Name: Onyeka Ngene

Course Tittle: Capstone Project

Date: 11/11/2023

INTRODUCTION

Examining the project's problem statement and features, it's crucial to recognize that some features contribute to the onset of diabetes, while in many cases, diabetes is not prevalent.

As per the World Health Organization (WHO), diabetes is a chronic metabolic disease characterized by elevated blood glucose levels, leading to serious damage to the heart, blood vessels, eyes, kidneys, and nerves over time. The most common type is type 2 diabetes, typically occurring in adults when the body becomes resistant to insulin or produces insufficient insulin. Over the past three decades, the prevalence of type 2 diabetes has sharply increased worldwide. Type 1 diabetes, formerly known as juvenile or insulin-dependent diabetes, is a chronic condition where the pancreas produces little or no insulin. Access to affordable treatment, including insulin, is crucial for the survival of individuals with diabetes. There is a global target to halt the rise in diabetes and obesity by 2025.

Diabetes affects about 422 million people globally, with the majority residing in low- and middle-income countries, and 1.5 million deaths are directly attributed to diabetes annually. Both the number of cases and the prevalence of diabetes have steadily increased in recent decades.

EXPLORATORY DATA ANALYSIS

We acknowledge an imbalance in our dataset, characterized by a significant difference in the number of patients with diabetes compared to those without. Nevertheless, certain features critical for diabetes development, identified in medical understanding, demonstrate inconsistencies such as missing values, skewed distribution, and data integrity issues. Given the healthcare context of the dataset, our approach to managing dimensionality is restricted.

In our exploration of the data, we observe 768 rows and 9 columns. Visual inspection highlights unusually high mean values for Glucose, Blood Pressure, Insulin, BMI, and Age, suggesting potential outliers indicated by respective standard deviations. The presence of zeros in minimum values for features like Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI indicates skewness.

Upon detailed examination, the dataset is evidently imbalanced, with 500 instances of zeros and 268 instances of ones in the target variable.

BIVARIATE ANLYSIS

The analysis uncovers key insights:

BMI and Skin Thickness: A moderate linear relationship exists between BMI and Skin Thickness. Increasing Skin Thickness and BMI correlate with a higher diabetes prevalence.

Glucose Levels and Blood Pressure: Diabetes prevalence rises with increased glucose levels and blood pressure.

Elevated plasma glucose plays a crucial role in diabetes onset, although with a weak correlation.

Glucose and BMI Interaction: Diabetes prevalence increases as Glucose scores rise across varying BMI levels.

Pregnancy and Age: The number of pregnancies (0.0 to 17.5) and age concentration (30 to 55) are associated with an elevated diabetes prevalence.

Glucose and Insulin: An upward trend in Glucose and Insulin corresponds to increased diabetes prevalence along the y-axis (Glucose).

BMI and Pregnancy History: Women with BMI above 40, including those never pregnant, show a higher diabetes prevalence.

BMI from 30 to 50, with multiple pregnancies, is linked to increased diabetes prevalence.

Plasma Glucose Levels: Individuals with plasma glucose levels (0 to 11), whether pregnant or not, exhibit diabetes when the glucose score is 120 or higher.

Diabetes Pedigree Function: Patients with a Diabetes Pedigree Function score (0.0 to 1.4) related to a glucose score of 120 and above tend to have diabetes.

Correlation Matrix: The correlation matrix indicates a moderate positive correlation between BMI and Skin Thickness, as well as Age and Pregnancies.

In summary, these findings enhance our understanding of the dataset by revealing intricate relationships between health metrics and diabetes prevalence.

MODEL APPLICATION

Because we are dealing with healthcare dataset, we are restricted in our approach to handling dimensionality. However, we shall be using Machine Learning Algorithms

The presence of outliers in our dataset holds significant importance; being a healthcare dataset we consider the relevance of these features otherwise, we might have considered options like outlier trimming, oversampling, or under sampling to achieve a balanced dataset. In this analysis, we applied several Machine Learning algorithms, including Logistic Regression, Random Forest Classifier, Decision Tree, and Support Vector Machine in comparison to K-Nearest Neighbors (KNN).

INSIGHT AND REPORT OF OUR MACHINE LEARNING ALGORITHM

Exploring different Machine Learning algorithms for our model has allowed us to identify the most effective algorithm by leveraging statistical metrics. This approach provides valuable insights into the model's performance. The

algorithms under consideration include Logistic Regression, Decision Tree, Random Forest Classifier, Support Vector Machine, and K-Nearest Neighbors.

We employed the above classification algorithms due to the nature of our dataset (binary). We shall be considering the performance metrics of these algorithm to determine the algorithm that fits our model best.

Comparing our Model Classification Report

ML Algorithm	Classes	Precision	Recall	F1-Score	Accuracy Score	Support
Logistic Regression	0	81%	84%	83%	77%	99
	1	69%	65%	67%		55
	Macro Avg.	75%	75%	75%		
	Weighted Avg.	77%	77%	77%		
Decision Tree Classifier	0	80%	77%	78%	73%	99
	1	61%	65%	63%		55
	Macro Avg.	71%	71%	71%		
	Weighted Avg.	73%	73%	73%		
Random Forest Classifier	0	80%	80%	80%	74%	99
	1	64%	64%	64%		55
	Macro Avg.	72%	72%	72%		
	Weighted Avg.	74%	74%	74%		
Support Vector Machine	**0**	78%	88%	83%	77%	99
	1	72%	56%	63%		55
	Macro Avg.	75%	72%	73%		
	Weighted Avg.	76%	77%	76%		
KNN	**0**	76%	68%	72%	66%	99
	1	52%	62%	56%		55

Macro Avg. Weighted Avg.		64% 67%	65% 66%	64% 66%		

The classification report table above is an important tool in evaluating the performance of our classification model. It provides a comprehensive summary of various performance metrics for each class in our classification problem. The classification report is particularly useful for understanding how well a model is performing across different classes and for gaining insights into its strengths and weaknesses. Here are some key aspects of the importance of a classification report:

Precision, Recall, and F1-Score: The classification report includes precision, recall, and F1-score for each class. These metrics give insights into the model's ability to correctly classify instances of each class, balance between false positives and false negatives, and overall performance.
In the analysis:

Logistic Regression:

Precision: 81%

Recall: 84%

F1-Score: 83%

K-Nearest Neighbor (KNN):

Precision: 76%

Recall: 68%

F1-Score: 72%

Analysis:

Precision:

Logistic Regression achieved a higher precision (81%) compared to KNN (76%), indicating a better ability to correctly identify true positive instances among all predicted positives.

Recall:

Logistic Regression has a higher recall (84%) compared to KNN (68%), suggesting a better ability to capture true positive instances among all actual positives.

F1-Score:

The F1-Score, which balances precision and recall, is higher for Logistic Regression (83%) than for KNN (72%), indicating a better overall trade-off between precision and recall.

Conclusion: The analysis suggests that, across multiple metrics, Logistic Regression outperformed KNN in evaluating the performance of the classification model.

The higher precision, recall, and F1-Score for Logistic Regression indicate its superior ability to correctly classify positive instances and capture all actual positives.

Further consideration of the specific goals and requirements of the task is recommended for a comprehensive evaluation of the models.

Accuracy: While accuracy is a commonly used metric, the classification report emphasizes that accuracy alone may not provide a complete picture, especially in imbalanced datasets Logistic Regression has a high accuracy score of 77% compared to KNN's 66%. Using the classification report we could say that Logistic Regression classified

Support: In the classification report, support is typically listed for each class, and it reflects the number of instances in the true class in the test set. It is an important metric because it provides information about the distribution of instances across different classes and helps interpret the significance of the performance metrics.

Macro and Weighted Averages: The classification report provides macro and weighted averages of precision, recall, and F1-score. These averages give an overall summary of the model's performance, accounting for the imbalance in class sizes.

ROC_AUC_SCORE

ROC AUC (Receiver Operating Characteristic Area Under the Curve) assesses the trade-off between true positive rate and false positive rate across various classification thresholds.

In the analysis:

K-Nearest Neighbor (KNN):

AUC Score: 0.73

Interpretation: Moderate discriminatory power, performing better than random chance (0.5) but not achieving perfect discrimination (1.0).

Logistic Regression:

AUC Score: 0.84

Interpretation: Higher AUC score (0.84) compared to KNN, indicating better discriminatory power and a stronger ability to distinguish between positive and negative instances.

Support Vector Machine (SVM):

AUC Score: 0.80

Interpretation: AUC score of 0.80 suggests good discriminatory power, falling between KNN and Logistic Regression.

Analysis:

The Logistic Regression model outperforms both KNN and SVM in terms of AUC score, indicating superior discriminatory ability.

KNN's AUC score of 0.73 suggests moderate performance, while SVM falls in between with an AUC score of 0.80.

In summary, the AUC scores provide insights into the discriminatory power of each model, with Logistic Regression exhibiting the highest performance among the three algorithms. Further analysis, considering other metrics and the specific requirements of the task, is recommended for a comprehensive evaluation.

CONFUSION MATRIX

Considering the Confusion Matrix metrics across our models, we shall be juxtaposing the performance of Logistic Regression with KNN classification to understand which model evaluation of the instance of the classes is considerable higher.

KNN (k-Nearest Neighbors):

False Positives (FP): 67

True Positives (TP): 34

Logistic Regression:

False Positives (FP): 83

True Positives (TP): 36

Analysis:

False Positives (FP):

KNN has fewer false positives (67) compared to Logistic Regression (83). A lower FP count suggests that KNN is making fewer incorrect positive predictions.

True Positives (TP)

KNN has a lower count of true positives (34) compared to Logistic Regression (36). This indicates that Logistic Regression correctly identifies more positive instances.

In summary, while KNN has a lower false positive count, Logistic Regression achieves a slightly higher true positive count.

In summary, Logistic Regression exhibited strong performance across all metrics when compared to the other models used.