RAPPORT DE MINI PROJET

Par

EL HASSNAOUI Salma

EL HILA Douae

MANSOUR Yousra

ZAIM Mohamed

Machine Learning

Prédiction des performances des élèves

Encadré par : Monsieur HAJA Zakaria

Année Universitaire 2023/2024

Contents

Introduction

Dans ce rapport, nous présenterons notre étude sur la prédiction des performances des élèves en utilisant des techniques d'apprentissage automatique.

Nous commencerons par définir l'objectif de notre projet, présenter la source des données utilisées, puis effectuer une analyse exploratoire des données. Ensuite, nous décrirons en détail la phase de prétraitement des données, suivie de l'approche algorithmique utilisée pour résoudre le problème de prédiction des performances des élèves.

Objectif

L'objectif de notre projet est de développer un modèle prédictif qui peut estimer les performances des élèves en fontion de diverses caractéristiques, telles que le sexe, l'âge, le niveau d'éducation des parents, le temps d'étude, etc.

Ce modèle pourrait être utile pour les écoles et les éducateurs afin d'identifier les élèves à risque et de mettre en place des interventions précoces pour les aider.

Source des données

Les données utilisées dans cette étude ont été obtenues à partir du jeu de données « Student Alcohol Consumption » disponible sur Kaggle.

Ce jeu de données contient des informations sur les performances scolaires de 395 élèves, ainsi que des données démographiques familiales et comportementales.

Analyse Exploratoire des données

Description du Jeu de Données

Le jeu de données "Student Alcohol Consumption" disponible sur Kaggle est une collection de données sur la consommation d'alcool chez les élèves de l'enseignement secondaire. Il contient des informations sur divers aspects de la vie des élèves, y compris leur consommation d'alcool, leurs performances scolaires, leur temps d'étude, etc.

Statistiques descriptives

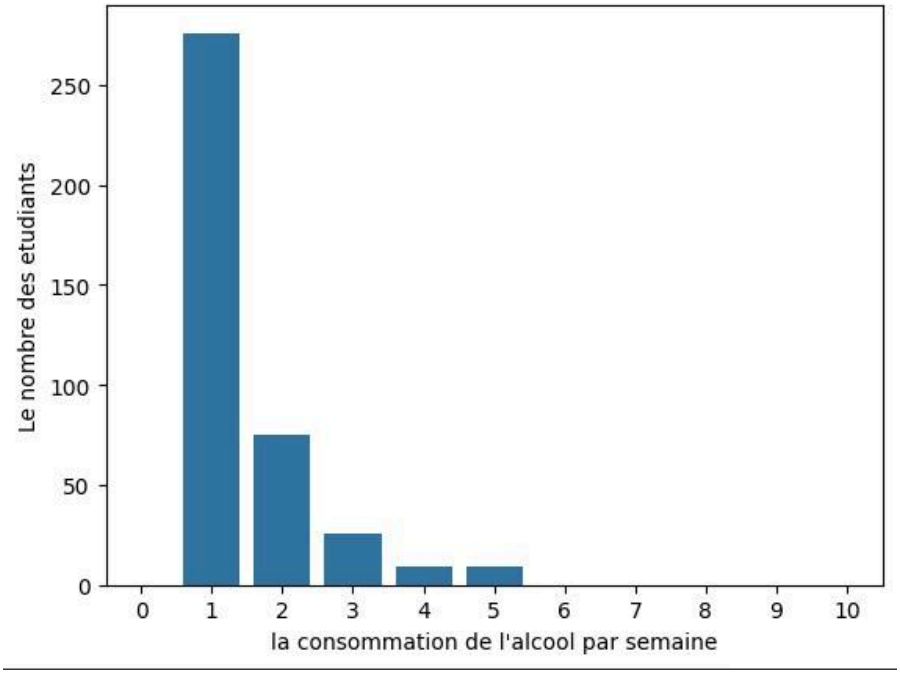
Le jeu de données comprend des informations sur les performances scolaires de 395 élèves.Les variables comprennent des caractéristiques telles que school, sex, age…

La moyenne d'âge des élèves est de 16.7 ans avec un écart-type de 1.27 ans.

La variable cible principale, "consommation d'alcool le week-end", a une moyenne de 2.3 verres et un écart type de 1.28.

Analyse des Graphiques

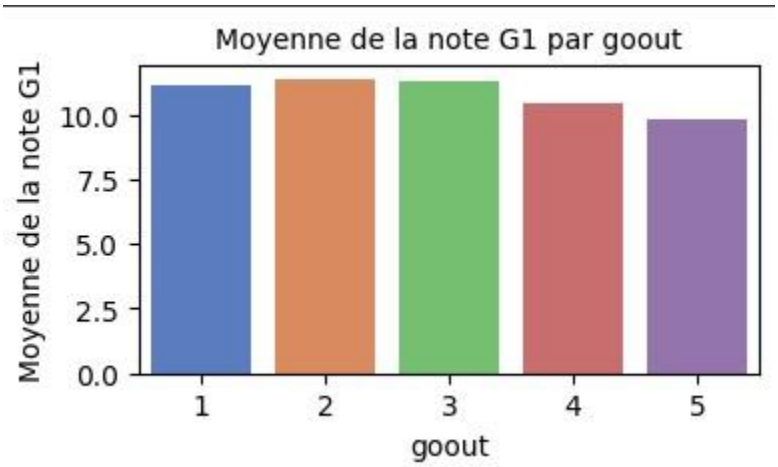
Nombre d'étudiants en fonction de la consommation d'alcool:



Le graphique illustre clairement que la consommation hebdomadaire d'alcool la plus fréquente parmi les étudiants est de consommer de l'alcool une fois par semaine.

Cela suggère que la majorité des étudiants ont tendance à limiter leur consommation à une fréquence hebdomadaire plutôt que de consommer de l'alcool plus souvent.

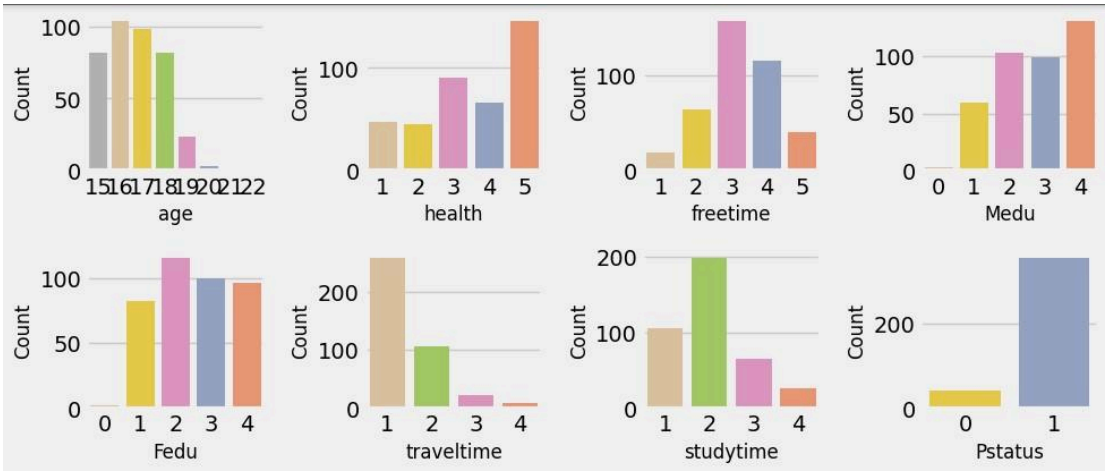
Moyenne de la note G1 en fonction de goout:



L'analyse montre que pour les étudiants ayant un nombre élevé de sorties (goout = 5), la moyenne de leur note G1 est inférieure par rapport à ceux ayant des niveaux de sorties inférieurs.

Cela suggère une corrélation négative entre le nombre de sorties et les performances académiques, indiquant que les étudiants qui sortent plus souvent ont tendance à avoir des résultats scolaires plus bas.

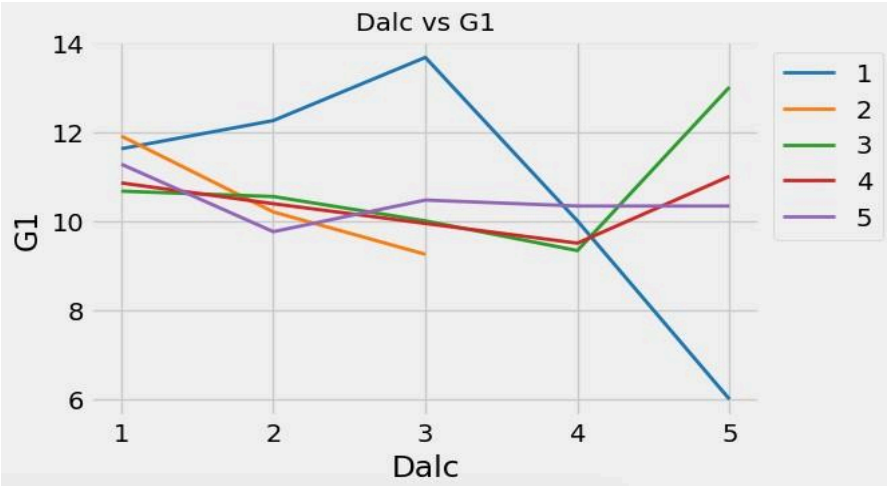
Facteurs divers:



En analysant différents facteurs tels que l'âge des étudiants, leur niveau de santé, le niveau d'éducation des parents, le temps de trajet pour se rendre à l'école, les heures d'étude par semaine et l'environnement familial, plusieurs observations significatives émergent.

Par exemple, il est notable que la plupart des élèves ont un temps de trajet court pour se rendre à l'école et étudient généralement pendant environ deux heures par semaine. De plus, la majorité des élèves semblent bénéficier d'un niveau de santé relativement bon et ont des parents avec un niveau d'éducation secondaire ou universitaire.

Notes des étudiants en fonction de l'utilisation de l'alcool pour chaque valeur de santé:



L'analyse du graphique de la note g1 en fonction de Dalc révèle deux tendances importantes :

- Pour les étudiants ayant une santé importante, on observe généralement que leurs notes restent stables ou légèrement fluctuent en fonction des changements dans la consommation d'alcool. Cela suggère que même si ces étudiants consomment plus ou moins d'alcool, leurs performances académiques ne varient pas de manière significative.

- En revanche, pour les étudiants ayant une santé basse, on constate que leurs notes ont tendance à diminuer à mesure que leur consommation d'alcool augmente. Cela indique une corrélation négative entre la santé des étudiants et leur performance académique en présence d'une consommation d'alcool accrue.

Prétraitement des données

Pour rendre notre jeu de données prêt à être utilisé pour l'analyse et la modélisation, nous avons entrepris des étapes de prétraitement :

-Suppression des Valeurs Doublantes et des Valeurs Manquantes

Deux fonctions ont été développées pour gérer les valeurs manquantes et les lignes dupliquées dans le dataset. Après leur application, il a été constaté que notre dataset ne contient ni lignes dupliquées ni valeurs manquantes, ce qui garantit la qualité et la complétude des données pour les analyses ultérieures.

-Conversion des Chaînes de Caractères en Données Numériques

Une fonction a été mise en place pour convertir les variables textuelles en variables numériques, facilitant ainsi l'analyse et l'utilisation des données dans les modèles de machine learning. Cette conversion est essentielle pour permettre aux algorithmes d'apprentissage automatique de traiter les données de manière efficace.

Approche Algorithmique

Bibliothèques utilisées

Nous avons utilisé les bibliothèques suivantes dans notre projet:

- **Scikit-learn** (sklearn): une bibliothèque d'apprentissage automatique en Python qui offre des outils simples et efficaces pour l'analyse prédictive.
  Fournit des outils pour le machine learning, y compris des modèles de régression, des méthodes de sélection d'hyperparamètres et des mesures d'évaluation.

  - train_test_split: Pour diviser les données en ensembles d'entraînement et de test.
  - GridSearchCV: Pour l'optimisation des hyperparamètres via la recherche en grille.
  - KNeighborsRegressor: Le modèle de régression K-Nearest Neighbors.
  - mean_squared_error: Pour calculer l'erreur quadratique moyenne (MSE).

- **Pandas**: une bibliothèque de manipulation et d'analyse des données en Python, utilisée pour lire, nettoyer et prétraiter nos données.
- **Numpy**: une bibliothèque Python qui ajoute un support pour les tableaux et les matrices multidimensionnels, utilisées pour effectuer des calculs numériques.
- **Matplotlib** : Une bibliothèque de visualisation de données en Python, utilisée pour créer des graphiques pour l'analyse exploratoire des données.

Régression Linéaire:

La régression linéaire est l'une des techniques les plus simples et les plus couramment utilisées en apprentissage supervisé pour modéliser la relation entre une variable dépendante continue et un ensemble de variables indépendantes. Dans le contexte de notre étude sur la prédiction des performances des élèves, nous avons opté sur l'utilisation de la régression linéaire en raison de sa simplicité et de son interprétabilité.

KNN (K-Nearest Neighbors):

KNN (K-Nearest Neighbors) est un algorithme d'apprentissage supervisé utilisé pour la régression et la classification. Il repose sur le principe selon lequel les points de données similaires tendent à être proches les uns des autres dans l'espace des caractéristiques. L'idée fondamentale de KNN est de trouver un nombre prédéfini de points de données les plus proches dans l'ensemble d'entraînement pour un point de données de test donné, puis de prédire la valeur de la variable cible en fonction de ses voisins les plus proches.

Dans le contexte de notre étude sur la prédiction des performances des élèves, nous avons choisi d'utiliser l'algorithme KNN pour sa capacité à capturer des relations complexes et non linéaires entre les caractéristiques des élèves et leurs notes finales.

SVM (Support Vector Machine):

SVM est un algorithme d'apprentissage supervisé utilisé pour la classification et la régression. Il vise à trouver l'hyperplan optimal qui sépare les données en deux classes ou qui prédit de manière optimale les valeurs numériques pour la régression. L'objectif principal de SVM est de maximiser la marge, c'est-à-dire la distance entre l'hyperplan de décision et les points de données les plus proches de chaque classe, tout en minimisant l'erreur de classification ou l'erreur de prédiction.

Dans notre étude sur la prédiction des performances des élèves, l'utilisation de SVM offre plusieurs avantages comme la gestion de données complexes, robustesse aux valeurs aberrantes.

Application à notre Projet :

Nous avons suivi les étapes suivantes pour appliquer les algorithmes choisis à notre projet :

1. **Prétraitement des Données** : Nous avons prétraité nos données en utilisant des techniques telles que l'encodage des variables catégorielles avec LabelEncoder et la séparation des données en variables prédictives (X) et la variable cible (y).

2. **Séparation des Données d'Entraînement et de Test :** Nous avons divisé nos données en un ensemble d'entraînement et un ensemble de test à l'aide de la fonction train_test_split de scikit-learn. Cela nous a permis d'évaluer les performances de notre modèle sur des données indépendantes.
3. **Entraînement du Modèle :** Nous avons instancié un objet de modèle de régression linéaire à l'aide de la classe LinearRegression de scikit-learn, puis nous avons ajusté le modèle aux données d'entraînement à l'aide de la méthode fit().
4. **Prédictions et Évaluation :** Nous avons utilisé le modèle entraîné pour faire des prédictions sur les données de test, puis nous avons évalué les performances du modèle en calculant le Mean Squared Error (MSE) à l'aide de la fonction mean_squared_error de scikit-learn.

Évaluation des Modèles

Nous avons évalué les trois modèles : la Régression Linéaire, le KNN (K-Nearest Neighbors) et le SVM (Support Vector Machine). Chaque modèle a été évalué en termes de temps d'entraînement moyen, temps de prédiction moyen, coefficient de détermination moyen (R^2), erreur quadratique moyenne négative (MSE) et erreur absolue moyenne négative (MAE).

Performances moyennes des modèles :

Régression Linéaire :

- **Temps de fit moyen** : 0.0103 secondes (± 0.0046)
- **Temps de score moyen** : 0.0059 secondes (± 0.0034)
- **R^2 moyen** : 0.7930 (± 0.0586)
- **Erreur quadratique moyenne négative (MSE)** : -4.1320 (± 1.5730)
- **Erreur absolue moyenne négative (MAE)** : -1.3257 (± 0.1325)

La régression linéaire s'est avérée rapide à l'entraînement et à la prédiction, avec un coefficient de détermination R^2 élevé. De plus, les erreurs quadratique et absolue moyennes négatives indiquent que le modèle surpasse les prédictions moyennes, bien que l'erreur puisse être légèrement sous-estimée.

KNN (K-Nearest Neighbors) :

- **Temps de fit moyen** : 0.0122 secondes (± 0.0032)
- **Temps de score moyen** : 0.0142 secondes (± 0.0033)
- **R^2 moyen** : 0.8008 (± 0.0510)
- **Erreur quadratique moyenne négative (MSE)** : -3.9210 (± 1.2335)
- **Erreur absolue moyenne négative (MAE)** : -1.3497 (± 0.1812)

Le modèle KNN a montré des performances similaires à la régression linéaire en termes de R^2, bien que légèrement plus lent à l'entraînement et à la prédiction. Les erreurs négatives indiquent également une

meilleure performance que les prédictions moyennes, bien que les prédictions aient tendance à sous-estimer légèrement les valeurs réelles.

SVM (Support Vector Machine) :

- **Temps de fit moyen** : 0.0213 secondes (± 0.0080)
- **Temps de score moyen** : 0.0095 secondes (± 0.0037)
- **R^2 moyen** : 0.7824 (± 0.0731)
- **Erreur quadratique moyenne négative (MSE)** : -4.5970 (± 2.0869)
- **Erreur absolue moyenne négative (MAE)** : -1.2702 (± 0.2154)

Le modèle SVM s'est avéré légèrement plus lent à l'entraînement, mais plus rapide que KNN à la prédiction. Bien que le R^2 soit légèrement inférieur aux deux autres modèles, les erreurs négatives indiquent également de bonnes performances prédictives.

En conclusion, tous les modèles ont montré des performances prometteuses, mais la régression linéaire semble être la plus rapide et offre des performances globalement cohérentes, tandis que KNN et SVM présentent des variations légères mais comparables dans leurs performances.

Conclusion

Ce projet avait pour objectif de développer et évaluer des modèles de régression pour prédire une variable cible à partir d'un ensemble de caractéristiques. Après une analyse exploratoire des données et un prétraitement, nous avons appliqué la Régression Linéaire, le KNN et le SVM. Tous les modèles ont montré des performances prometteuses, avec des scores R^2 élevés et des erreurs moyennes négatives. La régression linéaire était rapide et simple, tandis que le KNN et le SVM offraient des performances comparables.

En conclusion, ce projet a confirmé l'efficacité des modèles de régression pour la prédiction, ouvrant la voie à des applications plus larges dans l'analyse de données.