Rapport Technique Machine Learning et Evaluation
Projet EduPredictors

Equipe Data Science

14 février 2026

Table des matières

1 Contexte et objectif

Ce document centralise et aligne les résultats des notebooks :

- `code/ML_student_performance_VF.ipynb` : entraînement, comparaison et optimisation de modèles.
- `code/evaluation.ipynb` : èvaluation opérationnelle et visualisation.

Objectif principal : fournir un rapport unique, cohérent et exploitable pour la partie modélisation et èvaluation de la performance scolaire.

2 Données et périmètre

2.1 Jeu de données

- Fichier : `dataset/Morocco_Student_Data_Cleaned.csv`
- Taille observée : 10 000 lignes, 268 colonnes
- Type : données à dominante socio-économique, comportementale et académique

2.2 Variable cible (pipeline ML)

Dans le notebook ML, la cible principale est : `moyenne_annuelle`.

2.3 Scénarios de modélisation

Scénario 1 (début d'année)

- Variables explicatives socio-économiques et comportementales
- Exclusion de `moyenne_s1` et `moyenne_s2` pour réduire le risque de fuite de cible

Scénario 2 (mi-année)

- Même base que le scénario 1
- Ajout de `moyenne_s1`

3 Pipeline de modélisation (ML_student_performance_VF.ipynb)

3.1 Préparation

- Split train/test : `test_size=0.2`, `random_state=42`
- Préprocessing via `ColumnTransformer`
- Numériques : `StandardScaler`
- Catégorielles : `OneHotEncoder(handle_unknown="ignore")`

3.2 Modèles comparés (scénario 1)

1. Régression Linéaire
2. Ridge (L2)
3. Lasso (L1)
4. Random Forest
5. XGBoost
6. MLP Regressor

3.3 Résultats de base (avant tuning)

TABLE 1 – Comparatif des modèles – Scénario 1 (test)

Modèle	R^2 test	RMSE	MAE
Ridge (L2)	0.1777	2.0152	1.7137
XGBoost	0.1773	2.0157	1.7139
Régression Linéaire	0.1769	2.0162	1.7147
MLP	0.1585	2.0386	1.7351
Random Forest	0.1478	2.0515	1.7384
Lasso (L1)	0.1430	2.0573	1.7408

3.4 Optimisation hyperparamètres

Une recherche aléatoire (`RandomizedSearchCV`) est appliquée sur :
- Random Forest
- XGBoost

Résultats après tuning :
- Random Forest (Tuned) : $R^2 = 0.1806$, RMSE=2.0117, MAE=1.7110
- XGBoost (Tuned) : $R^2 = 0.1827$, RMSE=2.0090, MAE=1.7103

Meilleur modèle global (Scénario 1) : `XGBoost (Tuned)`.

3.5 Analyse résidus et interprétabilité

Le notebook inclut :
- Nuage *résidus vs prédictions*
- Distribution des résidus + référence normale
- Nuage *réel vs prédit*
- Test de Shapiro-Wilk sur les résidus (p-value très faible)
- Importance des features (modèles arbres)
- SHAP (si bibliothèque disponible)

4 Scénario 2 : ajout de `moyenne_s1`

TABLE 2 – Résultats – Scénario 2 (avec signal académique)

Modèle	R^2 test	R^2 train	RMSE
Random Forest (S1)	0.9809	0.9915	0.3070
XGBoost (S1)	0.9804	0.9903	0.3108

L'ajout de `moyenne_s1` augmente fortement la performance prédictive.

5 Evaluation opérationnelle (evaluation.ipynb)

5.1 Etat actuel du notebook

Le notebook d'évaluation utilise actuellement :
- Features : `X = [age, performance_cible]`

— Cible : `y = performance_cible` (via `target_numeric`)
— Modèle : `RandomForestRegressor`

Cette configuration réutilise la cible dans les entrées, ce qui explique un RMSE quasi nul et un graphique très serré autour de la diagonale.

5.2 Résultat observé

— RMSE affiché : $\approx 8.94 \times 10^{-6}$

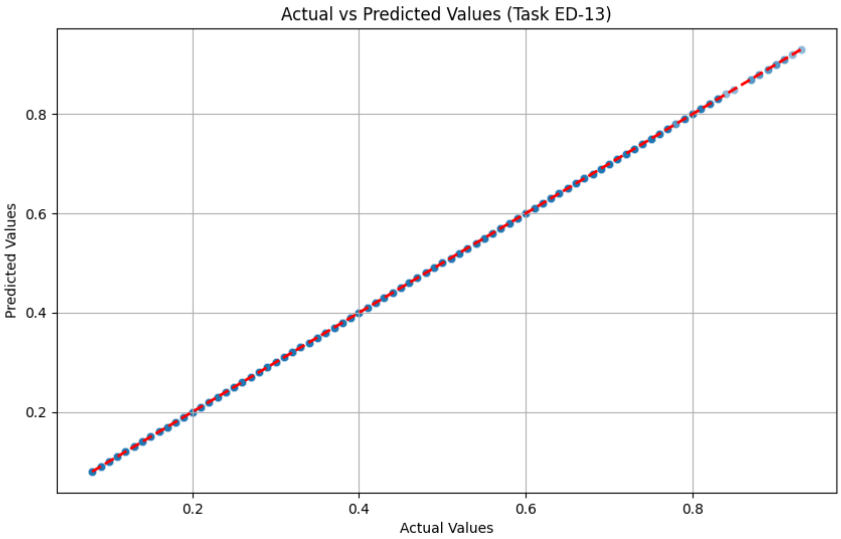

FIGURE 1 – Sortie visuelle de `evaluation.ipynb` (valeurs réelles vs prédites).

6 Artifacts et livrables

Fichier	Rôle
`code/best_model_student_prediction.pkl`	Modèle final sauvegardé
`code/model_features.pkl`	Liste des variables d'entrée du modèle
`code/model_evaluation_plot.png`	Figure d'évaluation
`rapport_par_partie/rapport_ML_DM.pdf`	Rapport PDF consolidé

7 Synthèse exécutive

— Sans information académique directe, la performance reste modérée ($R^2 \approx 0.18$).
— L'ajout de `moyenne_s1` permet d'atteindre $R^2 \approx 0.98$.
— Le meilleur modèle du pipeline principal (scénario 1) est `XGBoost (Tuned)`.
— L'évaluation actuelle est utile pour vérification d'exécution, mais doit être ajustée pour une comparaison méthodologique stricte.

Annexe : commandes de compilation

Depuis la racine du projet :

```
pdflatex -interaction=nonstopmode -halt-on-error \
  -output-directory rapport_par_partie \
  rapport_par_partie/rapport_ML_DM.tex
```