Clustering and Fitting
NAME:Ramya sree koka
ID: 23068394
Git Hub:https://github.com/rk24aao/NAME-Ramya-sree-koka.git

Introduction

This report aims at analyzing the use of clustering also known as classification and linear regression on diabetes dataset. It also involves visual developments such as bar graphs, scatter diagrams, confusion maps, and silhouette charts and then clustering and regression for outcomes of diabetes.

Bar Chart of Diabetic and Non-Diabetic Individuals

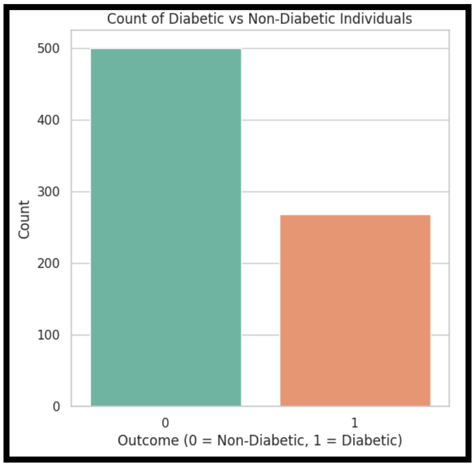


Figure 1: Count of Diabetic vs Non-Diabetic Individuals

The bar chart below shows the distribution of the diabetic cohort where Outcome = 1 and the non-diabetic cohort where Outcome = 0. A greater proportion of people are non-diabetic to diabetic population. It is also important to review the distribution of records primarily because one class may have significantly fewer records than another in any sample data provided for this dataset.

Scatterplot of Glucose Against BMI

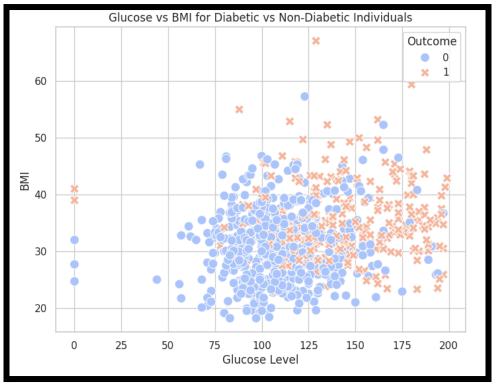


Figure 2: Glucose vs BMI for Diabetic vs Non-Diabetic Individuals

The following scatter plot is representative of glucose levels as well as BMI of diabetic and non-diabetic patients. The mean glucose level and BMI (Outcome = 1) are higher in the case of diabetic individuals. For clustering and classification, the plot supports the identification of the division between two distinct populations.

Confusion Matrix for Diabetes Prediction

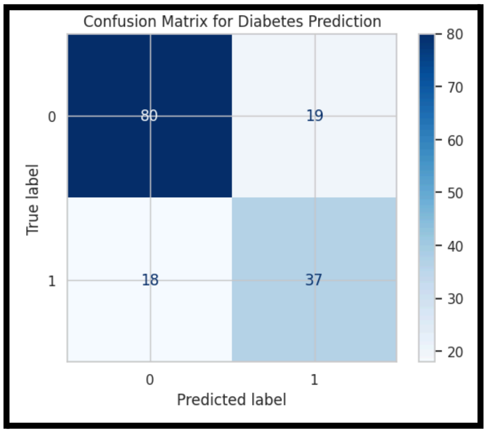
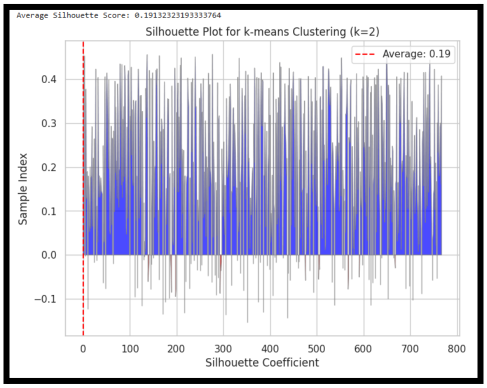


Figure 3: Confusion Matrix for Diabetes Prediction

This confusion matrix gives an account of the classification accuracy of the predicted model. How it helps: it gives real positive, real negative, false positive, and false negative. The non-diabetic people are predicted correctly most of the time, although there is some misclassification about the diabetics primarily owing to the problem of class imbalance that the current machine learning model faces.

Silhouette Plot for Clustering

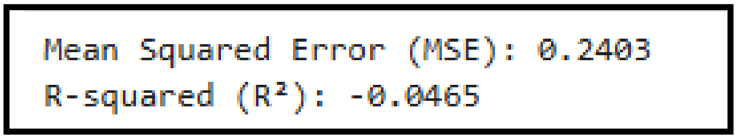


```
Average Silhouette Score: 0.19132323193333764
```

Figure 4: Silhouette Plot for K-Means Clustering

The silhouette plot assesses the performance of clustering of the k-means (k=2). A high positive silhouette coefficient is desirable while a low positive or a negative one points to misclassification of the data items. The smallest value of the silhouette coefficient with an average of 0.19 testifies to low clustering quality and can be explained by the interpenetration of the clusters of diabetic and non-diabetic patients in feature space.

Clustering Analysis



```
Mean Squared Error (MSE): 0.2403
R-squared (R²): -0.0465
```

Figure 5: K-Means Clustering Results

The following presents the performance measures of linear regression model. Mean Squared Error (MSE = 0.2403) shows the number of prediction errors and negative R-squared points to an improper fitting of the model.

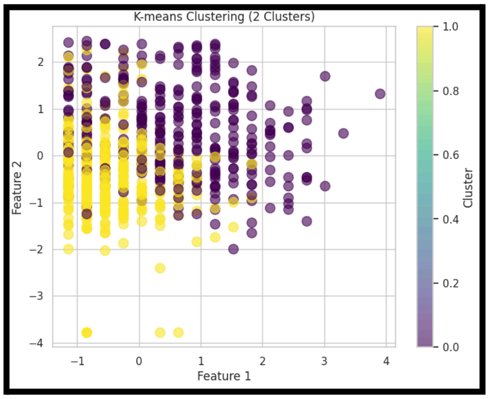
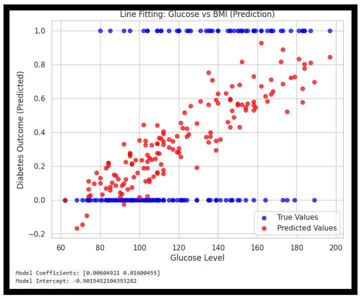


Figure 6: K-Means Clustering

Evidently, the scatterplot below represents k-means clustering outcomes by creating two clusters. They define data points placed into two separate clusters; yellow and purple according to feature value. The clustering can be observed, meaning the performance with the clusters is moderate with noticeable ambiguity when they overlap.

Linear Regression



```
Model Coefficients: [0.00604921 0.01600455]
Model Intercept: -0.9019452104393282
```

Figure 7: Linear Regression

This scatterplot also shows the Diabetes outcome model created with the Linear Regression comparing Glucose and BMI. The graphs show true values … and predicted values. The coefficients and intercept relate the input variables and response whereas the differences show that there is no complete explanation or prediction capability mainly owing to the characteristics of given datasets.

Conclusion

Lastly, this report used clustering and linear regression on the diabetes dataset to show the distribution of data and the performance of a model. Although there are some misclassified and prediction errors, the results offer the groundwork for enhancing the diabetes prediction models most successfully.

Reference

kaggle.com, 2024. Diabetes Dataset. Viewed on 9th Dec, 2024. From, https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset