Rapport Analyse de Performance Académique par Apprentissage Machine

1. Introduction et Objectifs

1.1 Contexte

Ce projet applique l'apprentissage automatique pour analyser les déterminants des performances scolaires d'élèves du secondaire aux États-Unis. Les données proviennent de Kaggle et contiennent 1 000 étudiants avec leurs scores académiques, informations démographiques, contexte familial et participation éventuelle à un cours de préparation au test.

1.2 Objectifs Spécifiques

Quatre axes d'analyse ont été étudiés :

1. Régression : prédire le score final global
2. Classification : prédire la participation au cours de préparation
3. Clustering : segmenter les profils d'étudiants
4. Méthodes d'ensemble : comparer avec les modèles individuels

2. Données et Préparation

2.1 Description du dataset

Le dataset contient 1 000 observations et 8 variables, aucune valeur manquante et aucun doublon.

2.2 Nature des variables

Variables catégorielles : gender, race/ethnicity, parental level of education, lunch, test preparation course
Variables numériques : math score, reading score, writing score

2.3 Prétraitement effectué

Encodage de variables catégorielles (One-Hot, Ordinal Encoding)
Standardisation pour modèles sensibles à l'échelle

Split train-test 80/20
Pipelines intégrés avec GridSearchCV et validation croisée

3. Méthodologie et Modèles

3.1 Pipeline général

Données brutes → Nettoyage & Encodage → Split train/test → Modélisation → Validation croisée → Analyse

3.2 Modèles implémentés

A. Classification
Objectif : prédire test preparation course
Modèles testés : SVM linéaire, Random Forest, Gradient Boosting, XGBoost, Bagging
Métriques : Accuracy, F1-score, Precision, Recall

B. Régression
Objectif : prédire le score final
Modèles testés : Linear Regression, Ridge, Lasso
Métriques : R^2, RMSE, MAE

C. Clustering
Algorithme utilisé : K-Means
Méthodes : courbe du coude + silhouette
Résultat optimal : K = 3

D. Méthodes d'ensemble
Bagging, Boosting et XGBoost comparés à des modèles individuels
Aucun stacking complet n'a été implémenté dans le notebook

4. Résultats et Interprétation

4.1 Classification

Meilleur modèle observé : SVM linéaire
Accuracy : 0.752
F1-score : 0.603
Analyse : performances modérées, peu de séparabilité entre les classes

4.2 Régression

Meilleur modèle observé : Lasso Regression
R^2 : 0.869
Bonne capacité de prédiction et généralisation

4.3 Clustering

Nombre de clusters optimal : 3
Score silhouette : 0.406
Interprétation
Cluster 0 : étudiants intermédiaires
Cluster 1 : élèves en difficulté
Cluster 2 : élèves performants

4.4 Méthodes d'ensemble

Gradient Boosting CV Accuracy ≈ 0.687 ± 0.016
Mais le SVM linéaire sur test surpasse les ensembles dans le notebook

5. Observations

5.1 Insights clés

Corrélations fortes lecture-écriture
Impact notable du niveau d'éducation parentale
Clustering cohérent mais faible séparation
Modèles d'ensemble ne surpassent pas systématiquement les modèles simples

5.2 Limitations

Jeu de données peu informative
Pas d'explicabilité avancée (SHAP, LIME)
Absence d'un vrai stacking classifier

5.3 Implications

Le dataset manque de variables prédictives riches
SVM illustre bien le fait qu'un modèle simple peut surpasser les ensembles

6. Conclusion

Analyse des performances étudiantes à travers une pipeline ML complete incluant preprocessing, validation et comparaison de modèles.

Résultats principaux

- Lasso Regression performant en régression
- SVM linéaire performant en classification
- K-Means identifie trois profils mais séparabilité modérée

Les ensembles n'apportent pas un gain significatif ici