DSA1101 Statistical Report

Introduction

This statistical report was conducted using a study on diabetes and its associated risk factors. The resulting dataset is entitled “Diabetes prediction dataset”, authored by Mohammed Mustafa.
The purpose of this report is to document the process of choosing a classification method for predicting diabetes status, and propose the most effective classifier based on the goodness of fit of the classiers fitted. There are a total of 9 variables (as detailed in Table 1) and 100,000 observations.

Variable Description	Variable Name in Dataset	Possible Values of Variable	Type of variable
Gender of the individual	gender	Female, Male, Other (including non-binary, gender fluid, etc)	Factor, 3 levels
Age of the individual	age	Range from 0.08[1] to 80	Double; Continuous
Whether the individual has hypertension	hypertension	0 = No; 1 = Yes	Factor, 2 levels
Whether the individual has heart disease	heart_disease	0= No; 1 = Yes	Factor, 2 levels
Whether the individual has smoked in the past, or is currently smoking	smoking_history	current = currently smokes; ever = smokes sometimes but not often former = smoked before but quitted completely; never = never smoked and no intention to smoke; not current = never smoked but unsure about the future; No Info = No data	Factor; 6 levels (including No Info)
Body Mass index of the individual	bmi	Range from 10.01 to 95.69	Double; Continuous
Individual’s average blood sugar level over the past 2-3 months	HbA1c_level	Range from 3.5 to 9	Double; Continuous
Amount of glucose in the individual’s bloodstream	blood_glucose_level	Range from 80 to 300	Integer; Continuous
Whether the individual has diabetes	diabetes	0= No; 1 = Yes	Factor, 2 levels (response variable)

Table 1: Overview of Variables

[1] Note: for individuals below the age of 2,
the age value includes months and years. Hence 0.08 is equivalent to an individual who is 1 month old.

Methods

Among females, 7.62% of them have diabetes. Among males, 9.75% of them have diabetes. Among transgender individuals, 0.00% of them have diabetes. Hence, there is no significant difference between males and females, but females and males are more likely to have diabetes compared to transgender individuals.

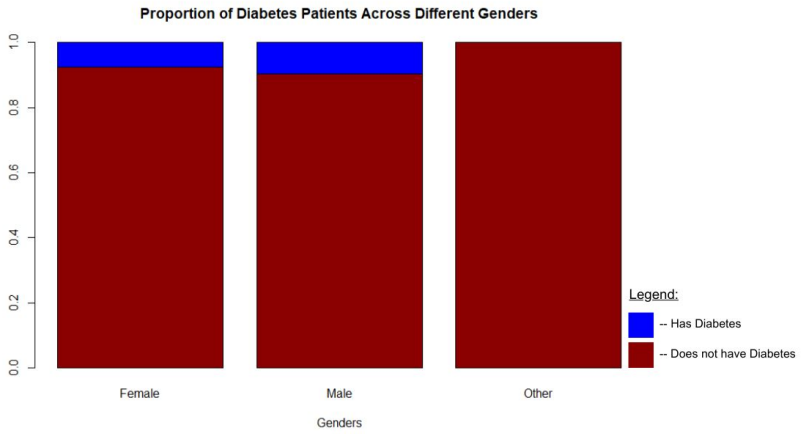


Figure 1: Proportion Bar Chart of Gender and Diabetes

Among those with hypertension, 27.9% of them have diabetes. Among those without hypertension, 6.93% of them have diabetes. The odds of having diabetes while having hypertension is 5.20 times the odds of having diabetes without having hypertension. Hence those with hypertension are significantly more likely to have diabetes.

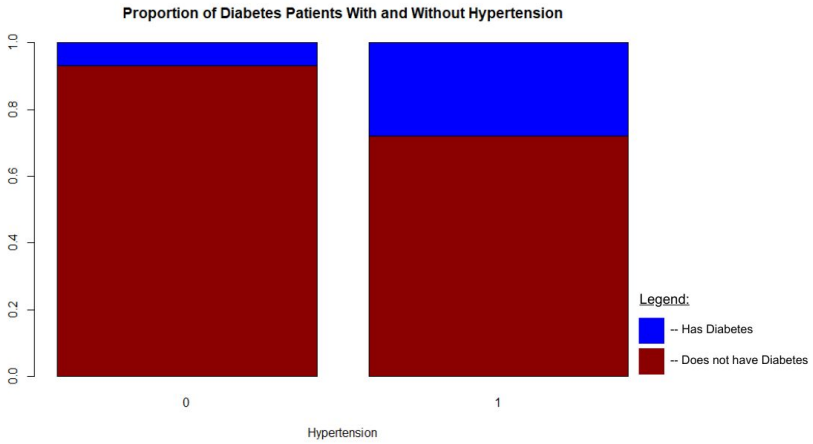


Figure 2: Proportion Bar Chart of Hypertension and Diabetes

Among those with heart disease, 32.1% of them have diabetes. Among those without heart disease, 7.53% of them have diabetes. The odds of having diabetes while having heart disease is 5.82 times the odds of having diabetes without having heart disease. Hence those with heart disease are significantly more likely to have diabetes.

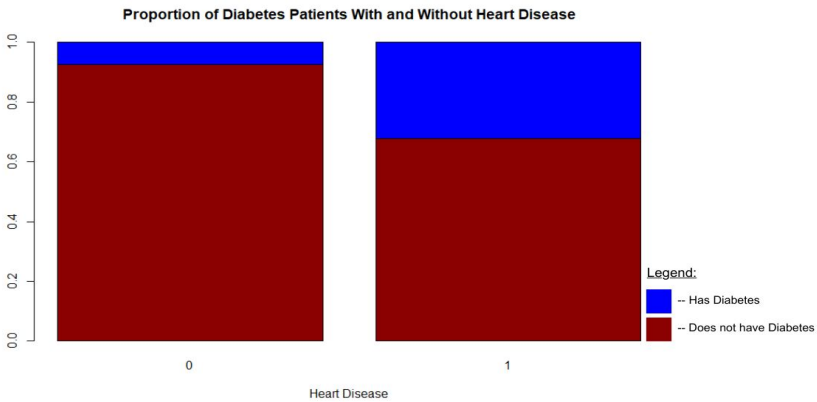


Figure 3: Proportion Bar Chart of Heart Disease and Diabetes

Among current smokers, 11.4% of them have diabetes. Among those who sometimes smokes (in the 'ever' category), 13.4% have diabetes. Among past smokers who currently do not smoke, 20.5% have diabetes. Among those who have never smoked and have no intention to, 10.5% have diabetes.

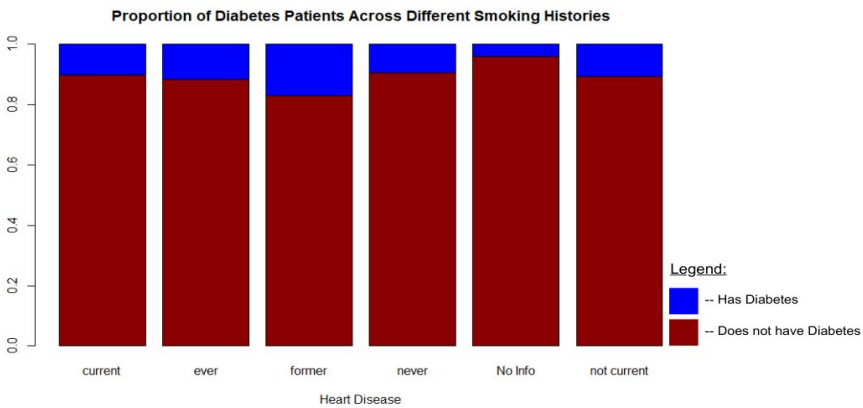


Figure 4: Proportion Bar Chart of Smoking History and Diabetes

Among those who have never smoked but might smoke in the future, 4.23% have diabetes. Among those with no data, 11.9% have diabetes. Hence past smokers who currently do not smoke are more likely to have diabetes, and those who have never smoked but might smoke in the future are less likely to have diabetes.

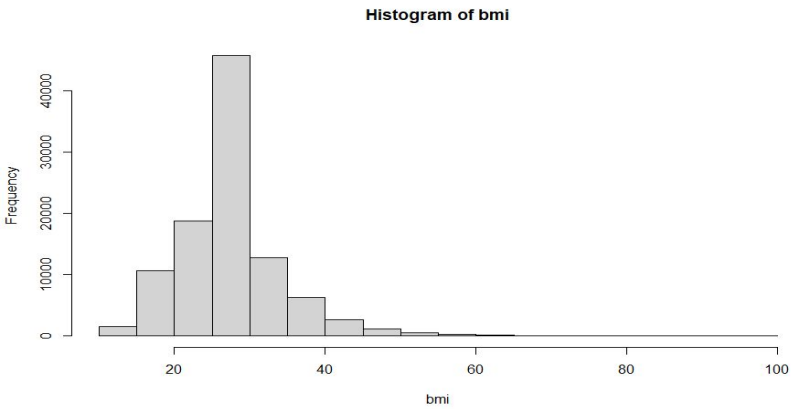


Figure 5: Histogram of BMI

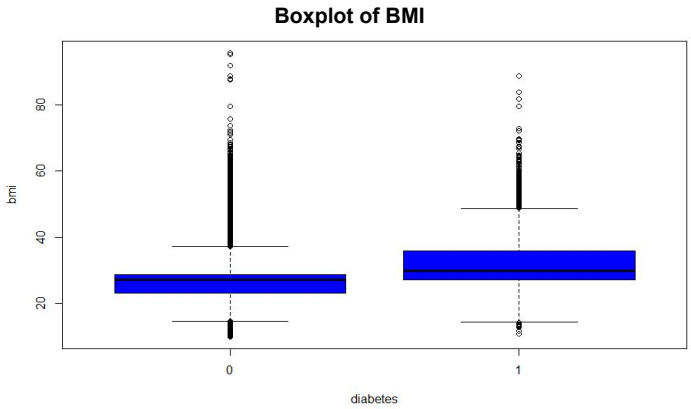


Figure 6: Boxplot of BMI

Across varying BMIs, those with diabetes have a slightly higher median BMI of 30.0 than those without diabetes, 27.3. In the box plot, both those with and without diabetes have many outliers (1694 outliers with diabetes and 6759 outliers without). Both groups are right-skewed, which follows the trend presented by the histogram. The boxes overlap, but despite having the same Q1, Q4 of diabetes present is significantly higher than the Q4 of no diabetes present.

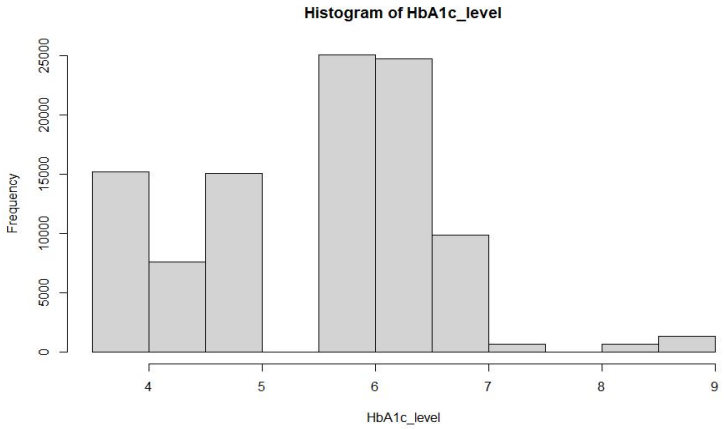


Figure 7: Histogram of HbA1c_level

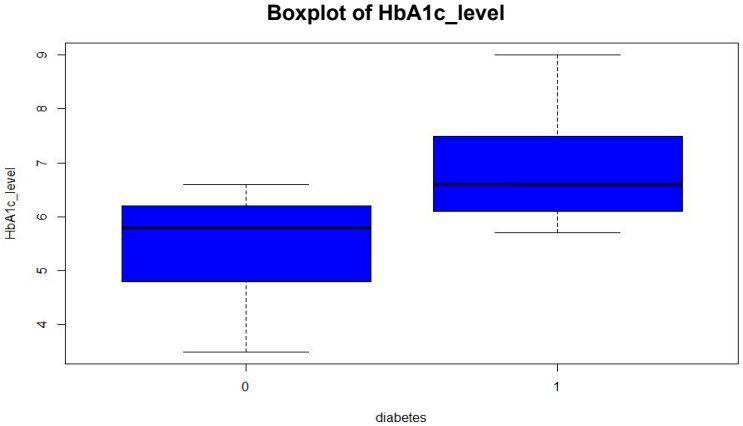


Figure 8: Boxplot of HbA1c_level

For HbA1c_level, there is no identifiable trend in the histogram, except that the median is around 5.5 to 6, and there are no values between 5 and 5.5, and between 7.5 and 8. Based on the boxplot, the median for those with diabetes is 6.6, which is higher than the median for those without diabetes, which is 5.8, indicating that HbA1c_level likely affects diabetes status. The boxes overlap slightly, but when looking at the box and whiskers together, there is a substantial overlap.

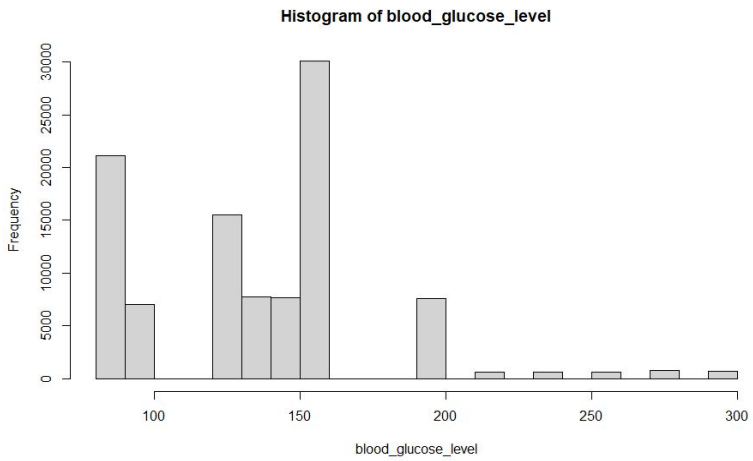


Figure 9: Histogram of Blood Glucose Level

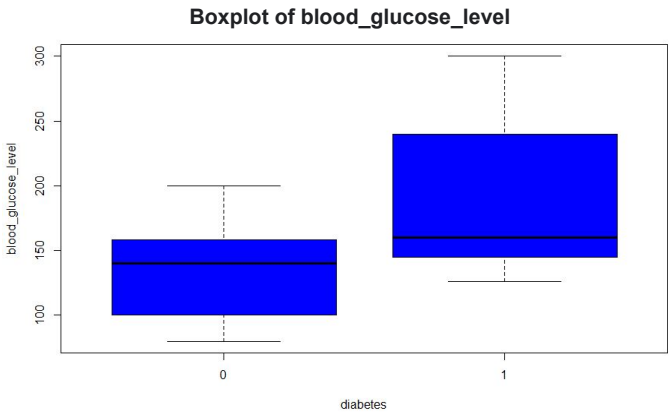


Figure 10: Boxplot of Blood Glucose Level

For blood_glucose_level, there is no identifiable trend in the histogram, except that the median is around 150 to 160, with a large frequency of 3000. Based on the boxplot, the median for those with diabetes is 160, which is higher than the median for those without diabetes, which is 140, indicating that blood glucose level likely affects diabetes status. The boxes overlap slightly, but when looking at the box and whiskers together, there is a substantial overlap.

I will use the models / classifiers: Decision Tree, Naive Bayes, and K-Nearest Neighbours. I standardised the variables age, bmi, HbA1c_level and blood_glucose_level, since the other variables are factor variables. I split the data into train and test sets, with a ratio of 8:2. The seed is set to 1101.

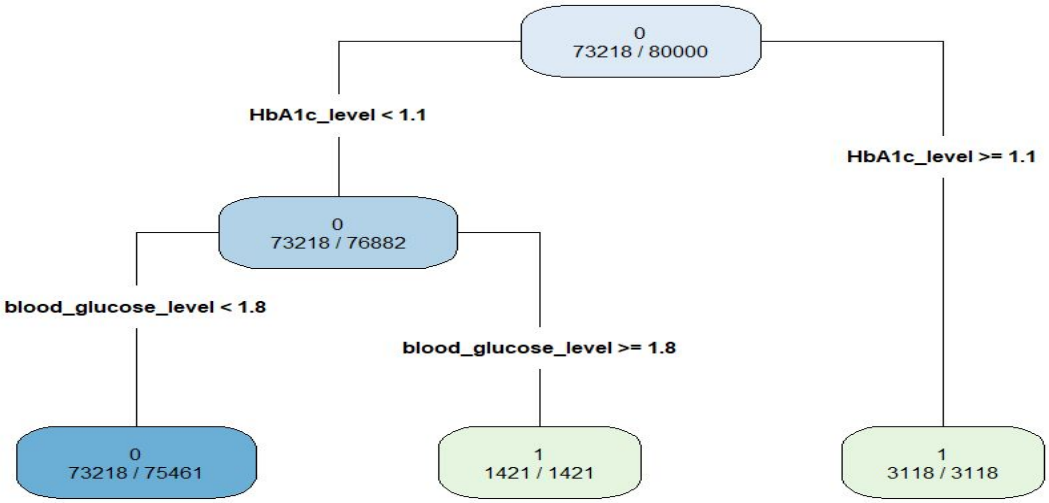
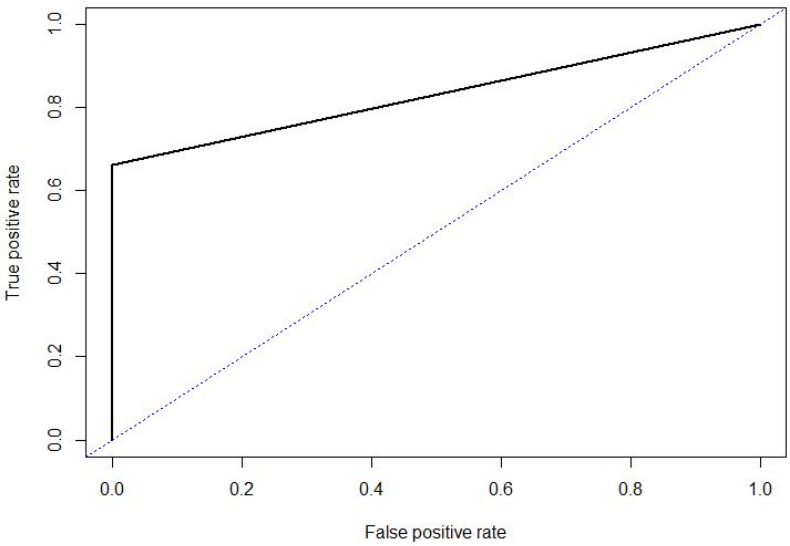


Figure 11: Plot of Decision Tree

Figure 12: ROC Curve for Decision Tree



Contingency Table / Confusion Matrix for Decision Tree		Prediction	
		0	1
test.Y (actual data)	0	18266	0
	1	584	1150

Table 2: Contingency table for Decision Tree

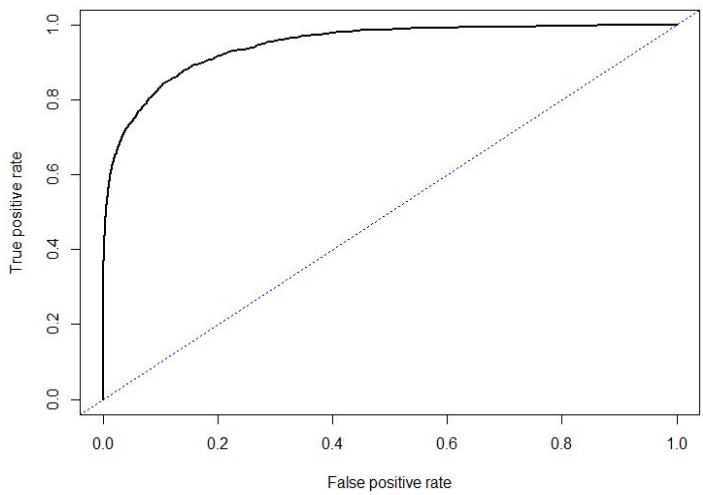
For **Decision Tree**, I used cp = 0.001, but it seemed no different than the default setting. Using cp = 0.001, and split by information, I plotted the above decision tree (Note: I standardised the variables beforehand for ease of use in the KNN classifier used later. I understand this was not necessary for decision trees.) I got an accuracy of 0.9708, false positive rate of 0 and false negative rate of 0.337. The attained AUC value is 0.832.

Contingency Table / Confusion Matrix for Naive Bayes		Prediction	
		0	1
test.Y (actual data)	0	17916	350
	1	625	1109

Table 3: Contingency table for Naive Bayes

For **Naive Bayes**, I got an accuracy of 0.951, false positive rate of 0.0192 and false negative rate of 0.360. The attained AUC value is 0.947

Figure 13: ROC Curve for Naive Bayes



For **K-Nearest Neighbours**, to determine k, I tried the following values: 20, 30, 40, 50, 100, 200, 300, 350, since there are a very large number of observations. **K = 30 gave the best results**, with an accuracy of 0.961, false positive rate of 0.000218 and false negative rate of 0.457. It seemed like as k increases, accuracy decreases, false positive rate decreases until stagnating at 0.00 from k=40 onwards, and false negative rate increases. Since diabetes should be treated as soon as it is detected, lest it remain untreated and become deadly, I tried keeping it low, below 50%. Since diabetes is a rather significant medical diagnosis and we do not want to cause the patients or their loved ones any unnecessary panic, I tried keeping the false positive rate fairly low. All the accuracies obtained are fairly high, all above 0.95, so I tried to pick one of the higher ones, while keeping in mind the false positive rate and false negative rate.

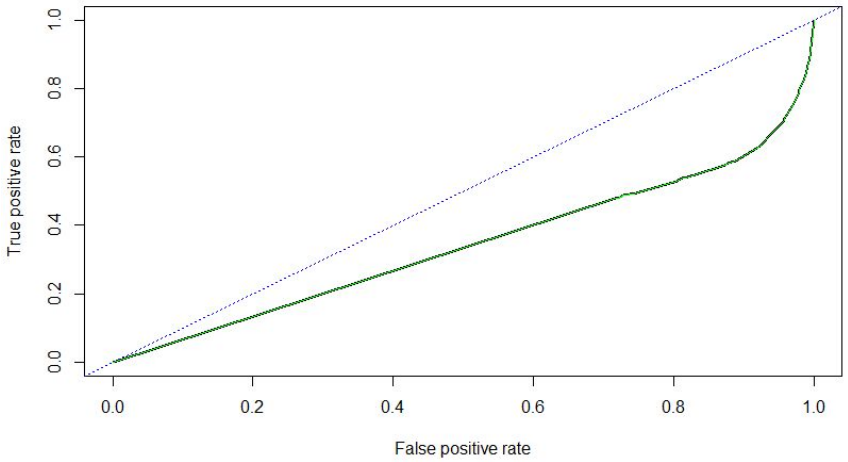
K-value	20	30	40	50	100	200	300	350
Accuracy	0.963	0.961	0.960	0.960	0.958	0.956	0.955	0.955
False Positive Rate	0.000546	0.000218	0.00	0.00	0.00	0.00	0.00	0.00
False Negative Rate	0.438	0.457	0.471	0.474	0.501	0.522	0.530	0.536

Table 4: K-values and Their Corresponding Metrics

Contingency Table / Confusion Matrix for K-Nearest Neighbours		Prediction	
		0	1
test.Y (actual data)	0	18310	3
	1	775	912

Table 5: Contingency table for K-Nearest Neighbours

Figure 14: ROC Curve for K-Nearest Neighbours



The attained AUC value is 0.341. This is extremely low but this may be because majority of diabetes values is 0, such that the classifier ‘over guesses’ the number of diabetes values that are 0. Hence, KNN has a relatively high accuracy for such a wide range of K values, but the false negative rate is perpetually high (0.438 to 0.546), which ultimately causes the AUC to be very low because AUC averages across all thresholds (such that KNN is seemingly worse than random guess).

Note that I converted all categorical variables to numeric variables, to work within the confines of the knn() function. This includes variables with more than 2 levels, which are gender and smoking_history. I used the library fastDummies to create new ‘dummy’ columns for each level, with values consisting of 1 and 0.

Evaluation of Models

Even though Naive Bayes has the best AUC value, Decision Tree has the highest accuracy of 0.9708, best false positive rate of 0, and best false negative rate of 0.337. Hence, **Decision Tree is the best model**. KNN might be the worst model because this dataset has imbalance response data, hence naturally there will be more '0' neighbour values.

The pro of Decision Tree is that It only takes into account the most salient variables, which increases the simplicity of the model, so it is easier and faster to implement. The con is that, the possible downside of only taking into account the most salient variables is that theoretically valid risk factors to contracting diabetes might be overlooked. For instance the proportion bar charts in page 2 shows that individuals with hypertension and heart disease do indeed have a higher chance of getting diabetes. However, this is not taken into account by the decision tree at all

The pros of Naive Bayes is, firstly, Naive Bayes is a very fast algorithm. This makes rapid modeling and testing less time consuming. Secondly, all the attributes are taken into account, so the model could be more holistic when making predictions. Thirdly, there can be more than 2 classes as the response variable. Hence, if the dataset were to be expanded to include individuals with pre-diabetes, or are at high risk of getting diabetes, Naive Bayes can still handle that. The con of Naive Bayes is that that all features are completely independent. This might not be true in practice, reducing the accuracy of the model. For example, and individuals' age and BMI are definitely correlated, since children tend to have lower BMIs than adults.

The pro of K-Nearest Neighbours is the K value can be tweaked to fit the context of the dataset and the purpose of the model. Hence, depending on which metric is the most important, the K value can be altered accordingly. This allows for a lot of flexibility when adapting the K-nearest neighbours classifier to its use case. The cons are that, firstly, performing the classifier is time consuming, especially when testing the different K values. knn() took the longest time to run compared to rpart() for decision tree and naiveBayes() for Naive Bayes. Secondly, when changing the categorical variables with more than 2 levels into dummy variables, the resulting model might become inaccurate