Diabetes Prediction on Female PIMA Indians

Mariano Tovar, Deepak Neelam, Abdullah Alharbi
The University of Texas at San Antonio, San Antonio TX, 78249

Abstract

- This study aims to predict the onset of diabetes in Pima Indians using advanced statistical and machine learning techniques. The dataset includes 768 records with features like number of pregnancies, glucose levels, body mass index, blood pressure, diabetes pedegree function and a binary outcome indicating the presence of diabetes.
- Data preprocessing involved checking correlations, scaling features, and selecting important features to improve model accuracy. Various predictive models were implemented, including Logistic Regression, LDA, QDA, KNN, Naive Bayes, Ridge, Lasso, and Elastic Net regression. Model performance was assessed on a 70/30 train-test split using prediction accuracy and misclassification error.
- Logistic Regression achieved the highest accuracy of 78.26%, with significant predictors identified. LDA followed closely with 77.83% accuracy. Regularized models like Lasso and Ridge Regression helped reduce overfitting and achieved minimal error rates. These findings highlight the importance of preprocessing, feature selection, and model evaluation for accurate diabetes prediction.

Introduction

- Diabetes is a chronic medical condition that affects millions worldwide, posing significant health and economic challenges. This project focuses on predicting the onset of diabetes in Pima Indians using 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, diabetes pedigree function, alongside the binary target variable indicating the presence or absence of diabetes.

The objectives of this project are to:

- Thoroughly clean and prepare the data for analysis.
- Test various predictive models to determine their accuracy and error rates.
- Identify the most promising methods for predicting diabetes.

Objective

- The key objective of this research is to determine the most effective predictive modeling method for forecasting the onset of diabetes in individuals of Pima Indian descent. To determine whether the subject is diabetes positive or not. To achieve this, we will focus on enhancing prediction accuracy by addressing missing data, scaling features, and employing robust modeling techniques.

Methods

Data Preprocessing:

- Performed feature scaling for standardization.
- Checked for multicollinearity using correlation matrices.
- Checked near zero variables
- Investigated possible transformations of data, such as BoxCox

Feature Selection:

- Logistic regression identified significant predictors (Pregnancies, Glucose, Blood Pressure, BMI, Diabetes Pedigree Function).

Modeling:

- **I**mplemented Logistic Regression, LDA, QDA, KNN, Naive Bayes, Ridge, Lasso, and Elastic Net regression.
- Models were evaluated on a 70/30 train-test split.

Evaluation Metrics:

- Accuracy and AUC(Area Under the Curve)

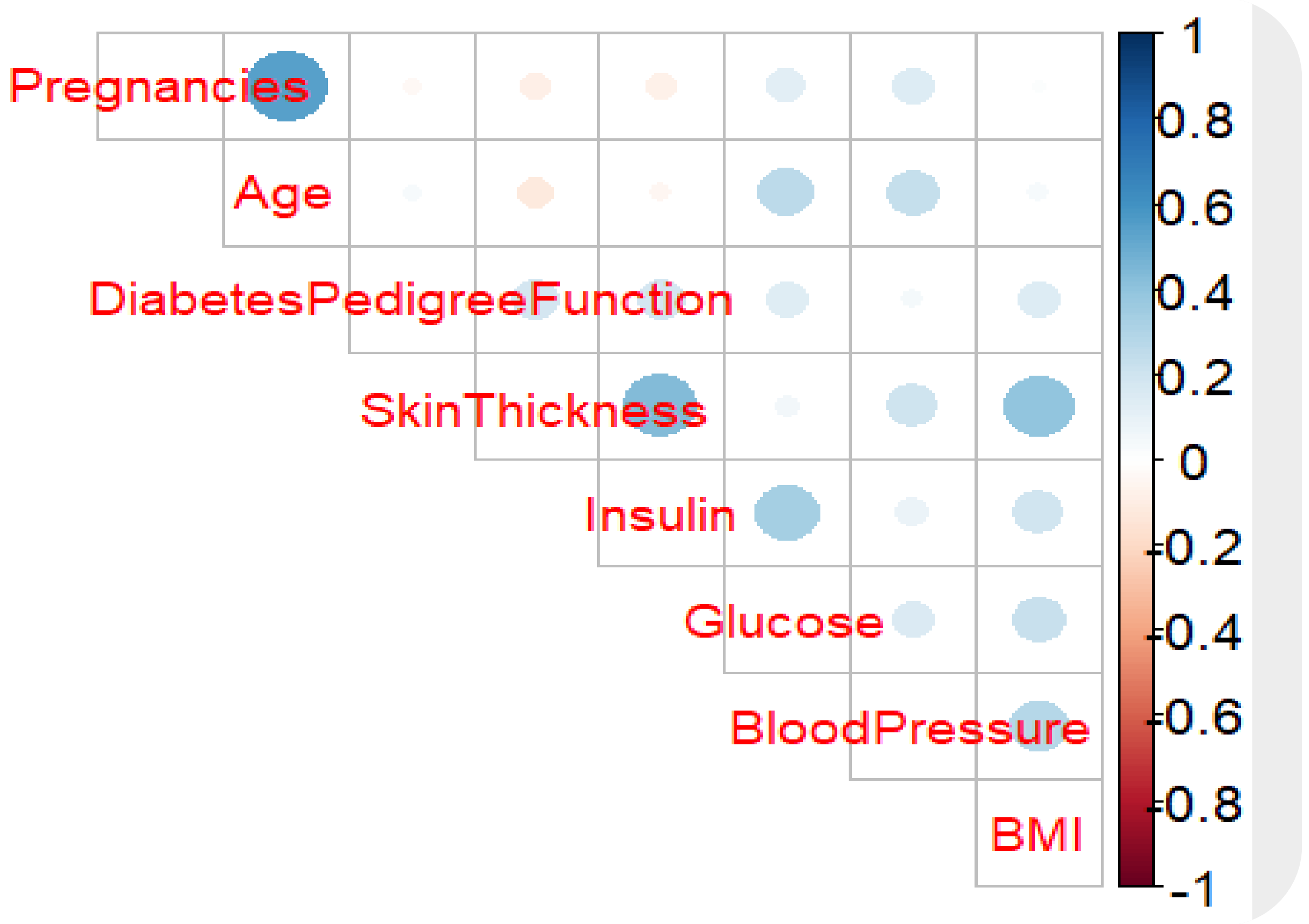

Figure 1. Chart to identify high and low correlations among the predictors in the diabetes dataset.

Results

Model	Accuracy (%)
Logistic Regression	**78.26**
Elastic Net Regression	77.9
LDA	77.83
Ridge Regression	77.4
Lasso Regression	77
Naive Bayes	76.51
QDA	76.09
KNN (K=10)	74.34
(K=50)	73.91
(K=1)	67.39

Table 1. Accuracy percentage of the statistical models that were tested.

Results - con't

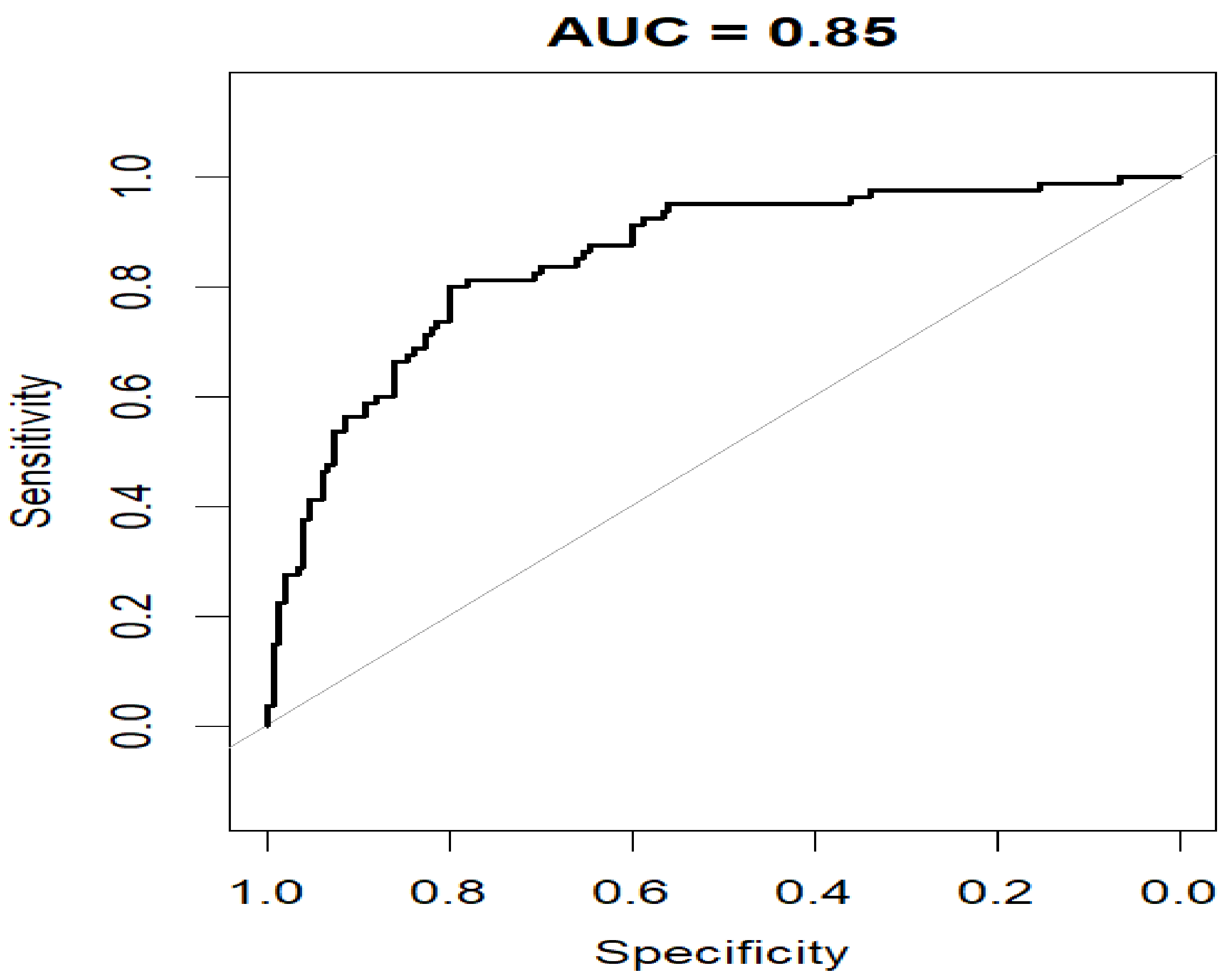

Figure 2. The ROC curve demonstrates that the logistic regression model exhibits strong performance, as indicated by an AUC of 0.85.
- AUC close to 1: Good model.
- AUC close to 0: Bad model.
- AUC equals 0.5: Model is as good as random.
- AUC greater than 0.8: Model is clinically considered useful.

Conclusions

- Logistic Regression emerged as the best model with 78.26% accuracy, leveraging significant predictors for diabetes prediction.
- Based on the AUC value of 0.85, the model is a good fit.
- If Doctors/Researchers give us information about number of pregnancies, glucose, BMI, Diabetes Pedigree Function and Blood Pressure then we can give the probability of subject having diabetes with an accuracy of 78.26.

References

- Kaggle. Pima Indians Diabetes Database. Dataset.
- Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling. Springer.

Acknowledgements

- Thank you Dr.Wang, all Professor from Statistics Department and group members for the contributions and support.