LINEAR DISCRIMINANT ANALYSIS AND ITS CLASSIFICATION EFFECT ON DATA

1. ABSTRACT

PURPOSE: To analyze the effectiveness of Linear Discriminant Analysis (LDA) and classifier assessment techniques in the context of health data related to the risk of early-stage diabetes, by implementing LDA, computing LDA scores, and the calculation of the ROC curves for the given health-related dataset.

METHODS: A comprehensive analysis takes place upon the diabetes (DM) and obesity (OB) datasets by employing robust methods for data preprocessing, utilizing PCA for dimensionality reduction and computing LDA axes and scores. Visualization was completed by including informative scatter plots of LDA scores and 2D representations of the original datasets. Along with this, the computation of Receiver Operating Characteristic (ROC) curves, determining Area Under the Curve (AUC), and identifying optimal thresholds for accuracy also takes place. Essential evaluation metrics, such as confusion matrices, AUC values, and optimal confusion matrices, are computed and displayed. This analysis primarily is based around diverse techniques for classification, dimensionality reduction, visualization, and performance evaluation, and thus, offers a thorough exploration and understanding of the characteristics of DM and OB datasets.

RESULTS: The optimal thresholds for the DB and OM datasets were -0.2985 and 1.2116 respectively. Meanwhile, the Area Under the Curve (AUC) for both the DB and OM datasets is 0.93086 and 0.6071 respectively. Through Recover Operating Characteristic (ROC) analysis, the values of the optimal confusion matrices are also received for both the DB and OM datasets, and the values of these matrices are [255 65; 4 196] and [49 39; 131 301] respectively.

CONCLUSIONS: The optimal decision thresholds maximize certain performance metrics, depending on the specific goal (e.g., accuracy, sensitivity, specificity), and the DB classifier performs well here, with an optimal threshold of -0.2985 and a high AUC of 0.93086, which indicates a better differentiation between positive and negative instances. Meanwhile, the OM classifier has a lower AUC, suggesting less effective discrimination, and its optimal threshold reflects a different trade-off between true positive and false positive rates. Overall, the consideration of both AUC and optimal confusion matrices provides a comprehensive evaluation of classifier performance for the dataset that’s being analyzed.

2. INTRODUCTION

The scientific objective of this assignment is to explore and apply Linear Discriminant Analysis (LDA) and classifier assessment techniques in the context of health data, specifically focusing on the risk of early-stage diabetes.

To understand this report, one must have knowledge with regards to some of the background topics addressed in this report. This report is based upon the relationship of health-related variables, such as measures or biomarkers like clinical obesity, gender, etc., that can be associated with diabetes risk. One of the most important topics that this report is based upon is that of Linear Discriminant Analysis (LDA), which is a dimensionality reduction and classification technique used in data analysis. Its primary goal is to find a linear combination of features that best separates two or more classes in a dataset. This is done by maximizing the distance between the means of different classes while minimizing the spread (variance) within each class, while also projecting the data onto a lower-dimensional space in such a way that the classes become more distinguishable. Another important topic that must be understood by someone reading this report would be that of Receiver Operating Characteristic (ROC) Curves. ROC curves are graphical representations commonly used to assess the performance of classification models, especially in binary classification problems. They evaluate and visualize the performance of classification models, particularly those predicting binary outcomes (positive or negative), while also providing a comprehensive view of the trade-offs between sensitivity and specificity across different classification thresholds. Both LDA and ROC curves are also based around the Confusion Matrix, which is a table that summarizes the performance of a classification model by comparing predicted and actual class labels. Its components include True Positives which are instances correctly predicted as positive, False Positives, which are instances incorrectly predicted as positive, True Negatives, which are instances correctly predicted as negative, and False Negatives, which are instances incorrectly predicted as negative. The LDA can contribute to the calculation of these components, while the confusion matrix provides the raw counts of correct and incorrect predictions, which can come quite handy for the ROC curves.

To test the scientific question, multiple steps took place. Firstly, the data set was analyzed, with an understanding that the matrix size was in a csv file, containing 521 rows and 17 columns. Within this data set, categorical binary values had been coded into +1 and -1. On this data set, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data to 2D, and then, LDA was applied to find the LDA axis and scores for both the "diabetes" and "obesity" labels. From here, separate ROC curves were calculated and plotted for the "diabetes" and "obesity" labels, and this was found by varying the threshold of LDA scores. The Area Under the Curve (AUC) is computed as a measure of the overall performance of the classifier, and the threshold selection takes place by considering the product of True Positive Rate (TPR) and True Negative Rate (TNR), from which the product of False Positive Rate (FPR) and False Negative Rate (FNR) had been subtracted. Lastly, the classification performance at a specific threshold were found through confusion matrices, which were computed as the best choices of thresholds for LDA scores.

3. METHODS

The first goal, as part of the task assigned, was to figure out as to how well the Linear Discriminant Analysis (LDA) separated the data when labelled for diabetes and for obesity. This required a standardization of data to 2D, and then a computation of the LDA axis and scores. To start this process off, the data is loaded and preprocessed. Data is read using the 'csvread' function to load the 521 rows and 17 columns of data from the 'dmrisk.csv' file. The file contains a header, and this header is omitted while reading the data. From here, columns related to diabetes and obesity are defined, in order to isolate specific features related to each condition. This allows for the extraction of the data matrices, which contain information about the respective conditions, and the label vectors, which represent the binary labels, i.e., 1 for positive instances and -1 for negative instances. Through this process, the data within the file is now ready to be analyzed, and the first step that is completed in order to analyze this data is dimensionality reduction by using Principal Components Analysis (PCA). This is achieved through the 'pca' function, which is applied upon the z-score normalized data. Eventually, the dimensionality of the data is reduced to 2D for both diabetes and obesity datasets, and the resulting 2D representations capture the most significant variations in the original data. The dimensionality reduction creates an opportunity for LDA, and therefore, the LDA axis and scores are calculated for both diabetes and obesity datasets. Furthermore, the LDA leads to the Fisher's linear discriminant axis being calculated based on the two datasets, which thus ensures maximum separation between the classes. With the LDA scores and axis now calculated, data visualization can be completed. This is done through scatter plots and 2D plots, wherein the scatter plots are generated to visually represent the LDA scores for both diabetes and obesity datasets, while the 2D plots for diabetes and obesity datasets provide a visual summary of the data after PCA and LDA. Distinct class representations are used to differentiate between positive and negative instances, and the various class representations are shown through different markers or colors.

The second goal in this task was to compute the ROC curves for the classifiers, and to compare the confusion matrices for the best choices of thresholds for the LDA scores. For each data label, diabetes and obesity, a classifier will have variable performance that depends on the selection of a threshold for the LDA scores of the data. Each threshold produces a confusion matrix, and each confusion matrix can be represented by its relative TPR and FPR values. Together, these rates can be plotted as an ROC curve. To implement this, the 'roccurve' function for both diabetes and obesity datasets was useful, as it allowed for the computation of the ROC curves and the AUC values. The ROC curve provided a graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1 - specificity) at various threshold settings. The AUC, meanwhile, was a scalar value representing the area under the ROC curve, which then provided a single metric to assess the classifier's performance. Another important aspect that was calculated by using the 'roccurve' function was that of optimal thresholds for accuracy, as these thresholds played a crucial role in binary classification. Through the optimal thresholds that had been calculated, the confusion matrices were deciphered, using the 'confmat' function. The confusion matrices provided a detailed and granular view of the classification results, as one was able to successfully receive the instances that were correctly or incorrectly labelled as positive or negative. The optimal thresholds ensured that the confusion matrices reflected the classifier's performance at the chosen decision points, as precision, recall, specificity, and F1 scores could be derived from the confusion matrix, which offer a more robust assessment of the classifier's behavior. These results of ROC curves were then visualized, by setting axis labels, titles, legends, and grid lines in the generated plots, and the use of color-coded markers or lines in the plots enhanced the interpretability of the results.

Overall, a comprehensive analysis of diabetes and obesity datasets is conducted, emphasizing discriminative power, classification performance, and scientific reproducibility. The tasks that were given employed Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimensionality reduction, enhancing class separation. Receiver Operating Characteristic (ROC) analysis compute curves, Area Under the Curve (AUC), and optimal thresholds, which thus evaluate discrimination ability. Confusion matrices provides a thorough assessment of the classifier's performance, and scientific reproducibility is given importance through clear documentation and internal code explanations. Visualizations, including scatter plots, facilitate the interpretation of reduced-dimensional datasets. Collectively, all of the topics addressed provide insights into data characteristics, classifier performance, while also ensuring transparency.

4. RESULTS

The results are primarily focusing on the LDA scores, and its primary objective is to find a direction (linear discriminant) that maximizes the separation between different classes in the dataset. In this case, they classify instances into positive or negative categories and represent the optimal decision thresholds determined through Receiver Operating Characteristic (ROC) analysis. Other metrics such as the Area Under the Curve (AUC), is derived from the ROC curve, and it indicates the model's ability to distinguish between positive and negative instances for the diabetes dataset. Lastly, the confusion matrices that are evaluated provide a detailed breakdown of the model's performance in terms of correct and incorrect classifications for each class in the diabetes and obesity dataset. These numerical values are then visualized using ROC curves, LDA scores, and 2D Data Plots. The ROC curves illustrated the trade-off between true positive rate and false positive rate at various decision thresholds, which hence revealed the optimal decision threshold. On the other hand, the LDA scores show how well the linear discriminant axis separates the classes, while the 2D data plots reveal patterns and clusters, making it easier to interpret class distribution after PCA. From a holistic perspective, these graphs collectively offer a visual understanding of the model's discriminatory ability, separation of classes, and the impact of dimensionality reduction on data distribution. They are crucial for assessing the model's performance, identifying optimal decision thresholds, and gaining insights into the inherent patterns within the datasets.

Table 1: The AUC and an "optimal" confusion matrix, computed using LDA, for the diabetes label and the obesity label

Label	AUC	True Positives (TP)	False Negatives (FN)	False Positives (FP)	True Negatives (TN)
Diabetes	0.93086	255	65	4	196
Obesity	0.6071	49	39	131	301

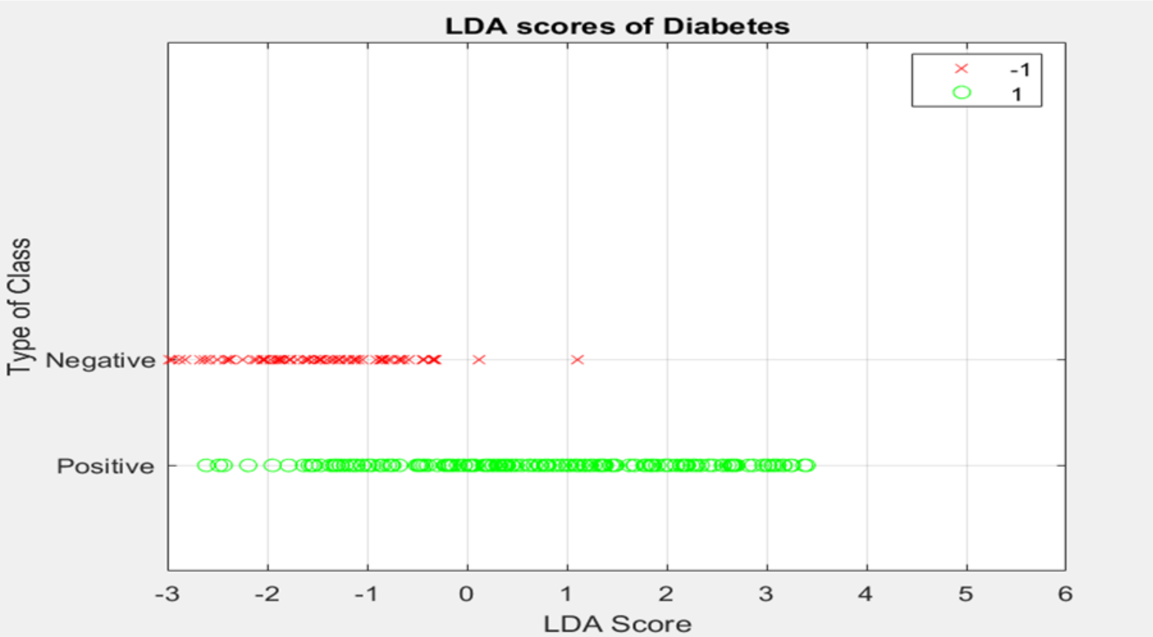

Figure 1: Scatter plot representing the LDA score for the Diabetes column. The y-axis, shows the type of class or the classification of positive and negative instances, while the x-axis, shows the LDA score of each instance.

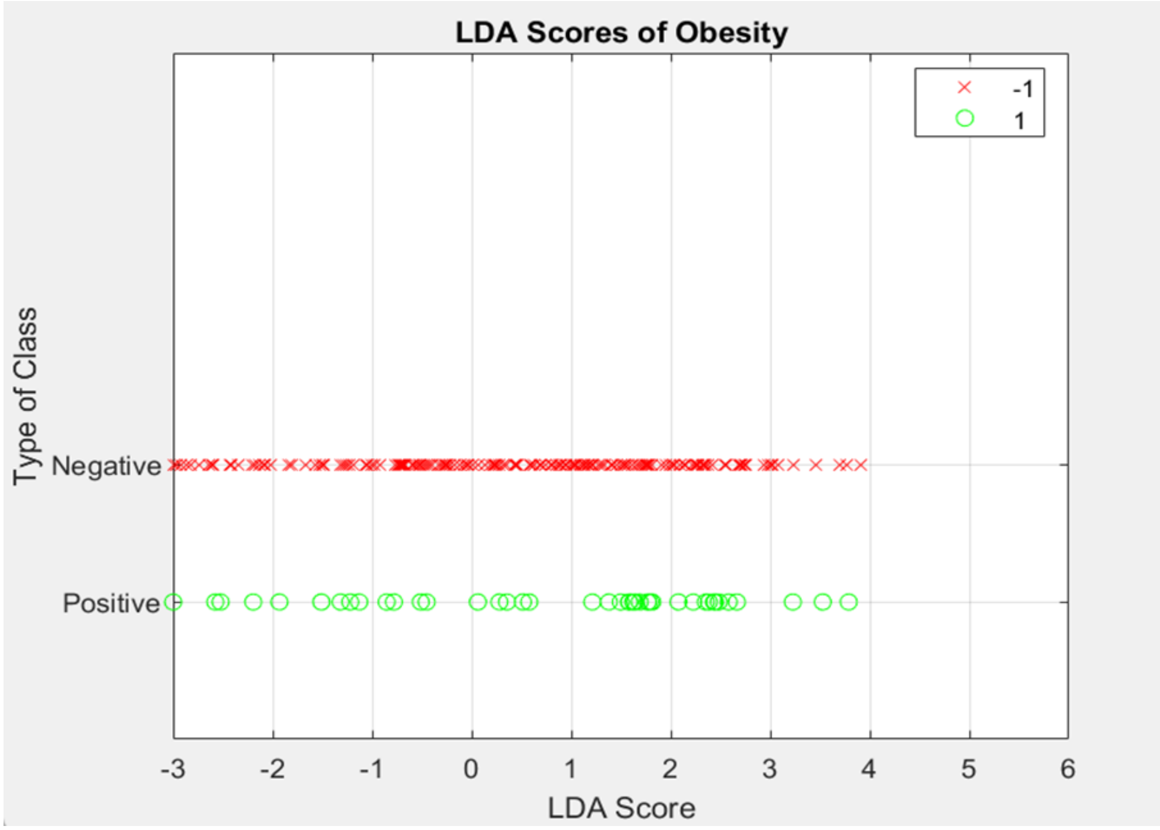

Figure 2: Scatter plot representing the LDA score for the Obesity column. The y-axis, shows the type of class or the classification of positive and negative instances, while the x-axis, shows the LDA score of each instance.

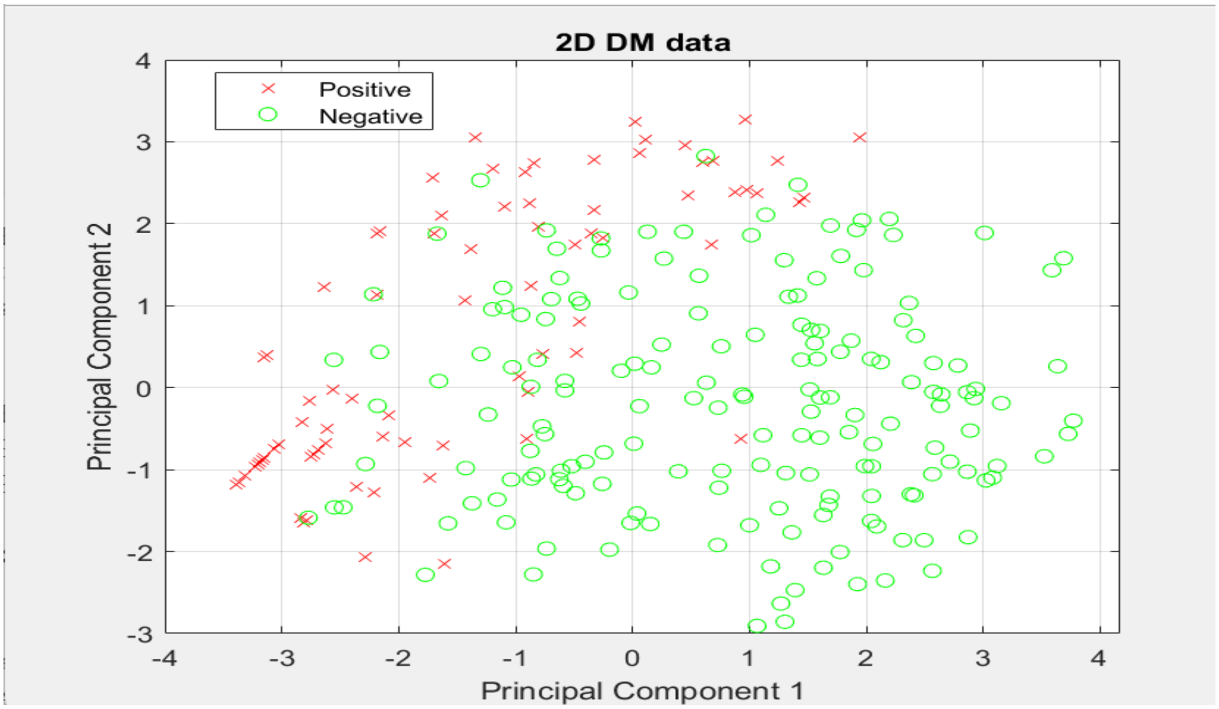

Figure 3: 2D plot representing the principal components of the Diabetes (DM) dataset after performing dimensionality reduction using Principal Component Analysis (PCA). One of these Principal Components is denoted on the x-axis, while the other Principal Component is denoted on the y-axis.

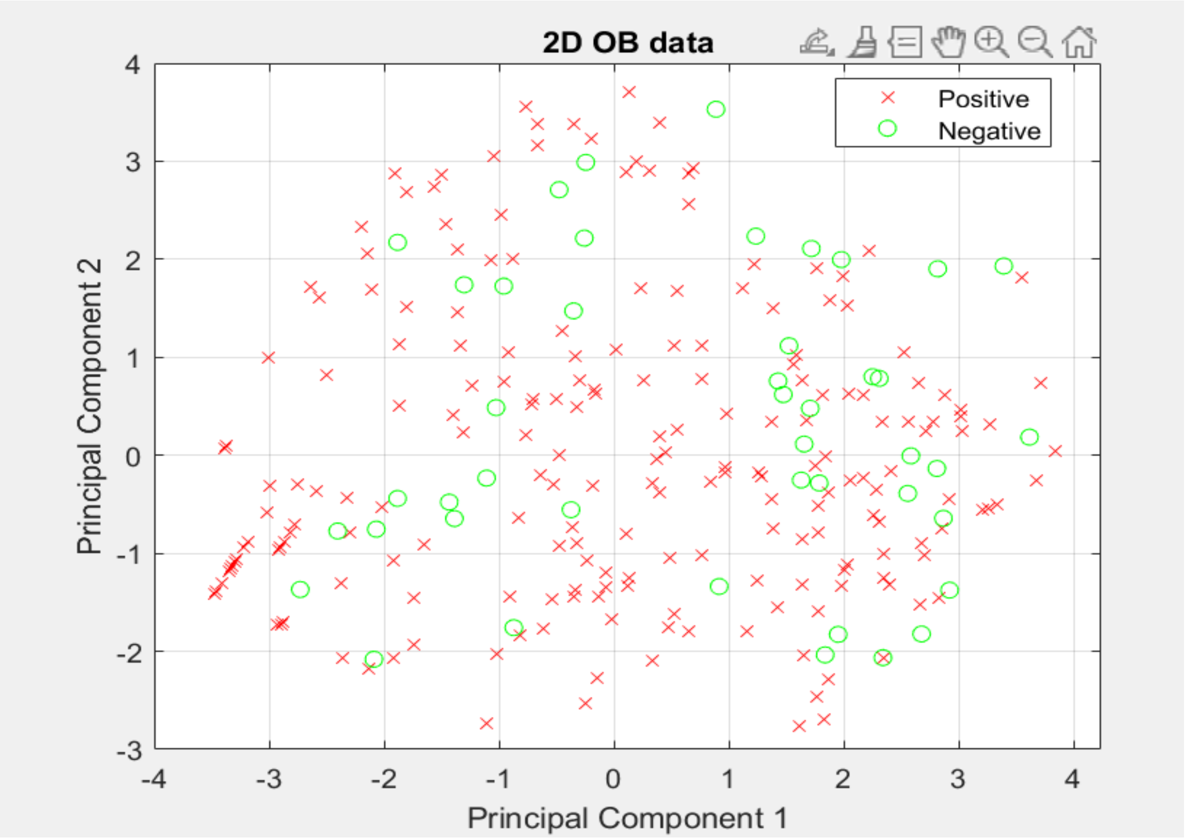

Figure 4: 2D plot representing the principal components of the Obesity (OB) dataset after performing dimensionality reduction using Principal Component Analysis (PCA). One of these Principal Components is denoted on the x-axis, while the other Principal Component is denoted on the y-axis.

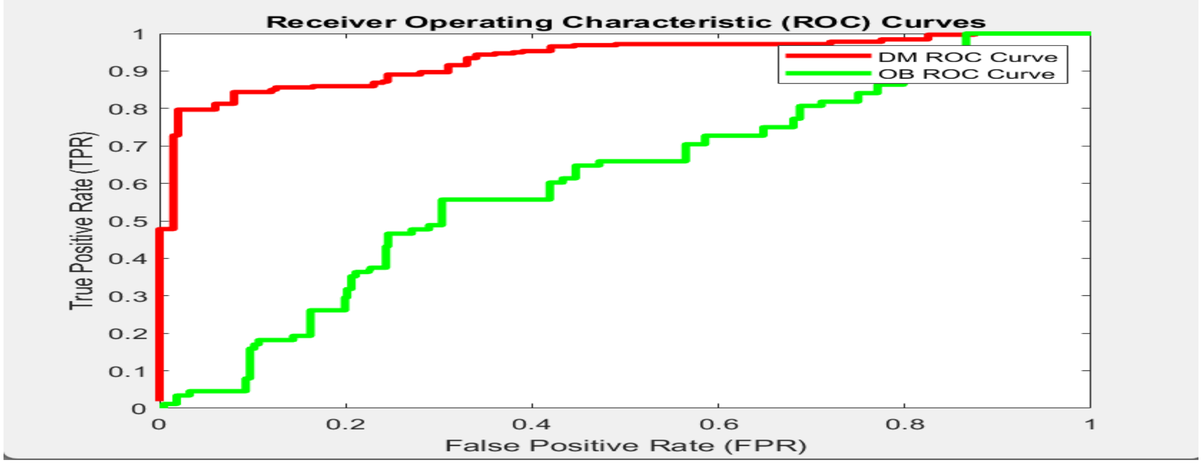

Figure 5: Receiver Operating Characteristic (ROC) curves for the Diabetes (DM) and Obesity (OB) datasets. The y-axis represents the True Positive Rate (TPR) for both of the datasets, while the x-axis represents the False Positive Rate (FPR) for both of the datasets.

5. DISCUSSION

To understand the relationship between the 17 health-related variables, and how they contribute to the labels of diabetes and obesity, the application of linear discriminant analysis (LDA) aimed to delve into the nuanced health-related data pertaining to the early-stage risk of diabetes and obesity. The process involved dimensionality reduction, transforming the original 17 variables into a 2D representation. LDA axes were computed for both diabetes and obesity labels, providing insight into the directions that contribute most significantly to the separation of the data. The subsequent visualization through 2D scatter plots allowed for a tangible understanding of the distribution and clustering of data points after the application of LDA. In the context of assessing classification performance, the generation of receiver operating characteristic (ROC) curves served as a pivotal tool. These curves portrayed the trade-off between true positive rate (sensitivity) and false positive rate, offering a comprehensive view of how varying threshold values for LDA scores influence the classification outcomes. The area under the ROC curve (AUC) quantified the discriminatory power, while the determination of optimal thresholds and the subsequent construction of confusion matrices provided a closer examination of the performance metrics at specific decision boundaries. This analysis revealed the delicate balance between correctly identifying positive instances (true positives) and inadvertently classifying negative instances as positive (false positives). The findings collectively underscore the challenges in achieving a clear demarcation between health categories based on the examined variables. To summarize, the analysis that was completed upon the dataset provided trends upon the LDA scores and LDA axis values, while providing valuable insights into the effectiveness of ROC curves for the comparing of the classifiers and also for comparing confusion matrices for the best choices of thresholds for the LDA scores., and the observations that can be made from these results will be discussed in the following paragraphs.

The first result that was received while completing this task was that the optimal threshold for the Diabetes column was approximately -0.2985. This value represents the specific point on the predicted score scale where the model makes the decision to classify instances into different classes. In binary classification problems, the algorithm assigns a probability or a score to each instance, indicating the likelihood of an instance belonging to one of the two classes (e.g., diabetic or non-diabetic). The choice of the optimal threshold involves a trade-off between sensitivity (true positive rate) and specificity (true negative rate). A lower threshold may increase sensitivity but decrease specificity, and vice versa. In medical contexts, finding the right balance is crucial as it determines how well the model identifies true positives while minimizing false positives and false negatives. The predicted scores generated by the model can be interpreted as a measure of confidence or probability that an instance belongs to a particular class. For the diabetes column, scores above this threshold are assigned to the negative class, and scores below it are assigned to the positive class. In this specific scenario, the negative class is classified as the one that's non-diabetic, while the positive class is the one that's classified as the diabetic class. This can also be seen through the plotting of the scatter plot that took place for the LDA Scores of Diabetes. Here, the negative and positive classes are differentiated from each other on the graph, and there are almost an equal number of negative (261) and positive (259) predicted instances, which can be confirmed through the scatter plot as well as through the confusion matrix. In a clinical setting, the choice of the optimal threshold has implications for patient care. For example, a lower threshold might be preferred if it is more critical to identify as many true positives (diabetic patients) as possible, even if it means accepting more false positives (non-diabetic patients misclassified as diabetic). The relevance of the above points also coincides with the optimal threshold of the Obesity column, which has a value of 1.2116. For the obesity column, since the sign of the optimal threshold value is positive, scores above this threshold are assigned to the positive class, and scores below it are assigned to the negative class. In this specific scenario, the negative class is classified as the one that's non-obese, while the positive class is the one that's classified as the obese class. When the scatter plot is created for the LDA Scores of Obesity, the negative and positive classes are also differentiated from each other on the graph. In this case, however, there are a greater number of negative (340) predicted instances, in comparison to positive (180) predicted instances, which can be confirmed through the scatter plot as well as through the confusion matrix once again. Similar to the diabetes column, selecting a lower threshold for the obesity column emphasizes the prioritization of true positives, indicating accurate identification of obese individuals. This choice is relevant when the primary concern is capturing as many cases of obesity as possible, even if it leads to an increase in false positives. This approach acknowledges the importance of identifying individuals with obesity, outweighing the potential cost of misclassifying non-obese individuals. These optimal threshold values that have been computed for both the Diabetes and Obesity columns also co-relate to the values of the Area Under the Curve (AUC) and the computation of the confusion matrices. The Area Under the Curve (AUC) of the ROC curve is a measure of the classifier's performance. An AUC close to 1 indicates a good classifier, while an AUC of 0.5 suggests a classifier performing no better than random chance. For the Diabetes column, the value of the AUC is 0.93086, which indicates high discriminatory power. As a result, it can be said that the classifier used here is highly effective in distinguishing between individuals with and without early-stage diabetes. When this is shown through a ROC curve, it further becomes clear as to why this is the case. The curve's proximity to the upper-left corner of the ROC space indicates a high true positive rate and a low false positive rate across a range of classification thresholds. This implies that the classifier achieves a strong balance between sensitivity and specificity, making it a reliable tool for diabetes classification. Another aspect by which one can confirm the high discriminatory power of the Diabetes column is through the computation of the confusion matrix that was completed for the Diabetes column. The confusion matrix that was received for the Diabetes data was [255, 65; 4, 196], which indicates that there were 255 correctly identified instances of diabetes (True Positives), 65 instances were falsely classified as non-diabetic (False Negatives), 4 instances were falsely classified as diabetic (False Positives), and 196 instances were correctly identified as non-diabetic (True Negatives). As there are a relatively low number of false positives and false negatives, and since the AUC is high, it is safe to say that the classifier for the Diabetes column performs well. On the other hand, this is not quite the case for the Obesity column. For the Obesity column, the value of the AUC is 0.6071, which indicates a mediocre discriminatory power in comparison to the Diabetes column. As a result, it can be said that the classifier used here is not quite as effective in distinguishing between individuals with and without early-stage obesity. On the ROC curve, the curve's proximity to the center of the ROC space indicates a low true positive rate and a high false positive rate across a range of classification thresholds. This implies that the classifier doesn’t have a strong balance between sensitivity and specificity, making it an untrustworthy tool for Obesity classification. Another aspect by which one can confirm the low discriminatory power of the Obesity column is through the computation of the confusion matrix that was completed for the Obesity column. The confusion matrix that was received for the Diabetes data was [49, 39; 131, 301], which indicates that there were 49 correctly identified instances of obesity (True Positives), 39 instances were falsely classified as non-obese (False Negatives), 131 instances were falsely classified as diabetic (False Positives), and 301 instances were correctly identified as non-diabetic (True Negatives). As there are a relatively high number of false positives and false negatives, and since the AUC is comparatively lower, it is safe to say that the classifier for the Obesity column performs poorly, with respect to the Diabetes column.

In conclusion, the LDA scores, for the Diabetes and Obesity columns, have a value of -0.2985 and 1.2116 respectively. These values assist in the differentiation of the positive and negative classes for both columns. Meanwhile, the AUC values of the Diabetes and Obesity columns is 0.93086 and 0.6071 respectively, which indicates an excellent classification and a mediocre classification for both of the columns respectively. The confusion matrices computed for both of these columns further provide a through analysis and proof of these classifications.

6. REFERENCES

1. Ellis RE. Patterns – Linear Discriminant Analysis, or LDA [Internet]. Queen’s University; 2021 [cited 2024 Mar 13]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class24.pdf
2. Ellis RE. Classification – Assessment With Confusion Matrix [Internet]. Queen’s University; 2021 [cited 2024 Mar 13]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class25.pdf

3. Ellis RE. Classification – Assessment With ROC Curve [Internet]. Queen's University; 2021 [cited 2024 Mar 13]. Available from: https://research.cs.queensu.ca/home/cisc271/pdf/Class26.pdf
4. What is linear discriminant analysis? | IBM [Internet]. www.ibm.com. Available from: https://www.ibm.com/topics/linear-discriminant-analysis#:~:text=Linear%20discriminant%20analysis%20(LDA)%20is
5. Classification: ROC Curve and AUC | Machine Learning Crash Course [Internet]. Google Developers. Available from: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=An%20ROC%20curve%20(receiver%20operating