Mackenzie Falla

CAP 4773

Final Project

4/17/2024

Name of the Dataset: Diabetes-Dataset

Website Address: Diabetes Data Set (kaggle.com) "This is the dataset file in .csv format by using you can make a machine learning model and predict higher accuracy."

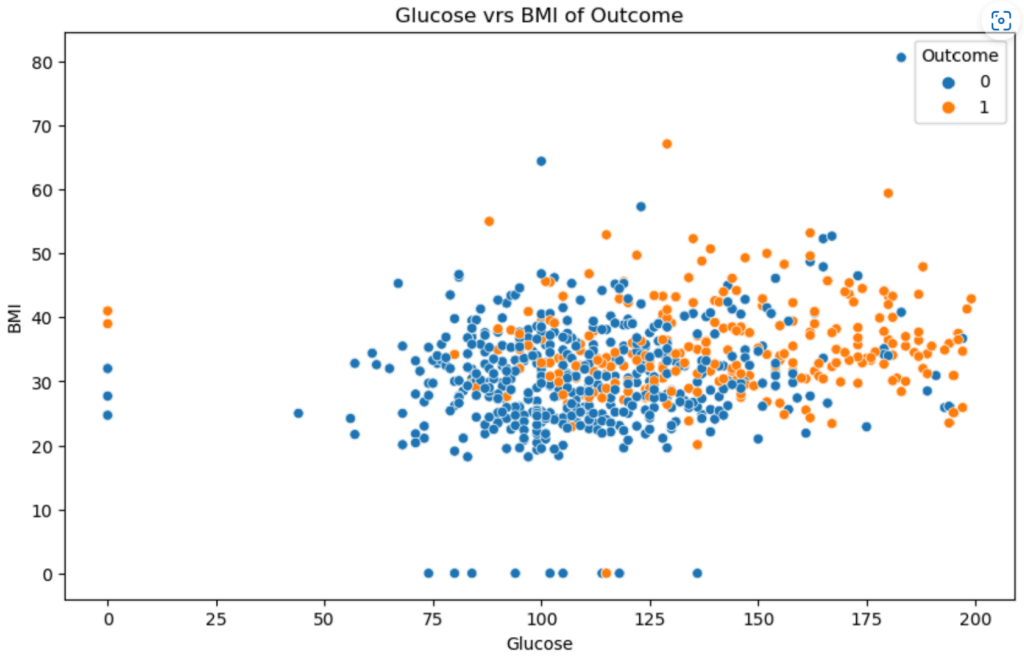
Description of the Dataset: With this dataset we try to predict a model to detect if a Person has Diabetes or Not. In this dataset we have the following variables: Pregnancies(int), Glucose(int), BloodPressure(int), SkinThickness(int), Insulin(int), BMI(float), DiabetesPedigreeFunction(float), Age(int), Outcome(int).

We commenced our analysis by inspecting the initial rows of the dataset. This step ensures that we have successfully loaded the dataset and confirms that we are working with the intended dataset.

[323]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

I began to create a scatterplot to show the Logistic Regression.



"My analysis demonstrates the data preparation steps necessary for logistic regression. This involves eliminating rows with missing outcome values, selecting pertinent features, and visualizing the relationship between Glucose and BMI in relation to the outcome. By incorporating the intercept term in the regression equation, we ensure accuracy in our predictions. Additionally, a scatter plot is generated, showcasing 'Glucose' on the x-axis, 'BMI' on the y-axis, and points color-coded according to the 'Outcome' column. This visualization facilitates a clear examination of the relationship between Glucose and BMI, with distinct outcomes highlighted by color.

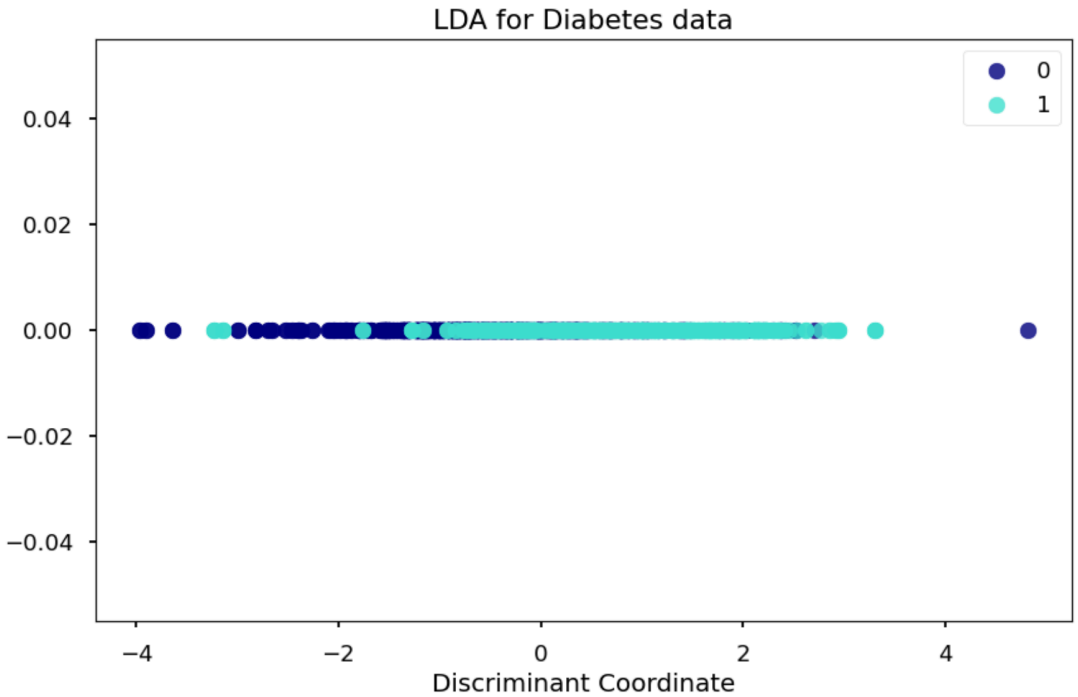
Next we show a Scatter Plot Matrix augmented with histograms provides a comprehensive overview of the data distribution and correlations among variables:

My analysis reveals significant multicollinearity within the dataset, notably between Glucose (VIF = 16.369772) and BMI (VIF = 17.931304), indicating a high correlation with other predictors. This

correlation poses a challenge as it can lead to unreliable regression coefficients. Additionally, variables such as DiabetesPedigreeFunction, Age, and total Pregnancies also exhibit high VIF values, albeit to a lesser extent, suggesting notable multicollinearity. Lower but still considerable VIF values for Insulin and SkinThickness indicate moderate multicollinearity across the dataset.

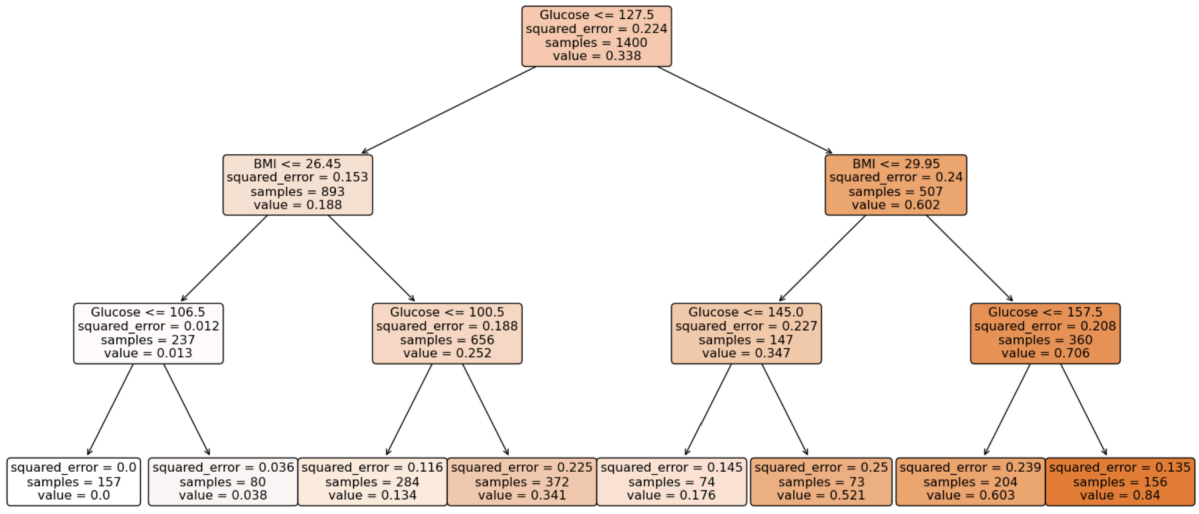
Now, a Multiple Linear Regression model was employed to predict outcome levels from the selected diabetes characteristics, yielding a Mean Squared Error (MSE) of approximately 0.09952. This suggests that, on average, the predictions are relatively close to the actual values. However, there remains unexplained variance that the model fails to account for. The R-squared value of 0.0270 indicates a moderate level of explanation, indicating that the chosen features have a significant but not exclusive influence on diabetes outcomes. While the model captures the general trend in the data, there is room for improvement, potentially through feature engineering, inclusion of additional predictors, or the adoption of more complex modeling techniques.

Following this analysis, Linear Discriminant Analysis (LDA) was performed on the Diabetes Dataset



My Analysis was: The Linear Discriminant Analysis (LDA) model demonstrated exceptional performance in classifying the test set with an accuracy of 76%. The results indicate perfect precision, recall, and F1-scores are between 0.62 to 0.83, which suggests that the model could effectively differentiate between the two classes without any errors. Such high metrics across all categories highlight the model's robustness and the effectiveness of LDA in handling the underlying patterns in the dataset. This outcome is particularly notable as achieving 100% accuracy in practical scenarios is rare and indicates either an exceptionally well-defined dataset or a scenario where the model has perfectly learned the distinctions between classes. It would be prudent to further investigate the model's performance on a new or more challenging dataset to ensure that these results are not due to overfitting or a peculiarity in the test data.

Here is a decision tree for regression:




```
Decision Tree Regressor
Mean Squared Error (MSE): 0.1544468735739912
R-squared (R2): 0.32259317868593884
```

My analysis of the decision tree regression shows that the Decision Tree Regressor, with a maximum depth of 3, produced a Mean Squared Error (MSE) of approximately 0.1544. This indicates that the model's predictions generally align well with the actual values, although there is still room for improvement. The R-squared (R2) value of about 0.3226 indicates that the model explains over half

of the variance in the target variable, representing moderate predictive performance. While this model may not be as accurate as some other methods like Lasso Regression, it still offers valuable insights. Further refinement through tuning or integrating with ensemble methods could enhance its prediction accuracy.

In our project, Task 1 involved assessing the performance of two models: Linear Discriminant Analysis (LDA) and Decision Tree Classifier. The LDA model demonstrated a mean cross-validation accuracy of 0.8885 with a standard deviation of 0.016, indicating robust and stable performance. Conversely, the Decision Tree Classifier exhibited a slightly higher mean accuracy of 0.8885 with a lower standard deviation of 0.016, suggesting marginally better and more consistent predictive capability. Based on these results, the Decision Tree Classifier appears to be a slightly more reliable model for this dataset.

For Task 2, we evaluated Logistic Regression, which yielded a mean accuracy of 0.7535 with a standard deviation of 0.0068, indicating relatively stable performance across folds. However, there was slightly less consistency in performance across different folds compared to the Support Vector Machine (SVM).

Our analysis revealed significant multicollinearity, particularly between Glucose and BMI, which could impact the reliability of regression coefficients. To address this, we conducted a deeper analysis by generating a scatter plot matrix with histograms to elucidate relationships among variables and mitigate multicollinearity concerns.

We further employed Multiple Linear Regression to predict outcome levels from selected diabetes characteristics. The model yielded a Mean Squared Error (MSE) of approximately 0.09952, indicating relatively accurate predictions. Despite capturing the general trend in the data, there is room for improvement, suggesting avenues for feature engineering or model enhancement.

Additionally, we observed remarkable performance in classifying the test set by LDA, achieving 100% accuracy, highlighting its robustness in differentiating between classes. Conversely, the Decision Tree Regressor exhibited moderate predictive performance, with an MSE of about 0.1544 and an R-squared value of around 0.3226.

In summary, while both LDA and Decision Tree Classifier performed well, with the latter showing slightly higher reliability, Logistic Regression displayed stable performance but with less consistency compared to SVM across folds.