Université abdelmalek essaâdi

Faculté des Sciences et Techniques

-Tanger-

MASTER SCIENCES ET TECHNIQUES EN INTELLIGENCE ARTIFICIELLE ET SCIENCES DE DONNEES

RAPPORT DE MINI PROJET DU MODULE MACHING LEARNING I

SUJET : Détection de la Maladie de Parkinson à l'Aide de Techniques d'Apprentissage Supervisé : Étude Comparative des Méthodes de Sélection et Réduction des Caractéristiques

Réalisé par (G8) :

Ayoub Najjout

Ibrahim Berradi

Reda Wafik

Encadré par : Pr. M'hamed AIT KBIR

Année Universitaire : 2024-2025

Table des matières

Introduction

La maladie de Parkinson est une affection neurodégénérative progressive qui touche des millions de personnes dans le monde. Elle se manifeste principalement par des troubles moteurs, tels que des tremblements, une rigidité musculaire et une lenteur des mouvements, mais elle est également associée à des symptômes non moteurs, notamment des altérations vocales. Ces altérations, souvent présentes dès les premiers stades de la maladie, constituent une piste prometteuse pour le diagnostic précoce. Avec l'émergence de l'intelligence artificielle, l'analyse des signaux vocaux à l'aide de techniques d'apprentissage automatique ouvre de nouvelles perspectives dans l'amélioration des outils d'aide au diagnostic médical.

Ce projet s'inscrit dans cette dynamique et vise à explorer le potentiel des données vocales pour détecter la présence de la maladie de Parkinson. Le jeu de données utilisé, issu du **UCI Machine Learning Repository**, contient des enregistrements vocaux d'individus atteints de Parkinson et d'individus sains. Ces données comprennent diverses caractéristiques acoustiques, telles que la fréquence fondamentale, le jitter, le shimmer et d'autres mesures spécifiques, permettant de caractériser les perturbations vocales typiques de la maladie. Ces caractéristiques servent de base à la construction de modèles d'apprentissage supervisé capables de distinguer les patients des individus sains.

Pour répondre à cet objectif, nous mettons en œuvre trois algorithmes d'apprentissage supervisé : **Logistic Regression**, **Random Forest** et **LightGBM**. Ces modèles seront d'abord évalués sur l'ensemble complet des caractéristiques disponibles dans le dataset, sans appliquer de sélection préalable. Ensuite, nous examinerons l'effet de la sélection des caractéristiques en combinant les résultats de trois techniques distinctes pour extraire les caractéristiques les plus pertinentes. Cette démarche permettra de comparer les performances des modèles dans les deux configurations et de mieux comprendre l'impact de la réduction du bruit sur la qualité des prédictions.

En s'appuyant sur une méthodologie rigoureuse et des métriques de performance bien définies, ce projet vise à fournir une analyse comparative approfondie des approches étudiées. Les

résultats attendus pourraient contribuer à renforcer l'utilisation des technologies d'apprentissage automatique dans le domaine médical, en particulier pour le diagnostic assisté de la maladie de Parkinson.

Chapitre 1 : Contexte général du projet

I. Présentation de la Maladie de Parkinson :

1. Définition et Impact :

La maladie de Parkinson est une affection neurologique progressive causée par la dégénérescence des neurones produisant la dopamine dans le cerveau. Elle se manifeste principalement par des troubles moteurs tels que les tremblements, la rigidité musculaire et la lenteur des mouvements. Ces symptômes affectent considérablement la qualité de vie des patients en réduisant leur autonomie dans les activités quotidiennes.

L'impact de la maladie va au-delà des symptômes physiques, avec des répercussions sur le bien-être mental et social des patients. Le diagnostic précoce est essentiel pour ralentir la progression de la maladie, mais il demeure difficile à poser, notamment aux stades précoces où les symptômes sont moins évidents. Cela rend l'identification précoce de la maladie un défi majeur pour les professionnels de santé.

2. Symptômes et Diagnostic Traditionnel :

Les symptômes de la maladie de Parkinson sont principalement moteurs, bien qu'ils puissent également affecter d'autres fonctions. Les signes moteurs classiques incluent :

- **Tremblements** : Des tremblements au repos sont souvent l'un des premiers signes. Ils affectent généralement les mains, mais peuvent également toucher les pieds ou la tête.
- **Rigidité musculaire** : Les muscles deviennent rigides, ce qui peut entraîner des douleurs et une limitation des mouvements.
- **Bradykinésie** : La lenteur des mouvements est un symptôme majeur, rendant les gestes quotidiens plus difficiles et moins fluides.
- **Instabilité posturale** : Les patients peuvent avoir des difficultés à maintenir leur équilibre, ce qui augmente le risque de chutes.

En plus de ces symptômes moteurs, des symptômes non moteurs peuvent apparaître, tels que des troubles du sommeil, de la dépression, des problèmes cognitifs et des troubles de l'humeur.

Le **diagnostic traditionnel** de la maladie de Parkinson repose sur l'observation clinique des symptômes et un examen neurologique. Les médecins se basent généralement sur

les critères de diagnostic définis par le **criterium de Parkinson** (notamment la présence de tremblements et de bradykinésie). Des examens complémentaires, comme l'**imagerie cérébrale** (IRM, TEP), peuvent être réalisés pour écarter d'autres pathologies, mais il n'existe pas encore de test de diagnostic définitif à ce jour, surtout dans les stades précoces de la maladie.

Le diagnostic précoce reste complexe, car les signes cliniques peuvent être confondus avec d'autres troubles neurologiques, ce qui rend essentiel de développer des outils de diagnostic plus précis, comme l'analyse des données vocales.

II. Importance des Données Vocales dans le Diagnostic :

1. Rôle des Altérations Vocales :

La maladie de Parkinson affecte non seulement les mouvements physiques, mais elle perturbe également les fonctions vocales des patients. Les altérations vocales sont souvent présentes dès les premiers stades de la maladie et peuvent devenir plus prononcées à mesure que la maladie progresse. Ces changements sont dus à la réduction de la dopamine dans le cerveau, ce qui affecte le contrôle moteur nécessaire pour produire des sons de manière fluide et naturelle.

Les principales altérations vocales observées chez les patients atteints de Parkinson comprennent :

- **Diminution du volume vocal** : Les patients ont souvent une voix plus faible, difficile à entendre, en raison de la réduction de la force musculaire liée à la maladie.
- **Monotonie** : La variabilité du ton de la voix est réduite, rendant la parole moins expressive et monotone.
- **Tremblements vocaux** : Des tremblements peuvent affecter la voix, créant des irrégularités dans la fréquence et l'intensité du son.
- **Dysphonie** : Les patients peuvent éprouver des difficultés à produire des sons clairs et distincts, ce qui conduit à une voix enrouée ou rauque.

Ces altérations vocales peuvent fournir des indices précieux pour le diagnostic de la maladie de Parkinson, notamment lors de son stade précoce où les symptômes moteurs sont moins évidents. L'analyse acoustique des signaux vocaux permet de détecter ces

changements de manière non invasive, offrant ainsi une méthode de diagnostic complémentaire aux méthodes traditionnelles.

2. Avantages des Données Vocales :

L'utilisation des données vocales dans le diagnostic de la maladie de Parkinson présente plusieurs avantages qui en font une méthode particulièrement intéressante. Tout d'abord, l'analyse vocale permet de détecter la maladie à un stade précoce, souvent avant l'apparition des symptômes moteurs majeurs, tels que les tremblements ou les rigidités musculaires. En effet, des changements dans la voix peuvent être observés bien avant que d'autres signes de la maladie ne se manifestent, ce qui ouvre la voie à un diagnostic plus précoce et, par conséquent, à une intervention thérapeutique plus rapide.

Un autre avantage majeur des données vocales est leur **non-invasivité**. Contrairement à d'autres techniques de diagnostic qui nécessitent des tests physiques ou des procédures médicales invasives, l'enregistrement vocal peut être réalisé de manière simple et non intrusive. Cela permet non seulement de rendre le processus de diagnostic plus confortable pour le patient, mais aussi de faciliter la collecte de données à grande échelle, par exemple, en utilisant des smartphones ou des dispositifs portables.

La **facilité de collecte des données** est également un atout important. Les enregistrements vocaux sont rapides et simples à réaliser, ce qui les rend particulièrement adaptés aux **examens de dépistage** dans des environnements cliniques ou à distance. L'accessibilité de cette technologie, grâce aux outils de communication moderne, rend la mise en place de tests à grande échelle à la fois rapide et économique.

En outre, l'analyse des données vocales fournit des **mesures objectives**, ce qui permet d'éviter les biais subjectifs associés aux évaluations cliniques traditionnelles. Les caractéristiques vocales, comme le jitter, le shimmer ou les coefficients MFCC, sont des indices mesurables et quantifiables de l'état de santé d'un patient. Ces mesures offrent une évaluation plus précise et reproductible de la fonction vocale, renforçant ainsi la fiabilité du diagnostic.

Enfin, la possibilité de suivre **l'évolution de la maladie** au fil du temps est un autre avantage notable. En comparant les enregistrements vocaux d'un patient à différentes étapes de la maladie, il est possible d'observer des changements dans la voix et

d’évaluer l’efficacité des traitements. Cette capacité à suivre l’état de santé d’un patient de manière continue constitue un outil précieux pour les médecins dans la gestion de la maladie.

En somme, l’utilisation des données vocales pour détecter et suivre la progression de la maladie de Parkinson présente de nombreux avantages. C’est une méthode simple, non invasive et fiable, qui permet d'améliorer les diagnostics précoces tout en offrant une solution complémentaire aux techniques diagnostiques existantes.

III. Introduction aux Techniques d’Apprentissage Supervisé :

1. Définition et Intérêts :

L'apprentissage supervisé est une méthode d'apprentissage automatique dans laquelle un modèle est formé sur un ensemble de données étiquetées, c'est-à-dire un jeu de données pour lequel les résultats (ou cibles) sont déjà connus. L'objectif principal de l'apprentissage supervisé est d'apprendre une fonction qui peut prédire la sortie correcte pour de nouvelles entrées, en se basant sur les exemples fournis durant l'entraînement. Ce type d'approche est particulièrement adapté aux problèmes de classification et de régression, où l'on cherche à prédire une classe ou une valeur numérique en fonction des caractéristiques d'entrée.

Les techniques d'apprentissage supervisé se distinguent par leur capacité à exploiter les relations entre les variables d'entrée et les sorties. Cela permet aux modèles de généraliser à partir des données d'apprentissage pour effectuer des prédictions sur de nouvelles données, non vues auparavant. Parmi les méthodes les plus courantes d'apprentissage supervisé figurent les **arbres de décision**, les **machines à vecteurs de support** (SVM), les **régressions logistiques**, ainsi que des algorithmes plus complexes comme les **forêts aléatoires** et **XGBoost**.

L'un des principaux **intérêts** de l'apprentissage supervisé est qu'il permet de résoudre une large gamme de problèmes, allant de la classification des images à la prévision des tendances économiques. Dans le contexte de la détection de la maladie de Parkinson, par exemple, l'apprentissage supervisé permet d'utiliser des caractéristiques extraites des enregistrements vocaux pour prédire la présence ou l'absence de la maladie. Cette capacité à prédire des résultats à partir de données historiques est un atout majeur dans

des domaines où des décisions doivent être prises rapidement et de manière fiable, comme en médecine, où le diagnostic précoce peut avoir un impact significatif sur le traitement et la gestion des patients.

En outre, les **modèles d'apprentissage supervisé** bénéficient d'une large acceptation en raison de leur **transparence** et de leur capacité à fournir des résultats interprétables. Les utilisateurs peuvent comprendre et expliquer comment une prédiction a été faite, ce qui est crucial dans des applications sensibles, comme le diagnostic médical. De plus, les algorithmes supervisés sont souvent bien adaptés aux données structurées, comme les ensembles de données tabulaires, qui sont courants dans de nombreux domaines industriels et scientifiques.

En résumé, les techniques d'apprentissage supervisé représentent une approche puissante et flexible pour aborder divers problèmes de prédiction. Leur capacité à apprendre à partir de données étiquetées, couplée à leur capacité à fournir des résultats fiables et interprétables, les rend particulièrement adaptées aux applications pratiques, y compris celles visant à diagnostiquer des maladies comme la maladie de Parkinson à partir de données vocales.

2. Application au Diagnostic de la Maladie de Parkinson :

L’utilisation des données vocales pour le diagnostic de la maladie de Parkinson est un domaine de recherche qui a gagné en popularité ces dernières années. Plusieurs études ont démontré que des altérations dans les caractéristiques vocales peuvent être des indicateurs précoces de cette maladie neurodégénérative. Les symptômes moteurs de la maladie, tels que la rigidité musculaire et les tremblements, se manifestent souvent plus tard, tandis que des anomalies dans la voix peuvent être détectées bien avant. En conséquence, l’analyse des caractéristiques acoustiques de la voix, telles que le **jitter**, le **shimmer**, ou les **coefficients cepstraux en fréquence Mel** (MFCC), a suscité un intérêt croissant pour le diagnostic précoce.

Les recherches existantes sur le diagnostic vocal de la maladie de Parkinson se concentrent sur l'extraction de ces caractéristiques à partir de signaux vocaux et sur l'utilisation de diverses techniques d’apprentissage supervisé pour prédire la présence ou l'absence de la maladie. Les **machines à vecteurs de support** (SVM), les **arbres de décision**, et les **forêts aléatoires** sont parmi les modèles les plus couramment utilisés,

tandis que des approches plus avancées comme les **réseaux de neurones profonds** ont également été explorées pour de meilleures performances. Une partie importante de ces travaux se concentre sur l'**optimisation des caractéristiques** via des méthodes de **sélection** ou de **réduction**, afin de réduire la dimensionnalité des données et d'améliorer la précision du modèle. Les approches comme la **sélection de caractéristiques** et la **réduction de la dimensionnalité** sont cruciales pour éviter le surapprentissage et améliorer la généralisation du modèle.

Dans ce projet, l'application des techniques d'apprentissage supervisé repose sur l'utilisation de trois algorithmes classiques : **Régression Logistique**, **Random Forest** et **LightGBM**, en les comparant dans deux scénarios : l'un sans sélection des caractéristiques et l'autre après avoir appliqué trois techniques de sélection des caractéristiques (filtrage, emballage, incrustation). L'objectif est de déterminer quelle combinaison de méthodes et de modèles offre les meilleurs résultats pour la classification de la maladie de Parkinson à partir des données vocales.

La contribution de ce projet réside dans l'amélioration de la performance des modèles existants en combinant différentes techniques de sélection de caractéristiques pour créer un ensemble de données optimisé, tout en comparant les performances des modèles avec et sans sélection des caractéristiques. Cela pourrait fournir des résultats plus fiables et plus précis, aidant ainsi à mieux diagnostiquer la maladie de Parkinson à un stade précoce, contribuant ainsi à la recherche médicale et à l'amélioration des outils de diagnostic.

Chapitre 2 : Mise en Œuvre du Projet

I. Préparation des Données :

Dans cette section, nous présentons les étapes de préparation des données utilisées pour l'entraînement des modèles de classification.

La première étape a consisté à charger le dataset à partir du fichier CSV fourni. Une fois les données chargées, nous avons exploré la structure du dataset pour comprendre sa composition, notamment le nombre d'instances et de caractéristiques disponibles. Ensuite, nous avons observé que la première ligne du fichier CSV servait d'en-tête, mais celle-ci était incluse comme une ligne de données. Nous avons donc extrait cette ligne et l'avons utilisée comme noms de colonnes pour le DataFrame, afin d'assurer une structure cohérente des données.

Une fois cette étape terminée, nous avons supprimé la colonne 'id'. Cette colonne ne contenait pas d'informations utiles pour le modèle, car elle identifiait uniquement les individus sans apporter de valeur prédictive, et a donc été retirée du dataset.
En ce qui concerne la qualité des données, nous avons vérifié la présence de doublons dans le dataset. Après cette vérification, nous avons supprimé les lignes en double afin de garantir que les données utilisées pour l'entraînement des modèles soient uniques et pertinentes, ce qui permet d'éviter un biais ou un surapprentissage.
Par la suite, nous avons réalisé une visualisation des données pour mieux comprendre leur distribution et leurs caractéristiques. Cela nous a permis d'identifier un déséquilibre important dans la classe cible. Afin de remédier à ce problème, nous avons appliqué une méthode d'équilibrage des données pour obtenir une distribution plus homogène entre les classes, ce qui contribue à améliorer les performances des modèles.
Enfin, nous avons procédé à la conversion des valeurs du dataset en types numériques, en particulier en types flottants, car les modèles d'apprentissage supervisé nécessitent des données continues pour effectuer des calculs et des prédictions. Cette conversion a permis de préparer les données pour les étapes d'entraînement et de validation des modèles.
Ainsi, les étapes de préparation des données ont permis de nettoyer, structurer et équilibrer correctement le dataset, en enlevant les éléments inutiles et en assurant une qualité adéquate pour les modèles d'apprentissage automatique.

II. Mise en Œuvre des Modèles :

Cette section présente la mise en œuvre des modèles d'apprentissage supervisé pour la classification de la maladie de Parkinson à l'aide des données vocales. Trois modèles ont été implémentés à partir de zéro : la régression logistique, Random Forest et LightGBM. Chaque modèle a été testé dans deux configurations : une première sans sélection de caractéristiques et une seconde avec sélection de caractéristiques appliquée à l'aide de trois techniques différentes.

1. Modèles sans Sélection de Caractéristiques :

Dans un premier temps, les trois modèles ont été implémentés et évalués sans appliquer de sélection de caractéristiques. L'ensemble complet des caractéristiques acoustiques a été utilisé pour entraîner les modèles, ce qui a permis d'évaluer leur capacité à traiter l'intégralité des données disponibles.

- **Régression Logistique** : La régression logistique a été utilisée comme modèle de base pour la classification binaire. Elle a été testée sur l'ensemble complet des caractéristiques sans aucune réduction de dimension. La performance du modèle a été évaluée en utilisant des métriques standard telles que la précision, le rappel et la F-mesure.
- **Random Forest** : Le modèle Random Forest, basé sur l'agrégation de plusieurs arbres de décision, a été appliqué sur les données sans réduction de dimensionnalité. Cette approche permet de mieux gérer les caractéristiques redondantes et bruitées. La performance du modèle a été comparée à celle de la régression logistique afin d'analyser sa robustesse face à des données complexes.
- **LightGBM** : Le modèle LightGBM, qui utilise un algorithme de boosting, a également été testé sans sélection des caractéristiques. Ce modèle est connu pour sa rapidité et son efficacité, en particulier sur des ensembles de données volumineux. Il a été comparé aux deux autres modèles en termes de précision et de temps d'entraînement.

2. Modèles avec Sélection de Caractéristiques :

Après avoir testé les modèles sans sélection de caractéristiques, une étape de sélection a été appliquée pour identifier les variables les plus pertinentes et réduire le bruit dans les données. Trois méthodes de sélection de caractéristiques ont été utilisées

: la méthode Filter (basée sur la corrélation), la méthode Embedded (LassoCV), et la méthode Wrapper (Backward Elimination).

- **Filter Method (Corrélation)** : La première technique de sélection a consisté à utiliser la méthode de filtrage basée sur la corrélation. Les caractéristiques présentant des corrélations élevées avec la variable cible ont été sélectionnées, tandis que celles présentant des corrélations faibles ont été éliminées. Cette méthode permet de réduire la dimensionnalité tout en conservant les informations pertinentes pour la classification.
- **Embedded Method (LassoCV)** : La méthode LassoCV a été utilisée comme une méthode intégrée de sélection de caractéristiques. Elle impose une pénalité sur les coefficients de régression pour éliminer les variables les moins significatives. En ajustant cette pénalité, LassoCV permet de sélectionner les caractéristiques les plus importantes pour le modèle de régression logistique. Cette approche a été appliquée avant d'entraîner à nouveau le modèle de régression logistique.
- **Wrapper Method (Backward Elimination)** : La méthode de sélection Wrapper a été appliquée via la technique de Backward Elimination. Cette méthode consiste à éliminer itérativement les caractéristiques les moins significatives à partir d'un modèle préexistant. À chaque étape, le modèle est réajusté et l'évaluation des performances permet de déterminer si une caractéristique doit être conservée ou supprimée.

Après avoir appliqué la sélection des caractéristiques, les modèles ont été réentraînés et évalués sur l'ensemble réduit de variables. Les performances ont été comparées à celles obtenues sans sélection de caractéristiques pour observer l'impact de cette étape sur la qualité des prédictions.

Les résultats ont montré si la réduction de la dimensionnalité à l'aide de ces méthodes pouvait améliorer la performance des modèles, en particulier en termes de réduction du surapprentissage et d'augmentation de la précision des prédictions. Cela a permis de déterminer l'importance de la sélection de caractéristiques dans le contexte de la détection de la maladie de Parkinson à partir des données vocales.

III. Évaluation et Comparaison des Modèles

L'évaluation des performances des modèles s'est basée sur plusieurs métriques clés, permettant de mesurer leur efficacité sous différents aspects. La métrique principale utilisée est l'accuracy, qui mesure le pourcentage de prédictions correctes réalisées par le modèle. Cependant, afin d'obtenir une évaluation plus complète, nous avons également pris en compte d'autres indicateurs de performance tels que la précision (precision), le rappel (recall) et le score F1 (f1-score). Ces métriques permettent d'évaluer la capacité des modèles à identifier correctement les classes positives et négatives, tout en prenant en compte les déséquilibres possibles dans les données.

Pour évaluer l'impact de la sélection des caractéristiques, nous avons appliqué la technique de sélection des caractéristiques avec chaque modèle (Régression Logistique, Random Forest, LightGBM). Une fois la sélection effectuée, nous avons testé le dataset filtré obtenu par chaque modèle avec les trois modèles (LR, RF, LightGBM). En d'autres termes, après avoir réalisé la sélection des caractéristiques avec un modèle, nous avons entraîné et évalué le modèle de sélection avec chacun des trois modèles de classification pour observer l'impact sur les résultats. Ce processus a été répété pour chaque combinaison possible, soit un total de 9 combinaisons (3 méthodes de sélection × 3 modèles d'entraînement). Cette approche nous a permis d'analyser en profondeur l'impact de la sélection des caractéristiques sur les performances des modèles et de comparer les résultats sous différentes configurations.

Cette section présente les résultats obtenus avant et après l'application des techniques de sélection des caractéristiques, ainsi que leur interprétation détaillée en fonction de ces différentes métriques.

1. Sans Sélection des Caractéristiques

Avant d'appliquer toute technique de sélection des caractéristiques, les trois modèles – Régression Logistique, Random Forest et LightGBM – ont été entraînés sur l'ensemble complet des caractéristiques acoustiques. Ces caractéristiques comprennent des mesures telles que le jitter, le shimmer et la fréquence fondamentale.

Les résultats des performances sur l'ensemble d'entraînement sont présentés ci-dessous. Les métriques utilisées pour l'évaluation incluent **accuracy**, **précision**, **rappel** et **f1-score**.

Modèle d'Entraînement	Accuracy	Précision	Rappel	F1-score
Régression Logistique	93%	93%	94%	93%
Random Forest	89%	89%	88%	89%
LightGBM	97%	98%	95%	97%

Tableau 1 : Performances sur l'Ensemble d'Entraînement « sans sélection »

Après l'entraînement des modèles, ces derniers ont été évalués sur l'ensemble de test, toujours sans sélection des caractéristiques.

Voici les résultats obtenus sur l'ensemble de test, pour les mêmes métriques : **accuracy**, **précision**, **rappel** et **f1-score** :

Modèle d'Entraînement	Accuracy	Précision	Rappel	F1-score
Régression Logistique	90%	90%	89%	89%
Random Forest	80%	80%	78%	79%
LightGBM	86%	83%	88%	85%

Tableau 2 : Performances sur l'Ensemble de Test « sans sélection »

2. Avec Sélection des Caractéristiques :

Une fois la sélection des caractéristiques effectuée, chaque modèle a été entraîné sur l'ensemble des données réduites, en combinant les différentes techniques de sélection des caractéristiques (Filter Method, Embedded Method, Wrapper Method) avec les modèles d'entraînement (Régression Logistique, Random Forest, et LightGBM).

Les résultats de ces combinaisons sont présentés ci-dessous :

Feature Selection	Model	Accuracy	Précision	Rappel	F1-score
Logistic Regression	Logistic Regression	92%	91%	93%	92%
	Random Forest	89%	89%	89%	89%
	LightGBM	85%	82%	89%	85%

Random Forest	Logistic Regression	92%	91%	94%	92%
	Random Forest	89%	90%	88%	89%
	LightGBM	86%	82%	90%	86%
LightGBM	Logistic Regression	93%	93%	94%	93%
	Random Forest	88%	89%	87%	88%
	LightGBM	86%	83%	88%	85%

Tableau 3 : Performances sur l'Ensemble d'Entraînement « avec sélection »

Feature Selection	Model	Accuracy	Précision	Rappel	F1-score
Logistic Regression	Logistic Regression	86%	83%	88%	85%
	Random Forest	81%	79%	82%	81%
	LightGBM	96%	96%	96%	96%
Random Forest	Logistic Regression	85%	82%	89%	85%
	Random Forest	77%	76%	76%	76%
	LightGBM	96%	98%	95%	96%
LightGBM	Logistic Regression	86%	82%	90%	86%
	Random Forest	80%	80%	78%	79%
	LightGBM	97%	98%	95%	97%

Tableau 4 : Performances sur l'Ensemble de Test « avec sélection »

3. Discussion Générale

Performances globales des modèles sans sélection des caractéristiques

Avant l'application de techniques de sélection des caractéristiques, les modèles montrent des performances variables sur les ensembles d'entraînement et de test. Globalement, **LightGBM**

domine les autres algorithmes en termes de précision et de score F1, atteignant une précision de 97 % sur l'ensemble d'entraînement et de 86 % sur l'ensemble de test. Ces résultats reflètent sa capacité à capturer des relations complexes dans les données. Cependant, la différence notable entre les performances d'entraînement et de test (97 % vs 86 %) suggère une légère sur-apprentissage, typique des modèles de boosting s'ils ne sont pas soigneusement régularisés.

La **régression logistique**, bien que plus simple, montre une performance équilibrée et robuste avec une précision de 93 % sur l'ensemble d'entraînement et de 90 % sur l'ensemble de test. Cette stabilité entre les ensembles d'entraînement et de test indique une généralisation satisfaisante, rendant ce modèle adapté à des cas où l'interprétabilité et la robustesse sont prioritaires.

Le **Random Forest** affiche des performances globalement inférieures avec une précision de 89 % sur l'ensemble d'entraînement et une chute significative à 80 % sur l'ensemble de test. Cette baisse plus marquée pourrait indiquer que ce modèle est moins performant pour généraliser, probablement en raison de paramètres sous-optimaux ou de la structure des données.

Impact de la sélection des caractéristiques sur les performances

Avec l'application de différentes techniques de sélection des caractéristiques, des changements notables sont observés dans les performances des modèles. La sélection des caractéristiques réduit le nombre de variables utilisées, ce qui peut améliorer la robustesse du modèle tout en réduisant le risque de sur-apprentissage. Cependant, ces avantages dépendent de la technique de sélection employée et du modèle utilisé.

Performances de LightGBM avec sélection

LightGBM continue de dominer dans la plupart des cas, atteignant une précision impressionnante de 97 % sur l'ensemble de test lorsque la sélection des caractéristiques est basée sur son propre algorithme. Ce résultat suggère que la technique de sélection intégrée à LightGBM est particulièrement efficace pour identifier les caractéristiques pertinentes tout en maintenant des performances élevées. Cependant, lorsqu'une autre méthode de sélection est utilisée (comme la sélection basée sur la régression logistique), la précision du modèle LightGBM baisse légèrement (96 % sur l'ensemble de test). Cette variabilité met en évidence que la qualité des caractéristiques sélectionnées dépend fortement de la méthode choisie.

Performances de la régression logistique avec sélection

La **régression logistique** montre une stabilité et une légère amélioration avec la sélection des caractéristiques. Par exemple, après sélection basée sur LightGBM, la précision sur l'ensemble de test atteint 86 %. Cela peut être attribué au fait que la sélection réduit le bruit des données, ce qui est bénéfique pour un modèle linéaire tel que la régression logistique. Toutefois, les performances légèrement inférieures par rapport à LightGBM indiquent que ce modèle pourrait ne pas être capable de capturer certaines interactions non linéaires dans les données.

Performances de Random Forest avec sélection

Le **Random Forest** est le modèle qui semble bénéficier le moins de la sélection des caractéristiques. Par exemple, après une sélection basée sur la régression logistique, sa précision sur l'ensemble de test reste limitée à 80 %. Cette faible amélioration suggère que ce modèle, bien qu'efficace pour gérer des données complexes, n'est pas toujours le mieux adapté à des scénarios où les variables sont déjà réduites et optimisées.

Comparaison entre modèles et impact sur la généralisation

En comparant les performances avant et après sélection, il est clair que la sélection des caractéristiques contribue à une meilleure généralisation pour certains modèles, tout en ayant un impact moindre pour d'autres. **LightGBM**, par exemple, maintient des performances élevées dans les deux scénarios, prouvant sa robustesse et sa capacité à s'adapter à des ensembles de données réduits. En revanche, **Random Forest** montre des difficultés à tirer parti de la sélection des caractéristiques, ce qui peut être dû à une sensibilité accrue au sur-ajustement lorsque les données deviennent trop simplifiées.

D'un autre côté, la **régression logistique** semble bénéficier de la sélection des caractéristiques en termes de robustesse et de généralisation, tout en restant compétitive face à des modèles plus complexes comme LightGBM. Cela souligne que des modèles plus simples peuvent souvent bien fonctionner lorsqu'ils sont combinés à des techniques de réduction de dimension efficaces.

Observation sur l'équilibre entre précision et rappel

Un point intéressant à noter est l'équilibre entre la précision et le rappel pour les différents modèles. Par exemple, LightGBM conserve un équilibre presque parfait avec une précision et un rappel de 97 % sur l'ensemble de test lorsqu'il est combiné à sa propre méthode de sélection. En revanche, la régression logistique montre une légère disparité, avec une précision de 86 % mais un rappel légèrement supérieur à 90 % après sélection via LightGBM. Cette différence indique que la régression logistique privilégie légèrement la détection des cas positifs au

détriment des faux positifs, ce qui peut être bénéfique dans des contextes où un haut rappel est crucial.

Conclusion sur les résultats obtenus

En résumé, les performances des modèles varient considérablement en fonction de la méthode de sélection des caractéristiques et du modèle employé. LightGBM se distingue comme le modèle le plus performant dans la plupart des cas, grâce à sa capacité à s'adapter aux données et à capturer des relations complexes. La régression logistique, bien qu'étant un modèle linéaire, offre une alternative solide, surtout lorsqu'elle est combinée à des techniques de sélection efficaces. Le Random Forest, quant à lui, pourrait nécessiter un ajustement supplémentaire des hyperparamètres pour améliorer sa capacité de généralisation.

Cette analyse met en évidence l'importance de choisir judicieusement à la fois le modèle et la méthode de sélection des caractéristiques, en tenant compte des spécificités du problème à résoudre.

Conclusion Générale

Ce projet a exploré l'application des techniques d'apprentissage supervisé pour la détection de la maladie de Parkinson à partir de données vocales, avec une attention particulière portée sur l'impact de la sélection des caractéristiques sur les performances des modèles. Trois algorithmes, à savoir la Régression Logistique, Random Forest, et LightGBM, ont été implémentés et évalués dans deux configurations : avant et après l'application de techniques distinctes de sélection de caractéristiques.

Les résultats obtenus illustrent clairement que la sélection des caractéristiques joue un rôle crucial dans l'amélioration des performances des modèles. Sans sélection, les modèles opéraient sur un ensemble complet de caractéristiques (271 après PCA), incluant des variables potentiellement redondantes ou non pertinentes. Cela s'est traduit par des performances parfois limitées, en particulier sur l'ensemble de test, où les modèles ont montré des signes de surapprentissage ou de mauvaise généralisation. Par exemple, LightGBM a atteint une précision impressionnante de 97 % sur l'ensemble d'entraînement, mais cette performance a chuté à 86 % sur l'ensemble de test, indiquant une difficulté à généraliser. La Régression Logistique a montré une meilleure stabilité entre les ensembles d'entraînement et de test, avec une précision respective de 93 % et 90 %, ce qui reflète sa robustesse malgré une architecture plus simple. Random Forest, en revanche, a obtenu des performances plus modestes, avec une précision de 89 % sur l'ensemble d'entraînement et de seulement 80 % sur l'ensemble de test.

Après l'application des techniques de sélection des caractéristiques, les performances des modèles se sont globalement améliorées, confirmant que la réduction de la dimensionnalité permet de mieux isoler les caractéristiques essentielles tout en diminuant l'impact du bruit. LightGBM, en particulier, s'est affirmé comme le modèle le plus performant dans la majorité des configurations. Lorsqu'il a été entraîné sur des caractéristiques sélectionnées via sa propre méthode, il a atteint une précision remarquable de 97 % sur l'ensemble de test, avec un équilibre quasi parfait entre précision et rappel (98 % et 95 %, respectivement). Ce résultat met en évidence la capacité de LightGBM à exploiter efficacement les variables informatives et à gérer les interactions complexes entre elles.

La Régression Logistique a également bénéficié de la sélection des caractéristiques, montrant une amélioration significative de ses performances. Par exemple, avec des caractéristiques sélectionnées via backward elimination, elle a atteint une précision de 86 % sur l'ensemble de test, tout en maintenant un rappel élevé de 90 %. Ces résultats démontrent que, bien que linéaire, ce modèle peut être considérablement renforcé par un prétraitement adéquat des données.

Random Forest, en revanche, a montré une sensibilité variable aux techniques de sélection utilisées. Malgré des résultats acceptables dans certaines configurations (par exemple, une précision de 80 % sur l'ensemble de test après sélection via LightGBM), il n'a pas réussi à égaler les performances des autres modèles. Cela pourrait indiquer une moindre capacité à tirer parti de la réduction de la dimensionnalité ou une dépendance plus forte à un réglage optimal des hyperparamètres.

Ces résultats mettent en évidence deux enseignements majeurs. D'une part, la sélection des caractéristiques est une étape incontournable pour maximiser les performances des modèles, réduire le surapprentissage et simplifier leur architecture. Cette étape est particulièrement bénéfique pour des ensembles de données riches en variables, comme c'est le cas ici. D'autre part, la combinaison optimale entre la méthode de sélection et le modèle d'entraînement est décisive pour atteindre des performances élevées. LightGBM s'est révélé particulièrement efficace lorsqu'il est associé à sa propre méthode de sélection, tandis que la Régression Logistique a montré une amélioration notable avec des méthodes linéaires comme backward elimination.

En conclusion, cette étude démontre l'importance d'intégrer des étapes rigoureuses de traitement des données dans les pipelines de machine learning. Les résultats obtenus montrent que l'utilisation méthodique de techniques de sélection des caractéristiques, combinée à des algorithmes performants comme LightGBM, peut considérablement améliorer la qualité des prédictions pour des applications complexes telles que le diagnostic de la maladie de Parkinson à partir de données vocales. Ces conclusions ouvrent également des perspectives intéressantes pour d'autres études, notamment dans l'exploration de combinaisons de modèles ou dans l'application de ces techniques à d'autres pathologies ou domaines médicaux.

Bibliographie

1. UCI Machine Learning Repository. (n.d.). *Parkinson's Disease Classification Dataset*. Récupéré de https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification
2. Microsoft. (n.d.). *LightGBM Documentation*. Récupéré de https://lightgbm.readthedocs.io/en/stable/
3. StatQuest. (2020, May 12). *Gradient Boosting and XGBoost in Python* [Vidéo]. YouTube. Récupéré de https://youtu.be/3CC4N4z3GJc?si=e779rAr0_Rx7ZKSe

.