Projet de Méthodologie de la recherche en informatique

Classification

Master 1 Langue & Informatique

Yeleen CANONGE et Camille FIEF

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).

Problématique et motivation

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

1. 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.
2. 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.

Objectifs du projet

Ce projet s'articule autour des objectifs suivants :

- 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 pouvant être comblées par des méthodes modernes, avec également une comparaison des classificateurs analysés.
- Proposer une classification avec une cohérence tout en déterminant quel classificateur est le plus adapté au jeu de données en fonction de celui-ci.
- Optimisation du traitement des données par les classificateurs

Description des données

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.

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.

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.

Etat de l'art

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.

2. Bank Marketing

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.

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.

Méthodologie base Wine

Dans ce projet, l'objectif est d'évaluer et de comparer les performances de plusieurs algorithmes de classification pour prédire les résultats de la manière qui fonctionne le mieux.
Pour le dataset Wine il s'agit de prédire la qualité du vin à partir du dataset "Wine Quality", disponible sur l'UCI Machine Learning Repository. Les étapes sont structurées autour de la préparation des données, de la modélisation, de l'optimisation des hyperparamètres et de l'interprétation des résultats.

La première étape consiste à importer les données et à transformer la variable cible pour en faire un problème de classification binaire : les vins ayant une note de qualité supérieure à 6 sont considérés comme "bons" (classe 1), tandis que les autres sont classés comme "non bons" (classe 0). Ensuite, un prétraitement est appliqué aux variables explicatives afin de garantir une échelle cohérente, en utilisant une standardisation via l'outil StandardScaler.

Les données sont ensuite divisées en deux ensembles distincts : un ensemble d'entraînement représentant 70 % des observations, et un ensemble de test composé des 30 % restants. Cette séparation permet de former les modèles tout en conservant un sous-ensemble de données pour évaluer leur capacité de généralisation. Une attention particulière est accordée à la distribution des classes pour éviter d'éventuels déséquilibres, pouvant biaiser les performances des modèles.

Différents algorithmes de classification sont mis en œuvre, notamment la régression logistique, les Random Forests, le Gradient Boosting, les SVM, ainsi que les approches Naïve Bayes (Gaussian et Multinomial). Chaque modèle est entraîné sur l'ensemble d'entraînement et testé sur l'ensemble de test. Les performances sont évaluées à l'aide de métriques standards, telles que la précision globale (accuracy), les matrices de confusion et des rapports détaillés sur les performances des classes. Ces métriques permettent de comprendre non seulement la capacité prédictive globale des modèles, mais aussi leur comportement sur chaque classe.

Pour améliorer les performances, une optimisation des hyperparamètres est réalisée sur le modèle Random Forest à l'aide de la méthode GridSearchCV. Cette étape consiste à explorer plusieurs combinaisons de paramètres, comme le nombre d'arbres, la profondeur maximale, et les critères de division, afin d'identifier la configuration offrant les meilleurs résultats. Les performances du modèle optimisé sont ensuite réévaluées et comparées à celles des versions non optimisées.

Dans un souci de robustesse, des ajustements sont également réalisés pour traiter le déséquilibre potentiel entre les classes. Par exemple, un modèle Random Forest est testé avec l'option class_weight='balanced', permettant d'attribuer un poids proportionnel à chaque classe en fonction de sa fréquence dans les données. Cette approche est utile pour éviter qu'une classe majoritaire n'écrase les performances globales.

Enfin, une analyse approfondie est menée pour interpréter les résultats. Les performances des modèles sont comparées à travers des visualisations, mettant en évidence la précision obtenue pour chaque méthode. Une attention particulière est portée au modèle Random Forest optimisé, avec une analyse des importances des caractéristiques, révélant quelles variables influencent le plus les prédictions.

Cette méthodologie repose sur l'utilisation de Python et de bibliothèques adaptées, notamment pandas et numpy pour la manipulation des données, scikit-learn pour la modélisation et les métriques, et matplotlib et seaborn pour la visualisation des résultats. L'ensemble des étapes a été conçu pour fournir une évaluation rigoureuse et complète, permettant d'identifier le classificateur le plus performant pour ce type de données.

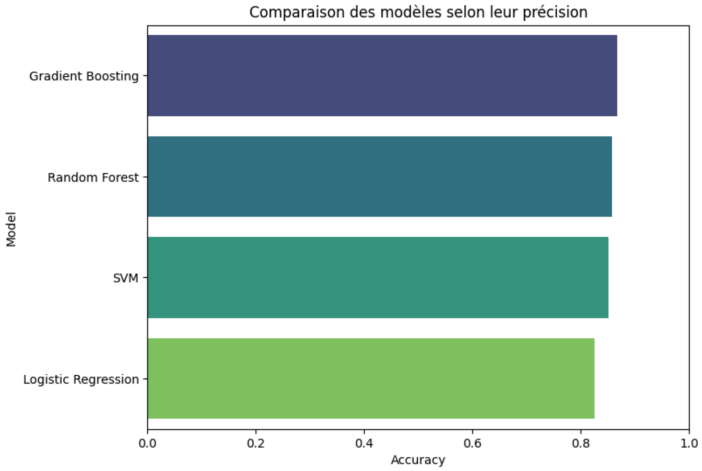
(La méthodologie ci dessus à été fait concernant le code Wine_quality, il faudra l'étendre aux autres dataset et la rendre moins spécifique et plus générique pour les différents dataset. Il faudra également l'adapté aux différentes méthodes utilisées pour chaque dataset et tout regrouper dans cette partie dès que les autres codes seront finis, nous allons également ajouter une parties sur le poids des caractéristiques et la possibilité d'optimisé ces poids.)

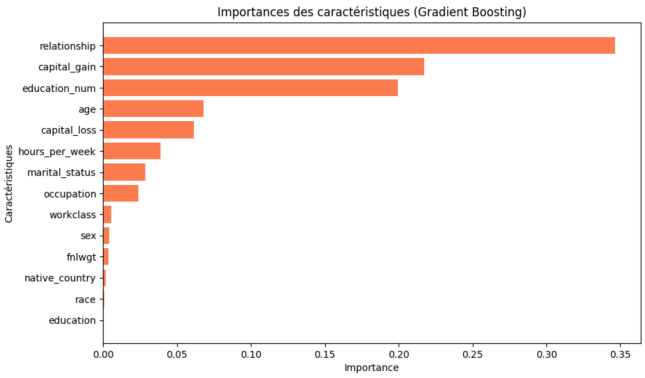
Interprétation (pour chaque dataset, caractéristique, poids des données à faire varié)

Nous allons interpréter dans cette partie les différents résultats obtenus pour les différents dataset, pour le moment nous avons seulement regroupé les différents graphiques générés, leurs interprétation sera faite prochainement.

Graphiques concernant le dataset adult/bank

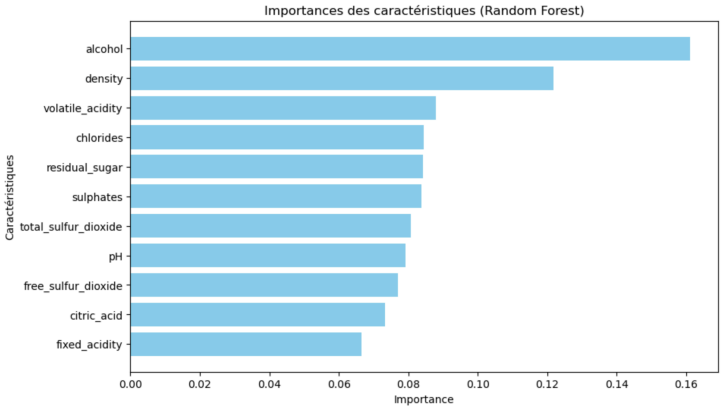
```
--- SVM ---
Accuracy: 0.85
              precision    recall  f1-score   support

           0       0.87      0.95      0.91      7455
           1       0.76      0.55      0.64      2314

    accuracy                           0.85      9769
   macro avg       0.81      0.75      0.77      9769
weighted avg       0.84      0.85      0.84      9769

--- Logistic Regression ---
Accuracy: 0.83
              precision    recall  f1-score   support

           0       0.85      0.94      0.89      7455
           1       0.71      0.45      0.55      2314

    accuracy                           0.83      9769
   macro avg       0.78      0.70      0.72      9769
weighted avg       0.81      0.83      0.81      9769

--- Random Forest ---
Accuracy: 0.86
              precision    recall  f1-score   support

           0       0.89      0.93      0.91      7455
           1       0.74      0.62      0.67      2314

    accuracy                           0.86      9769
   macro avg       0.81      0.78      0.79      9769
weighted avg       0.85      0.86      0.85      9769

--- Gradient Boosting ---
Accuracy: 0.87
              precision    recall  f1-score   support

           0       0.88      0.95      0.92      7455
           1       0.79      0.60      0.68      2314

    accuracy                           0.87      9769
   macro avg       0.84      0.78      0.80      9769
weighted avg       0.86      0.87      0.86      9769
```





Graphiques concernant le dataset wine_quality

```
Résumé des performances des modèles :
                Model  Accuracy
2       Random Forest  0.888205
3   Gradient Boosting  0.846154
0                 SVM  0.837436
1  Logistic Regression  0.826154
```



Conclusion V1

L'analyse comparative des classificateurs a révélé que le modèle Random Forest est celui qui offre les meilleures performances pour prédire la qualité des vins à partir des données étudiées. Avec une précision globale supérieure et une meilleure gestion des déséquilibres entre les classes, ce modèle s'est distingué par sa capacité à fournir des prédictions robustes et fiables. L'optimisation des hyperparamètres a permis d'améliorer encore davantage ses résultats, soulignant l'importance de l'étape de tuning pour exploiter pleinement son potentiel.

Par ailleurs, l'analyse des importances des caractéristiques a mis en évidence les variables les plus influentes dans la prédiction, ce qui offre des perspectives intéressantes pour des applications pratiques, notamment dans l'industrie vinicole. En revanche, bien que d'autres modèles comme le Gradient Boosting ou les SVM aient montré des performances respectables, ils n'ont pas égalé la polyvalence et l'efficacité du Random Forest sur ce dataset spécifique.

(La conclusion sera retravaillée afin d'être plus courte et synthétique, toutes les informations seront expliqué plus explicitement dans l'interprétation)

Aller plus loin…

*(**Dans cette partie nous allons parler des différents approfondissement qui pourraient être réalisés sur ce projet pour aller plus loin dans la recherche du meilleur résultat et également les différentes variations/modifications qu'il pourrait être intéressant d'explorer)***

Références

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553. Disponible à : https://repositorium.sdum.uminho.pt/bitstream/1822/10029/1/wine5.pdf

2. Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. Decision Support Systems, 62, 22-31. Disponible à : https://repositorium.sdum.uminho.pt/bitstream/1822/30994/1/dss-v3.pdf

3. Almeida, T. A., Hidalgo, J. M. G., & Yamakami, A. (2011). SMS spam filtering with big data frameworks. Proceedings of the ACM Symposium on Document Engineering (DocEng), 259-262. Disponible à : https://dl.acm.org/doi/abs/10.1145/2034691.2034742