

# DÉPRESSION

## étudiante

Hexabyte Team

ENCADRÉ PAR : MUSTAPHA TRABELSI



### PROBLÉMATIQUE

Comment identifier précocement les étudiants en situation de mal-être psychologique afin de prévenir la dégradation de leur santé mentale et de mettre en place des actions de soutien adaptées et efficaces ?

## **Introduction :**

La santé mentale des étudiants est devenue un sujet important ces dernières années. De nombreux étudiants font face à différentes sources de stress, comme la pression académique, les difficultés financières ou le manque de sommeil. Ces facteurs peuvent avoir un impact négatif sur leur bien-être et peuvent, dans certains cas, conduire à des problèmes de dépression.

Dans ce contexte, l'exploitation des données issues des environnements académiques et personnels offre une opportunité significative pour identifier précocement les étudiants à risque, mieux comprendre les facteurs influençant la dépression et proposer des actions de prévention ciblées.

Pour mener ce projet, nous avons suivi la méthodologie CRISP-DM. Cette méthodologie permet d'organiser le travail en plusieurs étapes, allant de la compréhension du problème métier jusqu'au déploiement des résultats.

Ce rapport présente les différentes phases du projet selon la méthodologie CRISP-DM :

- Compréhension métier et data
- Compréhension des données
- Préparation des données
- Modélisation
- Evaluation
- Déploiement

## **Problématique :**

Comment exploiter les données académiques, comportementales et personnelles des étudiants pour détecter précocement les risques de dépression, segmenter les profils à risque et proposer des actions de bien-être adaptées ?

## Chapitre 1 : Objectifs métier

Dans le cadre de ce projet, plusieurs objectifs métier ont été définis afin de répondre aux enjeux liés à la santé mentale des étudiants. Ces objectifs visent à améliorer la détection, la compréhension et la prévention de la dépression chez les étudiants, tout en facilitant la prise de décision des acteurs concernés.

### **1.Objectifs business**

#### **Détection des étudiants à risque de dépression :**

Identifier automatiquement les étudiants susceptibles de souffrir de dépression. Une détection précoce permettrait de mettre en place des actions de prévention et de soutien avant que la situation ne s'aggrave.

#### **Segmentation des profils d'étudiants :**

Regrouper les étudiants selon leurs habitudes de vie, leur niveau de stress et leur satisfaction académique. Cette segmentation permet de mieux comprendre les différents profils d'étudiants et d'identifier des groupes présentant des risques similaires.

#### **Recommandation d'actions de bien-être**

Proposer des actions de bien-être adaptées aux besoins des étudiants. En fonction de leur profil, il est possible de suggérer des conseils ou activités visant à réduire le stress et à améliorer la satisfaction et la qualité de vie des étudiants.

#### **Évaluation du niveau de sévérité de la dépression**

Estimer un niveau de sévérité (léger, modéré ou sévère) afin d'adapter les interventions et les ressources en fonction du degré de gravité.

#### **Suivi temporel des indicateurs de bien-être et système d'alertes**

Fournir à l'administration universitaire un outil de suivi temporel du bien-être étudiant, basé sur des indicateurs mensuels agrégés issus des données collectées. Ce dispositif permet de détecter automatiquement des périodes atypiques ou critiques grâce à des techniques de détection d'anomalies, afin de soutenir la prise de décisions préventives à l'échelle globale.

## 2.Traduction des objectifs métier en objectifs data science :

Afin de répondre aux objectifs métier précédemment définis, ceux-ci sont traduits en objectifs data science concrets et mesurables :

- **Prédiction de la dépression** : Mettre en place un modèle de classification supervisée permettant de prédire si un étudiant est en situation de dépression (oui/non) à partir de ses caractéristiques personnelles, académiques et comportementales.

Type d'analyse : Classification supervisée

Modèles utilisés : SVM ,Logistic Regression, Random Forest, XGBoost, tree decision, KNN

- **Segmentation des profils d'étudiants** : Appliquer des techniques de clustering non supervisé afin de regrouper les étudiants en profils similaires selon leurs habitudes de vie, leur niveau de stress et leur satisfaction académique, dans le but d'identifier des groupes à risque.

Type d'analyse : Clustering non supervisé

Modèles utilisés : K-means, DBSCAN, CAH

- **Recommandation d'actions de bien-être personnalisées** : Développer un système de recommandation capable de suggérer des actions personnalisées (amélioration du sommeil, activité physique, organisation du temps) en fonction du cluster auquel appartient chaque étudiant et des caractéristiques les plus à risque pouvant mener à une dépression.

Type d'analyse : Système de recommandation

Méthodes utilisées : recommandation basée sur les profils des clusters et des features influençant la dépression.

- **Prédiction du niveau de sévérité de la dépression** : Construire un modèle de régression permettant d'estimer le niveau de sévérité de la dépression (léger, modéré, sévère) à partir des caractéristiques académiques, personnelles et psychologiques des étudiants.

Type d'analyse : Régression supervisée

Modèles utilisés : Logistic regression , linear regression et RandomForest.

- **Suivi temporel des indicateurs de bien-être et détection d'anomalies :**

Mettre en place un système de suivi temporel basé sur des indicateurs mensuels agrégés (pression académique moyenne, durée moyenne de sommeil, taux de dépression, score

moyen de sévérité et répartition des profils étudiants), afin de détecter automatiquement des périodes atypiques ou critiques.

Type d'analyse : Détection d'anomalies non supervisée

Modèles utilisés : Z-Score (méthode statistique de référence), Isolation Forest

---

## Chapitre 2 : Compréhension des données

### 1. Présentation du dataset

Le projet s'appuie sur un dataset intitulé **Student Depression**, qui regroupe des informations relatives à la santé mentale, aux conditions académiques et aux habitudes de vie des étudiants. Les données incluent à la fois des variables quantitatives et qualitatives, permettant d'analyser différents facteurs pouvant être liés à la dépression.

```
#data shape (nbre de lignes et colonnes)
print(set.shape)
```

(34875, 18)

=> Chaque observation du dataset correspond à un étudiant, décrit par plusieurs caractéristiques personnelles, académiques et comportementales. La variable cible principale est Depression, indiquant si un étudiant est en situation de dépression ou non.

Student Depression																		Have you ever had suicidal thoughts ?		Work/Study Hours		Financial Stress		Family History of Depression Mental Illness														
id	Gender	Age	City	Profession	Academic Pressure		Work Pressure		Study Satisfaction		Job Satisfaction		Sleep Duration		Dietary Habits		Degree		Have you ever had suicidal thoughts ?		Work/Study Hours		Financial Stress		Family History of Depression Mental Illness													
					CGPA	SAT	GP	WP	SS	SAT	JD	SD	DH	DR	HS	WD	DS	DI	DU	ED	LLM	BBA	No	Yes	Hours	Score	Low	Medium	High	No	Yes							
I9842.0	Female	26.3	Agra	Student	4.9375	0.0	7.53	9.5	0.0	5-6 hours	Moderate	LLM	No	6.2	Low	No	0.0	More than 8 hours	Healthy	B.Ed	No	5.4	Low	No	0.0	Less than 5 hours	MD	No	1.0	Medium	No	0.0	ME	No	5.4	High	No	0.0
I9359.0	Female	25.6	Meerut	Student	5.2500	0.0	8.08	8.1	0.0	7-8 hours	Healthy	BBA	No	5.4	Low	No	0.0	More than 8 hours	Healthy	B.Ed	No	1.0	Medium	No	0.0	Less than 5 hours	MD	No	1.0	Medium	No	0.0	ME	No	5.4	High	No	0.0
I9886.0	Male	27.0	Ludhiana	Student	4.6250	0.0	6.47	6.3	0.0	More than 8 hours	Healthy	B.Ed	No	1.0	Medium	No	0.0	More than 8 hours	Healthy	B.Ed	No	1.0	Medium	No	0.0	Less than 5 hours	MD	No	1.0	Medium	No	0.0	ME	No	5.4	High	No	0.0
I0910.0	Male	32.9	Varanasi	Freelancer	5.1250	0.0	8.02	9.4	0.0	Less than 5 hours	Healthy	MD	No	1.0	Medium	No	0.0	More than 8 hours	Healthy	B.Ed	No	1.0	Medium	No	0.0	Less than 5 hours	MD	No	1.0	Medium	No	0.0	ME	No	5.4	High	No	0.0
I8887.0	Male	28.8	Vadodara	Student	6.4375	0.0	8.05	7.4	0.0	7-8 hours	Healthy	ME	No	5.4	High	No	0.0	More than 8 hours	Healthy	B.Ed	No	1.0	Medium	No	0.0	Less than 5 hours	MD	No	1.0	Medium	No	0.0	ME	No	5.4	High	No	0.0

### 2. Exploration du dataset (statistiques, visualisations) :

```
⌚ #Afficher les statistiques descriptives des colonnes numériques du Dataset
set.describe()
```

...	id	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Depression
count	33791.000000	33788.000000	33785.000000	33789.000000	34875.000000	34875.000000	33787.000000	33783.000000	34875.000000
mean	81363.012049	24.466238	5.986858	0.000503	7.193493	5.733091	0.010789	5.040346	0.709710
std	76766.706924	7.655751	1.008819	0.048351	1.964584	3.552985	0.076914	4.039432	0.917618
min	2.000000	17.000000	3.000000	0.000000	2.900000	1.000000	0.000000	1.000000	0.000000
25%	36216.000000	20.500000	5.187500	0.000000	6.220000	3.800000	0.000000	2.900000	0.000000
50%	72727.000000	23.400000	6.062500	0.000000	6.960000	5.200000	0.000000	4.500000	1.000000
75%	108631.500000	26.500000	6.750000	0.000000	7.720000	6.900000	0.000000	6.100000	1.000000
max	479431.875444	64.290257	8.000000	5.000000	17.346885	24.345029	4.000000	25.967775	5.522658

- L'âge moyen des étudiants est d'environ 24 ans, avec une majorité comprise entre 20 et 26 ans.
- La pression académique présente une moyenne proche de 6, indiquant un niveau de stress académique globalement élevé.
- Le CGPA moyen est d'environ 7,2, avec une dispersion modérée entre les étudiants.
- La satisfaction des études montre une forte variabilité, traduisant des perceptions très différentes selon les étudiants.
- Les heures de travail/études varient fortement, allant de quelques heures à plus de 25 heures.
- La variable Depression présente une forte dispersion, indiquant des niveaux de dépression très variés au sein de la population étudiée.

```
#Vérifier le nombre de lignes dupliquées dans le dataset
set.duplicated().sum()
```

```
np.int64(2648)
```

=>Le dataset contient **2 648 observations dupliquées**, détectées à l'aide de la fonction **duplicated()**.

```
# Afficher le nombre de chaque valeur unique dans la colonne 'Depression'
print(set['Depression'].value_counts())
```

```
Depression
1.000000    18343
0.000000    13067
5.522658     817
Name: count, dtype: int64
```

- La variable cible **Depression** présente trois valeurs distinctes.
- La classe **1** est majoritaire (56,9 %), suivie de la classe **0** (40,5 %).
- Une valeur atypique **5.522658** représente environ **2,5 %** des observations

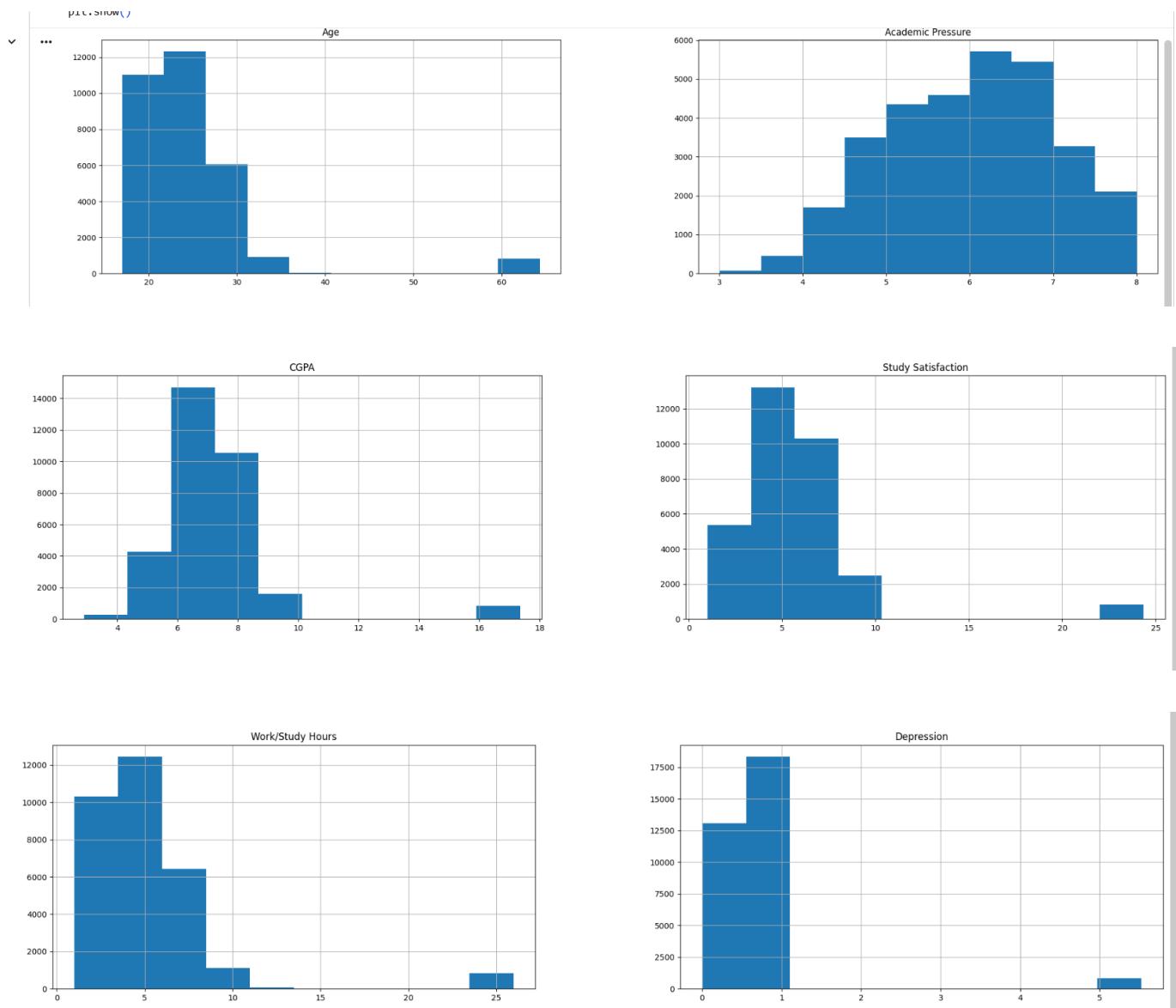
▶ # afficher les somme de valeurs nulls de chaque colonne  
`set.isna().sum()`

...	0
Gender	0
Age	1029
Profession	1030
Academic Pressure	1028
CGPA	0
Study Satisfaction	0
Sleep Duration	1022
Dietary Habits	0
Degree	0
Have you ever had suicidal thoughts ?	0
Work/Study Hours	1028
Financial Stress	0
Family History of Mental Illness	1031
Depression	817

- Plusieurs variables (**Age, Profession, Academic Pressure, Sleep Duration, Work/Study Hours, Family History of Mental Illness**) contiennent environ **1 000 valeurs manquantes**.
- La variable cible **Depression** présente **817 valeurs manquantes**, correspondant aux valeurs atypiques identifiées.
- Aucune valeur manquante n'est observée dans les autres variables principales du dataset.

### 3. Visualisation des données :

Des histogrammes ont été utilisés pour visualiser la distribution des variables numériques.



#### 4. Données pour le suivi temporel et la détection d'anomalies :

Pour répondre au cinquième objectif du projet, un jeu de données spécifique a été conçu afin de suivre l'évolution du bien-être étudiant dans le temps et d'identifier automatiquement des

périodes atypiques. Ce jeu de données adopte une approche **agrégée et temporelle**, où chaque ligne correspond à un mois donné, identifié par les variables *year* et *month*.

Les données sont obtenues par agrégation des soumissions individuelles des étudiants et permettent de calculer des indicateurs globaux tels que la pression académique moyenne, la durée moyenne de sommeil, le stress financier moyen, la satisfaction académique moyenne, ainsi que le taux de dépression et le score moyen de sévérité.

Shape: (25000, 13)													
	snapshot_id	year	month	avg_academic_pressure	avg_sleep_duration	avg_financial_stress	avg_study_satisfaction	depression_rate	high_risk_ratio	avg_severity_score	cluster_high_stress_ratio	cluster_stable_ratio	is_synthetic_anomaly
0	1	2018	1	5.549	2.848	1.842	5.872	0.4297	0.1195	0.4579	0.1468	0.4856	0
1	2	2018	1	6.141	2.493	2.061	5.721	0.6487	0.2691	0.7062	0.1967	0.4413	0
2	3	2018	1	5.277	2.900	1.942	6.583	0.4948	0.1329	0.5377	0.1624	0.4602	0
3	4	2018	1	5.326	3.029	2.061	6.628	0.3100	0.1235	0.2977	0.1501	0.5207	0
4	5	2018	1	6.199	2.671	1.995	5.956	0.4931	0.1446	0.5438	0.2327	0.4144	0

Des indicateurs issus de la segmentation des profils étudiants ont également été intégrés, notamment la proportion d'étudiants appartenant à des clusters à fort stress et à des profils stables, afin d'analyser l'évolution qualitative des groupes dans le temps.

Enfin, une variable d'anomalie synthétique est incluse à des fins d'évaluation, permettant de tester et comparer des méthodes de détection d'anomalies non supervisées telles que le Z-score et l'Isolation Forest.

## Chapitre 3 : Data Preparation

La préparation des données (*Data Preparation*) est une étape essentielle dans tout projet de *Machine Learning*. Elle consiste à transformer des données brutes en un format exploitable par les modèles d'apprentissage automatique.

Cette phase inclut plusieurs opérations telles que le nettoyage des données, le traitement des valeurs manquantes, la gestion des valeurs aberrantes, la normalisation, ainsi que la sélection et la transformation des variables.

=> Une bonne préparation des données permet d'améliorer significativement la performance, la robustesse et la capacité de généralisation des modèles.

En effet, même les algorithmes les plus avancés ne peuvent produire des résultats fiables si les données d'entrée sont de mauvaise qualité.

Nous allons maintenant vous présenter notre travail, en mettant l'accent sur les différentes étapes de la préparation des données réalisées dans le cadre de ce projet.

**NB :** La phase de préparation des données du deuxième dataset “`admin_feature_tracking_train_25000.csv`” assurant l’objectif **suivi temporel et la détection d’anomalies** a été minimale, les jeux de données utilisés étant déjà propres, complets et directement exploitables pour la phase de modélisation.

## **1. Traitement des valeurs nulles**

Nous vérifions tout d’abord les valeurs manquantes et décidons du traitement approprié selon la nature de chaque variable :

- **Remplacement par médiane** pour les variables numériques asymétriques.
- **Remplacement par moyenne** pour les variables numériques continues équilibrées.
- **Suppression** si le nombre de valeurs manquantes est très faible.

## **2. Traitement des lignes dupliquées**

```
#Vérifier le nombre de lignes dupliquées dans le dataset
set.duplicated().sum()
```

```
np.int64(2648)
```

```
#Suppression des lignes dupliquées dans le dataset
set.drop_duplicates(inplace=True)
```

```
#Verification de la suppression
set.duplicated().sum()
```

```
np.int64(0)
```

## **3.Traitement des features inutiles**

Le dataset comporte des colonnes inutiles, qui n’ont aucune valeur ajoutée pour le modèle, il suffit de les enlever afin d’éviter d’introduire du bruit inutile.

- **id** : n'apporte aucune information utile pour l'analyse ou la prédition, car elle ne contient pas de variable explicative réelle
- **Work Pressure / Job Satisfaction** : Toutes les lignes sont initialisées à 0, de plus le modèle se concentre sur des étudiants et pas des employés.
- **City** : Les noms de villes introduisent un risque de surapprentissage si le modèle les interprète comme des facteurs discriminants

#### Traitement des features inutiles

```
#supprimer la colonne id
set = set.drop(columns=["id"])

#supprimer la colonne Work Pressure
set = set.drop(columns=["Work Pressure"])

#supprimer la colonne Job Satisfaction
set = set.drop(columns=["Job Satisfaction"])

#supprimer la colonne City
set = set.drop(columns=["City"])
```

## 4. Traitement des valeurs manquantes par variable

Les valeurs manquantes ont été traitées individuellement pour chaque variable, en fonction de leur nature (numérique ou catégorielle) et de leur distribution :

- **Remplacement par la médiane (variables numériques)** :  
Les variables **Age**, **Work/Study Hours** et **Academic Pressure** présentent une distribution pouvant être asymétrique ou sensible aux valeurs extrêmes.

Les valeurs manquantes ont donc été remplacées par la médiane, afin de limiter l'influence des outliers et de préserver la distribution des données.

**Exemple:**

```
#Remplace les valeurs Null de colonne Academic Pressure par la médiane
set['Academic Pressure'].fillna(set['Academic Pressure'].median(), inplace=True)
```

- **Remplacement par la valeur la plus fréquente (variables catégorielles)** :  
Les variables **Profession**, **Family History of Mental Illness**, **Depression** et **Sleep Duration** sont de nature catégorielle.

Les valeurs manquantes ont été imputées par la modalité la plus fréquente (mode), ce qui permet de maintenir la cohérence des catégories dominantes dans le jeu de données.

## Exemple:

---

```
# Remplace les valeurs manquantes de la colonne 'Sleep Duration' par la valeur la plus fréquente de cette colonne.
set['Sleep Duration'] = set['Sleep Duration'].fillna(set['Sleep Duration'].mode()[0])
```

## Vérification du traitement des valeurs manquantes

Après l'application des méthodes de traitement des valeurs manquantes, une vérification a été effectuée afin de s'assurer de l'absence de valeurs nulles dans l'ensemble des variables du jeu de données.

=> Les résultats obtenus confirment que toutes les valeurs manquantes ont été correctement traitées.

---

```
#Vérifier les valeurs null
set.isna().sum()
```

	0
Gender	0
Age	0
Profession	0
Academic Pressure	0
CGPA	0
Study Satisfaction	0
Sleep Duration	0
Dietary Habits	0
Degree	0
Have you ever had suicidal thoughts ?	0
Work/Study Hours	0
Financial Stress	0
Family History of Mental Illness	0
Depression	0

---

```
dtype: int64
```

## 5. Traitement des Valeurs aberrantes

Le traitement des valeurs aberrantes constitue une étape importante de la préparation des données, visant à réduire l'impact des observations extrêmes susceptibles de fausser l'analyse statistique et les performances des modèles de *Machine Learning*. Cette étape

permet d'améliorer la qualité, la cohérence et la robustesse des données utilisées pour l'apprentissage.

### **Méthodes de traitement des valeurs aberrantes**

- Détection des valeurs aberrantes à l'aide de méthodes statistiques.
- Utilisation de l'intervalle interquartile (**IQR**) pour identifier les observations extrêmes.
- Remplacement des valeurs aberrantes par la **médiane** afin de réduire l'influence des valeurs extrêmes.
- Remplacement par la **moyenne** lorsque la distribution des données est équilibrée.
- Limitation des valeurs extrêmes par ajustement aux bornes définies (*clipping*) .
- Vérification du traitement à l'aide de statistiques descriptives et de visualisations graphiques.

#### **5.1. Traitement des valeurs aberrantes de la colonne Age**

1. les valeurs de la variable Age ont été arrondies à l'entier le plus proche ;
2. les observations pour lesquelles la valeur de Age est égale à 64 ont été supprimées, cette valeur ayant été identifiée comme aberrante ;
3. les valeurs aberrantes ont ensuite été détectées à l'aide de la méthode de l'intervalle interquartile (IQR) ;
4. un clipping a été appliqué afin de ramener les valeurs extrêmes dans les bornes inférieure et supérieure définies ;
5. La distribution finale a été vérifiée à l'aide de statistiques descriptives et d'un boxplot, confirmant l'efficacité du traitement.

```
# Arrondit les valeurs de la colonne 'Age' à l'entier le plus proche.
set['Age'] = set['Age'].round()

# Supprime toutes les lignes où la valeur de 'Age' est égale à 64 (valeur aberrante).
set.drop(set[set['Age'] == 64].index, inplace=True)

# Détection des valeurs aberrantes dans la colonne 'Age' à l'aide de l'IQR (Interquartile Range)
Q1 = set['Age'].quantile(0.25)
Q3 = set['Age'].quantile(0.75)
IQR = Q3 - Q1

lower = Q1 - 1.5 * IQR
upper = Q3 + 1.5 * IQR

outliers = set[(set['Age'] < lower) | (set['Age'] > upper)]
print("Valeurs aberrantes dans Age :")
print(outliers['Age'])

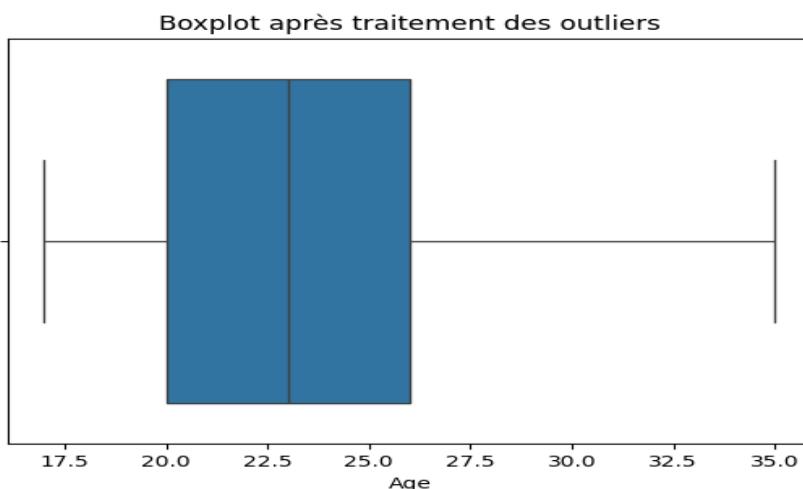
Valeurs aberrantes dans Age :
311      38.0
441      38.0
2417     36.0
2613     38.0
3315     38.0
...
31702    37.0
32175    36.0
32222    36.0
32502    36.0
33121    36.0
Name: Age, Length: 62, dtype: float64

#Limite inférieure et supérieure
set['Age'] = set['Age'].clip(lower, upper)
```

### Vérification de l'élimination des valeurs aberrantes dans la colonne age

Après le traitement des valeurs aberrantes, une analyse descriptive ainsi qu'un boxplot ont été réalisés afin de vérifier la distribution de la variable Âge. Les résultats montrent une distribution plus homogène, confirmant l'élimination efficace des valeurs extrêmes.

```
# vérification de l'élimination des valeurs aberrantes dans la colonne age
set['Age'].describe()
sns.boxplot(x=set['Age'])
plt.title("Boxplot après traitement des outliers")
plt.show()
```



## **5.2. Traitement des valeurs aberrantes de la colonne Academic Pressure**

1. l'analyse de la distribution de la variable Academic Pressure a révélé la présence de valeurs très peu fréquentes.
2. Les valeurs inférieures à 4 apparaissent avec un nombre d'occurrences très faible, ce qui les rend peu représentatives.
3. Ces valeurs ont été considérées comme atypiques et susceptibles d'introduire du bruit dans l'analyse.
4. Par conséquent, les observations correspondantes ont été supprimées afin d'améliorer la cohérence et la qualité du jeu de données.

---

```
# Supprime toutes les lignes où la valeur de 'Academic Pressure' est inférieure à 4.
set = set.drop(set[set['Academic Pressure'] < 4].index)
```

---

## **5.3. Traitement des valeurs aberrantes de la colonne Study Satisfaction**

---

```
# Affiche le nombre de lignes où la valeur de 'Study Satisfaction' est supérieure à 20.
set[set['Study Satisfaction'] > 20].shape[0]
```

---

545

---



---

```
# Remplacer les outliers > 10 par la médiane
median_val = set['Study Satisfaction'].median()
set.loc[(set['Study Satisfaction'] > 10), 'Study Satisfaction'] = median_val
```

---

1. l'analyse de la distribution de la variable Study Satisfaction a mis en évidence la présence de valeurs très élevées et peu fréquentes.
2. ces valeurs ont été considérées comme aberrantes et susceptibles de biaiser l'analyse.
3. Les observations présentant une valeur supérieure à 10 ont été identifiées comme des valeurs extrêmes.
4. Afin de limiter leur impact, ces valeurs aberrantes ont été remplacées par la médiane de la variable.

## **5.4. Traitement des valeurs aberrantes de la colonne Work/study Hours**

```

1 # Supprimer les valeurs qui sont supp à 24h
2 set = set[set['Work/Study Hours'] <= 24]

1
2 Q111 = set['Work/Study Hours'].quantile(0.25)
3 Q333 = set['Work/Study Hours'].quantile(0.75)
4 IQR1 = Q333 - Q111
5 borne_inf1 = Q111 - 1.5 * IQR1
6 borne_sup1 = Q333 + 1.5 * IQR1

7 set = set[(set['Work/Study Hours'] >= borne_inf1) & (set['Work/Study Hours'] <= borne_sup1)]
8 set['Work/Study Hours'] = set['Work/Study Hours'].clip(upper=10)

```

1. une analyse préliminaire de la variable Work/Study Hours a permis d'identifier des valeurs incohérentes dépassant le seuil maximal réaliste de 24 heures par jour ;
2. les observations présentant des valeurs supérieures à 24 ont été supprimées du jeu de données ;
3. la méthode de l'intervalle interquartile (IQR) a ensuite été utilisée pour détecter les valeurs aberrantes restantes ;
4. les observations situées en dehors des bornes inférieure et supérieure définies par l'IQR ont été filtrées ;
5. Enfin, un clipping a été appliqué afin de limiter les valeurs extrêmes restantes à une borne maximale, garantissant une distribution plus cohérente de la variable.

## **5.5. Traitement des valeurs aberrantes de la colonne CGPA**

```

# Calcul des bornes via la méthode IQR
Q1 = set['CGPA'].quantile(0.25)
Q3 = set['CGPA'].quantile(0.75)
IQR = Q3 - Q1
borne_inf = Q1 - 1.5 * IQR
borne_sup = Q3 + 1.5 * IQR

# Clipping des outliers
CGPA_clipped = set['CGPA'].clip(lower=borne_inf, upper=borne_sup)

# Mise à l'échelle [0, 4] (système international)
set['CGPA'] = 4 * (CGPA_clipped - CGPA_clipped.min()) / (CGPA_clipped.max() - CGPA_clipped.min())

```

1. les bornes inférieure et supérieure ont été calculées à l'aide de la méthode de l'intervalle interquartile (IQR) afin d'identifier les valeurs aberrantes ;
2. un clipping a été appliqué pour limiter les valeurs extrêmes de la variable CGPA à l'intérieur de ces bornes ;
3. après le traitement des valeurs aberrantes, la variable CGPA a été mise à l'échelle dans l'intervalle [0, 4], conformément au système international de notation ;

4. Cette transformation permet d'assurer la cohérence des données et de faciliter leur utilisation dans les modèles de Machine Learning.

## **6. Encodage des colonnes catégorielles**

### **6.1. Encodage binaire (Label Encoding binaire) :**

- la variable **Gender** a été encodée de manière binaire (*Male = 0, Female = 1*) ;
- la variable **Have you ever had suicidal thoughts ?** a été transformée en variable binaire (*No = 0, Yes = 1*).
- La variable **Family History of Mental Illness** a également été encodée de manière binaire (*No = 0, Yes = 1*).

→ ce type d'encodage est adapté aux variables à deux modalités.

### **6.2. Encodage nominal (Label Encoding) :**

- la variable **Profession** a été encodée en attribuant un identifiant numérique unique à chaque catégorie ;
- La variable **Degree** a été traitée de la même manière, chaque diplôme étant remplacé par un entier distinct.

→ cet encodage permet de convertir des variables catégorielles nominales en valeurs numériques exploitables.

### **6.3. Encodage ordinal (Ordinal Encoding) :**

- La variable **Sleep Duration** a été encodée selon un ordre croissant de durée du sommeil.
- La variable **Dietary Habits** a été encodée selon la qualité croissante des habitudes alimentaires.
- La variable **Financial Stress** a été encodée selon un niveau d'intensité croissant (*Low → High*).

→ cet encodage permet de préserver l'ordre logique entre les catégories.

...	Gender	Age	Profession	Academic Pressure	CGPA	Study Satisfaction	\
0	1	26.0	2	4.9375	2.392617	9.5	
1	1	26.0	2	5.2500	2.761745	8.1	
2	0	27.0	2	4.6250	1.681208	6.3	
3	0	33.0	1	5.1250	2.721477	9.4	
4	0	29.0	2	6.4375	2.741611	7.4	
Sleep Duration Dietary Habits Degree \							
0		2	2	13			
1		3	3	7			
2		4	3	3			
3		1	3	22			
4		3	3	23			
Have you ever had suicidal thoughts ? Work/Study Hours Financial Stress \							
0			0	6.2			1
1			0	5.4			1
2			0	1.0			2
3			0	1.0			2
4			0	5.4			3
Family History of Mental Illness Depression							
0			0	0.0			
1			0	0.0			
2			0	0.0			
3			0	0.0			
4			0	0.0			

## 7. Analyse en Composantes Principales (PCA / ACP)

L'Analyse en Composantes Principales (PCA – Principal Component Analysis) est une méthode statistique de réduction de dimension. Elle permet de transformer un ensemble de variables corrélées en un nouveau jeu de variables non corrélées appelées composantes principales (PC).

## Objectifs de la PCA

- Réduire le nombre de variables tout en conservant un maximum d'information
- Identifier les facteurs dominants expliquant la variabilité des données
- Faciliter la visualisation et l'interprétation des relations entre variables
- Déetecter les variables les plus influentes dans le phénomène étudié

Chaque composante principale est une combinaison linéaire des variables originales et explique une part décroissante de la variance totale.

### 7.1 Tableau de la variance expliquée par composante

VARIANCE EXPLIQUÉE par composante :			
PC	Variance	Cumulée	Signification
PC 1	0.2754	0.2754	Très importante
PC 2	0.0829	0.3583	Importante
PC 3	0.0825	0.4408	Moyennement importante
PC 4	0.0799	0.5207	Moyennement importante
PC 5	0.0764	0.5971	Moyennement importante
PC 6	0.0739	0.6710	Peu importante
PC 7	0.0712	0.7421	Peu importante
PC 8	0.0630	0.8051	Peu importante
PC 9	0.0601	0.8652	Peu importante
PC10	0.0528	0.9180	Peu importante
PC11	0.0458	0.9638	Peu importante
PC12	0.0362	1.0000	Peu importante

Le tableau de la variance expliquée montre que la première composante principale (PC1) concentre à elle seule 27,54 % de la variance totale, ce qui en fait la composante dominante et le principal axe de structuration des données.

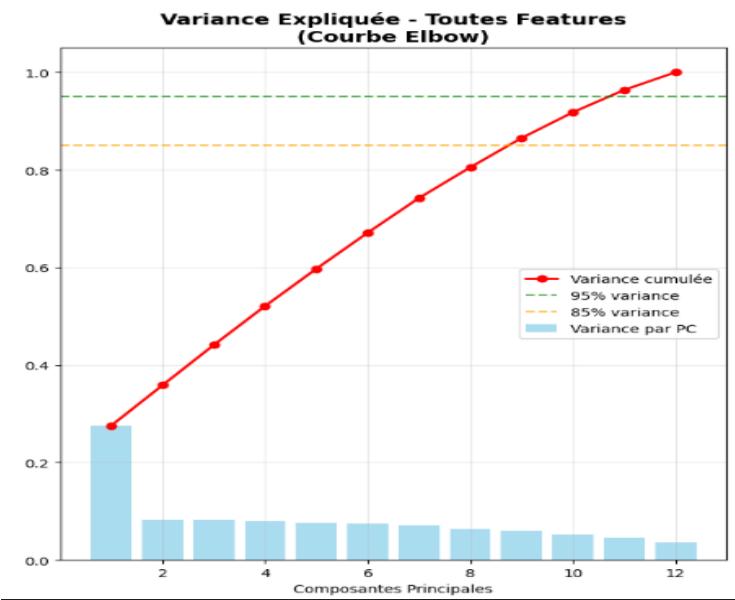
La deuxième composante (PC2) apporte une contribution supplémentaire de 8,29 %, portant la variance cumulée des deux premières composantes à 35,83 %.

Les composantes suivantes, de PC3 à PC5, expliquent une part modérée de la variance, tandis qu'à partir de PC6, l'apport informationnel devient de plus en plus faible.

Les dernières composantes (PC9 à PC12) contribuent très peu à la variance totale et peuvent être considérées comme marginales.

=> Ainsi, ces résultats indiquent qu'il est pertinent de conserver uniquement les premières composantes principales, ce qui permet de simplifier le modèle tout en préservant l'essentiel de l'information contenue dans les données.

## 7.2 Courbe Elbow – Variance expliquée cumulée



La courbe Elbow montre une forte contribution de la première composante principale, suivie d'une diminution progressive de l'apport des composantes suivantes, avec un point de coude visible autour de PC8–PC9.

Ainsi, le choix de 9 composantes représente un bon compromis entre simplicité et performance ( $\approx 85\%$  de variance expliquée), tandis que 11 composantes permettent de conserver presque toute l'information ( $\approx 95\%$ ).

## 7.3 Synthèse et recommandations finales

### 9. SYNTHÈSE ET RECOMMANDATIONS :

TOTAL FEATURES ANALYSÉES : 12

VARIANCE PC1 + PC2 : 35.8%

COMPOSANTES pour 85% variance : 9

COMPOSANTES pour 95% variance : 11

### TOP 5 VARIABLES les plus importantes :

1. Study Satisfaction (impact: 0.789)
2. Academic Pressure (impact: 0.756)
3. Sleep Duration (impact: 0.688)
4. Family History of Mental Illness (impact: 0.647)
5. Financial Stress (impact: 0.579)

### INTERPRÉTATION DES RÉSULTATS :

- PC1 explique 27.5% de la variance → Composante principale
- Les 2 premières PC expliquent 35.8% de la variance
- Study Satisfaction est la variable la plus influente

### RECOMMANDATIONS POUR LA SUITE :

- Pour 85% de variance : 9 composantes
- Pour 95% de variance : 11 composantes
- Ratio optimal qualité/performance : 9 composantes

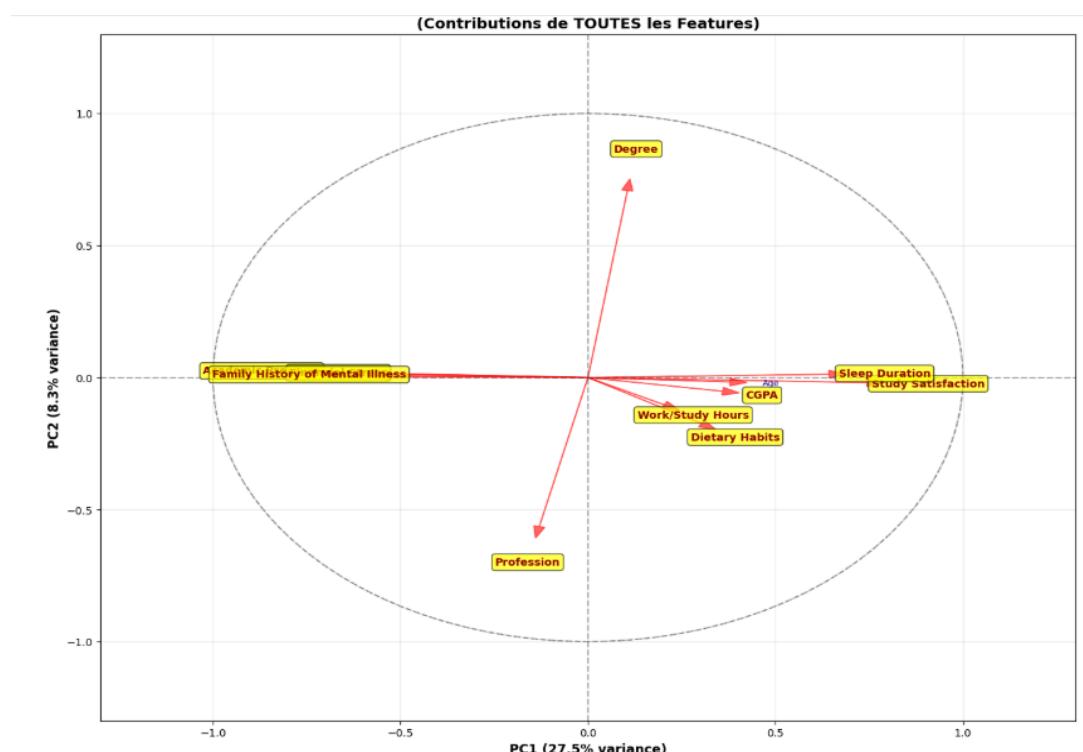
### Dans le cadre de cette analyse on a pu distinguer que :

- Les deux premières composantes principales (PC1 et PC2) expliquent à elles seules **35,8 % de la variance totale**, ce qui montre que l'essentiel de l'information est

concentré sur les premiers axes de l'analyse.

- Le seuil de **85 % de variance expliquée** est atteint avec **9 composantes principales**, qui ont été **retenues pour la suite de l'étude** afin de garantir un bon équilibre entre réduction de la dimensionnalité et conservation de l'information.
- Bien que **11 composantes** permettent d'expliquer environ **95 % de la variance**, ce choix n'a pas été retenu afin de privilégier la simplicité, la lisibilité et la parcimonie du modèle.
- Les variables les plus influentes dans la structuration des données sont la **satisfaction académique (Study Satisfaction)**, la **pression académique (Academic Pressure)**, la **durée du sommeil (Sleep Duration)**, les **antécédents familiaux de troubles mentaux (Family History of Mental Illness)** et le **stress financier (Financial Stress)**, soulignant l'importance des dimensions académiques et psychosociales.

#### 7.4. Cercle des corrélations (contributions des variables)



**Le cercle de corrélation** permet d'interpréter les relations entre les variables initiales en analysant leur projection sur les deux premières composantes principales et en mettant en évidence leurs corrélations, oppositions et contributions à la structure globale des données donc on peut noter que:

- Les variables orientées dans la même direction, comme Study Satisfaction et Sleep Duration, sont positivement corrélées et évoluent de manière similaire.
- Les variables opposées sur le cercle, par exemple Study Satisfaction et Academic Pressure, présentent une corrélation négative, indiquant des effets inverses.
- Les variables proches du cercle, telles que Financial Stress ou Family History of Mental Illness, contribuent fortement aux composantes principales.
- Les variables proches du centre du cercle ont une influence plus faible sur PC1/PC2.
- Globalement, le cercle de corrélation met en évidence une opposition entre bien-être académique et facteurs de stress, facilitant l'interprétation des relations entre les variables.

## **8. Standardisation**

RECOMMANDATIONS POUR LA SUITE :

- Pour 85% de variance : 9 composantes
- Pour 95% de variance : 11 composantes
- Ratio optimal qualité/performance : 9 composantes

Toutes les variables explicatives ont été standardisées à l'aide du StandardScaler afin de leur donner la même importance lors de l'entraînement du modèle. Cette étape est indispensable pour les algorithmes sensibles à l'échelle des données (SVM, régression logistique, KNN...).

**Conclusion:** Une fois le traitement des valeurs manquantes et aberrantes effectué, le jeu de données est prêt pour la phase de modélisation.

✓ DONNÉES OPTIMISÉES

✓ QUALITÉ AMÉLIORÉE :

- Les distributions sont maintenant plus représentatives
- Réduction du bruit dans les données
- Meilleure fiabilité pour l'entraînement des modèles

✓ IMPACT SUR LA MODÉLISATION :

- Modèles plus robustes et généralisables
- Réduction du risque de sur-apprentissage (overfitting)

## **Chapitre 4: Modélisation**

La phase de modélisation consiste à tester et comparer plusieurs modèles afin de répondre aux objectifs data science définis précédemment.

Les données ont été divisées en un **ensemble d'entraînement (70 %)** et un **ensemble de test (30 %)** à l'aide de la fonction `train_test_split`.

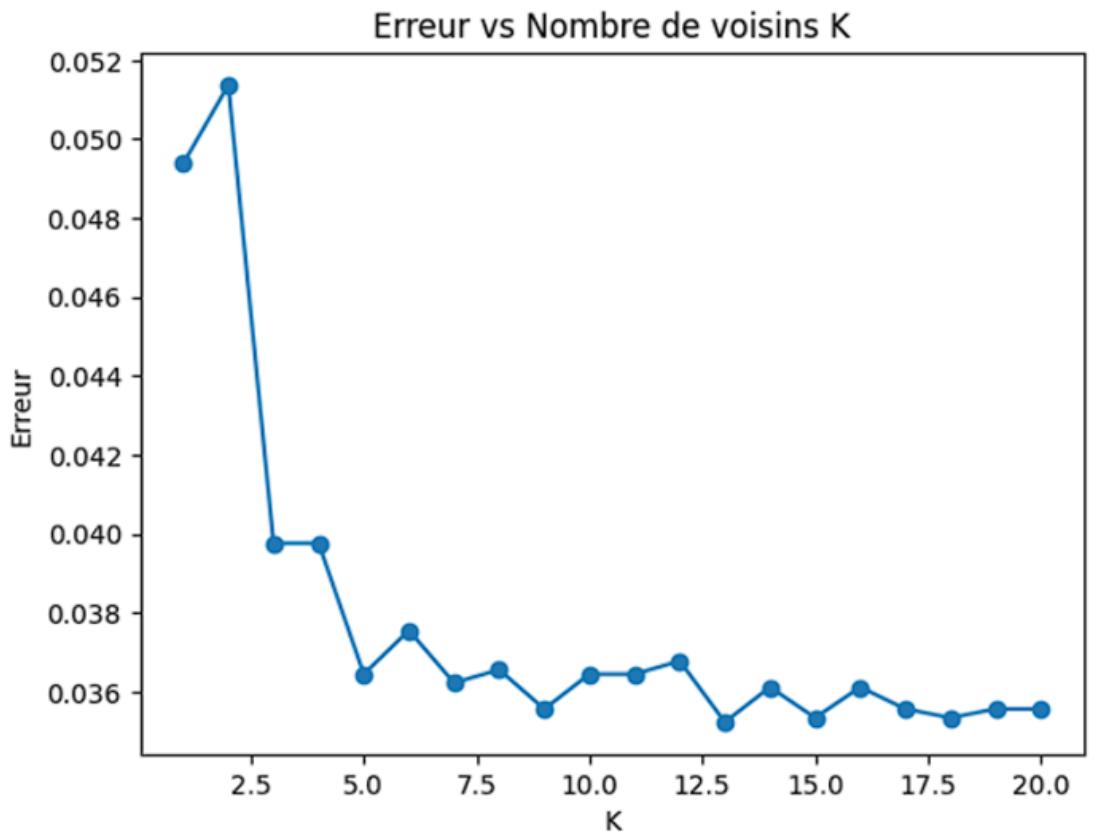
Les variables explicatives ont été **standardisées à l'aide du StandardScaler** afin de les ramener sur une même échelle : cette étape est indispensable pour les algorithmes sensibles à l'échelle des données, tels que le SVM, la régression logistique et le KNN.

## 1. Objectif 1 : Détection des étudiants à risque de dépression

- **Support Vector Machine (SVM)** cherche à définir une frontière de séparation optimale entre les classes, et il a été retenu pour sa capacité à bien fonctionner sur des problèmes de classification binaire avec des frontières complexes.
- **Random Forest** repose sur un ensemble d'arbres de décision (`n_estimators=200`) combinés entre eux, ce qui permet de modéliser des relations non linéaires complexes.
- **XGBoost** est un algorithme de boosting qui construit les modèles de manière séquentielle en corrigeant les erreurs précédentes avec **nombre d'arbres (n\_estimators)** et la **profondeur maximale des arbres (max\_depth)**,
- **Régression logistique** est un modèle de classification binaire qui estime la probabilité qu'un étudiant soit en situation de dépression à partir de ses caractéristiques, et elle a été choisie comme modèle de référence pour sa simplicité et sa bonne interprétabilité.
- **Decision Tree** est un modèle basé sur des règles de décision successives, ce qui permet de comprendre facilement le processus de prédiction. Il a été choisi pour sa capacité à fournir des résultats interprétables et intuitifs. Les hyperparamètres principaux sont la **profondeur maximale de l'arbre (max\_depth)**, le **nombre minimum d'échantillons par feuille (min\_samples\_leaf)** et le **critère de division (gini)**.
- **K-Nearest Neighbors (KNN)** est un algorithme basé sur la proximité entre les observations, qui classe un étudiant en fonction des étudiants les plus similaires.

Remarque : Le nombre de voisins **K** a été choisi en fonction du minimum de l'erreur observée sur le jeu de validation.

...



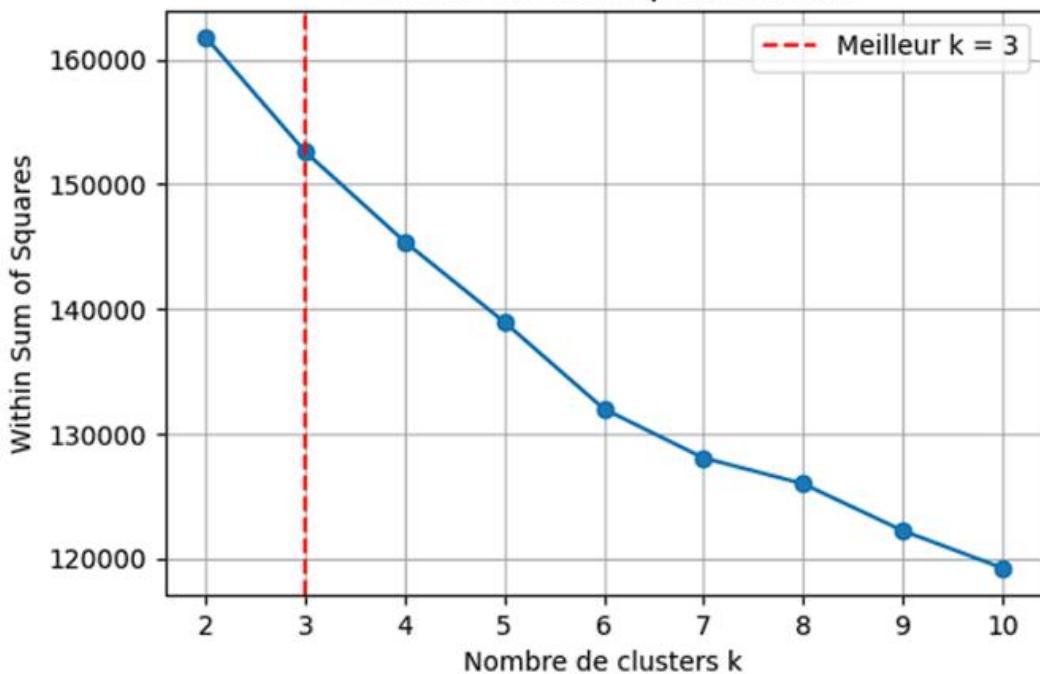
## 2.Objectif 2 : Segmentation des profils d'étudiants

### K-means :

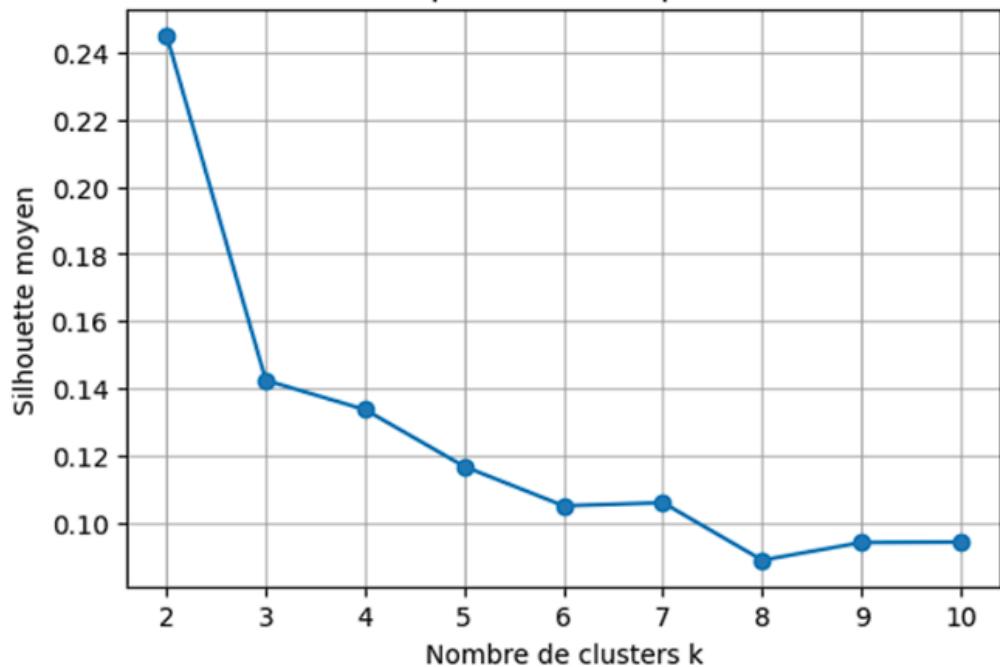
Avant d'appliquer l'algorithme de **K-means**, deux méthodes ont été utilisées afin de déterminer le nombre optimal de clusters :

- La méthode du coude (Elbow Method) permet d'analyser l'évolution de l'inertie en fonction du nombre de clusters et d'identifier un point à partir duquel l'amélioration devient marginale.
- Le score de silhouette a ensuite été utilisé pour évaluer la qualité de la séparation entre les clusters, en mesurant à quel point les observations sont bien regroupées dans leur cluster par rapport aux autres.

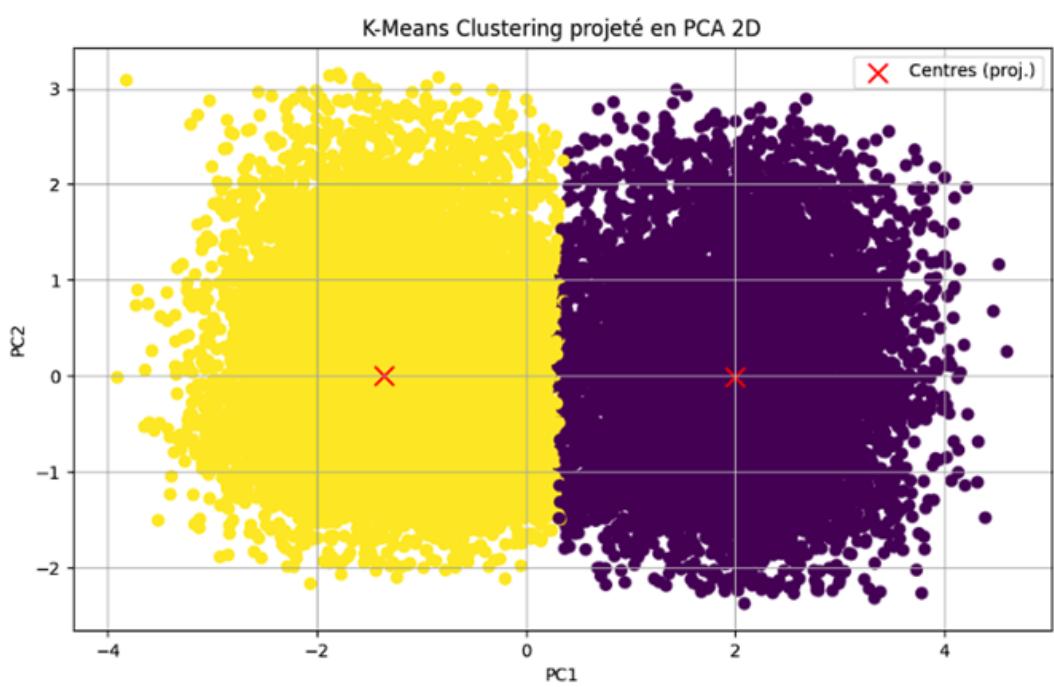
### Méthode du coude pour K-Means



### Choix de k par Silhouette pour K-Means



La méthode Silhouette évalue objectivement la qualité du clustering en mesurant à la fois la cohésion interne et la séparation entre groupes, ce qui en fait un critère plus fiable que la méthode du coude, souvent subjective et visuelle. Comme le score silhouette est maximal pour  $k = 2$ , cela signifie que les données forment deux groupes bien distincts et mieux structurés que pour  $k = 3$ .



=> Les points sont organisés en zones de couleurs dominantes, ce qui montre que K-Means parvient à regrouper les étudiants en profils relativement homogènes dans l'espace latent de la PCA.

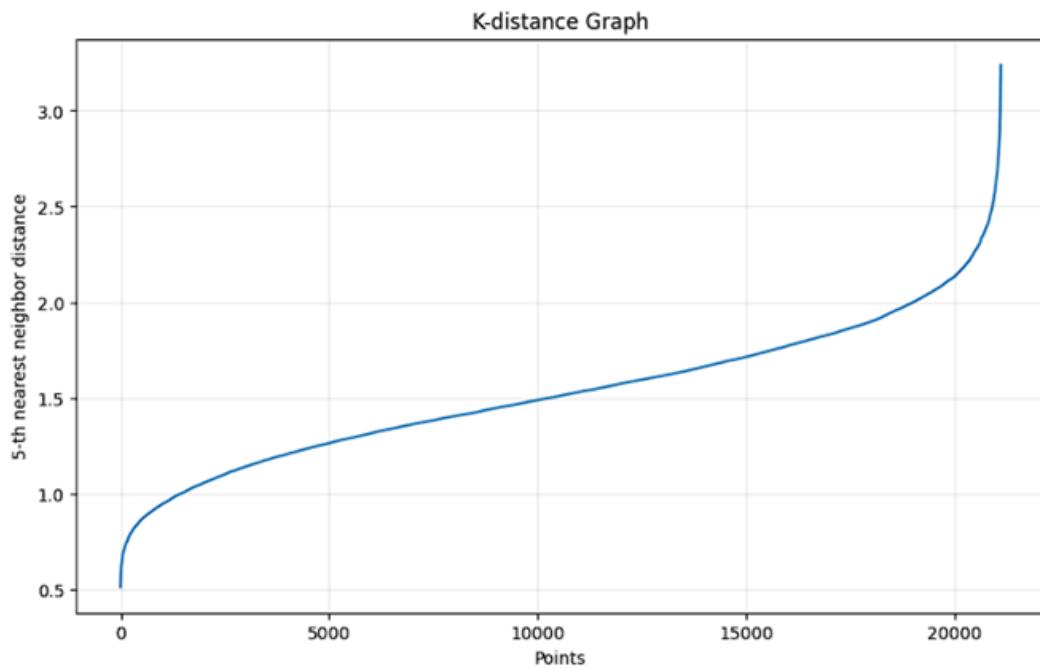
=> L'axe horizontal (PC1) sépare globalement des profils différents (par exemple, à gauche des profils plus "à risque" et à droite des profils plus "stables", selon les analyses précédentes de la PCA).

=> Les centres des clusters (croix rouges) se situent au cœur des zones de densité de chaque couleur : ils représentent les profils moyens typiques de chaque groupe d'étudiants.

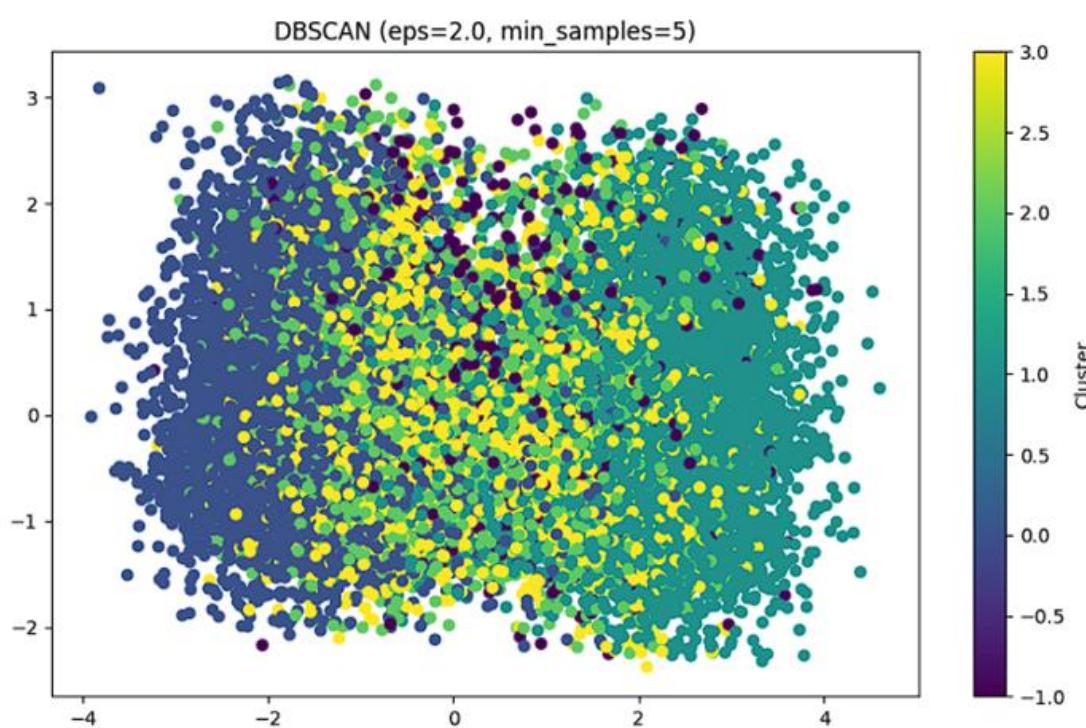
### DBSCAN :

un algorithme de clustering basé sur la densité, utilisé pour identifier des groupes d'étudiants de formes variées ainsi que des observations isolées considérées comme du bruit, sans avoir à fixer le nombre de clusters à l'avance.

Ce graphique suivant permet de choisir la valeur de **eps** en repérant le point où la distance augmente brusquement.

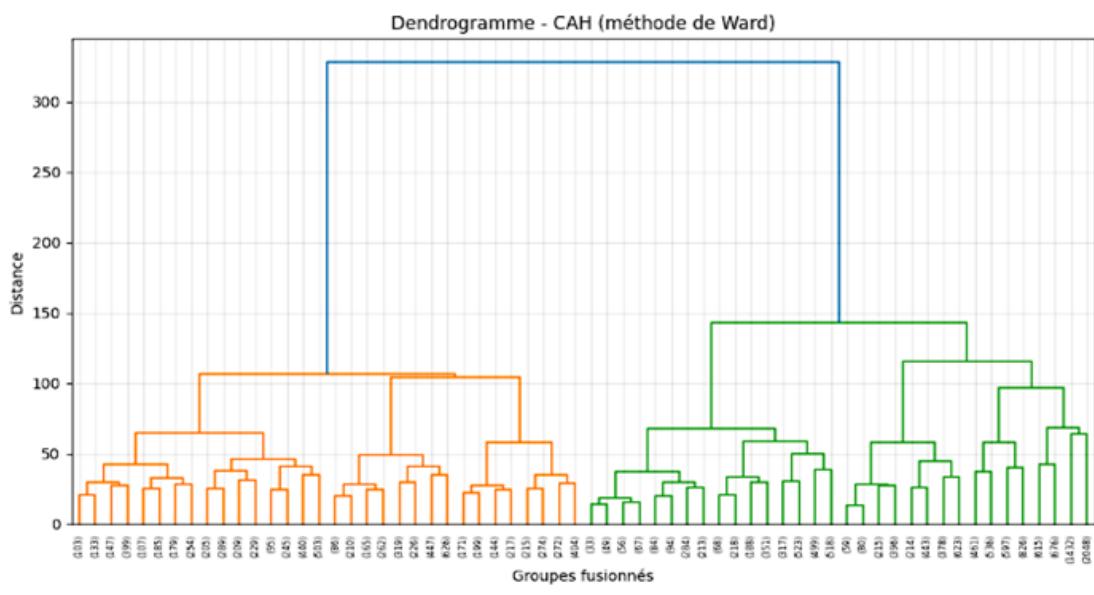


Le graphique montre la répartition des clusters identifiés par DBSCAN projetés en 2D via PCA.



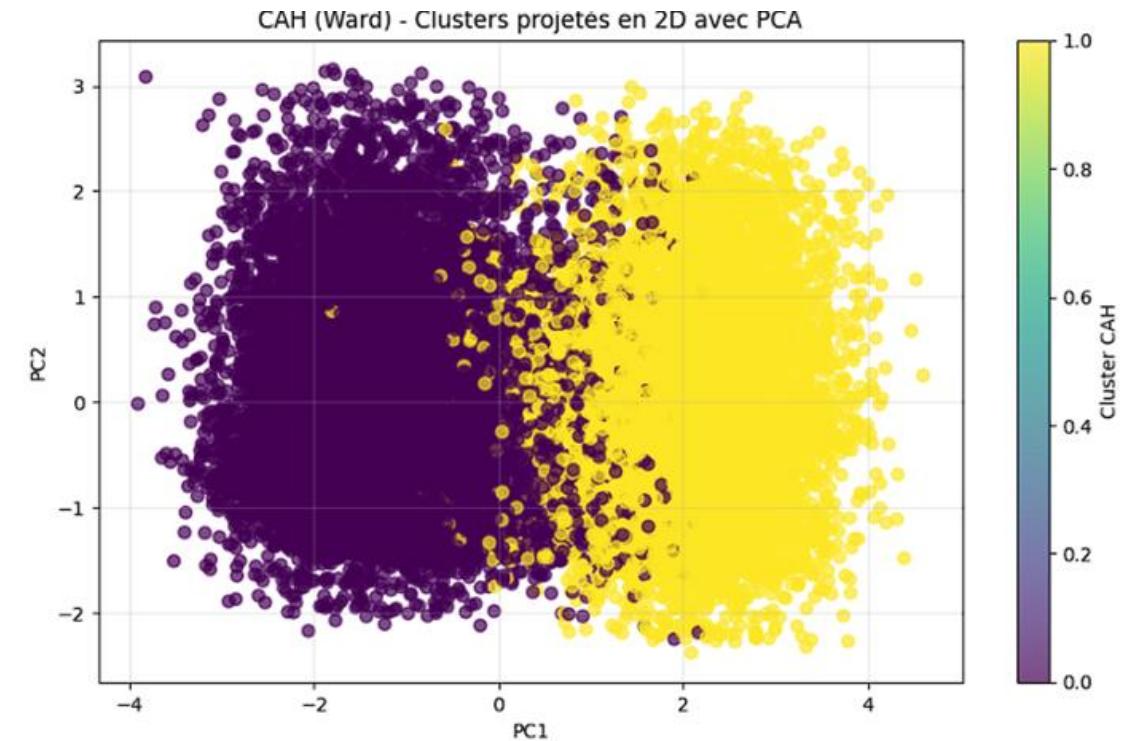
### CAH :

Construit une hiérarchie de clusters en fusionnant itérativement les observations les plus proches selon une mesure de distance et un critère de liaison, permettant d'analyser la structure globale des données à l'aide d'un dendrogramme ci dessous :



Le dendrogramme obtenu avec la CAH montre une séparation claire des observations en **deux groupes principaux**, ce qui confirme le choix d'un nombre de clusters égal à 2. Cette segmentation permet de distinguer deux profils d'étudiants aux caractéristiques différentes.

Le graphique issu de la CAH permet de visualiser la répartition des étudiants selon les clusters identifiés et confirme la séparation en deux groupes distincts, cohérente avec l'analyse du dendrogramme.



### 3.Objectif 3 : recommandation des actions de bien être :

#### 1ère méthode :

Le système de recommandation s'appuie sur les **résultats du clustering** obtenu lors de l'objectif 2.

Chaque étudiant est d'abord **associé à un cluster**, représentant un profil global (par exemple, risque élevé ou faible).

Les **variables individuelles à risque** (faible durée de sommeil, forte pression académique, stress financier élevé) sont ensuite analysées.

Des **règles conditionnelles** sont appliquées en fonction de ces caractéristiques afin de générer des recommandations adaptées à chaque étudiant.

	Academic Pressure	Sleep Duration	Financial Stress	Depression	cluster	cluster_recommendation
0	4.9375	2	1	0.0	3	Aucune intervention nécessaire (profil stable)
1	5.2500	3	1	0.0	3	Aucune intervention nécessaire (profil stable)
2	4.6250	4	2	0.0	1	Maintien des bonnes habitudes
3	5.1250	1	2	0.0	3	Aucune intervention nécessaire (profil stable)
4	6.4375	3	3	0.0	3	Aucune intervention nécessaire (profil stable)
5	6.0000	3	1	0.0	3	Aucune intervention nécessaire (profil stable)
6	5.8750	2	3	1.0	0	Activité sportive, gestion du stress, améliora...
7	6.2500	2	1	1.0	2	Soutien psychologique, gestion du stress, amél...
8	6.6250	1	2	1.0	0	Activité sportive, gestion du stress, améliora...
9	5.8125	4	1	0.0	3	Aucune intervention nécessaire (profil stable)

#### 2ème méthode :

Un algorithme SHAP est utilisé pour identifier les **variables qui contribuent le plus au risque de dépression pour chaque étudiant**. À partir de ces contributions, il est possible de cibler précisément les facteurs problématiques au niveau individuel.

\*\*\* Recommandations pour l'étudiant index 7 \*\*\*

- La caractéristique 'Financial Stress' (valeur = 2.0) contribue fortement à augmenter la probabilité de dépression (SHAP = 0.234). Dans ton dataset, cette variable est positif
- La caractéristique 'Sleep Duration' (valeur = 3.0) contribue fortement à augmenter la probabilité de dépression (SHAP = 0.156). Dans ton dataset, cette variable est négatif

### 4.Objectif 4 : Prédiction du niveau de sévérité de la dépression.

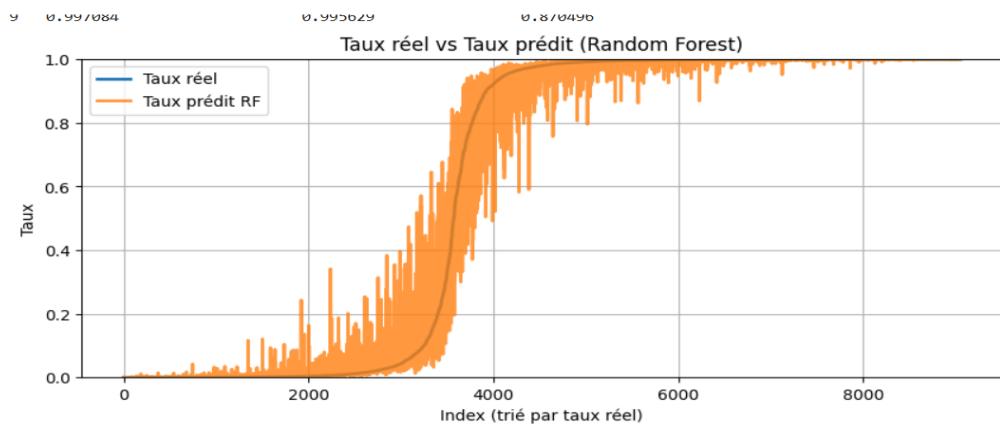
#### Étape 1 : Estimation du taux de dépression

Plutôt que d'utiliser directement la classe prédite (0 ou 1), la **probabilité de la classe positive** a été extraite. Cette probabilité est interprétée comme un **taux de dépression**, représentant le niveau de risque estimé pour chaque étudiant.

- Ce taux sert de **variable cible continue** pour les modèles de régression

## Étape 2 : Modélisation par régression

- **Régression linéaire** : sur le taux de dépression afin de modéliser une relation simple entre les caractéristiques des étudiants et le niveau de sévérité estimé.
- **Random Forest Regressor** : capture des relations non linéaires plus complexes entre les variables explicatives et le taux de dépression. Il est entraîné avec **n\_estimators = 200**



## 5. Objectif 5 : Suivi temporel des indicateurs et alertes

L'objectif de ce travail est de détecter automatiquement des situations anormales au niveau administratif à partir d'indicateurs agrégés liés au bien-être et au risque (pression académique, stress financier, sommeil, dépression, ratios de risque, etc.). En l'absence de données étiquetées, le problème est abordé comme une tâche de détection d'anomalies non supervisée, permettant à l'administration d'identifier les mois présentant des comportements atypiques, tels qu'un taux anormalement élevé d'étudiants en situation de dépression, et nécessitant une attention particulière.

### 1ère méthode : Z-score

est une méthode statistique basée sur la distance à la moyenne. Pour chaque variable, un score standardisé est calculé afin de mesurer l'écart d'une observation par rapport au comportement moyen. Une observation est considérée comme anormale lorsqu'au moins une variable dépasse un seuil critique, indiquant une valeur extrême

### 2ème méthode : Isolation Forest

est un algorithme non supervisé basé sur un ensemble d'arbres de décision construits aléatoirement. Son principe repose sur le fait que les anomalies, étant rares et différentes, sont isolées plus rapidement que les observations normales. Un score d'anomalie est attribué à chaque observation en fonction de la profondeur moyenne nécessaire pour l'isoler dans les arbres.

## Chapitre 5 : Evaluation

La phase Évaluation consiste à examiner et valider les résultats obtenus lors de la phase de modélisation afin de s'assurer qu'ils répondent correctement aux objectifs métiers définis au départ. Cette étape permet de vérifier la qualité, la pertinence et la fiabilité des modèles, d'identifier d'éventuelles limites, et de déterminer si les résultats sont suffisamment satisfaisants pour être interprétés ou déployés.

### Objectif 1 : Détection des étudiants à risque de dépression

#### 1.1. Classification par SVM

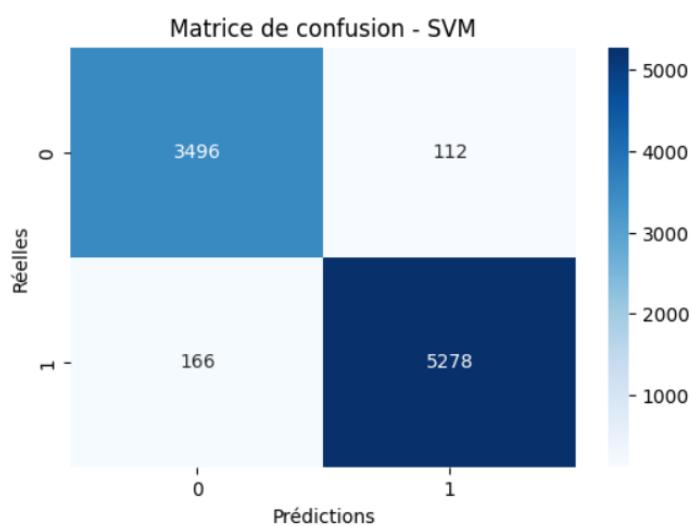
Résultats SVM :

Accuracy : 0.9692885550154662

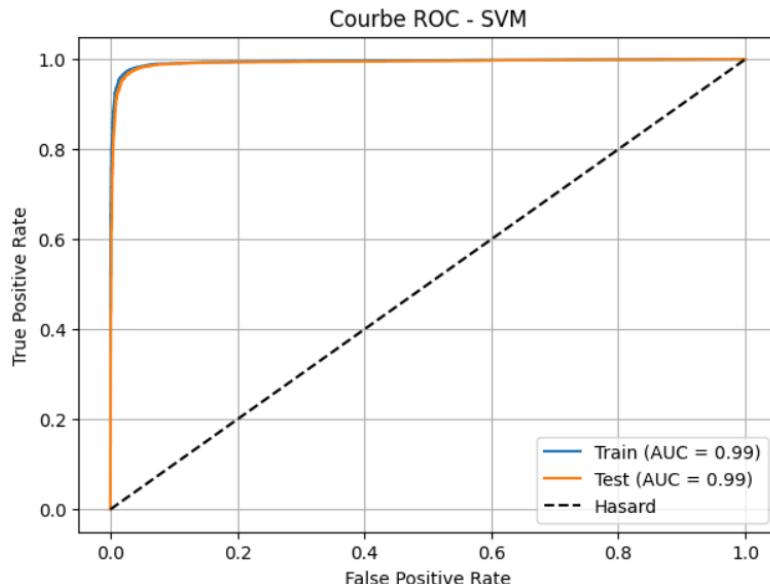
Rapport de classification :

	precision	recall	f1-score	support
0.0	0.95	0.97	0.96	3608
1.0	0.98	0.97	0.97	5444
accuracy			0.97	9052
macro avg	0.97	0.97	0.97	9052
weighted avg	0.97	0.97	0.97	9052

=> Ces résultats confirment que le modèle SVM, associé à la réduction dimensionnelle par PCA, est capable de prédire efficacement le risque de dépression, en équilibrant correctement la détection des deux classes malgré un léger déséquilibre dans le dataset.



=> La matrice de confusion du modèle **SVM** montre une majorité de prédictions correctes, avec **3496 vrais négatifs** et **5278 vrais positifs**, et un nombre limité d'erreurs (**112 faux positifs** et **166 faux négatifs**), ce qui traduit une bonne capacité du modèle à distinguer efficacement les deux classes.



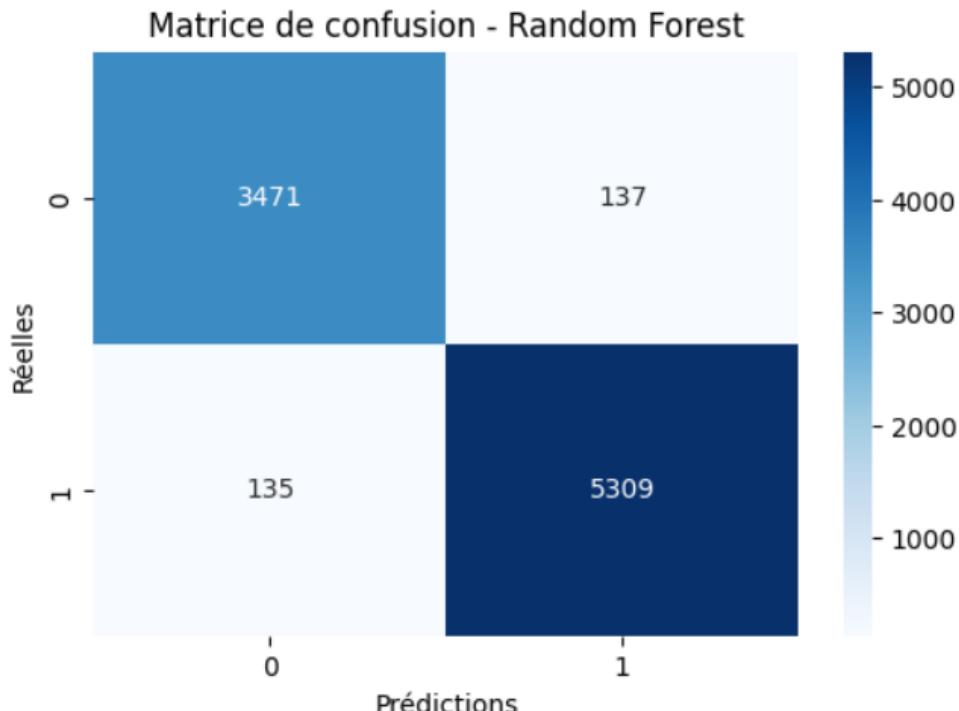
=> Le modèle SVM affiche une courbe ROC très performante, caractérisée par une **AUC de 0,99** aussi bien en apprentissage qu'en test, traduisant une capacité robuste à distinguer les deux classes.

## **1.2. Classification par Random Forest**

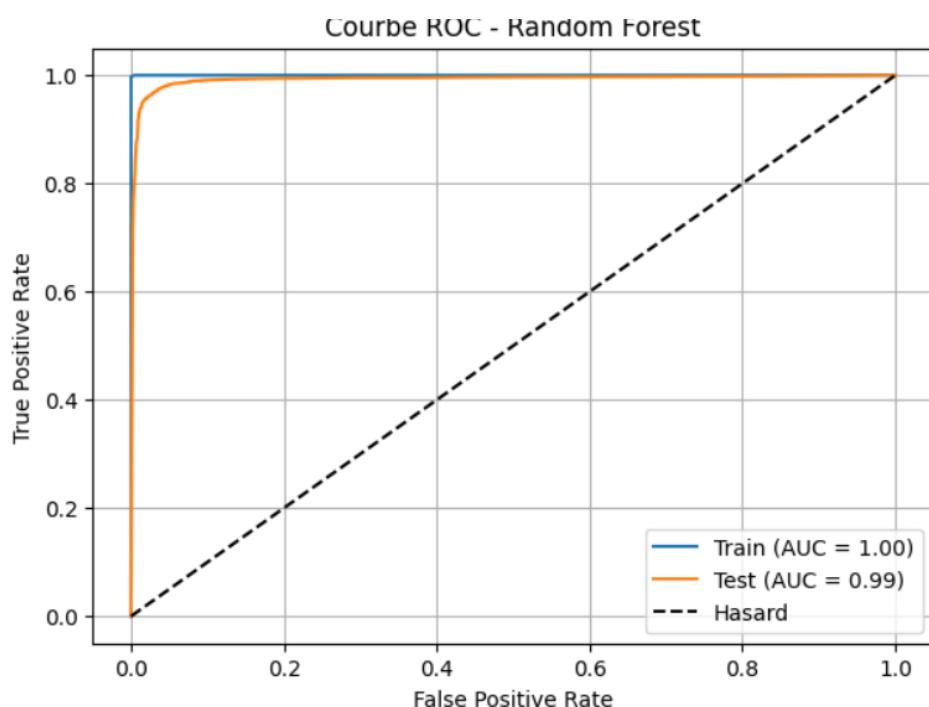
Accuracy Random Forest: 0.9699513919575784				
	precision	recall	f1-score	support
0.0	0.96	0.96	0.96	3608
1.0	0.97	0.98	0.98	5444
accuracy			0.97	9052
macro avg	0.97	0.97	0.97	9052
weighted avg	0.97	0.97	0.97	9052

---

=> Le modèle Random Forest a montré d'excellentes performances pour la prédiction du risque de dépression. Le modèle présente un équilibre confirmant sa capacité à généraliser et à capturer les relations complexes entre les facteurs psychologiques, académiques et environnementaux des étudiants.



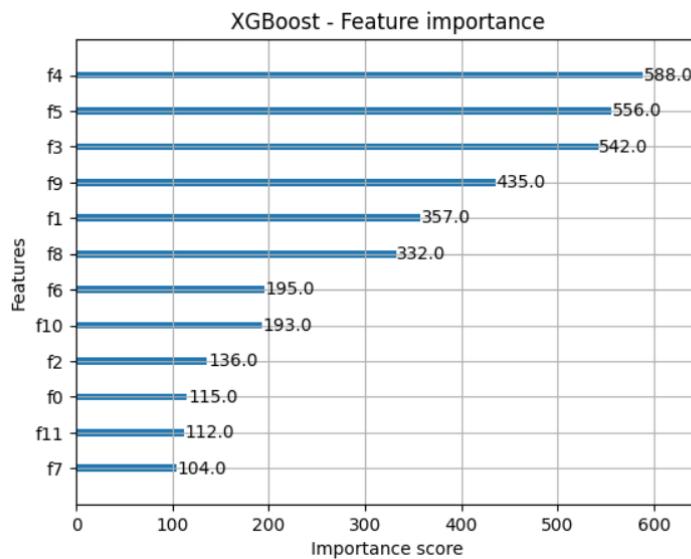
=> Le modèle *Random Forest* présente des résultats très satisfaisants, caractérisés par un nombre élevé de prédictions correctes (**3471 vrais négatifs et 5309 vrais positifs**) et un faible taux d'erreurs (**137 faux positifs et 135 faux négatifs**).



=> La courbe ROC du modèle Random Forest met en évidence des performances remarquables, avec une **AUC de 1,00 en entraînement** et **0,99 en test**, confirmant la robustesse et la fiabilité du modèle.

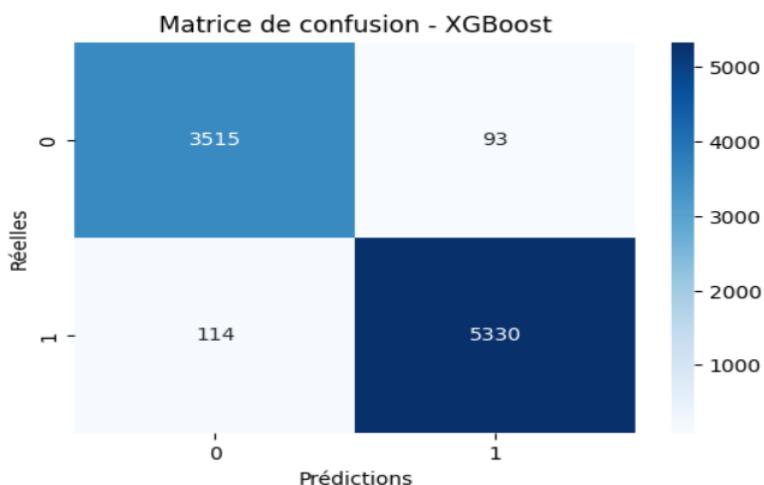
## 1.4. Classification par XGBoost

Accuracy XGBoost: 0.9771321254971277				
	precision	recall	f1-score	support
0.0	0.97	0.97	0.97	3608
1.0	0.98	0.98	0.98	5444
accuracy			0.98	9052
macro avg	0.98	0.98	0.98	9052
weighted avg	0.98	0.98	0.98	9052

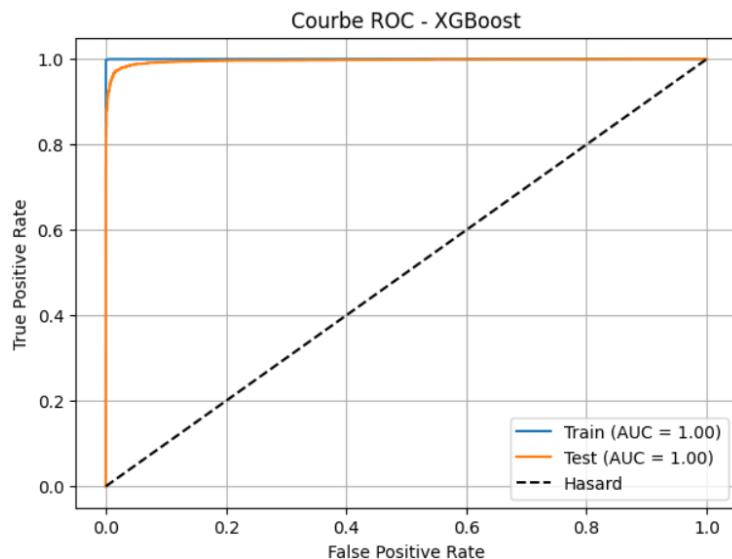


=> Ce modèle montre une excellente capacité à capturer les relations complexes entre les facteurs académiques, psychologiques et comportementaux, ce qui en fait un candidat très solide pour la prédition du risque de dépression.

=> L'analyse des importances de caractéristiques du modèle XGBoost montre que les facteurs académiques sont les plus déterminants dans la prédition du risque de dépression.



=>La matrice de confusion du modèle XGBoost révèle les meilleures performances parmi les modèles évalués, avec **3515 vrais négatifs et 5330 vrais positifs**, et un nombre d'erreurs particulièrement réduit (**93 faux positifs et 114 faux négatifs**).



=>Le modèle XGBoost obtient les meilleures performances, avec une **AUC égale à 1,00** sur les ensembles d'entraînement et de test, traduisant une capacité de discrimination quasi parfaite.

## 1.5. Classification par KNN

Évaluation du modèle KNN :

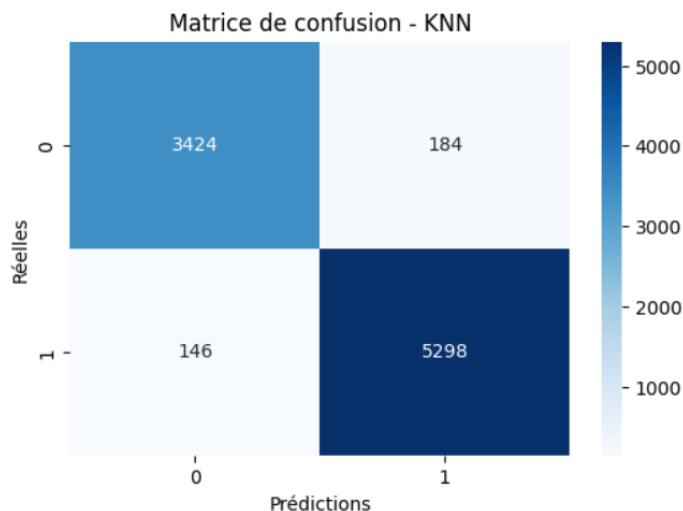
Accuracy : 0.9635439681838268

Rapport de classification :

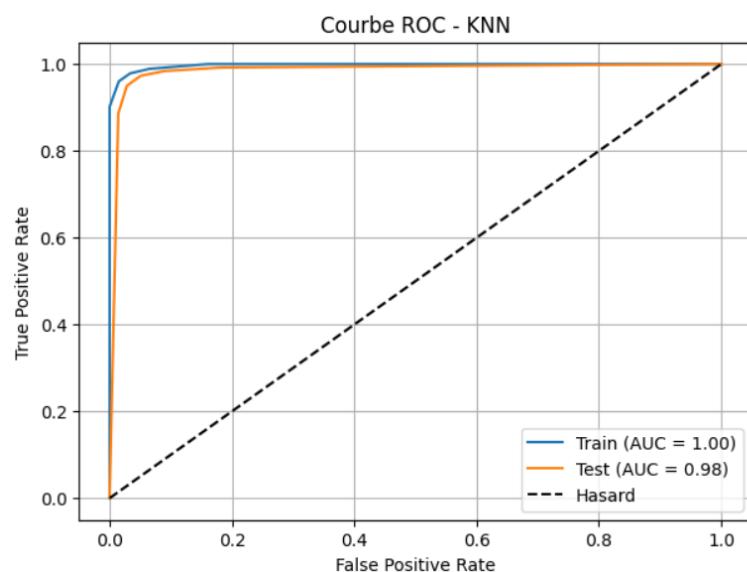
	precision	recall	f1-score	support
0.0	0.96	0.95	0.95	3608
1.0	0.97	0.97	0.97	5444
accuracy			0.96	9052
macro avg	0.96	0.96	0.96	9052
weighted avg	0.96	0.96	0.96	9052

---

=> Le modèle KNN classe chaque étudiant en fonction des profils qui lui ressemblent le plus dans le dataset (en termes de satisfaction d'étude, sommeil, pression académique, stress, etc.).



=> La matrice de confusion du modèle **KNN** indique de bonnes performances globales, avec **3424 vrais négatifs** et **5298 vrais positifs**, tandis que les erreurs restent limitées à **184 faux positifs** et **146 faux négatifs**, traduisant une capacité satisfaisante du modèle à classer correctement les observations.



=> La courbe ROC du modèle KNN montre une performance élevée, avec une **AUC de 1,00 sur l'ensemble d'entraînement** et **0,98 sur l'ensemble de test**, suggérant une légère perte de généralisation mais restant très satisfaisante.

## 1.6. Classification par Logistic Regression

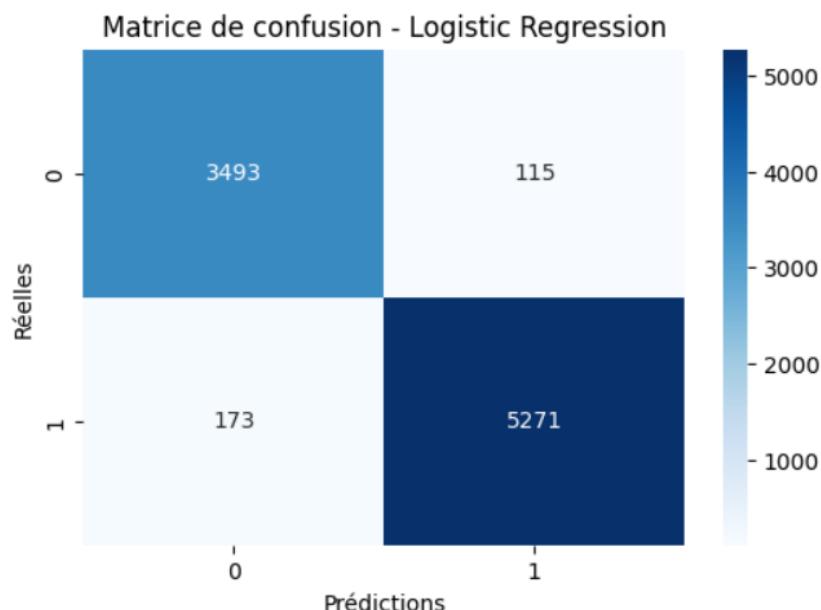
#### Évaluation du modèle Logistic Regression :

Accuracy : 0.9681838267786125

Rapport de classification :

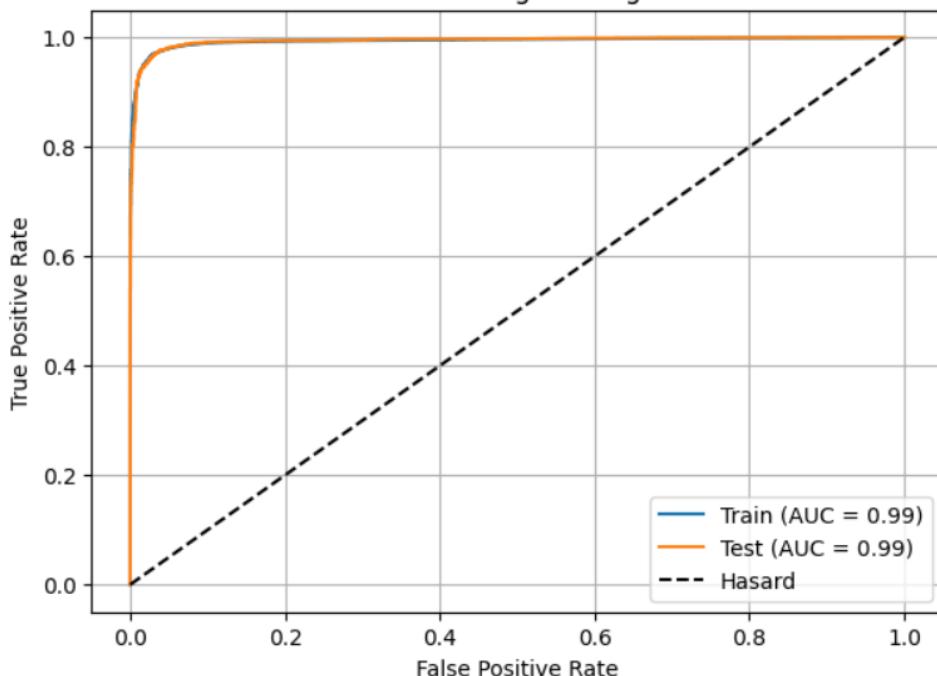
	precision	recall	f1-score	support
0.0	0.95	0.97	0.96	3608
1.0	0.98	0.97	0.97	5444
accuracy			0.97	9052
macro avg	0.97	0.97	0.97	9052
weighted avg	0.97	0.97	0.97	9052

=> Le modèle fournit une estimation probabiliste du risque de dépression en fonction des composantes principales. Cette approche linéaire offre un modèle simple, interprétable, et utile comme référence de base pour évaluer les performances des autres algorithmes plus complexes.



Le modèle de régression logistique affiche des performances solides, avec **3493 vrais négatifs** et **5271 vrais positifs**, bien que le nombre de faux négatifs (173) soit légèrement plus élevé comparativement aux modèles d'ensemble.

Course ROC - Logistic Regression



=> La courbe ROC du modèle de régression logistique présente une excellente capacité de discrimination, avec une **AUC de 0,99** sur les ensembles d'entraînement et de test, indiquant une très bonne stabilité du modèle.

## 1.7. Classification par Decision Tree

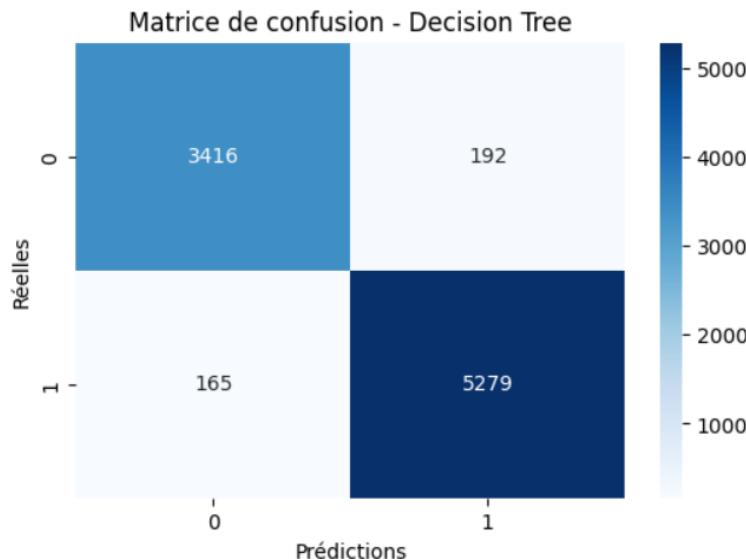
Évaluation du modèle Decision Tree :

Accuracy : 0.9605612019443217

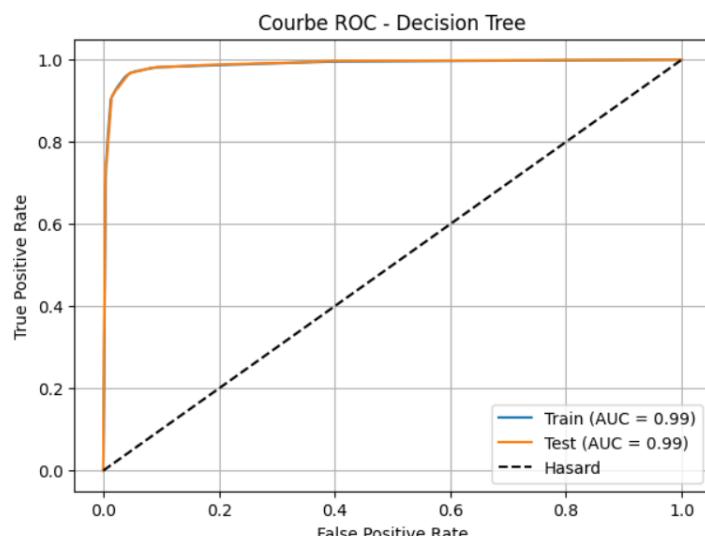
Rapport de classification :

	precision	recall	f1-score	support
0.0	0.95	0.95	0.95	3608
1.0	0.96	0.97	0.97	5444
accuracy			0.96	9052
macro avg	0.96	0.96	0.96	9052
weighted avg	0.96	0.96	0.96	9052

=> L'évaluation du modèle Decision Tree montre de très bonnes performances, avec une accuracy de 96,05 %, indiquant une capacité efficace et stable du modèle à discriminer correctement les observations.



=> La matrice de confusion du modèle *Decision Tree* met en évidence une bonne performance de classification, avec **3416 vrais négatifs** et **5279 vrais positifs**, tandis que les erreurs restent modérées (**192 faux positifs** et **165 faux négatifs**).



=> Le modèle *Decision Tree* présente une courbe ROC avec une **AUC de 0,99** sur les données d'entraînement et de test, indiquant une excellente capacité de classification et une bonne généralisation.

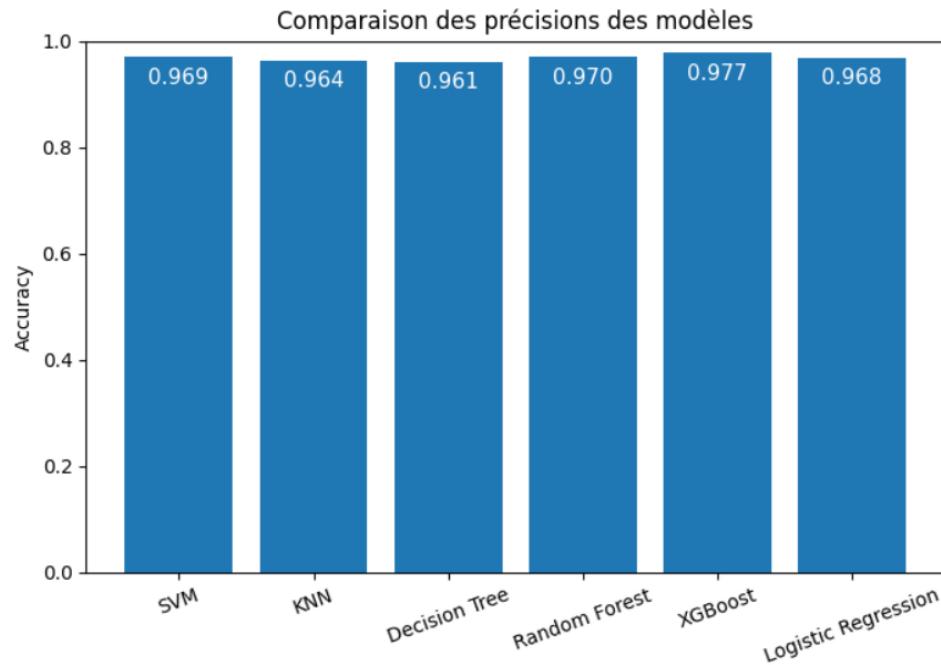
## 1.8. Cross validation

- ✓ ... Logistic Regression -> Scores : [0.96851698 0.97033969 0.96668876 0.96867749 0.97132913] , Accuracy moyenne: 0.9691  
 SVM -> Scores : [0.96835128 0.97199669 0.96784886 0.96917468 0.97265496] , Accuracy moyenne: 0.9700  
 KNN -> Scores : [0.96304888 0.96785418 0.96320849 0.96320849 0.96320849] , Accuracy moyenne: 0.9641  
 Random Forest -> Scores : [0.96835128 0.97199669 0.96818031 0.96900895 0.97149486] , Accuracy moyenne: 0.9698  
 Decision Tree -> Scores : [0.95907208 0.96205468 0.95525356 0.95956248 0.95657938] , Accuracy moyenne: 0.9585,  
 XGBoost -> Scores : [0.97663629 0.97928749 0.97596951 0.97845542 0.97596951] , Accuracy moyenne: 0.9773,

=> Les résultats de la validation croisée montrent que XGBoost obtient la meilleure performance globale avec une accuracy moyenne de 97,73 %, suivi par SVM (97,00 %), Random Forest (96,98 %) et régression logistique (96,91 %), tandis que KNN (96,41 %) et

Decision Tree (95,85 %) présentent des performances légèrement inférieures mais restent globalement satisfaisantes.

### **1.9. Choix du d'algorithme convenable pour la classification**



=> **XGBoost** se démarque comme le meilleur modèle, avec l'accuracy la plus élevée et le plus faible nombre d'erreurs dans la matrice de confusion. Random Forest arrive juste derrière, offrant une excellente robustesse. SVM et Regression Logistique fournissent aussi de très bons résultats, mais commettent davantage de faux négatifs. KNN et Decision Tree sont corrects mais moins performants, car ils gèrent moins bien les relations complexes présentes dans le dataset.

### **Objectif 2 : Segmentation des profils d'étudiants**

#### **Evaluation entre KMEANS et DBSCAN**

---

K-MEANS scores :

Silhouette Score : 0.2451345021256226

Calinski-Harabasz : 7500.496302617455

Davies-Bouldin Index : 1.6168425090652674

---

Labels DBSCAN uniques : [-1 0 1 2 3]

Clusters restants sans outliers : [0 1 2 3]

DBSCAN scores (sans outliers) :

Silhouette Score : 0.1121400237541291

Calinski-Harabasz : 2362.3614921604276

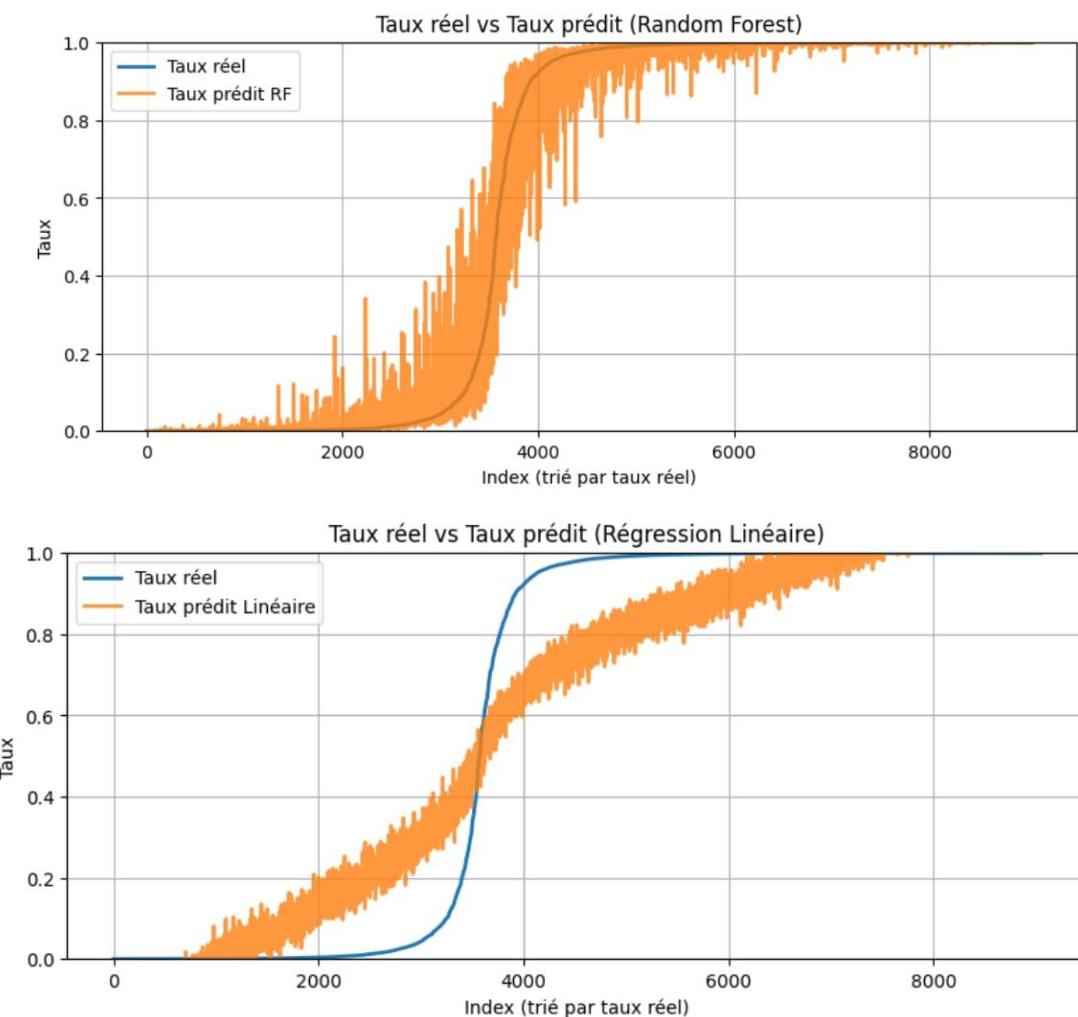
Davies-Bouldin Index : 2.5705683765019876

Les trois métriques internes indiquent systématiquement que K-Means sur les données projetées en PCA fournit un meilleur regroupement que DBSCAN :

- Le Silhouette Score, qui mesure la cohésion interne des clusters, est légèrement plus élevé pour K-Means.
- L'indice de Calinski–Harabasz, indicateur de la séparation entre clusters, est nettement supérieur pour K-Means.
- L'indice de Davies–Bouldin, où une valeur plus faible indique de meilleurs clusters, est également en faveur de K-Means.

Dans notre cas, ces résultats montrent que **la structure de nos données est mieux capturée par K-Means**, suggérant des clusters relativement compacts et bien séparés, ce qui correspond au type de structure que cet algorithme modélise le mieux.

#### Objectif 4 : Prédiction du niveau de sévérité de la dépression



Les résultats montrent que le modèle **Random Forest** suit plus fidèlement les valeurs réelles du taux de sévérité de la dépression que la régression linéaire, notamment dans les zones de variation rapide. Contrairement au modèle linéaire, le Random Forest capture mieux les relations non linéaires présentes dans les données. Pour cette raison, il a été choisi comme modèle final pour la prédiction du niveau de sévérité de la dépression.

#### **Objectif 5 : Suivi temporel des indicateurs et alertes**

### 5.1. Évaluation du Z-score

Z-score:

Precision: 0.8790123456790123

Recall : 0.2884927066450567

F1-score : 0.4344112263575351

---

La méthode du Z-score obtient une précision élevée ( $\approx 0,88$ ), indiquant que la majorité des anomalies détectées sont effectivement correctes. En revanche, son rappel est faible ( $\approx 0,29$ ), ce qui montre qu'elle ne détecte qu'une partie limitée des anomalies présentes dans les données.

Le F1-score relativement bas ( $\approx 0,43$ ) traduit ce déséquilibre entre précision et rappel. Cette méthode tend donc à être conservative, en générant peu de fausses alertes, mais au prix d'un nombre important d'anomalies non détectées.

- Le Z-score apparaît ainsi comme une méthode fiable mais incomplète, principalement adaptée à la détection de valeurs extrêmes simples et univariées.

### 5.2. Évaluation de l'Isolation Forest

Isolation Forest:

Precision: 0.5272

Recall : 0.534035656401945

F1-score : 0.5305958132045089

---

L'Isolation Forest présente une précision plus modérée ( $\approx 0,53$ ), mais un rappel nettement supérieur ( $\approx 0,53$ ) par rapport au Z-score. Cela indique une meilleure capacité à identifier les anomalies existantes, y compris celles résultant de combinaisons de plusieurs variables.

Son F1-score plus élevé ( $\approx 0,53$ ) reflète un meilleur équilibre global entre détection des anomalies et contrôle des fausses alertes. Les résultats montrent également que l'Isolation Forest permet d'identifier plusieurs mois atypiques, caractérisés par des scores d'anomalie plus faibles, signalant des situations potentiellement préoccupantes pour l'administration.

- Cette méthode est plus sensible et mieux adaptée à la détection d'anomalies multivariées, au prix d'une augmentation modérée des faux positifs.

### 5.3. Choix de la méthode retenue

Au regard des résultats obtenus, Isolation Forest est retenue comme méthode principale. Elle offre de meilleures performances globales, mesurées par le F1-score, et une capacité accrue à détecter des anomalies réalistes dans un contexte administratif, où les situations anormales résultent souvent de la combinaison de plusieurs indicateurs de risque.

	period	iforest_anomaly	iforest_score
<b>0</b>	2024-01	0	0.016595
<b>1</b>	2024-02	0	0.048419
<b>2</b>	2024-03	0	0.070149
<b>3</b>	2024-04	0	0.080515
<b>4</b>	2024-05	0	0.049688
<b>5</b>	2024-06	1	-0.015966
<b>6</b>	2025-01	1	-0.013902
<b>7</b>	2025-02	0	0.038347
<b>8</b>	2025-03	0	0.072379
<b>9</b>	2025-04	0	0.072596
<b>10</b>	2025-05	0	0.026319
<b>11</b>	2025-06	1	-0.040609

## Chapitre 6 : Déploiement

Ce chapitre est consacré au déploiement de notre modèle de Machine Learning développé initialement sous forme de notebook.

L'objectif principal est de transformer ce travail expérimental en une application web fonctionnelle et accessible aux utilisateurs.

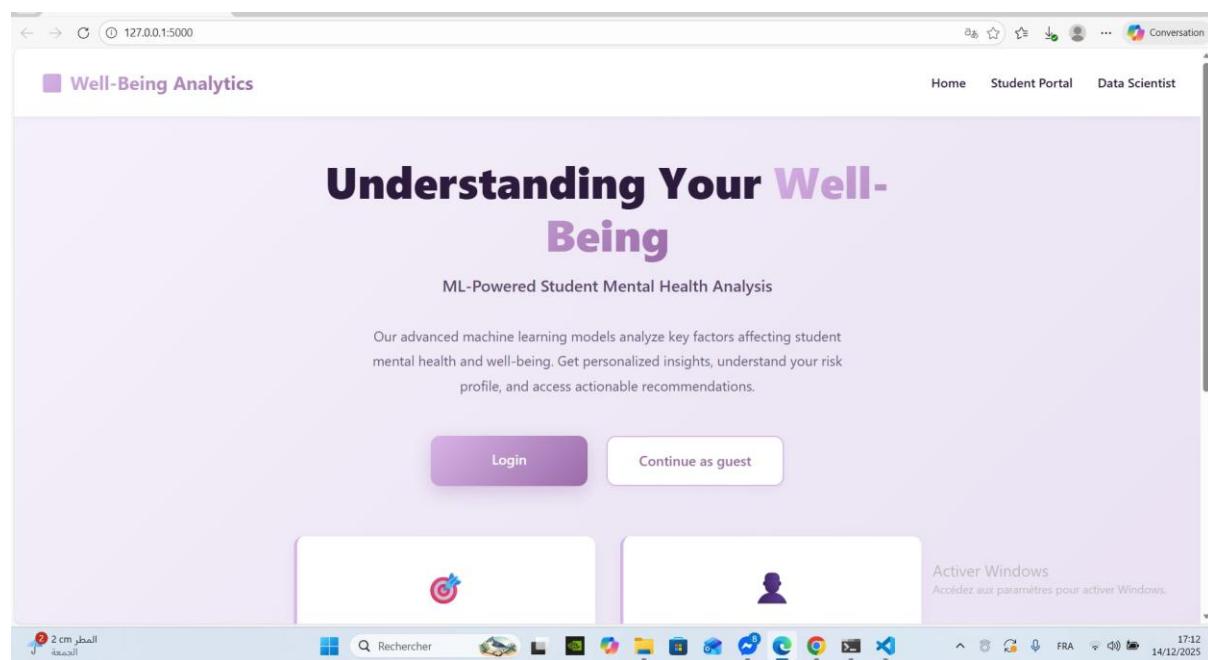
Pour cela, nous avons mis en place une architecture simple et efficace reposant sur un backend Flask chargé de l'exécution du modèle et de la gestion des requêtes, ainsi qu'un frontend développé en HTML et CSS permettant une interaction intuitive avec l'utilisateur. Les données sont stockées et manipulées à l'aide de fichiers CSV, servant de base de données légère pour l'application.

Cette approche permet de démontrer concrètement comment un modèle de Machine Learning peut être intégré dans un système web réel et exploité dans un contexte pratique.

**Cette section présente des captures d'écran de l'application web développée, illustrant l'interface utilisateur, le fonctionnement du site et l'affichage des résultats générés par le modèle de Machine Learning.**

## 6.1 Interface home

Cette page offre à l'utilisateur le choix entre une connexion au système ou un accès en mode invité.



The screenshot shows a web browser window titled "Well-Being Analytics - Student De..." with the URL "127.0.0.1:5000". The page has a light purple header with the title "Well-Being Analytics" and navigation links for "Home", "Student Portal", and "Data Scientist". Below the header are four white cards arranged in a 2x2 grid:

- Risk Detection**: Features a target icon and describes identifying depression risk factors through comprehensive analysis of sleep, academic pressure, and social support.
- Profile Clustering**: Features a person icon and describes discovering student profile clusters and connecting with similar students for peer support.
- Smart Recommendations**: Features a lightbulb icon and describes receiving personalized action plans based on specific risk profile and well-being factors.
- Severity Tracking**: Features a chart icon and describes monitoring well-being scores and tracking changes over time with visual progress indicators.

At the bottom right of the page, there is a link "Activer Windows" and a note "Accédez aux paramètres pour activer Windows.". The browser's taskbar at the bottom shows various pinned icons and the date "14/12/2025" at the bottom right.

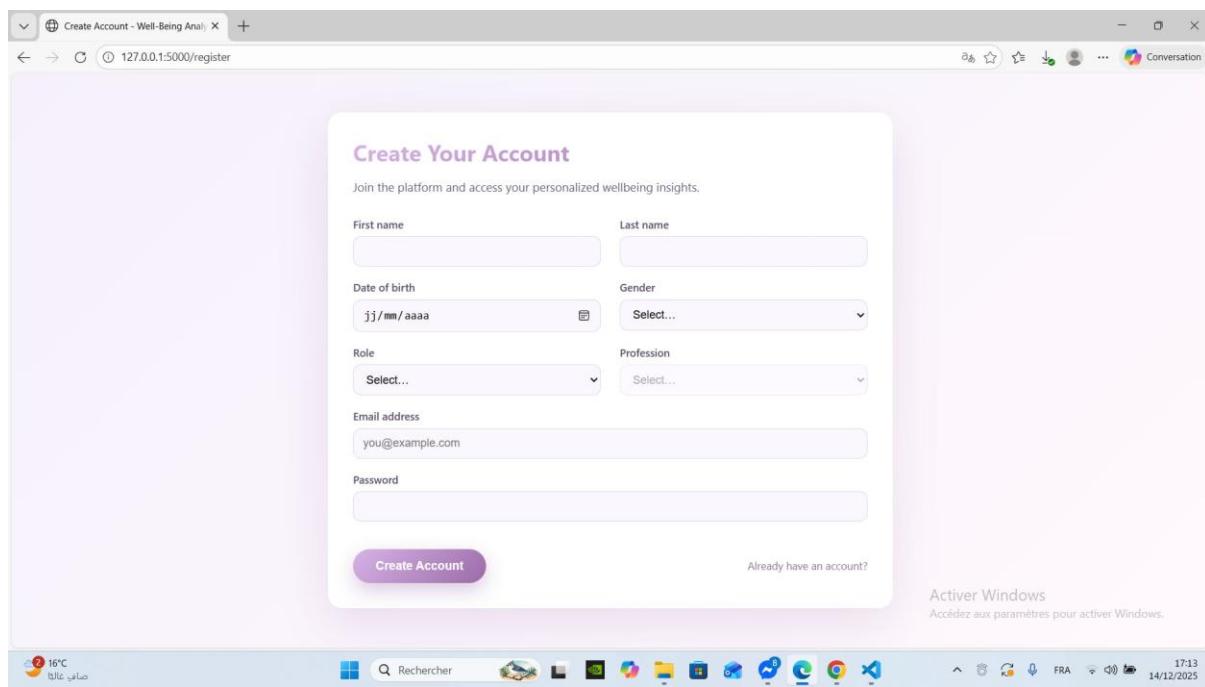
## 6.2 Interface Login

The screenshot shows a web browser window titled "Login - Well-Being Analytics" with the URL "127.0.0.1:5000/login". The page has a light purple header with the title "Login" and a sub-instruction "Log in to access your personalized wellbeing dashboard". Below the header is a form with two input fields:

- Email address**: A text input field containing "you@example.com".
- Password**: A text input field.

Below the form is a large purple "Login" button. At the bottom of the form area are two links: "Create an account" and "Back to home". At the very bottom of the page, there is a link "127.0.0.1:5000/register". The browser's taskbar at the bottom shows various pinned icons and the date "14/12/2025" at the bottom right.

## 6.3 Interface Sign Up



**Create Your Account**

Join the platform and access your personalized wellbeing insights.

First name

Last name

Date of birth  jj/mm/aaaa

Gender

Role

Profession

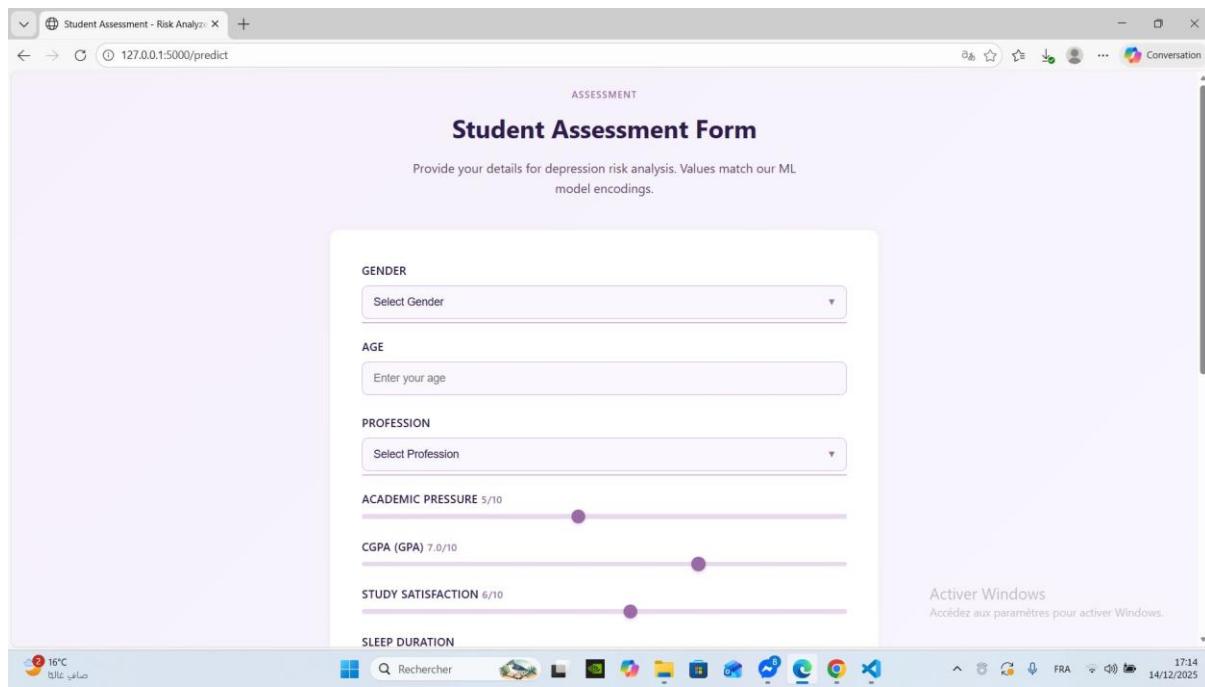
Email address  you@example.com

Password

**Create Account** Already have an account?

Activer Windows  
Accédez aux paramètres pour activer Windows.

## 6.4 Interface Formulaire pour le visiteur



**ASSESSMENT**

**Student Assessment Form**

Provide your details for depression risk analysis. Values match our ML model encodings.

**GENDER**

**AGE**

**PROFESSION**

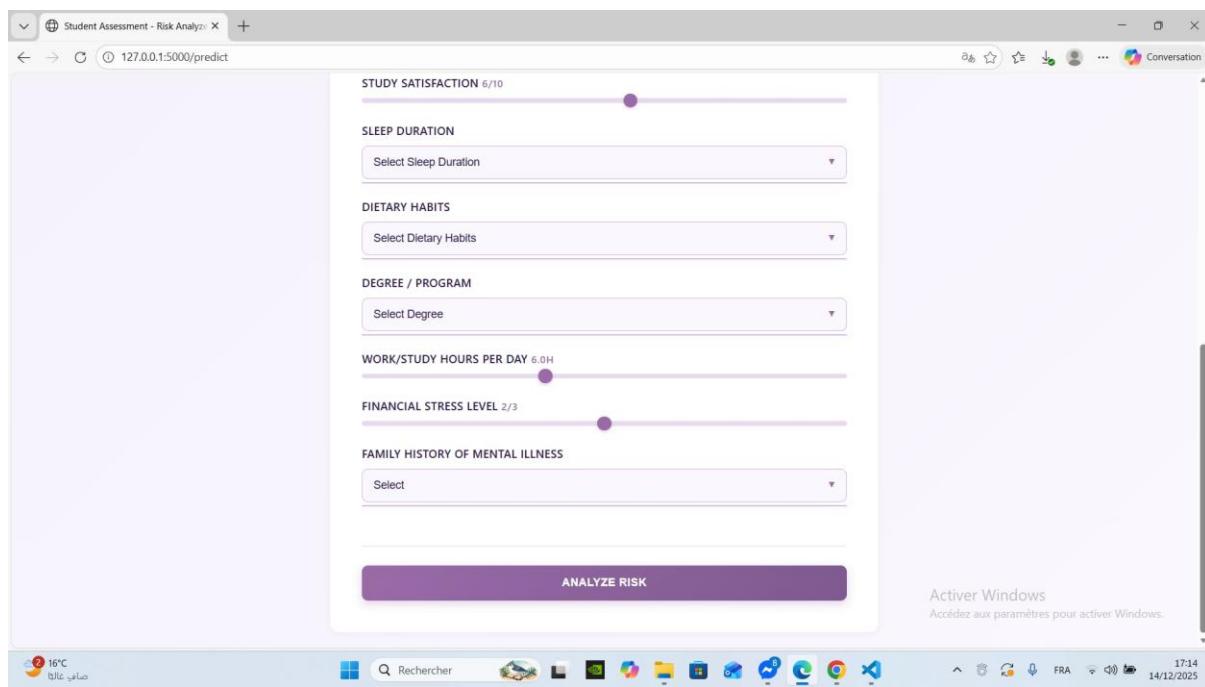
**ACADEMIC PRESSURE 5/10**

**CGPA (GPA) 7.0/10**

**STUDY SATISFACTION 6/10**

**SLEEP DURATION**

Activer Windows  
Accédez aux paramètres pour activer Windows.

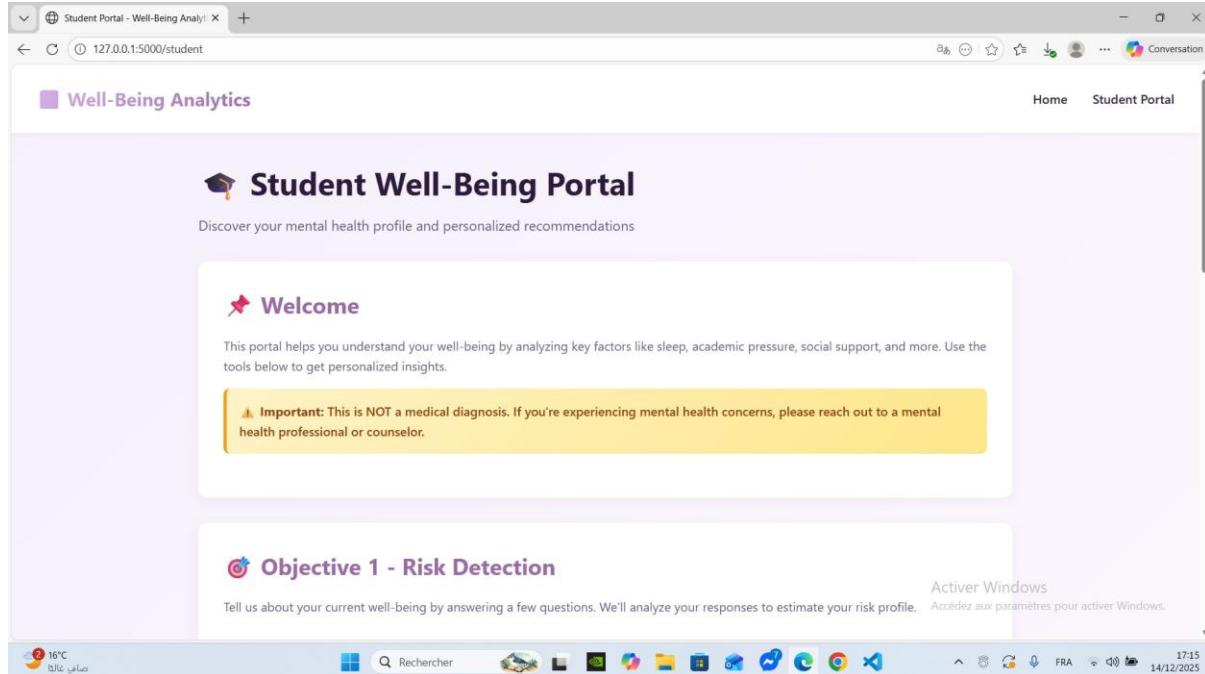


The screenshot shows a web application titled "Student Assessment - Risk Analysis". It contains several input fields:

- STUDY SATISFACTION**: A progress bar at 6/10.
- SLEEP DURATION**: A dropdown menu labeled "Select Sleep Duration".
- DIETARY HABITS**: A dropdown menu labeled "Select Dietary Habits".
- DEGREE / PROGRAM**: A dropdown menu labeled "Select Degree".
- WORK/STUDY HOURS PER DAY**: A progress bar at 6.0H.
- FINANCIAL STRESS LEVEL**: A progress bar at 2/3.
- FAMILY HISTORY OF MENTAL ILLNESS**: A dropdown menu labeled "Select".

A large purple button at the bottom center says "ANALYZE RISK".

## 6.4 Interfaces réservées pour un simple utilisateur



The screenshot shows the "Well-Being Analytics" section of the "Student Well-Being Portal".

**Student Well-Being Portal**

Discover your mental health profile and personalized recommendations

**Welcome**

This portal helps you understand your well-being by analyzing key factors like sleep, academic pressure, social support, and more. Use the tools below to get personalized insights.

**Important:** This is NOT a medical diagnosis. If you're experiencing mental health concerns, please reach out to a mental health professional or counselor.

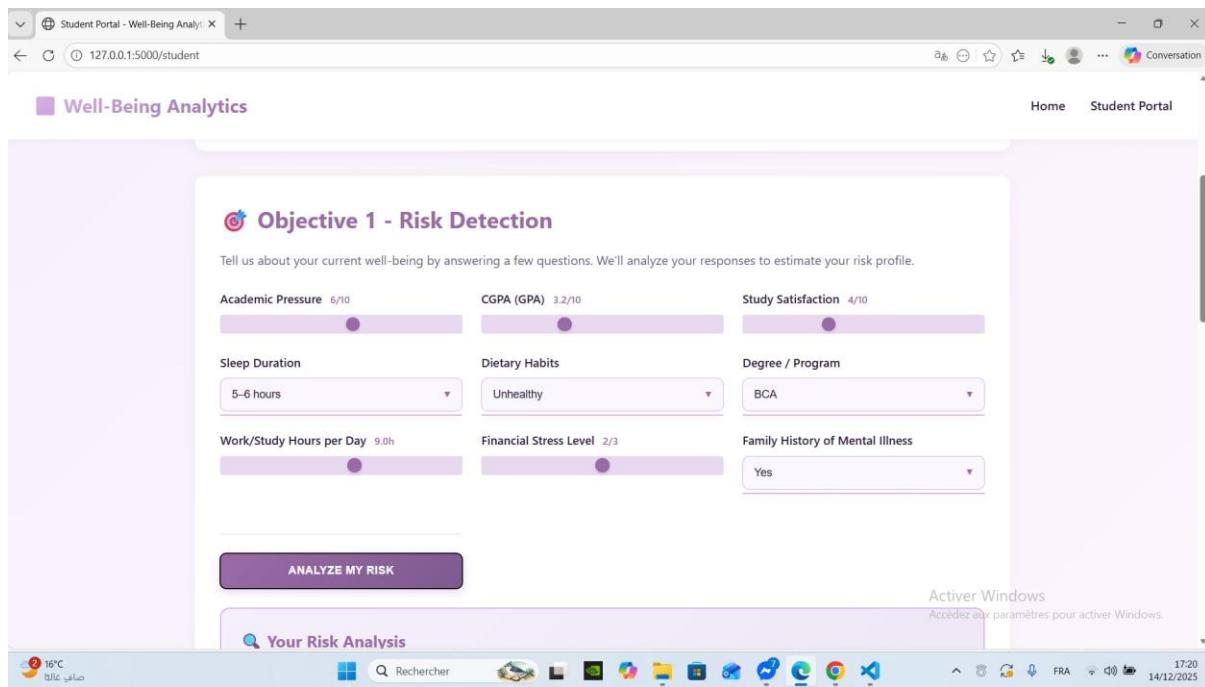
**Objective 1 - Risk Detection**

Tell us about your current well-being by answering a few questions. We'll analyze your responses to estimate your risk profile.

Activer Windows  
Accédez aux paramètres pour activer Windows.

## Formulaire – Détection du risque (Objective 1)

Cette interface permet à l'étudiant de saisir ses informations personnelles et académiques afin d'évaluer son niveau de risque.



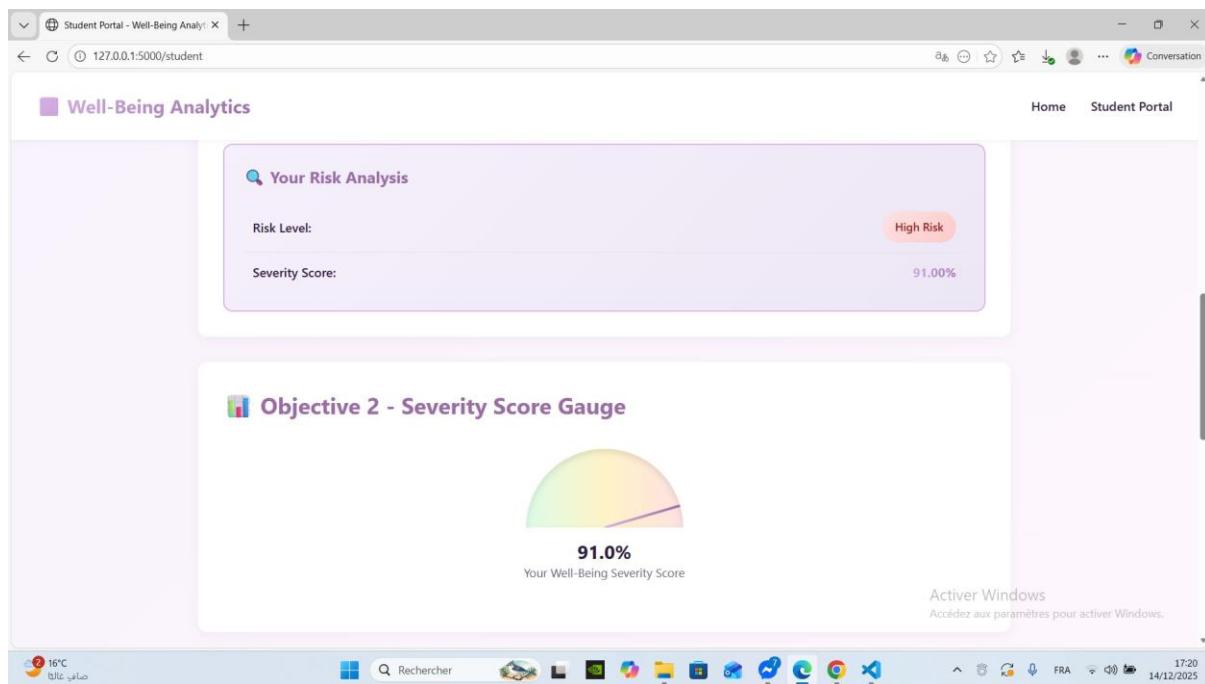
The screenshot shows a web-based form titled "Objective 1 - Risk Detection". The form includes several input fields and sliders:

- Academic Pressure: Slides from 6/10 to 10/10. Current value: 6/10.
- CGPA (GPA): Slides from 3.2/10 to 4.0/10. Current value: 3.2/10.
- Study Satisfaction: Slides from 4/10 to 10/10. Current value: 4/10.
- Sleep Duration: Selects "5-6 hours".
- Dietary Habits: Selects "Unhealthy".
- Degree / Program: Selects "BCA".
- Work/Study Hours per Day: Slides from 9.0h to 12.0h. Current value: 9.0h.
- Financial Stress Level: Slides from 2/3 to 3/3. Current value: 2/3.
- Family History of Mental Illness: Selects "Yes".

A large purple button labeled "ANALYZE MY RISK" is centered below the sliders. Below the analysis section, there is a link to "Your Risk Analysis".

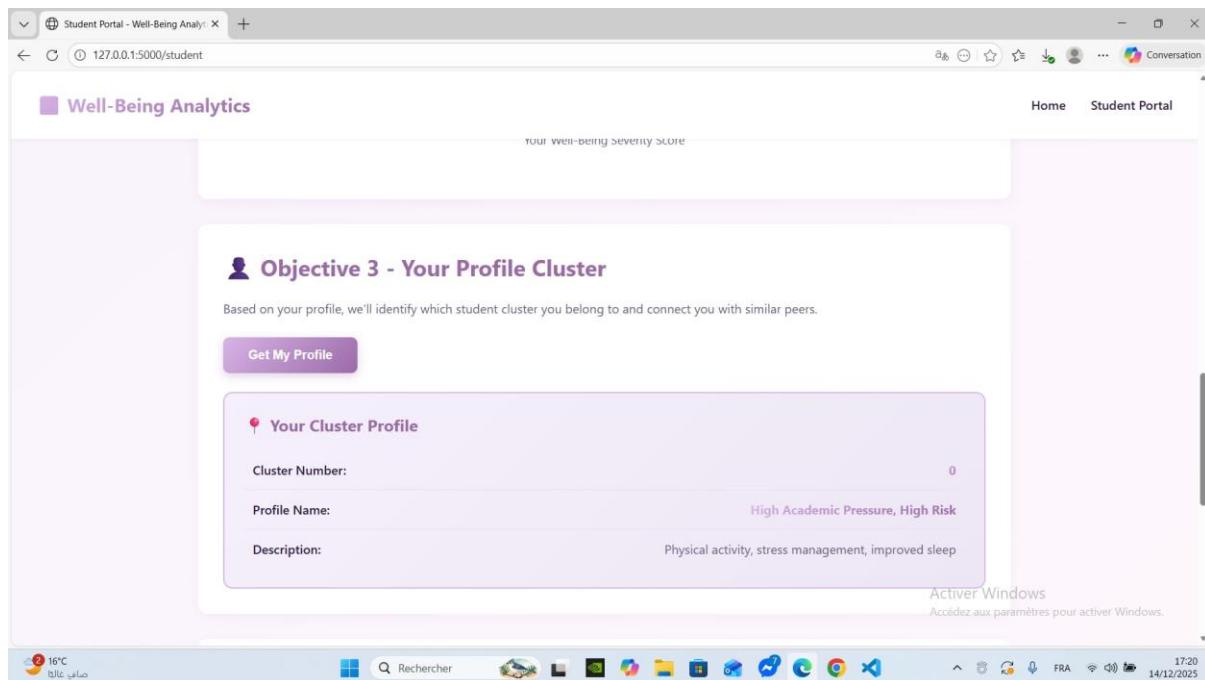
## Score de sévérité (Objective 2)

Cette section affiche le niveau de risque estimé ainsi que le score de sévérité calculé par le modèle.



## Profil de cluster (Objective 3)

Cette partie identifie le profil de l'étudiant en le classant dans un cluster spécifique basé sur ses caractéristiques.



Well-Being Analytics

your well-being severity score

**Objective 3 - Your Profile Cluster**

Based on your profile, we'll identify which student cluster you belong to and connect you with similar peers.

**Get My Profile**

**Your Cluster Profile**

Cluster Number: 0

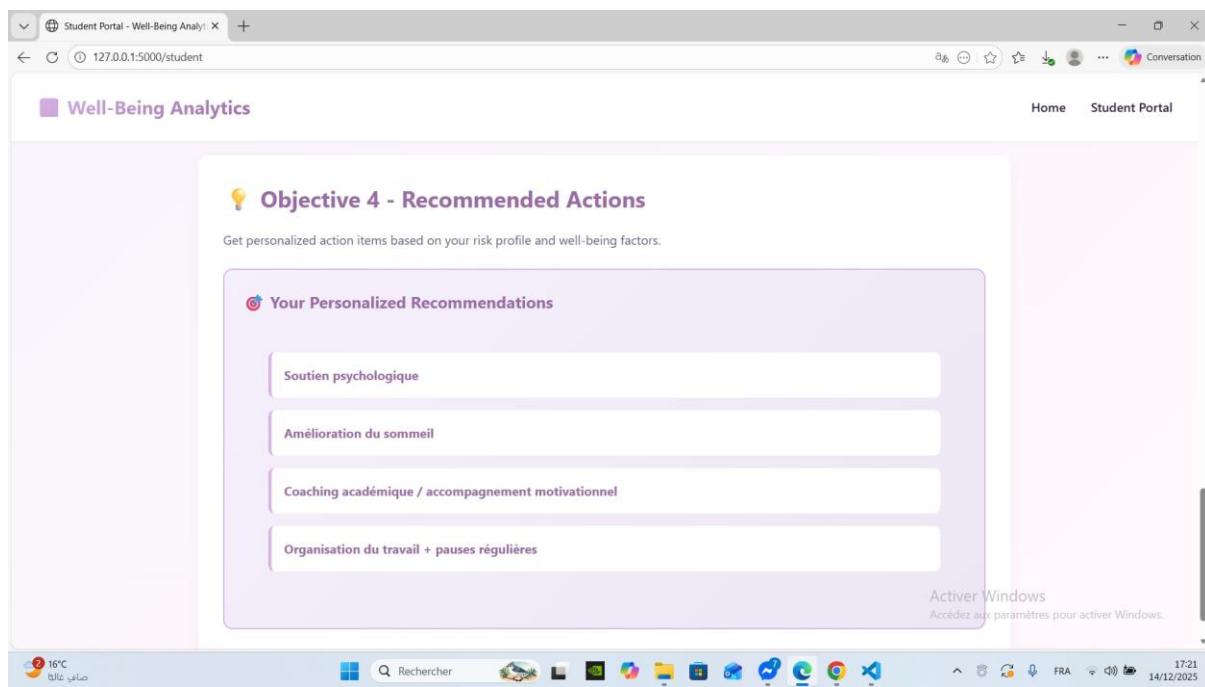
Profile Name: High Academic Pressure, High Risk

Description: Physical activity, stress management, improved sleep

Activer Windows  
Accédez aux paramètres pour activer Windows.

## Recommandations personnalisées (Objective 4)

Cette section propose des actions personnalisées visant à améliorer le bien-être de l'étudiant selon son profil de risque.



Well-Being Analytics

**Objective 4 - Recommended Actions**

Get personalized action items based on your risk profile and well-being factors.

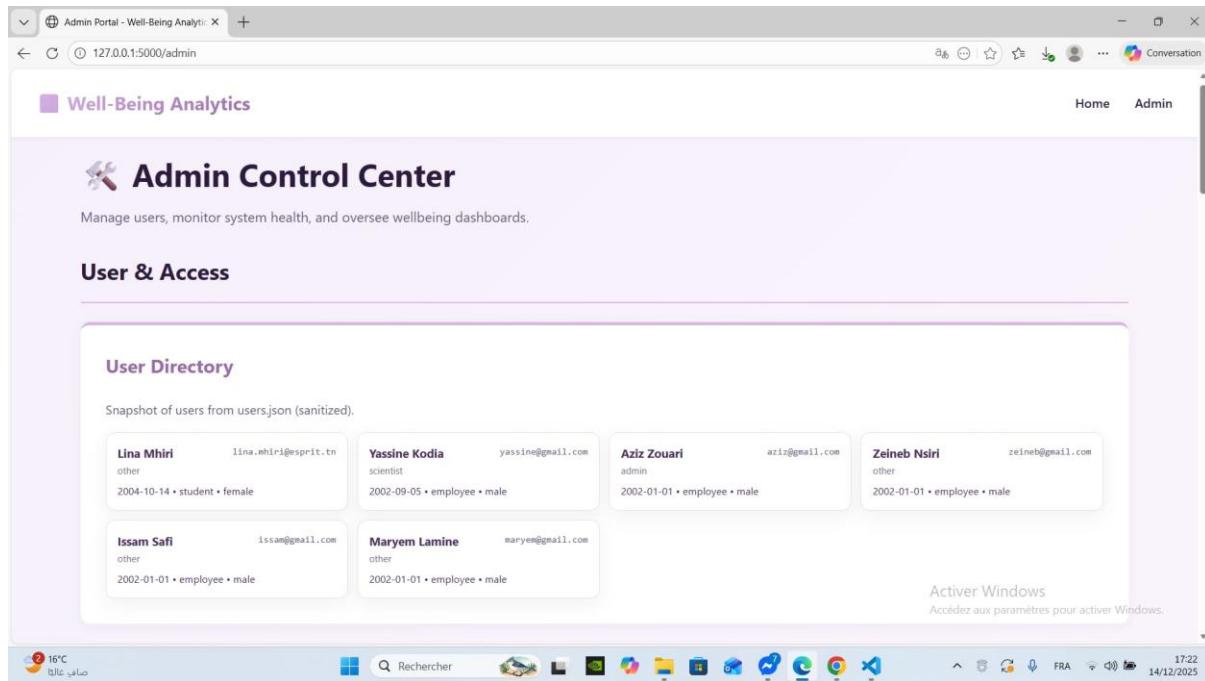
**Your Personalized Recommendations**

- Soutien psychologique
- Amélioration du sommeil
- Coaching académique / accompagnement motivationnel
- Organisation du travail + pauses régulières

Activer Windows  
Accédez aux paramètres pour activer Windows.

## 6.5 Interfaces réservées pour un administrateur

Cette page du panneau d'administration présente les utilisateurs enregistrés sur la plateforme ainsi que leurs informations principales.



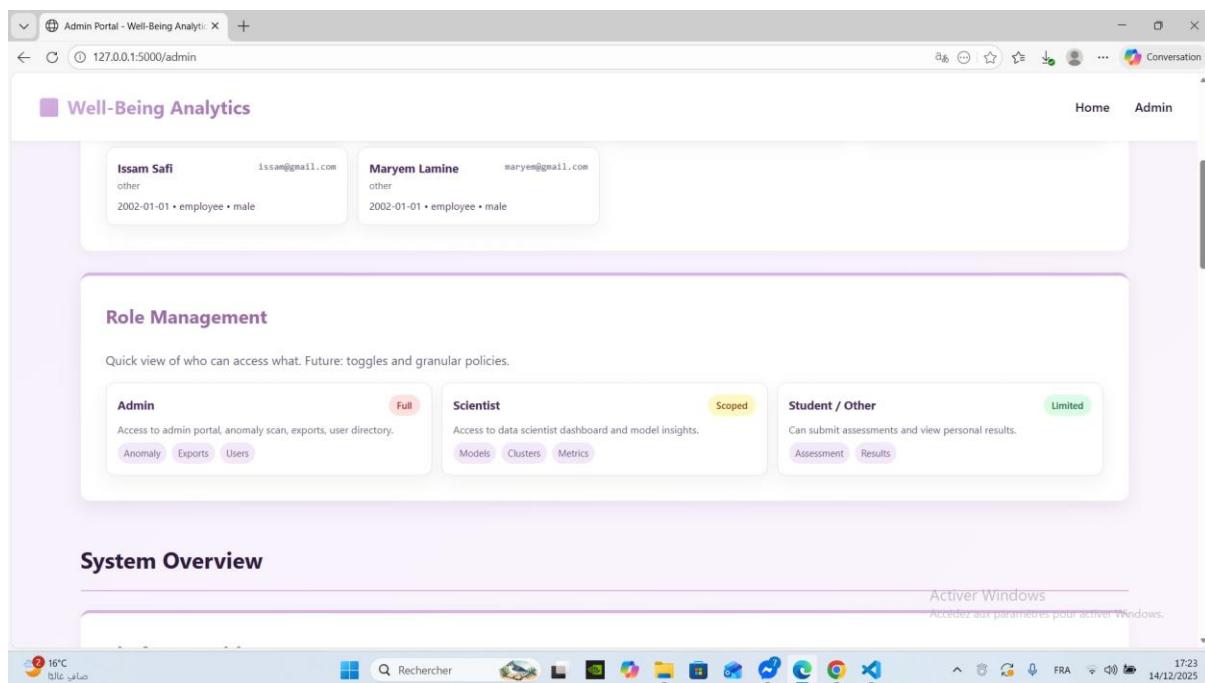
**User & Access**

**User Directory**

Snapshot of users from users.json (sanitized).

User	Email	Role	Details
Lina Mhiri	lina.mhiri@esprit.tn	other	2004-10-14 • student • female
Yassine Kodia	yassine@gmail.com	scientist	2002-09-05 • employee • male
Aziz Zouari	aziz@gmail.com	admin	2002-01-01 • employee • male
Zeineb Nsirri	zeineb@gmail.com	other	2002-01-01 • employee • male
Issam Safi	issam@gmail.com	other	2002-01-01 • employee • male
Maryem Lamine	maryem@gmail.com	other	2002-01-01 • employee • male

Activer Windows  
Accédez aux paramètres pour activer Windows.



**Role Management**

Quick view of who can access what. Future: toggles and granular policies.

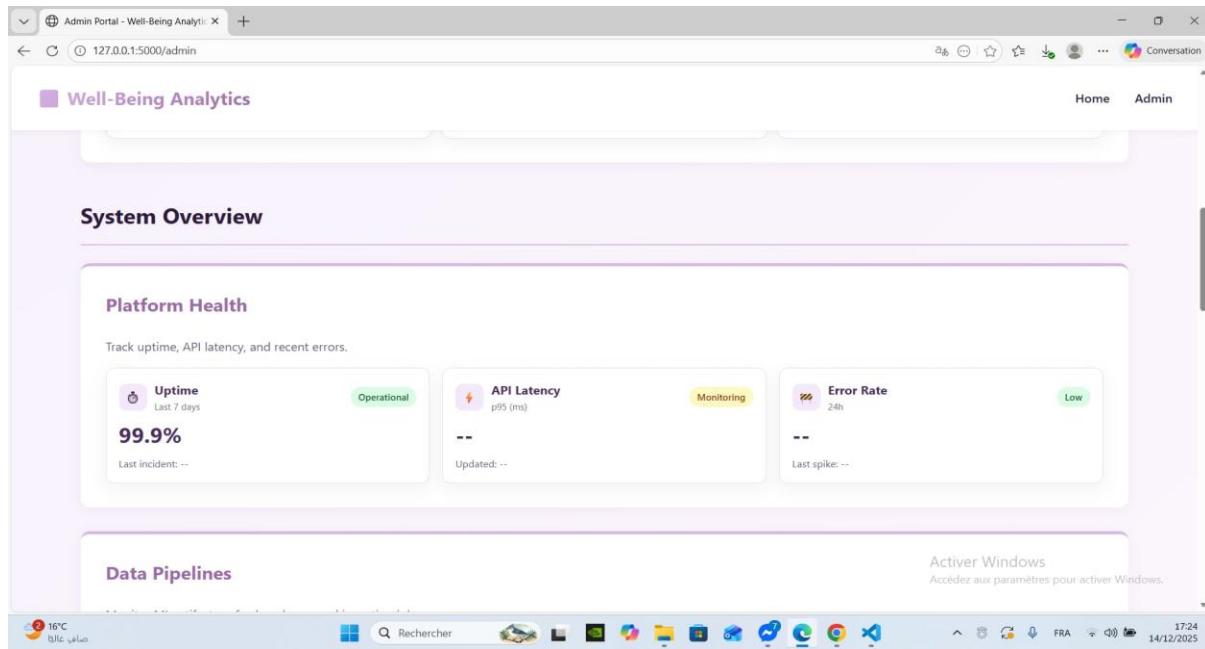
Role	Access Level	Description
Admin	Full	Access to admin portal, anomaly scan, exports, user directory. <a href="#">Anomaly</a> <a href="#">Exports</a> <a href="#">Users</a>
Scientist	Scoped	Access to data scientist dashboard and model insights. <a href="#">Models</a> <a href="#">Clusters</a> <a href="#">Metrics</a>
Student / Other	Limited	Can submit assessments and view personal results. <a href="#">Assessment</a> <a href="#">Results</a>

**System Overview**

Activer Windows  
Accédez aux paramètres pour activer Windows.

## System Overview – Platform Health

Cette section permet à l'administrateur de surveiller l'état général de la plateforme, incluant la disponibilité, la latence de l'API et le taux d'erreurs.

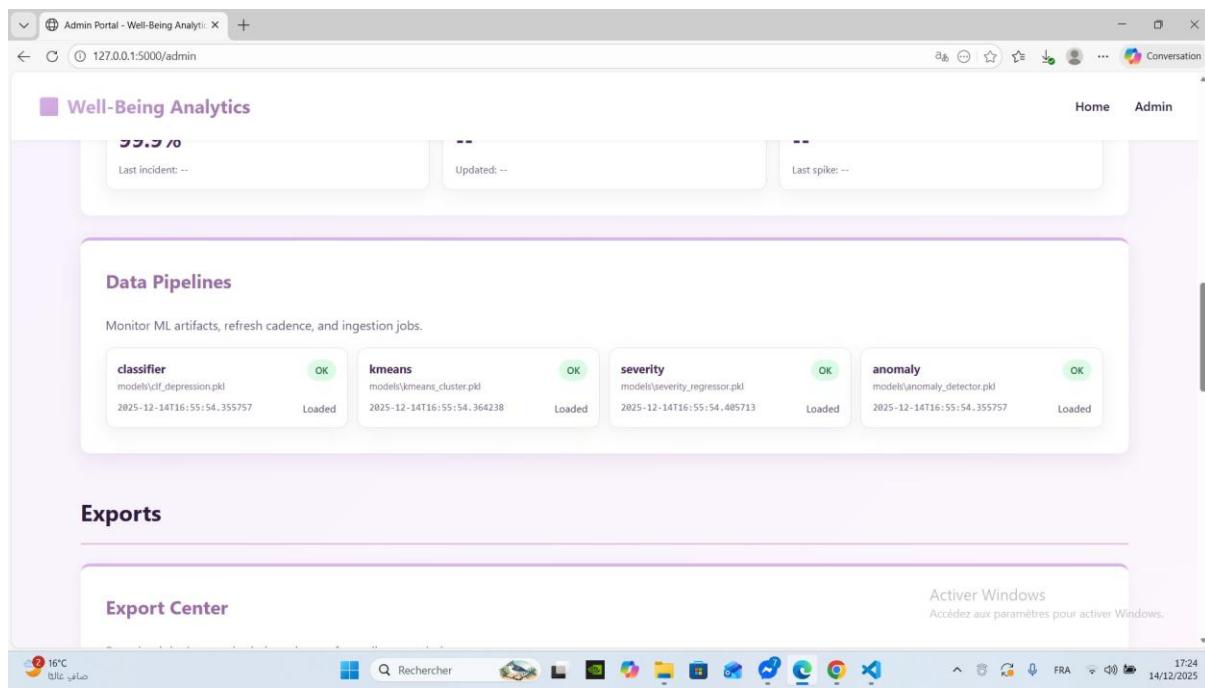


The screenshot shows the 'System Overview' section of the Well-Being Analytics admin portal. It displays three key metrics:

- Uptime:** Last 7 days, 99.9% (Operational)
- API Latency:** p95 (ms), -- (Monitoring)
- Error Rate:** 24h, -- (Low)

## Data Pipelines

Cette interface affiche l'état des pipelines de données et des modèles de Machine Learning utilisés par le système.

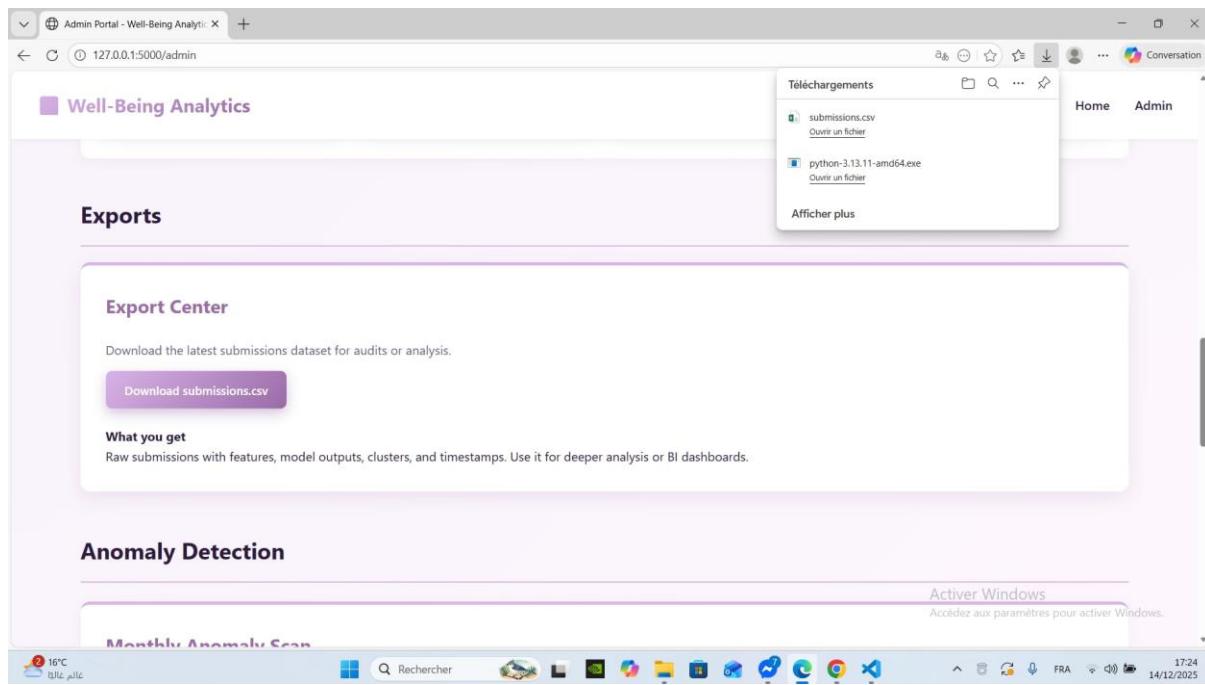


The screenshot shows the 'Data Pipelines' section of the Well-Being Analytics admin portal. It displays four ML artifacts:

- classifier**: models\clf\_depression.pkl, 2025-12-14T16:55:54, 355757 (OK, Loaded)
- kmeans**: models\kmeans\_cluster.pkl, 2025-12-14T16:55:54, 364238 (OK, Loaded)
- severity**: models\severity\_regressor.pkl, 2025-12-14T16:55:54, 405713 (OK, Loaded)
- anomaly**: models\anomaly\_detector.pkl, 2025-12-14T16:55:54, 355757 (OK, Loaded)

## Exportation des données des formulaires

L'administrateur dispose d'une option lui permettant d'exporter les données des formulaires pour analyse ou archivage.



**Exports**

**Export Center**

Download the latest submissions dataset for audits or analysis.

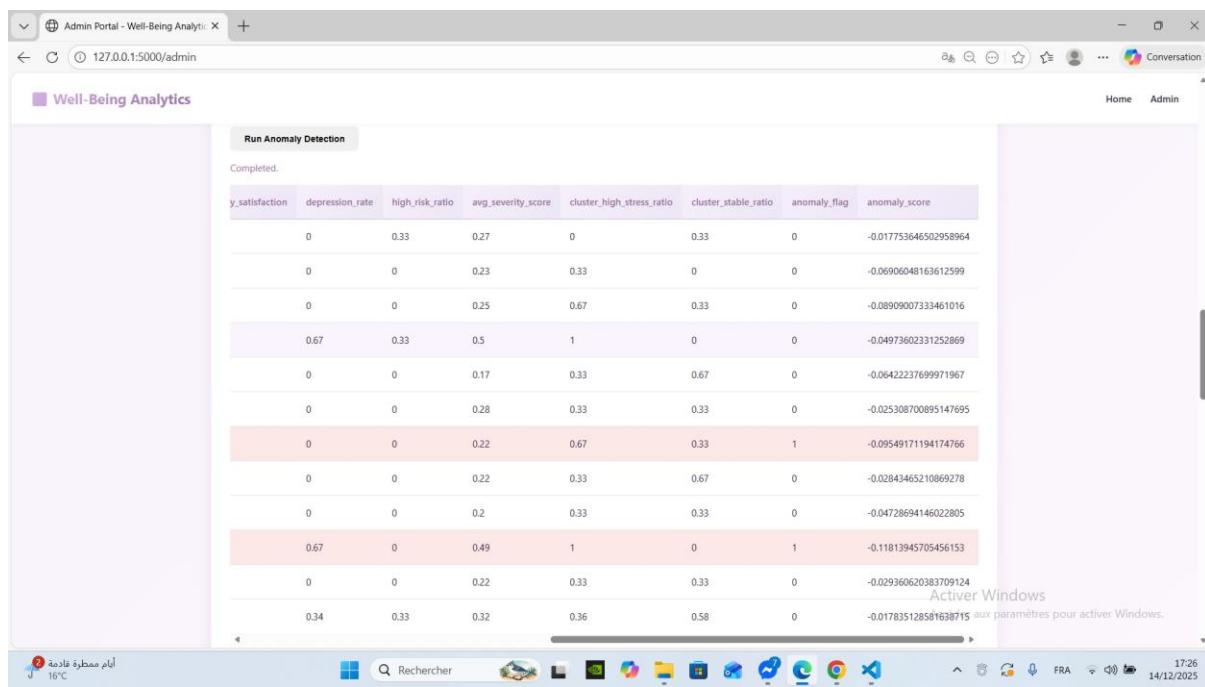
[Download submissions.csv](#)

**What you get**

Raw submissions with features, model outputs, clusters, and timestamps. Use it for deeper analysis or BI dashboards.

**Anomaly Detection**

Le *Monthly Anomaly Scan* permet à l'administrateur d'identifier, sur une base mensuelle, les anomalies détectées dans les indicateurs de bien-être.



y_satisfaction	depression_rate	high_risk_ratio	avg_severity_score	cluster_high_stress_ratio	cluster_stable_ratio	anomaly_flag	anomaly_score
0	0.33	0.27	0	0.33	0	-0.017753646502958964	
0	0	0.23	0.33	0	0	-0.06906048163612599	
0	0	0.25	0.67	0.33	0	-0.0890900733461016	
0.67	0.33	0.5	1	0	0	-0.04973602331252869	
0	0	0.17	0.33	0.67	0	-0.0642237699971967	
0	0	0.28	0.33	0.33	0	-0.025308700895147695	
0	0	0.22	0.67	0.33	1	-0.09549171194174766	
0	0	0.22	0.33	0.67	0	-0.02843465210869278	
0	0	0.2	0.33	0.33	0	-0.04728694146022805	
0.67	0	0.49	1	0	1	-0.11813945705456153	
0	0	0.22	0.33	0.33	0	-0.029360620383709124	
0.34	0.33	0.32	0.36	0.58	0	-0.017835128501638715	

Activer Windows

Accédez aux paramètres pour activer Windows.

Admin Portal - Well-Being Analytics

127.0.0.1:5000/admin

Well-Being Analytics

Run Anomaly Detection

Completed.

submission_year	submission_month	avg_academic_pressure	avg_sleep_duration	avg_financial_stress	avg_study_satisfaction	depression_rate	high_risk
2025	1	4.74	2.75	1.89	6.32	0	0.33
2025	2	4.36	2.67	1.86	6.52	0	0
2025	3	4.78	2.76	1.22	7.25	0	0
2025	4	7.18	3.03	2.16	6.66	0.67	0.33
2025	5	4.89	2.8	2.15	8.15	0	0
2025	6	4.6	3.25	1.92	6.36	0	0
2025	7	4.77	3.61	2.38	6.94	0	0
2025	8	5.14	2.82	1.79	7.03	0	0
2025	9	5.12	3.19	1.79	7.44	0	0
2025	10	8.23	3.86	2.82	7.16	0.67	0
2025	11	5.54	3.36	2.03	6	0	0
2025	12	4.84	1.93	2.03	5.12	0.34	0

Activer Windows  
Accédez aux paramètres pour activer Windows.

أيام ممطرة قادمة 16°C

Rechercher

17:26 14/12/2025

Cette interface lance et affiche les résultats de la détection mensuelle des anomalies à partir des données agrégées.

Admin Portal - Well-Being Analytics

127.0.0.1:5000/admin

Well-Being Analytics

Visual Insights

Trends and ratios derived from the anomaly dataset.

Anomaly Score by Month



High-Stress vs Stable Clusters

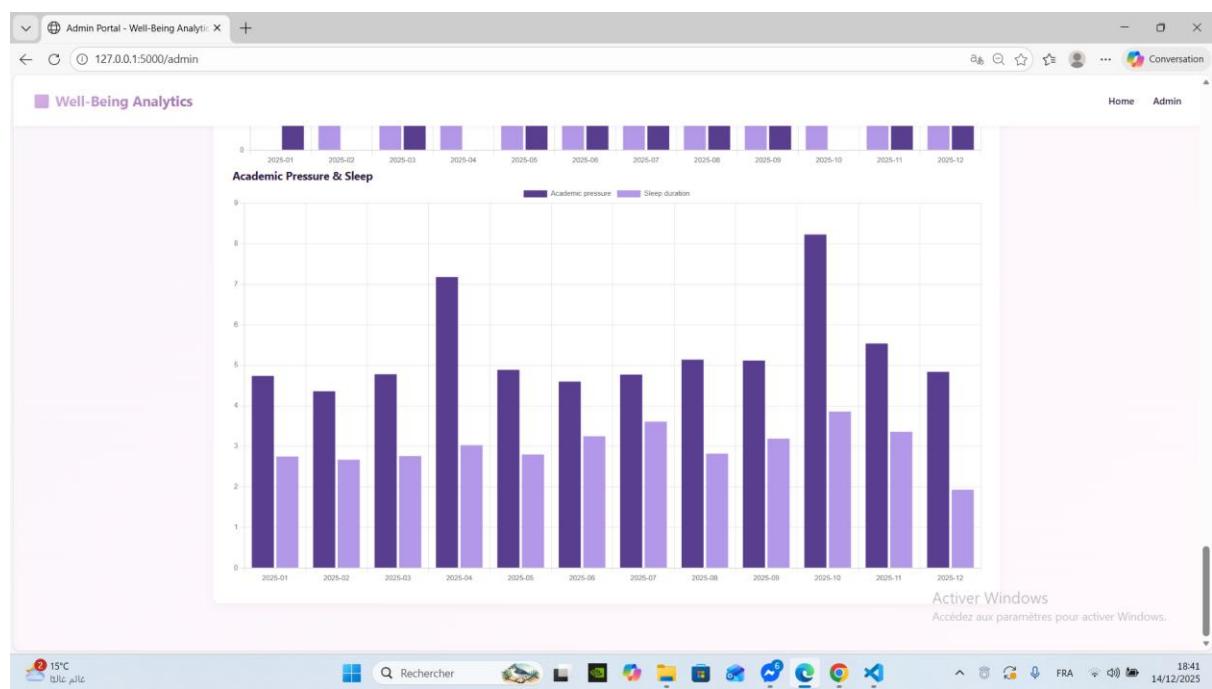
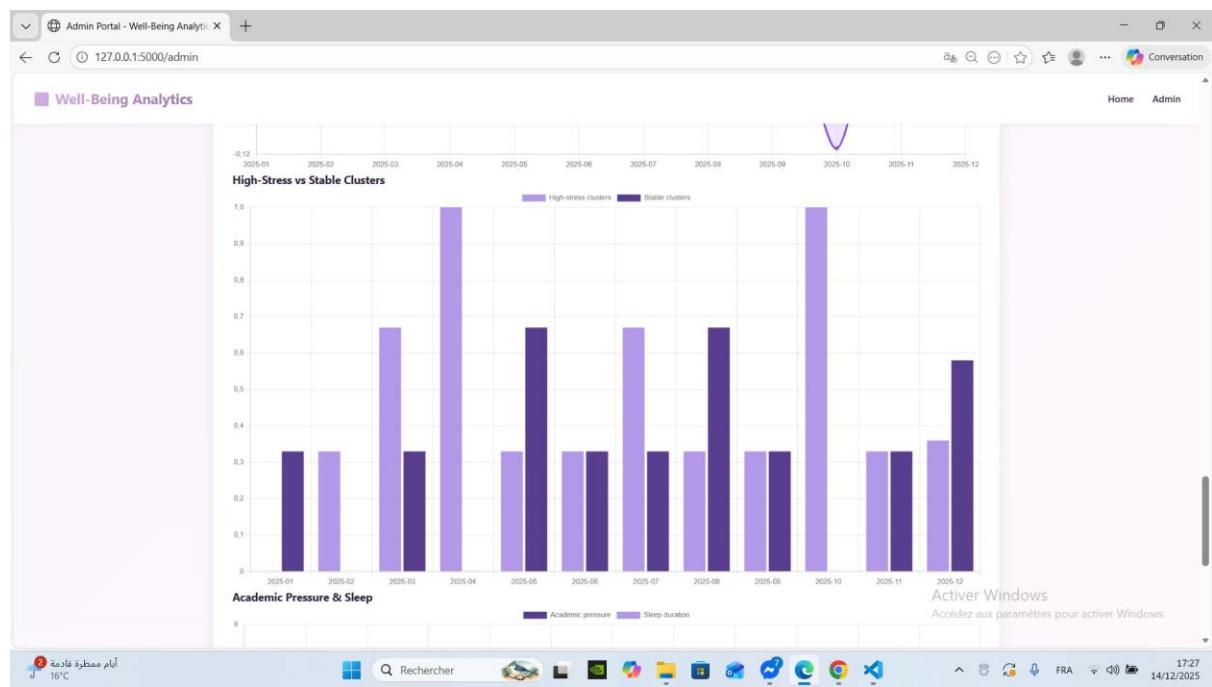
High-stress clusters    Stable clusters

Activer Windows  
Accédez aux paramètres pour activer Windows.

أيام ممطرة قادمة 16°C

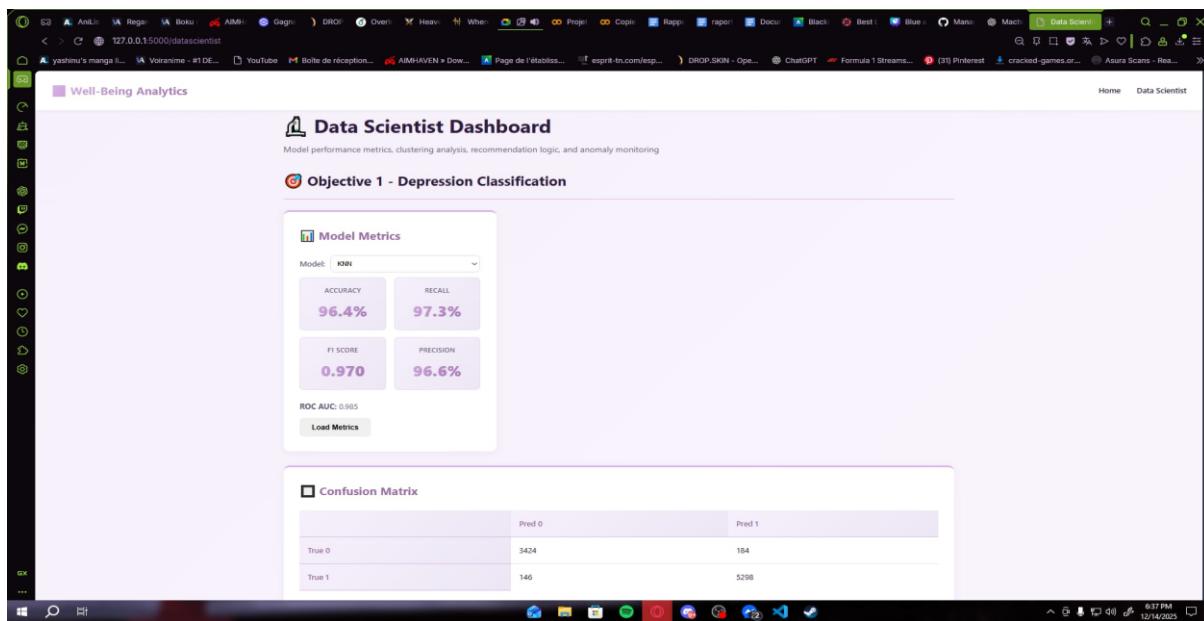
Rechercher

17:27 14/12/2025



## 6.5 Interfaces réservées pour un administrateur

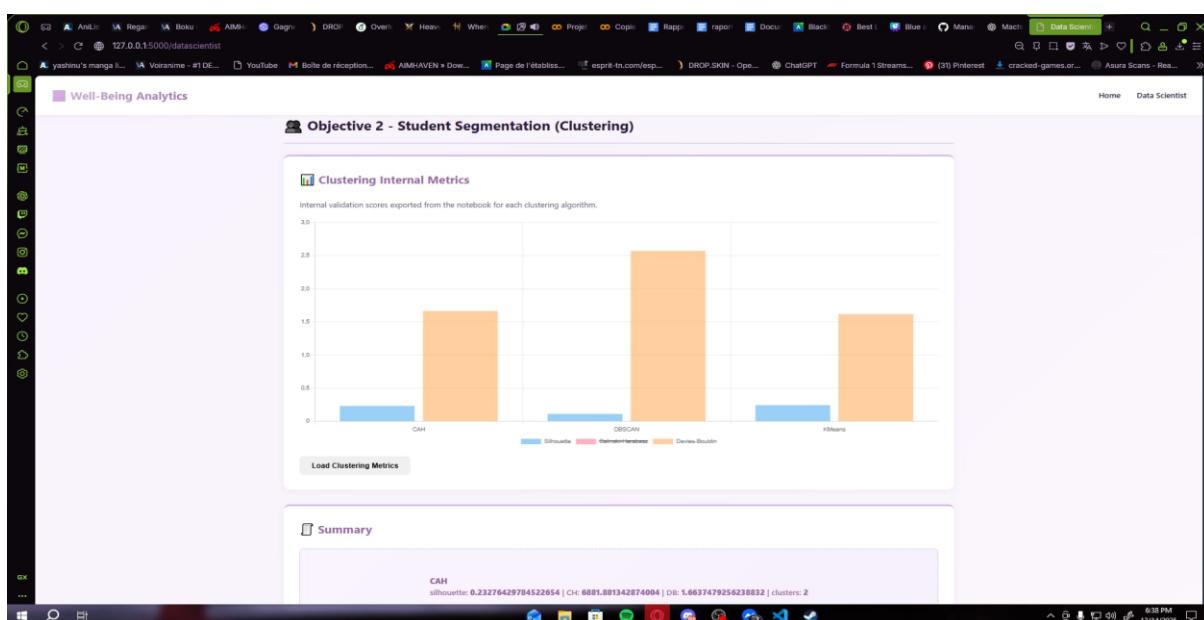
Ce tableau de bord permet au data scientist d'évaluer les performances des modèles de classification de la dépression à l'aide de métriques telles que l'accuracy, la précision, le rappel, le score F1 et la matrice de confusion. Il peut sélectionner l'algorithme de son choix (KNN, XGBoost, etc.), et les métriques ainsi que les résultats s'actualisent dynamiquement en fonction du modèle sélectionné.



The screenshot shows a dashboard titled "Data Scientist Dashboard" under the "Well-Being Analytics" tab. The top section displays "Model Metrics" for a KNN model, showing accuracy at 96.4%, recall at 97.3%, F1 score at 0.970, and precision at 96.6%. Below this is a "Confusion Matrix" table:

	Pred 0	Pred 1
True 0	3424	164
True 1	140	5298

Cette interface permet au data scientist de comparer les performances des algorithmes de clustering (CAH, DBSCAN, K-Means) à l'aide de métriques internes telles que le score de silhouette, l'indice de Calinski-Harabasz et l'indice de Davies-Bouldin, afin de sélectionner la méthode de segmentation la plus adaptée.



The screenshot shows a dashboard titled "Objective 2 - Student Segmentation (Clustering)" under the "Well-Being Analytics" tab. The top section displays "Clustering Internal Metrics" for CAH, DBSCAN, and KMeans algorithms, represented by a bar chart. The chart shows silhouette scores around 0.2 for CAH and KMeans, and a significantly higher score around 2.8 for DBSCAN. Below this is a "Summary" section showing CAH results:

CAH  
 silhouette: 0.23276429784522654 | CH: 6081.881342874004 | DI: 1.6637479256238832 | clusters: 2

## Summary

### CAH

silhouette: **0.23276429784522654** | CH: **6881.881342874004** | DB: **1.6637479256238832** | clusters: **2**

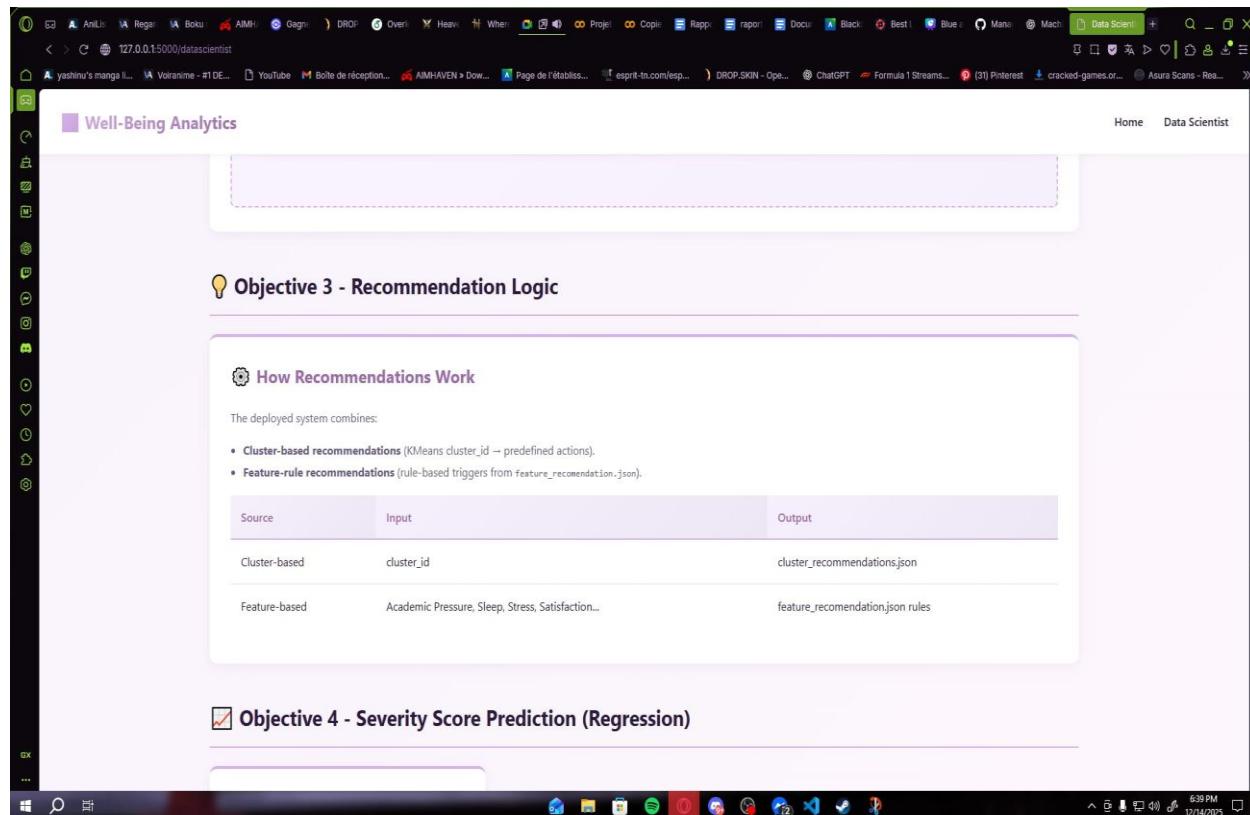
### DBSCAN

silhouette: **0.1121400237541291** | CH: **2362.3614921604276** | DB: **2.5705683765019876** | clusters: **4**

### KMeans

silhouette: **0.2451345021256226** | CH: **7500.496302617455** | DB: **1.6168425090652674** | clusters: **2**

Cette section décrit la logique de recommandation du système, basée à la fois sur le clustering des étudiants et sur des règles définies à partir des caractéristiques individuelles.



The screenshot shows a web-based analytics platform for student well-being. The main navigation bar includes links for Data Scientist, Home, and Data Scientist. The left sidebar contains various icons for file operations like Open, Save, Copy, Paste, etc. The main content area is titled "Well-Being Analytics".

### Objective 3 - Recommendation Logic

#### How Recommendations Work

The deployed system combines:

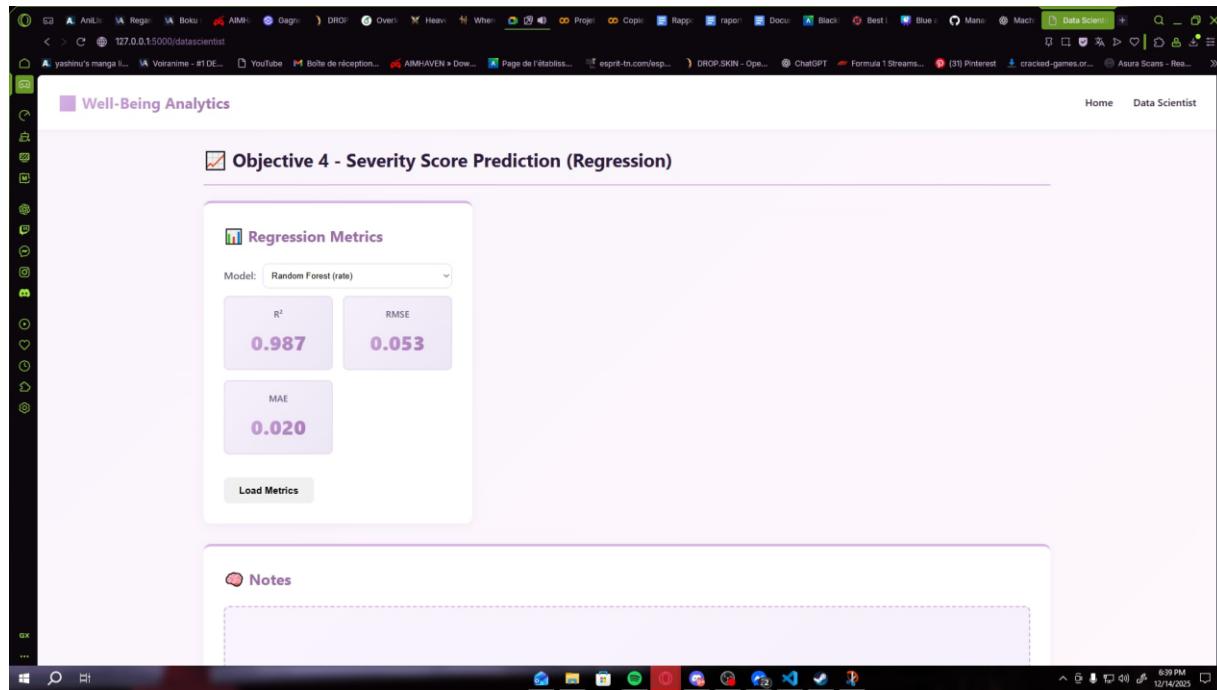
- Cluster-based recommendations (KMeans cluster\_id → predefined actions).
- Feature-rule recommendations (rule-based triggers from feature\_recommendation.json).

Source	Input	Output
Cluster-based	cluster_id	cluster_recommendations.json
Feature-based	Academic Pressure, Sleep, Stress, Satisfaction...	feature_recommendation.json rules

### Objective 4 - Severity Score Prediction (Regression)

The taskbar at the bottom of the screen shows several open applications, including a browser, a file manager, and communication tools. The system tray indicates the date (12/14/2025) and time (6:39 PM).

Cette interface permet d'évaluer les performances du modèle de régression sélectionné pour la prédiction du score de sévérité à l'aide de métriques telles que R<sup>2</sup>, RMSE et MAE ainsi que les résultats s'actualisent dynamiquement en fonction du modèle sélectionné.



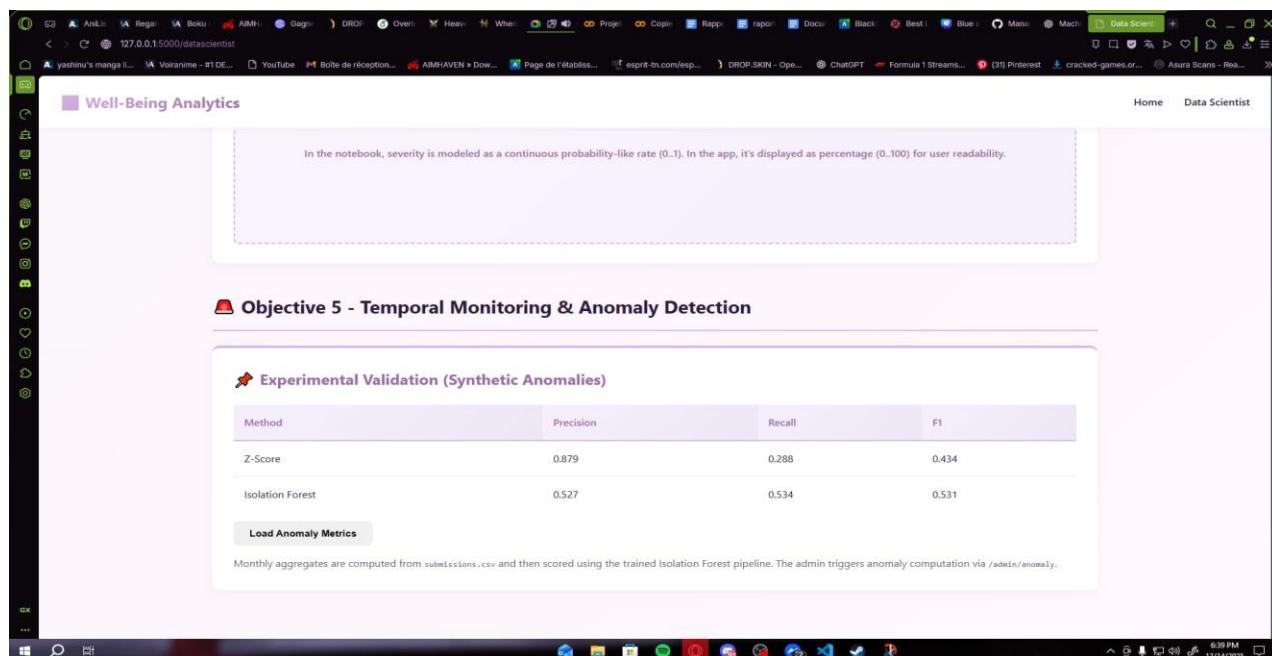
The screenshot shows a web-based application titled "Well-Being Analytics". The main section is titled "Objective 4 - Severity Score Prediction (Regression)". It displays three performance metrics for a "Random Forest (rate)" model:

Metric	Value
R <sup>2</sup>	<b>0.987</b>
RMSE	<b>0.053</b>
MAE	<b>0.020</b>

Below the metrics is a "Notes" section which contains the following text:

In the notebook, severity is modeled as a continuous probability-like rate (0..1). In the app, it's displayed as percentage (0..100) for user readability.

Cette section présente l'évaluation des méthodes de détection d'anomalies dans le temps, en comparant leurs performances à l'aide des métriques de précision, rappel et score F1.



The screenshot shows the same "Well-Being Analytics" application. The main section is titled "Objective 5 - Temporal Monitoring & Anomaly Detection". It displays experimental validation results for synthetic anomalies using two methods:

Method	Precision	Recall	F1
Z-Score	0.879	0.288	0.434
Isolation Forest	0.527	0.534	0.531

Below the table is a note about monthly aggregates and anomaly computation:

Monthly aggregates are computed from `submissions.csv` and then scored using the trained Isolation Forest pipeline. The admin triggers anomaly computation via `/admin/anomaly`.

## Conclusion générale

Ce projet a été réalisé en suivant les étapes de la méthodologie **CRISP-DM**, permettant une approche structurée allant de la compréhension du problème jusqu'au déploiement de la solution.

La phase de compréhension du métier a permis de définir clairement les objectifs liés à l'analyse du bien-être étudiant et à l'évaluation du risque de dépression.

Les étapes de compréhension et de préparation des données ont conduit à une exploration approfondie, au nettoyage et à la transformation des données, garantissant leur qualité pour la modélisation.

Plusieurs techniques de Machine Learning ont ensuite été appliquées lors de la phase de modélisation, incluant la classification, le clustering, la régression et la détection d'anomalies, avec une sélection des modèles les plus performants à l'issue de l'évaluation.

Enfin, la phase de déploiement a permis l'intégration des modèles dans une application web interactive, offrant des tableaux de bord adaptés aux profils étudiant, administrateur et data scientist.

Ce travail met en évidence l'intérêt de la méthodologie CRISP-DM pour développer une solution de Machine Learning complète, fiable et exploitable, tout en ouvrant la voie à des améliorations futures.