Statistical Report: Choosing a Classifier for Predicting Diabetes Status

Introduction and Objectives

As diabetes stands as one of the most prevalent chronic diseases worldwide, its impact on public health is far-reaching. The surge in diabetes cases has prompted an urgent need to comprehend the intricate interplay of factors contributing to its prevalence. In this analysis, we strive to examine its prevalence, risk factors, and associated health outcomes and ultimately, mitigate the escalating burden of diabetes on global health. Hence, we aim to achieve that by finding the best classifier in predicting the diabetes status of a person.

Data Description

In the dataset provided, there are 11 features and a response variable. According to the type of data, here is the breakdown of the 11 features and the response variable.

We extracted data by columns using the function 'data$variable'. Additionally, we employed the 'as.factor' function to declare categorical variables in R.

Response Variable: Diabetes_binary (0 = no diabetes; 1 = prediabetes or diabetes)

Categorical Variables: Diabetes_binary, HighBP *(0 = no high BP; 1 = high BP)*, HighChol (*0 = no high cholesterol; 1 = high cholesterol)*, CholCheck (*0 = no cholesterol checks in 5 years; 1 = has cholesterol check in 5 years)*, Stroke *(Ever told you had a stroke? 0 = No; 1 = Yes)*, HeartDiseaseorAttack *(coronary heart disease or myocardial infarction: 0 = no; 1 = yes),* PhysActivity *(physical activity in past 30 days, not including jobs: 0 = no; 1 = yes)*, Fruits *(consume fruits >=1 per day: 0 = no; 1 = yes)*, Veggies *(consume vegetables >=1 per day: 0 = no; 1 = yes)*, HvyAlcoholConsump *(adult men > 14 drinks per week and adult women >7 drinks per week: 0 = no; 1 = yes)*, AnyHealthcare *(Do you have any kind of health care coverage? 0 = no; 1 = yes)*, NoDocbcCost *(Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no; 1 = yes)*, DiffWalk *(Do you have serious difficulty walking or climbing stairs? 0 = no, 1 = yes)*, Sex *(0 = female; 1 = male)*

Quantitative Variables: BMI, MentHlth, PhysHlth, GenHlth, Age, Education, Income

Analysis on the Strength of Association between Input and Response Variable

1. **Boxplot**

   To assess the association strength between the categorical response variable, 'Diabetes_binary' and quantitative variables, we utilise side-by-side boxplots. In particular, we plotted x-axis to be 'Diabetes_binary' while y-axis is the quantitative variable. For the interpretation of boxplot, greater overlap in the

boxplots signifies weaker association, whereas reduced overlap suggests a stronger association.

For the feature, ‘BMI’, we observe that the median BMI for individuals with diabetes is higher than the median BMI for those without diabetes. This indicates a potential strong association between a higher BMI and the presence of diabetes.

For the feature, ‘MentHlth’, we observe that the median number of poor mental health days for individuals with diabetes is the same as for those without diabetes. However, the value of the 3rd quartile of those with diabetes is slightly higher than those without diabetes. This indicates a potential slight to no association between a higher number of poor mental health days and the presence of diabetes.

For the feature, ‘PhysHlth’, we observe that the median number of poor physical health days for individuals with diabetes is slightly higher as for those without diabetes. Similarly, the value of the 3rd quartile of those with diabetes is much higher than those without diabetes. This indicates a strong association between a higher number of poor physical health days and the presence of diabetes.

For the feature, ‘GenHlth’, we observe that the median score on general health for individuals with diabetes is much higher as for those without diabetes. The median score for individuals with diabetes is 4 (fair) while for those without diabetes it is 2 (very good). Similarly, the value of the 3rd quartile of those with diabetes is much higher than those without diabetes. This indicates a strong association between a higher score on general health and the presence of diabetes.

For the feature, ‘Age’, we observe that the median age for individuals with diabetes is higher as for those without diabetes. The median score for individuals with diabetes is 10 (age 65-70) while for those without diabetes it is 8 (age 55-59). Similarly, the value of the 3rd quartile of those with diabetes is higher than those without diabetes. This indicates a strong association between larger age and the presence of diabetes.

For the feature, ‘Education’, we observe that the median education level for individuals with diabetes overlaps and is the same as for those without diabetes. In the side-by-side boxplots, both plots are the exact same. This indicates no association between education level and the presence of diabetes.

For the feature, ‘Income’, we observe that the median income for individuals with diabetes is lower as for those without diabetes. Similarly, the value of the 1st quartile of those with diabetes is much lower than those without diabetes. This indicates a strong association between income and the presence of diabetes.

2. **Odds Ratio**

To assess the association strength between the categorical response variable, ‘Diabetes_binary’ and categorical variables, we utilise odds ratio. In particular, we create a contingency table using the function ‘table(variable, Diabetes_binary)’. In this case, the formula is given, odds ratio = (x[1,1]*x[2,2]) / (x[1,2]*x[2,1]).

For the interpretation of odds ratio, we have split the data into 2 categories — those with odds ratio larger than 1 and smaller than 1.

Using Table 1, we can interpret that the odds of having diabetes are *odds ratio* times higher among individuals with (*feature = 1)* as compared to those without (*feature = 0)*. For example, the odds of an individual having diabetes is 5.08 times higher among individuals with high blood pressure as compared to those without.

Feature	Odds Ratio
HighBP	5.088477
HighChol	3.296316
CholCheck	6.491553
Stroke	3.093272
HeartDiseaseorAttack	3.656197
DiffWalk	3.807365
Smoker	1.412383
AnyHealthcare	1.251644
NoDocbcCost	1.326217
Sex	1.195343

Table 1: Features with Odds Ratio larger than 1

Using Table 2, we can interpret that the odds of having diabetes are *odds ratio* times lower among individuals with *(feature = 1)* compared to those without (*feature = 0)*. For example, the odds of an individual having diabetes is 0.49 times lower among individuals who had physical activity as compared to those who did not.

Feature	Odds Ratio
PhysActivity	0.4939616
Veggies	0.6763796
HvyAlcoholConsump	0.3653163
Fruits	0.8007655

Table 2: Features with Odds Ratio less than 1

3. **Conclusion**

   As such, given that there is no association between education level and presence of diabetes, we will be omitting this feature in the dataset.

Use of N-Fold Cross Validation with Train and Test Data

In order to assess the performance of the classifiers for predicting diabetes status, we used a methodology involving N-fold cross-validation. The entire dataset was initially categorised into two distinct groups: individuals with diabetes and those without. Subsequently, for each category, 80% of the data was randomly sampled to construct the training dataset, while the remaining 20% constituted the test dataset. This was to ensure that both the training and testing sets maintained a representative distribution of individuals with and without diabetes. In this context, a 5-fold cross-validation approach was utilised. Through 5 iterative folds, metrics such as mean accuracy, error, True Positive Rate, False Positive Rate and False Negative Rate were collected for each fold. This systematic evaluation provided a robust assessment of the predictive models' effectiveness in predicting diabetes status across diverse subsets of the dataset.

Choice of Classification Methods

In this statistical report, we explored 3 different classifiers — Decision Tree, Naive Bayes and Logistic Regression. In the context of predicting the likelihood of an individual having diabetes, we will be focusing on 4 metrics — False Positive Rate (FPR), Precision, Area Under the ROC Curve (AUC) and Error. Our primary objective is to minimise both the FPR and overall Error. A low FPR is crucial because misdiagnosing individuals as having diabetes when they do not can lead to unnecessary treatments and potential health risks. Additionally, we strive for high Precision and Area

AUC values. Elevated Precision indicates accurate identification of positive cases, while a high AUC signifies a strong overall performance in predicting the likelihood of individuals having diabetes.

	FPR	Precision	AUC	Error
Decision Tree	0.2462518	0.7384448	0.7202934	0.2797066
Naive Bayes	0.2897922	0.71624	0.7208312	0.2791688
Logistic Regression	0.215272	0.7680005	0.7486136	0.2513864

1. **Decision Tree**
   For the Decision Tree classifier, the FPR and Error values are considerably low and Precision and AUC values are decently high. This suggests that the model is effective in identifying individuals with diabetes while minimising false alarms.

2. **Naive Bayes**
   For the Naive Bayes classifier, the FPR and Error values are low and Precision and AUC values are high. This suggests that the model is slightly effective in identifying individuals with diabetes while minimising false alarms.

3. **Logistic Regression**
   For the Naive Bayes classifier, the FPR and Error values are much lower and Precision and AUC values are much higher. This suggests that the model is very effective in identifying individuals with diabetes while minimising false alarms.

4. **Conclusion**
   Comparing the performance metrics across the three classifiers, it is evident that logistic regression stands out as the superior model. The FPR and Error values for logistic regression are notably lower compared to the other two classifiers. Additionally, when considering Precision and AUC values, logistic regression demonstrates the highest performance, outperforming the other classifiers.

   This collective assessment indicates that logistic regression is particularly effective in identifying individuals with diabetes while minimising the occurrence of false alarms. The combination of lower FPR and Error rates, coupled with higher Precision and AUC values, positions logistic regression as the most reliable model amongst the three for predicting diabetes status.

Comment on Logistic Regression Model

For logistic regression, the threshold is set to 0.482 as it was found to give the highest Precision and AUC values and lowest Error and FPR values. Diving into the logistic regression model, we created a new model without utilising the N-fold cross validation. With the help of the function “summary” and looking at the p-value, we found out that there were a few features that are insignificant to the model. In which, they are “Smoker”, “PhysActivity”, “Fruits”, “Veggies”, “AnyHealthcare”, “NoDocbcCost” and “MentHlth”. As such, we replaced the current data with a new one that does not have all these features and ran it with the 5-fold cross-validation again to find the different metrics.

	FPR	Precision	AUC	Error
Logistic Regression	0.215272	0.7680005	0.7486136	0.2513864
Logistic Regression (improved)	0.08894929	0.8610432	0.7311577	0.2688423

For the improved logistic regression, the threshold is set to be 0.32. In comparison to the previous model where we did not remove any additional features, in the improved version the FPR value has dropped significantly while the Precision value has increased greatly.

Equation:
Diabetes_binary = -7.104586 + 0.737517*HighBP + 0.585921*HighChol + 1.367341*CholCheck + 0.076190*BMI + 0.160773*Stroke + 0.250720*HeartDiseaseorAttack - 0.752663*HvyAlcoholConsump + 0.588440*GenHlth - 0.009264*PhysHlth + 0.117618*DiffWalk + 0.278435*Sex + 0.154517*Age - 0.066834*Income

Henceforth, logistic regression is the best classifier as it allows for interpretability of significance of each feature to the classifier. This allows for the model to be tweaked to achieve the best results. However, logistic regression assumes a linear relationship between the *features* and the log-odds of the response variable. This means the model assumes that the effect of a one-unit change in a *feature* and the log-odds of the response variable change linearly with the *feature*.