ITBA
Instituto Tecnológico de Buenos Aires

CASE STUDIES

-

ANALYSIS OF A DIABETES HEALTH INDICATORS DATA SET

December 2023

Thomas SIMON (65931)

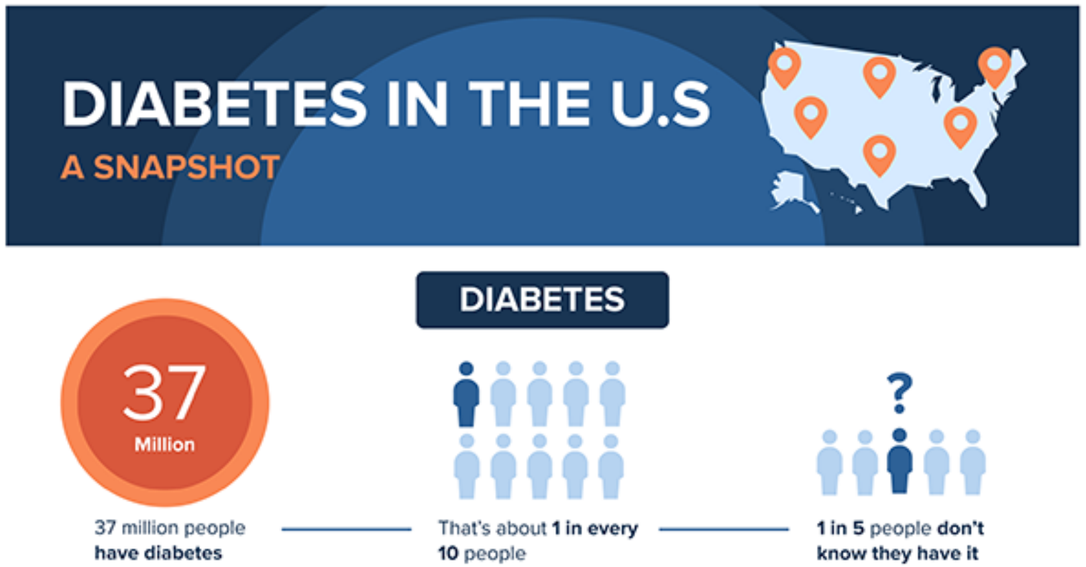
82.05 - Análisis Predictivo - EXAMEN FINAL

ITBA

AGENDA

1. Business Case
2. Dataset
3. Exploratory analysis
4. Evaluation
5. Reference model
6. Model selection
7. Description of the final model
8. Limitations and possible improvements
9. Conclusion

BUSINESS CASE



Profitability analysis

- Who we are ? An insurance company
- Our goal ? Assess and manage the risks associated with individuals with diabetes
- What for ? Evaluate the model's predictions to determine the risk associated with insuring each individual

ITBA

DATASET

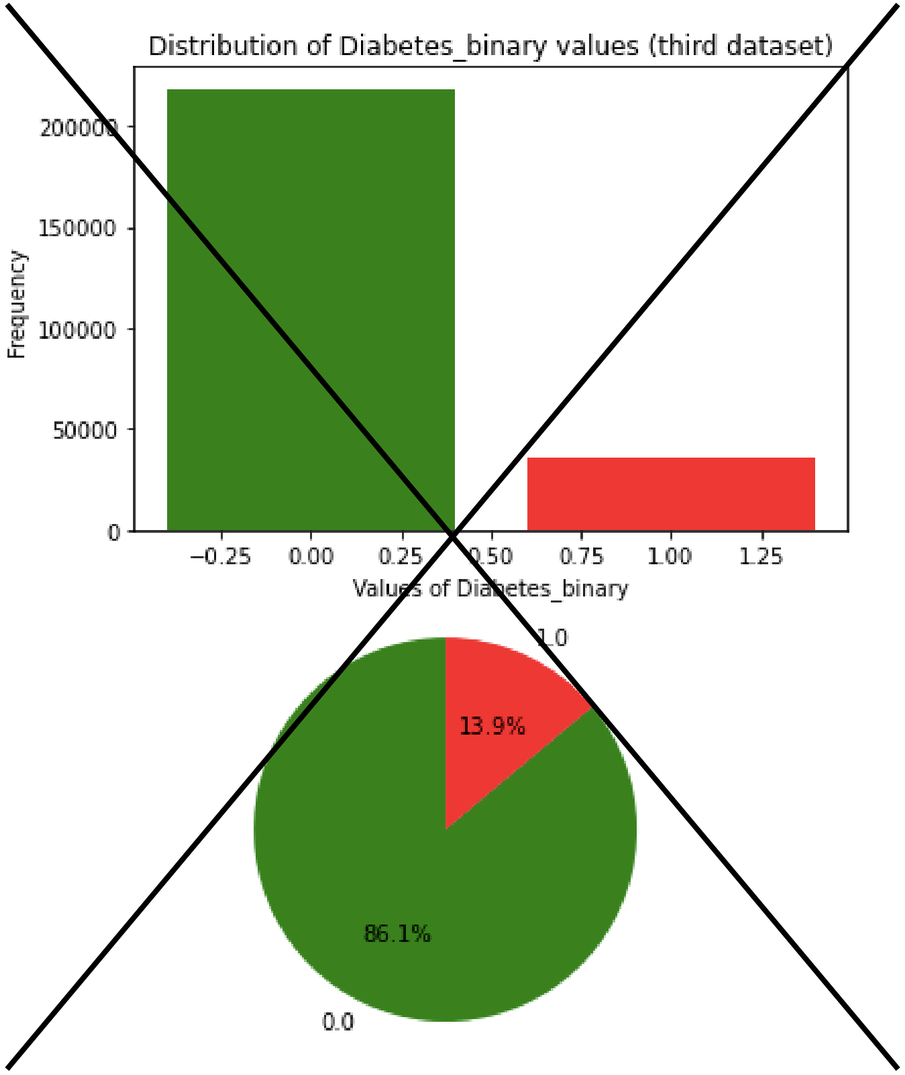
kaggle

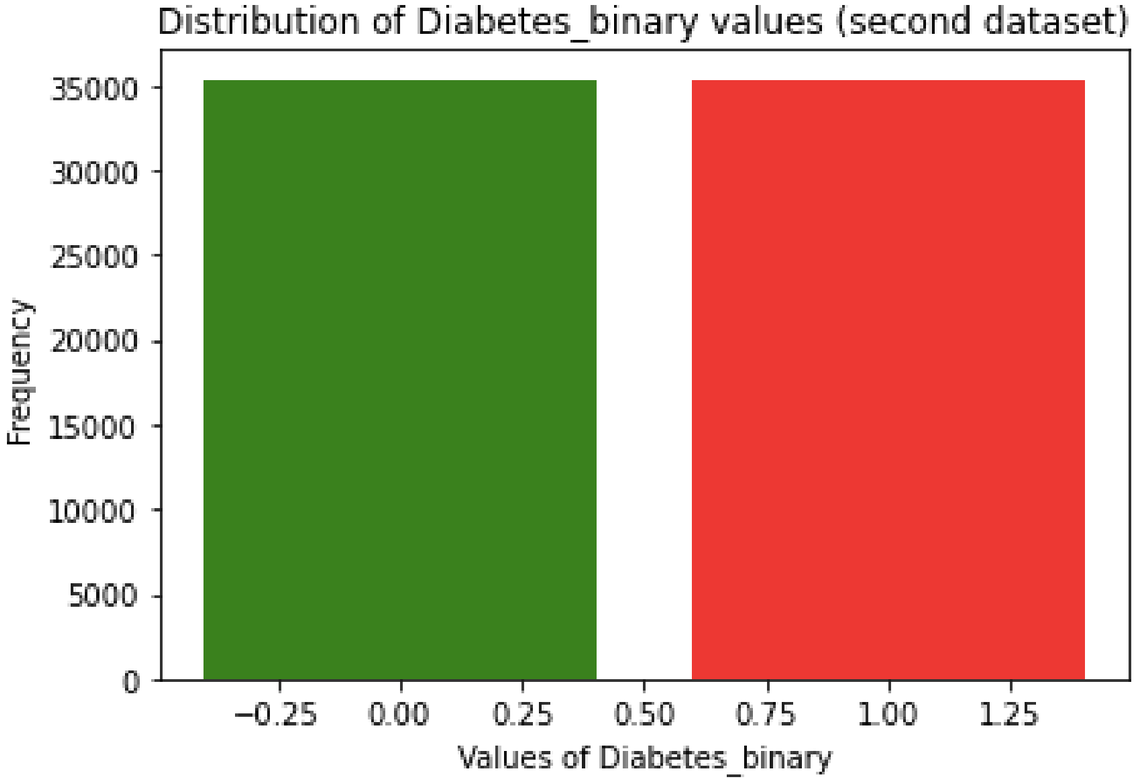
BRFSS™
Behavioral Risk Factor Surveillance System

CDC
CENTERS FOR DISEASE™
CONTROL AND PREVENTION

- 70,692 survey responses
- Equal 50-50 split of respondents with no diabetes and diabetes
- 21 feature variables
- Dataset is balanced

DATASET


Distribution of Diabetes_binary values (third dataset)
Frequency
200000
150000
100000
50000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Values of Diabetes_binary
1.0
13.9%
86.1%
0.0


Distribution of Diabetes_binary values (second dataset)
Frequency
35000
30000
25000
20000
15000
10000
5000
0
−0.25
0.00
0.25
0.50
0.75
1.00
1.25
Values of Diabetes_binary

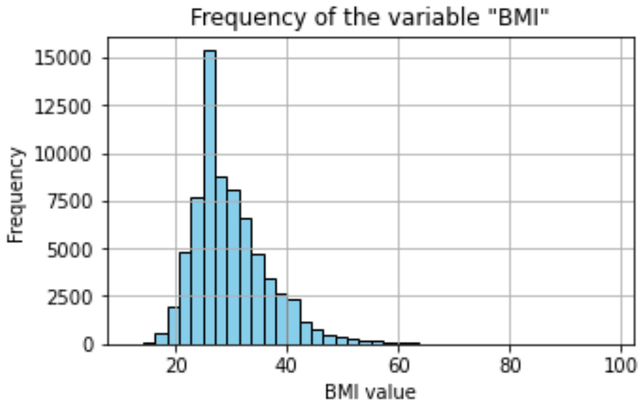
EXPLORATORY ANALYSIS

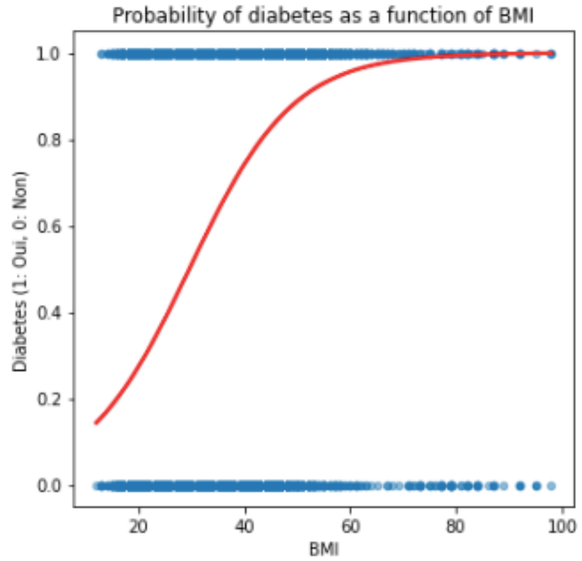
18 categorical variables

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

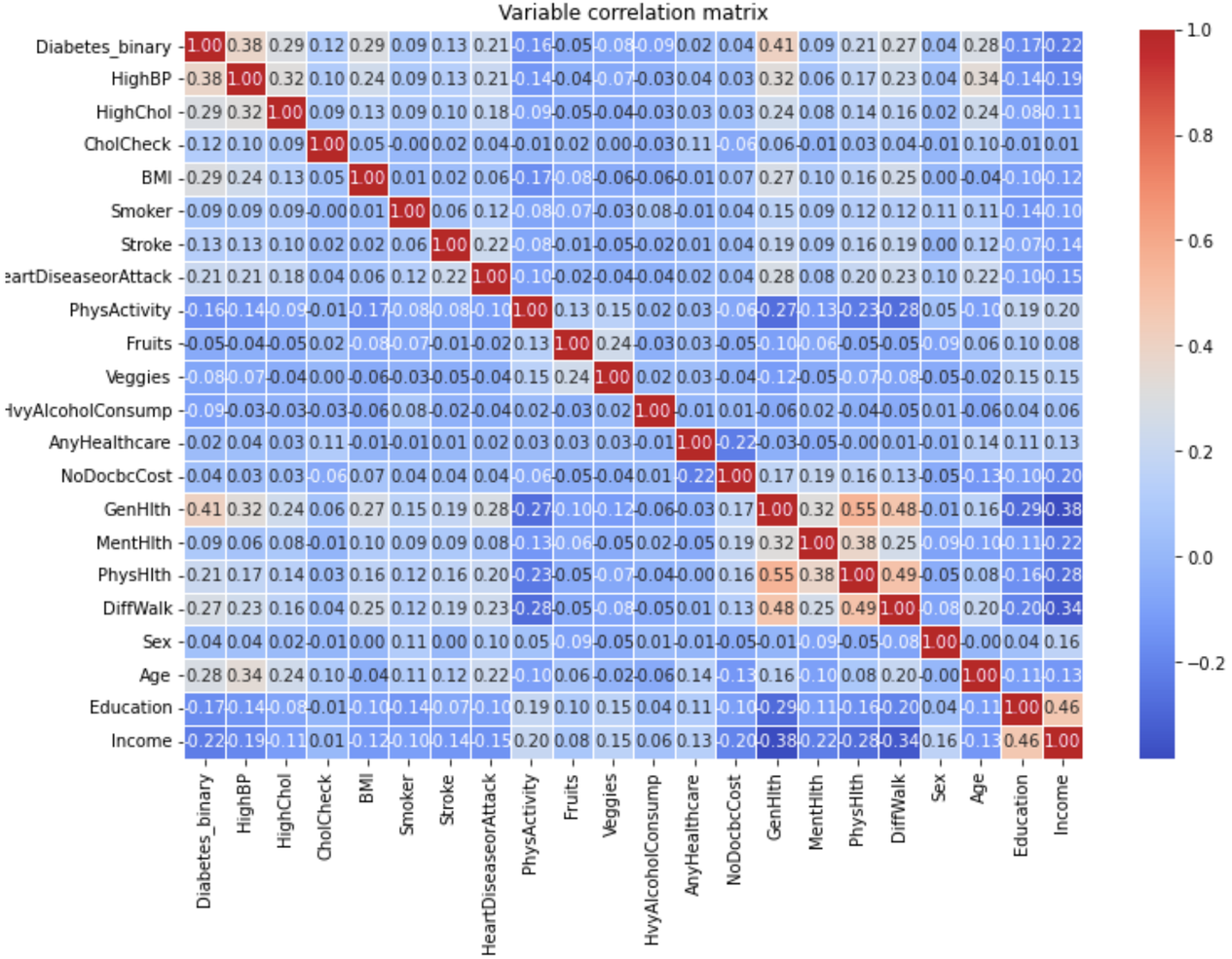
3 quantitative variables

'BMI', 'MentHlth', 'PhysHlth'





EXPLORATORY ANALYSIS



No significant correlation between the variables means no redundant variables

ITBA

EVALUATION

- Accuracy : it measures the overall correctness of the model's predictions.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

- Precision : it measures the accuracy of positive predictions.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

- Recall (Sensitivity) : it measures the ability of the model to capture all the positive instances.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

- F1 Score : it combines precision and recall, providing a balance between the two.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

REFERENCE MODEL

Logistic regression

- Simplicity
- Interpretability
- Linearity

Data split : 70-30

- X = 70%
- Y = 30%

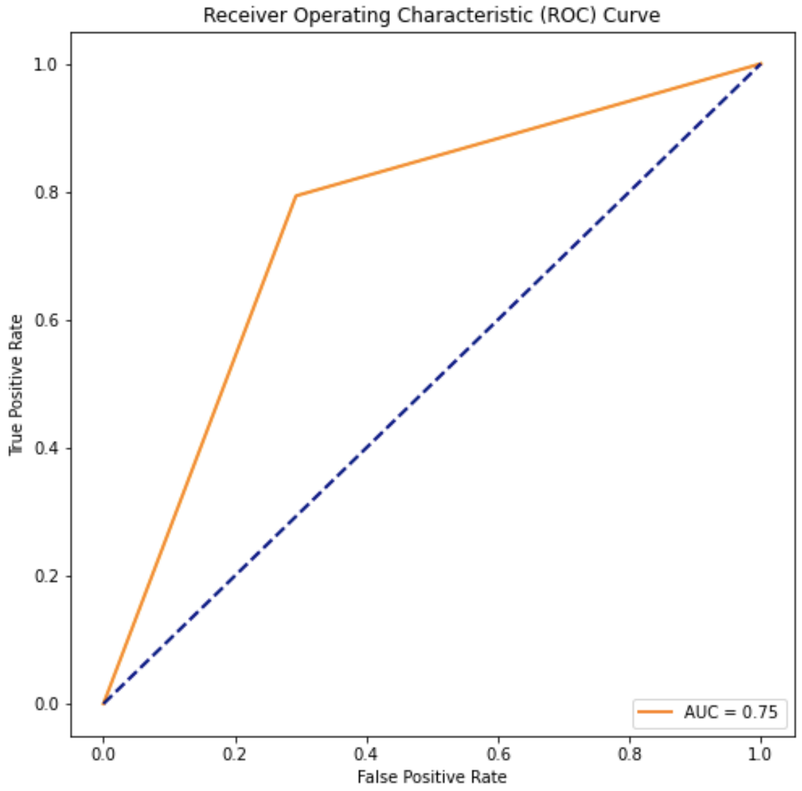
MODEL SELECTION

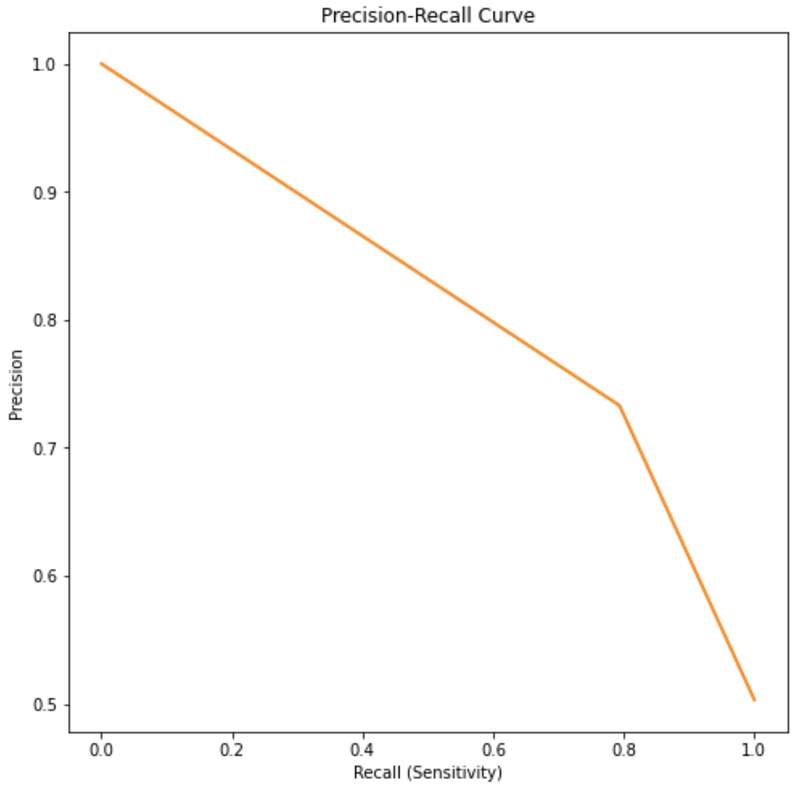
As an insurance company, we base our choice on the indicator **Precision**

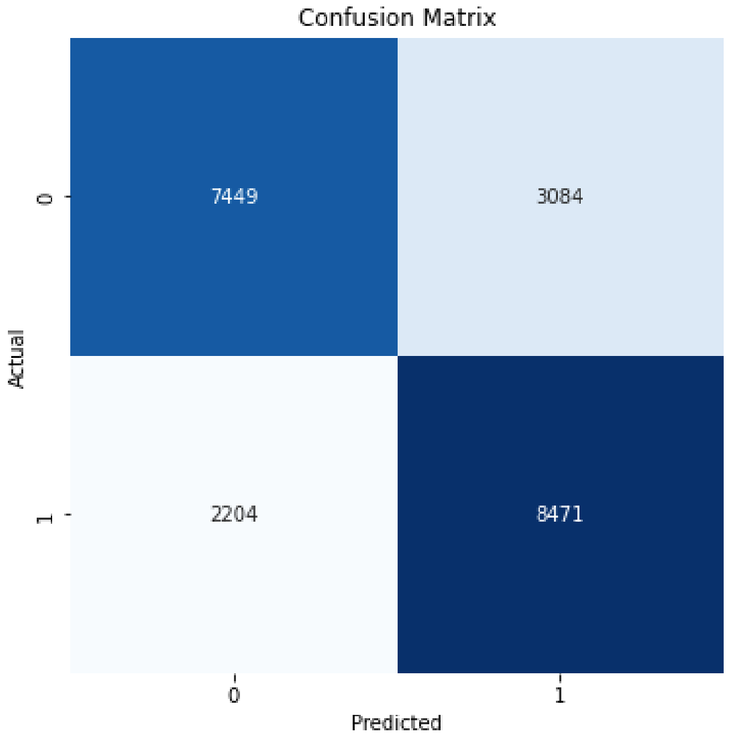
- Random Forest : Precision = 0.72
- Gradient Boosting (XGBoost) : Precision = 0.73
- Support Vector Machines (SVM) : Precision = 0.72

We choose **XGBoost**

EVALUATION

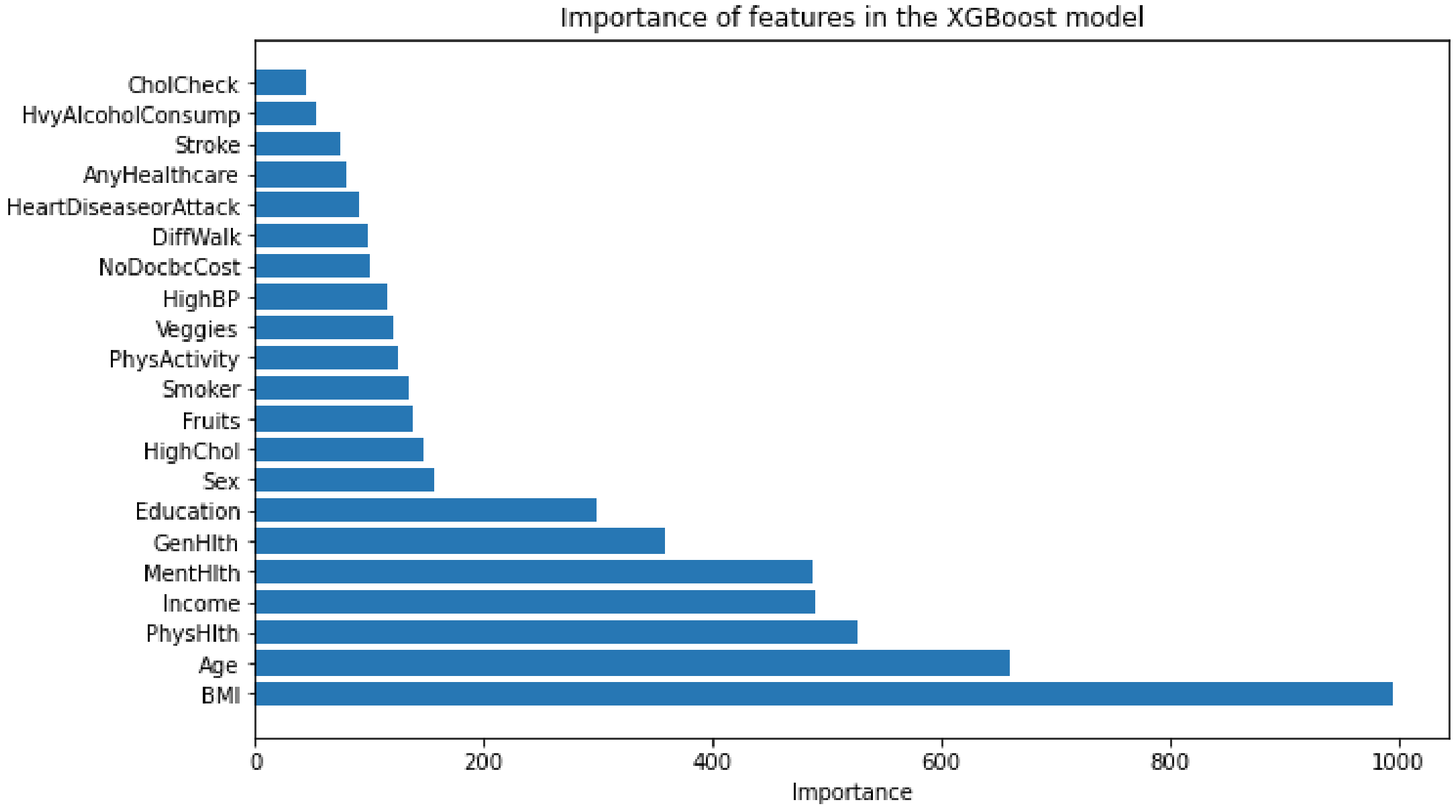




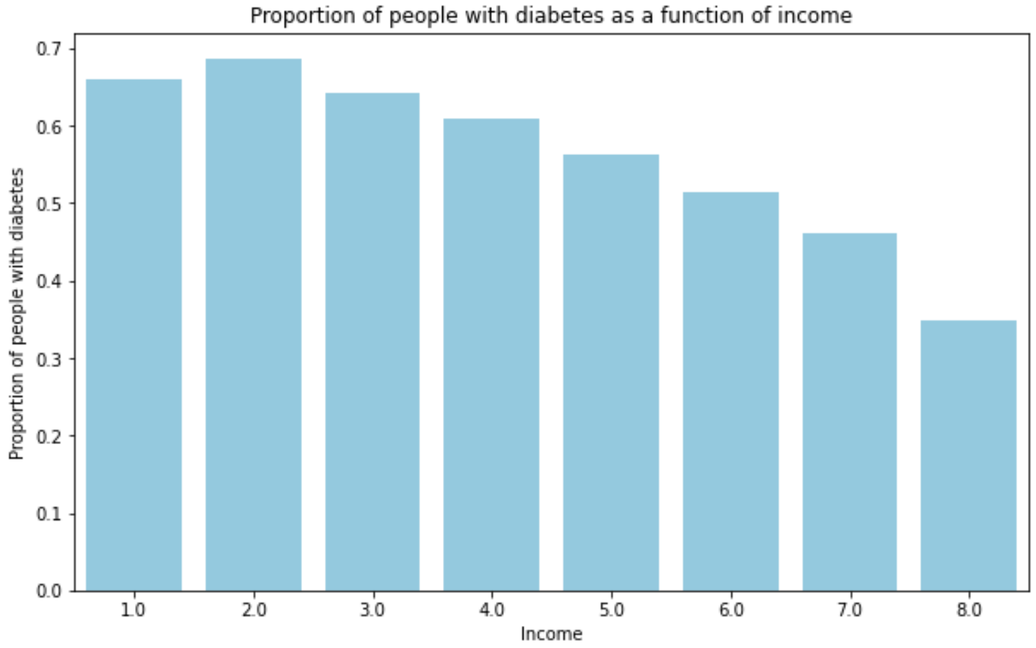


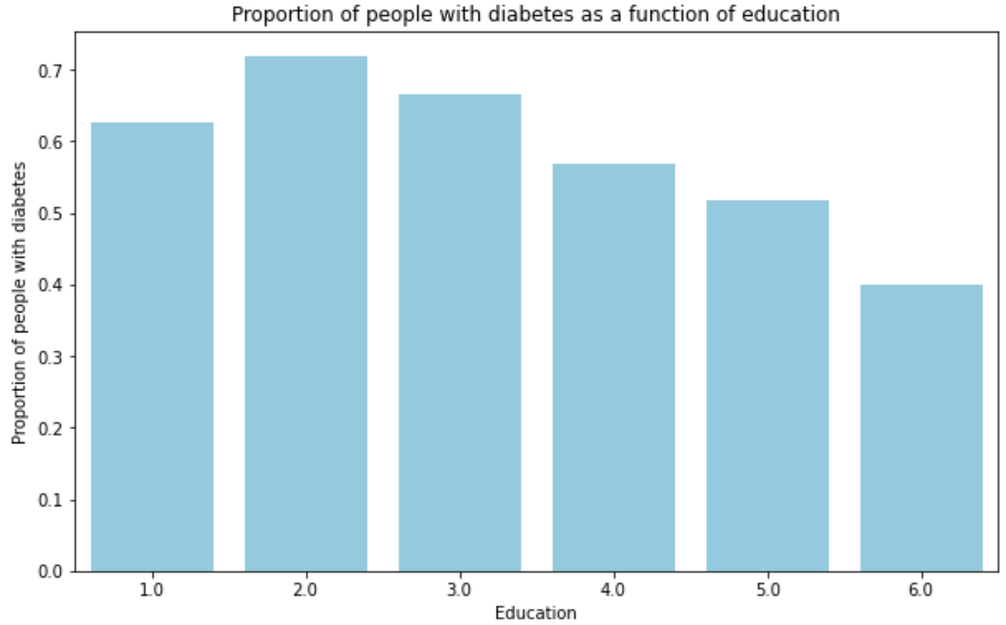
More false positives (3064) than false negatives (2260)

FINAL MODEL DESCRIPTION



FINAL MODEL DESCRIPTION





CONCLUSION

Limitations :

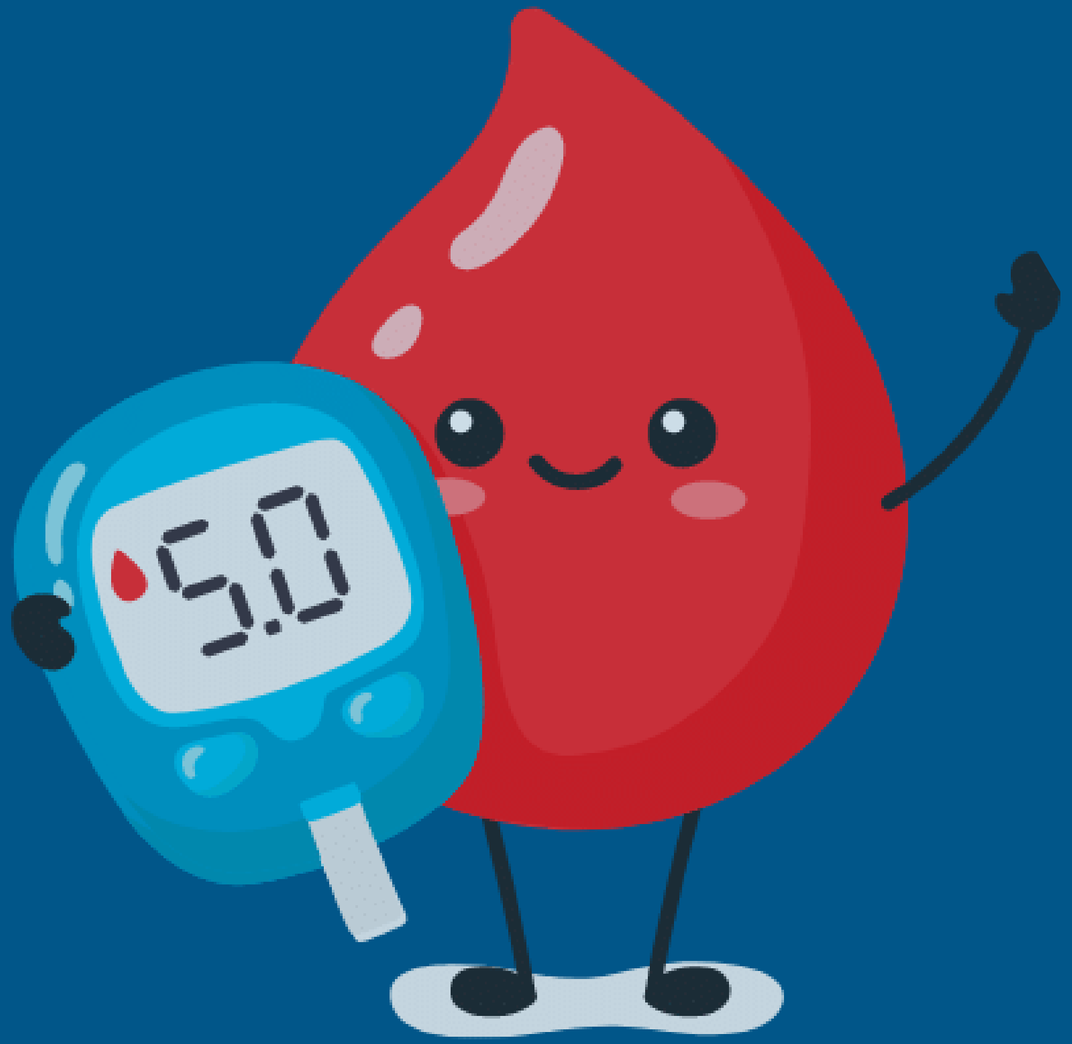
Potential lack of stability (XGBoost results can vary according to the training data)

Possible improvements :

New models based on non-balanced datasets

Review :

XGBoost model / BMI and age as most important features / Precision of 0.73%



Thanks !