STA 6543 Project

Abigail Dastur, Hannah Haley, Rachael Humphreys, and Karla Wiedmer

2022-07-28

## Introduction and Background:

The background of the study for the Pima Indian Dataset was conducted by scientists to analyze the significance of health related variables (predictors) to diabetes in Pima Indian Women. The samples’ population was made up of females aged 21 years and older. Alongside, the population had an heritage of diabetes and digestive and kidney diseases.

Type 2 diabetes is a lifelong and life altering chronic illness affecting 462 million people worldwide (~6.3% of the global population), in which the pancreas can no longer produce sufficient insulin to control blood sugar in the body, there is no cure . There are several factors that can lead to and increased risk of developing type 2 Diabetes, including but not limited to, genetic and environmental components.

The exponentially higher occurrence of diabetes in the Pima Indian population means the early detection of this disease in this underserved communities is extremely important due to the limited access to advanced health care. Earlier diagnosis, prevents the progression of the disease and escalation of the complexity of treatments needed to preserve the quality of life in the patients.

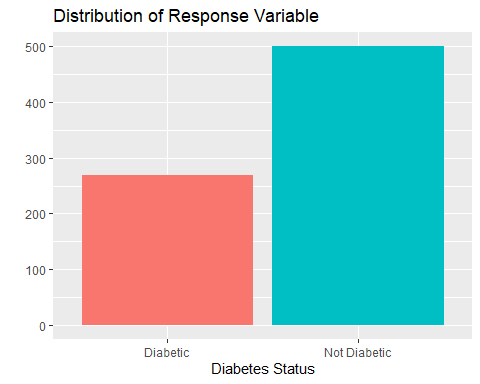
We will explore four different Machine Learning models to establish which one performs the best in predicting diabetes in Pima Indians. The machine learning methods used to construct this analysis were:K-nearest neighbors, Logistic Regression, Random Forest, and Support Vector Machine.

## Data Structure:

The structure of the data consists of 768 observations and 9 variables. Out of those 9 variables, there is one target variable, Outcome. The other eight variables are the predictor variables and includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on. Outcome was listed as 1 for a person with diabetes, and 0 for a person without diabetes. The data contains approximately twice as many examples of people without diabetes compared to people with diabetes.

## Pregnancies Glucose BloodPressure SkinThickness Insulin BMI  
## 1 6 148 72 35 0 33.6  
## 2 1 85 66 29 0 26.6  
## 3 8 183 64 0 0 23.3  
## 4 1 89 66 23 94 28.1  
## 5 0 137 40 35 168 43.1  
## 6 5 116 74 0 0 25.6

## DiabetesPedigreeFunction Age Outcome  
## 1 0.627 50 1  
## 2 0.351 31 0  
## 3 0.672 32 1  
## 4 0.167 21 0  
## 5 2.288 33 1  
## 6 0.201 30 0



## Statistical Learning Methods

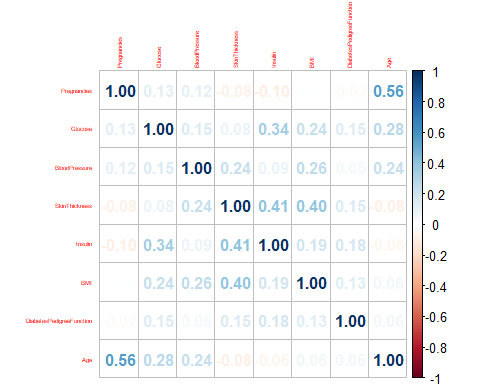
### Approach & Pre-processing

Before we could apply any methods to analyze our data, we needed to determine if there were any missing values, as well as check for correlation, skewness, and potential outliers. We also split the data into a training and testing set.

#### Near Zero Variance

**Observation:** It was observed that there are no predictors with near zero variance. Also, as there are a small amount of observations, splitting the data 25/75 is a balanced approach for testing/training sets.

#### Correlation

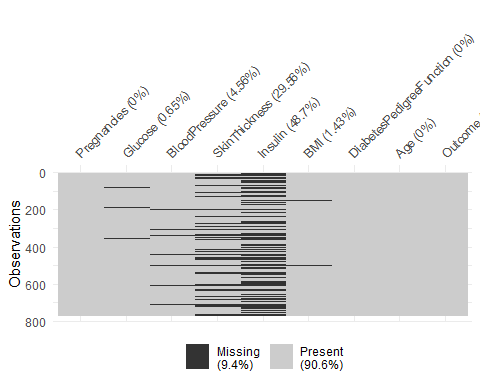


**Observation:** While some correlation was observed there were no variables that were removed using a 0.75 cutoff.

#### Missing Data

## Pregnancies Glucose BloodPressure   
## 111 5 35   
## SkinThickness Insulin BMI   
## 227 374 11   
## DiabetesPedigreeFunction Age   
## 0 0

**Analysis of missing values:**



**Observation:** Insulin and SkinThickness are missing a significant number of observations. Two approaches were tried. The first was replacing the missing values with the column mean, and the second was removing the Insulin and SkinThickness columns. Either approach gave similar model results so it was decided to move forward with replacing the missing values with column mean for all columns.

### Statistical Learning Methods & Models

### K-Nearest Neighbor

The KNN model is good for the analysis because it is a good classifier for small datasets.

Model results using training data.

## k-Nearest Neighbors   
##   
## 576 samples  
## 8 predictor  
## 2 classes: '0', '1'   
##   
## Pre-processing: centered (8), scaled (8)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 518, 519, 519, 518, 519, 517, ...   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 3 0.7135045 0.3500766  
## 4 0.7135942 0.3596110  
## 5 0.7534924 0.4300542  
## 6 0.7551288 0.4377841  
## 7 0.7431788 0.4123042  
## 8 0.7379751 0.4013158  
## 9 0.7482907 0.4168838  
## 10 0.7257226 0.3599522  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 6.

Model results using test data.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 100 37  
## 1 19 36  
##   
## Accuracy : 0.7083   
## 95% CI : (0.6386, 0.7716)  
## No Information Rate : 0.6198   
## P-Value [Acc > NIR] : 0.006386   
##   
## Kappa : 0.3502   
##   
## Mcnemar's Test P-Value : 0.023103   
##   
## Sensitivity : 0.4932   
## Specificity : 0.8403   
## Pos Pred Value : 0.6545   
## Neg Pred Value : 0.7299   
## Prevalence : 0.3802   
## Detection Rate : 0.1875   
## Detection Prevalence : 0.2865   
## Balanced Accuracy : 0.6667   
##   
## 'Positive' Class : 1   
##

**Observation:**

It was observed that the best k value was 6, which gave an Accuracy of 0.7551288 and a Kappa value of 0.4377841 on the training data. When the test data was used we saw an Accuracy of 0.7083 and a Kappa value of 0.3502.

### Support Vector Machines

Support Vector Machine is supervised learning and constructs hyper plane surfaces that predicts whether the examples fall into one class or another separated by a margin and classifies examples with the largest margin.

Model results using training data.

## Support Vector Machines with Radial Basis Function Kernel   
##   
## 576 samples  
## 8 predictor  
## 2 classes: 'Diabetic', 'NonDiabetic'   
##   
## Pre-processing: centered (8), scaled (8)   
## Resampling: Repeated Train/Test Splits Estimated (25 reps, 75%)   
## Summary of sample sizes: 433, 433, 433, 433, 433, 433, ...   
## Resampling results across tuning parameters:  
##   
## C Accuracy Kappa   
## 0.0625 0.7527273 0.4508137  
## 0.1250 0.7558042 0.4548174  
## 0.2500 0.7611189 0.4405442  
## 0.5000 0.7655944 0.4437280  
## 1.0000 0.7602797 0.4295368  
## 2.0000 0.7558042 0.4148726  
## 4.0000 0.7530070 0.4072010  
## 8.0000 0.7499301 0.3973407  
## 16.0000 0.7465734 0.3844308  
##   
## Tuning parameter 'sigma' was held constant at a value of 0.03412908  
## Accuracy was used to select the optimal model using the largest value.  
## The final values used for the model were sigma = 0.03412908 and C = 0.5.

Model results using test data.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Diabetic NonDiabetic  
## Diabetic 36 6  
## NonDiabetic 37 113  
##   
## Accuracy : 0.776   
## 95% CI : (0.7104, 0.8329)  
## No Information Rate : 0.6198   
## P-Value [Acc > NIR] : 2.744e-06   
##   
## Kappa : 0.4823   
##   
## Mcnemar's Test P-Value : 4.763e-06   
##   
## Sensitivity : 0.4932   
## Specificity : 0.9496   
## Pos Pred Value : 0.8571   
## Neg Pred Value : 0.7533   
## Prevalence : 0.3802   
## Detection Rate : 0.1875   
## Detection Prevalence : 0.2188   
## Balanced Accuracy : 0.7214   
##   
## 'Positive' Class : Diabetic   
##

**Observation:** It was observed that the best C value was .5, which gave an Accuracy of 0.7655944 and a Kappa value of 0.4437280 on the training data. When the test data was used we saw an Accuracy of 0.776 and a Kappa value of 0.4823.

### Random Forest

Random Forest is an extension of the decision tree and it produces different categories based on the predictors. This method is good for our analysis so we can categorize each health-related predictor.

Model results using training data.

## Random Forest   
##   
## 576 samples  
## 8 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 518, 519, 519, 518, 519, 517, ...   
## Resampling results across tuning parameters:  
##   
## mtry Accuracy Kappa   
## 8 0.7761804 0.4851823  
## 9 0.7657438 0.4615593  
## 10 0.7779641 0.4890388  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was mtry = 10.

## user system elapsed   
## 8.11 0.17 8.36

Model results using test data.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 109 51  
## 1 10 22  
##   
## Accuracy : 0.6823   
## 95% CI : (0.6114, 0.7475)  
## No Information Rate : 0.6198   
## P-Value [Acc > NIR] : 0.04244   
##   
## Kappa : 0.2438   
##   
## Mcnemar's Test P-Value : 3.032e-07   
##   
## Sensitivity : 0.3014   
## Specificity : 0.9160   
## Pos Pred Value : 0.6875   
## Neg Pred Value : 0.6813   
## Prevalence : 0.3802   
## Detection Rate : 0.1146   
## Detection Prevalence : 0.1667   
## Balanced Accuracy : 0.6087   
##   
## 'Positive' Class : 1   
##

**Observation:** It was observed that the best mtry value was 10, which gave an Accuracy of 0.7779641 and a Kappa value of 0.4890388 on the training data. When the test data was used we saw an Accuracy of 0.6823 and a Kappa value of 0.2438.

### Logistic Regression Model

Logistic Regression model is good for our analysis because it is supervised learning which is used to calculate the prediction of a binary event occurring. This is perfect for our analysis to determine if a patient has diabetes in relation to other health issues.

Model results using training data.

## Penalized Multinomial Regression   
##   
## 576 samples  
## 8 predictor  
## 2 classes: '0', '1'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 518, 519, 519, 518, 519, 517, ...   
## Resampling results across tuning parameters:  
##

## decay Accuracy Kappa   
## 0e+00 0.7603023 0.4291738  
## 1e-04 0.7603023 0.4291738  
## 1e-01 0.7620577 0.4293513  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was decay = 0.1.

Model results using test data.

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 110 29  
## 1 9 44  
##   
## Accuracy : 0.8021   
## 95% CI : (0.7386, 0.856)  
## No Information Rate : 0.6198   
## P-Value [Acc > NIR] : 4.02e-08   
##   
## Kappa : 0.5566   
##   
## Mcnemar's Test P-Value : 0.002055   
##   
## Sensitivity : 0.6027   
## Specificity : 0.9244   
## Pos Pred Value : 0.8302   
## Neg Pred Value : 0.7914   
## Prevalence : 0.3802   
## Detection Rate : 0.2292   
## Detection Prevalence : 0.2760   
## Balanced Accuracy : 0.7636   
##   
## 'Positive' Class : 1   
##

**Observation:** It was observed that the best decay value was 0.1, which gave an Accuracy of 0.7620577 and a Kappa value of 0.4293513 on the training data. When the test data was used we saw an Accuracy of 0.8021 and a Kappa value of 0.5566.

## Conclusion:

#### Summary of Test Results

## Accuracy Kappa  
## KNN 0.7083333 0.3501753  
## SVM 0.7760417 0.4823175  
## Random Forest 0.6822917 0.2438017  
## Logistic 0.8020833 0.5565820

#### ROC Curve

Chart, scatter chart

Description automatically generated

## Area Under Curve  
## KNN 0.7521008  
## SVM 0.8407966  
## Random Forest 0.8028664  
## Logistic 0.8624381

**Conclusion:**

Looking over all the metrics to evaluate performance it was observed that Logistic Regression had the best performance for both Accuracy and Kappa statistic with values of 0.8020833 and 0.5565820, as well as the best ROC curve with an area of 0.8624381.

Overall we would recommend the Logistic Regression model, as the Accuracy, Kappa, and ROC curve were observed to be the best performing. Additionally, the Outcome variable capturing Diabetes status is binary, which works well with a Logistic Regression model.

When looking at the number of False Negatives (predicting non-diabetic when diabetic), Logistic Regression still performs the best. The number of False Negatives was the lowest at 29 compared to the worst model having 51 False Negatives. Additionally, Logistic Regression performed second best when it came to False Positives at 9 compared to the best at 6. The overall number of incorrectly predicted outcomes is lowest with Logistic Regression compared to the other three models.