Unlocking Diabetes Prediction: Visual Insights and Analysis

Sowmya Sree Kemsaram, Narra Bhanu Prakash Reddy, Shreyas Shivaji Mali

1. Introduction

This analysis aims to predict diabetes in women based on health indicators such as insulin levels, blood pressure, body mass index (BMI), glucose levels, and age. The goal is to identify key risk factors and understand how these variables, both individually and in combination, influence the likelihood of developing diabetes. By examining these relationships, the study will highlight significant predictors like elevated glucose and high BMI, enabling early detection and targeted management strategies. The objective is to develop a predictive model that is both accurate and practical for healthcare settings, ensuring reliable diabetes risk assessments without complex computations. Ultimately, the analysis seeks to support early diagnosis and prevention, allowing healthcare providers to identify high-risk individuals, implement timely interventions, and offer personalized care plans to reduce the risk of diabetes.

2. Data Description

2.1 Dataset background:

The dataset used in this analysis originates from the National Institute of Diabetes and Digestive and Kidney Diseases. Its purpose is to predict whether a patient has diabetes based on diagnostic measurements. The data consists of 768 rows and 9 columns, with each row representing a female patient of Pima Indian heritage, at least 21 years old.

2.2 Response variable & predictor variables

The dataset includes several medical predictor variables and one Response variable, "Outcome."
The predictor variables are
Pregnancies: Number of times the patient has been pregnant.
Glucose: Blood glucose level.
BloodPressure:Blood pressure measurement.
SkinThickness: Thickness of the skin.
Insulin: Blood insulin level.
BMI: Body mass index.
DiabetesPedigreeFunction: A function indicating diabetes likelihood based on family history.
Age: Patient's age.
The response variable, “Outcome”, indicates whether the patient has diabetes (1 for Yes, 0 for No). The dataset provides a comprehensive basis for predicting diabetes risk using these health indicators.

An exploratory analysis showed a class imbalance in the "Outcome" variable, with a higher number of patients without diabetes compared to those with diabetes, as depicted in the bar chart. This imbalance should be considered when developing predictive models to ensure accurate results.

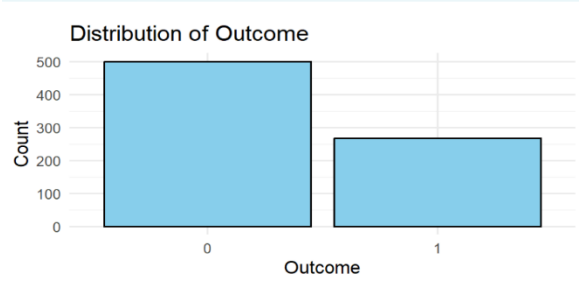

Figure 1: Distribution of response variable

2.3 Data Cleaning

We initiated the data cleaning process by checking any missing values (NA) and removing the outliers in the dataset.

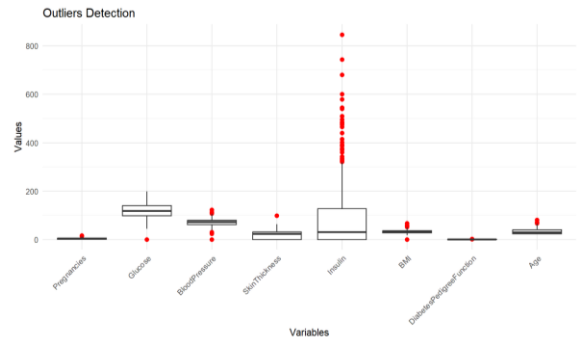

Figure 3: Box plot after removing outliers

After removing outliers, the variables show more compact distributions with fewer extreme values, especially for Insulin, which previously had values as high as 800. This adjustment made the data more normal and improved the distribution for variables like SkinThickness,BMI,DiabetesPedigreeFuncti -on, enhancing data quality and visualization. While there's a small risk of losing some rare but valid cases, the process improves the dataset's suitability for machine learning, balancing data integrity and model performance for more reliable analysis.

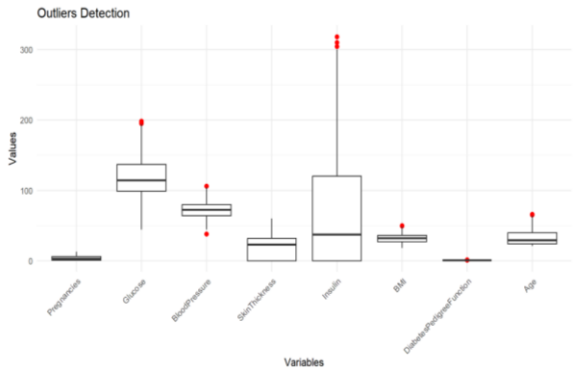

Figure 3: Box plot after removing outliers

3.METHODS AND RESULTS

3.1 Correlation matrix:

After cleaning the data, we examined the correlation between the variables, focusing on how the predictor variables relate to the response variable, Outcome. This step helps to identify which predictors are most strongly associated with diabetes and to understand the relationships among the variables.

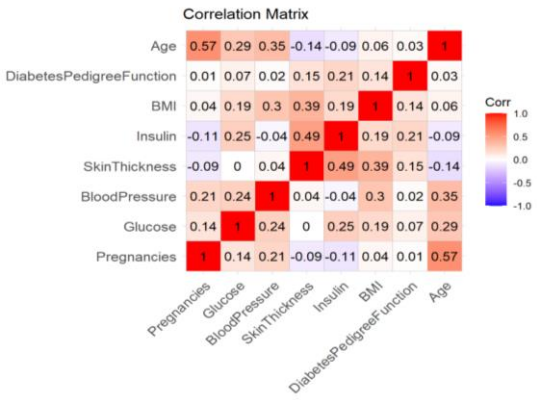

Figure 4: Correlation matrix

The correlation matrix shows that glucose has the strongest positive correlation with diabetes (0.49), highlighting its central role in the condition. BMI (0.27), age (0.26), and pregnancies (0.23) also show weak to moderate positive correlations, aligning with known risk factors for diabetes. Blood pressure (0.18), DiabetesPedigreeFunction (0.18), insulin (0.1) and skin thickness(0.03) show weak correlations, suggesting minimal direct relationships. Therefore, the variables of interest in predicting the outcome are BMI, age, pregnancies and glucose.

3.2 Random Forest Approach:

After removing the variables that were weakly correlated with the outcome, the dataset was split into a training set (70%) and a testing set (30%). A random forest model was then trained on the training data. The model's performance was evaluated using a confusion matrix to assess how accurately it predicts whether women have diabetes or not.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 114  27
         1  17  33

               Accuracy : 0.7696
                 95% CI : (0.7034, 0.8273)
    No Information Rate : 0.6859
    P-Value [Acc > NIR] : 0.006711

                  Kappa : 0.4401

 Mcnemar's Test P-Value : 0.174844

            Sensitivity : 0.8702
            Specificity : 0.5500
         Pos Pred Value : 0.8085
         Neg Pred Value : 0.6600
             Prevalence : 0.6859
         Detection Rate : 0.5969
   Detection Prevalence : 0.7382
      Balanced Accuracy : 0.7101

       'Positive' Class : 0
```

Figure 5: Confusion matrix

The confusion matrix shows that the model achieved an accuracy of 77%, meaning it correctly classified diabetes status most of the time. It has a high sensitivity of 87%, effectively identifying non-diabetes cases, but a lower specificity of 55%, indicating it struggled with identifying diabetes cases. The precision for predicting diabetes is 80.85%, while the negative predictive value is 66%, showing moderate reliability overall. The balanced accuracy, which accounts for both sensitivity and specificity, is 71%.

3.3 ROC Curve

The ROC curve shows the model's performance in distinguishing between diabetes and non-diabetes cases. It plots sensitivity (true positive rate) against 1-specificity(false positive rate), indicating how well the model balances correctly identifying positives while minimizing false positives. The Area Under the Curve (AUC)is 0.79, suggesting the model has a good ability to discriminate between the two classes, with 1.0 being perfect and 0.5 indicating no better than random guessing.
The higher AUC value reflects a reasonably strong performance in predicting diabetes.

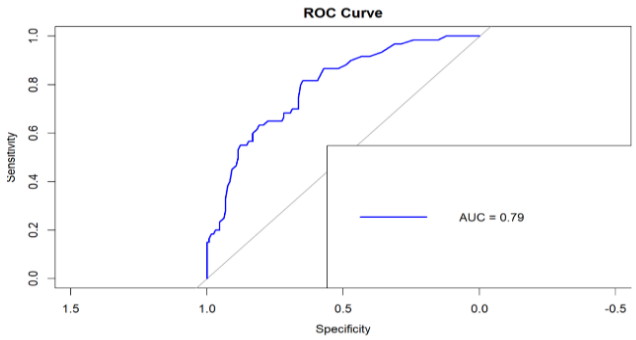

Figure 6: ROC curve

CONCLUSION

This analysis showed that predicting diabetes in women using health indicators like glucose, BMI, age, and pregnancies are feasible. Data cleaning improved model suitability, and correlation analysis identified key predictors. The random forest model achieved 77% accuracy with good sensitivity but lower specificity. The ROC curve's AUC of 0.79 indicated strong performance in distinguishing diabetes cases. Overall, the study underscores the value of these predictors and machine learning for early diagnosis, while suggesting further improvements, such as addressing data imbalance, to boost accuracy.

APPENDIX

Link for the code:
https://github.com/sowmya182/DiabetesPrediction.git

References:

1. https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset
2. https://ieeexplore.ieee.org/document/8878802
3. https://ieeexplore.ieee.org/document/8631968
4. https://shmpublisher.com/index.php/joiser/article/view/245