CLASSIFICATION

diabetes dataset

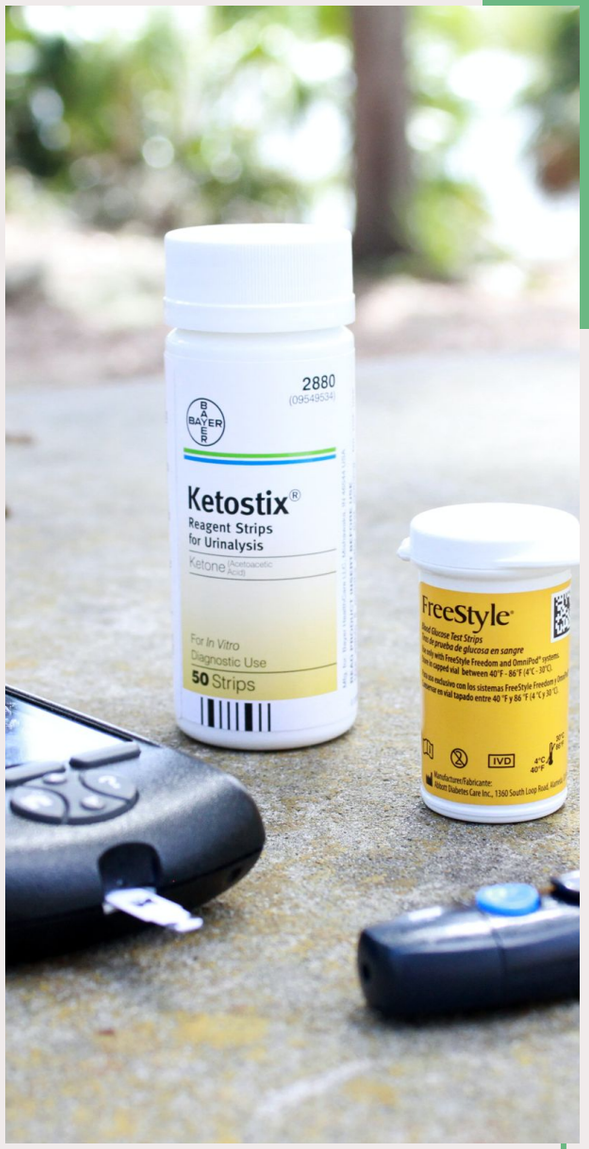
By samah alrefaei

TABLE OF CONTACT

PROBLEM

The objective of this diabetes dataset is to predict whether patient has diabetes or not

The datasets consist of several medical predictor (independent) variables and one target (dependent) variable, Outcome.



SOLUTION

in this project, I will try to build a machine learning classification model to accurately predict whether or not the patients in the dataset have diabetes or not?

Dataset description

downloaded from Kaggle

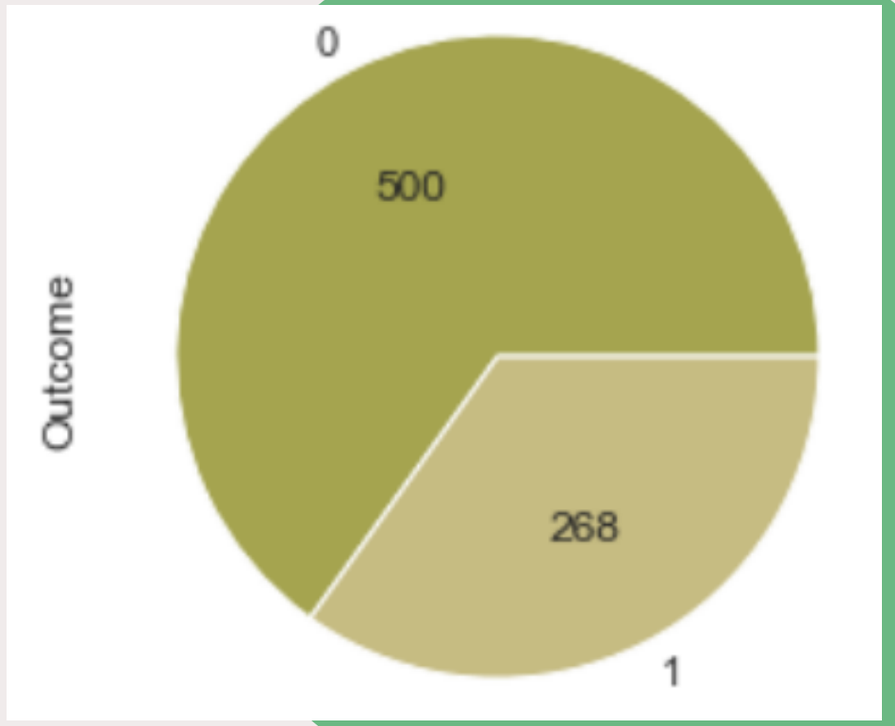
THERE ARE 768 OBSERVATIONS WITH 8 MEDICAL PREDICTOR FEATURES (INPUT) AND 1 TARGET **VARIABLE** (OUTPUT 0 FOR "NO" OR 1 FOR"YES")

FEATURE	DESCRIPTION
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance tes
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-Hour serum insulin (mu U/ml)
BMI	Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction	Diabetes pedigree function
AgeOutcame	Age (years)

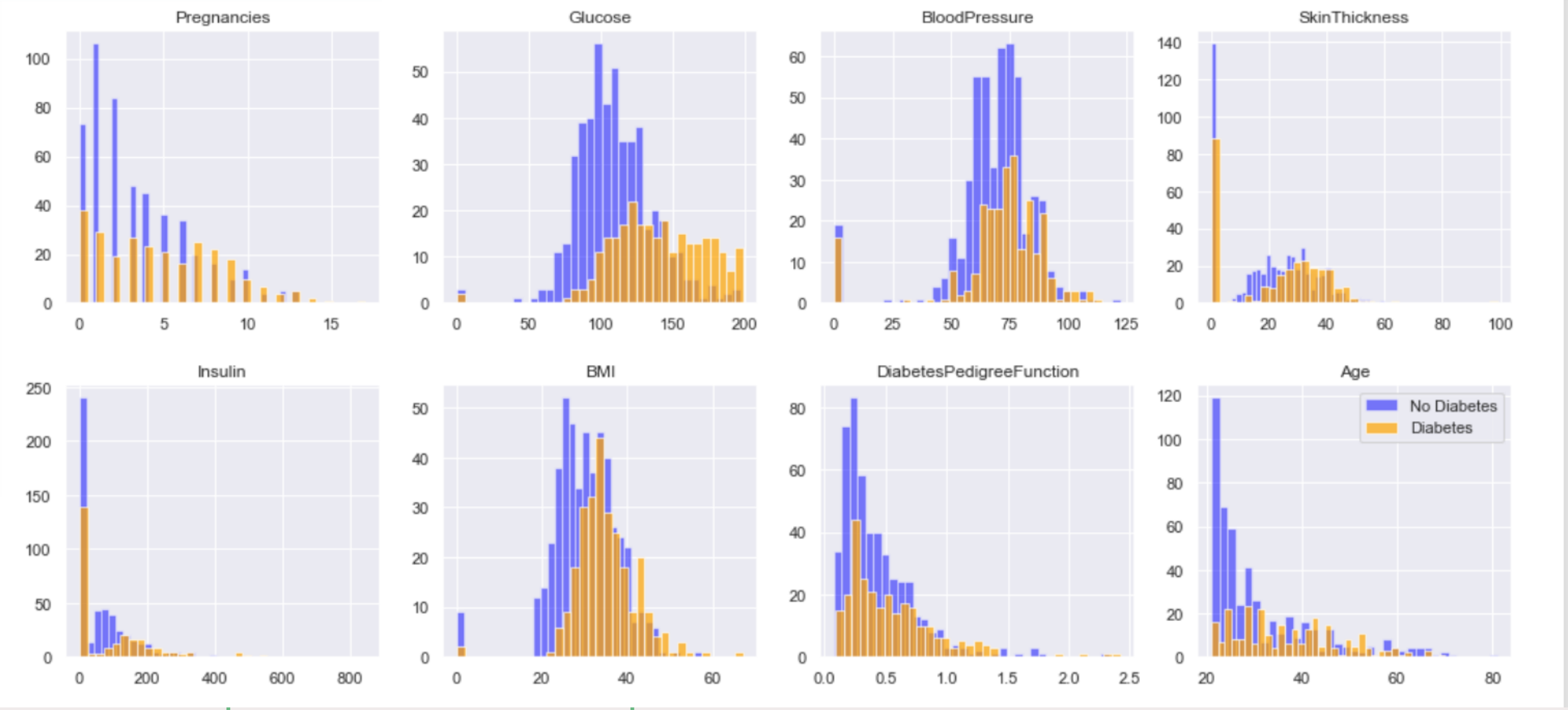
DATASET

Exploratory data analysis

(output 0 for ”no diabete” or 1 for”yes diabetes”)



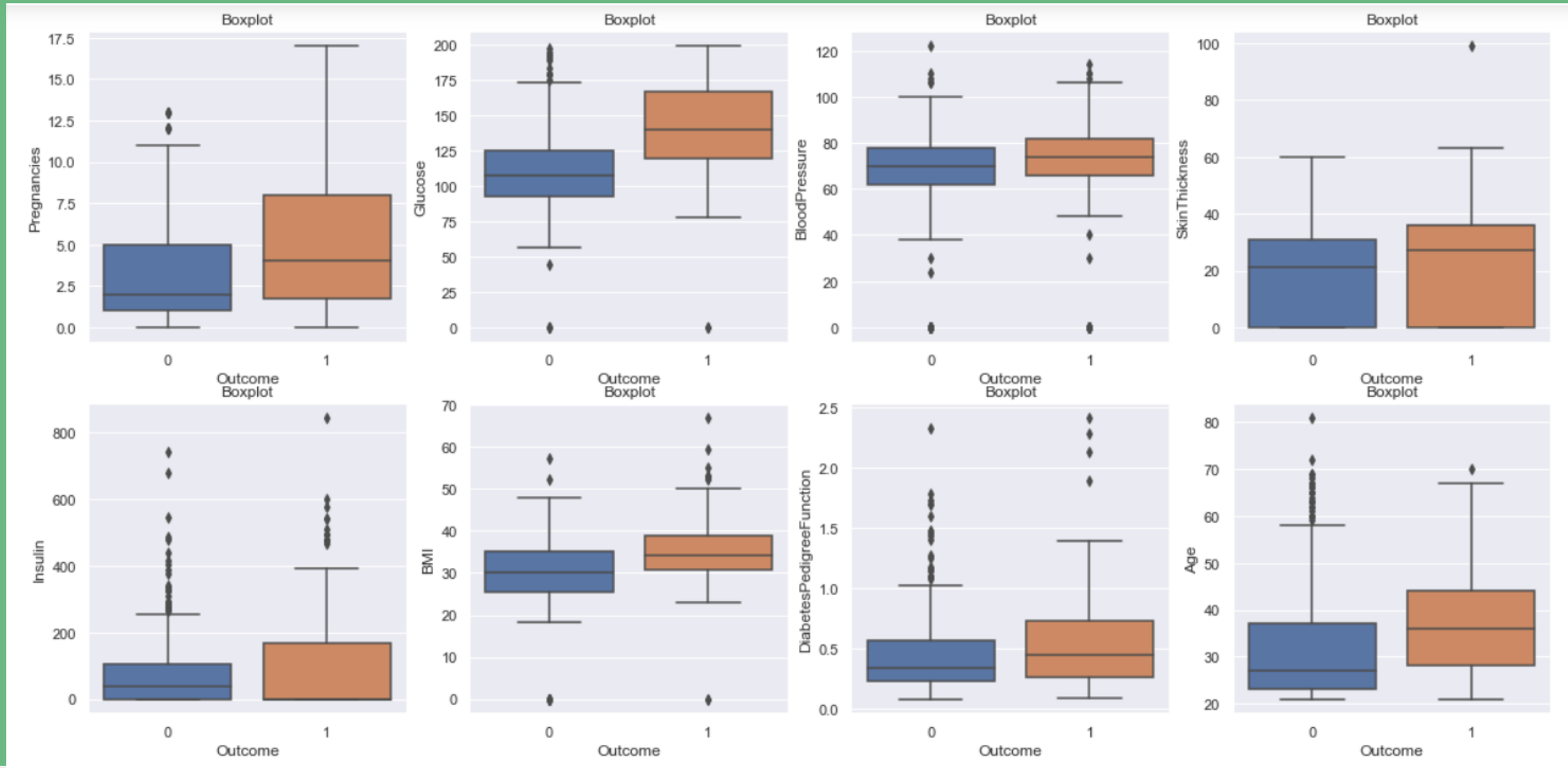
Exploratory Data Analysis (EDA)



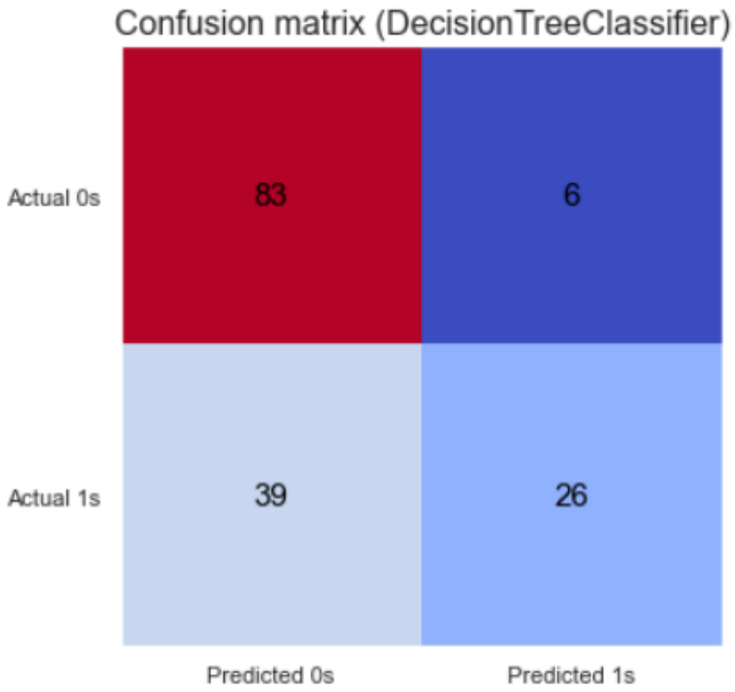
To analyse feature-outcome distribution in visualisation

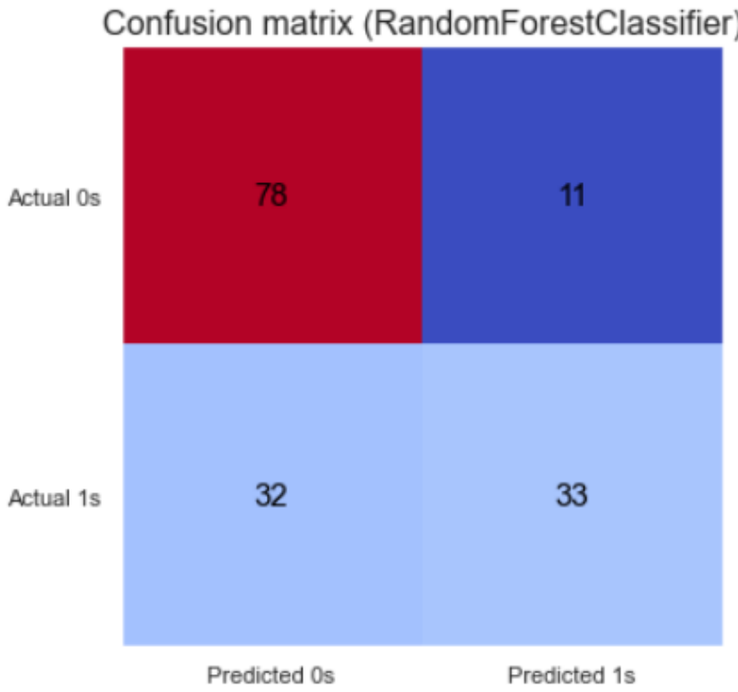
Features Correlation

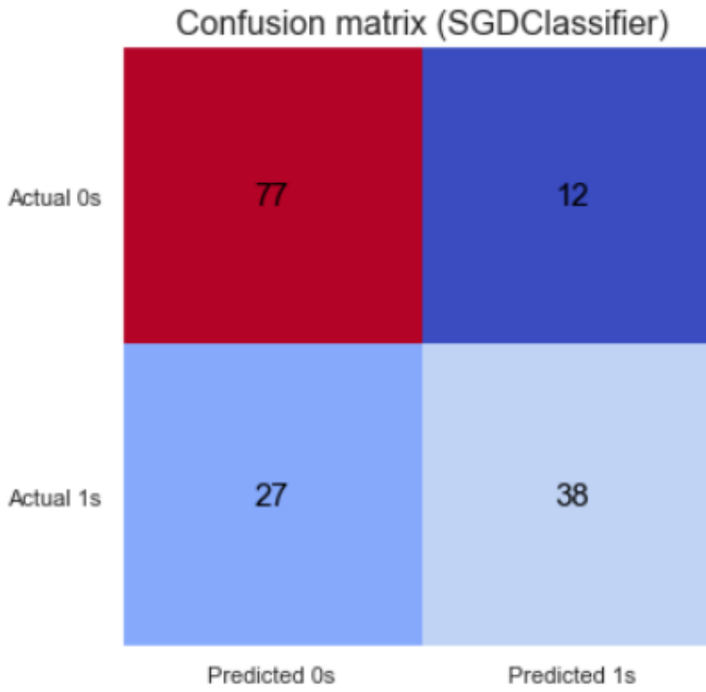
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.129459	0.141282	-0.081672	-0.073535	0.017683	-0.033523	0.544341	0.221898
Glucose	0.129459	1.000000	0.152590	0.057328	0.331357	0.221071	0.137337	0.263514	0.466581
BloodPressure	0.141282	0.152590	1.000000	0.207371	0.088933	0.281805	0.041265	0.239528	0.065068
SkinThickness	-0.081672	0.057328	0.207371	1.000000	0.436783	0.392573	0.183928	-0.113970	0.074752
Insulin	-0.073535	0.331357	0.088933	0.436783	1.000000	0.197859	0.185071	-0.042163	0.130548
BMI	0.017683	0.221071	0.281805	0.392573	0.197859	1.000000	0.140647	0.036242	0.292695
abetesPedigreeFunction	-0.033523	0.137337	0.041265	0.183928	0.185071	0.140647	1.000000	0.033561	0.173844
Age	0.544341	0.263514	0.239528	-0.113970	-0.042163	0.036242	0.033561	1.000000	0.238356
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356	1.000000


Boxplot
Pregnancies
Outcome
Boxplot
Glucose
Outcome
Boxplot
BloodPressure
Outcome
Boxplot
SkinThickness
Outcome
Boxplot
Insulin
Outcome
Boxplot
BMI
Outcome
Boxplot
DiabetesPedigreeFunction
Outcome
Boxplot
Age
Outcome

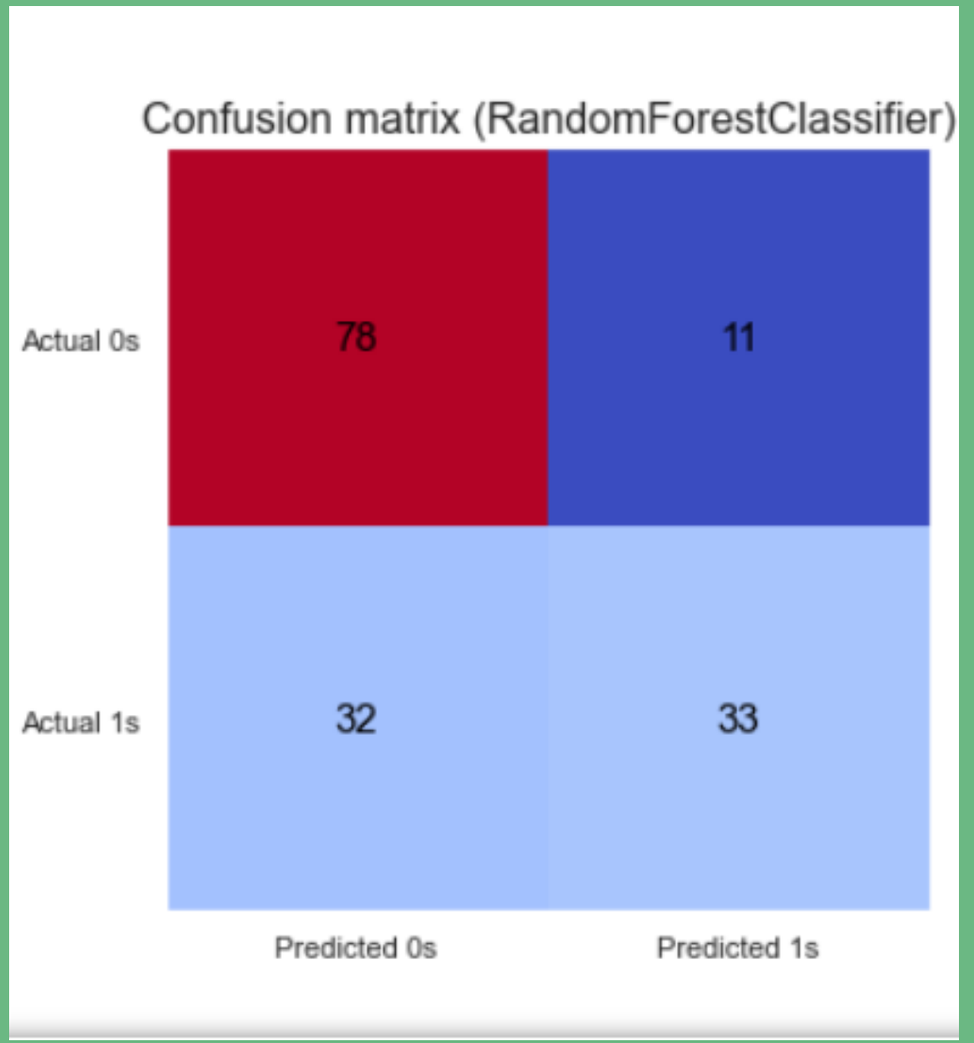
	Model	Score
4	RandomForestClassifier	0.794749
2	SVC	0.786645
0	LogisticRegression	0.783380
8	CatBoostClassifier	0.783327
7	XGBClassifier	0.780075
6	LGBMClassifier	0.776929
3	DecisionTreeClassifier	0.775250
1	KNeighborsClassifier	0.775237
5	SGDClassifier	0.760562


Confusion matrix (DecisionTreeClassifier)
Actual 0s
Actual 1s
83
6
39
26
Predicted 0s
Predicted 1s


Confusion matrix (RandomForestClassifier)
Actual 0s
Actual 1s
78
11
32
33
Predicted 0s
Predicted 1s


Confusion matrix (SGDClassifier)
Actual 0s
Actual 1s
77
12
27
38
Predicted 0s
Predicted 1s

The winning model is:



	Model	Score
4	RandomForestClassifier	0.794749
2	SVC	0.786645
0	LogisticRegression	0.783380
8	CatBoostClassifier	0.783327
7	XGBClassifier	0.780075
6	LGBMClassifier	0.776929
3	DecisionTreeClassifier	0.775250
1	KNeighborsClassifier	0.775237
5	SGDClassifier	0.760562

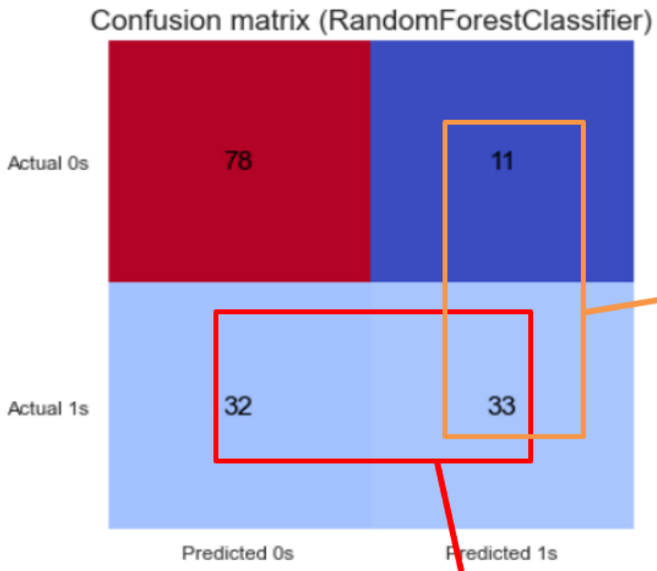
	model	accuracy	precision	recall	f1score
0	LogisticRegression	**0.6753**	0.7142	0.3846	0.5
1	KNN	0.6818	0.7352	0.384	0.5050
2	SVC	0.675	0.6923	0.415	0.51
3	DecisionTree	0.7077	0.8125	0.4	0.5360
4	RF	0.7207	0.75	0.507	0.60
5	SGD	0.727273	0.76	0.584	0.660

Conclusion

In this project, the **Random forest** model has achieved prediction (Recall) score of **79%**

Out of all diabetic patients,79% of them will be classified correctly using medical diagnostic measurements

Optimal threshold 0.207
F1 Score = 0.719



proportion of correct predictions out of all predicted diabetic cases

Precision = 33/44
= 0.75

Recall = 78/110
= 0.79

proportion of correct predictions out of all actual diabetic cases


2880
(09549534)
BAYER
Ketostix®
Reagent Strips
for Urinalysis
Ketone (Acetoacetic Acid)
For In Vitro
Diagnostic Use
50 Strips
FreeStyle
IVD
4°C
40°F

THANK YOU