DATA70121

Statistics and Machine Learning 1

EDA and Regression

User ID:

Data Information

PimaDiabetes is a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases in the US, containing 750 records of diagnostic tests for women, including Pregnancy, Blood Pressure, Glucose, Insulin, BMI, Skin Thickness, Age, and Diabetes Pedigree. The variable Outcome (1/0) indicates if the subject tested positive for diabetes. An Oral Glucose Tolerance Test (OGTT) measures plasma glucose concentration at 2 hours. Blood Pressure considers diastolic blood pressure. Skin Thickness stores the width of the skin over the triceps muscle. BMI measures weight and height. Insulin concentration is measured at 2 hours. A woman's diabetes pedigree score quantifies the genetic impact of her close relatives with and without diabetes, with higher scores indicating more diabetes diagnoses.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	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

Table 1: Data in PimaDiabetes dataset

Exploratory Data Analysis

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age	Outcome
count	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000	750.000000
mean	3.844000	120.737333	68.982667	20.489333	80.378667	31.959067	0.473544	33.166667	0.346667
std	3.370085	32.019671	19.508814	15.918828	115.019198	7.927399	0.332119	11.708872	0.476226
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.244000	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	36.500000	32.000000	0.377000	29.000000	0.000000
75%	6.000000	140.750000	80.000000	32.000000	129.750000	36.575000	0.628500	40.750000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Table 2: Central Tendency of PimaDiabetes dataset

The central tendency of different fields is displayed in Table 2. The fact that each field has a count of 750 indicates that there are no null values in the dataset. However, some fields, like BMI, Skin Thickness, Insulin, Blood Pressure, and Glucose, have minimum values of 0. These are considered as Null values since a value of 0 is not intended for them

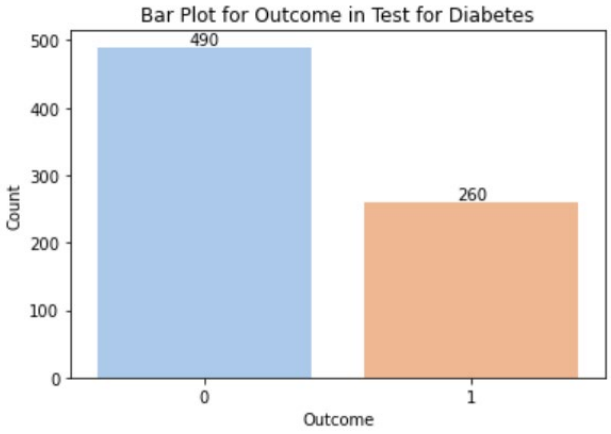

Figure 1: Bar Plot for Outcome in test for diabetes

Figure 1 shows that over 60% of the records in the dataset gives an outcome of 0 which shows that those people are not diagnosed with diabetes.

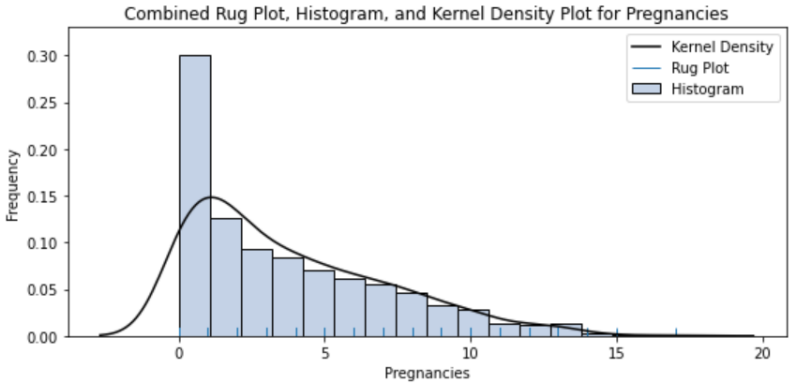

Figure 2: Combined Rug Plot, Histogram and Kernel Density Plot for Pregnancies

Figure 2 shows that the graph is right skewed and most of the women have less than 5 pregnancies in the dataset.

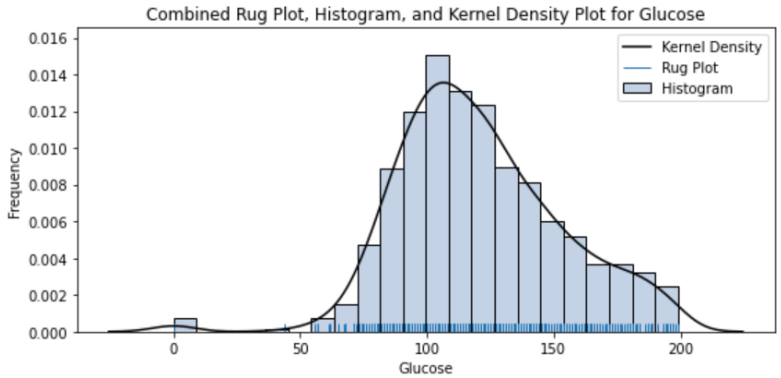

Figure 3: Combined Rug Plot, Histogram and Kernel Density Plot for Glucose

Figure 3 shows that Glucose contains some of the records with 0 as the value. The most frequent value comes in the range of 100 -150 mg/dl. The normal range of Glucose is found to be anywhere between 70 and 125 mg/dl. Figure 4 shows most women have normal diastolic pressure between 60 and 80. [1]

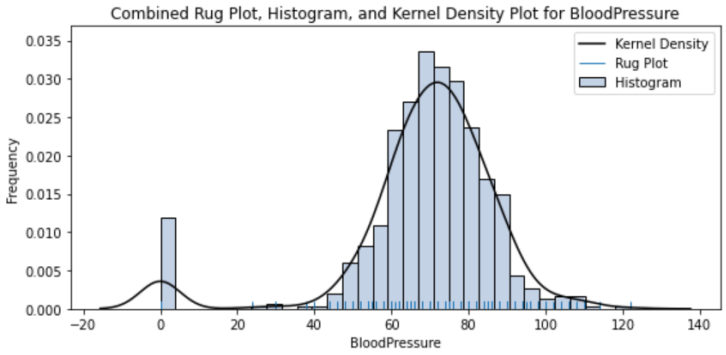

Figure 4: Combined Rug Plot, Histogram and Kernel Density Plot for Blood Pressure

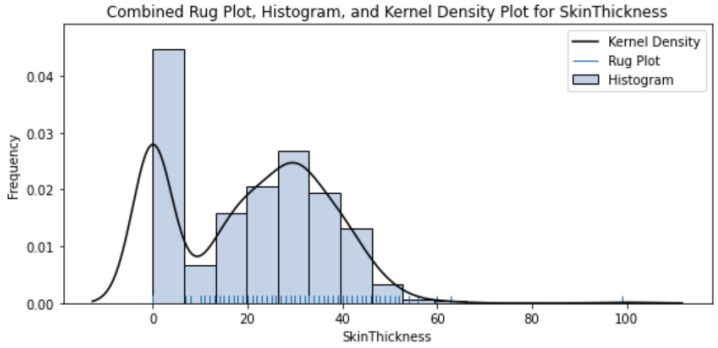

Figure 5: Combined Rug Plot, Histogram and Kernel Density Plot for Skin Thickness

When it comes to Skin Thickness, Figure 5 shows that it is bi-modal but the first peak in the graph is caused by records having a value of 0 which is inaccurate. Figure 6 shows that nearly 50% of data have a value of 0 and this should be imputed.

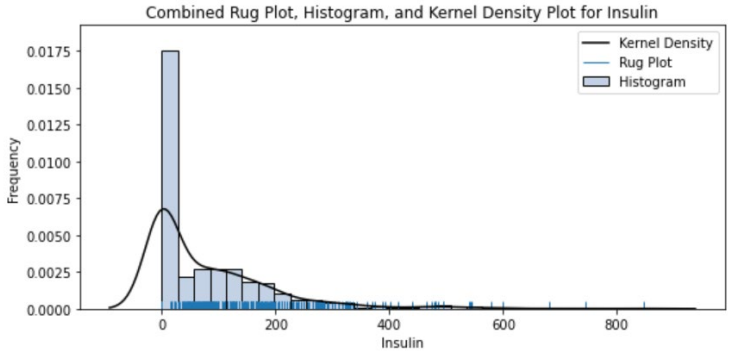

Figure 6: Combined Rug Plot, Histogram and Kernel Density Plot for Insulin

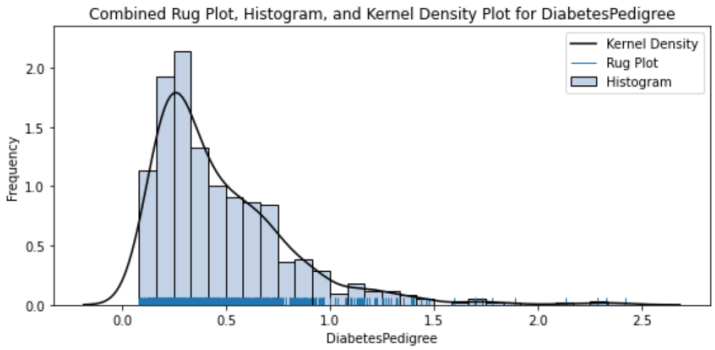

Figure 7: Combined Rug Plot, Histogram and Kernel Density Plot for Diabetes Pedigree

Figure 7 shows that most women have Diabetes Pedigree value between 0 and 0.5 which shows that most women do not have close relatives diagnosed with diabetes. From Figure 8 it is clear that most of the women fall under the age group between 20 and 40.

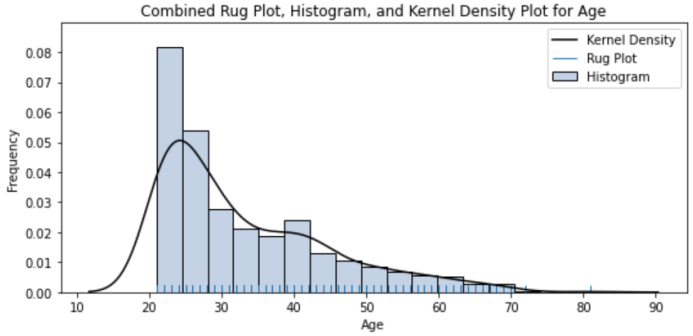

Figure 8: Combined Rug Plot, Histogram and Kernel Density Plot for Age

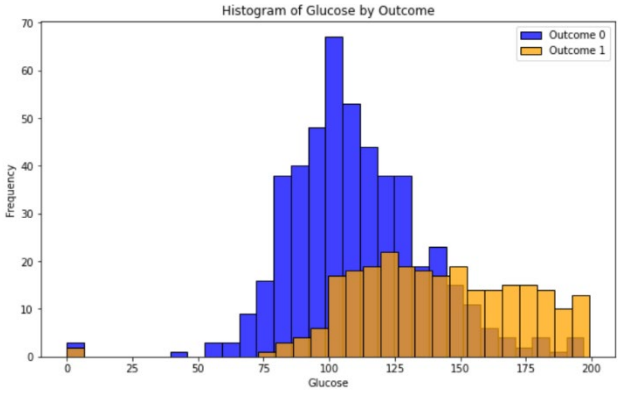

Figure 9: Histogram of Glucose by Outcome

Figure 9 shows that people who are diabetic have higher values of Glucose than non-diabetic people. From Figure 10 it is clear that people with greater diastolic pressure have diabetes.

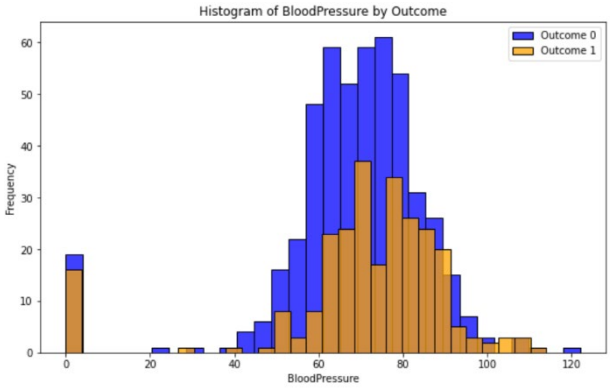

Figure 10: Histogram of Blood Pressure by Outcome

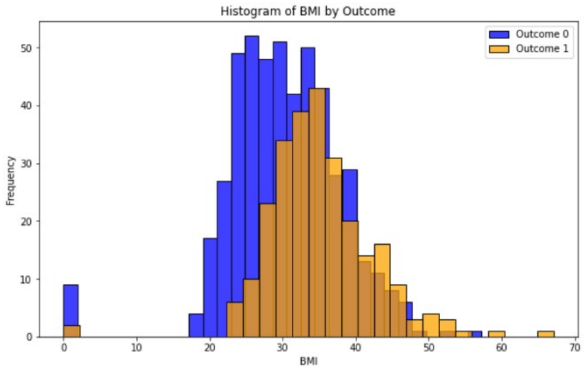

Figure 11: Histogram of BMI by Outcome

People who have a BMI between 30 and 40 are found to have diabetes from Figure 11. From Figure 12 it is clear that each column has outliers.[2] Hence pre-processing is required to get an optimized model to predict the Outcome.

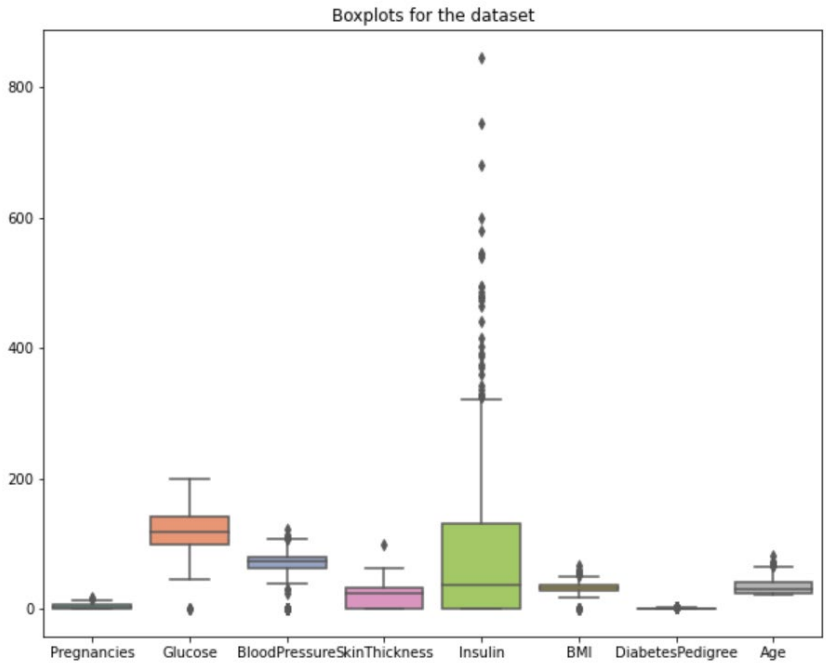

Figure 12: Boxplot for PimaDiabetes

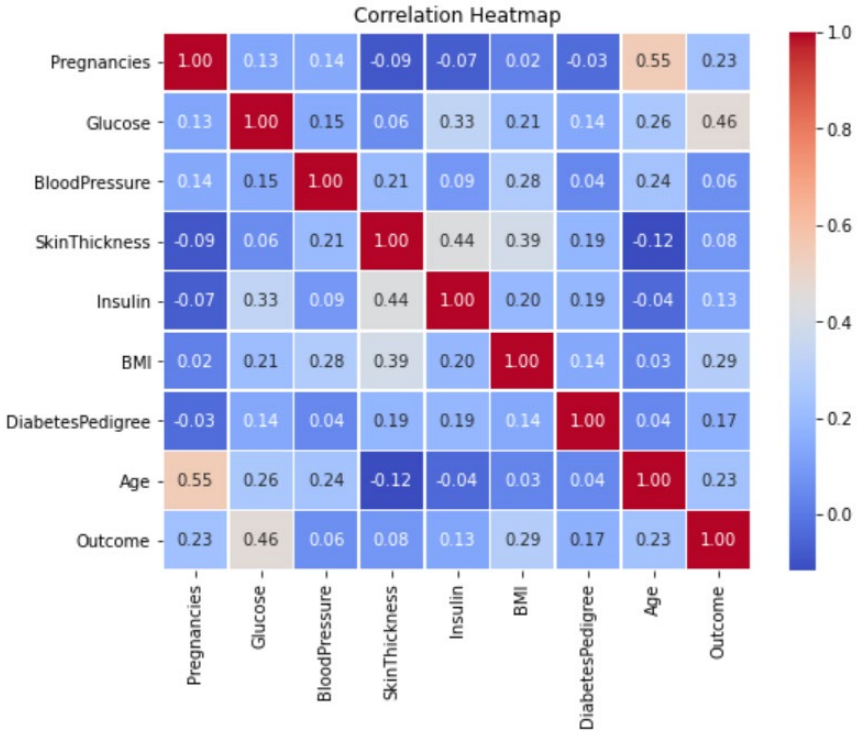

Figure 13: Heatmap of Correlation Matrix

The correlation heatmap from Figure 13 reveals that pregnancy and age have the highest correlation, while insulin and skin thickness and glucose and outcome have a stronger correlation.[3] Therefore, examining these relationships is crucial for a deeper understanding.

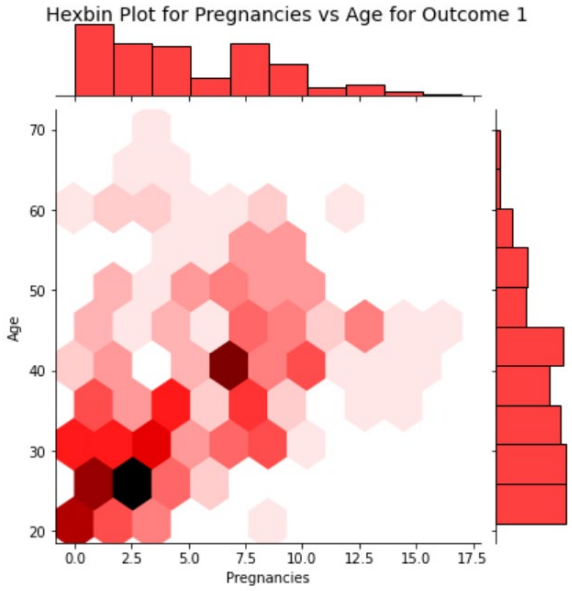

Figure 14: Hexbin plot for Pregnancies vs Age for Outcome 1

Based on the value of the outcome, a Hexbin is plotted after taking into account age and pregnancy.[4] Figure 14 shows that the greatest number of people with diabetes are found to be falling under the age group between 20 and 30 who have had less than 5 pregnancies. On the other hand, a lot of people between the age group of 20 and 30 with less than 4 pregnancies are also found to not have diabetes from Figure 15. However, both the graphs show that as age increases there is a tendency for the number of pregnancies also to increase.

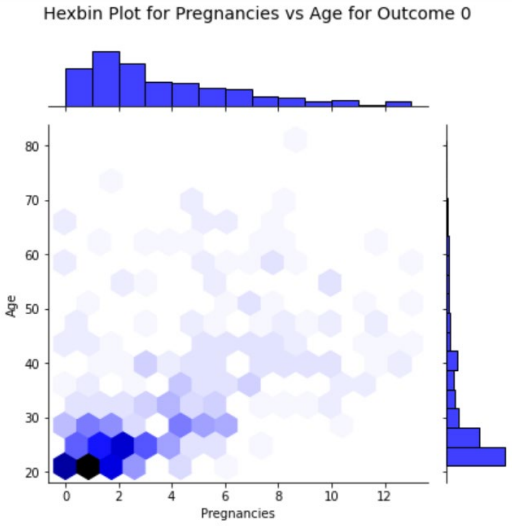

Figure 15: Hexbin plot for Pregnancies vs Age for Outcome 0

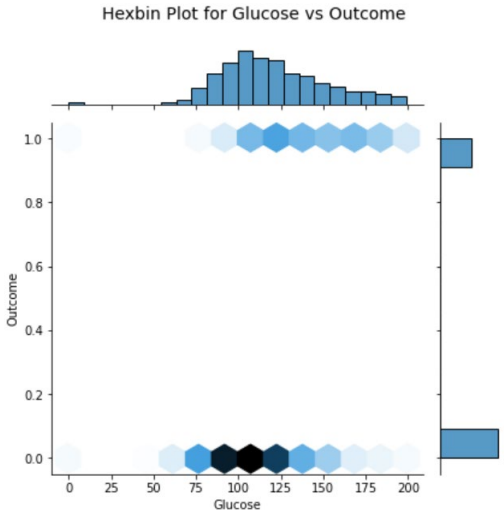

Figure 16: Hexbin plot for Glucose vs Outcome

From Figure 16, many of the women who do not have diabetes fall under the range of having glucose between 75 mg/dl and 125 mg/dl which is the normal range of glucose. Whereas, the greatest number of people with diabetes have over 125 mg/dl as their glucose level.

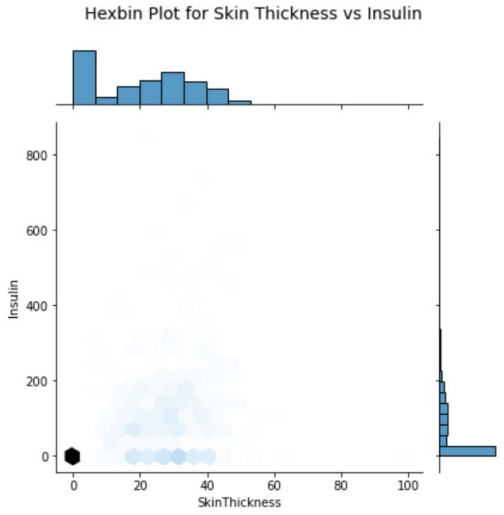

Figure 17: Hexbin plot for Skin Thickness vs Insulin

From Figure 17 it is clear that Skin Thickness and Insulin had higher correlation because of majority of their values in the records having a value of 0. This is inaccurate and these records should be imputed before being fitted in a model.

Probability of Developing Diabetes

A new field, SevenOrMorePregnancies, was created to determine the likelihood of diabetes in women with seven or more pregnancies. The data was divided into training and testing sets, and a Logistic Regression model was fitted using SevenOrMorePregnancies as the sole predictor. The model showed an accuracy of 0.6866 and an F-score of 0.405. The probability of developing diabetes was stored and the predict_proba function was used to determine the probability. The likelihood of getting diabetes with six or fewer pregnancies was 0.29464, and the probability of getting diabetes with seven or more pregnancies was 0.5787.

Regression Model for Predicting Outcome

The dataset was pre-processed using a KNN Imputer with a nearest neighbour value of 5 to impute 0 values for Glucose, Blood Pressure, Insulin, Skin Thickness, and BMI, and all fields except Outcome were standardized using a Standard Scaler. [5]

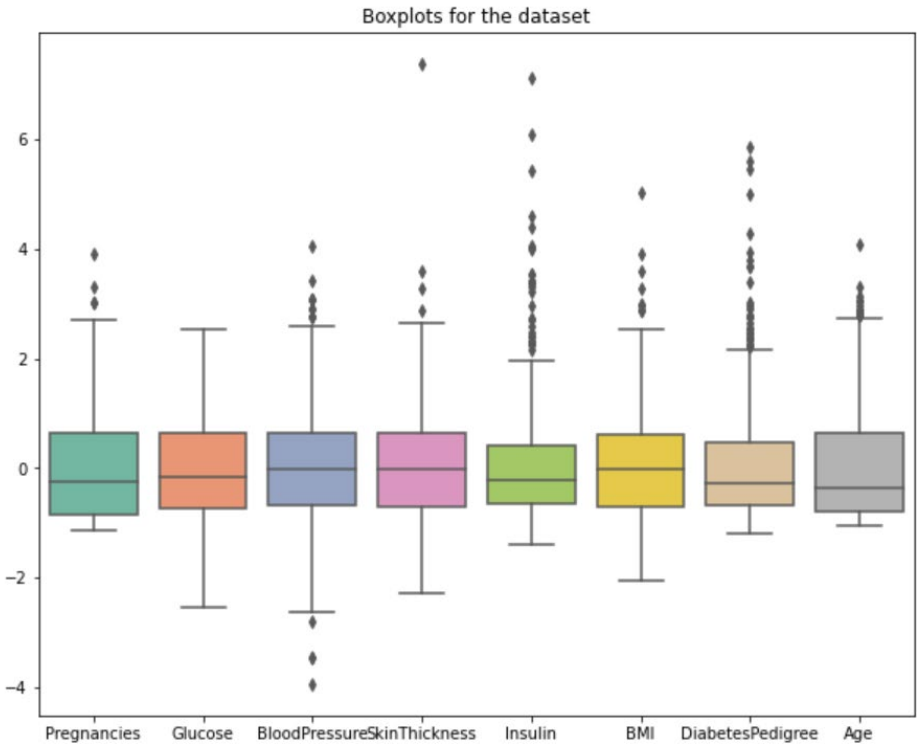

Figure 18: Boxplot for PimaDiabetes after Imputing and Scaling

Figure 18 displays a boxplot of fields post-imputing and scaling, revealing that all fields except field age have outliers, which requires removal.[6]

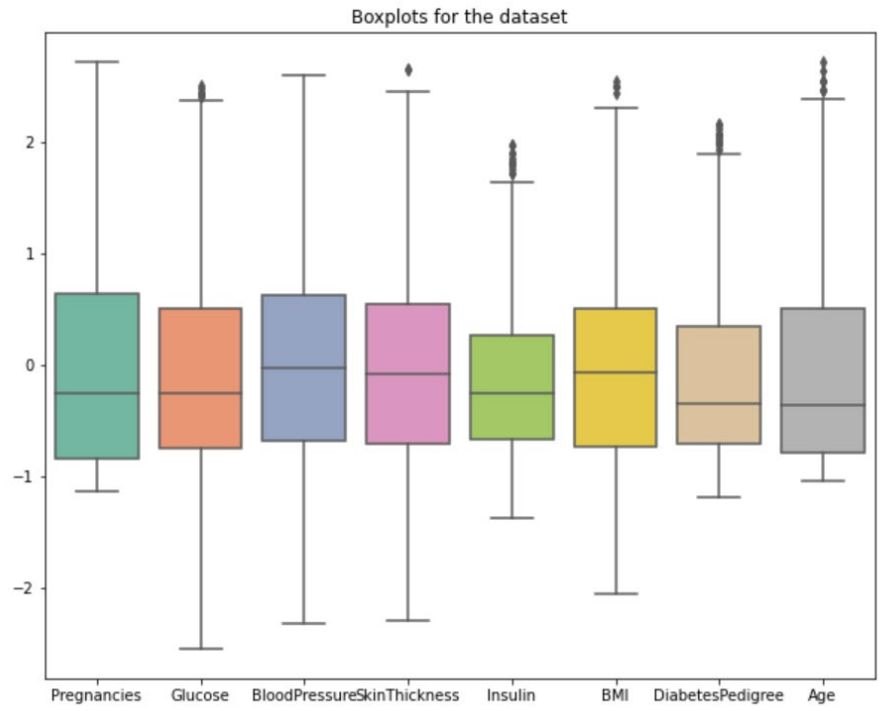

Figure 19: Boxplot after removing outliers from PimaDiabetes

Figure 19 displays a boxplot indicating a significant reduction in the number of outliers after removing them.

ToPredict is a dataset used for predicting outcomes and diabetes probability, consisting of 5 records with 0 values for Insulin and Skin Thickness in some of the records. Table 3 displays the entire dataset.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigree	Age
0	4	136	70	0	0	31.2	1.182	22
1	1	121	78	39	74	39.0	0.261	28
2	3	108	62	24	0	26.0	0.223	25
3	0	181	88	44	510	43.3	0.222	26
4	8	154	78	32	0	32.4	0.443	45

Table 3: ToPredict Dataset to predict the Outcome

The model uses KNN Imputer and Standard Scaler on the dataset, trained on the PimaDiabetes dataset, and logistic regression. The optimum feature combination is determined by using 80%

of the PimaDiabetes data as the training set and 20% as the testing set, with a feature significance graph initially produced. [7]

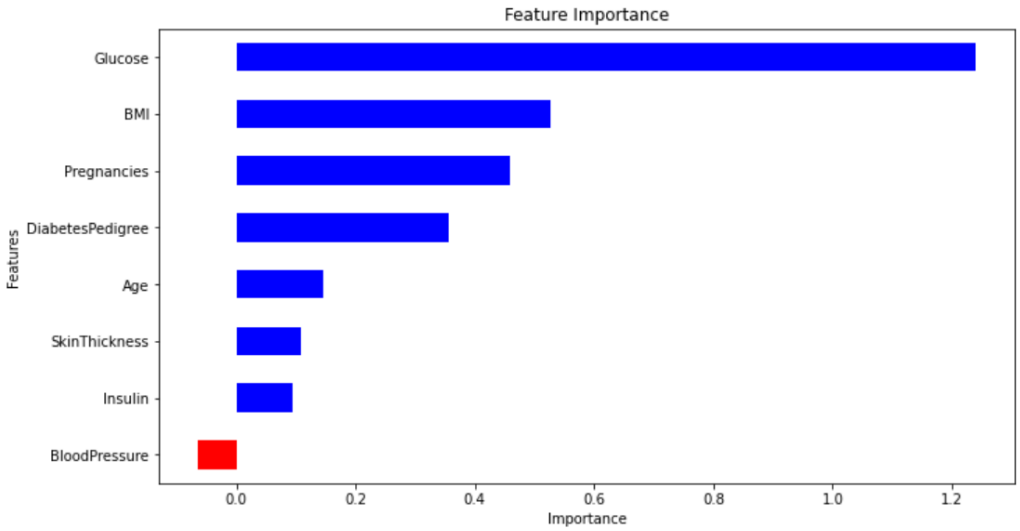

Figure 20: Feature Importance considering all variables as features

The model's accuracy is 0.712, with Blood pressure having a negative importance from Figure 20. Glucose, BMI, Pregnancies, Diabetes Pedigree, Age, and Insulin were used as features to increase accuracy to 0.719 for subsequent training. No feature has a negative impact on the model, as shown in Figure 21.

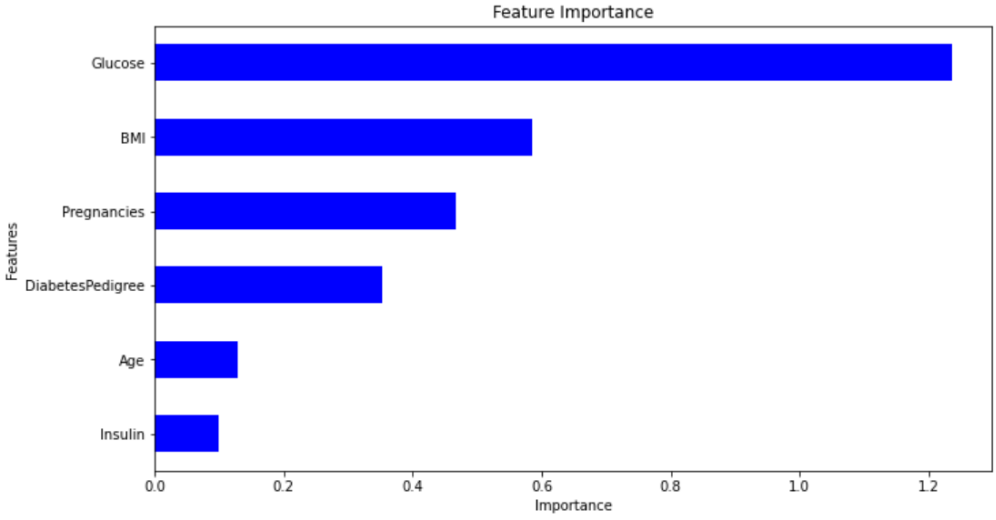

Figure 21: Feature Importance considering only Glucose, BMI, Pregnancies, DiabetesPedigree, Age and Insulin as Features

As a result, the model makes use of these features. To predict the outcome, it uses the ToPredict dataset after training on the whole PimaDiabetes dataset. The probability of developing diabetes was determined to be 0.389 using predict_proba ().

References

[1] National Institute on Aging. (n.d.). *High Blood Pressure and Older Adults*. [online] Available at: https://www.nia.nih.gov/health/high-blood-pressure/high-blood-pressure-and-older-adults#:~:text=Normal%20blood%20pressure%20for%20most.

[2] Nishida, K. (2019). *Introduction to Boxplot Chart in Exploratory*. [online] Medium. Available at: https://blog.exploratory.io/introduction-to-boxplot-chart-in-exploratory-255c316a01ca [Accessed 20 Nov. 2023].

[3] Szabo, B. (2020). *How to Create a Seaborn Correlation Heatmap in Python?* [online] Medium. Available at: https://medium.com/@szabo.bibor/how-to-create-a-seaborn-correlation-heatmap-in-python-834c0686b88e.

[4] He, S. (2023). *Mastering Hexbin Plotting in Python: A Beginner's Guide*. [online] Medium. Available at: https://levelup.gitconnected.com/mastering-hexbin-plotting-in-python-a-beginners-guide-3626e0389c37#:~:text=Hexbin%20plots%20are%20a%20versatile [Accessed 20 Nov. 2023].

[5] Bhanupsingh (2023). *Handling Missing Data with KNN Imputer*. [online] Medium. Available at: https://medium.com/@bhanupsingh484/handling-missing-data-with-knn-imputer-927d49b09015#:~:text=Applying%20KNN%20Imputer [Accessed 20 Nov. 2023].

[6] Bhandari, P. (2022). *How to Find Outliers | Meaning, Formula & Examples*. [online] Scribbr. Available at: https://www.scribbr.co.uk/stats/statistical-outliers/.

[7] Serengil, S. (2021). *Feature Importance in Logistic Regression for Machine Learning Interpretability*. [online] Sefik Ilkin Serengil. Available at: https://sefiks.com/2021/01/06/feature-importance-in-logistic-regression/.