Rapport Technique : Modélisation Prédictive

Projet : Prédiction de la Performance des Étudiants Marocains

Équipe : Squad 2

(Feature Engineering & Model Building)

10 Février 2026

Réalisé par : DOUAE MOUSSAOUI

Table des matières

Résumé

Ce rapport présente les travaux réalisés par le Squad 2 dans le cadre du projet de prédiction de la performance scolaire. Il détaille la méthodologie d'ingénierie des fonctionnalités (Feature Engineering), la sélection des variables, ainsi que l'entraînement et l'évaluation de modèles de Machine Learning (Régression Linéaire, Random Forest, XGBoost, MLP). Les résultats démontrent une capacité prédictive élevée avec un score R^2 supérieur à 0.87 pour les meilleurs modèles.

1 Introduction et Objectifs

L'objectif principal de cette phase du projet est de développer des modèles algorithmiques capables de prédire la note annuelle finale des étudiants du cycle secondaire qualifiant au Maroc. Cette prédiction vise à identifier les facteurs influençant la réussite scolaire et à fournir des outils d'aide à la décision pour les établissements.

À partir du jeu de données nettoyé et prétraité par l'équipe précédente (Squad 1), notre travail s'est concentré sur :

- L'ingénierie des fonctionnalités pour transformer les données brutes en indicateurs pertinents.
- La sélection des variables les plus corrélées avec la performance cible.
- L'entraînement et la comparaison rigoureuse de plusieurs modèles de régression.
- L'interprétation des performances et des résultats obtenus.

2 Ingénierie des Fonctionnalités (Feature Engineering)

Cette étape est cruciale pour adapter les données aux exigences des algorithmes d'apprentissage automatique et optimiser leur performance.

2.1 Sélection des Variables (Feature Selection)

Sur la base de l'analyse exploratoire (EDA) réalisée par le Squad 1, nous avons isolé les variables présentant la plus forte corrélation avec la performance cible (`moyenne_annuelle`), tout en considérant les recommandations de l'encadrant et en évitant la redondance. Les variables retenues sont :

- **Variables Académiques** : `moyenne_s1`, `moyenne_s2` (Notes des semestres précédents).
- **Habitudes d'Étude** : `heures_etude_jour`, `heures_etude_weekend`, `taux_assiduite`.
- **Démographie** : `age`.
- **Environnement socio-économique** : `zone` (Urbain/Rural), `soutien_familial` (Variable catégorielle).
- **Absences** : `absences_totales` (corrélation négative significative).
- **Distance École** : `distance_ecole_km`.

2.2 Prétraitement des Données (Preprocessing)

Afin d'harmoniser les échelles et de traiter les variables textuelles, un pipeline de transformation a été mis en place :

1. **Standardisation** : Les variables numériques ont été normalisées à l'aide du StandardScaler (moyenne centrée à 0, écart-type à 1). Cela empêche les variables à forte amplitude de biaiser le modèle.
2. **Encodage (One-Hot Encoding)** : Les variables catégorielles (`soutien_familial`, `zone`) ont été transformées en vecteurs binaires.
3. **Division du Dataset** : Les données ont été séparées en deux ensembles distincts : Entraînement (80%) et Test (20%).

3 Modélisation (Model Building) et Comparaison

Nous avons exploré et comparé quatre familles d'algorithmes pour ce problème de régression.

3.1 Modèles Testés et Justification

1. **Régression Linéaire** : Modèle de référence (baseline) simple et rapide.
2. **Random Forest Regressor** : Algorithme d'ensemble robuste, gérant bien les relations non-linéaires (200 arbres).
3. **XGBoost** : Algorithme de Gradient Boosting très performant, offrant une précision supérieure.
4. **MLP (Deep Learning)** : Réseau de neurones artificiels simple pour modéliser des relations complexes.

3.2 Métriques d'Évaluation

Pour évaluer les performances, nous avons utilisé :

— **RMSE (Root Mean Squared Error)** : $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

— **R^2 Score** : $1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

4 Résultats Actuels et Interprétation

4.1 Comparaison Quantitative des Modèles

Le tableau ci-dessous résume les métriques obtenues sur l'ensemble de test :

Modèle	**RMSE** (↓)	**R^2** (↑)
Régression Linéaire (Baseline)	0.0645	0.8786
Random Forest Regressor	0.0662	0.8720
XGBoost Regressor	0.0638	**0.8795**
MLP Regressor	0.0715	0.8650

TABLE 1 – Comparaison des performances des modèles sur l'ensemble de test.

4.2 Interprétation Claire des Performances

- **Performance Globale** : Les meilleurs modèles atteignent un score R^2 supérieur à 0.87.
- **Supériorité de XGBoost** : Il obtient le meilleur score R^2 et le RMSE le plus bas.
- **Performance de la Baseline** : La Régression Linéaire est très proche de XGBoost, indiquant une forte composante linéaire.
- **Impact des Modèles Complexes** : Random Forest et MLP n'apportent pas d'amélioration significative ici.

5 Problèmes Rencontrés et Limitations

- **Nature Synthétique des Données** : Les modèles peuvent performer de manière "trop parfaite".
- **Overfitting Potentiel du MLP** : Légère tendance au surapprentissage observée.
- **Manque de Données "Post-Bac"** : L'analyse se limite au baccalauréat.

6 Conclusion et Perspectives

Le travail mené par le Squad 2 a permis de développer un pipeline de modélisation prédictive robuste. L'utilisation combinée du Feature Engineering et de modèles comme XGBoost a porté ses fruits.

Pour les perspectives futures, nous recommandons :

- L'utilisation du modèle XGBoost sauvegardé (`best_model_student_prediction.pkl`).
- Le développement d'une interface utilisateur (Streamlit).
- L'exploration de techniques d'ingénierie plus avancées.
- La validation sur des données réelles.