2023-2024

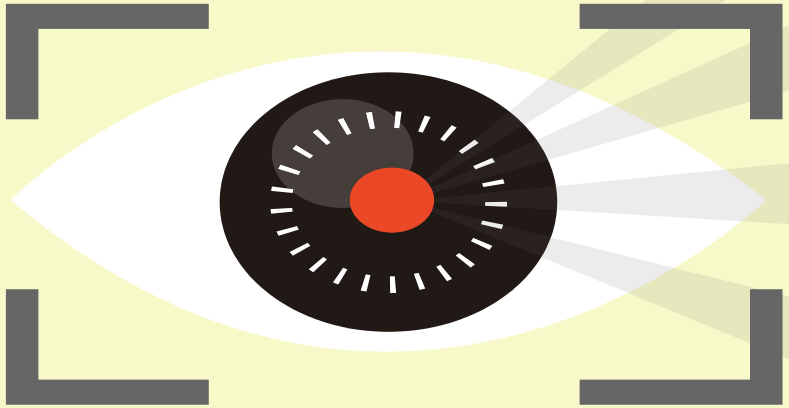
Machine Learning Project

DIABETES PREDICTION

Alè OUATTARA

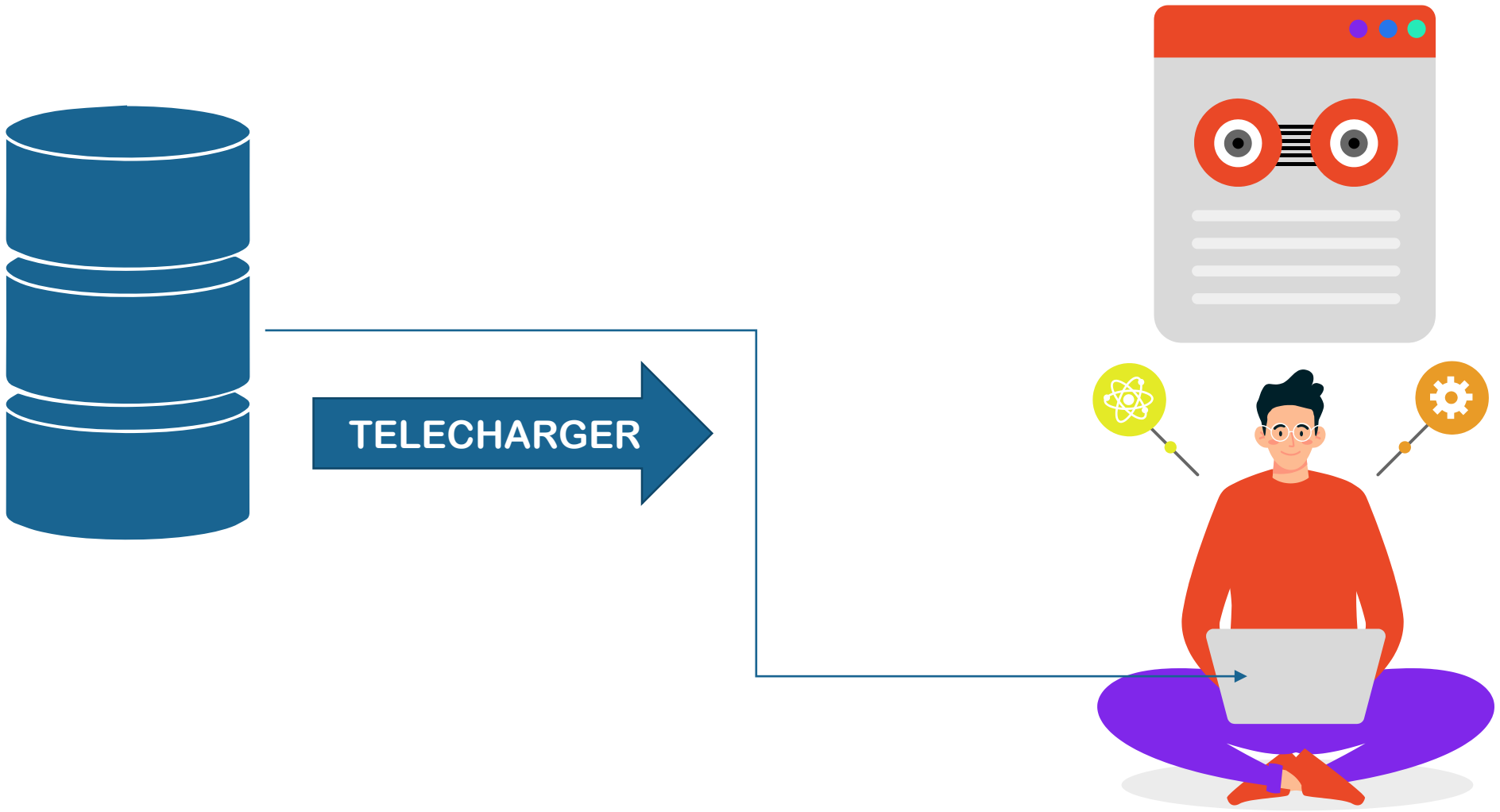
ETAPES DU PROJET

COMPREHENSION DU PROBLEME



- COLLECTE
- EXPLORATION
- PREPROCESSING
- MODEL- EVALUATION- OPTIMISATION

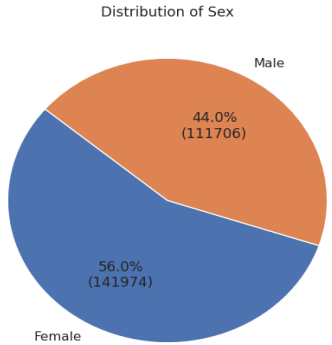
COLLECTE DES DONNEES

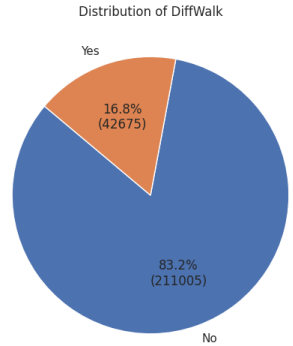


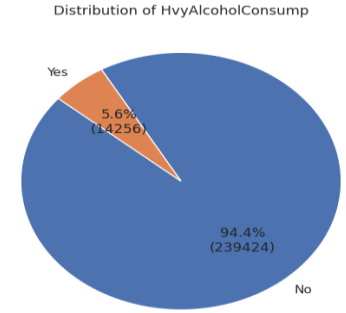
https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

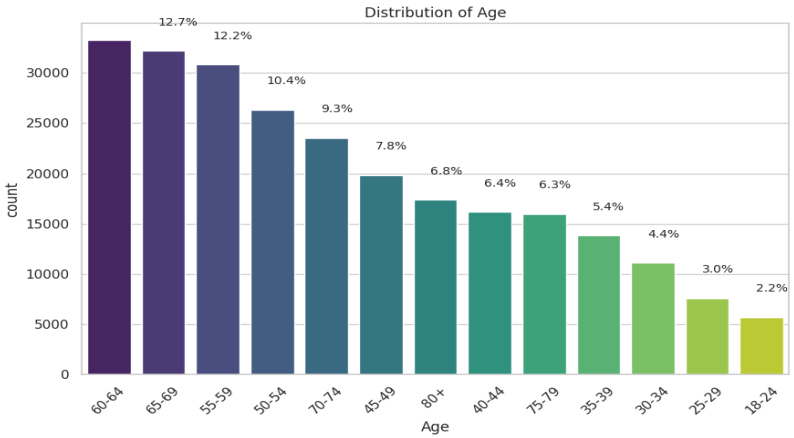
EXPLORATION DES DONNEES

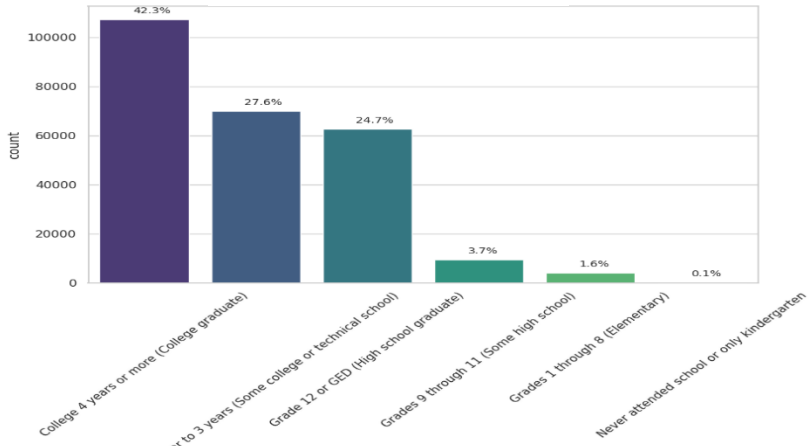
Statistiques descriptive :

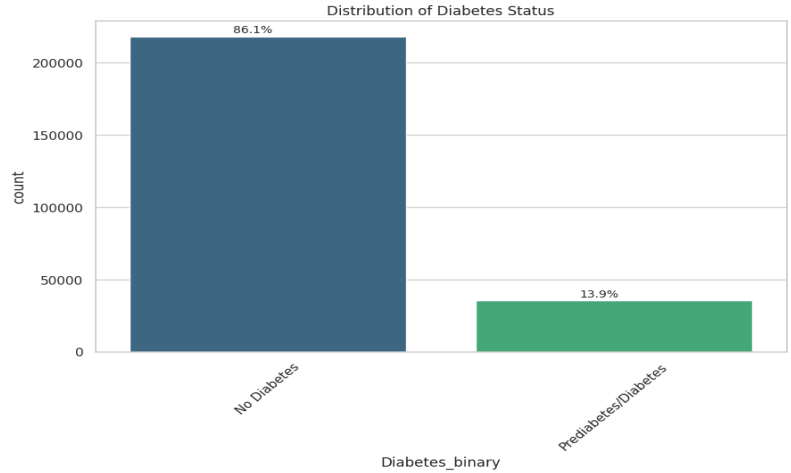


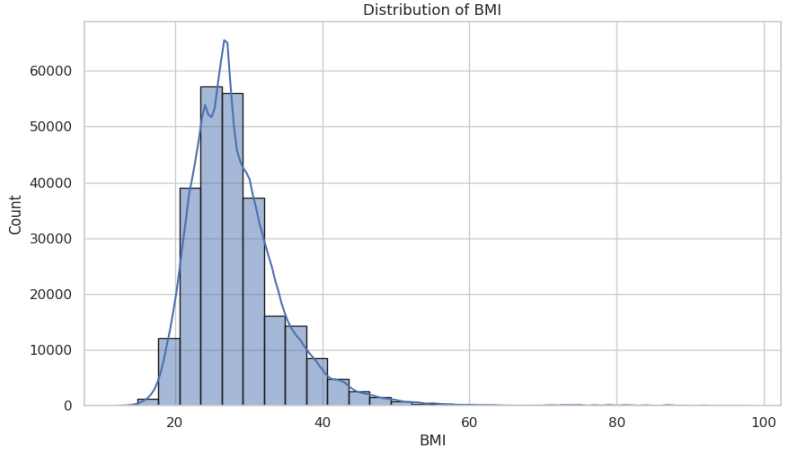


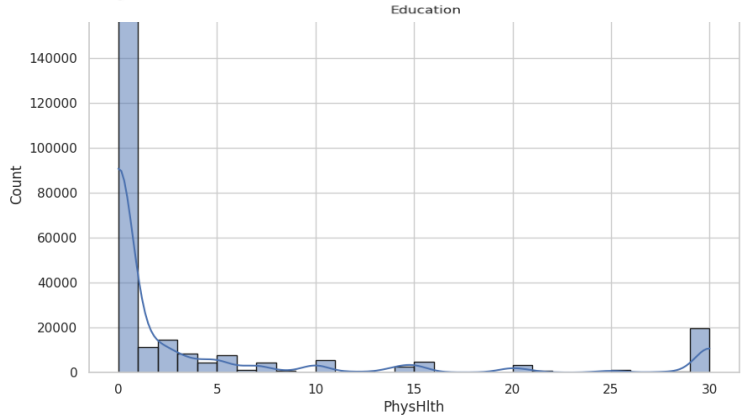


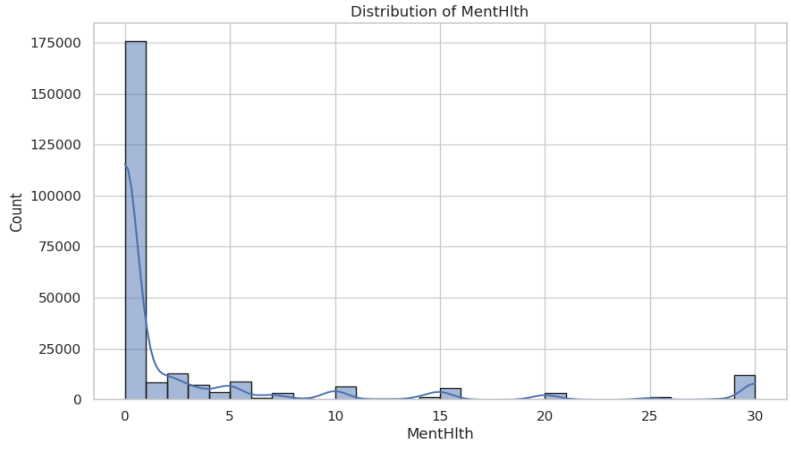












Statistiques descriptive :

Socio-Démographique

Genre

Femme = 56%

Education

Diplôme Secondaire = 94,6%

AGE

45-69 = 50%

Etat de santé

Santé Générale

82,9% affirme être au moins en bonne santé

Diabete

86,1% sans diabète

Facteurs de risque

IMC

IMC moyen 28,4

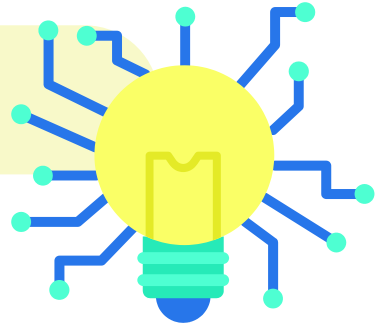
Activité physique

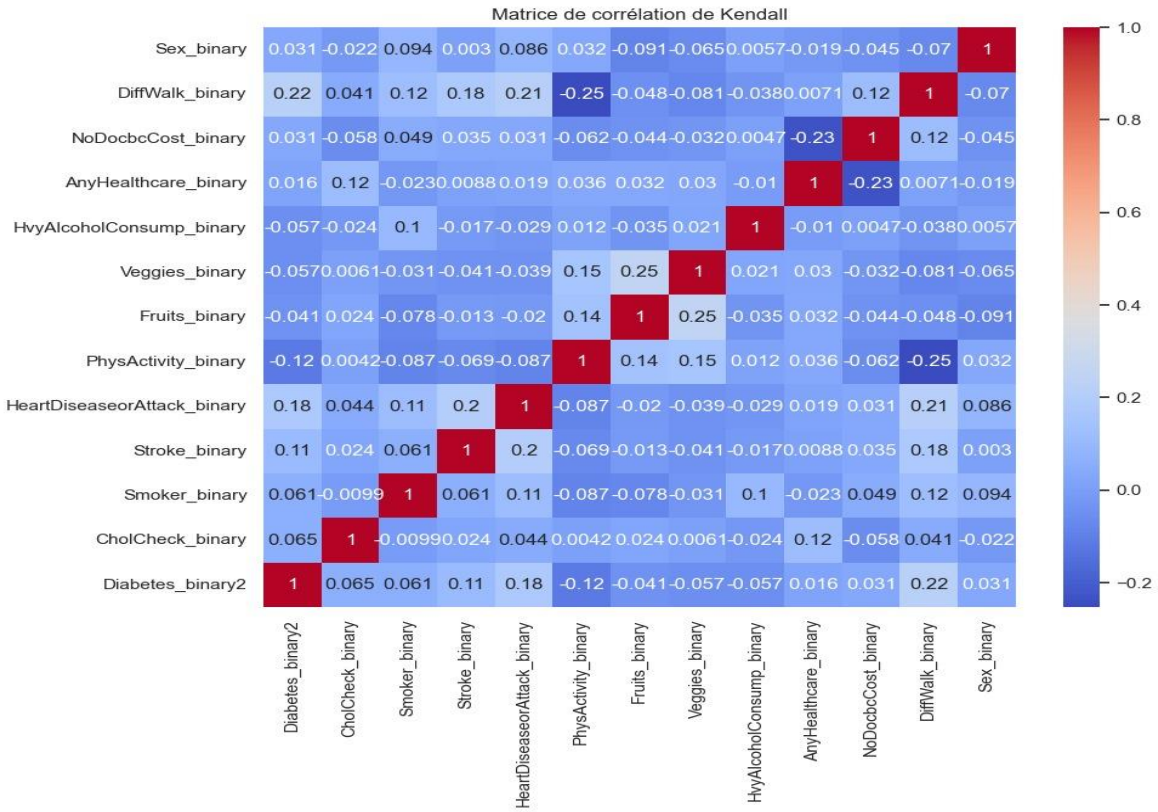
75,7% ont fait des activités physiques

Fumeur

55,7 % ne fument pas

CORELATION





	Coefficient de corrélation bisériale
MentHlth	0,069
PhysHlth	0,17
GenHlth	0,293
Income	-0,163
BMI	0,216

P R E P R O C E S S I N G

Encodage

OneHotEncoding

Normalisation

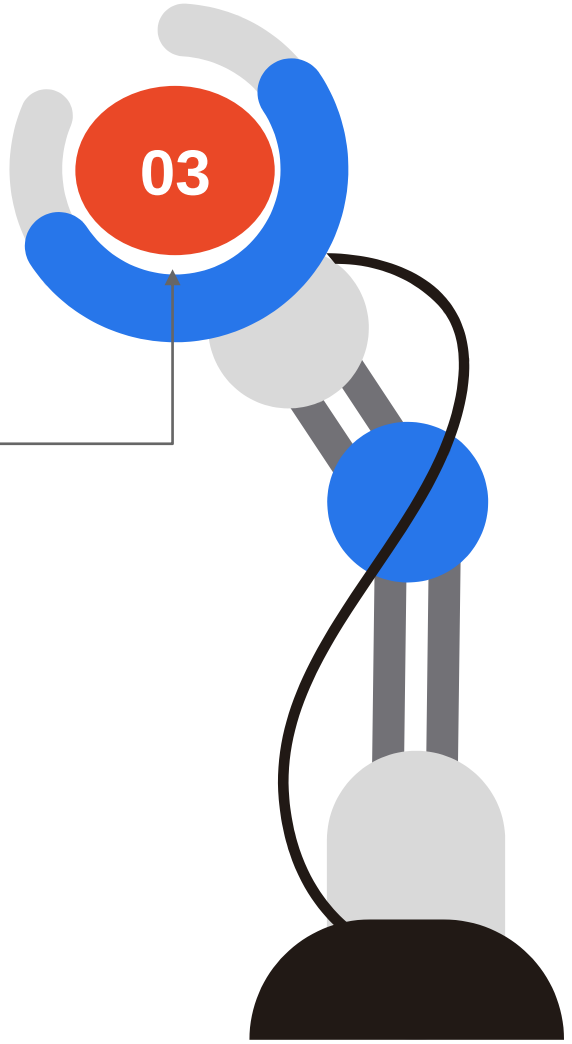
Standarscaler

Valeur manquante

01

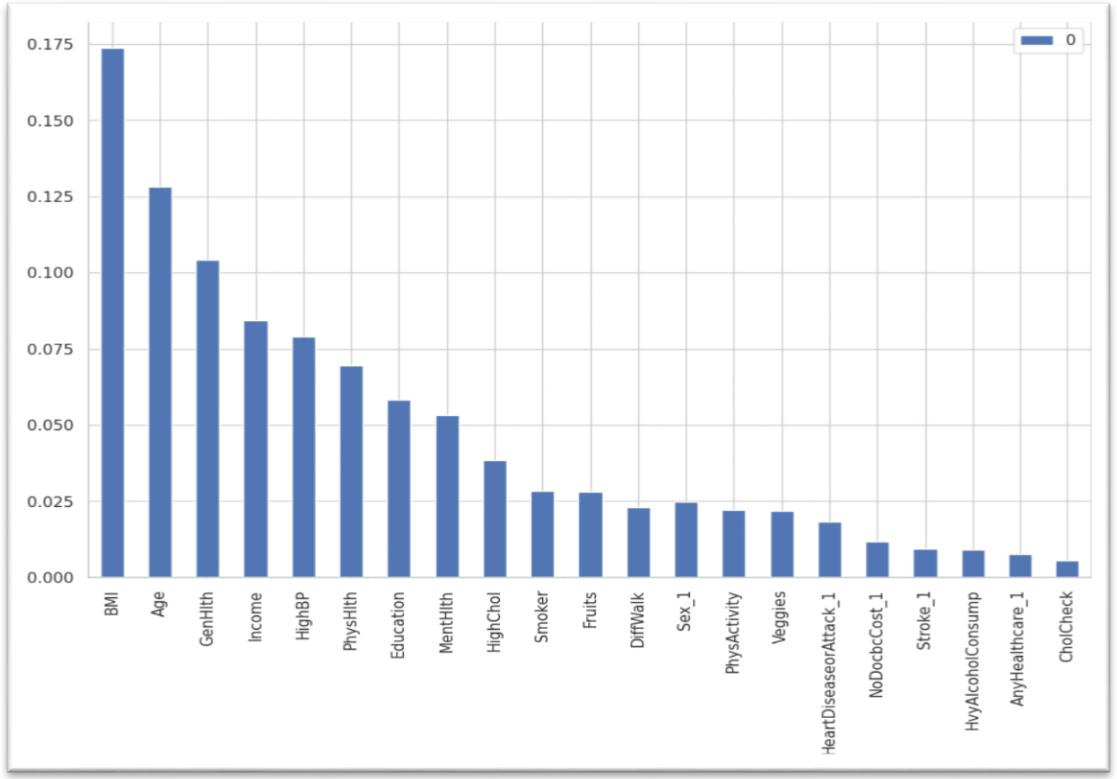
02

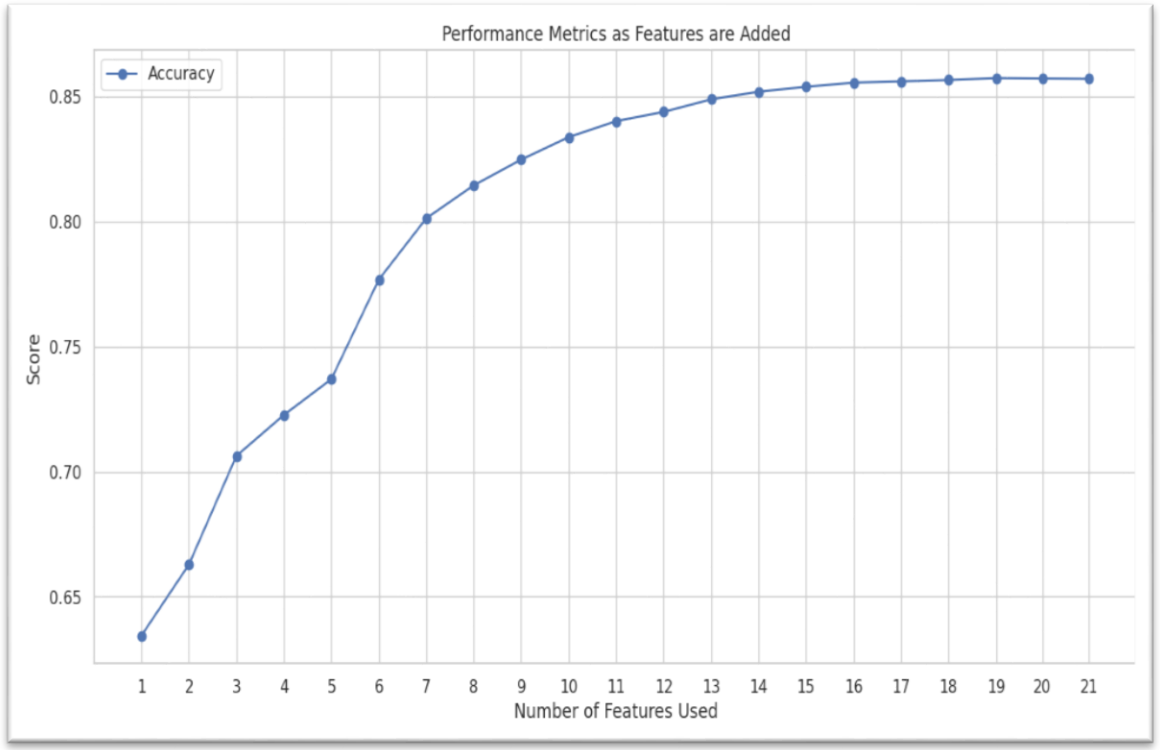
03



Data cleaning

Features Selection





Model- Evaluation

RandomForest
Adaboost
KNN

Choix du model

Split Data 80/20

Choix des métrics

Accurancy
Recall
F1_score

Modèle	Accuracy	Recall	F1 Score
Random Forest	0.8555	0.1534	0.2318
AdaBoost	0.7563	0.7032	0.4506
KNN	0.7074	0.6357	0.3818

Données déséquilibré : Oversampling

Modèle	Accuracy	Recall	F1 Score
Random Forest	0.8555	0.1534	0.2318
AdaBoost	0.8600	0.1900	0.2900
KNN	0.8500	0.2100	0.2800

O P T I M I S A T I O N

N_stimator = 200
[50, 100, 200]

Learning_rate = 1
[0.01, 0.1, 1]

Accuracy	Recall	F1 Score
0.7959	0.6039	0.4569

P R E D I C T I O N

Voir NoteBook


THANK YOU