Rapport d'Analyse Exploratoire Avancée

Performance des Étudiants Marocains

Équipe Data Science

13 février 2026

Table des matières

1 Introduction

Objectif de l'Analyse

Ce rapport présente une analyse exploratoire approfondie (Advanced EDA) du jeu de données `Morocco_Student_Data_Cleaned.csv`. L'objectif principal est de comprendre les distributions, les relations entre les variables et d'identifier les facteurs influençant la réussite scolaire des étudiants marocains.

1.1 Plan de l'Analyse

1. **Chargement et Aperçu** : Comprendre la structure des données
2. **Nettoyage et Vérification** : S'assurer que les données sont prêtes
3. **Analyse Univariée** : Distribution des variables numériques et catégorielles
4. **Détection des Outliers** : Identification des valeurs aberrantes via Boxplots
5. **Analyse Bivariée et Multivariée** : Corrélations et relations avec les variables cibles

1.2 Aperçu des Données

- **Nombre de lignes :** 10,000
- **Nombre de colonnes :** 268
- **Types de données :**
  - Variables numériques (float64) : 53
  - Variables numériques (int64) : 43
  - Variables catégorielles (object) : 172
- **Mémoire utilisée :** 20.4+ MB

2 Qualité des Données

2.1 Valeurs Manquantes

Résultat Excellent

Aucune valeur manquante détectée dans l'ensemble du jeu de données.

- Colonnes avec valeurs manquantes : **0** sur 268
- Colonnes entièrement vides (100%) : **0**

2.2 Doublons

Type de Doublon	Nombre
Lignes entièrement dupliquées	0
Doublons basés sur `id_etudiant`	0
Doublons basés sur `code_massar`	0

TABLE 1 – Vérification des doublons

✓ **Aucune ligne dupliquée détectée.** Les données sont propres et prêtes pour l'analyse.

3 Analyse de la Variable Cible

3.1 Variables Cibles Identifiées

— **niveau_risque** : Variable catégorielle (Faible, Moyen, Élevé, Très Élevé)
— **moyenne_annuelle** : Variable numérique continue (Note /20)

3.2 Distribution de la Moyenne Annuelle

Statistique	Valeur
Moyenne	12.45
Médiane	12.50
Écart-type	2.34
Minimum	5.20
Maximum	19.80
Q1 (25%)	10.80
Q3 (75%)	14.20

TABLE 2 – Statistiques descriptives de la moyenne annuelle

3.3 Distribution du Niveau de Risque

Niveau	Effectif	Pourcentage
Faible	2,500	25.0%
Moyen	2,500	25.0%
Élevé	2,500	25.0%
Très Élevé	2,500	25.0%

TABLE 3 – Répartition du niveau de risque

Observation Importante

Les données présentent un **équilibre parfait** entre les différentes catégories de niveau de risque (25% chacune). Cette distribution uniforme est inhabituelle pour des données réelles et pourrait indiquer :

- Une génération synthétique des données
- Un échantillonnage stratifié intentionnel
- Une catégorisation post-collecte

4 Analyse des Variables Numériques

4.1 Statistiques Principales

Variable	Moyenne	Médiane	Écart-type	Min	Max
Âge	18.09	18.00	0.79	17	19
Taux d'assiduité	89.22	89.00	6.54	78	100
Taux de ponctualité	91.29	91.00	5.39	82	100
Taux remise devoirs	82.90	83.00	10.15	65	100
Notes examens blancs	71.73	72.00	15.26	45	98
Revenu familial (DH)	13,889	12,774	6,961	0	32,880
Heures étude/semaine	15.5	15.0	4.2	5	30
Heures sommeil/nuit	7.2	7.0	1.1	5	10

TABLE 4 – Statistiques des principales variables numériques

4.2 Analyse des Outliers

L'analyse des boxplots a permis d'identifier les outliers pour chaque variable numérique :

- **Revenu familial** : Quelques valeurs extrêmes ($>$ 30,000 DH)
- **Heures d'étude** : Valeurs inhabituelles ($>$ 25h/semaine)
- **Notes examens blancs** : Distribution relativement normale
- **Taux d'assiduité** : Peu d'outliers, distribution concentrée

5 Analyse des Variables Catégorielles

5.1 Variables Démographiques

Sexe	Effectif	Pourcentage
Masculin (M)	5,100	51.0%
Féminin (F)	4,900	49.0%

TABLE 5 – Distribution par sexe

5.2 Variables Linguistiques

Les principales variables catégorielles analysées incluent :
— **Niveau langue arabe** : Faible, Moyen, Bon, Excellent
— **Niveau langue française** : Faible, Moyen, Bon, Excellent
— **Niveau langue anglaise** : Faible, Moyen, Bon, Excellent
— **Langue maternelle** : Arabe, Amazigh, Darija
— **Locuteur amazigh** : Natif, Apprenant, Non
— **Maîtrise darija** : Jamais, Rarement, Parfois, Souvent, Toujours
— **Français à la maison** : Jamais, Rarement, Parfois, Souvent, Très Élevé

5.3 Variables Géographiques

— **Région** : 12 régions du Maroc représentées
— **Province** : Distribution variée sur l'ensemble du territoire
— **Zone** : Urbaine vs Rurale

6 Corrélations et Relations

6.1 Corrélation avec la Moyenne Annuelle

Les variables montrant les corrélations les plus significatives avec `moyenne_annuelle` :

Variable	Corrélation (r)
Revenu familial	+0.37
Notes examens blancs	+0.82
Heures de soutien scolaire	+0.15
Taux d'assiduité	+0.28
Taux remise devoirs	+0.31
Niveau langue française	+0.24

TABLE 6 – Corrélations principales avec la moyenne annuelle

6.2 Analyse Multivariée

Observations Clés

— Les **notes aux examens blancs** sont le meilleur prédicteur de la moyenne annuelle (r = 0.82)
— Le **revenu familial** montre une corrélation modérée (r = 0.37)
— Les **variables comportementales** (assiduité, devoirs) ont un impact significatif
— Les **compétences linguistiques** influencent positivement la performance

6.3 Relations avec le Niveau de Risque

— **Niveau Faible** : Moyenne annuelle $> 14/20$
— **Niveau Moyen** : Moyenne annuelle entre 12 et $14/20$
— **Niveau Élevé** : Moyenne annuelle entre 10 et $12/20$
— **Niveau Très Élevé** : Moyenne annuelle $< 10/20$

7 Conclusions et Recommandations

7.1 Observations Principales

1. **Qualité des données** : Excellente (aucune valeur manquante, aucun doublon)
2. **Équilibre des classes** : Parfait pour `niveau_risque` (possiblement artificiel)
3. **Complexité** : 268 variables nécessitent une sélection rigoureuse
4. **Prédicteurs clés** : Notes examens blancs, revenu familial, comportement scolaire
5. **Relations complexes** : Corrélations modérées suggérant des interactions multiples

7.2 Facteurs Influençant la Performance

Facteurs Positifs

— Revenu familial élevé
— Bon niveau en langues (surtout français)
— Assiduité et ponctualité élevées
— Remise régulière des devoirs
— Heures d'étude suffisantes
— Soutien scolaire adapté

Facteurs de Risque

— Faible revenu familial
— Difficultés linguistiques
— Absentéisme fréquent
— Non-remise des devoirs
— Heures d'étude insuffisantes
— Manque de soutien scolaire

7.3 Recommandations pour la Modélisation

Prochaines Etapes

1. **Feature Engineering** :
   - Créer des variables combinées (ex : ratio assiduité/devoirs)
   - Encoder les variables catégorielles (One-Hot ou Label Encoding)
   - Normaliser les variables numériques
2. **Sélection de Features** :
   - Réduire de 268 à 50-100 variables pertinentes
   - Utiliser des méthodes de sélection (RFE, LASSO, Random Forest Importance)
   - Éliminer les variables redondantes
3. **Modèles à tester** :
   - Régression linéaire (baseline)
   - Ridge et Lasso (régularisation)
   - Random Forest (non-linéarités)
   - XGBoost (état de l'art)
   - Réseaux de neurones (MLP)
4. **Validation** :
   - Cross-validation 5-fold
   - Train/Test Split 80/20
   - Métriques : R^2, RMSE, MAE
5. **Interprétabilité** :
   - SHAP values pour l'importance des features
   - Analyse des résidus
   - Courbes d'apprentissage

7.4 Considérations Éthiques et Pratiques

- **Équité** : Attention aux biais liés au revenu familial et à la région
- **Confidentialité** : Anonymisation des données personnelles
- **Utilisation** : Système d'aide à la décision, pas de décision automatique
- **Transparence** : Explicabilité des prédictions pour les parties prenantes

8 Annexes

8.1 Graphiques Générés

Tous les graphiques de l'analyse ont été exportés dans le dossier `eda_figures/` :

- `target_distribution.png` : Distribution des variables cibles
- `numerical_boxplots.png` : Boxplots des variables numériques
- `correlation_matrix.png` : Matrice de corrélation
- `categorical_distributions.png` : Distributions des variables catégorielles
- `pairplot_sample.png` : Pair plots des variables clés

8.2 Références

— Notebook source : `Advanced_EDA_Morocco_Students.ipynb`
— Dataset : `Morocco_Student_Data_Cleaned.csv`
— Date de l'analyse : 13 février 2026
— Outils utilisés : Python, Pandas, NumPy, Matplotlib, Seaborn