India Student Depression

Jedha Projet final

Héloïse PELTIER TORROELLA

Olga GEOFFROY

Manon HAMEL

Présentation du sujet

Problématique ⚠️

En Inde, 33,6% des étudiants présentent des symptômes de dépression modérés à sévères, soulignant l'ampleur de la crise de santé mentale chez les jeunes adultes et la nécessité d'actions pour améliorer leur bien-être psychologique.**

Objectifs 🎯

#1 Réduire le temps de diagnostic et d'intervention pour venir en aide aux étudiants concernés.

#2 Identifier les principaux facteurs de risque pour permettre aux différents acteurs (établissement d'enseignement supérieur, gouvernement) de limiter les cas de dépression.

Implémentation ⚙️

#1 Modèle de classification de ML capable de prédire le risque de dépression pour un étudiant donné

#2 Analyse et classification des principaux facteurs de risque

** Mental Health, Suicidality, Health, and Social Indicators Among College Students Across Nine States in India, 2024: https://pubmed.ncbi.nlm.nih.gov/39564264/

Prise en main du Dataset

🔗 Student Depression Dataset : Analyzing Mental Health Trends and Predictors Among Students

	id	Gender	Age	City	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
0	2	Male	33.0	Visakhapatnam	Student	5.0	0.0	8.97	2.0	0.0	'5-6 hours'	Healthy	B.Pharm	Yes	3.0	1.0	No	1
1	8	Female	24.0	Bangalore	Student	2.0	0.0	5.90	5.0	0.0	'5-6 hours'	Moderate	BSc	No	3.0	2.0	Yes	0
2	26	Male	31.0	Srinagar	Student	3.0	0.0	7.03	5.0	0.0	'Less than 5 hours'	Healthy	BA	No	9.0	1.0	Yes	0
3	30	Female	28.0	Varanasi	Student	3.0	0.0	5.59	2.0	0.0	'7-8 hours'	Moderate	BCA	Yes	4.0	5.0	Yes	1
4	32	Female	25.0	Jaipur	Student	4.0	0.0	8.13	3.0	0.0	'5-6 hours'	Moderate	M.Tech	Yes	1.0	1.0	No	0

Forme initiale - 27901 lignes pour 18 colonnes

Quantitatives 🧮

id, Age, Work/Study Hours, CGPA

Qualitatives ordinales 📶

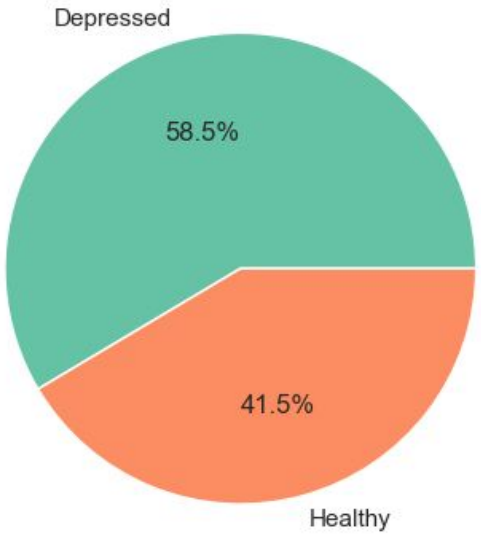
Academic Pressure, Work Pressure, Study Satisfaction, Job Satisfaction, Sleep Duration, Dietary Habits, Financial Stress

Qualitatives nominales 🏷️

Gender, City, Profession, Degree, Family History of Mental Illness, Depression, Have you ever had suicidal thoughts ?

Exploratory Data Analysis

🎯 Variable cible - Dépression

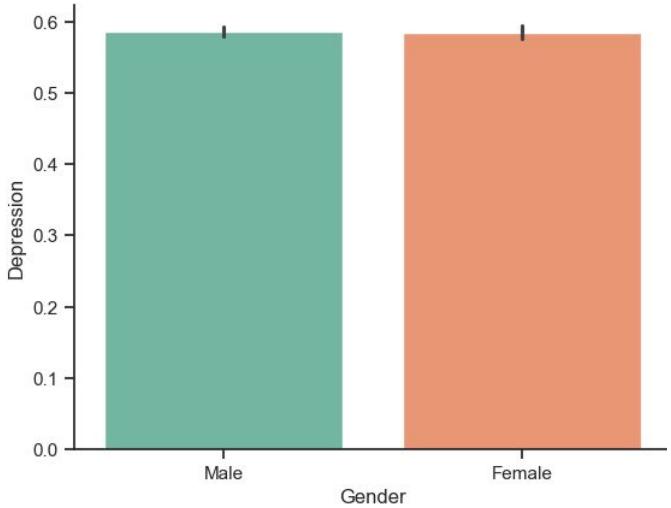


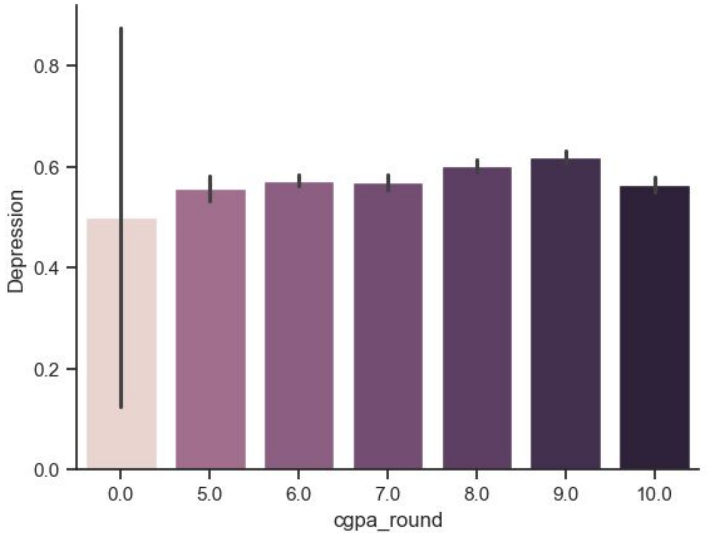
Notre échantillon est relativement équilibré avec 58.5% d'étudiants en dépression et 41.5% d'étudiants en bonne santé. Bien que ce ratio ne soit pas représentatif de la population réelle des étudiants, nous pouvons conserver ces proportions pour entraîner notre modèle.

Pas besoin d'*oversampling* ni d'*undersampling*.

Exploratory Data Analysis

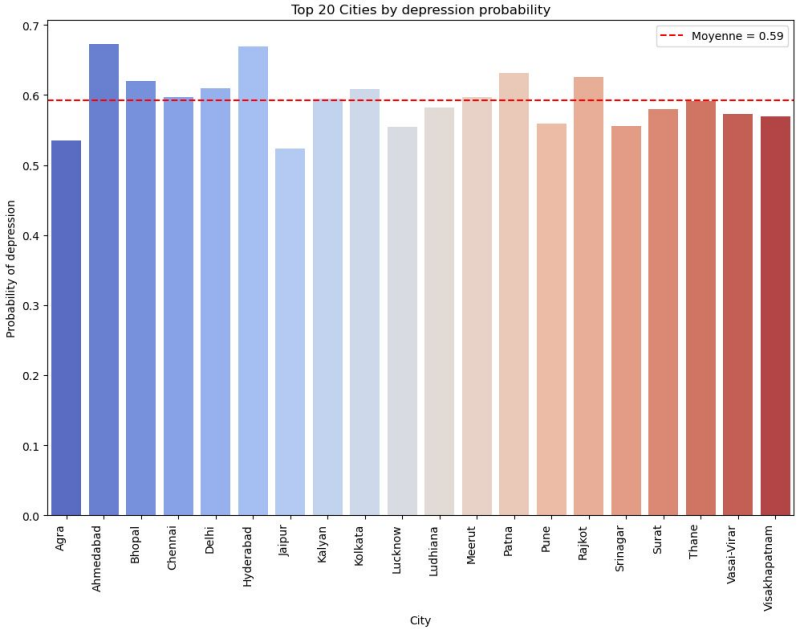
```
Nombre d'occurences pour chaque valeur :

Profession
Student                  27831
Architect                    8
Teacher                      6
'Digital Marketer'           3
'Content Writer'             2
Chef                         2
Doctor                       2
Pharmacist                   2
'Civil Engineer'             1
'UX/UI Designer'             1
'Educational Consultant'     1
Manager                      1
Lawyer                       1
Entrepreneur                 1
Name: count, dtype: int64

La part des étudiants et de : 99.9 %
```

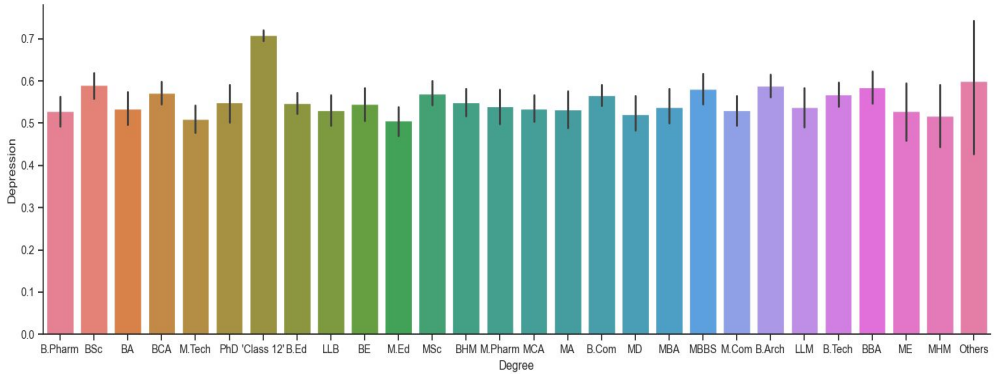




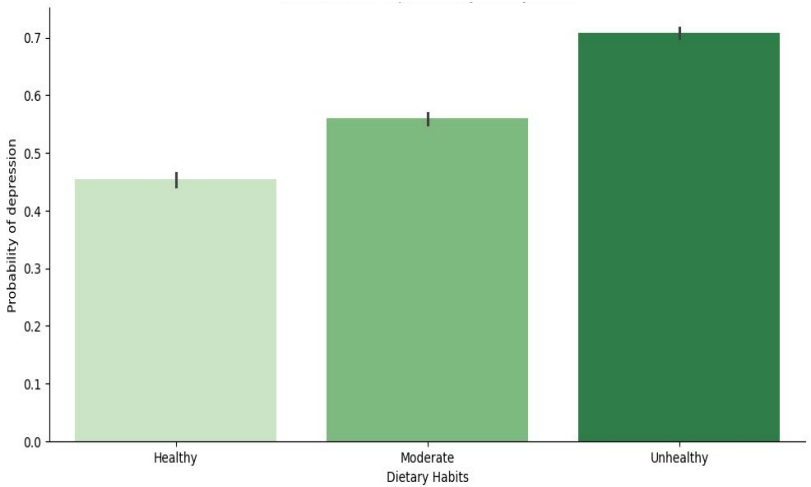

```
Nombre de lignes où 'Work Pressure' est égal à 0 : 27898
Nombre de lignes où 'Job Satisfaction' est égal à 0 : 27893
```

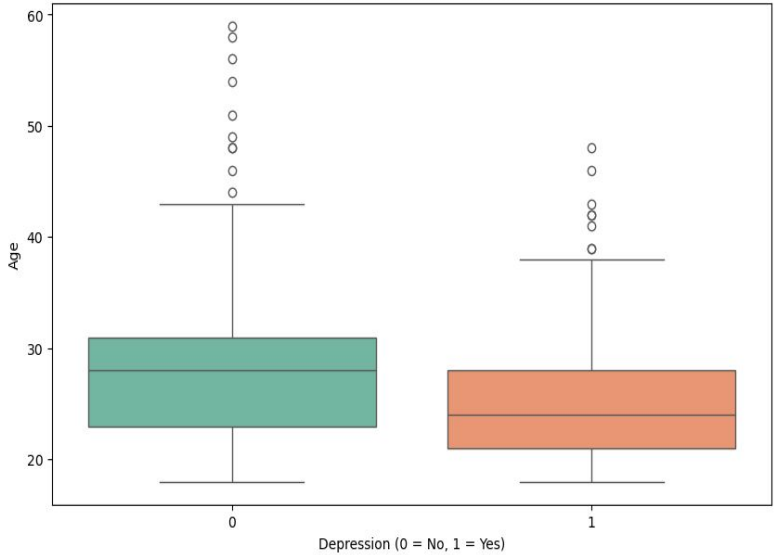
Exploratory Data Analysis



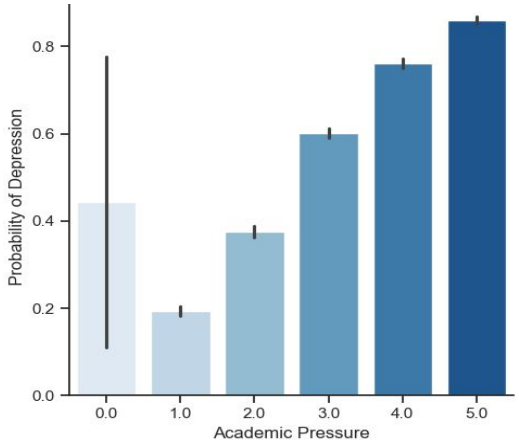


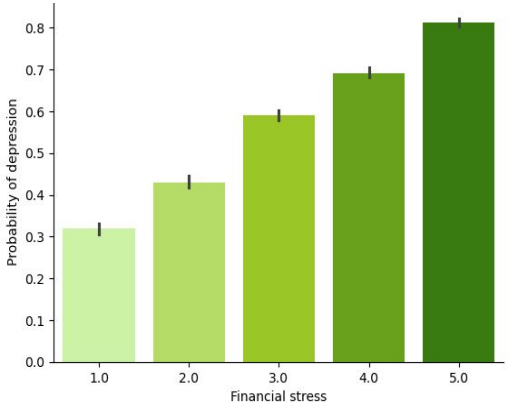
Exploratory Data Analysis

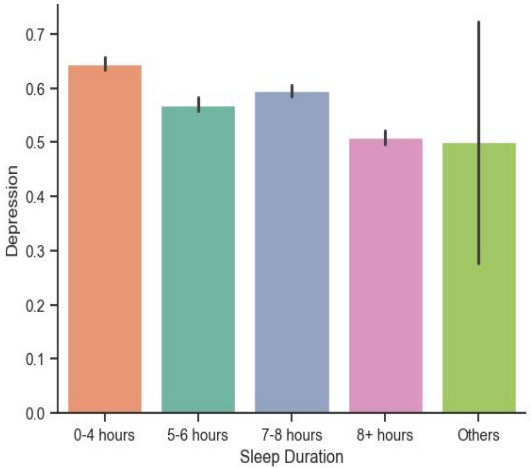




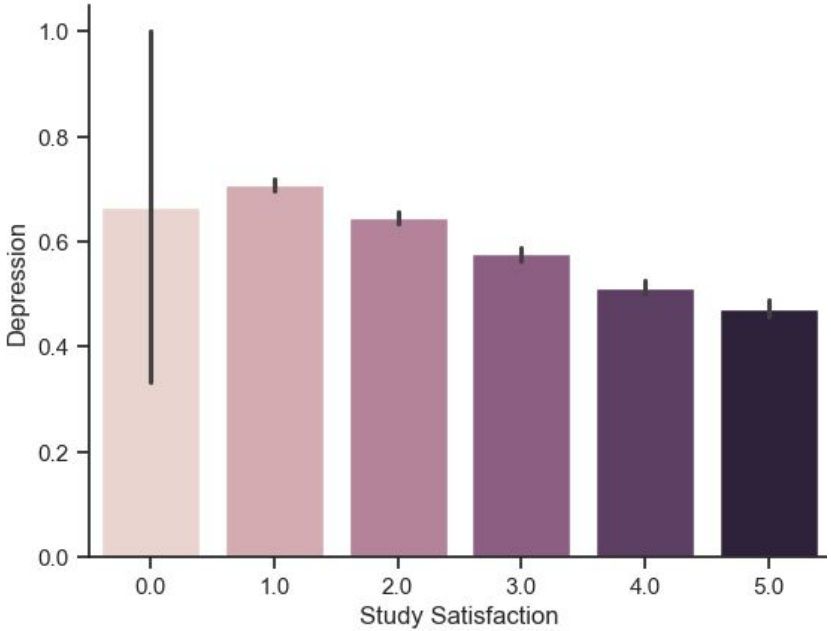
Exploratory Data Analysis

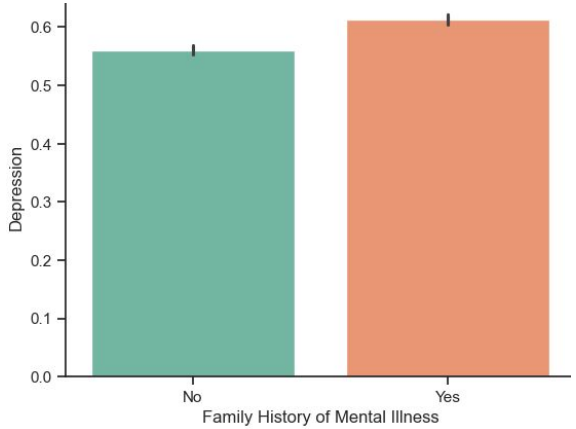


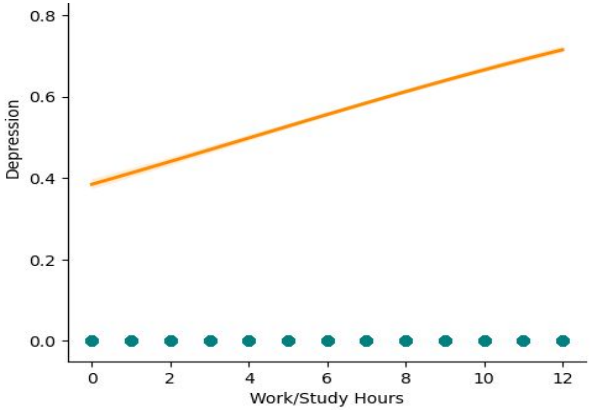




Exploratory Data Analysis







Feature Engineering

Data cleaning

- Vérifier l'existence des cellules vides
- Supprimer les données qui sortent des échelles: "?" et "Others" (0.24%)
- Renommer certaines colonnes et certaines valeurs pour plus de clarté
- Supprimer les outliers pour préserver la fiabilité du modèle (Âge)

DATA SET INITIAL

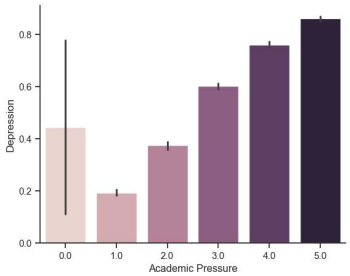
27901 lignes 18 colonnes

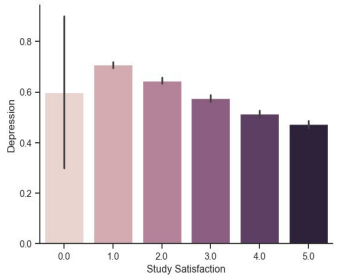
DATA SET FILTRÉ

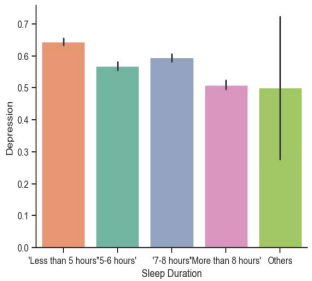
27856 lignes, 9 colonnes

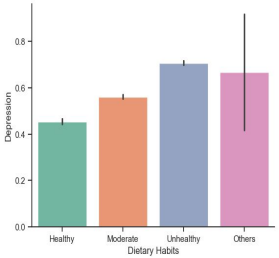
Feature Engineering

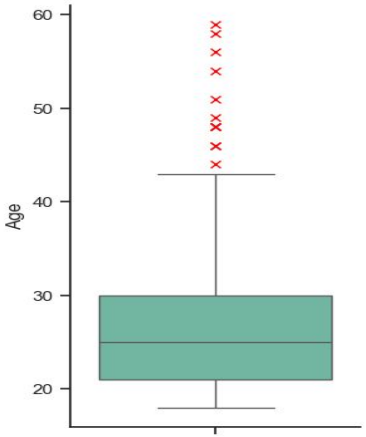
EDA nous a permis de choisir les données que nous allons analyser. Tool used - **Seaborn visualisation library**.

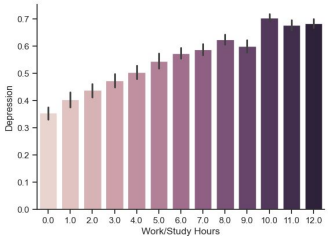


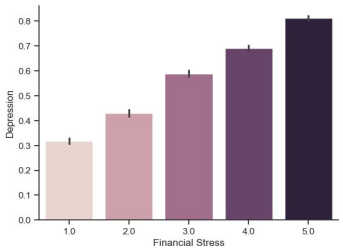


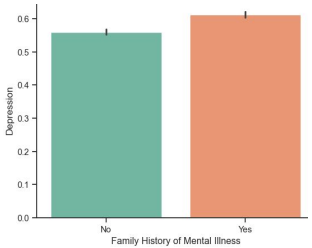












Feature Engineering

✨ **Dataset final**

1 . Division du dataset entre variable cible et variables explicatives

Variable cible (y)
'Dépression'

Variables explicatives (X)
'Pression', 'Satisfaction', 'Sommeil', 'Tps travail', 'Diff. financières', Alimentation', 'Antécédents', ‘Âge'

2 . Division du dataset en deux parties : train et test (80% - 20%)

3. Encodage des données :

- **StandardScaler** pour les variables numériques ('Pression', 'Satisfaction', 'Tps travail', 'Diff. financières', ‘Âge')
- **OneHotEncoder** pour les variables catégorielles nominales (Antécédents)
- **OrdinalEncoder** pour les variables catégorielles ordinales (Sommeil, Alimentation)

Machine Learning

Classification binaire - Apprentissage supervisé
L'étudiant est-il à risque de dépression ? (oui/non)

3 Modèles de classification

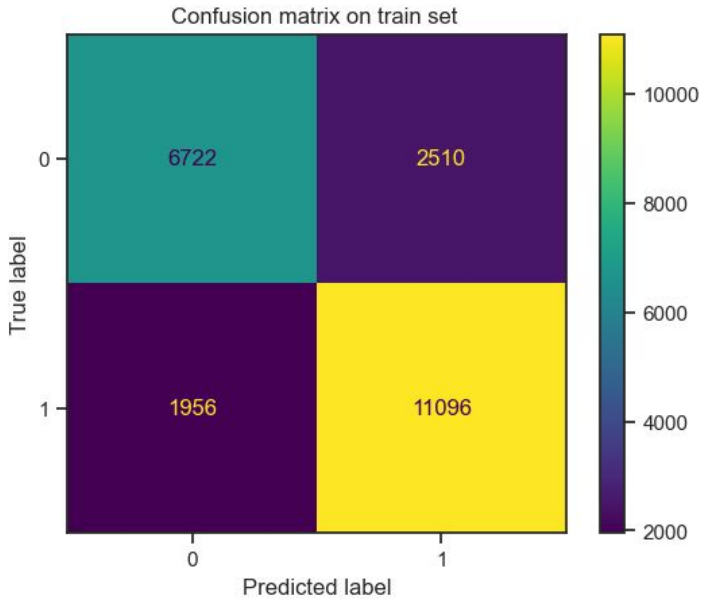
- Régression logistique
- Arbre de décision
- Forêt aléatoire

Évaluation des performances

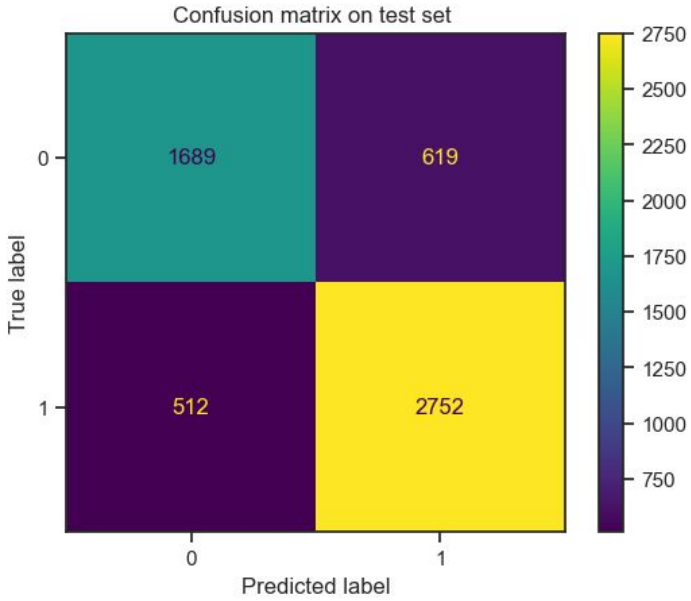
- Accuracy score : matrices de confusion
- Recherche d'hyperparamètres
- Precision // Recall

Machine Learning

📈 Régression logistique



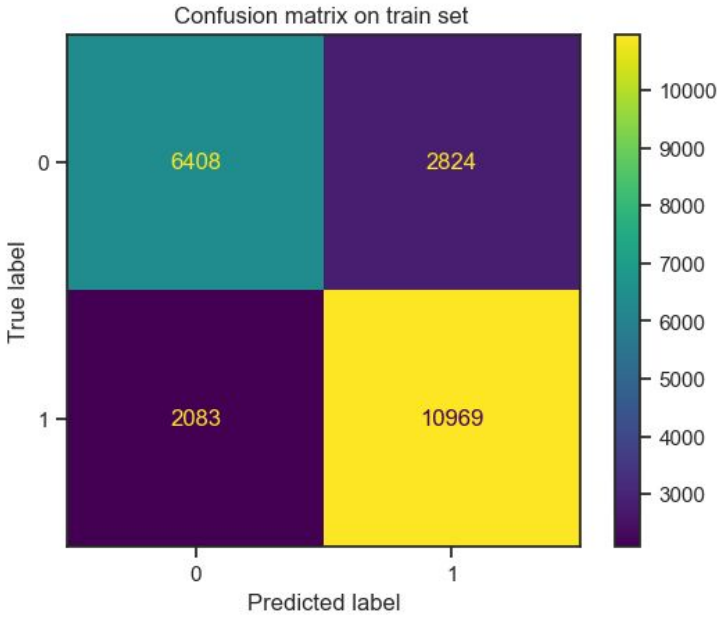
Train accuracy : 79,9 %



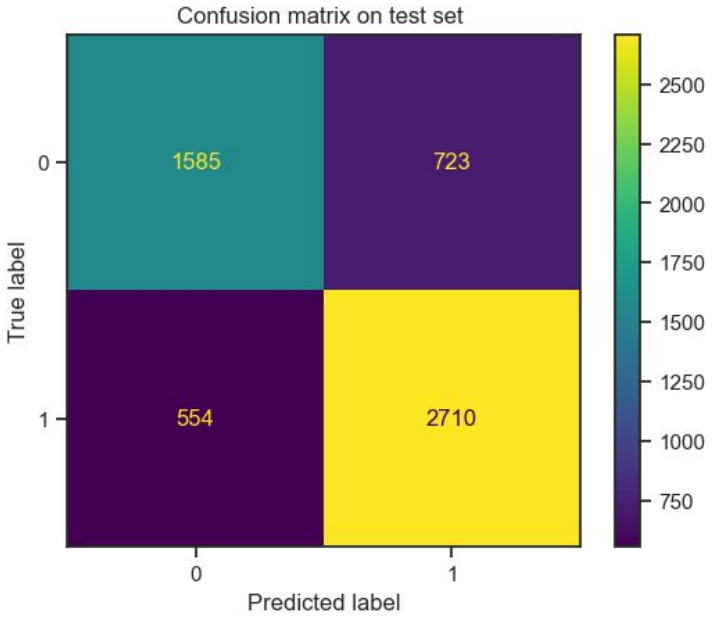
Test accuracy : 79,7 %

Machine Learning

🌳 **Arbre de décision** (*max_depth=5*)



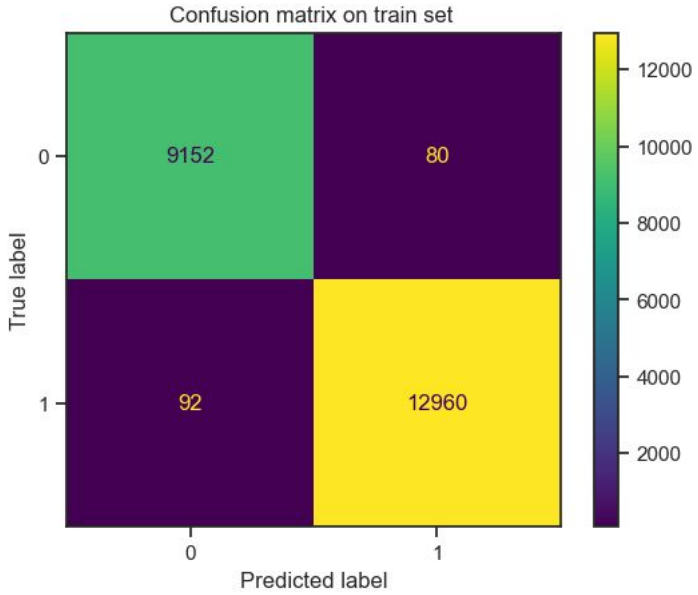
Train accuracy : 77,9 %

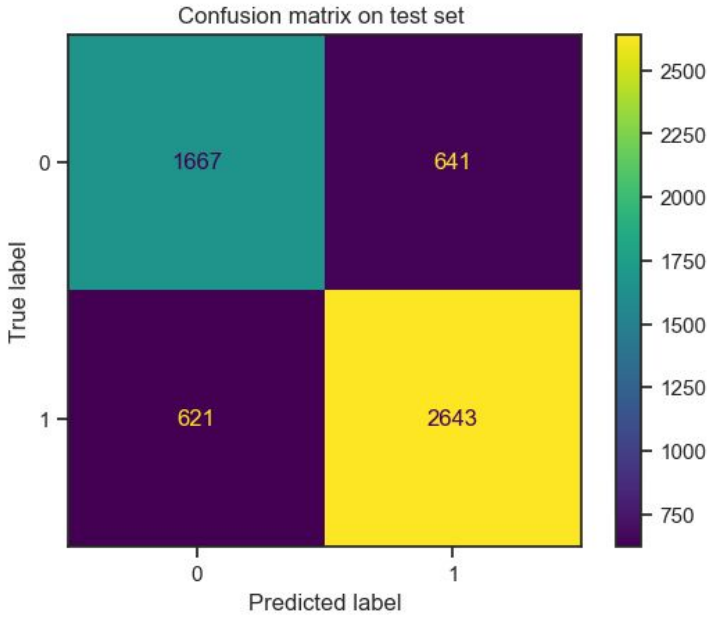


Test accuracy : 77,0 %

Machine Learning

🌳🌳🌳 Random Forest





Train accuracy : 99,2 %

⚠️ OVERFITTING ⚠️

Test accuracy : 77,3 %

Recherche d'hyper paramètres

🔎 Grid Search (nested for loops)

```
max_depth_to_test = [1, 2, 3]
n_estimators_to_test = [100, 1000, 5000]
```

```
____ Report model 1 ____
max_depth: 1
n_estimators: 100
Test recall    : 94.64 %
Test precision : 69.81 %
Test accuracy  : 72.88 %
Train accuracy : 72.90 %

____ Report model 2 ____
max_depth: 1
n_estimators: 1000
Test recall    : 95.71 %
Test precision : 68.64 %
Test accuracy  : 71.88 %
Train accuracy : 71.78 %

____ Report model 8 ____
max_depth: 3
n_estimators: 1000
Test recall    : 88.45 %
Test precision : 76.44 %
Test accuracy  : 77.26 %
Train accuracy : 77.67 %

____ Report model 9 ____
max_depth: 3
n_estimators: 5000
Test recall    : 88.73 %
Test precision : 76.33 %
Test accuracy  : 77.28 %
Train accuracy : 77.80 %
```

🏆 Sélection du modèle champion

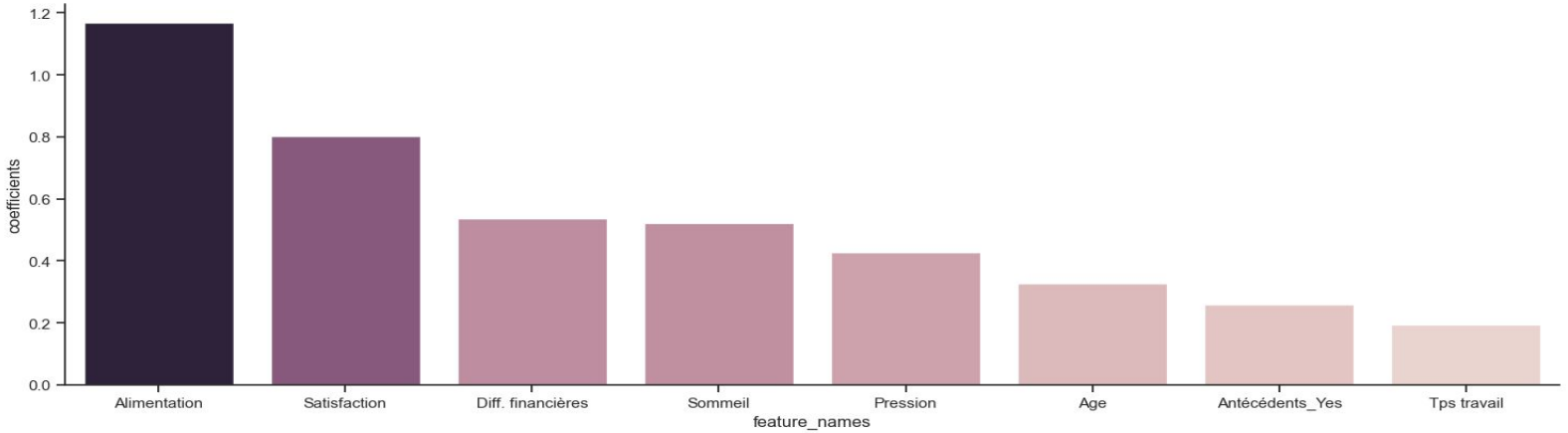
Faux positif ✅
Faux négatif ⚠️

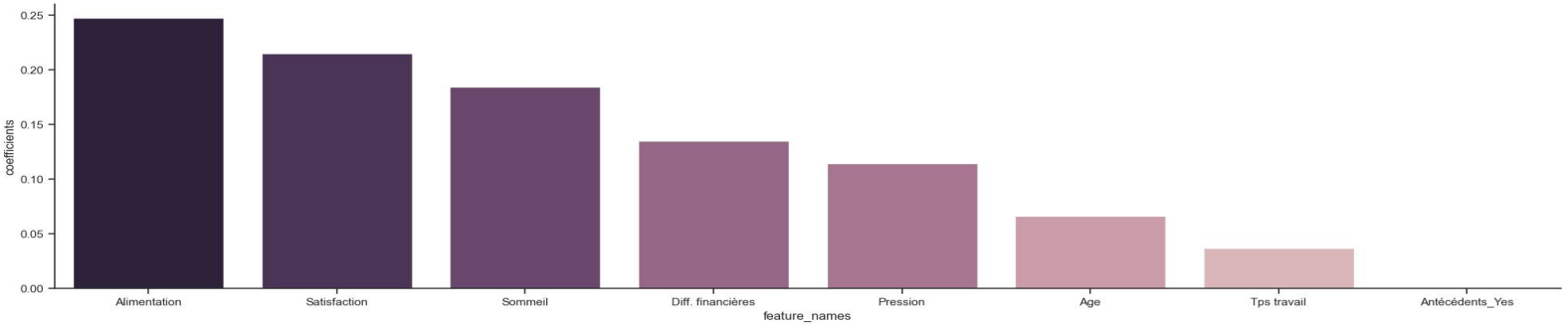
Recall > Accuracy > Precision

```
🥇____ Champion Model ____🥇
Parameters: {'max_depth': 1, 'n_estimators': 5000}
Test recall    : 96.38 %
Test precision : 68.05 %
Test accuracy  : 71.37 %
Train accuracy : 71.19 %
```

Feature importance

Régression logistique vs Random forest





Pistes de réflexions & améliorations 

Prise en compte de la variable ‘Pensées suicidaires’
Améliore significativement notre accuracy score mais il nous semble qu’il s’agit d’un symptôme de la dépression plutôt qu’un facteur de risque.

Revoir le feature engineering

- Supprimer les features avec un coef très bas (Antécédents)
- Ajouter une colonne binaire pour la variable ‘Class 12’ (0/1)

Revoir le sampling
50/50 dépressifs / non dépressifs

Traitement des outliers

- Ré-inclure les +43 ans pour éviter de manquer leur diagnostic
- Supprimer les lignes mentionnant une pression académique de 0.0 (0.03%)