Student Depression Study - L3 MIAGE - 2024-2025

Marius Adjakotan, Nesrine Allouche, Tiana Rasolonzatovo, Olesya ...

2025-05-17

Table of contents


```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

Introduction

La **santé mentale** des étudiants universitaires est une préoccupation croissante à l'échelle mondiale. Des études récentes indiquent qu'une part importante des jeunes adultes sont confrontés à des symptômes de dépression, liés à des facteurs variés : pression académique, incertitude financière, troubles du sommeil, ou encore problèmes personnels. Identifier précocement les étudiants à risque est crucial pour mettre en place des stratégies préventives et d'accompagnement.

Ce projet s'inscrit dans cette problématique : nous analysons un jeu de données regroupant plusieurs aspects de la vie étudiante afin de comprendre les déterminants de la dépression. Notre objectif est de modéliser la probabilité de dépression à partir de données académiques, personnelles et sociales, dans le but d'identifier les leviers d'action prioritaires pour les établissements d'enseignement.

Context / Dataset Description

Pour ce projet, nous utilisons un dataset intitulé "Student Depression Dataset", disponible sur Kaggle. Ce dataset vise à explorer les facteurs susceptibles d'influencer la santé mentale des étudiants, en particulier la dépression. Les données ont été collectées via des questionnaires auto-administrés, où chaque étudiant a décrit sa situation personnelle, académique et sociale.

Le jeu de données couvre plusieurs dimensions importantes :

- Aspects académiques :
  - Moyenne académique (CGPA) : indicateur global des performances scolaires.
  - Pression académique : ressenti du stress lié aux études (examens, devoirs…).
  - Satisfaction dans les études : niveau de contentement vis-à-vis de la filière suivie.
- Mode de vie et bien-être :
  - Durée moyenne du sommeil.
  - Habitudes alimentaires.
  - Pression professionnelle (pour ceux qui travaillent en parallèle des études).
  - Nombre d'heures de travail/études par jour.
  - Satisfaction professionnelle (pour ceux ayant un emploi).
- Facteurs personnels :
  - Genre, Âge, Ville de résidence.
  - Stress financier.
  - Profession / Diplôme préparé.
  - Antécédents familiaux de troubles mentaux.
  - Pensées suicidaires (question binaire : oui/non).

La variable d'intérêt principale est la présence ou non d'une dépression déclarée par l'étudiant (variable binaire).

☒ **À noter :**

Les données décrivent la situation des étudiants *au moment de l'enquête*, c'est-à-dire leur état de santé mentale et leurs conditions de vie *avant ou pendant* l'apparition d'une éventuelle dépression. Elles n'incluent pas les conséquences à long terme de la maladie ni de données cliniques approfondies.

Le dataset offre donc une photographie précieuse des *facteurs de risque potentiels* avant le début de symptômes dépressifs, ce qui permet d'identifier des profils à risque pour une détection précoce.

Il contient *27 901 enregistrements* et *18 variables*, chacun représentant un étudiant unique.

Description of the question

Notre question de recherche est formulée ainsi :

Quels sont les profils d'étudiants les plus vulnérables face à la dépression, et comment interagissent les facteurs académiques, sociaux et personnels dans la prédiction de ce risque ?

Cette question vise à :

- Identifier les **principaux facteurs prédictifs** de la dépression à partir des données collectées
- Explorer l'existence de **profils à risque** spécifiques (par exemple : étudiants cumulant stress académique élevé + mauvaise qualité de sommeil + difficultés financières)
- Analyser les **interactions potentielles** entre plusieurs facteurs pour comprendre comment ceux-ci se renforcent mutuellement dans l'augmentation du risque
- **Évaluer les disparités selon des sous-groupes :** par exemple en fonction du genre, de la ville, ou du niveau d'études, afin d'identifier des populations particulièrement vulnérables
- **Mettre en lumière des facteurs potentiellement "modifiables"** (ex : habitudes de sommeil, satisfaction, etc.) afin de proposer des pistes concrètes pour la prévention.
- **Identifier d'éventuelles incohérences ou paradoxes** (ex : des étudiants avec de bonnes notes mais malgré tout très dépressifs) pour enrichir la compréhension des dynamiques sous-jacentes

Ainsi, ce projet combine une approche descriptive (exploration des tendances globales et des disparités) et une approche prédictive.

Methodology

Data clean-up procedures

Avant de procéder à l'analyse, un nettoyage minutieux du jeu de données a été effectué afin d'assurer la qualité des résultats obtenus. Cela comprend les étapes suivantes :

- **Harmonisation des valeurs** : certaines colonnes comme `Sleep Duration` contenaient des guillemets superflus (ex : `'5-6 hours'`) qui ont été supprimés.
- **Typage correct des variables** : les colonnes catégorielles (ex. : `Gender`, `City`, `Profession`, `Degree`) ont été converties en facteurs. De même, les variables binaires (`Suicidal Thoughts`, `Family History of Mental Illness`, `Depression`) ont été recodées pour faciliter les analyses statistiques.
- **Renommage des colonnes** : certaines colonnes ont été renommées pour éviter les problèmes liés aux caractères spéciaux (comme les points ou espaces).
- **Contrôle des valeurs manquantes** : aucune valeur manquante n'a été détectée dans le jeu de données, ce qui a permis de conserver l'ensemble des 27 901 observations.
- **Préparation à l'analyse exploratoire** : les données sont désormais prêtes à être analysées en profondeur, notamment via des résumés statistiques, des visualisations, et des modèles prédictifs.

Ces étapes permettent de garantir la robustesse des analyses statistiques ultérieures et d'éviter les biais dus à des erreurs de codage ou à des valeurs aberrantes.

```
[1] 3
```