Analyse des données d'étudiants
Stress, anxiété et performance académique
Modélisation prédictive et recommandations basées sur les données

RAVELOJOELITAFIKA Tommy Alan M1 MAFI/MATH STRUCT

Décembre 2025

Résumé

Ce rapport présente une analyse complète des données d'étudiants issues du notebook `Student_stress_performance_prediction_recomandation.ipynb`. L'étude comprend une analyse exploratoire, un prétraitement des données, l'entraînement de plusieurs modèles d'apprentissage automatique pour la prédiction du niveau de stress et de la performance académique, ainsi qu'un module de recommandations. Ce module simule des interventions comportementales — augmentation du temps de sommeil et d'activité physique, réduction du temps d'écran — afin d'estimer leur impact potentiel sur le stress et la performance académique.

1 Introduction

Le stress académique et l'anxiété avant les examens constituent des facteurs déterminants de la réussite universitaire. L'objectif du notebook analysé est double :

— prédire automatiquement le niveau de stress et les variations de performance académique à partir de données comportementales ;
— transformer ces prédictions en recommandations concrètes et interprétables destinées à améliorer le bien-être étudiant.

2 Description du jeu de données

Le jeu de données utilisé dans le notebook contient des observations individuelles d'étudiants, chaque ligne correspondant à un étudiant. Les variables principales sont :

— `id` (identifiant anonyme),
— `Gender`, `Age`, `Education Level`,
— `Screen Time (hrs/day)`,
— `Sleep Duration (hrs)`,
— `Physical Activity (hrs/week)`,
— `Stress Level`,
— `Anxious Before Exams`,
— `Academic Performance Change`.

Le dataset exploité dans le notebook comporte environ 100 observations après chargement et sélection des variables pertinentes.

3 Prétraitement et nettoyage des données

Les étapes de prétraitement implémentées dans le notebook sont les suivantes :

— suppression des variables non pertinentes ou identifiantes (ex. noms) afin de garantir l'anonymat ;

— conversion explicite des variables quantitatives (âge, temps d'écran, durée de sommeil, activité physique) en format numérique ;
— encodage des variables catégorielles (genre, niveau d'éducation, stress, performance) ;
— vérification de l'absence de valeurs manquantes ou traitement approprié lorsque nécessaire ;
— séparation des données en ensembles d'entraînement, de validation et de test.

4 Statistiques descriptives clés

Mesure	Valeur
Nombre d'échantillons	1000.
Moyenne d'âge	20.342 ans.
Moyenne temps écran (hrs/day)	6.9092 h/j.
Moyenne durée de sommeil (hrs)	6.4508 h.
Moyenne activité physique (hrs/week)	5.0176 h.
Distribution `Stress Level` (modalité la plus fréquente)	`Medium` (492 occurrences).

5 Analyse exploratoire des données

L'analyse exploratoire réalisée dans le notebook met en évidence plusieurs tendances :
— un temps d'écran quotidien élevé est fréquemment associé à des niveaux de stress plus importants ;
— une durée de sommeil plus longue et l'activité physique hebdomadaire sont corrélés à une diminution du stress et de l'anxiété et à une meilleure stabilité de la performance académique ;

Les corrélations linéaires observées restent globalement faibles, ce qui suggère des relations non linéaires ou multifactorielles entre les variables comportementales et les cibles étudiées.

6 Tâches de modélisation

Le notebook définit principalement deux tâches de classification :

1. la prédiction du `Stress Level` (Low / Medium / High) ;
2. la prédiction du `Academic Performance Change` (Improved / Same / Decreased).
3. 55% train, 25% val ,teste = 20%

Les données sont divisées en ensembles d'entraînement, de validation et de test afin d'évaluer la capacité de généralisation des modèles.

7 Modèles entraînés

Plusieurs algorithmes d'apprentissage automatique sont implémentés et comparés dans le notebook :
— régression logistique ;
— arbre de décision ;
— forêts aléatoires ;
— gradient boosting ;
— machines à vecteurs de support (SVM) ;
— k plus proches voisins (KNN) ;
— Naive Bayes.

— XGBOOST

Les performances sont évaluées à l'aide de métriques standards : accuracy, précision, rappel et score F1.

8 Résultats expérimentaux

Les résultats obtenus montrent des performances globalement modestes, avec des scores F1 typiquement autour de 0.4 pour la prédiction du stress et de la performance académique. Cette limite s'explique par :

— la taille relativement réduite du jeu de données ;
— le déséquilibre entre certaines classes ;
— le caractère complexe et multifactoriel du stress académique.

Prédiction de `academic performance` (exemples) :

— Naive Bayes : F1 : 0.4146 | Accuracy : 0.4200
— Gradient Boosting : F1 : 0.4115 | Accuracy : 0.4480
— K-Nearest Neighbors : F1 : 0.3993 | Accuracy : 0.4040

Prédiction de `Stress Level` (exemples) :

— Naive Bayes : F1 : 0.4146 | Accuracy : 0.4200
— Gradient Boosting : F1 : 0.4115 | Accuracy : 0.4480
— K-Nearest Neighbors : F1 : 0.3993 | Accuracy : 0.4040

Néanmoins, les modèles parviennent à capter des tendances cohérentes exploitables pour la recommandation.

9 Fonction de recommandation basée sur les modèles

Une contribution centrale du notebook est l'intégration d'une fonction de recommandation. Cette fonction procède comme suit :

1. une observation représentant un étudiant est sélectionnée ;
2. certaines variables comportementales sont modifiées de manière contrôlée ;
3. le modèle prédit à nouveau le niveau de stress et la performance académique ;
4. la variation des prédictions est interprétée comme un effet potentiel de l'intervention.

9.1 Interventions simulées

Les interventions étudiées dans le notebook incluent :

— augmentation du temps de sommeil (+1 à +2 heures par nuit) ;
— augmentation de l'activité physique hebdomadaire ;
— réduction du temps d'écran quotidien.

Ces simulations montrent, dans de nombreux cas, une diminution du stress prédit et une amélioration ou stabilisation de la performance académique.

10 Conclusion

Le notebook analysé démontre la faisabilité d'une approche intégrée combinant apprentissage automatique et recommandations comportementales pour le stress et la performance académique. Des améliorations futures incluent l'augmentation du volume de données, l'enrichissement des variables et l'intégration de modèles causaux.