ITBA
Instituto Tecnológico
de Buenos Aires

CASE STUDIES
-
ANALYSIS OF A DIABETES HEALTH INDICATORS DATA SET

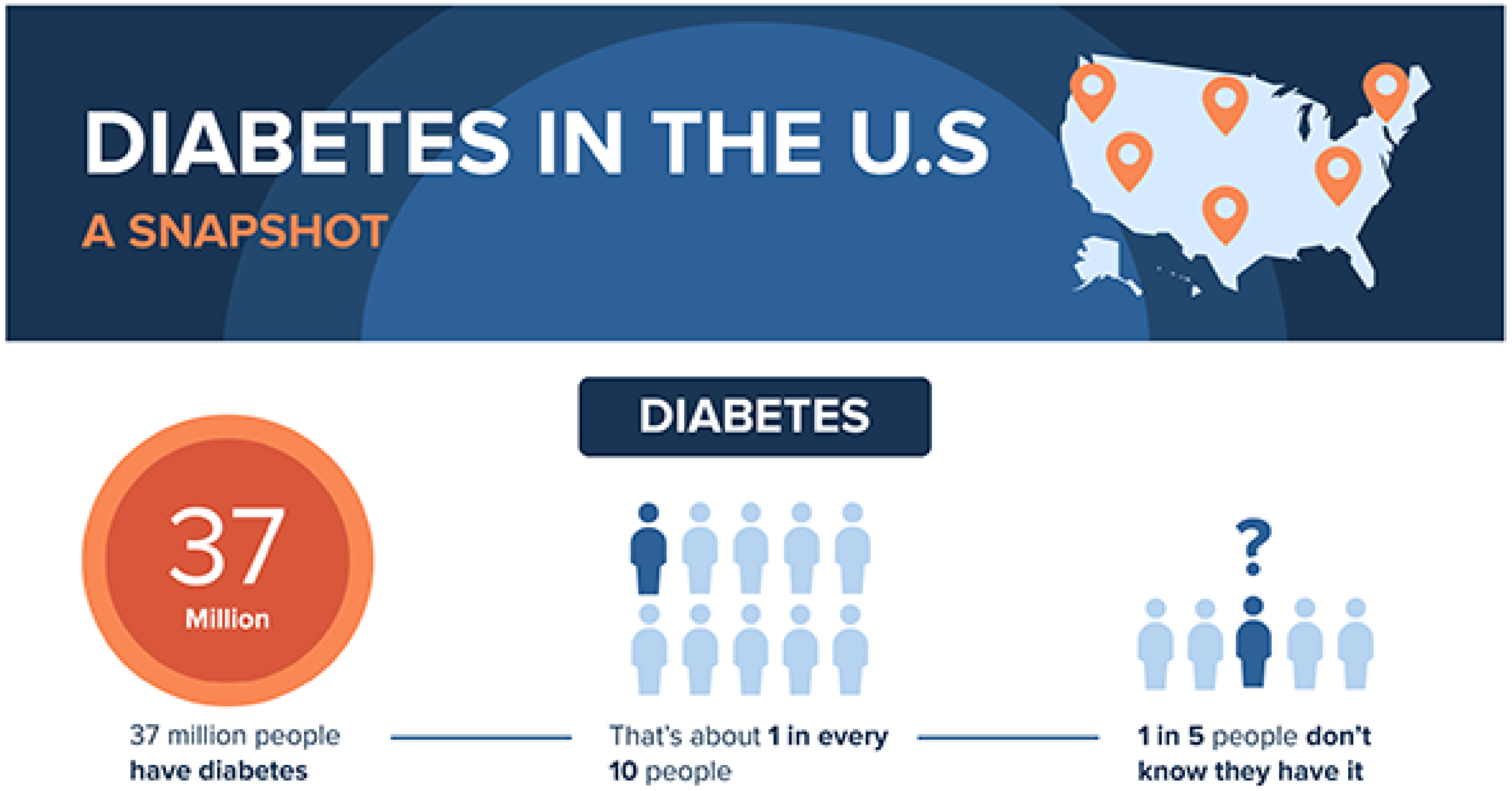
13/12/2023

82.05 - Análisis Predictivo - Final

Fanny LATRON (65998)

ITBA

CASE STUDY



DIABETES IN THE U.S

A SNAPSHOT

DIABETES

37 Million

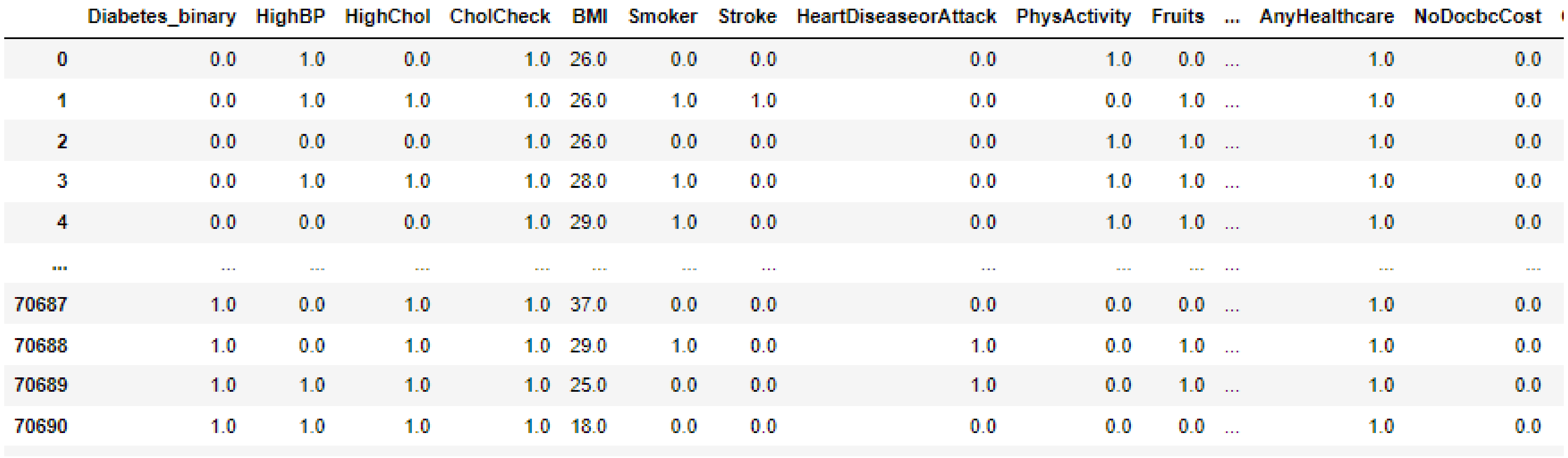
37 million people have diabetes

That's about 1 in every 10 people

1 in 5 people don't know they have it

ITBA

DATASET



	Diabetes_binary	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost
0	0.0	1.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	0.0	...	1.0	0.0
1	0.0	1.0	1.0	1.0	26.0	1.0	1.0	0.0	0.0	1.0	...	1.0	0.0
2	0.0	0.0	0.0	1.0	26.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0
3	0.0	1.0	1.0	1.0	28.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0
4	0.0	0.0	0.0	1.0	29.0	1.0	0.0	0.0	1.0	1.0	...	1.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
70687	1.0	0.0	1.0	1.0	37.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0
70688	1.0	0.0	1.0	1.0	29.0	1.0	0.0	1.0	0.0	1.0	...	1.0	0.0
70689	1.0	1.0	1.0	1.0	25.0	0.0	0.0	1.0	0.0	1.0	...	1.0	0.0
70690	1.0	1.0	1.0	1.0	18.0	0.0	0.0	0.0	0.0	0.0	...	1.0	0.0

21 feature variables and 70692 survey responses

VARIABLES

Variables :

Categorical :

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

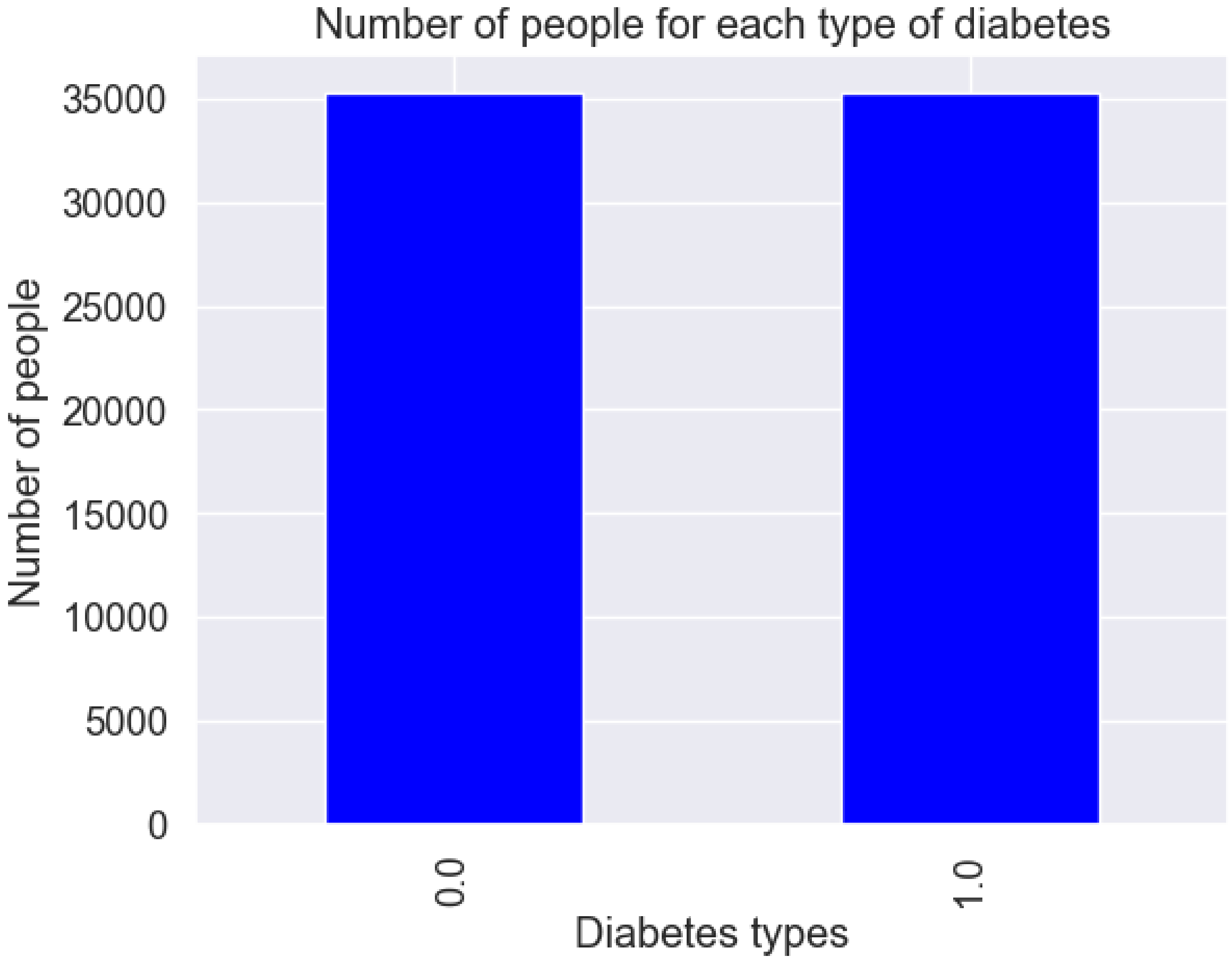
Numerical :

- MentHlth
- PhysHlth
- BMI

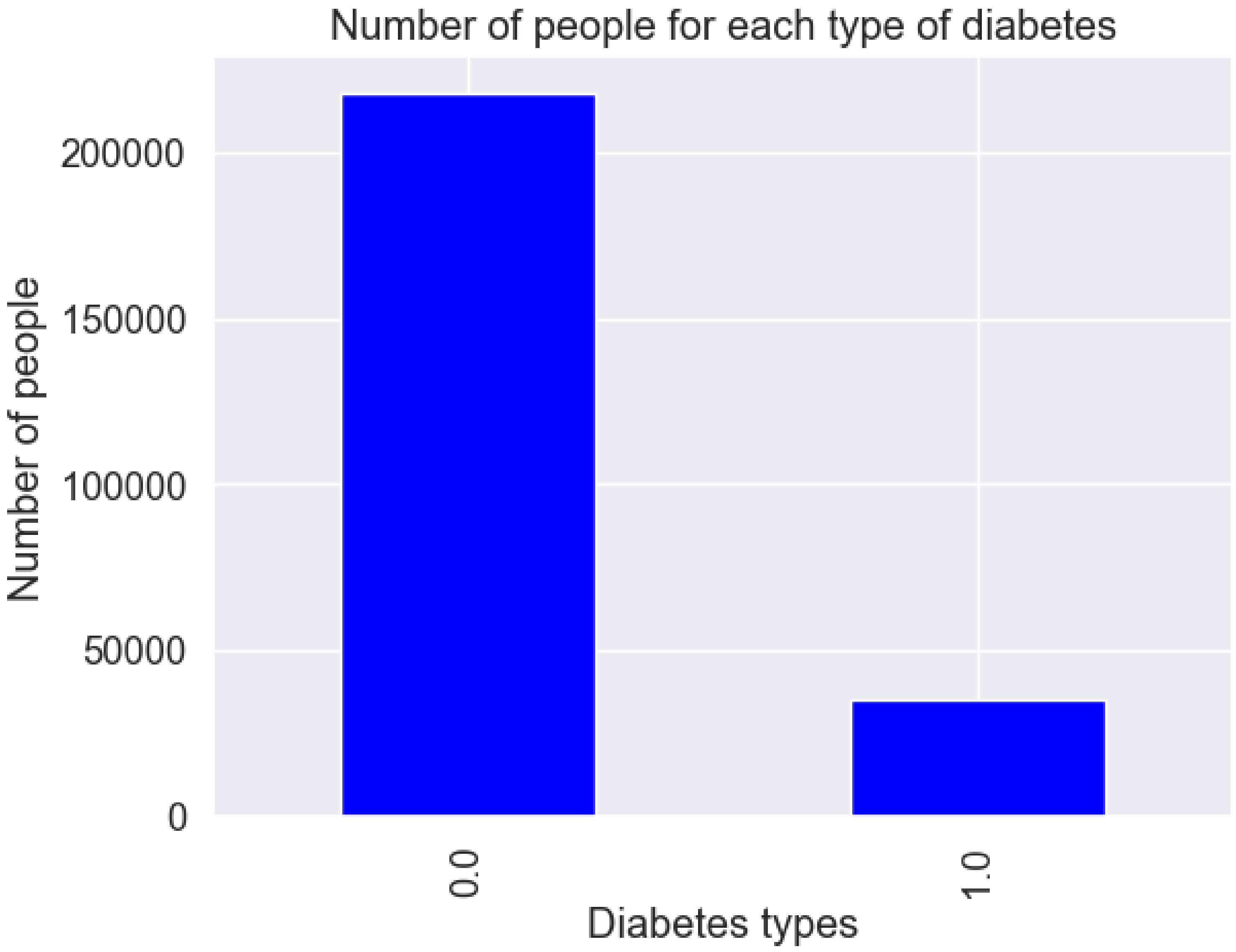
VARIABLES

Response variable

Training dataset



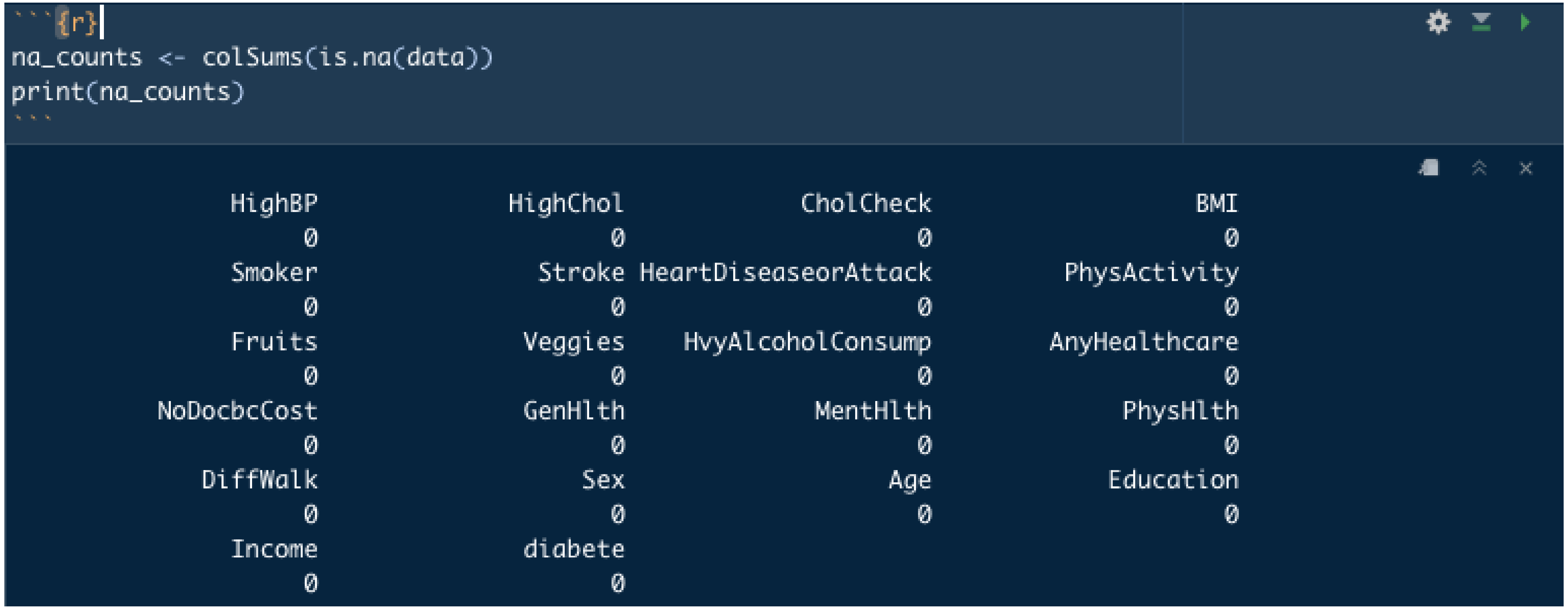
Test dataset



VARIABLES

No missing values / No outliers

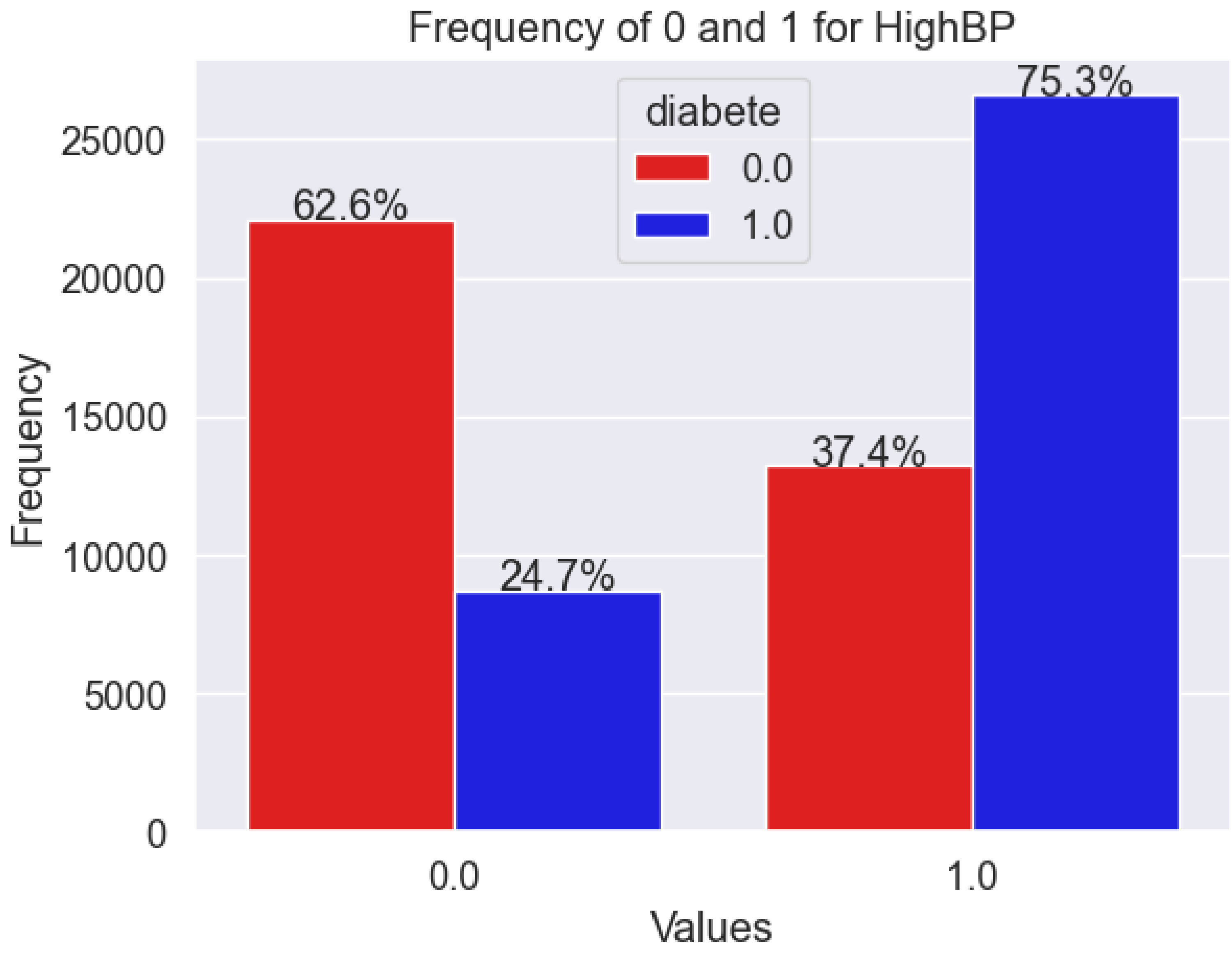
→ Dataset already cleaned

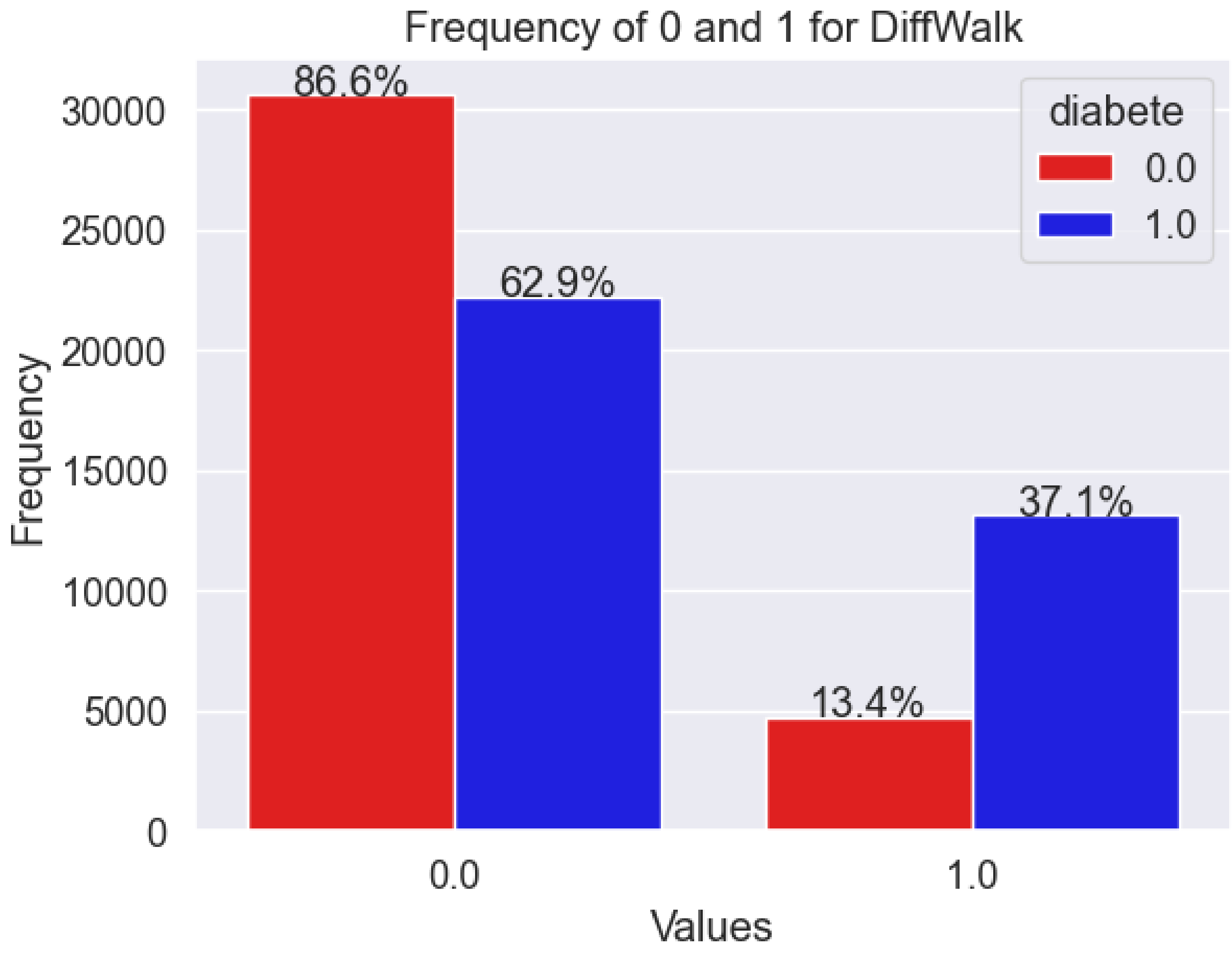


~~~
```{r}
na_counts <- colSums(is.na(data))
print(na_counts)
```
~~~

```
              HighBP             HighChol            CholCheck                  BMI 
                   0                    0                    0                    0 
              Smoker               Stroke HeartDiseaseorAttack         PhysActivity 
                   0                    0                    0                    0 
              Fruits              Veggies    HvyAlcoholConsump        AnyHealthcare 
                   0                    0                    0                    0 
         NoDocbcCost              GenHlth             MentHlth             PhysHlth 
                   0                    0                    0                    0 
            DiffWalk                  Sex                  Age            Education 
                   0                    0                    0                    0 
              Income              diabete 
                   0                    0 
```

EXPLORATION

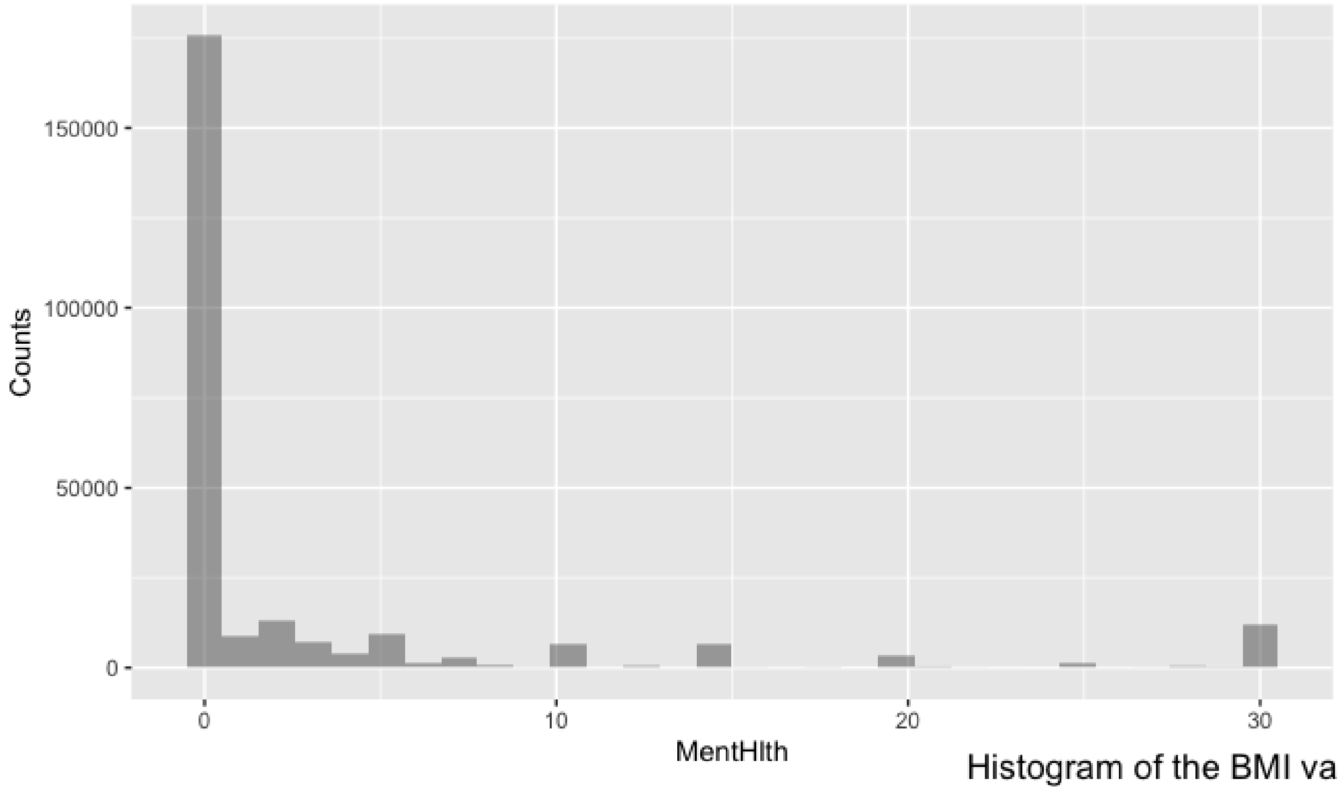




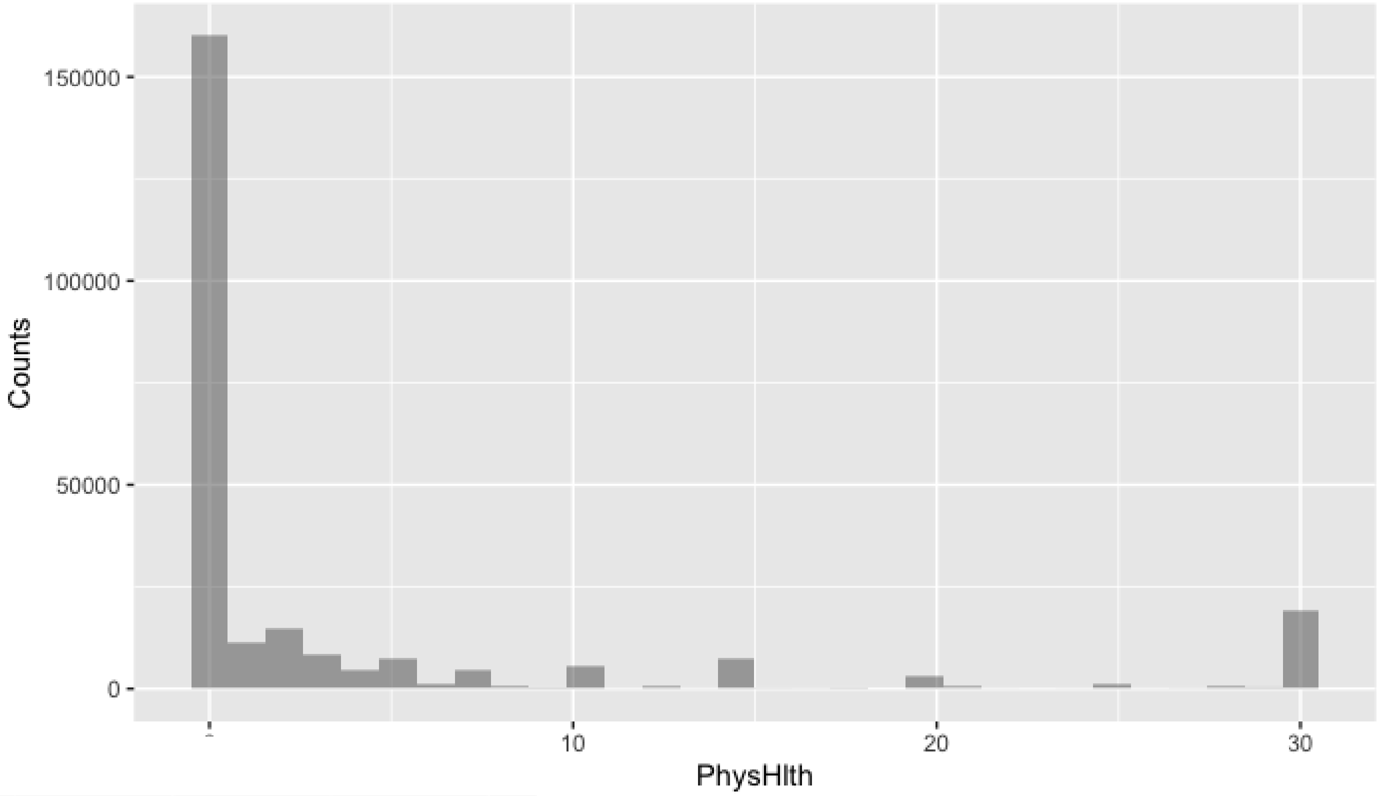
ITBA

EXPLORATION

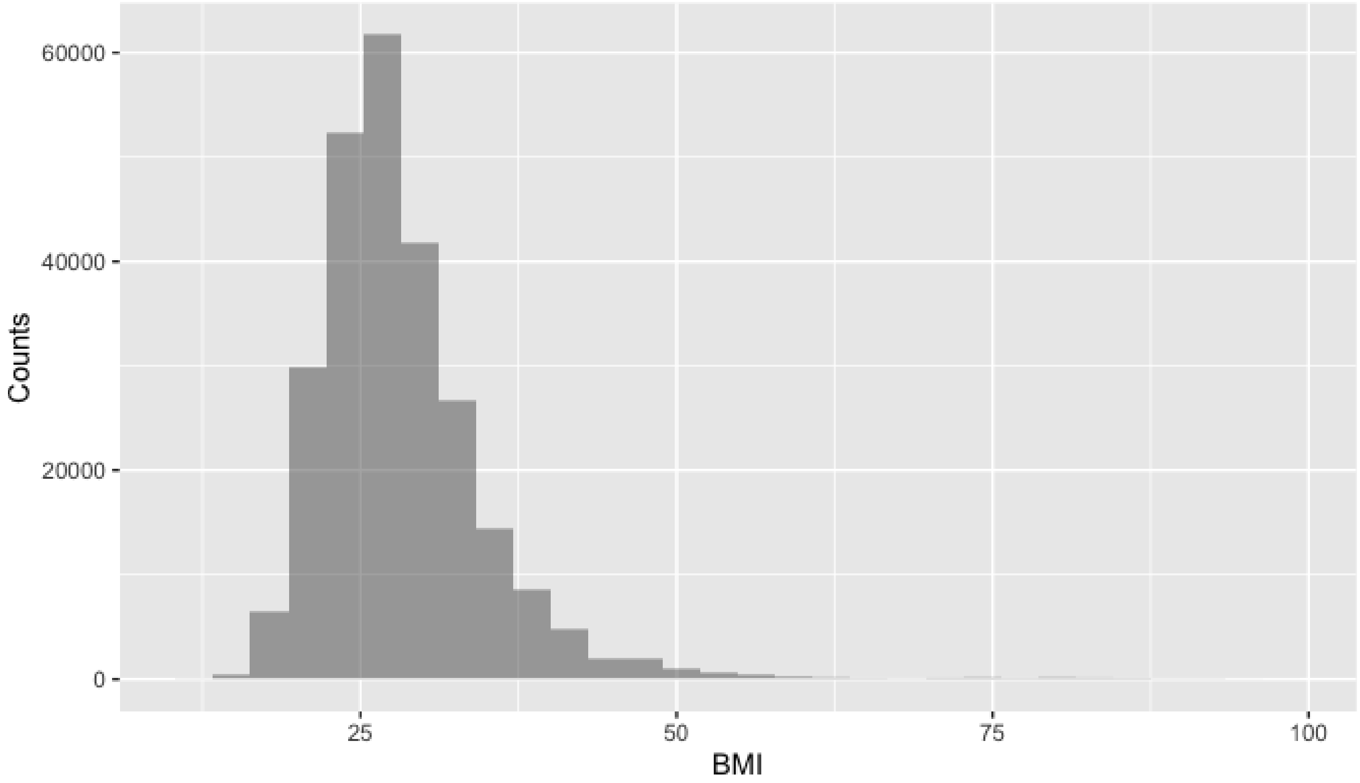
Histogram of the MentHlth variable



Histogram of the PhysHlth variable

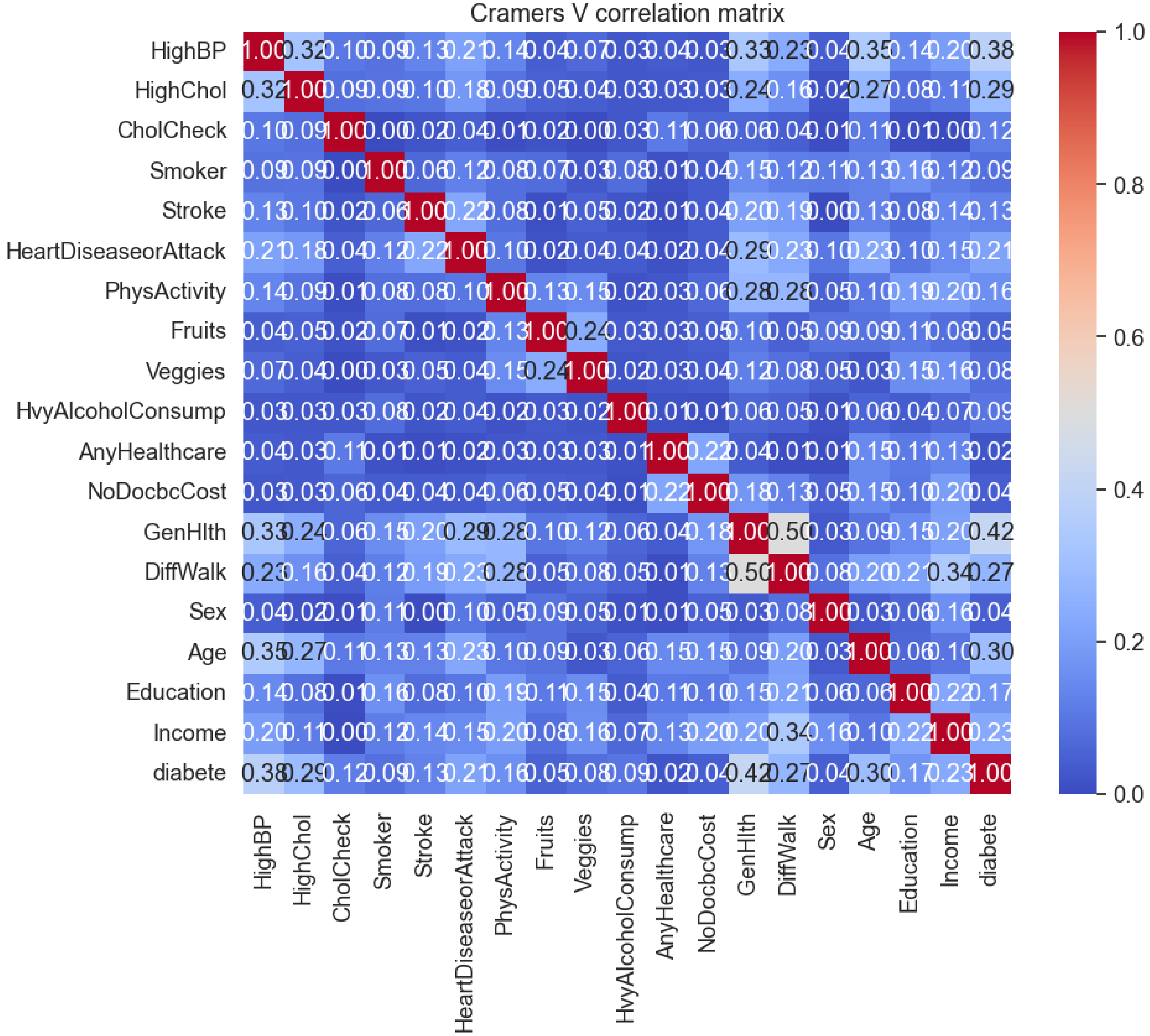


Histogram of the BMI variable



ITBA

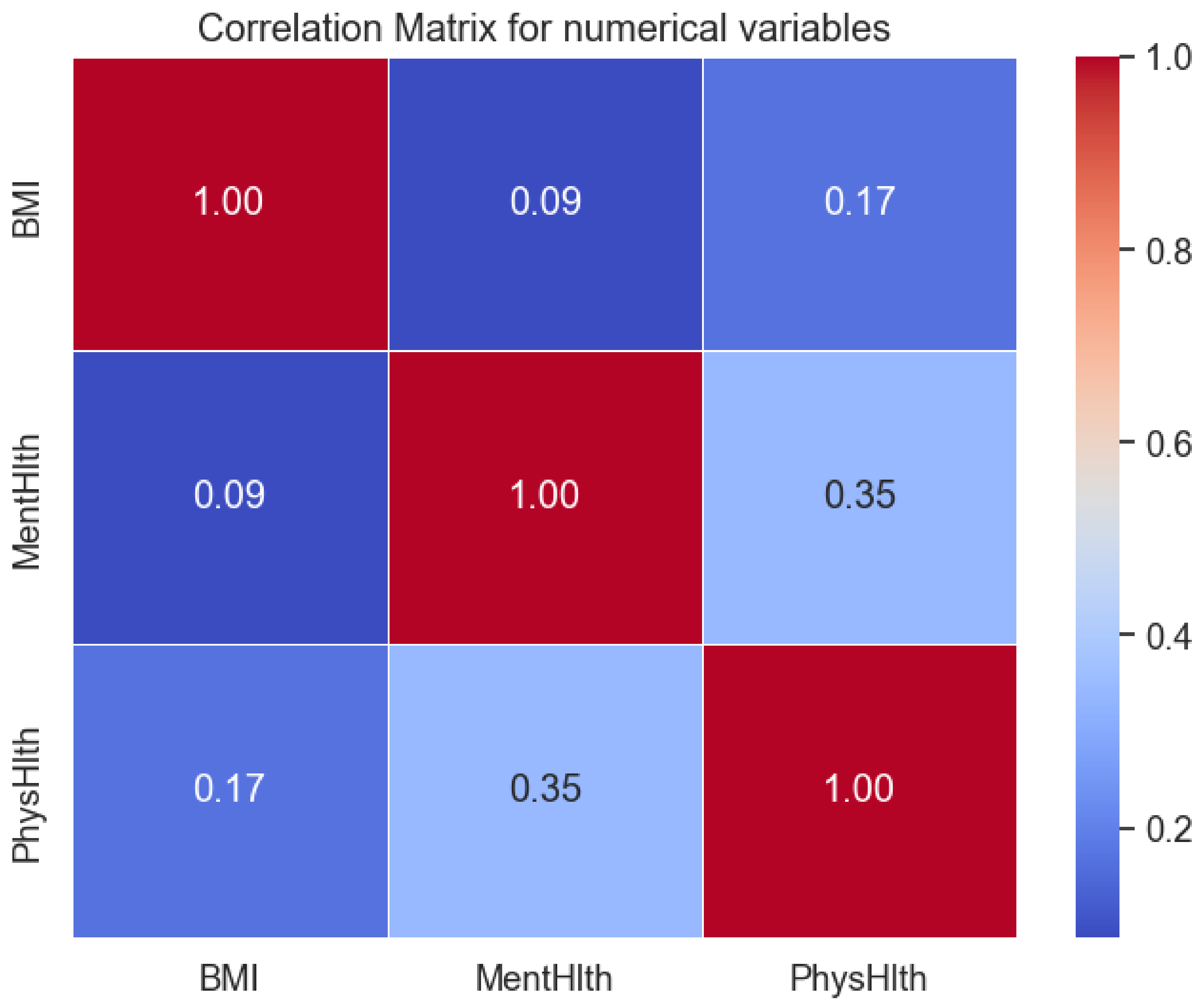
CORRELATIONS



Cramers V correlation matrix

	HighBP	HighChol	CholCheck	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	HvyAlcoholConsump	AnyHealthcare	NoDocbcCost	GenHlth	DiffWalk	Sex	Age	Education	Income	diabete
HighBP	1.00	0.32	0.10	0.09	0.13	0.21	0.14	0.04	0.07	0.03	0.04	0.03	0.33	0.23	0.04	0.35	0.14	0.20	0.38
HighChol	0.32	1.00	0.09	0.09	0.10	0.18	0.09	0.05	0.04	0.03	0.03	0.03	0.24	0.16	0.02	0.27	0.08	0.11	0.29
CholCheck	0.10	0.09	1.00	0.00	0.02	0.04	0.01	0.02	0.00	0.03	0.11	0.06	0.06	0.04	0.01	0.11	0.01	0.00	0.12
Smoker	0.09	0.09	0.00	1.00	0.06	0.12	0.08	0.07	0.03	0.08	0.01	0.04	0.15	0.12	0.11	0.13	0.16	0.12	0.09
Stroke	0.13	0.10	0.02	0.06	1.00	0.22	0.08	0.01	0.05	0.02	0.01	0.04	0.20	0.19	0.00	0.13	0.08	0.14	0.13
HeartDiseaseorAttack	0.21	0.18	0.04	0.12	0.22	1.00	0.10	0.02	0.04	0.04	0.02	0.04	0.29	0.23	0.10	0.23	0.10	0.15	0.21
PhysActivity	0.14	0.09	0.01	0.08	0.08	0.10	1.00	0.13	0.15	0.02	0.03	0.06	0.28	0.28	0.05	0.10	0.19	0.20	0.16
Fruits	0.04	0.05	0.02	0.07	0.01	0.02	0.13	1.00	0.24	0.03	0.03	0.05	0.10	0.05	0.09	0.09	0.11	0.08	0.05
Veggies	0.07	0.04	0.00	0.03	0.05	0.04	0.15	0.24	1.00	0.02	0.03	0.04	0.12	0.08	0.05	0.03	0.15	0.16	0.08
HvyAlcoholConsump	0.03	0.03	0.03	0.08	0.02	0.04	0.02	0.03	0.02	1.00	0.01	0.01	0.06	0.05	0.01	0.06	0.04	0.07	0.09
AnyHealthcare	0.04	0.03	0.11	0.01	0.01	0.02	0.03	0.03	0.03	0.01	1.00	0.22	0.04	0.01	0.01	0.15	0.11	0.13	0.02
NoDocbcCost	0.03	0.03	0.06	0.04	0.04	0.04	0.06	0.05	0.04	0.01	0.22	1.00	0.18	0.13	0.05	0.15	0.10	0.20	0.04
GenHlth	0.33	0.24	0.06	0.15	0.20	0.29	0.28	0.10	0.12	0.06	0.04	0.18	1.00	0.50	0.03	0.09	0.15	0.20	0.42
DiffWalk	0.23	0.16	0.04	0.12	0.19	0.23	0.28	0.05	0.08	0.05	0.01	0.13	0.50	1.00	0.08	0.20	0.21	0.34	0.27
Sex	0.04	0.02	0.01	0.11	0.00	0.10	0.05	0.09	0.05	0.01	0.01	0.05	0.03	0.08	1.00	0.03	0.06	0.16	0.04
Age	0.35	0.27	0.11	0.13	0.13	0.23	0.10	0.09	0.03	0.06	0.15	0.15	0.09	0.20	0.03	1.00	0.06	0.10	0.30
Education	0.14	0.08	0.01	0.16	0.08	0.10	0.19	0.11	0.15	0.04	0.11	0.10	0.15	0.21	0.06	0.06	1.00	0.22	0.17
Income	0.20	0.11	0.00	0.12	0.14	0.15	0.20	0.08	0.16	0.07	0.13	0.20	0.20	0.34	0.16	0.10	0.22	1.00	0.23
diabete	0.38	0.29	0.12	0.09	0.13	0.21	0.16	0.05	0.08	0.09	0.02	0.04	0.42	0.27	0.04	0.30	0.17	0.23	1.00

CORRELATIONS



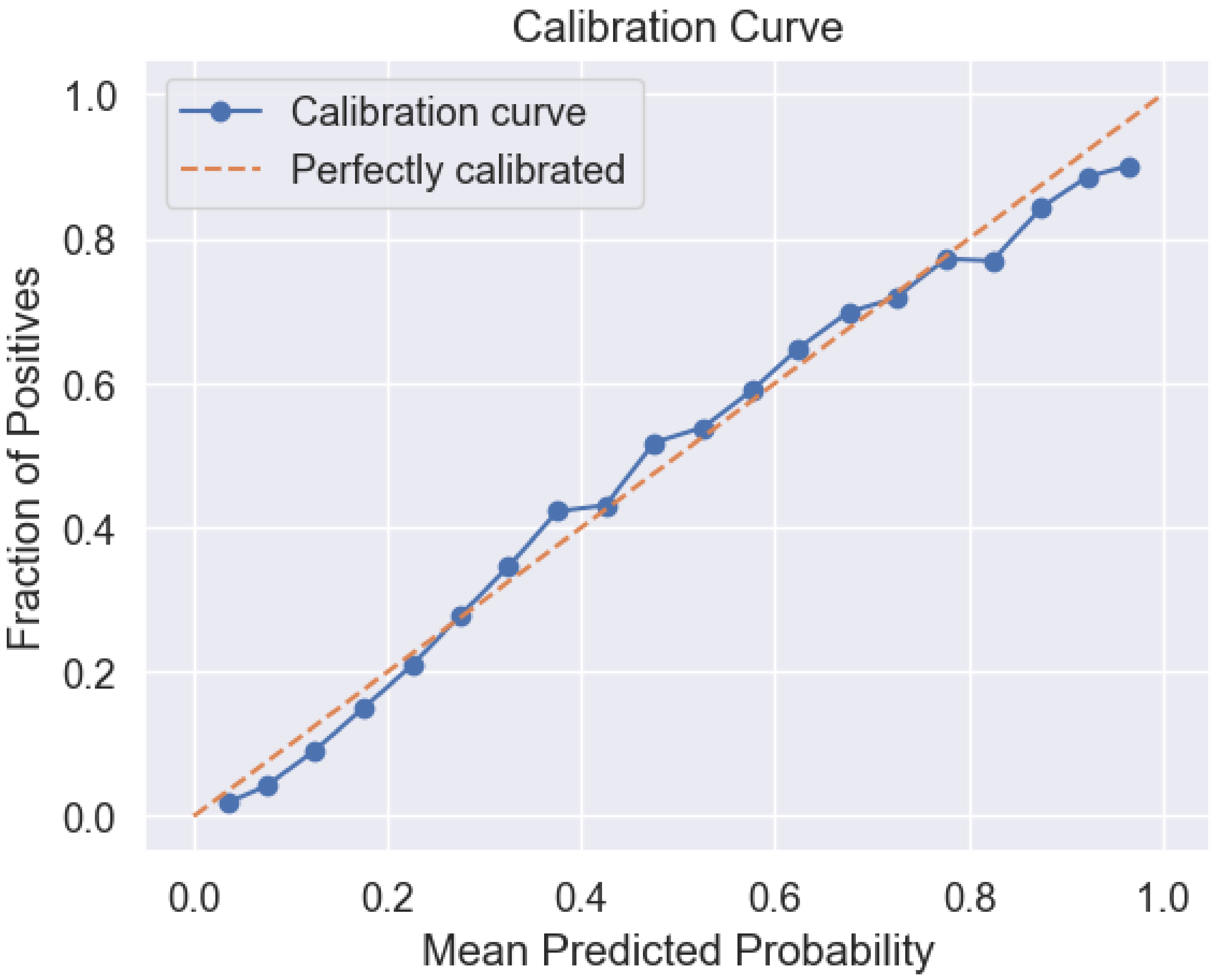
Correlation Matrix for numerical variables

	BMI	MentHlth	PhysHlth
BMI	1.00	0.09	0.17
MentHlth	0.09	1.00	0.35
PhysHlth	0.17	0.35	1.00

MODELS

- **First Model : Logistic regression**

  → ROC-AUC Score: 0.8153
  → Brier Score: 0.1827

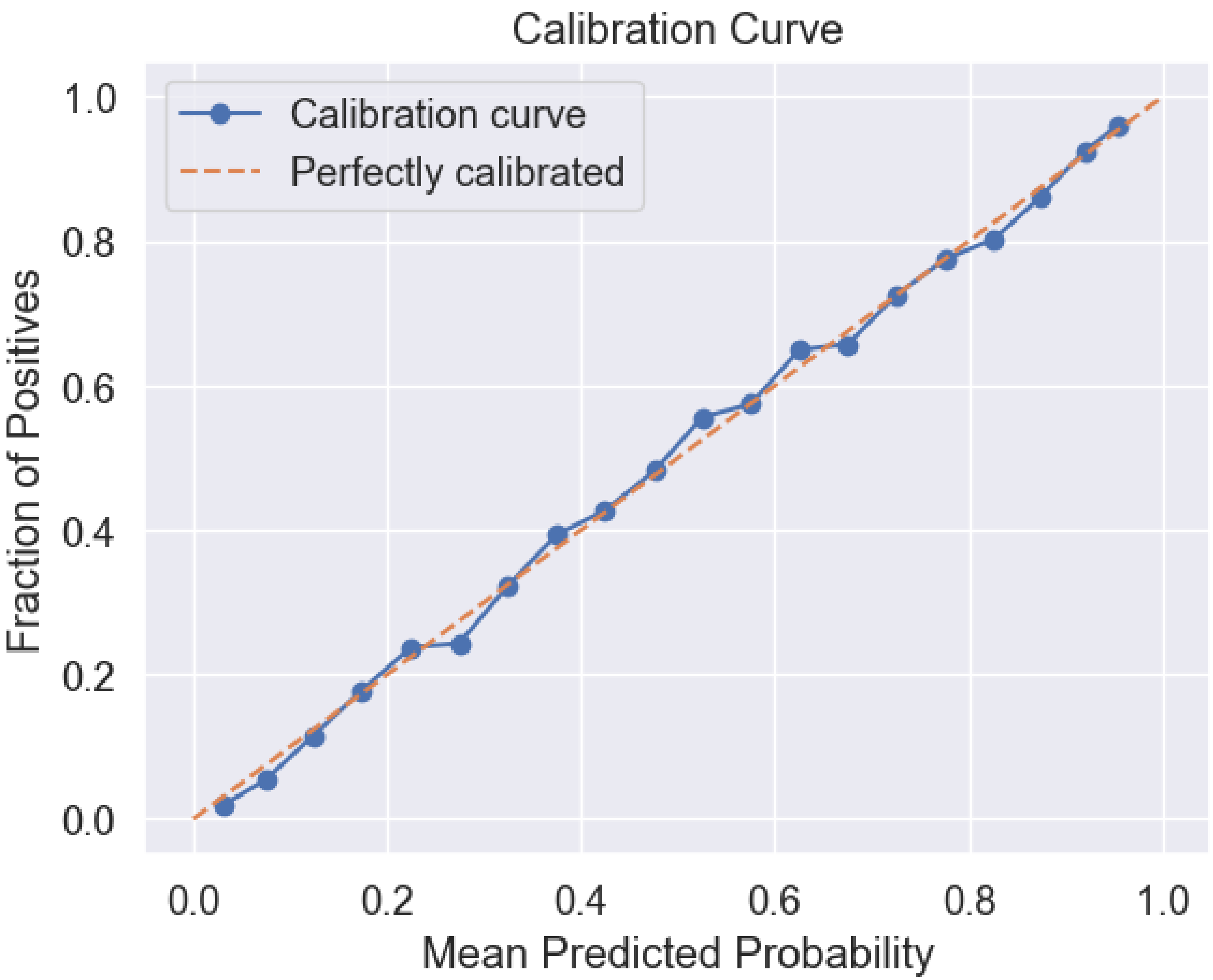


MODELS

- **Second Model : XGBoost**

  → ROC-AUC Score: 0.8322
  → Brier Score: 0.1760
  → Parameters :
  'learning_rate': 0.15,
  'max_depth': 2,
  'n_estimators': 200

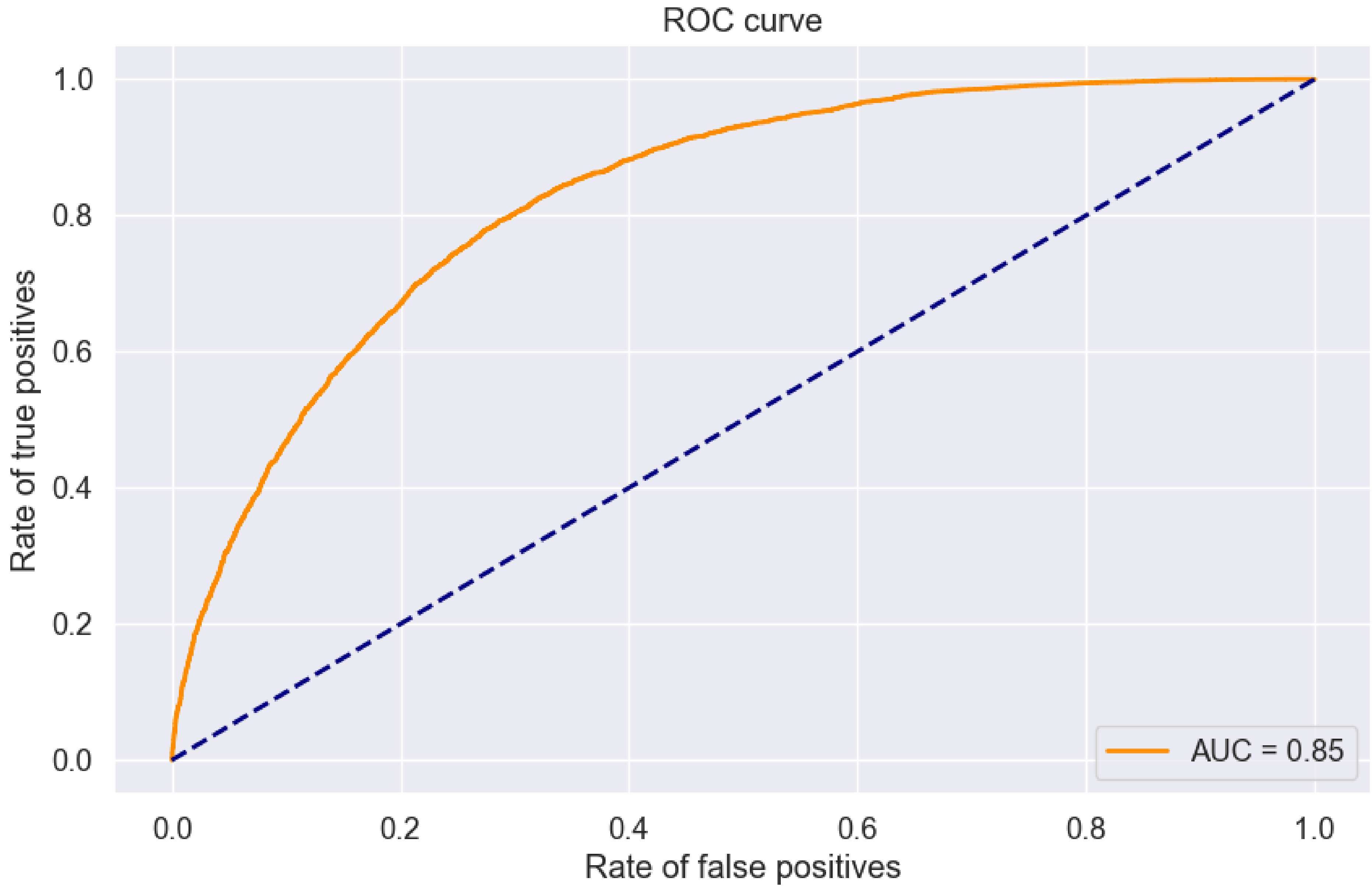
- **Other Model : SVM**



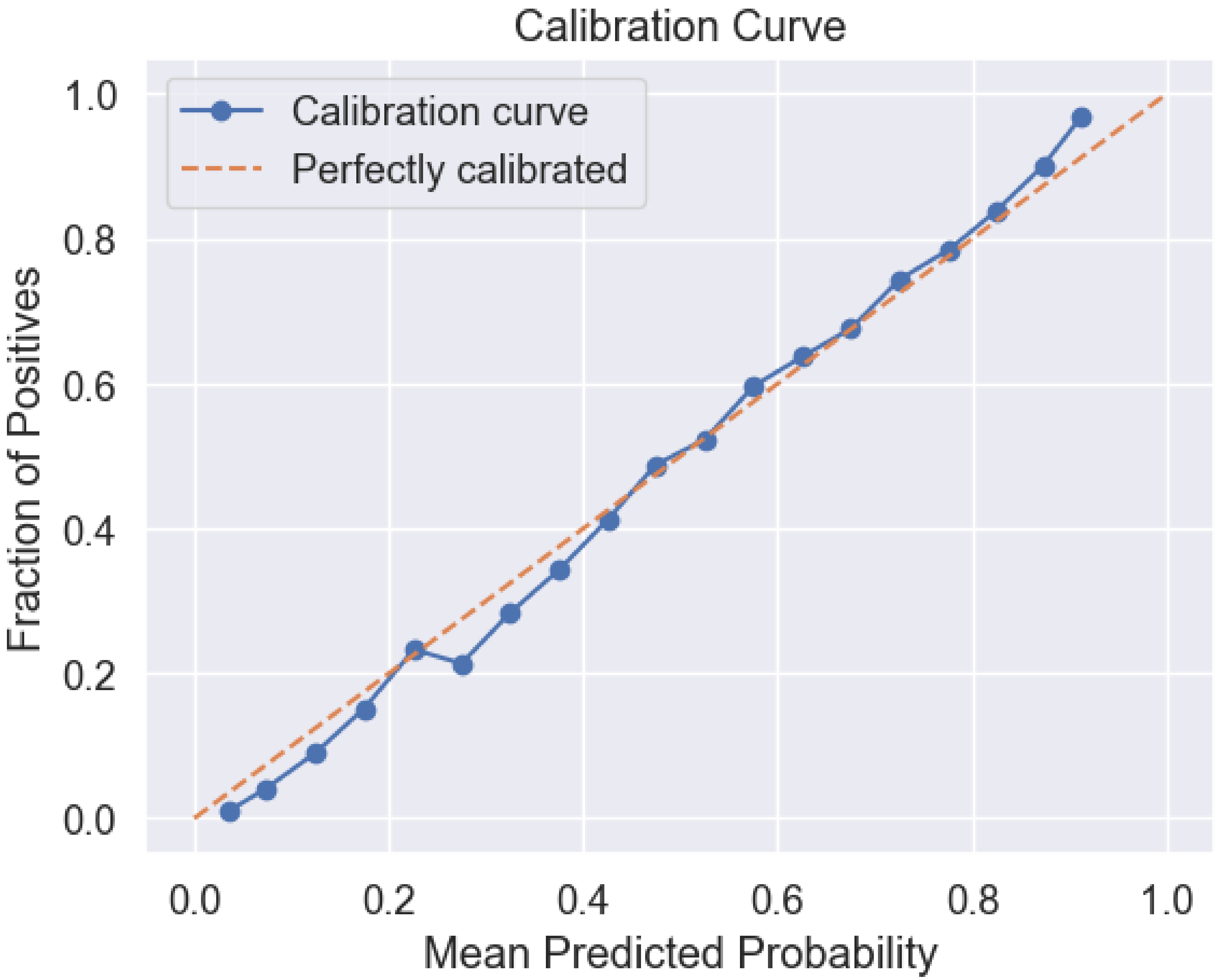
FINAL MODEL

- **Final Model : Random Forest**

→ ROC-AUC Score: 0.8461
→ Brier Score: 0.1739
→ Parameters :
'max_depth': 12,
'min_samples_leaf': 3,
'min_samples_split': 3,
'n_estimators': 210



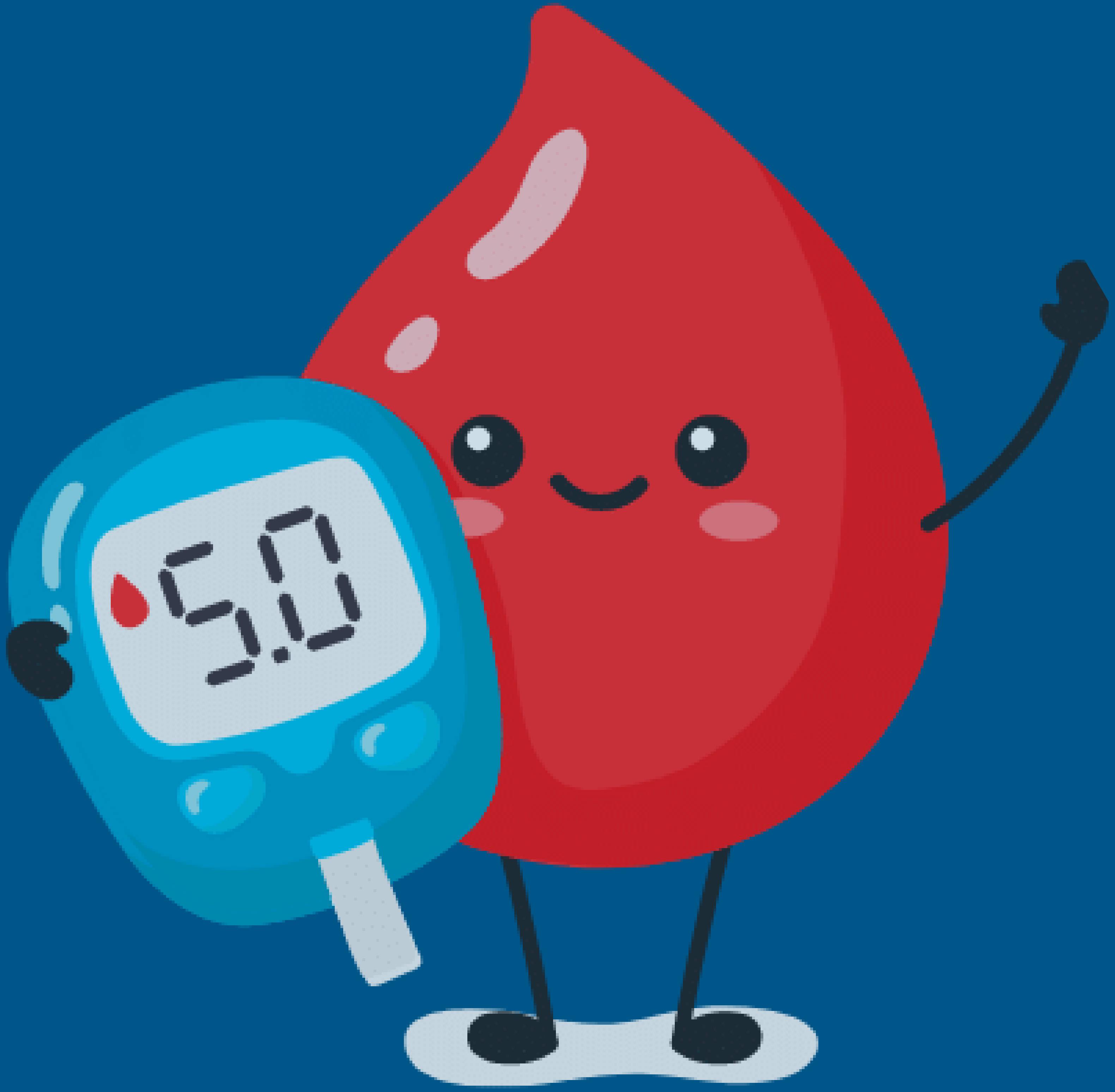
FINAL MODEL


Calibration Curve
Calibration curve
Perfectly calibrated
Fraction of Positives
Mean Predicted Probability
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

ITBA

CONCLUSION

- Could possibly find a better model but still efficient
- Low execution time
- Gives an indication close enough to the truth



Thanks !