trouvé qui créera un ensemble d'entrées-sorties avec le moins d'erreurs possibles.

Dans cette étude, l'algorithme ANN se compose de huit paramètres d'entrée, à savoir G, Cu, Cc, LL, PL, PI, OMC et MDD, et deux paramètres de sortie qui sont CBR imbibé et non imbibé. Le programme Matlab ANN Toolbox a été utilisé pour effectuer cette opération.

Les performances des modèles sont évaluées en suivant les valeurs de R et la meilleure valeur est considérée comme étant proche de 1.

Après avoir expérimenté 18 réseaux différents, le meilleur réseau a été sélectionné avec une valeur R de 0,998.

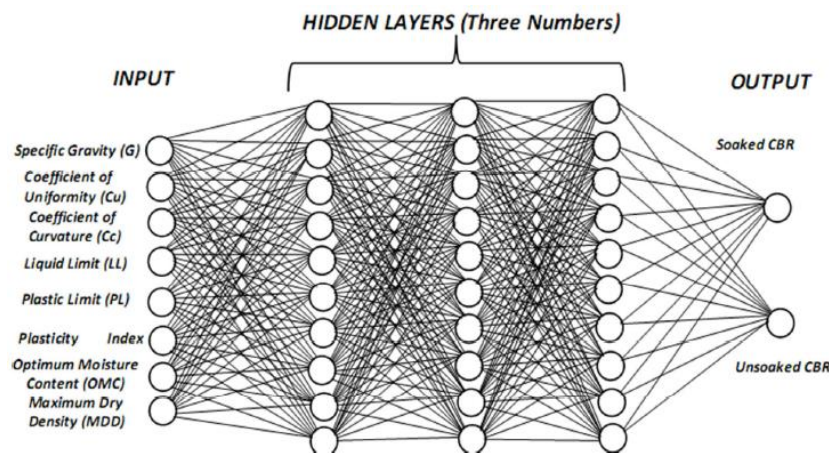


Figure 1.14. Le meilleur réseau pour prédire le CBR imbibé et non imbibé.

III.4.2.3. La méthode de Kriggen : Sacks et al. (Kaymaz, 2005) ont proposé que la réponse F(d) contienne une partie spécifique et une partie non-confirmée qui peut être définie comme suit :

$$F(d) = g^T(d)\alpha + G(d) \dots (1.11)$$

$$\text{où } g(d) = [g_1(d), \dots, g_n(d)]^T \dots (1.12)$$

$$\text{et } \alpha = [\alpha_1, \dots, \alpha_n]^T \dots (1.13)$$

n sont les fonctions de base dans le modèle de régression, $g(d)$ est une fonction connue pour d , α est un vecteur pour les coefficients de régression et $G(d)$ est une constante gaussienne avec une moyenne de '0' et une variance de σ^2 ainsi que la matrice de corrélation K . Les chercheurs ont également fourni une matrice de corrélation R dirigée par le coefficient k . Il est important de noter que cette étude a adopté la fonction de corrélation exponentielle $e^{(-\theta |d|)}$ pour prédire le CBR à partir des propriétés d'indice du sol en utilisant la méthode Kriging DACE A MATLAB Krigingtoolbox (Lophaven, Nielsen et Søndergaard, 2012).

III.4.3. Résultats et discussion : Les résultats qui obtenus sont résumées dans les deux figures suivants et elles sont représenté les valeurs d'évaluation statistiques par une comparaison entre leurs résultats et les résultats des recherche précédents comme Taskiran et Gunaydonn.

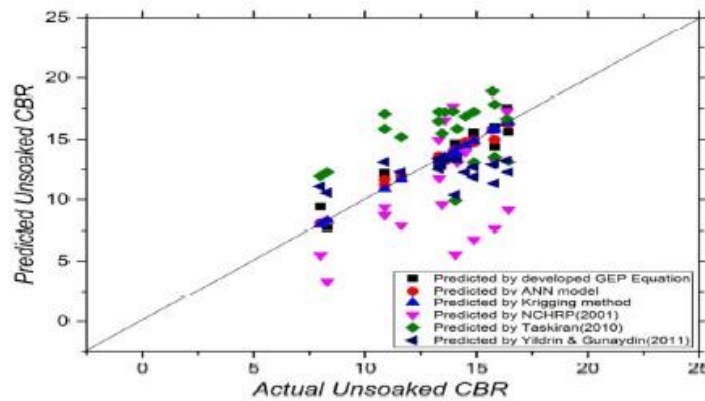


Figure 1.15. Comparaison du CBR non-imbibé expérimental et prédit.

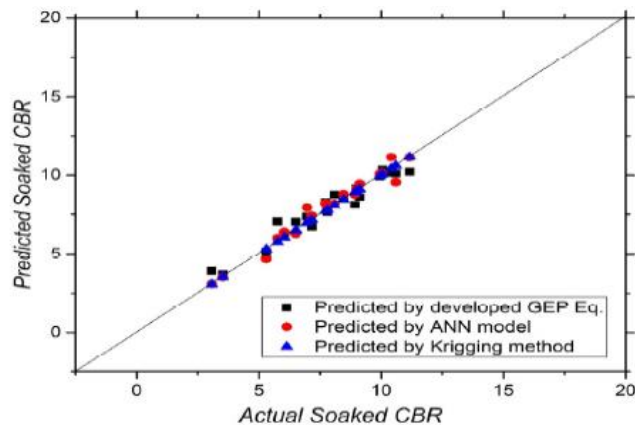


Figure 1.16. Comparaison du CBR imbibé expérimental et prédit.

Les chercheurs ont comparé les résultats des modèles Taskiran de Turquie 2010, Gunaydin 2011 et les modèles de NCHRP 2001 avec leurs propres modèles dans cette étude. Il a été démontré que le modèle de méthode de Kriging avait la meilleure performance avec une valeur de r égale à 0,998. Il a également été constaté que les valeurs CBR prévues dans le modèle de NCHRP 2001 étaient presque identiques aux valeurs réelles, à l'exception de quatre échantillons, qui pourraient être dus à des différences dans les propriétés du sol. De même, dans le modèle Taskiran 2010, il y avait une convergence entre les deux résultats, à l'exception de l'échantillon N5.

Conclusion : Cette étude visait à établir une relation entre les valeurs CBR et les différentes propriétés des indicateurs de sol tels que la densité (G), le coefficient d'uniformité (Cu), le coefficient de courbure (Cc), la limite liquide (LL), la limite plastique (PL), l'indice de plasticité (PI), l'humidité optimale (OMC) et la densité sèche maximale (MDD) en utilisant la programmation génétique, le modèle de réseau neuronal artificiel et la méthode Krigging. Vingt échantillons de sol ont été testés dans le sud-ouest du Bengale occidental, en Inde, et les valeurs CBR saturées et non saturées ainsi que les propriétés des indicateurs de sol ont été évaluées en laboratoire. Les résultats ont montré que ces trois méthodes peuvent prédire les valeurs CBR à partir des propriétés des indicateurs de sol mieux que les autres relations disponibles. Une optimisation multi-objectifs a également été réalisée pour obtenir les valeurs CBR saturées et non saturées maximales à partir des propriétés des indicateurs de sol du sol. L'étude a conclu que ces méthodes peuvent très bien prédire les valeurs CBR saturées et non saturées à partir des propriétés des indicateurs de sol du sol, ce qui aide les autorités locales à avoir une idée initiale des propriétés de la force du sol. D'autres études peuvent être menées en collectant un plus grand nombre d'échantillons de sol de la région du Bengale occidental. [4]

III.5. Taskiran 2010 : Le coût de la recherche dépend de l'accès aux études antérieures sur l'importance du CBR dans la détermination de la flexibilité et de la résistance des chaussées ou des sous-couches, qui sont essentielles pour la construction de routes et de pistes d'atterrissage. Cela se fait en comparant la force des matériaux de la sous-couche avec celle de la roche concassée, exprimée en pourcentage. Le chercheur a mentionné les difficultés auxquelles les ingénieurs sont confrontés pour déterminer ces valeurs, c'est pourquoi il a salué le rôle de l'intelligence artificielle dans la prédiction de ces valeurs. Plusieurs chercheurs ont tenté d'établir des relations entre le CBR et son influence sur les caractéristiques des sols, notamment les travaux de Black qui ont développé une corrélation entre le CBR et l'indice de plasticité PI des sols argileux.

Ghonim et Aggarwal ont également essayé de développer une équation reliant le CBR à la limite de liquidité LL ou la limite plastique LP, ou encore l'indice de plasticité PI, mais ils ont échoué à trouver une corrélation.

Au lieu de cela, ils ont trouvé une corrélation améliorée avec Wopt et LL, comme indiqué dans l'équation suivante :

$$CBR = 2.0 - 16.0 \times \log(Wopt) + 0.07 \times LL \quad \dots (1.14)$$

Le chercheur a mentionné les résultats du programme NCHRP aux États-Unis, qui a établi des liens entre les indicateurs du sol et le CBR. Deux équations ont été proposées, l'une pour les sols granulaires non plastiques contenant 12% de fines, comme suit :

$$CBR = 28.09 (D_{60})^{0.358} \quad \dots (1.15)$$

Cependant, lors de l'évaluation des recherches précédentes, il a été constaté qu'il n'y avait pas de corrélation satisfaisante dans de nombreux cas.

III.5.1. La base des données et la région d'échantillonnage :

a) La région d'étude : Dans cette étude, 151 données de test CBR des sols fins provenant de la Direction Régionale des Routes du Diyarbakır ont été utilisées. Ces données ont été analysées pour évaluer les matériaux de la chaussée/sous-couche des routes régionales dans quatre villes du sud-est de la Turquie (Diyarbakır, Mardin, Siirt et Şanlıurfa).

b) La base des données : 151 échantillons de sols fins ont été sélectionnés parmi un total de 354 collectés. Des tests ont été effectués pour évaluer l'adéquation du sol pour une utilisation en tant que

matériau de base et sous-couche au laboratoire central de la neuvième direction régionale des routes à Diyarbakir.

Les entrées, ils ont adopté 7 entrées présentés dans les lignes suivantes:

Les valeurs ont été déterminées en utilisant ASTM D-422 pour l'analyse mécanique et on prend.

1) La quantité de gravie G, 2) La quantité de sable S et 3) La quantité de limon-argile C+M.

Les limites d'Atterberg **4) Limite de liquidité LL et 5) Indice de plasticité IP** : Ils ont déterminés selon ASTM D-4318.

6) Le teneur en eau optimal W_{opt} et 7) la densité sèche optimale γ_d : Ils ont déterminés d'après l'essai de Proctor selon ASTM D-698 pour les tests de pression avec l'utilisation d'un pressiomètre standard Bruceton.

Les sortes, CBR : Les tests CBR ont été effectués conformément à AASHTO T-193 en comprimant les échantillons imbibés pendant quatre jours avec des charges de forme de disque pesant 2 kg pour simuler le poids des matériaux de revêtement qui couvrent la couche de base.

Classification des échantillons de sol, Les sols ont été classés selon le système américain de classification AASHTO M-147 et les résultats ont confirmé les types précis de sols des échantillons prélevés, à savoir **A-4, A-5, A-6 et A-7**.

III.5.2. Comparaison de performance de modèle ANN et de GEP pour prédire le CBR :

III.5.2.1. Les réseaux de neurones artificiels (ANN) : Le chercheur a constaté la difficulté de trouver des solutions pour les systèmes complexes d'ingénierie, ce qui a poussé les chercheurs à étudier les réseaux de neurones artificiels ANN, qui sont inspirés du comportement du cerveau humain et du système nerveux. Ils se composent principalement de trois couches : un ensemble de nœuds d'entrée, qui sont au nombre de 7 dans cette étude (la quantité de gravie G, la quantité de sable S, la quantité de limon-argile C+M, limite de liquidité LL, indice de plasticité IP, le teneur en eau optimal W_{opt} et la densité sèche optimale γ_d), une ou plusieurs couches cachées, qui comprennent 4 nœuds dans cette étude, et un ensemble de nœuds de sortie, qui représentent le CBR à prédire dans cette étude. Les cellules nerveuses ou les nœuds utilisent des algorithmes de calcul parallèle avec des poids de connexion ajustables. La méthodologie ANNs repose sur l'apprentissage à partir d'un ensemble de données présentées par la couche d'entrée et testées avec un autre ensemble de données de vérification. Les fonctions d'adaptation sont utilisées pour simuler le réseau, tandis qu'il est mis à jour à chaque pas temporel pour le vecteur d'entrée. Les fonctions de performance sont utilisées pour évaluer les modèles.

Dans ce modèle, la régression graduelle avec impulsion, le taux d'apprentissage adaptatif train_gdx , la régression graduelle avec impulsion pour le poids et la fonction d'apprentissage pour l'écart-type learn_gdm et l'erreur quadratique moyenne MSE ont été utilisés.

Au stade de l'apprentissage, le réseau commence par assigner les poids et les seuils au hasard. Les données sont divisées en 21% pour les tests et 79% pour l'entraînement, soit 32 résultats de test de vérification et 119 autres pour l'entraînement. L'outil ANN a été adopté dans le programme MATLAB pour effectuer les calculs nécessaires. De nombreux réseaux à couches multiples avec différentes fonctions de transfert sont testés et modifiés jusqu'à ce que la convergence soit atteinte dans les valeurs MSE.

Les valeurs du coefficient de détermination R^2 , de la moyenne des carrés moyens MSE, de la moyenne X_{mean} et de l'écart-type S_x sont adoptées comme outil de vérification statistique.

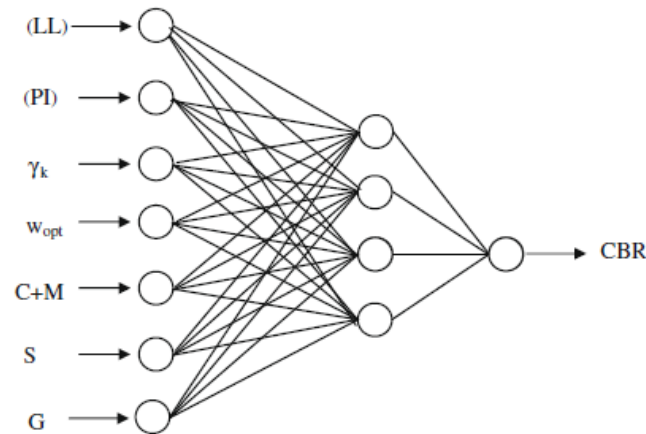


Figure 1.17. L'architecture des réseaux de neurones artificiels à propagation-avant avec sept entrées.

III.5.2.2. La programmation de l'expression génétique (GEP) : Est un algorithme basé sur des algorithmes génétiques et la programmation génétique, proposé pour la première fois par Ferreira. Son objectif principal est de développer une fonction mathématique adaptée à un ensemble de données fournies au modèle GEP. Elle exécute la régression symbolique à travers la plupart des opérateurs génétiques de GA. La différence entre GEP et GA réside dans le fait que le premier représente les individus comme des chaînes simples de longueur fixe exprimées par des arbres d'expression, tandis que le second représente des chaînes symboliques de longueur fixe.

Cinq paramètres sont utilisés dans l'analyse GEP : l'ensemble des gènes, la fonction de fitness, les paramètres de contrôle et la condition d'arrêt. Les valeurs attendues dans ce modèle sont comparées aux valeurs réelles et le fitness de chaque chromosome est calculé. Si cela est satisfaisant, l'algorithme se termine, sinon des chromosomes sont choisis au hasard et transformés en nouvelles générations.

a) Éléments de GEP : GEP se compose de deux éléments principaux : les chromosomes et les arbres d'expression. Le chromosome est composé d'un ou plusieurs gènes et représente une équation mathématique. Deux langages binaires et complets appelés Karva (langage des gènes) et Expression Trees (ET) sont utilisés pour coder n'importe quelle équation mathématique. Le gène GEP se compose de deux parties appelées tête et queue. La tête est utilisée dans le gène pour coder la fonction, mais la queue est utilisée comme symboles supplémentaires en cas d'insuffisance de symboles finaux dans la tête pour coder la fonction.

b) Opérateurs de GEP : Dans GEP, les opérations de sélection, mutation, transposition et croisement sont les principales opérations. Les chromosomes sont modifiés pour obtenir une meilleure aptitude pour la génération suivante en utilisant ces opérations. Les taux d'opérations sont définis au début de la construction du modèle et une certaine probabilité est attribuée à chaque chromosome. En général, le taux de mutation est recommandé de varier entre 0,001 et 0,1. D'autre part, le taux de transposition est recommandé d'être de 0,1 et le taux de croisement de 0,4.

Dans cette étude, différents modèles de GEP ont été développés en utilisant différents nombres de variables d'entrée, qui sont les mêmes que celles utilisées dans le modèle ANN (LL PI γ_d Wopt C+S S et G). Pour développer le meilleur modèle GEP, des paramètres tels que le nombre de chromosomes entre 25 et 35, le nombre de gènes entre 3 et 6 et la taille de la tête entre 7 et 12 ont été modifiés.

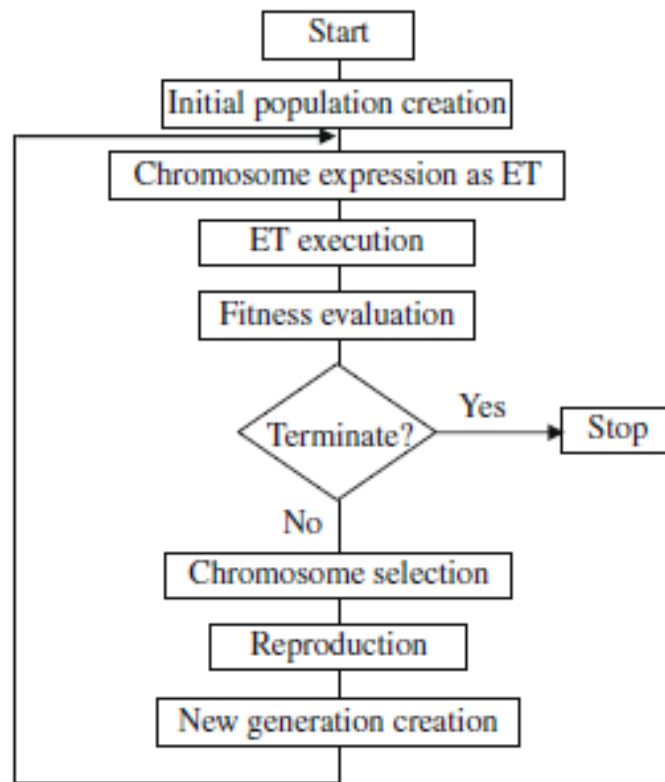
Les résultats ont montré que le premier et le cinquième modèle avec quatre variables d'entrée sont les meilleurs performants. Les fonctions mathématiques pour les deux modèles sont présentées dans les équations ci-dessous et pour éviter toute confusion, le contenu de sable et de gravier est désigné par F.

Modèle I :

$$\begin{aligned} \text{CBR} = & ((\gamma_d/\text{PI})/(((S+F) - (\text{LL} - 31.99)) - ((\text{LL} + 11.99) \\ & + 31.99))) + \text{sqrt}((((\text{LL} - w_{\text{opt}}) + G) - \text{PI}) * ((\text{PI} / -24.55) \\ & + \gamma_d))) + ((\gamma_d - ((\gamma_d * (w_{\text{opt}} - \gamma_d))/(F - \text{PI}) - 11.30))) * \gamma_d) \\ & + (((((44.14 - G) - F) - (G + \text{PI})) - (G/(18.91 \\ & - \text{PI}))))/18.91) + \text{sqrt}(((F/(((F + 1.04) - (S + \text{PI})) * (-22.56 \\ & + S))) + S)) \end{aligned}$$

Modèle V :

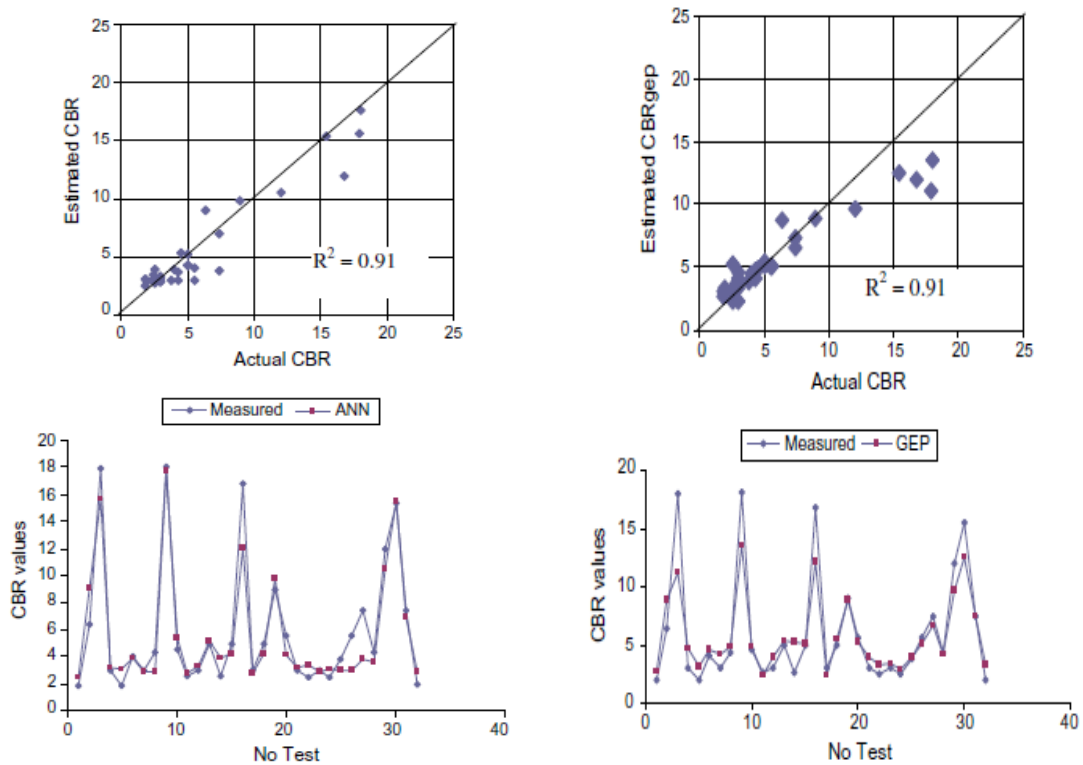
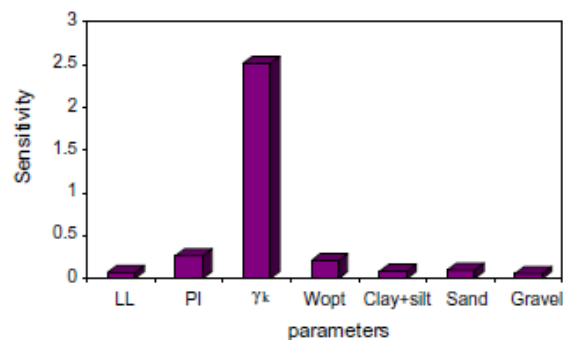
$$\begin{aligned} \text{CBR} = & (\text{sqrt}(\text{sqrt}(S))/(((F - 34.75) + (\gamma_d * \text{PI})) + (\text{sqrt}(8.28) - F))) \\ & + ((\gamma_d * ((S/\gamma_d) - (\gamma_d * 21.49)))/(\gamma_d + ((\gamma_d - 21.49)/\gamma_d))) \\ & + ((-2.091/\gamma_d)/(((2.37/\text{PI}) + \gamma_d) + (2.37 * \text{PI}) - F)) + ((S \\ & + (((S * \gamma_d) + 19.35) - 2\text{PI}) * \gamma_d)/23.02) \\ & + (\text{sqrt}(\text{sqrt}(\text{PI}))/(((-1.81 - F) + (\text{PI} * \gamma_d)) - ((-1.81 - \gamma_d) \\ & - 36.31))) \end{aligned}$$

**Figure 1.18.** Algorithme de GEP.

III.5.3. Evaluation des résultats d'étude de Taskiran 2010 : En créant sept modèles différents pour les deux modèles ANN et GEP en variant les entrées parmi les sept entrées originales. Les résultats pour les deux modèles sont donnés dans le tableau 1.6 statistique et comparatif qui représente les valeurs d'évaluation (la régression R^2 , l'erreur MSE, la moyenne des valeurs prédit X_{mean} et la déviation standard S_x) et les images suivantes montrent le graphe de régression R^2 dans l'étape de test ainsi que les résultats de test de sensibilité.

Tableau 1.6. La performance statistique de modèles ANN/GEP.

Modèle	ANN/GEP	R ²	MSE	Xmean	Sx
I	ANN	0.910	2.207	5.641	4.265
	GEP	0.918	1.287	5.750	3.110
II	ANN	0.641	8.679	5.428	4.275
	GEP	0.882	4.360	5.231	3.394
III	ANN	0.523	29.581	6.445	7.887
	GEP	0.812	4.770	5.493	3.699
IV	ANN	0.807	4.432	5.802	4.208
	GEP	0.821	5.125	5.611	3.317
V	ANN	0.838	4.271	5.978	5.205
	GEP	0.911	3.491	5.696	3.357
VI	ANN	0.885	2.869	5.818	4.071
	GEP	0.879	3.691	5.632	3.532
VII	ANN	0.681	8.147	5.526	3.143
	GEP	0.617	9.120	5.346	3.306

**Figure 1.19.** Comparaison entre les valeurs estimées (par ANN ou GEP) et les valeurs mesurées de CBR.**Figure 1.20.** Analyse de sensibilité pour les résultats de CBR.

Comme mentionné par le chercheur, l'évaluation des performances des modèles dans les deux méthodes se fait en suivant les variations des valeurs de (la régression R^2 , l'erreur MSE, la moyenne des valeurs prédit X_{mean} et la déviation standard S_x). Un modèle avec une valeur de $R^2 \geq 0,8$, et une faible valeur de MSE est considérée comme le meilleur en termes de performance.

Dans cette étude, il a été démontré que le modèle I avec toutes les entrées et le modèle V en supprimant Wopt et LL sont les meilleurs en termes de performance avec la méthode GEP. De plus, le modèle I avec toutes les entrées et le modèle VI avec les trois variables PI, γ_d et C+M sont les meilleurs en termes de performance avec la méthode ANN. Les résultats de l'évaluation sont présentés dans le tableau 5 ci-dessus.

Après avoir comparé les valeurs prédites avec les modèles les plus performants, l'efficacité des modèles sélectionnés a été confirmée et leur proximité avec les résultats réels mesurés a été constatée.

En suivant les valeurs de R^2 pour les deux modèles, leur efficacité a été démontrée, tous deux offrant de bonnes performances avec ANN et GEP. Cependant, les autres paramètres statistiques tels que MSE, S_x et X_{mean} étaient meilleurs et plus faibles avec le modèle ANN, ce qui suggère l'efficacité du modèle ANN par rapport à GEP. L'étude a également réalisé un test de sensibilité pour déterminer l'entrée ayant le plus d'impact sur la sortie CBR.

Les résultats ont montré que le poids unitaire sec était le plus efficace sur le CBR parmi les autres entrées, comme le montre dans la figure 1.20.

Conclusion sur l'étude de Taskiran : Lorsque les résultats sont évalués, il peut être conclu que l'ANN et le GEP sont capables d'apprendre la relation entre le CBR et les propriétés de base des sols et peuvent être utilisés pour la prédiction des valeurs de CBR des sols. De plus, l'analyse de sensibilité a montré que le poids unitaire sec γ_d est le paramètre le plus efficace sur le CBR, suivi par PI, wopt, S, C + S, LL et G respectivement. Étant donné que les tests de sol CBR sont assez difficiles, longs et coûteux, il peut être souligné que l'utilisation de modèles ANN et GEP pour estimer le CBR, en termes de paramètres de sol, pourrait être un outil utile à utiliser pour l'identification préliminaire du matériau ou/et une base de jugement pour la validité des valeurs de CBR. [5]

IV. Conclusion et suggestions :

Le rôle du CBR en tant qu'indicateur de la rigidité des matériaux, en particulier dans les structures de chaussées, est significatif dans toutes les études de recherche précédentes. La prédiction de ses valeurs sans recourir à des techniques de laboratoire est possible et a été étudiée par de nombreux chercheurs.

L'utilisation de des modèles d'apprentissage machine en tant que modèle de prédiction pour les valeurs de CBR a donné des résultats satisfaisants dans toutes les études de recherche précédentes.

Cela nous encourage à adopter cette technique en tant qu'outil de prédiction dans notre étude et l'appliquer à notre base de données.

La plupart des études de recherche précédentes ont été menées sur des sols de faible portance et ont donné des résultats satisfaisants. Cela nous donne confiance dans le succès du processus de formation sur notre base de données, qui est principalement constituée de sols fins.

Le processus d'évaluation en suivant les valeurs des coefficients de détermination R^2 et les moyennes d'erreurs MSE, RMSE et MAE a donné une satisfaction considérable, ce qui nous encourage à les utiliser comme moyen d'évaluer le succès du modèle développé.

Chapitre 2

*C*ollection et préparation de la base de données

I. Introduction : Il est indiqué que la nécessité est la mère de l'invention, et dans notre étude, notre besoin est de trouver des solutions innovantes pour résoudre les problèmes expérimentaux de régulation des valeurs CBR qui représentent un défi pour les chercheurs et nous poussent à explorer l'expérimentation numérique sur ce sujet.

L'intelligence artificielle est la dernière découverte de la science moderne, elle est similaire et se rapproche de l'esprit humain dans sa performance en utilisant l'apprentissage et la vérification.

Le travail de l'intelligence artificielle repose principalement sur la formation sur une base de données spécifiquement définie, quel que soit le domaine suivi (prédiction, classification, contrôle, etc.).

La base de données est constituée de valeurs expérimentales obtenues par des laboratoires ou des professionnels sur le terrain par des méthodes conventionnelles. Il est préférable qu'elle contienne des échantillons précisément classés, sélectionnés et soumis à des contrôles.

Dans cette étude, notre objectif est de présenter la base de données utilisée pour former les modèles prédictifs des valeurs CBR. Dans notre étude, nous avons utilisé une base de données contenant 1123 échantillons de sol prélevés de plusieurs villes de la région nord-est de l'Algérie.

II. L'étude de terrain : La collecte des données expérimentales a été effectuée par observation et questionnaire au niveau des archives du laboratoire de sols et des matériaux de l'Organisme National de Contrôle Technique des Travaux Publics CTTTP, basée à Alger, sous la supervision du chef de laboratoire, l'ingénieur Maiza Hamza (ingénieur d'État en travaux publics). Les données expérimentales ont été collectées pour la période 2011-2021.

II.1. Carte d'identification par laboratoire d'étude : Le laboratoire de sols et des matériaux de l'Organisme Nationale de Contrôle Technique des Travaux Publics CTTTP, basé à Alger, est l'un des laboratoires les plus importants et reconnus pour l'étude des sols en Algérie. Il se spécialise dans les domaines suivants :

- Routes et ouvrages d'art
- Aéroportuaire et maritime
- Géotechnique
- Recherche et développement

Adresse : BP n° 206 Kouba, 16230 Alger (route de Bach-Djerrah)-Algérie.

Téléphone : 021284432 / 021284436.

Fax : 021284436.

Adresse e-mail : cttp@cttp.de.

Site web : <http://www.cttp.dz>.



[6]

III. La base de données : Notre base de données comprenait 1123 échantillons de sol provenant de 14 wilayas du nord-est de l'Algérie, sur un total de 1680 échantillons préalablement collectés. Nous avons sélectionné des échantillons de sol avec une texture fine, contenant uniquement de l'argile et du limon, ainsi que des échantillons de sol provenant de sondages contenant du sols-supports.

III.1. La structure des chaussées (La couche sélectionnée) :

Les routes sont l'une des infrastructures les plus importantes dans le monde et sont considérées comme l'un des principaux indicateurs de progrès et de développement. C'est pourquoi les chercheurs, les ingénieurs et même les travailleurs s'efforcent de garantir une infrastructure géotechnique appropriée.

La structure des chaussées se compose généralement de couches de sol support, compacté, porteur ou friable, avec une couche de gravier et de bitume comme illustré dans la figure 2.1.

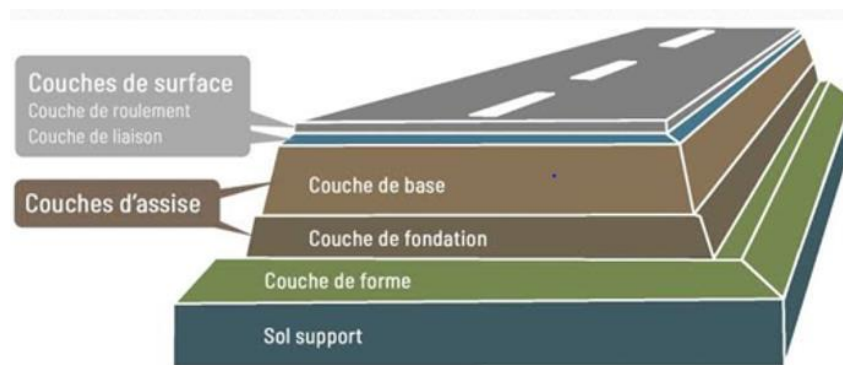


Figure 2.1. Coupe transversale dans une structure des chaussées.

III.1.1. Sol-support : La couche de sol-support est la couche la plus profonde dans la structure des chaussées. Elle est située directement sur le sol naturel et a pour rôle de supporter le poids de la chaussée ainsi que les charges des véhicules qui y circulent. La qualité du sol-support est donc essentielle pour assurer la stabilité et la durabilité de la chaussée.

Le choix des matériaux pour la couche de sol-support dépend des caractéristiques géologiques et hydrologiques du site. Les matériaux les plus couramment utilisés sont le sable, le gravier, les pierres concassées (GNT) ou les matériaux granulaires stabilisés. Ces matériaux sont compactés pour former une couche dense et résistante qui peut supporter les charges des véhicules.

Il est important que la couche de sol-support soit bien drainée pour éviter l'accumulation d'eau qui pourrait affaiblir la structure de la chaussée, (des techniques de drainage peuvent être utilisées pour assurer une bonne évacuation de l'eau).

III.1.2. Couche de forme : La couche de forme est un élément crucial dans la structure des chaussées car elle assure la répartition des charges des véhicules sur toute la surface de la chaussée. Elle est située juste au-dessus du sol-support et doit être suffisamment résistante pour éviter les déformations permanentes et les fissures. Les matériaux utilisés pour la couche de forme peuvent varier en fonction des conditions géologiques et climatiques de la région où se trouve la chaussée.

En général, la couche de forme est composée de matériaux granulaires tels que le gravier, le sable ou le concassé de roche. Ces matériaux sont compactés pour former une couche solide et résistante qui peut supporter les charges des véhicules. La qualité de cette couche est essentielle pour garantir la durabilité et la stabilité de la chaussée. Une couche conçue et bien construite peut prolonger la durée de vie de la chaussée et réduire les coûts d'entretien à long terme.

III.1.3. Couches d'assise :

a) Couche de fondation : La couche de fondation est mise en place sur la plate-forme support de la chaussée par compactage avec humidification et est exécutée sur toute la largeur de la plate-forme. Les matériaux couramment utilisés dans cette couche sont les tufs d'encroûtement.

La couche de fondation apporte la résistance mécanique aux charges verticales, répartissant les pressions sur la plate-forme afin de limiter les déformations à ce niveau au seuil admissible.

b) Couche de base : La couche de base est souvent composée par le gravé concassé, est aussi compactée avec humidification et exécutée sur toute la largeur de la plate-forme, avec des élargissements sur certaines sections (passages submersibles, remblais de grandes hauteurs, etc.).

III.1.4. Couches de surfaces :

a) Couche de liaison (imprégnation) : Préalablement à la réalisation d'une couche de roulement, les matériaux de la couche de base doivent recevoir une couche d'imprégnation bitumineuse de but :

- Assurer l'accrochage de la couche de roulement sur un matériau souvent très fermé en surface.
- Servir éventuellement de couche de roulement provisoire.
- Protéger la couche de base d'éventuelles précipitations.

L'imprégnation est réalisée sur toute la largeur de la plate-forme, le plus souvent au cut-back 0/1 et parfois au cut-back 10/15.

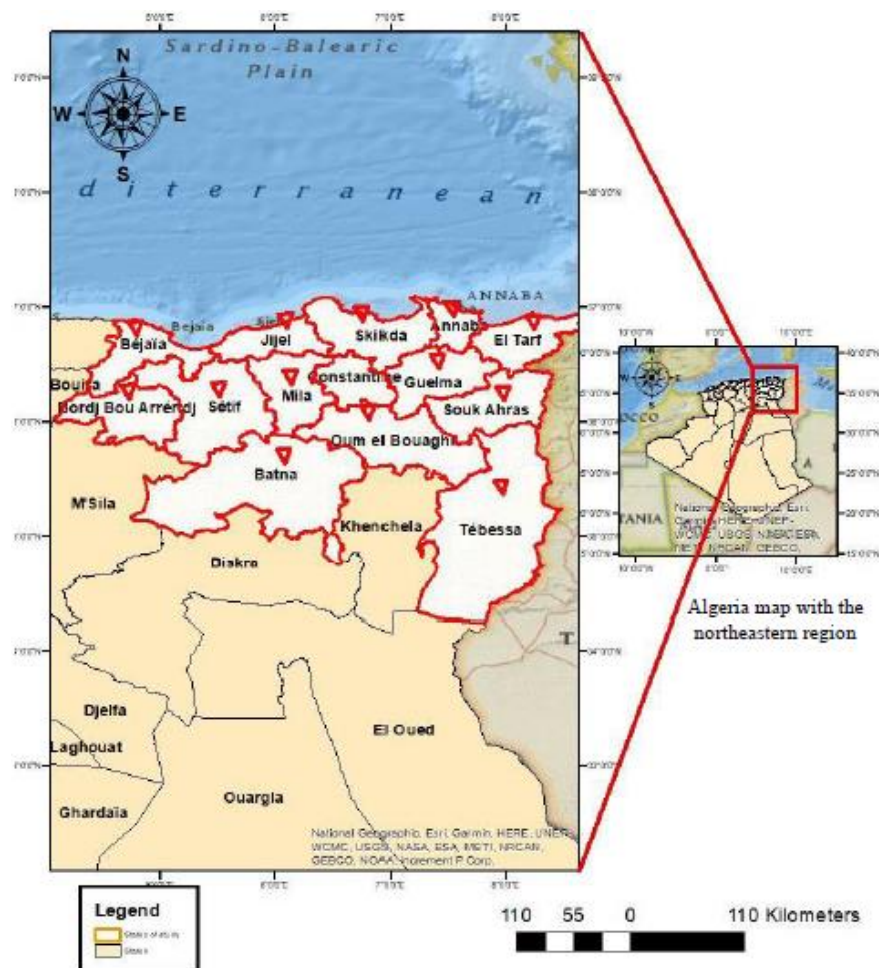
b) Couche de roulement : Selon l'intensité du trafic, la couche de roulement sera constituée par un enduit superficiel, un enrobé à froid, un enrobé à chaud ou un sable bitume.

Cette couche assure les fonctions d'adhérence avec les pneumatiques, elle doit résister aux actions directes du trafic et du climat. [7]

III.2. Les zones d'échantillonnage : Les laboratoires évaluent les sols avant la réalisation des projets à travers tout le territoire algérien, du nord au sud, de l'est à l'ouest.

Au départ, nous avons collecté 1680 échantillons de sols variés (sables, graviers 'GNT ou TVO', tufs et limono-argileux) provenant de plusieurs wilayas du pays. Ensuite, nous avons réduit le nombre à 1123 en nous concentrant uniquement sur les sols et limono-argileux, pour la période allant de 2011 à 2021, sous la supervision de l'ingénieur responsable du suivi des réalisations pendant cette période.

Les 1123 échantillons sont répartis dans 14 wilayas du nord-est de l'Algérie, couvrant les hauts plateaux et la côte, ces zones indiquées sur une photo bien déterminée à l'aide de programme ArcGIS 10.8, version 2020, comme mentionné dans la figure 2.2.



Algerian states, which samples have taken from:

Oum el Bouaghi – Batna – Béjaia – Tébessa – Jijel – Setif – Skikda – Annaba
– Guelma – Constantine – Bordj Bou Arreridj – El Tarf – Souk Ahras – Mila

Figure 2.2. Répartition des zones de collecte d'échantillons selon les wilayas Algériennes.

III.3. Les protocoles expérimentaux pour les tests de laboratoire CTTP : Concernant la base de données ont été utilisés pour collecter la base de données à partir des résultats expérimentaux sur les sols fins de la couche de sols-supports des régions du nord-est de l'Algérie.

La base de données sélectionnée se présente sous la forme de 12 entrées et sorties uniques CBR répétées successivement avec 1123 échantillons, comme indiqué dans le tableau 2.1, qui montre les paramètres du sol et ce qu'ils représentent, ainsi que les tests utilisés pour collecter ces éléments.

Tableau 2.1. Résumé des symboles d'entrée et de sortie et de leurs significations et expériences.

Type	Symbole	Signification	Essai
Entrées	G	Pourcentage de gravier.	Analyse granulométrique et l'essai de sédimentométrie.
	S	Pourcentage de sable.	
	FC	Pourcentage de sol-fin (limon-argile).	
	Cu	Coefficient de uniformité (Hazen), $Cu = \frac{D_{60}}{D_{10}} \dots (2.1)$	
	Cc	Coefficient de courbure, $Cc = \frac{D_{30}^2}{D_{10} \cdot D_{60}} \dots (2.2)$ avec: D_{60} , D_{30} et D_{10} représente les diamètres des particules pour lequel (60, 30 et 10)% en poids de l'échantillon passent à travers de ces ouvertures consécutivement.	
	MC	La teneur en eau dans chaque CC (10, 25 et 55).	L'essai de Proctor.
	DD	La densité sèche dans chaque CC (10, 25 et 55).	
	OMC	La teneur en eau optimal.	
	MDD	La densité sèche optimale.	
	LL	Limite de liquidité.	Limites d'Atterberg.
	LP	Limite de plasticité.	
	PI	L'indice de plasticité.	
Sorte	CBR	California Bearing Ratio.	L'essai de CBR.

Les normes NF P ont été adoptées dans le laboratoire des sols et des matériaux de l'institution CTTP pour réglementer les essais sur les sols. Ces normes sont basées sur notre base de données d'étude et incluent des protocoles d'essais approuvés :

III.3.1. L'analyse granulométrique : Selon les normes NF P 94-056 est une méthode pour déterminer la distribution des tailles de particules dans un échantillon de sol. Cette méthode implique le tamisage de l'échantillon à travers une série de tamis de tailles croissantes, en utilisant un dispositif appelé tamiseur mécanique. Les tamis sont agités pendant un temps déterminé, puis le poids de chaque fraction tamisée est mesuré. Les résultats sont ensuite utilisés pour tracer la courbe granulométrique, qui montre la proportion de chaque taille de particule dans l'échantillon. Cette analyse est importante pour déterminer les propriétés physiques et mécaniques du sol, telles que la perméabilité, la compressibilité, etc. Elle est utilisée dans de nombreux domaines, y compris la construction, l'agriculture, l'environnement, etc.

III.3.1.1. Objectif de l'essai : Déterminer la répartition des différentes tailles de particules dans un échantillon de sol.

III.3.1.2. Principe de l'essai : Le sol est tamisé à travers une série de tamis de tailles de mailles différentes, puis la masse de chaque fraction est mesurée pour déterminer la distribution granulométrique.

III.3.1.3. Matériel requis :

- Sol à tester.
- Tamiseur mécanique avec tamis de tailles croissantes (généralement de 80 μm à 80 mm).
- Balance de précision.
- Échantillon de sol (environ 500 g).
- Récipients pour les fractions tamisées.
- Brosse et spatule pour nettoyer les tamis.



Figure 2.3. Appareillage de l'essai d'analyse granulométrique. [8]

III.3.1.4. Exécution de l'essai :

1. Peser l'échantillon de sol (environ 500 g) et noter la masse.
2. Installer les tamis sur le tamiseur mécanique, en commençant par le plus gros tamis en bas et en finissant par le plus petit tamis en haut.
3. Placer l'échantillon de sol sur le tamis supérieur.
4. Mettre en marche le tamiseur mécanique pendant un temps déterminé (généralement 10 à 15 minutes).
5. Retirer chaque tamis du tamiseur mécanique et peser la fraction tamisée.
6. Noter la masse de chaque fraction tamisée et la placer dans un récipient approprié.
7. Répéter les étapes 3 à 6 pour chaque tamis.
8. Calculer la masse cumulée de chaque fraction tamisée (c'est-à-dire la somme des masses des fractions tamisées à travers tous les tamis plus petits que celui considéré).
9. Calculer la proportion de chaque taille de particule dans l'échantillon en divisant la masse cumulée de chaque fraction tamisée par la masse totale de l'échantillon.
10. Tracer la courbe granulométrique en représentant les proportions de chaque taille de particule en fonction de la taille des tamis.
11. Calculer les deux coefficients d'uniformité (Hazen) :

$$Cu = \frac{D_{60}}{D_{10}} \quad \dots (2.1)$$

Et de courbure :

$$Cc = \frac{D_{30}^2}{D_{10} * D_{60}} \quad \dots (2.2) \text{ avec :}$$

D_{10} , D_{30} et D_{60} représente les diamètres des particules pour lequel (10, 30 et 60)% en poids de l'échantillon passent à travers de ces ouvertures consécutivement.

III.3.2. L'essai sédimentométrique : Selon la norme NF P94-057 L'analyse des sédiments, ou sédimentométrie, est une méthode standard permettant de déterminer la distribution de la taille des particules et leur densité dans un échantillon de sol. Cette technique, largement utilisée en géologie, en agriculture et en ingénierie environnementale, consiste d'abord à préparer l'échantillon en le séchant et en le broyant afin d'obtenir une matière homogène. L'échantillon est ensuite mélangé à un liquide, généralement de l'eau, puis agité pour mettre les particules en suspension. Après agitation, le mélange est laissé au repos afin que les particules se déposent sous l'effet de la gravité selon leur taille et leur masse, les plus grosses se déposant plus rapidement que les plus fines. Les mesures de la vitesse ou de la profondeur de sédimentation permettent ensuite de déterminer la granulométrie des particules et d'établir leur répartition dans l'échantillon. En complément, l'analyse peut également fournir des informations sur la densité relative, la forme et la cohésion des particules. Les résultats obtenus sont essentiels pour évaluer les propriétés physiques des sols, notamment leur texture, leur capacité de rétention d'eau et leur comportement dans différentes conditions environnementales. Ainsi, la sédimentométrie constitue un outil fiable et reproductible pour l'étude des sols, la gestion des ressources naturelles, l'évaluation des phénomènes d'érosion, le choix des pratiques agricoles et la réhabilitation des sites contaminés.

III.3.2.1. Objectif de l'essai : Déterminer la distribution granulométrique et la densité relative des particules d'un échantillon de sol.

III.3.2.2. Principe de l'essai : Le principe de l'essai consiste à tamiser l'échantillon de sol sur une série de tamis de mailles différentes pour séparer les particules en fonction de leur taille. Les fractions de particules retenues sur chaque tamis sont pesées avec précision pour calculer la distribution granulométrique en pourcentage de la masse totale. La mesure de la densité relative permet d'estimer la forme et la compacité des particules.

III.3.2.3. Matériel requis :

- Sol à tester.
- Un tamiseur mécanique avec une série de tamis de mailles différentes.
- Une balance de précision.
- Des récipients pour prélever les échantillons.
- Un agitateur mécanique.
- Un densimètre ou un pycnomètre pour mesurer la densité relative.
- Des accessoires de manipulation (pince, spatule, entonnoir, etc.).
- Un four pour sécher les échantillons si nécessaire.
- Des produits chimiques pour éventuellement effectuer des analyses complémentaires (par exemple, pour déterminer la teneur en carbonate).



Figure 2.4. Appareillage de l'essai de sédimentométrique. [9]

III.3.2.4. Exécution de l'essai :

1. Préparation de l'échantillon : L'échantillon de sol est séché, désagrégié et pesé avec précision. La fraction fine destinée à l'analyse sédimentométrique est ensuite prélevée et mélangée avec un agent défloculant afin de disperser les particules.
2. Préparation de la suspension : Le mélange sol-eau est introduit dans une éprouvette graduée, puis agité énergiquement afin d'obtenir une suspension homogène avant le début des mesures.
3. Mesures sédimentométriques : Le sédimentomètre est plongé dans la suspension et des lectures sont effectuées à des intervalles de temps déterminés. La température de la suspension est également relevée afin d'apporter les corrections nécessaires.
4. Calcul des résultats : Les lectures du sédimentomètre sont utilisées pour déterminer le diamètre des particules en suspension et calculer leur pourcentage en fonction du temps de décantation, conformément à la loi de Stokes.
5. Établissement de la courbe granulométrique : Les résultats obtenus sont exploités pour tracer la courbe granulométrique de la fraction fine du sol.
6. Rapport : Un rapport est rédigé pour présenter les résultats de l'essai sédimentométrique, comprenant les mesures effectuées, la courbe granulométrique et l'interprétation des résultats.
7. Calcul des coefficients : Les coefficients d'uniformité (C_u) et de courbure (C_c) sont déterminés à partir de la courbe granulométrique afin de caractériser la répartition des dimensions des particules: (voir les équations (2.1) et (2.2)).

III.3.3. L'essai de limites d'Atterberg : Selon les normes NF EN ISO 17892-12: 2018 est une méthode pour déterminer les caractéristiques plastiques et liquides des sols. Cet essai permet de mesurer la teneur en eau à laquelle un sol passe de l'état plastique à l'état liquide et vice versa.

III.3.3.1. Objectif de l'essai : Déterminer les limites de plasticité et de liquidité d'un sol.

III.3.3.2. Principe de l'essai : Le sol est soumis à des essais permettant de déterminer la limite de liquidité et la limite de plasticité. Les résultats obtenus servent à évaluer la consistance du sol et à le classer selon son indice de plasticité.

III.3.3.3. Matériel requis :

- Sol à tester.
- Récipient en métal ou en verre.
- Balance précise, plaque en verre et spatule.
- Coupelle de Casagrande ou cône de pénétration (selon la méthode utilisée).
- Eau distillée et un tige de roulage ou les mains pour le façonnage des boudins.



Figure 2.5. Appareillage de l'essai de limites d'Atterberg. [10]

III.3.3.4. Exécution de l'essai :

1. Prélever un échantillon de sol d'environ 100 g.
2. Sécher l'échantillon au four à une température de 105°C pendant environ 24 heures.
3. Peser l'échantillon sec et noter la masse (m_1).

4. Ajouter de l'eau distillée à l'échantillon sec en petites quantités à l'aide d'une pipette ou d'un compte-gouttes, tout en mélangeant avec une spatule, jusqu'à ce que le sol atteigne l'état plastique.
5. Prélever une petite quantité du sol plastique et former une boule d'environ 3 cm de diamètre.
6. Placer la boule sur une plaque en verre ou en métal et l'aplatir avec un rouleau à pâtisserie jusqu'à obtenir une épaisseur d'environ 3 mm.
7. Utiliser un emporte-pièce pour découper un disque de 4 cm de diamètre dans la boule aplatie.
8. Peser le disque obtenu et noter la masse (m_2).
9. Placer le disque dans une bouteille graduée de 100 ml et ajouter de l'eau distillée jusqu'à atteindre un volume total de 100 ml.
10. Agiter vigoureusement la bouteille pour que le sol se dissolve complètement dans l'eau.
11. Laisser reposer la bouteille pendant 24 heures pour que les particules fines se déposent au fond.
12. Prélever avec précaution une éprouvette graduée de 50 ml du liquide clair situé au-dessus du dépôt de particules fines et noter le volume (V_1).
13. Répéter les étapes 4 à 12 avec des teneurs en eau croissantes jusqu'à ce que le sol atteigne l'état liquide.
14. Peser à nouveau l'échantillon restant après chaque étape et noter la masse (m_3).
15. Calculer les limites d'Atterberg du sol en utilisant les formules suivantes :

- Limite de liquidité :

$$(LL) = (V_1/V_2) \times 100 \quad \dots (2.3) \text{ où :}$$

V_2 est le volume total de la bouteille graduée (100 ml).

- Limite de plasticité :

$$(LP) = (m_1 - m_3)/m_2 \times 100 \quad \dots (2.4)$$

- Indice de plasticité :

$$(IP) = LL - LP \quad \dots (2.5)$$

III.3.4. L'essai de Proctor : Selon les normes NF P 94-093 est une méthode pour déterminer la teneur en eau optimal et la densité sèche optimale d'un sol pour la construction de routes, de pistes d'aéroport, de digues, etc. L'essai Proctor existe en deux types : le Proctor normal (standard), utilisé pour les sols soumis à des conditions de compactage ordinaires, et le Proctor modifié, destiné aux ouvrages nécessitant un compactage plus important grâce à une énergie de compactage plus élevée. Dans notre étude, nous avons réalisé l'essai Proctor modifié, car il est le plus adapté aux exigences de notre travail.

III.3.4.1. Objectif de l'essai : Déterminer la densité maximale d'un sol et son humidité optimale pour une compaction optimale.

III.3.4.2. Principe de l'essai : Le sol est compacté dans un moule cylindrique en acier en utilisant un nombre défini de coups d'un marteau standard. La densité et l'humidité sont mesurées pour chaque essai, et les résultats sont utilisés pour tracer une courbe de compaction.

III.3.4.3. Matériel requis :

- Sol à tester.
- Moule cylindrique en acier de 1000 cm³ de volume.
- Marteau de 2,5 kg avec une hauteur de chute de 30 cm.
- Balance précise.
- Bouteille graduée de 100 ml.
- Éprouvette graduée de 50 ml.



Figure 2.6. Appareillage de l'essai de Proctor. [8]

III.3.4.4. Exécution de l'essai :

1. Prélever un échantillon de sol d'environ 2,5 kg.
2. Sécher l'échantillon au four à une température de 105°C pendant environ 24 heures.
3. Tamiser l'échantillon séché à travers un tamis de 20 mm pour éliminer les grosses particules.
4. Peser 1000 g du sol tamisé et noter la masse (m_1).
5. Ajouter de l'eau distillée à l'échantillon sec en petites quantités à l'aide d'une pipette ou d'un compte-gouttes, tout en mélangeant avec une spatule, jusqu'à ce que le sol atteigne une teneur en eau proche de la limite de liquidité déterminée par l'essai d'Atterberg.
6. Préparer le moule cylindrique en acier en le nettoyant et en le remplissant avec le sol humide jusqu'à environ un tiers de sa hauteur.
7. Compacter le sol en utilisant une dame de 2,5 kg en le laissant tomber d'une hauteur de 30 cm sur chaque couche de sol ajoutée.
8. Ajouter des couches de sol compactées successives jusqu'à remplir complètement le moule.
9. Peser le moule rempli et noter la masse (m_2).
10. Retirer le moule et mesurer la hauteur du sol compacté à l'aide d'une règle.
11. Répéter les étapes 5 à 10 avec des teneurs en eau croissantes jusqu'à ce que la courbe de compaction soit obtenue.
12. Calculer la compacité sèche maximale (ρ_{max}) et la teneur en eau optimale (w_{opt}) en utilisant les formules suivantes :

$$\rho_{max} = m_2 / (1000 \times h) \quad \dots (2.6) \text{ avec :}$$

h est la hauteur du sol compacté en cm.

$$w_{opt} = [(m_1 - m_3) / m_3] \times 100 \quad \dots (2.7) \text{ où :}$$

m_3 est la masse de sol restante après compaction.

III.3.5. L'essai de CBR (California Bearing Ratio) : Selon les normes NF P 94-078 est une méthode pour déterminer la résistance d'un sol en conditions humides et sèches. L'essai CBR peut être réalisé selon deux modes: non imbibé, effectué sur un échantillon compacté sans immersion préalable, et imbibé, où l'échantillon est immergé dans l'eau pendant 4 jours afin de simuler les conditions les plus défavorables en présence d'humidité. Dans notre étude, nous avons retenu l'essai CBR imbibé, car il permet d'évaluer le comportement du sol dans des conditions plus représentatives.

III.3.5.1. Objectif de l'essai : Déterminer la capacité portante d'un sol en conditions humides et sèches.

III.3.5.2. Principe de l'essai : Le sol est compacté dans un moule cylindrique en acier, puis saturé en le laissant tremper dans de l'eau. Une charge est appliquée sur le sol à remplacer par 'Un piston cylindrique normalisé (section de 19.35 cm²) pénètre le sol à vitesse constante de 1.27 mm/min'. La presse enregistre l'effort nécessaire pour atteindre des enfoncements précis, en particulier 2.5 mm et 5 mm. L'indice CBR est un pourcentage obtenu en comparant l'effort mesuré sur le sol à un effort de référence obtenu sur un matériau concassé de haute qualité en Californie.

III.3.5.3. Matériel requis :

- Sol à tester.
- Moule cylindrique en acier de 150 mm de diamètre et 175 mm de hauteur.
- Piston en acier de 50 mm de hauteur et de 50 mm de diamètre, lors de l'essai, le piston de pénétration est déplacé à une vitesse constante de 1,27 mm/min. Le piston utilisé possède une surface de 19,35 cm² (diamètre de 50 mm) et les masses de surcharge appliquées sur l'échantillon sont généralement de 2,5 kg chacune, afin de reproduire les conditions de confinement du sol sur le terrain.
- Masse de 2,5 kg pour appliquer une pression sur le piston.
- Dispositif de mesure de la déformation verticale.
- Balance précise, éprouvette graduée de 100 ml et spatule.



Figure 2.7. Appareillage de l'essai de CBR. [11]

III.3.5.4. Exécution de l'essai :

1. Prélever un échantillon de sol d'environ 4,5 kg.
2. Sécher l'échantillon au four à une température de 105°C pendant environ 24 heures.
3. Tamiser l'échantillon séché à travers un tamis de 20 mm pour éliminer les grosses particules.
4. Peser 3000 g du sol tamisé et noter la masse (m1).
5. Ajouter de l'eau distillée à l'échantillon sec en petites quantités à l'aide d'une pipette ou d'un compte-gouttes, tout en mélangeant avec une spatule, jusqu'à ce que le sol atteigne une teneur en eau proche de la limite de liquidité déterminée par l'essai d'Atterberg.
6. Préparer le moule cylindrique en acier en le nettoyant et en le remplissant avec le sol humide jusqu'à environ la moitié de sa hauteur.
7. Compacter le sol en utilisant le piston en acier en appliquant une pression de 4,5 kg pendant 5 secondes.
8. Ajouter des couches de sol compactées successives jusqu'à remplir complètement le moule.
9. Peser le moule rempli et noter la masse (m2).
10. Retirer le piston en acier et mesurer la hauteur du sol compacté à l'aide d'une règle.
11. Immerger le moule rempli dans de l'eau pendant 4 jours pour saturer le sol.
12. Placer le moule sur le dispositif d'essai et installer le piston cylindrique qui pénètre le sol à vitesse constante.
13. Appliquer la pénétration du piston à vitesse constante tout en enregistrant la charge correspondante.
14. Mesurer la charge nécessaire pour obtenir des pénétrations de 2,5 mm et de 5 mm.
15. Poursuivre l'essai jusqu'à la pénétration requise en relevant les valeurs de charge à intervalles réguliers.
16. Calculer le CBR en divisant la charge nécessaire pour produire une pénétration de 2,5 mm (ou de 5 mm si elle est plus représentative) par la charge de référence correspondant à la même pénétration sur un matériau standard, puis exprimer le résultat en pourcentage.

III.4. Classification des sols : Les échantillons de sol ont été classés en fonction de leur texture selon les systèmes mondial USCS (Unified Soil Classification System) et américain AASHTO (American Association of State Highway and Transportation Officials).

Dans notre document, nous nous sommes concentrés uniquement sur les sols fins.

III.4.1. Classification selon USCS : Le système de classification unifié des sols répond aux exigences d'un système de, ce système est présenté dans la figure 2.8.

Les classifications sont basées sur les sols à grains fins et grossiers, et conservent les quatre groupes communs de sols - gravier, sable, limon et argile, les symboles sont facilement associés à la classification, étant simplement la première lettre du type de sol (sauf pour le limon, qui a la désignation M, du mot suédois pour limon), le raffinement dans le regroupement est basé sur la granulométrie d'un sol grossier qui est bien ou mal classé et la plasticité d'un sol à grains fins qui est élevée ou faible, ce qui semble être un grand nombre de groupes sont en fait des catégories très logiques. L'expérience a prouvé que le personnel non formé apprend très rapidement le système et l'utilise avec précision. Le système unifié inclut l'utilisation d'un graphique de plasticité pour aider à la classification des sols à grains fins.

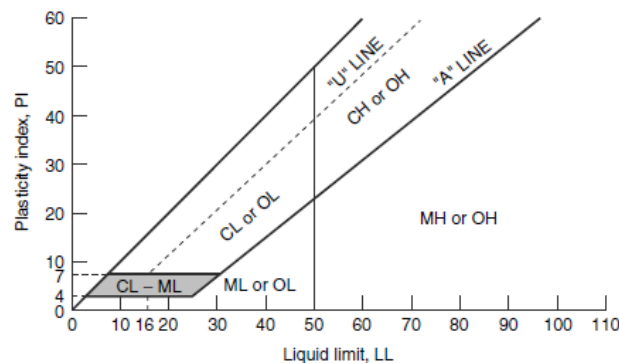


Figure 2.8. Plan de classification de sols d'après 'limite de liquidité LL et l'indice de plasticité PI' selon USCS système. [12]

Avec U-LINE c'est le droit d'équation :

$$PI = 0.9 (LL - 8) \quad \dots (2.8)$$

Et A-LINE d'équation :

$$PI = 0.73 (LL - 20) \quad \dots (2.9)$$

Le sol est classé en fonction de son point d'intersection d'abscisse LL avec l'ordonnée PI, ainsi que de sa position par rapport à la courbe linéaire définie par les équations U-LINE et A-LINE. On peut également déterminer le type de sol en fonction du pourcentage de passage des particules et du degré de plasticité. Les symboles de classification sont les suivants : C pour l'argile, M pour le limon ; et le suffixe H, I ou L est utilisé pour indiquer le degré de plasticité de l'échantillon : H pour les sols à forte plasticité, I pour les sols à plasticité moyenne et L pour les sols à faible plasticité.

Les classifications pour cette étude sont : **CL**, **CI**, **CH**, **MI** et **MH**, et d'autres échantillons présentent un mélange d'argile et de limon : **CL-ML**, **CI-MI** et **CH-MH**.

III.4.2. Classification selon AASHTO : Le système de classification AASHTO est largement utilisé au niveau national à les Etats-Unis pour les projets de construction liés aux autoroutes et aux transports, ce système classe les matériaux du sol en huit groupes, avec une affectation à un groupe

particulier en fonction de la distribution de la taille des particules, de la limite de liquidité et de l'indice de plasticité, ainsi que de la présence de matière organique. Les groupes de sol sont basés sur des critères de performance qui concernent le soutien de charge pour les revêtements de chaussées, comparé au système unifié, le système AASHTO met davantage l'accent sur la présence et la plasticité de la partie à grains fins d'un sol.

Un tableau de plasticité aide à identifier la fraction à grains fins, en général, plus le numéro de groupe est bas, meilleur sera le matériau pour le support de chaussée. Les comparaisons des sols dans le même groupe se font à partir de valeur d'indice de groupe GI (une valeur numérique calculée à l'aide de la formule empirique incluse dans l'équation suivante :

$$GI = (F-35)[0.2+0.005(LL-40)+0.01(F-15)(PI-10)] \dots (2.10), \text{ avec :}$$

F c'est le pourcentage passant au tamis 0,075 mm (No. 200).

Les critères pour les regroupements sont logiques, évaluant essentiellement l'aptitude à être utilisé comme matériau de base ou de sous-couche pour les chaussées, mais des tests en laboratoire sont nécessaires pour déterminer une classification, une comparaison généralisée des systèmes unifié et AASHTO est présentée dans le tableau 2.3.

Tableau 2.2. Regroupe des sols d'après la granulométrie selon AASHTO système.

General Classification	Granular Materials (35 Percent or Less Passing 0.075 mm)			Silt-Clay Materials (More than 35 Percent Passing 0.075 mm)			
Group Classification	A-1	A-3 ^a	A-2	A-4	A-5	A-6	A-7
Sieve analysis, percent passing:							
2.00 mm (No. 10)	—	—	—	—	—	—	—
0.425 mm (No. 40)	50 max	51 min	—	—	—	—	—
0.075 mm (No. 200)	25 max	10 max	35 max	36 min	36 min	36 min	36 min
Characteristics of fraction passing 0.425 mm (No. 40)							
Liquid limit	—	—	—	40 max	41 min	40 max	41 min
Plasticity index	6 max	N.P	b	10 max	10 max	11 min	11 min
General rating as subgrade	Excellent to good			Fair to poor			

III.4.2.1. Groupe des sols fins :

Groupe A-4 : Le matériau typique de ce groupe est un sol limoneux non plastique ou modérément plastique ayant généralement 75% ou plus qui passe le tamis de 0,075 mm (#200). Le groupe comprend également des mélanges de sol limoneux fin et jusqu'à 64% de sable et de gravier retenus sur le tamis de 0,075 mm (#200).

Groupe A-5 : Le matériau typique de ce groupe est similaire à celui décrit dans le groupe A-4, sauf qu'il est généralement de nature diatomée ou micacée et peut être très élastique comme indiqué par la limite de liquidité élevée.

Groupe A-6 : Le matériau typique de ce groupe est un sol argileux plastique ayant généralement 75% ou plus qui passe le tamis de 0,074 mm (#200). Le groupe comprend également des mélanges de sol argileux fin et jusqu'à 64% de sable et de gravier retenus sur le tamis de 0,074 mm (#200). Les matériaux de ce groupe ont généralement une forte variation de volume entre les états humide et sec.

Groupe A-7 : Le matériau typique de ce groupe est similaire à celui décrit dans le groupe A-6, sauf qu'il a les limites de liquidité élevées caractéristiques du groupe A-5 et peut être élastique ainsi que sujet à une forte variation de volume.

Sous-groupe A-7-5 : Comprend les matériaux avec des indices de plasticité modérés par rapport aux limites de liquidité et qui peuvent être très élastiques ainsi que sujet à une variation de volume considérable.

Sous-groupe A-7-6 : Comprend les matériaux avec des indices de plasticité élevés par rapport aux limites de liquidité et qui sont sujets à une variation de volume extrêmement élevée.

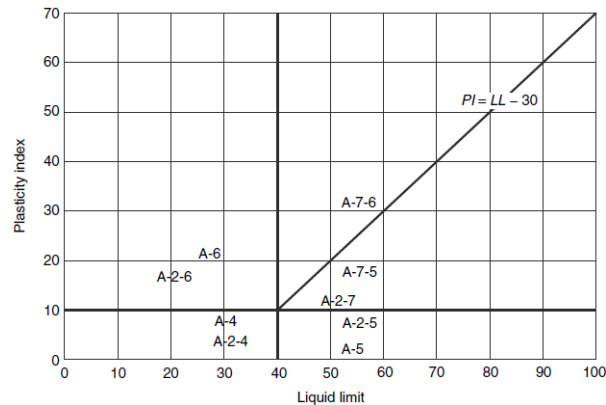


Figure 2.9. Plan de classification de sols d'après 'limite de liquidité LL et l'indice de plasticité PI' selon AASHTO système.

On peut classer le sol aussi d'après le point d'intersection de virgule LL avec l'ordonnée PI, ainsi que de sa position par rapport à la courbe linéaire définie par l'équation :

$$PI = LL - 30 \quad \dots (2.11), \text{ voir la figure 2.9.}$$

Les classifications pour cette étude sont : **A-4, A-5, A-6, A-7-5 et A-7-6.**

Tableau 2.3. Identification entre les deux systèmes de classification de sol AASHTO et USCS.

AASHTO	USCS
A-4	ML, OL
A-5	MH
A-6	CL
A-7-5	CL, OL
A-7-6	CH, OH

Avec le symbole O identifie les sols organiques. [12]

III.5. Exemple de base de données : Dans le tableau ci-dessous tableau 2.4, un exemple de la base de données sur laquelle l'étude a été menée et les modèles prédictifs ont été développés à travers l'entraînement et le test, représentant 1123 échantillons avec des coups de compactage différents entre 10, 25 et 55 provenant de différentes régions du nord-est de l'Algérie (des haute plateaux et les côtes).

La base de données repose sur 12 entrées comme mentionné précédemment et la sortie cible pour prédire la valeur CBR, tandis que le tableau montre également la classification des sols selon les systèmes USCS et AASHTO ainsi que l'index de groupe GI qui renforce la classification AASHTO du sol. Le tableau montre également la wilaya (région) du sondage (d'échantillonnage).

Tableau 2.4. Exemple de base des données.

Wilaya	C/C	G	S	FC	Cu	Cc	LL	LP	PI	Class- USCS	Class- AASHTO	GI	MC	DD	OMC	MDD	CBR
Batna	10	0.01	0	99.99	1.47	0.68	42.7	21.5	21.2	CI	A-7-6	23	14	1.66	13.9	1.78	1.3
	25	0.01	0	99.99	1.47	0.68	42.7	21.5	21.2	CI	A-7-6	23	14	1.7	13.9	1.78	1.98
	55	0.01	0	99.99	1.47	0.68	42.7	21.5	21.2	CI	A-7-6	23	14	1.77	13.9	1.78	2.97
	10	11	23	66	4.38	1.43	26.44	18.54	7.9	CL-ML	A-4	3	5.85	1.73	6.1	2.04	2.76
Annaba	25	11	23	66	4.38	1.43	26.44	18.54	7.9	CL-ML	A-4	3	5.85	1.93	6.1	2.04	4.29
	55	11	23	66	4.38	1.43	26.44	18.54	7.9	CL-ML	A-4	3	5.85	2.04	6.1	2.04	4.59
	10	5	27	68	21.53	0.59	31.7	20.52	11.18	CL	A-6	6	12.47	1.81	12.5	1.92	2.49
	25	5	27	68	21.53	0.59	31.7	20.52	11.18	CL	A-6	6	12.47	1.86	12.5	1.92	3.74
Sétif	55	5	27	68	21.53	0.59	31.7	20.52	11.18	CL	A-6	6	12.47	1.91	12.5	1.92	5.77
	10	2	23	75	2.93	0.34	92.8	40.52	52.28	MH	A-7-5	44	17.94	1.62	17.4	1.73	0.69
	25	2	23	75	2.93	0.34	92.8	40.52	52.28	MH	A-7-5	44	17.94	1.71	17.4	1.73	1.36
	55	2	23	75	2.93	0.34	92.8	40.52	52.28	MH	A-7-5	44	17.94	1.73	17.4	1.73	1.71
Mila	10	4	12	84	20.25	2.71	56.7	24.93	31.77	CH	A-7-6	29	14.21	1.55	14.1	1.71	9.48
	25	4	12	84	20.25	2.71	56.7	24.93	31.77	CH	A-7-6	29	14.21	1.64	14.1	1.71	12.01
	55	4	12	84	20.25	2.71	56.7	24.93	31.77	CH	A-7-6	29	14.21	1.71	14.1	1.71	15.96
	10	4	10	86	2.93	0.34	32.26	14.71	17.55	CL	A-6	14	12.63	1.68	12.6	1.86	1.63
BBA	25	4	10	86	2.93	0.34	32.26	14.71	17.55	CL	A-6	14	12.63	1.77	12.6	1.86	2.37
	55	4	10	86	2.93	0.34	32.26	14.71	17.55	CL	A-6	14	12.63	1.87	12.6	1.86	3.67
	10	7	12	81	3.57	0.28	37.2	17.44	19.76	CI	A-6	15	8.25	1.81	8.3	1.94	2.51
	25	7	12	81	3.57	0.28	37.2	17.44	19.76	CI	A-6	15	8.25	1.88	8.3	1.94	4.1
Béjaia	55	7	12	81	3.57	0.28	37.2	17.44	19.76	CI	A-6	15	8.25	1.94	8.3	1.94	5.85

III.6. Etude statistique de base de données : L'étude statistique de base de données est une étape cruciale dans le développement de modèles d'IA pour la prédiction, elle implique l'analyse des données collectées à partir de différentes sources et la détermination des valeurs maximales, minimales et moyennes de chaque variable d'entrée, le coefficients de corrélation est également utilisé pour identifier les relations entre les variables d'entrée et la variable de sortie cible.

Ces informations sont ensuite utilisées pour développer des modèles d'IA, les modèles sont entraînés sur un ensemble de données préalablement collectées et analysées.

La validité des modèles est vérifiée en utilisant des données qui n'ont pas été utilisées lors de l'entraînement (80% pour l'entraînement et 20% pour le test de validation).

III.6.1. Un aperçu statistique des valeurs de variables : Les valeurs maximales 'Max', minimales 'Min', moyennes 'Mean' et l'écart type 'St' sont toutes des mesures statistiques importantes qui sont utilisées pour analyser les données, la valeur maximale est la plus grande valeur observée dans un ensemble de données, tandis que la valeur minimale est la plus petite valeur observée, la moyenne est la somme de toutes les valeurs divisée par le nombre total de valeurs, tandis que l'écart type mesure la répartition d'une variable aléatoire autour de sa moyenne, le tableau ci-dessous tableau 2.5 est présente ces valeurs dans notre base de données.

Tableau 2.5. Les valeurs statistiques de base de données à la partie (d'entraînement et de test).

	Echnt	G	S	FC	Cu	Cc	LL	LP	PI	MC	DD	OMC	MDD	CBR
Train	Min	0.0000	0.0000	33.0000	1.0000	0.0100	21.1000	11.3200	6.3000	3.5700	1.4000	4.0000	1.5200	0.6900
	Max	16.0000	63.0000	100.0000	366.6600	2.88000	92.8000	40.5200	52.2800	18.3300	2.1300	18.5000	2.2000	26.7000
	Mean	4.0786	17.8110	78.1103	19.8926	0.6326	39.3106	20.0863	19.2242	11.3422	1.8176	11.4190	1.8810	6.5847
	St	4.6520	16.1182	18.6024	36.1508	0.7246	10.4159	4.4178	6.9504	2.8066	0.1281	2.7959	0.1119	5.8924
Test	Min	0.0000	0.0000	40.0000	2.6700	0.0200	24.8000	11.3200	8.5200	4.6100	1.6100	4.6100	1.7300	1.3600
	Max	16.0000	49.0000	99.9900	103.3300	2.6000	92.8000	40.5200	52.2800	18.5000	2.2000	18.5000	2.2000	25.5400
	Mean	3.6297	19.2868	77.0834	19.2954	0.5837	40.3228	20.6125	19.7102	11.6411	1.8602	11.7982	1.9117	8.1314
	St	4.9701	15.3293	17.3299	19.5774	0.7689	11.7998	5.1231	7.3827	2.9451	0.1278	2.9079	0.1215	7.5448

Ces valeurs sont importantes car elles permettent de comprendre la distribution des données et de détecter les valeurs aberrantes qui pourraient affecter les résultats de l'analyse, par exemple, si une valeur maximale est beaucoup plus grande que les autres valeurs, cela pourrait indiquer une erreur ou une anomalie dans les données.

III.6.2. Le coefficient de corrélation (Pearson) r : Le coefficient de corrélation r est une mesure statistique qui permet de déterminer la force et la direction de la relation entre deux variables ou plus dans une base de données. Il est souvent utilisé pour évaluer la relation entre des variables indépendantes (X) et des variables dépendantes (Y).

Lorsque deux phénomènes ont une évolution commune, nous disons qu'ils sont « corrélés ». La corrélation simple mesure le degré de liaison existant entre ces deux phénomènes représentés par des variables. Si nous cherchons une relation entre trois variables ou plus, nous ferons appel alors à la notion de corrélation multiple.

Nous pouvons distinguer la corrélation linéaire, lorsque tous les points du couple de valeurs (x,y) des deux variables semblent alignés sur une droite, de la corrélation non linéaire lorsque le couple de valeurs se trouve sur une même courbe d'allure quelconque,

Le coefficient de corrélation peut prendre des valeurs allant de -1 à 1, deux variables peuvent être :

- En corrélation positive (r approche ou égale à 1), on constate alors une augmentation (ou diminution, ou constance) simultanée des valeurs des deux variables.

- En corrélation négative (r approche ou égale à -1), lorsque les valeurs de l'une augmentent, les valeurs de l'autre diminuent.
- Non corrélées (r égale à 0), il n'y a aucune relation entre les variations des valeurs de l'une des variables et les valeurs de l'autre.

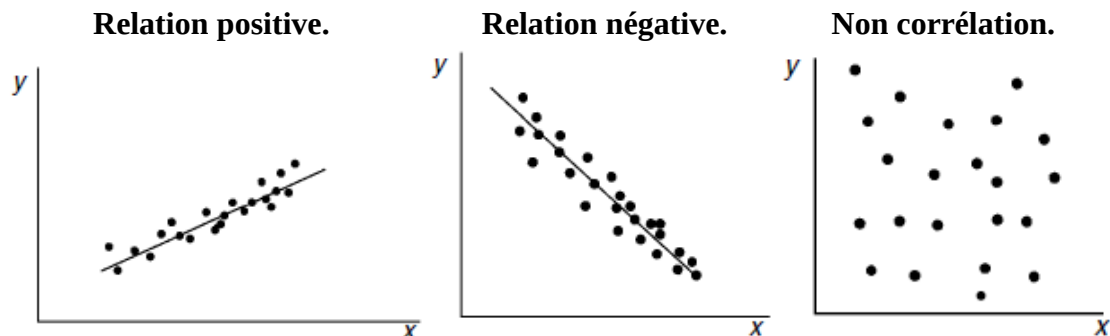


Figure 2.10. Les types de relation de corrélation linéaire.

Pour déterminer le coefficient de corrélation r , on utilise généralement la formule de Pearson, qui est la plus couramment utilisée, cette formule consiste à diviser la covariance entre les deux variables par le produit de leurs écarts types.

$$r = \frac{\text{cov}(X,Y)}{(\sigma_x * \sigma_y)} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad \dots (2.12) \quad \text{avec :}$$

$\text{cov}(x,y)$ = covariance entre x et y .

σ_x et σ_y = écart type de x et écart type de y .

n = nombre d'observations.

X_i = la valeur de variables X dans l'échantillon.

$\bar{X}$ = la valeur moyenne de variables X dans l'échantillon.

Y_i = la valeur de variables Y dans l'échantillon.

$\bar{Y}$ = la valeur moyenne de variables Y dans l'échantillon. [13]

Une fois que le coefficient de corrélation a été calculé, il peut être interprété pour déterminer la force et la direction de la relation entre les deux variables. Un coefficient de corrélation proche de 1 ou -1 indique une forte corrélation entre les variables, tandis qu'un coefficient proche de 0 indique une faible corrélation.

III.6.3. La matrice de corrélation (Pearson) : La matrice de corrélation de Pearson est une mesure statistique qui permet de mesurer la relation entre deux variables continues, elle est souvent utilisée dans l'analyse de données pour déterminer si deux variables sont liées et dans quelle mesure, la matrice de corrélation de Pearson est une matrice carrée qui contient les corrélations entre toutes les paires de variables dans un ensemble de données. Les valeurs de la matrice de corrélation de Pearson varient entre -1 et 1 , une valeur de 1 indique une corrélation positive parfaite, ce qui signifie que les deux variables augmentent ou diminuent ensemble, une valeur de -1 indique une corrélation négative parfaite, ce qui signifie que les deux variables se déplacent dans des directions opposées et une valeur de 0 indique qu'il n'y a pas de corrélation entre les deux variables.

Les fonctions de la matrice de corrélation de Pearson dans les bases de données sont multiples, elle peut être utilisée pour identifier les relations entre les variables, pour sélectionner les variables les plus importantes pour une analyse, pour détecter les valeurs aberrantes et pour évaluer la qualité des modèles de prévision, elle peut également être utilisée pour réduire le nombre de variables dans un ensemble de données en éliminant les variables qui ne sont pas corrélables, en somme, la matrice de corrélation de Pearson est un outil important pour l'analyse des données et la prise de décision en entreprise. La figure 2.11 représente la matrice de corrélation de notre base de données.

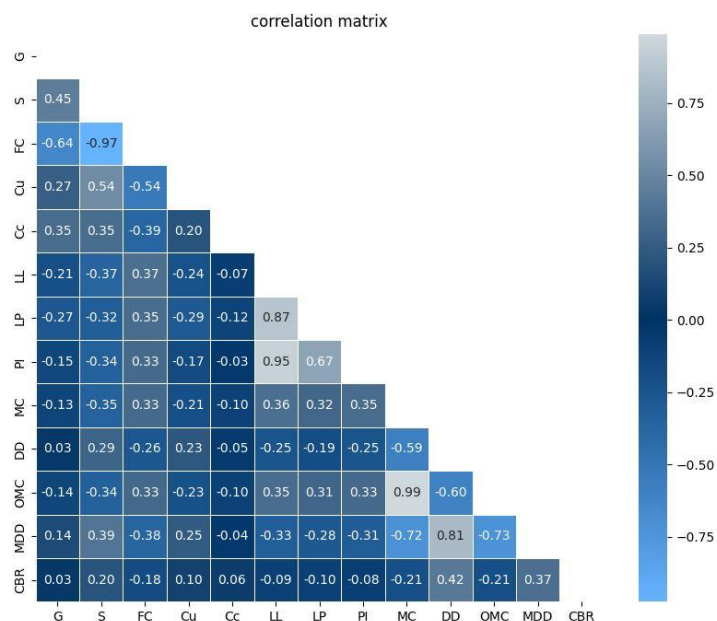


Figure 2.11. Matrice de corrélation (Pearson).

Cette matrice offre une vue d'ensemble des corrélations linéaires entre les caractéristiques géotechniques des sols obtenues à partir des données d'essai. Les coefficients de corrélation, compris entre -1 et 1, permettent d'évaluer à la fois l'intensité et le sens des relations entre les différents paramètres. Cette analyse est essentielle pour mieux comprendre les interactions entre les propriétés du sol dans les études géotechniques.

Des corrélations positives significatives sont observées entre certains paramètres, mettant en évidence leur comportement interdépendant. La relation entre la densité sèche (DD), la densité sèche maximale (MDD) et l'indice CBR (coefficients de corrélation de Pearson $r = 0,42$ et $r = 0,37$) montre qu'une densité plus élevée est généralement associée à des valeurs plus importantes du CBR, ce qui est conforme aux résultats de recherches antérieures. De plus, la corrélation entre le CBR et les caractéristiques des fractions grossières (G % et S %) souligne le rôle essentiel des particules grossières dans l'amélioration de la capacité portante du sol. Ce résultat confirme que la présence d'éléments grossiers contribue à accroître la résistance du sol aux charges. De même, la corrélation entre l'indice de plasticité (PI) et la teneur en fines (FC) ($r = 0,33$) met en évidence l'influence de la composition granulométrique sur le comportement du sol vis-à-vis de l'humidité, en montrant qu'une augmentation de la fraction fine est associée à une plus grande plasticité. Des corrélations modérées apportent également des informations importantes sur le comportement géotechnique. La relation négative entre la densité sèche maximale (MDD) et l'indice de plasticité (PI) ($r = -0,31$) indique que les sols les plus plastiques présentent généralement des densités maximales plus faibles, ce qui est un élément important pour l'optimisation des méthodes de compactage.

Enfin, bien que plus faible, la corrélation négative entre le CBR et la teneur en eau optimale (OMC) ($r = -0,21$) suggère que l'augmentation de l'humidité peut réduire la résistance du sol. La carte thermique (heat map) constitue ainsi un outil efficace pour identifier les relations significatives entre les principaux paramètres géotechniques. La compréhension de ces interactions est essentielle pour prévoir le comportement des sols, optimiser le choix des matériaux et améliorer la conception des ouvrages géotechniques.

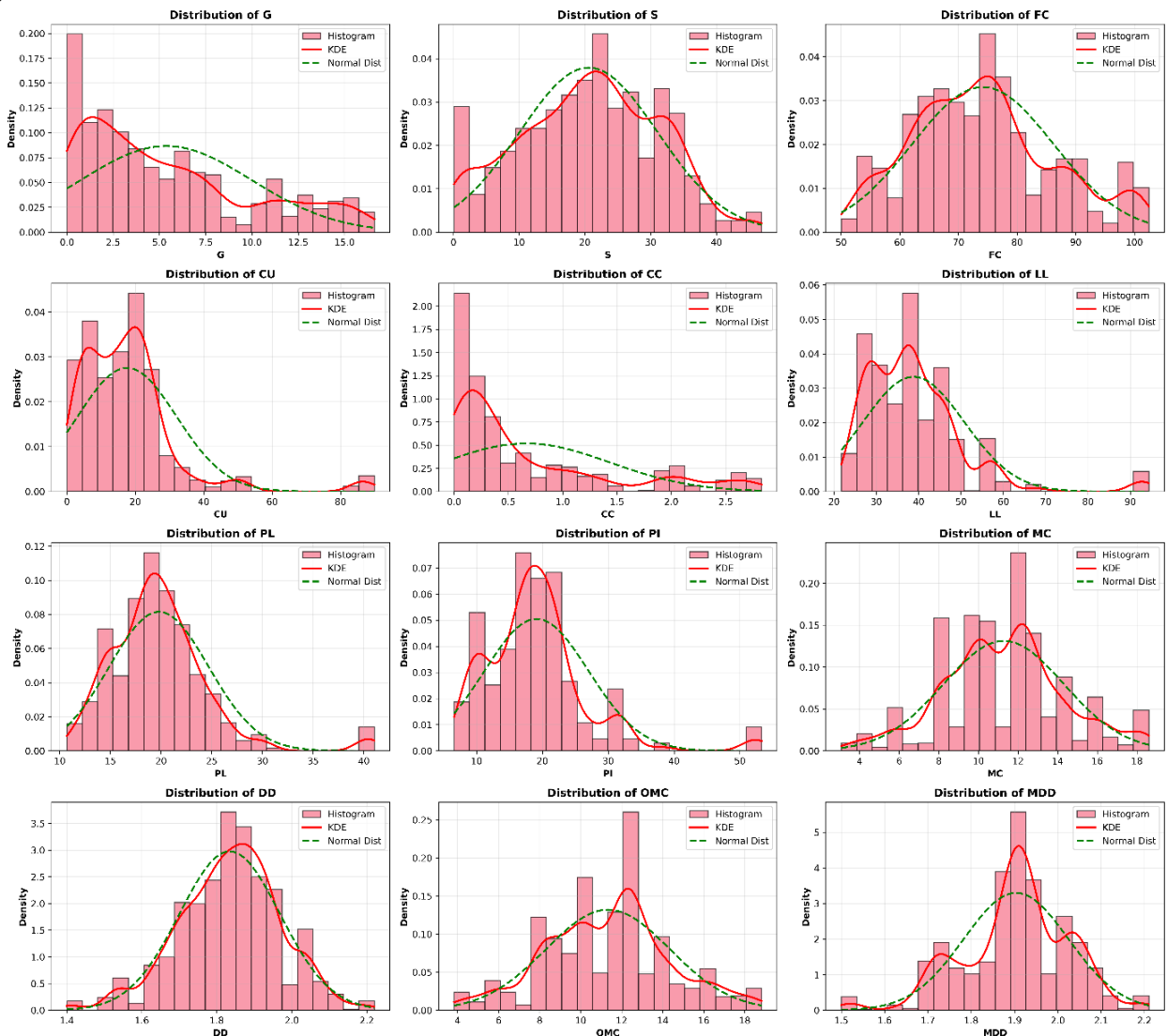
III.6.4. L'histogramme d'erreur : Un histogramme d'erreur (error histogram en anglais) est un graphique qui montre la distribution des erreurs dans un ensemble de données, il est souvent utilisé pour évaluer la qualité d'un modèle de prédiction ou d'une méthode d'estimation en comparant les valeurs prédites ou estimées avec les valeurs réelles, l'histogramme d'erreur peut aider à identifier les types d'erreurs les plus courants et à déterminer si les erreurs sont aléatoires ou systématiques.

Pour tracer l'histogramme d'erreur voici les étapes générales à suivre :

1. Collecter les données réelles et les données prédites ou estimées.
2. Calculer l'erreur pour chaque observation en soustrayant la valeur réelle de la valeur prédite ou estimée.
3. Regrouper les erreurs en intervalles (ou classes) de même amplitude, en choisissant le nombre et la taille des intervalles en fonction de la distribution des erreurs.
4. Compter le nombre d'observations dans chaque intervalle.
5. Tracer l'histogramme en représentant les intervalles sur l'axe horizontal et le nombre d'observations dans chaque intervalle sur l'axe vertical.
6. Ajouter une légende, un titre et des étiquettes d'axe pour rendre l'histogramme plus clair et compréhensible.

Il est également possible d'ajouter des informations supplémentaires à l'histogramme d'erreur, telles que la moyenne, la médiane, l'écart-type ou d'autres mesures de dispersion, ces informations permettent de mieux comprendre la distribution des erreurs et de comparer différents modèles ou méthodes d'estimation.

a)



b)

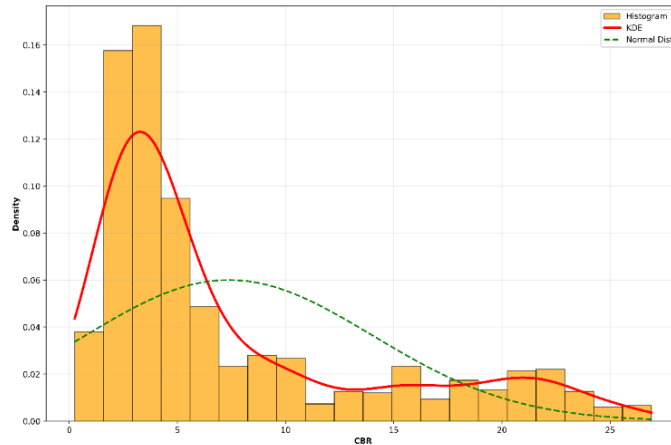


Figure 2.12. Histogramme avec une distribution probable des erreurs.
a) pour les 12 entrées; **b)** pour la sortie CBR.

La figure 2.12 représente la distribution de l'histogramme d'erreur pour toutes les 12 entrées et la sortie CBR., trois représentations : l'histogramme, la courbe de densité KDE (Kernel Density Estimation) et la courbe de distribution normale. L'histogramme montre la fréquence des observations, la KDE fournit une estimation lissée de la distribution réelle des données, tandis que la distribution normale sert de référence pour évaluer leur normalité. La comparaison montre que plusieurs variables présentent une asymétrie, une dispersion inégale ainsi que quelques valeurs extrêmes, ce qui traduit une distribution non parfaitement gaussienne. De plus, la courbe KDE suit généralement plus fidèlement la forme de la distribution que la courbe normale, mettant en évidence des écarts liés à la variabilité naturelle des propriétés géotechniques. Certaines variables présentent également plusieurs pics (multimodalité), suggérant l'existence de sous-populations ou de conditions expérimentales différentes. Ces observations confirment que les données sont hétérogènes et justifient le recours à des méthodes d'apprentissage automatique, capables de modéliser efficacement des relations complexes sans supposer une distribution normale des données.

IV. Conclusion :

Cette étude a présenté la collecte et la préparation de la base de données destinée à l'entraînement des modèles de prédiction du CBR, d'après la méthode d'observation et questionnaire au niveau de l'archive de laboratoire de sol et de matériaux au CTTP d'Alger, une base de données diversifiée de 1680 échantillons de différent types de sols et de couches de sondage a été collectée, dont 1123 sont limités aux sols fins entre limon et argile et de couche de sols-supports du nord-est de l'Algérie.

La base de données est composée de 12 entrées et d'une sortie CBR destinée à prédire sa valeur, tous les paramètres de sol sont issus d'essais en laboratoire pour la période 2011-2021 et conformément aux protocoles des normes françaises NF P.

Les entrées comprennent le pourcentage de gravier G, le pourcentage de sable S et le pourcentage des fines FC dans les échantillons, ainsi que le coefficient de courbure Cc et le coefficient d'uniformité Cu, les limites de liquidité LL et de plasticité LP ainsi que l'indice de plasticité PI, la teneur en eau MC et la densité sèche DD, la teneur en eau optimal OMC et la densité sèche optimale MDD, tous ces paramètres sont déterminés d'après les essais d'analyse granulométrique et de sédimentométrique, limites d'Atterberg, et l'essai de Proctor et de CBR.

Les valeurs statistiques des paramètres ont été déterminées dans la base de données (Max, Min, Mean et St), ainsi que la matrice et le coefficient de corrélation (Pearson) r, et un aperçu de l'histogramme avec une distribution probable des erreurs pour les 12 entrées et le sorte CBR ont été présenté.

De plus, la base de données a été divisée en 80% pour l'entraînement et 20% pour la validation.

Chapitre 3

***A**rchitecture des modèles*

I. Introduction : L'intelligence artificielle est devenue l'outil le plus efficace au monde en raison des résultats satisfaisants qu'elle a récemment obtenus. Peut-être que l'efficacité des modèles dans les études précédentes nous donne une motivation pour choisir ces mêmes modèles pour prédire les valeurs CBR dans cette étude et les comparer à d'autres modèles que nous choisirons.

Les types d'applications de l'intelligence artificielle sont multiples, allant de la classification à la prédiction en passant par le contrôle des outils, ...etc. Dans notre recherche, nous avons utilisé la prédiction en utilisant les réseaux de neurones ANN et l'apprentissage automatique ML.

Bien sûr, les réseaux de neurones sont de types différents (RBF, perceptron simple et MLP), tout comme l'apprentissage automatique qui comporte plusieurs modèles tels que (arbre de décision DT, forêts aléatoires RF, SVR, GPR, boosting, méthode kriging...etc.).

Il est évident que les modèles d'intelligence artificielle ont une structure théorique et mathématique, et leur modélisation n'est pas toujours conventionnelle. Ils peuvent être soumis à des optimisations par des techniques spéciales et peuvent être hybridés par des méthodes différentes dans le but d'ajuster leurs hyper-paramètres et d'atteindre le noyau optimal dans un court laps de temps, tout en prenant en compte l'augmentation des performances. C'est ce qui sera abordé dans les sections de ce chapitre.

Les programmes de développement et de contrôle dans le domaine de la modélisation sont multiples, tels que MATLAB le plus courant, R et PYTHON, ...etc.

II. La prédiction : La prédiction est une technique qui consiste à estimer ou à anticiper un événement futur en se basant sur des données historiques ou des tendances actuelles. Elle est utilisée dans de nombreux domaines tels que la finance, l'économie, la météorologie et la médecine pour aider à prendre des décisions éclairées.

Historiquement, la prédiction a une longue histoire qui remonte à l'Antiquité, où les astrologues utilisaient les étoiles pour prédire l'avenir. Au fil du temps, la prédiction est devenue plus sophistiquée et est devenue une discipline scientifique à part entière. Les mathématiques ont joué un rôle important dans le développement de la prédiction en fournissant des outils pour analyser et modéliser les données historiques. Les statistiques, par exemple, ont été utilisées pour identifier les tendances et les corrélations dans les données, tandis que les modèles mathématiques ont été utilisés pour simuler les scénarios futurs.

Mathématiquement, la prédiction utilise des techniques mathématiques pour analyser et modéliser les données historiques afin de prédire les événements futurs. Les mathématiques fournissent des outils pour identifier les tendances et les corrélations dans les données, ainsi que pour créer des modèles mathématiques qui peuvent simuler les scénarios futurs. Les statistiques sont utilisées pour quantifier la probabilité d'un événement futur en se basant sur les données historiques et les modèles mathématiques. Les équations différentielles, par exemple, sont utilisées pour modéliser les systèmes dynamiques qui évoluent dans le temps, tandis que les méthodes numériques sont utilisées pour résoudre ces équations et prédire l'évolution future du système.

Un exemple d'utilisation d'équations différentielles pour la prédiction est la modélisation de la croissance de la population. Supposons que nous avons des données historiques sur la taille de la population d'une ville et que nous voulons prédire sa croissance future. Nous pouvons utiliser une équation différentielle pour modéliser l'évolution de la population en fonction du temps. Par exemple, nous pouvons utiliser l'équation différentielle suivante :

$$dP/dt = rP(1-P/K) \quad \dots (3.1)$$

Où P est la taille de la population, t est le temps, r est le taux de croissance de la population et K est la capacité de charge de l'environnement (c'est-à-dire le nombre maximum de personnes que l'environnement peut soutenir). Cette équation décrit le taux de croissance de la population en fonction du nombre de personnes présentes et de la capacité de charge de l'environnement.

En résolvant cette équation, nous pouvons prédire la taille future de la population en fonction des données historiques et des hypothèses sur les taux de croissance et les capacités de charge.

Avec l'avènement de l'informatique, les modèles de prédiction sont devenus plus avancés et ont permis de traiter des quantités massives de données en temps réel. Aujourd'hui, la prédiction est une discipline multidisciplinaire qui utilise des techniques mathématiques, statistiques, informatiques et d'autres disciplines pour fournir des prévisions précises dans de nombreux domaines différents.

Les modèles de prédiction, tels que le modèle ANN, sont des outils importants pour la prédiction car ils peuvent analyser des données complexes pour fournir des prévisions précises. Cependant, il est important de noter que les prédictions ne sont pas toujours exactes et peuvent être influencées par des facteurs imprévus ou inconnus. Par conséquent, il est important d'utiliser les prévisions avec prudence et de les considérer comme un guide plutôt que comme une certitude absolue.

III. L'environnement de la modélisation des modèles : L'environnement de modélisation fait référence aux logiciels informatiques (langages de programmation) utilisés pour développer des modèles, qui sont actuellement multiples. L'intelligence artificielle utilise des algorithmes programmés dans un format d'application basé sur le programme MATLAB, qui varient en fonction du modèle, tels qu'ANN avec tous les noyaux (kernels), les modèles de génétique des populations GEP, les modèles d'apprentissage profond PL ainsi que les modèles d'apprentissage automatique ML ainsi que de classification et de contrôle, etc. Ou bien, ils utilisent des algorithmes disponibles dans les bibliothèques des entreprises émises pour ces programmes, comme c'est le cas avec le programme PYTHON. Le navigateur Google Colab nous permet de créer une version virtuelle programme PYTHON avec une grande capacité de stockage et une meilleure performance de processeur.

III.1. Programme PYTHON : PYTHON est un langage de programmation interprété, créé par le Néerlandais Guido van Rossum en 1991. Il a été conçu pour être facile à lire et à écrire, avec une syntaxe simple et claire, ce qui en fait un choix populaire pour les débutants en programmation.

Depuis sa création, PYTHON a connu une croissance exponentielle en popularité et est devenu l'un des langages de programmation les plus utilisés dans le monde. [14]

En ce qui concerne l'intelligence artificielle IA, PYTHON est devenu un choix populaire en raison de sa simplicité et de sa flexibilité. Il est utilisé pour créer des algorithmes d'apprentissage automatique ML, des réseaux de neurones artificiels ANN, des systèmes de reconnaissance d'image et de langage naturel, et bien plus encore. De plus, PYTHON dispose d'un grand nombre de bibliothèques open source, telles que **TensorFlow**, **Keras**, **Scikit-Learn** et **PyTorch**, qui facilitent la création de modèles d'IA.

PYTHON est également utilisé pour le traitement des données, ce qui est essentiel pour l'IA. Les bibliothèques telles que **Pandas**, **NumPy** et **SciPy** sont largement utilisées pour la manipulation et l'analyse des données. [15]

Le site web officiel de Python, <http://docs.python.org> offre une documentation complète pour les développeurs qui souhaitent apprendre et utiliser ce langage de programmation puissant et polyvalent.

IV. Choix des modèles et de ses architectures :

IV.1. Division de la base de données : En ce qui concerne les techniques d'apprentissage automatique, elles reposent sur PYTHON et utilisent une méthode d'entraînement-test. Elle consiste à entraîner un ensemble d'échantillons avec un pourcentage donné, puis à valider sur les échantillons restants. Ces pourcentages sont déterminés par une division aléatoire ajustée pour obtenir la meilleure performance des modèles.

En d'autres termes, les proportions des échantillons d'entraînement et de test et validation sont déterminées en fonction du résultat optimal de la performance du modèle final.

Dans notre étude, nous avons utilisé un ratio de 80% pour l'entraînement et de 20% pour le test et la validation.

IV.2. La méthode de validation croisée K-fold: La méthode de K-fold est une technique d'évaluation de la performance des modèles d'apprentissage automatique, elle consiste à diviser les données en k sous-ensembles, appelés « folds », où k est un nombre entier déterminé à l'avance. Ensuite, le modèle est entraîné sur k-1 folds et évalué sur le fold restant, cette procédure est répétée k fois, chaque fois en utilisant un fold différent comme ensemble de test et validation. Finalement, les résultats sont moyennés pour donner une estimation de la performance du modèle.

La 'K-fold cross validation' est une technique populaire car elle permet d'utiliser tous les échantillons de données pour l'entraînement et le test, ce qui peut améliorer la précision de l'estimation de la performance du modèle. De plus, elle permet de détecter les problèmes de sur-apprentissage (over-fitting) en évaluant la performance du modèle sur des données qui n'ont pas été utilisées pour l'entraînement. [16]

Il existe plusieurs fonctions en Python pour implémenter la K-fold cross validation, notamment la fonction K-Fold de la bibliothèque **Scikit-Learn**, cette fonction permet de spécifier le nombre de folds à utiliser, ainsi que d'autres paramètres tels que le mélange aléatoire des données avant la division en folds. [17]

Dans notre étude, la base de données est divisée en 10 parties représentées par le chiffre k. On s'entraîne sur les 9 premières parties puis on vérifie sur la 10^{ème}. Ensuite, on s'entraîne sur les 8 premières parties et la 10^{ème} et on vérifie sur la 9^{ème} et ainsi de suite jusqu'à ce que toutes les possibilités soient couvertes, cela nous permet de diviser la base de données pour l'entraînement et la vérification de manière optimale.

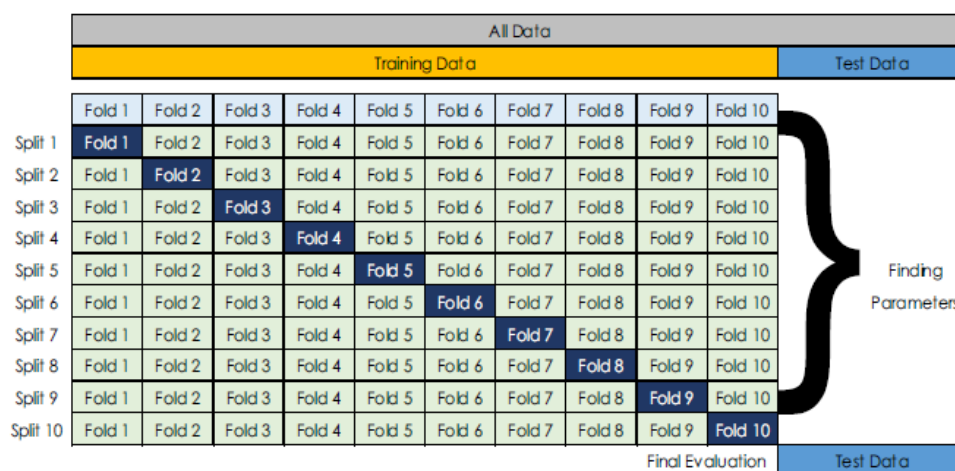


Figure 3.1. La méthode de K-fold cross validation. [18]

IV.3. L'apprentissage automatique (Machine Learning ML) et ces types :

L'apprentissage automatique, également connu sous le nom de machine Learning, est une branche de l'intelligence artificielle qui se concentre sur le développement de techniques permettant aux machines d'apprendre à partir des données et d'améliorer leurs performances au fil du temps sans être explicitement programmées. Il existe plusieurs types d'apprentissage automatique, notamment :

IV.3.1. L'apprentissage supervisé : Dans ce type d'apprentissage, un modèle est entraîné à partir d'un ensemble de données étiquetées, où chaque exemple est associé à une étiquette ou une classe, le modèle apprend à prédire les étiquettes pour de nouvelles données non étiquetées.

Les algorithmes couramment utilisés en apprentissage supervisé sont les réseaux de neurones, les arbres de décision et les machines à vecteurs de support.

IV.3.2. L'apprentissage non supervisé : Contrairement à l'apprentissage supervisé, l'apprentissage non supervisé ne nécessite pas d'étiquettes pour les données d'entraînement. Au lieu de cela, il identifie des structures ou des modèles intrinsèques dans les données. Les algorithmes d'apprentissage non supervisé comprennent la classification automatique, la réduction de dimensionnalité et la détection d'anomalies.

IV.3.3. L'apprentissage par renforcement : Dans ce type d'apprentissage, un agent apprend à prendre des décisions en interagissant avec un environnement, l'agent reçoit des récompenses ou des punitions en fonction de ses actions, ce qui lui permet d'apprendre quelles actions sont bénéfiques ou préjudiciables. Les algorithmes de renforcement les plus couramment utilisés sont la programmation dynamique, la recherche taboue et les algorithmes génétiques.

IV.4. Choix de la méthode d'apprentissage automatique : Le choix de la méthode d'apprentissage machine se fait en trois étapes :

IV.4.1. Identification du format de la variable de sortie : En effet, la variable de sortie oriente le choix du problème d'apprentissage machine. Celle-ci est définie à partir de l'objectif fixé par l'opérateur. Il existe deux types de données de sortie : les données quantitatives et les données qualitatives. Lorsque l'objectif de l'opérateur est de prédire une valeur spécifique, la variable de sortie est de type quantitatif. En revanche, lorsque l'objectif est de classer une donnée ou un paramètre, la variable de sortie est de type qualitatif. Par exemple : prédiction de la qualité (bonne ou mauvaise) les données sont de type qualitatif.

IV.4.2. Identifier le type de problème d'apprentissage machine : En se basant sur le format de la variable de sortie définie. En effet, si la variable de sortie est de type qualitatif, le problème d'apprentissage machine est un problème de classification ou de regroupement ou de partitionnement.

Dans le cas contraire, lorsque la variable de sortie est de type quantitatif, le problème d'apprentissage automatique est un problème de régression.

VI.4.3. Choix de la méthode : Celle-ci est fonction de la connaissance des informations contenues dans l'objectif de départ (type de données, connaissance des classes ...etc.).

Il existe trois grands groupes de méthodes d'apprentissage machine :

- Les méthodes supervisées.
- Les méthodes non supervisées.
- Les méthodes d'apprentissage par renforcement. [19]

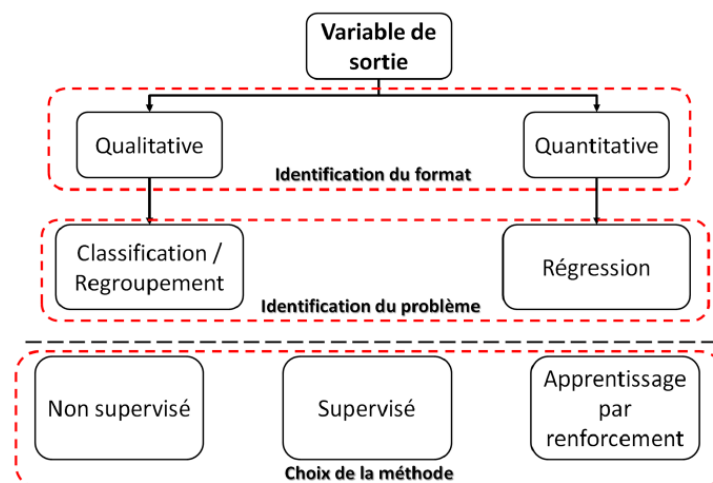


Figure 3.2. Étapes d'identification de la méthode d'apprentissage automatique. [19]

Dans notre étude nous avons une sortie sous forme d'une prédiction de valeur (valeur de CBR) donc notre parcours est dans cette forme :



Figure 3.3. Parcours de choix de méthode d'apprentissage automatique.

Vous pouvez prendre plus d'informations sur Machine Learning d'après les articles spécialisés de ces références : [16, 20, 21, 22 et 23].

IV.5. La régression : La régression est une technique d'apprentissage supervisé utilisée pour prédire des valeurs numériques en se basant sur des données d'entraînement. Nous pouvons effectuer une régression linéaire dans différentes formes, en fonction de la nature de la relation. Si la régression est linéaire, nous pouvons utiliser soit la régression linéaire simple, soit la régression linéaire multiple.

IV.5.1. Régression linéaire : La régression linéaire cherche à établir une relation linéaire entre les variables d'entrée et la variable cible. L'objectif est de trouver la meilleure ligne (ou hyperplan dans un espace de dimension supérieure) qui minimise l'erreur entre les valeurs prédites et les valeurs réelles.

La régression linéaire est largement utilisée en statistiques et en apprentissage automatique pour modéliser des relations simples entre les variables.

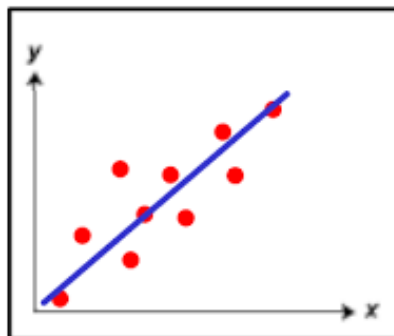


Figure 3.4. Régression linéaire.

IV.5.1.1. Régression simple SLR : La régression simple est utilisée lorsque nous voulons prédire une variable cible en fonction d'une seule variable d'entrée. Par exemple, prédire le prix d'une maison en fonction de sa surface habitable.

Dans ce cas, la relation entre la variable cible (le prix) et d'entrée (la surface habitable) est modélisée à l'aide d'une équation linéaire simple, telle que :

$$y = \beta_0 + \beta_1 x \quad \dots (3.2) \text{ où :}$$

y est la variable cible, x est la variable d'entrée, β_0 est l'ordonnée à l'origine et β_1 est la pente de la droite de régression.

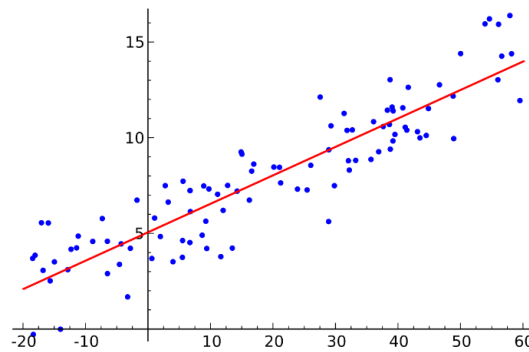


Figure 3.5. Exemple de régression linéaire simple SLR. [24]

IV.5.1.2. Régression multiple MLR : La régression multiple est utilisée lorsque nous voulons prédire une variable cible en fonction de plusieurs variables d'entrée. Par exemple, prédire le rendement d'une culture en fonction de la quantité de pluie, de l'ensoleillement et de la température.

Dans ce cas, la relation entre la variable cible et les variables d'entrée est modélisée à l'aide d'une équation linéaire multiple, telle que :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n \dots (3.3) \text{ où :}$$

y est la variable cible, $x_1, \dots, x_n$ sont les variables d'entrée, β_0 est l'ordonnée à l'origine et $\beta_1, \dots, \beta_n$ sont les coefficients associés à chaque variable d'entrée. [25] [26]

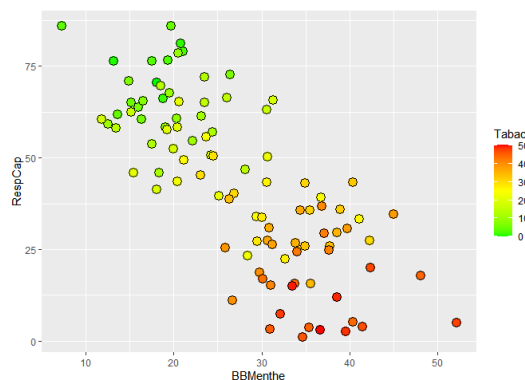


Figure 3.6. Exemple de régression linéaire multiple MLR. [27]

La figure illustre un exemple de prédiction de l'indice de plasticité (PI) à l'aide de modèle MLR. Dans cet exemple, les variables d'entrée sont représentées sur l'axe des abscisses (X), tandis que la variable cible ou de sortie (PI) est représentée sur l'axe des ordonnées (Y). Chaque point correspond à une observation de la base de données, permettant de visualiser la relation entre les variables explicatives et la variable à prédire ainsi que la tendance générale capturée par le modèle MLR.

En résumé, la régression simple est utilisée avec une seule variable d'entrée, tandis que la régression multiple est utilisée avec plusieurs variables d'entrée pour prédire une variable cible.

On prend ce type d'apprentissage (MLR) dans notre étude comme un modèle conventionnel (par défaut c.-à-d. non optimisé).

IV.5.1.3. L'hyper-paramètre de la régression linéaire : Le seul hyper-paramètre commun entre la régression simple SLR et multiple MLR est la force de régularisation α (il est parfois utilisé).

α : (*Ridge* (α) / *Lasso* (λ)) $\rightarrow$ réduit sur-apprentissage et multi-colinéarité.

Il y a des plusieurs recherches qui applique ce type d'apprentissage pour développer des modèle prédictifs dans le domaine de génie civil, voir ces références : [26, 28, 29, 30 et 31].

IV.5.2. Régression non linéaire : Contrairement à la régression linéaire, la régression non linéaire permet de modéliser des relations plus complexes entre les variables d'entrée et la variable cible. Elle peut être réalisée à l'aide de techniques telles que les polynômes, les fonctions exponentielles, les fonctions logarithmiques, ou en utilisant des modèles plus avancés tels que les réseaux de neurones ou les méthodes d'interpolation.

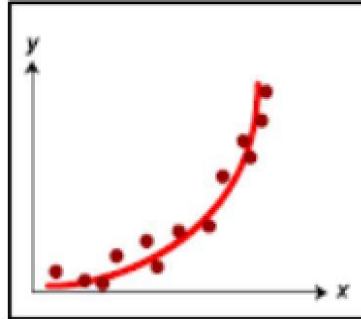


Figure 3.7. Régression non linéaire.

IV.6. Théories de modèles :

IV.6.1. Réseaux de Neurones Artificiel ANN (Artificial Neural Network) : Le modèle ANN (Artificial Neural Network), également connu sous le nom de réseau de neurones artificiels, est un type de modèle d'apprentissage automatique basé sur le fonctionnement du cerveau humain, il est constitué de plusieurs couches de neurones interconnectés (entrées, cachée et les sorties), chacune effectuant des calculs mathématiques pour produire une sortie.

Les réseaux de neurones ont été développés dans les années 1950 et 1960, mais ils ont connu une popularité croissante ces dernières années en raison des avancées dans la puissance de calcul et des données massives disponibles pour l'entraînement des modèles.

Les réseaux de neurones peuvent être utilisés pour une variété de tâches, notamment la classification d'images, la reconnaissance de la parole, la prédiction et la détection d'anomalies ...etc., ils sont également utilisés dans des domaines tels que la finance, la médecine et l'industrie manufacturière ...etc.

Les réseaux de neurones sont entraînés à l'aide d'un ensemble de données d'entraînement, où les entrées sont présentées au réseau et les sorties attendues sont fournies, le réseau ajuste ses poids et biais pour minimiser l'erreur entre les sorties produites et les sorties attendues, une fois entraîné, le réseau peut être utilisé pour prédire des sorties pour de nouvelles entrées.

Les réseaux de neurones peuvent être très complexes, avec des milliers ou des millions de neurones et de connexions, cependant, des techniques telles que la régularisation et le dropout peuvent être utilisées pour éviter le sur-apprentissage (over-fitting) et améliorer la généralisation du modèle. [32]

IV.6.1.1. Lecture biologique sur le modèle d'ANN : Les réseaux de neurones sont des modèles informatiques inspirés par le fonctionnement du cerveau humain, en biologie, les neurones sont des cellules spécialisées dans la transmission des signaux électriques et chimiques, ces cellules sont reliées entre elles par des connexions appelées synapses.

De manière similaire, les réseaux de neurones artificiels sont composés de "neurones" artificiels reliés entre eux par des connexions pondérées, ces connexions permettent le transfert d'informations sous forme de signaux électriques simulés.

En biologie, les neurones communiquent entre eux en transmettant des signaux électriques le long de leurs axones, qui sont ensuite convertis en signaux chimiques au niveau des synapses, dans un réseau de neurones artificiels, les informations sont transmises à travers les connexions pondérées

entre les neurones, et ces connexions sont ajustées pour optimiser la capacité du réseau à apprendre à partir des données.

De plus, en biologie, les réseaux de neurones dans le cerveau humain sont capables d'apprendre et de s'adapter en modifiant la force des connexions synaptiques, de manière similaire, les réseaux de neurones artificiels utilisent des algorithmes d'apprentissage pour ajuster les poids des connexions entre les neurones en fonction des données d'entrée et de sortie.

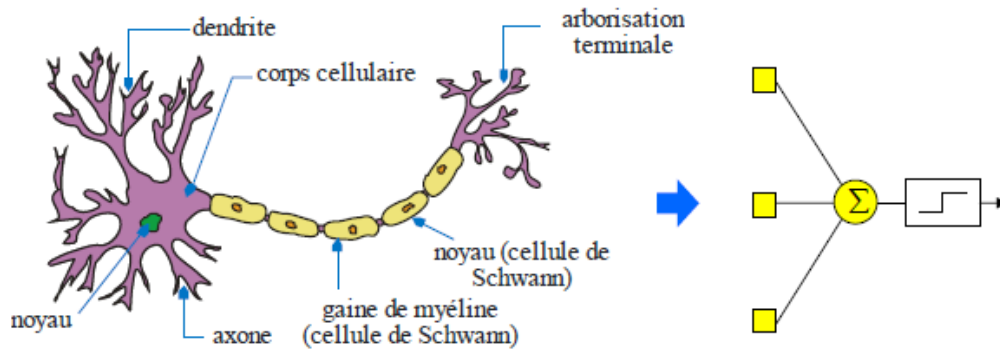


Figure 3.8. Simulation entre la structure biologique d'un neurone et ANN. [33]

IV.6.1.2. Bref historique sur le développement de modèle ANN :

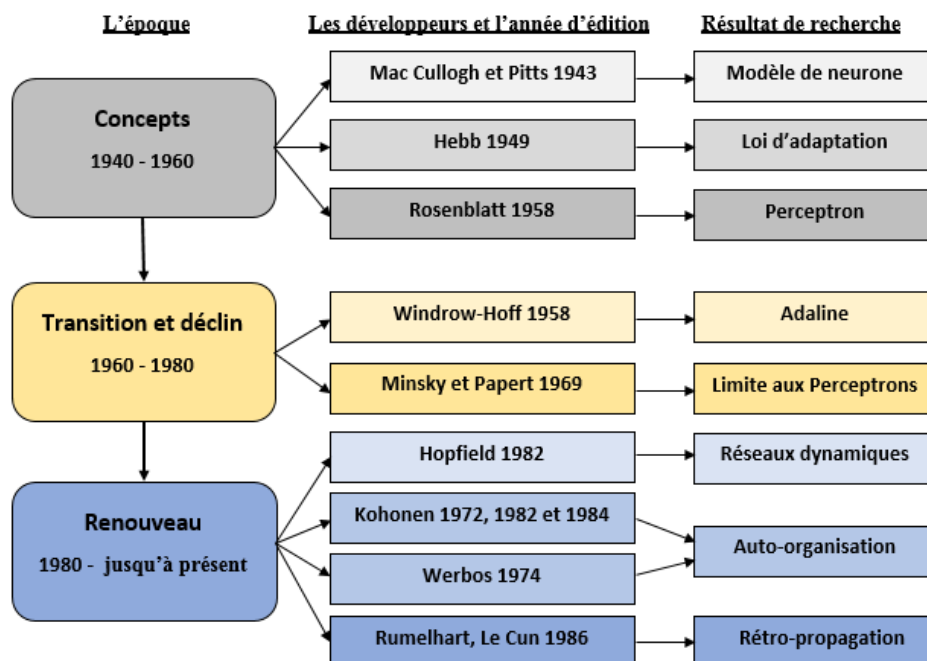


Figure 3.9. Bref historique sur les développements scientifiques de modèle ANN.

IV.6.1.3. Neurone formel : Les neurones formels, également connus sous le nom de "neurones logiques" ou "neurones à seuil", sont des modèles de base de l'intelligence artificielle qui ont été développés dans les années 1940 et 1950, ces modèles ont jeté les bases de la théorie des réseaux de neurones artificiels et ont contribué de manière significative au développement initial de l'apprentissage automatique.

Un neurone formel est un modèle mathématique simplifié d'un neurone biologique, il prend en entrée un vecteur de données pondérées par des poids, puis applique une fonction d'activation pour produire une sortie, la fonction d'activation (3.4) compare la somme pondérée des entrées à un seuil prédéfini, et si cette somme dépasse le seuil, le neurone produit une sortie; sinon, il ne produit pas de sortie.

$$y = f(w_0 + x_1w_1 + x_2w_2 + \dots + x_nw_n) = f(w_0 + w^T x) \dots (3.4)$$

Correction de poids synaptiques :

$$w_{k+1} = w_k + \Delta w_k \dots (3.5)$$

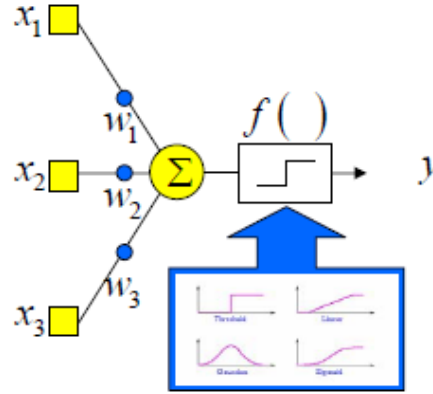


Figure 3.10. Exemple de trois entrées de neurone formel. [33]

Le neurone formel a été développé pour modéliser le comportement des neurones biologiques dans le cerveau, en particulier en ce qui concerne la transmission et le traitement des signaux électriques, il a été largement utilisé pour développer des modèles simples de classification et de prédiction, et a posé les bases pour le développement ultérieur de réseaux de neurones plus complexes. Cependant, les neurones formels présentent des limitations importantes, notamment leur incapacité à modéliser des relations non linéaires complexes, cela a conduit au développement de modèles plus avancés tels que le **perceptron** et les réseaux de neurones multicouches (**perceptron multicouches MLP**), qui ont permis de surmonter certaines de ces limitations et ont ouvert la voie à des applications plus puissantes de l'intelligence artificielle. [33]

IV.6.1.3.1. Perceptron : Le perceptron a été proposé par Frank Rosenblatt en 1957. C'est un modèle simple de réseau de neurones composé d'une seule couche de neurones, avec des connexions pondérées entre les entrées et la sortie. Le perceptron est capable d'apprendre à classer des données linéairement séparables en ajustant les poids des connexions en fonction des erreurs de prédiction.

Le perceptron prend en entrée n données binaires, chacune de ses entrées i est pondérée par un poids noté w_i , le neurone peut prendre les états "1" ou "0" (respectivement actif ou non-actif) en fonction de ses entrées pondérées et d'un biais noté ($\beta \in \mathbb{R}$), cet état représente la sortie du modèle, il est donc possible de représenter le perceptron comme une fonction paramétrique.

$$f_{\theta} : \{0, 1\}^n \rightarrow \{0, 1\} \dots (3.6) \text{ avec :}$$

$$\theta \text{ est l'ensemble de ses paramètres, c'est-à-dire le biais } \beta, \text{ et les poids } w = (w_1, \dots, w_n) \dots (3.7)$$

La sortie d'un perceptron pour un vecteur $x \in \{0, 1\}^n$ en entrée est calculée tel que :

$$f(x, w) = H(z(x, w) + \beta) \dots (3.8) \text{ avec :}$$

$H(t)$ la fonction de Heaviside définie pour tout $(t \in \mathbb{R})$ comme $H(t) = 1_{\{t > 0\}}$ et $z(x, w)$ la somme pondérée des entrées :

$$Z(x, w) = w^T x = \sum_{i=1}^n w_i x_i \dots (3.9)$$

Intuitivement, les poids $(w_1, \dots, w_n)$ représentent l'importance accordée à chaque entrée pour l'activation du perceptron, pour rappel, 1_A est la fonction indicatrice qui est égale à 1 si la condition A est vérifiée et 0 sinon.

Le biais β peut être vu comme l'ajout d'un seuil à la difficulté d'activation du perceptron. En effet, si la somme pondérée $z(x, w)$ dépasse $-\beta$ (l'opposé du biais), le perceptron s'active, sinon il reste inactif.

Pour simplifier les notations, nous allons inclure le biais β dans le vecteur de poids en ajoutant une constante en entrée $x_0 = 1$ (le biais devient donc w_0). Nous obtenons la fonction paramétrique.

$$f_\theta : \{0, 1\}^{n+1} \times \mathbb{R}^{n+1} \rightarrow \{0, 1\} \quad \dots (3.10) \text{ telle que :}$$

$$f(x, w) = H(z(x, w)) = H\left(\sum_{i=0}^n w_i x_i\right) \quad \dots (3.11) [34]$$

IV.6.1.3.2. Perceptron multicouches MLP (multi-layer perceptron) : Est une variante du réseau de neurones artificiels qui utilise des couches de neurones entièrement connectées, dans un MLP, les entrées sont présentées à la première couche de neurones, qui effectue des calculs mathématiques pour produire une sortie, cette sortie est ensuite transmise à la couche suivante, et ainsi de suite jusqu'à ce que la sortie finale soit produite comme présente dans figure 3.11.

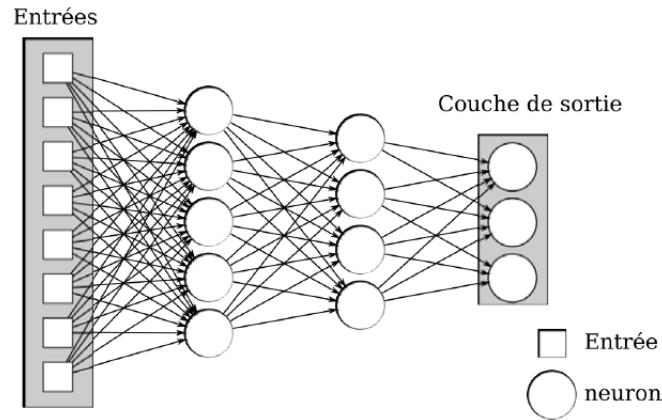


Figure 3.11. Exemple d'une représentation d'un MLP. [34]

Comme montré sur la figure 3.11, chaque neurone de la première couche prend en entrée, l'entrée du MLP, chaque couche suivante reçoit en entrée les valeurs d'activation de la couche précédente (c'est-à-dire, le vecteur contenant les valeurs de chaque neurone de la couche précédente).

La sortie du MLP est composée de la valeur d'activation de chaque neurone de la dernière couche, appelée couche de sortie, le vecteur d'activation de la couche l , composé de s neurones peut être calculé suivant le vecteur d'entrées $e \in \mathbb{R}^p$ (Les entrées du MLP ou le vecteur d'activation de la couche précédente), de la manière suivante :

$$a^{(l)} = f^{(l)}(e W^{(l)}, \beta^{(l)}) = \phi^{(l)}(W^{(l)} e + \beta^{(l)}) \quad \dots (3.12) \text{ avec :}$$

$W = (w_1, \dots, w_s)^T \in \mathbb{R}^{s \times p}$ et $\beta^{(l)} = (\beta_1, \dots, \beta_s)^T \in \mathbb{R}^s$ avec respectivement w_i , le vecteur des poids du neurone i et β_i son biais. La fonction $\phi^{(l)}$ est une fonction d'activation appliquée individuellement à chaque neurone de la couche l , dans la section précédente, nous avons vu la fonction de Heaviside et la fonction sigmoïde, dans un MLP, tous les neurones d'une même couche ont la même fonction d'activation $\phi^{(l)}$. [34]

IV.6.1.3.3. Correction des erreurs : Systématiquement, s'appuyer sur quelque chose de probabiliste pour estimer une sortie qui peut comporter des erreurs dans la collecte des probabilités, c'est ce qui a poussé les fondateurs de la théorie et des modèles initiaux des réseaux neuronaux à inventer des solutions plus appropriées dans leur temps, ce qui a nécessité le développement des techniques comme la descente de **gradient** et la méthode de **rétro-propagation**.

La descente de gradient et la rétro-propagation sont deux concepts distincts, mais ils sont souvent utilisés ensemble dans le contexte de l'apprentissage automatique et des réseaux de neurones.

* **La descente de gradient :** En résumé, la descente de gradient est utilisée pour mettre à jour les paramètres du modèle dans la direction opposée du gradient de la fonction de coût.

La descente de gradient a été développée pour la première fois par Cauchy en 1847, cependant, la version moderne de l'algorithme de descente de gradient utilisée en apprentissage automatique et en intelligence artificielle a été popularisée par plusieurs chercheurs au fil du temps, les travaux de Richard Bellman, David E. Rumelhart, Geoff Hinton, Yann Le Cun, Yoshua Bengio et d'autres ont contribué à son développement et à son application dans ces domaines.

La descente de gradient est un algorithme d'optimisation utilisé pour minimiser une fonction de coût en ajustant itérativement les poids du réseau, l'idée est de calculer le gradient de la fonction de coût par rapport aux poids, puis de déplacer les poids dans la direction opposée du gradient pour atteindre un minimum local ou global, cela se fait en utilisant la règle de mise à jour des poids, qui est généralement de la forme :

$$W_{new} = W_{old} - learning_rate \times gradient \quad \dots (3.13) \text{ où :}$$

learning_rate est un hyper-paramètre qui contrôle la taille des pas effectués dans la direction du gradient.

Il existe différentes variantes de la descente de gradient, notamment :

- La descente de gradient stochastique.
- La descente de gradient par mini-lots.
- La descente de gradient par lots.

L'équation de la descente de gradient (fonction d'erreur) est donnée par :

$$E = \frac{1}{2} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \dots (3.14)$$

Ajustement des poids :

$$\nabla W_j = -\varepsilon \frac{\partial E}{\partial W_j} = -\varepsilon \sum_{i=1}^n (y_i - \hat{y}_i) x_{ij} \quad \dots (3.15) \text{ [35]}$$

* **La méthode de Rétro-propagation :** En résumé, La rétro-propagation ou "back-propagation" en anglais est une méthode initialement développée par Rumelhart et Le Cun en 1986, en moderne est un algorithme spécifique qui permet de calculer efficacement le gradient de la fonction de coût par rapport aux poids du réseau, elle repose sur l'application de la règle de la chaîne pour propager l'erreur à travers le réseau, en commençant par la couche de sortie et en remontant vers les couches cachées ainsi que la couche d'entrées, cela permet de calculer les gradients partiels par rapport aux poids à chaque couche du réseau, ce qui est essentiel pour la mise à jour des poids lors de la descente de gradient.

- **Rétro-propagation de l'erreur (forme de Hebb 1949) :** Il s'agit de l'algorithme du gradient stochastique, pas de fonction d'erreur dans les couches cachées donc la rétro-propagation d'erreur selon Hebb est donnée par :

$$\forall w_{ki} = -\varepsilon \frac{\partial E}{\partial w_{ki}} = \varepsilon \delta_i x_k \dots (3.16) \text{ avec :}$$

E est l'erreur instantanée.

$\delta_i = (y_i - \hat{y}_i) \sigma'(u_i)$, pour les cellules de sortie.

$\delta_i = (\sum_{j=1}^J \delta_j w_{ij}) \sigma'(u_i)$, pour les cellules cachées. [35]

- **Rétro-propagation de l'erreur dans MLP :** On présente un vecteur d'entrée pour lequel on détermine la sortie du réseau. L'ensemble des poids des liaisons synaptiques détermine le fonctionnement du réseau de neurones. On compare les sorties des neurones de la couche de sortie avec les valeurs modèles (= sorties désirées) et on calcule l'erreur de chacun.

L'erreur d'un neurone k dans une couche ' j ' est calculée à partir des erreurs des neurones de la couche ' $j+1$ ' pondérées par les poids :

$$\varepsilon_k^{(j)} = f(\sum_{i=1}^J x_i w_{ij}) \cdot \sum_{i \in j+1} \varepsilon_i^{j+1} \cdot w_{ik}^{j+1} \dots (3.17)$$

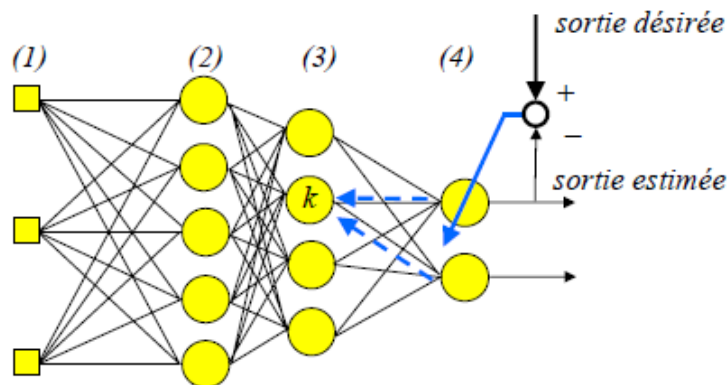


Figure 3.12. Forme explicative de rétro-propagation de l'erreur dans MLP. [33]

IV.6.1.4. Les hyper-paramètres de modèle de MLP d'ANN :

a) Nombre des couches cachées (hidden layers) : Ajoutent de la profondeur → captent plus de complexité.

b) Nombre des neurones dans les couches : Plus de neurones = plus de capacité mais risque sur-apprentissage.

c) La vitesse d'apprentissage : $learning_{rate}$ → trop grand = instable, trop petit = lent.

d) La taille des mini-lots : Batch size pour mise à jour des poids.

e) Fonction d'activation : Ajoute de la non-linéarité (ReLU rapide, sigmoid lisse).

f) Nombre d'itérations : nb de passes sur tout le dataset.

g) Les poids élevés (weight decay) : Pénalise les poids élevés → améliore la généralisation.

IV.6.2. Les forêts aléatoires (RF) : Ou Random Forests en anglais, les forêts aléatoires (RF) sont des modèles d'apprentissage automatique qui combinent plusieurs arbres de décision pour effectuer des prédictions. Chaque arbre de décision est construit en utilisant un sous-ensemble aléatoire des données d'entraînement et des caractéristiques disponibles, lors de la prédiction, les résultats de chaque arbre sont combinés pour obtenir une prédiction finale plus robuste et précise.

IV.6.2.1. Bref historique sur le développement de modèle de RF : Le modèle de Random Forest a été introduit pour la première fois par Leo Breiman et Adele Cutler en 2001. Il s'agit d'une technique d'apprentissage automatique qui repose sur l'agrégation de multiples arbres de décision pour effectuer des prédictions, cette approche combine la puissance de la modélisation par arbre de décision avec la diversité et la robustesse apportées par l'agrégation. [36] Certains chercheurs qui ont contribué de manière significative au développement et à l'amélioration du modèle de Random Forest incluent Leo Breiman lui-même, ainsi que Tin Kam Ho, qui a travaillé sur des variantes du modèle, des chercheurs tels que Pierre Geurts, Gilles Louppe et Louis Wehenkel ont également apporté des contributions importantes à la compréhension théorique et à l'amélioration des performances du modèle.

IV.6.2.2. Arbre de décision DT : Ou Decision Tree en anglais, Les arbres de décision (DT) sont une méthode populaire en apprentissage automatique pour la modélisation et la prédiction de données, ils sont largement utilisés dans de nombreux domaines tels que la classification, la régression et le clustering. Les arbres de décision sont des modèles qui utilisent une structure arborescente pour représenter un ensemble de décisions à prendre en fonction des caractéristiques des données.

IV.6.2.3. Bref historique sur le développement de modèle de DT : Le concept des arbres de décision remonte aux années 1960 et 1970, avec des travaux pionniers réalisés par des chercheurs tels que Edward F. Moore et Robert C. Prim. Cependant, c'est en 1986 que le chercheur Ross Quinlan a développé l'algorithme ID3 (Itérative Dichotomiser 3) qui a marqué un tournant majeur dans l'histoire des arbres de décision. L'algorithme ID3 est un algorithme d'apprentissage automatique qui construit un arbre de décision à partir d'un ensemble de données d'entraînement en utilisant une approche de type "diviser pour régner". Ross Quinlan a également développé d'autres algorithmes d'arbres de décision tels que C4.5, qui est une extension de l'algorithme ID3 et qui a introduit des améliorations significatives, ces algorithmes ont grandement contribué à populariser l'utilisation des arbres de décision dans le domaine de l'apprentissage automatique. [37]

Depuis lors, de nombreux chercheurs ont continué à travailler sur les arbres de décision, développant de nouvelles variantes et techniques pour les rendre plus efficaces et adaptés à une grande variété de problèmes, aujourd'hui, les arbres de décision sont largement utilisés dans de nombreuses applications pratiques, et ils restent un sujet actif de recherche dans le domaine de l'apprentissage automatique.

IV.6.2.4. La structure de modèle de RF et (DT) : Les forêts aléatoires sont composées de multiples arbres de décision, où chaque arbre est construit de manière indépendante comme suivante :

* **Structure de modèle de DT** : Un DT est composé de nœuds, de branches, d'une racine et de feuilles.

1- Nœuds et branches : Un arbre de décision est composé de nœuds qui représentent des tests sur les caractéristiques des données, et de branches qui relient ces nœuds, chaque nœud teste une caractéristique spécifique des données et chaque branche représente le résultat de ce test.

2- Racine et feuilles : L'arbre commence par un nœud racine qui teste la caractéristique la plus importante, suivie de nœuds internes qui testent d'autres caractéristiques, et se termine par des feuilles qui contiennent les prédictions finales.

D'autre part, une forêt aléatoire est un ensemble d'arbres de décision qui travaillent ensemble pour effectuer des prédictions, chaque arbre de décision dans la forêt aléatoire est construit en utilisant un sous-ensemble aléatoire des données d'entraînement et un sous-ensemble aléatoire des caractéristiques. Lorsqu'il est temps de faire une prédiction, chaque arbre de décision de la forêt aléatoire donne sa propre prédiction, et le résultat final est déterminé par un vote majoritaire (pour le cas de classification), ou une moyenne des prédictions de tous les arbres (pour le cas de régression comme dans notre étude), on peut donner une forme exemplaire dans la figure 3.13 suivante :

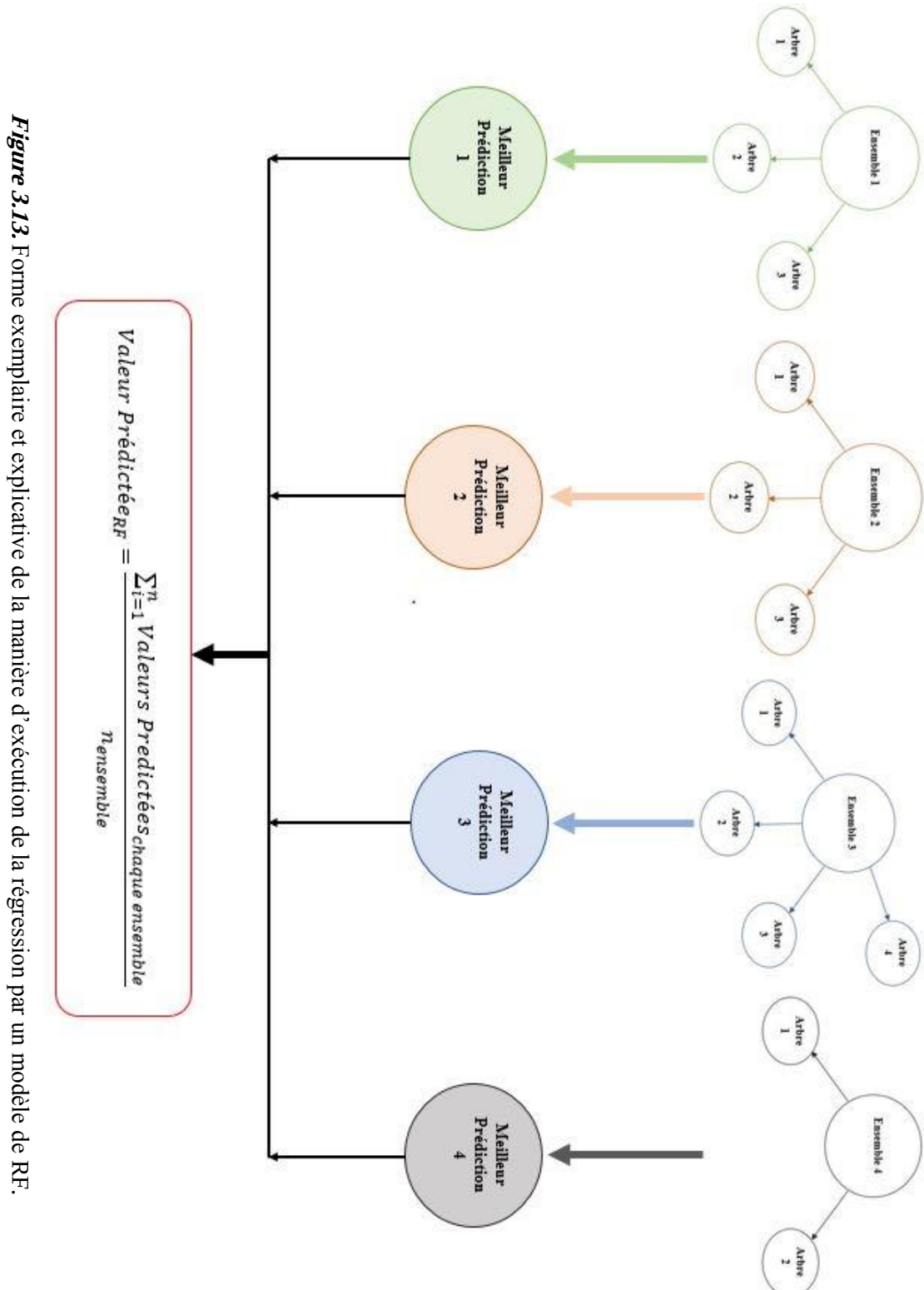


Figure 3.13. Forme exemplaire et explicative de la manière d'exécution de la régression par un modèle de RF.

IV.6.2.5. Processus de division des données : L'algorithme de construction de l'arbre de décision divise récursivement les données en sous-ensembles plus petits en fonction des caractéristiques, jusqu'à ce que les données soient suffisamment homogènes ou qu'une certaine condition d'arrêt soit atteinte.

*** Critère de division :** Pour chaque nœud, un critère est utilisé pour déterminer comment diviser les données en sous-ensembles. Dans la méthode de l'arbre de décision en apprentissage (régression), les données sont divisées de manière itérative en utilisant un critère appelé l'erreur quadratique moyenne (MSE), le MSE mesure la différence entre les valeurs prédites par l'arbre de décision et les valeurs réelles des données d'entraînement, l'objectif est de trouver les divisions qui minimisent cette erreur et permettent de prédire au mieux les valeurs de sortie.

L'algorithme de l'arbre de décision utilise différents critères, tels que la réduction de le MSE, l'**indice de Gini** ou l'**entropie**, pour choisir la meilleure division à chaque étape de construction de l'arbre, cela permet de créer un arbre de décision qui peut être utilisé pour prédire les valeurs de sortie pour de nouvelles données.

a) Critère de l'indice de Gini : L'indice de Gini mesure la probabilité qu'un élément choisi au hasard dans un ensemble soit incorrectement classé lorsqu'il est classé de manière aléatoire selon la distribution des étiquettes de classe dans cet ensemble, plus l'indice de Gini est faible, plus l'ensemble de données est pur. Pour calculer l'indice de Gini I_{Gini} , tu peux utiliser la formule suivante :

$$I_{Gini} = 1 - (p_1^2 + p_2^2 + \dots + p_n^2) \dots (3.18)$$

Dans cette formule, $p_1, p_2, \dots, p_n$ représentent les proportions des différentes classes dans l'ensemble de données, plus précisément, chaque p_i est le nombre d'éléments appartenant à la classe i divisé par le nombre total d'éléments dans l'ensemble de données.

b) Critère de l'entropie : L'entropie, quant à elle, mesure le désordre ou l'incertitude dans un ensemble de données, elle est calculée en utilisant la distribution des étiquettes de classe dans cet ensemble, plus l'entropie est élevée, plus l'ensemble de données est impur. Pour calculer l'entropie, tu peux utiliser la formule suivante :

$$Entropie = - [p_1 \times \log_2(p_1) + p_2 \times \log_2(p_2) + \dots + p_n \times \log_2(p_n)] \dots (3.19)$$

Dans cette formule, $p_1, p_2, \dots, p_n$ représentent également les proportions des différentes classes dans l'ensemble de données et le $\log_2(p)$ est le logarithme en base 2 de la proportion p .

Remarque : L'indice de Gini I_{Gini} et l'entropie sont tous deux des mesures de l'impureté des données, lorsque ces valeurs sont plus faibles, cela signifie que l'ensemble de données est plus pur et plus facile à diviser pour construire un arbre de décision précis.

IV.6.2.6. Les hyper-paramètres de modèle de RF et (DT) :

a) Nombre d'arbres (n_estimators) : Ce paramètre contrôle le nombre d'arbres de décision utilisés dans la forêt aléatoire. Un plus grand nombre d'arbres peut conduire à une meilleure généralisation, mais cela augmente également le temps de calcul.

b) Profondeur maximale de l'arbre (max_depth) : Ce paramètre limite la profondeur maximale des arbres de décision individuels dans la forêt, une profondeur plus grande peut conduire à un sur-

apprentissage (over-fitting), tandis qu'une profondeur trop petite peut conduire à un sous-apprentissage (under-fitting).

c) Nombre minimum d'échantillons requis pour diviser un nœud interne (min_samples_split) :

Ce paramètre contrôle le nombre minimum d'échantillons requis pour diviser un nœud interne de l'arbre, une valeur plus élevée peut aider à prévenir le sur-apprentissage en limitant la croissance de l'arbre.

d) Nombre minimum d'échantillons requis pour être une feuille (min_samples_leaf) :

Ce paramètre spécifie le nombre minimum d'échantillons requis pour former une feuille de l'arbre, une valeur plus élevée peut aider à lisser les prédictions et à réduire le sur-apprentissage.

e) Nombre maximal de caractéristiques à considérer pour diviser un nœud (max_features) :

Ce paramètre contrôle le nombre maximal de caractéristiques à considérer lors de la recherche de la meilleure division d'un nœud, il introduit de l'aléatoire dans la construction de chaque arbre, ce qui aide à réduire la corrélation entre les arbres et à améliorer la généralisation.

En ajustant ces paramètres, on peut contrôler la complexité des forêts aléatoires, éviter le sur-apprentissage et améliorer leurs performances prédictives.

f) La critère de division: L'indice de Gini I_{Gini} et l'entropie

Pour obtenir plus d'informations sur les modèles RF et DT, les références spécialisés sont disponible : [36, 37 et 38].

IV.6.3. Régression par Processus Gaussien GPR : Ou bien Gaussian Process Regression en anglais, le modèle GPR est une méthode d'apprentissage automatique utilisée pour modéliser des relations complexes entre des variables d'entrée et de sortie.

Le GPR repose sur le concept de processus gaussien, qui est une généralisation de la distribution normale à des fonctions aléatoires, en utilisant des processus gaussiens, le GPR peut capturer la structure de dépendance entre les données d'entrée et de sortie, tout en fournissant des estimations de l'incertitude associée à ces prédictions.

IV.6.3.1. Bref historique sur le développement de modèle GPR : Pendant les années 1980 et 1990, les recherches sur le modèle GPR ont connu un développement significatif, l'approche probabiliste de la régression gaussienne a été explorée en détail par des chercheurs tels que Christopher K. I. Williams et David Barber, ils ont travaillé sur l'élaboration des fondements théoriques du modèle GPR, en mettant l'accent sur ses capacités de prédiction et sa capacité à gérer l'incertitude.

Leurs travaux ont contribué à établir le modèle GPR comme un outil puissant pour la modélisation statistique, l'apprentissage automatique et la prédiction dans un large éventail de domaines d'application.

Les années 1980 et 1990 ont ainsi été marquées par des avancées significatives dans la compréhension et l'utilisation du modèle GPR, jetant les bases de son utilisation généralisée dans la recherche et l'industrie. [39]

IV.6.3.2. Processus Gaussien GP : Ou bien Gaussian Process en anglais, les processus gaussiens sont une méthode pour effectuer une régression probabiliste. À l'origine développés en géophysique [Matheron, 1963] et en statistiques [O'Hagan et Kingman, 1978], ils ont depuis acquis une forte popularité dans le domaine de l'apprentissage automatique. [40] [41]

Un processus gaussien (GP) est une collection de variables aléatoires telles que tout sous-ensemble fini d'entre elles a une distribution gaussienne conjointe (Rasmussen et Williams, 2008).

Soit : $\{Y(x); x \in D\}$ un GP. Alors, Y est entièrement défini par sa fonction moyenne μ et sa fonction de covariance k :

$$Y \sim \mathcal{GP}(\mu, k) \quad \dots (3.20) \text{ où}$$

$$\mu(x) = E\{Y(x)\} \text{ et } k(x, x_0) = cov\{Y(x), Y(x_0)\} \dots (3.21) \text{ pour } (x, x_0) \in D.$$

L'opérateur E représente l'espérance des variables aléatoires, et l'opérateur de covariance est donné par :

$$cov\{Y(x), Y(x_0)\} = E\{[Y(x) - \mu(x)][Y(x_0) - \mu(x_0)]\} \quad \dots (3.22)$$

Pour faciliter la notation, nous nous concentrons sur les hypothèses a priori des processus gaussiens centrés, c'est-à-dire que Y a une fonction moyenne μ égale à zéro, pour un nombre fini donné de points : $(x_1, ..., x_n)$, le vecteur avec les variables aléatoires : $Y_n = [Y(x_1), ..., Y(x_n)]$ satisfait.

$$Y_n \sim N(0, K) \quad \dots (3.23), \text{ avec :}$$

K une matrice de covariance : $K = (cov\{Y(x_i), Y(x_j)\}), (1 \leq i, j \leq n)$. [42]

En raison de leur facilité d'analyse, il est pratique de modéliser une collection finie de fonctions réelles à l'aide de distributions gaussiennes multi-variées. En pratique, on peut considérer un processus gaussien comme un vecteur gaussien multi-varié très long indexé par un espace d'indices (par exemple, le temps, l'espace, l'hyperespace...). Les données d'entraînement seront certaines dimensions de ce vecteur que l'on a déjà observées, et l'on souhaite faire des prédictions sur les dimensions qui n'ont pas été observées.

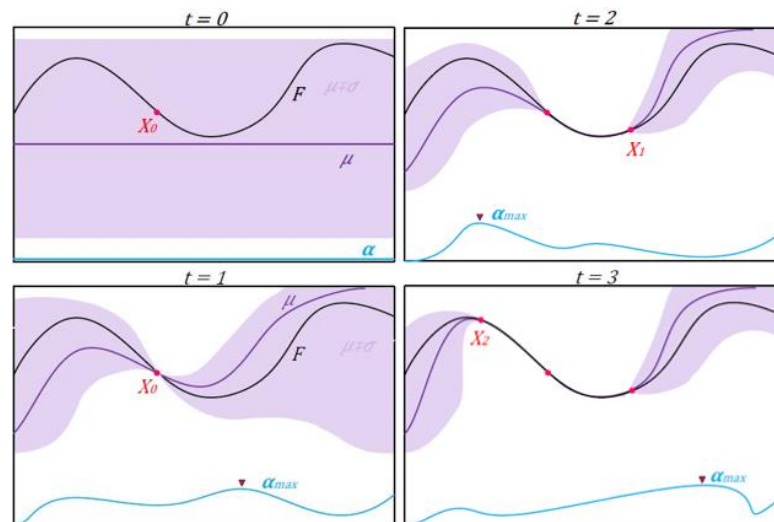


Figure 3.14. Exemple de changement entre le priori et le posteriori de Processus Gaussien en raison de change des paramètres entre deux phases. [43]

La ligne en gras et en noir représente la moyenne, qui dans notre exemple est constamment zéro, les lignes colorées sont des fonctions échantillonnées à partir de ce processus gaussien, la zone grise représente un écart-type par rapport à la moyenne à chaque point d'entrée et le panneau inférieur représente le posteriori, les points sur la ligne noire sont nos observations, les lignes colorées sont des fonctions échantillonnées à partir de la distribution posteriori.

La zone grise représente un écart-type par rapport à la moyenne, remarquez que lorsque nous avons nos observations, nous sommes plus certains de la valeur de la fonction autour des points observés, cela est illustré par la réduction de la zone grise dans le panneau inférieur.

Lorsque de nouveaux points de données arrivent, l'incertitude de la distribution posteriori diminue, ainsi, la prédiction devient de plus en plus confiante lorsque le modèle observe de plus en plus de données.

Pour les méthodes paramétriques traditionnelles, le modèle a tendance à sur-ajuster les données d'entraînement lorsque la taille de l'ensemble d'entraînement augmente, cependant, pour les processus gaussiens, étant donné qu'on ne suppose aucune forme spécifique de la fonction prédite, on n'a pas à se soucier du sur-ajustement (over-fitting) des données d'entraînement, il reste encore une certaine flexibilité (spécifiée par la variance de la distribution gaussienne posteriori), même si de nombreux points de données d'entraînement ont été observés. [42]

IV.6.3.3. Présentation théorique de l'algorithme de GPR : La Régression par Processus Gaussien (GPR) est un algorithme d'apprentissage automatique non paramétrique fondé sur les processus gaussiens, qui sont des généralisations de la distribution normale à des fonctions. Ce modèle est particulièrement adapté à des tâches de régression non linéaire, notamment lorsqu'on dispose de peu de données et que celles-ci sont de faible dimensionnalité. [44] [45]

Contrairement à des méthodes de régression classiques comme les régressions linéaires ou polynomiales qui nécessitent de définir explicitement la forme fonctionnelle reliant les entrées aux sorties, la GPR considère que l'espace des fonctions possibles est infini et explore cet espace à l'aide de distributions de probabilité définies par des processus gaussiens. [46]

Elle est donc très utilisée dans des domaines critiques où les données peuvent être rares, coûteuses à obtenir, ou bruitées, comme : la prédiction de l'état de santé des batteries lithium-ion [44], ou la prévision de la production d'énergie solaire [45], les systèmes physiques complexes nécessitant des estimations probabilistes robustes.

a) Fondements mathématiques du Processus Gaussien GP : Un processus gaussien (GP) est une généralisation de la loi normale multi-variée à un espace de fonctions. Il est défini comme un ensemble de variables aléatoires, dont toute sous-collection finie suit une loi normale multi-variée. Il est noté :

$$f(x) \sim GP[m(x), k(x, x')] \quad \dots (3.24) \text{ où:}$$

$x, x' \in \mathbb{R}^n$ sont des variables aléatoires.

$f(x)$ est une fonction aléatoire indexé par l'entré.

$m(x)$ est une fonction moyenne tel que:

$$m(x) = E[f(x)] \quad \dots (3.25)$$

$k(x, x')$ est la fonction de covariance ou fonction noyau (kernel), donnée par:

$$k(x, x') = E[(f(x) - m(x))(f(x') - m(x')))] \quad \dots (3.26)$$

Dans la majorité des applications pratiques, on suppose souvent que la fonction moyenne est nulle : $m(x) = 0$. [45]

b) Prise en compte du bruit dans les données : Les observations réelles ne sont pas directement égales à la fonction, mais incluent un bruit aléatoire, souvent modélisé par une variable gaussienne à moyenne nulle et variance:

$$y = f(x) + \varepsilon \quad \dots (3.27) \text{ avec :}$$

$\varepsilon \sim N(0, \sigma_n^2)$, par conséquent, la distribution a priori de y , avant d'observer les données est une loi normale multi-variée :

$$y \sim N(0, k(x, x') + \sigma_n^2 I_n) \dots (3.28) \text{ où :}$$

I_n est la matrice identité de dimension $n \times n$ et de nombre des points n , σ_n^2 représente la variance du bruit des observations.

c) Distribution conjointe a priori (Entraînement + Test) : Soit : X le jeu d'entraînement et X^* le jeu de test. La distribution conjointe a priori de (valeurs observées) et de (valeurs prédites) est :

$$\begin{bmatrix} y \\ y^* \end{bmatrix} \sim N \left(0, \begin{bmatrix} K(X, X) + \sigma_n^2 I_n & K(X, X^*) \\ K(X^*, X) & K(X^*, X^*) \end{bmatrix} \right) \dots (3.29) \text{ où:}$$

$K(X, X)$ est la matrice de covariance des entrées d'entraînement.

$K(X, X^*)$ et $K(X^*, X)$ sont les matrice de covariance entre les entrées d'entraînement et de test.

$K(X^*, X^*)$ est la matrice de covariance des entrées de test

d) Distribution a posteriori (Prédiction) : À partir des données observées, la distribution a posteriori de y^* conditionnée par les données d'entraînement, est :

$$y^* | X, y, X^* \sim N(\bar{y}^*, \text{cov}(y^*)) \dots (3.30) \text{ avec :}$$

$$\bar{y}^* = K(X^*, X) [K(X, X) + \sigma_n^2 I_n]^{-1} Y$$

y^* est la moyenne prédite et $\text{cov}(y^*)$ est la covariance prédite, tel que:

$$\text{cov}(y^*) = K(X^*, X^*) - K(X^*, X) [K(X, X) + \sigma_n^2 I_n]^{-1} K(X, X^*) \dots (3.31)$$

Ces expressions permettent de faire des prédictions probabilistes, i.e., obtenir une estimation de la moyenne et de l'incertitude associée à chaque prédiction. [47]

IV.6.3.4. Les noyaux de modèle de GPR : En anglais GPR kernels, les noyaux (Kernels) constituent la pierre angulaire de ces processus (gaussien), car ils définissent les propriétés de corrélation et de similarité entre les points dans l'espace d'entrée, influençant ainsi directement le comportement et la capacité du modèle à capturer des motifs complexes dans les données.

Le choix d'un noyau approprié dépend de la nature du problème et des hypothèses concernant la fonction à modéliser, et la disponibilité de différentes classifications de noyaux offre une grande flexibilité pour adapter le modèle à divers types de données. [42]

IV.6.3.4.1. Noyau Square Exponentiel (SE) / (RBF) : Le noyau de covariance Square Exponentiel (SE), également connu sous le nom de Radial Basis Function (RBF) ou noyau Gaussien, est l'une des fonctions les plus utilisées et les plus importantes dans le domaine des Processus Gaussiens (GP), et il est classé comme un noyau stationnaire.

Historiquement, ce noyau a fait l'objet d'études intensives et d'une utilisation généralisée dans des domaines tels que la géostatistique et la modélisation probabiliste, son application dans les modèles de régression gaussienne remontant au moins aux années 1980.

La formule précise de ce noyau exprime la relation entre deux points dans l'espace d'entrée, x et x' en fonction de la distance qui les sépare. Il peut être représenté par la formule suivante :

$$k(x, x') = \sigma^2 e^{\left(-\frac{|x-x'|^2}{2l^2} \right)} \dots (3.32)$$

En raison de ses propriétés de capture de la continuité et de la douceur, le noyau RBF est largement employé dans les applications nécessitant une modélisation précise de données présentant régularité et continuité, telles que :

- Modélisation du climat: évoluent continuellement dans le temps et l'espace.
- Prévision financière: pour modéliser des séries temporelles qui affichent souvent des tendances.
- Reconnaissance de formes: pour distinguer des caractéristiques de nature continue.

Dans le contexte des modèles GP, le noyau RBF influence directement la forme à la fois de la distribution a priori et de la distribution a posteriori. La douceur qu'il confère se reflète dans la nature des fonctions supposées générer les données et dans la distribution de probabilité de la fonction en de nouveaux points après observation des données.

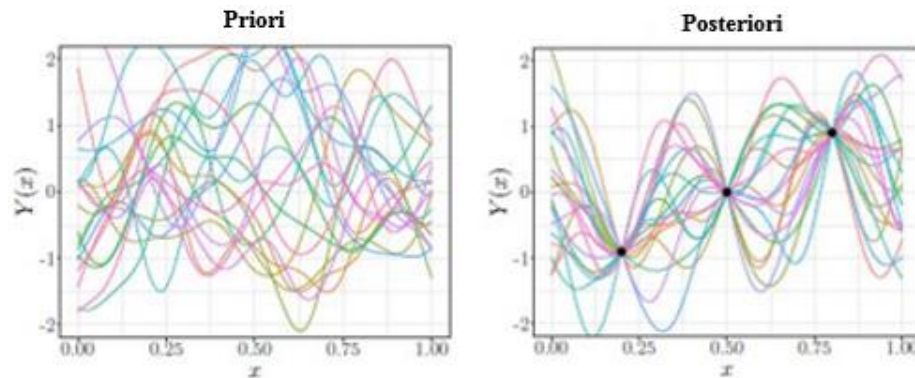


Figure 3.15. Le changement entre le priori et le posteriori du noyau SE (RBF).

IV.6.3.4.2. Noyau Matern : Le noyau Matern constitue une généralisation importante du noyau de la Radial Basis Function (RBF) ou Squared Exponential (SE). À l'instar du noyau RBF, le noyau Matern est classé comme un noyau stationnaire, ce qui signifie que ses propriétés ne dépendent pas de la position absolue des points mais de la distance relative entre eux. Cependant, le noyau Matern offre un avantage supplémentaire : la présence d'un paramètre de douceur additionnel (ν) qui permet un contrôle précis du niveau de douceur de la fonction résultante des Processus Gaussiens (GP). Historiquement, le noyau Matern a été développé par le statisticien suisse Matern dans les années 1960 et a depuis été largement adopté dans divers domaines, notamment la géostatistique, l'apprentissage automatique et la modélisation probabiliste.

Sur le plan des fondements mathématiques et de l'expression précise, la formule générale du noyau Matern repose sur la fonction Gamma (Γ) et la fonction de Bessel modifiée de seconde espèce (K_ν). Pour un noyau Matern avec le paramètre ν , il est exprimé mathématiquement comme suit :

$$k(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x-x'|}{l} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}|x-x'|}{l} \right) \dots (3.33) \text{ où:}$$

ν est le paramètre de douceur, qui détermine le degré de douceur de la fonction il prend généralement des valeurs de (1/2, 3/2 ou 5/2).

K_ν est la fonction de Bessel modifiée de seconde espèce.

Le noyau Matern est largement utilisé dans :

- La géostatistique : pour modéliser les distributions spatiales de ressources naturelles ou de phénomènes environnementaux.
- ML : comme fonction de covariance dans les processus gaussiens pour des applications de régression et de classification, en particulier lorsque le degré de douceur des données est incertain.
- La modélisation probabiliste : pour construire des modèles qui prennent en compte l'incertitude avec un niveau d'ordre spécifique.

Le noyau Matern est considéré comme un choix robuste et important dans la boîte à outils de modélisation des processus gaussiens. Il offre une généralisation du noyau RBF avec une capacité améliorée à contrôler la douceur de la fonction résultante, ce qui le rend adapté à un large éventail d'applications pratiques nécessitant un équilibre entre la flexibilité mathématique et la représentation réaliste des données.

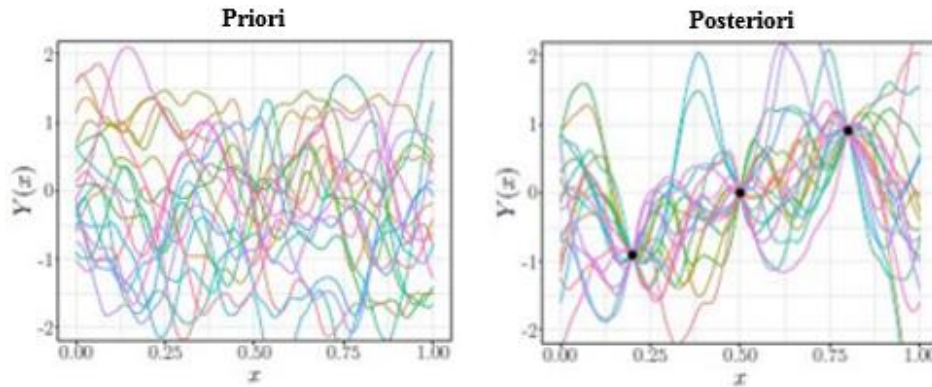


Figure 3.16. Le changement entre le priori et le posteriori du noyau Matern.

IV.6.3.4.3. Noyau Exponentiel : Ce noyau a également été étudié dans le contexte de la géostatistique et de l'apprentissage automatique, son utilisation remonte également à plusieurs décennies, son expression exact est donner par :

$$k(x, x') = \sigma^2 e^{-\frac{|x-x'|}{l}} \quad \dots (3.34)$$

Ce noyau est similaire au noyau RBF mais utilise une distance euclidienne exponentielle pour calculer la similarité entre les points, il est souvent utilisé pour modéliser des données où la décroissance de la similarité avec la distance est plus rapide que celle du noyau RBF, par conséquent, ce noyau peut être plus adapté pour des données où les variations locales sont importantes et où une décroissance rapide de la similarité est souhaitée.

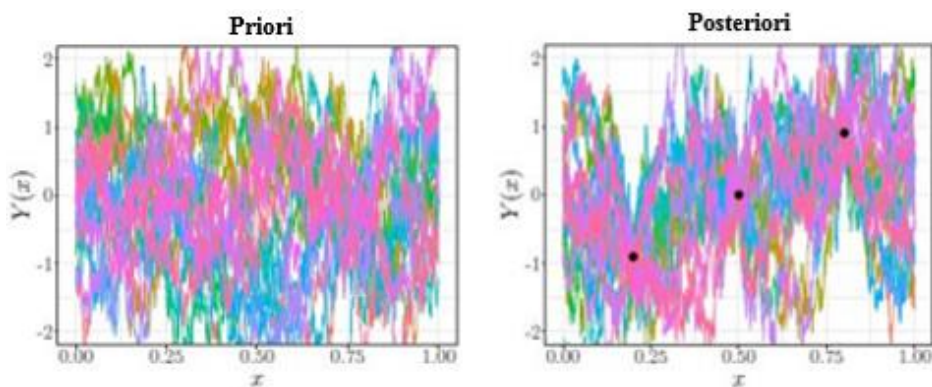


Figure 3.17. Le changement entre le priori et le posteriori du noyau Exponentiel.

IV.6.3.4.4. Noyau Exp-Square-Sine (Périodique) : Le noyau Exp-Square-Sine, également connu sous le nom de noyau périodique, est un outil puissant dans le domaine des Processus Gaussiens (GP) spécialement conçu pour modéliser des fonctions qui présentent un comportement périodique ou répétitif. Contrairement au noyau RBF qui suppose une continuité et une douceur absolues, ce noyau permet de capturer la variation et l'instabilité périodique dans les données. Le noyau Exp-Square-Sine s'appuie sur la fonction sinus pour capturer la répétition. Sa formule générale, en considérant que le support actuel est limité à la variante isotrope, est la suivante :

$$k(x, x') = e^{\left(\frac{-2 \sin^2 \left(\frac{\pi(x-x')}{p} \right)}{l^2} \right)} \quad \dots (3.35) \text{ où :}$$

p est le paramètre de périodicité, un hyper-paramètre positif ($p > 0$) qui détermine la longueur de la période ou la fréquence de la fonction. Plus la valeur de p est petite, plus la fonction ne se répète fréquemment.

$\sin^2$ est la fonction sinus au carré, qui garantit que la valeur du noyau est toujours comprise entre 0 et 1 et reflète la nature oscillante.

Le noyau Exp-Square-Sine se caractérise par les propriétés fondamentales suivantes :

- Périodicité: La caractéristique la plus remarquable est sa capacité à modéliser des dépendances qui se répètent à intervalles réguliers.
- Non-Stationnarité: Contrairement au noyau RBF, le noyau Exp-Square-Sine peut ne pas être entièrement stationnaire en soi, car la périodicité dépend de la position absolue et pas seulement de la distance. Cependant, il peut être combiné avec un autre noyau stationnaire (comme le RBF) pour créer un noyau composé qui réunit les caractéristiques des deux types.
- Douceur: Le degré de douceur des fonctions générées dépend de la relation entre l et p . Si l est petit par rapport à p , les oscillations sont aiguës. Si l est grand, les oscillations sont plus douces.

Le noyau Exp-Square-Sine est utilisé efficacement dans les scénarios nécessitant la modélisation de phénomènes périodiques, tels que :

- Analyse de signaux : Modélisation de signaux audio ou physiques récurrents.
- Prédiction financière : Pour capturer les cycles saisonniers sur les marchés.
- Robotique et planification : Modélisation de trajectoires de mouvement périodiques.

Le noyau Exp-Square-Sine offre une méthode efficace pour représenter les dépendances périodiques dans les données. Grâce à ses deux paramètres, longueur d'échelle et périodicité, ce noyau permet de modéliser un large éventail de phénomènes caractérisés par la répétition et la régularité, ce qui en fait un outil indispensable dans de nombreuses applications pratiques.

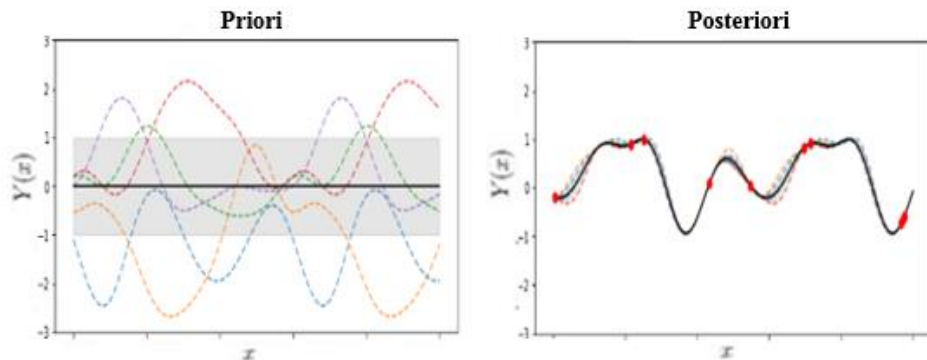


Figure 3.18. Le changement entre le priori et le posteriori du noyau Exp-Square-Sine (périodique).

IV.6.3.4.5. Noyau de covariance Rational Quadratique (RQ) : La fonction de covariance Rational Quadratique (RQ) est l'un des noyaux les plus utilisés dans les Gaussian processus pour modéliser la similarité entre deux points x et x' . Elle s'écrit comme suit:

$$k(x, x') = (1 + |x - x'|^2)^{-\alpha} \alpha \geq 0 \quad \dots (3.36)$$

Dans cette expression, α est un paramètre qui régule la forme et la flexibilité du noyau. Lorsque α est grand, la fonction se rapproche du noyau exponentiel quadratique (RBF), imposant un lissage

fort et régulier. À l'inverse, pour des valeurs petites de α , le modèle autorise des fluctuations rapides et plus complexes.

Ce noyau possède trois caractéristiques majeures :

- Flexible : il s'adapte aussi bien aux données lisses qu'aux données plus irrégulières, offrant une grande capacité de modélisation.
- Multi-échelle : il peut être interprété comme une superposition de plusieurs noyaux RBF avec différentes longueurs d'échelle, ce qui lui permet de capturer simultanément des tendances lentes (long terme) et des variations rapides (court terme).
- Probabiliste : dans le cadre des Gaussien processus, il ne fournit pas seulement une prédiction moyenne, mais également une estimation de l'incertitude (variance), ce qui est essentiel pour l'analyse bayésienne.

Et comme des exemples d'applications de ce noyau :

- Prédiction des prix sur les marchés financiers: les cours des actions ou des devises combinent des tendances lentes (long terme) et des fluctuations rapides (court terme). Ce noyau permet de modéliser ces deux niveaux de variation.
- Traitement d'images et de signaux: les signaux acoustiques ou les images contiennent à la fois des détails locaux (texture, tonalité) et une structure globale (forme, mélodie). Ce noyau modélise efficacement ces différents niveaux.
- Analyse de textes et traitement du langage naturel (NLP): dans les textes, on observe des structures globales (thème, sujet) ainsi que des variations locales (choix lexical, syntaxe). Le noyau RQ capture ces deux échelles, améliorant la classification et la traduction. [48] [49] [50] [51]

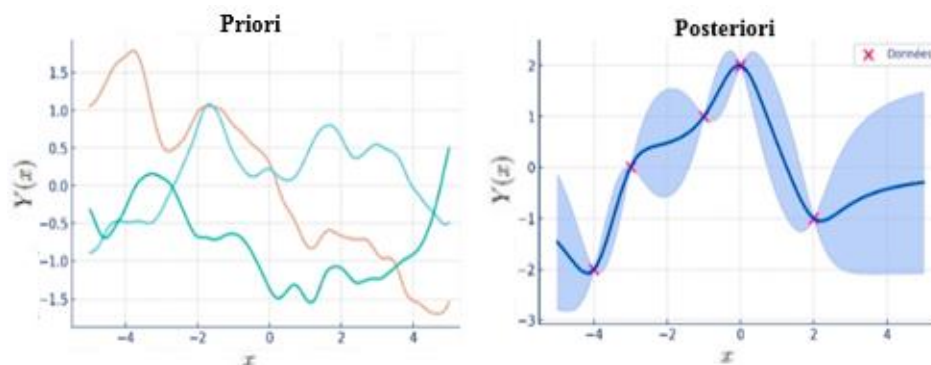


Figure 3.19. Le changement entre le priori et le posteriori du noyau Rational Quadratique (RQ).

IV.6.3.4.6. Noyau Linéaire / Dot-Product (Quadratique) : La fonction Dot-Product est un noyau non stationnaire, dérivé de la régression linéaire. Il s'obtient en posant des priors (lois a priori) sur les coefficients du modèle et un autre prior sur le biais, ce noyau est invariant aux rotations autour de l'origine, mais pas aux translations, il est paramétré par un coefficient (σ^2 , ou variance).

- Lorsque le paramètre de biais (c) = 0 $\rightarrow$ on obtient le noyau linéaire homogène.
- Lorsque le biais (c) $\neq$ 0 $\rightarrow$ il devient inhomogène (quadratique). La forme générale est :

$$k(x, x') = \sigma^2(x \cdot x' + c) \quad \dots (3.37) \text{ où } c \text{ est le biais.}$$

On combine souvent ce noyau avec une exponentiation (ex. puissance 2) pour modéliser des relations plus complexes. Les caractéristiques principales de ce noyau sont :

- Type : non stationnaire.
- Invariance : rotations mais pas translations.
- Lien : équivalent à une régression linéaire bayésienne avec priors.
- Simplicité : adapté aux données linéaires ou faiblement non-linéaires.

Et comme des exemples de ces utilisations :

- Analyse économique : prédire une variable (ex. revenu) en fonction de caractéristiques linéaires (âge, expérience, diplôme).
- Vision par ordinateur : classification basique d'images avec caractéristiques vectorielles (pixels moyens, histogrammes simples).
- Bio-informatique : relation approximativement linéaire entre expression génétique et un phénotype.

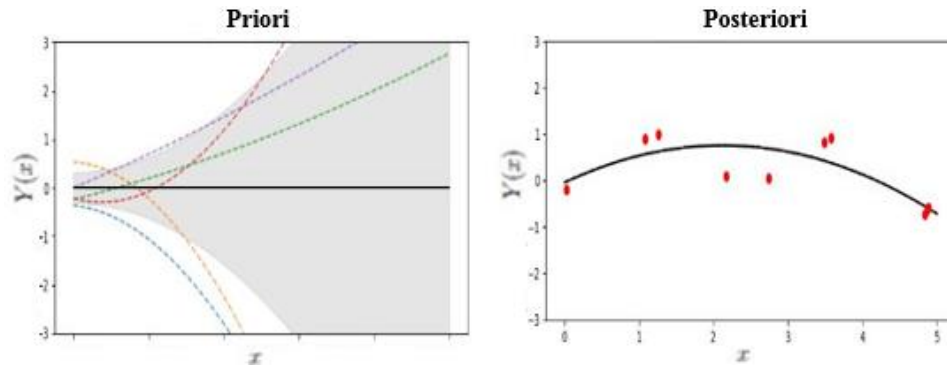


Figure 3.20. Le changement entre le priori et le posteriori du noyau Noyau Linéaire / Dot-Product. [50] [51]

Pour tous ces noyaux il nécessite indique que :

$x - x'$ représente la distance d euclidienne entre les points x et x' .

l est le paramètre de longueur d'échelle, est un hyper-paramètre positif.

$l > 0$ qui contrôle la "portée" de la similarité entre les points. Plus l est grand, plus les points sont considérés comme similaires sur de plus grandes distances, conduisant à une fonction de covariance qui décroît lentement. Inversement, une petite valeur de l implique que seuls les points adjacents sont considérés comme similaires.

σ^2 est le paramètre de variance qui détermine l'amplitude globale de la sortie attendue du processus gaussien. [48] [49]

IV.6.3.5. Résoudre du bruit dans les données (Le noyau de bruit blanc) : Le GPR est un modèle statistique utilisé pour estimer des fonctions inconnues à partir de données limitées. Elle permet à la fois : de prédire des valeurs, et d'estimer l'incertitude associée à ces prédictions.

La GPR repose sur l'utilisation de noyaux (kernels), et Autrement dit, le noyau décrit la structure ou la corrélation entre les points. Dans les données réelles, les observations ne sont jamais parfaites, elles sont presque toujours affectées par du bruit (erreur de mesure, perturbation, etc.). Il est donc nécessaire d'en tenir compte dans le modèle GPR.

- **Le noyau de bruit blanc (white kernel) :** Le noyau de bruit blanc est un noyau que l'on ajoute au noyau de base du modèle GPR pour modéliser le bruit présent dans les données. Sa représentation mathématique est :

$$k_{white}(x_i, x_j) = \sigma^2 \cdot \delta_{ij} \quad \dots (3.38) \text{ telles que:}$$

σ^2 : représente la variance du bruit, c'est-à-dire l'intensité du bruit supposé dans les données.

δ_{ij} : est la fonction delta de Kronecker, telle que :

$$\delta_{ij} = \begin{cases} 1, & \text{si } i = j \\ 0, & \text{si } i \neq j \end{cases} \quad \dots (3.39)$$

Cela signifie que cette fonction vaut 1 uniquement lorsque les points x_i et x_j sont exactement les mêmes, sinon elle vaut 0, pour cela en écrivant :

$$k_{\text{white}}(x_i, x_j) = \begin{cases} \sigma^2, & \text{si } x_i = x_j \\ 0, & \text{si } x_i \neq x_j \end{cases} \dots (3.40)$$

En d'autres termes, le white kernel n'ajoute du bruit que sur la diagonale de la matrice de covariance, car le bruit est supposé indépendant pour chaque observation individuelle.

- Pourquoi ajouter un white kernel dans un modèle GPR ?

Parce que dans les données réelles, les observations sont souvent bruitées. Le white kernel permet donc de :

- Représenter ce bruit aléatoire.
- Supposer que ce bruit est indépendant d'un point à l'autre.
- Quantifier son intensité par: σ^2

- **Forme complète du noyau dans GPR avec bruit :** Dans un modèle GPR, on combine souvent : un noyau principal (ex. : RBF, Matern, ...etc.) qui modélise la structure globale des données, et un white kernel qui modélise le bruit. La formule devient alors :

$$k(x_i, x_j) = k_{\text{base}}(x_i, x_j) + k_{\text{white}}(x_i, x_j) = k_{\text{base}}(x_i, x_j) + \sigma^2 \cdot \delta_{ij} \dots (3.41)$$

La matrice de covariance (ou kernel matrix) sera :

$$K_{(i,n)} = \begin{bmatrix} k(x_i, x_i) + \sigma^2 & \dots & k(x_i, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_i) & \dots & k(x_n, x_n) + \sigma^2 \end{bmatrix} \dots (3.42)$$

On voit que le bruit est ajouté uniquement sur la diagonale, pour représenter l'incertitude propre à chaque observation. [48] [49] [50]

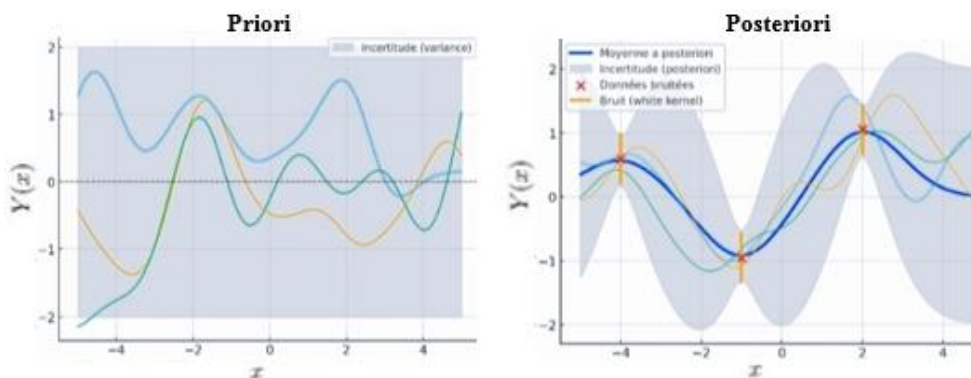


Figure 3.21. L'effet de white kernel sur les noyaux de GPR entre le priori et le posteriori.

Comme une interprétation sur la figure 3.21, et comment le noyau blanc (white kernel) résout ou bien indiquer le bruit dans les noyaux de GPR :

À gauche (a priori) : avant d'observer des données, le processus gaussien suppose que la fonction inconnue peut prendre de nombreuses formes possibles.

- Les courbes colorées représentent des fonctions tirées de la distribution a priori.
- La zone grise illustre une grande incertitude (variance élevée).

On remarque que les fonctions oscillent librement autour de la moyenne (ici ≈ 0).

À droite (a posteriori, avec bruit) : après avoir observé quelques points :

- Les points rouges sont les données bruitées.
- La courbe bleue correspond à la moyenne a posteriori (prédiction moyenne du GP).
- La zone grise est l'intervalle de confiance, qui devient étroit près des données et s'élargit loin des données.
- Les barres orange verticales mettent en évidence l'effet du white kernel : il ajoute une variance seulement sur la diagonale de la matrice de covariance, ce qui revient à dire que chaque observation possède un bruit indépendant (incertitude locale).

☞ Le contraste est donc :

- À priori : incertitude globale, sans contrainte.
- À posteriori : incertitude réduite grâce aux données, mais toujours du bruit local dû au white kernel.

IV.6.3.6. Les hyper-paramètres de modèle de GPR :

a) Longueur d'échelle l : Détermine à quelle distance deux points restent corrélés, $l > 0$ qui contrôle la "portée" de la similarité entre les points. Plus l est grand, plus les points sont considérés comme similaires sur de plus grandes distances, conduisant à une fonction de covariance qui décroît lentement. Inversement, une petite valeur de l implique que seuls les points adjacents sont considérés comme similaires.

b) Le paramètre de douceur ν : C'est qui détermine le degré de douceur de la fonction il prendre généralement des valeurs de (1/2, 3/2 ou 5/2). Il contrôle la régularité de la fonction (faible ν = plus rugueuse).

c) Le période p : Définit la fréquence de répétition des motifs (utile pour signaux périodiques).

d) Le paramètre de variance σ^2 : Détermine l'amplitude globale de la sortie attendue du processus gaussien.

d) Le paramètre de la régularité α : Mélange de plusieurs longueurs d'échelle pour gérer différentes variations.

d) Le paramètre de biais c : Si $c = 0 \rightarrow$ noyau linéaire homogène (strictement linéaire).
Si $c \neq 0 \rightarrow$ noyau inhomogène (ajout d'un terme constant $\rightarrow$ quadratique). [52] [53]

Certaines références peuvent être utiles concernant le modèle de GPR : le travail développé par [42], la bibliothèque Scikit-Learn (GPR) [50] et les articles [16, 54, 55, 56 et 57].

IV.6.4. Régression à Vecteurs de Support SVR : Ou bien Support Vector Regression en anglais, le modèle SVR est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression, il fonctionne en trouvant une frontière de décision optimale entre les différentes classes de données.

IV.6.4.1. Bref historique sur le développement de modèle de SVR : La Régression à vecteurs de support (SVR) ont été développées dans les années 1990 par les chercheurs Vladimir Vapnik et Alexey Chervonenkis, l'idée principale derrière les SVR est de trouver une frontière de décision optimale pour séparer les données en différentes classes.

Les SVR ont été initialement utilisées pour la classification binaire, où elles ont montré de très bons résultats en termes de précision et de généralisation, au fil du temps, les SVR ont été étendues pour prendre en charge la classification multi-classes et la régression.

Depuis leur développement initial, les SVR ont été largement utilisées dans de nombreux domaines tels que la vision par ordinateur, l'apprentissage automatique, la bio-informatique et bien d'autres, leur capacité à gérer des données non linéaires en utilisant des noyaux a également contribué à leur popularité.

Les SVR sont appréciées pour leur capacité à gérer des ensembles de données complexes, à résoudre des problèmes de classification et de régression, et à fournir des frontières de décision claires et interprétables.

IV.6.4.2. Présentation théorique de l'algorithme de SVR : L'algorithme de Support Vector Regression (SVR) est une méthode générale de classification et de transformation basée sur la théorie de l'apprentissage statistique. [58] Cette méthode a été étendue pour résoudre les problèmes de transformation en utilisant la fonction de perte "epsilon-insensitive" connue sous le nom de Support Vector Regression (SVR). SVR utilise le principe de la réduction du risque structurel (SRM) pour améliorer sa capacité de généralisation, même avec un nombre limité de données d'apprentissage.

Cette réduction vise à obtenir une fonction qui a une erreur inférieure à ε pour toutes les données réelles d'apprentissage. Comme il est impossible d'éviter les erreurs, l'objectif est de maintenir l'erreur dans une plage de valeurs spécifique (ε). La relation entre les variables peut être exprimée par les coefficients de l'équation (3.43). [59]

$$k(x) = \{w \cdot \phi(x)\} + b \quad \dots (3.43)$$

Dans l'équation (3.43) $x = (x_1, x_2, x_3, \dots, x_n)$ représente la valeur d'entrée, et y_i représente la valeur de sortie et $\phi(x)$ désigne une fonction non linéaire pour mapper les données d'entrée vers un espace de dimensions supérieures. De plus, w représente le vecteur de poids qui détermine la direction du plan de séparation ($w \in \mathbb{R}^n$), $b \in \mathbb{R}$ représente le seuil numérique associé au terme de biais, et n est la taille des données d'apprentissage. La fonction de perte "epsilon-insensitive" est définie comme suit : [58]

$$|y - k(x)|_\varepsilon = \max\{0, |y - k(x)| - \varepsilon\}, \varepsilon > 0 \quad \dots (3.44)$$

Dans l'équation (3.43), le niveau d'aplatissement dépend de la plus petite valeur de w . En pratique, il n'est pas possible pour la fonction de fournir une erreur inférieure à ε pour tous les points de données. Pour permettre une certaine marge d'erreur, les variables de relaxation ξ_i et ξ_i^* sont introduites. Ainsi, la fonction d'optimisation de SVR peut être écrite comme suit :

Minimiser :

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \quad \dots (3.45)$$

Sous réserve de :

$$\begin{cases} y_i - \{w \cdot \phi(x_i) + b\} \leq \varepsilon + \xi_i \\ \{w \cdot \phi(x_i) + b\} - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \quad \dots (3.46)$$

Où ξ_i et ξ_i^* sont les variables de relaxation, b est une constante, $\phi(x_i)$ est une fonction de mappage non linéaire, et C est le facteur de pénalité. Une valeur plus élevée de C indique que des pénalités plus élevées seront imposées aux échantillons présentant une erreur d'entraînement supérieure à ε . ε définit les exigences de la fonction de régression.

Dans l'équation (3.45), on peut exprimer la solution finale du problème d'optimisation en la transformant en une forme duale en utilisant la méthode des multiplicateurs de Lagrange qui peut être écrit comme :

$$-\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(x_i, x_j) - \varepsilon \sum_{i=1}^n (\alpha_i - \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) \quad \dots (3.47)$$

Sous réserve de :

$$\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0; \alpha_i, \alpha_i^* \in [0, C] \quad \dots (3.48) \text{ où :}$$

$k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ est définie comme une fonction de noyau.

Après avoir résolu l'équation (3.48) pour les valeurs α_i et α_i^* , on peut écrire la forme finale de l'équation (3.43) comme suit :

$$k(x) = \sum_{j=1}^n (\alpha_j - \alpha_j^*) k(x, x_j) + b \quad \dots (3.49)$$

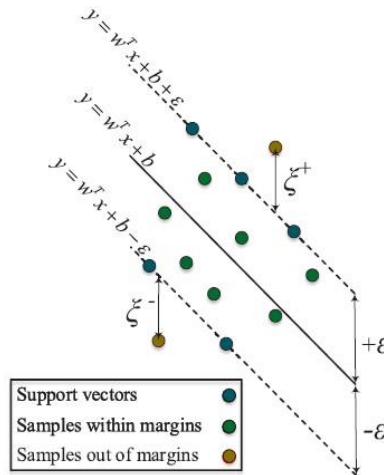


Figure 3.22. Représentation schématique de la régression à vecteurs de support SVR.

IV.6.4.3. Les noyaux de modèle de SVR : En anglais SVR kernels, les noyaux de SVR sont des fonctions qui permettent de transformer les données d'entrée dans un espace de dimension supérieure, où il est plus facile de trouver une séparation linéaire, certains exemples de noyaux SVR couramment utilisés sont le noyau linéaire, le noyau polynomial, le noyau gaussien (ou RBF) et le noyau sigmoïde, chaque noyau a ses propres caractéristiques et est adapté à différents types de données.

IV.6.4.3.1. Noyau linéaire : Le noyau linéaire est utilisé lorsque les données peuvent être séparées par une ligne droite. Il calcule simplement le produit scalaire entre les vecteurs d'entrée. Cela signifie qu'il mesure la similarité entre deux vecteurs en regardant combien ils se ressemblent en termes de direction et de magnitude. Mathématiquement, cela peut être représenté comme suit :

$$K(x, y) = x \cdot y \quad \dots (3.50)$$

Ce noyau est efficace pour des relations simples et linéaires, et il est appliqué dans :

- La prédiction du prix de l'immobilier en fonction de la surface et du nombre de pièces.
- L'estimation de la consommation de carburant en fonction du poids et de la taille du véhicule.
- La prévision des ventes commerciales à partir des dépenses publicitaires.

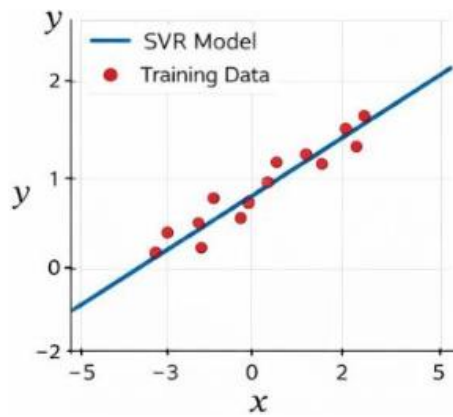


Figure 3.23. Représentation de noyau linéaire de modèle SVR.

IV.6.4.3.2. Noyau polynomial : Le noyau polynomial est utilisé lorsque les données ne peuvent pas être séparées par une ligne droite, mais peuvent être séparées par une courbe de degré supérieur. Il calcule le produit scalaire entre les vecteurs d'entrée, l'ajoute à une constante et élève le résultat à une puissance donnée. Cela permet de créer des frontières de décision plus complexes et non linéaires.

L'équation de calcul pour le noyau polynomial est donnée par :

$$K(x, y) = (x \cdot y + c_0)^d \dots (3.51)$$

Dans cette équation, ' c_0 ' est un terme de décalage (biais) et ' d ' est le degré du polynôme, ces paramètres peuvent être ajustés pour obtenir une meilleure séparation des données, donc la complexité dépend du degré choisi. Ce noyau permet de représenter des relations non linéaires par une expansion polynomiale, peut introduire un risque d'over-fitting pour des degrés trop élevés, exemples de ces applications :

- Modélisation de la croissance démographique avec des courbes quadratiques ou cubiques.
- Estimation de la consommation énergétique d'un véhicule en fonction de la vitesse.
- Prédiction du rendement agricole selon la quantité d'engrais et d'eau.

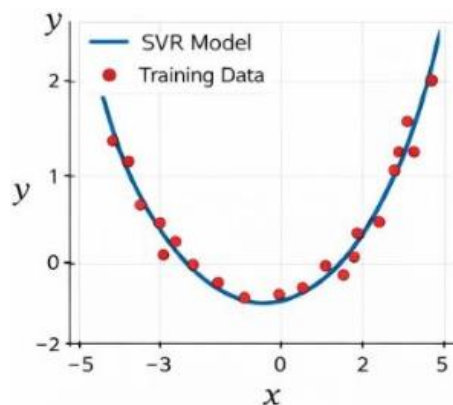


Figure 3.24. Représentation de noyau polynomial de modèle SVR.

IV.6.4.3.3. Noyau Gaussien (ou RBF) : Le noyau gaussien, également appelé RBF (Radial Basis Function), est utilisé pour modéliser des frontières de décision très complexes et non linéaires, il mesure la similarité entre deux vecteurs en utilisant une fonction exponentielle de la distance euclidienne entre eux. Cela signifie qu'il attribue une valeur plus élevée aux vecteurs qui sont proches les uns des autres et une valeur plus faible aux vecteurs qui sont éloignés. L'équation de calcul pour le noyau gaussien est donnée par :

$$k(x, y) = e^{(-\gamma \|x - y\|^2)} \dots (3.52)$$

Dans cette équation, ' γ ' est un paramètre qui contrôle l'influence de chaque exemple d'entraînement sur la frontière de décision, une valeur plus élevée de gamma donne une frontière de décision plus complexe. Le RBF est le plus utilisé et le plus flexible, et il est capable de modéliser des relations fortement non linéaires. On peut rencontrer un risque de sur-apprentissage si les hyper-paramètres ne sont pas bien réglés. Exemples d'applications :

- Prédiction météorologique (température, humidité, précipitations).
- Modélisation financière (prédiction du prix des actions ou indices boursiers).
- Estimation de la production industrielle selon divers paramètres interdépendants.

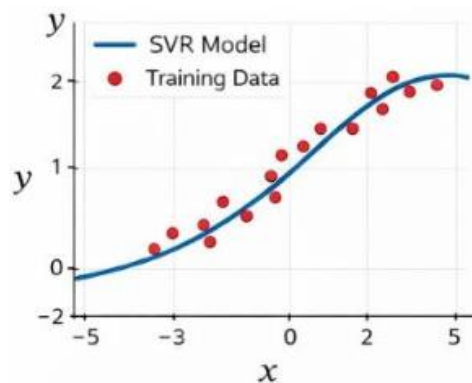


Figure 3.25. Représentation de noyau RBF (Gaussien) de modèle SVR.

IV.6.4.3.4. Noyau sigmoïde : Le noyau sigmoïde est utilisé pour modéliser des frontières de décision non linéaires similaires à celles des réseaux de neurones, il calcule le produit scalaire entre les vecteurs d'entrée, le multiplie par une constante et applique ensuite la fonction tangente hyperbolique. Cela permet de créer des frontières de décision plus complexes et non linéaires. L'équation de calcul pour le noyau sigmoïde est donnée par :

$$K(x, y) = \tanh(ax \cdot y + c_0) \quad \dots (3.53)$$

Dans cette équation, ' α ' est un paramètre de mise à l'échelle et ' c_0 ' est un terme de décalage (biais).

Ce noyau est caractérisé par :

- Inspiré des fonctions d'activation des réseaux de neurones ($\tanh$).
- Peut représenter des relations non linéaires avec une forme en S (S-shape).
- Moins stable que le noyau RBF dans la pratique, et donc moins utilisé.

Et comme des exemples d'applications :

- Analyse du comportement des consommateurs (probabilité d'achat en fonction du prix).
- Modélisation des courbes d'apprentissage (progrès rapide puis plateau).
- Prédiction de la densité du trafic routier aux heures de pointe.

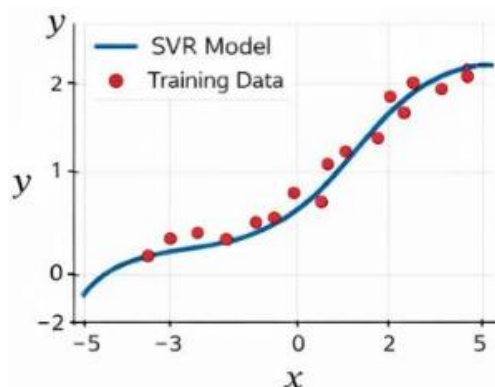


Figure 3.26. Représentation de noyau sigmoïde de modèle SVR.

IV.6.4.4. Les hyper-paramètres de modèle de SVR :

a) Le paramètre de régularisation C : Le paramètre C est trouvé dans tous les noyaux de modèle de SVR, est un paramètre qui contrôle la marge d'erreur tolérée dans la formation du modèle SVR.

Une valeur faible de C permet une marge d'erreur plus large, ce qui peut conduire à une frontière de décision plus douce, mais également à une plus grande tolérance aux erreurs d'entraînement, une valeur élevée de C réduit la marge d'erreur tolérée, ce qui peut conduire à une frontière de décision plus rigide, mais également à une plus faible tolérance aux erreurs d'entraînement.

b) Le paramètre ϵ : Le paramètre epsilon est utilisé dans les SVR à marge douce pour définir la tolérance aux erreurs d'optimisation, il contrôle la précision avec laquelle le modèle SVR doit être optimisé.

Une valeur faible d'epsilon signifie une optimisation plus précise, mais cela peut également augmenter le temps de calcul, une valeur élevée d'epsilon permet une optimisation plus rapide, mais avec une précision légèrement réduite.

c) Le paramètre de contrôle l'influence γ : Le paramètre gamma on le trouve dans le modèle de RBF seulement et de fonction de contrôle l'influence des exemples d'entraînement sur la formation de la frontière de décision.

Une valeur faible de gamma signifie que les exemples d'entraînement ont une influence plus étendue, ce qui peut donner une frontière de décision plus lisse mais potentiellement moins précise, une valeur élevée de gamma signifie que les exemples d'entraînement ont une influence plus locale, ce qui peut donner une frontière de décision plus précise mais potentiellement plus complexe.

De plus, on trouve dans le noyau polynomial le degré du polynôme est un autre paramètre à prendre en compte dans ce noyau, il détermine la complexité de la frontière de décision.

d) Le degré d : Ce degré il donne un aperçu sur la complexité du noyau polynomial.

$d \rightarrow$ plus grand = plus de courbure.

e) Le terme de décalage c_0 (biais) : c_0 est une constante indépendante (similaire à le biais c de noyau Dot Product / Quadratique de GPR), et ce rôle est ajouter une partie fixe qui permet de modéliser des relations plus riches.

Si $c_0 = 0 \rightarrow$ le noyau dépend uniquement du produit scalaire (partie linéaire).

Si $c_0 > 0 \rightarrow$ on introduit une constante qui améliore la flexibilité du noyau.

Pour obtenir plus d'informations sur le modèle de SVR, les références spécialisés sont disponible : les articles [60, 61, 62, 63] et la bibliothèque Skit-Learn (SVR) [64].

Tableau 3.1. Résumé de modèles sélectionnés, les noyaux, les hyper-paramètres et les (+ avantages, – inconvenants) de chacun noyau.

Modèle	Noyau	Hyper-paramètres de noyau		Caractéristiques de noyau	
		Communs	Spécifiques	+ Avantages	- Inconvénients
Régression	SLR MLR	α : Ridge(α) = $\frac{[10^{-4}, 10^3]}{[10^{-4}, 10^3]}$ Lasso (α) = $[10^{-4}, 10^3]$		+ Simple et interprétable. + Efficace pour relations linéaires.	- Faible pour non-linéarité. - Sensible aux outliers.
ANN	MLP		nb couches cachées : [1; 10] nb neurones/couche : [10; 500] Learning rate : $[10^{-5}, 1]$ fonction d'activation : {ReLU, tanh, sigmoïde}	+ Puissant pour non-linéarité. + Bonne généralisation.	- Temps d'entraînement élevé. - Risque d'over-fitting.
Arbre	DT	max_depth : [1; + ∞] min_samples_split : [2; 20] min_samples_leaf : [1; 20] critère de division : {Gini, entropie} n_estimators : [10; 1000] max_features : $[\sqrt{p}; \log_2(p); p]$		+ Facile à interpréter. + Pas besoin de normalisation.	- Over-fitting. fréquent. - Instable
	RF			+ Réduit over-fitting vs DT. + Bonne précision.	- Moins interprétable. - Plus lent à entraîner.
GPR	RBF	ℓ (longueur) $[10^{-2}; 10]$, Le bruit $[10^{-4}, 1]$	σ^2 (variance) : $[10^{-3}, 10^2]$	+ Lisse et flexible. + Bonne approximation.	- Faible pour périodicité. - Coût élevé.
	Exponentiel		/	+ Capture variations brusques. + Tolère valeurs extrêmes.	- Non lisse. - Approximations grossières.
	Matern		ν (lissage) : $\{1/2, 3/2, 5/2\}$	+ Bon compromis RBF. + Très flexible.	- Réglage ν délicat. - Calcul plus lourd.
	Exp-Sqr-Sin		p (périodicité) : $[10^{-2}, 10^2]$	+ Excellent pour données périodiques. + Capture saisonnalité.	- Mauvais pour non-périodiques. - Sensible à périodicité.
	RQ		α (échelle mixture) : $[10^{-2}, 10]$	+ Très flexible. + Combine plusieurs échelles.	- Réglage complexe. - Plus lourd que RBF.
SVR	Dot-Product	C : $[10^{-5}, 100]$ ϵ : $[10^{-5}, 10^5]$	σ^2 (variance) : $[10^{-3}, 10^2]$ c (biais) : $\begin{cases} c = 0, \text{ noyau linéaire} \\ c \neq 0, \text{ quadratique} \end{cases}$	+ Simple. + Bon pour relations linéaires.	- Faible pour non-linéaire. - Précision limitée.
	Linéaire		/	+ Rapide et simple. + Bon en haute dimension.	- Faible pour non-linéarité. - Peu flexible.
	RBF		γ : $[10^{-5}, 10^5]$	+ Très puissant. + Haute flexibilité.	- Sensible à γ . - Coût élevé.
	Sigmoïde		c_0 (biais) : ≥ 0 degré (d) : [2; 10]	+ Proche des ANN. + Adapté à certains cas.	- Instable. - Performance faible souvent
	Polynomial			+ Représente polynômes. + Plus expressif que linéaire.	- Risque over-fitting. - Complexité rapide.

V. Sélection des caractéristiques d'entrées (Features selection technique) :

C'est le processus de choix des meilleures caractéristiques d'entrées pour le modèle. Ce processus peut différer d'une technique à l'autre, mais l'objectif principal est de déterminer quelles caractéristiques d'entrées ont le plus d'impact sur le modèle.

Il existe deux grandes catégories de techniques pour cela :

1. Techniques de sélection de caractéristiques supervisées.

2. Techniques de sélection de caractéristiques non supervisées.

- Pourquoi devrions-nous faire une sélection de caractéristiques ?

Parce que parfois, avoir trop de caractéristiques peut nuire au modèle d'apprentissage automatique. Comment ?

Il peut y avoir plusieurs raisons. Par exemple, ces caractéristiques peuvent être corrélées entre elles, ce qui peut entraîner une multicollinéarité, ruinant ainsi la performance de modèle.

Un autre problème potentiel est lié à la puissance de calcul. La présence de trop nombreuses caractéristiques nécessite plus de puissance de calcul pour exécuter la tâche simultanément, ce qui pourrait nécessiter plus de ressources et, par conséquent, augmenter les coûts.

Il peut certainement y avoir d'autres raisons également. Mais ces exemples devraient vous donner une idée générale des problèmes potentiels. Cependant, il y a un aspect encore plus important à comprendre avant d'approfondir ce sujet.

- Quelle méthode de sélection de caractéristiques sera la meilleure pour mon modèle ?

Le choix du modèle de sélection de caractéristiques dépend du type de données que vous avez et des objectifs de votre projet, par exemple, les méthodes basées sur des filtres telles que le test du chi carré ou le gain d'information mutuelle sont généralement utilisées pour la sélection de caractéristiques dans des données catégorielles. Les méthodes basées sur des wrappers (Wrapper-based) comme la sélection avant ou arrière sont adaptées aux données numériques.

Cependant, il est bon de savoir que de nombreuses méthodes de sélection de caractéristiques peuvent traiter à la fois des données catégorielles et numériques, par exemple, la régression lasso, les arbres décisionnels et les forêts aléatoires peuvent gérer ces deux types de données assez bien.

En termes de sélection de caractéristiques supervisée et non supervisée, les méthodes supervisées comme l'élimination récursive des caractéristiques ou les arbres décisionnels sont bonnes pour les données étiquetées. Les méthodes non supervisées comme l'analyse en composantes principales (ACP) ou l'analyse en composantes indépendantes (ICA) sont utilisées pour les données non étiquetées. En fin de compte, le choix de la méthode de sélection de caractéristiques doit être basé sur les caractéristiques spécifiques de données et les objectifs de projet.

V.1. Techniques de sélection de caractéristiques supervisées : Les stratégies de sélection de caractéristiques en apprentissage supervisé visent à découvrir les caractéristiques les plus pertinentes pour prédire la variable cible en utilisant la relation entre les caractéristiques d'entrée et la variable cible. Ces stratégies peuvent aider à améliorer la performance du modèle, réduire le sur-apprentissage et diminuer le coût computationnel d'entraînement du modèle, voici un aperçu des techniques de sélection de caractéristiques supervisées dont nous allons parler.

V.1.1. Approche basée sur les filtres : Les approches de sélection de caractéristiques basées sur des filtres reposent sur des attributs intrinsèques des données tels que la corrélation des caractéristiques ou des statistiques. Ces approches évaluent la valeur de chaque caractéristique seule ou par paires sans tenir compte des performances d'un algorithme d'apprentissage particulier.

Les approches basées sur des filtres sont efficaces sur le plan computationnel et peuvent être utilisées avec une variété d'algorithmes d'apprentissage. Cependant, comme elles ne tiennent pas compte des interactions entre les caractéristiques et la méthode d'apprentissage, elles ne capturent pas toujours le sous-ensemble idéal de caractéristiques pour un certain algorithme.

Aperçu des approches basées sur des filtres, prenons un moment pour examiner l'aperçu des approches basées sur des filtres, puis nous discuterons de chacune d'elles.

V.1.1.1. Gain d'information : Le gain d'information est une statistique qui mesure la réduction de l'entropie (incertitude) pour une caractéristique spécifique en divisant les données selon cette caractéristique. Il est souvent utilisé dans les algorithmes d'arbres de décision et possède également des caractéristiques utiles. Plus le gain d'information d'une caractéristique est élevé, plus elle est utile pour la prise de décision.

V.1.1.2. Test du Chi-carré : Le test du Chi-carré est un test statistique utilisé pour évaluer la relation entre deux variables catégorielles. Il est utilisé dans la sélection de caractéristiques pour analyser la relation entre une caractéristique catégorielle et la variable cible. Un score de Chi carré plus élevé montre un lien plus fort entre la caractéristique et la cible, indiquant que la caractéristique est plus importante pour la tâche de classification. Bien que le test du Chi carré soit une méthode de sélection de caractéristiques couramment utilisée, il est généralement appliqué aux données catégorielles, où les caractéristiques et les variables cibles sont discrets.

V.1.1.3. Score de Fisher : Le ratio discriminant de Fisher, communément connu sous le nom de Score de Fisher, est une approche de sélection de caractéristiques qui classe les caractéristiques en fonction de leur capacité à différencier diverses classes dans un ensemble de données. Il peut être utilisé pour des caractéristiques continues dans un problème de classification. Le Score de Fisher est calculé comme le ratio de la variance entre classes à la variance intra-classe. Un Score de Fisher plus élevé implique que la caractéristique est plus discriminante et précieuse pour la classification. Pour utiliser le Score de Fisher pour la sélection de caractéristiques, calculez un score pour chaque caractéristique continue et classez-les selon leurs scores. Le modèle considère les caractéristiques avec un Score de Fisher plus élevé comme plus importantes.

V.1.1.4. Ratio de valeurs manquantes : Le Ratio de valeurs manquantes est une méthode simple de sélection de caractéristiques qui prend des décisions en fonction du nombre de valeurs manquantes dans une caractéristique. Les caractéristiques ayant une proportion significative de valeurs manquantes peuvent être peu informatives et nuire à la performance du modèle. Vous pouvez filtrer les caractéristiques avec trop de valeurs manquantes en spécifiant un seuil pour le ratio acceptable de valeurs manquantes. Pour utiliser le Ratio de valeurs manquantes pour la sélection de caractéristiques, suivez ces étapes :

1. Calculez le ratio de valeurs manquantes pour chaque caractéristique en divisant le nombre de valeurs manquantes par le nombre total d'instances dans l'ensemble de données.
2. Définissez un seuil pour le ratio acceptable de valeurs manquantes (par exemple, 0,8, ce qui signifie qu'une caractéristique ne doit avoir au maximum 80 % de ses valeurs manquantes pour être considérée).
3. Filtrez les caractéristiques ayant un ratio de valeurs manquantes supérieur au seuil.

V.1.2. Approche basée sur des Wrappers : Les approches de sélection de caractéristiques basées sur des wrappers consistent à évaluer l'importance des caractéristiques en utilisant un algorithme d'apprentissage automatique spécifique. Elles cherchent le meilleur sous-ensemble de caractéristiques en expérimentant différentes combinaisons et en évaluant leurs performances avec la méthode sélectionnée. En raison du grand nombre de sous-ensembles de caractéristiques disponibles, les

approches basées sur des wrappers peuvent être coûteuses en termes computationnels, surtout lorsqu'on travaille avec des ensembles de données à haute dimension. Cependant, elles surpassent souvent les approches basées sur des filtres car elles prennent en compte la relation entre les caractéristiques et l'algorithme d'apprentissage.

V.1.2.1. Sélection avant : Dans la sélection avant, vous commencez avec un ensemble vide de caractéristiques et ajoutez itérativement des caractéristiques à cet ensemble. À chaque étape, vous évaluez la performance du modèle avec l'ensemble actuel de caractéristiques et la caractéristique supplémentaire. La caractéristique qui entraîne la meilleure amélioration des performances est ajoutée à l'ensemble. Le processus se poursuit jusqu'à ce qu'aucune amélioration significative des performances ne soit observée ou qu'un nombre prédéfini de caractéristiques soit atteint.

Exemple : Le graphique (ci-dessous) linéaire fournit des informations supplémentaires sur la performance du modèle avec différents nombres de caractéristiques. Il montre que :

- Avec seulement 1 caractéristique, le modèle atteint une précision d'environ 91.
- L'ajout d'une deuxième caractéristique augmente la précision à 94.
- Avec 3 caractéristiques, la précision s'améliore encore pour atteindre 95.
- En incluant 4 caractéristiques, la précision dépasse légèrement les 95 %. Au-delà de 4 caractéristiques, les améliorations de la précision deviennent moins significatives.

Ces informations peuvent vous aider à prendre des décisions éclairées sur les compromis entre la complexité du modèle et les performances prédictives. Sur la base de ces résultats, vous pourriez décider d'utiliser seulement 3 ou 4 caractéristiques dans votre modèle pour équilibrer précision et simplicité.

Selected Features: ('0', '1', '4', '21', '22')
Accuracy: 0.9548417721518987

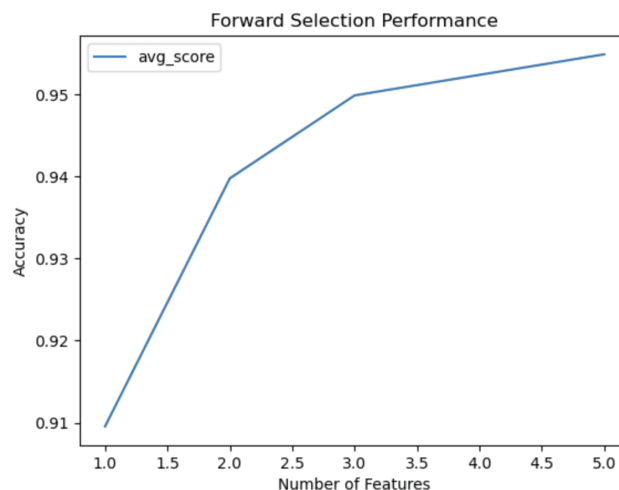


Figure 3.27. Exemple sur la performance de modèle avec la méthode de sélection avant.

V.1.2.2. Sélection arrière : L'opposé de la sélection avant est la sélection arrière. Vous commencez avec l'ensemble complet des caractéristiques et éliminez progressivement des caractéristiques. À chaque étape, vous mesurez la performance du modèle avec l'ensemble actuel de caractéristiques moins la caractéristique à supprimer. La caractéristique qui entraîne le moins de réduction de performance est éliminée de l'ensemble. La procédure est répétée jusqu'à ce qu'il n'y ait pas d'augmentation substantielle de la performance ou qu'un nombre prédéfini de caractéristiques soit atteint. Les sélections avant et arrière sont classées comme une sélection séquentielle de caractéristiques ; vous pouvez en apprendre davantage ici.

V.1.2.3. Sélection Exhaustive des caractéristiques (EFS) : La sélection Exhaustive des caractéristiques compare la performance de tous les sous-ensembles de caractéristiques possibles et choisit le sous-ensemble ayant la meilleure performance. Cette approche est coûteuse en termes computationnels, surtout pour les grands ensembles de données, mais elle garantit le meilleur sous-ensemble de caractéristiques, (c'est la méthode qu'on a adopté avec le modèle SVR dans notre étude).

V.1.2.4. Élimination Récursive des caractéristiques (RFE) : L'élimination Récursive des caractéristiques commence avec l'ensemble complet des caractéristiques et élimine les caractéristiques de manière répétée en fonction de leur pertinence telle que jugée par l'algorithme d'apprentissage. La caractéristique la moins importante est supprimée à chaque étape, et le modèle est ré-entraîné. La méthode est répétée jusqu'à ce qu'un nombre prédéterminé de caractéristiques soit atteint, (c'est la méthode qu'on a adopté avec le modèle GPR dans notre étude).

V.1.3. Approche intégrée : Les approches de sélection intégrée des caractéristiques incluent le processus de sélection des caractéristiques comme partie intégrante de l'algorithme d'apprentissage. Cela implique qu'au cours de la phase d'entraînement, l'algorithme d'apprentissage non seulement optimise les paramètres du modèle, mais sélectionne également les caractéristiques les plus importantes. Les méthodes intégrées peuvent être plus efficaces que les méthodes basées sur des wrappers, car elles ne nécessitent pas une procédure externe de sélection des caractéristiques.

V.1.3.1. Régularisation : La régularisation est une méthode qui ajoute un terme de pénalité à la fonction de perte pour prévenir le sur-apprentissage dans les modèles d'apprentissage automatique. Les méthodes de régularisation, telles que le lasso (régularisation L1) et le ridge (régularisation L2), peuvent être utilisées en conjonction avec la sélection des caractéristiques pour diminuer les coefficients des caractéristiques moins significatives vers zéro, choisissant ainsi un sous-ensemble des caractéristiques les plus pertinentes.

V.1.3.2. Importance des forêts aléatoires : La forêt aléatoire est une approche d'apprentissage par ensemble qui combine les prédictions de plusieurs arbres décisionnels. La forêt aléatoire calcule un score d'importance des caractéristiques pour chaque caractéristique dans le cadre du processus de construction des arbres, ce qui peut être utilisé pour classer les caractéristiques en fonction de leur pertinence. Le modèle considère les caractéristiques avec des notes d'importance plus élevées comme étant plus significatives.

V.2. Techniques de sélection de caractéristiques non supervisées : Lorsqu'aucune variable cible n'est disponible, les approches de sélection de caractéristiques non supervisées peuvent être utilisées pour réduire la dimensionnalité du jeu de données tout en conservant sa structure sous-jacente. Ces méthodes consistent souvent à transformer l'espace des caractéristiques initiales en un nouvel espace de dimension inférieure dans lequel les caractéristiques modifiées capturent la majorité de la variation dans les données.

V.2.1. Analyse en Composantes Principales (ACP) : L'ACP est une méthode linéaire de réduction de dimensionnalité qui convertit l'espace des caractéristiques d'origine en un nouvel espace orthogonal défini par les composantes principales. Ces composantes sont des combinaisons linéaires des caractéristiques d'origine choisies pour capturer le plus haut niveau de variance dans les données. L'ACP peut être utilisé pour sélectionner les k premières composantes principales représentant la majorité de la variation, réduisant ainsi la dimensionnalité du jeu de données.

V.2.2. Analyse en Composantes Indépendantes (ACI) : L'ACI est une méthode permettant de diviser un signal multidimensionnel en ses composantes. Dans le contexte de la sélection de

caractéristiques, l'ACI peut être utilisée pour transformer l'espace des caractéristiques d'origine en un nouvel espace caractérisé par des composantes statistiquement indépendantes. Vous pouvez réduire la dimensionnalité du jeu de données tout en préservant la structure sous-jacente en choisissant les k premières composantes indépendantes.

V.2.3. Factorisation de Matrice Non Négative (NMF) : La factorisation de matrice non négative (NMF) est une approche de réduction de dimensionnalité qui approxime une matrice de données non négative comme le produit de deux matrices non négatives de dimension inférieure. La NMF peut être utilisée dans le contexte de la sélection de caractéristiques pour extraire un nouvel ensemble de caractéristiques de base qui capturent la structure importante des données d'origine. Vous pouvez minimiser la dimensionnalité du jeu de données tout en maintenant la contrainte de non-négativité en choisissant les k premières caractéristiques de base.

V.2.4. Intégration Stochastique à Proximité t-distribuée (t-SNE) : Le t-SNE est une méthode non linéaire de réduction de dimensionnalité qui tente de préserver la structure du jeu de données en réduisant la différence entre les distributions de probabilité par paires dans des emplacements à haute et basse dimension. Le t-SNE peut être appliqué dans la sélection de caractéristiques pour projeter l'espace des caractéristiques d'origine dans un espace de dimension inférieure qui maintient la structure des données, permettant ainsi une meilleure visualisation et évaluation.

V.2.5. Auto-encodeur : Un auto-encodeur, un type de réseau neuronal artificiel, apprend à encoder les données d'entrée en une représentation de dimension inférieure puis à les décoder pour revenir à la version originale. La représentation de dimension inférieure de l'auto-encodeur peut être utilisée pour produire un autre ensemble de caractéristiques qui capturent la structure sous-jacente des données d'origine.

En conclusion, la sélection des caractéristiques est essentielle en apprentissage automatique. Elle aide à réduire la dimensionnalité des données, minimise le risque de sur-apprentissage et améliore la performance globale du modèle.

Le choix de la méthode de sélection des caractéristiques dépend du problème spécifique, de l'ensemble de données et des exigences de modélisation. [65]

Dans notre travail, notre base de données est moyenne et de 1123 échantillons mais contient un nombre limité de caractéristiques (douze caractéristiques), ce qui rend faisable la mise en œuvre des techniques de basée sur des Wrappers (du méthode d'élimination Récursive des caractéristiques (RFE) avec le modèle GPR, et du méthode de sélection Exhaustive des caractéristiques (EFS) avec le modèle SVR). On obtient les deux identifications (noms symboliques) suivants: GPR-RFE et SVR-EFS.

Ces deux techniques visent à déterminer les sous-ensembles optimaux de caractéristiques à partir du jeu de données initial, combiné avec les deux modèles hybrides : TPE-GPR-RFE et TPE-SVR-EFS. (on va s'expliqué dans le prochaine point).

Les deux sous-ensembles optimaux que nous avons obtenus contiennent seulement sept caractéristiques pour chacun, en utilisant les deux techniques de basée sur des Wrappers (élimination Récursive des caractéristiques (RFE) et sélection Exhaustive des caractéristiques (EFS)) avec une validation croisée à 10 plis.

- Le premier sous-ensemble optimal est déterminé avec la technique RFE qui hybridé avec le modèle de GPR, comme consistant en les sept caractéristiques suivantes : (G, FC, PI, MC, DD, OMC et MDD).

- Le deuxième sous-ensemble optimal est déterminé avec la technique EFS qui hybridé avec le modèle de SVR, comme consistant en les sept caractéristiques suivantes : (G, S, PL, PI, DD, OMC et MDD).

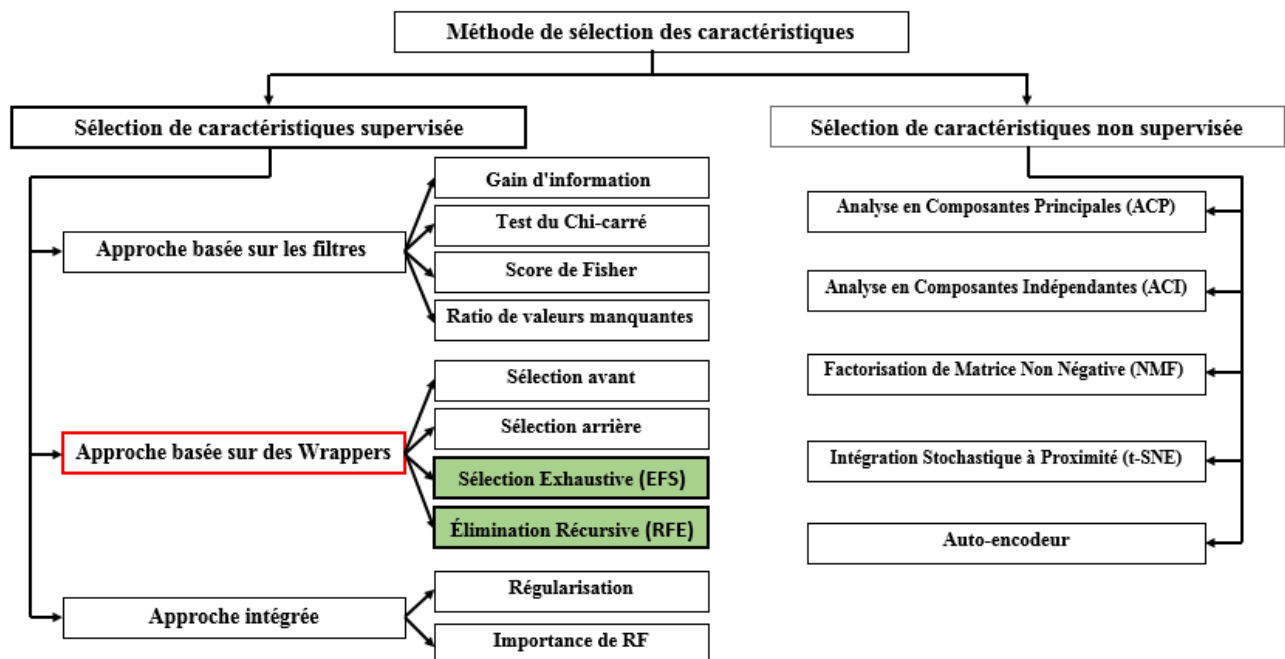


Figure 3.28. Organigramme explicatif pour les types de méthode de features selection, en montrant les techniques adoptées dans notre étude.

Pour obtenir plus d'informations sur la méthode de sélection des entrées (features selection method), les références spécialisés sont disponible : [66, 67 et 68].

VI. Méthode d'optimisation : L'optimisation des modèles est une méthode qui vise à trouver les meilleures valeurs pour les paramètres d'un modèle afin d'améliorer ses performances, l'objectif est d'optimiser une fonction de coût ou une mesure de performance spécifique, pour y parvenir, différentes techniques d'optimisation sont utilisées par exemple :

a) Descente de gradient et rétro-propagation : Ces méthodes sont utilisées pour optimiser les modèles de MLP en corrigeant les erreurs et en ajustant les poids de manière itérative.

b) Optimisation bayésienne : Cette approche utilise un algorithme spécial pour déterminer le meilleur noyau et préciser les hyper-paramètres du modèle, cela permet d'obtenir des résultats plus précis et d'améliorer les performances du modèle.

c) Algorithme de Boruta : Il est utilisé pour éliminer les entrées ayant une corrélation élevée entre elles et ayant peu d'impact sur la sortie, cela permet de simplifier le modèle et d'améliorer son efficacité.

d) Validation croisée avec K-Fold : Cette technique permet de diviser la base de données de manière optimale entre les deux phases d'entraînement et de test, cela garantit une évaluation fiable des performances du modèle.

VI.1. Hyper-paramètres : Les hyper-paramètres sont des variables de configuration externes que les scientifiques des données utilisent pour gérer l'entraînement des modèles de Machine Learning.

Parfois appelés hyper-paramètres du modèle, les hyper-paramètres sont définis manuellement avant la formation d'un modèle. Ils sont différents des paramètres, qui sont des paramètres internes dérivés automatiquement au cours du processus d'apprentissage et non fixés par les scientifiques des données.

Le nombre de nœuds et de couches dans un réseau neuronal et le nombre de branches dans un arbre de décision sont des exemples d'hyper-paramètres. Les hyper-paramètres déterminent des caractéristiques clés telles que l'architecture du modèle, le taux d'apprentissage et la complexité du modèle.

VI.1.1. Identification des hyper-paramètres : La sélection du bon ensemble d'hyper-paramètres est importante en termes de performance et de précision du modèle. Malheureusement, il n'y a pas de règles établies sur les hyper-paramètres qui fonctionnent le mieux, ni sur leurs valeurs optimales ou par défaut. Vous devez faire des expériences pour trouver l'ensemble optimal d'hyper-paramètres.

Cette activité est connue sous le nom de réglage des hyper-paramètres ou d'optimisation des hyper-paramètres.

VI.1.2. Réglage des hyper-paramètres : Lorsque vous entraînez des modèles de Machine Learning, chaque jeu de données ainsi que chaque modèle nécessite un ensemble différent d'hyper-paramètres constituant une sorte de variable. La seule façon de les déterminer consiste à effectuer plusieurs expériences en sélectionnant un ensemble d'hyper-paramètres pour les exécuter dans votre modèle.

C'est ce qu'on appelle le réglage des hyper-paramètres. Ainsi, vous entraînez votre modèle de manière séquentielle avec différents ensembles d'hyper-paramètres. Ce processus peut être manuel ou vous pouvez choisir l'une des nombreuses méthodes de réglage automatique des hyper-paramètres.

Quelle que soit la méthode que vous utilisez, il est essentiel de suivre les résultats de vos expériences. Vous devrez appliquer une forme d'analyse statistique, telle que la fonction de perte, afin de déterminer l'ensemble d'hyper-paramètres donnant le meilleur résultat. Le réglage des hyper-paramètres est un processus d'importance et d'intensité de calcul élevé.

- Pourquoi le réglage des hyper-paramètres est-il important ?

Les hyper-paramètres contrôlent directement la structure, la fonction et la performance du modèle. Le réglage des hyper-paramètres permet aux spécialistes des données d'ajuster les performances du modèle pour obtenir des résultats optimaux. Ce processus est une partie essentielle du Machine Learning, et le choix des valeurs appropriées des hyper-paramètres est crucial pour le succès. Par exemple, supposons que vous utilisiez le taux d'apprentissage du modèle comme hyper-paramètre. Si la valeur est trop élevée, le modèle peut converger trop rapidement avec des résultats sous-optimaux. Alors que si le taux est trop faible, la formation prend trop de temps et les résultats peuvent ne pas converger. Un choix judicieux et équilibré des hyper-paramètres permet d'obtenir des modèles précis et d'excellentes performances.

VI.1.3. Fonctionnement de réglage des hyper-paramètres : Comme indiqué précédemment, le réglage des hyper-paramètres peut être manuel ou automatisé. Bien que le réglage manuel soit lent et fastidieux, l'avantage est que vous comprenez mieux comment les pondérations des hyper-paramètres affectent le modèle. Mais dans la plupart des cas, vous utiliserez normalement l'un des algorithmes d'apprentissage à hyper-paramètres bien connus.

Le processus de réglage des hyper-paramètres est itératif, et vous essayez différentes combinaisons de paramètres et de valeurs. Vous commencez généralement par définir une variable cible, telle que la précision, comme métrique primaire, et vous souhaitez maximiser ou de minimiser cette variable. Il est bon d'utiliser des techniques de validation croisée, afin que votre modèle ne soit pas centré sur une seule partie de vos données.

VI.1.4. Les techniques d'ajustement des hyper-paramètres : Il existe de nombreux algorithmes de réglage des hyper-paramètres, mais les types les plus couramment utilisés sont l'optimisation bayésienne, la recherche dans la grille et la recherche aléatoire.

VI.1.4.1. Optimisation Bayésienne : L'optimisation bayésienne est une technique basée sur le théorème de Bayes, qui décrit la probabilité qu'un événement se produise en fonction des connaissances actuelles. Lorsqu'il est appliqué à l'optimisation des hyper-paramètres, l'algorithme construit un modèle probabiliste à partir d'un ensemble d'hyper-paramètres qui optimise une métrique spécifique. Il utilise l'analyse de régression pour choisir de manière itérative le meilleur ensemble d'hyper-paramètres.

VI.1.4.2. Recherche dans la grille : Avec la recherche par grille, vous spécifiez une liste d'hyper-paramètres et une mesure de performance, et l'algorithme examine toutes les combinaisons possibles pour déterminer le meilleur ajustement. La recherche par grille fonctionne bien, mais elle est relativement fastidieuse et exigeante en termes de calcul, surtout avec un grand nombre d'hyper-paramètres.

VI.1.4.3. Recherche aléatoire : Bien que basée sur des principes similaires à ceux de la recherche sur grille, la recherche aléatoire sélectionne des groupes d'hyper-paramètres de manière aléatoire à chaque itération. Elle fonctionne bien lorsqu'un nombre relativement faible d'hyper-paramètres détermine principalement le résultat du modèle. [69]

VI.2. L'Optimisation Bayésienne BOA : L'optimisation bayésienne en Machine Learning est une méthode d'optimisation qui utilise des modèles probabilistes pour trouver de manière efficace les hyper-paramètres d'un modèle. En d'autres termes, c'est une technique mathématique dont l'objectif est de trouver la combinaison d'hyper-paramètres optimale.

L'optimisation bayésienne fonctionne en construisant une distribution postérieure de fonctions (Processus Gaussien) qui décrit le mieux la fonction que nous souhaitons optimiser. À mesure que le nombre d'observations augmente, la distribution postérieure s'améliore et l'algorithme devient plus sûr des régions de l'espace des paramètres qui méritent d'être explorées et de celles qui ne le sont pas, comme le montre la figure 3.29.

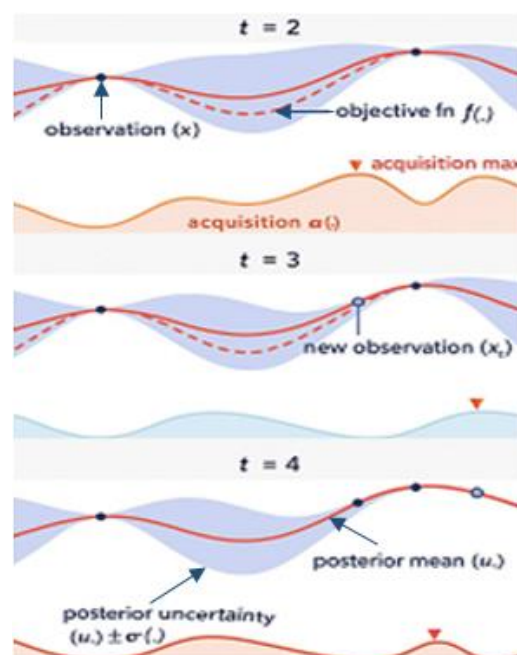


Figure 3.29. Exemple d'utilisation de l'optimisation bayésienne BOA. [70]

Au fur et à mesure des itérations, l'algorithme équilibre ses besoins d'exploration et d'exploitation en tenant compte de ce qu'il sait sur la fonction cible. À chaque étape, un processus gaussien est ajusté aux échantillons connus (points précédemment explorés), et la distribution postérieure, associée à une stratégie d'exploration (telle que la limite de confiance supérieure ou l'amélioration attendue) est utilisée pour décider du prochain point qui devrait être exploré.

Ce processus est conçu pour minimiser le nombre d'étapes nécessaires pour trouver une combinaison de paramètres proche de la combinaison optimale. Pour ce faire, cette méthode utilise un problème d'optimisation de proxy (trouver le maximum de la fonction d'acquisition) qui, bien qu'il reste un problème difficile, est moins coûteux (au sens de calcul) et que des outils communs peuvent être utilisés. Par conséquent, l'optimisation bayésienne convient le mieux aux situations dans lesquelles l'échantillonnage de la fonction à optimiser est une tâche très coûteuse. [71] [72]

VI.2.1. Principaux composants (stratégies) de l'Optimisation Bayésienne BOA :

VI.2.1.1. Le modèle surrogate : (Ou modèle de substitution) permet d'estimer la manière dont les différentes valeurs des hyper-paramètres affectent les performances du modèle.

Il utilise les données des expérimentations précédentes pour prédire les performances potentielles de nouvelle combinaison d'hyper-paramètres.

Ainsi, le modèle surrogate guide le choix des prochaines configurations à évaluer.

VI.2.1.2. La fonction d'acquisition : C'est une mesure mathématique qui évalue l'intérêt (ou l'utilité) d'évaluer une configuration d'hyper-paramètres donnée. Elle tient compte à la fois de l'incertitude du modèle surrogate et de la recherche d'une amélioration des performances du modèle.

Les fonctions d'acquisition couramment utilisées incluent :

- Expected Improvement.
- Expected Optimism.
- UCB – Upper Confidence Bound.

VI.2.1.3. La stratégie d'équilibre entre exploration et exploitation : C'est l'approche utilisée pour décider si l'on doit explorer de nouvelles configurations d'hyper-paramètres pour découvrir des améliorations potentielles (exploration) ou exploiter les configurations actuellement connues considérées comme étant les meilleures (exploitation).

Cette stratégie est essentielle pour trouver efficacement la meilleure combinaison d'hyper-paramètres. Pour réaliser cela, elle ajuste progressivement son approche pour choisir les prochaines combinaisons à évaluer en fonction de l'objectif d'optimisation.

VI.2.1.4. L'estimateur de Parzen structuré arborescent (TPE) : L'algorithme Tree-structured Parzen Estimator (TPE) est une méthode avancée d'optimisation des hyper-paramètres largement utilisée dans l'apprentissage automatique. Contrairement aux approches classiques de recherche aléatoire ou exhaustive, le TPE repose sur la modélisation probabiliste des espaces de recherche.

Il construit deux distributions distinctes : l'une pour les configurations d'hyper-paramètres associées à de bonnes performances, et l'autre pour celles conduisant à des résultats moins satisfaisants. Les nouvelles valeurs sont ensuite échantillonnées de manière préférentielle dans les régions jugées prometteuses, ce qui rend la recherche plus efficace et ciblée.

Cette approche permet non seulement de réduire considérablement le temps de calcul, mais également d'explorer des espaces complexes et de grande dimension de façon plus structurée.

En comparaison avec l'optimisation bayésienne classique, qui s'appuie généralement sur un processus gaussien (GP) pour modéliser la fonction objectif, le TPE présente l'avantage d'une meilleure adaptabilité aux fonctions fortement non linéaires et à haute dimension, tout en maintenant une robustesse et une rapidité accrues dans le processus d'optimisation.

VI.2.2. Le Processus Itératif de l'Optimisation Bayésienne BOA : En pratique, l'optimisation bayésienne BOA se déroule itérativement :

- 1- Utiliser le modèle surrogate pour prédire la performance de différentes configurations d'hyper-paramètres.
- 2- Utiliser la fonction d'acquisition pour choisir la configuration d'hyper-paramètres la plus prometteuse à évaluer ensuite.
- 3- Évaluer la configuration d'hyper-paramètres choisie.
- 4- Mettre à jour le modèle surrogate avec les nouveaux résultats.
- 5- Répéter le processus jusqu'à ce qu'un critère d'arrêt soit atteint, tel qu'un nombre maximal d'itérations ou un niveau de performance satisfaisant. [73]

VI.2.3. Bonnes pratiques pour le réglage des hyper-paramètres : Pour garantir un réglage efficace des hyper-paramètres, il est essentiel de suivre quelques bonnes pratiques :

1- Définissez un espace de recherche raisonnable : L'espace de recherche doit être défini sur la base d'une connaissance et d'une compréhension préalables du problème. L'inclusion d'un large éventail de valeurs peut aider à explorer différentes possibilités, mais elle doit être limitée pour éviter les valeurs non pertinentes ou extrêmes.

2- Utiliser la validation croisée : La validation croisée est cruciale pour estimer les performances de différentes configurations d'hyper-paramètres. En divisant les données en plusieurs volets et en évaluant le modèle sur chaque repli, il fournit une estimation plus fiable des capacités de généralisation du modèle.

3- Commencez par une recherche grossière : Il est recommandé de commencer par une recherche grossière pour identifier rapidement les régions prometteuses de l'espace des hyper-paramètres. Cela peut être suivi d'une recherche plus fine dans ces régions pour affiner davantage les valeurs des hyper-paramètres.

4- Régulariser pour éviter le sur-apprentissage (over-fitting) : Les techniques de régularisation telles que la régularisation L1 et L2 peuvent aider à prévenir le sur-apprentissage. L'inclusion d'hyper-paramètres de régularisation dans le processus de réglage peut conduire à une meilleure généralisation et à de meilleures performances du modèle. [74]

VI.2.4. Présentation théorique d'Algorithme d'Optimisation Bayésien BOA et TPE :

L'optimisation bayésienne (BOA) constitue l'un des outils les plus puissants dans le domaine de l'intelligence artificielle et de l'apprentissage automatique, en particulier pour la tâche cruciale de l'optimisation des hyper-paramètres (Hyper-parameter Optimization - HPO). Cette étape est souvent considérée comme l'une des plus complexes dans la conception des modèles.

L'avantage majeur de la BOA réside dans son fondement statistique rigoureux basé sur le théorème de Bayes, qui permet d'exploiter les connaissances a priori et les observations passées pour mettre à jour la distribution des résultats attendus. Contrairement aux approches classiques telles que la recherche par grille ou la recherche aléatoire, BOA repose sur deux composantes principales :

1. Un modèle de Surrogate servant à approximer la fonction objectif (black-box function).
2. Une fonction d'acquisition qui détermine le prochain point à évaluer afin de maximiser l'information acquise.

La règle de Bayes s'écrit comme suit : [75]

$$P(y|x) = \frac{P(x|y)p(y)}{p(x)} \quad \dots (3.54) \text{ où :}$$

Notez que w désigne une valeur non observée, $p(y)$ désigne la distribution précédente, $p(x/y)$ désigne la probabilité, et $p(y/x)$ indique la distribution postérieure, la règle de Bayes couvre la connaissance préalable pour déterminer la possibilité postérieure, ce qui signifie que lors du choix des valeurs pour la prochaine itération, les résultats des itérations précédentes seront utilisés. Par conséquent, il peut approcher le point optimal de manière plus efficace qu'une sélection arbitraire BOA peut être utilisée avec deux modèles de sous-ensemble, à savoir le remplacement et l'acquisition. Le modèle de remplacement évalue la fonction objective en utilisant le processus gaussien (GP), qui est une alternative courante à la modélisation de la fonction objectif.

GP est une extension générale de la distribution gaussienne en général, GP définit une distribution a priori sur la fonction, puis met à jour cette distribution en fonction des observations, cela peut être transformé en une distribution ultérieure sur les fonctions.

Dans cette technique, il est supposé que la fonction $f(x)$ est une réalisation de GP avec une fonction moyenne $m(x)$ et une fonction de variance $k(x, x_i)$ comme indiqué ci-dessous, où x est la valeur de la fonction avec n'importe quelle paire potentielle de (x_i, x_j) dans l'espace d'entrée, les détails peuvent être trouvés ailleurs : [76]

$$f(x) \sim \text{GP}[m(x), k(x, x')] \quad \dots (3.55)$$

Chaque entrée de cette fonction est une variable liée aux autres variables de l'espace d'entrée, telles que celles déterminées par la fonction de variance.

La fonction de variance est un noyau qui contrôle la douceur et l'extension des échantillons du processus gaussien, d'un autre côté, la fonction d'acquisition dans BOA est basée sur les observations précédentes et cherche à maximiser sa valeur au fil des itérations.

Le modèle d'acquisition propose le point suivant de l'itération en utilisant les résultats du modèle alternatif. Mathématiquement, l'optimisation du hyper-paramètre peut être formulée en utilisant BOA comme l'équation (3.56).

$$g^* = \arg \min_{g \in G} f(g) \quad \dots (3.56)$$

Ici, (g) représente le degré de l'objectif de minimiser l'erreur (MSE ou RMSE), g^* est l'ensemble des hyper-paramètres qui génèrent la plus petite valeur de degré, et g peut être n'importe quelle valeur dans l'espace G .

- Modèles de surrogate (substitution) dans BOA : Comparaison entre GP et TPE.

a) Processus Gaussien (Gaussian Process - GP) : Le GP est largement utilisé en tant que modèle de substitution. Supposant l'équation (3.55). Après chaque observation, le modèle est mis à jour via la règle de Bayes. Pour sélectionner les prochains points, l'une des fonctions d'acquisition les plus connues est l'Amélioration Attendue (Expected Improvement - EI) :

$$EI_{y^*}(x) = \int_{-\infty}^{+\infty} \max(y^* - y, 0) P(y|x) dy \quad \dots (3.57) \text{ où :}$$

$EI_{y^*}(x)$ désigne la meilleure performance observée jusqu'à présent.

Cette formulation illustre l'équilibre entre exploitation des régions prometteuses et exploration de nouvelles zones de l'espace.

b) Estimateur de Parzen structuré arborescent (Tree-structured Parzen Estimator - TPE) : Le TPE constitue une alternative performante au GP, particulièrement efficace dans les espaces de grande dimension ou présentant des fonctions objectif complexes. Sa particularité est de reformuler directement la règle de Bayes : (voir l'équation (3.54)), où la densité conditionnelle est modélisée par deux distributions distinctes :

$$P(x|y) = \begin{cases} l(x), & \text{if } y < y^* \\ g(x), & \text{if } y \geq y^* \end{cases} \quad \dots (3.58) \text{ avec :}$$

$l(x)$: distribution des « bons résultats » (faible perte).

$g(x)$: distribution des « mauvais résultats » (forte perte).

Le seuil est déterminé par un quantile γ , $p(y < y^*) = \gamma$. Ainsi la fonction d'acquisition devient :

$$EI_{y^*}(x) = \int_{-\infty}^{y^*} (y^* - y) P(y|x) dy = \frac{\gamma y^* l(x) - l(x) \int_{-\infty}^{y^*} p(y) dy}{\gamma l(x) + (1-\gamma)g(x)} \alpha \left(\frac{1}{\gamma + \frac{(1-\gamma)g(x)}{l(x)}} \right) \quad \dots (3.59)$$

Cette formulation favorise les points ayant un faible rapport, ce qui accroît les chances d'amélioration substantielle.

- Applications Réelles du TPE : L'efficacité du TPE a été largement démontrée dans divers contextes:

- Détection de fraude bancaire. [77]
- Systèmes de détection d'intrusion IoT sur données déséquilibrées. [78]
- Prédiction de la résistance du béton recyclé. [79]
- Optimisation de la locomotion robotique via apprentissage par renforcement. [80]
- Développement de récepteurs acoustiques adaptatifs pour communications sous-marines. [81]
- Prédiction spatiale de l'ammoniac dans les sols. [82]
- Analyse épi-génétique du cerveau avec amélioration des performances prédictives. [83]

Ces résultats confirment la capacité du TPE à traiter des problèmes complexes et réels.

- Pourquoi BOA et TPE Sont Préférés à d'Autres Méthodes ?

Selon [84] et [85], la BOA se distingue dans les cas où : la fonction objectif est coûteuse à évaluer (par ex. entraînement de réseaux profonds), l'espace de recherche est de grande dimension, aucune information analytique sur les gradients n'est disponible. En pratique, BOA combinée au TPE surpasse: la recherche par grille, la recherche aléatoire, l'optimisation manuelle, ainsi que des méta-heuristiques comme PSO.

Dans notre étude, BOA de type (stratégie) TPE a même été intégrée aux deux modèles GPR et SVR avec une validation croisée en 10 plis (10-fold cross-validation), fournissant ainsi un cadre robuste et fiable pour l'optimisation globale des hyper-paramètres.

VI.2.5. Les étapes d'exécution d'un Algorithme d'Optimisation Bayésien BOA / TPE :

VI.2.5.1. BO avec (Processus Gaussien - GP) :

- Initialisation :

- Sélection d'un ensemble initial: Choisir un petit nombre de points dans l'espace de recherche.
- Évaluation de la fonction objective: Calculer les valeurs de la fonction objective pour ces points.

- Construction du modèle :

- Modèle probabiliste: Construire un modèle, souvent un processus gaussien, pour approximer la fonction objective.
- Estimation de la moyenne et de l'incertitude: Évaluer la moyenne et l'incertitude sur toute la surface de l'espace de recherche.

- Fonction d'acquisition :

- Calcul de la fonction d'acquisition: Utiliser des méthodes comme (EI) ou (UCB).
- Équilibre exploration/exploitation: La fonction d'acquisition doit équilibrer l'exploration des zones peu connues et l'exploitation des zones prometteuses.

- Choix du point suivant :

- Maximisation de la fonction d'acquisition: Sélectionner le point qui maximise cette fonction.
- Évaluation de la fonction objective: Calculer la valeur de la fonction objective à ce nouveau point.

- Mise à jour :

- Mise à jour du modèle: Incorporer les nouvelles données pour mettre à jour le processus gaussien.
- Répétition: Continuer le processus jusqu'à atteindre un critère d'arrêt (nombre d'itérations ou budget. [86])

VI.2.5.2. TPE (Tree-structured Parzen Estimator) :**- Initialisation :**

- Génération d'un ensemble initial: Effectuer un échantillonnage aléatoire pour obtenir un ensemble de points.
- Évaluation de la fonction objective: Calculer les valeurs de la fonction objective pour ces points.

- Construction des distributions :

- Division des observations: Séparer les configurations en deux groupes :
 - $l(x)$: avec de bonnes performances (meilleures qu'un seuil).
 - $g(x)$: Configurations avec de moins bonnes performances.
- Estimateurs de densité: Construire des estimateurs de densité (Parzen) pour chaque groupe.

- Sélection des candidats :

- Génération de nouveaux candidats: Proposer de nouveaux points à partir des zones les plus prometteuses.
- Évaluation de leur probabilité: Calculer la probabilité des nouveaux candidats sous $l(x)$ et $g(x)$.

- Choix du point :

- Maximisation du rapport: Sélectionner le candidat qui maximise le rapport $\left(\frac{l(x)}{g(x)}\right)$.
- Évaluation de la fonction objective: Calculer la valeur de la fonction objective à ce nouveau point.

- Mise à jour :

- Mise à jour des distributions: Mettre à jour les estimateurs $l(x)$ et $g(x)$ avec les nouvelles données.
- Répétition: Continuer le processus au fil des itérations. [87]

Ces étapes décrivent le fonctionnement général de chaque méthode d'optimisation, mettant en avant leurs approches respectives pour explorer et exploiter l'espace de recherche.

VI.2.6. Comparaison entre les deux stratégies de BOA (avec GP) et TPE : De manière générale, le TPE (Tree-structured Parzen Estimator) et le GP (Gaussian Process) sont des stratégies d'Optimisation Bayésienne BOA. Bien qu'ils puissent exister des divergences dans leurs spécificités, leur point commun réside dans la création d'un modèle de la fonction objective. Au final, les deux visent à optimiser les hyper-paramètres des modèles.

Le tableau 3.2 et les images de figure 3.30 (ci-dessous/jointes) illustrent une comparaison entre ces deux stratégies.

Tableau 3.2. Comparaison entre les deux stratégies de BOA (GP et TPE).

Critère	TPE	BOA (avec GP)
Concepts de base	- Utilise deux distributions : $l(x)$ et $g(x)$. - Sélectionne les points qui maximisent le ratio $\left(\frac{l(x)}{g(x)}\right)$.	- Modélise la probabilité inversée : $P(x, y)$.
Performance et efficacité	- Meilleure performance pour les espaces complexes (catégoriels, conditionnels, etc.). - Rapide et facilement parallélisable. - Moins précis pour modéliser des fonctions lisses.	- Très précis pour les fonctions coûteuses et lisses. - Moins performant avec des dimensions élevées. - Plus lent et difficile à paralléliser.
Applications adaptées	- Optimisation d'hyper-paramètres avec beaucoup de variables. - Espaces contenant des paramètres catégoriels ou conditionnels. - Outils Auto-ML.	- Optimisation de fonctions coûteuses à évaluer. - Convient aux espaces de faible dimension, continus et lisses. - Utilisé en recherche et simulations scientifiques.
Outils et bibliothèques	- Optuna, Hyperopt, Ray Tune. - Très utilisé dans les Framework Auto-ML. - Supporte très bien le parallélisme.	- Scikit-Optimize, GPyOpt, Spearmint, BoTorch. - Moins adapté au parallélisme. - Plus complexe à configurer dans certains cas.
Résumé de la comparaison	✓ Très bon pour les espaces complexes. ✓ Rapide, flexible et simple à implémenter. ✗ Moins précis dans certains cas.	✓ Excellente précision pour les problèmes lisses à faible dimension. ✗ Moins adapté aux variables catégorielles ou aux hautes dimensions.

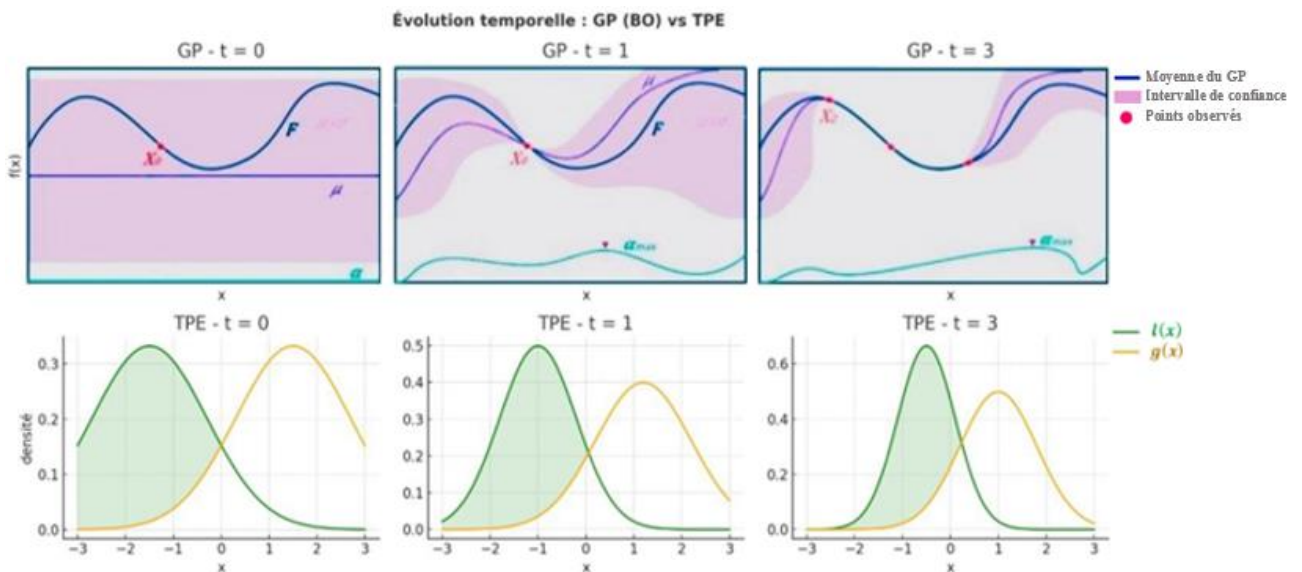


Figure 3.30. Exemple d'évolution temporelle de l'optimisation bayésienne BOA : GP vs TPE.

👉 Conclusion de la comparaison :

- Bayesian Optimization avec GP est préférable lorsque la fonction objectif est lisse, coûteuse à évaluer, et définie sur un espace continu de faible dimension. Construit un modèle probabiliste de la fonction cible (ex: Processus Gaussien). Utilise une fonction d'acquisition (ex: EI, UCB) pour guider l'exploration. Équilibre entre exploration (tester des zones inconnues) et exploitation (tester près des meilleures solutions connues).
- TPE est un excellent choix pour l'optimisation dans des espaces hétérogènes, avec des variables complexes ou catégorielles, et dans les systèmes Auto-ML. Sépare les observations en "bonnes" et "mauvaises". Construit deux modèles de densité (Parzen Estimators) : un pour les bonnes $l(x)$ et un pour les mauvaises $g(x)$. Choisit le prochain point à tester en maximisant le ratio $\left(\frac{l(x)}{g(x)}\right)$. Peut-être plus performant dans les espaces à haute dimension.

Le sujet de l'optimisation des modèles inspire de nombreuses personnes, voici les références principales pour ce sujet :

- La bibliothèque PyPi [86] et la bibliothèque GPyOpt [88] pour l'optimisation bayésienne.
- La bibliothèque HyperOpt [87] et un article de [89] pour l'implémentation de TPE.

VI.3. Hybridation des modèles : L'hybridation des modèles est un processus contrôlé et dirigé visant à améliorer les modèles par l'intégration d'algorithmes d'optimisation, de techniques appropriées de division des données, et l'optimisation de la base de données.

- Dans cette étude, les modèles GPR et SVR ont été sélectionnés pour subir ce processus.
 - Premièrement, la base de données a été divisée en un ratio de 80 % pour l'entraînement et 20 % pour le test.
 - Ensuite, la technique de K-fold Cross-Validation a été intégrée afin d'éviter le sur-apprentissage (over-fitting).
- Les deux modèles, GPR et SVR, ont été associés à des techniques de sélection de caractéristiques avancées de type "Wrapper-based" :
 - La technique RFE avec GPR.
 - La technique EFS avec SVR.

Le résultat a été la réduction de la base de données à 7 entrées à partir des 12 initiales. Il en a résulté l'obtention de deux modèles initiaux : **GPR-RFE** et **SVR-EFS**.

- Comme étape finale, un travail a été réalisé pour ajuster les hyper-paramètres des deux modèles (GPR-RFE et SVR-EFS) en les hybridant avec l'une des stratégies (types) de l'Algorithme d'Optimisation Bayésienne BOA, le TPE (Tree-structured Parzen Estimator). Ceci permet de déterminer le noyau (kernel) optimal et le plus efficace pour le développement et l'utilisation, ainsi que d'ajuster les paramètres représentant chaque noyau des deux modèles.

Par conséquent, cela permet de gagner du temps lors de l'entraînement et d'améliorer les performances. Finalement, les deux modèles résultant de cette hybridation et association sont:

TPE-GPR-RFE et **TPE-SVR-EFS**.

(Les détails du développement des deux modèles, les résultats de performance, ainsi que les applications, seront détaillés ultérieurement.)

VII. Critères d'évaluation des modèles : Dans le cadre de cette étude, afin de quantifier la performance prédictive des modèles, plusieurs mesures d'évaluation ont été sélectionnées : l'Erreur Quadratique Moyenne (MSE), la Racine Carrée de l'Erreur Quadratique Moyenne (RMSE), l'Erreur Absolue Moyenne (MAE), l'Erreur Absolue Moyenne en Pourcentage (MAPE) et le Coefficient de Détermination (R^2).

VII.1. Erreur Quadratique Moyenne (MSE) : Le MSE est une métrique de risque qui calcule la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles. Il s'agit d'une mesure fondamentale de la qualité d'un estimateur. Sa valeur est toujours non négative, et des valeurs plus proches de zéro indiquent une meilleure performance.

$$MSE = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i)^2 \quad \dots (3.60)$$

« Mauvaise valeur $+\infty$, meilleure valeur 0 »

VII.2. Racine Carrée de l'Erreur Quadratique Moyenne (RMSE) : Le RMSE est indicatif de la dispersion des erreurs. Il est obtenu en prenant la racine carrée du MSE. Comme le MSE, il est sensible aux grandes erreurs et des valeurs plus faibles sont préférables.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - Y_i)^2} \quad \dots (3.61)$$

« Mauvaise valeur $+\infty$, meilleure valeur 0 »

VII.3. Erreur Absolue Moyenne (MAE) : Le MAE représente la moyenne des déviations absolues entre les valeurs attendues (réelles) et les valeurs prédites. Il fournit une indication directe de l'ampleur des erreurs de prédiction, sans accentuer les grandes erreurs comme le fait le MSE/RMSE.

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_i - Y_i| \quad \dots (3.62) [90]$$

« Mauvaise valeur $+\infty$, meilleure valeur 0 »

VII.4. Erreur Absolue Moyenne en Pourcentage (MAPE) : Le MAPE est une mesure statistique de l'exactitude des prévisions. Il exprime la magnitude moyenne de l'erreur générée par le modèle prédictif, exprimée en pourcentage des valeurs observées réelles. Des valeurs de MAPE plus faibles sont souhaitables, car elles indiquent une plus grande précision prédictive.

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{X_i - Y_i}{Y_i} \right| \times 100\% \quad \dots (3.63) [91]$$

« Mauvaise valeur proche à 100%, meilleure valeur en voisin de 0% »

VII.5. Coefficient de Détermination (R^2) : Le R^2 mesure la proportion de la variance de la variable dépendante (sortie) qui est prédictible à partir des variables indépendantes (entrées). Il quantifie la qualité de l'ajustement du modèle aux données. Une valeur de R^2 approchant 1 indique un excellent ajustement du modèle, signifiant que le modèle explique une grande partie de la variabilité des données. Inversement, un R^2 proche de 0 suggère que le modèle n'explique que peu de la variance, et un R^2 négatif indiquerait que le modèle est pire qu'un simple modèle prédisant la moyenne.

L'équation de R^2 est reliée généralement au coefficient de corrélation r , tel que :

$$R^2 = (r)^2 = \left(\frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{[\sum_{i=1}^n (X_i - \bar{X})^2][\sum_{i=1}^n (Y_i - \bar{Y})^2]}} \right)^2 \dots (3.64)$$

« Mauvaise valeur $-\infty$, meilleure valeur $+1$, ou dépasse le seuil 80% »

L'équation (3.64) pour R^2 est souvent présentée sous une autre forme alternative, équivalente, basée sur la somme des carrés des résidus (SSR) et la somme totale des carrés (SST), tel que :

$$R^2 = 1 - \left(\frac{SSR}{SST} \right) \dots (3.65) \text{ où :}$$

$SSR = \sum_{i=1}^n (\hat{Y}_i - Y_i)^2$ et $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$, donc l'équation de R^2 se sera :

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \dots (3.66) [13]$$

Où, pour toutes les équations des erreurs et de R^2 :

X_i représente la valeur prédite pour l'observation i .

Y_i représente la valeur mesurée (réelle) pour l'observation i .

$\bar{X}$ représente la moyenne des valeurs prédites sur l'ensemble des n observations.

$\bar{Y}$ représente la moyenne des valeurs mesurées (réelles) sur l'ensemble des n observations.

$\hat{Y}_i$ représente les valeurs estimées (expliquées) par le modèle de régression.

n dénote le nombre total d'enregistrements de données (observations).

Cette dernière formulation de l'équation (3.66) met en évidence que R^2 mesure la réduction proportionnelle de l'erreur (variance non expliquée) obtenue par le modèle par rapport à un modèle naïf qui prédit simplement la moyenne.

Pour l'étude statistique et les métriques d'évaluation de performances des modèles, un livre de Régis Bourbonnais qui parle sur l'économétrie est proposé [13].

VIII. Analyse de sensibilité (Impact des entrées sur la sortie CBR) :

Dans le domaine de l'apprentissage automatique, lorsqu'un modèle prédictif est construit (tel que TPE-GPR-RFE ou TPE-SVR-EFS dans notre exemple), il utilise un ensemble d'entrées (features) pour prédire une sortie spécifique (par exemple, la valeur du CBR dans notre cas). Le test de sensibilité constitue une méthodologie essentielle visant à quantifier et comprendre comment les variations de chaque entrée individuelle affectent la prédiction finale du modèle.

Cette analyse est cruciale pour plusieurs raisons :

- Interprétabilité du modèle : identification des facteurs les plus déterminants dans les prédictions, ce qui rend les modèles plus transparents.
- Fiabilité des prédictions : détection d'une éventuelle dépendance excessive à une entrée peu fiable.
- Optimisation des caractéristiques : guidage dans la sélection ou l'ingénierie des variables les plus pertinentes.

- Confiance dans le modèle : meilleure compréhension du comportement du modèle, renforçant ainsi la confiance dans ses résultats.

L'analyse de sensibilité est une discipline qui étudie comment les variables d'entrée (inputs) influencent les sorties (outputs) d'un modèle mathématique, d'une simulation ou d'un modèle d'apprentissage automatique. [92]

- **L'idée principale** : Si une petite variation d'une entrée entraîne une forte variation de la sortie → la variable est fortement influente. Sinon → la variable est moins importante.

VIII.1. Types d'Analyse de Sensibilité :

VIII.1.1. Analyse de sensibilité locale :

- Étudie les effets autour d'un point spécifique dans l'espace des variables, exemple : calcul de dérivées partielles en un point donné.
- Utile pour une compréhension locale, mais limitée pour une vision globale. [93]

VIII.1.2. Analyse de sensibilité globale :

- Évalue l'impact des variables sur l'ensemble du domaine des valeurs possibles.
- Prend en compte les interactions entre variables.
- Bien adaptée aux modèles complexes et non linéaires. [94]

VIII.2. Effet d'ordre 1 et Effet d'interaction :

a) Effet d'ordre 1 : c'est l'impact direct d'une seule variable d'entrée sur la sortie, exemple : si l'indice de sensibilité pour une entrée X_1 est : $S\{X_1\} = 0.40$, donc $S\{X_1, Y \text{ (la sortie)}\} = 0.40$.

b) Effet d'interaction : c'est l'influence indirecte issue de l'interaction entre deux variables ou plus, exemple : si l'interaction entre deux entrées X_1 et X_2 est forte, cela signifie que la variation de la sortie Y est due à leur effet conjoint, et non à X_1 ou X_2 pris séparément.

- Effet d'ordre 1 → effet individuel et direct.
- Effet d'interaction → effet collectif issu de la combinaison de variables.

Donc, dans l'analyse de sensibilité : on calcule l'indice d'ordre 1 pour mesurer l'effet direct, et on calcule l'indice d'effet total pour inclure l'effet direct + les interactions. La différence entre les deux représente l'effet d'interaction. [94]

VIII.3. Méthodologies et les techniques principales d'analyse de sensibilité :

VIII.3.1. One-At-a-Time (OAT) : Cette méthode implique la variation d'une seule variable à la fois, en maintenant les autres constantes, afin d'évaluer l'effet de cette variable sur le résultat. Bien que cette approche soit simple, elle ne permet pas de détecter les interactions entre les variables, ce qui peut être un inconvénient majeur dans des systèmes complexes. [95]

VIII.3.2. Méthodes basées sur les dérivées : Ces méthodes font usage des dérivées partielles pour évaluer l'impact des petites modifications sur les sorties du modèle. Elles s'avèrent pertinentes pour des changements minimes, mais présentent des limitations lorsque le modèle est fortement non linéaire, car elles peuvent ne pas capturer la complexité des relations entre les variables. [96]

VIII.3.3. Méthodes de screening (ex. méthode de Morris) : Les méthodes de screening, telles que celle proposée par Morris, sont particulièrement efficaces lorsque le nombre de variables à évaluer est élevé. Elles visent à identifier rapidement les variables les plus influentes, réduisant ainsi le besoin d'effectuer des analyses exhaustives et en profondeur pour chaque variable. [97]

VIII.3.4. Méthodes basées sur la variance (ex. indices de Sobol) : Ces techniques consistent à décomposer la variance totale de la sortie en contributions attribuables à chaque variable ainsi qu'à leurs interactions. Les indices de Sobol permettent de quantifier non seulement l'effet d'ordre un (première ordre) de chaque variable, mais également l'effet d'interaction, offrant ainsi une vue d'ensemble complète des dynamiques en jeu dans le modèle. [98]

VIII.3.5. Méthodes modernes liées au ML :

VIII.3.5.1. SHAP (SHapley Additive exPlanations) : La méthode de SHAP est une technique d'interprétation des modèles de ML qui permet de quantifier l'impact de chaque variable sur une prédiction spécifique. Elle repose sur le concept des valeurs de Shapley, issues de la théorie des jeux coopératifs, afin d'évaluer de manière systématique la contribution individuelle de chaque variable à la sortie du modèle. Cela permet d'identifier quelles variables influencent le plus la décision du modèle, tout en fournissant une interprétation transparente des résultats.

VIII.3.5.2. Importance par permutation : La méthode d'importance par permutation consiste à mesurer la dégradation des performances d'un modèle après avoir mélangé aléatoirement les valeurs d'une variable spécifique. En comparant la performance du modèle avant et après cette permutation, il est possible d'évaluer la contribution de la variable à la précision globale du modèle. Cette méthode est simple à mettre en œuvre et permet d'obtenir des indicateurs d'importance des variables tout en respectant l'intégrité du modèle original.

VIII.3.5.3. LIME (Local Interpretable Model-agnostic Explanations) : LIME est une technique d'explication localisée qui vise à rendre les prédictions des modèles de ML plus compréhensibles. Elle fonctionne en perturbant les données d'entrée d'un exemple spécifique et en observant comment cela affecte la sortie du modèle. LIME construit ensuite un modèle interprétable simple, tel qu'un modèle linéaire, pour approximativement représenter le comportement du modèle complexe autour de cette instance. Cela permet d'obtenir des explications locales sur la prédiction et d'identifier les caractéristiques qui ont conduit à cette sortie, tout en n'étant pas limité à un type de modèle particulier.

VIII.4. Application à la prédiction du CBR : Dans cette étude, l'analyse de la sensibilité de la sortie CBR aux 12 variables d'entrée a été effectuée à l'aide de deux approches distinctes :

1. La méthode de SHAP avec le modèle TPE-GPR-RFE. Cette approche permet de déterminer quelle variable a le plus grand impact sur la prédiction du CBR tout en expliquant les prédictions de manière transparente.
2. Analyse de sensibilité de Sobol avec le modèle TPE-SVR-EFS. Cette méthode complémentaire permet d'évaluer l'impact de chaque variable sur la sortie tout en tenant compte de la robustesse et de la cohérence des résultats obtenus.

Ces approches sont cruciales pour identifier les entrées les plus influentes et offrir des insights pratiques sur la prédiction du CBR.

VIII.5. Présentation théorique de la méthode de SHAP : L'interprétabilité est un aspect crucial des modèles d'apprentissage automatique. À cet égard, la méthode de SHAP, proposée par Lundberg et ses collègues en 2017, s'avère particulièrement efficace pour analyser ces modèles. [99]

L'analyse de SHAP est dérivée des valeurs de Shapley issues de la théorie des jeux coopératifs. Les valeurs de Shapley sont utilisées pour mesurer l'impact marginal de chaque caractéristique sur le résultat de la prédiction. Plus précisément, Shapley a défini la valeur d'une caractéristique f comme la contribution marginale moyenne de f sur tous les sous-ensembles possibles de caractéristiques et a proposé une formule pour cette valeur.

En pratique, la valeur de Shapley est résolue par approximation, tandis que SHAP est un algorithme d'optimisation qui estime cette valeur.

La méthode de SHAP représente le modèle d'explication basé sur les valeurs de Shapley sous la forme d'une attribution additive des caractéristiques, comme montré ci-dessous :

$$g(z') = \psi_0 + \sum_{j=1}^M \psi_j z'_j \quad \dots (3.67) \text{ où :}$$

$g(z')$ représente la fonction d'interprétation de z' .

z' est un vecteur binaire (0/1) de longueur M , où 1 indique la présence d'une caractéristique et 0 son absence. M correspond au nombre de caractéristiques dans le modèle.

ψ_0 représente le terme constant, c'est-à-dire la valeur de prédiction moyenne sur tous les échantillons.

ψ_j représente la valeur de Shapley (l'impact moyen pondéré) pour la caractéristique j .

La formule fondamentale pour calculer la valeur de Shapley ψ_j est la suivante :

$$\psi_j = \sum_{S \subseteq \{x_1, x_2, \dots, x_p\}} \frac{|S|!(p-|S|-1)!}{p!} (f_x(S \cup \{x_j\}) - f_x(S)) \quad \dots (3.68) \text{ avec :}$$

$\{x_1, x_2, \dots, x_p\}$ représente l'ensemble de toutes les caractéristiques d'entrée.

p représente le nombre total de caractéristiques d'entrée.

S représente un sous-ensemble des caractéristiques d'entrée (excluant x_j).

$f_x(S \cup \{x_j\})$ représente la valeur prédite lorsque la caractéristique x_j est incluse avec le sous-ensemble S .

$f_x(S)$ représente la valeur prédite pour le sous-ensemble de caractéristiques S seul.

La somme est effectuée sur tous les sous-ensembles possibles S qui n'incluent pas x_j . Le terme factoriel pondère chaque sous-ensemble pour garantir une attribution juste.

La méthode SHAP est une approche d'attribution largement reconnue dans le domaine de l'apprentissage automatique interprétable ces dernières années. Comparée à d'autres méthodes, elle offre l'avantage d'être agnostique au modèle. [100]

Cela signifie qu'elle peut théoriquement expliquer n'importe quel modèle d'apprentissage automatique, indépendamment de son architecture interne. Cette caractéristique confère à SHAP une flexibilité et une scalabilité considérables, lui permettant d'interpréter les algorithmes d'apprentissage automatique sous divers angles. [101] [102] [103]

VIII.6. Présentation théorique de la méthode d'analyse de Sobol : La méthode porte le nom du mathématicien Russe Ilya Mikhailovich Sobol (1926–2023). Spécialiste des méthodes de Monte Carlo, il a introduit une approche novatrice permettant de réaliser une analyse de sensibilité globale. Cette technique vise à quantifier, de manière rigoureuse, l'influence des variables d'entrée sur la variance d'un modèle complexe.

L'idée principale de cette technique, on considère un modèle (ou une fonction) :

$$Y = f(X_1, X_2, \dots, X_n) \quad \dots (3.69) \text{ avec :}$$

X_i sont les variables d'entrée (features),
 Y est la sortie du modèle.

L'objectif de la méthode de Sobol est de décomposer la variance totale de la sortie en contributions provenant, de :

- Chaque variable considérée individuellement.
- Interactions entre plusieurs variables.

Formellement :

$$Var(Y) = \sum_i V_i + \sum_{i < j} V_{ij} + \sum_{i < j < k} V_{ijk} + \dots + \sum_{i < j < \dots < n} V_{ij\dots n} \quad \dots (3.70) \text{ tel que :}$$

V_i est un part de variance expliquée par la variable seule.

V_{ij} est un part de variance due à l'interaction entre X_i et X_j ..., etc. [98]

VIII.6.1. Les indices de Sobol :

a) Indice de premier ordre :

$$S_i = \frac{V_i}{Var(Y)} \quad \dots (3.71)$$

b) Indice d'ordre total :

$$S_{T_i} = \frac{V_i + \sum_{i \neq j} V_{ij} + \dots + \sum_{i \neq n} V_{in}}{Var(Y)} \quad \dots (3.72)$$

c) Indice d'interaction secondaire :

$$S_{ij} = \frac{V_{ij}}{Var(Y)} \quad \dots (3.73) [98]$$

VIII.7. Comparaison entre les deux méthodes (SHAP et Sobol) :

- SHAP :

- Outil d'explication des prédictions des modèles de machine learning ML.
- Attribue une contribution à chaque variable pour chaque prédiction, facilitant l'interprétation locale.
- Permet de visualiser l'impact positif ou négatif des caractéristiques sur la décision du modèle.

- Sobol :

- Méthode d'analyse de sensibilité globale, mesurant l'impact des variables d'entrée sur la variance des résultats.
- Fournit des indices qui montrent l'interaction entre les variables.
- Utile pour obtenir une idée d'ensemble de la sensibilité du modèle, mais peut nécessiter des échantillons plus importants pour des systèmes complexes.

IX. Organigrammes de modélisation :

IX.1. Modélisation des modèles conventionnels (par défaut) : Les modèles de (CML) sélectionnés dans cette étude sont : régression multilinéaire (MLR), Multicouches Perceptron (MLPNN) de ANN et les forêts aléatoires (RF).

Le schéma illustré à la figure 3.31 montre les étapes de modélisation des modèles non hybrides, c'est-à-dire ceux développés sans optimisation ni ajustement de leurs hyper-paramètres, et sans les combiner avec des techniques de sélection avancée de caractéristiques.

- À l'exception du fait qu'ils ont été intégrés avec la technique de k-fold Cross Validation à 10 plis pour éviter le sur-apprentissage (over-fitting).
- La base de données a été divisée entre 80 % pour l'entraînement et 20 % pour le test.

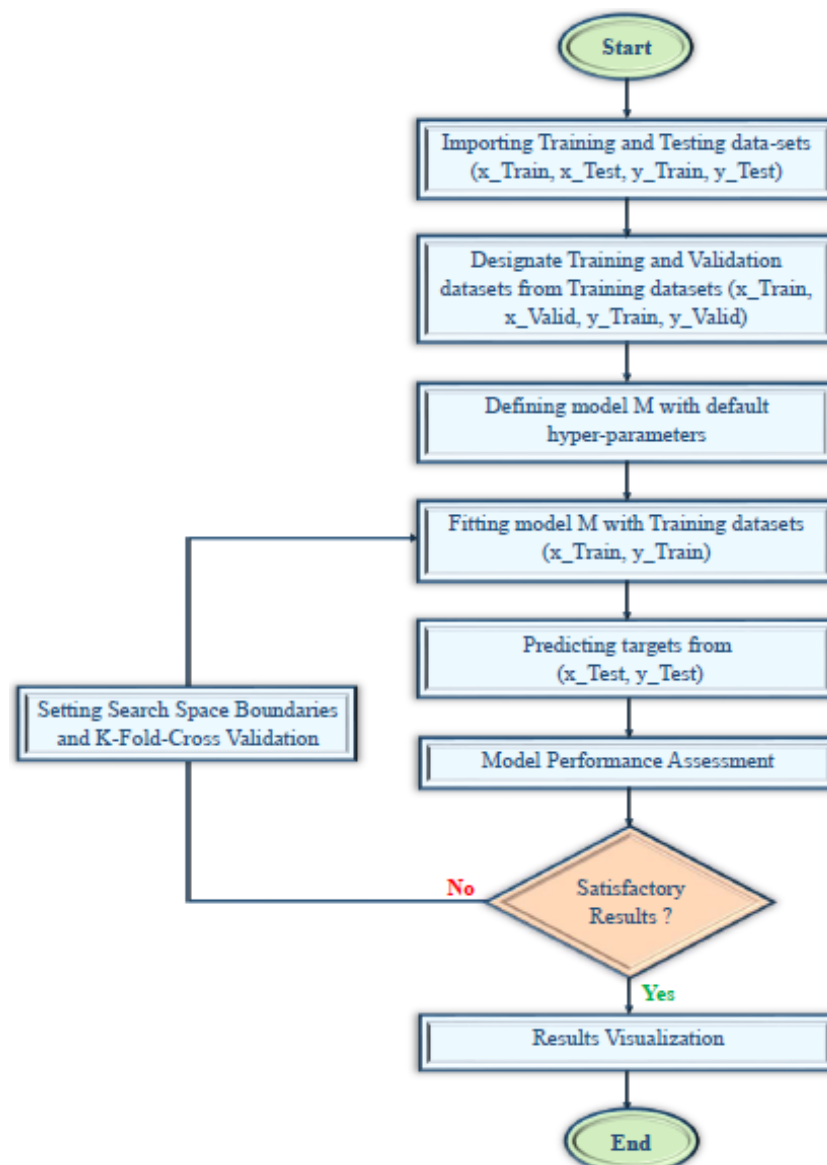


Figure 3.31. Organigramme de la modélisation des modèles conventionnels (CML) dans notre étude.

IX.2. Modélisation des modèles hybrides : Ce point concerne la modélisation des modèles GPR et SVR. Il s'agit d'un processus hiérarchique séquentiel commençant par :

- Diviser la base de données en 80 % pour l'entraînement et 20 % pour le test.
- Ensuite, incorporer la technique de k-fold Cross Validation à 10 plis pour ajuster l'apprentissage sans mémorisation, évitant ainsi le sur-apprentissage (over-fitting).
- Intégrer les méthodes de sélection avancée de caractéristiques de type Wrapper-based par l'intégration du technique RFE avec le modèle GPR et du technique EFS avec le modèle SVR, dans le but de réduire la taille base de données de 12 entrées à seulement 7. Le résultat est deux modèles initiaux : **GPR-RFE** et **SVR-EFS**.
- En dernière étape, les deux modèles initiaux (GPR-RFE et SVR-EFS) ont été associés à la technique TPE (Tree-structured Parzen Estimator), l'une des techniques d'optimisation bayésienne (BOA), dans le but d'ajuster ses hyper-paramètres et de sélectionner leurs meilleurs noyaux pour cette étude.

IX.2.1. Algorithme d'exécution des techniques RFE et EFS : La figure 3.32, représentant les étapes d'exécution des techniques RFE et EFS.

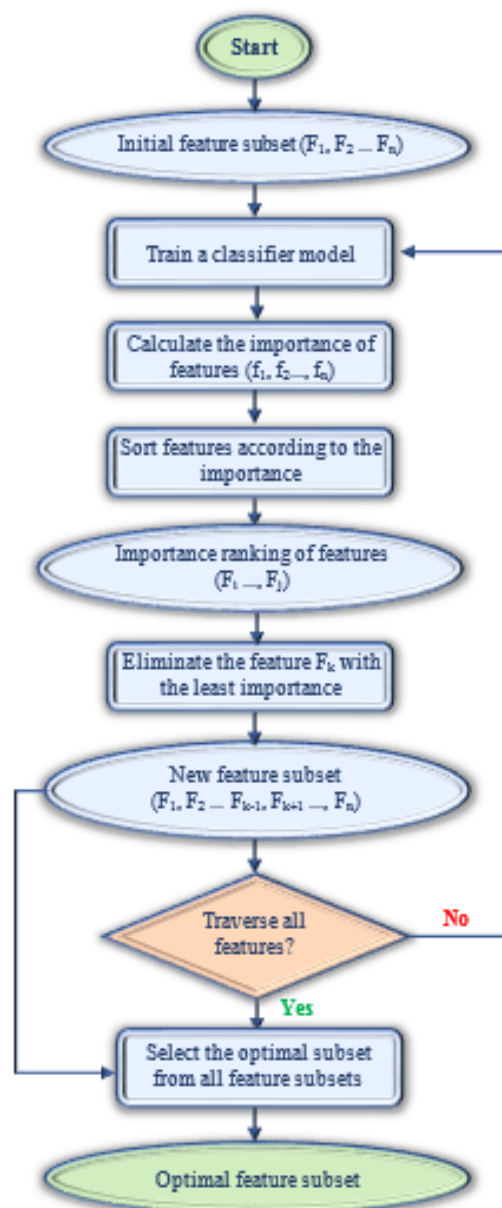


Figure 3.32. Organigramme d'exécution des techniques de RFE et EFS.

IX.2.2. Algorithme d'exécution de technique d'optimisation TPE : La figure 3.33, représentant les étapes d'exécution de technique d'optimisation TPE l'un des types de l'optimisation bayésienne BOA.

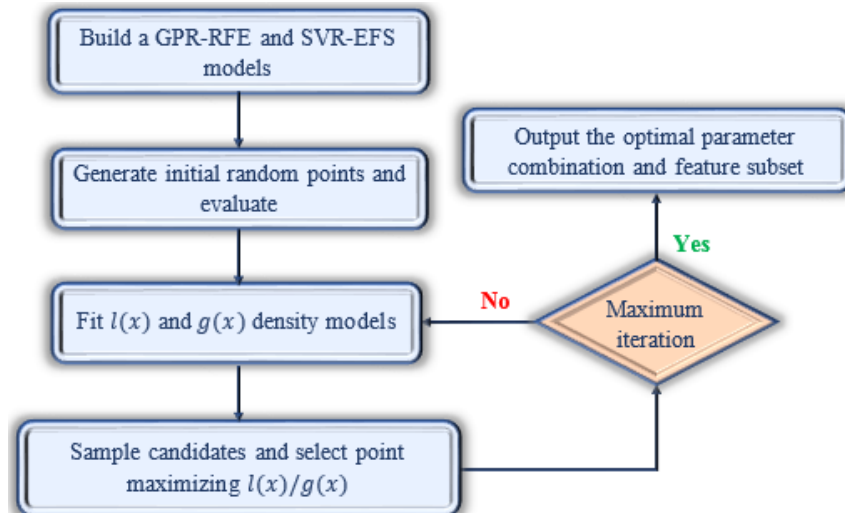


Figure 3.33. Organigramme d'exécution de technique d'optimisation TPE.

IX.2.3. Algorithme global de l'étude : La figure finale 3.34 représente une vue d'ensemble des étapes dans notre étude.

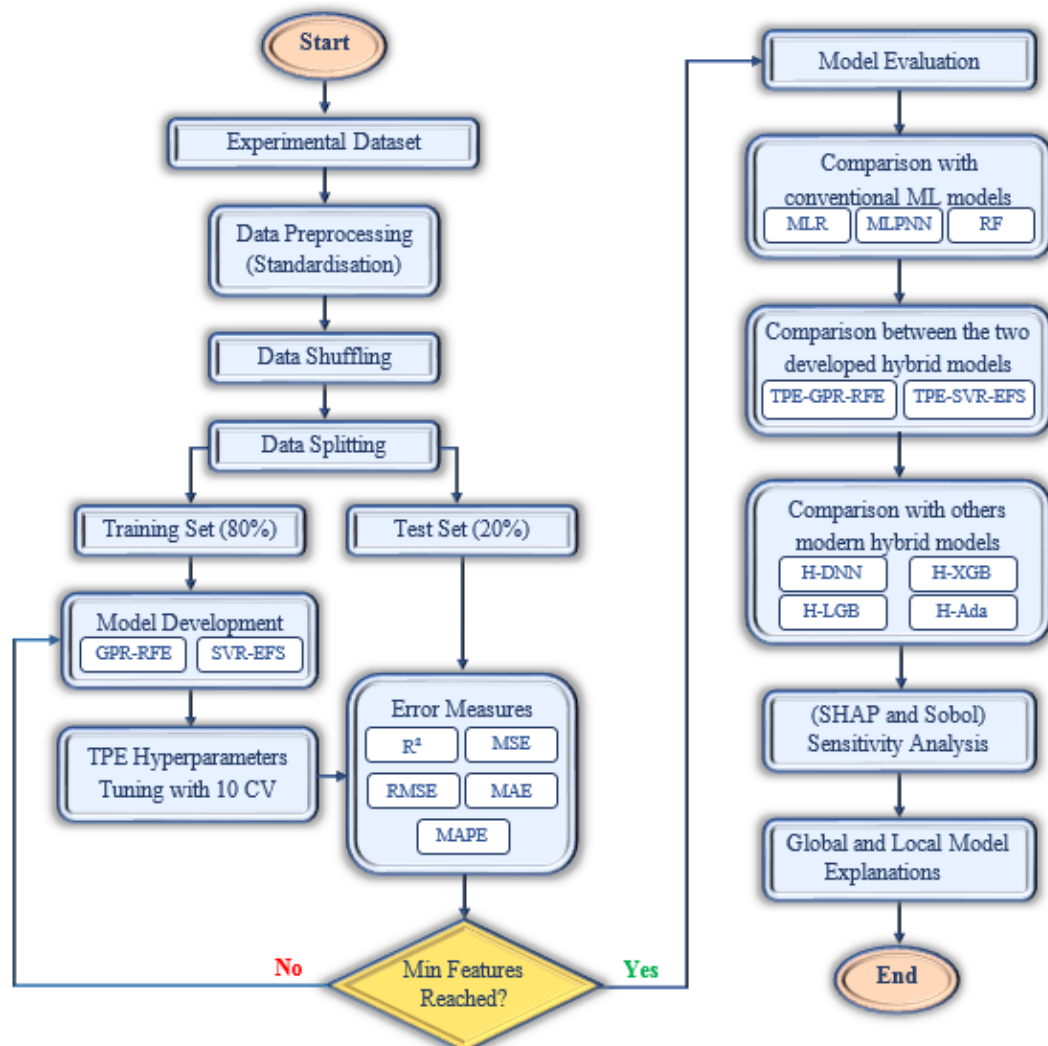


Figure 3.34. Organigramme de la méthodologie de cette étude.

X. Conclusion :

Le développement des modèles nécessite un environnement de programmation varié, nous avons adopté le langage de programmation Python comme outil pour l'apprentissage des modèles.

Les modèles d'apprentissage automatique sont nombreux, mais notre objectif était de prédire les valeurs de CBR nous sommes donc sous la catégorie de l'apprentissage supervisé (régression). En plus, chaque modèle a sa propre architecture, noyau et théories qui le distinguent des autres modèles, et chaque modèle contient des hyper-paramètres lorsqu'ils sont modifiés, affectent les performances du modèle.

Initialement en modélisant divers modèles d'apprentissages automatiques conventionnels CML (par défaut et sans optimisations ou hybridations), ces modèles sont : régression multilinéaire (MLR), Multicouches Perceptron (MLPNN) d'ANN, Forêts Aléatoires (RF) et même le modèle de GPR par défaut.

Il est possible de constater que l'aspect mathématique théorique des modèles d'apprentissage automatique remonte à des décennies, certains remontant même à deux siècles, comme le modèle ANN, tandis que parmi les plus récents se trouve le modèle RF.

L'intégration de la technique de features selection (Wrapper-based) de type Récursive (RFE) avec le modèle de GPR et Exhaustive (EFS) avec le modèle de SVR a permis de minimiser la taille de la base de données et optimiser la performance des modèles à l'entraînement et le test.

L'intégration de la technique de validation croisée K-fold améliore les performances du modèle en le dirigeant vers une division optimale de données entre l'entraînement et le test, de même l'hybridation des modèles avec l'algorithme d'optimisation bayésienne (BOA) de stratégie de l'estimateur de Parzen structuré arborescent (TPE) permet d'améliorer les performances des modèles et aide à déterminer le noyau optimal du modèle tout en ajustant ses hyper-paramètres à des valeurs optimales.

Finalement pour déterminer la sensibilité entre les entrées et la sortie ou bien l'impact des entrées sur la sortie, divers méthodes et techniques d'analyse de sensibilité existent, en choisissant la méthode de SHAP pour le modèle de GPR et l'analyse de Sobol pour le modèle de SVR.

Chapitre 4

***R**ésultats et optimisation*

I. Introduction : Dans cette étude, nous présentons une analyse complète des résultats de modélisation des trois modèles conventionnels sélectionnés (MLR, MLPNN de ANN, RF et même le GPR par défaut) et de modélisation des deux modèles hybridés (TPE-GPR-RFE et TPE-SVR-EFS), ces deux modèles ont été hybride avec l'algorithme d'optimisation Bayésienne de stratégie TPE (l'estimateur de Parzen structuré arborescent) en intégrant la technique de K-fold cross validation.

Egalement en présentant les résultats de ces deux modèles (TPE-GPR-RFE et TPE-SVR-EFS) avec la base de données sélectionnée, améliorée par un traitement utilisant deux techniques de sélection de caractéristiques supervisées du type Wrapper-based (RFE avec le modèle GPR et EFS avec le modèle SVR), qui a réduit les entrées de 12 à seulement 7 pour les deux modèles.

Les modèles sont évalués comme mentionné précédemment à l'aide du coefficient de détermination (R^2) et des valeurs d'erreur (MSE, RMSE, MAE et MAPE).

Les résultats comprennent également le choix optimal du noyau pour les deux modèles GPR et SVR, ainsi que l'ajustement de leurs hyper-paramètres qui correspondent à la performance optimale lors des phases d'entraînement et de test.

Les résultats montrent également des représentations graphiques des valeurs de R^2 et les résidus pour différents modèles.

Les commentaires d'évaluation sur les résultats de modélisation donnent un aperçu des performances et fournissent une image détaillée de chaque modèle, tandis que l'étude comparative des résultats distingue le modèle offrant les meilleures performances.

Enfin, nous avons réalisé des tests de sensibilité avec les deux modèles hybridés (TPE-GPR-RFE et TPE-SVR-EFS) pour évaluer l'impact des entrées sur la sortie souhaitée, le coefficient CBR.

II. Développement de modèles conventionnels (CML) : Les modèles conventionnels (CML) ne sont pas nécessairement inefficaces pour la prédiction, ni ne réagissent négativement aux améliorations et aux hybridations proposées dans cette étude. Ils ont été utilisés dans cette étude par défaut pour confirmer l'efficacité des autres modèles et leur réactivité aux techniques et méthodes d'optimisation proposées. Les modèles MLR, MLPNN et RF ont été utilisés dans leurs formes par défaut, ce qui signifie qu'ils ont été affinés manuellement en sélectionnant un noyau parmi tous les noyaux de modèles et en réglant automatiquement les hyper-paramètres pour évaluer leur capacité prédictive sans compromettre leurs éléments fondamentaux. De plus, le modèle GPR a également été utilisé dans sa forme par défaut pour examiner la réponse du modèle GPR hybride aux méthodes d'optimisation proposées dans cette étude.

II.1. Développement de modèle MLR : Le modèle de régression multilinéaire (MLR) est un modèle statistique utilisé pour analyser la relation entre une variable dépendante (variable cible) et plusieurs variables indépendantes explicatives, en établissant une équation linéaire.

Dans le cadre du développement d'un modèle MLR pour prédire les valeurs du CBR (California Bearing Ratio), il est nécessaire de définir clairement la variable dépendante (Y) et les variables indépendantes (X).

- **Variable dépendante (Y) :** Il s'agit de la valeur du CBR que l'on souhaite prédire.

- **Variables indépendantes (X) :** Ce sont les variables qui influencent la valeur du CBR. Dans ce cas, ce sont les 12 variables explicatives considérées comme entrées du modèle.

- **Formule du modèle MLR :** En fonction des variables disponibles, le modèle MLR peut s'écrire comme suit :

$$CBR = \beta_0 + \beta_1.G + \beta_2.S + \beta_3.FC + \beta_4.Cu + \beta_5.Cc + \beta_6.LL + \beta_7.LP + \beta_8.PI + \beta_9.MC + \beta_{10}.DD + \beta_{11}.OMC + \beta_{12}.MDD + \varepsilon \quad \dots (4.1) \text{ où :}$$

β_0 (Intercepte) : Représente le coefficient constant. C'est la valeur attendue du CBR lorsque toutes les variables indépendantes sont nulles, autrement dit, le point d'intersection de la droite de régression avec l'axe des ordonnées.

$\beta_1, \beta_2 \dots, \beta_{12}$: Ce sont les coefficients de régression associés à chacune des variables indépendantes. Par exemple :

- β_1 indique l'effet de la teneur en gravier G sur le CBR. Une valeur positive de β_1 signifie qu'une augmentation de la teneur en gravier G entraîne une augmentation du CBR, et inversement si β_1 est négatif.
- β_{12} quantifie l'impact de la densité sèche optimale MDD sur le CBR.
- Les autres coefficients représentent respectivement l'effet des autres variables indépendantes sur le CBR.

ε : C'est le terme d'erreur aléatoire ou résiduel, qui correspond à la différence entre la valeur observée du CBR et la valeur prédite par le modèle. Il reflète les influences non capturées par le modèle ainsi que les erreurs de mesure.

Dans cette étude, le modèle MLR est développé sans optimisation supplémentaire, mais la technique de validation croisée en 10-plis (K-fold cross validation) est intégrée pour assurer une évaluation plus robuste et fiable des performances du modèle.

- **Suggestion** : Les paramètres de ce régression $\beta_0, \beta_1 \dots, \beta_{12}$ et l'erreur ε , sont déterminés de manière computationnelle par des relations statistiques-mathématiques. Pour cette raison, les étapes peuvent être suivies en détail (le livre de Régis Bourbonnais [13] est proposé comme référence).

II.2. Développement de modèle MLPNN d'ANN :

Le modèle de réseaux de neurones ANN a été développé après l'intégration de la méthode de K-fold cross validation à 10-plis et après le choix optimal de manière manuelle de noyau et des hyper-paramètres de ce modèle (c'est-à-dire la fonction de noyau, nombre de couches, nombre de neurones dans chaque couche et la fonction d'activation) comme spécifié dans le tableau 4.1.

Après l'évaluation de la performance on a adopté que le noyau de Multicouches Perceptron d'ANN (MLPNN) a donné des meilleur résultat avec une couche d'entrée composée de 12 entrées (neurones) et une couche de sortie unique qui donne la valeur CBR, ainsi qu'une couche cachée de 8 neurones, comme illustré dans la figure 4.1.

La fonction sélectionnée pour le modèle MLPNN d'ANN est la sigmoïde.

Tableau 4.1. Les hyper-paramètres de modèle MLPNN d'ANN utilisé.

Paramètre	Paramètre optimal
Fonction de noyau	MLPNN
Nombre de couches cachées	1
Nombre de neurones	8
Fonction d'activation	Sigmoïde

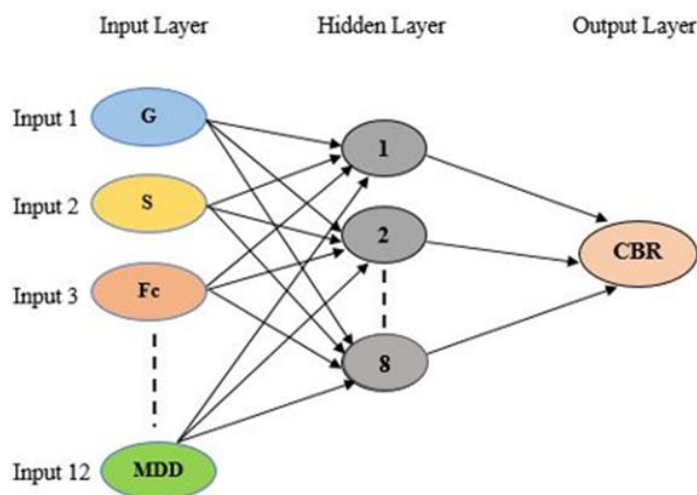


Figure 4.1. L'architecture de modèle MLPNN 12-8-1 d'ANN.

II.3. Développement de modèle RF : Le modèle de forêt aléatoire RF n'est pas comme les autres modèles, il possède des kernels multiples ou spécifiques et des hyper-paramètres particuliers. Il est construit sur la base du modèle d'arbre de décision (DT) et prend ses hyper-paramètres comme base, par exemple (profondeur de l'arbre, nombre de caractéristiques utilisées, nombre d'échantillons pour construire l'arbre ou la feuille de chaque modèle).

Le modèle RF basé sur le modèle DT est développé sans optimisation supplémentaire, mais la technique de validation croisée en 10-plis (K-fold cross validation) est intégrée pour éviter le sur-apprentissage (over-fitting).

III. Développement des modèles hybrides :

III.1. Cadre méthodologique pour le développement des modèles hybrides (TPE-GPR-RFE et TPE-SVR-EFS) : Dans cette partie de la recherche, un cadre méthodologique intégré a été développé pour la construction de modèles hybrides combinant les techniques d'apprentissage automatique et les méthodes d'optimisation Bayésienne probabiliste, dans le but d'obtenir des prédictions précises et fiables des valeurs du California Bearing Ratio (CBR), une propriété mécanique fondamentale utilisée pour évaluer les performances du sol sous charge.

Les modèles de régression par processus gaussiens (GPR) et de machines à vecteurs de support pour la régression (SVR) ont été choisis en raison de leur forte capacité à représenter les relations non linéaires entre les variables et de leur flexibilité dans la gestion des structures complexes via différentes fonctions noyau (kernels).

Chaque modèle possède un ensemble d'hyper-paramètres influençant directement sa précision et son efficacité computationnelle, tels que le paramètre du noyau, le coefficient de régularisation et la marge d'insensibilité pour le SVR, ainsi que les paramètres du bruit et du noyau blanc pour le GPR.

Afin d'éviter tout biais et d'assurer une comparaison scientifique rigoureuse entre les deux approches, la même architecture générale a été adoptée pour les deux modèles, mais avec des méthodes de sélection de variables différents.

Ainsi, le GPR a utilisé la méthode d'élimination récursive des variables (Recursive Feature Elimination – RFE), tandis que le SVR a employé la sélection exhaustive des variables (Exhaustive Feature Selection – EFS).

Dans les deux cas, l'algorithme d'estimateur de Parzen structuré arborescent (Tree-Structured Parzen Estimator (TPE)) a été utilisé dans le cadre de l'optimisation bayésienne pour ajuster intelligemment et efficacement les hyper-paramètres.

III.2. Développement du modèle hybride TPE-GPR-RFE :

▪ **Étape 1 :** Préparation et partition des données.

- Une base de données expérimentale a été constituée à partir de mesures de laboratoire précises des valeurs de CBR et des propriétés de plasticité, de compactage et de granulométrie des sols.
- Les données ont été divisées en 80 % pour l'entraînement et 20 % pour le test, permettant au modèle d'apprendre sur la majorité de l'échantillon tout en conservant une partie indépendante pour évaluer sa capacité de généralisation.
- Les données ont été prétraitées afin d'uniformiser les échelles et d'éliminer les valeurs aberrantes susceptibles d'affecter la stabilité statistique du modèle.

▪ **Étape 2 :** Validation croisée (10-Fold Cross-Validation).

- Une validation croisée à dix plis a été adoptée pour garantir la stabilité du modèle et réduire le sur-apprentissage (over-fitting). À chaque itération, le modèle est entraîné sur neuf sous-ensembles et testé sur le dixième, puis la moyenne des indicateurs de performance tels que MSE et R^2 est calculée. Cette approche permet une évaluation objective et équilibrée des performances du modèle sur l'ensemble des données.

▪ **Étape 3 :** Élimination des variables (RFE).

- L'algorithme RFE (Recursive Feature Elimination) a été appliqué pour sélectionner les variables les plus influentes sur le CBR. À chaque itération, le modèle est entraîné sur l'ensemble des variables restantes, puis la variable la moins significative est éliminée progressivement jusqu'à obtenir la combinaison optimale. Où la performance est idéale, c'est-à-dire que R^2 est élevé et MSE est faible. Il est certain que ces valeurs se sont manifestées avec la septième entrée, après quoi la performance a décliné et les valeurs ont commencé à diminuer.
- Les résultats ont montré que sept variables constituaient le jeu optimal avant que les indicateurs de performance ne se dégradent à cause du sur-apprentissage.

Ces variables sont : G, FC, PI, MC, DD, OMC, MDD (voir la figure 4.2 dans la page suivante).

▪ **Étape 4 :** Ajustement des hyper-paramètres (TPE – Optimisation bayésienne).

- Après sélection des variables optimales, les paramètres du GPR ont été ajustés via l'optimisation bayésienne utilisant l'algorithme TPE. Ce processus a permis d'optimiser les hyper-paramètres du modèle de GPR-RFE, le bon choix de noyau de modèle et même ajuster le bruit et le noyau blanc afin de minimiser les erreurs et maximiser le R^2 simultanément.
- Cent itérations ont été réalisées, mettant à jour à chaque fois les distributions de probabilité pour explorer efficacement l'espace des paramètres.

Le tableau 4.2 présente les domaines de variation des hyper-paramètres et les types de noyaux testés, ainsi que le noyau optimal et les valeurs des hyper-paramètres ajustées obtenues par l'optimisation de TPE pour le modèle de GPR.

Tableau 4.2. Les hyper-paramètres optimisés pour le modèle GPR utilisé.

Paramètre	Plage / type	Paramètres optimisés
Fonction de noyau	RBF, RQ, Matern, ...etc.	Rational Quadratique RQ
Longueur d'échelle (l)	$[10^{-2}, 10]$	0.6192
Echelle mixture (α)	$[10^{-2}, 10]$	1.5966
Le bruit	$[10^{-4}, 1]$	0.0406

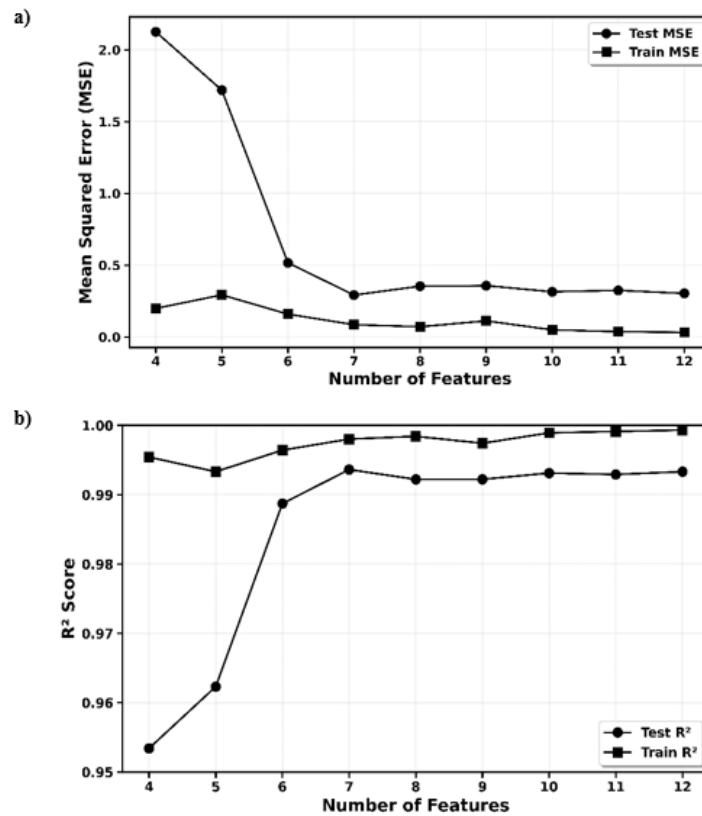


Figure 4.2. Résultats de la recherche du sous-ensemble optimal de caractéristiques
a) Résultats de MSE ; b) Résultats de R².

▪ **Étape 5 : Évaluation des performances et estimations d'incertitude.**

- Le modèle a été évalué à l'aide d'indicateurs globaux de précision et de variabilité :

MSE (Mean Squared Error) : moyenne des carrés des écarts entre valeurs réelles et prédites.

RMSE (Root Mean Squared Error) : écart-type moyen des erreurs.

MAE (Mean Absolute Error) : moyenne des écarts absolus entre valeurs prédites et observées.

MAPE (Mean Absolute Percentage Error) : erreur moyenne relative en pourcentage.

R² (Coefficient of Determination) : mesure du lien statistique entre valeurs réelles et estimées.

- Les résultats d'estimations de l'incertitude ont montré une forte cohérence entre valeurs réelles et prévues, avec une faible incertitude dans les zones à forte densité de données. Ce comportement indique que le modèle distingue clairement les zones de haute et de faible confiance, lui conférant une dimension interprétative et probabiliste dépassant les modèles purement déterministes.

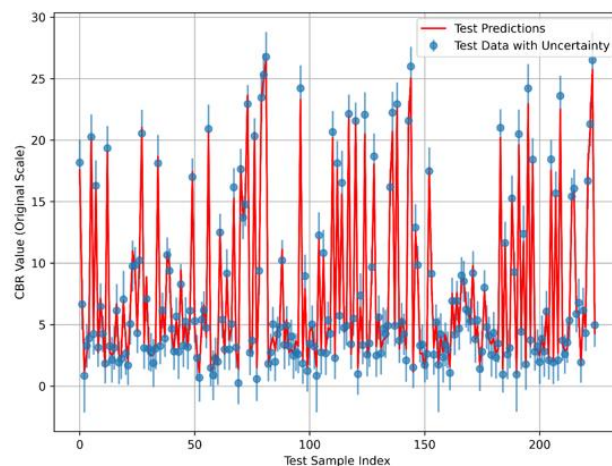


Figure 4.3. Test des prédictions avec estimations d'incertitude pour les valeurs CBR.

- **Conclusion partielle** : Le modèle TPE-GPR-RFE a atteint un équilibre optimal entre précision prédictive et interprétation probabiliste de l'incertitude, démontrant sa capacité à fournir des prédictions fiables tout en estimant le niveau de confiance associé à chaque prédiction. Ce modèle constitue la base probabiliste du cadre hybride global et sert de référence méthodologique pour le développement du modèle équivalent TPE-SVR-EFS ou d'autres modèles en phase d'hybridation ou d'optimisation.

III.3. Développement du modèle hybride TPE-SVR-EFS :

- **Étape 1** : Préparation et partition des données.
- Le modèle TPE-SVR-EFS a été développé en tenant compte de la nature fortement dépendante des données des modèles d'intelligence artificielle. Afin d'assurer une fiabilité élevée, des essais expérimentaux en laboratoire ont été réalisés pour générer une base de données représentant les propriétés des sols (plasticité, compactage, granulométrie) ainsi que les mesures de CBR.
- Parallèlement, le modèle a permis d'établir des statistiques descriptives et des distributions de résultats basées sur les valeurs mesurées et prédites. Ces analyses statistiques comparatives sont présentées dans le tableau 4.3 suivant :

Tableau 4.3. Résumé statistique des valeurs mesurées et des prédictions du modèle TPE-SVR-EFS.

Paramètre	Mesurés	TPE-SVR-EFS
Moyenne	26.63	26.65
Std. Déviation	12.91	13.68
Min	8.57	8.60
Max	70.20	89.23

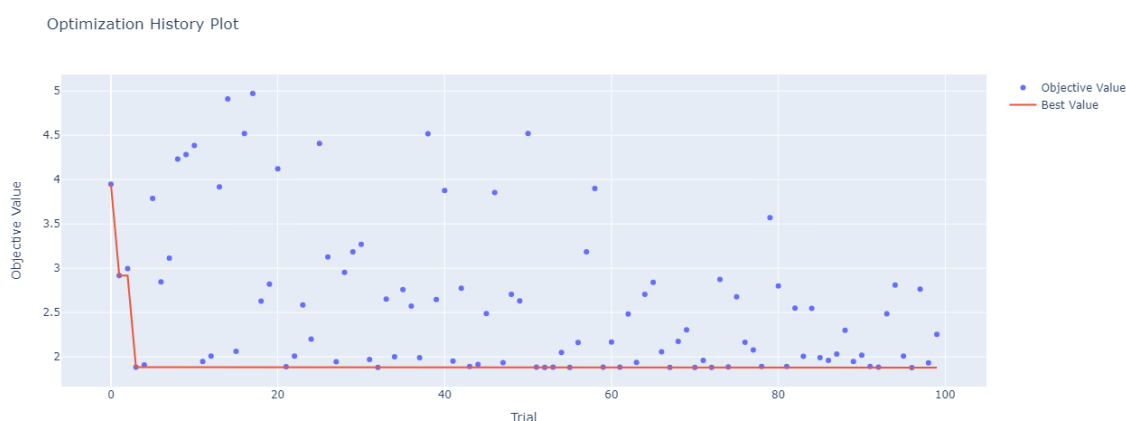
- **Étape 2** : Validation croisée (10-Fold Cross-Validation).
- Comme pour le GPR, une validation croisée à dix plis a été appliquée pour garantir la stabilité du modèle pendant l'apprentissage, réduire la variance due à la division des données et éviter le sur-apprentissage (over-fitting). Cette procédure a permis de construire un modèle robuste et généralisable à grande échelle.
- **Étape 3** : Sélection des variables (EFS).
- Le modèle repose sur l'algorithme EFS (Exhaustive Feature Selection), qui examine toutes les combinaisons possibles d'entrées pour identifier l'ensemble optimal minimisant le MSE.
- Les résultats ont montré que cinq des sept variables identifiées par le modèle GPR formaient également le jeu optimal pour le SVR.
Les sept variables sélectionnés sont : G, S, PL, PI, DD, OMC, MDD, confirmant ainsi la cohérence structurelle des données et la stabilité des corrélations entre variables.
- **Étape 4** : Ajustement des hyper-paramètres (TPE – Optimisation bayésienne).
- Les hyper-paramètres essentiels du SVR (le coefficient de régularisation (C), la marge d'insensibilité (ϵ), paramètre du noyau (γ) et la fonction de noyau (Kernel)), ont été ajustés via l'optimisation bayésienne à l'aide de la stratégie TPE, dans des intervalles préalablement définis.

Ces hyper-paramètres sont présentés dans le tableau 4.4 ci-dessous.

Tableau 4.4. Les hyper-paramètres optimisés pour le modèle SVR utilisé.

Paramètre	Plage / type	Paramètres optimisés
Fonction de noyau	RBF, Polynomial ...etc.	RBF
La marge d'insensibilité (ϵ)	$[10^{-5}, 10^5]$	0.001
Paramètre du noyau (γ)	$[10^{-5}, 10^5]$	8310
Régularisation (C)	$[10^{-5}, 100]$	0.003

- L'objectif principal était de minimiser le RMSE tout en tenant compte du temps de calcul et de l'efficacité algorithmique. Après 100 itérations, la meilleure configuration obtenue a conduit à la fonction objective (RMSE) égale à 1.88, a été observée après seulement 3 secondes indiquant une excellente efficacité computationnelle du modèle (voir la figure 4.4).

**Figure 4.4.** Le progrès de l'optimisation bayésienne des hyper-paramètres pour le modèle SVR.

- **Étape 5 : Évaluation des performances et analyse de distribution statistique.**
 - Les mêmes indicateurs que pour le modèle GPR ont été utilisés : les moyennes d'erreurs (MSE, RMSE, MAE, MAPE) et le coefficient de détermination (R^2).
 - Les résultats ont montré une forte concordance entre valeurs mesurées et prédites, attestant la capacité du modèle à capter à la fois les tendances globales et les détails fins des valeurs de CBR. Pour l'analyse visuelle des distributions des données mesurées et prévues, des diagrammes en boîtes (Box plots) ont été utilisées.

* **Diagramme en boîte (Box plot)** : Est un outil statistique introduit par John Tukey (1970) permettant de représenter la distribution d'un ensemble de données de manière synthétique et complète. Ce graphique repose sur les quartiles statistiques, où la boîte représente les valeurs comprises entre le premier et le troisième quartile (50 % des valeurs centrales), tandis que les "moustaches" indiquent les valeurs extrêmes inférieure et supérieure. Cette méthode permet une visualisation rapide de la tendance centrale, de la dispersion et des valeurs aberrantes (Outliers), complétant ainsi les indicateurs numériques (moyenne, écart type, coefficient de détermination).

- La figure 4.5 présente un diagramme en boîte comparant les valeurs mesurées et prédites du CBR obtenues par le modèle SVR. Dans chaque boîte, le bord inférieur correspond au premier quartile (Q1), le bord supérieur au troisième quartile (Q3) et la ligne horizontale au centre représente la médiane (Q2). Les moustaches indiquent l'étendue des valeurs minimales et maximales (hors valeurs aberrantes), tandis que les points situés au-delà des moustaches représentent les valeurs atypiques (outliers). On observe un fort recouvrement entre les deux boîtes, avec des médianes très proches et des intervalles interquartiles similaires, ce qui traduit une bonne concordance entre les

valeurs mesurées et prédites. La faible dispersion des données confirme également la stabilité et la fiabilité des prédictions fournies par le modèle SVR.

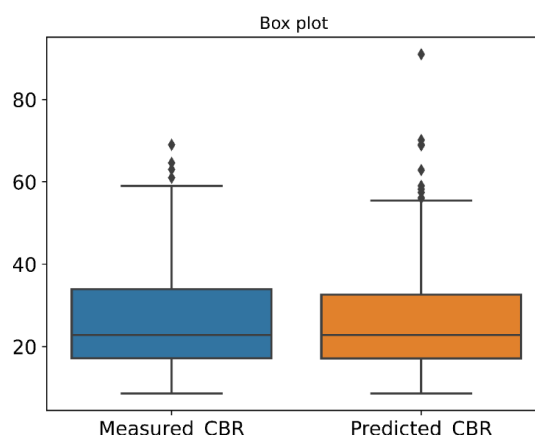


Figure 4.5. Diagramme en boîte des valeurs mesurées et prédites du CBR pour le modèle SVR.

- **Conclusion partielle :** Le modèle TPE-SVR-EFS a démontré une très haute précision prédictive combinée à une efficacité computationnelle remarquable, ainsi qu'une excellente concordance avec les données expérimentales mesurées. L'analyse statistique a révélé une quasi-identité des distributions, confirmant la capacité du modèle à représenter fidèlement les données réelles, faisant de lui un complément idéal au modèle TPE-GPR-RFE, spécialisé dans la représentation de l'incertitude.

III.4. Intégration méthodologique des deux modèles : L'intégration des modèles TPE-GPR-RFE et TPE-SVR-EFS met en évidence un équilibre méthodologique entre la représentation probabiliste (dans le GPR) et la précision numérique (dans le SVR). Le premier fournit une vision interprétative basée sur l'estimation de la confiance dans les prédictions, tandis que le second offre des performances prédictives optimisées avec une précision quantitative élevée.

Ce cadre dual combine ainsi la prédiction statistique et la certitude numérique, constituant un outil analytique intégré applicable aux domaines géotechniques et environnementaux sensibles, permettant d'évaluer à la fois la valeur prédite et le niveau de confiance associé.

IV. Évaluation des résultats de performance :

IV.1. Comparaison du modèle GPR hybride avec le modèle GPR par défaut :

La performance des modèles hybrides TPE-GPR-RFE et GPR par défaut a été vérifiée dans cette étude en mesurant le coefficient de détermination (R^2), qui est l'un des indicateurs les plus importants de l'efficacité statistique. Il est supposé que sa valeur dépasse le seuil de 80 % pour indiquer la fiabilité du modèle. Les métriques d'erreur ont également été évaluées pour s'assurer que leur valeur se rapprochait de zéro, reflétant ainsi la précision des prévisions.

Le modèle a montré une performance satisfaisante, avec une valeur initiale du coefficient de détermination (R^2) de 96 % lors de la validation, avec une diminution claire des erreurs même dans son état par défaut (GPR par défaut), qui s'appuie sur 12 entrées sans appliquer d'outils ou de techniques d'optimisation. Les figures 4.8 et 4.9 (avec une distribution des points sous forme des cercles rouges ● pour R^2 , et en forme de losanges gris ◆ pour les résidus) illustrent cette performance, à travers la convergence des distributions des points expérimentaux et prédictifs autour de la ligne de 45° dans le diagramme de R^2 et de la ligne zéro dans la distribution des résidus.

- **Remarque :** Les deux figures 4.8 et 4.9 ci-dessous dans la partie de la comparaison de performance de GPR hybride avec les modèles conventionnels.

Suite à l'intégration de la technique de sélection des entrées (RFE), qui a réduit le nombre d'entrées à seulement sept, et à l'hybridation du modèle avec une optimisation bayésienne utilisant la stratégie TPE, une amélioration notable de la performance du modèle a été enregistrée, avec une valeur de (R^2) portée à 99 %, et une baisse supplémentaire des métriques d'erreur MSE, RMSE, MAE et MAPE durant la phase de test. Ces résultats indiquent une amélioration qualitative de la performance du modèle hybride, comme la montre le tableau 4.5 détaillant cela pour les phases d'entraînement et de test de manière comparative pour les deux modèles, par défaut et hybride.

En termes de représentation visuelle, le modèle développé TPE-GPR-RFE a été utilisé pour estimer les valeurs CBR des sols de fondations combinés. La figure 4.6 montre une ligne de 45° représentant l'ajustement idéal, reflétant le lien fort entre les valeurs expérimentales et prédictives de CBR. Les valeurs du coefficient de détermination (R^2) pour les ensembles d'entraînement et de test ont atteint 99 %, ce qui indique une haute efficacité et adéquation du modèle.

Tableau 4.5. Les indices de performance comparée entre le modèle de GPR par défaut et l'hybride TPE-GPR-RFE à la partie d'entraînement et de test.

			Entrainement	Test
Conventionnel	GPR (Par défaut)	R^2	1.0000	0.9640
		MSE	0.0000	1.6400
		RMSE	0.0000	1.2806
		MAE	0.0000	0.9400
		MAPE	0%	23%
Hybride	TPE-GPR-RFE	R^2	0.9982	0.9935
		MSE	0.0768	0.2974
		RMSE	0.2771	0.5453
		MAE	0.2107	0.4364
		MAPE	5.32%	13.79%

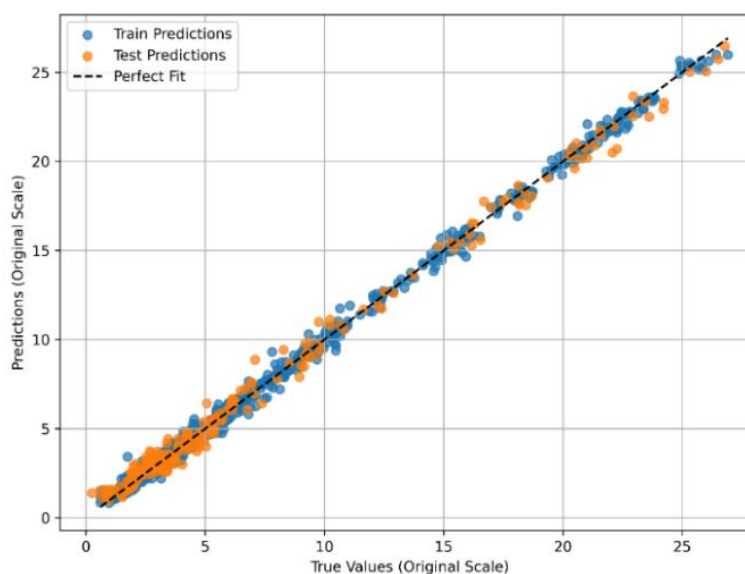


Figure 4.6. Le graphique de la ligne ajustée des valeurs prédites et mesurées du CBR pour le modèle hybride TPE-GPR-RFE.

Les valeurs expérimentales concordent largement avec les valeurs prédites, comme le montre l'analyse des résidus dans la figure 4.7, où la proximité des points à la ligne zéro renforce la fiabilité du modèle. De plus, des valeurs faibles ont été atteintes pour toutes les métriques d'erreur (MSE, MAE, RMSE, MAPE) dans les ensembles d'entraînement et de test, ce qui confirme la forte capacité prédictive du modèle et sa grande précision avec $R^2 \approx 1$.

Ainsi, les prévisions du modèle hybride TPE-GPR-RFE peuvent être considérées comme très précises, proche de la correspondance parfaite avec les résultats expérimentaux.

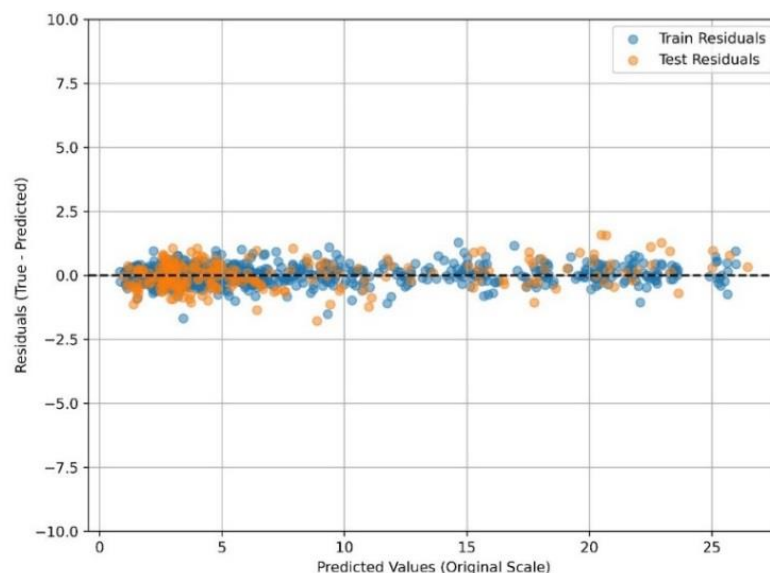


Figure 4.7. Nuage de points des résidus par rapport aux valeurs prédites pour le modèle hybride TPE-GPR-RFE.

IV.2. Comparaison de performance du modèle hybride TPE-GPR-RFE avec des modèles conventionnels (CML) : Après avoir démontré l'efficacité du modèle hybride TPE-GPR-RFE par rapport à son homologue par défaut (GPR par défaut), l'évaluation a été élargie pour inclure une comparaison de sa performance avec certains modèles d'apprentissage machine conventionnels (CML) développés par défaut, afin de confirmer sa supériorité en matière de prévision.

À cet effet, trois modèles supplémentaires développés dans les étapes précédentes ont été évalués: forêt aléatoire (RF), multicouches perceptron de réseaux de neurones (MLPNN) et multilinéaire régression (MLR).

Le tableau 4.6 ci-dessous présente les valeurs de R^2 et les moyennes d'erreurs (MSE, RMSE, MAE et MAPE) pour les deux phases d'entraînement et de test de ces trois modèles, de manière comparative avec le modèle TPE-GPR-RFE hybride.

Tableau 4.6. Les indices de performance comparée entre le modèle hybride TPE-GPR-RFE et les autres modèles conventionnels (CML) à la partie d'entraînement et de test.

			Entraînement	Test
Conventionnels	MLR	R^2	0.2334	0.1495
		MSE	31.5100	38.7800
		RMSE	5.6134	6.2274
		MAE	4.7000	4.8600
		MAPE	103.40%	96.53%
	MLPNN	R^2	0.9316	0.9189
		MSE	2.9900	3.7000
		RMSE	1.7292	1.9235
		MAE	1.1927	1.2400
		MAPE	27.44%	28.28%
	RF	R^2	0.9942	0.9558
		MSE	0.2520	2.0100
		RMSE	0.5020	1.4177
		MAE	0.2840	0.8400
		MAPE	7.01%	23.94%
Hybride	TPE-GPR-RFE	R^2	0.9982	0.9935
		MSE	0.0768	0.2974
		RMSE	0.2771	0.5453
		MAE	0.2107	0.4364
		MAPE	5.32%	13.79%

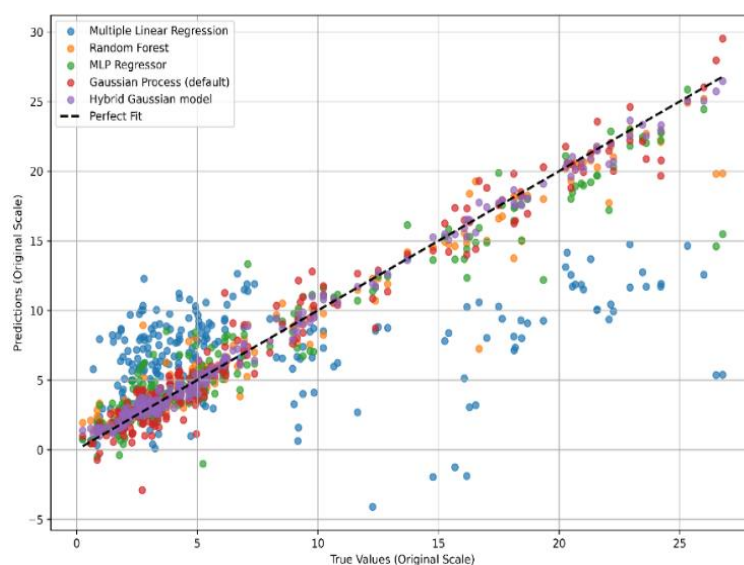


Figure 4.8. Comparaison des performances des modèles hybrides et conventionnels : valeurs CBR prédites vs valeurs CBR réelles.

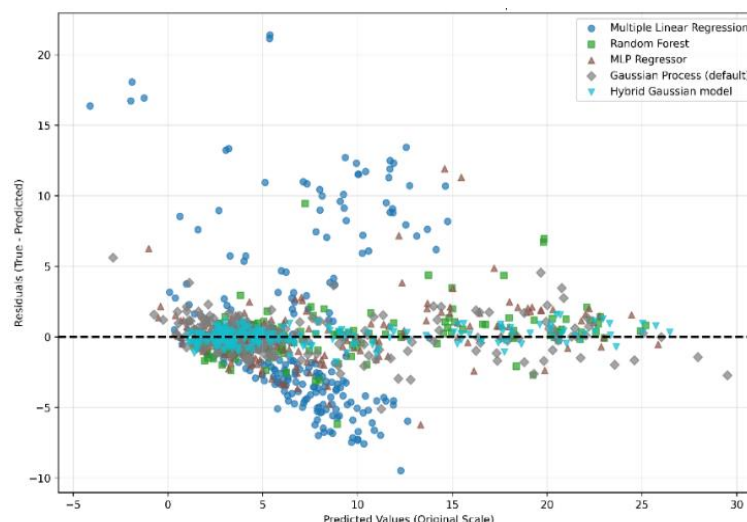


Figure 4.9. Analyse des résidus des modèles : résidus de prédiction à travers les cinq modèles de Machine Learning (hybride et CML).

IV.2.1. Performance du modèle de MLR : Le modèle MLR a montré des limites dans le traitement des relations non linéaires, ce qui a eu un impact négatif sur ses résultats ; en effet, la valeur R^2 n'a atteint que 14,95 % à la partie de test, bien en deçà du seuil minimum acceptable de 80 %. Le modèle a également enregistré des taux d'erreur élevés lors de le test (voir le tableau 4.6), ce qui indique la présence d'un bruit élevé et d'une dispersion dans les résultats. Cela est confirmé par la représentation visuelle avec une distribution des points sous forme des cercles bleus ● dans la figure 4.8 pour le coefficient de détermination (R^2), et ● dans la figure 4.9 pour les résidus, où l'on observe l'écart des points par rapport aux lignes de référence (45° et 0) respectivement.

Comparé au modèle hybride TPE-GPR-RFE, il est clair que ce dernier se distingue nettement en termes de précision et de stabilité.

IV.2.2. Performance du modèle de MLPNN : Le modèle MLPNN (Multilayer Perceptron Neural Network) est l'un des modèles de réseaux de neurones artificiels (ANN). En utilisant 8 neurones dans la couche cachée et une fonction d'activation Sigmoidale, le modèle a pu surmonter les limites de la linéarité et s'adapter aux données de l'étude avec une plus grande efficacité. Cela s'est traduit par une bonne performance, avec une valeur de R^2 atteignant 91 % et dépassant le seuil de référence (80%), et une diminution significative des valeurs d'erreur comme indiqué dans le tableau 4.6.

La figure 4.8 (en vert ●) montre la proximité des points prédictifs par rapport à la ligne de 45° , tandis que la figure 4.9 (avec une distribution des triangles marrons ▲) illustre la régularité des résidus autour de la ligne zéro, renforçant ainsi la fiabilité du modèle. Bien que sa performance soit satisfaisante sans optimisations supplémentaires, le modèle hybride GPR-TPE-RFE reste néanmoins supérieur en termes de précision globale.

IV.2.3. Performance du modèle de RF : Le modèle Random Forest (RF) se distingue par sa capacité à représenter les relations non linéaires grâce à sa multitude d'arbres et à ses processus d'agrégation. Ce modèle a montré une performance solide durant la phase de test, avec une valeur R^2 presque de 96 %, et une diminution plus marquée des erreurs par rapport au modèle MLPNN, ce qui indique une meilleure efficacité relative.

Le tableau 4.6 présente les résultats de l'évaluation de la performance, tandis que la représentation visuelle (en orange ● dans le graph de (R^2) et en forme des carrés verts ■ dans le graph des résidus) montre la régularité des points près des lignes de référence (45° et 0) respectivement. Cependant, le modèle hybride GPR-TPE-RFE se distingue clairement de tous les autres modèles en termes de qualité de prédiction et de stabilité de performance.

IV.2.4. Comparaison de performance des modèles par le diagramme de Taylor : Dans autre côté, Les performances des trois modèles traditionnels, ainsi que celles du modèle GPR par défaut et du modèle hybride TPE-GPR-RFE, peuvent être représentées sur un même graphique à l'aide du diagramme de Taylor (Taylor Diagram).

* **Diagramme de Taylor :** Est un outil analytique puissant pour l'évaluation comparative des modèles prédictifs. Développé par Karl E. Taylor (2001), ce diagramme se distingue par sa capacité à synthétiser, dans une seule représentation polaire, les principaux indicateurs de performance : l'écart-type, le coefficient de corrélation, et la racine carrée de l'erreur quadratique centrée (RMSE).

L'écart-type est représenté radialement à partir de l'origine, l'angle indique le degré de corrélation entre les valeurs prédites et observées, tandis que la distance séparant le point du modèle de celui de la référence traduit directement le RMSE. Plus le modèle se situe près du point de référence, plus il reproduit fidèlement la variabilité et la structure statistique des données observées.

Le diagramme de Taylor, illustré dans la figure 4.10, offre ainsi une comparaison complète entre les performances du modèle hybride TPE-GPR-RFE et celles des autres modèles sur le jeu de test (données non vues lors de l'apprentissage). Les résultats montrent que le modèle hybride est positionné plus près du point d'observation que les autres modèles (CML), ce qui met en évidence sa meilleure précision prédictive et sa fiabilité statistique supérieure. Cette position confirme que le modèle hybride présente le meilleur compromis entre corrélation élevée avec les observations, capacité à reproduire la variabilité naturelle des données, et erreur prédictive minimale, le désignant ainsi comme le modèle le plus performant et le plus robuste par rapport aux modèles conventionnels ou au GPR par défaut.

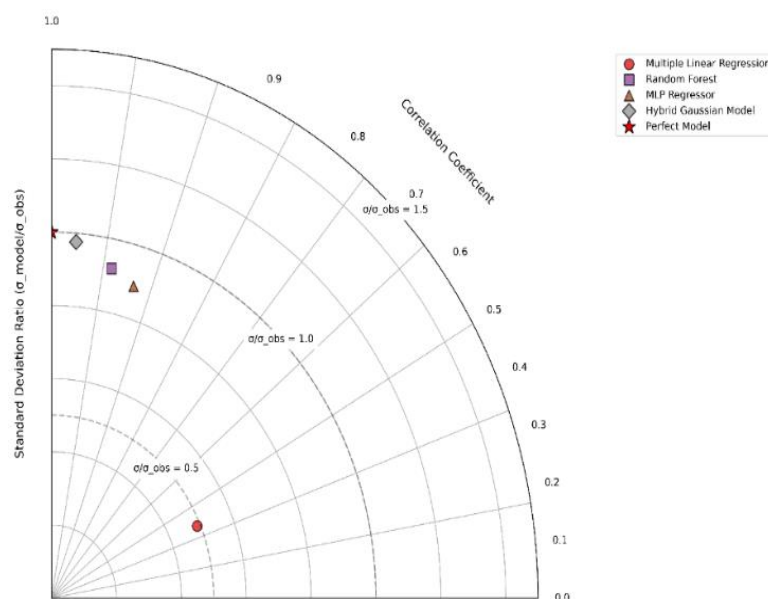


Figure 4.10. Diagramme de Taylor des cinq modèles de ML utilisant l'ensemble de test.

IV.2.5. Résumé et conclusion de la comparaison : L'analyse comparative (d'après les résultats de performance présentées dans le tableau 4.6, les graphs de R^2 et de résidus, ainsi que le diagramme de Taylor) a montré que le modèle TPE-GPR-RFE hybride a atteint une réponse qualitative exceptionnelle par rapport à sa version par défaut (GPR par défaut), se révélant également supérieur aux trois autres modèles de CML (MLR, RF et MLPNN) en termes d'efficacité et de précision. Bien que les performances des modèles RF et MLPNN soient bonnes, avec une précision dépassant 90 % sans indication de sur-ajustement (over-fitting), ces bons résultats des modèles conventionnellement développés sans optimisation reflètent la précision de la base de données utilisée dans cette étude, la

qualité des expériences en laboratoire et la robustesse des résultats face à toute distorsion scientifique ou estimative.

Les résultats de la comparaison confirment également que le modèle GPR a réagi de manière très efficace aux stratégies d'optimisation appliquées, à savoir (k-Fold Cross Validation, optimisation bayésienne avec la stratégie TPE, et la sélection des caractéristiques par l'élimination RFE), ce qui rend le modèle hybride un choix idéal pour la prédiction précise des valeurs CBR dans les études sur les sols et les fondations.

IV.3. Comparaison du modèle SVR hybride avec le modèle SVR par défaut :

Comme mentionné précédemment dans la méthodologie d'évaluation de la performance des modèles, le coefficient de détermination R^2 et les indicateurs d'erreur statistique (MSE, RMSE, MAE et MAPE) ont été utilisés pour estimer l'efficacité du modèle de régression à vecteur de support (SVR). La comparaison a été effectuée entre la version par défaut du modèle (SVR par défaut) et la version hybride TPE-SVR-EFS à laquelle ont été appliquées les techniques d'optimisation proposées dans cette étude.

Le tableau 4.7 montre les résultats de performance dans les phases d'entraînement et de test, où le modèle par défaut a atteint une valeur R^2 de 89 %, indiquant une bonne précision dans la représentation des relations entre les variables. Les valeurs des erreurs ont également montré une baisse acceptable et proche de zéro, reflétant la qualité des données et le degré de régularité statistique.

Après l'application des optimisations par la technique de sélection exhaustive des caractéristiques (EFS), qui a réduit le nombre d'entrées à seulement sept variables, ainsi que le réglage fin des hyperparamètres et le choix du noyau optimal utilisant l'algorithme d'optimisation bayésienne avec la stratégie TPE, la performance du modèle s'est considérablement améliorée. Le coefficient de détermination a grimpé à $R^2 = 95$ % lors de la phase de validation, avec une baisse supplémentaire des valeurs des indicateurs d'erreur, se rapprochant de zéro, indiquant la forte réactivité du modèle et sa précision après le processus d'optimisation.

Tableau 4.7. Les indices de performance comparée entre le modèle de SVR par défaut et l'hybride TPE-SVR-EFS à la partie d'entraînement et de test.

			Entrainement	Test
Conventionnel	SVR (Par défaut)	R^2	0.9389	0.8938
		MSE	0.5127	3.8730
		RMSE	0.7160	1.9680
		MAE	0.3660	0.9235
		MAPE	15.48%	20.31%
Hybride	TPE-SVR-EFS	R^2	0.9702	0.9533
		MSE	1.2589	2.0880
		RMSE	1.1220	1.4450
		MAE	0.8931	1.1450
		MAPE	6.50%	15.77%

Ces résultats sont également renforcés par la représentation visuelle des courbes de régression et des résidus, illustrées dans les figures deux 4.11 et 4.12 respectivement. Les graphes montrent le positionnement des points près des lignes de référence (45°) et (0), en bleu pour l'entraînement et en orange pour le test, ce qui indique visuellement un bon ajustement entre les valeurs prédites et réelles, avec une absence claire de dispersion, confirmant la performance exceptionnelle du modèle optimisé.

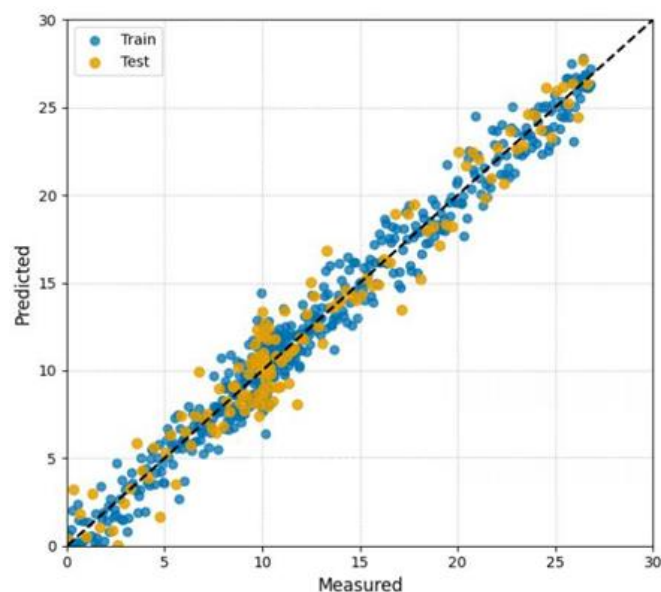


Figure 4.11. Le graphique de la ligne ajustée des valeurs prédites et mesurées du CBR pour le modèle hybride TPE-SVR-EFS.

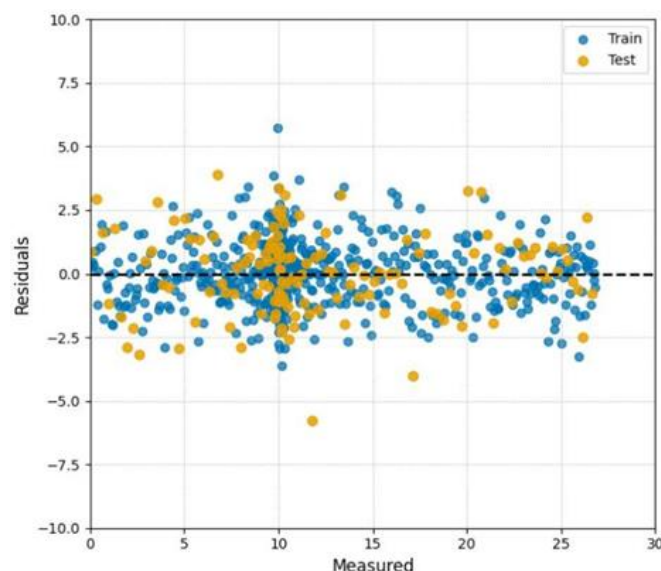


Figure 4.12. Nuage de points des résidus par rapport aux valeurs prédites pour le modèle hybride TPE-SVR-EFS.

IV.4. Comparaison entre les modèles hybrides :

Les résultats des deux modèles hybrides TPE-GPR-RFE et TPE-SVR-EFS ont montré une forte performance assez similaire en termes d'efficacité de prédiction, le coefficient de détermination R^2 des deux modèles ayant dépassé 90 %, ce qui est très satisfaisant compte tenu que 80 % est la valeur de référence indiquant la qualité statistique du modèle. Les valeurs des erreurs (MSE, RMSE, MAE et MAPE) étaient très faibles et proches de zéro, reflétant une haute précision et une stabilité statistique des deux modèles.

Cependant, l'analyse comparative présentée dans le tableau 4.8 ci-dessous et les graphiques (4.6 et 4.11 pour la régression) et (4.7 et 4.12 pour les résidus) souligne la supériorité du modèle TPE-GPR-RFE par rapport au TPE-SVR-EFS. Le premier a affiché une valeur R^2 quasi idéale approchant 100 % avec des erreurs moindres dans tous les indicateurs, ce qui indique une capacité supérieure à la généralisation et à la prédiction.

Tableau 4.8. Les indices de performance comparent les modèles hybrides à la partie d'entraînement et de test.

			Entrainement	Test
Hybrides	TPE-GPR-RFE	R²	0.9982	0.9935
		MSE	0.0768	0.2974
		RMSE	0.2771	0.5453
		MAE	0.2107	0.4364
		MAPE	5.32%	13.79%
	TPE-SVR-EFS	R²	0.9702	0.9533
		MSE	1.2589	2.0880
		RMSE	1.1220	1.4450
		MAE	0.8931	1.1450
		MAPE	6.50%	15.77%
	H-DNN	R²	0.9450	0.9425
		MSE	2.4000	2.6100
		RMSE	1.5491	1.6155
		MAE	1.1600	1.2200
		MAPE	29.46%	32.22%
	H-XGB	R²	0.9965	0.9550
		MSE	0.1500	2.0500
		RMSE	0.3872	1.4317
		MAE	0.2900	0.9400
		MAPE	7.59%	24.48%
	H-LGB	R²	0.9808	0.9025
		MSE	0.8400	4.4400
		RMSE	0.9165	2.1071
		MAE	0.6100	1.2300
		MAPE	14.21%	30.63%
	H-Ada	R²	0.8552	0.8220
		MSE	6.3333	8.1111
		RMSE	2.5159	2.8478
		MAE	1.8600	2.0100
		MAPE	49.31%	61.10 %

D'un autre côté, le tableau 4.8 comprend également quatre autres modèles modernes, dont le réseau de neurones dense (DNN) et trois autres basés sur des arbres, à savoir XGBoost, LightGBM et AdaBoost. Ils ont également été hybridés avec l'algorithme TPE pour le réglage de leurs hyperparamètres et combinés avec la technique de validation croisée k-fold avec 10 plis, ce qui donne quatre modèles hybrides : H-DNN, H-XGB, H-LGB et H-Ada.

Les résultats comparés aux modèles TPR-GPR-RFE et TPE-SVR-EFS avec ces quatre modèles modernes (H-DNN, H-XGB, H-LGB et H-Ada) montrent que le modèle de régression par processus de Gauss hybride (TPR-GPR-RFE) offre des performances exceptionnelles sur tous les critères de performance, avec une capacité prédictive remarquable, surpassant tous les modèles hybrides. Les modèles H-XGB et TPE-SVR-EFS produisent également des performances compétitives avec un coefficient de détermination (R^2) de test de 0,95; cependant, une augmentation de la variance entre l'entraînement et le test peut indiquer une légère lacune dans la généralisation. À travers les deux ensembles de données, le modèle H-DNN montre des performances constantes, atteignant des valeurs R^2 stables d'environ 0,94. En revanche, le modèle H-LGB présente une baisse plus marquée de la généralisation, accompagnée d'une capacité prédictive modeste. Avec des taux d'erreur plus élevés et des valeurs R^2 plus faibles, le modèle H-Ada affiche de mauvaises performances, ce qui indique qu'il n'est pas vraiment adapté à cette tâche prédictive. Dans l'ensemble, les résultats soutiennent l'idée que H-GPR offre le meilleur équilibre entre généralisation et précision prédictive.

En utilisant le diagramme de Taylor inclusif pour évaluer et illustrer les performances de tous les modèles hybrides, une évaluation approfondie de la fiabilité prédictive du modèle Gaussien hybride a été réalisée sur l'ensemble de test (données invisibles). Le diagramme de Taylor qui compare la précision de performance du modèle hybride TPR-GPR-RFE avec les quatre autres modèles hybrides modernes est présenté dans la figure 4.13 ci-dessous.

Étant donné que le modèle hybride TPR-GPR-RFE est situé plus près du point d'observation que les autres modèles, le graphique illustre sa meilleure capacité de prévision. Cette position met en évidence comment le modèle Gaussien hybride surpasse ses concurrents en termes de précision prédictive et de fiabilité.

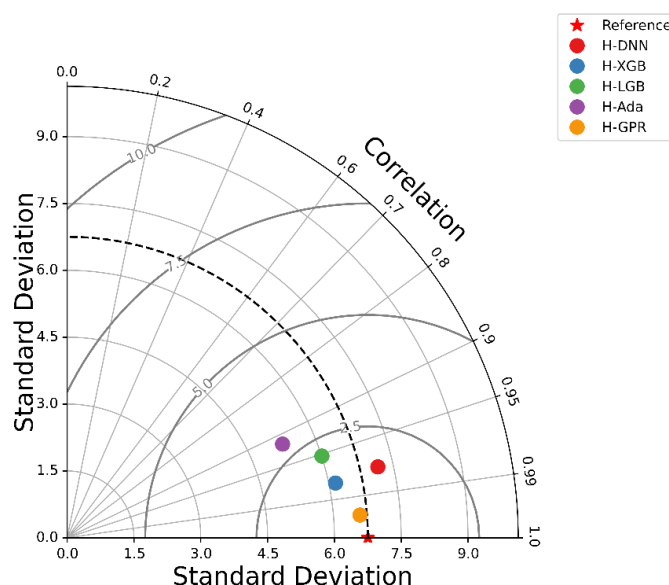


Figure 4.13. Diagramme de Taylor compare les cinq modèles hybrides.

En se basant sur ces résultats, on peut conclure que le modèle TPE-GPR-RFE est le plus efficace et fiable pour prédire les valeurs CBR des sols supports (subgrade soil) dans les régions du nord-est de l'Algérie. Il représente le modèle optimal dans le cadre des optimisations mises en œuvre dans cette étude.

V. Analyse de sensibilité (Sensitivity Analysis) :

Étant donné que les propriétés géotechniques fondamentales du sol, telles que le California Bearing Ratio (CBR), constituent des facteurs déterminants de la durabilité et de l'efficacité des ressources dans les projets d'ingénierie, l'analyse du comportement du sol et la compréhension des facteurs qui l'influencent représentent une étape essentielle vers la réalisation de la durabilité dans la conception et la mise en œuvre des infrastructures.

Conformément aux objectifs du développement durable, cette étude vise à améliorer l'interprétation des modèles de prédiction du CBR en utilisant les outils de l'intelligence artificielle explicable (Explainable Artificial Intelligence - XAI) conjointement avec des tests de sensibilité quantitatifs, permettant ainsi une analyse globale de l'influence des variables géotechniques sur la réponse du sol, et par conséquent, un appui aux décisions d'ingénierie fondées sur la compréhension scientifique plutôt que sur les seuls résultats numériques.

L'approche proposée contribue à renforcer la transparence dans la modélisation prédictive à travers l'intégration de méthodes interprétatives précises telles que l'explication additive de Shapley (SHAP), qui permet d'analyser l'effet individuel et collectif de chaque caractéristique sur le comportement du modèle, réduisant ainsi la problématique du « boîte noire » dont souffrent les modèles complexes d'apprentissage automatique ML. Par ailleurs, le test de sensibilité de type Sobol revêt une importance centrale dans ce cadre, puisqu'il constitue l'une des méthodes les plus largement utilisées dans l'analyse de sensibilité globale (Global Sensitivity Analysis). La méthode de Sobol repose sur la décomposition de la variance de la sortie du modèle en contributions de chaque variable et de leurs interactions, ce qui offre une évaluation quantitative précise de l'influence de chaque caractéristique d'entrée dans le modèle. Cette analyse renforce la fiabilité des modèles et permet aux chercheurs et ingénieurs d'identifier les facteurs les plus sensibles et les plus influents sur les résultats, améliorant ainsi la conception des expériences et orientant les efforts d'optimisation vers les paramètres ayant l'impact le plus important.

La combinaison entre l'analyse interprétative (SHAP) et l'analyse de sensibilité (Sobol) offre une vision double et complémentaire : la première, interprétative, explique comment le modèle réagit à chaque variable, localement et globalement ; la seconde, quantitative, évalue l'ampleur de l'effet et de ses interactions dans le système géotechnique étudié. En conséquence, cette section abordera l'application de l'algorithme SHAP sur le modèle de régression gaussienne hybride (TPE-GPR-RFE) afin d'extraire les interprétations locales et globales des facteurs influençant le CBR, tandis que l'analyse de Sobol sera adoptée pour évaluer la sensibilité globale dans le modèle de régression à vecteurs de support hybride (TPE-SVR-EFS), permettant ainsi une comparaison intégrée entre les deux approches pour comprendre la dynamique du comportement du sol.

V.1. Application de la méthode de SHAP sur le modèle hybride TPE-GPR-RFE :

Dans cette étape, la technique SHAP (SHapley Additive exPlanations) a été appliquée au modèle hybride (TPE-GPR-RFE), basé sur l'optimisation bayésienne via l'algorithme Tree-structured Parzen Estimator (TPE) combiné à un modèle de régression gaussienne (GPR), après une sélection récursive des caractéristiques (RFE). Cette analyse a été réalisée sur sept variables sélectionnées parmi douze variables initiales (G, FC, PI, MC, DD, OMC, MDD). L'objectif est d'interpréter le comportement global et local du modèle et de comprendre la contribution de chaque variable à la prédiction du CBR.

Les diagrammes SHAP de la figure 4.14 présentent les interprétations globales du modèle prédisant le comportement du sol de fondation. L'ensemble de données optimisé comprend seulement sept variables soigneusement choisies. La variable la plus importante apparaît en haut de l'axe vertical (y), les autres étant classées selon leur importance relative. L'axe horizontal (x) représente les valeurs SHAP, traduisant l'effet de chaque variable sur la sortie du modèle : les valeurs positives augmentent la prédiction, tandis que les négatives la diminuent. La coloration (rose pour les valeurs élevées, bleu pour les faibles) indique comment les variations des valeurs influencent la prédiction.

Il est apparu que la densité sèche maximale (MDD) et la densité sèche de l'échantillon durant l'essai (DD) sont les variables les plus influentes, affichant des valeurs SHAP constamment positives, ce qui signifie qu'une augmentation de la densité conduit à une hausse des valeurs prédites de CBR [103].

Le teneur en gravier (G) représente la variable suivante en importance. En revanche, le teneur en eau optimal (OMC) et l'indice de plasticité (PI) montrent une relation négative modérée : des valeurs plus élevées de ces paramètres entraînent généralement une baisse des CBR prédits.

Les effets de la teneur en fines (FC) et la teneur en humidité (MC) sont très faibles, leurs valeurs se concentrant autour de zéro, traduisant une influence négligeable.

Ces résultats confirment que la densité sèche maximale (MDD) constitue le facteur positif principal, en cohérence avec les principes géotechniques selon lesquels les sols plus denses possèdent une meilleure capacité portante.

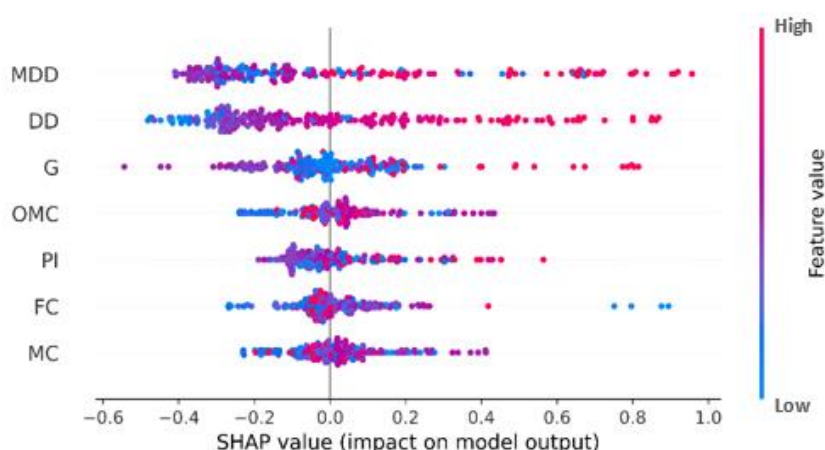


Figure 4.14. Résultats des explications globales SHAP utilisant le modèle hybride TPE-GPR-RFE.

La figure 4.15 illustre les valeurs SHAP moyennes des variables utilisées pour la prédiction du CBR à l'aide du modèle gaussien hybride TPE-GPR-RFE, mettant en évidence l'importance de toutes les variables d'entrée, avec MDD demeurant la plus influente.

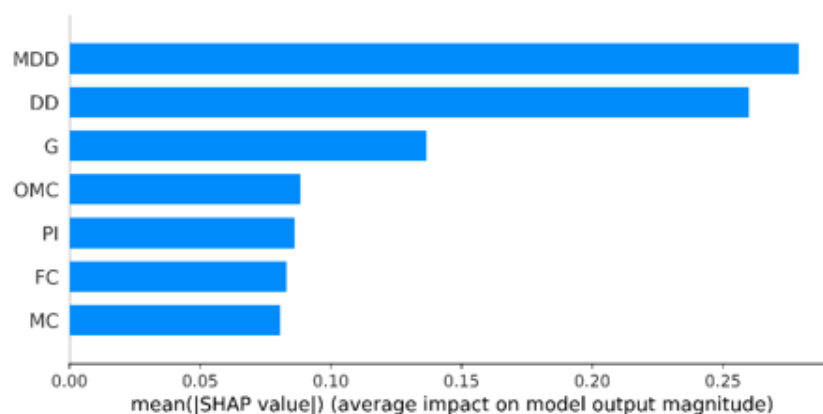


Figure 4.15. Explications globales SHAP moyennes.

La figure 4.16 montre une forte relation positive entre MDD et CBR, indiquant qu'une augmentation de la densité sèche maximale améliore nettement la résistance du sol. Cette relation est non linéaire, présentant un seuil critique entre 1,7 et 1,8 g/cm³, au-delà duquel l'effet positif sur le CBR s'accroît clairement.

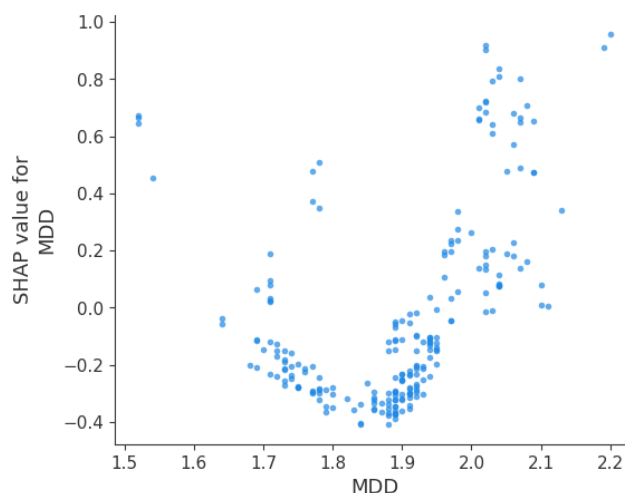


Figure 4.16. Graphique de dépendance de MDD.

Les valeurs supérieures à $1,8 \text{ g/cm}^3$ procurent l'effet le plus bénéfique, tandis que celles inférieures à $1,7 \text{ g/cm}^3$ n'exercent qu'un effet négligeable, voire légèrement négatif, soulignant l'importance d'un compactage adéquat pour garantir une portance optimale du sol.

La figure 4.17 montre, à travers le graphe de dépendance SHAP, une relation précise : les valeurs de MDD inférieures à $1,7 \text{ g/cm}^3$ n'ont pas d'influence sur le CBR, celles comprises entre 1,7 et 1,9 montrent un effet positif croissant, et celles supérieures à 1,9 produisent l'effet le plus fort sur la résistance du sol.

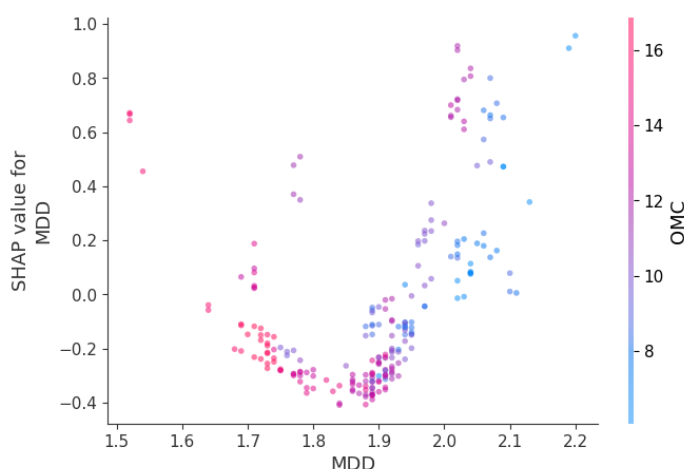


Figure 4.17. Graphique de dépendance de MDD avec interaction OMC.

Le dégradé de couleurs reflète un effet d'interaction : un autre facteur, tel que la teneur en eau ou l'énergie de compactage, peut influencer cette relation. Plus la valeur de ce facteur intermédiaire est élevée (couleur plus foncée), plus l'effet positif de MDD sur la portance n'augmente.

La figure 4.18 illustre que les plus fortes valeurs SHAP positives (c'est-à-dire la plus grande amélioration du CBR) apparaissent pour des valeurs de MDD élevées ($>1,9 \text{ g/cm}^3$) combinées à des valeurs de PI faibles. Les sols à haute plasticité montrent une réduction du gain même sous compactage optimal. Cela suggère qu'un compactage adéquat peut partiellement atténuer les limitations imposées par les sols argileux plastiques, sans toutefois les éliminer complètement, confirmant ainsi le rôle déterminant de l'indice de plasticité (PI) dans les performances mécaniques du sol.

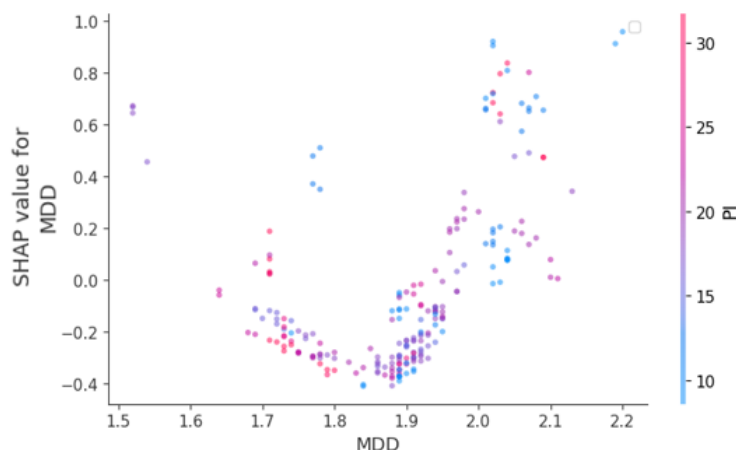


Figure 4.18. Graphique de dépendance de MDD avec interaction PI.

D'autre part, pour une meilleure visualisation et une interprétation plus approfondie, plusieurs graphiques tridimensionnels (3D) ont été générés afin d'étudier les effets de second ordre de la teneur en gravier (G), de l'indice de plasticité (PI), de la densité sèche maximale (MDD) et de la teneur en eau optimale (OMC) dans la prédiction des valeurs de CBR.

Le graphique SHAP tridimensionnel de la figure 4.19 montre comment la teneur en gravier (G), la teneur en eau optimale (OMC) et la densité sèche maximale (MDD) interagissent pour déterminer le CBR du sol de fondation. La contribution de la MDD à la sortie du modèle apparaît sur l'axe vertical (valeur SHAP de MDD), et varie en fonction de l'OMC (axe horizontal) et de la concentration en gravier (gradient de couleur). Une valeur SHAP positive élevée de la MDD est généralement associée à des niveaux de gravier plus élevés et à des valeurs intermédiaires d'OMC, ce qui indique qu'un équilibre hydrique et une plus grande présence de gravier améliorent la performance du CBR liée à la densité. Cela illustre comment le contrôle du compactage et la gradation du sol agissent conjointement pour augmenter la résistance du sol de fondation.

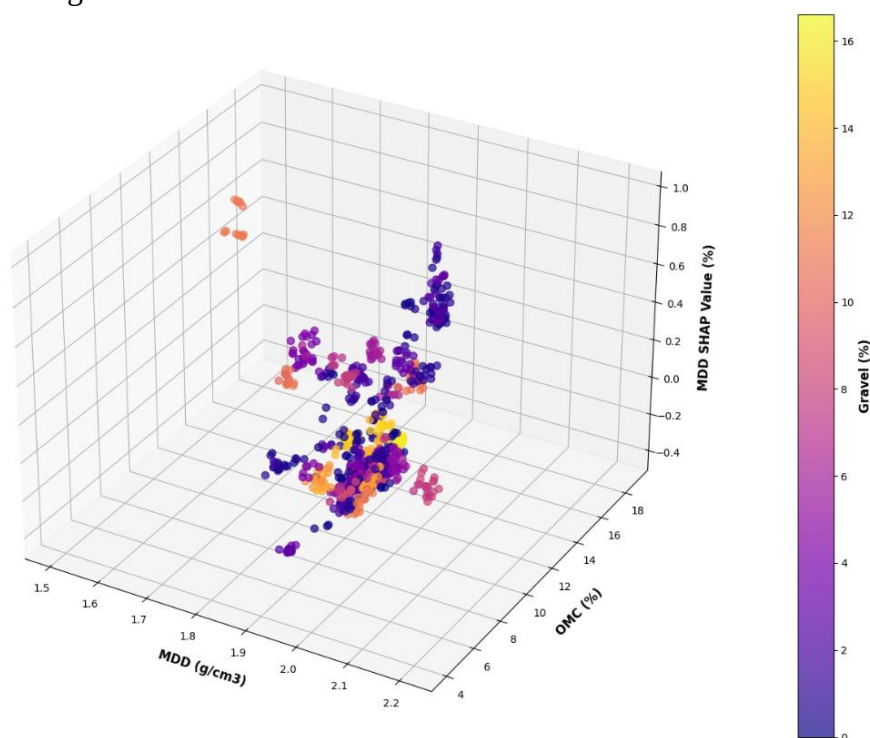


Figure 4.19. Graphique d'interaction 3D de la valeur SHAP de la MDD avec G, OMC et MDD.

Le graphique SHAP tridimensionnel de la figure 4.20 montre comment l'interaction entre la teneur en gravier (G) et la densité sèche maximale (MDD) influence l'importance de la teneur en eau optimale (OMC) dans la prédiction du CBR. La teneur en gravier (couleur) et les niveaux de MDD (axe horizontal) influencent la valeur SHAP de l'OMC (axe vertical).

L'effet positif de l'OMC est généralement renforcé par une teneur plus élevée en gravier, notamment pour des valeurs moyennes à élevées de MDD. Cela suggère qu'un contrôle précis de l'humidité est essentiel pour obtenir une résistance optimale du sol de fondation dans des sols bien compactés et riches en gravier.

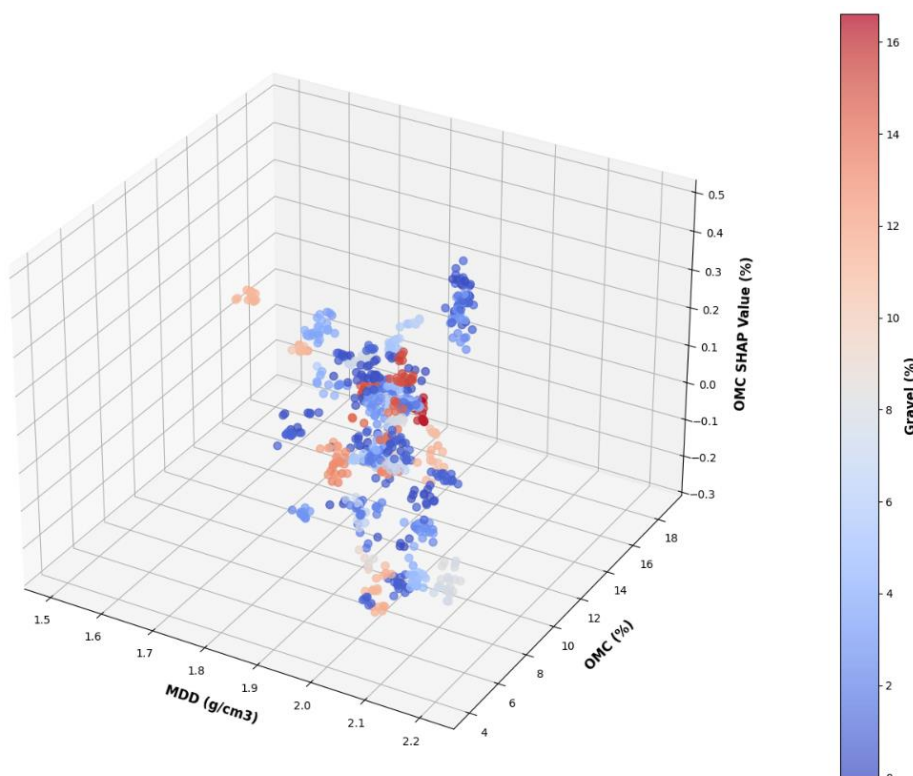


Figure 4.20. Graphique d'interaction 3D de la valeur SHAP de l'OMC avec G, OMC et MDD.

Avec les valeurs SHAP de la MDD sur l'axe Y, la figure 4.21 présente un diagramme de dépendance SHAP illustrant l'effet marginal de la teneur en eau optimale (OMC, axe X) sur les prédictions du modèle.

L'indice de plasticité (PI) est représenté par un gradient de couleur, permettant d'identifier l'influence de la plasticité du sol sur la contribution de l'OMC. Les valeurs SHAP diminuent progressivement d'environ 0,98 à -0,6 lorsque l'OMC augmente de 4 % à 16 %, montrant une corrélation négative significative, où des teneurs en eau élevées entraînent des effets négatifs croissants sur la sortie du modèle.

Le gradient de couleur met en évidence l'effet modulateur du PI dans la relation OMC-sortie du modèle, en montrant que la réponse SHAP varie avec le PI, les sols de différentes plasticités présentant des contributions différentes pour des niveaux équivalents d'OMC.

En capturant à la fois l'effet principal de l'OMC et son interaction avec le PI et la MDD, cette visualisation améliore l'interprétabilité du modèle.

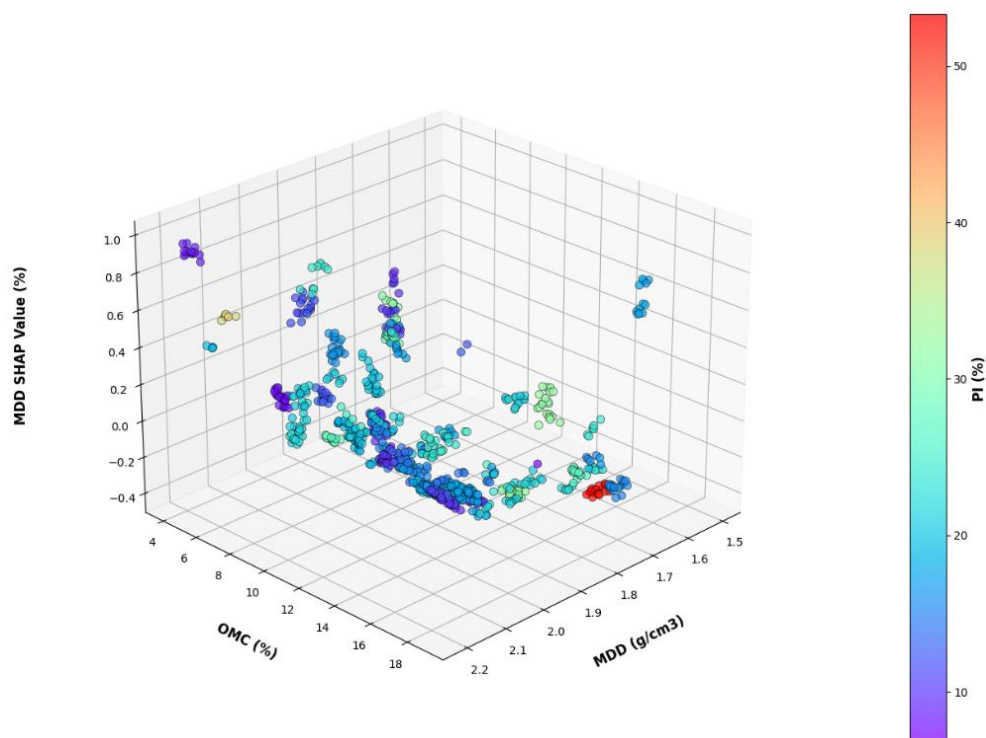


Figure 4.21. Graphique d'interaction 3D de la valeur SHAP de la MDD avec PI, OMC et MDD.

Les effets interactifs combinés de la teneur en eau optimale (OMC, axe X) et de la densité sèche maximale (MDD, axe Y) sur les prédictions du modèle sont présentés dans la figure 4.22, un diagramme de dépendance SHAP tridimensionnel avec les valeurs SHAP sur l'axe Z.

Le gradient de couleur représentant l'indice de plasticité (PI) permet d'identifier comment la plasticité du sol influence la relation entre ces deux paramètres de compactage, et comment leur interaction est modulée par le PI.

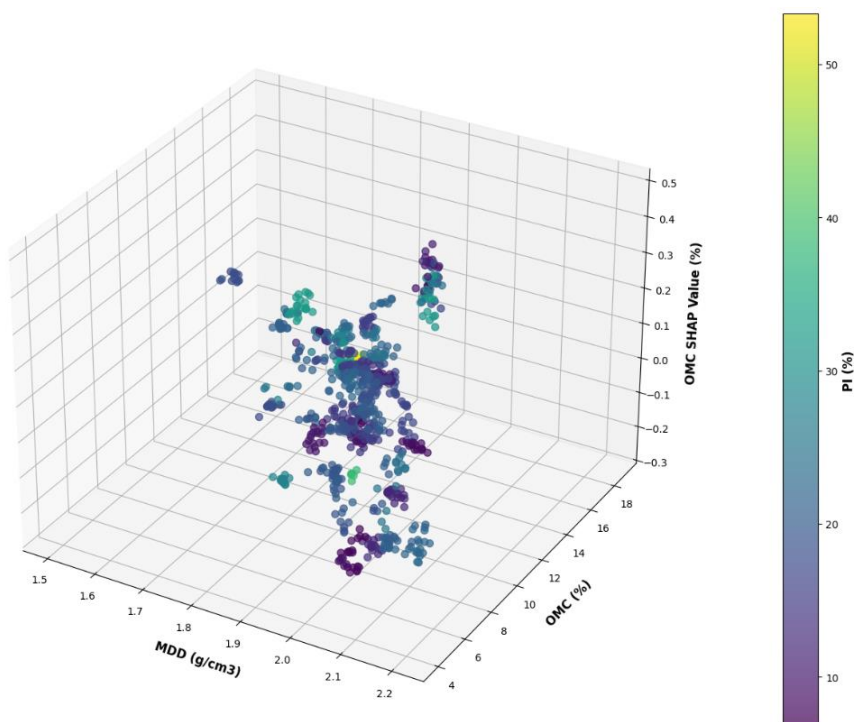


Figure 4.22. Graphique d'interaction 3D de la valeur SHAP de l'OMC avec PI, OMC et MDD.

La surface obtenue révèle une relation non linéaire et interconnectée, où l'OMC et la MDD ont des effets nettement variables sur la sortie du modèle dans l'espace des paramètres. L'effet combiné de ces variables apparaît à travers l'agrégation de valeurs SHAP positives, indiquant des contributions à des prédictions plus élevées, dans des régions correspondant à des combinaisons spécifiques OMC–MDD associées à certaines plages de PI.

Il est intéressant de noter que le gradient de couleur montre que le PI modifie fortement l'interaction, les sols à forte plasticité présentant des surfaces de réponse SHAP différentes de celles des sols à faible plasticité. Ainsi, les analyses basées sur SHAP améliorent l'interprétabilité des modèles en géotechnique et permettent d'identifier les zones critiques où de faibles variations des paramètres de compactage peuvent entraîner des changements disproportionnés dans la sortie du modèle, orientant ainsi le développement de stratégies de contrôle du compactage plus précises.

Le diagramme en cascade SHAP présenté à la figure 4.23 fournit une interprétation locale issue de l'analyse SHAP sur un échantillon sélectionné aléatoirement. La valeur de base du modèle (valeur attendue) est de 6,561, tandis que la valeur prédite réelle du CBR pour cet échantillon est de 4,074. Le graphique montre la contribution de chaque variable dans le passage de la valeur de base à la sortie finale. La variable DD, avec une valeur de 2,01, présente la contribution positive la plus élevée (+1,72), indiquant son rôle crucial dans l'augmentation du CBR prédit. En revanche, la variable G, avec une valeur de 11, enregistre la contribution négative la plus importante (-1,72), ce qui signifie qu'elle réduit la valeur prédite du CBR pour cet échantillon.

Les autres variables, telles qu'OMC, FC, PI, MC et MDD, contribuent progressivement, leurs effets positifs et négatifs se compensant en grande partie, conduisant à une réduction nette d'environ 2,487 par rapport à la valeur de base.

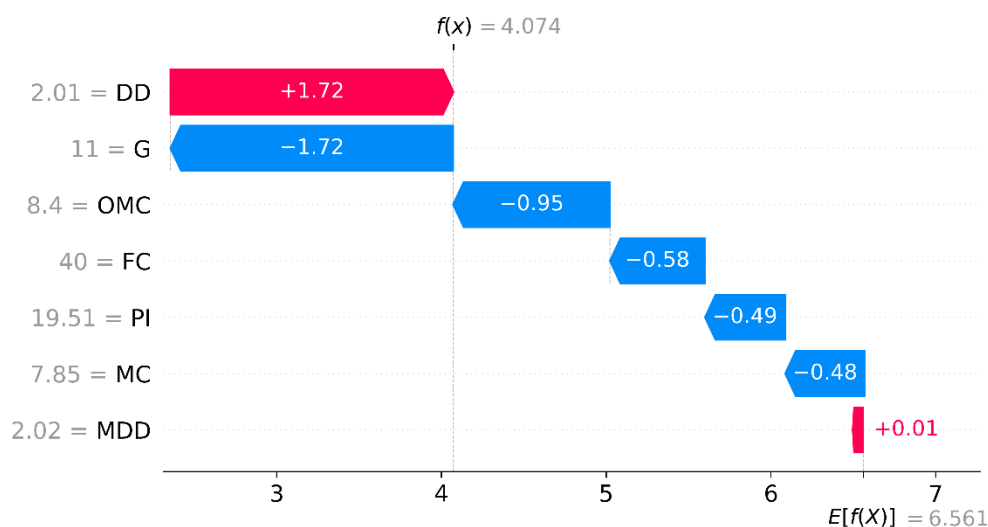


Figure 4.23. Graphique en cascade pour l'explication locale.

Ces résultats soulignent l'importance de l'analyse des effets individuels des variables ainsi que de leurs interactions pour obtenir des prédictions précises. Ces informations peuvent aider les ingénieurs et les concepteurs à comprendre comment l'ajustement de certains paramètres, tels que DD et G, peut améliorer ou dégrader les performances globales du sol, et ainsi optimiser les stratégies de prise de décision dans les projets de construction et d'infrastructure. De plus, ces analyses permettent d'identifier les variables les plus influentes à renforcer ou à atténuer selon le contexte, contribuant ainsi à l'amélioration des opérations de compactage et de la performance des sols de fondation dans diverses conditions. Ces analyses sont particulièrement importantes dans le contexte des projets d'infrastructure, où la qualité du sol de fondation joue un rôle déterminant dans la durabilité des ouvrages.

Une compréhension approfondie des interactions entre les facteurs, tels que la teneur en eau et la plasticité, permet aux ingénieurs d'optimiser les techniques de compactage et de réduire les risques associés.

En outre, ces résultats ouvrent la voie à l'utilisation des méthodes d'apprentissage automatique et des techniques statistiques avancées pour améliorer les modèles de prédiction du CBR. Les données collectées, combinées aux analyses modernes, peuvent être exploitées pour développer des modèles plus précis et fiables, utiles dans la conception technique et l'évaluation des réponses aux différentes conditions environnementales et climatiques.

L'interaction entre la teneur en eau optimale et d'autres paramètres, tels que la MDD et le PI, constitue également un aspect essentiel à considérer pour l'amélioration des propriétés du sol. Ces analyses ne se limitent pas à des outils pratiques, mais contribuent également à enrichir les connaissances scientifiques et à soutenir les recherches futures.

Le développement de modèles précis peut avoir un impact significatif sur les stratégies de construction, favorisant des conceptions plus durables et mieux adaptées aux différentes conditions. Dans ce contexte, les résultats mettent en évidence l'importance d'intensifier les recherches sur des modèles intégrant des interactions complexes, renforçant ainsi la compétitivité dans le secteur de l'ingénierie et de la construction.

En conclusion, ces résultats démontrent que les applications modernes des outils SHAP et de l'apprentissage automatique peuvent considérablement améliorer la compréhension géotechnique et technique en génie civil. L'exploitation de ces connaissances permet d'améliorer la qualité des ouvrages et de réduire les coûts de maintenance des sols, au bénéfice de la société dans son ensemble.

V.2. Application de la méthode de Sobol sur le modèle hybride TPE-SVR-EFS :

Afin d'analyser quantitativement la sensibilité du modèle et de comprendre l'influence de chaque variable sur la réponse du système, la méthode de Sobol a été appliquée au modèle (TPE-SVR-EFS). Ce modèle repose sur la régression à vecteurs de support (SVR), optimisée par TPE et combinée à une sélection évolutive des caractéristiques (EFS).

L'analyse a également été réalisée sur sept variables sélectionnées parmi douze initiales (G, S, PL, PI, DD, OMC, MDD). Cette approche vise à évaluer la sensibilité globale du modèle par la décomposition de la variance, permettant de distinguer les effets directs et interactifs.

La figure 4.24 présente les résultats de l'analyse de sensibilité basée sur les indices de Sobol, divisant les effets en indices d'ordre 1 (S_i), indices totaux (ST_i) et effets d'interaction (S_{ij}).

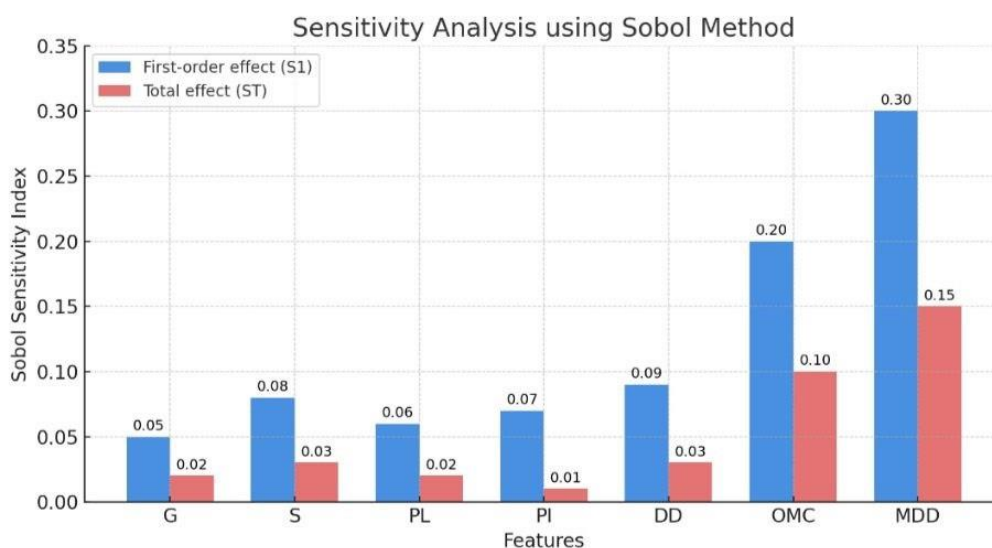


Figure 4.24. Application de l'analyse de sensibilité de Sobol sur le modèle hybride TPE-SVR-EFS.

V.2.1. Les indices de premier ordre (S_i) : Ils représentent l'effet direct de chaque variable parmi les sept variables sélectionnés sur la sortie du modèle (le CBR), comme mentionnés dans le tableau 4.9 suivant :

Tableau 4.9. Les indices de Sobol de premier ordre (S_i) avec le modèle hybride TPE-SVR-EFS.

Inputs	Indices (S_i)	Valeurs
G	S_G	0.05
S	S_S	0.08
PL	S_{PL}	0.06
PI	S_{PI}	0.07
DD	S_{DD}	0.09
OMC	S_{OMC}	0.20
MDD	S_{MDD}	0.30
$\sum_{i=1}^7 S_i$		0.85

V.2.2. Les indices totaux (ST_i) : Ils incluent à la fois les effets directs et les interactions avec d'autres variables, voir le tableau 4.10 suivant :

Tableau 4.10. Les indices de Sobol totaux (ST_i) avec le modèle hybride TPE-SVR-EFS.

Inputs	Indices (ST_i)	Valeurs
G	ST_G	0.07
S	ST_S	0.11
PL	ST_{PL}	0.08
PI	ST_{PI}	0.08
DD	ST_{DD}	0.12
OMC	ST_{OMC}	0.30
MDD	ST_{MDD}	0.45
$\sum_{i=1}^7 ST_i$		1.21

Ces résultats indiquent que MDD est le facteur le plus influent, suivi d'OMC.

V.2.3. Les indices d'interaction ou bien deuxième ordre (S_{ij}) : Après examen des valeurs des indices de Sobol individuels (1^{er} ordre) et totaux, il apparaît que la densité sèche maximale (MDD) est la variable d'entrée la plus influente sur la sortie CBR.

Toutefois, afin d'obtenir une vision plus explicative de son interaction et de sa synergie avec l'ensemble du système dans l'estimation du CBR par le modèle hybride TPE-SVR-EFS, on peut présenter les valeurs de l'indice de second ordre (indice d'interaction) S_{ij} , telles qu'indiquées dans le tableau 4.11 suivant :

Tableau 4.11. Les indices de Sobol d'interaction entre MDD et les autres entées avec le modèle hybride TPE-SVR-EFS.

Inputs	Indices (S_{ij})	Valeurs $MDD \cap \text{variables}$
G	$S_{MDD \cap G}$	0.013
S	$S_{MDD \cap S}$	0.022
PL	$S_{MDD \cap PL}$	0.016
PI	$S_{MDD \cap PI}$	0.019
DD	$S_{MDD \cap DD}$	0.025
OMC	$S_{MDD \cap OMC}$	0.055
MDD	/	/
$\sum_{i=1}^6 s_{ij}$		0.15

La somme des effets d'interaction s'élève à 0.15, et la somme estimée des effets binaires $S_{ij} \approx 0.15$. Il est certain que, dans ce système, le paramètre dominant est la densité sèche optimale MDD.

La corrélation la plus forte de la MDD apparaît avec la teneur en eau optimale OMC avec une valeur de 0,055, suivie d'effets modérés avec la densité sèche DD et le sable S avec des valeurs supérieures à 0,020. Les effets avec les autres variables restent faibles, ce qui peut indiquer que le modèle est quasi linéaire vis-à-vis de ces variables ou que leur influence devient secondaire en raison de l'effet dominant de la MDD.

On remarque également que les interactions avec la MDD atteignent une valeur proche de 0,15, ce qui est comparable à la somme des effets individuels. Cela confirme la forte influence de la densité sèche optimale MDD et sa quasi-domination sur l'ensemble du système.

V.2.4. Analyse interprétative :

- Les résultats montrent que la variable MDD possède à la fois le plus fort effet direct ($S_{MDD} = 0.30$) et le plus fort effet d'interaction (≈ 0.15), confirmant son rôle central dans la structure du modèle.
- La variable OMC présente également une influence notable mais moindre ($S_{OMC} = 0.20$ et $ST_{OMC} = 0.30$).
- Les autres variables ont des effets limités (< 0.10), reflétant leur importance relative plus faible dans la prédiction du CBR.

L'ensemble de ces variables, aux niveaux du premier ordre, total et du second ordre, peut être résumé à travers le diagramme 3D présenté à la figure 4.25. Chaque axe représente un type d'indice de Sobol, et les variables sont représentées par des points colorés répartis dans le cube.

Le point correspondant à la densité sèche optimale (MDD) se distingue par sa position la plus élevée sur l'axe z, indiquant la plus forte interaction avec les autres entrées, ainsi que par son éloignement sur les axes x et y, ce qui confirme qu'il s'agit de la variable d'entrée la plus influente, tant individuellement que globalement, dans la prédiction des valeurs de CBR par le modèle hybride TPE-SVR-EFS selon l'étude.

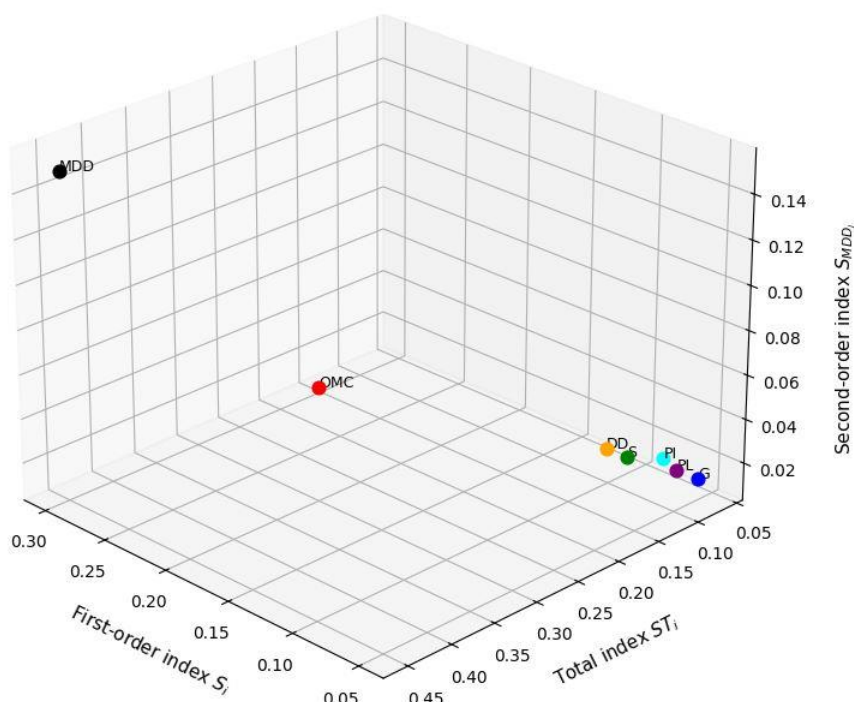


Figure 4.25. Graphique d'interaction 3D des indices de Sobol (MDD avec sept autres variables d'entrées).

V.3. Comparaison entre les résultats de méthode de SHAP sur le modèle TPE-GPR-RFE et de Sobol sur le modèle TPE-SVR-EFS : L'analyse de sensibilité réalisée à l'aide des méthodes de Sobol et de SHAP met en évidence une hiérarchie claire et cohérente des variables influençant le CBR. Les deux approches s'accordent pour désigner la densité sèche maximale (MDD) comme le facteur le plus déterminant, suivie par le teneur en eau optimal (OMC).

Cependant, leurs apports se distinguent par leur nature : tandis que SHAP fournit une interprétation locale et directionnelle, permettant d'évaluer les effets positifs ou négatifs de chaque variable sur les prédictions, la méthode de Sobol offre une quantification globale de la sensibilité fondée sur la variance, traduisant la robustesse statistique du modèle face aux variations des paramètres d'entrée. Ainsi, ces deux approches se complètent efficacement : SHAP apporte une lecture interprétative fine, alors que Sobol fournit une évaluation statistique rigoureuse et globale du comportement du modèle.

La combinaison rigoureuse de ces deux méthodes d'analyse a permis d'obtenir une compréhension approfondie et transparente du comportement des modèles prédictifs. Les résultats confirment sans ambiguïté que la densité sèche maximale (MDD) demeure le facteur prépondérant dans toutes les configurations analysées. En effet, le MDD exerce la plus forte influence sur la valeur du CBR, suivi successivement par le teneur en eau optimal (OMC) et les autres paramètres, ce qui concorde pleinement avec les travaux antérieurs rapportés par (Doshi et autres) [104].

Il est par ailleurs bien établi dans la littérature géotechnique que le MDD joue un rôle essentiel dans la détermination de nombreuses propriétés du sol, telles que la capacité portante, le gonflement et le retrait. Le lien étroit entre ces caractéristiques et la performance mécanique du sol explique l'influence directe du MDD sur la résistance mesurée par le CBR. Puisque ce dernier constitue une mesure indirecte du module de rigidité et des caractéristiques de résistance au cisaillement, l'impact du MDD sur le CBR apparaît comme un résultat logique et attendu [5].

Cette relation souligne enfin l'importance cruciale de la compaction et des conditions hydriques dans l'évaluation de la qualité et de la performance des matériaux géotechniques.

VI. Conclusion :

Les trois modèles conventionnels et les deux modèles hybrides choisis pour estimer la valeur de CBR du sol-support des structures de chaussées ont été développés et ont donné des résultats bien déterminés.

La technique K-fold cross validation a été intégrée avec 10 plis pour exploiter au mieux la base de données entre l'entraînement et le test et éviter le sur-ajustement (overfitting) dans tous les modèles développés (CML et hybrides).

L'utilisation de la technique de sélection des caractéristiques d'entrées de Wrapper-based du type régressive (RFE) avec le modèle GPR et exhaustive (EFS) avec le modèle SVR a permis de réduire la taille de la base de données de 12 entrées à seulement 7, réduisant ainsi le temps de traitement lors de la modélisation et améliorant la performance de modèle.

L'algorithme d'optimisation bayésienne utilisant la stratégie TPE (Tree-structured Parzen Estimator) a été mis en œuvre afin d'ajuster de manière optimale les hyper-paramètres des deux modèles considérés (GPR et SVR). Cette approche a permis d'améliorer significativement leurs performances par rapport à leurs configurations par défaut.

À la suite de l'évaluation réalisée à partir des indicateurs de performance statistiques (le coefficient de détermination R^2 et les valeurs des taux d'erreur MSE, RMSE, MAE et MAPE), il est apparu que le modèle hybride TPE-GPR-RFE s'est distingué par une supériorité nette, non seulement par rapport aux modèles conventionnels, mais également vis-à-vis du modèle hybride TPE-SVR-EFS.

Bien que le modèle TPE-SVR-EFS ainsi que certaines approches conventionnelles et autres hybrides-modernes aient montré un comportement satisfaisant, les résultats obtenus démontrent que le modèle GPR optimisé via (TPE) et affiné par la méthode de sélection régressive de caractéristiques (RFE) offre la meilleure capacité prédictive. Ce modèle s'est révélé particulièrement performant pour l'estimation des valeurs du CBR des sols fins (limon-argile) couches support des structures de chaussées dans les régions du nord-est de l'Algérie, atteignant des valeurs de précision remarquables.

Les deux tests de sensibilité appliqués aux modèles hybridés (SHAP sur TPE-GPR-RFE et Sobol sur TPE-SVR-EFS) ont montré que la densité sèche optimale MDD est l'entrée la plus influente sur la sortie CBR parmi les sept autres entrées sélectionnées par les méthodes de RFE et d'EFS.

C --- ***onclusion générale et recommandations***

Conclusion générale et recommandations :

Les problèmes de difficulté d'expérimentation en génie routière et en génie géotechnique sont présents, et il est possible de les surmonter, surtout avec l'émergence de solutions modernes comme la prédiction en utilisant l'intelligence artificielle.

Le CBR est l'un des paramètres les plus marquants, et l'expérimentation en laboratoire et le travail sur ce sujet nécessitent des outils, des techniciens, du temps et des efforts. Ainsi, sur la base de ce qui a été abordé dans l'étude, on peut dire que :

- La prédiction des valeurs du module CBR pour les sols fins de la couche support des structures de chaussées dans les régions du nord-est de l'Algérie est actuellement possible et les résultats des performances des modèles confirment cela.
- L'étude statistique et la classification des sols pour la sélection d'échantillons présentant des valeurs proches et des caractéristiques similaires aident à rendre la base de données plus efficace, fluide et réactive à l'apprentissage.
- Le mode d'apprentissage automatique varie en fonction de l'objectif de l'étude, entre classification et régression, quantitatif et qualitatif, ainsi que s'il est supervisé ou non supervisé. Si l'objectif est de prédire des valeurs comme dans cette étude (valeurs de CBR), nous travaillons sur des modèles d'apprentissage automatique pour une régression, quantitative et supervisée.
- Il existe de nombreux modèles d'apprentissage automatique, les plus célèbres étant les réseaux neuronaux. L'historique mathématique et théorique de certains modèles remonte à des décennies voire au milieu du 16^{ème} siècle, depuis lors, les chercheurs travaillent à développer ces modèles jusqu'à l'émergence de leurs formes modernes avec l'évolution de la technologie aujourd'hui.
- L'évaluation statistique des modèles à travers le suivi des valeurs de régression R^2 et de coefficients des erreurs (MSE, RMSE, MAE et MAPE) montre clairement la performance des modèles entre ce qui est satisfaisant et ce qui est faible.
- L'inclusion de la technique de validation croisée K-fold permet d'améliorer la performance des modèles en optimisant la division optimale des données entre l'entraînement et le test, ce qui permet d'éviter le sur-ajustement (over-fitting).
- L'utilisation de la technique de sélection des caractéristiques d'entrées (Wrapper-based) du type récurive RFE avec le modèle de GPR et exhaustive EFS avec le modèle de SVR a permis de réduire la taille de la base de données de 12 entrées à seulement 7 entrées, réduisant ainsi le temps de traitement lors de la modélisation et améliorant la performance de ces deux modèles.
- L'amélioration des modèles par l'hybridation avec l'algorithme d'optimisation bayésienne BOA de la stratégie de l'estimateur de Parzen structuré arborescent TPE a permis de choisir de manière optimale les noyaux des modèles, compte tenu de la multiplicité des noyaux des modèles qui peuvent être perdus dans l'étude. De plus, cette hybridation a permis d'ajuster les hyper-paramètres de chaque modèle avec des valeurs optimales donnant une meilleure performance, ce qui a donné plus de satisfaction.
- Les résultats d'entraînement des modèles conventionnels sélectionnés (MLR, MLPNN d'ANN, RF et même les deux modèles GPR et SVR par défaut) ainsi que les modèles hybridés TPE-GPR-RFE

et TPE-SVR-EFS et mêmes des autres modèles modernes et hybridés ont montré une performance satisfaisante dans une certaine mesure. Ils ont varié en fonction du modèle, certains montrant des valeurs de régression R^2 proches ou dépassant 90% et des erreurs presque nulles, qui sont les valeurs de référence pour juger de la satisfaction de la performance et autre modèle (MLR) de faible performance.

- Par la comparaison de la performance des modèles TPE-GPR-RFE vs (GPR par défaut, CML modèles, SVR par défaut, TPE-SVR-EFS et même des autres modèles modernes-hybridés (H-DNN, H-XGB, H-LGB et H-Ada), il ressort que le modèle de Gaussian Process Regression GPR traité et hybridé avec l'algorithme d'optimisation bayésienne de stratégie TPE et soumis à l'intégration de la méthode de la sélection des caractéristiques d'entrées de type d'élimination récursive RFE, avec noyau Rational Quadratique RQ et ses hyper-paramètres ajustés ($l = 0.6192$, $\alpha = 1.5966$ et $\text{bruit} = 0.0406$ dans le cas de notre base de données) offre une meilleure performance que les autres modèles avec des valeurs sont données dans les paires (Entraînement/Test) pour la régression et les coefficients des erreurs comme suit:

$R^2 = 0.9982/0.9935$, $MSE = 0.0768/0.2974$, $RMSE = 0.2771/0.5453$, $MAE = 0.2107/0.4364$ et $MAPE(\%) = 5.32/13.79$.

- Cette comparaison met en évidence l'efficacité du couplage entre l'optimisation bayésienne et les techniques d'apprentissage automatique, confirmant la pertinence du modèle TPE-GPR-RFE. Ce dernier combine la puissance analytique du GPR, capable de modéliser des relations non linéaires complexes, avec la robustesse de la stratégie TPE pour la recherche d'hyper-paramètres optimaux. À titre de comparaison, le modèle hybride TPE-SVR-EFS, bien qu'efficace, présente une performance légèrement inférieure, en raison de la nature déterministe du SVR face à la flexibilité probabiliste du GPR. Ainsi, l'intégration synergique des approches bayésiennes et des méthodes de sélection de variables améliore significativement la précision et la stabilité des prédictions du CBR, renforçant la pertinence de ces modèles hybrides pour les applications en ingénierie géotechnique.
- La combinaison des approches SHAP et Sobol a permis d'interpréter plus finement les modèles hybrides et de réduire le caractère de « boîte noire » des prédictions. L'application de l'analyse SHAP sur le modèle TPE-GPR-RFE et de Sobol sur TPE-SVR-EFS a montré que la densité sèche maximale (MDD) exerce l'influence la plus marquée sur la sortie CBR, suivie par le teneur en eau optimale (OMC). Ces résultats, cohérents avec la littérature géotechnique, confirment le rôle déterminant du MDD dans la résistance mécanique des sols et soulignent l'importance de la compaction et des conditions hydriques pour la performance des matériaux géotechniques.
- Il est évident que les modèles GPR et SVR répond à cette norme d'optimisation (BOA/TPE, K-fold cross validation et feartures selection methods RFE/EFS) et donne des bons résultats en termes de performance, avec une nette supériorité du modèle hybride TPE-GPR-RFE.

On peut donc en conclure que le modèle hybride TPE-GPR-RFE est le modèle le plus performant pour prédire les valeurs de CBR du sol-support des structures de chaussées de texture fin (limon/argile) dans les régions nord-est de l'Algérie.

En guise de recommandations, on peut retenir les points suivants :

- Élargir la base de données et améliorer sa diversité, en incluant d'autres entrées telles que le module VBS, le groupe index GI et le poids spécifique G pourrait améliorer la performance.

- Inclure d'autres algorithmes d'apprentissage automatique ML et les améliorer soit avec l'algorithme d'optimisation bayésienne BOA/TPE, soit avec un autre algorithme d'optimisation pourrait améliorer la capacité de prédiction des modèles d'apprentissage machine.
- Il existe encore de nombreuses autres techniques d'analyse de sensibilité telles que Morris, LIME et l'Importance par permutation, qui peuvent être appliquées dans des études similaires. Elles peuvent être plus concluantes ou concordantes dans l'identification des variables d'entrée les plus influentes sur les sorties dans les études prédictives ou classificatoires.

R --- ***éférences bibliographiques***

Références bibliographiques :

- [1] Semachew Molla Kassa and Betelhem Zewdu Wubineh, Use of Machine Learning to Predict California Bearing Ratio of Soils, *Hindawi, Advances in Civil Engineering*, Article ID 8198648, page 11, January 25, 2023.
- [2] Yusuf Erzin, Ali Firat CABALAR, Zuheir KARABASH and Selin ERZIN, Artificial Neural Network Modeling For California Bearing Ration (CBR) Prediction in Clayey Soil Stabilised with Type Buffings and Lime, *Research Square*, 2022.
- [3] Van Quan Tran, Hai Quan Do, Prediction of California Bearing Ratio (CBR) of Stabilized Expansive Soils with Agricultural and Industrial Waste Using Light Gradient Boosting Machine, *Journal of Science and Transport Technology*, University of Transport Technology, JSST, 1 (1), page 1-9, 2021.
<https://doi:10.58845/jstt.utt.2021.en3>
- [4] SK Kamrul Alam, Abhijit Mondal and Amit Shiuly, Prediction of CBR Value of Fine Grained Soils of Bengal Basin by Genetic Expression Programming, *Artificial Neural Network and Krigging Method, JOURNAL* Vol. 95, page 190-196, February 2020.
- [5] T. Taskiran, Prediction of California bearing ratio (CBR) of fine-grained soils by AI methods, *Advances in Engineering Software*, 41(6), page 886-892, June 2010.
<https://doi:10.1016/j.advengsoft.2010.01.003>
- [6] Collecté d'après le site web de CTPP : <https://www.ctpp.dz> de visite de 04/09/2023 à 14 :09.
- [7] KHELLOU Abderrezak, Modélisation de comportement de matériaux sable gypso-calcaires, université de Ouargla, page 33-34, 2011.
- [8] Bedioun Zaid, Rapport de stage d'étudiant à l'école polytechnique Maurice Audin Oran, qui viste labo de LTP de Sétif en 2023.
- [9] Collecté d'après le site web : <http://fr.scribd.com> de visite de 09/09/2023 à 11 :57.
- [10] Collecté d'après le site web : <http://fr.bakhtissoftware.com> de visite de 07/09/2023 à 16 :17.
- [11] https://www.researchgate.net/figure/appareil-de-lessai-CBR_fig7_344220484
- [12] David F. McCarthy, Essentials of Soil Mechanics and Foundations: Basic Geotechnics, *Seventh Edition 'Pearson New International Edition'*, page 99-100 et 104 -108, 2014.
- [13] Régis Bourbonnais, Econométrie cours et exercices corrigés, 9^{ème} édition 'DUNOD', 2015.
- [14] Guido van Rossum, Introduction de la première édition du livre Programming Python de Mark Lutz [archive], 1996, la visite de 02/10/2023 à 15 :02.
- [15] Collecté d'après le site web : <http://docs.python.org> de visite de 02/10/2023 à 20:40.
- [16] Hastie, T, Tibshirani, R., & Friedman, J, The elements of statistical learning: data mining, inference and prediction. *Springer Science & Business Media*, 2009.

- [17] Scikit-learn documentation: https://scikit-learn.org/stable/modules/cross_validation.html
- [18] Vimal Rathakrishnan, Salmia Bt. Beddu and Ali Najah Ahmed, Predicting compressive strength of high-performance concrete with high volume ground granulated blast-furnace slag replacement using boosting machine learning algorithms, *scientific reports*, 2022.
- [19] Stéphanie ALLOGBA, Modèles de classification et de prédiction de performance des signaux optiques WDM basés sur les algorithmes d'apprentissage machine, université Québec, Montréal novembre, page 16-17, 2021.
- [20] Christopher M. Bishop, Pattern Recognition and Machine Learning.
- [21] Ian Goodfellow, Yoshua Bengio et Aaron Courville, Deep Learning.
- [22] Géron, A, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, 2019.
- [23] Kevin P. Murphy, Machine Learning: A Probabilistic Perspective, 2012.
- [24] Collecté d'après le site web : <https://wisecattv.wordpress.com> de visite le lundi 25/12/2023 à 18.40.
- [25] B. B. Pule, J. A. Yendaw, R. S. Jamisola Jr, O. Saubi. Predicting California Bearing Ratio Using Machine Learning to Model Soil Behavior for Road Construction in Tshimoyapula. BIUST Teaching, Research, & Innovation Symposium 2023, ISSN: 2521-2293, Botswana
- [26] A. M. Al-Swaidani; et al. Building MLR, ANN and FL models to predict the strength of problematic clayey soil stabilized with a combination of nano lime and nano pozzolan of natural sources for pavement construction. International Journal of Geo-Engineering, (2024) 15:2. <https://doi.org/10.1186/s40703-023-00201-1>
- [27] Collecté d'après le site web : <https://delladata.fr> de visite de 25/12/2023 à 18.53.
- [28] Mata J (2011) Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. Eng Struct 33(3):903–910
- [29] Khademi F, Akbari M, Jamal SM, Nikoo M (2017) Multiple linear regression, artificial neural network, and fuzzy logic prediction of 28 days compressive strength of concrete. Front Struct Civ Eng 11(1):90–99
- [30] Harini SN (2014) Prediction CBR of fine grained soils by artificial neural network and multiple linear regression. Int J Civil Eng Technol 5(2):119–126
- [31] Al-Swaidani A, Khweis W, Al-Bali M, Lala T (2022) Development of multiple linear regression, artificial neural networks and fuzzy logic models to predict the efficiency factor and durability indicator of nano natural pozzolana as cement additive. J Build Eng 52:104475
- [32] Schmidhuber, J, Deep learning in neural networks: An overview, *Neural networks*, 61, page 85-117, 2015.

- [33] Patrice Wira, Réseaux neurones artificiels : architectures et applications, université Haute Alsace, page 9,14 et page 49-50, Avril 2009.
- [34] Corentin Hardy, Contribution au développement de l'apprentissage profond dans les systèmes distribués, université de Rennes 1, page 13-14, 20-21, avril 2019.
- [35] Marie Cottrell, Les réseaux de neurones historique, méthodes et applications, université Paris 1 Sorbonne, CNRS UMR 8595.
- [36] Breiman, L, Random Forests, *Machine Learning*, 45(1), page 5-32, 2001.
- [37] Quinlan, J. R, Induction of Decision Trees, *Machine Learning*, 1(1), page 81-106, 1986.
- [38] Liaw, A & Wiener. M, Classification and regression by random forest, *R news*, 2(3), page 18-22, 2002.
- [39] Barber, D., & Williams, C. K. I. (1996). Gaussian Processes for Bayesian Classification via Hybrid Monte Carlo. In Advances in Neural Information Processing Systems NIPS 9.
- [40] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246–1266.
<https://doi.org/10.2113/gsecongeo.58.8.1246>
- [41] O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1), 1–42
- [42] Carl Edward Rasmussen & Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, 2006.
- [43] Hamoudi, Y; et al. Hyperparameter Bayesian Optimization of Gaussian Process Regression Applied in Speed-Sensorless Predictive Torque Control of an Autonomous Wind Energy Conversion System. *Energies*, MDPI, 2023, 16, 4738.
<https://doi.org/10.3390/en16124738>
- [44] Sheng, H.M.; Xiao, J.; Cheng, Y.H.; Ni, Q.; Wang, S. Short-Term Solar Power Forecasting Based on Weighted Gaussian Process Regression. *IEEE Trans. Ind. Electron* 2018, 65, 300–308.
- [45] Liu, K.L.; Hu, X.S.; Wei, Z.B.; Li, Y.; Jiang, Y. Modified Gaussian Process Regression Models for Cyclic Capacity Prediction of Lithium-Ion Batteries. *IEEE Trans. Transp. Electr* 2019, 5, 1225–1236.
- [46] Zhou, Y; et al. A novel combined multi-task learning and Gaussian process regression model for the prediction of multi-timescale and multi-component of solar radiation. *J. Clean. Prod* 2021, 284, 124710.
- [47] Li, X.Y.; Yuan, C.G.; Li, X.H.; Wang, Z.P. State of health estimation for Li-Ion battery using incremental capacity analysis and Gaussian process regression. *Energy* 2020, 190, 116467.

- [48] Khatti, J.; Grover, K.S. Determination of the optimum performance AI model and methodology to predict the compaction parameters of soils. *ICTACT Journal of Soft Computing* 2022, 12(3), 2640–2650.
<https://doi.org/10.21917/ijsc.2022.0368>
- [49] Ludkovski, M.; Risk, J. Gaussian Process models for Quantitative Finance. In *Springer Briefs in quantitative finance* 2025.
<https://doi.org/10.1007/978-3-031-80874-6>
- [50] Scikit-learn documentation: https://scikit-learn.org/stable/modules/gaussian_process.html
- [51] David Duvenaud, “The Kernel Cookbook: Advice on Covariance functions”, 2014
<https://www.cs.toronto.edu/~duvenaud/cookbook/>
- [52] Myren, S.; Lawrence, E. A comparison of Gaussian processes and neural networks for computer model emulation and calibration. *Stat. Anal. Data Min. ASA Data Sci. J.* 2021, 14, 606–623.
<https://doi.org/10.1002/sam.11507>
- [53] Galuzzi, B.G.; Giordani, I.; Candelieri, A.; Perego, R.; Archetti, F. Hyperparameter optimization for recommender systems through Bayesian optimization. *Comput. Manag. Sci.* 2020, 17, 495–515.
<https://doi.org/10.1007/s10287-020-00376-3>
- [54] Williams, C. K. I., & Rasmussen, C. E. (1996). Gaussian Processes for regression. *Advances in Neural Information Processing Systems*, 8, 514-520.
- [55] Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning: Adaptive Computation and Machine Learning*. Cambridge, MA: MIT Press. ISBN 978-0-262-18253-9.
- [56] Gelman, A; John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, et Donald B. Rubin, *Bayesian Data Analysis*, 2003.
- [57] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press. ISBN 978-0-262-01802-9
- [58] V.N. Vapnik, *The Nature of Statistical Learning Theory*, Springer New York, 1995.
<http://dx.doi.org/10.1007/978-1-4757-2440-0>
- [59] H. Su, X. Li, B. Yang and Z. Wen, Wavelet support vector machine-based prediction model of dam deformation, *Mech-Syst. Signal Process*, (110), page 412–427, 2018.
<http://dx.doi.org/10.1016/j.ymssp.2018.03.022>
- [60] Andrea Vedaldi et Andrew Zisserman, *Support Vector Machines: History and Applications*.
- [61] Cristianini. N. et Shawe-Taylor. J, *Support Vector Machines: A Survey*.
- [62] Gustavo Isaza et al, *A Gentle Introduction to Support Vector Machines in Biomedicine: Theory and Methods*.

- [63] Corinna Cortes et Vladimir Vapnik, Support Vector Machines. (Le travail original sur les SVR).
- [64] Scikit-learn documentation:
https://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html
- [65] Nate. Rosidi, Advanced Feature Selection Techniques for Machine Learning Models, *KDnuggets Market Trends & SQL Content Specialist*, June 6, 2023.
<https://www.kdnuggets.com/2023/06/advanced-feature-selection-techniques-machine-learning-models.htm>
- [66] Theng. D, Bhoyar. K.K, Feature selection techniques for machine learning: a survey of more than two decades of research, *Knowl Inf Syst*, 66, page 1575–1637, 2024.
<https://doi.org/10.1007/s10115-023-02010-5>
- [67] Horn. F, A Practitioner’s Guide to Machine Learning, *Book*, 2024.
ranziskahorn.de/mlbook_all.html
- [68] Murty. NM, Ananthanarayana, VS. Machine Learning: Theory and Practice, *Book*, ISBN: 9789393330697, Paperback, page: 372, 2024.
- [69] Collecté d’après le site web : <https://aws.amazon.com/fr/what-is/hyperparameter-tuning/> de visite de 22/09/2024 à 10 :02.
- [70] E. Brochu, V.M. Cora et de N. Freitas, A Tutorial on Bayesian Optimization of Expensive Cost Functions with Application to Active User Modeling and Hierarchical Reinforcement Learning, *arXiv*, page 1012-2599 [cs], 2010. Récupéré le 2019-01-05 de
<http://1/arxiv.org/abs/1012.2599>
- [71] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, in: *Adv. Neural Inf. Process. Syst*, page 2951–2959, 2012.
- [72] Faten Belarbi, OPTIMISATION DES HYPERPARAMÈTRES POUR LA CLASSIFICATION DES VIRUS, Mémoire présentée comme exigence partielle de la maîtrise en informatique, UNIVERSITÉ DU QUÉBEC À MONTRÉAL, page 47-48, Juillet 2019.
- [73] Collecté d’après le site web : <https://inside-machinelearning.com/optimisation-bayesienne/> de visite de 22/09/2024 à 11:22.
- [74] Collecté d’après le site web : <https://fastercapital.com/fr/contenu/Reglage-des-hyperparametres-optimisation-des-hyperparametres-avec-des-modeles-par-default.html/> de visite de 22/09/2024 à 11:40.
- [75] B. Shahriari, K. Swersky, Z. Wang, R.P. Adams and N. De Freitas, Taking the human out of the loop: A review of Bayesian optimization, *IEEE Conference Proceeding*, 104, p. 148-175, 2016.
<https://dx.doi.org/10.1109/JPROC.2015.2494218>
- [76] D.T. Chang, Bayesian hyperparameter optimization with BoTorch, GPyTorch and Ax, *Computer Science-Machine learning*, 2019.
<https://doi.org/10.48550/arXiv.1912.05686>

- [77] Lim, Z.; Pang, Y.; Kamarudin, K.; Yin, O.; Hiew, F. Bayesian optimization driven strategy for detecting credit card fraud with extremely randomized trees. *Methodsx* 2024, 13, 103055.
<https://doi.org/10.1016/j.mex.2024.103055>
- [78] Xu, B.; Sun, L.; Mao, X.; Ding, R.; Liu, C. Iot intrusion detection system based on machine learning. *Electronics* 2023, 12(20), 4289.
<https://doi.org/10.3390/electronics12204289>
- [79] Zhang, X.; Dai, C.; Li, W.; Chen, Y. Prediction of compressive strength of recycled aggregate concrete using machine learning and Bayesian optimization methods. *Frontiers in Earth Science* 2023, 11.
<https://doi.org/10.3389/feart.2023.1112105>
- [80] Aydoğmuş, Ö.; Yılmaz, M. Comparative analysis of reinforcement learning algorithms for bipedal robot locomotion. *Ieee Access* 2024, 12, 7490-7499.
<https://doi.org/10.1109/access.2023.3344393>
- [81] Melainine, C.; Socheleau, F.; Jarrot, A. Self-configurable receiver for underwater acoustic communications using bayesian optimization. 2024, 1-5.
<https://doi.org/10.1109/ucomms64662.2024.10847986>
- [82] Song, Y.; et al. Tree-structured parzan estimator–machine learning–ordinary kriging: an integration method for soil ammonia spatial prediction in the typical cropland of Chinese yellow river delta with sentinel-2 remote sensing image and air quality data. *Remote Sensing* 2023, 15(17), 4268.
<https://doi.org/10.3390/rs15174268>
- [83] Huang, Y.; et al. A machine learning approach to brain epigenetic analysis reveals kinases associated with alzheimer’s disease. *Nature Communications* 2021, 12(1).
<https://doi.org/10.1038/s41467-021-24710-8>
- [84] J. Mockus, *Bayesian Approach to Global Optimization*, Springer. Dordrecht, page 270, 2012.
<http://dx.doi.org/10.1007/978-94-009-0909-0>
- [85] Snoek, J.; Larochelle, H.; Adams, R.P. Practical Bayesian optimization of machine learning algorithms. *Adv Neural Inf Process Syst* 2012, 25.
- [86] Bibliothèque PyPi: <https://pypi.org/>
- [87] Bibliothèque HyperOpt: <https://hyperopt.github.io/hyperopt/>
- [88] Bibliothèque GpyOpt: <https://pypi.org/project/GPyOpt/>
- [89] Bergstra, J; Bardenet, R; Bengio, Y and Kégl, B. Algorithms for Hyper-Parameter Optimization. NIPS, 2011.
- [90] Sandjak, K., Ouanani, M. Bayesian optimization algorithm based support vector regression analysis for estimation of resilient modulus of crushed rock materials for pavement design. *MATEC Web of Conferences WMCAUS*, Prague, Czech Republic. 2024, 396, pp. 05-16.

- [91] Wu, X; Lu, F and He, T. Exploring the Potential of Machine Learning in Predicting Soil California Bearing Ratio Values. *Periodica Polytechnica Civil Engineering*, 69(2), pp. 551-566, 2025.
<https://doi.org/10.3311/PPci.38678>
- [92] Hamby, D. M. (1994). A review of techniques for parameter sensitivity analysis of environmental models. *Environmental Monitoring and Assessment*, 32(2), 135–154.
<https://doi.org/10.1007/BF00547132>
- [93] Cacuci, D. G. (2003). *Sensitivity & Uncertainty Analysis — Volume 1: Theory*. Chapman & Hall / CRC (book).
<https://doi.org/10.1201/9780203498798>
- [94] Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons (book), pp. 232, ISBN 0-470-87093-1, Wiley book identifier.
<https://doi.org/10.1002/0470870958>
- [95] Campolongo, F., Cariboni, J., & Saltelli, A. (2007). An effective screening design for sensitivity analysis of large models. *Environmental Modelling & Software*, 22(10), 1509-1518.
<https://doi.org/10.1016/j.envsoft.2006.10.004>
- [96] Helton, J. C. (1993). Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive-waste disposal. *Reliability Engineering & System Safety*, 42(2-3), 327-367.
[https://doi.org/10.1016/0951-8320\(93\)90097-I](https://doi.org/10.1016/0951-8320(93)90097-I)
- [97] Morris, M. D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics*, 33(2), 161–174.
<https://doi.org/10.1080/00401706.1991.10484804>
- [98] Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation*, 55(1–3), 271-280.
[https://doi.org/10.1016/S0378-4754\(00\)00270-6](https://doi.org/10.1016/S0378-4754(00)00270-6)
- [99] Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst* 2017, 30.
- [100] Lundberg, S.M.; et al. From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell* 2020, 2, pp. 56–67.
- [101] Gowda, M. K. S.; Murthy, Y. I.; Gupta, A. Explainable AI based support vector machine models for soaked CBR prediction. *International Journal of Pavement Research and Technology* 2025.
<https://doi.org/10.1007/s42947-025-00558-9>
- [102] Moncada-Torres, A.; et al. Explainable Machine Learning Can Outperform Cox Regression Predictions and Provide Insights in Breast Cancer Survival. *Sci. Rep* 2021, 11, 6968.
- [103] Thapa, I.; Ghani, S. Explainable AI-Enhanced machine learning for CBR prediction in stabilized and unstabilized subgrade soils. *Applied Soft Computing* 2025, 113275.

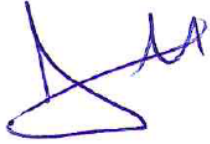
<https://doi.org/10.1016/j.asoc.2025.113275>

- [104] Doshi SN, Mesdary MS, Guirguis HR, Conference: statistical study of laboratory CBR for Kuwaiti soils, In: *Fourth conference of the road engineering association of Asia and Australia*, vol. 2, page 43-51, Jakarta, Indonesia, 1983.

Le président de jury de soutenance :

Dr. Attia Ahmed

Fait à : Djelfa, le : 14/07/2026



- **Version finale corrigée et prête à être déposée.**