Diabetes
Awareness

NOV 2021

PREDICTING DIABETES

Presented By:

Sultanah Aldossari
Rawan Alharbi

CONTENT OUTLINE

- Introduction
- Dataset
- Expolarity Data Analysis
- Data Pre-processing
- Model Building and Evaluation
- Conclusion

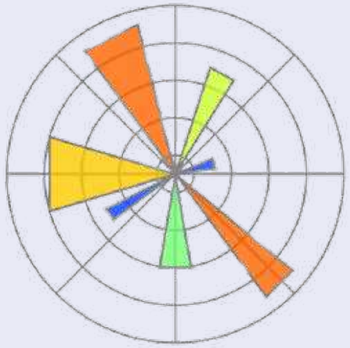
PROBLEM STATEMENT

- Diabetes is a common chronic disease and poses a great threat to human health.
- It can lead to chronic damage and dysfunction of various tissues. Therefore, The earlier diagnosis is obtained, the much easier we can control diabetes.
- Our aim is to predict diabetes using Classification techniques.

DATASET

Pregnancies	Glucose	Blood Pressure
Skin Thickness	Insulin	BMI
Diabetes Pedigree Function	Age	Outcome

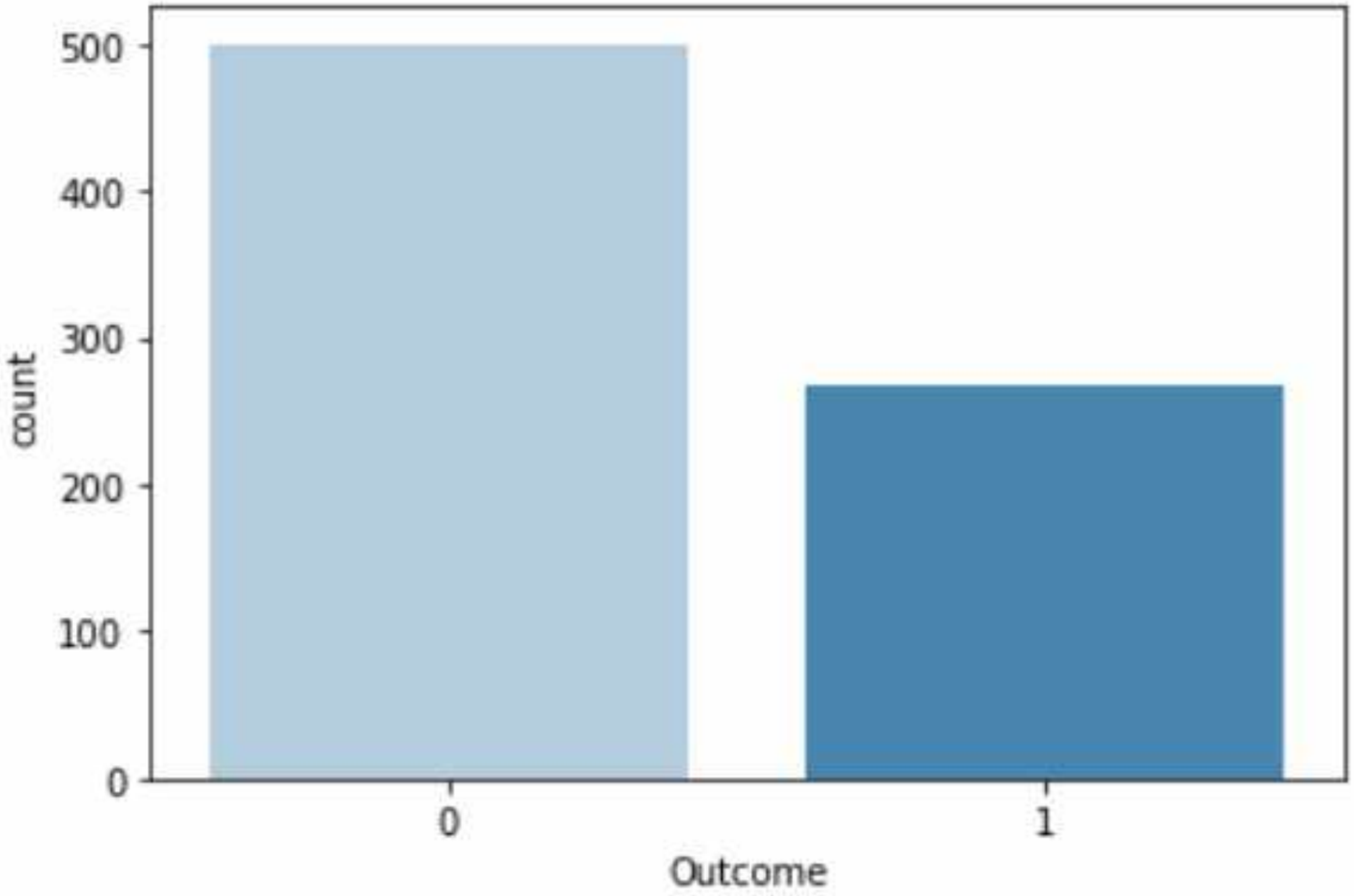
TOOLS



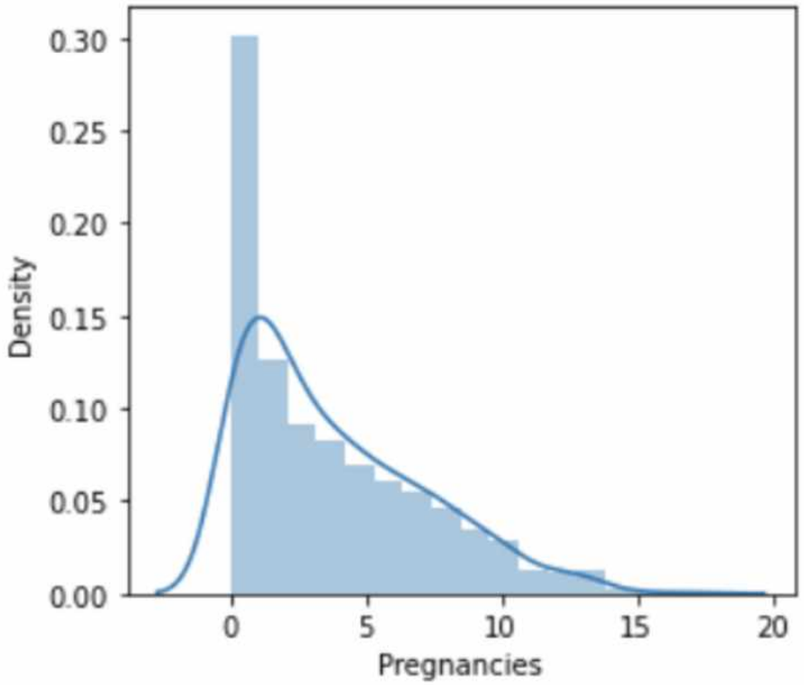
matplotlib

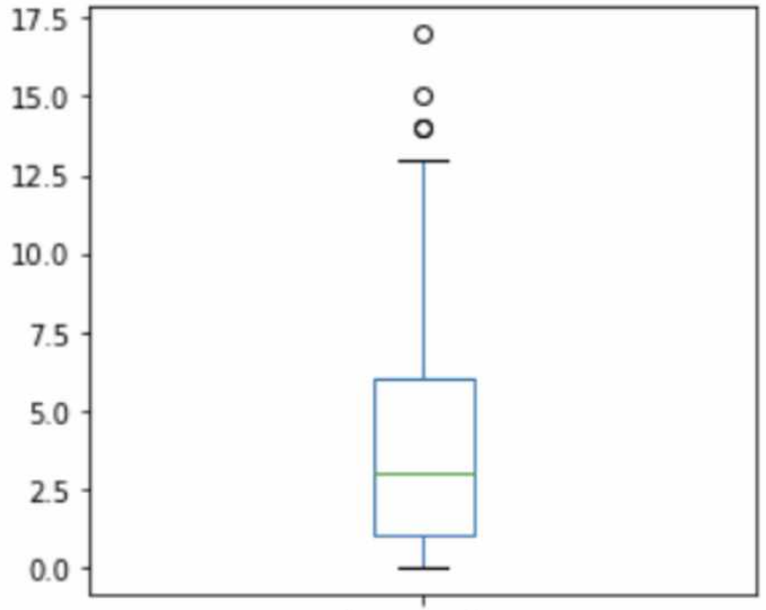
Pandas

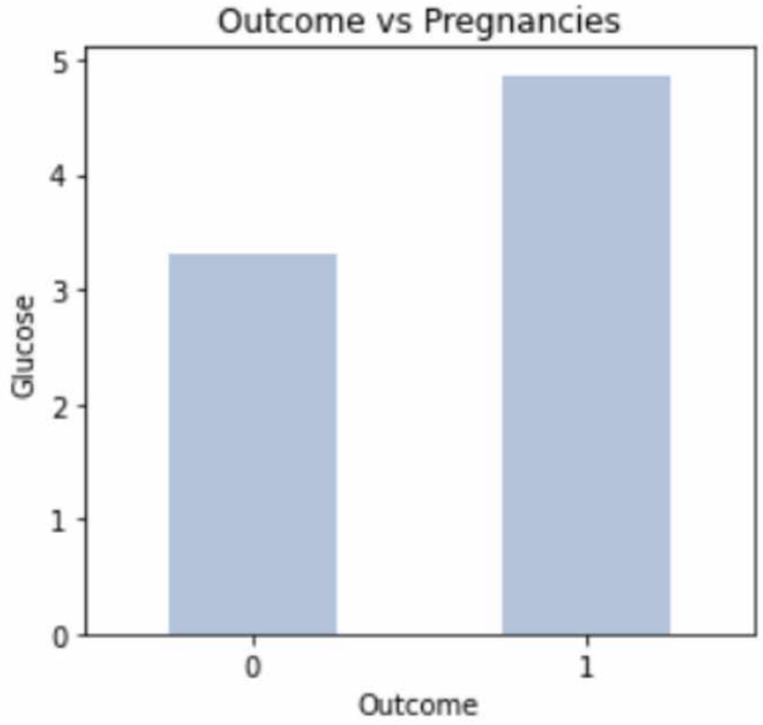
EXPOLARITY DATA ANALYSIS



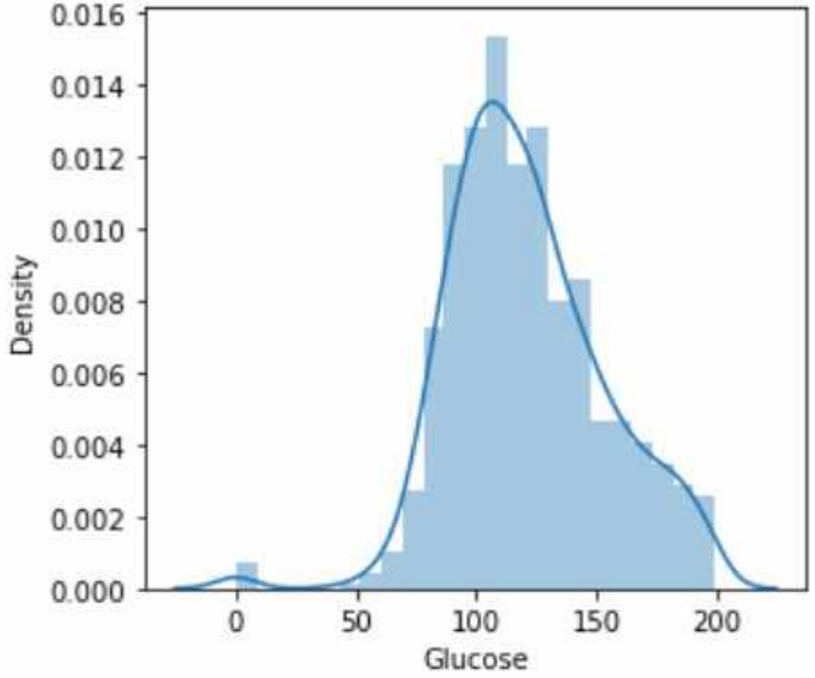
EXPOLARITY DATA ANALYSIS

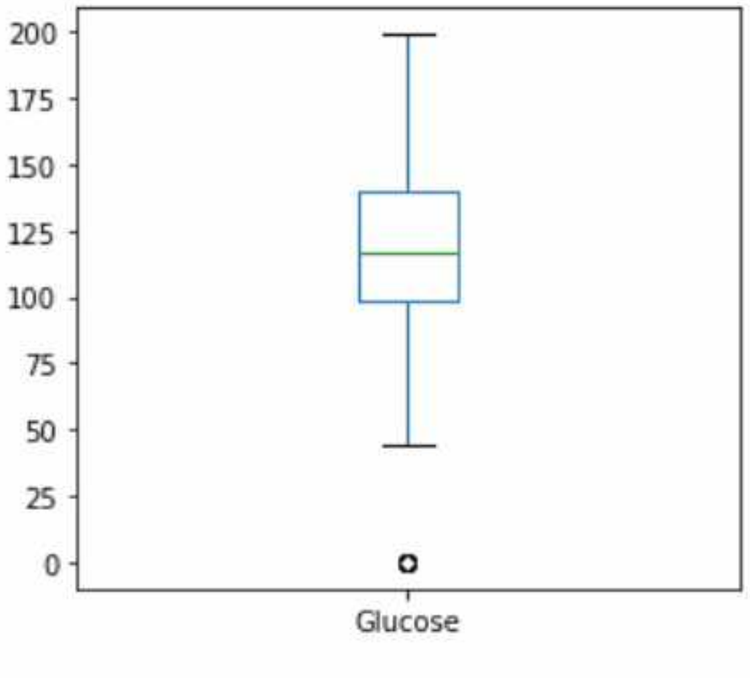


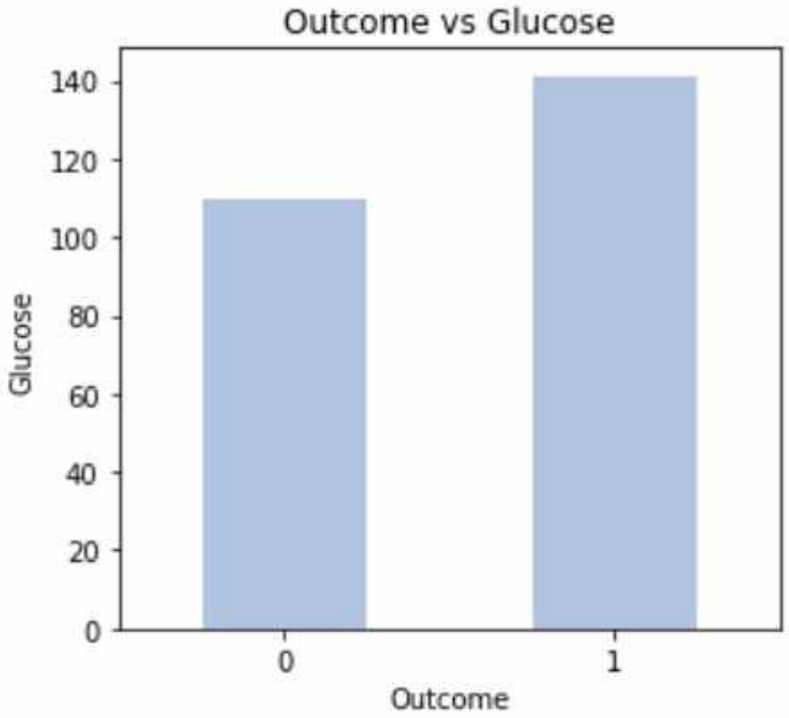




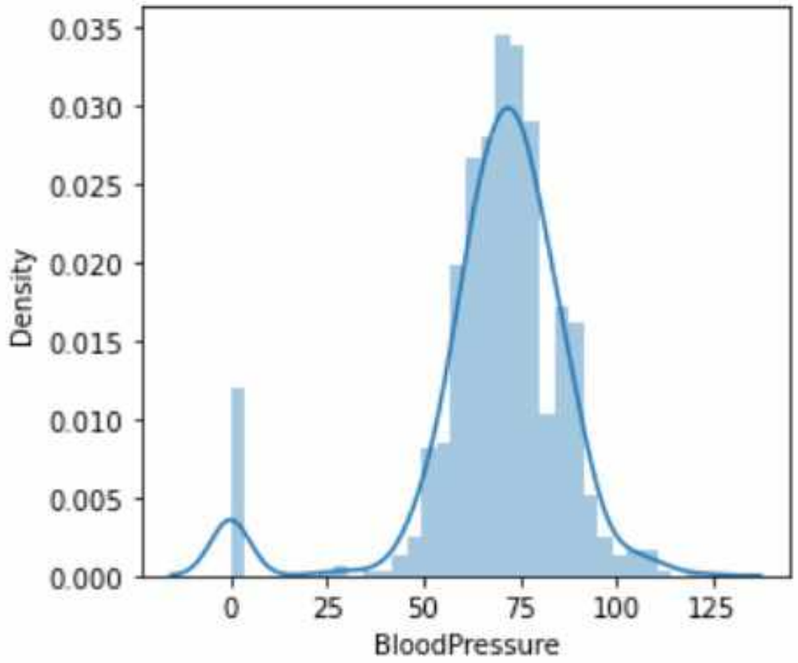
EXPOLARITY DATA ANALYSIS

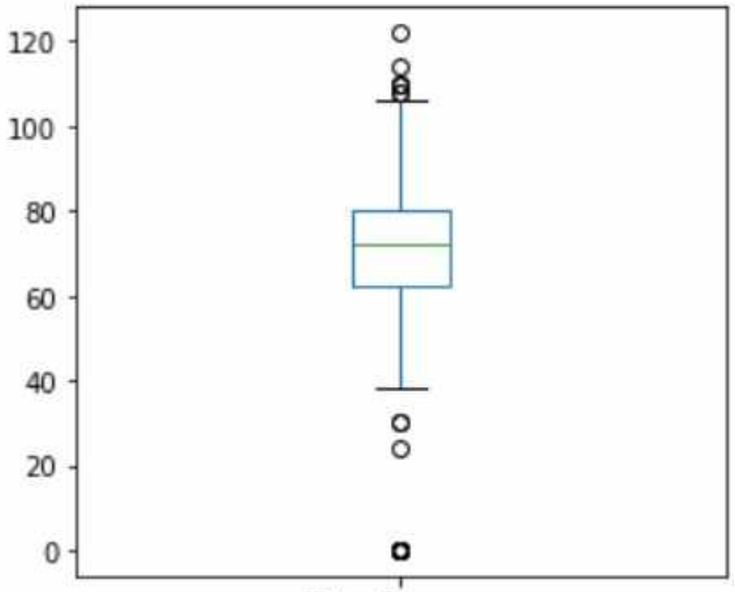


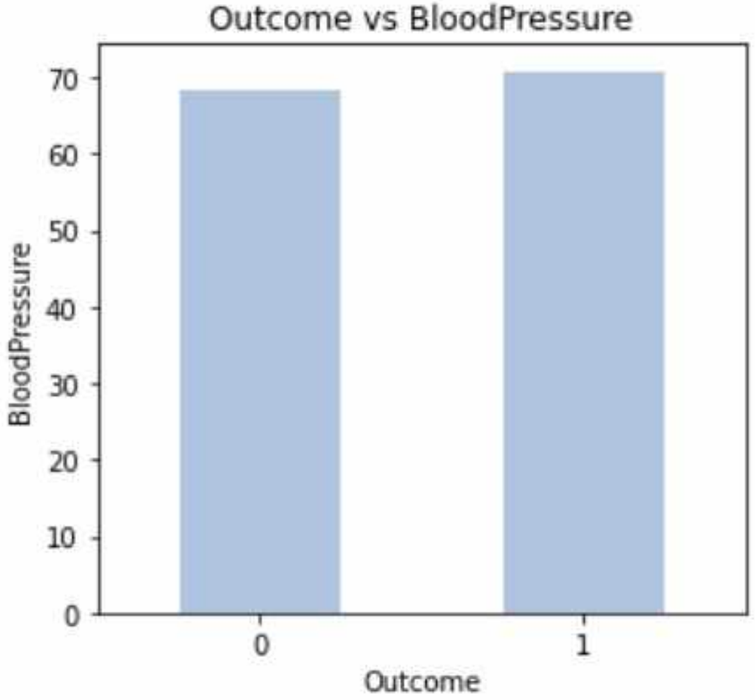




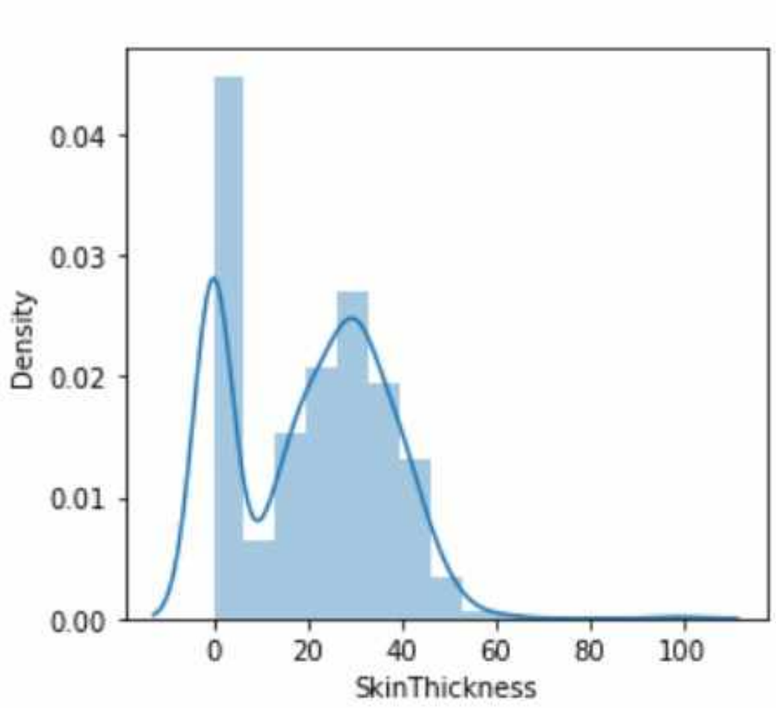
EXPOLARITY DATA ANALYSIS

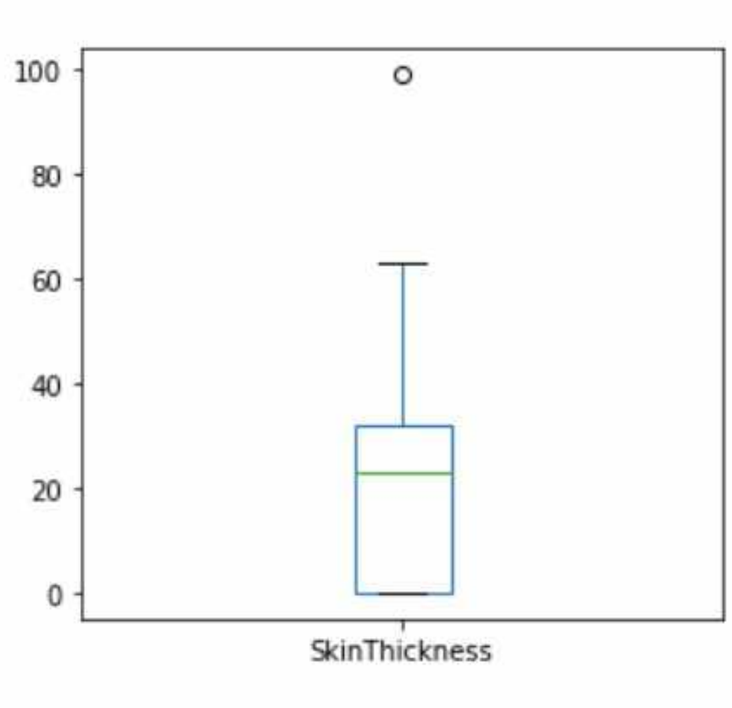


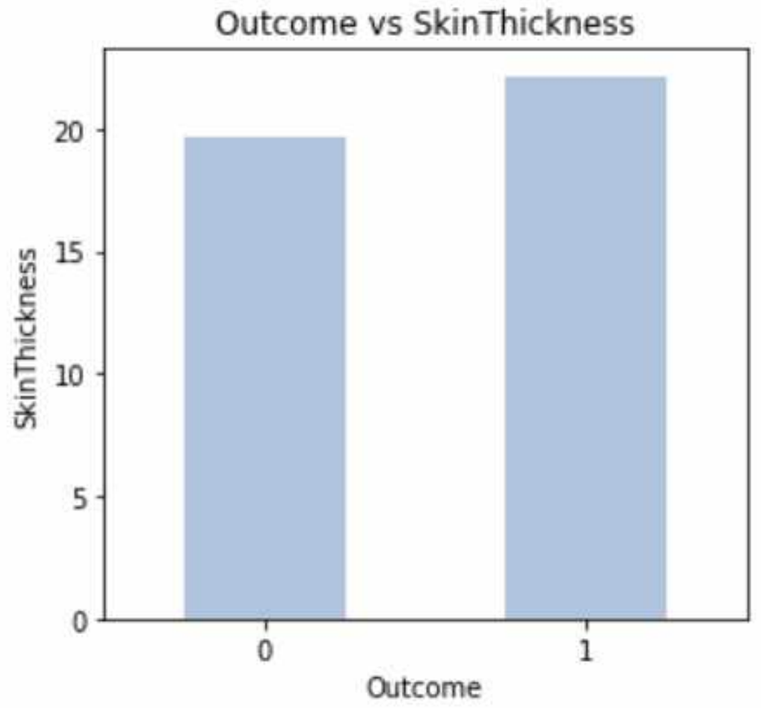




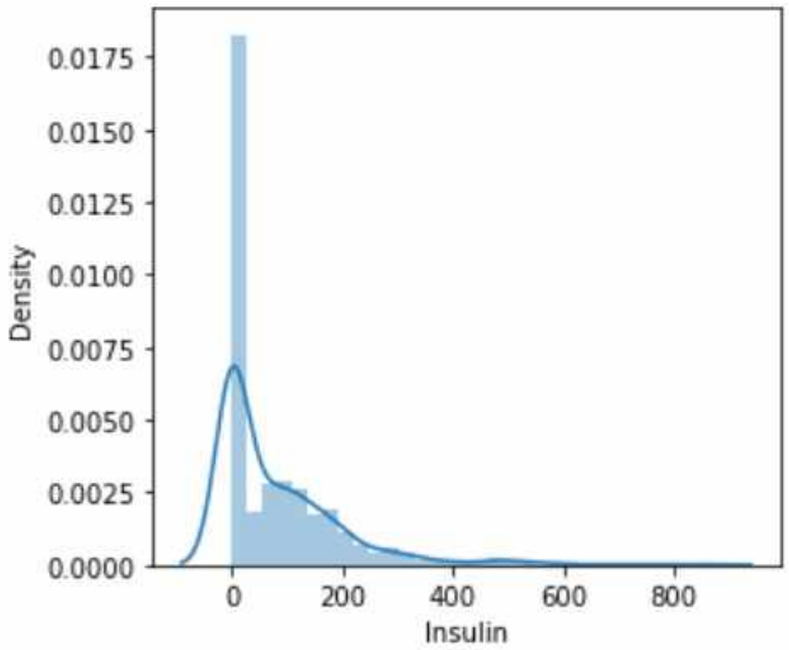
EXPOLARITY DATA ANALYSIS

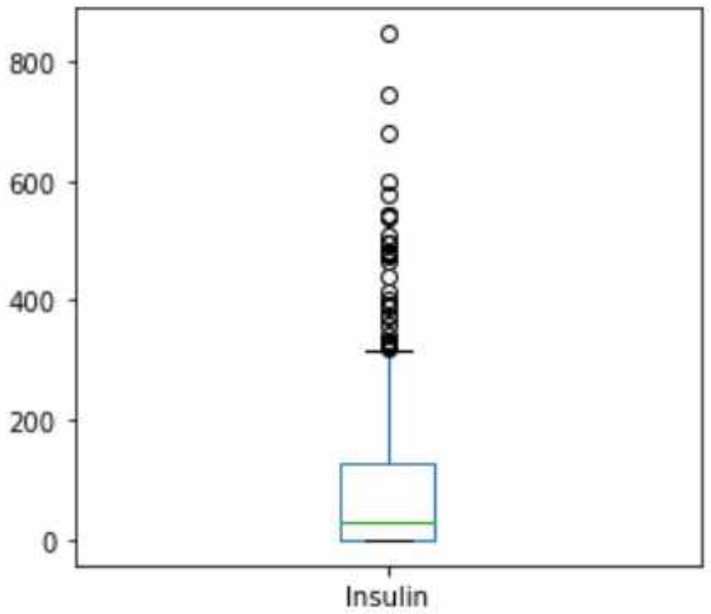


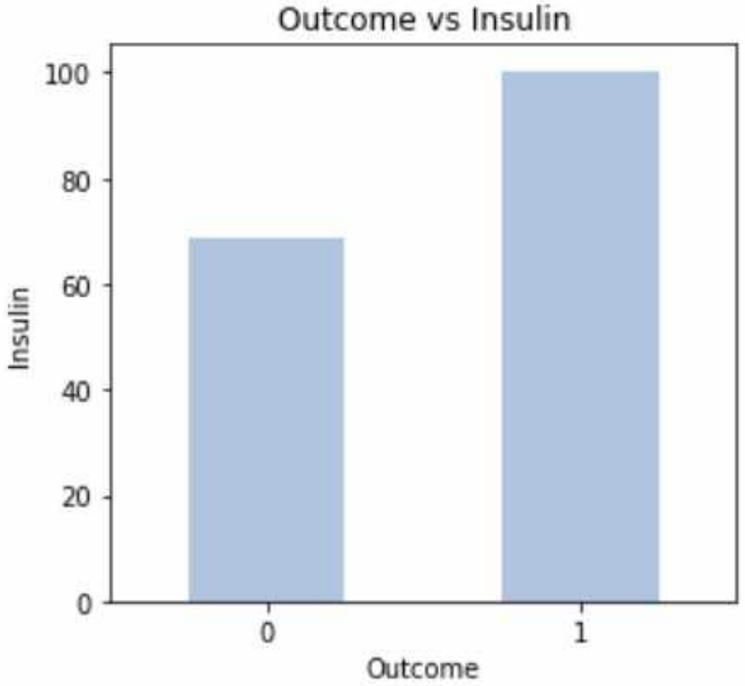




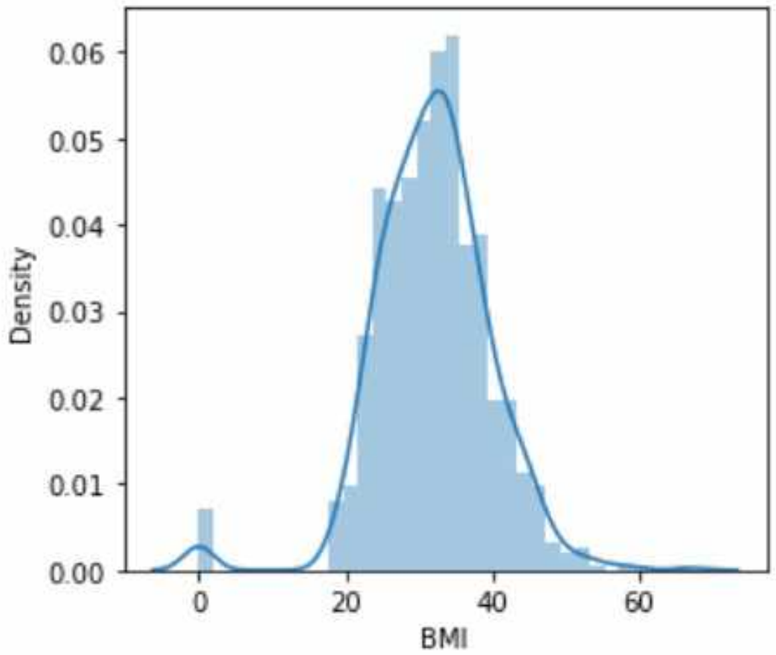
EXPOLARITY DATA ANALYSIS

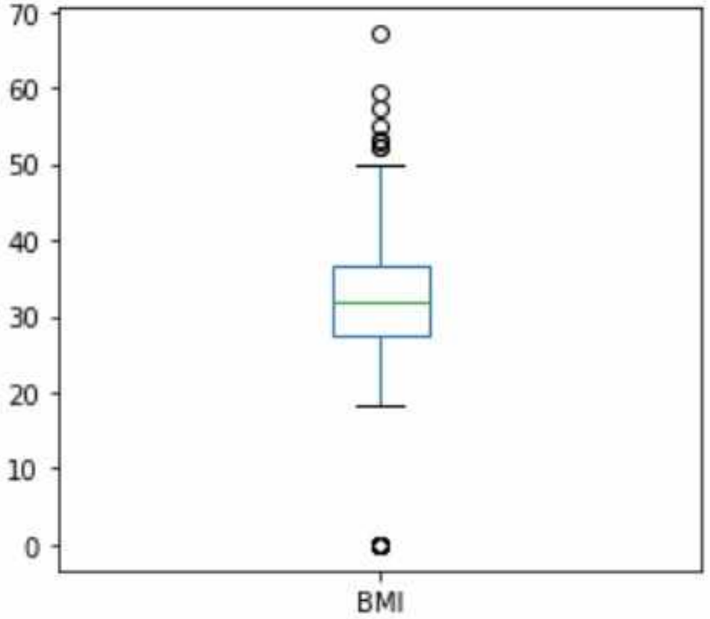


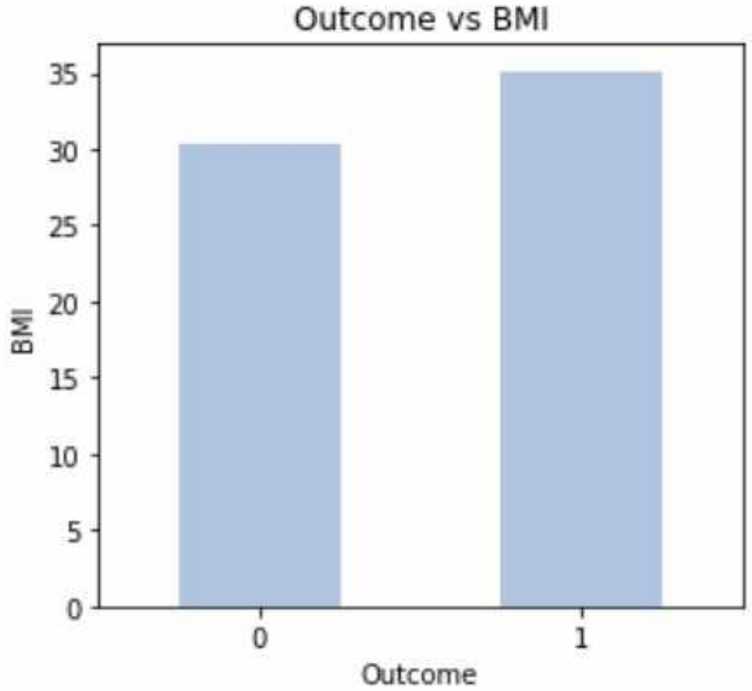




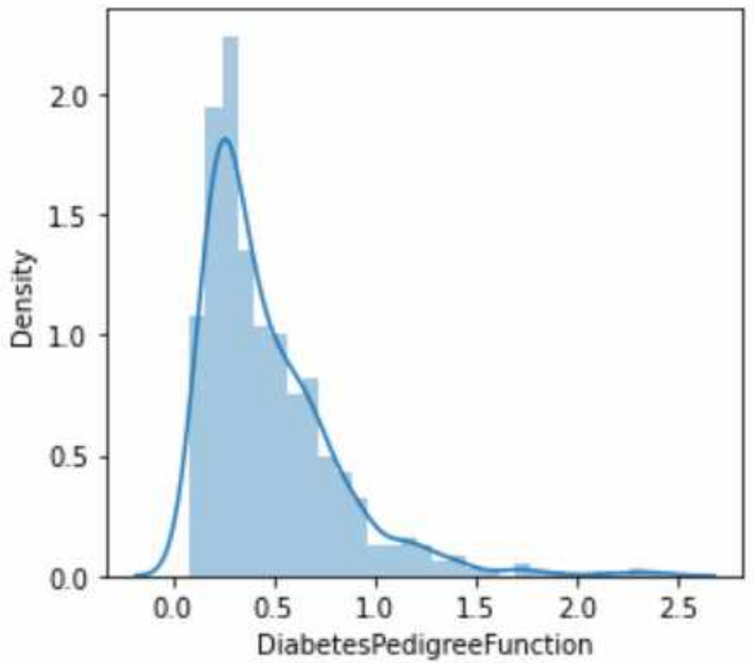
EXPOLARITY DATA ANALYSIS

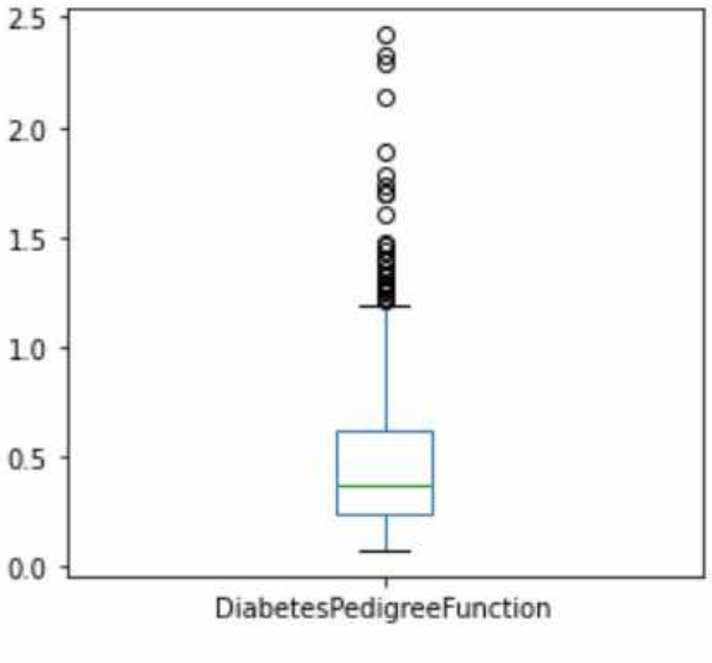


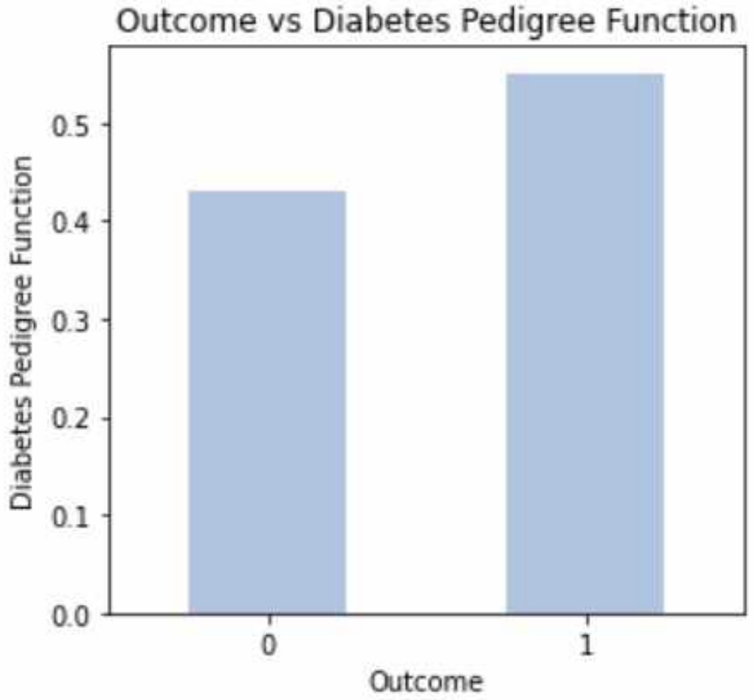




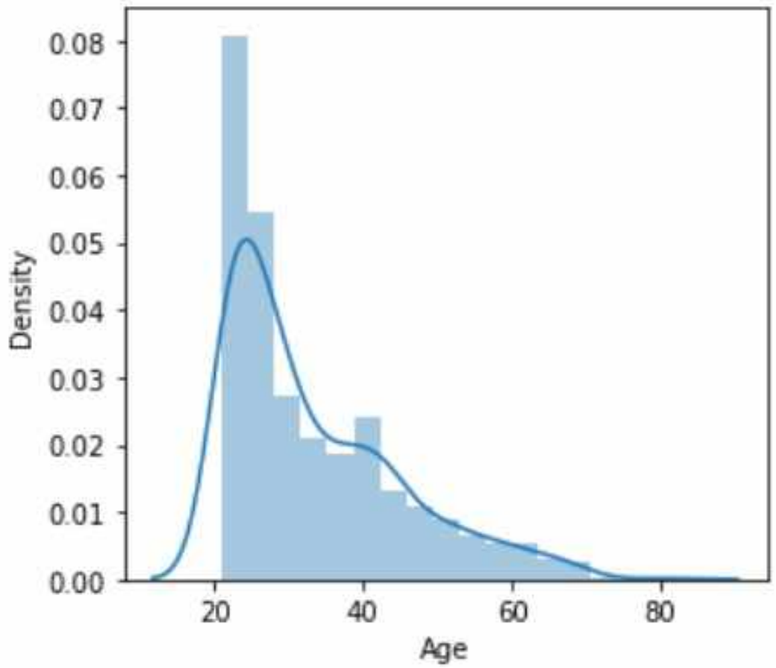
EXPOLARITY DATA ANALYSIS

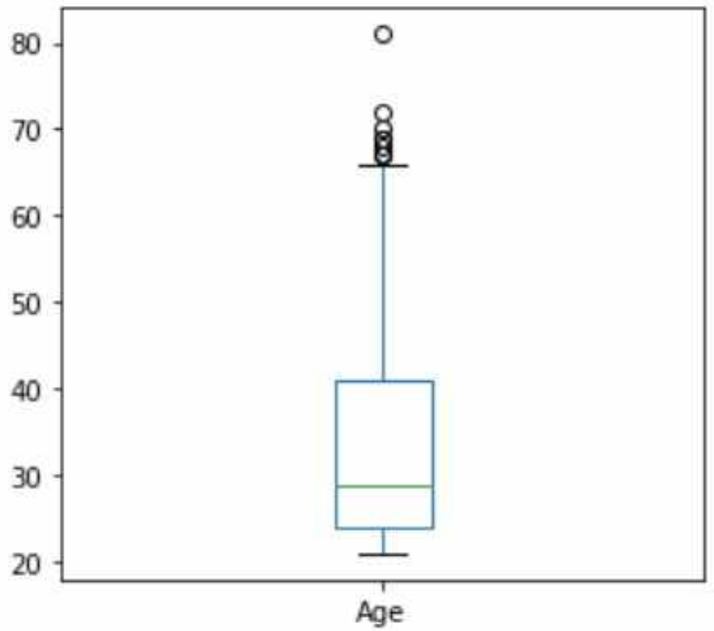


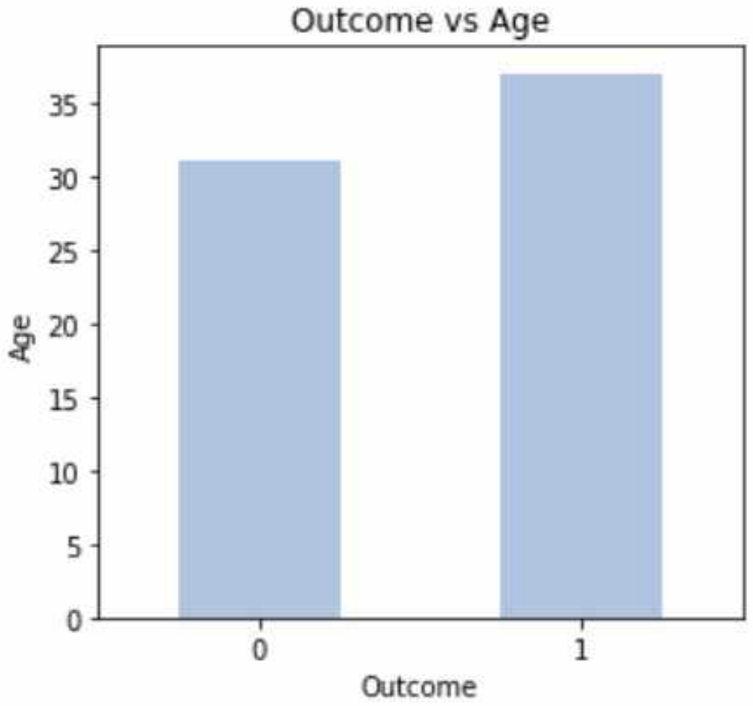




EXPOLARITY DATA ANALYSIS







CONT: EDA

Relationship between target and predictors

DATA PRE-PROCESSING

Outliers Handling

Handle Zeros

Scaling Data

Split into Train and Test

MODEL EXPERIMENTS:

	Algorithm Used	Accuracy	Recall	Precision	F1-Score	AUC
0	K-Nearest Neighbours	0.785088	0.682353	0.725000	0.703030	0.764253
1	K-Nearest Neighbours Tuned	0.789474	0.635294	0.760563	0.692308	0.758206
2	Logistic Regression	0.824561	0.752941	0.771084	0.761905	0.810037
3	Logistic Regression with Ridge Regression	0.807018	0.811765	0.711340	0.758242	0.807980
4	Logistic Regression with Lasso	0.811404	0.811765	0.718750	0.762431	0.811477
5	Random Forest	0.785088	0.647059	0.743243	0.691824	0.757096
6	Random Forest Tuned	0.780702	0.776471	0.680412	0.725275	0.779844
7	Decision Tree	0.710526	0.682353	0.597938	0.637363	0.704813
8	Decision Tree Tuned	0.750000	0.600000	0.689189	0.641509	0.719580
9	Gradient Boosting Classifier	0.767544	0.764706	0.663265	0.710383	0.766968
10	Ada Boost Classifier	0.754386	0.705882	0.659341	0.681818	0.744550
11	Ada Boost Classifier Tuned	0.736842	0.835294	0.606838	0.702970	0.756808
12	Bagging for Decision Tree	0.824561	0.752941	0.771084	0.761905	0.810037
13	Support Vector Classifier	0.802632	0.647059	0.785714	0.709677	0.771082
14	Support Vector Classifier Tuned	0.807018	0.788235	0.720430	0.752809	0.803209
15	Bagging for Support Vector Classifier	0.811404	0.800000	0.723404	0.759777	0.809091

MODEL SELECTION:

	Algorithm Used	Accuracy	Recall	Precision	F1-Score	AUC
0	K-Nearest Neighbours	0.785088	0.682353	0.725000	0.703030	0.764253
1	K-Nearest Neighbours Tuned	0.789474	0.635294	0.760563	0.692308	0.758206
2	Logistic Regression	0.824561	0.752941	0.771084	0.761905	0.810037
3	Logistic Regression with Ridge Regression	0.807018	0.811765	0.711340	0.758242	0.807980
4	Logistic Regression with Lasso	0.811404	0.811765	0.718750	0.762431	0.811477
5	Random Forest	0.785088	0.647059	0.743243	0.691824	0.757096
6	Random Forest Tuned	0.780702	0.776471	0.680412	0.725275	0.779844
7	Decision Tree	0.710526	0.682353	0.597938	0.637363	0.704813
8	Decision Tree Tuned	0.750000	0.600000	0.689189	0.641509	0.719580
9	Gradient Boosting Classifier	0.767544	0.764706	0.663265	0.710383	0.766968
10	Ada Boost Classifier	0.754386	0.705882	0.659341	0.681818	0.744550
11	Ada Boost Classifier Tuned	0.736842	0.835294	0.606838	0.702970	0.756808
12	Bagging for Decision Tree	0.824561	0.752941	0.771084	0.761905	0.810037
13	Support Vector Classifier	0.802632	0.647059	0.785714	0.709677	0.771082
14	Support Vector Classifier Tuned	0.807018	0.788235	0.720430	0.752809	0.803209
15	Bagging for Support Vector Classifier	0.811404	0.800000	0.723404	0.759777	0.809091

BEST PREFORMED ALGORITHM:

Ada Boost

Recall = 84%
AUC = 76%

Logistic Regression L2

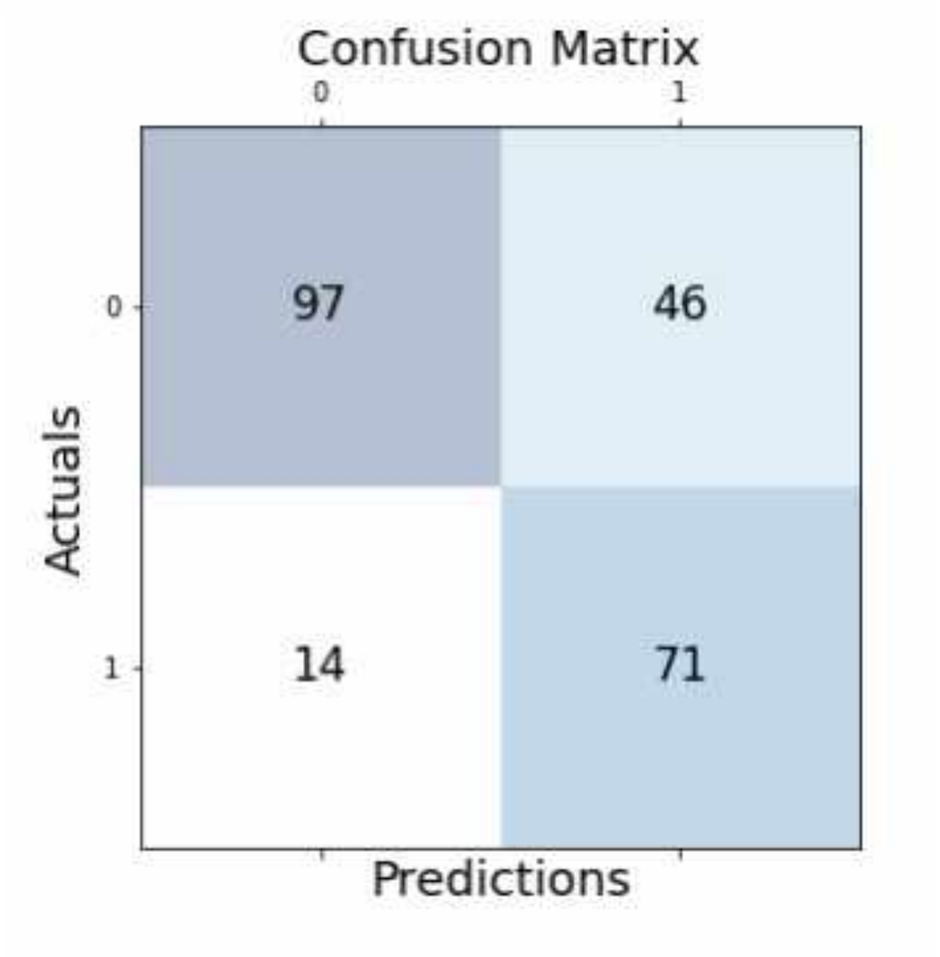
Recall = 81%
AUC = 80%

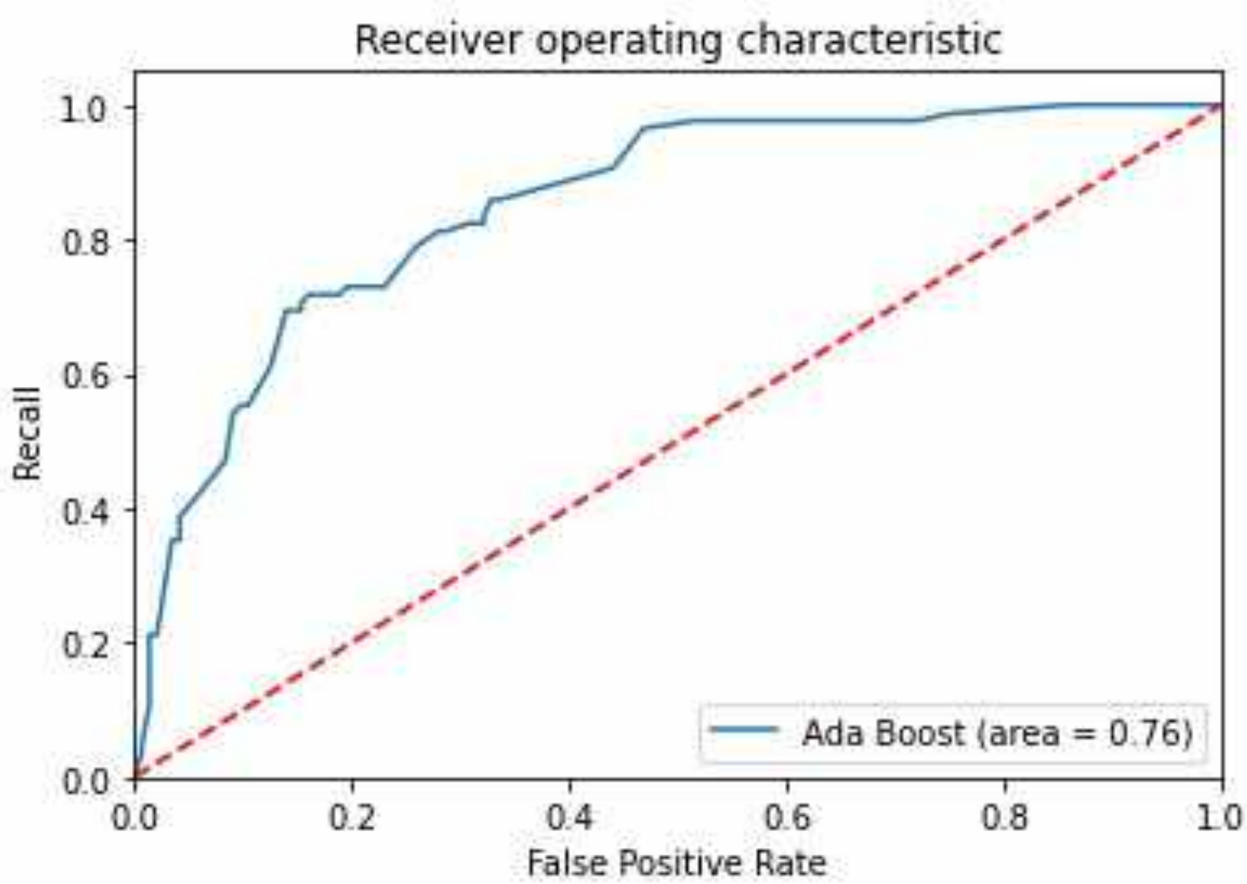
Logistic Regression L1

Recall = 81%
AUC = 81%

ADA BOOST CLASSIFIER

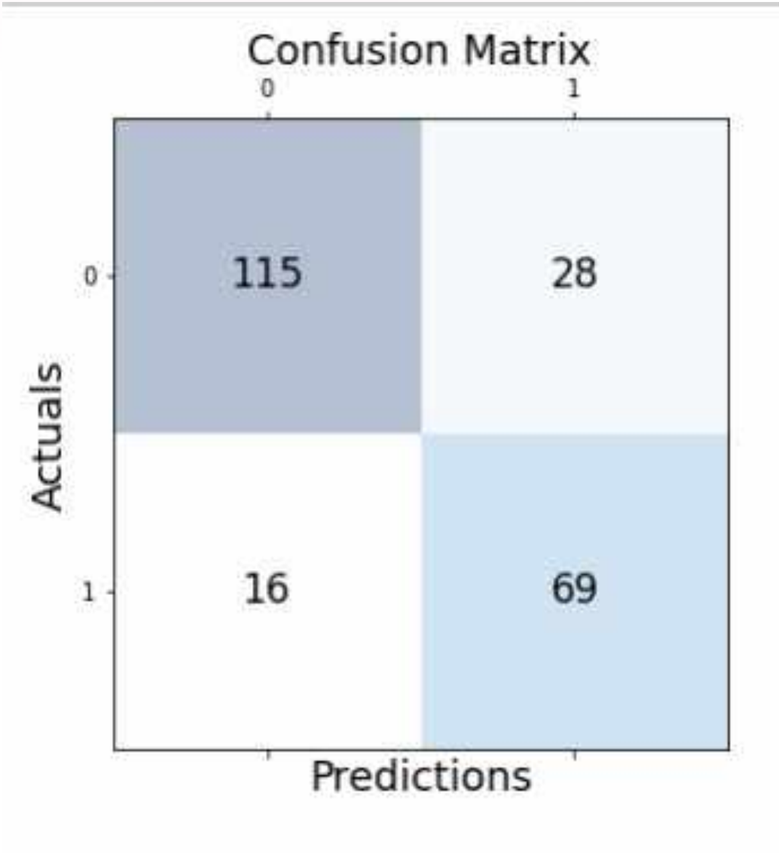
Confusion matrix & ROC Curve Plot

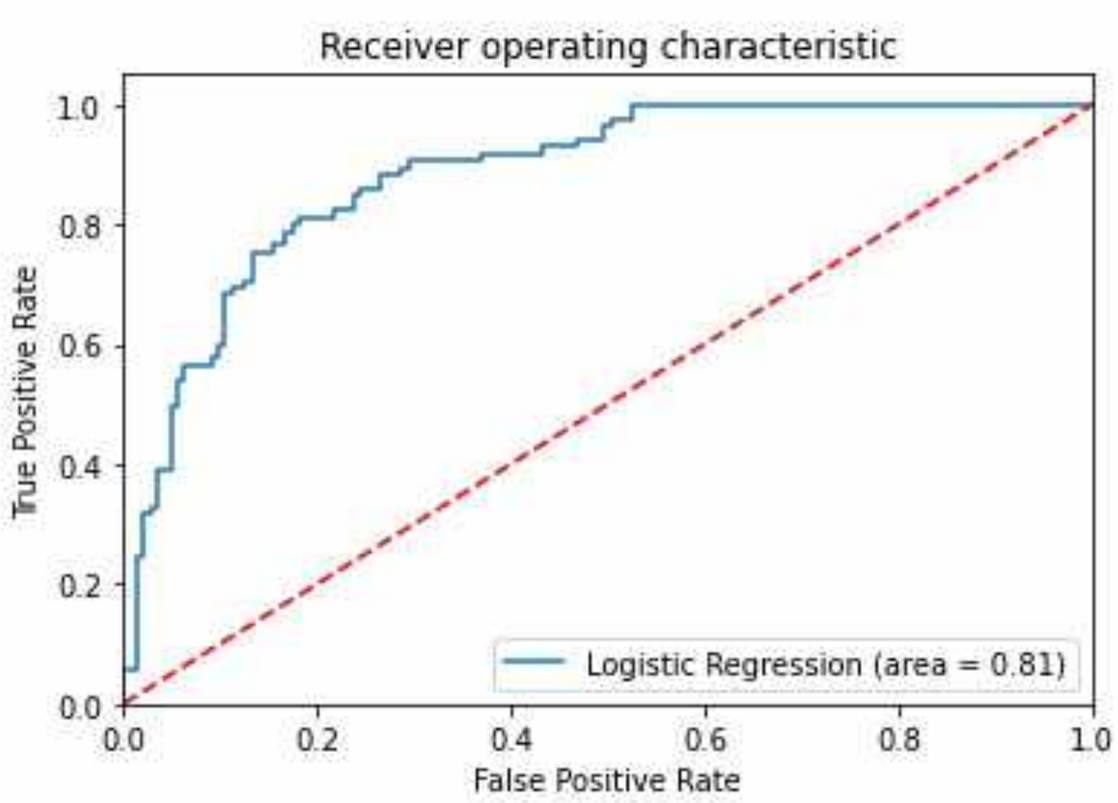




LOGISTIC REGRESSION L2:

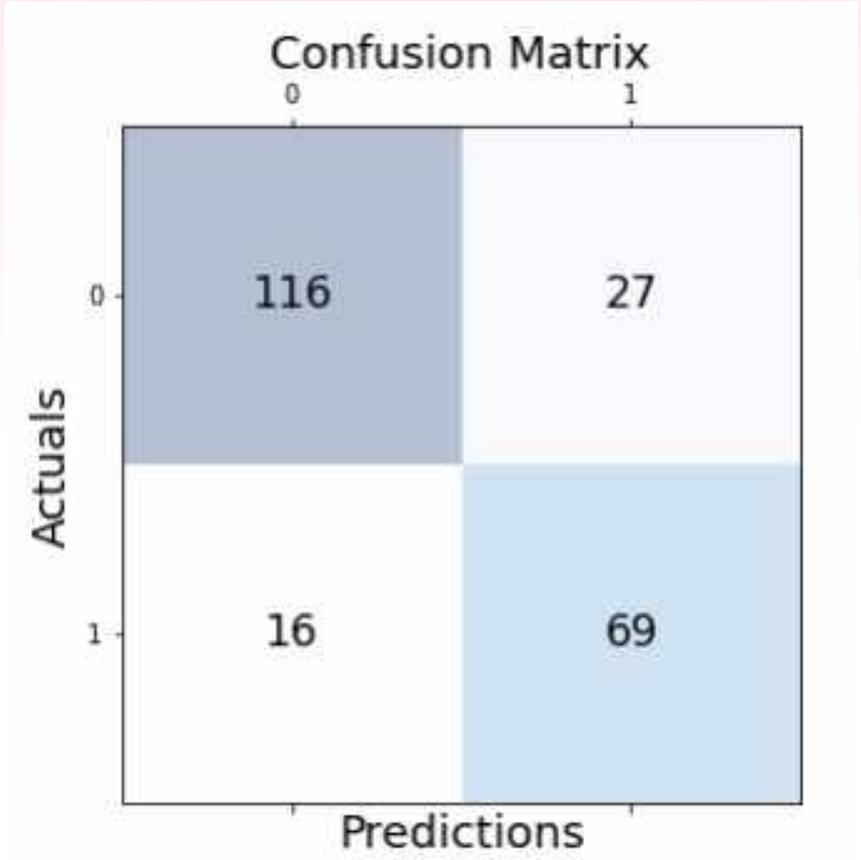
Confusion matrix & ROC Curve Plot

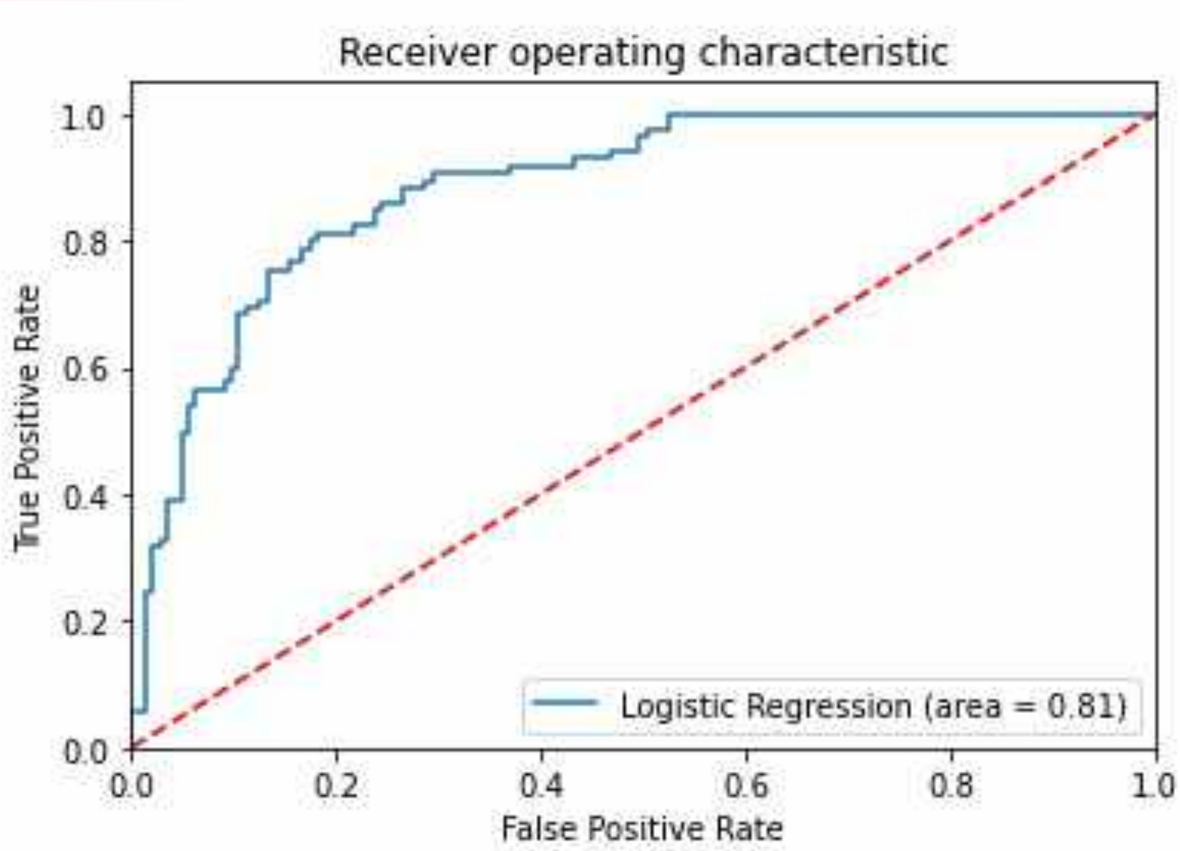




LOGISTIC REGRESSION L1

Confusion matrix & ROC Curve Plot





DIABETES PREDICTION APP

FUTURE WORK

Collect more data about diabetes and analyze them

For better predictions

Build a system that helps people in getting an initial information

CONCLUSION

DIABETES

THANK YOU FOR LISTENING!