RAPPORT : PREDICTION DE LA MALADIE DE PARKINSON AVEC L'APPRENTISSAGE AUTOMATIQUE

Réalise par:

Imad El Maftouhi

Zitan Houssam

Amgrout Zakaria

Sous-encadrement :

Professeur Ait Kbir M’hamed

Le 01, Decembre 2024

MST IASD/S1 2024-2025

Contents

ABSTRACT

La maladie de Parkinson (MP) est une affection neurodégénérative progressive qui provoque des symptômes moteurs et non moteurs.

Ses symptômes se développent lentement, ce qui rend difficile une identification précoce. L'apprentissage automatique a un potentiel important pour prédire la maladie de Parkinson à partir des caractéristiques cachées dans les données vocales.

Ce travail a pour but d'identifier les caractéristiques les plus pertinentes à partir d'un ensemble de données à haute dimension, ce qui permet de classer avec précision la maladie de Parkinson avec un temps de calcul réduit. Un ensemble de données individuels comportant diverses caractéristiques médicales basées sur la voix ont été analysés dans ce travail. Une technique d'algorithme de sélection de caractéristiques d'ensemble (EFSA) basée sur des algorithmes de filtrage, d'enveloppement et d'intégration qui sélectionnent les caractéristiques les plus pertinentes pour la classification de la maladie de Parkinson. Ces techniques peuvent réduire le temps de formation pour améliorer la précision du modèle et minimiser l'ajustement excessif.

1. INTRODUCTION

Les modèles d'apprentissage automatique (AA) ont montré leur efficacité dans l'analyse de données médicales comme le texte, la voix et les images pour diagnostiquer des maladies telles que la maladie de Parkinson. Leur performance repose sur des ensembles de données d'entraînement conséquents pour éviter le surapprentissage.

Des techniques d'échantillonnage, comme le sur- et sous-échantillonnage, équilibrent les données pour améliorer les résultats et réduire les biais. L'Algorithme de Sélection d'Ensemble de Caractéristiques (EFSA) combine des méthodes de filtrage, d'enveloppement et d'intégration pour optimiser la sélection des caractéristiques, augmentant la précision tout en réduisant le temps de calcul.

Cette étude utilise des classificateurs AA appliqués à des données vocales de patients atteints ou non de Parkinson, mettant en avant des caractéristiques acoustiques comme la hauteur et la gigue. L'approche EFSA vise à développer un modèle de diagnostic rapide et précis pour des applications de télémédecine, promettant des bénéfices significatifs pour les patients et les médecins.

2. Méthodologie de travail :

Dans ce projet, nous avons adopté une méthodologie structurée combinant plusieurs étapes clés pour garantir des résultats fiables et pertinents. Tout d'abord, nous avons procédé à une exploration approfondie des données afin de comprendre leurs caractéristiques et de détecter d'éventuelles anomalies. Ensuite, nous avons appliqué un algorithme de sélection de caractéristiques, EFSA, pour identifier les variables les plus pertinentes liées à la maladie de Parkinson. Cette étape a impliqué l'utilisation de méthodes combinées (filtre, wrapper et embedded) afin de tirer parti des points forts de chaque approche. Une fois les caractéristiques sélectionnées, nous avons entraîné et évalué plusieurs modèles de classification (régression logistique, Random Forest,LightGBM) en ajustant leurs hyperparamètres pour optimiser leurs performances. Enfin, nous avons intégré les meilleurs modèles au sein d'un classificateur à vote d'ensemble, tout en évaluant leur précision à l'aide de mesures rigoureuses. Cette méthodologie intégrative a permis de développer un cadre robuste pour le diagnostic prédictif.

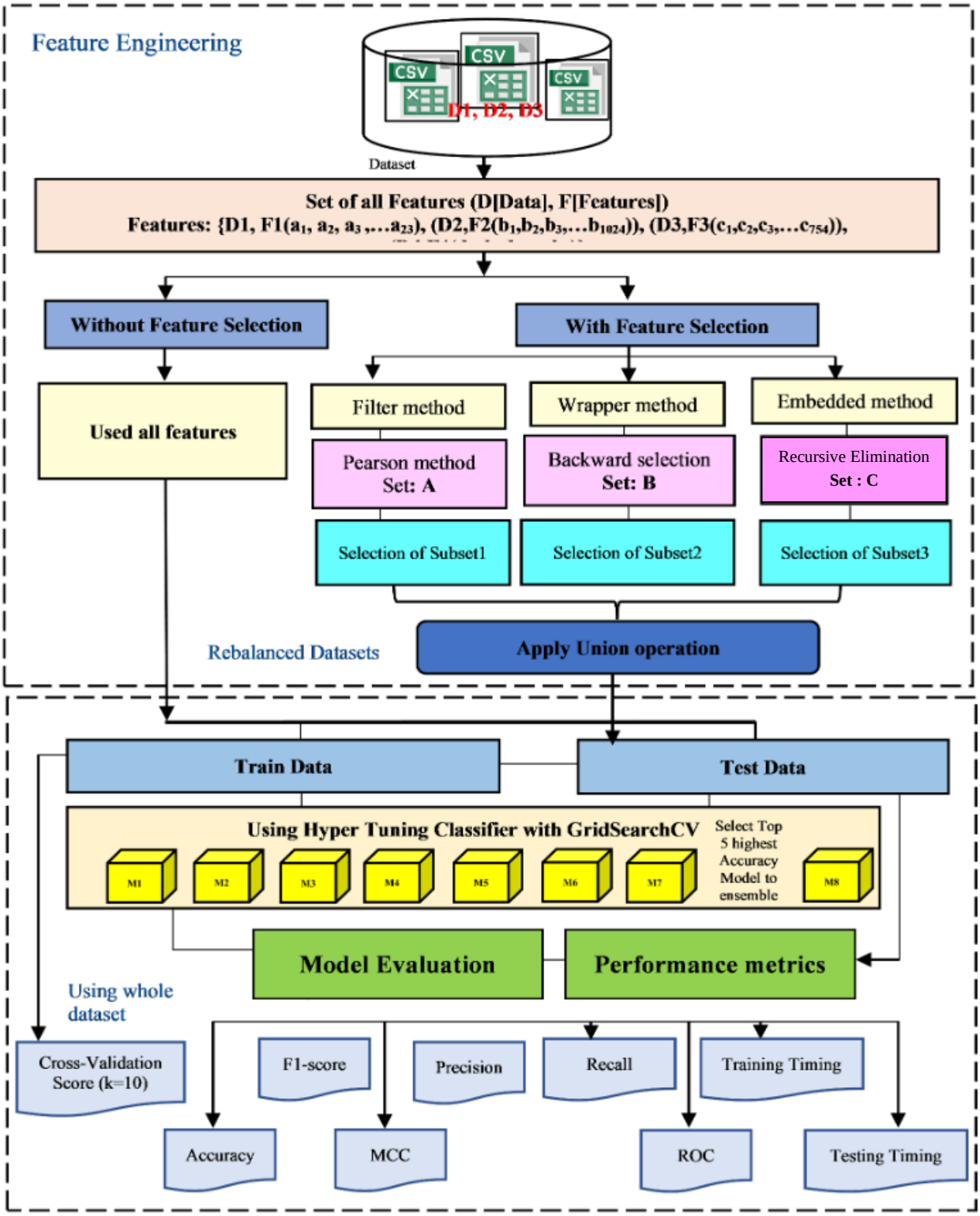


Figure 1 : méthodologie de travail adoptéeEnsembe de Jeu

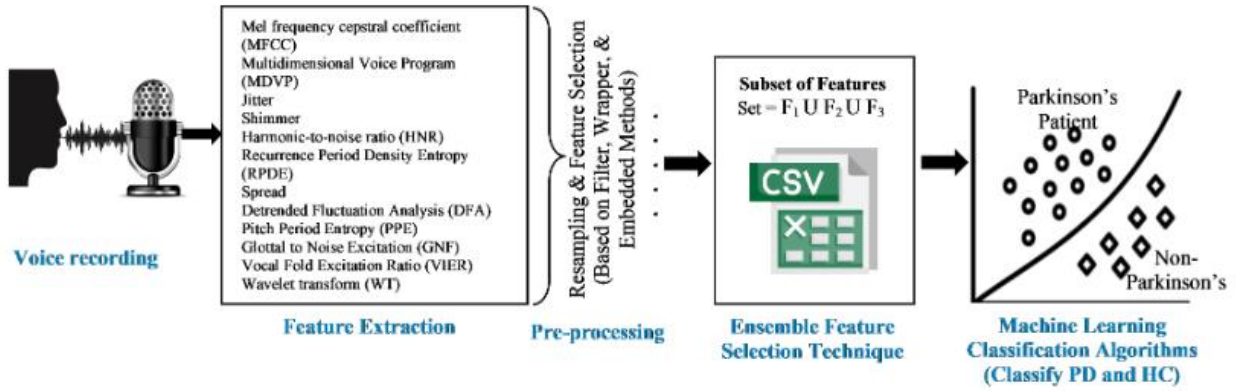


Figure 2 : Processus d'ensemble features selection algorithms

---

Algorithm 1: EFSA

Entrées :

- **DI** : Un jeu de données avec des caractéristiques numériques réelles F avec

  $F = \{f_1, f_2, \dots, f_k\}$, avec $k = 1,2,3$ une variable catégorique de classe.

- **R** : Un seuil de pertinence des caractéristiques qui sépare les caractéristiques pertinentes des non pertinentes.

Étape CDA1

- Calculer la corrélation entre chaque caractéristique et la variable cible.
- Sélectionner les caractéristiques ayant une corrélation supérieure au seuil RR. **Retourner :** Les caractéristiques pertinentes AA.

Étape CDA2

- Appliquer une méthode d'élimination arrière (Backward Elimination) basée sur un wrapper :

  1. Initialiser **features** avec la liste des colonnes **data.columns.**
  2. Initialiser **remaining_features** avec la liste des colonnes **features.**
  3. Calculer la précision initiale **acc** en utilisant toutes les colonnes avec le modèle **Logistic Regression**

4. Pour chaque colonne dans **features :**
    - Retirer la colonne courante de **remaining_features** .
    - Calculer la précision **temp_acc** en utilisant les colonnes restantes.
    - Si **temp_acc** < **acc** : rajouter la colonne courante.
5. Si le nombre de colonnes dans **remaining_features** reste identique après un tour, arrêter la boucle.

Retourner : Les caractéristiques sélectionnées BB.

Étape CDA3

- Appliquer une sélection basée sur SelectFromModel et **RFE*** :
    1. Ajuster un modèle **Logistic Regression**
    2. Calculer les coefficients absolus des caractéristiques.
    3. Identifier les indices des **n** caractéristiques les plus importantes.
    4. Extraire les noms des colonnes correspondant aux indices des caractéristiques les plus importantes.

Retourner : Les caractéristiques sélectionnées CC.

Résultat Final : Combiner les ensembles A, B, et C : $PEF = A \cup B \cup C$

**RFE = Recursive Feature Elimination algorithm*

6. Méthodes et Ensemble de Jeu

Aperçu d'ensemble de jeux

Nom du Jeu de Données	Source	Nombre de Caractéristiques	Nombre d'Instances	Répartition des Classes	Description
Dataset-II	UCI Machine Learning Repository Dataset	754	756	PD: 188 (107 hommes, 81 femmes), HC: 64	Jeu de données vocales déséquilibré avec un grand nombre de caractéristiques, nécessitant une pré-sélection.

Sélection des caractéristiques :

L'approche EFSA innovante combine trois stratégies de sélection de caractéristiques : un filtre de corrélation de Pearson, une technique d'élimination à rebours et RFE. En fusionnant ces méthodes complémentaires, l'algorithme identifie un sous-ensemble de caractéristiques vocales diversifié et ciblé, améliorant significativement la prédiction du modèle pour le diagnostic de la maladie de Parkinson.

Module EFSA = feature_selection.py

Prétraitement des données

Le prétraitement des données en Machine Learning est crucial pour améliorer l'analyse et la précision des algorithmes. Les valeurs nulles résultent souvent de lacunes dans la collecte ou l'observation des données médicales. Les techniques utilisées incluent :

- Transformation des caractéristiques avec StandardScaler()

- Mise à l'échelle de chaque caractéristique dans une fourchette prédéterminée

- Division de l'ensemble de données en 80% d'entraînement et 20% de test

Cette approche permet de standardiser les données, réduire les biais et préparer efficacement le jeu de données pour l'analyse.

Module preprocessing : preprocessing.py

Balancement des classes

La technique SMOTE (Synthetic Minority Over-sampling Technique) résout le déséquilibre des classes en générant des échantillons synthétiques pour la classe minoritaire. Le processus consiste à :

1. Sélectionner chaque instance minoritaire

2. Choisir un voisin minoritaire proche

3. Créer une nouvelle instance synthétique entre ces deux points

L'objectif est d'équilibrer la distribution des classes, améliorant ainsi la performance et la représentativité du modèle d'apprentissage automatique.

Tableau 1 : Description ensemble de jeux

Dataset	Forme	Avant Pré-traitement		Après pré-traitements		Avec EFSA	
		Entrainement	Test	Entrainement	Test	Entrainement	Test
Dataset-II	756, 754	529, 754	227, 754	788, 754	338, 754	788, 137	338, 137

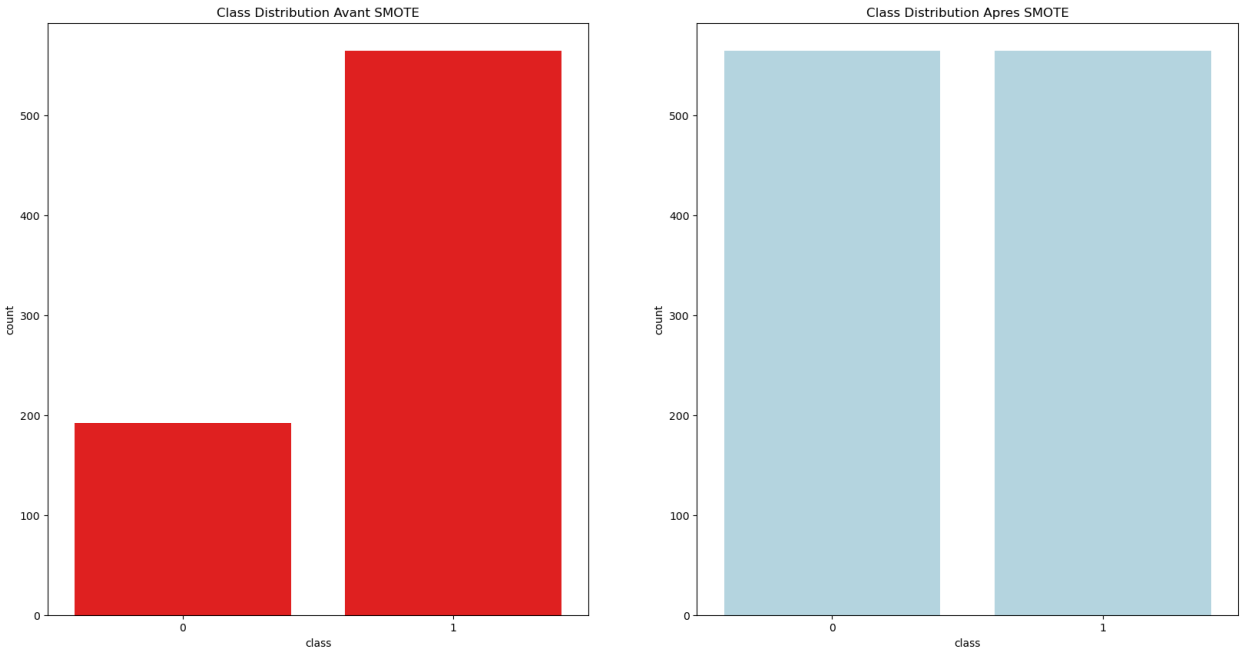


Figure 3 : Rééchantillonnage avec SMOTE

7. Les modèles Développe

Decision Tree:

Dans ce projet, nous avons adopté l'**arbre de décision** comme modèle de classification pour prédire la maladie de Parkinson. L'arbre de décision est une méthode d'apprentissage supervisé qui divise les données en sous-groupes basés sur des critères de décision simples, à chaque nœud de l'arbre, pour aboutir à une prédiction finale. Dans notre approche, l'algorithme utilisé est basé sur la méthode **C4.5**, qui permet de construire des arbres optimisés en utilisant la **mesure d'entropie** et le **gain d'information** pour déterminer les meilleurs critères de séparation à chaque étape. Ce modèle est facile à interpréter et offre l'avantage d'identifier les interactions clés entre les caractéristiques vocales des patients, ce qui le rend particulièrement adapté pour des applications en santé, où l'interprétabilité est essentielle.

Random Forest :

Le Random Forest est une méthode d'apprentissage supervisé qui combine plusieurs arbres de décision pour améliorer la précision. Elle repose sur la sélection aléatoire de sous-ensembles de caractéristiques, l'utilisation de la technique de Bootstrap pour créer des sous-échantillons, la construction d'arbres avec l'algorithme C4.5, et une prédiction finale réalisée par vote majoritaire.

Régression Logistique :

La régression logistique est une méthode de classification binaire qui prédit la probabilité d'appartenance à une classe. Elle utilise la fonction sigmoïde pour transformer les sorties continues en probabilités entre 0 et 1. Son fonctionnement repose sur une combinaison linéaire des caractéristiques d'entrée, la minimisation de la fonction de log-vraisemblance et l'estimation des paramètres du modèle.

LightGBM :

LightGBM (Light Gradient Boosting Machine) est un algorithme de gradient Boosting rapide et efficace, idéal pour traiter de grands ensembles de données. Il construit des arbres de décision séquentiels, chaque arbre corrigeant les erreurs des précédents, avec une approche feuille par feuille pour accélérer l'apprentissage. Il gère efficacement des jeux de données volumineux tout en évitant le surapprentissage via un ajustement dynamique des paramètres. Sa rapidité, sa flexibilité et ses performances en font un choix privilégié pour les problèmes complexes et les compétitions de Machine Learning

8. Expérimente et résultats

Tableau d'évaluation

Tableau 2 : Tableau d'évaluation de performance des models

Sans EFSA							
Model	**Test Accuracy**	**F1-score**	**Précision**	**Recall**	**MCC %**	**Training Time (sec)**	**Testing time (sec)**
Logistic Regression	90.265 %	90.228 %	90.279 %	90.191	80.47 %	6.51	0.001158
Random Forest	86.726 %	86.659 %	88.71 %	87.348 %	76.045 %	9.608	0.228
LightGBM	87.610 %	87.595 %	87.575 %	87.66 %	75.245 %	2382	0.029
Avec EFSA							
Model	**Test Accuracy**	**F1-score**	**Precision**	**Recall**	**MCC %**	**Training Time (sec)**	**Testing time (sec)**
Logistic Regression	89.823 %	89789 %	89.811 %	89.771 %	79.582 %	4.725	0.0
Random Forest	86.268 %	86.22 %	88.092	86.881 %	74.963 %	5.21	0.256
LightGBM	87.168 %	87.165 %	87.265 %	87.344 %	74.61 %	16.805	0.0198

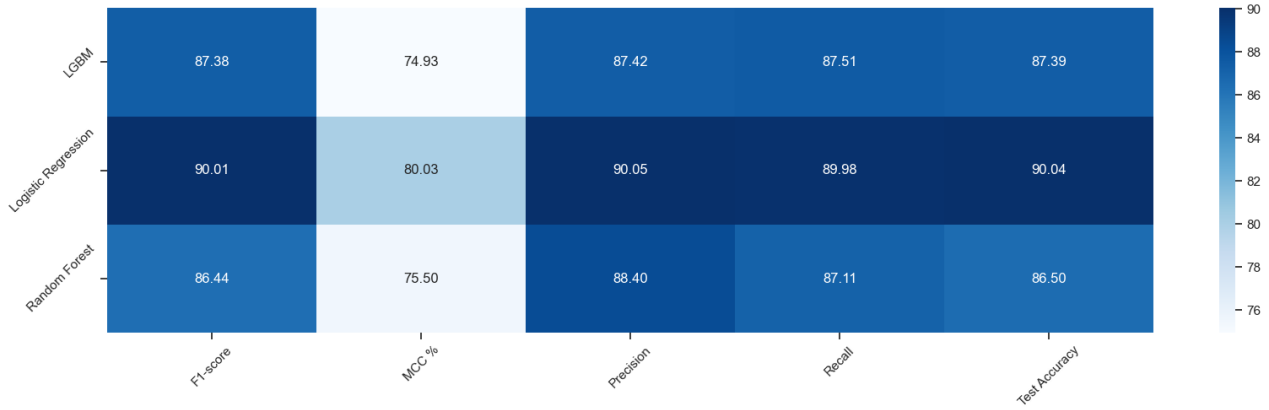


Figure 4 : Evaluation des models avec EFSA

L'efficacité du modèle proposé est évaluée par l'application de cinq formules distinctes, comme l'illustrent les équations (1), (2), (3), (4) et (5). Les performances des algorithmes de catégorisation sont évaluées sur la base de trois mesures clés : la précision, la sensibilité et la spécificité.

Accuracy d'un algorithme de classification peut être définie comme la somme du nombre de vrais positifs (TP) et de vrais négatifs (TN) sur le total des vrais positifs (TP), des faux positifs (FP), des faux négatifs (FN) et des vrais négatifs (TN) est donné par l'équation suivante : $\frac{TP+TN}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}+\mathrm{TN}}$ (1)

Recall est défini comme la proportion d'instances positives correctement identifiées, exprimée sous forme de ratio. $\frac{TP}{\mathrm{TP}+\mathrm{FN}}$ (2)

Precision est définie comme le rapport entre le nombre de vrais positifs et le nombre total de positifs sur la somme des vrais positifs et des faux positifs $\frac{TP}{TP+FP}$(3).

Le **score F1** est calculé en multipliant la valeur de précision par la valeur de rappel, ce qui donne une valeur égale a : $\frac{2\times(\text{précision}\times\text{rappel})}{(\text{précision}+\text{rappel})}$ (4)

L'équation suivante représente le calcul du coefficient de corrélation de Matthews (MCC) :

MCC = $\frac{\mathrm{TP}*\mathrm{TN}-\mathrm{FP}*\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})*(TP+FN)*(TN+FP)*(TN+FN)}}$ (5)

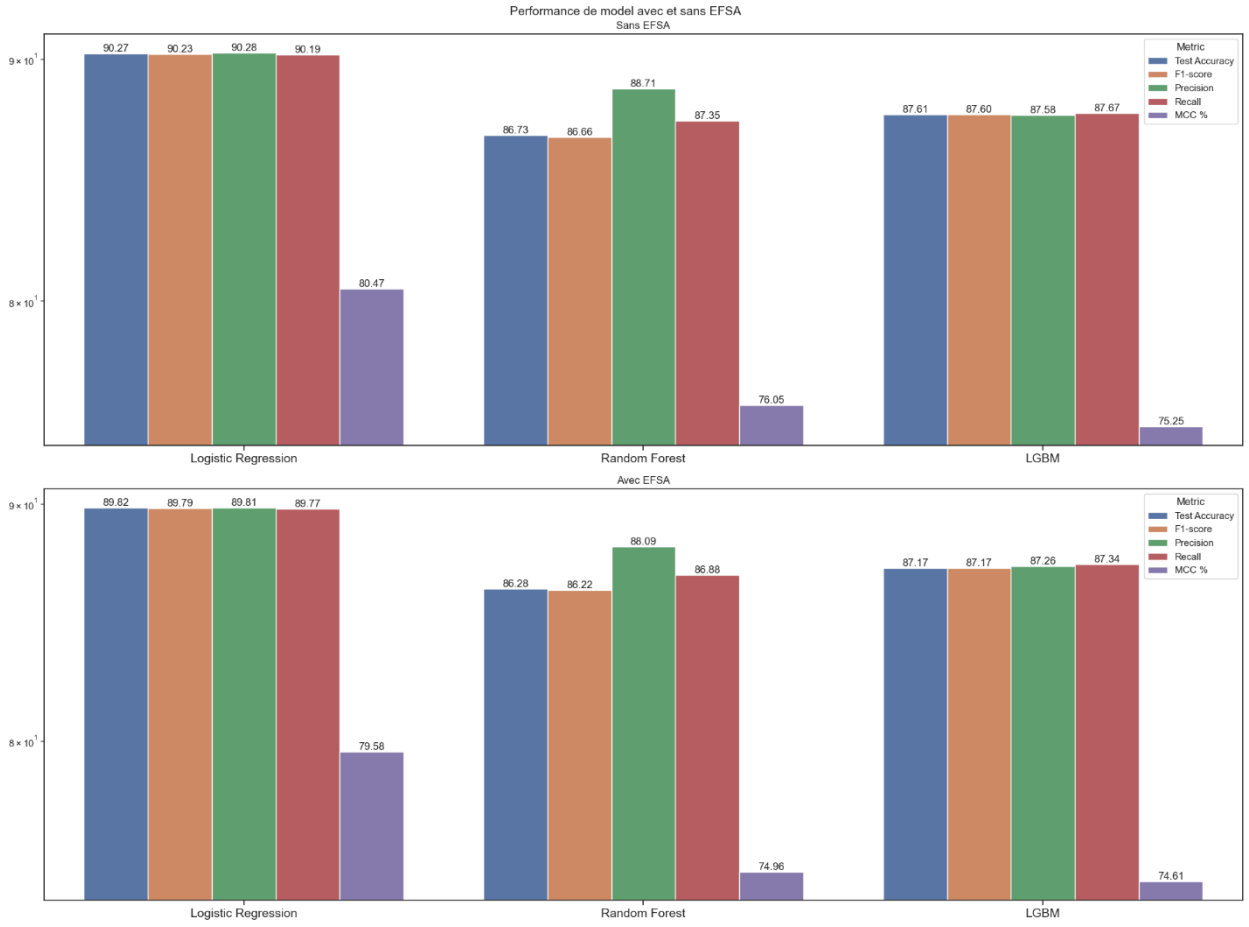


Figure 5 : Performances des models avec et sans EFSA

9. Optimisation par GridSearch et Cross-Validation

Grid Search

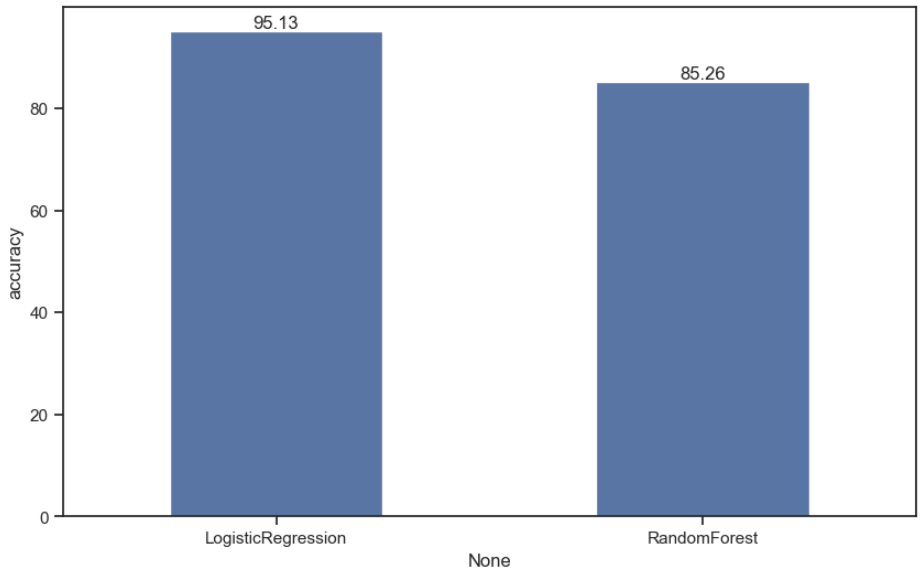
- **Logistic Regression**:
  - Un tableau de paramètres a été défini pour le taux d'apprentissage, epsilon et le nombre maximum d'itérations.
  - Chaque combinaison de paramètres a été testée pour déterminer la meilleure précision.
- **Random Forest**:
  - Les paramètres optimaux ont été recherchés, incluant le nombre d'arbres, la profondeur maximale des arbres, et le nombre minimum de données par arbre.
  - Une recherche en grille a été réalisée en utilisant une validation croisée K-Fold (5 plis) pour évaluer la robustesse du modèle.

Évaluation des Modèles

Les performances des modèles ont été mesurées par leur précision moyenne.

	accuracy	**params**
LogisticRegression	95.132743	(0.1, 1e-05, 1500)
RandomForest	85.255851	(15, 50, class, 30)

5. Résultats Finaux



10. Classificateur propose

Après avoir déterminé les paramètres optimaux pour chaque modèle et réduit la dimensionnalité de notre ensemble de données grâce à l'algorithme EFSA, nous avons conçu une méthode d'ensemble pour combiner les forces des trois modèles testés : la régression logistique, la forêt aléatoire, et LightGBM.

La méthode repose sur un vote majoritaire entre les prédictions des trois modèles. Pour chaque échantillon, les prédictions individuelles sont agrégées, et la classe qui obtient le plus grand nombre de votes est sélectionnée comme la classe finale. Cela permet de tirer parti de la complémentarité des modèles :

Avantages de la méthode d'ensemble

- **Robustesse accrue :** La combinaison des modèles diminue l'impact des éventuels biais ou faiblesses spécifiques à un modèle.
- **Amélioration des performances :** Les résultats montrent que le vote majoritaire améliore légèrement la précision globale et le F1-score comparé à l'utilisation individuelle des modèles.
- **Adaptabilité :** Cette approche peut facilement être adaptée à d'autres ensembles de données et modèles.

Résultats

En appliquant ce classificateur d'ensemble, nous avons atteint une précision finale de **85,4%**, démontrant que la méthode d'ensemble peut exploiter efficacement les contributions des différents modèles pour fournir une prédiction cohérente et fiable.

Perspectives

Bien que la méthode de vote majoritaire soit simple et efficace, elle pourrait être enrichie en pondérant les votes en fonction des performances spécifiques de chaque modèle sur les données d'entraînement ou en adoptant des méthodes d'ensemble plus avancées telles que le stacking ou le boosting.

11. Conclusion

Dans ce travail, nous avons démontré l'efficacité de l'algorithme de sélection d'ensemble de caractéristiques (EFSA) dans le contexte de la classification de la maladie de Parkinson. En intégrant les approches Filter, Wrapper et Embedded, EFSA a permis de sélectionner les caractéristiques les plus pertinentes, réduisant ainsi la dimensionnalité tout en améliorant les performances des modèles. Les résultats montrent que les modèles de régression logistique, forêt aléatoire et LightGBM bénéficient de cette méthode, comme en témoignent les améliorations des métriques telles que le F1-score et la précision. Cependant, l'impact d'EFSA varie selon les algorithmes, la régression logistique affichant les meilleurs résultats globaux. Ces résultats soulignent l'importance de la sélection de caractéristiques pour traiter les ensembles de données complexes et déséquilibrés, en vue d'optimiser les performances des modèles d'apprentissage automatique.

Bibliographie

Articles principaux liés au projet

1. **Singh, N., & Tripathi, P. (2024).**
   An ensemble technique to predict Parkinson's disease using machine learning algorithms. *Speech Communication*, 159, 103067.
   DOI: 10.1016/j.specom.2024.103067.
   Résumé : Cet article propose une méthode basée sur un algorithme d'ensemble pour la sélection de caractéristiques (EFSA) et un classificateur de vote pour prédire la maladie de Parkinson. Les résultats montrent une amélioration significative des performances des modèles grâce à la réduction dimensionnelle et l'utilisation de techniques d'optimisation des paramètres.

Autres références clés citées dans l'article

2. **Little, M.A., McSharry, P.E., Hunter, E.J., Spielman, J., & Ramig, L.O. (2009).**
   Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4), 1015–1022.
   DOI: 10.1109/TBME.2008.2005954.
   Résumé : Étude pionnière sur l'utilisation de caractéristiques vocales pour surveiller la maladie de Parkinson à distance.

3. **Sakar, B.E., & Kursun, O. (2010).**
   Telediagnosis of Parkinson's disease using measurements of dysphonia. *Journal of Medical Systems*, 34(4), 591–599.
   DOI: 10.1007/s10916-009-9268-2.
   Résumé : Introduction des caractéristiques basées sur l'information mutuelle pour améliorer la précision des modèles.

4. **Lamba, R., Gulati, T., & Jain, A. (2022).**
   A hybrid feature selection approach for Parkinson's detection based on mutual information gain and recursive feature elimination. *Arabian Journal for Science and Engineering*, 47(8), 10263–10276.
   DOI: 10.1007/s13369-021-06544-0.
   Résumé : Présentation d'approches hybrides combinant des techniques de sélection et d'extraction des caractéristiques.

5. **Das, R. (2010).**
   A comparison of multiple classification methods for diagnosis of Parkinson disease. *Expert Systems with Applications*, 37(2), 1568–1572.
   DOI: 10.1016/j.eswa.2009.06.040.

6. **Hussain, M.M., Weslin, D., Kumari, S., et al. (2023).**
Enhancing Parkinson's disease identification using ensemble classifier and data augmentation techniques. *Clinical eHealth*, 6, 150–158.
DOI: 10.1016/j.ceh.2023.11.002.
Résumé : Approche combinée utilisant des classificateurs ensemblistes et des techniques d'augmentation de données pour améliorer les prédictions.

7. **El-Hasnony, I.M., Barakat, S.I., & Mostafa, R.R. (2020).**
Optimized ANFIS model using hybrid metaheuristic algorithms for Parkinson's disease prediction in IoT environment. *IEEE Access*, 8, 119252–119270.
DOI: 10.1109/ACCESS.2020.3005614.
Résumé : Modèle ANFIS optimisé pour un diagnostic précis en temps réel dans des environnements IoT.