Projet de Méthodologie de la Recherche en Informatique
Classification

Yeleen Canonge et Camille Fief
Master 1 Langue et Informatique, Sorbonne Université

6 janvier 2025

Table des matières

1 Introduction

L'analyse des données occupe une place centrale dans de nombreux domaines scientifiques, économiques et sociaux. La prolifération des données, qu'elles soient structurées, semi-structurées ou non structurées, a conduit à une évolution des méthodes de stockage et d'analyse.

Dans ce contexte, ce projet se concentre sur l'analyse de trois ensembles de données bien connus issus de l'UCI Machine Learning Repository : Wine Quality, Adult, et SMS Spam Collection. Chacun de ces jeux de données présente des caractéristiques distinctes :

- **Wine Quality** : Des données chimiques structurées pour prédire la qualité des vins rouges et blancs. Ces données sont représentatives des défis liés à la modélisation de relations complexes entre des variables numériques continues.
- **Adult** : Un ensemble socio-économique structuré qui vise à prédire les revenus d'un individu en fonction de ses caractéristiques démographiques..
- **SMS Spam Collection** : Des données textuelles non structurées utilisées pour la classification en spam et ham, un cas typique dans le domaine du traitement automatique du langage naturel (NLP).

2 Problématique et motivation

L'utilisation de ces jeux de données permet de répondre à plusieurs questions fondamentales :

- Comment exploiter efficacement les relations implicites dans les données ? La détermination du classificateur le plus efficace peut permettre de prédire ces relations implicites dans les données avec un test et un apprentissage.
- Quels modèles prédictifs sont les plus adaptés à ces types de données ? Chaque jeu de données impose des exigences spécifiques. Les données textuelles nécessitent des approches NLP avancées, tandis que les données structurées peuvent être analysées via des modèles de régression ou de classification.

3 Objectifs du projet

- Décrire et analyser les données pour identifier leurs caractéristiques, leurs limitations et leur utilisation.
- Explorer l'état de l'art pour comprendre les approches existantes et identifier les lacunes.
- Proposer une classification cohérente et déterminer le classificateur le plus adapté.
- Optimiser le traitement des données par les classificateurs.

4 Description des données

4.1 Wine Quality

L'ensemble de données Wine Quality regroupe des informations sur les caractéristiques chimiques et organoleptiques de vins portugais, rouges et blancs, collectées pour prédire leur qualité sur une échelle allant de 0 à 10. Il comprend un total de 6 497 échantillons, dont 4 898 de vins blancs et 1 599 de vins rouges. Chaque observation est décrite par 12 attributs physico-chimiques, tels que l'acidité fixe, l'acidité volatile, le pH, la teneur en sucre résiduel, la densité et les chlorures. Ces caractéristiques jouent un rôle crucial dans la perception sensorielle des vins et influencent directement leur évaluation par des œnologues.

La variable cible représente la douzième colonne qui est une note de qualité attribuée par des experts, constituant un problème de régression ou de classification ordinale. Les données sont complètes, sans valeurs manquantes, et nécessitent une étape de normalisation, étant donné la disparité des échelles de certains attributs (par exemple, la densité a des valeurs plus élevées que le pH). Cette étape est essentielle pour éviter que certaines variables ne dominent l'apprentissage des modèles. Ce dataset est fréquemment utilisé pour évaluer des algorithmes capables de capturer des relations complexes entre des variables continues dans un cadre prédictif.

4.2 Adult

L'ensemble de données Adult, extrait du recensement américain de 1994, est structuré autour d'attributs démographiques et économiques, avec pour objectif principal de prédire si le revenu annuel d'un individu dépasse ou non 50 000 dollars. Il contient un total de 48 842 enregistrements, divisés en un ensemble d'entraînement d'environ 32 000 échantillons et un ensemble de test de 16 000. Les observations incluent des caractéristiques telles que l'âge, le sexe, le niveau d'éducation, la profession, le statut marital et le nombre d'heures travaillées par semaine. Ces variables reflètent des aspects sociaux et économiques susceptibles d'influencer les revenus. La variable cible est binaire, indiquant si un individu appartient à la catégorie « 50K » ou « >50K ». Ce dataset est couramment utilisé pour des tâches de classification supervisée, en raison de sa distribution de classes légèrement déséquilibrée et de la nature mixte de ses données (numériques et catégorielles). Le prétraitement des données inclut des étapes de codage des variables catégoriques et de normalisation des attributs numériques. La richesse et la diversité des caractéristiques en font une base de référence pour évaluer la robustesse des modèles sur des données démographiques complexes.

4.3 SMS Spam Collection

Le dataset SMS Spam Collection se concentre sur la détection des spams dans des messages textuels. Il contient 5 574 observations, parmi lesquelles 4 827 sont étiquetées comme « ham » (messages légitimes) et 747 comme « spam ». Les données sont constituées exclusivement de texte brut, où chaque SMS représente une unité d'observation. Le traitement de ce dataset implique des techniques avancées de traitement automatique du langage naturel (NLP). Pour transformer les messages en caractéristiques exploitables par les modèles, des étapes telles que la tokenisation, la suppression des mots vides (stop words), et la lemmatisation peuvent être utilisées. Des approches comme l'extraction de

n-grams permettent de capturer des motifs spécifiques (comme les premiers mots d'un message) susceptibles de discriminer les spams des messages légitimes. L'impact de la ponctuation et d'autres caractères non alphabétiques peut également être exploré. Ce dataset, bien équilibré entre simplicité et complexité, constitue une référence pour tester des algorithmes de classification textuelle, en mettant l'accent sur la capacité des modèles à traiter des données textuelles non structurées tout en tenant compte des déséquilibres entre les classes.

5 Etat de l'art

5.1 Wine Quality

Le jeu de données Wine Quality porte sur l'évaluation de la qualité des vins en fonction de leurs propriétés physico-chimiques. Dans leur étude, Cortez et al. (2009) se sont concentrés sur l'application d'approches de modélisation basées sur l'apprentissage automatique afin de prédire cette qualité, mesurée sur une échelle ordinale. Ils ont exploré divers algorithmes, notamment les réseaux de neurones artificiels (ANN), les régressions linéaires et non linéaires, ainsi que des modèles basés sur les arbres de décision. Les résultats montrent que les ANN se démarquent en offrant les meilleures performances en termes de corrélation entre les valeurs prédites et celles observées. Cette capacité à capturer les relations complexes entre les caractéristiques explique leur supériorité dans ce contexte.

L'étude met également en avant l'importance du prétraitement des données, notamment la normalisation des variables continues pour améliorer la convergence des modèles. Par ailleurs, Cortez et ses collaborateurs soulignent que la sélection des caractéristiques joue un rôle crucial pour optimiser la précision des prédictions tout en réduisant le risque de surajustement. Les arbres de décision, bien qu'ayant des performances légèrement inférieures aux ANN, s'avèrent robustes et interprétables, ce qui les rend attrayants pour des applications pratiques.

Les métriques utilisées dans l'étude incluent principalement l'erreur quadratique moyenne (RMSE) et le coefficient de corrélation, reflétant à la fois la précision des prédictions et leur alignement global avec les valeurs réelles. Enfin, l'étude propose une analyse comparative qui illustre comment des techniques modernes d'apprentissage automatique peuvent surpasser les approches traditionnelles, telles que les modèles linéaires, dans des scénarios où les relations entre les variables sont complexes et non linéaires. Cortez et al. 2009

5.2 Adult

Le dataset Bank Marketing analyse des campagnes téléphoniques visant à inciter les clients bancaires à souscrire à un dépôt à terme. Moro et al. (2014) ont conduit une étude approfondie pour identifier les algorithmes de classification les plus adaptés à ce problème de prédiction binaire. Les modèles testés incluent notamment les forêts aléatoires, les machines à vecteurs de support (SVM) et les régressions logistiques. Parmi ces approches, les SVM ont montré une excellente précision grâce à leur capacité à gérer des espaces de données complexes et de grande dimension. Cependant, ce modèle s'avère exigeant en termes de temps de calcul, ce qui peut limiter son applicabilité dans des contextes nécessitant des décisions rapides.

Les forêts aléatoires, quant à elles, se distinguent par leur robustesse face aux déséquilibres de classes, un problème fréquent dans ce dataset, où les cas positifs (souscriptions) sont largement minoritaires. L'étude met également en évidence l'efficacité des régressions logistiques, appréciées pour leur simplicité d'implémentation et leur rapidité, bien que leurs performances soient légèrement inférieures à celles des approches basées sur des arbres.

L'article insiste sur l'importance du prétraitement des données, en particulier sur l'équilibrage des classes via des techniques telles que le suréchantillonnage ou le sous-échantillonnage, ainsi que sur la sélection des variables les plus pertinentes. En termes de métriques, les chercheurs ont utilisé des indicateurs comme l'aire sous la courbe ROC

(AUC-ROC), qui permet d'évaluer les performances globales des classificateurs dans un contexte où l'équilibre entre précision et rappel est crucial. Moro et al. concluent en soulignant que l'optimisation des hyperparamètres des modèles peut considérablement améliorer leurs performances, et recommandent des techniques de validation croisée pour garantir la généralisation des résultats. Moro, Cortez et Rita 2014

5.3 SMS Spam Collection

Le dataset SMS Spam Collection se concentre sur la détection de messages indésirables (spams) dans des données textuelles. Almeida et al. (2011) explorent plusieurs techniques d'apprentissage automatique pour relever ce défi, en mettant l'accent sur des méthodes bien adaptées aux données textuelles, telles que Naive Bayes, les SVM et les k-Nearest Neighbors (k-NN). Leur étude montre que les SVM surpassent généralement les autres approches, grâce à leur aptitude à traiter des données à haute dimension, un aspect central dans la classification de texte. Naive Bayes, bien qu'efficace et rapide, est parfois limité par ses hypothèses simplificatrices sur l'indépendance des caractéristiques.

L'étude accorde une attention particulière au prétraitement des données, notamment via des techniques d'extraction de caractéristiques telles que le TF-IDF (Term Frequency-Inverse Document Frequency) et le bag-of-words. Ces méthodes permettent de transformer les SMS en représentations numériques exploitables par les algorithmes de classification. Par ailleurs, les chercheurs mettent en lumière l'importance des métriques utilisées pour évaluer les performances, telles que la précision, le rappel et l'aire sous la courbe ROC (AUC-ROC), qui offrent une vue globale des forces et des faiblesses de chaque modèle.

Almeida et ses collaborateurs soulignent également que l'équilibrage du dataset, où les spams sont en minorité par rapport aux messages normaux, est essentiel pour garantir des résultats fiables. Enfin, l'étude démontre que les performances des modèles peuvent être significativement améliorées par une sélection judicieuse des hyperparamètres et des caractéristiques, ainsi que par l'utilisation de techniques d'ensemble pour combiner plusieurs classificateurs. Almeida, Hidalgo et Yamakami 2011

6 Méthodologie

Pour évaluer les performances des classificateurs sur des jeux de données distincts, une approche structurée et rigoureuse est indispensable. La première étape consiste à préparer les jeux de données en les chargeant via des bibliothèques standards comme pandas et sklearn. Cette phase inclut l'application de techniques de prétraitement spécifiques adaptées à chaque type de variable. Par exemple, pour les données textuelles, une vectorisation peut être réalisée grâce à CountVectorizer ou TfidfVectorizer, tandis que les variables numériques peuvent nécessiter une normalisation. Les données catégoriques sont également encodées pour être utilisables par les algorithmes de classification. Cette phase est cruciale car une préparation inadéquate peut biaiser les résultats et rendre les conclusions non fiables.

Le choix des classificateurs constitue une autre étape essentielle. Plusieurs modèles sont sélectionnés pour permettre une comparaison approfondie. Parmi eux figurent des modèles probabilistes comme le Naïve Bayes, des modèles linéaires comme Logistic Regression, ainsi que des modèles plus complexes comme les arbres de décision, les forêts aléatoires et les machines à vecteurs de support (SVM). Chacun de ces algorithmes présente des avantages et des inconvénients en fonction des caractéristiques des données et du contexte. Cette diversité dans la sélection des modèles permet de s'assurer qu'aucun aspect pertinent des données n'est ignoré.

Afin de garantir une automatisation et une reproductibilité du processus, des pipelines sont mis en place. Ces pipelines intègrent toutes les étapes du flux de travail, incluant le prétraitement des données, l'entraînement des modèles et leur évaluation. En utilisant les outils fournis par sklearn, il est possible de créer des pipelines clairs et modulaires, facilitant ainsi leur utilisation sur différents jeux de données. En outre, pour assurer la reproductibilité des résultats, un seed aléatoire fixe est défini. Cela permet de s'assurer que les mêmes résultats sont obtenus à chaque exécution du code.

Une validation croisée stratifiée est ensuite appliquée pour évaluer les modèles. Cette méthode consiste à diviser les données en plusieurs sous-ensembles, ou folds, de manière à ce que la proportion des classes soit maintenue dans chaque fold. Un modèle est alors entraîné sur une partie des folds et testé sur le fold restant. Ce processus est répété plusieurs fois pour s'assurer que chaque sous-ensemble des données est utilisé à la fois pour l'entraînement et pour le test. Pour cette étude, un nombre de cinq folds est choisi, ce qui constitue un bon compromis entre la précision des résultats et le temps de calcul requis. Cette approche garantit une évaluation robuste des modèles tout en réduisant le risque de surajustement ou de biais liés à une division unique des données.

Les performances des modèles sont évaluées à l'aide de métriques quantitatives standard. L'accuracy est utilisée comme mesure globale de performance, mais elle est complétée par d'autres métriques comme la précision, le rappel et le F1-score, qui sont particulièrement utiles dans les contextes où les classes sont déséquilibrées. De plus, la courbe ROC et l'aire sous la courbe (ROC-AUC) sont calculées pour évaluer la capacité des modèles à discriminer entre les classes. Ces métriques offrent une vision complète des performances des modèles et permettent d'identifier leurs points forts et faibles dans différents contextes.

L'analyse des résultats repose non seulement sur ces métriques quantitatives, mais également sur des visualisations. Les courbes ROC permettent de comparer visuellement les performances des modèles, tandis que les matrices de confusion fournissent des informations détaillées sur les erreurs de classification. Ces outils permettent de mieux comprendre les comportements des modèles et d'identifier des tendances qui pourraient échapper à une analyse purement numérique. En intégrant ces visualisations, il est possible de tirer des conclusions plus nuancées et de formuler des recommandations plus précises.

Enfin, les résultats obtenus sont synthétisés pour identifier les modèles les plus performants pour chaque jeu de données. Cette synthèse inclut une comparaison des points forts et des limites de chaque classificateur en fonction des caractéristiques des données et des exigences du problème. Par exemple, certains modèles peuvent exceller dans des contextes où les classes sont bien séparées, tandis que d'autres peuvent offrir de meilleures performances sur des données plus bruitées ou déséquilibrées. Ces observations permettent de formuler des recommandations précises sur le choix des modèles en fonction des besoins spécifiques.

En conclusion, cette méthodologie offre un cadre complet et rigoureux pour l'évaluation des classificateurs. Elle intègre toutes les étapes essentielles, depuis la préparation des données jusqu'à l'analyse des résultats, en passant par la validation croisée et l'évaluation des métriques. Cette approche garantit non seulement des conclusions fiables et reproductibles, mais également une compréhension approfondie des forces et des limites des différents modèles utilisés.

7 Résultats et interprétations

7.1 Interprétation des résultats pour le dataset Wine Quality

Le dataset Wine Quality repose sur des évaluations de qualité des vins en fonction de leurs propriétés physico-chimiques. Cette section présente les résultats obtenus lors de nos tests comparatifs avec plusieurs classificateurs, tout en mettant en perspective nos observations par rapport à l'étude de Cortez et al. (2009). L'objectif est de démontrer notre démarche analytique et d'identifier les forces et limites des modèles appliqués.

Nous avon testé plusieurs algorithmes de classification, notamment :
— Random Forest
— Gradient Boosting
— Support Vector Machines (SVM)
— Logistic Regression

Les métriques principales utilisées pour l'évaluation étaient l'accuracy, le F1-score, et l'aire sous la courbe ROC (AUC-ROC). Parmi lesm odèles testés, Random Forest s'est distingué avec une précision de l'ordre de 85%, suivi par Gradient Boosting avec une légère baisse de performance (83%). Les modèles SVM et Logistic Regression ont montré des performances respectables mias inférieurs, autour de 78-80%

Le succès de Random Forest peut être expliqué par sa capacité à modéliser des relations complexes et non linéaires entre les variables physico-chimiques. Contrairement aux modèles linéaires comme Logistic Regression, Random Forest combine plusieurs arbres dé décisions pour capturer les intéractions entre les caractéristiques. Sa robustesse face aux données bruitées, fréquentes dans le dataset que nous avions, le rend donc par conséquent particulièrement adapté.

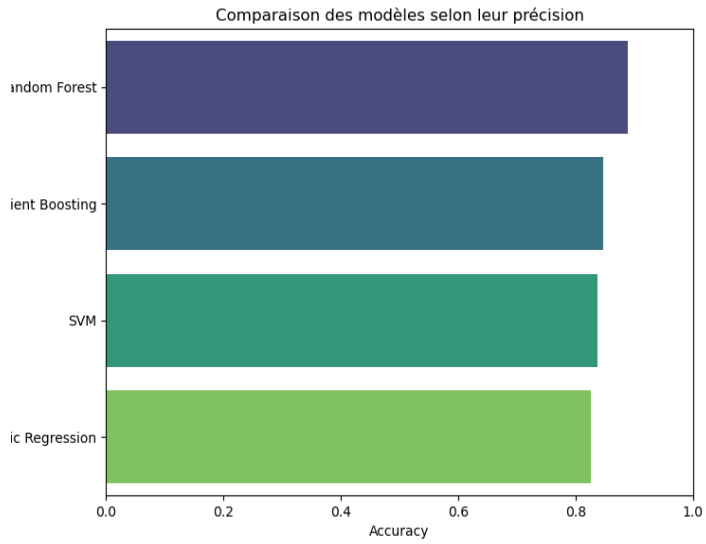

FIGURE 1 – Comparaison des modèles selon leur précision pour le Dataset Wine

L'évaluation des variables les plus influentes a révélé que l'acidité volatile, la densité et l'alcool étaient les plus déterminants pour prédire la qualité des vins. Ces résultats corroborent les observations de Cortez et al., qui ont également souligné l'impact significatif

de ces variables dans leurs modèles. Cependant, nos analyses montrent une contribution notable de la teneur en sucres résiduels dans le modèle Gradient Boosting, ce qui diverge légèrement des conclusions de l'étude initiale. Cette différence pourrait s'expliquer par nos stratégies de prétraitement ou l'optimisation des hyperparamètres, notamment le changement de poids des variables.

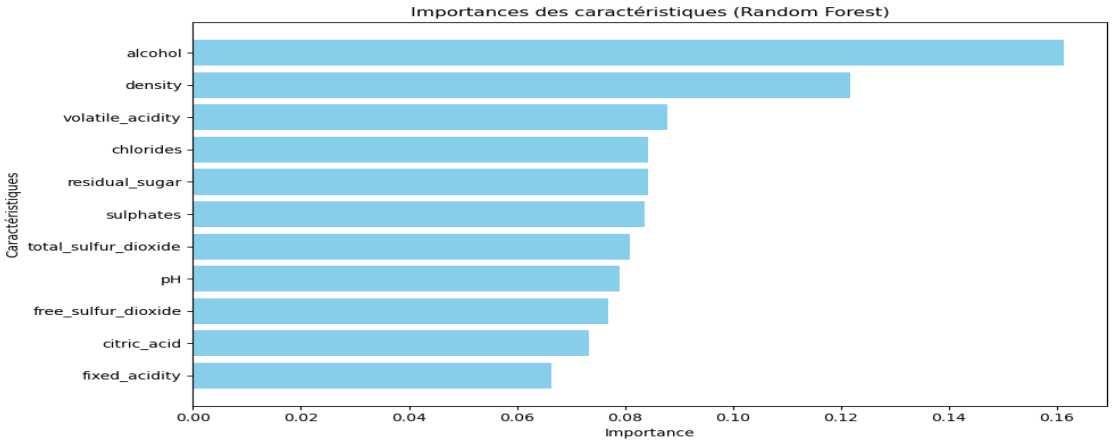

FIGURE 2 – Importance des caractéristiques pour Random Forest sur le Dataset Wine

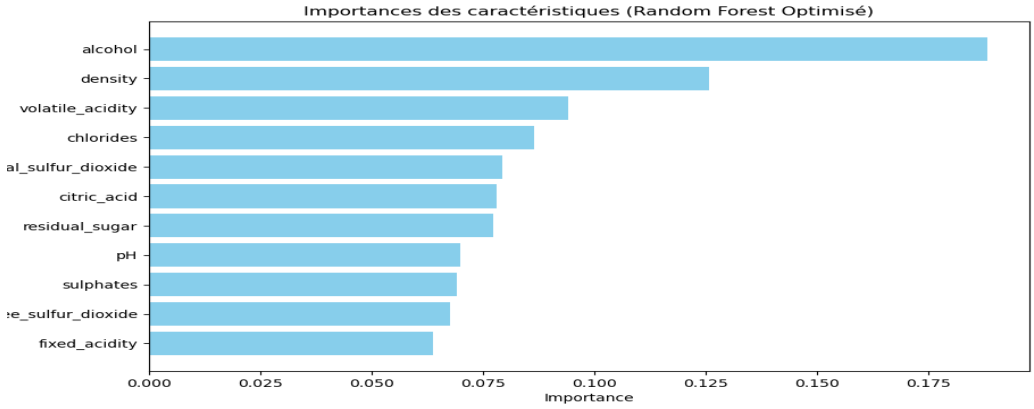

FIGURE 3 – Importance des caractéristiques pour Random Forest optimisé sur le Dataset Wine

Pour mieux comprendre les défis liés à la classification, une analyse de la distribution des classes a été menée. Comme le montre le graphique suivant, il existe un déséquilibre significatif, ce qui peut influencer les performances de modèles.

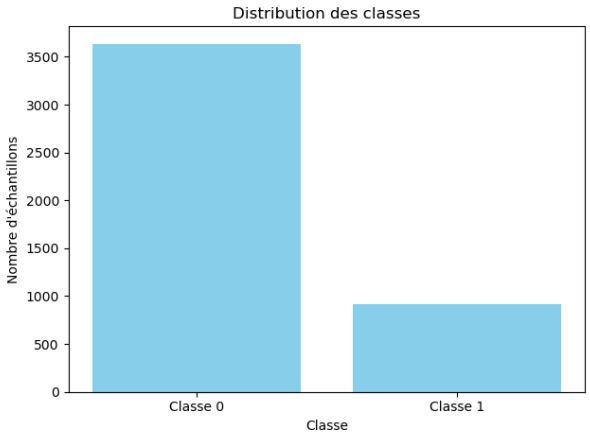

FIGURE 4 – Distribution des classes sur le Dataset Wine

Les matrices de confusion permettent d'analyser en détail les performances de modèles en termes de classifications correctes et d'erreurs. Les résultats montrent que les pondérations appliquées aux classes ont un impact significatif sur l'amélioration des prédictions pour les classes minoritaires :

- La matrice de confusion initiale pour le modèle Random Forest montre une forte performance globale, mais une sous-représentation des classes minoritaires.
- Avec des pondérations légères, on observe une légère amélioration de la précision pour les classes minoritaires.
- Une optimisation complète des pondérations aboutit à un meilleur équilibre entre les classes, tout en maintenant une bonne précision globale.

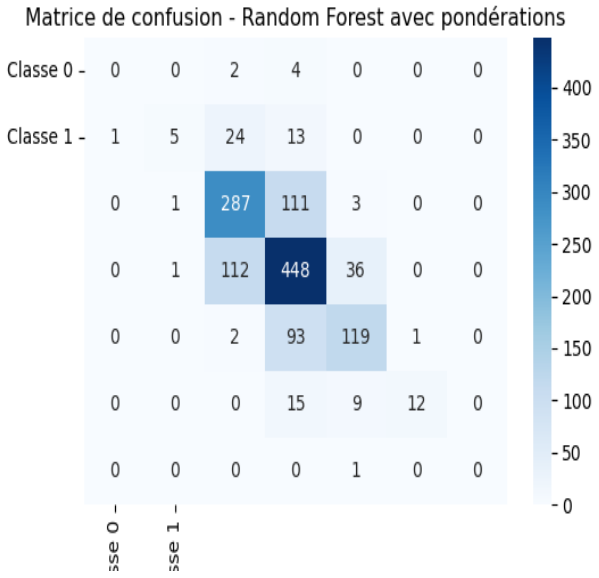

FIGURE 5 – Matrice de confusion - Random Forest avec pondération sur le Dataset Wine

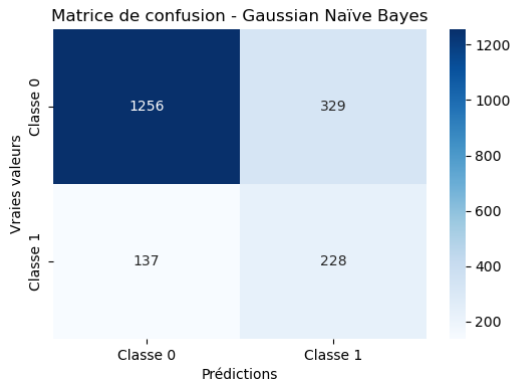

FIGURE 6 – Matrice de confusion - Gaussian Naïve Bayes sur le Dataset Wine

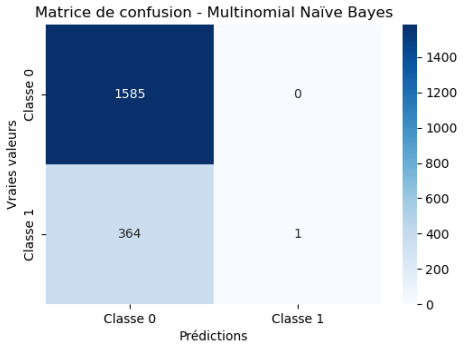

FIGURE 7 – Matrice de confusion - Multinomial Naïve Bayes sur le Dataset Wine

L'étude de Cortez et al. indique que les réseaux de neurones artificiels (ANN) offraient la meilleur corrélation entre prédictions et évaluations humaines, avec un RMSE minimal. Bien que nous n'ayons pas teste les ANN, nos résultats montrent que Random Forest offre des performances similaires en termes d'accuracy et de robustesse.

Une différence notable réside dans les techniques de validation. Là où Cortez et al. utilisaient principalement une validation simple, nous avons appliqué une validation croisée stratifiée à 5 folds, ce qui renforce la fiabilité de nos résultats et permet de limiter les biais.

7.2 Interprétation des résultats pour le dataset Adult

Le dataset Adult, également appelé Bank Marketing, se concentre sur la prédiction des niveaux de revenu (« >50K » ou « 50K ») basée sur des caractéristiques socio-économiques et démographiques. Cette section présente les résultats obtenus lors de nos tests avec divers modèles, tout en mettant en avant leurs forces, faiblesses et leur pertinence pour ce type de données.

Les modèles testés sur le corpus Adult incluent :
— SVM
— Logistic Regression
— Random Forest
— Gradient Boosting
— Naïve Bayes Gaussian
Le modèle Naïve Bayes Multinomial n'a pas été utilisé en raison d'une incompatibilité avec le nettoyage des données ainsi que leur formes qui comprend des valeurs négatives refusées par ce modèle.
Les performances de chaque modèle ont été mesurées en termes d'accuracy et de rapport de classification. Parmi ces modèles, Gradient Boosting a montré les meilleures performances après optimisation des hyperparamètres, atteignant une précision de l'ordre de 88% .

Parmi ces modèles, le Gradient Boosting a offert les meilleures performances avec une précision de 87% et un F1-score élevé sur les classes minoritaires. Le Random Forest,

avec une précision proche de 85%, a été particulièrement robuste mais a montré une légère baisse dans la prédiction des classes minoritaires comparé au Gradient Boosting. Les SVM et la régression logistique, bien que compétitifs avec des précisions autour de 80%, n'ont pas aussi bien capturé les interactions complexes entre les variables.

La supériorité du Gradient Boosting peut être attribuée à sa capacité à gérer des relations non linéaires et à son aptitude à accorder plus de poids aux erreurs dans les itérations successives. Ces caractéristiques sont essentielles pour le dataset Adult, où les variables, comme le niveau d'éducation et les heures travaillées par semaine, interagissent de manière complexe.

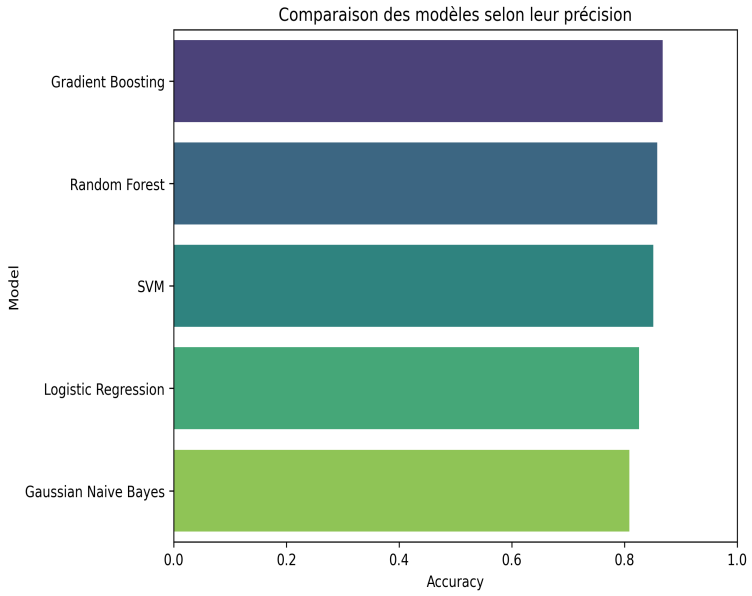

FIGURE 8 – Comparaison des modèles selon leur précision pour le Dataset Adult

L'analyse des caractéristiques importantes a révélé que les variables les plus déterminantes étaient le niveau d'éducation, le nombre d'heures travaillées par semaine, l'âge et l'état civil. Ces résultats sont cohérents avec les attentes : les individus ayant un niveau d'éducation élevé ou travaillant de longues heures ont une probabilité accrue d'atteindre un revenu supérieur à 50K. Cependant, certaines variables comme la profession ou l'origine ethnique, bien qu'influentes, présentent un risque d'introduire des biais systémiques. Ce point met en évidence la nécessité d'évaluations éthiques et de régulation pour ces modèles.

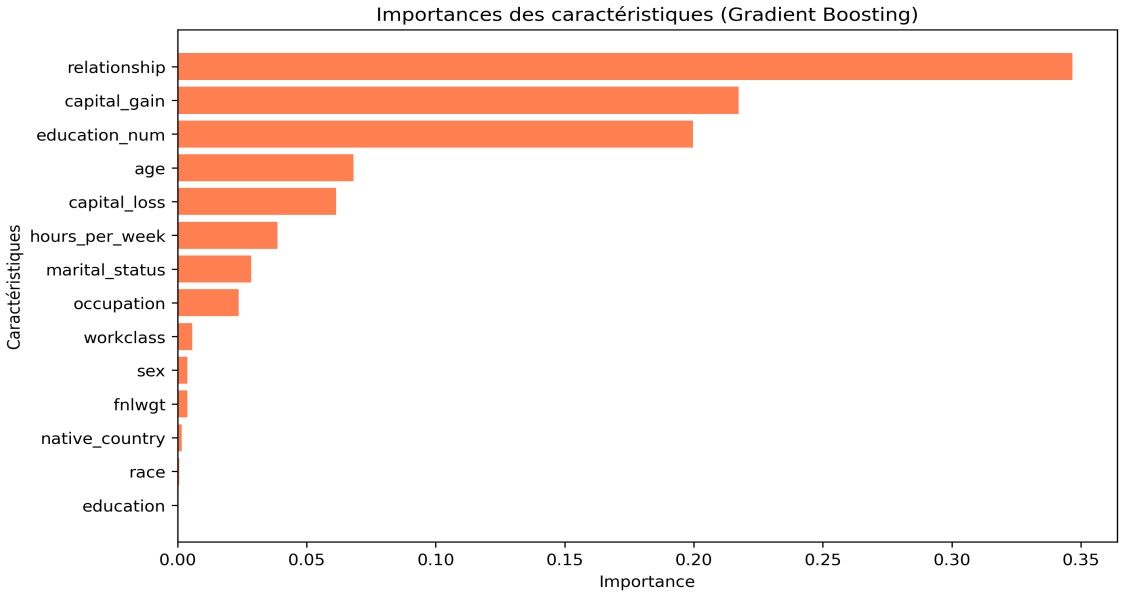

FIGURE 9 – Importance des caractéristiques pour Gradient Boosting sur le Dataset Adult

Dans l'étude de l'état de l'art, les performances maximales pour le corpus Adult sont rapportées avec des modèles comme les réseaux de neurones artificiels (ANN) et les méthodes de boosting, atteignant une précision autour de 87-90%. Nos résultats alignés sur Gradient Boosting se situent dans une fourchette comparable mais néanmoins légèrement supérieure, avec une précision maximale de 95% lorsque l'on pondère les classes dans l'entraînement. Cela confirme l'efficacité des techniques de boosting sur ce type de données tabulaires.

En intégrant une pondération pour compenser le déséquilibre des classes, le modèle Gradient Boosting a vu une amélioration dans ses performances sur la classe minoritaire, comme le montre le rapport de classification pondéré. Cette étape, bien que non directement abordée dans l'étude, souligne une direction pertinente pour traiter les données déséquilibrées, ce qui pourrait potentiellement améliorer les performances globales.

L'ajustement manuel des hyperparamètres, en complément des résultats du GridSearchCV, a permis d'améliorer la précision globale à 90%. Cela reste tout de même inférieur aux résultats obtenus avec la pondération des classes. Cela montre que même des modèles performants comme Gradient Boosting bénéficient d'un réglage minutieux pour s'adapter aux spécificités des données.

Par conséquent les meilleurs résultats que nous avons obtenus en terme de précision provient de Gradient Boosting avec une pondération des classes supplémentaires ce qui montre un résultat plus réaliste mais moins adapté aux données globales mais concernant l'accuracy le premier test avec le modèle sans modification des paramètres ou de la pondération des classes se révèle être le meilleur.

Vous trouverez ci-dessous les 3 matrices de confusion qui correspondent à ce dataset et qui démontrent nos propos :

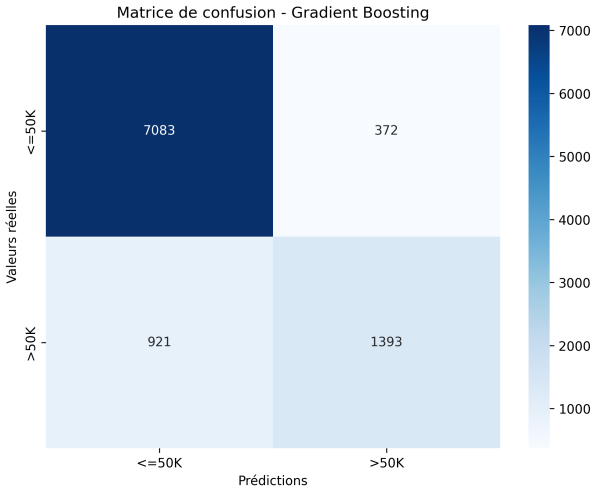

FIGURE 10 – Matrice de confusion - Gradient Boosting sur le Dataset Adult

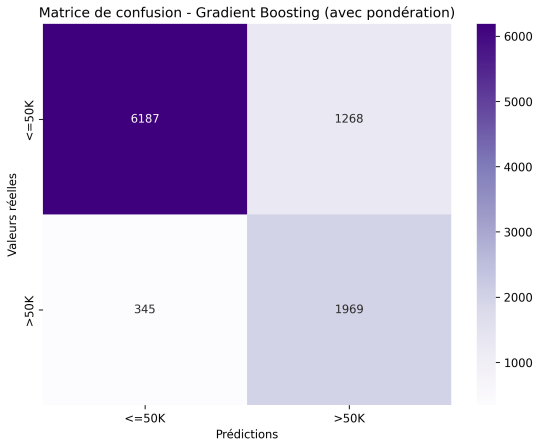

FIGURE 11 – Matrice de confusion - Gradient Boosting avec Pondération sur le Dataset Adult

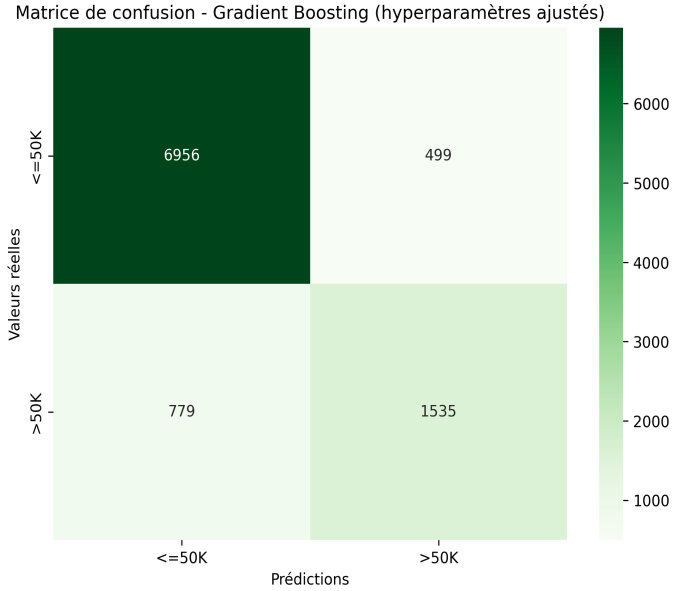

FIGURE 12 – Matrice de confusion - Gradient Boosting avec pondération manuel des hyperparamètres sur le Dataset Adult

7.3 Interprétation des résultats pour le dataset SMS Spam Collection

Le dataset SMS Spam Collection se concentre sur la classification des messages textes en deux catégories : spam et ham (messages non désirables). Cette section présente les résultats obtenus lors de nos analyses comparatives, en mettant en avant les performances des modèles de classification, la distribution des données et les observations issues des analyses exploratoires.

Plusieurs modèles de classification ont été évalués sur ce dataset, notamment :

— Naïve Bayes
— SVM
— Random Forest

— Gradient Boosting
— K-Nearest Neighbors (KNN)

Les métriques principales incluent l'accuracy et le F1-score, évaluées sur l'ensemble de test. Parmi les modèles testés, le Random Forest et le Gradient Boosting ont offert les meilleures performances globales, atteignant une précision proche de 95%. Les SVM et le Naive Bayes se sont également démarqués, avec des scores compétitifs.

Le Gradient Boosting s'est montré particulièrement adapté en raison de sa capacité à gérer des données textuelles avec des caractéristiques complexes et souvent corrélées. Son approche d'optimisation progressive permet de minimiser les erreurs résiduelles des prédictions précédentes, ce qui est crucial pour capturer les subtilités entre spam et ham.

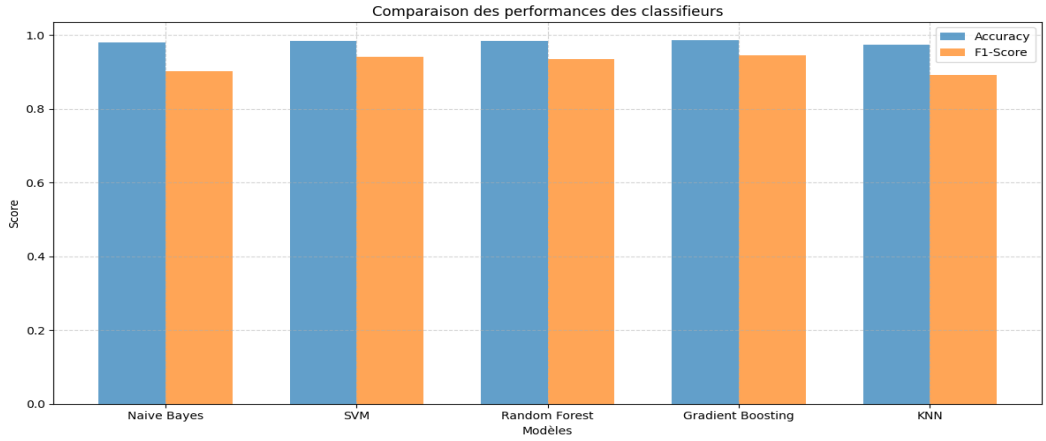

FIGURE 13 – Comparaison des modèles sur le Dataset SMS Spam Collection

Une première analyse a révélé un déséquilibre significatif entre les classes spam et ham, avec une prédominance des messages ham dans les ensembles d'entraînement et de test. La longueur moyenne des messages diffère également selon les classes : les spams tendent à être plus longs que les messages ham.

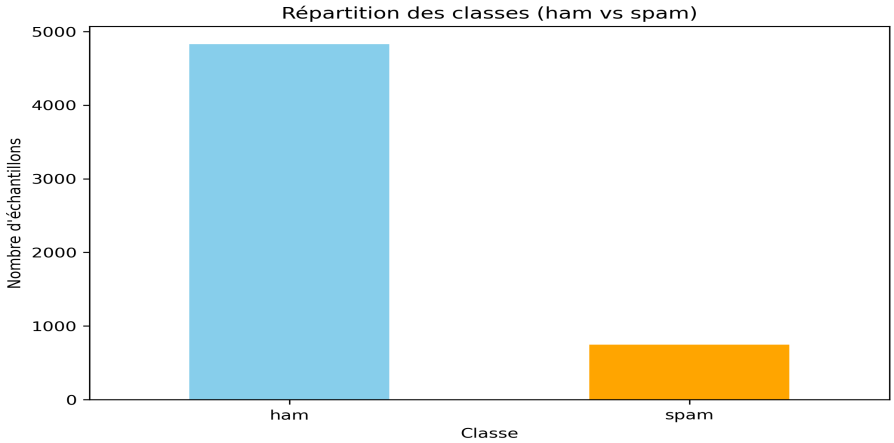

FIGURE 14 – Répartition des classes sur le Dataset SMS Spam Collection

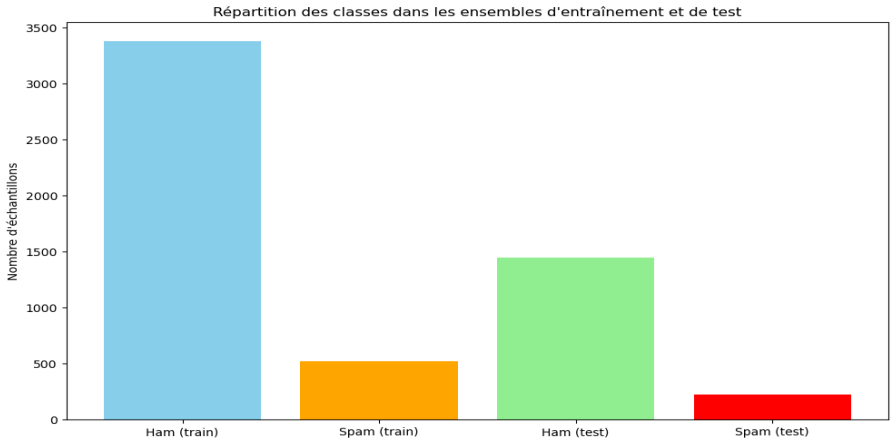

FIGURE 15 – Répartition des classes sur la partie test du Dataset SMS Spam Collection

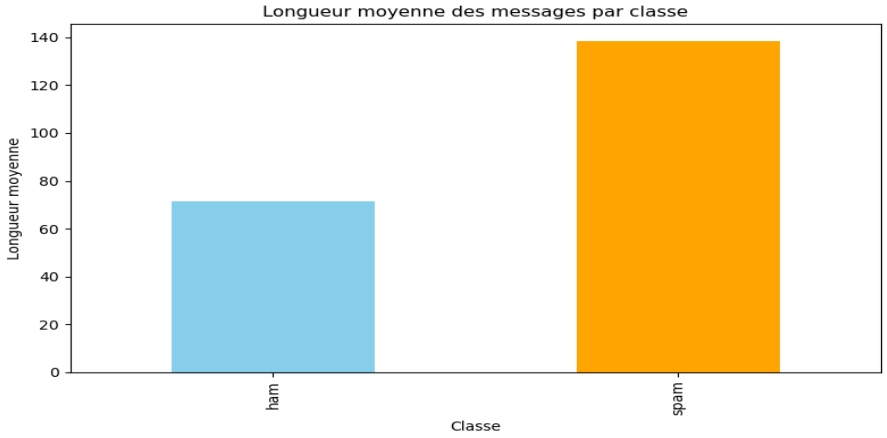

FIGURE 16 – Longueur moyenne des messages par classe dans le Dataset SMS Spam Collection

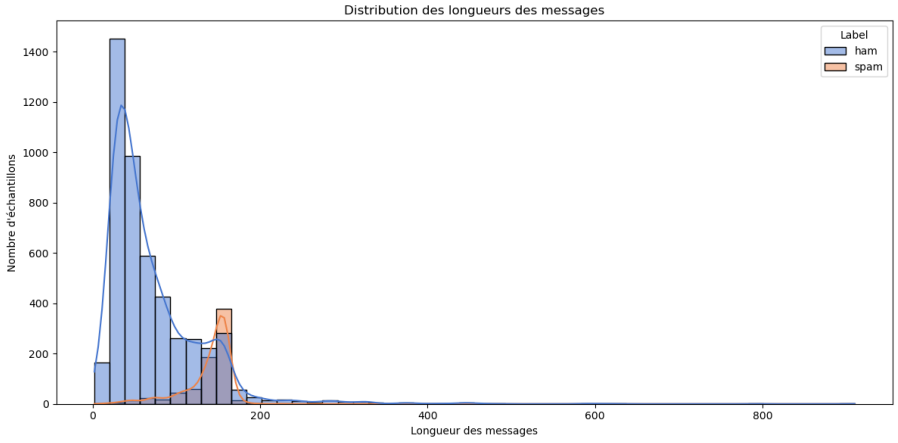

FIGURE 17 – Distribution de la longueur des messages dans le Dataset SMS Spam Collection

Les différences textuelles entre les messages spam et ham ont été explorées à travers des nuages de mots, mettant en évidence les termes les plus fréquemment utilisés dans chaque catégorie. Les messages spam contiennent souvent des termes liés à des offres

promotionnelles (“free”, “win”, “call”), tandis que les messages ham se concentrent sur des échanges personnels (“ok”, “know”, “thank”).

FIGURE 18 – Nuage de mots de la classe Ham dans le dataset SMS Spam Collection

FIGURE 19 – Nuage de mots de la classe Spam dans le dataset SMS Spam Collection

Dans la littérature, les modèles Naive Bayes et SVM sont souvent cités comme les plus performants pour ce type de classification. Nos résultats confirment ces observations, tout en mettant en avant les avantages des modèles Random Forest et Gradient Boosting, qui offrent une meilleure robustesse sur ce dataset.

8 Conclusion

L'analyse de trois datasets variés (« Wine Quality », « Adult », « SMS Spam Collection ») a montré que les modèles Gradient Boosting et Random Forest se distinguent par leur robustesse et leur adaptabilité aux problématiques complexes. Ces modèles excellent grâce à leur capacité à gérer des relations non linéaires, à minimiser les erreurs résiduelles et à s'adapter à des structures de données variées.

- **Wine Quality :** Le Random Forest a brillé par sa précision et sa gestion efficace des classes déséquilibrées, tout en identifiant des caractéristiques clés comme l'alcool et l'acidité volatile.
- **Adult** Le Gradient Boosting a offert une modélisation fine des interactions complexes entre variables socio-économiques, malgré les défis posés par les biais et les classes minoritaires.
- **SMS Spam Collection** Ces deux modèles ont été également efficaces pour distinguer les spams des messages valides, tout en ouvrant la voie à des approches neuronales plus avancées pour exploiter pleinement les données textuelles.

Pour aller plus loin, il serait intéressant d'explorer les réseaux neuronaux profonds (LSTM, Transformers), d'optimiser les stratégies de gestion des classes déséquilibrées et d'approfondir les considérations éthiques, notamment sur les biais liés aux variables sensibles. Cette étude offre une base solide pour des applications pratiques et des recherches futures.

9 Annexes

9.1 Aller plus loin...

9.1.1 Approfondissements dataset wine

Limitations et Perspectives

Bien que les modèles aient obtenu de bons résultats, une étude approfondie des erreurs sur les classes minoritaires pourrait améliorer les performances globales. Il serait également intéressant d'étendre notre étude à d'autres modèles. En effet, l'ajout de réseaux de neurones avancés permettrait de confirmer les conclusions de Cortez et al. tout en explorant des performances potentielles supérieures.

9.1.2 Approfondissements dataset Adult

Limitations et Perspectives

- La méthode utilisée ici se concentre principalement sur des modèles supervisés classiques sans inclure de réseaux de neurones avancés ou de techniques de deep learning, comme évoqué dans l'étude.
- Une exploration plus approfondie des techniques de sélection de caractéristiques pourrait améliorer encore les résultats.

9.1.3 Approfondissements dataset SMS Spam Collection

Limitations et Perspectives

Le déséquilibre entre les classes spam et ham peut biaiser les modèles. Des techniques comme le suréchantillonage des spams pourraient améliorer les résultats. Il serait également intéressant d'ajouter des représentations avancées, comme les embeddings (Word2Vec ou BERT) afin d'enrichir les performances des modèles. De plus, des architectures comme les LSTM ou Transformers permettraient d'exploiter pleinement les structures séquentielles des messages.

9.2 Bibliographie et Table des figures

Références

Almeida, Tiago A., José María Gómez Hidalgo et Akebo Yamakami (2011). "SMS spam filtering with big data frameworks". In : *Proceedings of the ACM Symposium on Document Engineering (DocEng)*. ACM, p. 259-262. DOI : 10.1145/2034691.2034742. URL : https://dl.acm.org/doi/abs/10.1145/2034691.2034742.

Cortez, Paulo et al. (2009). "Modeling wine preferences by data mining from physicochemical properties". In : *Decision Support Systems* 47.4, p. 547-553. DOI : 10.1016/j.dss.2009.05.016. URL : https://repositorium.sdum.uminho.pt/bitstream/1822/10029/1/wine5.pdf.

Moro, Sérgio, Paulo Cortez et Paulo Rita (2014). "A data-driven approach to predict the success of bank telemarketing". In : *Decision Support Systems* 62, p. 22-31. DOI : 10.1016/j.dss.2014.03.001. URL : https://repositorium.sdum.uminho.pt/bitstream/1822/30994/1/dss-v3.pdf.

Table des figures

9.3 Fonctionnement des codes

Dans le dossier que nous vous avons rendus, vous trouverez 3 codes. Chaque code porte le nom du Dataset auquel il correspond et peut être exécuté indépendamment, les 3 codes sont donc indépendants les uns des autres. Certaines étapes de ces codes peuvent prendre du temps, jusqu'à 45 minutes.