Diabetes Prediction on Female PIMA Indians

Mariano Tovar

Deepak Neelam

Abdullah Alharbi

Department of Management Sciences & Statistics, University of Texas at San Antonio

STA 6923: Introduction to Statistical Learning

Dr. Min Wang

December 12, 2024

List of Figures

Introduction

Diabetes is a chronic medical condition that affects millions worldwide, as a result causing significant health and economic challenges. This project focuses on predicting the onset of diabetes in Pima Indians using data pre-processing techniques followed by application of statistical and machine learning methods. The Pima Indians Diabetes Database, consisting of 768 observations and 8 clinical features, serves as the basis for this analysis. The features include medical metrics such as number of pregnancies, glucose levels, body mass index, blood pressure, and diabetes pedigree function, alongside the outcome variable indicating the presence or absence of diabetes.

Objective

This study aims to predict the onset of diabetes in Pima Indians using advanced statistical and machine learning techniques.

The objectives of this project are to:

- Thoroughly clean and prepare the data for analysis.
- Evaluate the performance of various predictive models to improve predictive accuracy and minimize error rates.
- Identify the most promising methods for predicting diabetes based on Accuracy, Misclassification error and AUC.

Data Pre-Processing

Data preprocessing involved checking correlations, scaling features, and selecting important features to improve model accuracy. Model performance was assessed on a 70/30 train-test split using prediction accuracy and misclassification error.

1. Correlation:

   Our 1st step as part of data preprocessing was checking the correlation heat map for all variables. From figure 1 mentioned below, we can see rly see high correlation between Age and Pregnancy. Also, there is a significant correlation between in Thickness and Insulin, Glucose, and Insulin which is mainly because of the presence of missing values in these variables which are being interpreted as zeros.se correlations were not above 0.75 so they were not excluded.

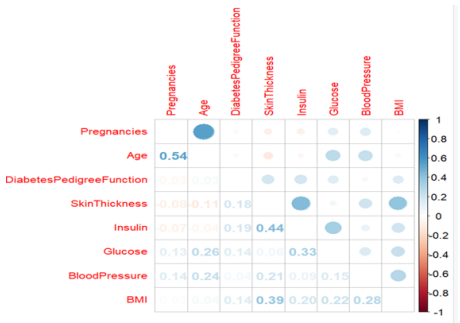

Figure 1

2. Presence of Zeros:

   It was clear that there were significant number of missing values in the data. Since the number was high and this is medical data, imputing the data at such extent would alter the integrity of the analysis. So, we have decided not to do any imputation until we are directed on how to handle those discrepancies.

3. Scaling:

   All variables were standardized by transforming them on to similar scale using scaling.

4. Near Zero Variance:

   We checked for near **zero** variance variables. Fortunately, there were no presence of such variables.

5. BoxCox Transformation:

   No variables needed BoxCox Transformation.

6. Train/Test Data Split:

   Based on the distribution of the outcome variable, it was decided to split the Train and Test Data in the ratio of 70% - 30%. Which also had a significant impact on the model accuracies.

7. Feature Selection:

   An initial study was performed using Logistic Regression, to identify the features that are not significant based on the 5% significance level. Based on the output from Figure 3, it can be determined that variables Skin Thickness, Insulin and Age are not significant. Consequently, these variables were excluded.

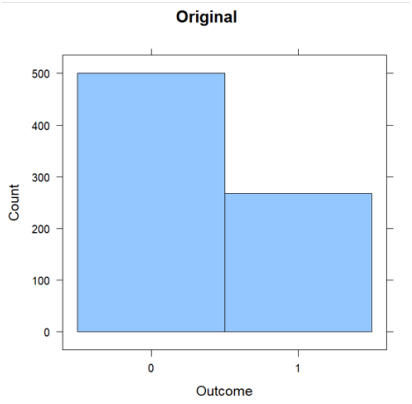

Figure 2

```
Call:
glm(formula = Outcome ~ ., family = binomial, data = Data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.871102   0.096942  -8.986  < 2e-16 ***
Pregnancies              0.415072   0.108088   3.840 0.000123 ***
Glucose                  1.124276   0.118577   9.481  < 2e-16 ***
BloodPressure           -0.257346   0.101301  -2.540 0.011072 *
SkinThickness            0.009874   0.110060   0.090 0.928515
Insulin                 -0.137336   0.103861  -1.322 0.186065
BMI                      0.707217   0.118953   5.945 2.76e-09 ***
DiabetesPedigreeFunction 0.313165   0.099116   3.160 0.001580 **
Age                      0.174863   0.109779   1.593 0.111192
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 723.45  on 759  degrees of freedom
AIC: 741.45

Number of Fisher Scoring iterations: 5
```

Figure 3

Statistical Modeling

After pre-processing the data, it was decided to start the statistical study with a Logistic Regression model including only significant predictor variables. The next step will be to perform the same test, using the Logistic Regression Model but only with the predictors that are significant, which are Pregnancies, Glucose, Blood Pressure, BMI and Diabetes Pedigree Function. Various predictive models were implemented, including Logistic Regression, LDA, QDA, KNN, Naive Bayes, Ridge, Lasso, and Elastic Net regression.

A. Logistic Regression

```
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.8505     0.1131  -7.517 5.60e-14 ***
Pregnancies                0.4310     0.1099   3.922 8.77e-05 ***
Glucose                    1.0681     0.1283   8.323  < 2e-16 ***
BloodPressure             -0.1028     0.1172  -0.878  0.38020
BMI                        0.6609     0.1369   4.828 1.38e-06 ***
DiabetesPedigreeFunction   0.3362     0.1109   3.032  0.00243 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 696.28  on 537  degrees of freedom
Residual deviance: 520.65  on 532  degrees of freedom
AIC: 532.65

Number of Fisher Scoring iterations: 5
```

Figure 4

From Figure 4, it can be observed that Blood Pressure now is the only one that shows as not significant based on a 5% significance level.

Based on confusion matrix for Logistic Regression from Figure 5, it can be determined that the prediction accuracy of this statistical model is 78.26%.

```
             Data.Outcome
glm.pred    0   1
       0 133  33
       1  17  47
```

Figure 5

B. Linear Discriminant Analysis (LDA):

Based on confusion matrix for Linear Discriminant Analysis from Figure 6, it can be determined that the prediction accuracy of this statistical model is 77.83%.

```
                Data.Outcome
lda.Dataclass    0   1
            0 133  34
            1  17  46
```

Figure 6

C. Navie Bayes:

Based on confusion matrix for Linear Discriminant Analysis from Figure 7, it can be determined that the prediction accuracy of this statistical model is 76.52%.

	Data.Outcome	
nb.Dataclass	0	1
0	130	34
1	20	46

Figure 7

D. K-Nearest Neighbors KNN (K=1):

The next statistical model that was tested is the K-Nearest Neighbors with K=1. Based on confusion matrix for K-Nearest Neighbors from Figure 8, it can be determined that the prediction accuracy of this statistical model is 69.57%.

	Data.Outcome	
knn.Datapred	0	1
0	120	40
1	30	40

Figure 8

E. K-Nearest Neighbors KNN (K=10):

The next statistical model that was tested is the K-Nearest Neighbors with K=10. Based on confusion matrix for K-Nearest Neighbors from Figure 9, it can be determined that the prediction accuracy of this statistical model is 75.65%.

	Data.Outcome	
knn.Datapred_10	0	1
0	132	38
1	18	42

Figure 9

F. K-Nearest Neighbors KNN (K=50):

The next statistical model that was tested is the K-Nearest Neighbors with K=50. Based on confusion matrix for K-Nearest Neighbors from Figure 10, it can be determined that the prediction accuracy of this statistical model is 75.65%.

	Data.Outcome	
knn.Datapred_50	0	1
0	137	43
1	13	37

Figure 10

G. QDA:

The next statistical model that was tested is the QDA. Based on confusion matrix for QDA from Figure 11, it can be determined that the prediction accuracy of this statistical model is 76.09%.

	Data.Outcome	
qda.class	0	1
0	129	34
1	21	46

Figure 11

H. Lasso:

The next statistical model that was tested is the Lasso. Based on Lambda from Figure 12, one standard deviation from the minimum, the misclassification error is **22.6%.**

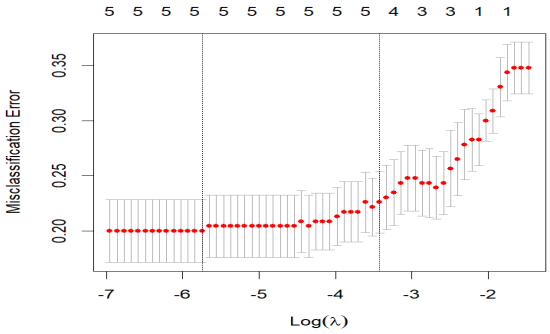

Figure 12

I. Ridge Regression:

The next statistical model that was tested is the Ridge. Based on Lambda from figure 13, one standard deviation from the minimum, the misclassification error is 21.74**%.**

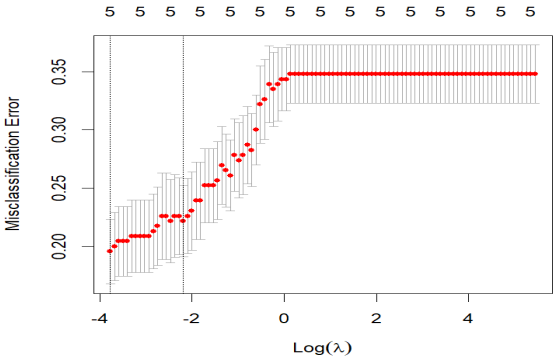

Figure 13

Model Performance

Model	Accuracy (%)	Error Rate %
Logistic Regression	**78.26**	**21.74**
Ridge Regression	*78.26*	*21.74*
LDA	*77.83*	*22.17*
Lasso Regression	*77.4*	*22.6*
Naive Bayes	*76.52*	*23.48*
QDA	*76.09*	*23.91*
KNN (K=10)	*75.65*	*24.35*
(K=50)	*75.65*	*24.35*
(K=1)	*69.57*	*30.43*

Figure 14

The Logistic and Ridge Regressions gave the maximum accuracy among all the other Statistical methods. Since there is no presence of high multi-collinearity, it is decided to choose Logistic Regression as our primary model.

AUC (Area under Curve)

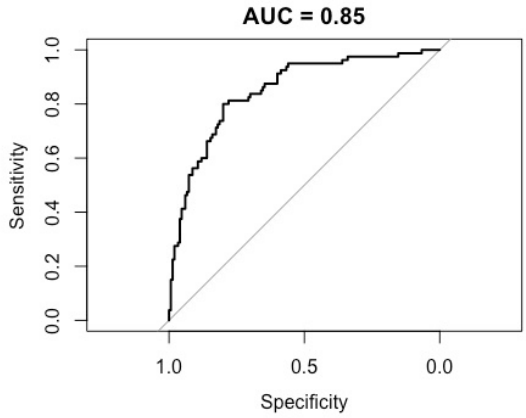

Figure 15

The ROC curve demonstrates that the logistic regression model exhibits strong performance, as indicated by an AUC of 0.85.

• AUC close to 1: Good model.

• AUC close to 0: Bad model.

• AUC equals 0.5: Model is as good as random.

• AUC greater than 0.8: Model is clinically considered useful.

Doctors can calculate the probability of having diabetes by plugging the coefficients from the logistic models and choosing the values of all the predictors of a subject by using proper units.

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k)}}$$

Figure 16

Conclusions

- Logistic Regression emerged as the best model with 78.26% accuracy and AUC of 0.85.
- Inconsistent Data might affect the performance of the model.
- It is recommended that data is examined due to the presence of inconsistencies.
- A study more focused on Type-II error is recommended once all the discrepancies in the data have been properly handled.

References

• Kaggle. Pima Indians Diabetes Database. Dataset.

• Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.