Rapport de Projet

Projet Machine learning

19 May 2024

Analyse de la Performance des Élèves du Secondaire en Mathématiques

Réalisé par:
Jidar Ahmed Amine
Khartouch Mouad
Khribech Hamza
Ankri Mohamed Khalil

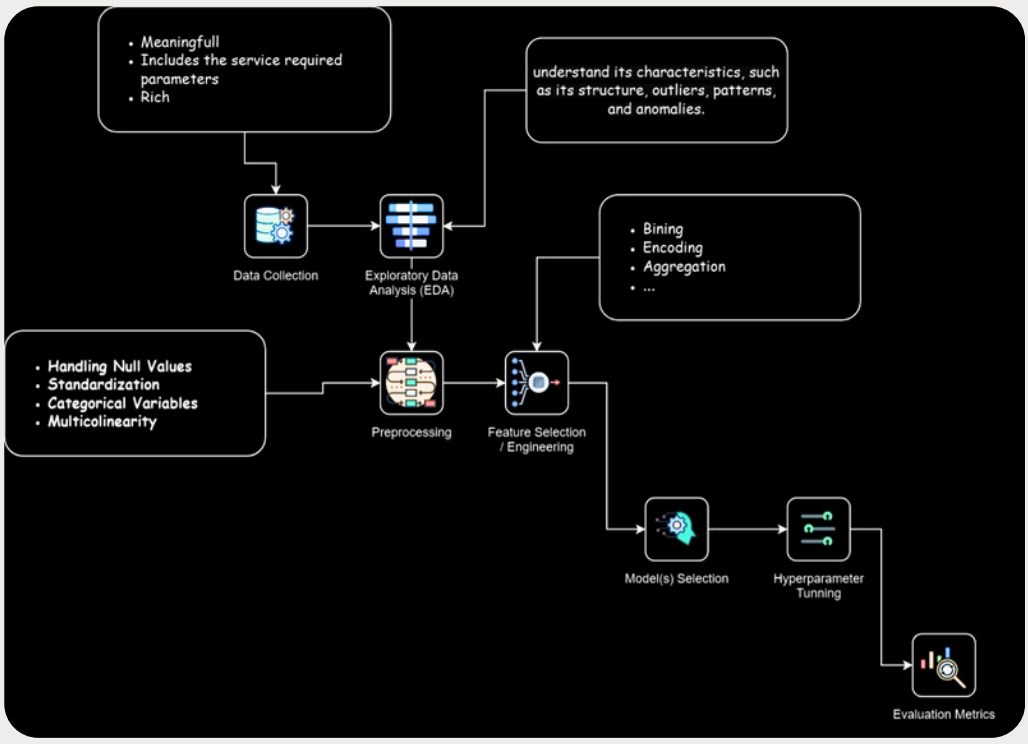
Encadré par:
Mr. Haja Zakaria

Sommaire

1 Définition de l'objectif du projet

L'objectif principal de ce projet est d'analyser les facteurs influençant les performances des élèves du secondaire en mathématiques et de prédire leurs scores à l'aide de modèles de machine learning. Les facteurs considérés incluent le genre, l'origine ethnique, le niveau d'éducation des parents, l'accès aux repas gratuits ou à prix réduit, et la participation à des cours de préparation aux tests. Ce projet vise à répondre aux questions de recherche suivantes :

- Quel est l'impact du niveau d'éducation des parents sur les performances en mathématiques des élèves ?
- Les élèves ayant suivi des cours de préparation aux tests obtiennent-ils de meilleurs résultats en mathématiques ?
- Les caractéristiques démographiques influencent-elles de manière significative les scores en mathématiques ?



2 Source des données

Le jeu de données utilisé dans ce projet contient des informations sur les performances des élèves de trois lycées aux États-Unis. Les données comprennent les scores des élèves en mathématiques, lecture et écriture, ainsi que des informations démographiques et académiques. Les colonnes du jeu de données sont les suivantes :

- **Genre** : Le genre de l'élève (masculin/féminin)
- **Race/ethnicité** : L'origine raciale ou ethnique de l'élève (Asiatique, Afro-Américain, Hispanique, etc.)
- **Niveau d'éducation des parents** : Le plus haut niveau d'éducation atteint par les parents ou tuteurs de l'élève
- **Repas** : Si l'élève reçoit des repas gratuits ou à prix réduit (oui/non)
- **Cours de préparation aux tests** : Si l'élève a suivi un cours de préparation aux tests (oui/non)
- **Score en mathématiques** : Le score de l'élève à un test standardisé de mathématiques
- **Score en lecture** : Le score de l'élève à un test standardisé de lecture
- **Score en écriture** : Le score de l'élève à un test standardisé d'écriture

https://www.kaggle.com/datasets/rkiattisak/student-performance-in-mathematics/data

3 Analyse Exploratoire des Données (EDA)

L'analyse exploratoire des données (EDA) vise à comprendre la distribution des données, à identifier les relations potentielles entre les variables, et à détecter les anomalies ou valeurs manquantes. Voici les principales étapes de notre EDA :

Distribution des variables

Analyse des distributions des scores en mathématiques, lecture et écriture.

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler , OneHotEncoder , MinMaxScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error

df = pd.read_csv('exams.csv')
df.head()
```

Python

	gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
0	female	group D	some college	standard	completed	59	70	78
1	male	group D	associate's degree	standard	none	96	93	87
2	female	group D	some college	free/reduced	none	57	76	77
3	male	group B	some college	free/reduced	none	70	70	63
4	female	group D	associate's degree	standard	none	83	85	86

Relations entre variables

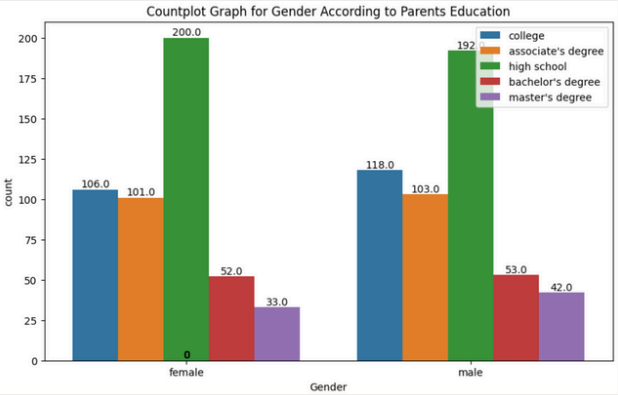
Analyse des corrélations entre les scores des tests et les caractéristiques démographiques.

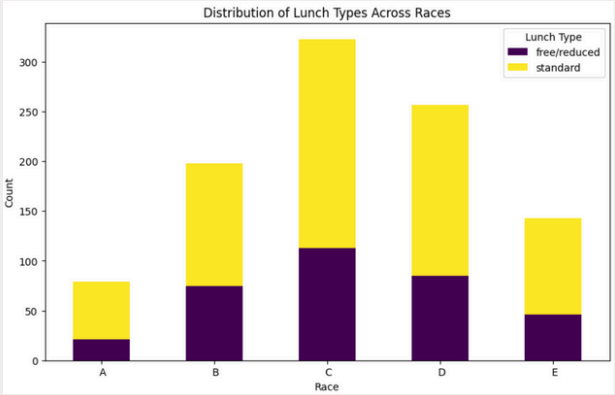
	math score	reading score	writing score
count	1000.000000	1000.000000	1000.000000
mean	67.810000	70.382000	69.140000
std	15.250196	14.107413	15.025917
min	15.000000	25.000000	15.000000
25%	58.000000	61.000000	59.000000
50%	68.000000	70.500000	70.000000
75%	79.250000	80.000000	80.000000
max	100.000000	100.000000	100.000000

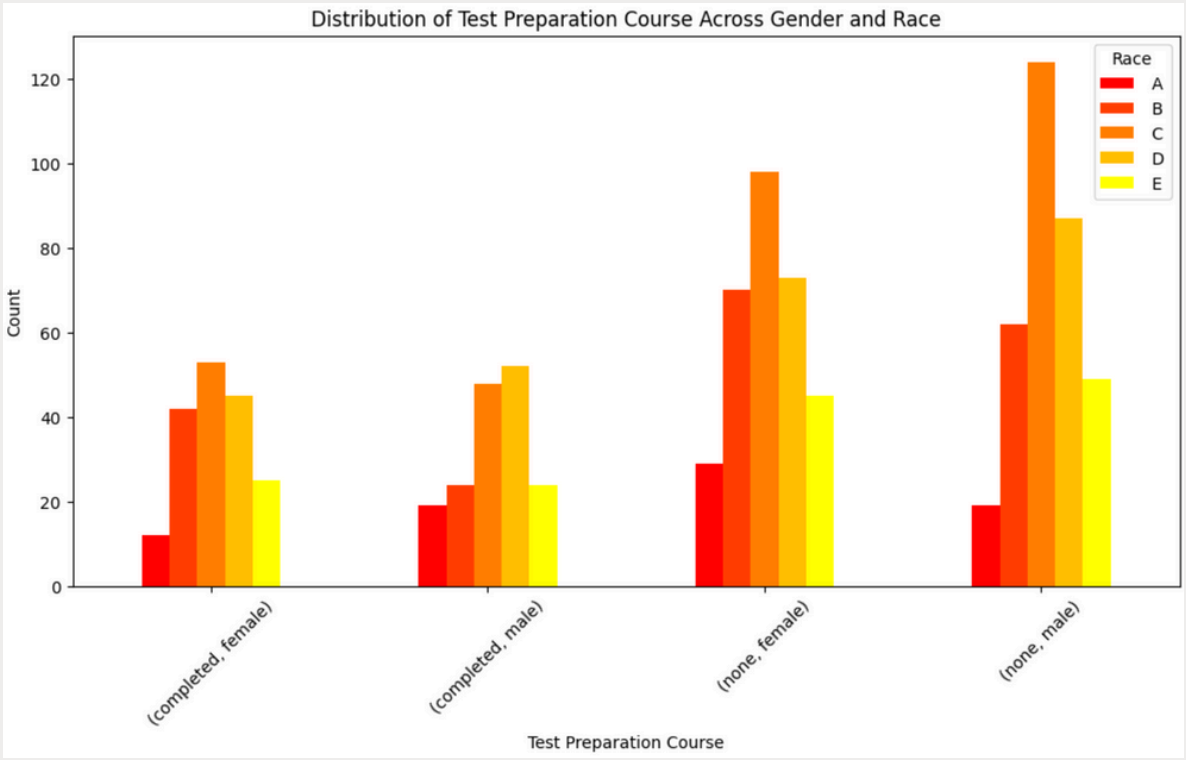
3 Analyse Exploratoire des Données (EDA)

Visualisation des données

Utilisation de graphiques pour représenter les distributions et les relations entre les variables.







4

Prétraitement des Données

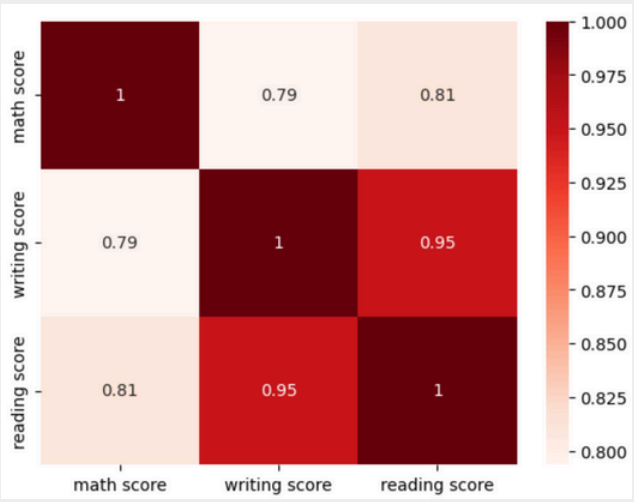
Objectifs du Prétraitement

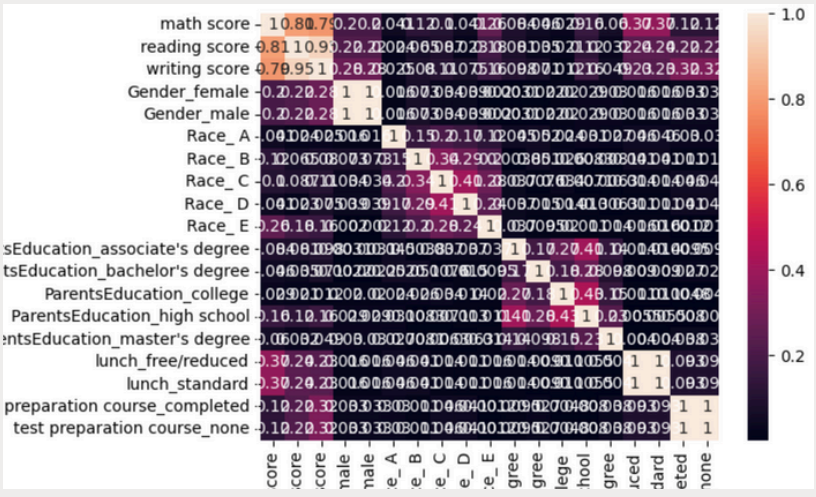
Le prétraitement des données est une étape cruciale pour préparer les données brutes en vue de leur utilisation dans des modèles de machine learning. Les étapes principales incluent :

- Gestion des valeurs manquantes : Remplacement ou suppression des données manquantes.
- Encodage des variables catégorielles : Transformation des variables catégorielles en variables numériques.
- Normalisation : Mise à l'échelle des données pour assurer une distribution uniforme.

Étapes du Prétraitement

1. Gestion des Valeurs Manquantes : Les données manquantes ont été remplacées par la moyenne (pour les scores) ou par le mode (pour les variables catégorielles).
2. Encodage des Variables : Les variables catégorielles telles que le genre, la race/ethnicité, et les repas ont été encodées en utilisant la méthode One-Hot Encoding.
3. Normalisation des Données : Les scores ont été normalisés pour avoir une moyenne de 0 et un écart-type de 1.
4. Matrice de Corrélation





5 Sélection des Modèles

Algorithmes Utilisés

Nous avons testé plusieurs algorithmes de machine learning pour prédire les scores en mathématiques :

- Régression Linéaire : Pour une première approche simple et interprétable.
- Arbre de Décision : Pour capturer les relations non linéaires entre les variables.
- Random Forest : Pour améliorer la précision en réduisant la variance des prédictions.
- Gradient Boosting : Pour obtenir des prédictions robustes en combinant plusieurs modèles faibles.

6 Tuning des Hyperparamètres

- Objectifs du Tuning :

Optimiser les performances des modèles en ajustant les hyperparamètres pour chaque algorithme. Nous avons utilisé la validation croisée pour évaluer les performances des modèles et sélectionner les meilleurs paramètres.

- Méthodes Utilisées

Méthodes Utilisées

1. Grid Search : Exploration exhaustive d'un espace de paramètres prédéfini.
2. Random Search : Exploration aléatoire pour une recherche plus efficace dans un espace de paramètres large.
3. Validation Croisée : Pour évaluer la robustesse des modèles avec les différents ensembles de validation.

7 Évaluation des Modèles

Métriques d'Évaluation

Les performances des modèles ont été évaluées à l'aide des métriques suivantes :

- Erreur Quadratique Moyenne (MSE) : Pour mesurer la précision des prédictions.
- R^2 (Coefficient de Détermination) : Pour évaluer la proportion de la variance expliquée par le modèle.
- MAE (Erreur Absolue Moyenne) : Pour comprendre l'erreur moyenne en termes absolus.

8 Conclusion

Ce projet d'analyse des performances des élèves du secondaire en mathématiques a permis d'explorer divers facteurs influençant leurs scores, notamment le genre, l'origine ethnique, le niveau d'éducation des parents, les repas scolaires et la participation à des cours de préparation aux tests. L'analyse exploratoire des données a révélé des relations significatives entre ces variables et les scores en mathématiques, soulignant l'importance de facteurs socio-économiques et éducatifs.

Les résultats de ce projet fournissent des informations précieuses pour les éducateurs et les décideurs politiques. Ils mettent en évidence la nécessité de soutenir les élèves issus de milieux défavorisés et de promouvoir des programmes de préparation aux tests pour améliorer les performances académiques.

En conclusion, ce projet démontre le potentiel des techniques de machine learning pour analyser et prédire les performances scolaires, offrant ainsi des outils puissants pour l'amélioration de l'éducation et l'égalité des chances.