FINAL EXAM
-
PREDICTIVE MODEL FOR DIABETES

13/12/2023

82.05 - Análisis Predictivo - Final

Elsa DOYEN (65990)

AGENDA

01
Study case and data presentation

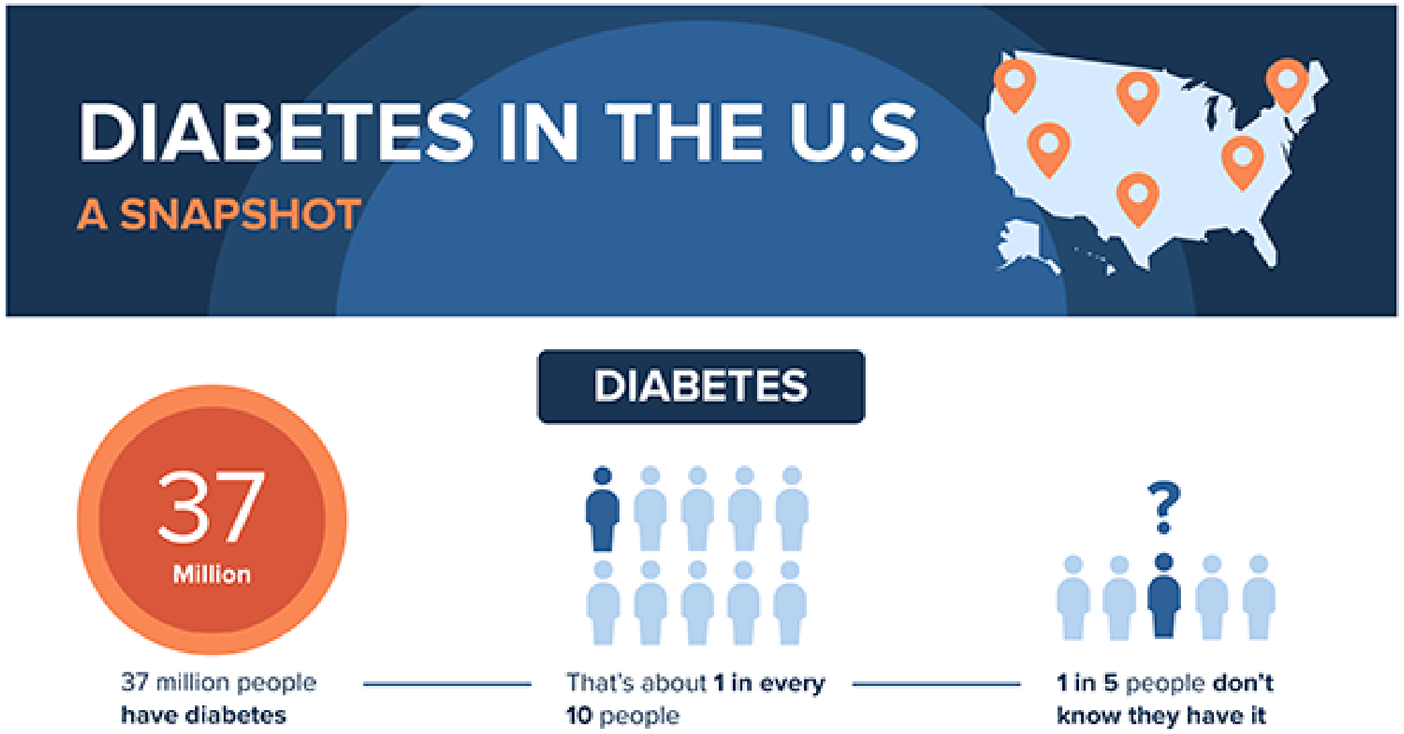
02
Exploratory analysis

03
Testing several models

04
Model selection

05
Conclusion

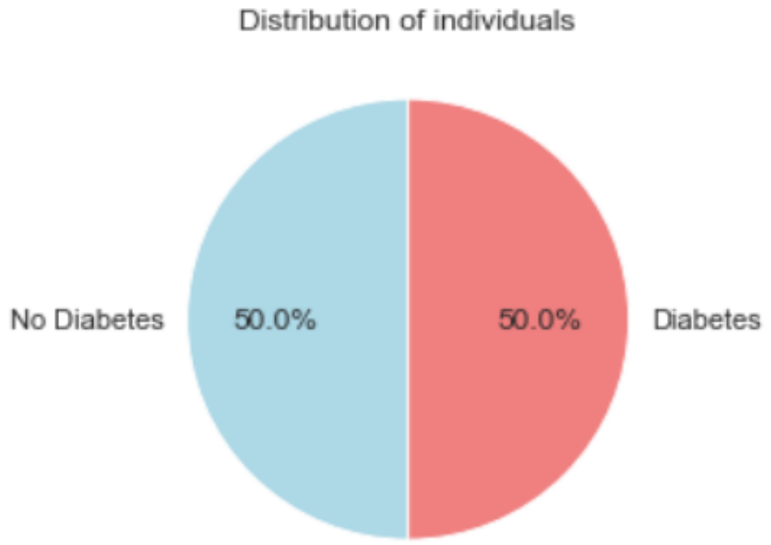
CASE STUDY



- Insurance company
- Know the percentage of risk that an individual has diabetes
- Prediction model from a database

DATASET PRESENTATION

- Response variable : Diabetes (0 or 1)
- 21 feature variables
- 70 692 survey responses



Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9.0	4.0	3.0
0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7.0	6.0	1.0
0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9.0	4.0	8.0
0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11.0	3.0	6.0
0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11.0	5.0	4.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	0.0	1.0	0.0	2.0	0.0	2.0	0.0	1.0	10.0	6.0	8.0
0.0	1.0	0.0	1.0	30.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	3.0	0.0	14.0	0.0	0.0	9.0	6.0	7.0
0.0	1.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	3.0	0.0	0.0	1.0	0.0	11.0	4.0	4.0

DATASET PRESENTATION

Variables :

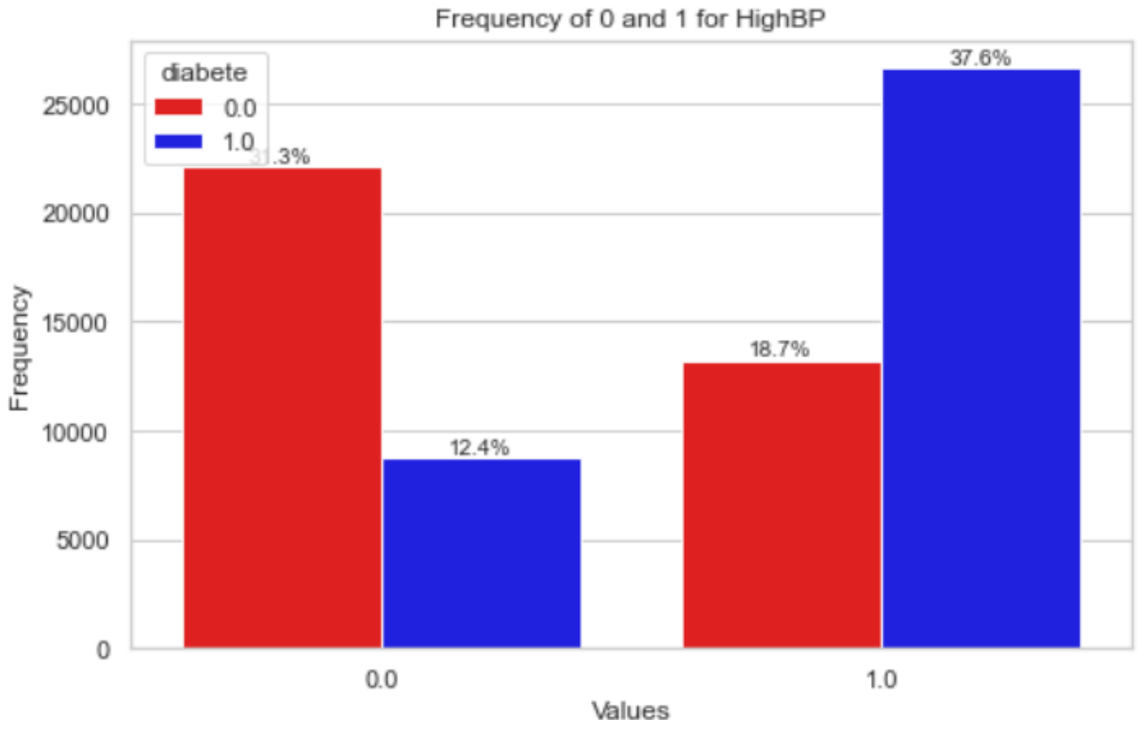
Categorical :

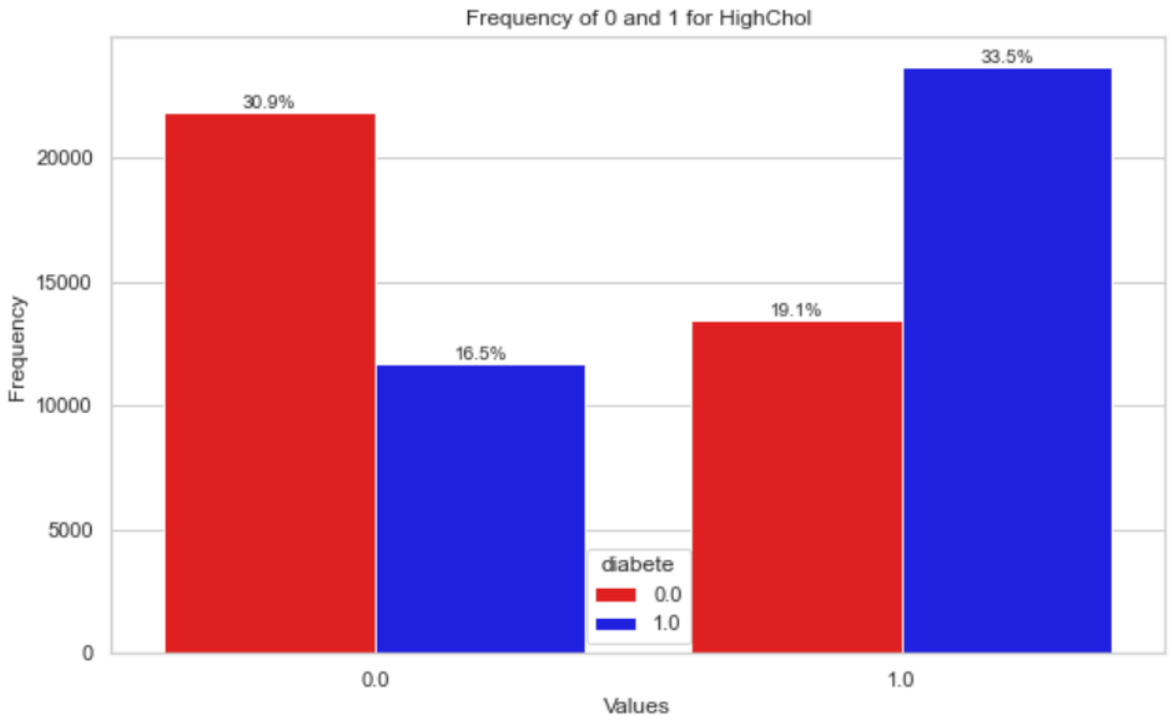
- HighBP
- HighChol
- CholCheck
- Smoker
- Stroke
- HeartDiseaseorAttack
- PhysActivity
- Fruits
- Veggies
- HvyAlcoholConsump
- AnyHealthcare
- NoDocbcCost
- DiffWalk
- Sex
- GenHlth
- Age
- Education
- Income

Numerical :

- MentHlth
- PhysHlth
- BMI

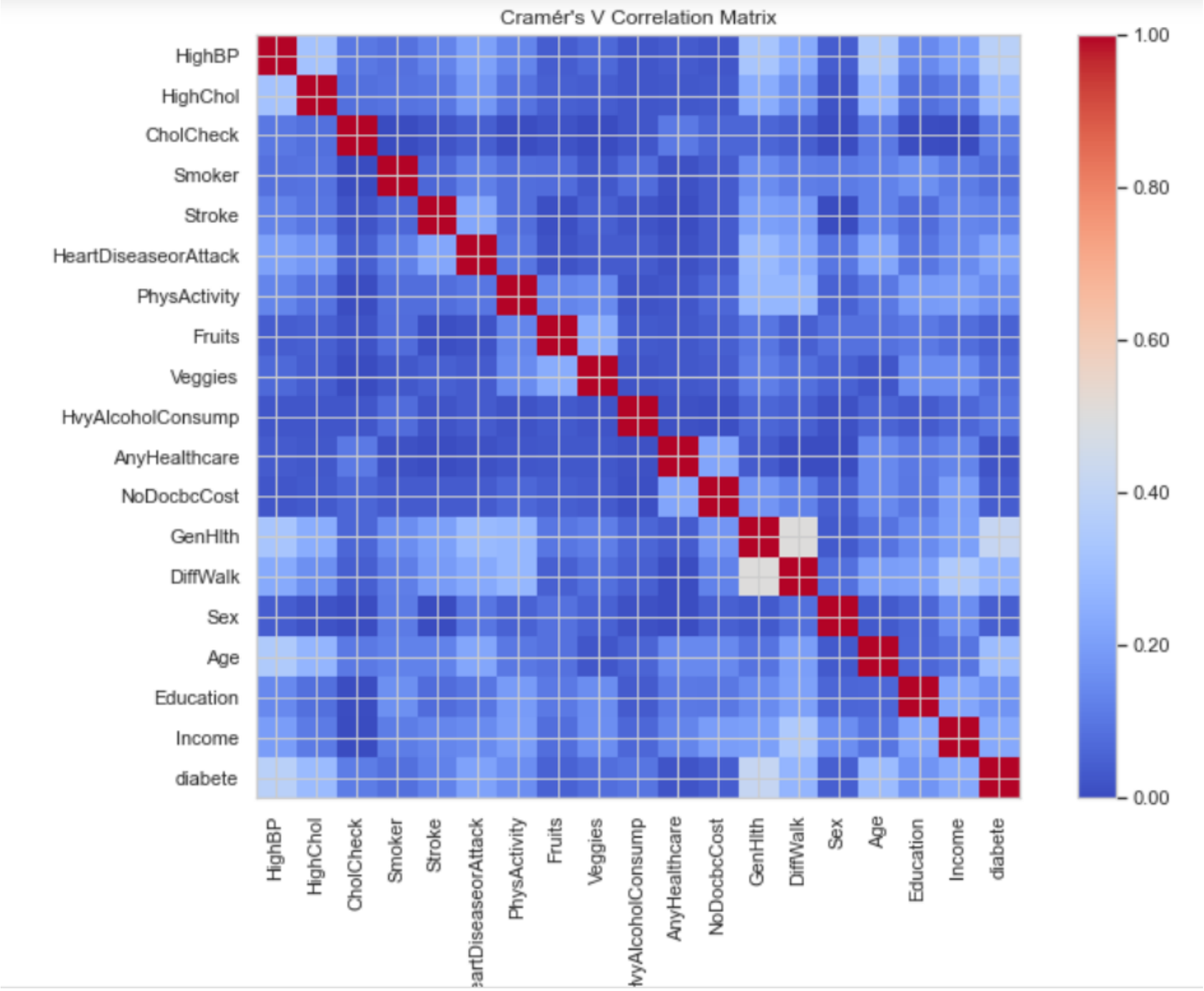
EXPLORATORY ANALYSIS





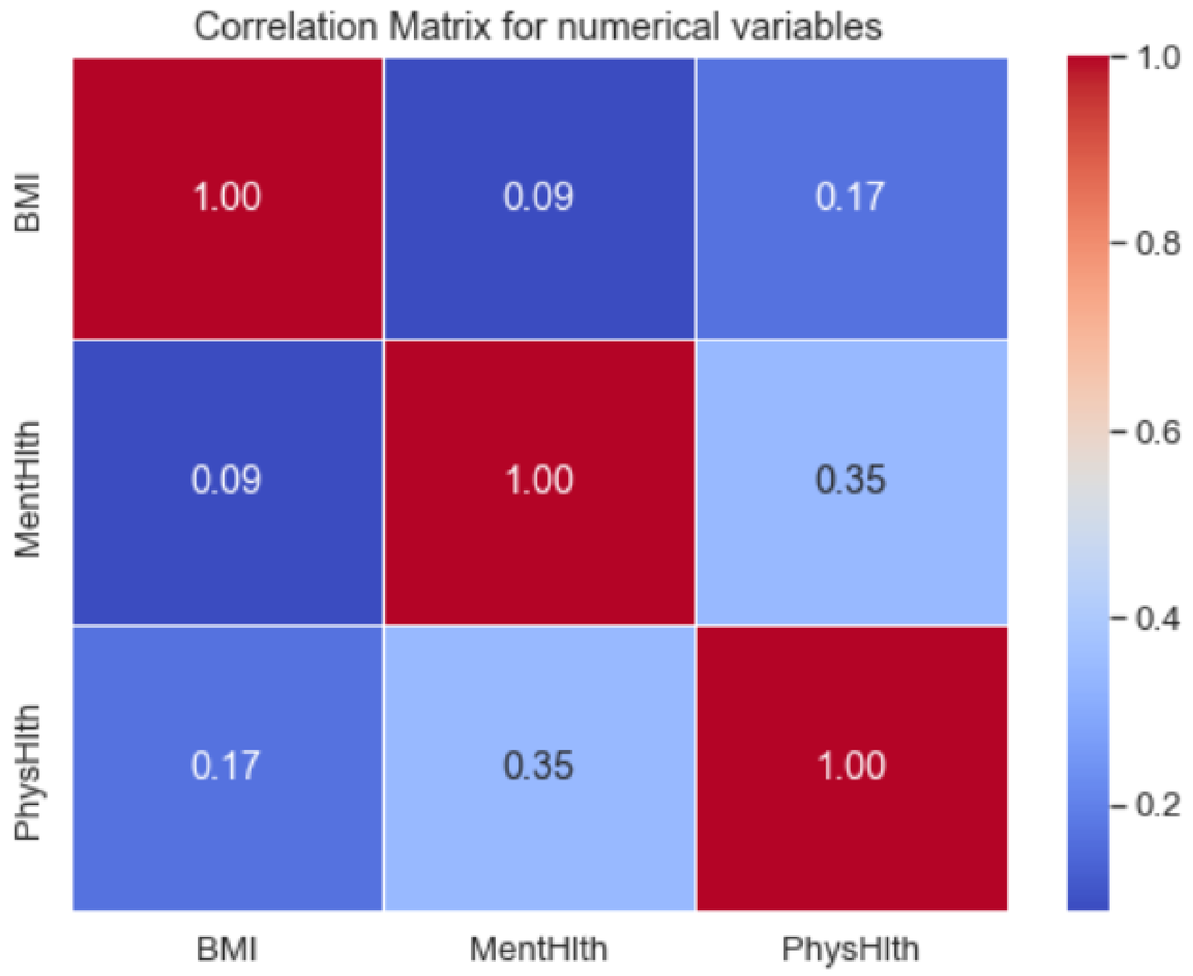
EXPLORATORY ANALYSIS

Correlation



EXPLORATORY ANALYSIS

Correlation



EVALUATION

- Score chosen for tests : **ROC AUC score**

- **Split database**
  - training data 80%
  - test data 20 %

→ **Prediction of the probabilities to have diabetes**

TESTING SEVERAL MODELS

Model	ROC AUC score	Time (sec)
Logistic Regression	0.81746	0.4135
Random Forest	0.81800	2.0975
XGBoost	0.82431	0.7825
KNN	0.76441	0.0190

→ Selection of XGBoost model

MODEL SELECTION

XGBoost model

Hyperparameters to test :

- n_estimators
- max_depth
- learning_rate

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'max_depth': [3, 4, 5],
    'learning_rate': [0.1, 0.01, 0.001]
}
```

```
Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 300}
ROC AUC: 0.8315537059643547
```

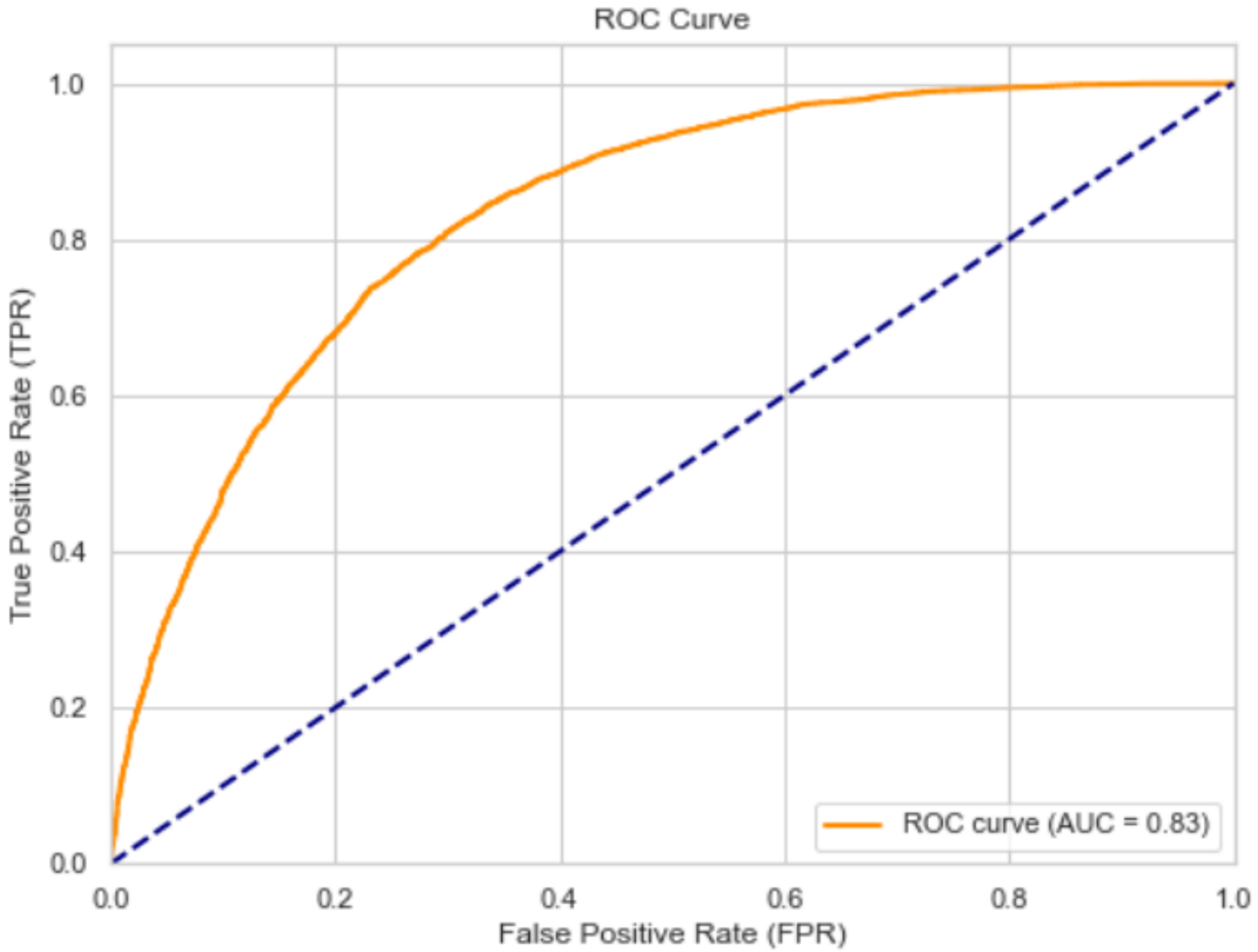
MODEL SELECTION

XGBoost model

Choice :

- n_estimators = 300
- max_depth = 3
- learning_rate = 0.1

ROC AUC score = 0.83155

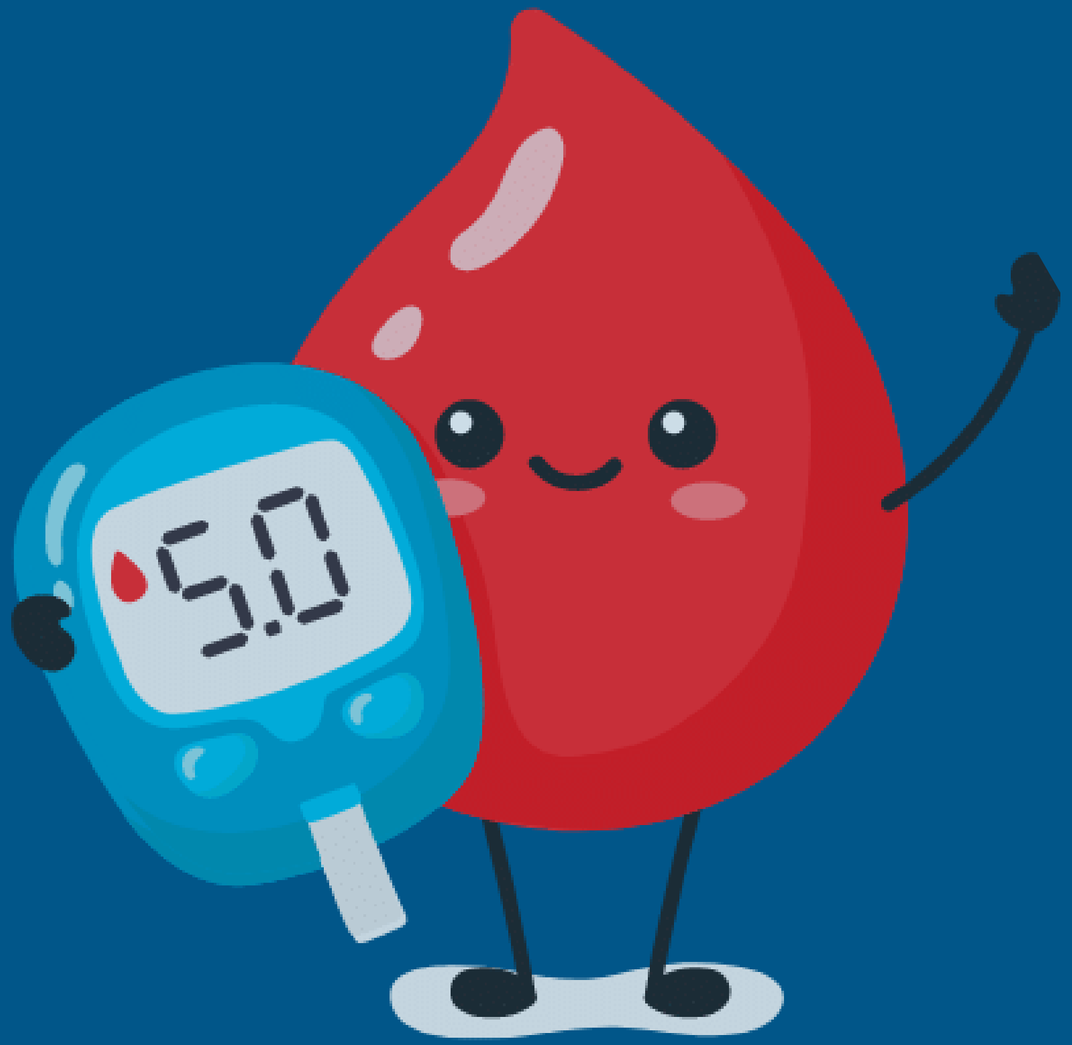


CONCLUSION

- Easy-to-use dataset (no cleaning)
- Test different models
- Selection of XGBoost model and its hyperparameters
- Final AUC ROC score of 0.83

Limits and improvements :

- Correct score but can be improved
- Other possible models
- Other hyperparameters possible



Thanks !