Rapport de Création du Data Pool

Prédiction de la Performance des Étudiants Marocains

Lycées Qualifiants - Tous les Niveaux (Tronc Commun, 1BAC, 2BAC)

Projet : Prédiction de la Performance des Étudiants
Date : 5 Février 2026

Résumé

Ce rapport documente le processus de création d'un ensemble de données synthétiques complet pour la prédiction de la performance des étudiants marocains du cycle secondaire qualifiant. Le dataset généré contient **10 000 enregistrements** d'étudiants avec **286 variables (features)** couvrant tous les aspects pertinents pour l'analyse et la prédiction de la réussite scolaire.

Table des matières

1 Introduction

1.1 Contexte du Projet

Le projet de prédiction de la performance des étudiants vise à développer des modèles de machine learning capables d'identifier les facteurs influençant la réussite scolaire des lycéens marocains et de prédire leur probabilité de réussite au baccalauréat.

1.2 Objectif du Data Pool

La création de ce data pool répond à plusieurs objectifs :

— Fournir une base de données réaliste et représentative de la population estudiantine marocaine

— Inclure une diversité de profils socio-économiques et académiques

— Couvrir toutes les régions du Maroc

— Intégrer des variables prédictives pertinentes basées sur la littérature scientifique

2 Description du Dataset

2.1 Caractéristiques Générales

Caractéristique	**Valeur**
Nombre d'enregistrements	10 000 étudiants
Nombre de variables	286 features
Format du fichier	CSV (UTF-8)
Taille du fichier	∼17 MB
Langue des données	Français
Couverture géographique	12 régions du Maroc

TABLE 1 – Caractéristiques générales du dataset

2.2 Catégories de Variables

Le dataset est organisé en plusieurs catégories thématiques :

2.2.1 1. Informations Personnelles et Démographiques (20 variables)

— Identifiant étudiant (id_etudiant)

— Prénom, Nom, Nom complet

— Sexe (M/F)

— Date de naissance, Âge

— Code Massar (identifiant national)

— Région, Province, Commune

- Zone (Urbain/Semi-Urbain/Rural)
- Adresse, Code postal
- Téléphone, Email

2.2.2 2. Informations Scolaires (15 variables)

- Nom de l'établissement
- Type d'établissement (Lycée Qualifiant)
- Secteur (Public/Privé)
- Académie régionale
- Direction provinciale
- Niveau (Tronc Commun / 1BAC / 2BAC)
- Filière et Spécialité
- Classe
- Année d'inscription et année scolaire

2.2.3 3. Informations Familiales (25 variables)

- Statut des parents (vivant/décédé)
- Informations sur le père : prénom, nom, niveau d'éducation, profession, secteur d'activité, revenu
- Informations sur la mère : prénom, nom, niveau d'éducation, profession, secteur d'activité, revenu
- Information sur le tuteur
- Revenu familial total et source de revenu
- Nombre de frères et surs, rang dans la fratrie
- Statut parental (Mariés/Divorcés/Veuf)

2.2.4 4. Conditions Socio-économiques (20 variables)

- Type de logement (Villa/Appartement/Maison/Maison Traditionnelle)
- Statut de propriété (Propriétaire/Locataire)
- Nombre de pièces
- Accès aux services (électricité, eau, internet)
- Équipements disponibles (chambre personnelle, bureau, ordinateur, laptop, tablette, smartphone)
- Livres à la maison
- Distance à l'école et moyen de transport
- Bourses et aides (Tayssir, bourses sociales, bourses d'excellence)

2.2.5 5. Informations de Santé (15 variables)

— Assurance maladie et type
— Maladies chroniques
— Handicap
— Port de lunettes, problèmes auditifs, allergies
— État de santé général
— IMC, taille, poids

2.2.6 6. Performance Académique (50+ variables)

Notes pour chaque matière (Semestre 1, Semestre 2, Moyenne annuelle) :
— Arabe
— Français
— Anglais
— Mathématiques
— Physique-Chimie
— Sciences de la Vie et de la Terre (SVT)
— Histoire-Géographie
— Éducation Islamique
— Philosophie
— Éducation Physique
— Informatique
— Économie, Comptabilité, Gestion (pour les filières économiques)

Indicateurs globaux :
— Moyenne générale (S1, S2, Annuelle)
— Rang dans la classe
— Effectif de la classe
— Absences (totales, justifiées, non justifiées)
— Retards
— Avertissements et sanctions
— Comportement

2.2.7 7. Habitudes d'Étude (25 variables)

— Heures d'étude par jour et weekend
— Cours particuliers et matières
— Coût du soutien scolaire
— Lieu et moment préféré d'étude
— Étude en groupe
— Utilisation des ressources en ligne
— Taux de remise des devoirs
— Participation en classe
— Prise de notes
— Utilisation de la bibliothèque

2.2.8 8. Activités Parascolaires (20 variables)

- Activités sportives et type de sport
- Activités artistiques
- Clubs et associations
- Bénévolat
- Participation aux événements scolaires
- Membre du conseil des élèves

2.2.9 9. Orientation et Aspirations (15 variables)

- Aspiration de carrière
- Université souhaitée
- Domaine d'étude souhaité
- Projet d'études à l'étranger
- Pays cible
- Connaissance de l'orientation
- Séances d'orientation suivies

2.2.10 10. Facteurs Psychologiques (20 variables)

- Niveau de motivation
- Confiance en soi
- Niveau de stress et d'anxiété
- Anxiété aux examens
- Satisfaction scolaire
- Relations avec les pairs
- Soutien familial
- Implication parentale
- Attentes et pression parentales

2.2.11 11. Mode de Vie (15 variables)

- Petit déjeuner et repas par jour
- Heures de sommeil (semaine/weekend)
- Heure de coucher et de réveil
- Activité physique
- Temps d'écran
- Réseaux sociaux
- Jeux vidéo
- Lecture
- Consommation de caféine
- Travail à temps partiel

2.2.12 12. Compétences et Aptitudes (20 variables)

- Efficacité d'auto-apprentissage
- Capacité à fixer des objectifs
- Gestion du temps
- Organisation
- Résolution de problèmes
- Pensée critique
- Communication
- Style d'apprentissage (Visuel/Auditif/Kinesthésique)
- Intelligence dominante
- Créativité, Adaptabilité, Résilience

2.2.13 13. Compétences Linguistiques (10 variables)

- Niveau en arabe (C2-A1)
- Niveau en français (C1-A2)
- Niveau en anglais (B2-A1)
- Niveau en espagnol
- Niveau en allemand
- Autres langues
- Locuteur amazigh

3 Couverture Géographique

Le dataset couvre les 12 régions administratives du Maroc :

1. Rabat-Salé-Kénitra
2. Casablanca-Settat
3. Fès-Meknès
4. Marrakech-Safi
5. Tanger-Tétouan-Al Hoceima
6. Souss-Massa
7. Oriental
8. Beni Mellal-Khenifra
9. Draa-Tafilalet
10. Laayoune-Sakia El Hamra
11. Guelmim-Oued Noun
12. Dakhla-Oued Ed-Dahab

Filière	Spécialités
Sciences Expérimentales	Sciences Physiques (PC), SVT
Sciences Mathématiques	Maths A (SMA), Maths B (SMB)
Sciences Économiques	Gestion Comptable (SGC), Sciences Économiques (SE)
Lettres et Sciences Humaines	Lettres, Sciences Humaines
Sciences et Technologies	Sciences Mécaniques (STM), Sciences Électriques (STE)

TABLE 2 – Filières et spécialités du baccalauréat couvertes

4 Filières Couvertes

5 Méthodologie de Génération

5.1 Approche Utilisée

Les données ont été générées de manière synthétique en utilisant Python avec les principes suivants :

- **Réalisme** : Les valeurs sont basées sur des distributions réalistes correspondant au contexte marocain
- **Cohérence** : Les corrélations logiques entre variables sont respectées (ex : revenu familial et équipements disponibles)
- **Diversité** : Large éventail de profils socio-économiques et académiques
- **Représentativité** : Distribution proportionnelle entre zones urbaines, semi-urbaines et rurales

6 Analyse et Nettoyage des Données

Avant de procéder à l'exploration des données, une phase initiale d'analyse et de nettoyage a été réalisée pour garantir la qualité des données.

6.1 Analyse Initiale

L'examen initial du dataset a révélé les caractéristiques suivantes :

- **Dimensions** : 10 000 enregistrements et 286 variables.
- **Types de données** : 65 variables flottantes (float64), 48 entiers (int64) et 173 objets (object).
- **Valeurs manquantes** : Identification de colonnes entièrement vides (100% manquantes) telles que `type_handicap, economie_s1, remarques`, et d'autres partiellement renseignées comme `etablissement_precedent` (5925 manquantes) et `annees_redoublees` (4537 manquantes).

6.2 Actions de Nettoyage

Les étapes suivantes ont été appliquées pour produire le fichier final : `Morocco_Student_Data_Cleaned.csv`

1. **Suppression des colonnes vides** : Retrait des 17 colonnes ne contenant aucune donnée.
2. **Traitement des valeurs manquantes** : Imputation des années de redoublement par 0 et des activités par "Aucun".
3. **Formatage des données** : Standardisation du format des dates de naissance et de collecte.

7 Utilisation du Dataset

7.1 Applications Possibles

— Développement de modèles de prédiction de réussite scolaire
— Identification des facteurs de risque d'échec
— Analyse des inégalités éducatives
— Conception de systèmes d'alerte précoce
— Études sur l'impact des facteurs socio-économiques

7.2 Variable Cible

La variable cible à prédire est **moyenne_annuelle** : la moyenne générale de l'étudiant pour l'année scolaire en cours. Il s'agit d'un problème de **régression**.

7.3 Modèles Recommandés

— **Régression Linéaire** : Modèle de base pour établir une référence
— **Random Forest Regressor** : Capture les relations non-linéaires
— **XGBoost / Gradient Boosting** : Performance élevée pour les données tabulaires
— **Deep Learning (MLP)** : Réseaux de neurones pour les relations complexes

8 Conclusion

Ce data pool constitue une ressource complète et réaliste pour le développement de modèles de prédiction de la performance des étudiants marocains. Avec ses 10 000 enregistrements et 286 variables (269 après nettoyage), il offre une base solide pour l'analyse exploratoire, le feature engineering, et l'entraînement de modèles de machine learning.

Fichiers générés :
`Morocco_Student_Data_Pool.csv` (Données brutes)
`Morocco_Student_Data_Cleaned.csv` (Données nettoyées)
`analyze_data.ipynb` (Notebook d'analyse)
`clean_data.ipynb` (Notebook de nettoyage)
Date de création : 5 Février 2026
Encodage : UTF-8