Statistical Report

Calvin Soe Frederick

November 10, 2023

1 Introduction

Diabetes is among the most prevalent chronic diseases in the world. The data set being used in this report is a clean data set of 70,692 survey responses from a survey conducted in the US in 2015. This report aims to showcase different classification methods for predicting diabetes status and propose the best classifier based on the investigations of the goodness of fit of that classifier. "Diabetes binary" would be the response variable in this report.

2 Methods

In analysing the dataset, different visualisation and statistical techniques were applied, aligning with the variable types. The categorical response variable was summarised using a barplots, while discrete input variables were similarly depicted to examine their distribution. Continuous variables were explored through histograms and described by variance, standard deviation, and interquartile range metrics.

Feature selection was guided by assessing the association strength between inputs and the response variable. Odds ratios from frequency tables were calculated for categorical inputs and boxplots for continuous inputs highlighted their relationships with the outcome. These steps informed the inclusion of variables in the subsequent classification models.

A decision tree algorithm was also implemented to refine feature selection, ensuring only variables with significant influence were retained. The classification models — comprising Decision Tree, Logistic Regression, Naive-Bayes, and K-Nearest Neighbours — was evaluated using ROC curves, AUC scores, and Accuracy measures to determine model fitness and predictive performance.

3 Results

3.1 Summary of dataset

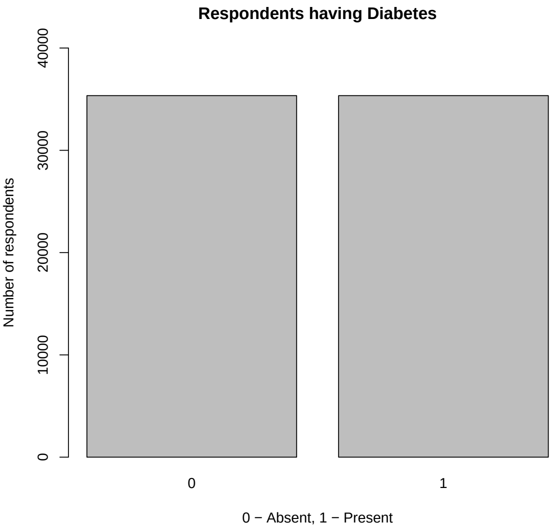


Figure 1: Barplot of respondents having diabetes

Equal proportions of respondents indicated the presence of diabetes/prediabetes, and having no diabetes at all. (Figure 1)

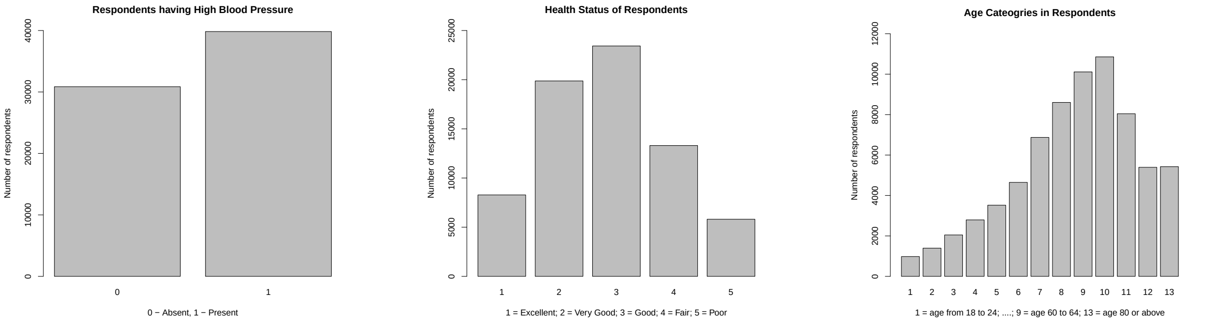
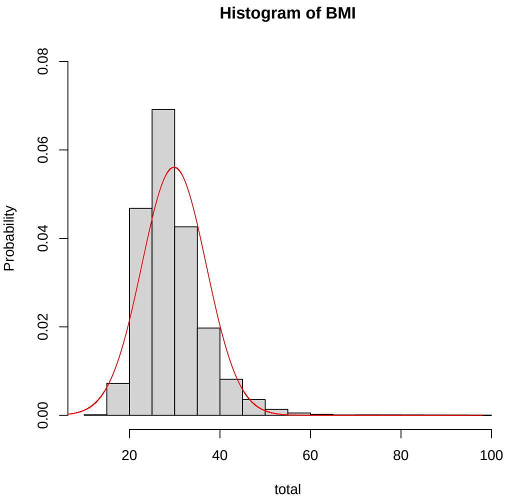
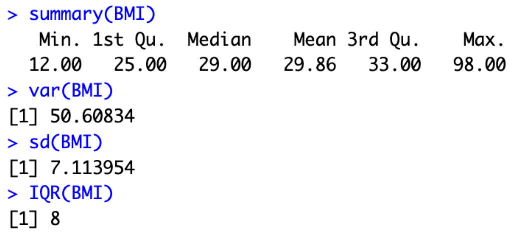


Figure 2: Different types of graphs to present data. Left to right: Bar chart of respondents having high blood pressure, Bar chart of health status of respondents: 1 = excellent; 2 = very good; 3 = good; 4 = fair; 5 = poor, Bar chart of the age categories of respondents

Respondents with higher blood pressure outnumbered those without, as depicted in the bar chart (Figure 2). The distribution of respondents' health status revealed a predominance of 'Good' health, followed by 'Very Good', 'Fair', 'Excellent', and finally 'Poor' (Figure 2). Age analysis indicated a majority within the 65-70 years category, suggesting an aging demographic in the study (Figure 2).



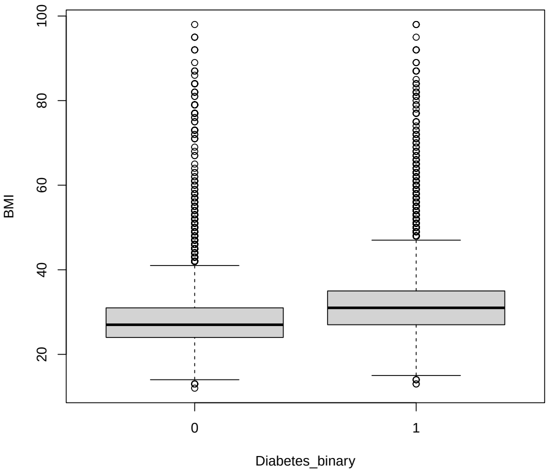


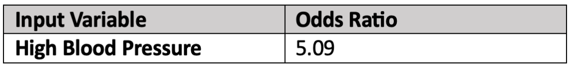
```
> summary(BMI)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.00   25.00   29.00   29.86   33.00   98.00 
> var(BMI)
[1] 50.60834
> sd(BMI)
[1] 7.113954
> IQR(BMI)
[1] 8
```

Figure 3: Graphical and numerical summaries. Left to right: Histogram of BMI in respondents, Numerical summary of BMI, including, Variance, Standard Deviation and Interquartile range

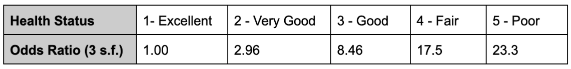
The BMI histogram (Left) illustrates a right-skewed distribution, with a unimodal peak, comparing with overlaid normal density curve. This skewness is evidenced by a mean BMI (29.9) that exceeds the median (29.0), indicating a tail of higher values (Figure 3). The numerical summary details a range from a minimum of 12.0 to a maximum of 98.0, with a variance of approximately 50.6 and a standard deviation of around 7.11, signifying substantial dispersion. The interquartile range (IQR) of 8.00 points to variability in the central distribution of BMI values.

3.2 Feature Selection





Input Variable	Odds Ratio
High Blood Pressure	5.09



Health Status	1- Excellent	2 - Very Good	3 - Good	4 - Fair	5 - Poor
Odds Ratio (3 s.f.)	1.00	2.96	8.46	17.5	23.3

Figure 4: Strength of associations between input and response variable. Left: Box plot of BMI and Diabetes. Right top: Odds ratio of High BP and diabetes. Right bottom: Odds ratio of health status and diabetes.

The box plot analysis (Figure 4: Left) reveals a trend whereby higher BMI is associated with an increased likelihood of diabetes, indicating BMI as a significant factor in diabetes risk.

The table (Figure 4: Right top) indicates an odds ratio of 5.09 for the association between high blood pressure and the likelihood of diabetes. This suggests that individuals with high blood pressure are approximately five times more likely to have diabetes compared to those without high blood pressure, signifying a strong positive association between high blood pressure and the incidence of diabetes. This finding highlights the including blood pressure as part of the model.

For health status (Figure 4: Right bottom) as the ratings progress from 'Excellent' to 'Poor', the odds ratios increase from 1.00 to 23.3, indicating a higher likelihood of diabetes with declining health status. These figures underscore the significant relationship between poorer health status and increased diabetes risk.

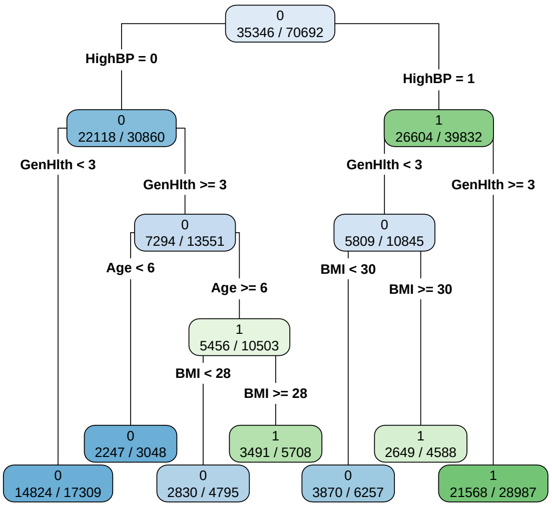
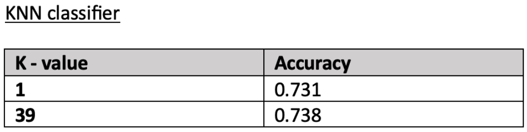


Figure 5: Decision Tree Model

Additionally, the decision tree analysis (Figure 5) has underscored the significance of four key predictors: 'HighBP', 'GenHlth', 'Age', and 'BMI'. These variables have been identified as influential in determining the likelihood of diabetes, each contributing distinctively to the model's predictive accuracy. The inclusion of these features in the classification model leverages their statistical importance, enhancing the model's capability to discern the presence of diabetes effectively.

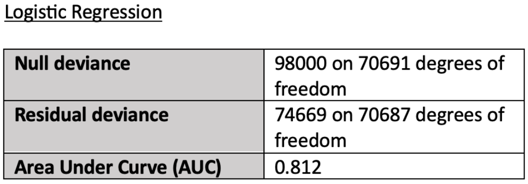
4 Classification Models



KNN classifier

K - value	Accuracy
1	0.731
39	0.738

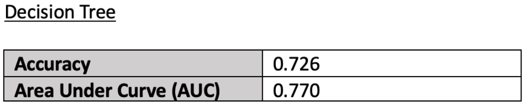
Figure 6: KNN



Logistic Regression

Null deviance	98000 on 70691 degrees of freedom
Residual deviance	74669 on 70687 degrees of freedom
Area Under Curve (AUC)	0.812

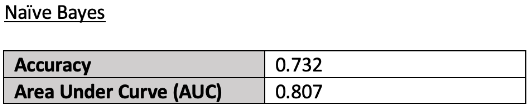
Figure 7: Logistic Regression



Decision Tree

Accuracy	0.726
Area Under Curve (AUC)	0.770

Figure 8: Decision Tree



Naïve Bayes

Accuracy	0.732
Area Under Curve (AUC)	0.807

Figure 9: Naive Bayes

The KNN analysis (Figure 6) with k = 1 yielded an accuracy of 0.731, indicating a moderate predictive power. Optimization of the parameter revealed k = 39 as the most effective, enhancing accuracy to 0.738. This increment contributes to improved model performance.

The logistic regression analysis (Figure 7) a significant improvement in the model fit when incorporating predictors—general health, high blood pressure, BMI, and age. The logistic regression model effectively distinguished between the presence and absence of the diabetes, achieving an AUC of 0.812, which suggests a strong predictive capability. This performance significantly surpasses a baseline AUC of 0.5, denoting more than random classification accuracy.

The decision tree model (Figure 8) constructed to predict diabetes status based on general health, high blood pressure, BMI, and age, demonstrated a strong ability to discriminate between outcomes, with an AUC value of 0.770. The chosen control parameter with a minimum split size of 7000, representing 10 percent of the observations, suggests a model geared towards capturing broader patterns rather than overfitting to the training data. The use of the 'information' split criterion aimed to maximize the information gain at each decision node. The model also provides an accuracy of 0.726, evidencing its predictive ability.

The Naive Bayes model (Figure 9) trained to predict diabetes using general health, high blood pressure, BMI, and age as predictors, achieved an accuracy of 0.732, indicating a substantial level of correct classifications. Moreover, the model's ability to distinguish between patients with and without diabetes was further evidenced by an AUC of 0.807, demonstrating a strong predictive performance. This AUC value significantly exceeds the threshold of 0.5, typically associated with random chance, underscoring the model's effectiveness.

5 Analysis

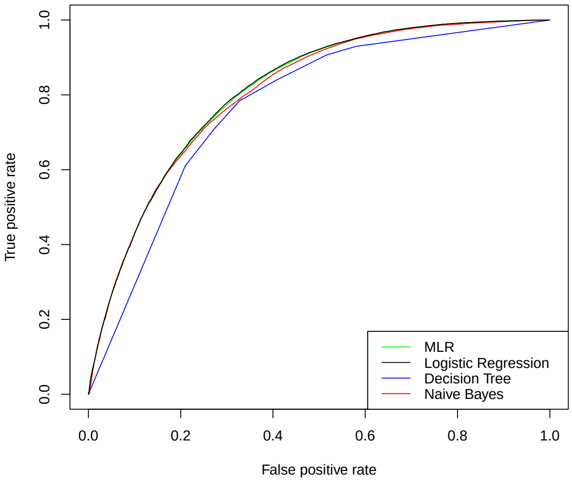


Figure 10: ROC,AUC curves of the different models

Considering the provided results for each classifier, we can compare their performance based on the reported accuracy and AUC values, which are common metrics used to evaluate the effectiveness of classification models.

Multiple Linear Regression Model is unsuitable for binary outcomes as it predicts values outside the 0 and 1 range, violating the binary distribution. Logistic Regression Model, tailored for binary data, models the probability within the 0 to 1 range, adhering to the binary nature of the response variable.

Given these results, the Logistic Regression model is the best classifier for predicting diabetes. (Figure 10) The justification for this selection is primarily its highest AUC value of 0.812, which indicates the best predictive ability among all models. An AUC closer to 1 signifies that the model has a high true positive rate and a low false positive rate, making it a reliable predictor. The significant reduction in deviance also suggests that the logistic regression model has a good fit to the data.

While accuracy is an important measure, it is not the sole criterion for model selection, especially when dealing with imbalanced datasets where the prevalence of one class significantly outweighs the other. In such cases, accuracy can be misleading, and AUC becomes a more critical measure as it considers the balance between true positive and false positive rates. Therefore, despite the KNN classifier having a slightly lower but comparable accuracy to Naive Bayes, and the Decision Tree model also having a reasonably good AUC, the Logistic Regression model's superior AUC value makes it the most robust classifier among those evaluated.

Additionally, logistic regression's parametric nature allows for better generalisation, which can be a deciding factor when predictive performance on unseen data is a priority.

In conclusion, the Logistic Regression model is recommended as the most suitable classifier for predicting diabetes in this scenario due to its highest AUC value.