Pima Indian Diabetes Analysis Report

STA 6923
Robert Santos

I. Introduction

The purpose of this analysis was to determine a means of predicting whether a patient has diabetes through machine learning methods. The analysis was performed using data about the Pima Indian population obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. The linear and nonlinear methods used in the analysis were Logistic regression performed with both subset selection and shrinkage methods, Linear discriminant analysis, Quadratic discriminant analysis, Classification Tree, and Random Forest. Results suggest that a logistic regression with the elastic net shrinkage method produced the most accurate predictions based on predictive accuracy, the AUC value, and the number of false negatives in the confusion matrix. Being able to achieve a highly accurate diabetes prediction model will be useful for at-risk populations and future studies. I was able to achieve a fairly accurate predictive model through the logistic regression with elastic net shrinkage.

A. Background

The Pima Indians are a group of Native Americans currently living on reservations in Arizona. They have the highest rate of diabetes in the world with more than half being diagnosed with type 2 diabetes (Diabetes, 2023). This population has been the subject of many longitudinal diabetes and obesity studies since 1965. There seems to be some genetic component contributing to their extremely high rates of diabetes. Drastic lifestyle changes between historic and modern Pima Indians in terms of food availability and physical activity levels is a leading theory in how their modern metabolisms make them susceptible to developing type 2 diabetes (Schulz & Chaudhari, 2015) . Many of them develop diabetes after the age of 35, and it affects their long-term health, including a comorbid risk of developing kidney disease.

B. Objective and Hypotheses

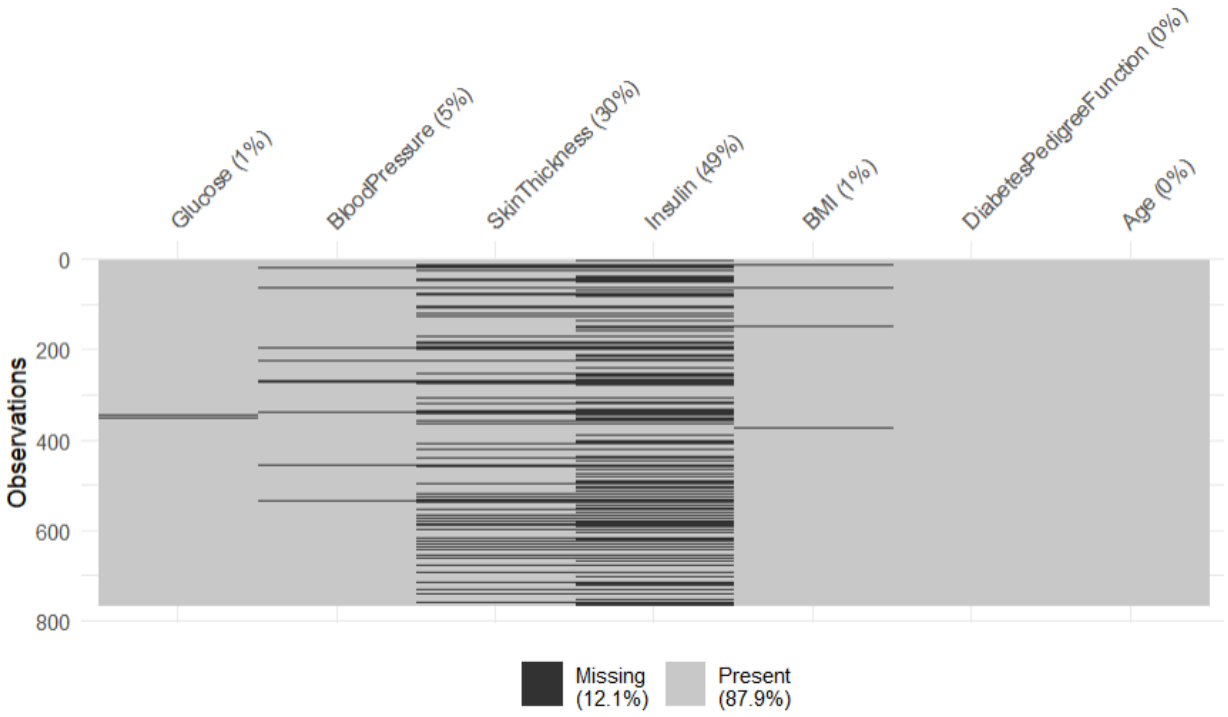
The purpose of this analysis was to build a model to accurately predict whether a patient has diabetes based on several medical predictor variables.
I have two hypotheses:
-Logistic regression with the subset selection method will provide the most accurate predictions.
-Glucose Levels, Insulin Levels, BMI, and Diabetes Pedigree Function will be the best variables to include in the model for accurate predictions.

II. Methods

A. Exploratory Analysis

The original dataset included 768 observations of 7 predictor variables and one outcome variable. The outcome variable was a binary variable indicating whether someone had diabetes (1) or not (0). The 7 predictor variables consisted of Glucose, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Pregnancies, BloodPressure, and Age. The 768 observations were data from 768 Pima Indian female patients who were at least age 21.

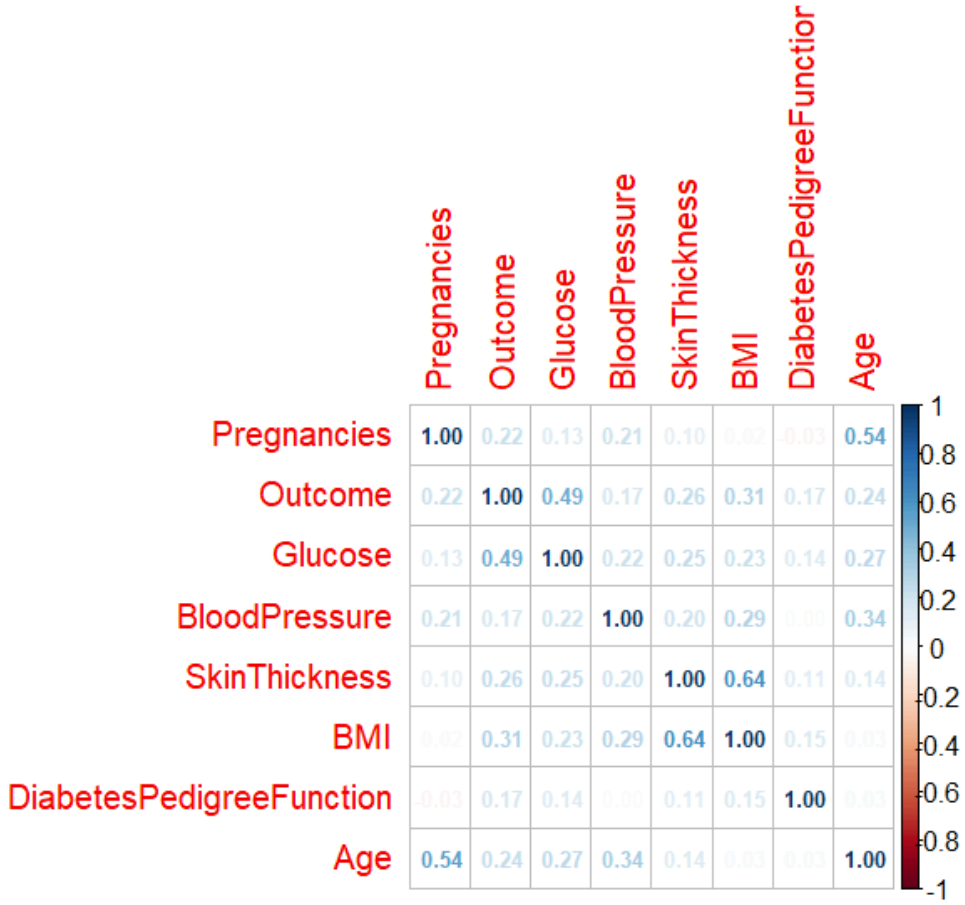
There were no missing values in the dataset, however there were several zeroes in rows of columns I deemed inappropriate. For example, a blood pressure value of zero does not make any medical sense. So I first changed the zero values to na values in order to visualize how many missings were in the data frame, which can be seen in the figure below.



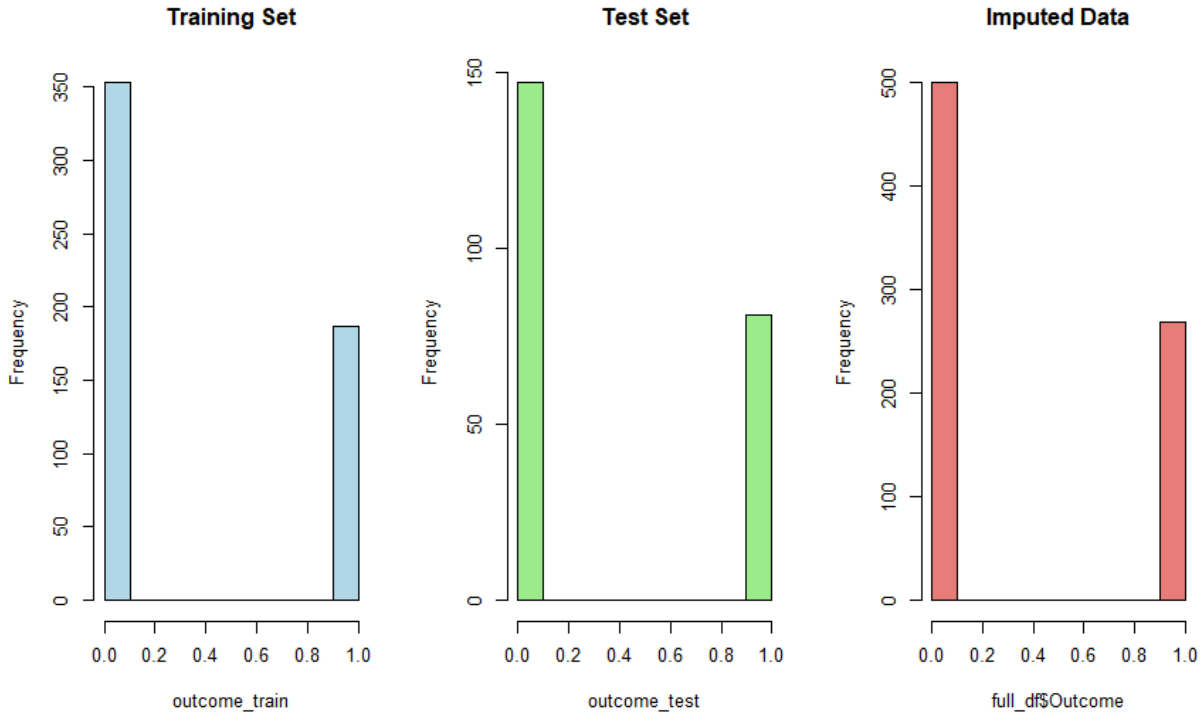
SkinThickness was missing about 30% of its values, and Insulin was missing nearly half of its values. I decided to only delete Insulin from analysis and to impute the missing values for the rest of the predictor variables.

B. Data Preprocessing

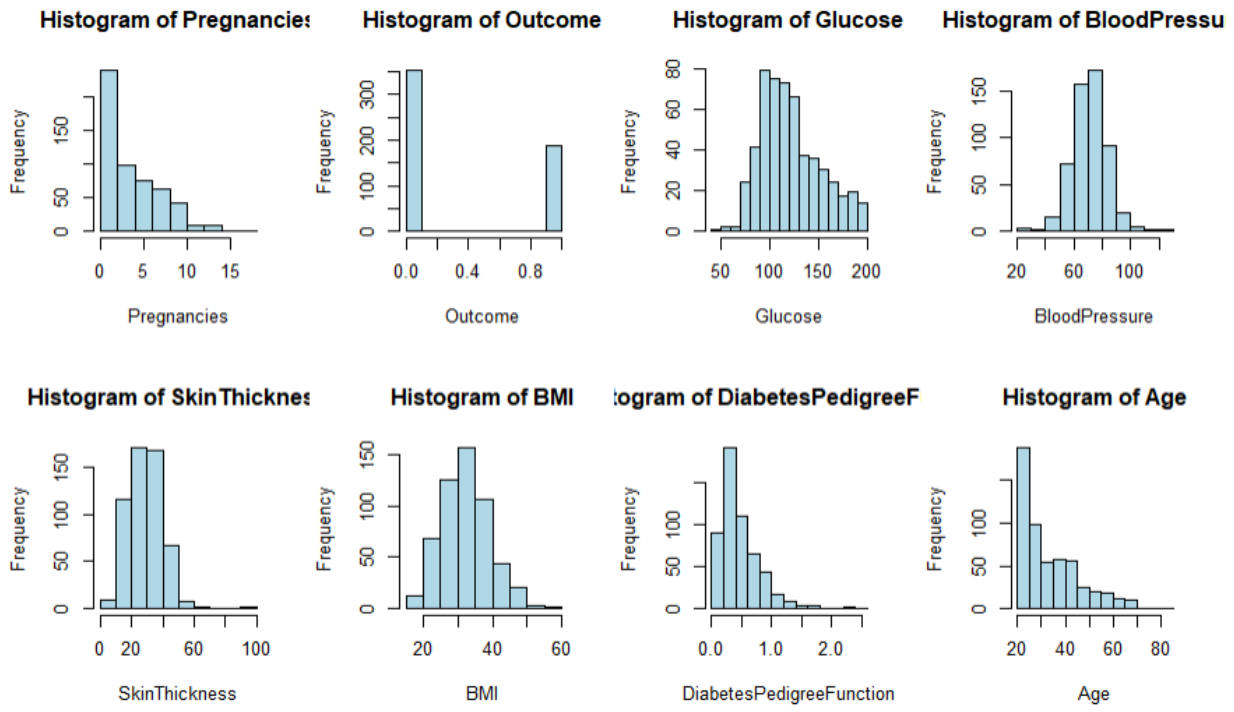
I first checked the correlations of the variables to determine if there was any multicollinearity. As you can see in the correlation plot below, there was no indication of multicollinearity between the variables, and so I did not remove any from analysis.



I then split the data into a training and test data set in a 70/30 ratio, so that I could fit a model to the training data and test its predictive accuracy with the test data. I confirmed that the imbalance of values in the Outcome variable from the original data set was preserved in both the training and test data sets, as you can see in the below figures.



I then checked the distributions of the variables to see if any transformations would be necessary. As can be seen below, Pregnancies, Age, and DiabetesPedigreeFunction all did not follow a normal distribution, and appeared to be right skewed. I decided to only try a transformation of DiabetesPedigreeFunction in an attempt to preserve interpretability of the predictions in terms of the other two skewed variables. A log transform of the DiabetesPedigreeFunction was considered in later analysis.



C. Model Building

In order to create an optimal model, I performed stepwise selection on a logistic regression model that included the predictors that remained after the imputation step (all but Insulin). I also performed stepwise selection on a different logistic regression model with the same predictors (but with a log-transformed DiabetesPedigreeFunction predictor). Both of these stepwise selection methods suggested that Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction should be kept in the model for analysis.

III. Model Validation and Results

I performed one logistic regression on all of the predictors (minus Insulin), one only including the subset of Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction, and one only including the subset of Pregnancies, Glucose, BMI, and a log-transformed DiabetesPedigreeFunction. For the logistic regression that had the log-transformed DiabetesPedigreeFunction in the training set, I also performed the same transformation on the test set. Out of these three logistic regressions, the model with the subset with no transformations performed had the best prediction accuracy at 76.83%, based on the confusion matrix below.

```
                      outcome_test
logistic_predfinal     0   1
                   0 150  44
                   1  13  39
```

I then ran LDA and QDA with the same variables (the subset with no transformations) and got a prediction accuracy of 76.42% and 75.61%, respectively.

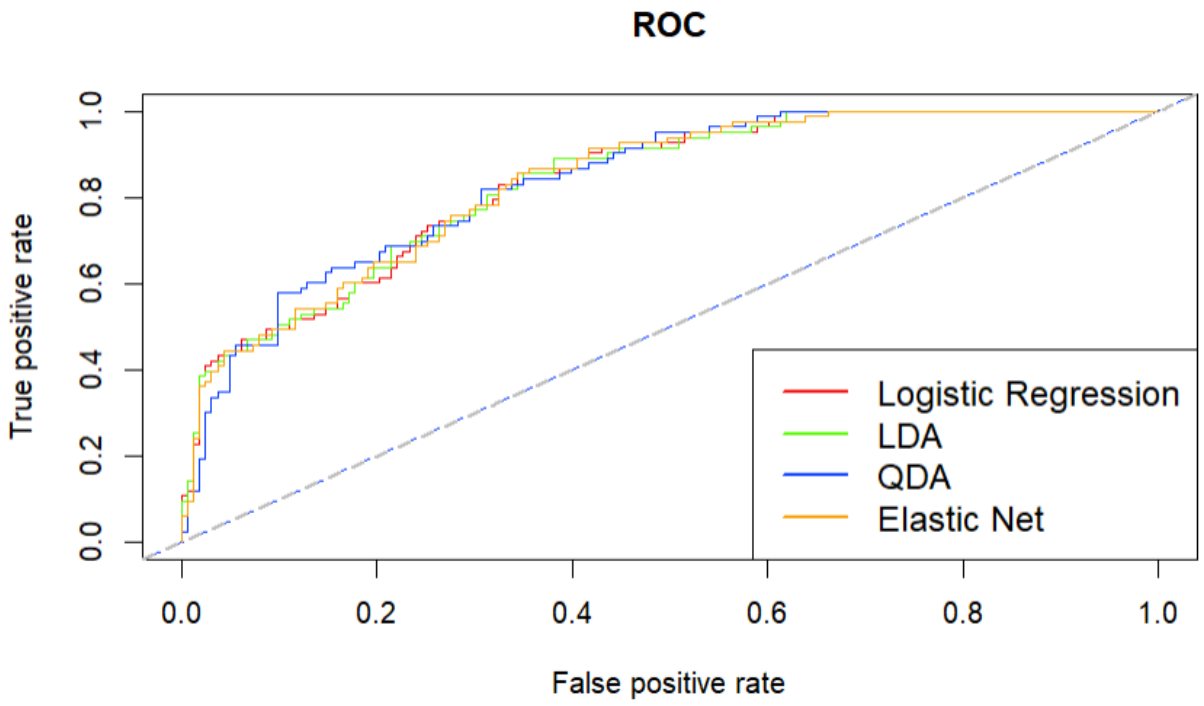
```
  outcome_test
     0   1
0 149  44
1  14  39
```

```
  outcome_test
     0   1
0 147  44
1  16  39
```

Lastly, I performed a logistic regression with the elastic net shrinkage method and achieved a prediction accuracy of 76.83%, which matches the logistic regression with the subset selection method.

```
                      true_outcomes
binary_predictions     0   1
                   0 149  43
                   1  14  40
```

Below are the ROC curves of the four models that I fit to the data.



The four models all performed relatively similarly, but they did have different AUC values. The logistic regression with subset selection had an AUC of .8338; LDA had an AUC of .8335; QDA had an AUC of .8369; and logistic regression with the elastic net shrinkage method had an AUC of .8348. The larger the AUC (area under the ROC curve) value, the more accurate the model is at predicting. Based on these values, the QDA model had the best predictive accuracy, and the elastic net model had the second best predictive accuracy.

The logistic regression with subsetting, LDA, and the QDA model were all fit with the variables chosen by stepwise selection (Pregnancies, Glucose, BMI, and DiabetesPedigreeFunction). I used an elastic shrinkage method (which is able to do variable selection) on another logistic regression model that was fit with the full amount of imputed predictors. Below are the variables and coefficients from the elastic net model. The model did keep all of the variables, but the ones with the largest coefficients are the same four predictors that were used in the other three models.

name <chr>	coefficient <dbl>
(Intercept)	-8.636793323
Pregnancies	0.110812533
Glucose	0.035079845
BloodPressure	-0.008790632
SkinThickness	0.010635275
BMI	0.082650532
DiabetesPedigreeFunction	0.413672490
Age	0.015297459

IV. Conclusions

Based on my analysis, I would suggest identifying Pregnancies, Glucose, BMI, and the DiabetesPedigreeFunction as the most important variables for predicting whether someone in the selected population has diabetes. I would also suggest fitting future data from the population with an elastic net model because it had the highest prediction accuracy of the models I fit. The elastic net model did not have the largest AUC value, but when you look at the confusion matrices, it had the least amount of false negatives (43). False negatives are when we predict that someone does not have diabetes based on the predictors, but they actually do have diabetes. This could be a detrimental mistake by a predictive model that you would want to minimize for this population.

Lastly, I would also suggest making sure to obtain clean data, especially for insulin, which I had to drop from analysis due to the large number of missings. The relationship between Insulin and the Outcome variable of this dataset requires future analysis in order to include it in future predictive models.

References

Schulz LO, Chaudhari LS. High-Risk Populations: The Pimas of Arizona and Mexico. Curr Obes Rep. 2015 Mar;4(1):92-8. doi: 10.1007/s13679-014-0132-9. PMID: 25954599; PMCID: PMC4418458.

Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

(2023). Diabetes Education. National Indian Council on Aging, Inc. https://www.nicoa.org/diabetes-still-highest-among-ai-an/#:~:text=The%20Pima%20Indians%20of%20Arizona,develop%20it%20earlier%20in%20life.