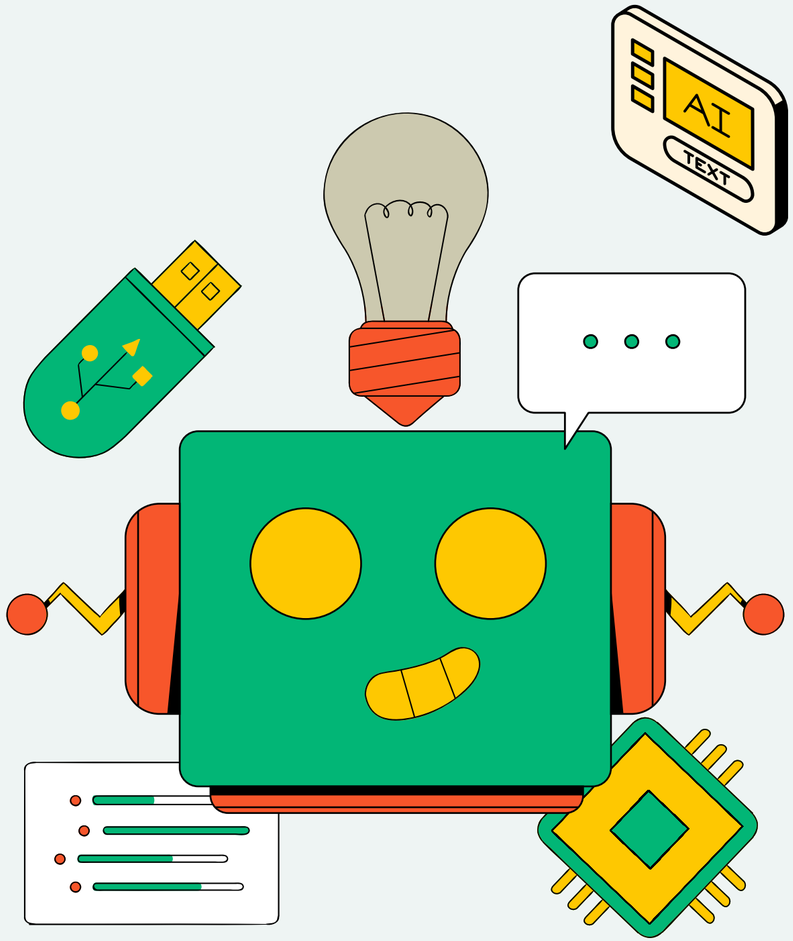
SCIENCES SORBONNE UNIVERSITÉ

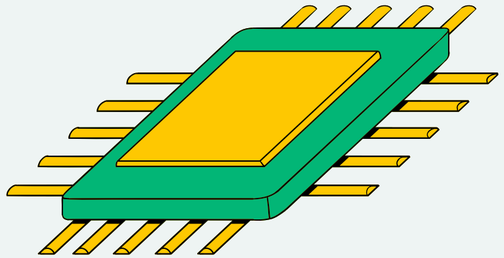


SPLEX PROJECT

LIFE SCIENCES DATA EXPLORATION

PRESENTED BY:

OCÉANE LI & ADAM BOUMESSAOUD

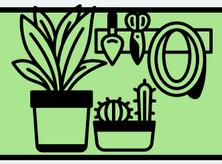
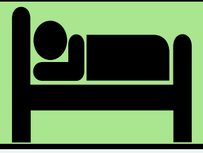


INTRODUCTION

Student Lifestyle Dataset

2000 individus

No missing values

ID						GPA	Stress
127	8.8	2.2	8.9	2	2.1	3.16	High
...	float	float	float	float	float	float	object

INTRODUCTION

Peut-on identifier les facteurs qui influencent le stress des étudiants et les utiliser pour prédire leur niveau de stress ?

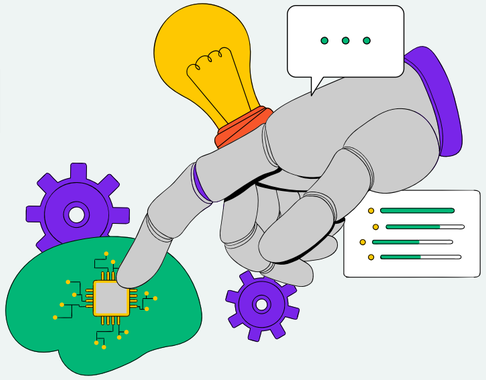
GPA

Modèle de prédiction

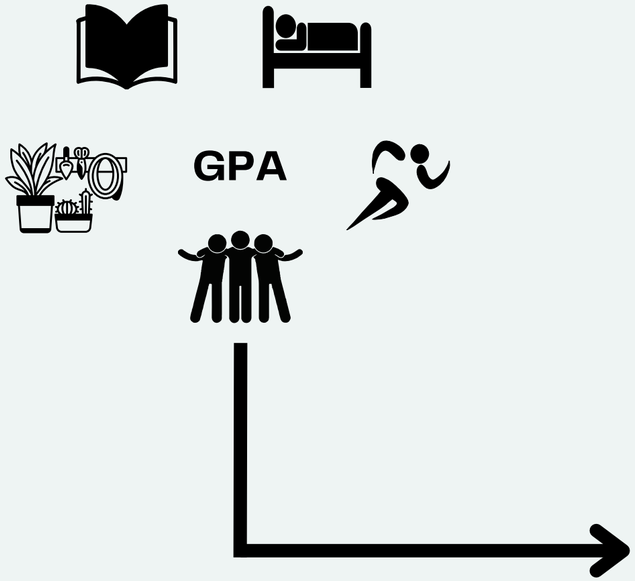
?

Stress Level

PRINCIPAL COMPONENT ANALYSIS (PCA)



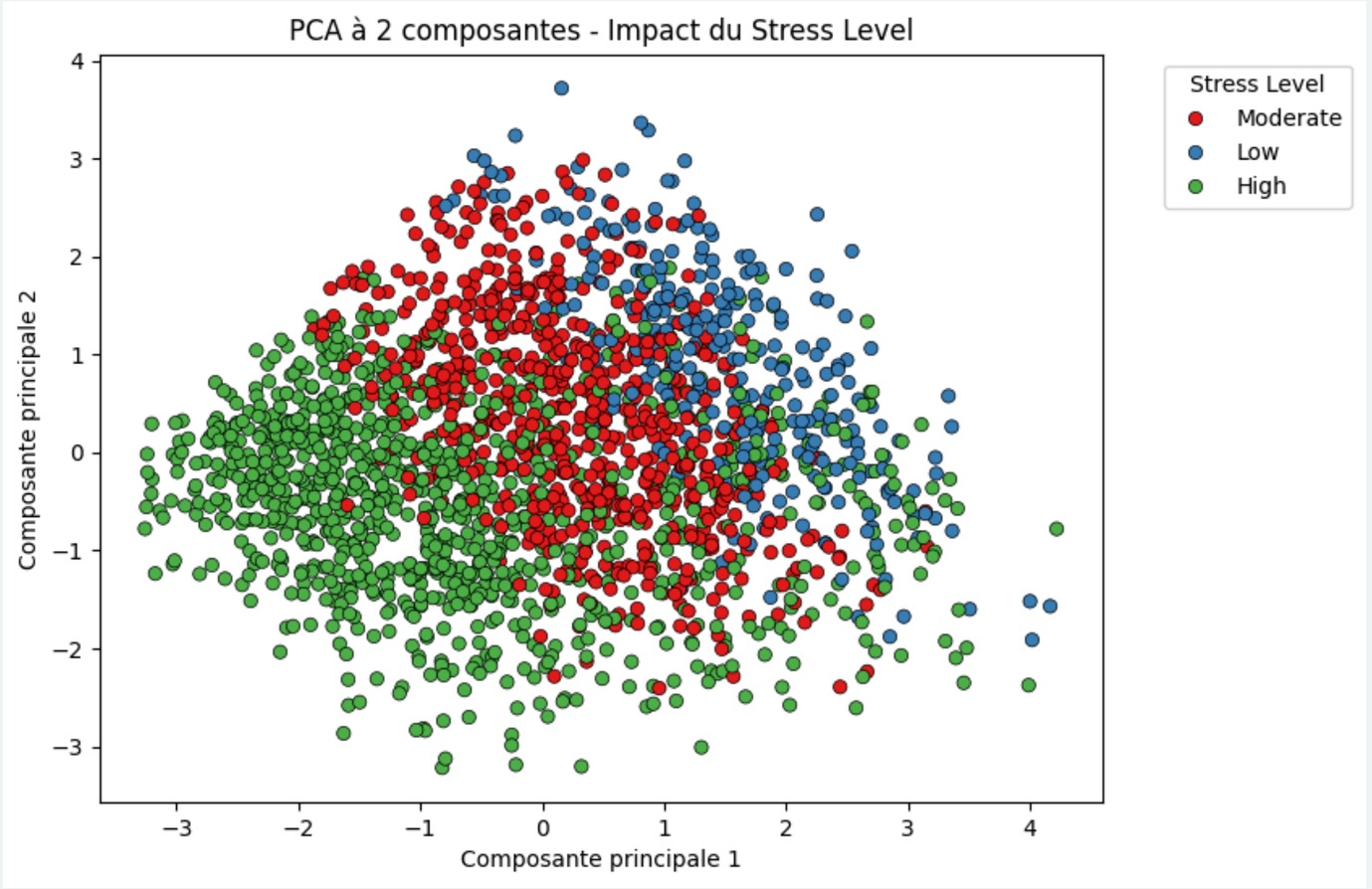
Variables numériques



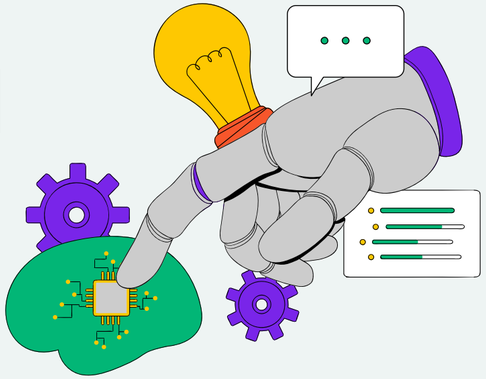
Standardiser les données

moyenne = 0

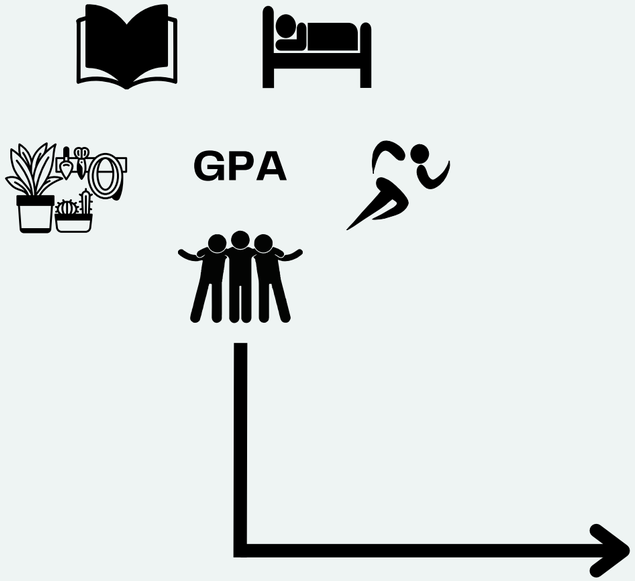
écart-type = 1



PRINCIPAL COMPONENT ANALYSIS (PCA)



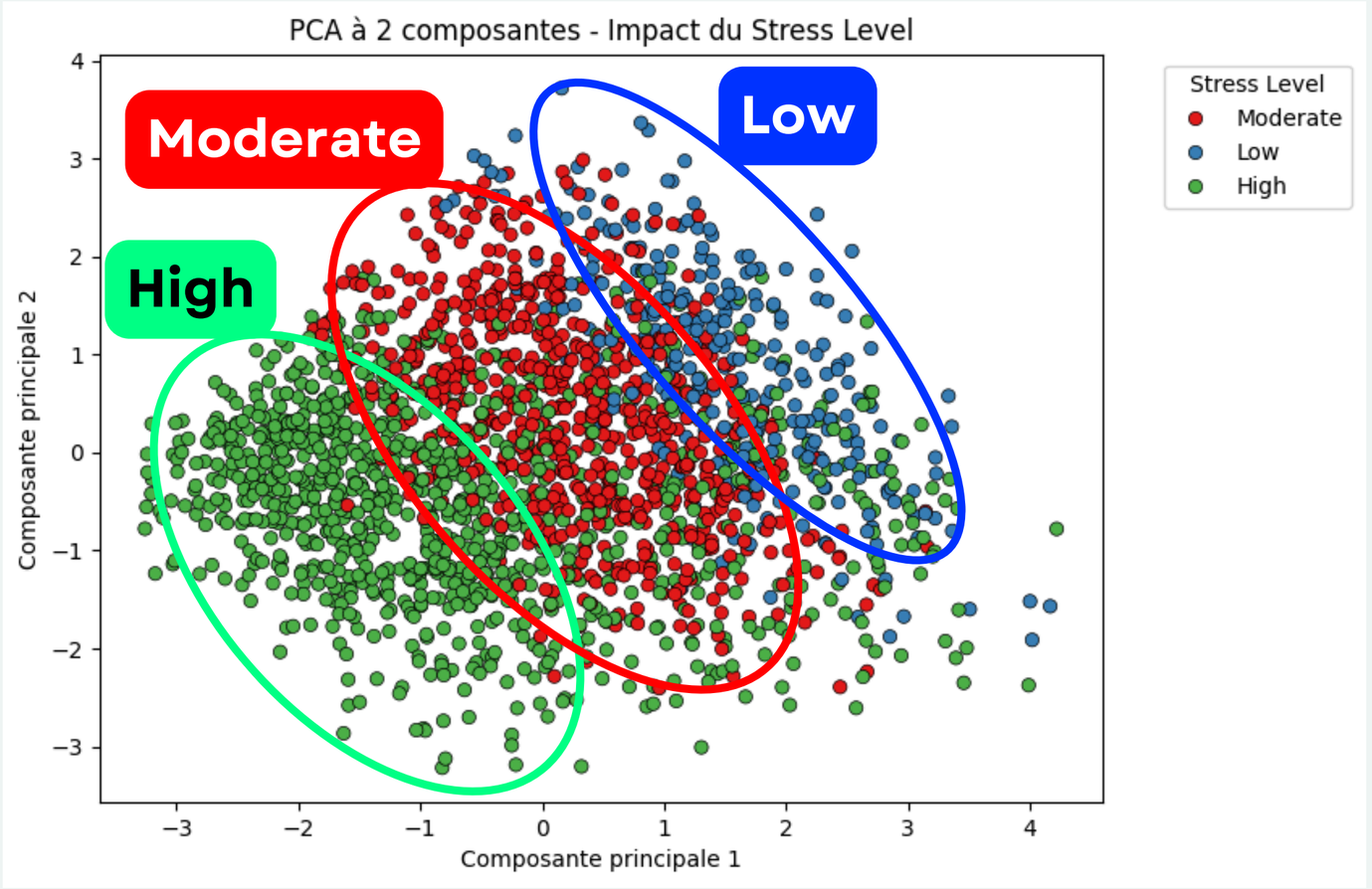
Variables numériques



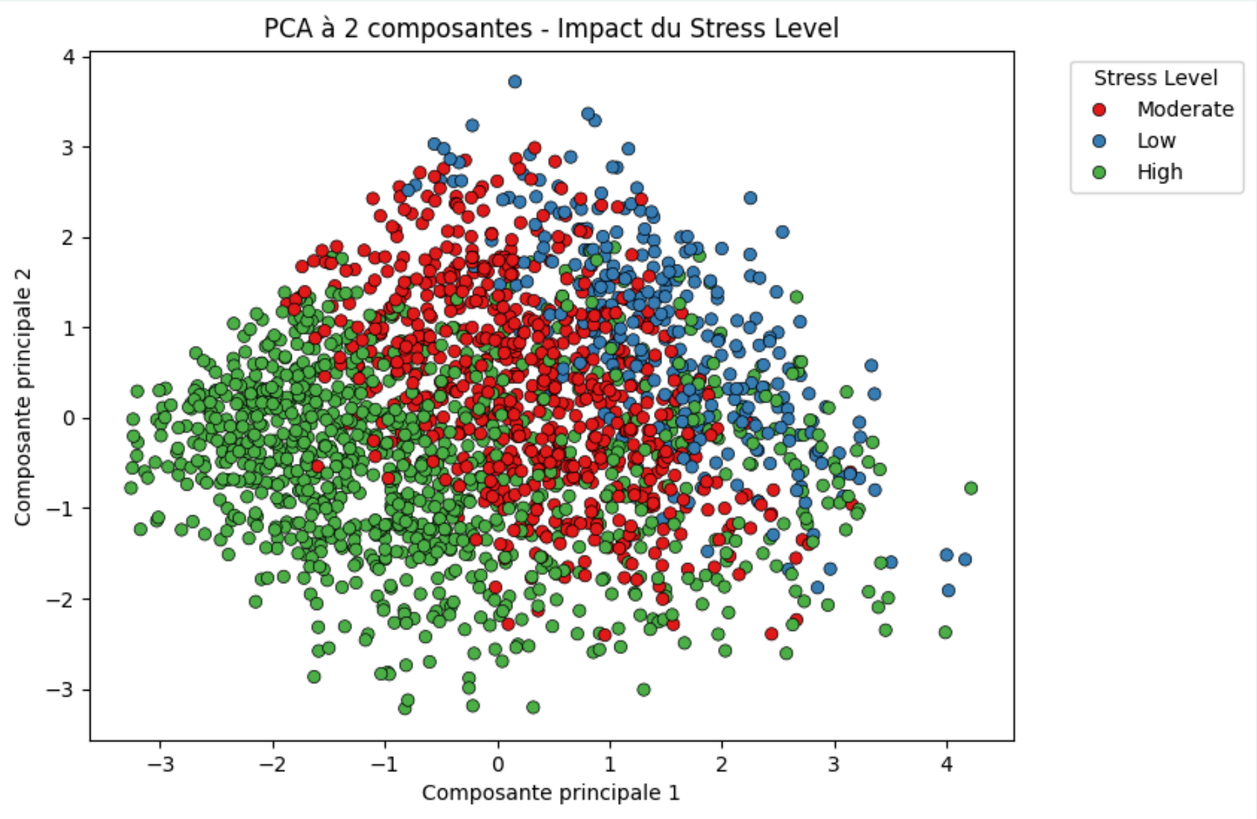
Standardiser les données

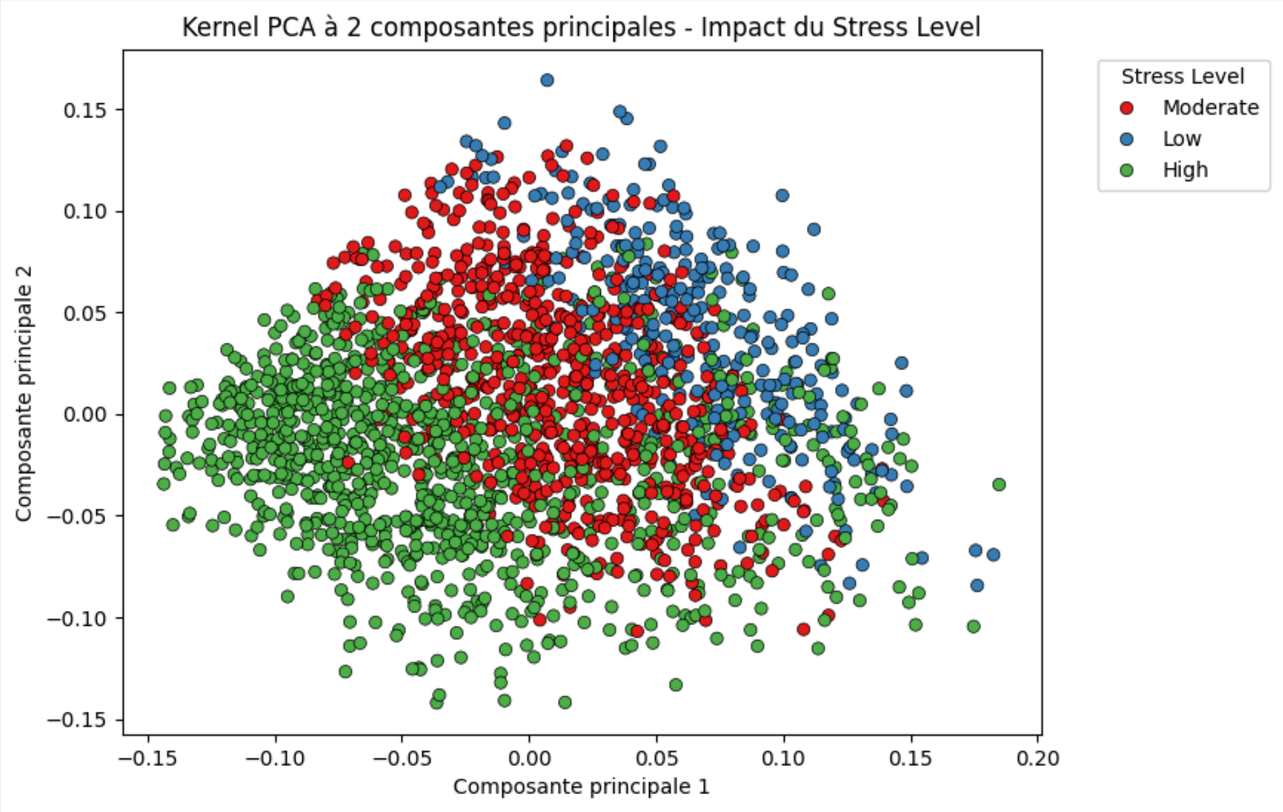
moyenne = 0

écart-type = 1



KERNEL PCA





PCA Standard

Kernel PCA

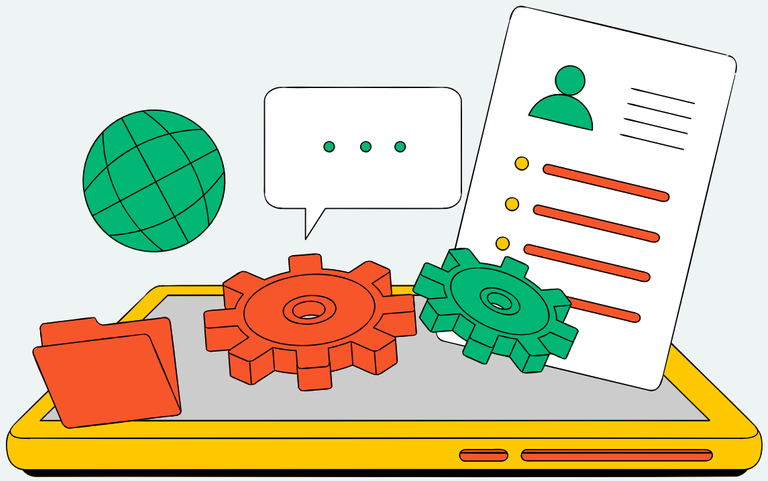
MÉTHODES NON SUPERVISÉE

K-Means

ARI : 0.25

Spectral clustering

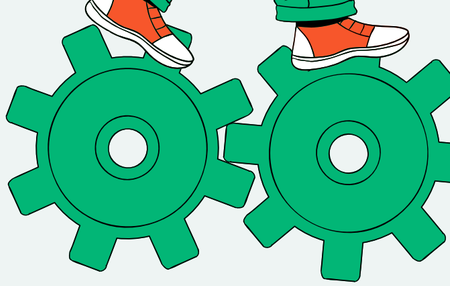
ARI : 0.19



Hierarchical clustering

ARI (ward) : 0.14
ARI (average) : 0.24
ARI (complete) : 0.13

CLASSIFICATION SUPERVISÉE

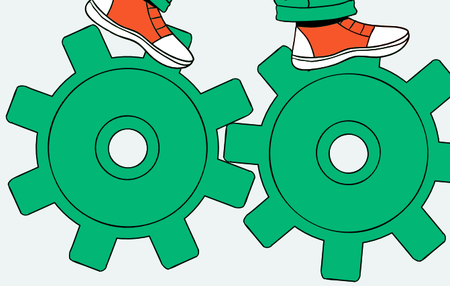


Data

80% training

20% test

CLASSIFICATION SUPERVISÉE



Régression Logistique

Random Forest

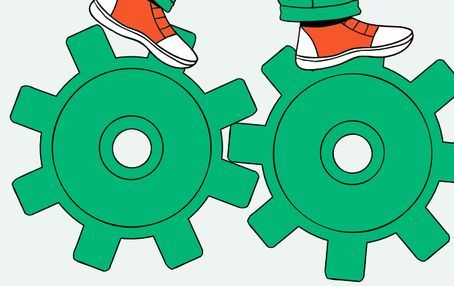
Arbre de décisions

Support Vector Machine

K-Nearest Neighbors

Le meilleur ?

CLASSIFICATION SUPERVISÉE



Accuracy : 0.75

Support Vector Machine

Noyau RBF

Paramètres

Régularisation	C = 1
Influence des points d'entraînement individuels	Gamma = 1
Noyau	RBF

DEEP LEARNING

- Encoding
- Scaling
- Train/Test Split
- Training
- Testing

DEEP LEARNING

Input	Input
Hidden Layer 1	ReLu activation, 30% dropout
Hidden Layer 2	ReLu activation, 30% dropout
Hidden Layer 3	ReLu activation, 20% dropout
Output	Softmax activation

CATBOOSTCLASSIFIER

1

Encoding des variables non numériques

2

Calcul des splits

3

Choix de la structure de l'arbre

4

Boosting

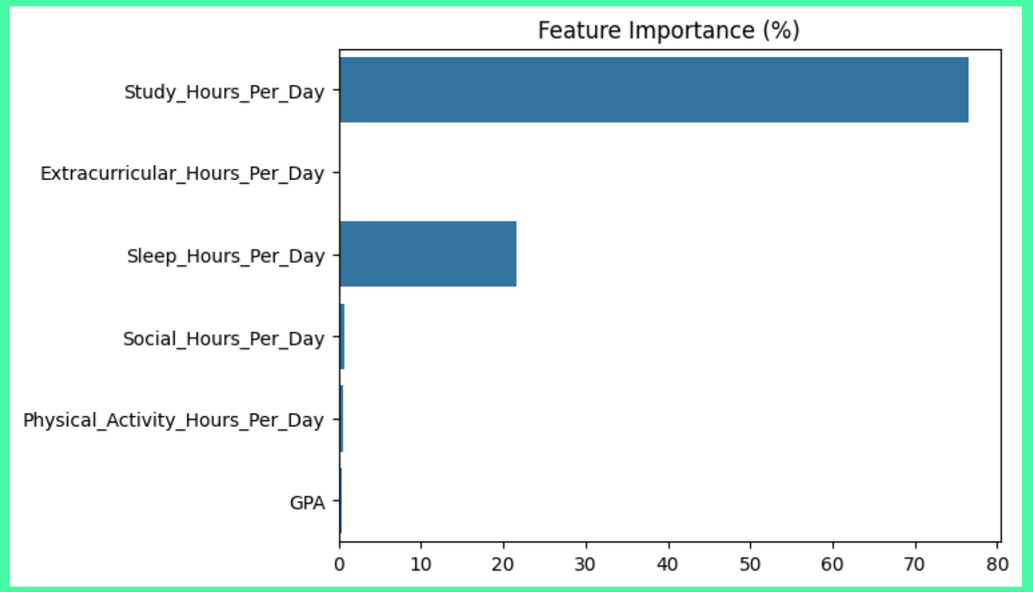
5

Repeat

6

Overfitting detector

CATBOOSTCLASSIFIER

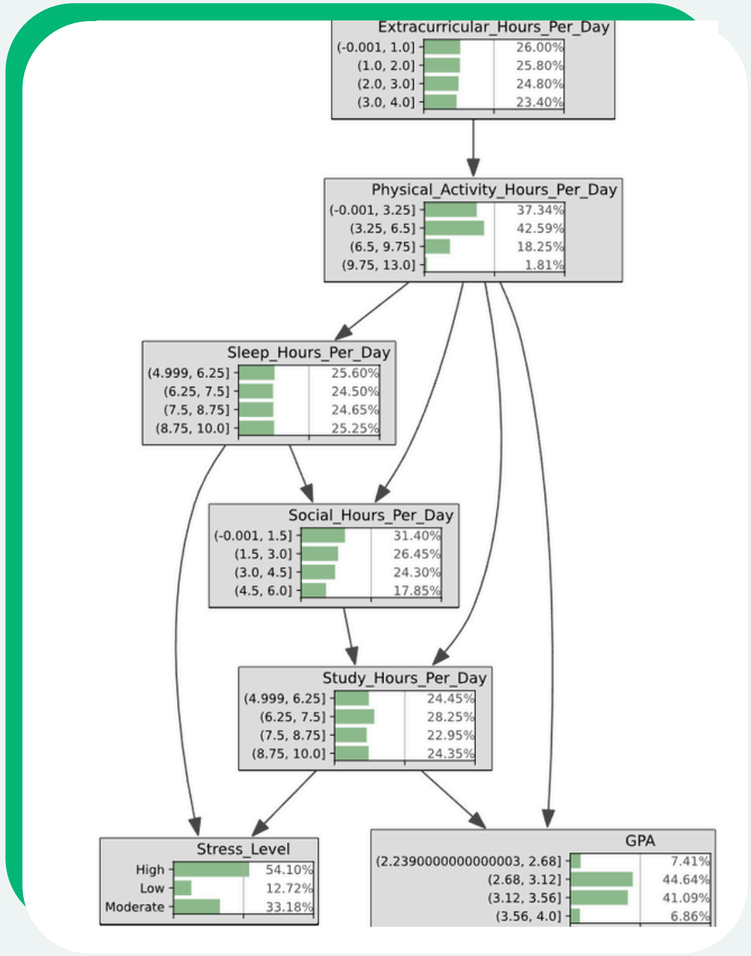


Résultats parfaits et sans overfitting

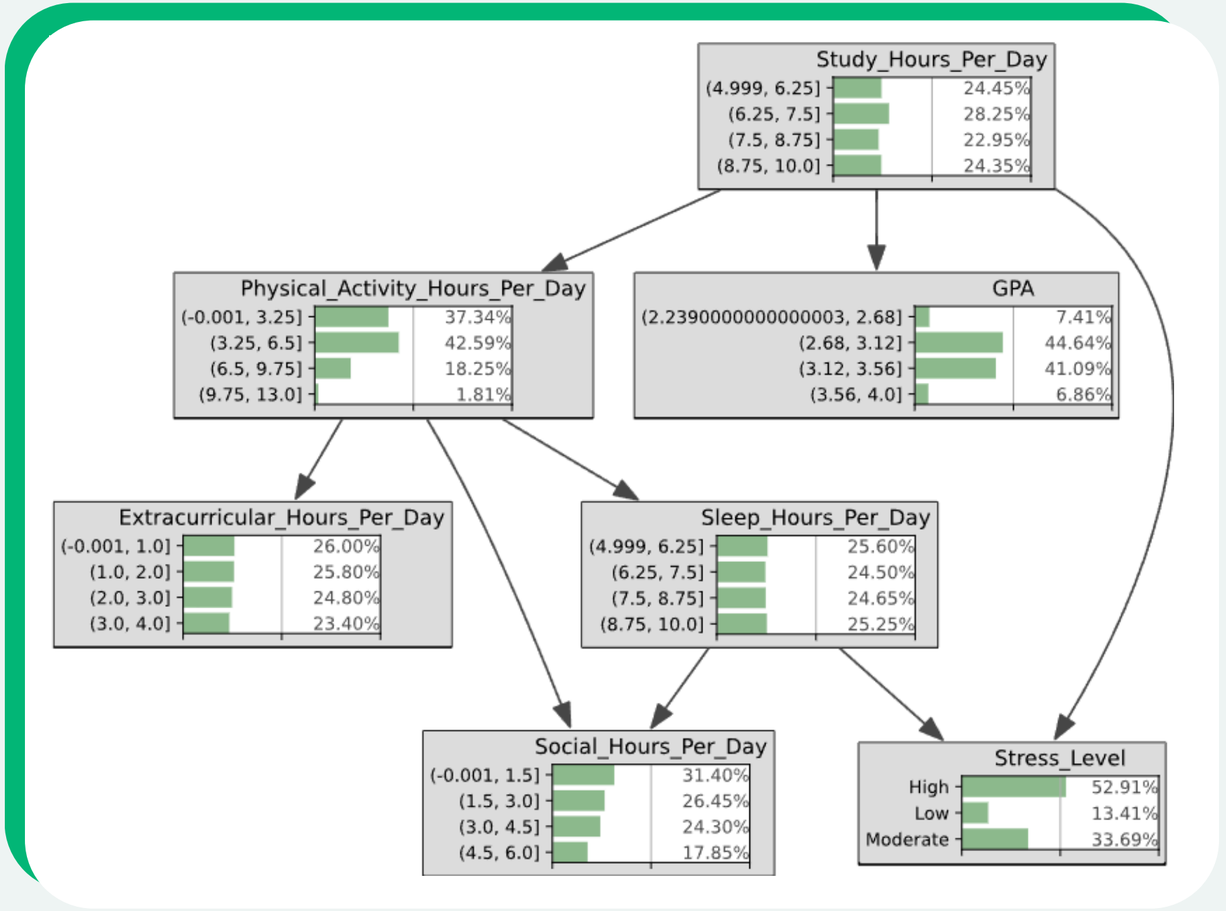
	precision	recall	f1-score	support
High	1.00	1.00	1.00	194
Low	1.00	1.00	1.00	60
Moderate	1.00	1.00	1.00	146
accuracy			1.00	400
macro avg	1.00	1.00	1.00	400
weighted avg	1.00	1.00	1.00	400

GRAPH ORIENTÉ ACYCLIQUE

LocalSearchWithTabuList()



GreedyHillClimbing()

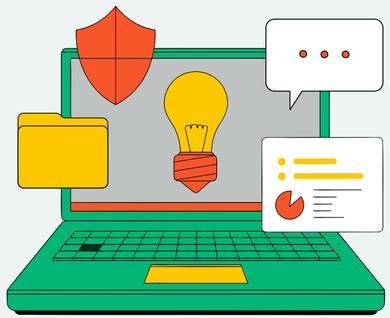
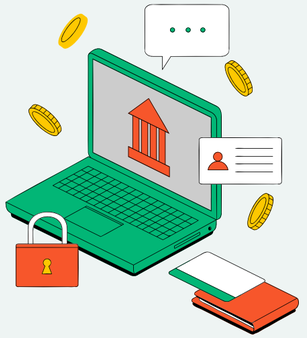
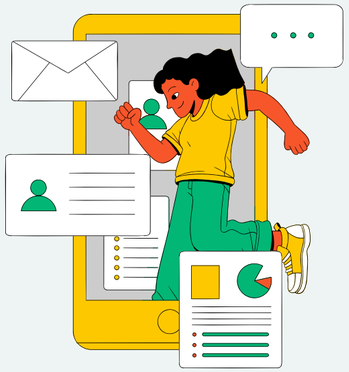


GRAPH ORIENTÉ ACYCLIQUE

LocalSearchWithTabuList()

		Stress_Level		
Sleep_Hours_Per_Day	Study_Hours_Per_Day	High	Low	Moderate
(4.999, 6.25]	(4.999, 6.25]	0.7396	0.1952	0.0652
	(6.25, 7.5]	0.7886	0.0001	0.2112
	(7.5, 8.75]	0.9271	0.0002	0.0727
	(8.75, 10.0]	0.9996	0.0002	0.0002
(6.25, 7.5]	(4.999, 6.25]	0.0002	0.7279	0.2720
	(6.25, 7.5]	0.0001	0.0001	0.9997
	(7.5, 8.75]	0.5900	0.0002	0.4098
	(8.75, 10.0]	0.9996	0.0002	0.0002
(7.5, 8.75]	(4.999, 6.25]	0.0002	0.7325	0.2673
	(6.25, 7.5]	0.0002	0.0002	0.9997
	(7.5, 8.75]	0.6632	0.0002	0.3366
	(8.75, 10.0]	0.9997	0.0002	0.0002
(8.75, 10.0]	(4.999, 6.25]	0.0002	0.7719	0.2280
	(6.25, 7.5]	0.0002	0.0002	0.9997
	(7.5, 8.75]	0.6248	0.0002	0.3750
	(8.75, 10.0]	0.9997	0.0002	0.0002

GreedyHillClimbing()

		Stress_Level		
Sleep_Hours_Per_Day	Study_Hours_Per_Day	High	Low	Moderate
(4.999, 6.25]	(4.999, 6.25]	0.7396	0.1952	0.0652
	(6.25, 7.5]	0.7886	0.0001	0.2112
	(7.5, 8.75]	0.9271	0.0002	0.0727
	(8.75, 10.0]	0.9996	0.0002	0.0002
(6.25, 7.5]	(4.999, 6.25]	0.0002	0.7279	0.2720
	(6.25, 7.5]	0.0001	0.0001	0.9997
	(7.5, 8.75]	0.5900	0.0002	0.4098
	(8.75, 10.0]	0.9996	0.0002	0.0002
(7.5, 8.75]	(4.999, 6.25]	0.0002	0.7325	0.2673
	(6.25, 7.5]	0.0002	0.0002	0.9997
	(7.5, 8.75]	0.6632	0.0002	0.3366
	(8.75, 10.0]	0.9997	0.0002	0.0002
(8.75, 10.0]	(4.999, 6.25]	0.0002	0.7719	0.2280
	(6.25, 7.5]	0.0002	0.0002	0.9997
	(7.5, 8.75]	0.6248	0.0002	0.3750
	(8.75, 10.0]	0.9997	0.0002	0.0002

MERCI DE VOTRE ATTENTION