

PLAN PRÉVISIONNEL

Contexte

On veut créer une solution qui aide à repérer des étudiants qui pourraient être en difficulté (risque d'échec/décrochage) en lien avec le sommeil, le stress et le bien-être. L'objectif est la **prévention** : repérer tôt pour proposer un accompagnement.

Dataset retenu

On utilise un dataset d'environ **2000 lignes** avec des variables comme : sommeil, stress, bien-être, performance scolaire, activité physique et une variable catégorielle (Gender). Après nettoyage, on crée un fichier unique **final.csv**.
Ce fichier sert de base commune : tout le monde travaille sur les mêmes données.

Modèle envisagé (algorithme)

CatBoost, c'est quoi ?

CatBoost est un algorithme de Machine Learning (un modèle) qui sert à prédire quelque chose à partir de données en tableau.

Exemple dans ton projet

Tu as des infos comme :

durée de sommeil , stress , activité physique , genre ...

CatBoost apprend des "règles" à partir de ces variables pour répondre à une question du type :

"Est-ce que cet étudiant est à risque ?" ou "Quel est son score de risque ?"

Pourquoi CatBoost est intéressant ?

Parce qu'il est souvent très bon sur les données "tableaux" (comme Excel) et surtout il gère bien les variables catégorielles (des mots) comme :

Gender = Male/Female

Beaucoup de modèles demandent de transformer ces mots en chiffres avant. CatBoost le fait très bien et proprement.

SHAP, c'est quoi ?

SHAP n'est pas un modèle. C'est un outil pour expliquer les décisions d'un modèle.

Pourquoi on en a besoin ?

Un modèle comme CatBoost peut donner une prédiction, mais le prof (et le métier) veut savoir :

“Pourquoi tu dis que cet étudiant est à risque ?”

“Qu’est-ce qui a le plus influencé la décision ?”

SHAP répond à cette question.

SHAP “global” vs “local” (très important)

1) Explication globale (global feature importance)

Ça répond à :

“Dans l’ensemble, quelles variables comptent le plus pour le modèle ?”

Exemple :

Stress = très important

Sommeil = important

Activité physique = moins important

Donc global = vue générale sur tous les étudiants.

2) Explication locale (local explanation)

Ça répond à :

“Pour CET étudiant précis, pourquoi le modèle a prédit ‘à risque’ ?”

Exemple :

Stress très élevé → augmente le risque

Sommeil faible → augmente le risque

Bien-être faible → augmente le risque

Activité physique correcte → diminue un peu le risque

Donc local = explication personnalisée pour une seule prédiction

On teste un modèle récent (exemple : **CatBoost**) parce qu’il peut être performant sur des données “tableau” et qu’il gère bien les variables catégorielles comme Gender. L’objectif du modèle est de produire un **score** ou une **classe** (“à risque” / “pas à risque”).

Références bibliographiques

Le projet s’appuie sur 2–3 références qui expliquent :

- le problème étudié (sommeil/stress et performance),
- et l’algorithme choisi (ex : CatBoost) + l’explicabilité (ex : SHAP).

Démarche de test (preuve de concept)

On va comparer deux choses :

- une **baseline** (modèle simple) pour avoir une référence,
- le **nouvel algorithme** (modèle récent) pour voir s'il fait mieux.

On mesure les performances avec une métrique adaptée au besoin : en prévention, on veut éviter de rater un étudiant à risque, donc on suit surtout le **recall** (et aussi précision/F1).

CatBoost: unbiased boosting with categorical features

A Unified Approach to Interpreting Model Predictions

<https://data.mendeley.com/datasets/5mvr4v62z/3>

<https://arxiv.org/abs/>

<https://www.youtube.com/watch?v=cQK5ipnhoTY>

je me suis basé sur cet vidéo pour en apprendre plus sur le projet et le sujet.

Planning

Semaine 1 : Comprendre + préparer les données

- Étape 1 : Cadrage (objectif, ce qu'on veut prédire, contraintes)
 - Livrable : texte "contexte + objectif"
- Étape 2 : Analyse du dataset (EDA : valeurs manquantes, incohérences, graphiques)
 - Livrable : notebook EDA
- Étape 3 : Nettoyage + création du fichier final.csv
 - Livrable : final.csv

Semaine 2 : Tester les modèles (POC)

- Étape 4 : Baseline (modèle simple)
 - Livrable : résultats baseline (tableau métriques)
- Étape 5 : Modèle récent (ex : CatBoost) + optimisation
 - Livrable : résultats modèle récent + comparaison

Semaine 3 : Expliquer + présenter

- Étape 6 : Explicabilité (feature importance globale + locale)
 - Livrable : graphiques + interprétation
- Étape 7 : Dashboard KPI (tableaux + graphiques compréhensibles)
 - Livrable : notebook dashboard + interprétations
- Étape 8 : Rédaction finale (méthodo + résultats + limites)
 - Livrable : rapport final