Assignment 4.1

Project-1: Final Presentation/ Milestone – 3

Sucharitha Puppala

Data Science, Bellevue University

DSC680-T301 Applied Data Science (2233-1)

Prof. Catherine Williams

January 8, 2023.

Project-1

Topic:

Prediction of Diabetes in female patients, based on certain diagnostic measurements from the dataset.

Business Problem:

Diabetes is one of the biggest health problems that affect millions of people across the world. Uncontrolled diabetes can increase the risk of heart attack, cancer, kidney damage, blindness, and other illnesses. Researchers are motivated to create a Machine Learning methodology that can predict diabetes in the future. Exploiting Machine Learning Algorithms (MLA) is essential if healthcare professionals are able to identify diseases more effectively. In order to improve the medical diagnosis of diabetes this research explored and contrasts various MLA that can identify diabetes risk early.

The project mainly focuses on the Prediction of diabetes in female patients.

The topic raises the following questions:

1. Are pregnant women more prone to diabetics or not?
2. What are the levels of glucose that cause diabetics?
3. What is the BMI level in people having diabetics?

Datasets:

The dataset used for this project is collected from Kaggle.com.

The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The data in the dataset is related to all female patients.

Project-1

Information about dataset attributes -

1. Pregnancies: To express the Number of pregnancies
2. Glucose: To express the Glucose level in blood
3. BloodPressure: To express the Blood pressure measurement
4. SkinThickness: To express the thickness of the skin
5. Insulin: To express the Insulin level in blood
6. BMI: To express the Body mass index
7. DiabetesPedigreeFunction: To express the Diabetes percentage
8. Age: To express the age
9. Outcome: To express the final result 1 is Yes and 0 is No

Methods:

As the problem is a classification problem I would like to use the following four models Logistic Regression model, K- Nearest Neighbor Classifier, Decision Tree Classifier, Support Vector Machine (SVM). All the models are created, trained, and fit with the test (20%) and train (80%) datasets respectively. For the model evaluation, evaluation metrics Accuracy, Precision, Recall, and F1 Score are calculated with the test data. Confusion Matrices that summarizes the performance of the model build are plotted for each model respectively.

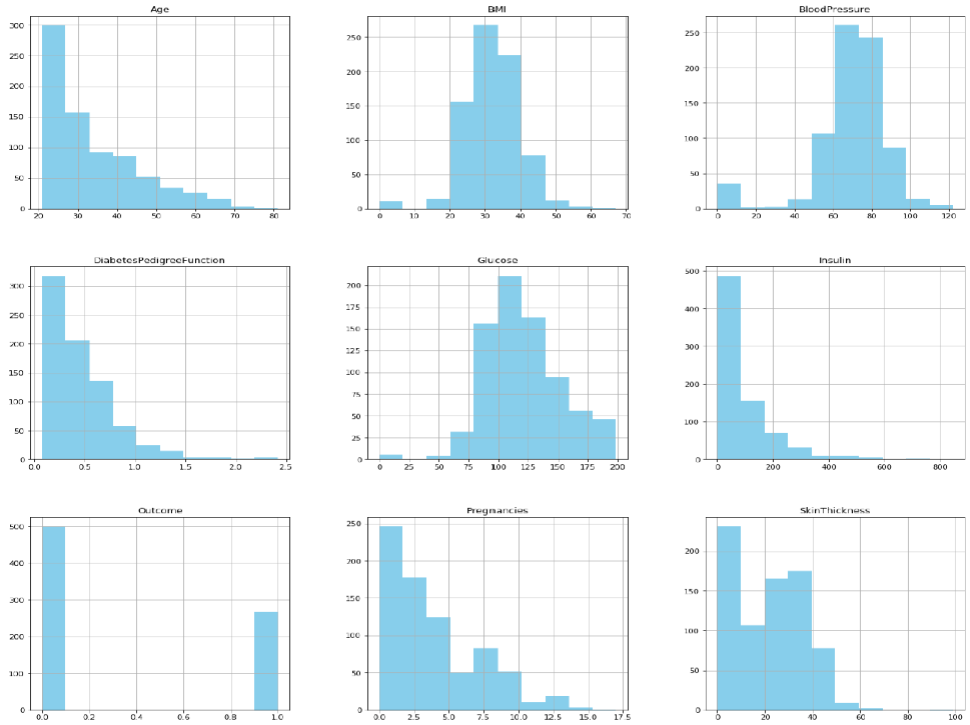
Analysis:

The analysis of the project is done using the CRISP DM methodology. The data file diabetes.csv is loaded for analysis. The shape of the dataset is obtained, and the dataset is of size 768*9 (i.e. 768 rows and 9 columns). Initially the dataset is checked for the null values if any present in the dataset, when observed there are no null values in the dataset selected, then check for duplicate values is done, when observed

there are no duplicate values in the dataset. The unique values present in the dataset are extracted and the dataset data types are obtained, all the data types are of integer data types. In the Exploratory Data Analysis (EDA) Histograms, bar plots and Box plots are used for the analysis.

The following figure shows the histogram plot of all the numerical data types present in the data set.

Figure 1: Histograms of numerical features of the dataset.



From the above histogram plot we see that some features are skewed.

Count plots are plotted to know the distribution of the target variable “Outcome” with the features available from the dataset. Appendix –A gives the count plots of all the features with the target variable.

When observed the count plots it is observed that the female patients who tested positive for the diabetes have high values of BMI, BloodPressure, DibetesePedigreeFunction, Glucose, Insulin,Skinthickness. To know the correlation of the variables in the dataset the correlation heat map

and pair plots are plotted. Appendix- B gives the heat map and the pair plot of the dataset. From the correlation heat map we can say that the features Glucose, BloodPressure, Age are having good correlation but with very low value of correlation coefficient. Boxplots are plotted to identify the outliers. Appendix – C provides the Box plots. When observed the box plot we can see that the features Glucose, Blood Pressure, Skin Thickness, Insulin are having some outliers.

The outliers are removed by IQR (Inter Quartile Range Method) and zeros are identified in the Glucose and Skinthickness columns and are replaced with the median values respectively. The data set shape is obtained and the correlation matric is being plotted again to understand the correlation of the variables of the dataset.

The dataset is split into test and train datasets initially the dataset is split into features(X) with the target variable dropped from the dataset and target(y) variable by considering only the 'Outcome' column from the dataset. The data set is to be split into training and test datasets using 'Outcome' as the target variable. Here the dataset is split into 80% training dataset and 20% test dataset, taking the test size as 0.2 and the random_state as 42. X_train, X_test, y_train, y_ test are created using train_test_split()

The train and test data sets shapes are obtained for better understanding the split of the dataset.

Principal Component Analysis (PCA) is applied and second set of train and test data sets are created, with this the features have been reduced from 8 to 7.

The features train and test datasets are standardized using StandardScaler. The X_train dataset is transformed and fit to the standard scaler and the X_test dataset is just transformed but not fit.

With this the data is ready for model building.

All the 4 models i.e. Logistic Regression model, K- Nearest Neighbor Classifier, Decision Tree Classifier, Support Vector Machine (SVM) is built. All the models are created, trained, and fit with the train

datasets and PCA applied test and train datasets respectively. For the selection of the best model that fits our data, evaluation metrics Accuracy, Precision, Recall and F1 Score are calculated with the test data and the PCA applied test data. Confusion Matrices that summarizes the performance of the model build are plotted for each model respectively. After observing the evaluation metrics calculated for each model with the test dataset and the PCA applied test and train datasets, the best model that fits the dataset, is selected.

When observed the results we can say that the model Logistic Regression, performed well with an accuracy of about 77%. Support Vector Machine performed better with the dataset having an accuracy of 75%. The KNN classifier performed with low accuracy score of 0.66 but the accuracy of the trained dataset is 0.80 indicating there is some over fit with the trained dataset. The Decision Tree Classifier model also shows there is over fit with the trained data set. The table for the results can be seen in Appendix –D, Table 1.

When observed the results of the PCA applied test and trained datasets the Logistic regression model performed with accuracy 76% and Support Vector Machine model also performed with an accuracy of 76%. The KNN classifier performed the same even after the feature reduction. The summary of the evaluation metrics of the models using the PCA transformed test and training sets can be seen in Appendix – D Table 2.

Conclusion:

There are several factors that can cause diabetes in the Female patients. From the above analysis we can conclude that the Logistic Regression Model and Support Vector Machine model performed well for the diabetes dataset and from the EDA we can say that the high values of the features Glucose, Insulin, BMI, and SkinThickness are more prone to be tested positive for diabetes. The Female patients of age 70 years are more prone to be testing positive for diabetics. The female patients having BMI level above 30

is being tested positive for diabetics. The glucose levels are above 120 in the female patients having diabetics. The female patients being pregnant for more times are being tested positive for diabetics.

Assumptions:

Here in this analysis few values of the features are having zero values and cannot be considered as zero. Hence the median of the respective column is calculated and replaced the zeros with the median value for better analysis.

Limitations:

The study's limitations are observed in the heterogeneity between the models that difficult to compare them. This heterogeneity is present in many aspects; the main is the populations and the number of samples used in each model. Another significant limitation is when the model predicts diabetes complications, not diabetes.

Challenges/ Issues:

The amount of uncertainty in any model is difficult to know exactly, but there are specific types of uncertainties which can affect the model's performance. There are many factors that can cause diabetes, which when added can help in analyzing more in deep.

Future Uses/Additional Applications

As part of additional applications, I would like to use the functions and Pipelines for transforming the data and to improve the code readability. There are few other features that can be added to the dataset, and more samples can be added to the dataset which helps in better prediction of the Diabetes in Female patients.

Recommendations

For future recommendations of this project, I would like to perform the hyper parameter tuning on the models and check the model performance, and for evaluation purpose Cross Validation can be included for better model evaluation.

Implementation Plan

Though the Logistic Regression model and the Support Vector Machine models performed well with the features available in the dataset, the models are not recommended for deployment, as there are a lot of other features that are to be added to the dataset in prediction of Diabetes in Female patients like family history, diet, daily active hours, habit of smoking, habit of alcohol consumption etc. also play an important role in the prediction of Diabetes in female patients. However the models can be tested with the real life data collected from the hospitals and helps understand the features that are having more influence in a female patient tested positive for diabetes.

Ethical Considerations:

Ethical considerations while dealing the data related to patients is not to misuse the data collected and misrepresentation of the facts that are collected. I have taken care to not to misuse the data related to the patients in the project analysis.

References

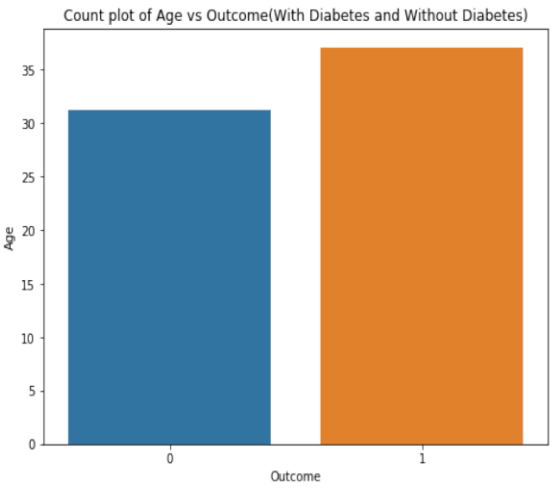
Kaggle.com, Diabetes dataset, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

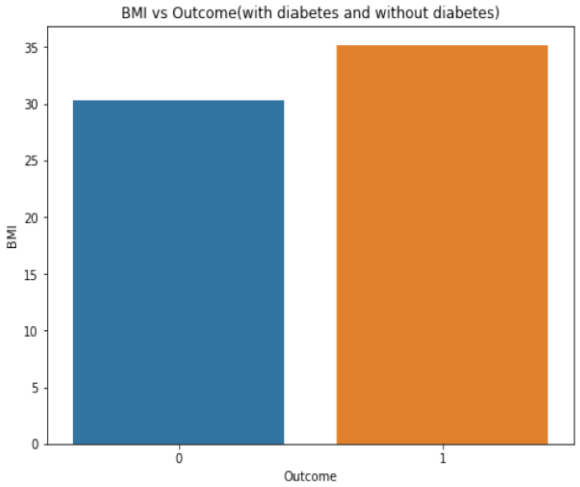
Hindawi, Prevalence, and Early Prediction of Diabetes Using Machine Learning in North Kashmir: A Case Study of District Bandipora, 29 August 2022, https://www.hindawi.com/journals/cin/2022/2789760/

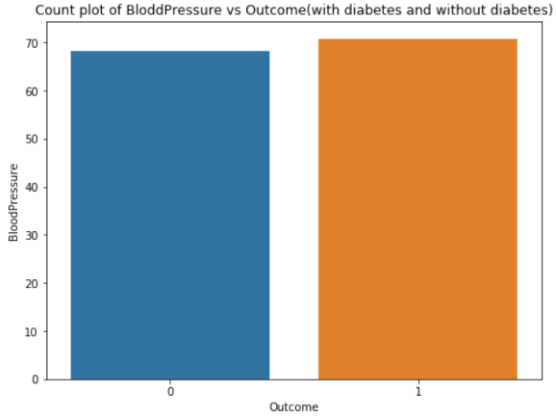
Google, Machine learning and deep learning predictive models for type 2 diabetes: a systematic review, 20 December 2021, https://dmsjournal.biomedcentral.com/articles/10.1186/s13098-021-00767-9

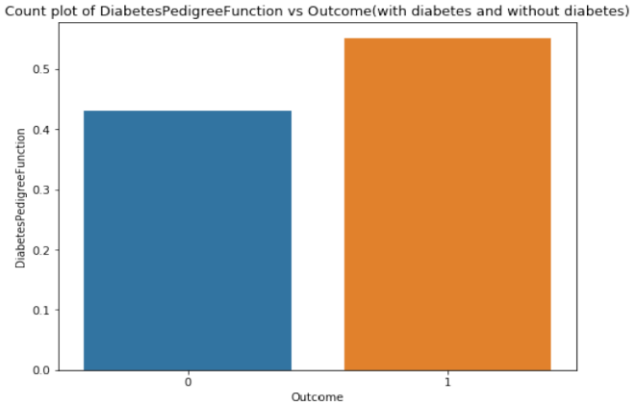
Appendix –A :

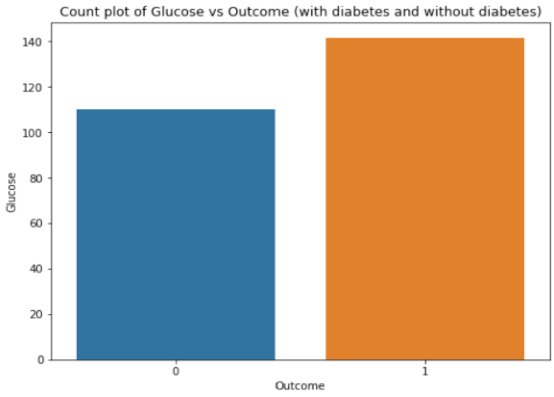
Figure 2: Count Plots of the distribution of features with the target variable

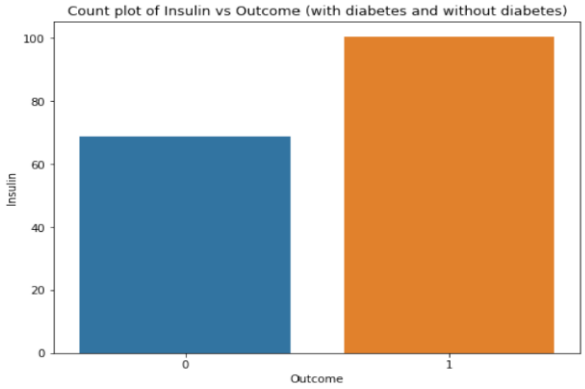




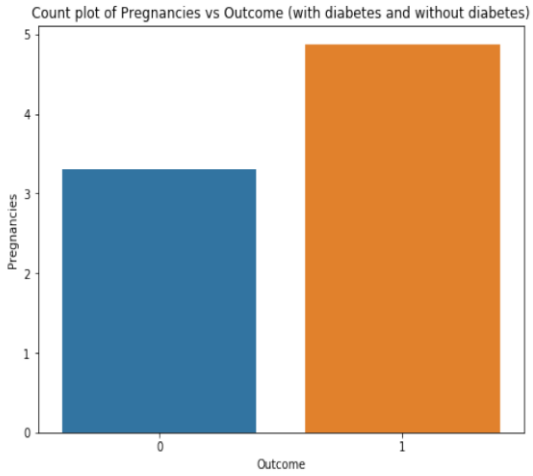


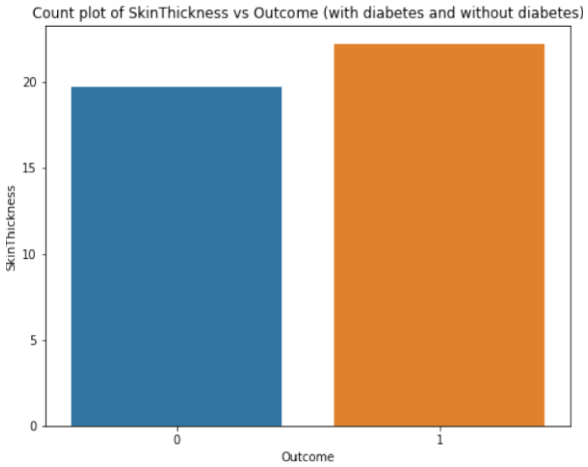


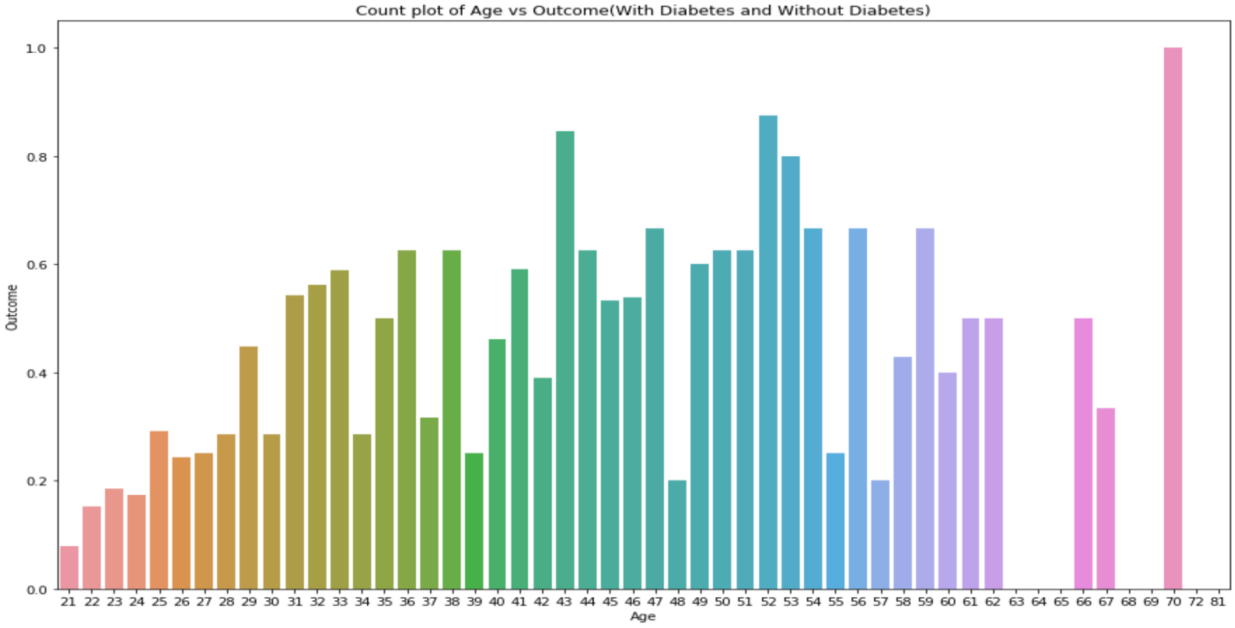


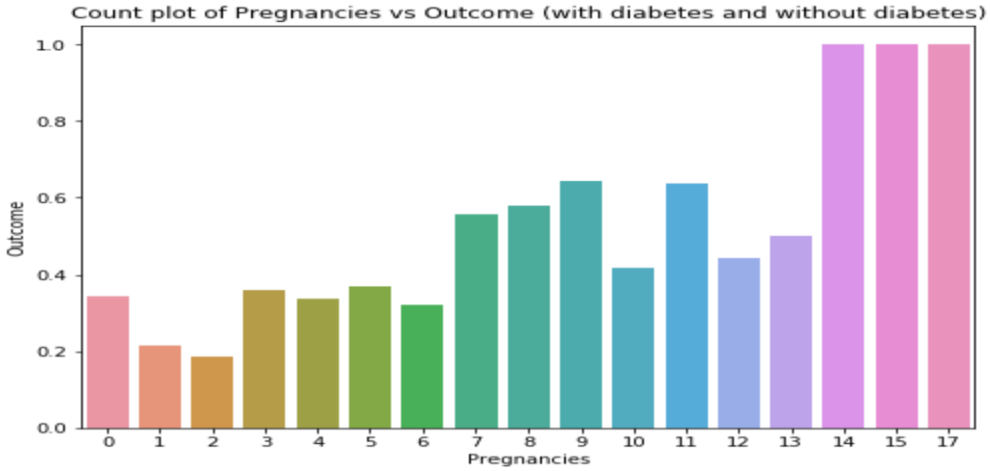


Project-1


Count plot of Pregnancies vs Outcome (with diabetes and without diabetes)
Pregnancies
Outcome
0
1
2
3
4
5


Count plot of SkinThickness vs Outcome (with diabetes and without diabetes)
SkinThickness
Outcome
0
5
10
15
20


Count plot of Age vs Outcome(With Diabetes and Without Diabetes)
Outcome
Age
0.0
0.2
0.4
0.6
0.8
1.0


Count plot of Pregnancies vs Outcome (with diabetes and without diabetes)
Outcome
Pregnancies
0.0
0.2
0.4
0.6
0.8
1.0

Appendix – B:

Figure 3 : Pair Plot of the diabetes dataset.

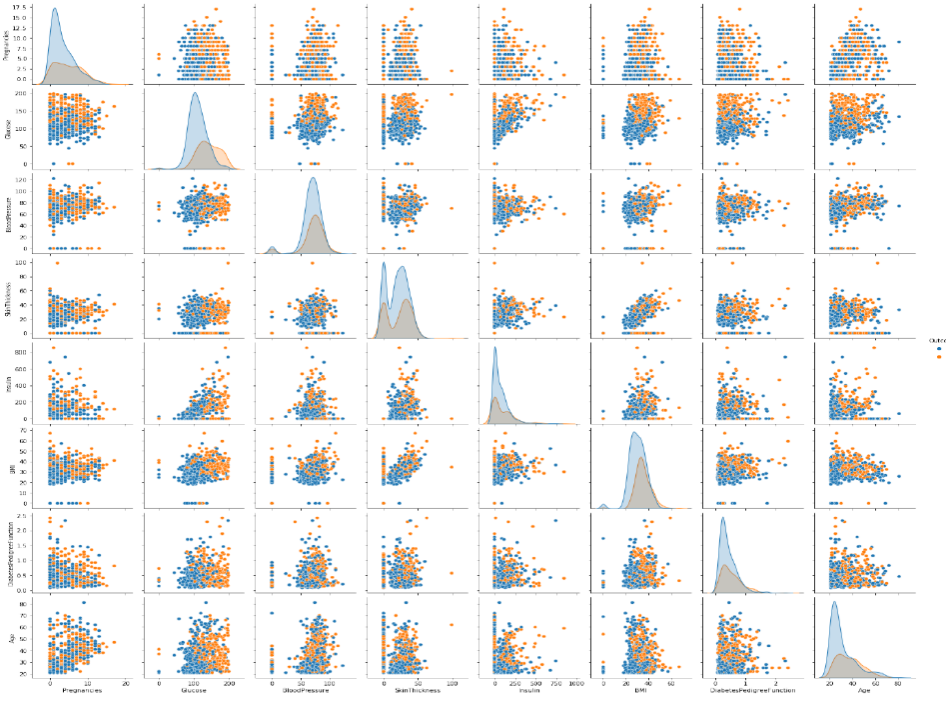
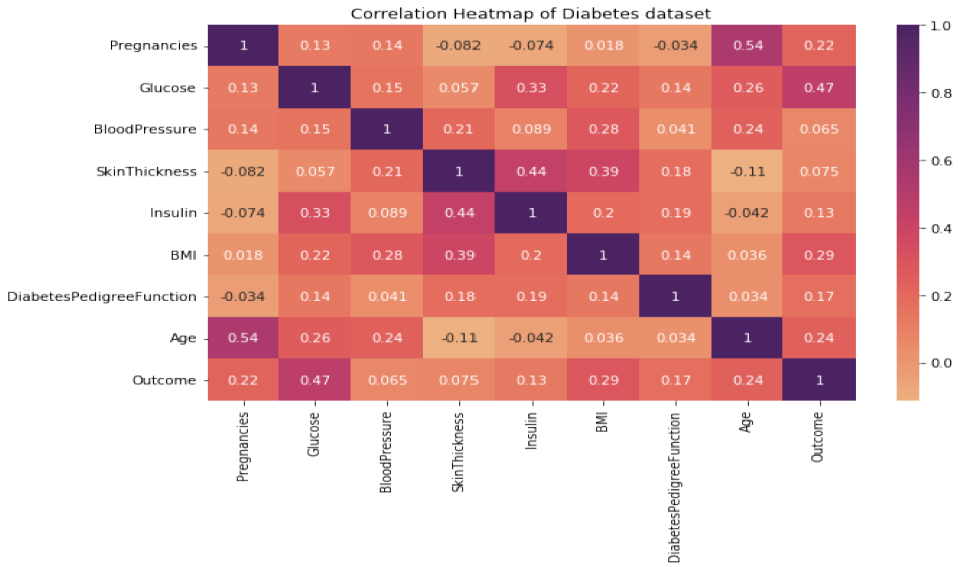
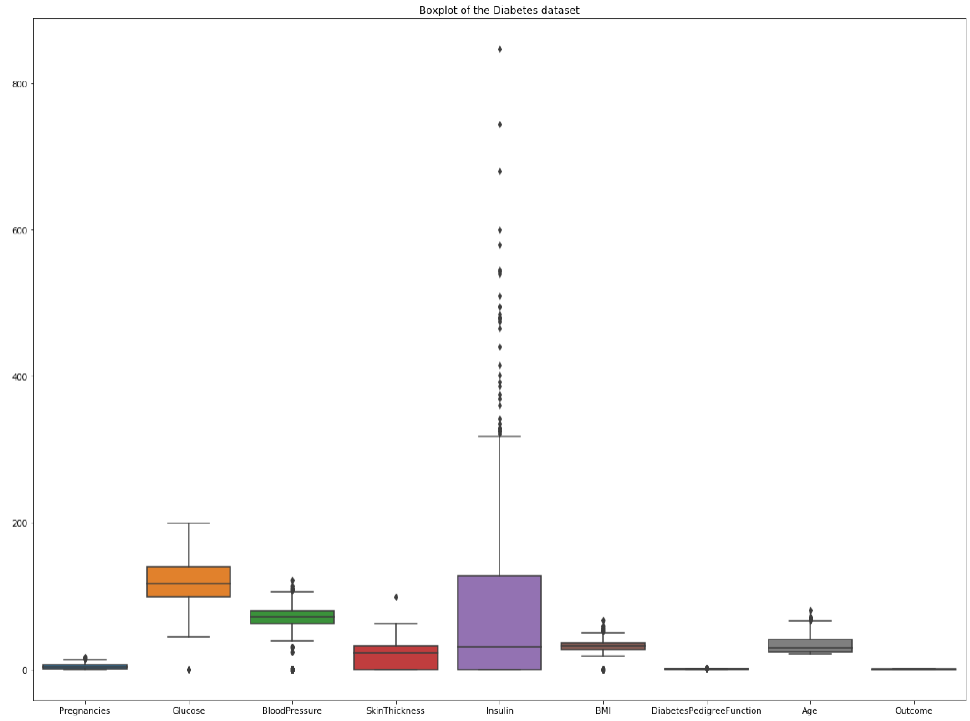


Figure 4: Correlation heat map of the diabetes dataset



Appendix – C:

Figure 5: Boxplot for identifying the outliers



Appendix – D:

Table – 1 : Summary of the evaluation metrics of the models without PCA

	Logistic Regression Model	KNN Classifier	Decision Tree Classifier Model	Support Vector Machine Model
Model	Logistic Regression	KNN Classifier	Decision Tree Classifier	Support Vector Machine Model
Accuracy (test)	0.779221	0.668831	0.74026	0.753247
Accuracy (train)	0.783388	0.809446	1	0.773616
Precision score	0.698	0.532	0.615	0.655
Recall score	0.673	0.6	0.727	0.655
F1 Score	0.685	0.564	0.667	0.655

Table- 2 : Summary of the evaluation metrics of the models with PCA

	Logistic Regression Model	KNN Classifier	Decision Tree Classifier Model	Support Vector Machine Model
Model	Logistic Regression(PCA)	KNN Classifier(PCA)	Decision Tree Classifier(PCA)	Support Vector Machine Model(PCA)
Accuracy (test)	0.766234	0.694805	0.675325	0.766234
Accuracy (train)	0.770358	0.819218	1	0.771987
Precision score	0.673	0.574	0.541	0.679
Recall score	0.673	0.564	0.6	0.655
F1 Score	0.673	0.569	0.569	0.667