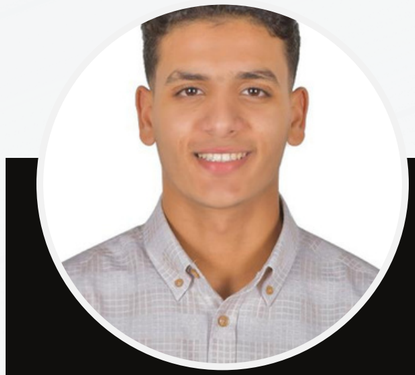
DIABETES DETECTION

CONTENT

OUR TEAM

Omar Ashraf

Ahmed Hassan

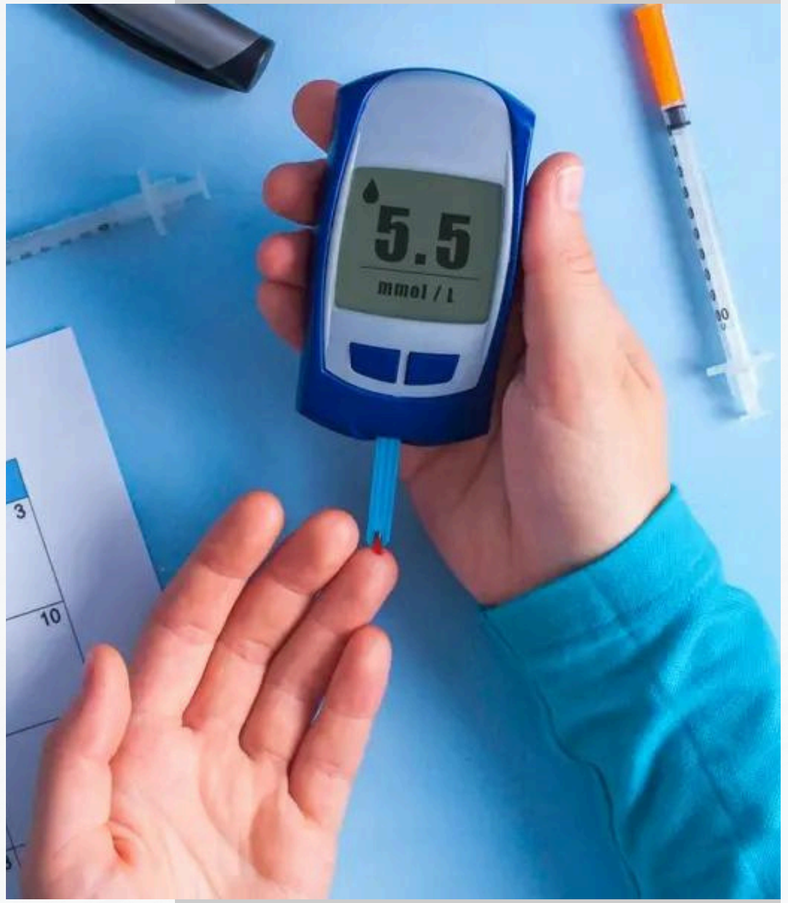


Moaz Atef

INTRODUCTION

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces.

The IDF Diabetes Atlas (2021) reports state that 10.5% of the adult population (20-79 years) has diabetes, with almost half unaware that they are living with the condition.

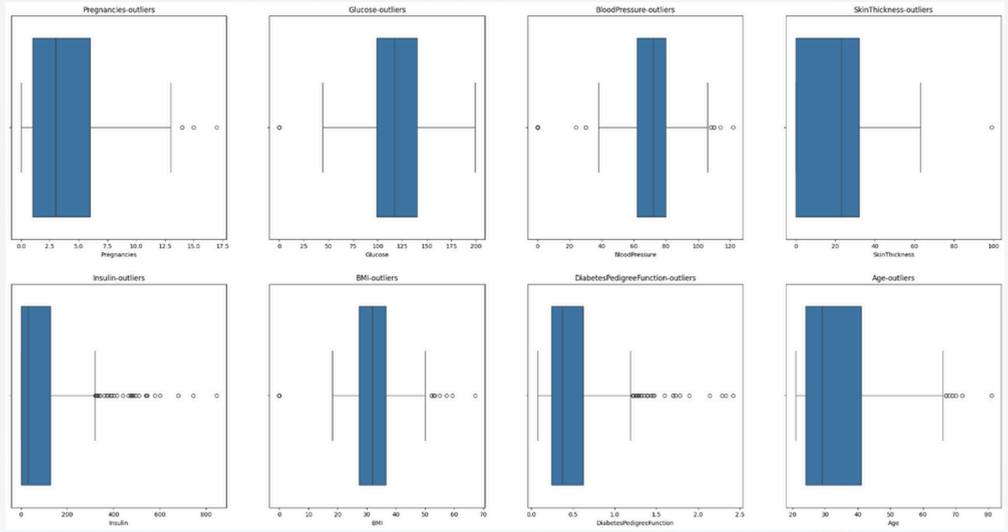


CHOOSING THE DATASET

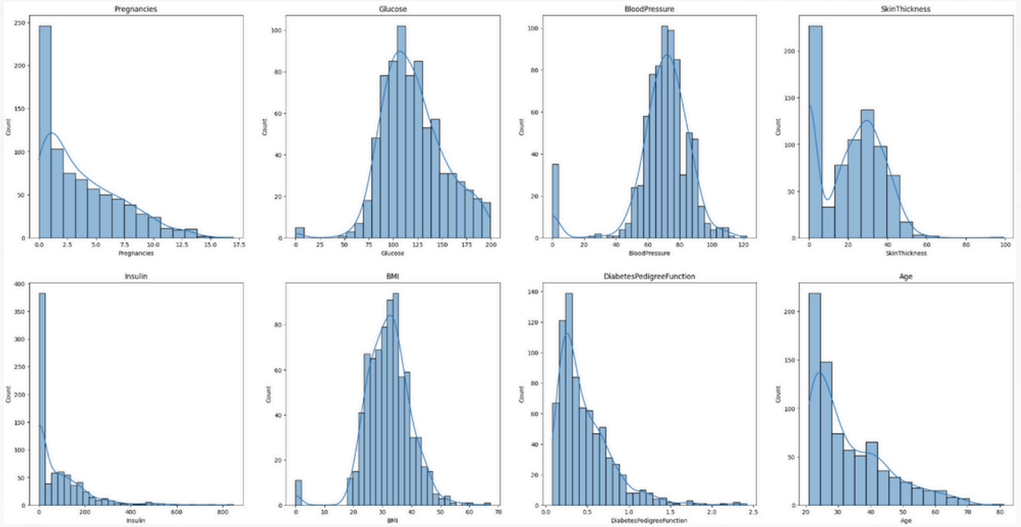
	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...

```
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

PREPROCESSING AND EDA



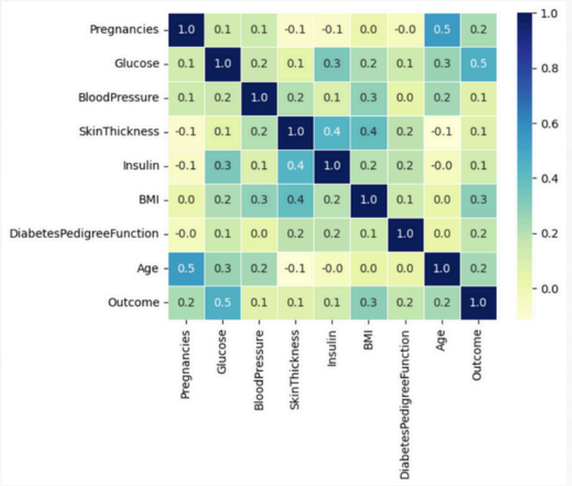
- The data didn’t contain null or duplicated values but we found that the data contained many outliers so we had to remove the outliers before we applied any model.



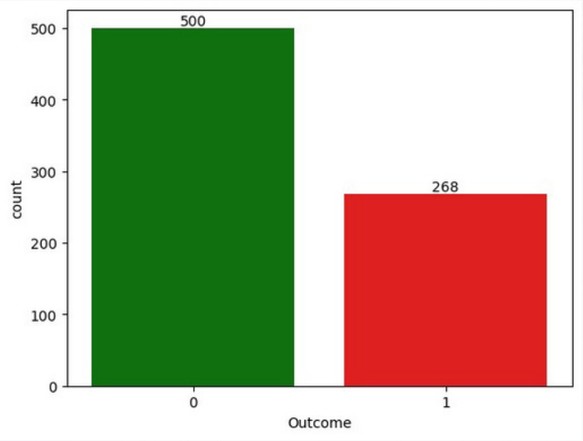
- The dataset indicates significant risk factors for diabetes, with many patients displaying elevated glucose levels and BMI. The skewed distributions for insulin and skin thickness suggest potential data quality issues, which should be considered in the analysis.

PREPROCESSING AND EDA

- The Heatmap describes the relations between the features to indicate which features are related to each other or doesn’t relate.

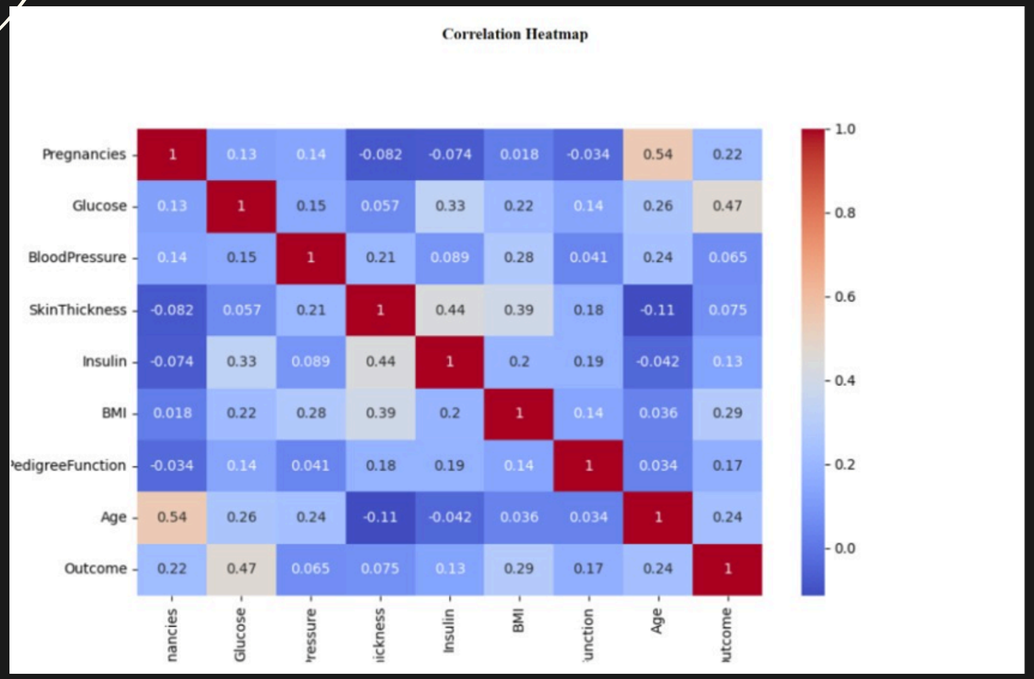


- this count plot shows the number of persons who have diabetes described as ‘1’ and and the other who doesn’t have diabetes described as ‘0’



DASHBOARD

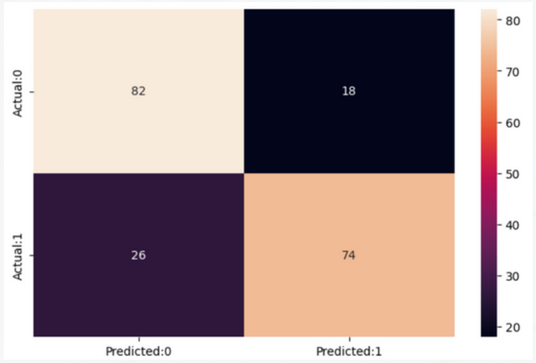

Diabetes Dataset Insights Dashboard
Distribution of Glucose
Distribution of BMI
Distribution of Age
count
Glucose
BMI
Age


Correlation Heatmap
Pregnancies 1 0.13 0.14 -0.082 -0.074 0.018 -0.034 0.54 0.22
Glucose 0.13 1 0.15 0.057 0.33 0.22 0.14 0.26 0.47
BloodPressure 0.14 0.15 1 0.21 0.089 0.28 0.041 0.24 0.065
SkinThickness -0.082 0.057 0.21 1 0.44 0.39 0.18 -0.11 0.075
Insulin -0.074 0.33 0.089 0.44 1 0.2 0.19 -0.042 0.13
BMI 0.018 0.22 0.28 0.39 0.2 1 0.14 0.036 0.29
edigreeFunction -0.034 0.14 0.041 0.18 0.19 0.14 1 0.034 0.17
Age 0.54 0.26 0.24 -0.11 -0.042 0.036 0.034 1 0.24
Outcome 0.22 0.47 0.065 0.075 0.13 0.29 0.17 0.24 1
1.0
0.8
0.6
0.4
0.2
0.0

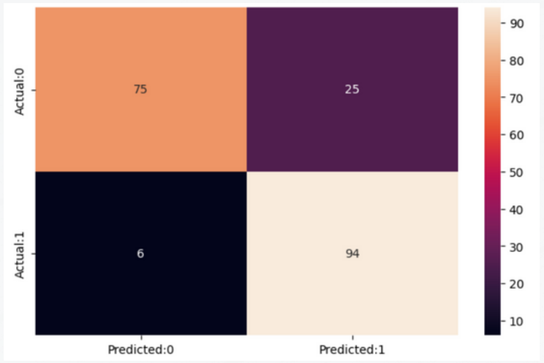

Feature Importance
Feature Importance for Diabetes Prediction
Importance
Breakdown by Outcome
Glucose by Outcome
BMI by Outcome
Age by Outcome
Glucose Levels by Outcome
Outcome

CHOOSING THE RIGHT MODEL

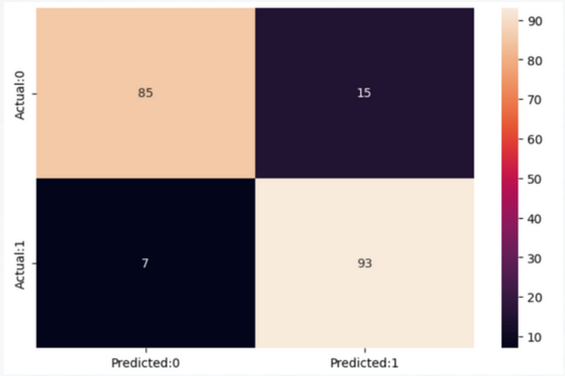
LOGISTIC REGRESSION



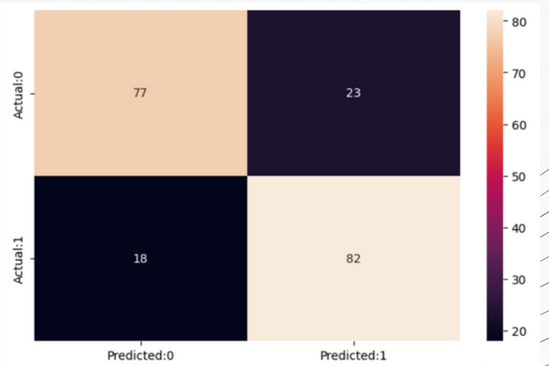
KNN



RANDOM FOREST



SVM



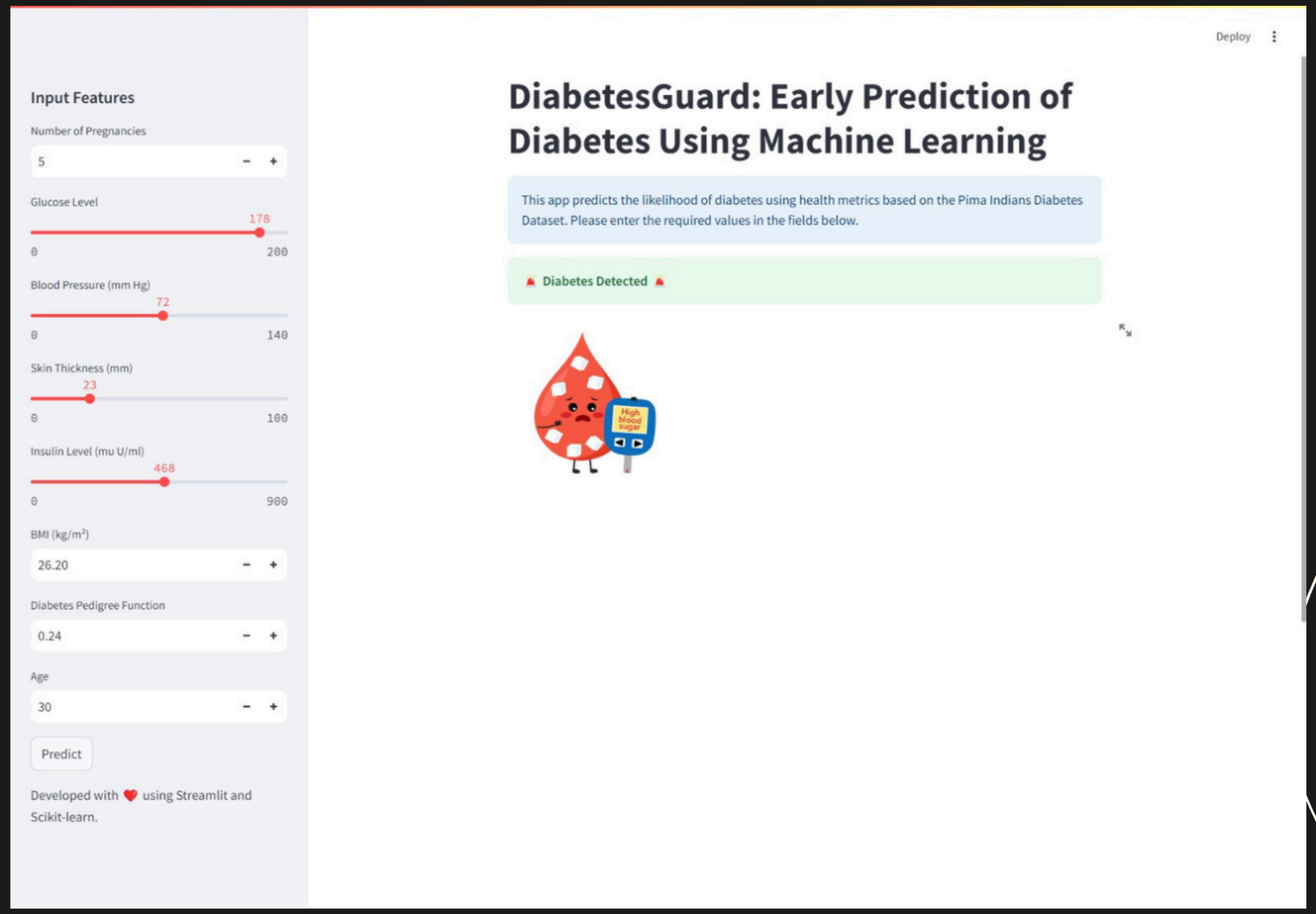
BEST MODEL

89%

After Applying the four models on the dataset and comparing their accuracies, it appears that the best model for our data is the RandomForest with 81%

	Estimators	Accuracy
0	RandomForest	0.890
2	K-Nearest Neighbor	0.845
3	Support Vector Machine	0.795
1	Logistic Regression	0.780

DEPLOYMENT



We used streamlit for deployment and this is the view of the API that appears for the user to check if he has diabetes or not.

THANK'S FOR WATCHING