Decoding Diabetes: Logistic Regression and KNN

Sridhar Malladi
The University of Texas at San Antonio, San Antonio TX, 78249

Abstract

This poster dives into the predictive power of statistical learning techniques for identifying the onset of diabetes in Pima Indians population. By using the data provided by the Pima Indians Diabetes Database, we can employ logistic regression and K-nearest neighbors (KNN) as our primary analytical tool. The study not only aims to develop accurate predictive models but also seeks to unravel the critical features influencing diabetes prediction. The results provide insights into the comparative effectiveness of the logistic regression and KNN in decoding the complexities of the diabetes onset.

.

Introduction

Diabetes poses a significant global health challenge, which makes it necessary for effective predictive tools for early detection. The Pima Indians Diabetes Database offers a unique opportunity to explore the intricate relationship between diagnostic measures and diabetes onset. In this analysis, we focus on two powerful statistical analysis: Logistic Regression and KNN. Logistic Regression provides framework for understanding the impact of individual features while KNN helps at capturing complex patterns in the data. By combining these methods we aim to uncover robust predictive models that contribute to the ongoing efforts In diabetes research and management.

Purpose

The primary purpose of this analysis is to develop accurate predictive models for the onset of diabetes in the Pima Indians population and to evaluate the performance of the both the algorithms used in the analysis. Compare the effectiveness of logistic regression and KNN in predicting diabetes, considering their strengths and limitations. Provide insights into how each models contributes to the overall predictive accuracy. Unravelling the critical features influencing diabetes predictions using logistic regression and KNN. Understanding the importance of specific diagnostic measures in determining the likelihood of diabetes onset. Finally, to contribute valuable insights to the broader field of diabetes research and management. Also, enhancing the understanding of statistical learning techniques in the context of diabetes prediction, potentially influencing future approaches and interventions.

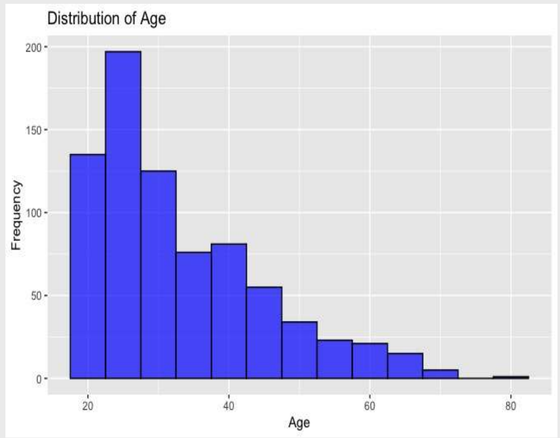
Data structure

The dataset comprises 768 observations with 9 variables, including features like Pregnancies, Glucose, and BMI, alongside the binary Outcome variable indicating diabetes presence. There are no missing values. However, variables like SkinThickness and Insulin exhibit zero values. The Dataset displays numeric and integer types, with the outcome variable representing a class imbalance. Summary statistic reveal variable distributions, and the exploration sets the stage for applying logistic regression and KNN. Addressing zero values and class imbalance is crucial for accurate model development and interpretation

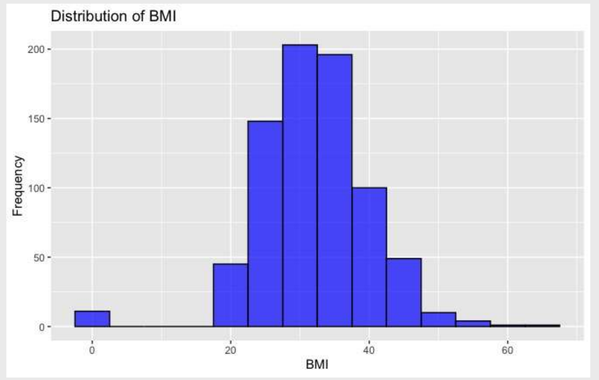
Data Structure:-

```
'data.frame':    768 obs. of  9 variables:
 $ Pregnancies              : int  6 1 8 1 0 5 3 10 2 8 ...
 $ Glucose                  : int  148 85 183 89 137 116 78 115 197 125 ...
 $ BloodPressure            : int  72 66 64 66 40 74 50 0 70 96 ...
 $ SkinThickness            : int  35 29 0 23 35 0 32 0 45 0 ...
 $ Insulin                  : int  0 0 0 94 168 0 88 0 543 0 ...
 $ BMI                      : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
 $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
 $ Age                      : int  50 31 32 21 33 30 26 29 53 54 ...
 $ Outcome                  : int  1 0 1 0 1 0 1 0 1 1 ...
              Pregnancies                     Glucose               BloodPressure
                        0                           0                           0
            SkinThickness                     Insulin                         BMI
                        0                           0                           0
DiabetesPedigreeFunction                         Age                     Outcome
                        0                           0                           0
```

Results: Histogram



The histograms shows distributions of the Age, Glucose and the BMI of the people. It is a start of understanding of the predictive model.



Results: Logistic regression

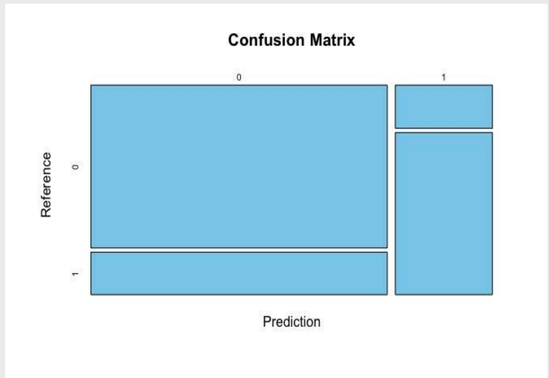
Aspects	Value
Accuracy	0.81
Kappa	0.51
Sensitivity	0.92
Specificity	0.555
PosPredValue	0.7931
NegPredValue	0.7895
Prevalence	0.6494
Detection rate	0.5974
Balanced Accuracy	0.7378

The model correctly classified 78.9% of the data points.

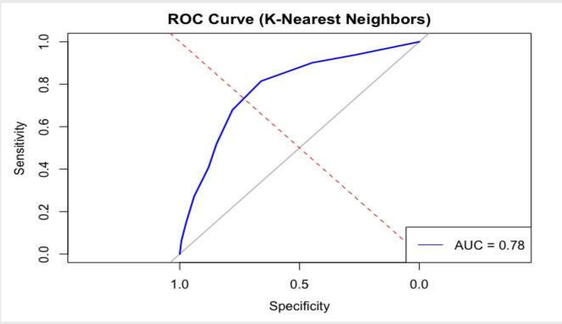
55.6% of the data points that the model predicted to be positive were actually positive.

78.9% of the data points that were actually positive were correctly classified as positive by the model.

Overall, the model is performing well. It has a high accuracy, precision, and recall. However, there is still room for improvement as the model is not 100% perfect.



Results: KNN



The ROC curve shows that the KNN model has an AUC of 0.80.

This is a good score, indicating that the model is able to distinguish between individuals with and without diabetes with a fair degree of accuracy.

The TPR curve increases steadily as the FPR increases and vice versa.

The ROC curve shows a sharp kink around the FPR value of 0.2. This suggests that there is a specific threshold value at which the classifier's performance significantly changes.

Aspects	Value
Accuracy	0.7316
Kappa	0.3826
Sensitivity	0.8467
Specificity	0.5185
PosPredValue	0.7651
NegPredValue	0.6462
Prevalence	0.6494
Detection rate	0.5498
Balanced Accuracy	0.7126

Conclusions

- Logistic Regression excelled in providing interpretable result with a notable accuracy of approximately 81%, contributing insights on feature and its impact on diabetes. Understanding the contribution of each measure is important for targeted screening.
- This study enhances understanding of statistical learning techniques, particularly logistic regression for diabetes prediction. This knowledge can contribute to methods of diabetes research and may influence future approaches.
- The data driven approach provides a valuable tool to identify individuals at risk, informing personalized interventions and contributing to more effective diabetes prevention strategies.
- The model demonstrated high sensitivity, effectively identifying individuals with diabetes, and specificity, correctly identifying non-diabetic cases.
- The balanced accuracy of 73.78% reflects a satisfactory trade-off between sensitivity and specificity, contributing to the overall effectiveness of the model in classifying both positive and negative cases.