FCMB GROUP 7

SC1015 MINI PROJECT: DIABETES

Chan Yu Yan U2322215K
Daryl Poh Wei Xuan U2322272C
Eldon Lim Kai Jie U2320323F

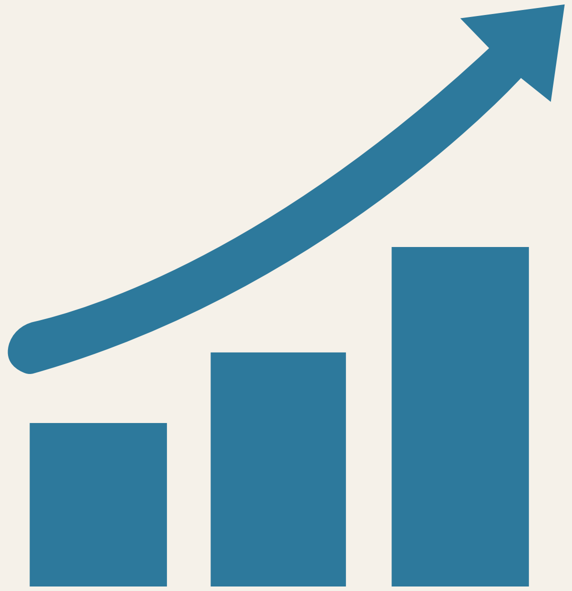
TABLE OF CONTENTS

- Problem Statement
- Exploratory Data Analysis and Data Preparation
- Analytic Visualisation
- Machine Learning
- Neural Network Model
- Conclusion and Insights

PROBLEM STATEMENT

PRACTICAL MOTIVATION

Approx. 537mil adults (aged 20-79) worldwide are diabetic, as of 2021



Expected to rise to 643mil by 2030

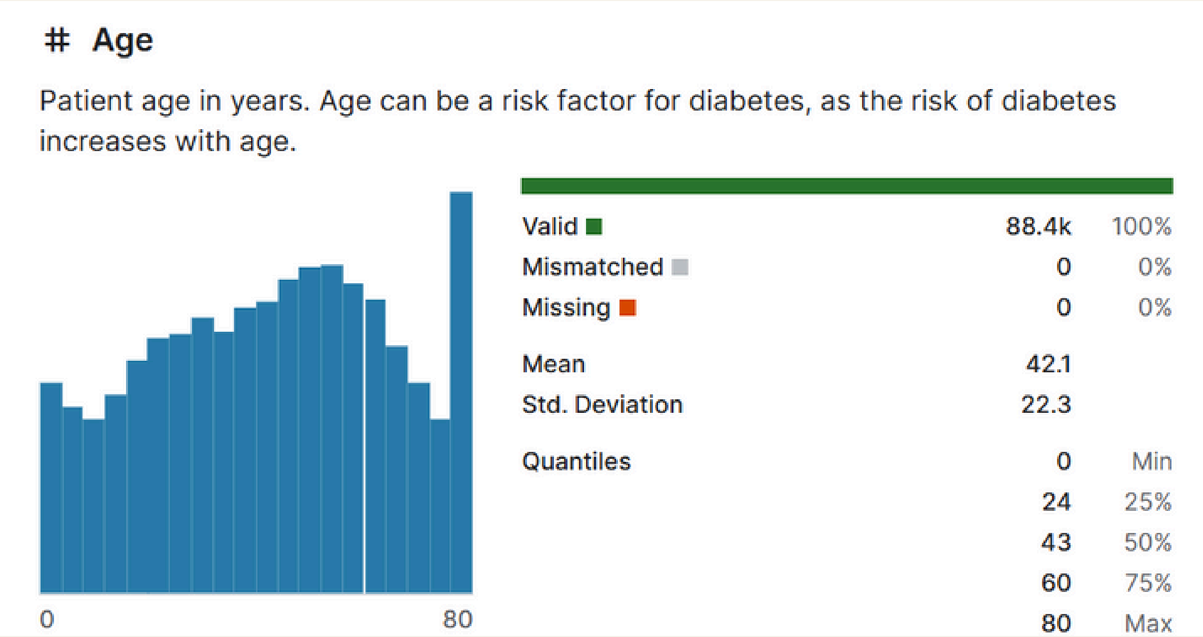
Long term implications for individuals and healthcare systems

Importance of early intervention

PROBLEM FORMULATION

Which variable or combination of variables is most effective in predicting diabetes diagnosis accurately?

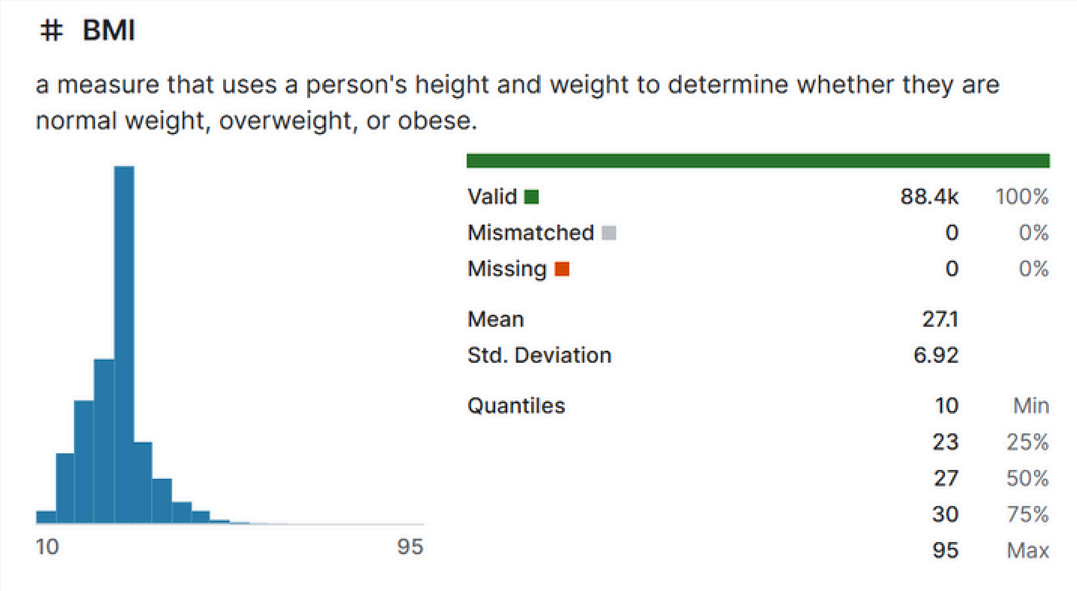
SAMPLE COLLECTION



Age

Patient age in years. Age can be a risk factor for diabetes, as the risk of diabetes increases with age.

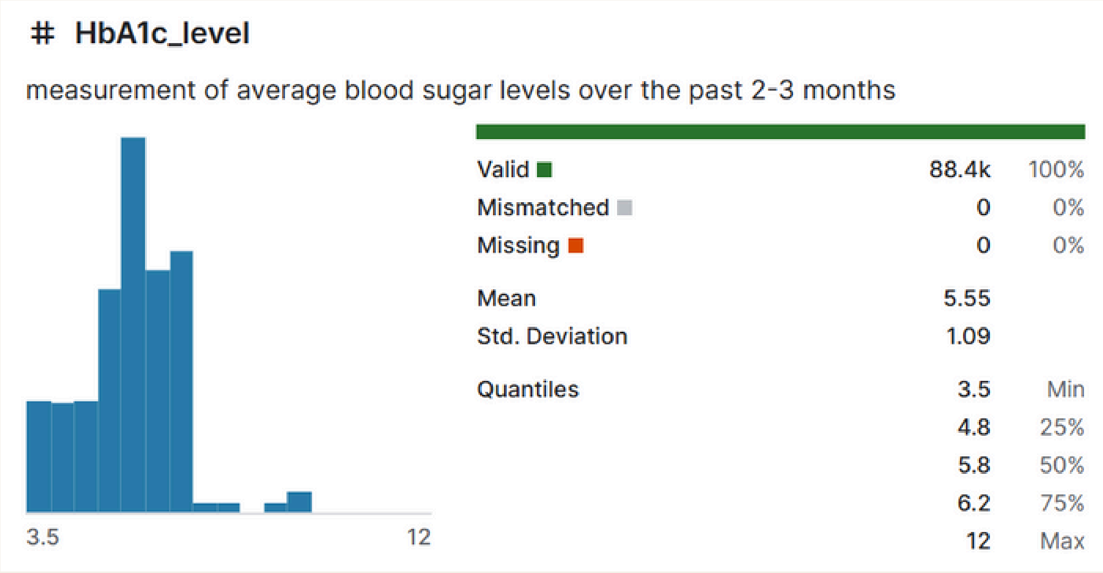
Valid 🟩	88.4k	100%
Mismatched ⬜	0	0%
Missing 🟥	0	0%
Mean	42.1	
Std. Deviation	22.3	
Quantiles	0	Min
	24	25%
	43	50%
	60	75%
	80	Max



BMI

a measure that uses a person's height and weight to determine whether they are normal weight, overweight, or obese.

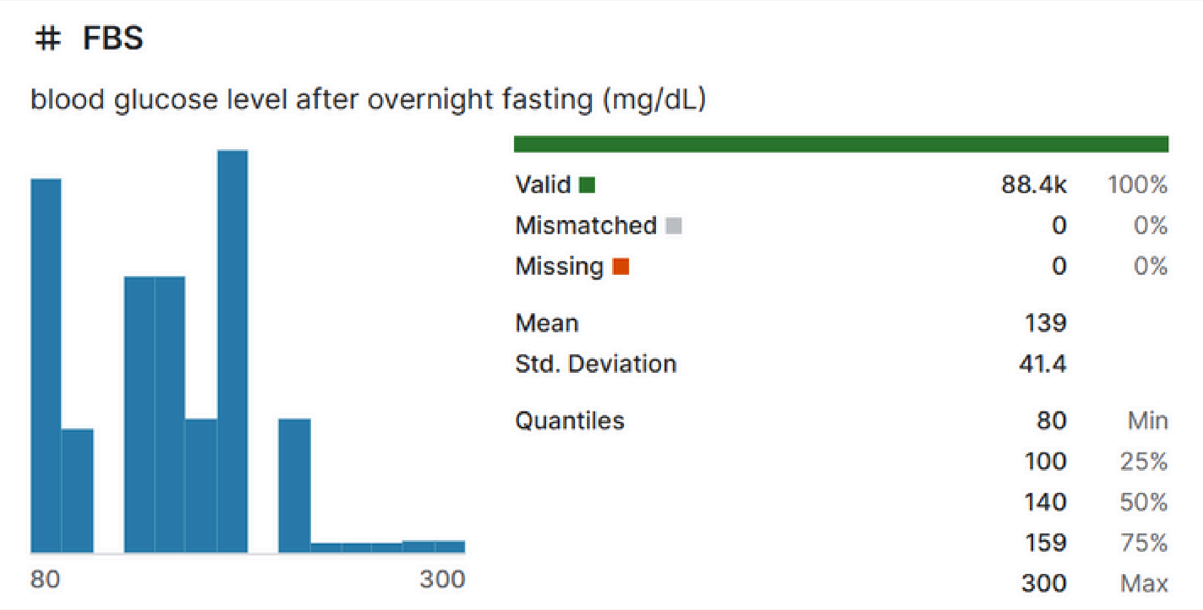
Valid 🟩	88.4k	100%
Mismatched ⬜	0	0%
Missing 🟥	0	0%
Mean	27.1	
Std. Deviation	6.92	
Quantiles	10	Min
	23	25%
	27	50%
	30	75%
	95	Max



HbA1c_level

measurement of average blood sugar levels over the past 2-3 months

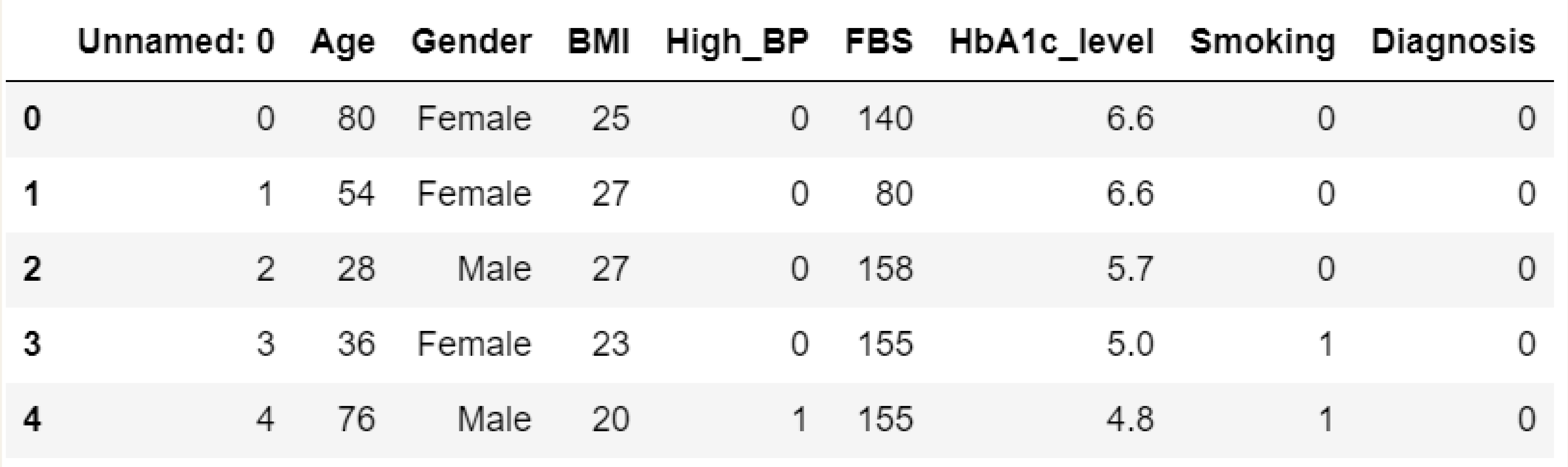
Valid 🟩	88.4k	100%
Mismatched ⬜	0	0%
Missing 🟥	0	0%
Mean	5.55	
Std. Deviation	1.09	
Quantiles	3.5	Min
	4.8	25%
	5.8	50%
	6.2	75%
	12	Max



FBS

blood glucose level after overnight fasting (mg/dL)

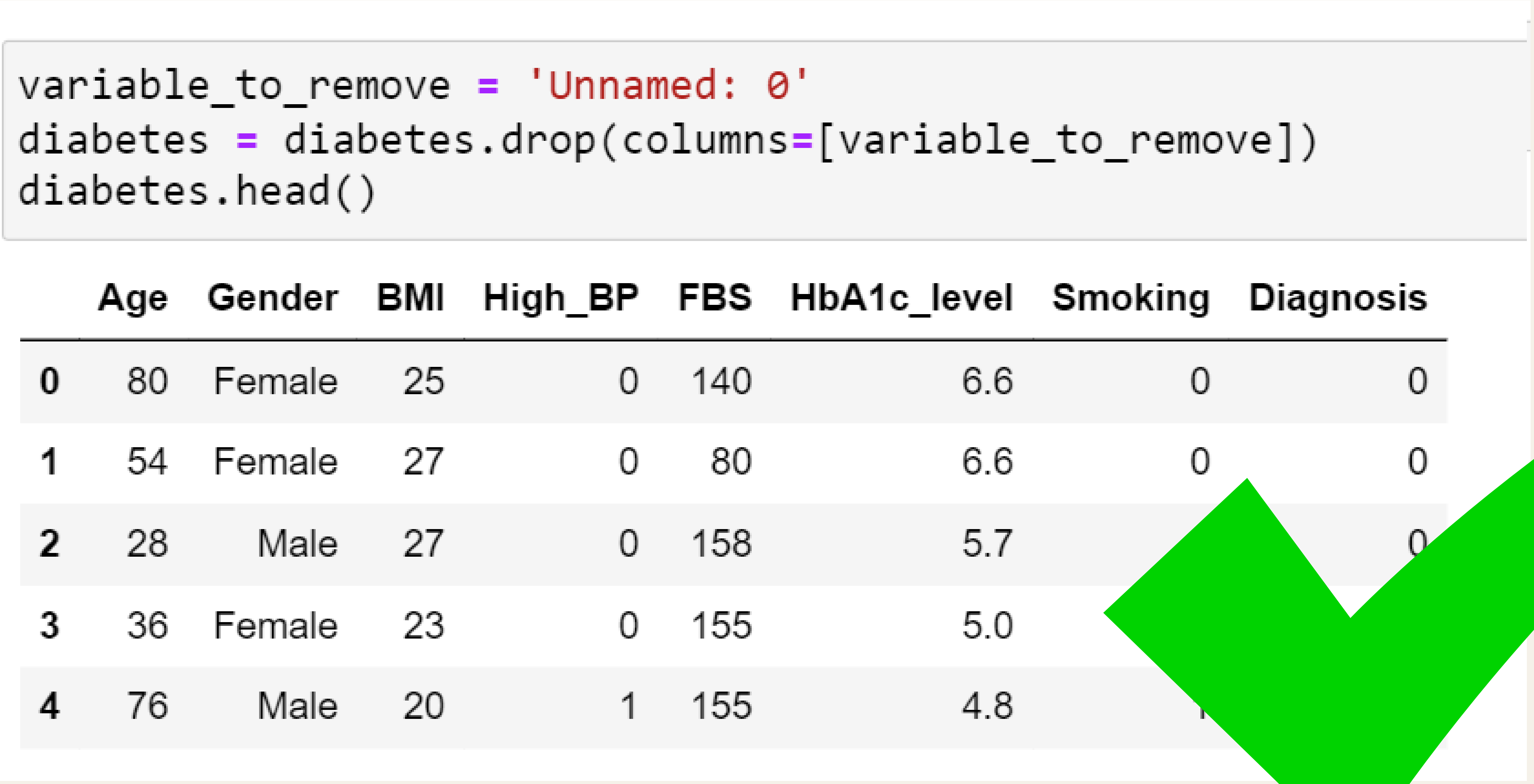
Valid 🟩	88.4k	100%
Mismatched ⬜	0	0%
Missing 🟥	0	0%
Mean	139	
Std. Deviation	41.4	
Quantiles	80	Min
	100	25%
	140	50%
	159	75%
	300	Max



	Unnamed: 0	Age	Gender	BMI	High_BP	FBS	HbA1c_level	Smoking	Diagnosis
0	0	80	Female	25	0	140	6.6	0	0
1	1	54	Female	27	0	80	6.6	0	0
2	2	28	Male	27	0	158	5.7	0	0
3	3	36	Female	23	0	155	5.0	1	0
4	4	76	Male	20	1	155	4.8	1	0

EXPLORATORY DATA ANALYSIS

PRELIMINARY EXPLORATION & DATA CLEANING

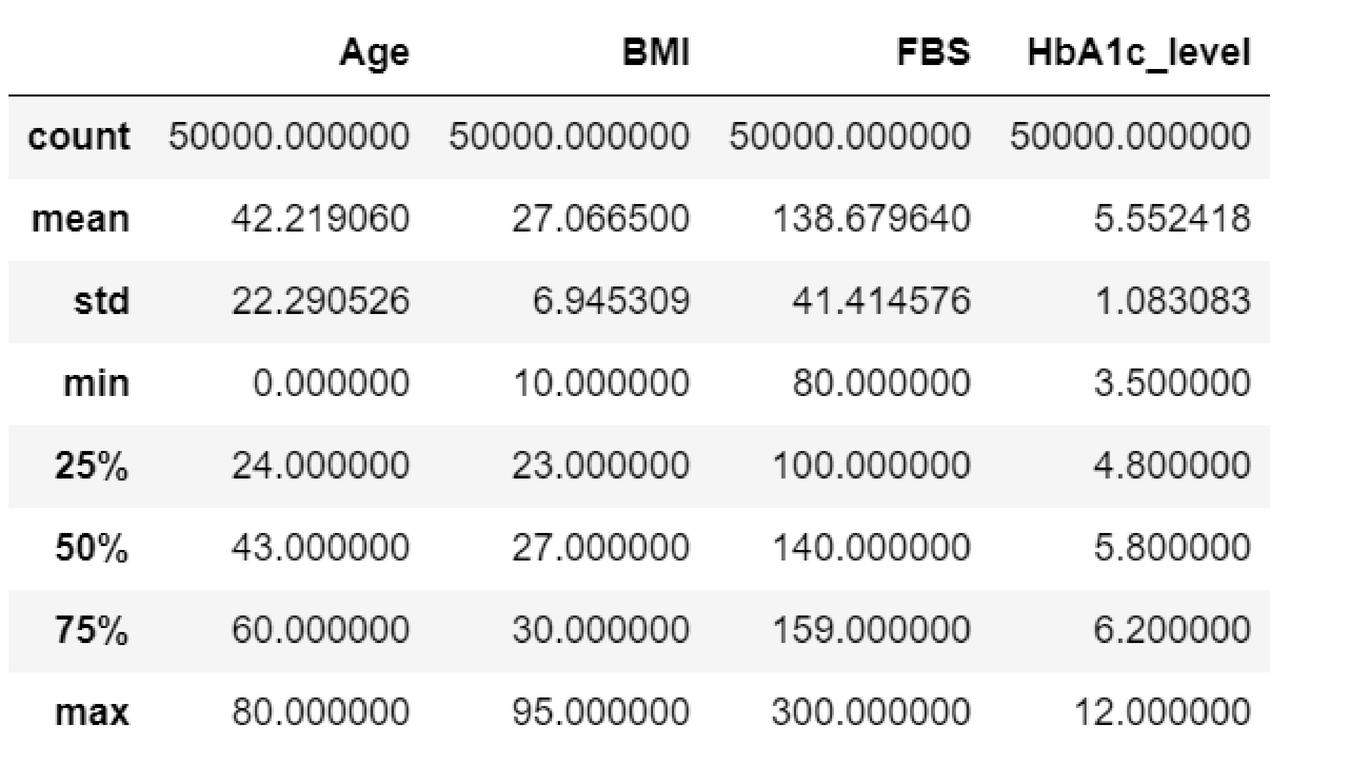


```
variable_to_remove = 'Unnamed: 0'
diabetes = diabetes.drop(columns=[variable_to_remove])
diabetes.head()
```

	Age	Gender	BMI	High_BP	FBS	HbA1c_level	Smoking	Diagnosis
0	80	Female	25	0	140	6.6	0	0
1	54	Female	27	0	80	6.6	0	0
2	28	Male	27	0	158	5.7	[illegible]	0
3	36	Female	23	0	155	5.0	[illegible]	[illegible]
4	76	Male	20	1	155	4.8	1	[illegible]

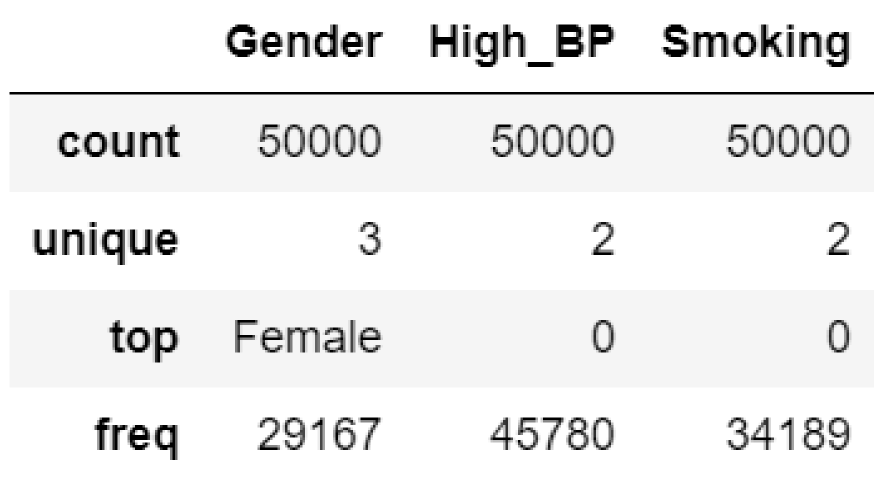
PRELIMINARY EXPLORATION & DATA CLEANING

Numeric predictors:

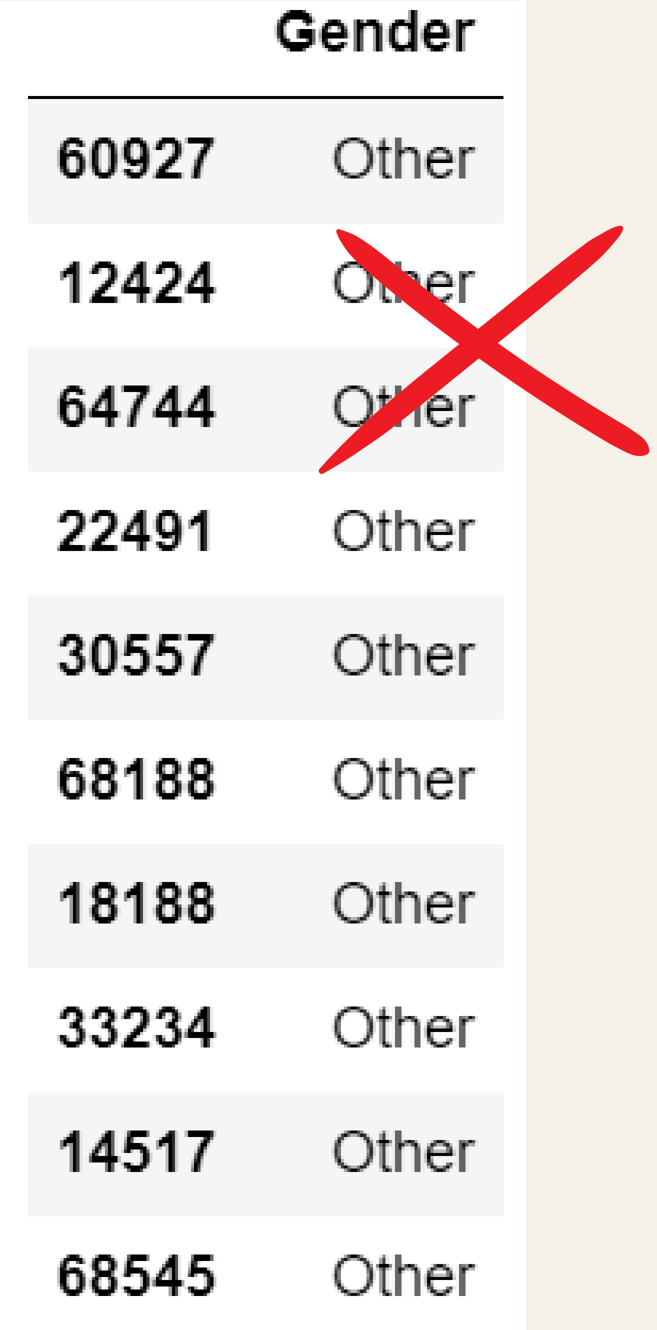


	Age	BMI	FBS	HbA1c_level
count	50000.000000	50000.000000	50000.000000	50000.000000
mean	42.219060	27.066500	138.679640	5.552418
std	22.290526	6.945309	41.414576	1.083083
min	0.000000	10.000000	80.000000	3.500000
25%	24.000000	23.000000	100.000000	4.800000
50%	43.000000	27.000000	140.000000	5.800000
75%	60.000000	30.000000	159.000000	6.200000
max	80.000000	95.000000	300.000000	12.000000

Categorical predictors:



	Gender	High_BP	Smoking
count	50000	50000	50000
unique	3	2	2
top	Female	0	0
freq	29167	45780	34189



	Gender
60927	Other
12424	Other
64744	Other
22491	Other
30557	Other
68188	Other
18188	Other
33234	Other
14517	Other
68545	Other

ANALYTIC VISUALISATION

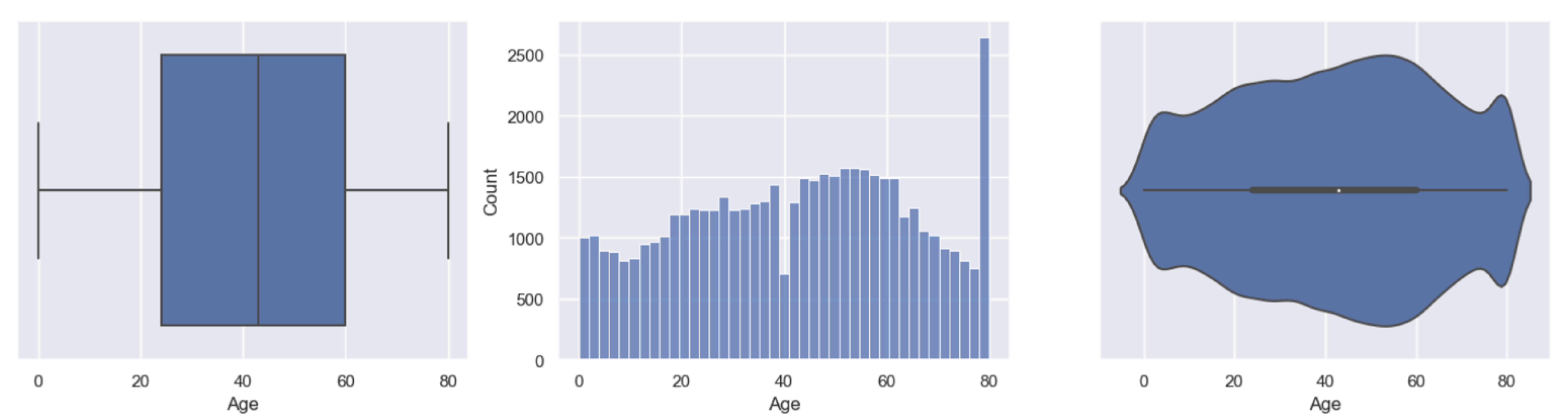
VISUALISATION

Box Plot

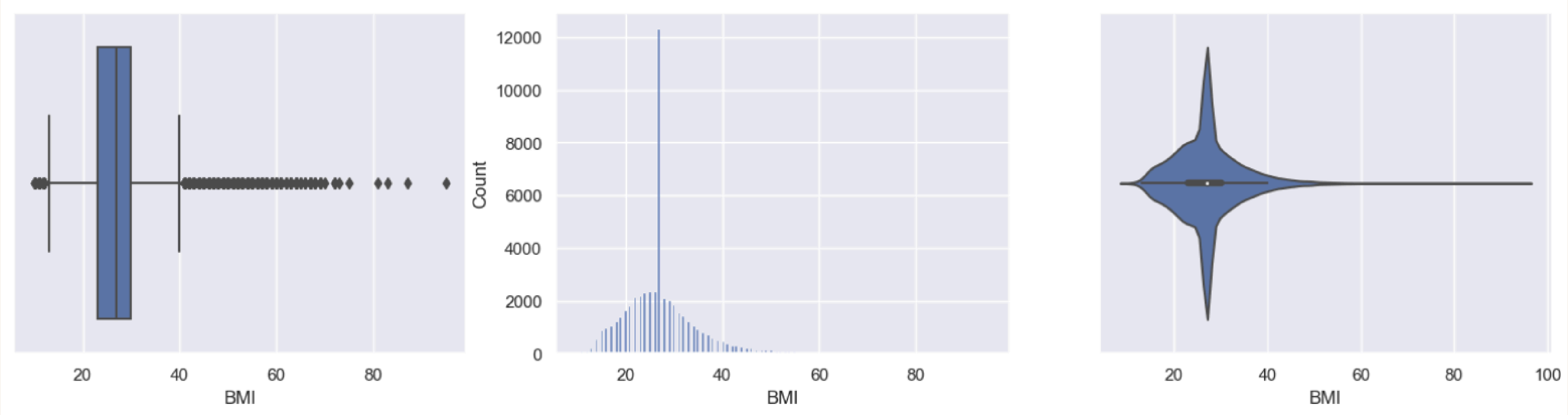
Histogram

Violin Plot

Age:



BMI:



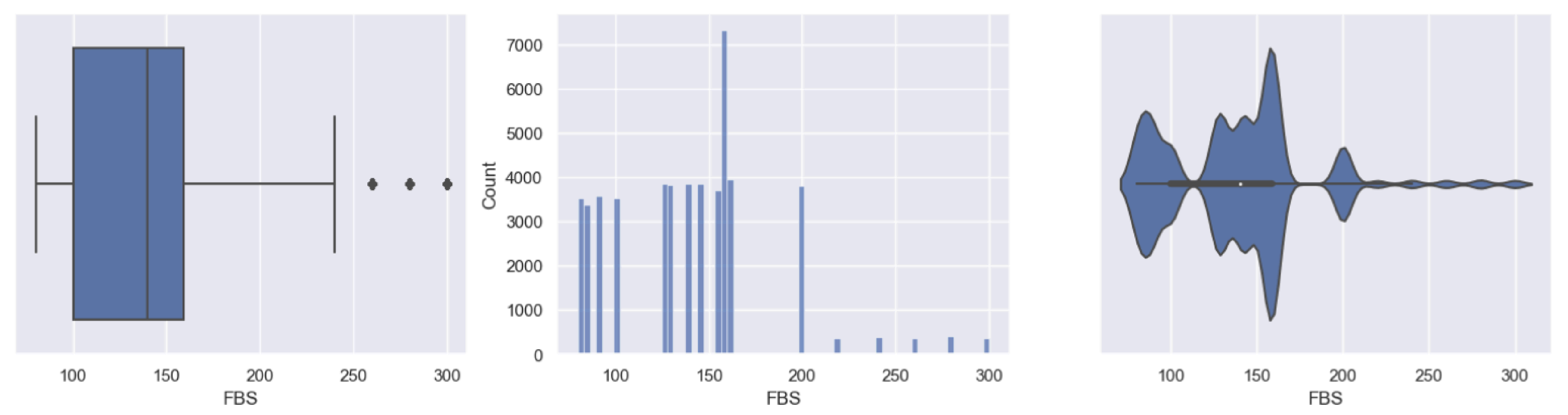
VISUALISATION

Box Plot

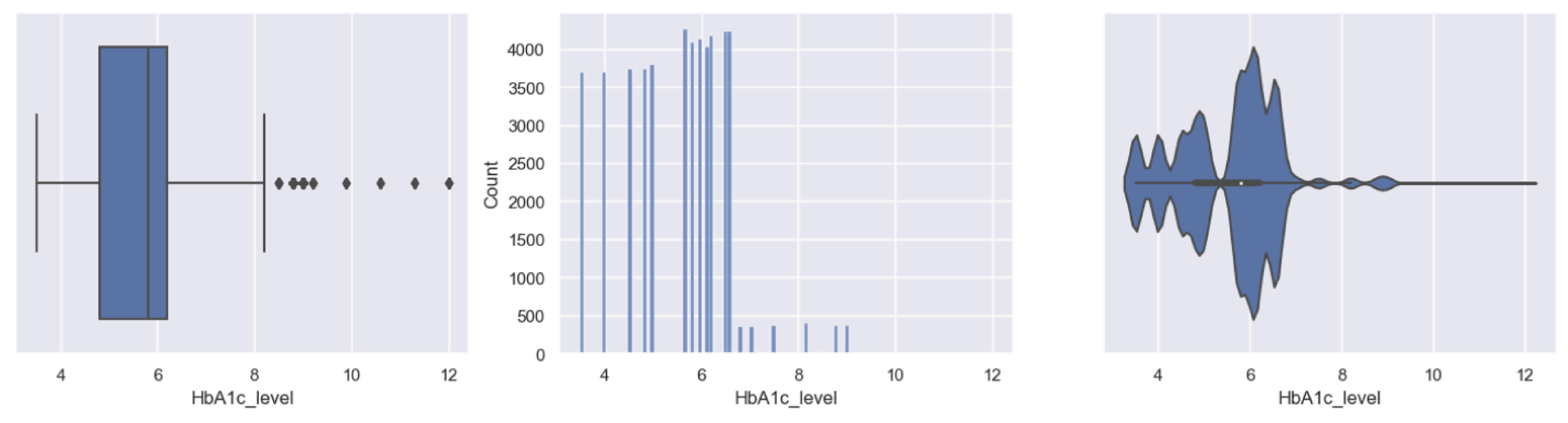
Histogram

Violin Plot

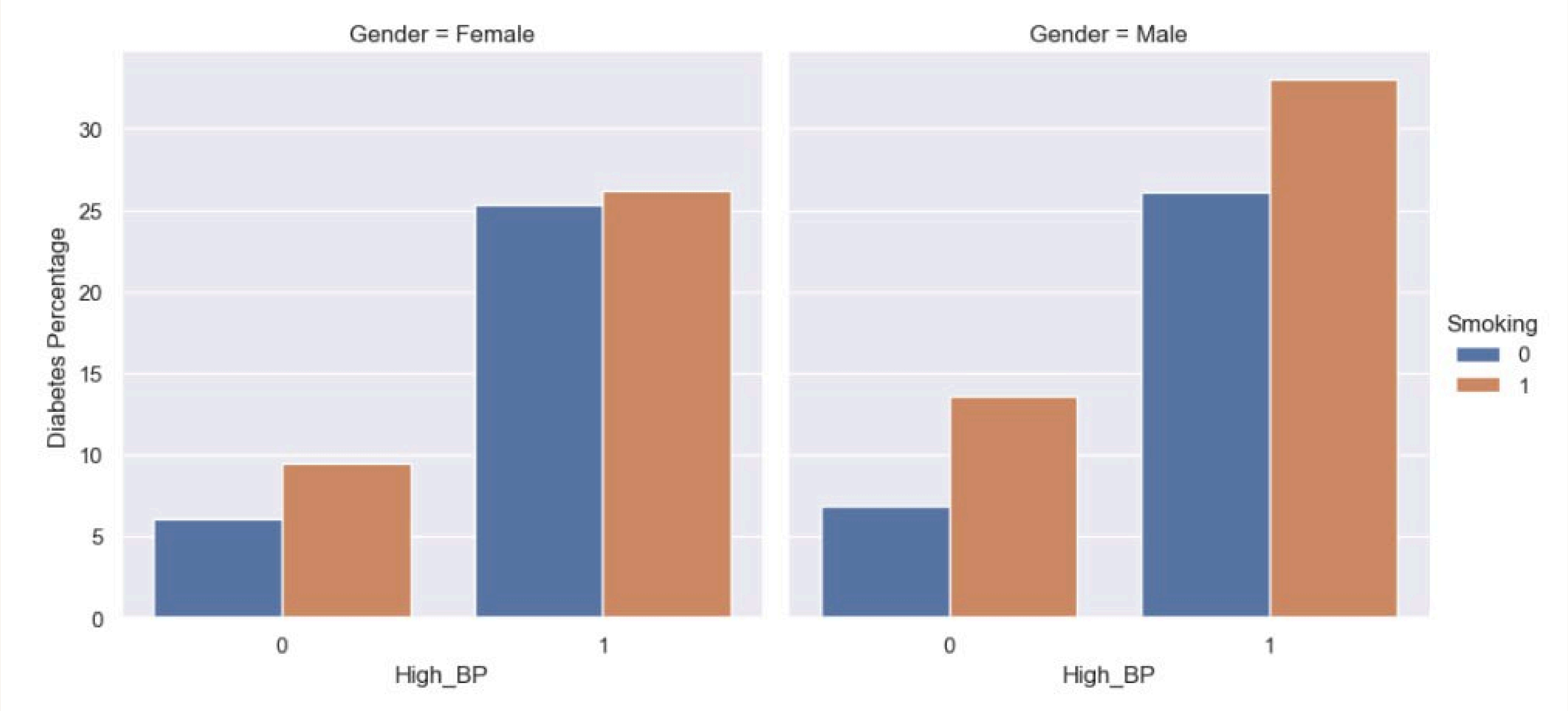
FBS:



HbA1c_level:

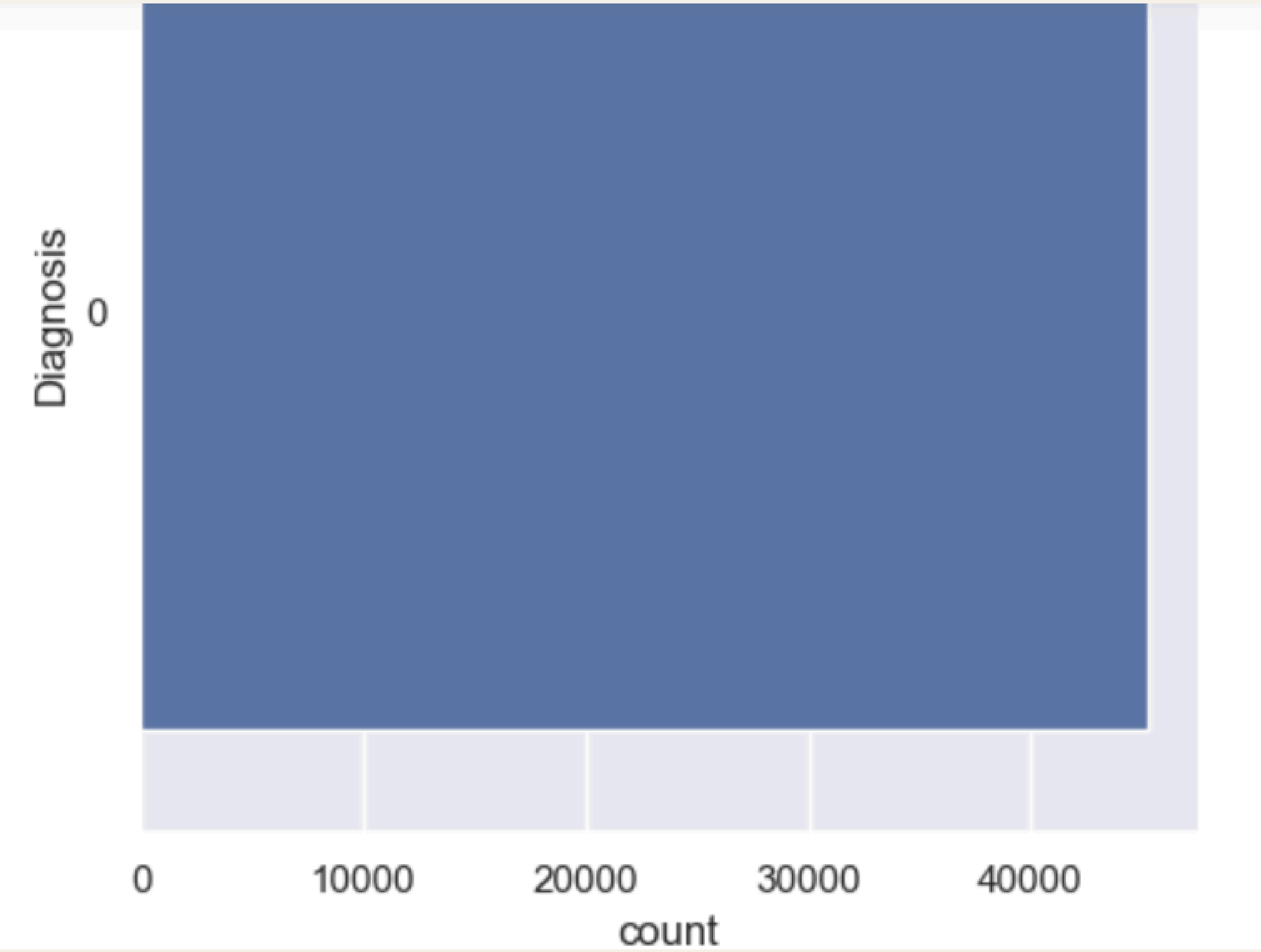


VISUALISATION



- Those with high blood pressure have higher probability to get diabetes.
- Diabetes percentage in male is higher than female.
- Those who smoke have higher probability to get diabetes.

After removing outliers for each of the numerical variables...



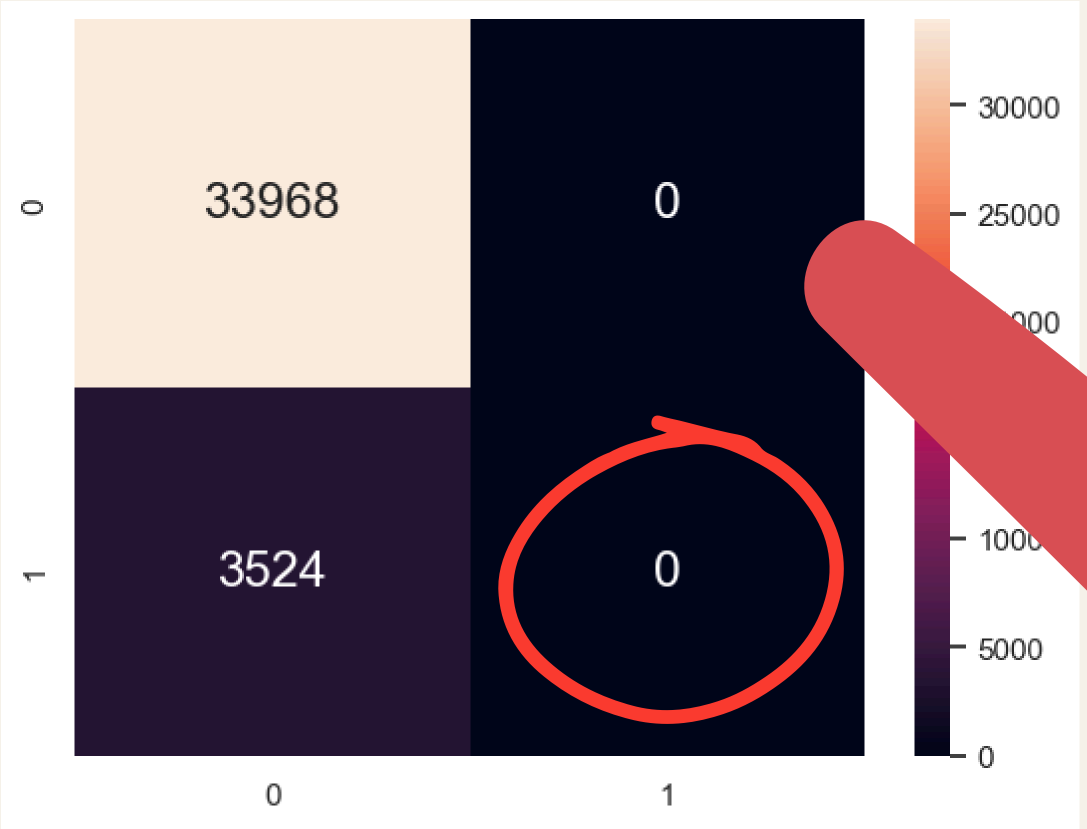
No positive cases of diabetes left in the dataset ?! How could this be ??

We decided not to remove the outliers !!

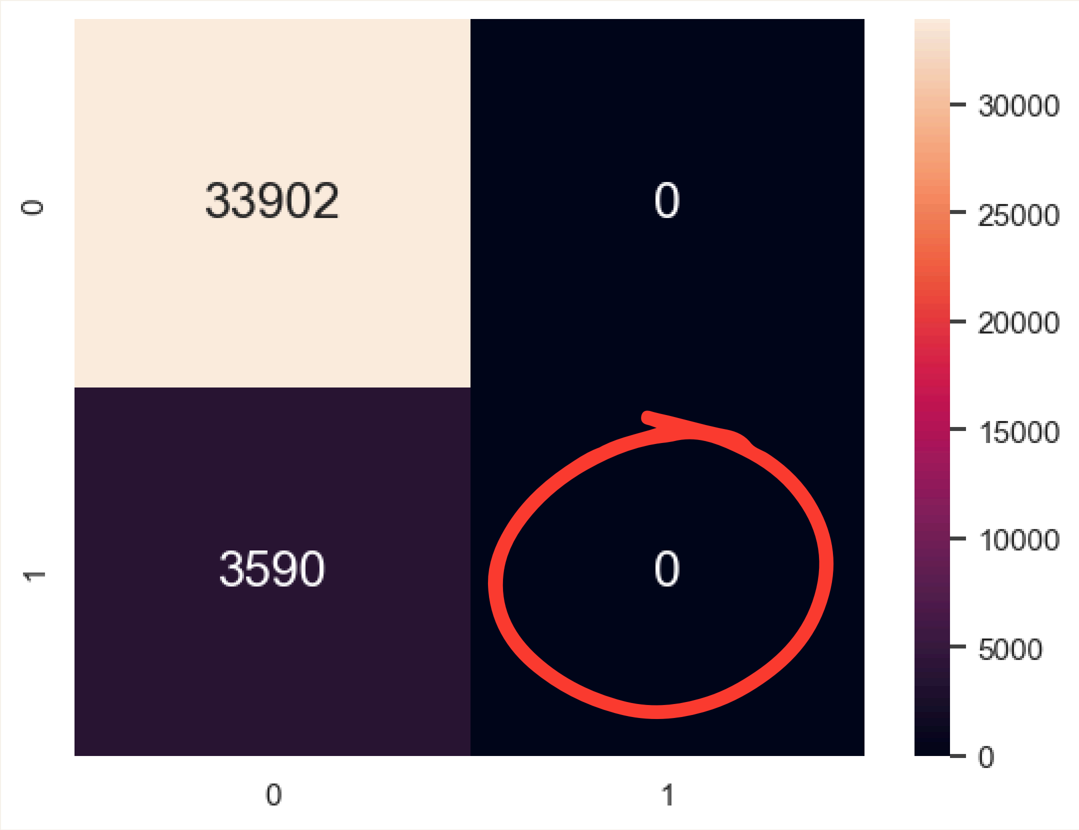
MACHINE LEARNING

UNI-VARIATE CLASSIFICATION

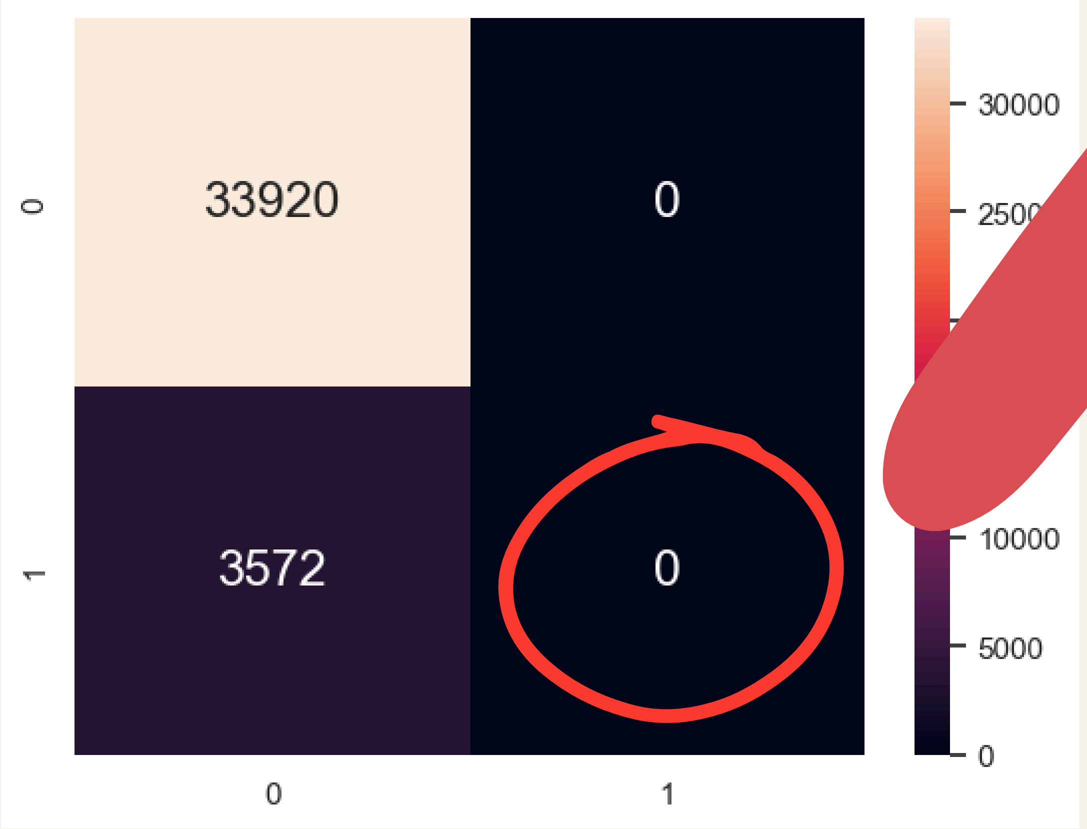
Age:



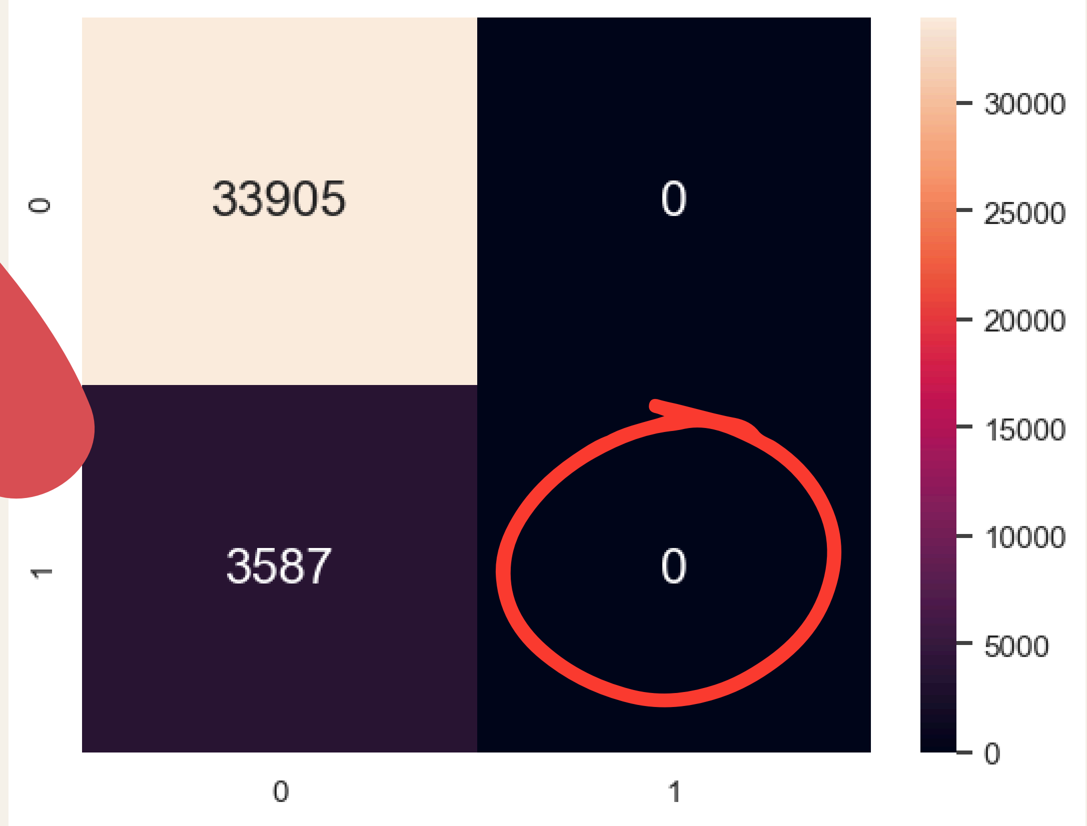
BMI:



FBS:

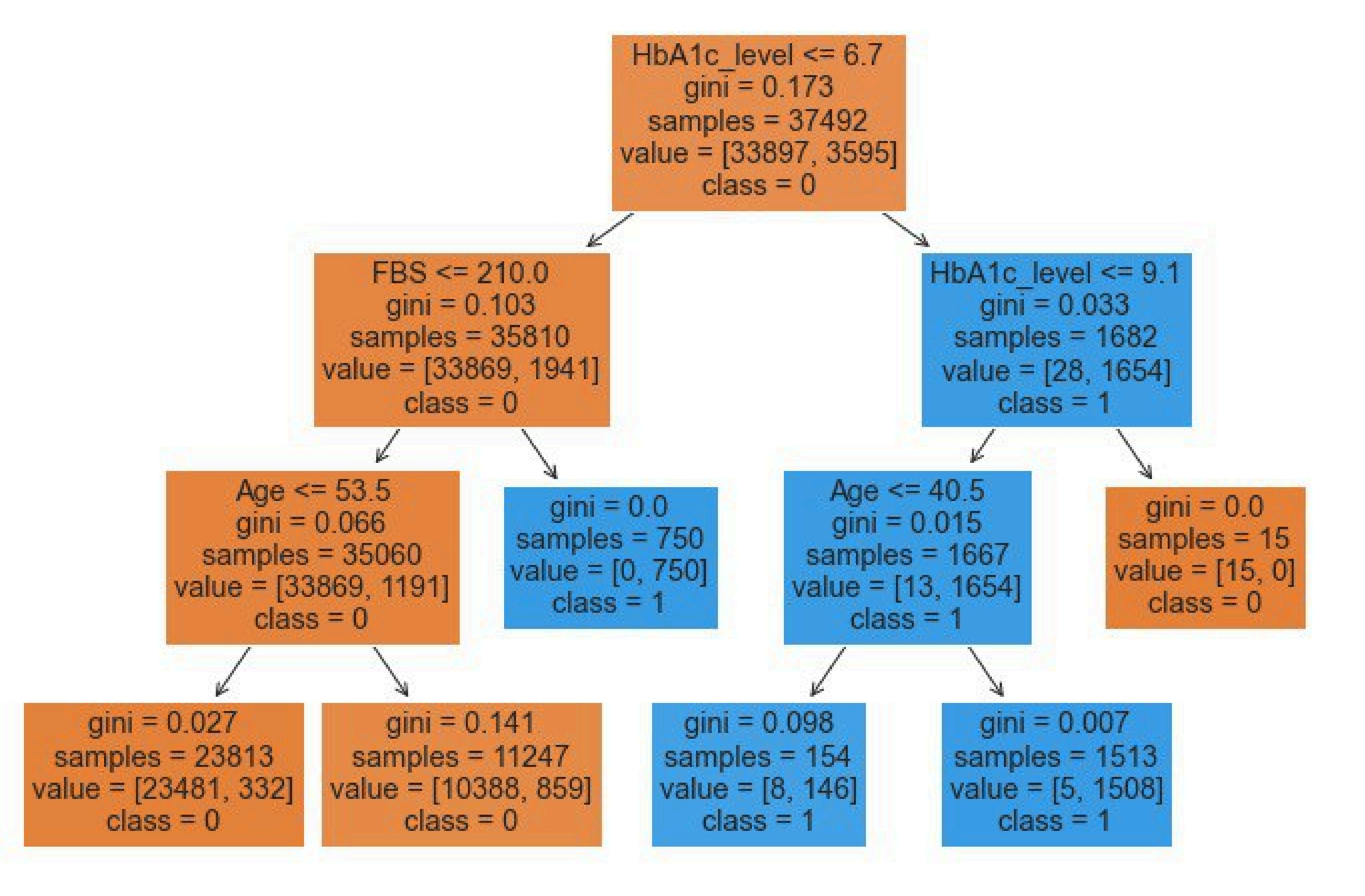


HbA1c_level:



MULTI-VARIATE CLASSIFICATION

DEPTH 3

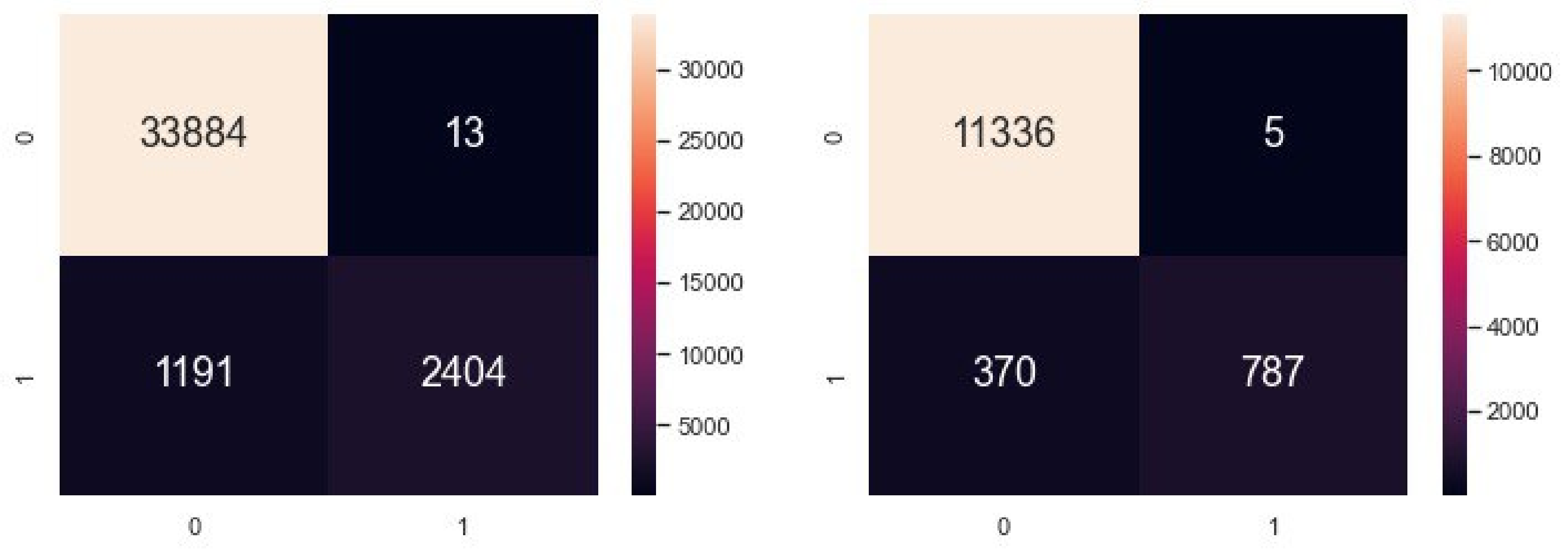


Goodness of Fit of Model Train Dataset Classification Accuracy :

0.9678864824495892

Goodness of Fit of Model Test Dataset Classification Accuracy :

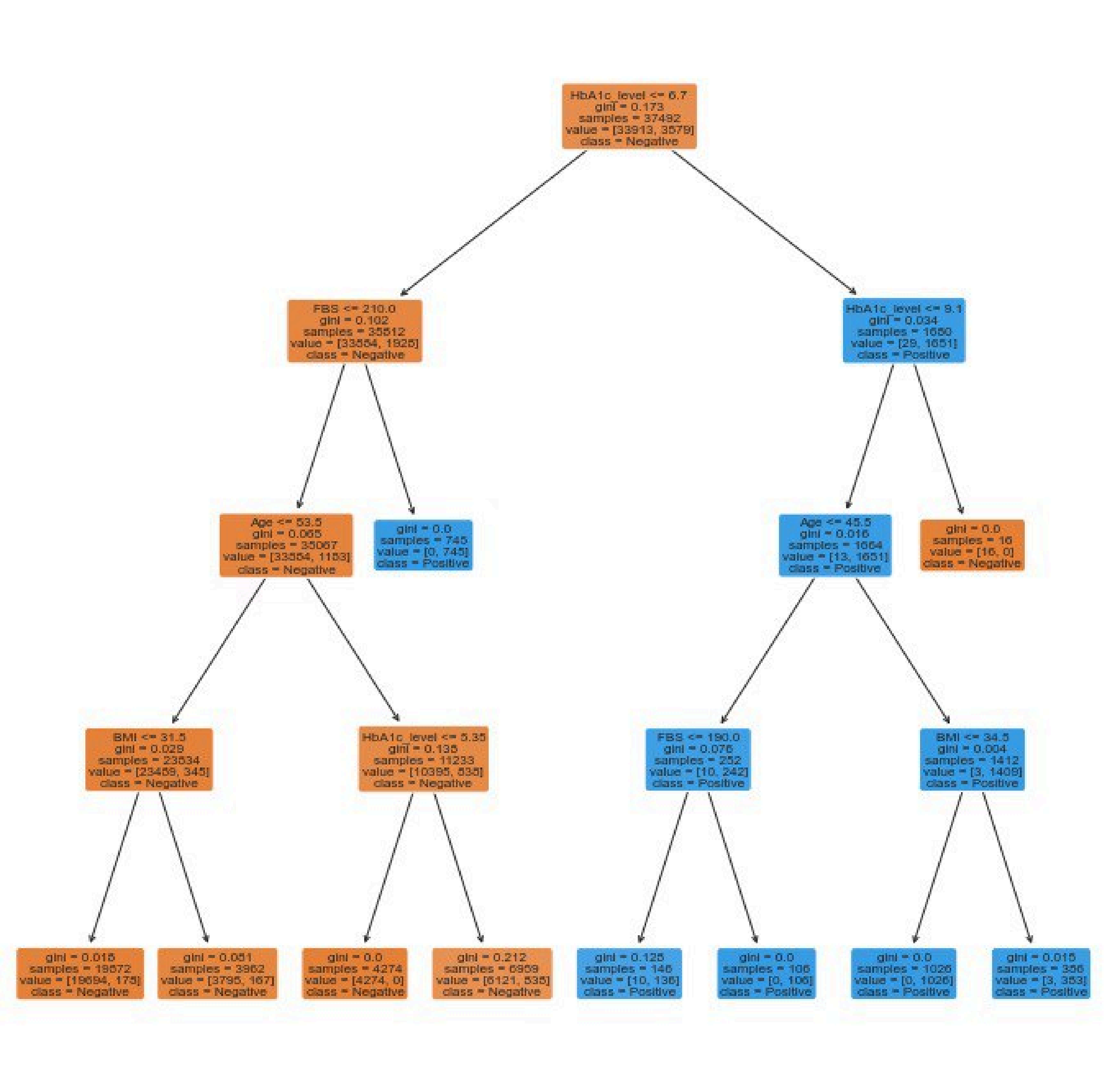
0.969995199231877



True Positive Rate for both train and test: 0.6802

MULTI-VARIATE CLASSIFICATION

DEPTH 4

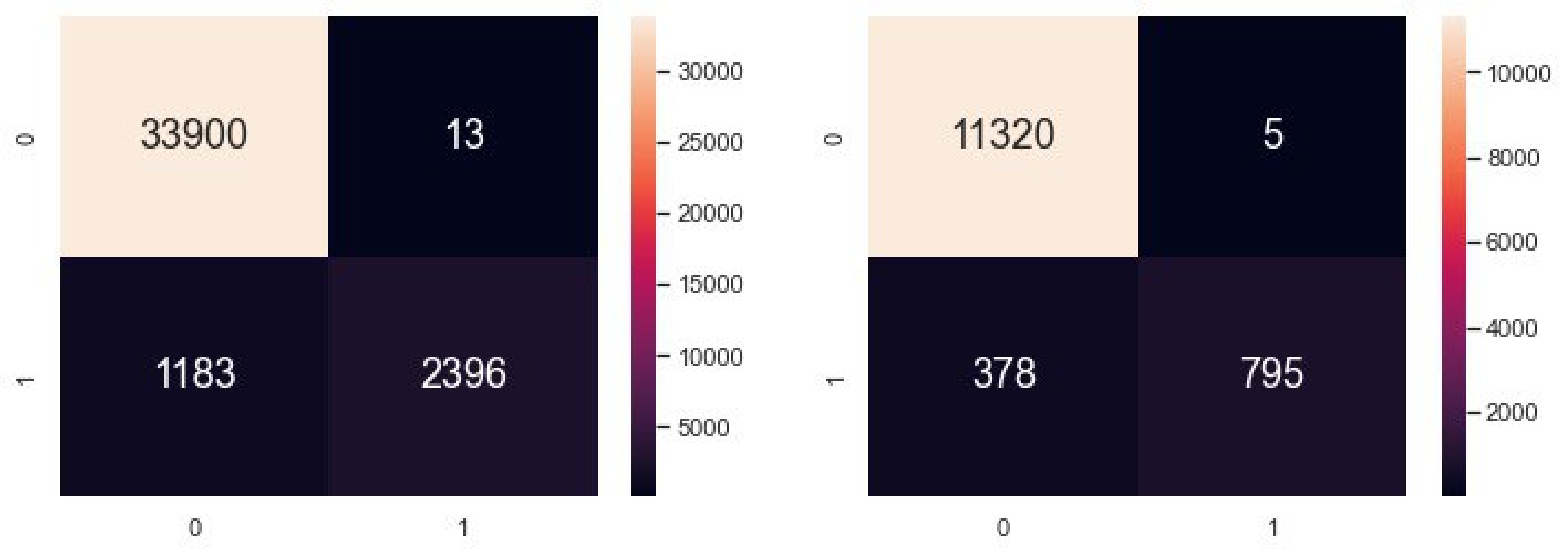


Goodness of Fit of Model Train Dataset Classification Accuracy :

0.9680998613037448

Goodness of Fit of Model Test Dataset Classification Accuracy :

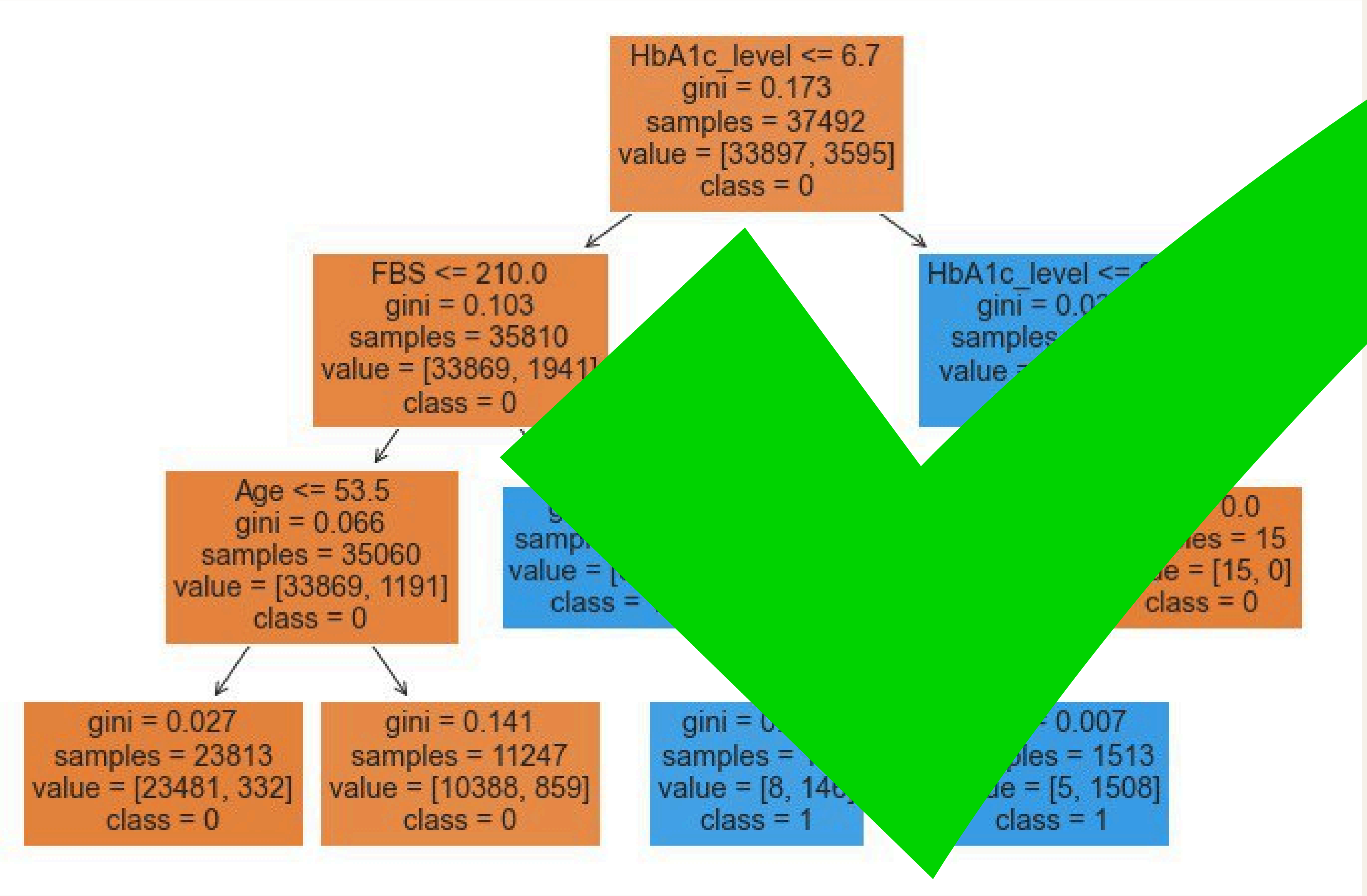
0.9693550968154905



True Positive Rate for both train and test: 0.6714

MULTI-VARIATE CLASSIFICATION

DEPTH 3

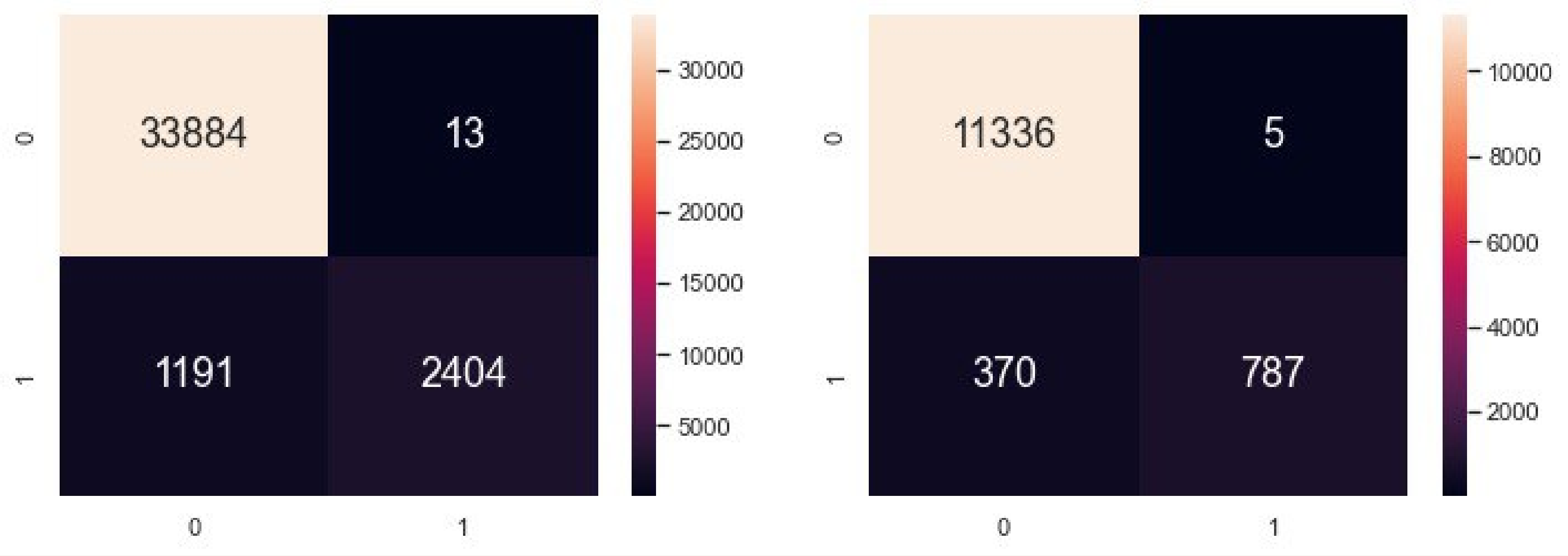


G it of Model Train Dataset Classification Accuracy :

0.9678864824495892

oodness of Fit of Model Test Dataset Classification Accuracy :

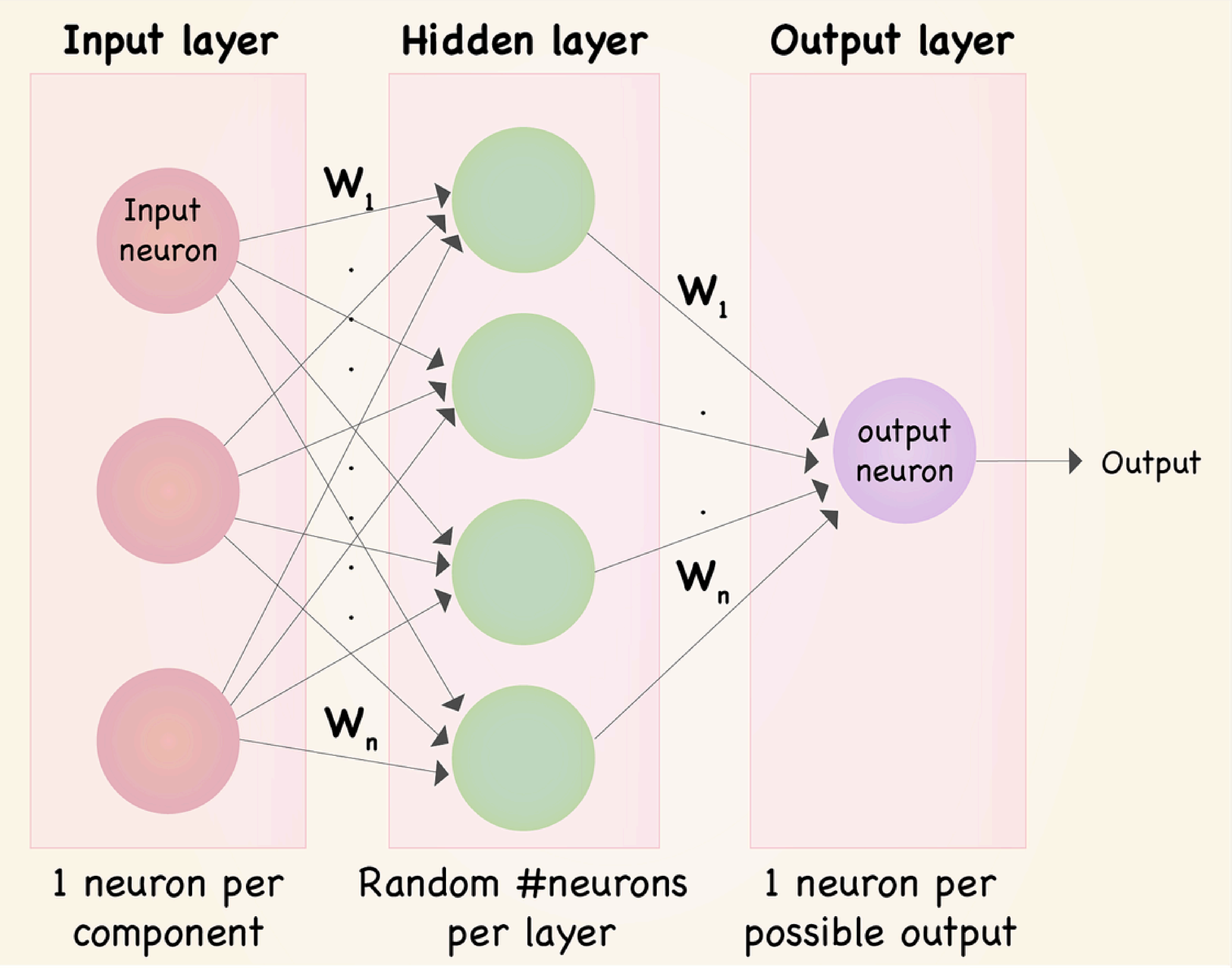
0.969995199231877



True Positive Rate for both train and test: 0.6802

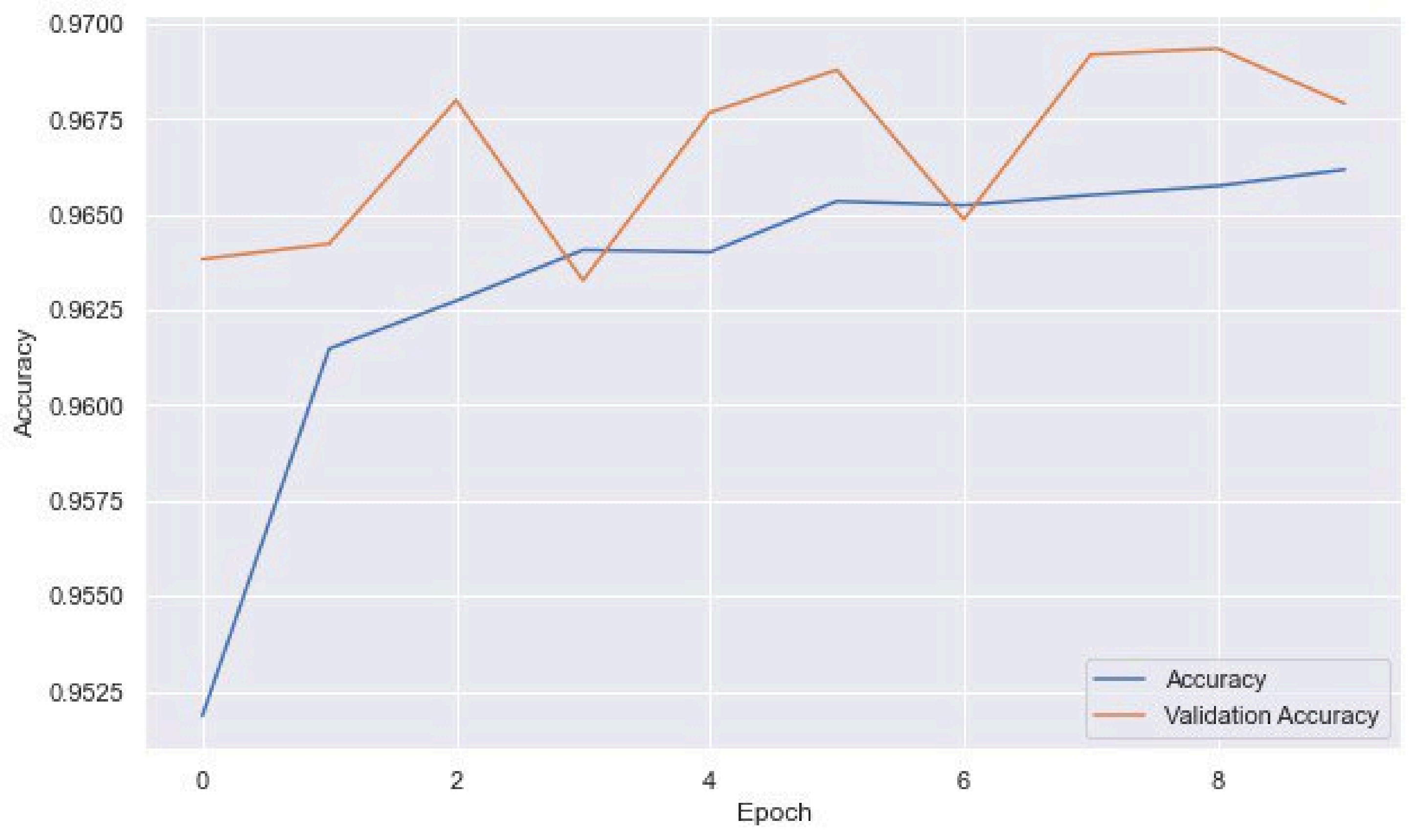
NEURAL NETWORK MODEL

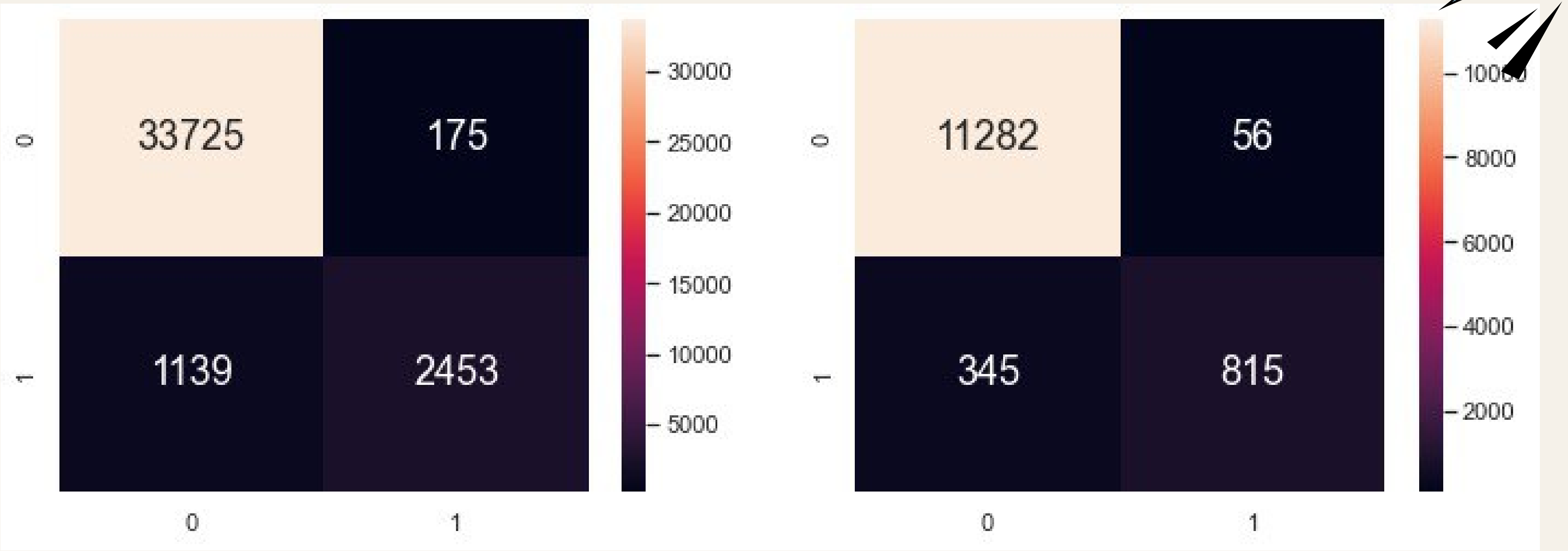
NEURAL NETWORK MODEL



- Excel at earning **complex** patterns from data
- Ability to capture non-linear relationships
- Allows for **comprehensive** evaluation of predictive performance

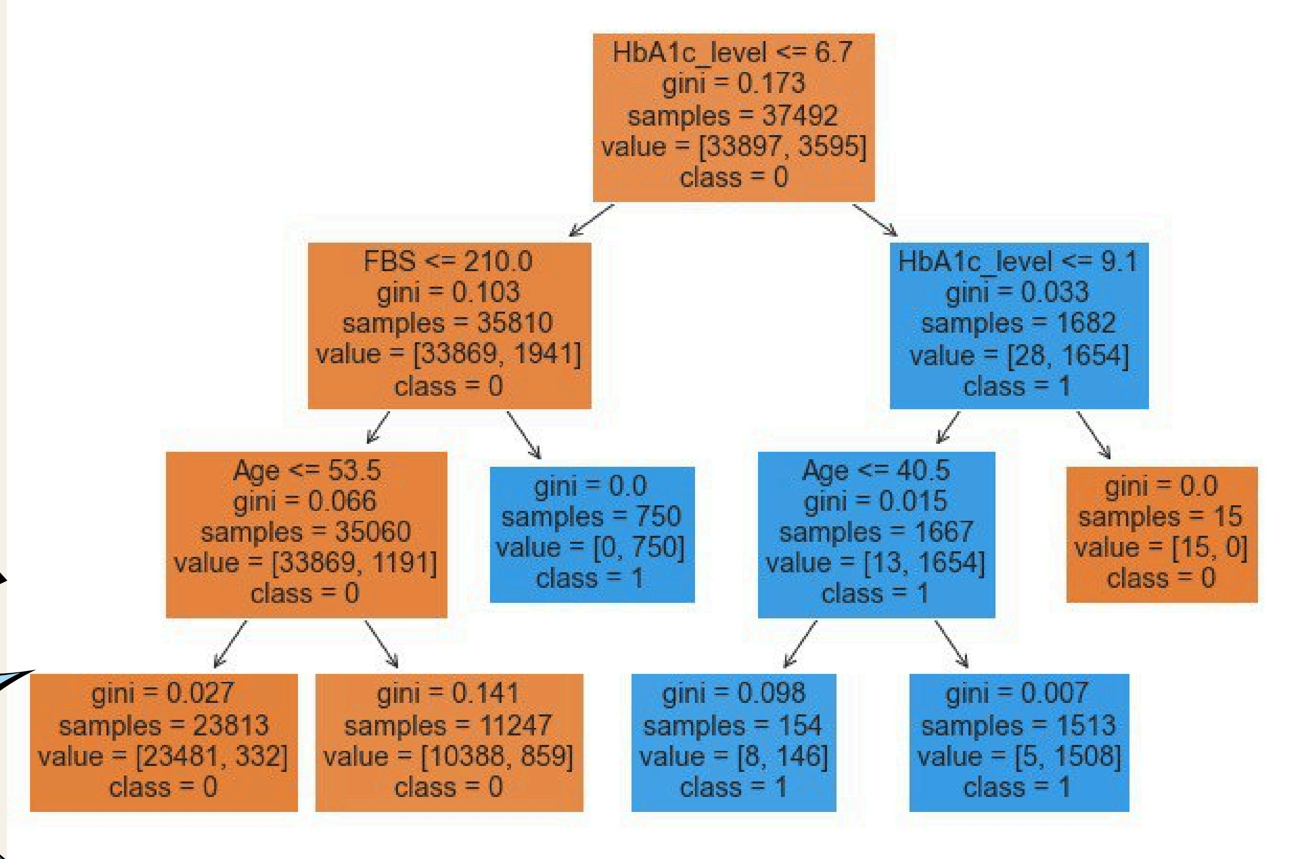
NEURAL NETWORK MODEL





True Positive Rate (TPR) for Train Data: 0.6829064587973274

True Positive Rate (TPR) for Test Data: 0.7025862068965517



HbA1c_level <= 6.7
gini = 0.173
samples = 37492
value = [33897, 3595]
class = 0

FBS <= 210.0
gini = 0.103
samples = 35810
value = [33869, 1941]
class = 0

HbA1c_level <= 9.1
gini = 0.033
samples = 1682
value = [28, 1654]
class = 1

Age <= 53.5
gini = 0.066
samples = 35060
value = [33869, 1191]
class = 0

gini = 0.0
samples = 750
value = [0, 750]
class = 1

Age <= 40.5
gini = 0.015
samples = 1667
value = [13, 1654]
class = 1

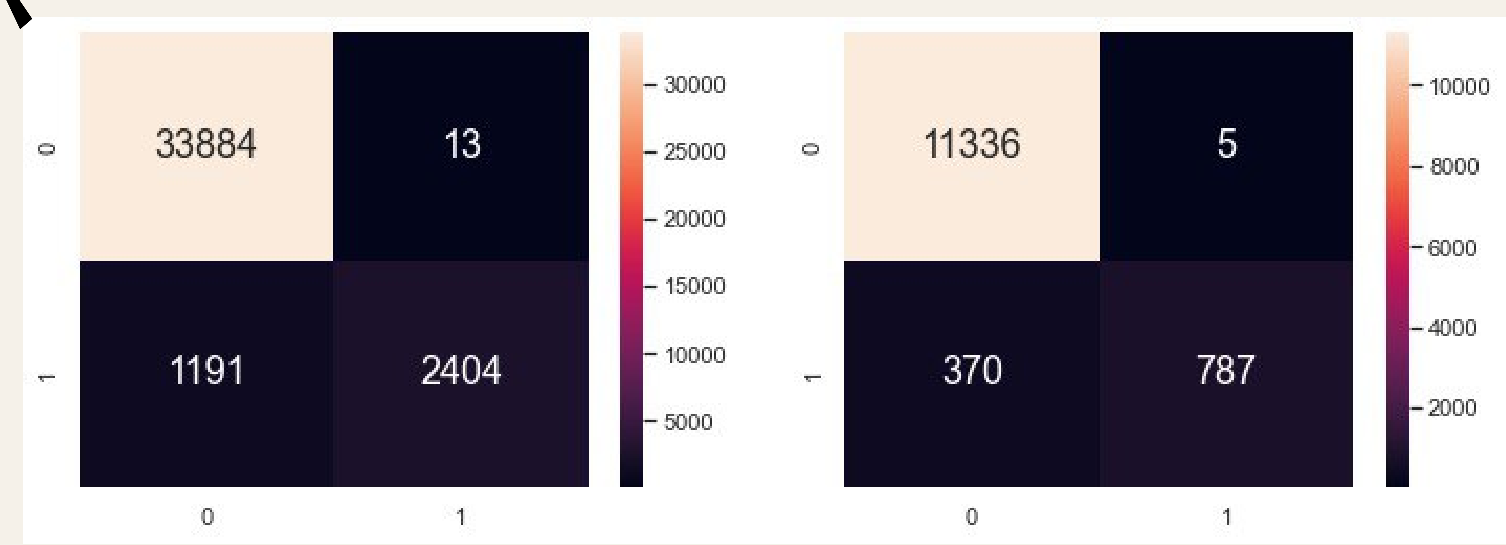
gini = 0.0
samples = 15
value = [15, 0]
class = 0

gini = 0.027
samples = 23813
value = [23481, 332]
class = 0

gini = 0.141
samples = 11247
value = [10388, 859]
class = 0

gini = 0.098
samples = 154
value = [8, 146]
class = 1

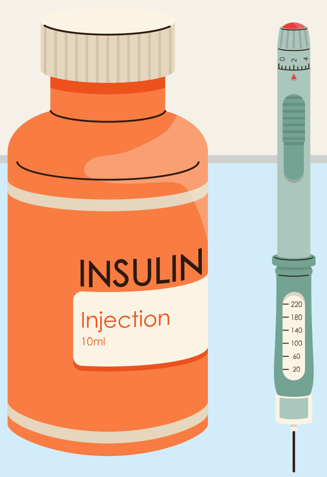
gini = 0.007
samples = 1513
value = [5, 1508]
class = 1



True Positive Rate for both train and test: 0.6802

CONCLUSION

INSIGHTS



Complexitites inherent in diabetes prediction

Importance of adopting a multivariate approach

Limitations of uni-variate classification

Diabetes diagnosis is influenced by multiple factors -> multi-variate approach

Neural network model enhanced the predictive performance

Moving forward, by embracing a multidisciplinary approach, we can continue to advance our understanding of diabetes and develop innovative solutions to improve prevention, diagnosis, and management strategies.

THANK YOU!

REFERENCES

Simpe feature to detect diabetes. (n.d.). Www.kaggle.com. Retrieved April 20, 2024, from https://www.kaggle.com/datasets/simaanjali/diabetes-simple-diagnosis/data

Mesquita, D. (n.d.). Python AI: How to Build a Neural Network & Make Predictions – Real Python. Realpython.com. https://realpython.com/python-ai-neural-network/

A, A. (2019, January 13). First neural network for beginners explained (with code). Medium; Towards Data Science. https://towardsdatascience.com/first-neural-network-for-beginners-explained-with-code-4cfd37e06eaf