Analyse et Prédiction de la Dépression par Machine Learning

N'dje Jules Geraud Odje

16 février 2025

1 Introduction

1.1 Contexte

Cette étude analyse 140,700 observations comportant 20 variables pour prédire la dépression chez les étudiants et professionnels. L'objectif est de développer un modèle de détection précoce des risques de dépression.

1.2 Analyse Préliminaire

L'analyse des corrélations avec la dépression révèle :
- Indicateurs forts positifs : pression académique (0.59), stress financier (0.29), pression travail (0.11)
- Facteurs protecteurs (corrélation négative) : âge, satisfaction études/travail

1.3 Structure du Dataset

Le dataset présente une particularité : certaines variables dépendent du statut professionnel :

Variables étudiant-spécifiques :
- Academic Pressure
- CGPA
- Study Satisfaction

Variables professionnel-spécifiques :
- Work Pressure
- Job Satisfaction

2 Méthodologie

2.1 Prétraitement des Données

2.1.1 Gestion des Valeurs Manquantes

- Analyse NA selon statut

— Remplacement stratégique : 0 pour variables non applicables
— Moyenne pour valeurs manquantes dans variables applicables

2.1.2 Préparation des Features

— Standardisation des variables numériques
— Sélection features basée sur corrélations
— Création versions complète (4 features) et simplifiée (2 features)

2.2 Modélisation

2.2.1 Régression Logistique

— Choix : modèle baseline, interprétable
— Performance (train) :
  — Accuracy : 0.77
  — Recall (dépression) : 0.84
  — ROC-AUC : 0.758

2.2.2 Random Forest

— Choix : capture relations complexes
— Performance identique :
  — Accuracy : 0.77
  — Recall (dépression) : 0.84
  — ROC-AUC : 0.760

2.2.3 XGBoost

— Choix : performant sur données tabulaires
— Résultats similaires :
  — Accuracy : 0.77
  — Recall (dépression) : 0.84
  — ROC-AUC : 0.759

3 Résultats et Analyses

3.1 Performance Comparée

3.1.1 Modèle Complet

Variables : Academic Pressure, Financial Stress, Work Pressure, Work/Study Hours

— Accuracy maintenue : 0.77
— Recall élevé : 0.84
— ROC-AUC stable : 0.76

3.1.2 Modèle Simplifié

Variables : Academic Pressure, Financial Stress
— Légère baisse accuracy : 0.76
— Recall maintenu : 0.82
— ROC-AUC acceptable : 0.74

3.2 Prédictions sur Test Set

— Distribution constante : 62% dépression
— Cohérence entre modèles
— Random Forest sélectionné pour déploiement

4 Déploiement

4.1 Stratégie

— Sélection Random Forest
— Sauvegarde modèle et scaler
— Fonction prédiction opérationnelle

4.2 Application Pratique

— Détection précoce risques dépression
— Utilisation features facilement mesurables
— Possibilité intégration systèmes existants

5 Conclusion

Cette étude démontre l'efficacité d'une approche machine learning pour la détection de la dépression. Le modèle simplifié, basé sur la pression académique et le stress financier, offre un excellent compromis performance/complexité. Le maintien d'un recall élevé assure une bonne détection des cas positifs, prioritaire dans ce contexte médical.