# Analysis of Pima Indians Diabetes Database

**Yuanyuan Gao, Yirong Xu, Faith Liu, Sabrina Zhu, Shu Hong**
University of California, Davis
STA 135 - Multivariate Data Analysis
Professor Xiucai Ding
June 13, 2023

## 1 Introduction

The Pima Indians of the United States have an unusually high prevalence of diabetes [1]. Predicting the onset of diabetes in this population can help identify high-risk groups and implement focused prevention programs. The goal of the study is to use medical record data to predict the onset of diabetes in female Pima Indians. We will use Box's M test, Quadratic Discriminant Analysis (QDA), decision boundary analysis, and Principal Component Analysis (PCA) to assess covariance differences, classify individuals, visualize decision boundaries, and identify variables influencing diabetes predictions. The findings of this study could help to target interventions and improve healthcare for this high-risk population.

## 2 Research Question

1. Which statistical method is the most appropriate the dataset classification?
2. How can we simplify the dataset without losing much information? How to determine which variables we attain and which ones should be dropped? What is the reduced dataset looks like?
3. What is the decision boundary we use to classify the dataset? How to compute the boundary?
4. What is the confusion matrix, and what insights does it provide about the model's performance?

## 3 Data Description and Data Visualization

Our dataset is sourced from the "PimaIndiansDiabetes2" built-in package in R. It is consist of 768 observations and 9 variables: Pregnant, Glucose, Pressure, Triceps, Insulin, Mass, Pedigree, Age, and an indicator variable for Diabetes. The target variable, Diabetes, is a binary categorical variable with values 1 (negative for diabetes) and 2 (positive for diabetes). The remaining variables, except for Diabetes, are numerical.

According to R Documentation, we know that "while the UCI repository index claims that there are no missing values, closer inspection of the data shows several physical impossibilities, e.g., blood pressure or body mass index of 0. In PimaIndiansDiabetes2, all zero values of glucose, pressure, triceps, insulin and mass have been set to NA, see also Wahba et al (1995) and Ripley (1996)" [2]. Therefore, to ensure data integrity and accuracy, we first remove any rows containing missing values (NA values) from the dataset since it would allows us to work with a complete and reliable set of data for our analysis.

| Attribute | Description |
|---|---|
| Pregnant | Number of times pregnant. |
| Glucose | Plasma glucose concentration (mg/dL). |
| Pressure | Diastolic blood pressure (mmHg). |
| Triceps | Triceps skinfold thickness (mm). |
| Insulin | 2-Hour serum insulin (μU/ml). |
| Mass | Body mass index (BMI) is defined as the ratio of weight in kilograms to the square of height in meters ($kg/m^2$). |
| Pedigree | Diabetes pedigree function. |
| Age | Age in years. |
| Indicator Variable Diabetes | Whether the individual has diabetes or not (Negative in group 1 and Positive in group 2). |

Figure 1: Dataset Description

Then, we use this dataset for performing classification tasks and predictive modeling, specifically to determine whether an individual has diabetes based on the given factors. To describe the relationships among the 8 explanatory variables, we plotted the paired panel plot (Figure 2), commonly known as a scatter matrix. Within this matrix, the diagonal shows histograms for each variable, while the lower triangle portion provides scatter plots illustrating the relationship between pairs of variables, and the upper triangle portion shows the correlation values between the variables.

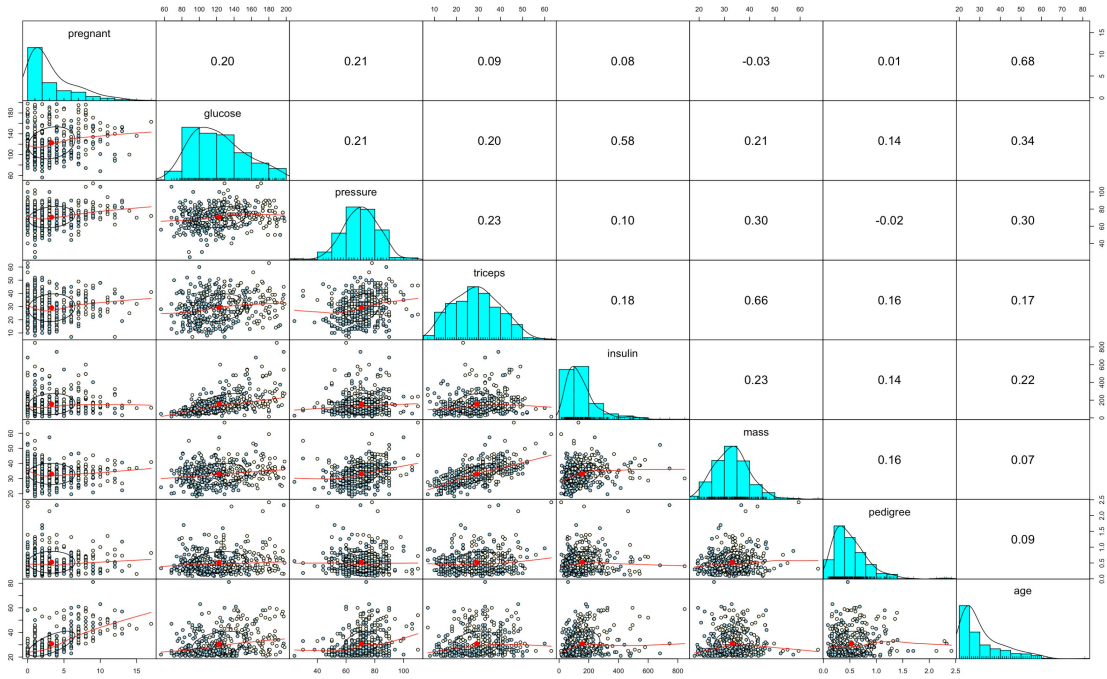


Figure 2: Paired Panel Plot

We conducted an analysis of the paired panel plot (Figure 2), focusing on three portions: histogram, scatter plot, and correlation matrix. Regarding the histograms, we observed that most variables displayed a normal distribution or a right-skewed distribution. The scatter plots did not exhibit any obvious patterns and there were no clear clusters in the data. We further investigated the correlation matrix and found that none of the variables exhibited a high correlation (above 0.7). This suggests a lack of strong correlation among the explanatory variables. This implies that the explanatory variables are not strongly correlated with each other. Consequently, there is no need to remove any variables. We will conduct further research to gain a deeper understanding of the dataset.

# 4 Box's M test

After observing the relationships and correlations in the paired panel plot, we have decided to conduct a Box's M test to get valuable insight to our analysis. This statistical test will enable us to further investigate the dataset and assess the equality of covariance matrices among groups or variables. Box's M test is appropriate in this case to determine whether the co-variances (denoted as $\sum_i$) of the two groups (negative diabetes 1 and positive diabetes 2) are equal. We firstly conducted a hypothesis test, with the null hypothesis being $H_0$: $\sum_1 = \sum_2$, and the alternative hypothesis being $H_a$: $\sum_1 \neq \sum_2$. We then obtained a chi square value of 164.12 and a p-value of $2.82 \times 10^{-18}$ based on the Box's M test. Since our p-value is extremely close to zero, we would reject the null hypothesis and conclude that the co-variances are different.

**Rule:** If the covariance matrices are found to be significantly different, we apply Quadratic Discriminant Analysis (QDA). In contrast, if the covariance matrices do not exhibit significant differences, we use Linear Discriminant Analysis (LDA), which will be elaborated in the next section.

# 5 Quadratic Discriminant Analysis (QDA)

## 5.1 Confusion Table

Next, we continue to use discriminant analysis methods to solve classification problems. Considering the two methods, linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), we would choose to perform QDA. This decision is based on the results of the previous Box's M test, where we concluded that the covariance is different. This is consistent with the assumptions required for QDA, making it a better choice in this context. We first split the data into $70\%$ training set and $30\%$ testing set. The 70% of the data is used for the QDA classification model. To evaluate the performance of the model, we examined the error rates calculated by the confusion matrices (Table 1 and Table 2) for training and testing data. The error rate is computed through the function:

$$ErrorRate = 1 - \frac{n_{11} + n_{22}}{n_1 + n_2}$$

The error rate for the training data is $0.2071$. This indicates that the model performs reasonably on the training data as it is neither too large to suggest inaccuracies nor too small to imply overfitting. In general, this error rate is acceptable. The remaining 30% of the data is used as the test set, which allows us to determine the error rate of the model. We observed the error rate to be $0.1786$, which is small and closed to that of the training data. This indicates that the model is effective in making accurate predictions beyond the training data.

| | | Predicted Group | |
|---|---|---|---|
| Actual Group | Observations | 1 | 2 |
| 1 | 195 ($n_1$) | 162 ($n_{11}$) | 33 ($n_{12}$) |
| 2 | 85 ($n_2$) | 25 ($n_{21}$) | 60 ($n_{22}$) |

Table 1: Training Confusion Table

| | | Predicted Group | |
|---|---|---|---|
| Actual Group | Observations | 1 | 2 |
| 1 | 75 ($n_1$) | 65 ($n_{11}$) | 10 ($n_{12}$) |
| 2 | 37 ($n_2$) | 10 ($n_{21}$) | 27 ($n_{22}$) |

Table 2: Testing Confusion Table

## 5.2 Partition Plot

We will endeavor to lower the error rate to improve the accuracy of our prediction. We want to see if we can get a smaller error rate if we use only two of these eight variables to predict. The partition plot in Figure 3 displays the figure and classification error rate for each pair of variables. We mainly focused on numbers in black color and numbers in red color. The black numbers in each sub-graphs represent correct classification, while the red ones represent the incorrect classification. By observing the sub-plots in Figure 3, we find that the amount of red numbers is not small. Each small graph has an error rate at the top of graph, and the error rates are all around 0.25-0.3, which are greater than the error rate of the QDA model, that is 0.1786. As a result, we may conclude that the QDA model, which utilizes $70\%$ of the data, is ideal in this case.

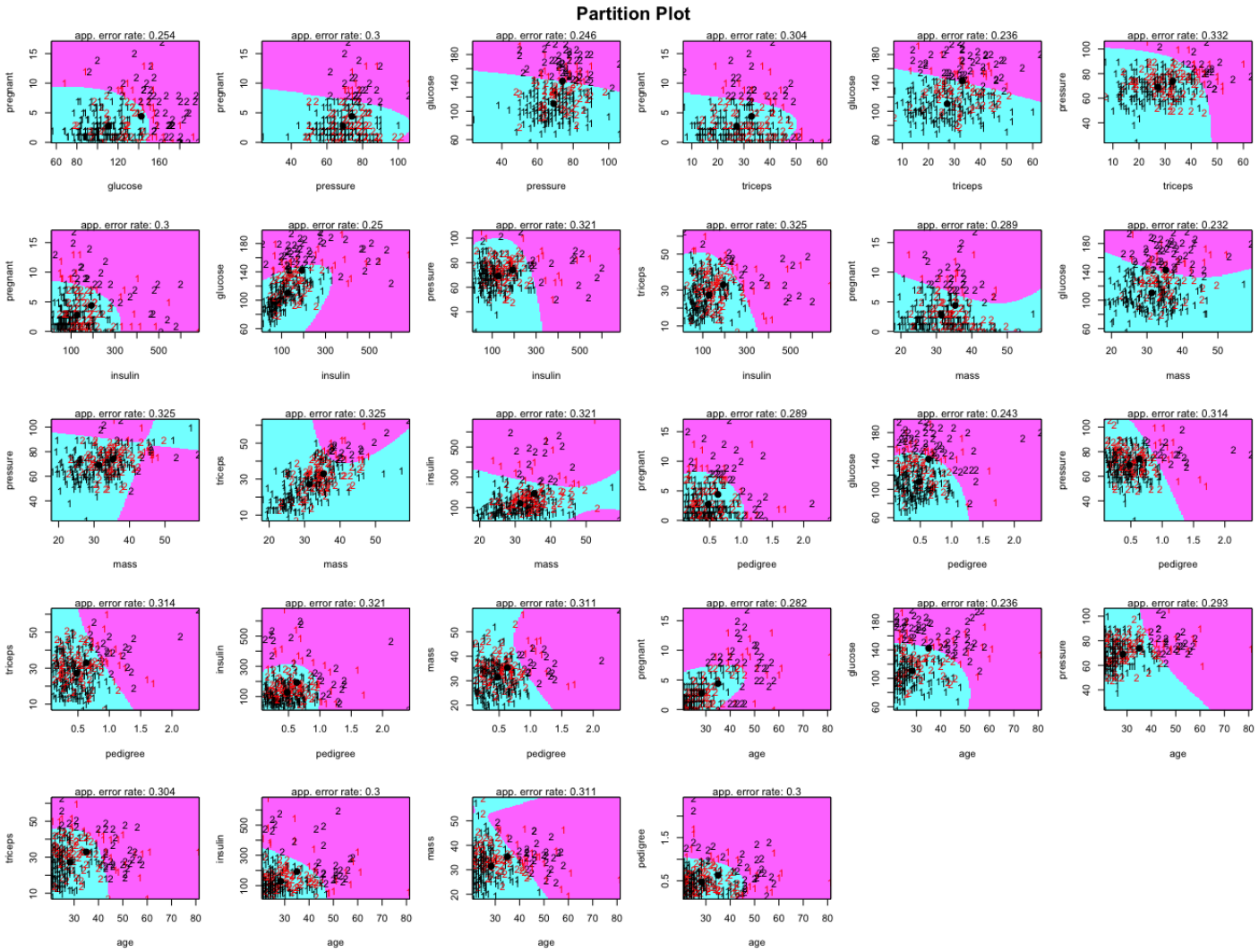


Figure 3: The Partition Plot

### 5.3 Decision Boundary $S_1(X) - S_2(X)$

Having applied the QDA model to analyze the dataset, we now proceed to derive the decision boundary, which can classify new individuals into appropriate groups based on their variable values. To represent the new individual, we'll use vector x, where x holds the values of the 8 variables in this case. The decision boundary is as follows.

$$S_1(X) - S_2(X) = \frac{1}{2}X^T * (\Sigma_2^{-1} - \Sigma_1^{-1}) * X + X^T * (\Sigma_1^{-1} * \mu_1 - \Sigma_2^{-1} * \mu_2) + log\frac{\pi_1}{\pi_2} +$$

$$\frac{1}{2} * (\mu_2^T * \Sigma_2^{-1} * \mu_2 - \mu_1^T * \Sigma_1^{-1} * \mu_1) + \frac{1}{2} * log\frac{|\Sigma_2|}{|\Sigma_1|}$$

where we use

$$\hat{\pi_1} = \frac{n_1}{n_1 + n_2}; \hat{\pi_2} = \frac{n_2}{n_1 + n_2}; \hat{\mu_1} = \bar{X_1}; \hat{\mu_2} = \bar{X_2}; \hat{\Sigma_1} = S_1; \hat{\Sigma_2} = S_2$$

to estimate $\pi_1$, $\pi_2$, $\mu_1$, $\mu_2$, $\Sigma_1$, $\Sigma_2$ . Then we define:

$$L_2 = \frac{1}{2} * (\mu_2^T * \Sigma_2^{-1} * \mu_2 - \mu_1^T * \Sigma_1^{-1} * \mu_1) = 1.68; L_1 = \frac{1}{2} * log\frac{|\Sigma_2|}{|\Sigma_1|} = 5.07; a = log\frac{\pi_1}{\pi_2} = 0.70$$

then plug them all into the equation:

$$S_1(X) - S_2(X) = \frac{1}{2}X^T * (\Sigma_2^{-1} - \Sigma_1^{-1}) * X + X^T * (\Sigma_1^{-1} * \mu_1 - \Sigma_2^{-1} * \mu_2) + a + L_2 + L_1$$
$$= X^T * m_1 * X + X^T * m_2 + 7.45$$

where

$$m_1 = \begin{bmatrix} -0.24 & 2.89e-04 & 0.0016 & 2.11e-05 & 2.94e-04 & -0.01 & -0.18 & 0.05 \\ 2.89e-04 & -0.0021 & 0.0001 & -0.00016 & 0.00037 & -0.00014 & -0.01 & 0.0012 \\ 0.0016 & 0.0001 & -0.0048 & -0.00094 & 8.29e-05 & 0.0032 & -0.01 & 0.00066 \\ 2.11e-05 & -0.00016 & -0.00094 & -0.0086 & -0.00012 & 0.011 & -0.01 & 0.004 \\ 2.94e-04 & 0.00037 & 8.29e-05 & -0.00012 & -0.00013 & 0.00051 & 0.011 & -0.00034 \\ -0.0082 & -0.00014 & 0.0032 & 0.011 & 0.00051 & -0.026 & -0.01 & -0.00058 \\ -0.18 & -0.013 & -0.014 & -0.014 & 0.011 & -0.01 & -8.84 & 0.065 \\ 0.05 & 0.0012 & 0.00066 & 0.004 & -0.00034 & -0.00058 & 0.065 & -0.018 \end{bmatrix}$$

$$m_2 = \begin{bmatrix} 0.32 & -0.0034 & -0.0033 & 0.01 & -0.0021 & -0.0041 & 0.38 & -0.05 \\ -0.11 & 0.12 & 0.02 & 0.03 & -0.03 & -0.02 & 1.65 & -0.09 \\ -0.10 & 0.0064 & 0.04 & 0.10 & -0.0114 & -0.10 & -0.19 & 0.07 \\ 0.09 & 0.0080 & 0.05 & -0.10 & 0.0055 & -0.03 & 1.29 & -0.15 \\ -0.10 & -0.03 & -0.03 & 0.03 & 0.0070 & -0.07 & -1.81 & 0.09 \\ 0.09 & -0.0031 & -0.04 & -0.07 & -0.0161 & 0.10 & 1.16 & -0.0088 \\ 0.07 & 0.0070 & -0.0048 & 0.02 & -0.0062 & 0.02 & 1.64 & -0.0228 \\ -0.80 & -0.02 & 0.04 & -0.16 & 0.02 & -0.01 & -1.40 & 0.16 \end{bmatrix}$$

Then when we plug in any value of $\mathbf{X}$, where $\mathbf{X}$ has to be an $8 \times 1$ vector,
if $S_1(\mathbf{X}) - S_2(\mathbf{X}) > 0$, then it is predicted to be in class one, positive for diabetes, and
if $S_1(\mathbf{X}) - S_2(\mathbf{X}) < 0$, then it is predicted to be in class two, negative for diabetes.

## 6 Principal Component Analysis (PCA)

PCA is a method to reduce the dimensions of a large data set so that it's easier to analyze the data. The goal of the method is to reduce the number of variables but still contain most information in the original data set to preserve the accuracy of our analysis and prediction.

Principal components are linear combinations of the variables in the original data set, for each contains specific percentage of initial information. To preserve as much information as possible and reduce variables, principal components with more information should be attained and the ones with less information should be reduced. The covariance matrix of the original data set is generated to construct uncorrelated principal components, which avoids redundant information, and the eigenvalues and associated eigenvectors of the covariance matrix are computed to generate principal components. Therefore, the number of the principal components is the same as the number of eigenvalues of the covariance matrix, or the number of initial variables. We further generated two plots in our analysis: Scree plot and Cumulative percentage plot. These plots contribute to our understanding of the data by observing patterns and characteristics.

### 6.1 Plot Analysis and Interpretation

**The scree plot** (see Figure 4) displays the values or variances of the eight eigenvalues. The eigenvalues on the Scree Plot are typically plotted in decreasing order. This arrangement helps visualize the cumulative contribution of the components or factors, as the ones with the largest eigenvalues and explained the most variance are located towards the left side of the plot. In other words, larger eigenvalues indicate that the corresponding principal components capture more information from the data, making them more valuable for analysis. According to the scree plots of our dataset, the values and variances of the first five principle components are relatively larger than the last three components, showing that the last three components have relatively low information. In this case, the last three components can be dropped from our dataset without losing too much information, which simplifies the dataset and makes it easier to analyze.

In the **cumulative percentage plot** (see Figure 5), the x-axis represents the number of principal components, while the y-axis represents the cumulative percentage of variance explained. By analyzing the cumulative percentage plot, we can determine how many principal components are needed to capture a certain amount of variance in the data. In our case, it would be preferred to achieve an overall accuracy rate above 85% since 85% would be enough to explain all

the variables. Therefore, we would select the first five variables, as they contribute to an accuracy rate of 87%. This information helps us make informed decisions about the number of principal components to select for dimensional reduction or feature extraction.

| Exact Value of Cumulative Percentage | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Index | 1 | 2 | 3 | 4 | **5** | 6 | 7 | 8 |
| Percentage | 31.99% | 51.45% | 66.44% | 78.38% | **87.39%** | 92.40% | 96.29% | 100% |

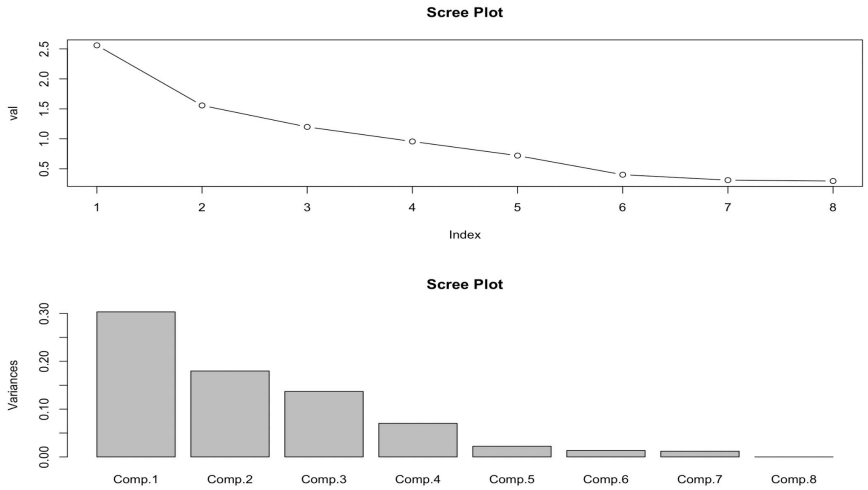


Figure 4: Scree Plot

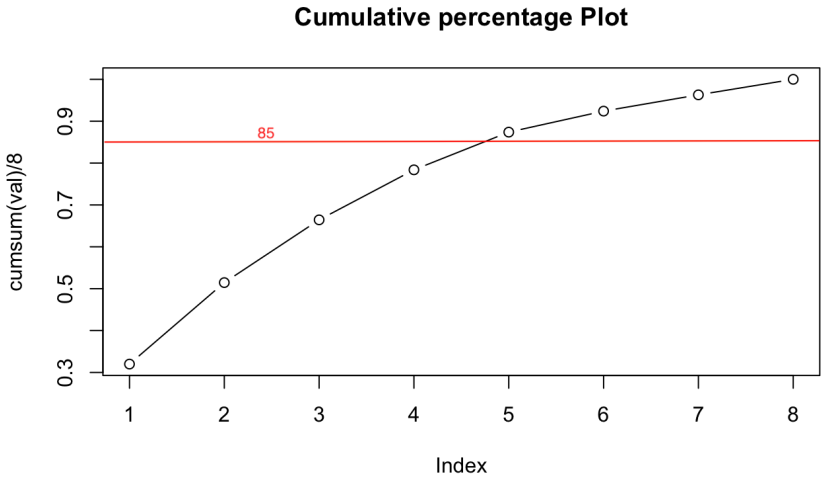


Figure 5: Cumulative Percentage Plot

## 7 Conclusion

In conclusion, our analysis consisted of the use of Box's M test, QDA, decision boundary analysis, and PCA to gain valuable insights from the data. Box's M test confirmed the presence of significant differences in covariance matrices, supporting the application of QDA for classification analysis. Using the QDA model, we classified individuals as having or not having diabetes, achieving a practical error rate of 0.1786. The derived decision boundary allows for predicting the diabetes status of new individuals. Furthermore, we applied PCA to reduce the dimensionality of the dataset from 8 variables to 5 variables. The transformed dataset retained over 87% of the original information, preserving the accuracy of our predictions with the reduced model. This implies that the variables of Pregnant, Glucose, Pressure, Triceps, and Insulin are sufficient for predicting the onset of diabetes in female Pima Indians. Therefore, these techniques and analyses provide insights into datasets that enable informed conclusions about the onset of diabetes in the study population. However, in order to achieve a comprehensive analysis, it is indeed essential to include factor analysis in our methodology. In the future, we will consider factor analysis of our dataset to identify underlying factors or structures that explain the pattern of relationships between observed variables. This may lead to a more robust understanding of the factors that influence the onset of diabetes in this population.

## 8 Acknowledgments

We would like to thank our TA, Rui Hu, for his guidance and support throughout this project. Additionally, we extend our appreciation to our groupmates for their invaluable contributions towards the success of this project.

## 9 Reference

[1] (Bennett, n.d.)R: Pima Indians Diabetes Database. (n.d.). Search in: R. Retrieved June 13, 2023, from https://pubmed.ncbi.nlm.nih.gov/7468572/.

[2] R: Pima Indians Diabetes Database. (n.d.). Search in: R. Retrieved June 13, 2023, from https://search.r-project.org/CRAN/refmans/mlbench/html/PimaIndiansDiabetes.html.



| Author | Contribution | Email |
|---|---|---|
| Yuanyuan Gao | PCA, QDA, Report Layout | yuygao@ucdavis.edu |
| Yirong Xu | PCA, QDA, Report Layout | lyrxu@ucdavis.edu |
| Faith Liu | Box's M Test, QDA | yyxliu@ucdavis.edu |
| Sabrina Zhu | PCA, Report Layout | wtzzhu@ucdavis.edu |
| Shu Hong | QDA, Report Layout | shuhong@ucdavis.edu |

**Code Appendix**

```
knitr::opts_chunk$set(echo = FALSE)
library(mlbench)
library(tidyverse)
library(gridExtra)
library(corrplot)
library(MVA)
library(ggplot2)
library(ggcorrplot)
library(visreg)
library(heplots)
library(caret)
library(e1071)
library(mvtnorm)
###Real
library(klaR)
library(psych)
library(MASS)
#library(ggord)
library(devtools)
library(heplots)
library(ade4)
data(PimaIndiansDiabetes2)
data <- PimaIndiansDiabetes2
data <- na.omit(data)

# Check for NULL Values/Missing values
data <- data[rowSums(is.na(data)) != ncol(data), ]        # Drop empty rows
any(is.na(data))
table(is.na(data))
str(data)
colnames(data)
table(PimaIndiansDiabetes2$diabetes)
# group by
data$diabetes <- unclass(data$diabetes)
data$diabetes = as.numeric(data$diabetes)
data # diabetes become 1/2 not pos/neg

# correlation
ggcorrplot(cor(data[1:9]))
ggcorrplot(cor(data[1:9]),
           hc.order = TRUE,
           type = "lower",
           lab = TRUE)


# Box's M Test (check condition for LDA)
# default method
res <- boxM(data[, 1:8], data[, "diabetes"])
res[3]
#summary(res)
pairs.panels(data[1:8],
             gap = 0,
```

```
             bg = c("lightblue", "lightyellow")[data$diabetes],
             pch = 21)

set.seed(123)
ind <- sample(2, nrow(data),
              replace = TRUE,
              prob = c(0.7, 0.3))
training <- data[ind==1,]
testing <- data[ind==2,]
dim(data)[1]
dim(training)[1]
dim(testing)[1]


quadratic <- qda(diabetes~., training)
quadratic
data$diabetes <- as.factor(data$diabetes)
#partimat(diabetes~., data = training, method = "qda")

p1 <- predict(quadratic, training)$class
tab <- table(Predicted = p1, Actual = training$diabetes)
tab

p2 <- predict(quadratic, testing)$class
tab1 <- table(Predicted = p2, Actual = testing$diabetes)
tab1

1-sum(diag(tab))/sum(tab)
1-sum(diag(tab1))/sum(tab1)


# Create a confusion matrix
conf_matrix <- confusionMatrix(tab)

# Plot the confusion matrix
plot(conf_matrix$table, col = conf_matrix$byClass,
     main = "Confusion Matrix",
     xlab = "Actual",
     ylab = "Predicted",
     color = "chartreuse4")

c1 <- subset(data, data$diabetes==1)[,1:8]
c2 <- subset(data, data$diabetes==2)[,1:8]
l1 = 0.5*log(det(cov(c2))/det(cov(c1)))
l2=0.5*(t(colMeans(c2))%*%solve(cov(c2))%*%(colMeans(c2))-t(colMeans(c1))%*%solve(cov(c1))%*%(colMeans(
n1 = dim(c1)[1]
n2 = dim(c2)[1]
n = n1+n2
pi1 = n1/n
pi1
pi2 = n2/n
pi2
solve(cov(c2))
```

```r
solve(cov(c1))
a = log(pi1/pi2)
a

l1 = 0.5*log(det(cov(c2))/det(cov(c1)))
l2=0.5*(t(colMeans(c2))%*%solve(cov(c2))%*%(colMeans(c2))-t(colMeans(c1))%*%solve(cov(c1))%*%(colMeans(c
# descionB = 0.5*t(x)%*%(solve(cov(c2))-solve(cov(c1)))%*%x + t(x)%*%(solve(cov(c1))*colMeans(c1)-solve

m1 = .5*solve(cov(c2))-solve(cov(c1))
m2 = solve(cov(c1))*colMeans(c1)-solve(cov(c2))*colMeans(c2)

a+l1+l2
mydata <- data[,1:8]
dim(mydata)
str(mydata)
pca_a=mydata
cov(pca_a)
ev=eigen((cov(pca_a)))$values
avg_ev = sum(ev)/8
eve=eigen(cov(pca_a))$vectors
pca_l1=eve[1:8,1]%*%t(pca_a)
pca_l2=eve[1:8,2]%*%t(pca_a)
plot(pca_l1[1,1:392], -pca_l2[1,1:392], main = "Scatter Plot",
                                        xlab = "First Principle Component",
                                        ylab = "Second Principle Component")
plot(sort(pca_l1[1,1:392]), main = "Sorted plot of First Principal Component",
                            ylab = "Sorted First Principle Component")

S <- cor(pca_a)
S
val<-eigen(S)$values
val
vec<-eigen(S)$vectors
vec
par(mfrow=c(2,1))
plot(val,type="b",main="Scree Plot")
screeplot(princomp(S),main="Scree Plot")
cumsum(val)/8
par(mfrow=c(1,1))
plot(cumsum(val)/8,type="b",main="Cumulative percentage Plot")

PCA<-princomp(pca_a,cor=T)
PCA
loadings(PCA)
summary(PCA)
```