

Rapport de Creation du Data Pool

Prediction de la Performance des Etudiants Marocains

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

Projet : Prediction de la Performance des Etudiants

Date : 5 Fevrier 2026

Résumé

Ce rapport documente le processus de creation d'un ensemble de donnees synthetiques complet pour la prediction de la performance des etudiants marocains du cycle secondaire qualifiant. Le dataset genere contient **10 000 enregistrements** d'etudiants avec plus de **250 variables (features)** couvrant tous les aspects pertinents pour l'analyse et la prediction de la reussite scolaire.

Table des matières

1	Introduction	3
1.1	Contexte du Projet	3
1.2	Objectif du Data Pool	3
2	Description du Dataset	3
2.1	Caracteristiques Generales	3
2.2	Categories de Variables	3
2.2.1	1. Informations Personnelles et Demographiques (20 variables)	3
2.2.2	2. Informations Scolaires (15 variables)	4
2.2.3	3. Informations Familiales (25 variables)	4
2.2.4	4. Conditions Socio-economiques (20 variables)	4
2.2.5	5. Informations de Sante (15 variables)	5
2.2.6	6. Performance Academique (50+ variables)	5
2.2.7	7. Habitudes d'Etude (25 variables)	5
2.2.8	8. Activites Parascolaires (20 variables)	6
2.2.9	9. Orientation et Aspirations (15 variables)	6
2.2.10	10. Facteurs Psychologiques (20 variables)	6
2.2.11	11. Mode de Vie (15 variables)	6
2.2.12	12. Competences et Aptitudes (20 variables)	7
2.2.13	13. Competences Linguistiques (10 variables)	7
3	Couverture Geographique	7
4	Filières Couvertes	8
5	Methodologie de Generation	8
5.1	Approche Utilisee	8
6	Analyse et Nettoyage des Données	8
6.1	Analyse Initiale	8
6.2	Actions de Nettoyage	8
7	Utilisation du Dataset	9
7.1	Applications Possibles	9
7.2	Variable Cible	9
7.3	Modeles Recommandes	9
8	Conclusion	9

1 Introduction

1.1 Contexte du Projet

Le projet de prediction de la performance des etudiants 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 étudiante 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	250+ 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 Demographiques (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
- Telephone, Email

2.2.2 2. Informations Scolaires (15 variables)

- Nom de l'établissement
- Type d'établissement (Lycee Qualifiant)
- Secteur (Public/Prive)
- Academie regionale
- Direction provinciale
- Niveau (Tronc Commun / 1BAC / 2BAC)
- Filiere et Specialite
- Classe
- Annee d'inscription et annee scolaire

2.2.3 3. Informations Familiales (25 variables)

- Statut des parents (vivant/decede)
- Informations sur le pere : prenom, nom, niveau d'éducation, profession, secteur d'activite, revenu
- Informations sur la mere : prenom, nom, niveau d'éducation, profession, secteur d'activite, revenu
- Information sur le tuteur
- Revenu familial total et source de revenu
- Nombre de freres et soeurs, rang dans la fratrie
- Statut parental (Maries/Divorces/Veuf)

2.2.4 4. Conditions Socio-economiques (20 variables)

- Type de logement (Villa/Appartement/Maison/Maison Traditionnelle)
- Statut de propriete (Proprietaire/Locataire)
- Nombre de pieces
- Acces aux services (electricite, eau, internet)
- Equipements disponibles (chambre personnelle, bureau, ordinateur, laptop, tablette, smartphone)
- Livres a la maison
- Distance a l'école et moyen de transport
- Bourses et aides (Tayssir, bourses sociales, bourses d'excellence)

2.2.5 5. Informations de Sante (15 variables)

- Assurance maladie et type
- Maladies chroniques
- Handicap
- Port de lunettes, problemes auditifs, allergies
- Etat de sante general
- IMC, taille, poids

2.2.6 6. Performance Academique (50+ variables)

Notes pour chaque matiere (Semestre 1, Semestre 2, Moyenne annuelle) :

- Arabe
- Francais
- Anglais
- Mathematiques
- Physique-Chimie
- Sciences de la Vie et de la Terre (SVT)
- Histoire-Geographie
- Education Islamique
- Philosophie
- Education Physique
- Informatique
- Economie, Comptabilite, Gestion (pour les filieres economiques)

Indicateurs globaux :

- Moyenne generale (S1, S2, Annuelle)
- Rang dans la classe
- Effectif de la classe
- Absences (totales, justifiees, non justifiees)
- Retards
- Avertissements et sanctions
- Comportement

2.2.7 7. Habitudes d'Etude (25 variables)

- Heures d'etude par jour et weekend
- Cours particuliers et matieres
- Cout du soutien scolaire
- Lieu et moment prefere d'etude
- Etude en groupe
- Utilisation des ressources en ligne
- Taux de remise des devoirs
- Participation en classe
- Prise de notes
- Utilisation de la bibliotheque

2.2.8 8. Activites Parascolaires (20 variables)

- Activites sportives et type de sport
- Activites artistiques
- Clubs et associations
- Benevolat
- Participation aux evenements scolaires
- Membre du conseil des eleves

2.2.9 9. Orientation et Aspirations (15 variables)

- Aspiration de carriere
- Universite souhaitee
- Domaine d'etude souhaite
- Projet d'etudes a l'étranger
- Pays cible
- Connaissance de l'orientation
- Seances 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 dejeuner et repas par jour
- Heures de sommeil (semaine/weekend)
- Heure de coucher et de reveil
- Activité physique
- Temps d'écran
- Réseaux sociaux
- Jeux vidéo
- Lecture
- Consommation de cafeine
- Travail à temps partiel

2.2.12 12. Competences et Aptitudes (20 variables)

- Efficacite d'auto-apprentissage
- Capacite a fixer des objectifs
- Gestion du temps
- Organisation
- Resolution de problemes
- Pensee critique
- Communication
- Style d'apprentissage (Visuel/Auditif/Kinesthesique)
- Intelligence dominante
- Creativite, Adaptabilite, Resilience

2.2.13 13. Competences Linguistiques (10 variables)

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

3 Couverture Geographique

Le dataset couvre les 12 regions administratives du Maroc :

1. Rabat-Sale-Kenitra
2. Casablanca-Settat
3. Fes-Meknes
4. Marrakech-Safi
5. Tanger-Tetouan-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

Filiere	Specialites
Sciences Experimentales	Sciences Physiques (PC), SVT
Sciences Mathematiques	Maths A (SMA), Maths B (SMB)
Sciences Economiques	Gestion Comptable (SGC), Sciences Economiques (SE)
Lettres et Sciences Humaines	Lettres, Sciences Humaines
Sciences et Technologies	Sciences Mecaniques (STM), Sciences Electriques (STE)

TABLE 2 – Filières et spécialités du baccalaureat couvertes

4 Filières Couvertes

5 Methodologie de Generation

5.1 Approche Utilisée

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

- **Realisme** : 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 `établissement_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

- Developpement de modeles de prediction de reussite scolaire
- Identification des facteurs de risque d'echeec
- Analyse des inegalites educatives
- Conception de systemes d'alerte precoce
- Etudes sur l'impact des facteurs socio-economiques

7.2 Variable Cible

La variable cible a predire est **moyenne_annuelle** : le moyenne generale de l'étudiant pour l'année scolaire en cours. Il s'agit d'un probleme de **regression**.

7.3 Modeles Recommandes

- **Regression Lineaire** : Modele 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 plus de 250 variables, 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 creation : 5 Février 2026

Encodage : UTF-8