University Of Prince Edwards Island

Faculty of Computer Science, Data Analysis

STAT-4660 DATA VISUALIZATION AND MINING

Diabetes Dataset

Student Name

1. Mira Samer
2. Jessy Samer
3. Nour Mansour

Lecturer:

Dr. Amira El Ayouti

December 24, 2023

1 Overview of the Dataset

Because the diabetes dataset is pertinent to treating a serious and common health issue, we selected it for our study. Diabetes affects millions of people worldwide and is a common chronic illness with major social ramifications. We hope to learn more about the variables impacting the course of diabetes and create prediction models that will help with early identification and individualized treatment plans by examining and evaluating the diabetes dataset. Comprehensive analysis is facilitated by the extensive supply of information provided by the dataset, which includes patient characteristics, medical measures, and outcomes.

1.1 Information about dataset attributes

Our chosen dataset has nine columns, including the target variable, as listed below:

- Pregnancies: To express the Number of pregnancies
- Glucose: To express the Glucose level in blood
- BloodPressure: To express the Blood pressure measurement
- SkinThickness: To express the thickness of the skin
- Insulin: To express the Insulin level in blood
- BMI: To express the Body mass index
- DiabetesPedigreeFunction: To express the Diabetes percentage
- Age: To express the age
- Outcome: To express the final result 1 is Yes and 0 is No

2 Research Questions

We've decided to use two analysis methods in this project: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). PCA is beneficial as it lowers the dimensionality of the dataset by converting the original features into a new collection of uncorrelated variables (principal components). This reduction reduces problems associated with the curse of dimensionality, improves the interpretability of the model, and helps control computing complexity. Regarding LDA, when the conditions of normality and equal covariance matrices are satisfied, it is known to operate successfully. In some situations, it can perform better than other classification methods, especially if the dataset's dimensionality isn't too large.

3 Visualization Tools

3.1 Histograms

The first visualization tool used was histograms, as they are a graphical representation of the distribution of the dataset. It provides a visual summary of the underlying frequency distribution of continuous or discrete data. Additionally, it helps identify the skewness of the data distribution. A normal distribution has a balanced shape, also known as a bell shape, while skewed distributions are asymmetrical [1]. Regarding the histograms presented in Figures 1 and 2, it is evident that the variables pregnancies, skin thickness, insulin, age, and Diabetes Pedigree Function exhibit right-skewed distributions. In contrast, glucose, BMI, and blood pressure manifest approximately normal distributions. An attempt to address the skewness involved log-transforming the skewed variables. However, it was observed that this transformation adversely impacted the dataset during subsequent analyses utilizing PCA and LDA. Specifically, a few observations underwent a transformation resulting in infinite values, introducing complexities in the analytical procedures.

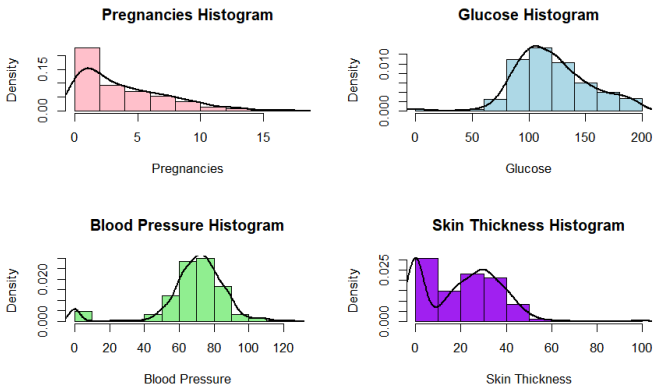

Figure 1: Histograms of Pregnancies, Glucose, Blood Pressure, and Skin Thickness

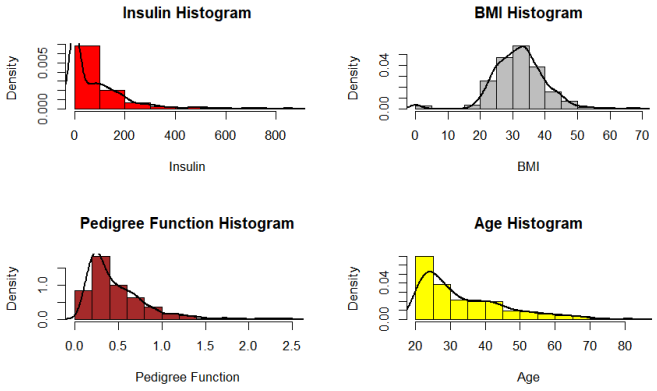

Figure 2: Histograms of Insulin, BMI, Pedigree Function, and Age

3.2 Transformed Histograms

The histograms in Figure 3 show the distribution characteristics of the transformed variables, specifically, insulin, skin thickness, and pedigree function. Following the application of log transformation, discernible changes in the distribution patterns are observed, resulting in a shift towards a more symmetrical and approximately normal distribution for all three variables.

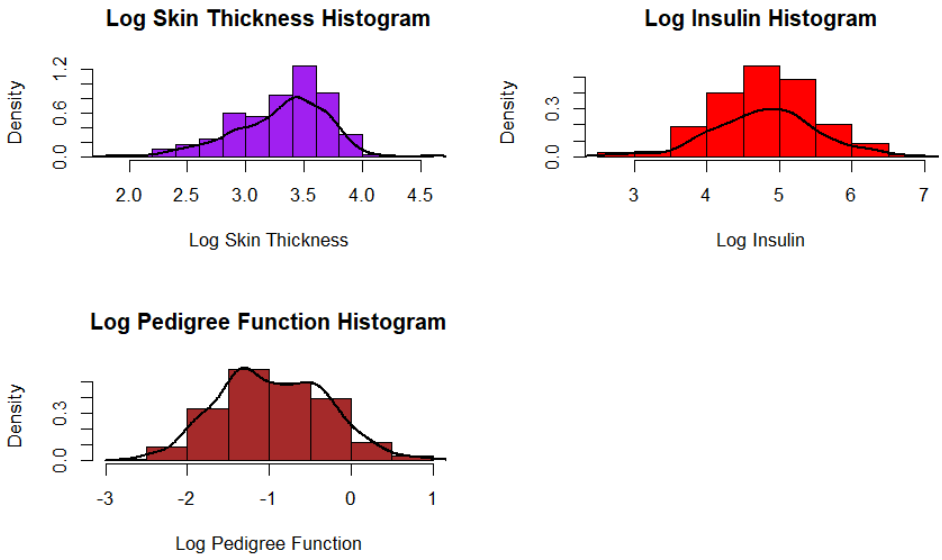

Figure 3: Transformed Histograms of Skin Thickness, Insulin, and Pedigree Function

3.3 Boxplots

We then used boxplots as shown in Figure 5 and Figure 4 as they serve as informative tools for summarizing the distribution of data, revealing central tendencies, and identifying potential outliers, making them widely used in exploratory data analysis and statistical comparisons [2].

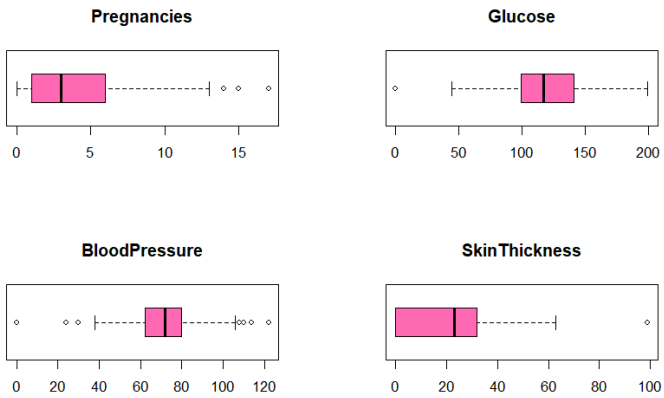

Figure 4: Boxplots of Pregnancies, Glucose, Blood Pressure, and Skin Thickness

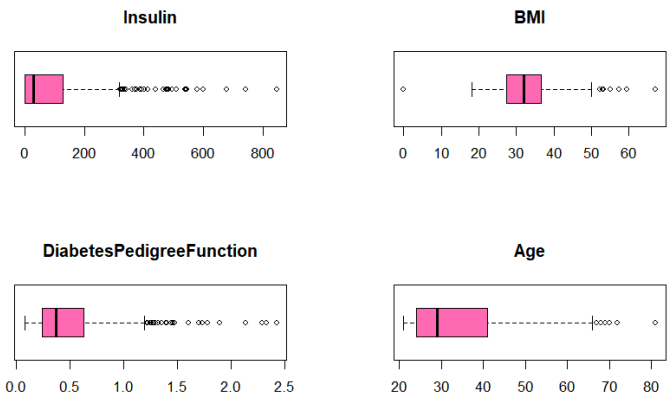

Figure 5: Boxplots of Insulin, BMI, Pedigree Function, and Age

3.4 Pie Chart

This pie chart in Figure 6 depicts the target variable, which is the result, as well as how many observations are diabetic and how many are not. Furthermore, 65.1% were categorized as non-diabetic (Outcome = 0), whereas 34.9% were classified as diabetes (Outcome = 1).

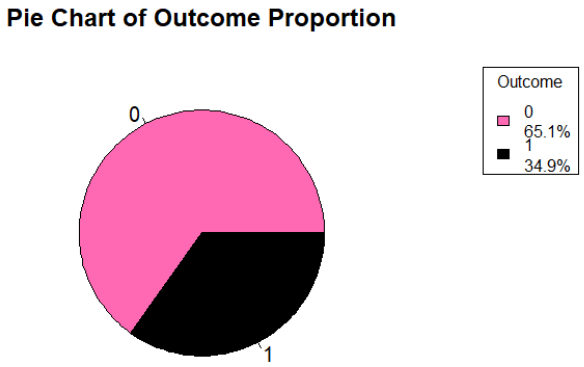

Figure 6: Outcome Pie chart

3.5 Pairwise Scatter Plot

The pairwise plot shown in figure 7 depicts the relationship between all of the features, while the figure below reveals that the bulk of the plots do not correlate. However, skin thickness and BMI show a strong positive relationship.

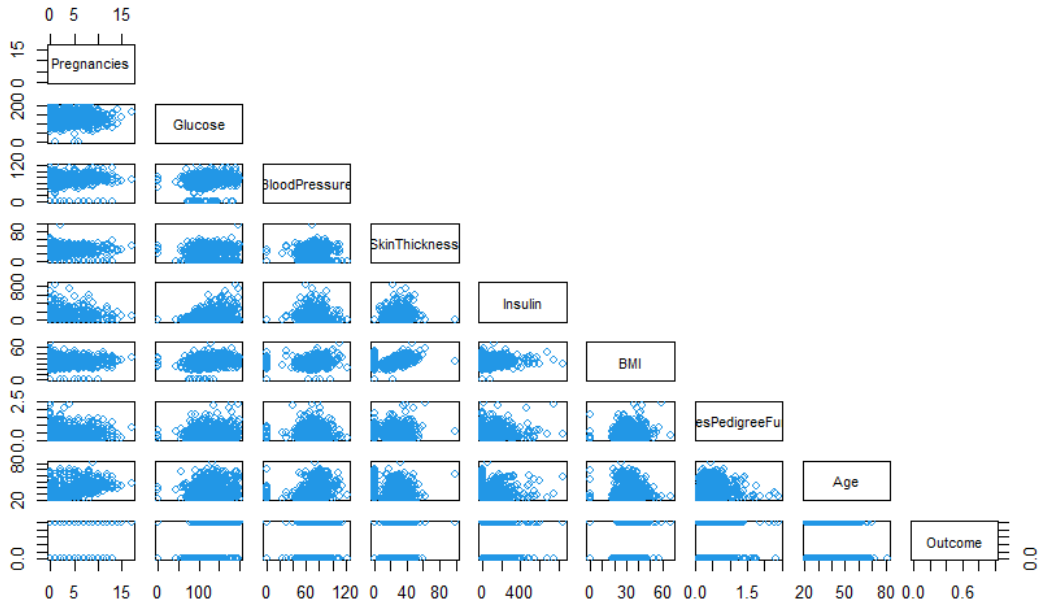

Figure 7: Pairwise Scatter Plot

3.6 Correlation Matrix

The correlation matrix was created to guarantee two things: first, that each feature has a high correlation with the target variable, which in our case is the outcome (the presence or absence of diabetes); as we can see from the correlation matrix, the feature that has the highest influence on the target variable is glucose, with a value of 0.47. The second thing is to check for multicollinearity which is the dependence between the features, It is crucial to select features that provide new information rather than redundant information, and as Figure 8 shows none are dependent on each other.

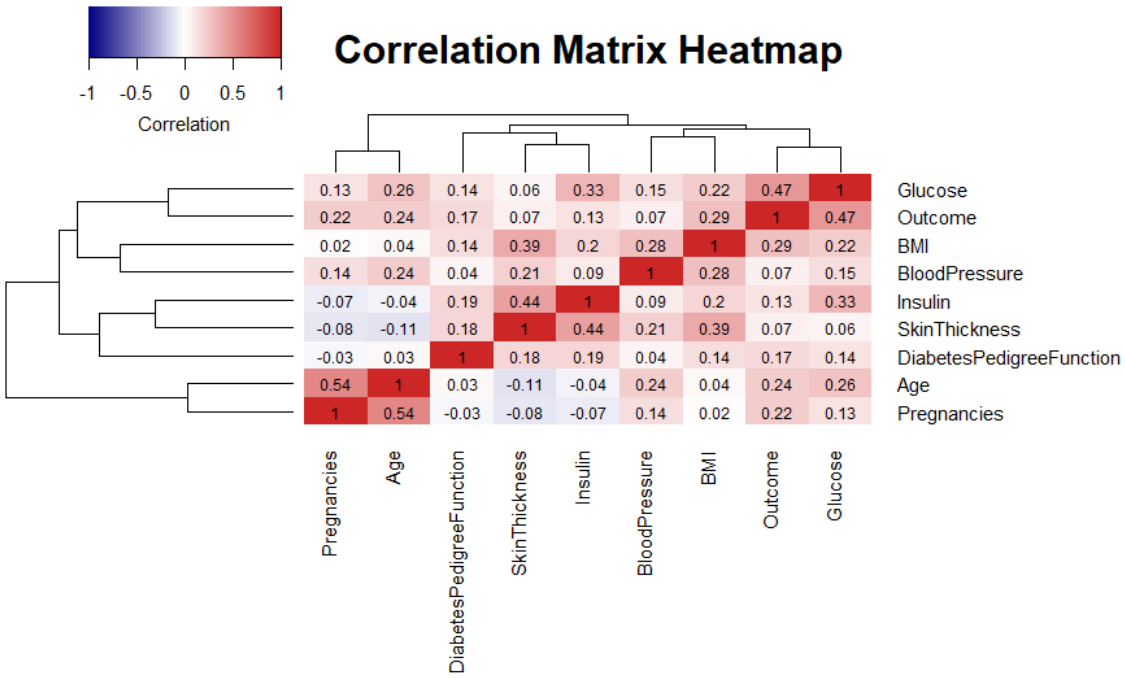

Figure 8: Correlation Matrix

4 Principle Component Analysis

The scale of the variables is taken into consideration by the covariance matrix. Since the variables in the diabetes dataset are measured in different units or scales, using the covariance matrix may result in components that are more affected by variables with larger variances. In contrast, the correlation matrix standardizes the variables by dividing each element by the product of their standard deviations. The correlation matrix becomes scale-independent as a result, and variables are viewed as equal in terms of their contribution to the principal components. Thus, influencing our choice of using the correlation matrix in performing PCA. After applying PCA on the correlation matrix, we can see in Figure 9 that we can stop at Principle Component 3 as it retains more than 60% of the variance in the dataset or in other words, this represents the proportion of total variance in the data that is explained by each principal component. Additionally, the loadings are the weights applied to each original variable in the principle component construction. These weights represent each variable's contribution to the primary component. The direction of the link between the original variables and the principal component is shown by positive and negative loadings. Variables with high positive or negative loadings have a significant impact on the main component. For example, Principle Component (PC) 1, seems to be a very weighted average component, with all loadings being around the same range. Variables pregnancies and age have the highest effect on PC 2. To summarize, the first few components (PC 1 and PC 2) explain a significant portion of the total variance, and the cumulative proportion increases steadily until reaching PC 3 as it explains a substantial amount of variance.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.4471973 1.3157546 1.0147068 0.9356971 0.87312335
## Proportion of Variance 0.2617975 0.2164013 0.1287037 0.1094411 0.09529305
## Cumulative Proportion  0.2617975 0.4781988 0.6069025 0.7163436 0.81163667
##                            Comp.6     Comp.7     Comp.8
## Standard deviation     0.82621328 0.64793223 0.63597331
## Proportion of Variance 0.08532855 0.05247702 0.05055776
## Cumulative Proportion  0.89696522 0.94944224 1.00000000

##
## Loadings:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## Pregnancies               0.128  0.594                0.476  0.194  0.589
## Glucose                   0.393  0.174 -0.468 -0.404 -0.466
## BloodPressure             0.360  0.184  0.535        -0.328 -0.634  0.192
## SkinThickness             0.440 -0.332  0.238         0.488        -0.282
## Insulin                   0.435 -0.251 -0.337 -0.350  0.347 -0.271  0.132
## BMI                       0.452 -0.101  0.362        -0.253  0.685
## DiabetesPedigreeFunction  0.271 -0.122 -0.433  0.834 -0.120
## Age                       0.198  0.621                0.109        -0.712
##                          Comp.8
## Pregnancies               0.118
## Glucose                   0.450
## BloodPressure
## SkinThickness             0.566
## Insulin                  -0.549
## BMI                      -0.342
## DiabetesPedigreeFunction
## Age                      -0.212
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.125  0.125  0.125  0.125  0.125  0.125  0.125  0.125
## Cumulative Var  0.125  0.250  0.375  0.500  0.625  0.750  0.875  1.000
```

Figure 9: Principal Component Analysis

4.1 Scree Plot

According to the scree plot in Figure 10, the elbow shape occurs at a point where the eigenvalue size starts decreasing at a very small rate. The red line represents the occurrence of the elbow shape thus choosing our principle components to be three.

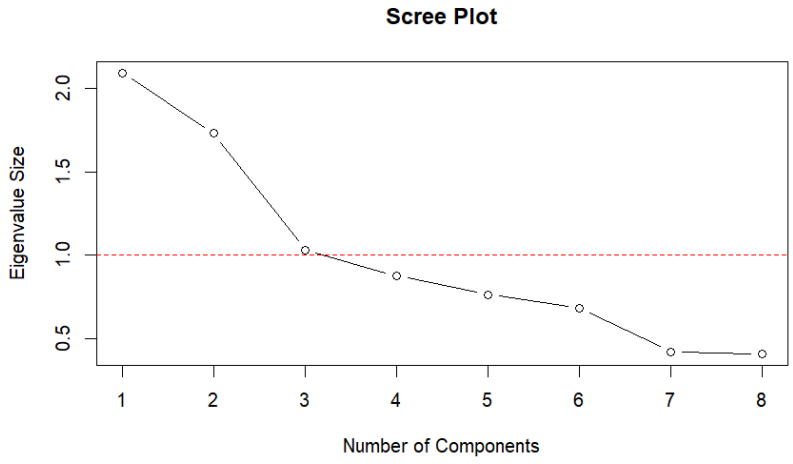

Figure 10: Scree Plot

4.2 Biplot

The biplot in Figure 11 shows that age, pregnancies, blood pressure, and glucose are all positively loaded on principle components 1 and 2. Furthermore, the upper region in the first quarter of the data reveals high levels of pregnancies and high age, the observations in the lower region of the first quadrant show low levels of glucose and blood pressure. Moreover, BMI, DiabetesPedigreeFunction, Insulin, and skin thickness are positively loaded on component 1 and negatively loaded on component 2. Observations that occur in the fourth quadrant have high levels of these variables.

The biplot in Figure 12 shows that blood pressure, BMI, and skin thickness are all positively loaded on principle components 1 and 3 while, pregnancies and age are only loaded on component 1. On the other hand, insulin, glucose, and diabetes pedigree function are negatively loaded on component 3 and positively loaded on component 1. Furthermore, the observations in the first quadrant show high levels of blood pressure, BMI, and skin thickness.While the observations in the fourth quadrant show high levels of insulin, glucose, and diabetes pedigree function.

The biplot in Figure 13 shows that blood pressure is positively loaded on principle components 2 and 3, while pregnancies and age are only loaded on Component 2. In contrast, the variables insulin and diabetes pedigree function are negatively loaded on both components. Furthermore, skin thickness and BMI are negatively loaded on Component 2 and positively loaded on Component 3, and vice versa for glucose. Furthermore, the upper region in the first quarter of the data reveals high levels of blood pressure, and the observations in the lower region of the fourth quadrant show low glucose levels. Observations that occur in the second quadrant have high levels of BMI and skin thickness. Last but not least, high levels of insulin and diabetes pedigree function are evident in observations found in the third quadrant.

The 3Dbiplot in Figure 14 shows that the most influential component in classifying an observation as diabetic or not is component 3 and we chose to retain the first three components only as they retain most of the information of the data while reducing the dimensionality from 8 to 3.

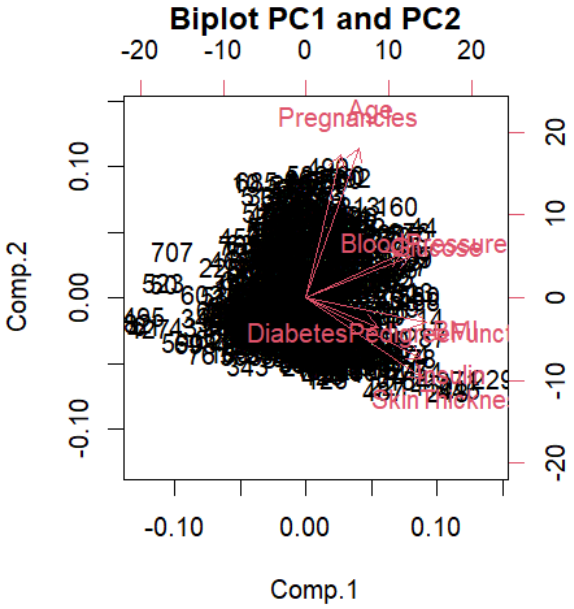

Figure 11: Biplot of PC1 and PC2

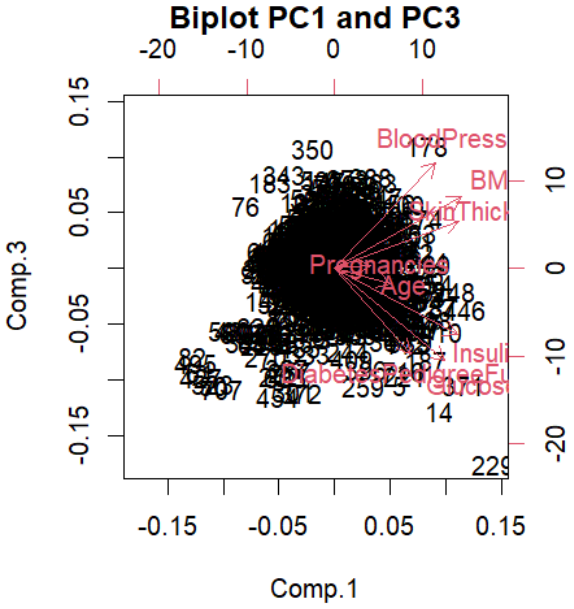

Figure 12: Biplot of PC1 and PC3

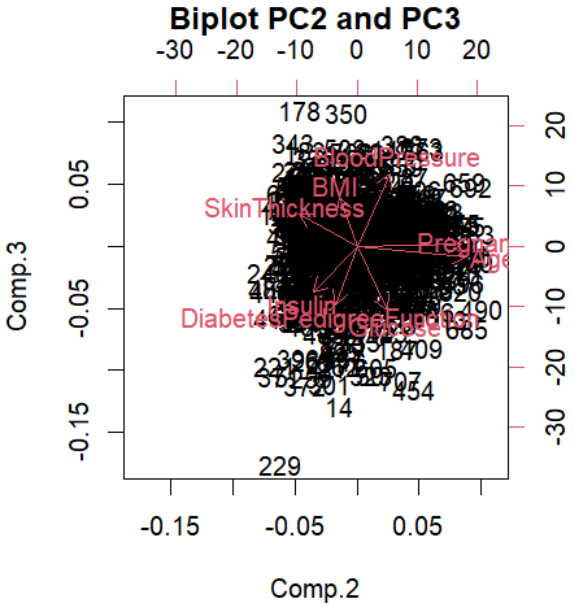

Figure 13: Biplot of PC2 and PC3

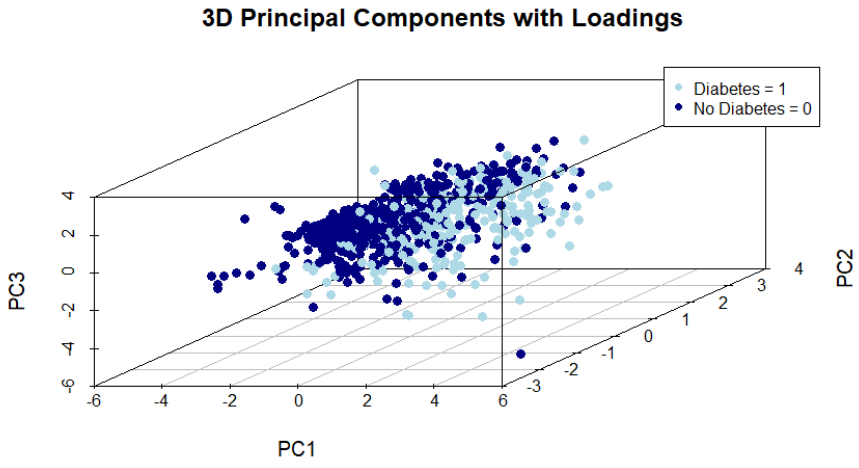

Figure 14: 3D Biplot of PC1, PC2, and PC3

5 Linear Discriminant Analysis

Before performing Linear Discriminant Analysis, two assumptions must be checked, normality and equal covariance matrices. This section represents the chi-squared plot to test the first assumption, which is the multivariate normality of the data. However, as Figure 15 demonstrates, the normality assumption is invalid, however further analysis was made using LDA to stay within the scope of this course.

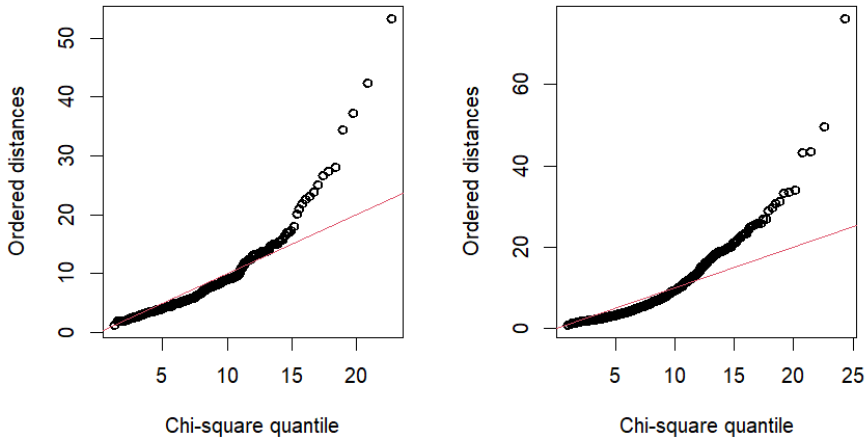

Figure 15: Chi-squared Plot

The second assumption to be checked is the equality of the covariance matrices. As shown in Figure 16, the p-value is less than alpha (0.5), we reject the null hypothesis, meaning that the assumption of the equality of the covariance matrices is violated.

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  diabetes[, -9]
Chi-Sq (approx.) = 226.71, df = 36, p-value < 2.2e-16
```

Figure 16: BoxM Test

Before performing LDA, splitting the data into training and testing sets was done to avoid the underestimation of the misclassification rate, which usually happens when we test on the same data used in creating the classification function.

After performing LDA, the following results were achieved:

- True Positive (TP): 31 instances were correctly predicted as class 1.
- True Negative (TN): 88 instances were correctly predicted as class 0.
- False Positive (FP): 24 instances were incorrectly predicted as class 1.

- False Negative (FN): 11 instances were incorrectly predicted as class 0.

The overall accuracy is a measure of how often the model makes correct predictions. It is calculated as the sum of TPs and TNs divided by the total number of instances. Our findings in Figure 17 show that the model correctly predicted the class for approximately 77.2% of the instances in the dataset. This Linear Discriminant Analysis (LDA) was performed using equal priors.

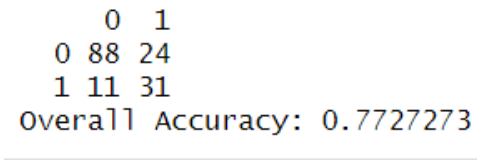

Figure 17: LDA Using Equal Priors

- True Positive (TP): 28 instances were correctly predicted as class 1.
- True Negative (TN): 93 instances were correctly predicted as class 0.
- False Positive (FP): 22 instances were incorrectly predicted as class 1.
- False Negative (FN): 11 instances were incorrectly predicted as class 0.

Our findings in Figure 18 show that the model correctly predicted the class for approximately 78.6% of the instances in the dataset, giving us the highest accuracy.

This Linear Discriminant Analysis was performed using prior proportions of the data, as shown in Figure 19. The figure also shows that glucose has the highest effect on this model for classification, given that it has the highest coefficient.

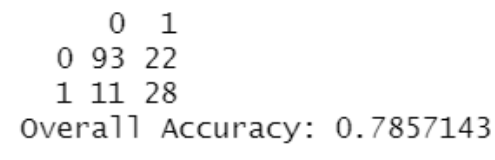

Figure 18: LDA Using Default Priors

```
Call:
lda(Outcome ~ ., data = train_data)

Prior probabilities of groups:
        0         1
0.6416938 0.3583062

Group means:
  Pregnancies    Glucose BloodPressure SkinThickness     Insulin
0  -0.2101142 -0.3586068   -0.07206152   -0.01199085 -0.05426568
1   0.2995904  0.6112053    0.04738932    0.11852988  0.18609190
         BMI DiabetesPedigreeFunction        Age
0 -0.1972006               -0.1009109 -0.2423441
1  0.4156970                0.2132354  0.2918175

Coefficients of linear discriminants:
                                LD1
Pregnancies               0.30066693
Glucose                   0.85975360
BloodPressure            -0.20292935
SkinThickness             0.00452411
Insulin                  -0.11662112
BMI                       0.47661297
DiabetesPedigreeFunction  0.17961870
Age                       0.21148754
```

Figure 19: LDA Using Default Priors Analysis

5.1 Quadratic Discriminant Analysis

Since the assumption of equal covariance matrices was violated we attempted the quadratic discriminant analysis. Similar to LDA we performed the analysis on both equal priors and the actual priors. The results are shown in Figure 20.

```
Call:
qda(Outcome ~ ., data = train_data, prior = c(0.5, 0.5))

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
  Pregnancies    Glucose BloodPressure SkinThickness    Insulin        BMI DiabetesPedigreeFunction
0  -0.1518643 -0.3277927   -0.03649958   -0.06275647 -0.0914089 -0.1987302               -0.1283325
1   0.2869423  0.5800537    0.03341761    0.09318312  0.1800613  0.4072257                0.2539950
         Age
0 -0.1514672
1  0.2849313
```

Figure 20: QDA Analysis

- True Positive (TP): 32 instances were correctly predicted as class 1.
- True Negative (TN): 85 instances were correctly predicted as class 0.
- False Positive (FP): 18 instances were incorrectly predicted as class 1.
- False Negative (FN): 19 instances were incorrectly predicted as class 0.

Our findings in Figure 21 show that the model correctly predicted the class for approximately 76% of the instances in the dataset.

This Quadratic Discriminant Analysis was performed using equal prior proportions.

Figure 21: QDA Using Equal Priors

- True Positive (TP): 34 instances were correctly predicted as class 1.
- True Negative (TN): 82 instances were correctly predicted as class 0.
- False Positive (FP): 16 instances were incorrectly predicted as class 1.
- False Negative (FN): 22 instances were incorrectly predicted as class 0.

Our findings in Figure 22 show that the model correctly predicted the class for approximately 75.3% of the instances in the dataset.

This Quadratic Discriminant Analysis was performed using equal prior proportions.

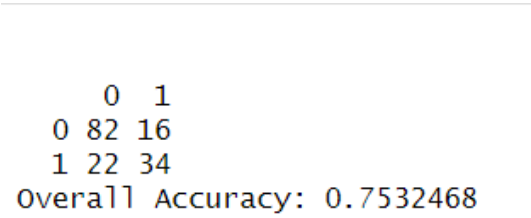

Figure 22: QDA Using Default Priors

6 Comparison

Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are both dimensionality reduction techniques, although they serve different purposes and have significant distinctions. To begin, their primary goals differ: PCA concentrates on maximizing the variance of the data, seeking to preserve as much information as possible within fewer dimensions, whereas LDA prioritizes class separation, improving discriminatory power for classification tasks. Second, PCA is an unsupervised approach that evaluates the full dataset regardless of class labels. In contrast, LDA is a supervised method that considers class details, making it particularly useful for classification tasks [3]. Last but not least, the two techniques give different results: PCA creates uncorrelated principle components, each reflecting a linear combination of the original characteristics, while LDA produces linear combinations that maximize the differences between class means while minimizing variances within classes.

7 Conclusion

In conclusion, using a mix of exploratory data analysis and advanced statistical approaches, we were able to get useful insights from our comprehensive examination of the dataset, which focused on predicting diabetes outcomes. The dataset comprised nine attributes, including key factors such as pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI, diabetes pedigree function, and age, all contributing to the binary outcome variable indicating the presence or absence of diabetes.

Our research topics focused on the use of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) to effectively reduce dimensionality and improve model interpretability. Furthermore, visualization methods such as histograms, boxplots, pie charts, and pairwise plots were useful in analyzing variable distribution patterns and relationships.

Principal Component Analysis (PCA) revealed that the top three components captured more than 60% of the variation in the dataset, resulting in a more streamlined and interpretable representation of the data. The loadings on these components revealed the influential characteristics that were driving the principal sources of variation.

Linear Discriminant Analysis (LDA) has developed as an effective classification method. Despite encountering a violation of the assumption of equal covariance matrices, LDA with the actual proportions as priors demonstrated the highest testing accuracy of 78.6%.

Visualization approaches such as the scree plot and biplot were critical in finding the ideal number of components for PCA and comprehending variable correlations, respectively. Furthermore, utilizing the chi-squared plot to investigate the dataset's multivariate normality yielded crucial insights, even though the assumption was deemed invalid. As an alternative, Quadratic Discriminant Analysis (QDA) was examined, however it did not outperform the chosen LDA model.

Our findings highlight the value of a methodical and iterative approach to data analysis, emphasizing the practical necessity of identifying and changing underlying assumptions. As we move forward, these findings can help to feed

future research and perhaps advise healthcare practitioners in building diabetes predicting models based on similar datasets.

8 References

[1] "What are histograms?" [Online]. Available: https://asq.org/quality-resources/histogram

[2] Admin, "Box plot (definition, parts, distribution, applications amp; examples)," Mar 2021. [Online]. Available: https://byjus.com/maths/box-plot/

[3] S. K. Vungarala, "Pca vs lda - no more confusion!" Apr 2023. [Online]. Available: https://medium.com/@seshu8hachi/pca-vs-lda-no-more-confusion-fc21fb8d06e9#:~:text=PCA%20is%20an%20unsupervised%20method%20that%20aims%20to%20find%20the,the%20characteristics%20of%20the%20dataset.