**Concept et Plan de Mise en Œuvre du Projet**

**Titre du projet**

**Prédire la réussite académique à partir des habitudes quotidiennes des étudiants par l’apprentissage automatique**

**Membres de l’équipe – Groupe ML-Académique**

* Abate Yacoub
* Bassim
* Daoud

**🧭 Vue d’ensemble du projet**

La réussite académique est un enjeu central dans les Objectifs de Développement Durable (ODD), notamment l’ODD 4 : *"Assurer une éducation équitable, inclusive et de qualité pour tous"*. Malgré les avancées dans la digitalisation de l’enseignement, de nombreux établissements peinent à identifier, à temps, les étudiants à risque d’échec.

Ce projet vise à concevoir un modèle intelligent, capable de prédire le score académique des étudiants en s’appuyant sur leurs habitudes de vie (sommeil, activité physique, temps d’étude, médias sociaux…). En combinant **Machine Learning**, **exploration de données comportementales** et **visualisation intelligente**, notre solution entend outiller les enseignants, conseillers pédagogiques et étudiants pour favoriser une réussite personnalisée.

**🎯 Objectifs du projet**

L’objectif principal est de construire un système prédictif robuste et éthique permettant de :

1. **Modéliser le lien entre habitudes quotidiennes et performance académique** à l’aide de données comportementales.
2. **Identifier les variables les plus influentes** sur le succès scolaire (étude, sommeil, santé mentale...).
3. **Détecter les étudiants à risque** de contre-performance et recommander des actions préventives.
4. **Fournir une solution interprétable et reproductible**, utilisable dans divers contextes éducatifs.

**🧱 Contexte et justification**

Traditionnellement, la prédiction de la réussite se basait sur les notes antérieures, le sexe ou l’origine sociale. Or, de récentes recherches (Cao et al., 2017 ; Kim et al., 2021) ont montré que des variables comportementales comme la régularité du sommeil, l’activité physique ou le temps passé à étudier sont des prédicteurs plus fiables.

Grâce aux progrès du **machine learning supervisé**, il est désormais possible d’exploiter ces données pour modéliser de façon fine la réussite étudiante, à condition de disposer d’un jeu de données cohérent et bien prétraité. Notre approche entend démontrer cela en se basant sur un dataset public complet et éthique.

**🧠 Méthodologie**

**Architecture du projet**

Le projet repose sur une architecture modulaire en 5 étapes :

1. **Prétraitement des données**
   * Nettoyage des NA, encodage, normalisation
   * Feature engineering (ex. : score de routine)
2. **Exploration de données (EDA)**
   * Visualisation des corrélations
   * Analyse des distributions
3. **Modélisation machine learning**
   * Régression linéaire
   * Random Forest
   * XGBoost
   * Évaluation croisée (KFold)
4. **Interprétation des modèles**
   * Feature importance
   * SHAP values
5. **Visualisation des résultats**
   * Dashboards de performance
   * Matrice de confusion, courbe d’apprentissage

**🧠 Diagramme de conception**

Données brutes Kaggle

|

v

[Prétraitement] --> [EDA] --> [Modélisation ML]

| |

| v

[Nettoyage + encodage] [Random Forest / XGBoost]

|

v

[Interprétation] --> [Rapport + Visualisation]

**🗂 Source des données**

* **Nom** : *Student Habits vs Academic Performance*
* **Plateforme** : Kaggle
* **Lien** : [Kaggle Dataset](https://www.kaggle.com/datasets/jayaantanaath/student-habits-vs-academic-performance)
* **Taille** : 1000 observations, 15 variables
* **Contenu** : Sommeil, étude, activité physique, alimentation, part-time job, etc.
* **Prétraitement effectué** :
  + Suppression de student\_id
  + Encodage des variables catégorielles (gender, diet\_quality…)
  + Corrélation entre chaque variable et le score (exam\_score)
  + Visualisations avec Seaborn et Matplotlib

**📚 Revue de littérature (résumé)**

La littérature récente converge vers l’idée que les **habitudes de vie** influencent fortement la performance académique. Des travaux empiriques (Kim et al., 2021 ; Grossman et al., 2019) montrent que le sommeil, les heures d’étude ou le stress sont des prédicteurs significatifs.  
L’apprentissage automatique, et en particulier des algorithmes comme les **Random Forests** ou les **réseaux neuronaux multicouches**, permettent de capturer ces relations complexes et non linéaires avec des performances supérieures à 85 % de précision.

**🛠 Stack Technologique**

* **Langage principal** : Python 3.11
* **Prétraitement et visualisation** : Pandas, NumPy, Seaborn, Matplotlib
* **Machine learning** : Scikit-learn, XGBoost
* **Interprétation** : SHAP, Feature Importance
* **IDE** : Google Colab / Jupyter Notebook
* **Suivi collaboratif** : Trello, Google Drive, GitHub

**🗓 Chronologie du projet (Gantt simplifié)**

| **Semaine** | **Activité principale** |
| --- | --- |
| S1 | Collecte & exploration des données |
| S2 | Prétraitement, visualisation, corrélation |
| S3 | Implémentation des modèles ML |
| S4 | Évaluation, tuning, interprétation |
| S5 | Rapport final + présentation orale |

**📌 Jalons principaux**

* ✅ Données prétraitées et visualisées (corrélations fortes identifiées)
* 🔄 En cours : Entraînement des modèles (régression linéaire, XGBoost…)
* 🔜 À venir : Évaluation finale et recommandations d’action

**🧠 Défis et solutions proposées**

| **Défi** | **Solution** |
| --- | --- |
| Risque de surapprentissage | Validation croisée, simplification du modèle |
| Données déséquilibrées ou bruitées | Analyse exploratoire poussée, standardisation |
| Faible interprétabilité | Utilisation de modèles interprétables ou SHAP |
| Données sensibles | Respect des normes RGPD, anonymisation intégrale des données |

**🧠 Considérations éthiques**

* **Consentement et vie privée** : Aucune donnée nominative n’est utilisée.
* **Biais algorithmiques** : Nous évaluerons les biais potentiels selon le genre ou la situation socio-économique.
* **Usage éthique** : Le modèle est conçu comme un **outil d’aide à la décision**, non un système de notation automatisée.

**📎 Références (sélection)**

* Cao, R. et al. (2017). *Orderness predicts academic performance…*
* Kim, Y. et al. (2021). *Smartwatch-based monitoring…*
* Aya Nabil, R. et al. (2022). *Machine learning for academic performance prediction…*
* Grossman, M. et al. (2019). *Sleep and student achievement…*
* MDPI. (2021). *Educational Data Mining in Higher Education*