STATISTICAL REPORT: PREDICTING DIABETES STATUS

PART 0: INTRODUCTION

Diabetes is the most prevalent chronic disease in the world which arises from various biological & lifestyle factors. This report aims to explore, test & evaluate the performance of various classification models when predicting the diabetes status of a respondent. The data set used for this analysis is a clean data set containing 100000 survey responses, provided by the author Mohammed Mustafa. This report looks into 3 classification models (K-Nearest Neighbours, Decision Tree & Logistic Regression) & makes use of the data set to come to a conclusion on the best classifier to predict diabetes status.

PART I: EDA (EXPLORING VARIABLES & ASSOCIATION)

The dataset contains 100000 responses, each consisting of 4 quantitative variables *(Age, BMI, HbA1c levels & blood glucose level)* & 5 categorical variables *(Gender, Hypertension status, Heart disease status, Smoking history & Diabetes status)*. To gain better insight into the dataset & each variable, plots were used to visualise the distribution of each variable & investigate the possible association between each feature variable & the response variable.

1.1: Observing distribution of variables

The proportions of each categorical variable can be visualised with pie charts. The plotted pie charts are as shown in Fig. 1.

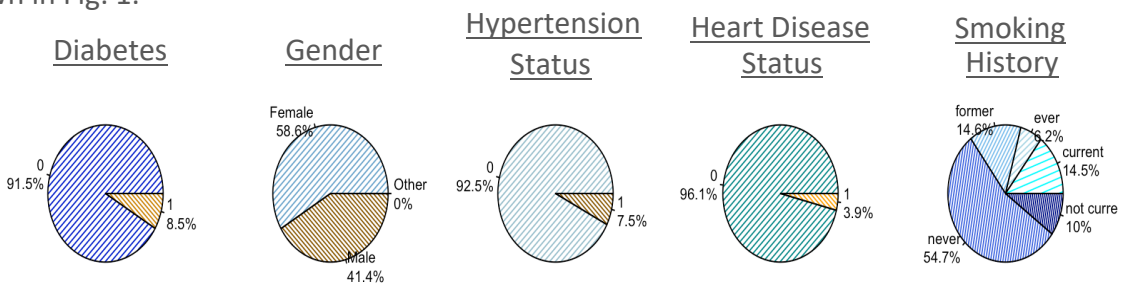

Fig. 1: Pie charts showing the distribution of each categorical variable

From Fig.1, the following it can be observed that the data is very unbalanced, with the majority value of each categorical variable constituting a significantly large proportion of the variable. For *"Smoking history"*, it contains a category *"No Info"* which makes up 35.8% of the data set. Omitting these rows is not justifiable as detail would be lost within the data set.

For quantitative variables, histograms were used as aid to check if the variables are normally distributed. On top of histograms, density lines were included to have a better idea of properties such as symmetry & skewness. The plotted histograms & respective density lines are as shown in Fig. 2.

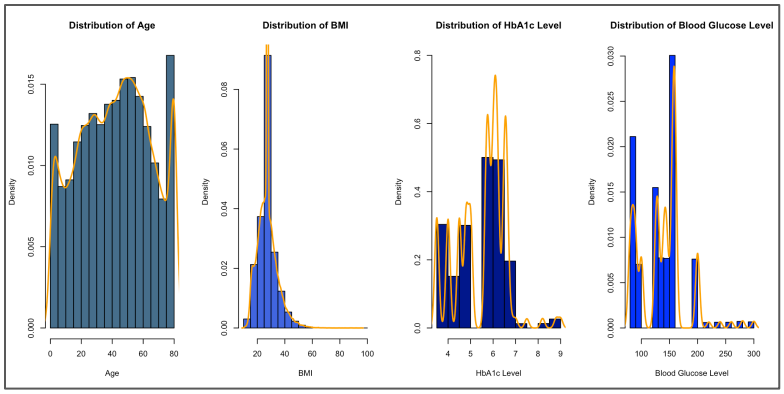

Fig. 2: Histograms showing the distribution of each quantitative variable

From Fig. 2, it can be inferred that the distribution of each variable is skewed. This suggests that they are unlikely to be normally distributed. Among the variables, the distribution of BMI is unimodal while the others are multimodal.

1.2: Investigating association between variables

For categorical features, bar plots are used to compare the proportion of respondents with diabetes with respect to each categorical feature & their categories. The bar plots are as shown in Fig. 3.

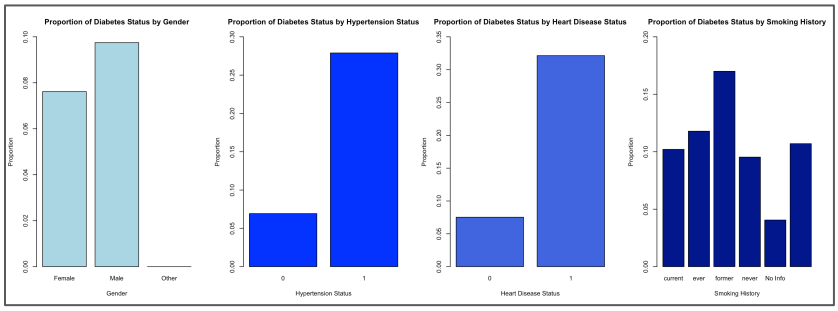

Fig. 3: Bar plots to investigate association between categorical features & response variable

From Fig. 3, it shows that for each categorical feature, there exists a category for each feature which consists of a higher proportion of respondents with diabetes. This suggests the possible association between each categorical feature & the diabetes status of the respondents.

For quantitative features, box plots were used to compare the values of each feature with regards to the diabetes status of the respondents. Fig 4 illustrates the box plots used for comparison.

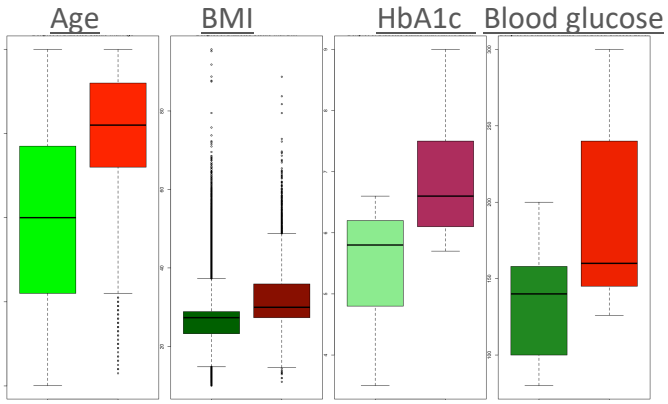

Fig. 4: Box plots to identify association between quantitative features & response variable

From Fig 4, it can be observed that the median value of each feature is consistently larger when comparing the respondents with diabetes & those without. This shows as evidence that there may be an association between each quantitative feature & the diabetes status of the respondents.

1.3: Key findings

After exploring the data, it can be concluded that each feature variable could have some form of association with the response variable, *diabetes status*. In the medical context, HbA1c & blood glucose levels are used in practice as indicators of diabetes. Hypertension & heart disease are commonly known to co-occur with diabetes due to common risk factors between these medical conditions. Factors due to lifestyle habits, such as BMI & smoking history, as well as biological factors, such as age & gender, are known to be indirectly related to diabetes status. Therefore, it is appropriate to consider every variable during feature selection. It is also important to notice that since some variables are indirectly related to one another (e.g. BMI & heart disease), the variables cannot be assumed to be independent of one another & that classification models such as Naïve Bayes should not be considered.

1.4: Handling *"No Info"* category in *Smoking history*

From the data set, 35.8% of respondents indicated *"No Info"* as their response for smoking history. Smoking history does have an implication on blood nicotine level & more intense smokers tend to have a higher blood nicotine level. Unlike *NA* rows, *"No Info"* is a valid category due to the sensitive nature of smoking history. However, it does not provide any insight into the blood nicotine levels of the respondents. In other words, there is missing information with regards to the blood nicotine level in the respondents. A possible approach to handle these rows would be to carry out multiple imputation to provide valid inferences of smoking history for the *"No Info"* rows. A chi-square test of independence can be carried out to observe the missingness mechanism of *"No Info"* as a category of *smoking history.* After carrying out a chi-square test of independence on "*diabetes"* & *"smoking history"*, the resulting p-value is very low (p-value < 2.2e-16). This suggests that there is sufficient evidence to conclude that there is strong association between *diabetes* & *smoking history.* From this we can assume that the missingness mechanism of *"No Info"* should be considered as Missing at Random & not Missing Completely at Random. This justifies the suggested approach to carry out multiple imputation on the *"No Info"* rows within *"smoking history"*.

1.5: Feature selection

To determine the final features to be included in our models, we construct a baseline logistic regression model with all feature variables & observe the significance of each feature in the model. The less significant features are collated into Table 1.
To address the 3 categories for smoking history, we can consider the relation between blood nicotine levels & diabetes status. Higher blood nicotine levels from smoking cigarettes puts people at a higher risk of becoming diabetic. This suggests that it may be more impactful to categorise the responses in the data set into 3 categories instead based on estimated blood nicotine level from their

Feature & Category	p-value
Gender = "Other"	**0.92664**
Smoking history = "ever"	**0.58154**
Smoking history = "former"	**0.12203**
Smoking history = "not current"	**0.01115**

Table 1: p-value of less significant categories

smoking intensity. The categories would be namely: "Former", "Current" & "Never". For *"gender = 'Other'"*, the high p-value could be due to its low proportion in the data set. The biological profile of respondents who select this category are likely to be much more complex than that of *"Male" & "Female"* respondents, hence it would be kept as its own category. Therefore, the features to be included in our model construction are *Gender, Age, Hypertension, Heart Disease, Smoking History ("Former", "Current" & "Never"), BMI, HbA1c level & Blood glucose level.*

PART II: METHODS

To produce the best classifier to predict diabetes status, supervised learning methods such as K-Nearest Neighbours (KNN), Decision Tree & Logistic Model can be used. To construct the most optimal model, a 5-fold cross validation will be carried out on a subset of the data set to obtain the best parameters for our KNN & decision tree classifiers. Firstly, the data set is split into 5 folds, ensuring that the proportion of each variable & category is representative of the actual proportions in the original data set. 1 fold would then be selected to train the models. The selected fold is further divided into 5 sub-folds containing a proportional amount of each variable & category, which will be used for hyper parametric tuning of the KNN & decision tree classifiers. These classifiers are tuned such that they produce the highest True Positive Rate (TPR). When predicting someone's diabetes status, it is important that it is detected early & accurately such that healthcare services can provide timely interventions & valuable resources are not wasted. Hence, TPR is prioritised for determining the best parameters.

2.1: K-Nearest Neighbour (KNN) Classification

For a specific value K, KNN algorithm classifies a new point as the majority class amongst the K nearest neighbours (data points). These neighbours are determined based on a distance metric. The algorithm takes in quantitative features to determine the class of the response variable. In this case, the quantitative variables *Gender, Age, BMI, HbA1c level & Blood glucose level* were used as inputs for the KNN algorithm. Since hypertension & heart disease share the same risk factors as diabetes, they have been included into the algorithm with quantitative values 0 & 1 representing "0" & "1" respectively. Since KNN is a distance-based algorithm, it is vital that the input features are standardised to reduce biasness when selecting the K neighbours. By collating the mean TPR per fold for each odd value of K from 1 to 141, it can be concluded that the highest TPR obtained from the KNN classifier is 0.6745021 which occurs when K is 1. Fig. 5 shows a graph illustrating the TPR for each value of K.

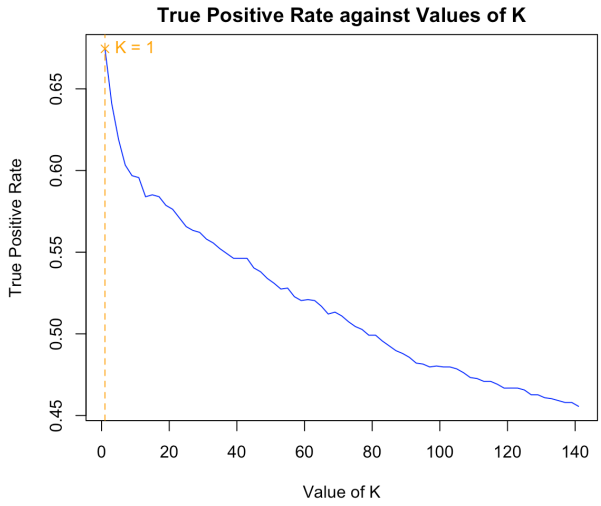

Fig. 5: Graph of TPR against values of K

2.2: Decision Tree Classification

Based on a set of input features, the decision tree algorithm uses a tree structure which depicts a sequence of decisions leading to a class label. A range of values for complexity parameter (cp) (10^-5 to 10^-3), minsplit (1 to 50) & maxdepth (2 to 10) were used to tune the model. The mean TPR per fold

is recorded along with the corresponding combination of parameters used. In conclusion, the highest TPR obtained from the decision tree classifier is 0.6833159 which occurs when cp = 10^-5, minsplit = 14 & maxdepth = 10. Fig. 6 visualises the various TPR values for different values of the respective parameters. The plot for the best decision tree constructed in terms of TPR is as shown in Fig. 7.

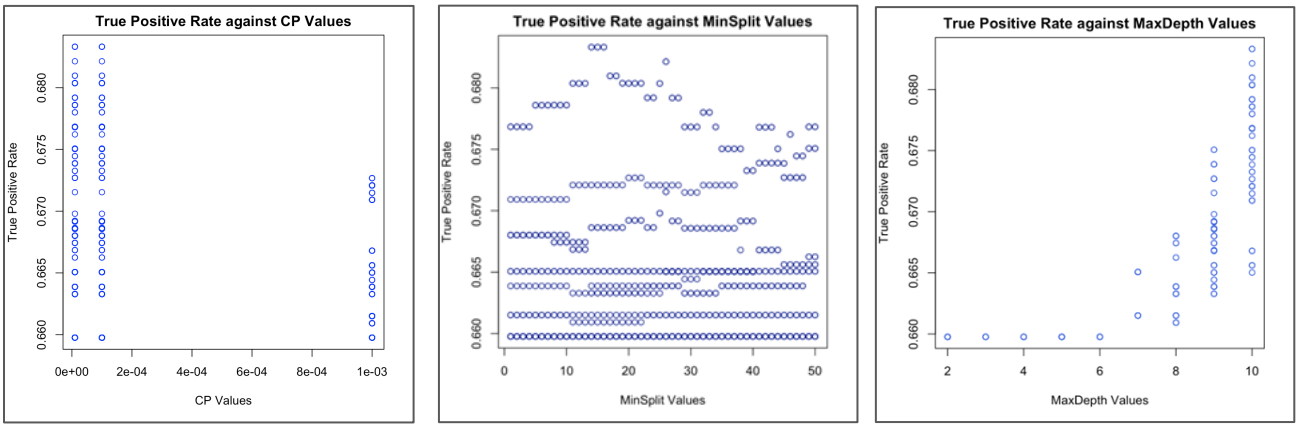

Fig. 6: TPR varying with each parameter

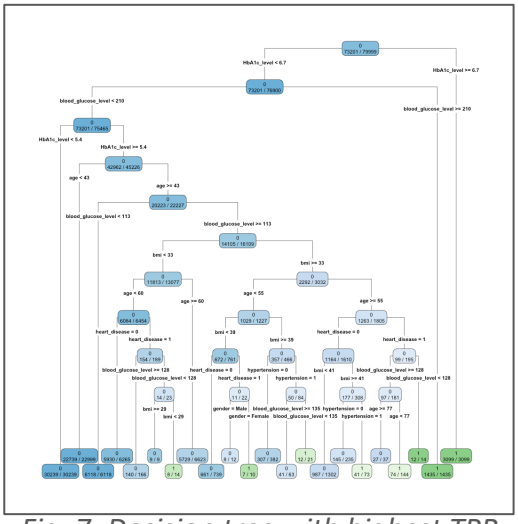

Fig. 7: Decision tree with highest TPR

2.3: Logistic Model Classification

The logistic regression model uses a logistic function to predict the probability that an input belongs to a particular class, & classifying the input based on a pre-determined threshold. The model uses all input features to predict the diabetes status of a respondent. All input features are statistically significant (p-value < 0.01) except the indicator variable *"gender = 'Other'"* which was previously addressed in **PART I**. The equation of the logistic regression model is as shown in Fig. 8.

$$\begin{aligned}\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = {} & -27.25 + 0.2502I(gender = \text{Male}) - 9.088I(gender = \text{Other}) + 0.04639(age) \\ & + 0.7962I(hypertension = 1) + 0.7967I(heart\_disease = 1) \\ & - 0.2028I(smoking\_history = \text{former}) \\ & - 0.2085I(smoking\_history = never) + 0.092(bmi) \\ & + 2.336(HbA1c\_level) + 0.0332(blood\_glucose\_level)\end{aligned}$$

Fig. 8: Logistic regression model equation, where $\hat{p}$ denotes the probability that a respondent has diabetes

2.4: Determining the best model

To observe the performance of each model, 4 folds from the data set is used as the train set while the last fold is used as the test set, inherently portioning the data set into subsets of 80% & 20% respectively. The models were constructed with the best parameters, determined previously, & the train set. Thereafter, each classifier was used to predict the diabetes status based on the input features from the test set. These predictions are compared against the *diabetes* column in the test set. The TPR, precision & AUC of each model can be observed in Table 2 below. The ROC curve for each model can be seen from Fig. 9.

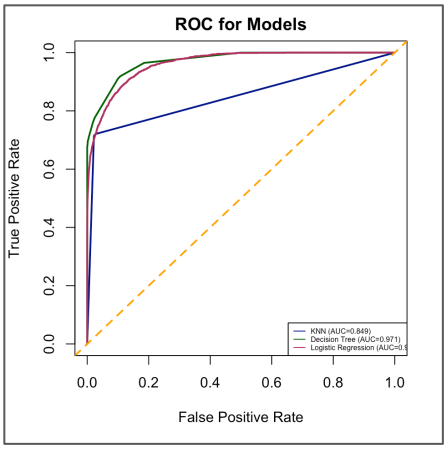

Fig. 9: ROC curve for each model

Model	Parameters	TPR	Precision	AUC
KNN	$K = 1$	0.7168038	0.7393939	0.8486042
Decision Tree	cp = 10^-5 minsplit = 14 maxdepth = 10	0.6980024	0.9690049	0.9710294
Logistic Regression	-	0.6263220	0.8737705	0.9623711

Table 2: Performance metrics of each model

PART III: CONCLUSION

3.1: Comparing performance metrics

Logistic regression has the lowest TPR & precision. This could be due to the innate assumption within the logistic regression model that there exists a linear relationship between the variables, which would be inappropriate in the context of diabetes prediction. The decision tree classifier has the highest precision & highest AUC. This is typical of tree models due to their hierarchical structure. The KNN classifier has the lowest AUC which could be because it is an instance-based learning algorithm. However, it produces the highest TPR due to the algorithm's method of predicting class labels based on the class of the K nearest neighbours.

3.2: Best classifier

For having the highest precision & AUC, the best classifier in this analysis would be the decision tree classifier. This aligns with current practices in the medical field, where decision tree classifiers can be used for applications such as diagnosis of medical conditions. However, the pitfalls of using a decision tree classifier should not be overlooked. The algorithm would create biased predictions if given a biased training set. Furthermore, interpreting complex decision trees with multiple branches may be challenging for complex models. Nonetheless, from the report's evaluation & analysis, it can be concluded that the decision tree classifier is the best classifier, serving as an effective & valuable tool in the medical field.