Data Powered Stroke Prevention

Motolani Ojo-Bello

2022-11-03

Contents

INTRODUCTION

INTRODUCTION

A medical patient is said to have had a stroke when there is damage to the brain due to an interruption in its blood supply. According to a report from the World Health Organization (WHO) on the 28th of October 2021, stroke is the second leading cause of death and the third leading cause of disability globally. Also, based on the journal "Long-Term Survival and Causes of Death After Stroke" by Henrik Brønnum-Hansen et al in 2001, within a year of a stroke about 40% of patients die, the journal goes on to say that "The estimated cumulative risk for death was 60%, 76%, and 86% at 5, 10, and 15 years after the in stroke, respectively." Furthermore, beyond the risk of death, strokes can also lead to several other health conditions like post-stroke seizures, urinary incontinence, bowel incontinence, cognitive impairment and hemiplegic shoulder pains amongst other things. Taking all this into consideration, stroke prevention where possible is much preferred to "the cure". The objective of this project is to build a model that would be able to predict whether or not a person is likely to have a stroke or not so that they can take steps to adjust their lifestyles and prevent a stroke from actually happening.

The data used to build test and validate the model was taken from kaggle: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset I downloaded it into a temporary file within R, then I used "read_csv" and "unzip" to read the compressed csv file into a data frame.

The data frame contained data for patients who have and have not had strokes. This data is represented with the following columns and their corresponding values for each entry:

1. id: A unique ID for the entry
2. gender: Whether the patient was Male, Female or Other
3. age: The age of the patient
4. hypertension: Whether the patient has had hypertension in the past with a value of 1 or they have not with a value of 0
5. heart_disease: Whether the patient has a heart disease with a value of 1 or not with a value of 0
6. ever_married: "Yes" or "No"
7. work_type: Their employment status, whether they are children, work a government job, have never worked, work a private job or are self-employed.
8. residence_type: Whether they live in a rural or urban area
9. avg_glucose_level: Average glucose level in the patients blood
10. bmi: Their body mass index
11. smoking_status: Whether they have formerly smoked, never smoked, smoke or if their smoking status is unknown
12. stroke: the value we are trying to predict, Whether the patient had a stroke or not Numbers one to eleven are our predictors, and twelve is the variable we want to be able to predict.

KEY STEPS

The key steps taken to complete this project included, analyzing each predictor and their potential usefulness accompanied with visualization, ensuring each of these predictors was in the right format, training and testing the selected models on the data and observing the resulting performance metrics from these models.

METHODOLOGY

MODIFICATION and ANALYSIS OF EACH PREDICTOR

After importing the data, I went through each predictor to make sure they were in the right format and subsequently carried out analysis on them one by one.Before that though, for the purpose of the analysis

I also created another data frame "had_stroke" containing data for patients that actually had a stroke to support the analysis of the data.

ID

The ID's provide unique identification for each patient, there are no duplicate ID's(Each row in the data corresponds to a specific ID). Since there are no duplicates and we want a model that predicts the stroke rate for people in general, the IDs will not provide much significance to the construction of our model. So I removed this predictor

GENDER

The gender column was a character vector containing the gender of the patient. Firstly, looking at the distribution of the genders as shown in the table below, there is only one entry with gender "Other" because it is only one entry and because of the lack of context to exactly what "Other" entails I removed that single entry from the data set. After that, I converted the gender vector from a character vector to a factor vector.

gender	proportion	no_of_entries
Male	0.4138943	2115
Female	0.5859100	2994
Other	0.0001957	1

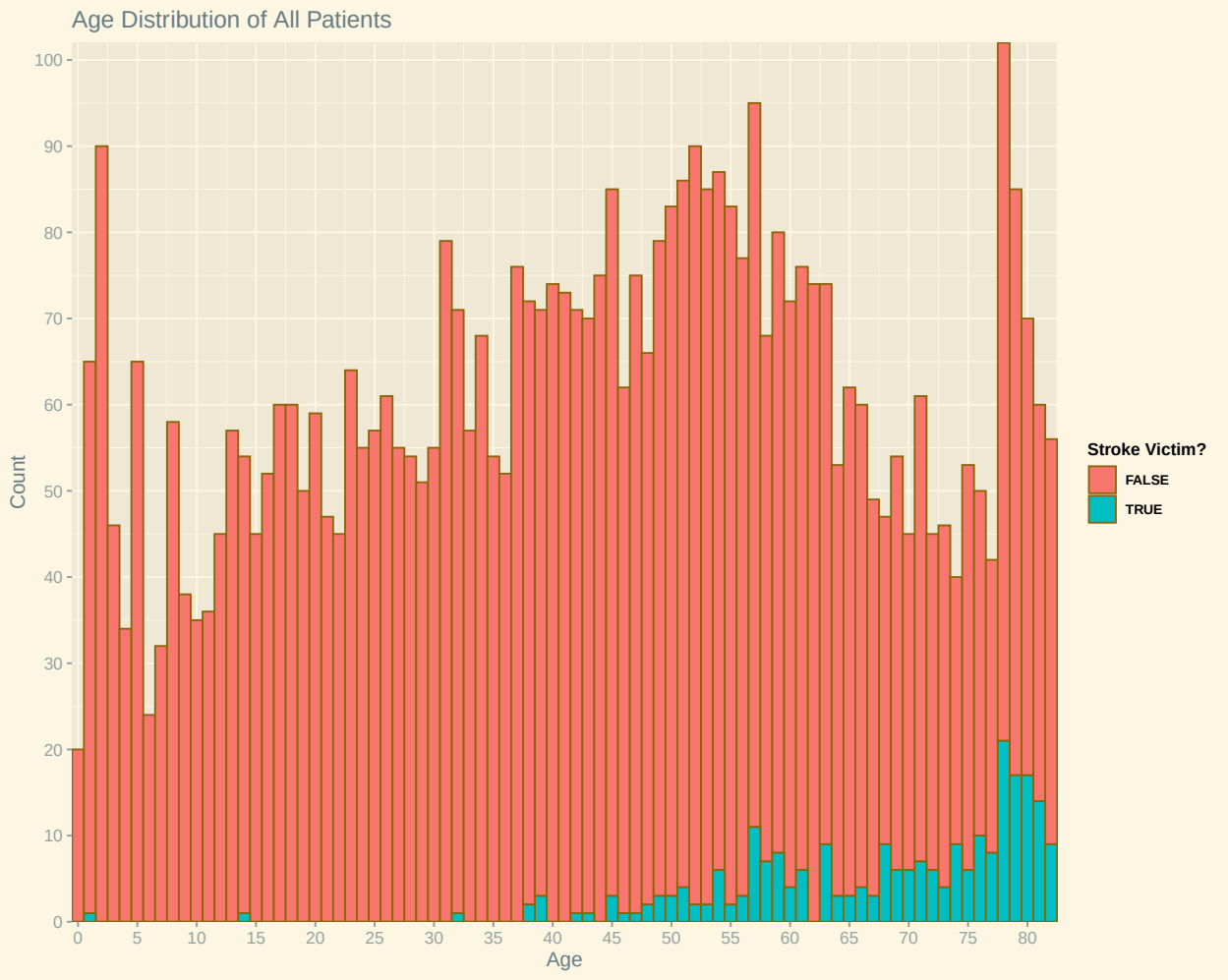
I also looked at the distribution of genders for the patients that had a stroke and as shown in the table below, there is a good proportion of both Men and women who had strokes in our data with slightly more female victims. This checks out with information on strokes across genders that women are more likely to have strokes because of their longer life expectancy and higher incidence at older ages.

gender	proportion
Male	0.4337349
Female	0.5662651

AGE

Looking at all the possible age values in our data, the ages range from `0.08` to `82` years old. Also, only `2.2509297`% of age values are reported as decimals so I rounded them up and converted the ages from numeric to integers. The minimum age of a patient with stroke in our data set was `1.32` just over one years old, this is a rare condition called pediatric stroke which according to Hopkins medicine happens to one in every 4000 children, by contrast the average age for our stroke victims was `67.7281928` which is more in the range that would be expected.

I also looked at the age distribution for all patients grouped by whether or not they had a stroke and from the resulting histogram below, we can see that the data has purposely been compiled so that there is a wide range of varying patient ages in our data set but despite this, most of the patients who had strokes are well above their 60s which is in line with medical theory, 90% of strokes happen to people over the age of 65 years old.



HYPERTENSION AND HEART DISEASE

Hypertension is a condition where the pressure flow of blood in a patient's arteries is too high (Also called high blood pressure), over time hypertension increases the risk of having both heart diseases and stroke. Heart diseases refer to any ailments that affect the heart of the patient and as earlier discussed, a stroke is a condition that affects the brain, according to medicinal science, either condition (heart disease or stroke) can lead to the other. This could lead to confounding within our model, however with the data collected, we know that the patients who had stroke either had heart diseases or they did not before the stroke happened. The other potential confounding issue is with hypertension and heart diseases, since hypertension causes heart diseases over time, however looking at the data there isn't a strong correlation between both predictors. Just a correlation of 0.1082925.

HYPERTENSION: The hypertension predictor has 1 or 0 values for whether the patient is hypertensive or not. There are 498 hypertensive patients in our data set, out of those 66 had strokes, so 13.253012% of hypertensive patients had strokes. Also, 26.5060241% of the stroke victims were hypertensive compared to 9.7475044% of all the patients in the data showing that there is some correlation between hypertension and stroke

HEART DISEASE: This predictor also has 1 or 0 values but for whether the patient has a heart disease or not. There are 276 patients with heart diseases in our data set 47 had strokes, so 17.0289855% of patients

with heart diseases had strokes. In addition, 18.875502% of the stroke victims in the data had heart diseases compared to 5.4022314% of all the patients showing that there is a relationship between heart diseases and likelihood of a stroke.

EVER MARRIED

This predictor gives a Yes or No reply to whether or not the patient has ever been married, in the total data set 65.6292817% of patients are married of course it is worth noting that 0.7036602% of all our patients are also above or exactly 30 years old so the high proportion of married patient is understandable. Looking at the stroke patients, 88.3534137percent of them are married, this may seem like a big jump, but it is worth noting that the older people are, the more likely they are to be married. So I looked at the correlation between age and marriage status (represented by 1 for married and 0 otherwise) and I obtained a correlation of 0.6790819 which is fairly high. Thus I didn't use the marriage status for building my model.

WORK TYPE

This predictor contains the patients employment status (children, Government job, never worked, Private, Self employed) . We can see the distribution of work types in the table below.

Work Type	Number of Entries
children	687
Govt_job	657
Never_worked	22
Private	2924
Self-employed	819

Firstly I looked at the 22 entries that were classed as never worked as shown in the table below,, 21 of them were in their teens and there was one 23 year old because 22 is a very small sample size for young adults, I combined the children with the never worked into one class never worked

Age	Work Type
14	Never_worked
23	Never_worked
19	Never_worked
13	Never_worked
17	Never_worked
17	Never_worked
13	Never_worked
16	Never_worked
14	Never_worked
17	Never_worked
15	Never_worked
16	Never_worked
18	Never_worked
14	Never_worked
17	Never_worked
15	Never_worked
16	Never_worked
18	Never_worked
13	Never_worked

Age	Work Type
17	Never_worked
18	Never_worked
16	Never_worked

Also, I looked at all the three employment possibilities (Government work, Privat or Self-employed) and the proportion of each of them that had stroke as shown in the table below and there was no significant difference in the likelihood of having a stroke across them so I combined them into one class "working"

	% with stroke
Government Work	5.022831
Private	5.095759
Self-employed	7.936508

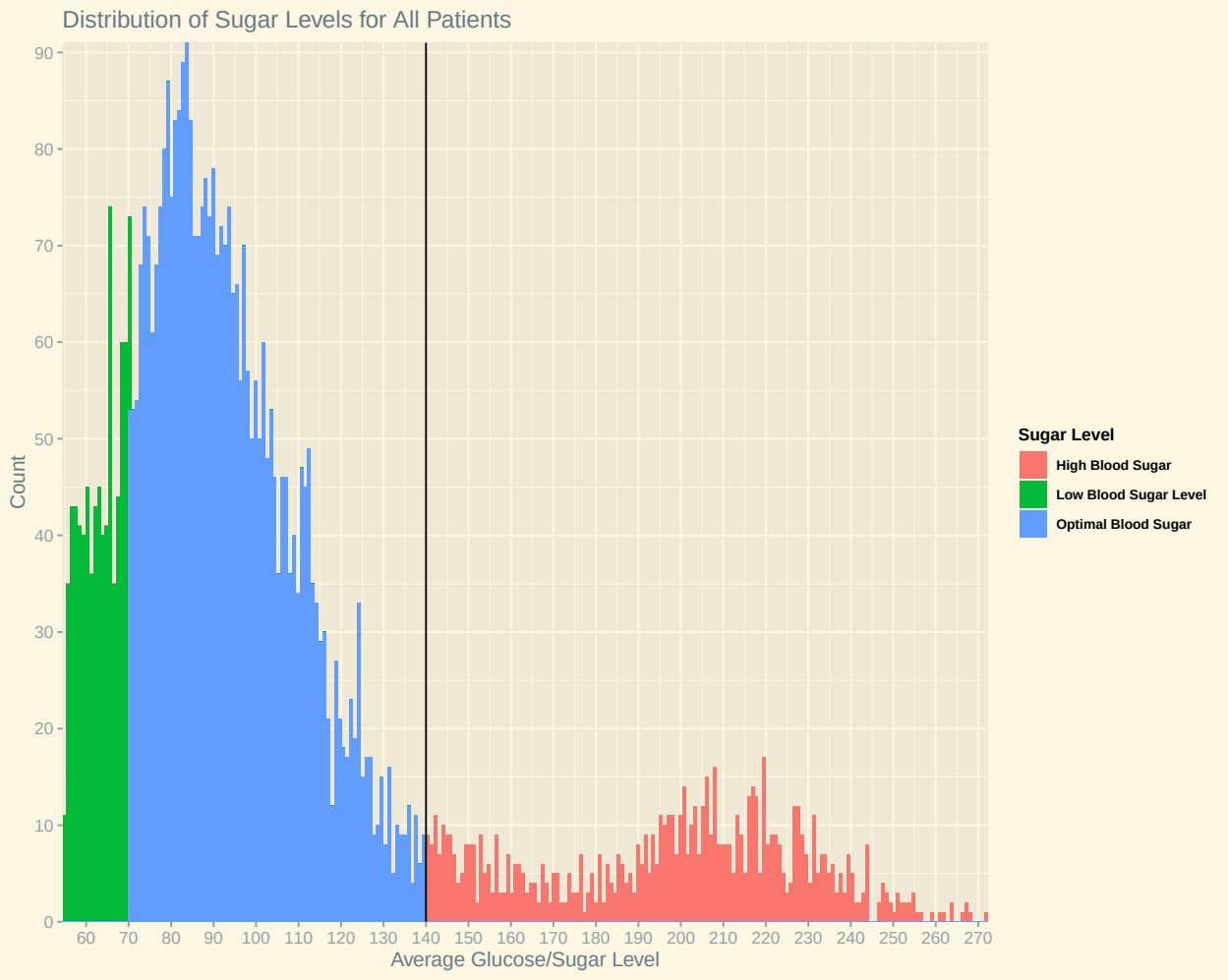
So, our work type now has 0 for unemployed and 1 for employed. However, as you would expect, there is a fairly high correlation(0.6408574) between a patient's age and their employment status, so I removed this predictor.

RESIDENCE TYPE

This variable has two possible values for the type of community the patient lives in, rural or urban. According to The CANHEART Stroke Study stroke incidence is supposedly higher in rural areas, however in the data set, the distribution of stroke victims is fairly even in both rural and urban areas, with 45.7831325% in rural and 54.2168675% in urban areas. Also, our full patient data is fairly evenly distributed between rural and urban areas with 49.187708% in rural areas and 50.812292% in urban areas. The only change I made here was to make the residence type a factor rather than a character vector.

GLUCOSE LEVEL

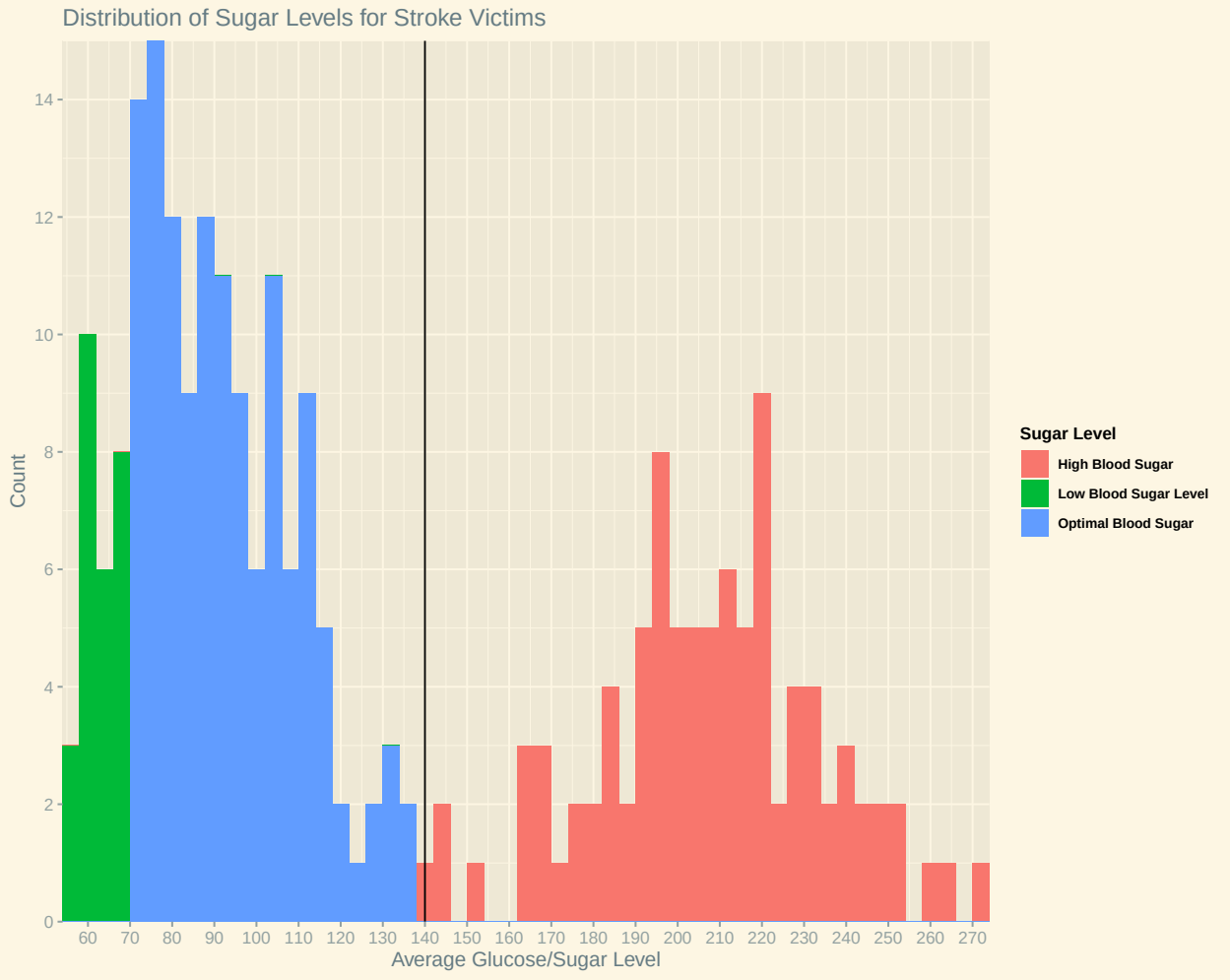
The glucose level, also called blood sugar level is a measure of the concentration of sugar in a patient's blood stream.High blood sugar, above or equal to 140 milligrams per deciliter mg/dL) usually happens when the patient's body can no longer make or use insulin properly causing the build up of sugar in the bloodstream. Over time, this can damage the body's blood vessels increasing the chance of stroke. Research states that diabetic patients (patients whose bodies can no longer produce insulin) are twice as likely to have a stroke. Low blood sugar levels(below 70 mg/dL) are not known to increase chances of stroke. The plot below shows the distribution of glucose levels in our data set grouped by whether the patient has Low, Optimal or High sugar levels. As we can see in the plot, majority of the patients in our data set are below the black vertical line(Crossover point for high blood sugar 140 mg/dl)



The table below confirms this, only 16.0501077 percent of our patients have high blood sugar levels

Blood_Sugar_Level	percentage
High Blood Sugar	16.05011
Low Blood Sugar Level	14.79742
Optimal Blood Sugar	69.15248

Looking at the same plot for the patients that suffered stroke, we can see that a much larger proportion of the stroke victims have high blood sugar compared with the full data set.



The table below confirms this, 37.3493976% of stroke victims had high blood sugar levels compared to just 16.0501077% in our total data.

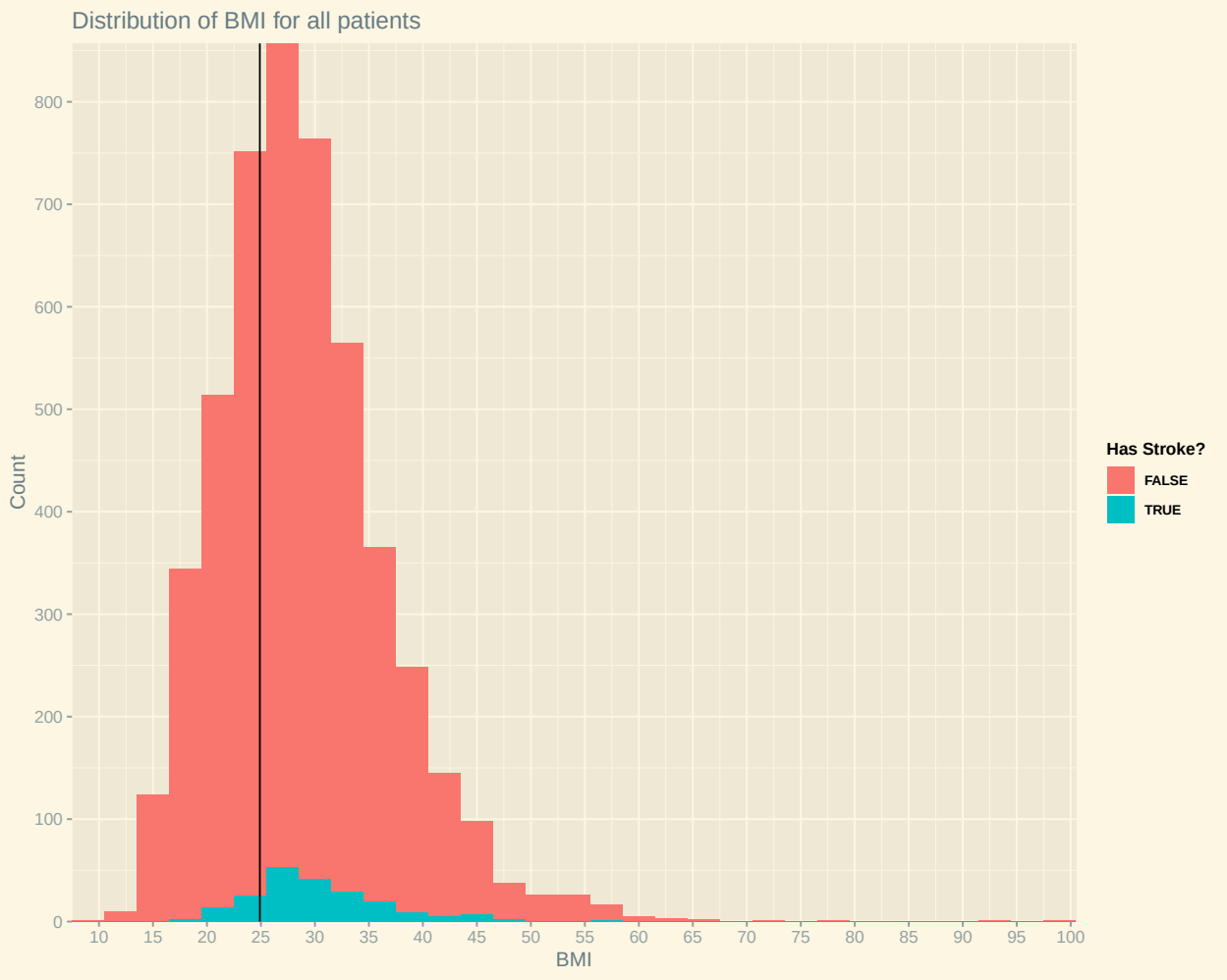
Blood_Sugar_Level	percentage
High Blood Sugar	37.34940
Low Blood Sugar Level	10.84337
Optimal Blood Sugar	51.80723

BMI

The BMI(Body Mass Index) is a metric used to measure a person's body fat based on their height and weight. The BMI is defined as the body mass divided by the square of the body height, and is expressed in units of kg/m^2. The healthy range for the BMI is between 18.5 to 24.9. Loosely speaking a patient with a BMI under 18.5 is underweight and a patient with a BMI over 24.9 is overweight. While being underweight slightly increases the likelihood of stroke compared to being in the healthy range, overweight patients have the highest risk factor.

The bmi data was initially a character vector with either bmi values or N/A, I converted this to numeric and looking at the percentage of NA's just 3.9342337% I decided to remove rows with unavailable bmi values. Looking at the distribution of the BMI values in our data, grouped by whether the patients had strokes or not, the increased stroke risk for underweight patients doesn't really show up, likely because

of the few number of underweight patients in our data 6.8663407%. However, we can clearly see that majority of the stroke victims have BMI's above 24.9 (shown by the black vertical line on the graph)



SMOKING STATUS

Smoking is known to increase the likelihood of a patient having a stroke. This predictor tells us whether the patient smokes, has never smoked or used to smoke. If this information is not known, it is also stated here. In our data 0.3204971% of all patients were either smokers or had smoked before but amongst stroke victims that rises to 0.4593301%. Also, I converted this predictor from a character to a factor.

STROKE

This is the value we are trying to predict, 1 for patients who had stroke and 0 for those that did not because of the nature of this problem (classification) I converted the stroke vector from numeric to categorical. It is worth noting that the data only contains 4.2583537% of stroke victims.

RESULTS

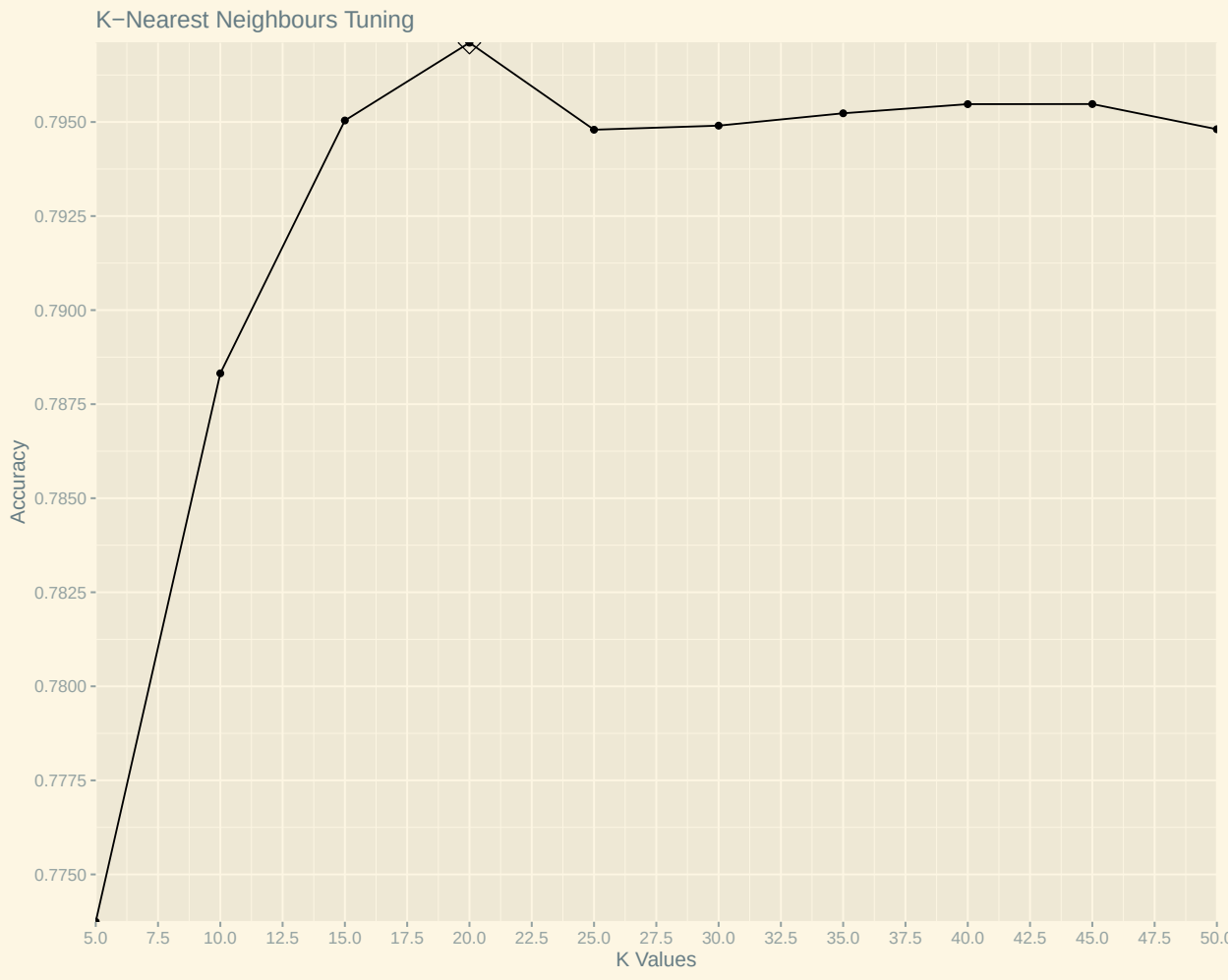
MODEL SELECTION

After preparing all the predictors, the next step was to select and build the models. This problem is a binary classification problem, we want to know whether a patient is likely to have a stroke or not. However, before selecting the models, there is one issue, prevalence. Only 4.2583537% of the patients in our data actually had a stroke, the prevalence of patients who didn't have stroke is 0.9574165, so a model run using this data will mostly just predict low risk of stroke for most patients to maximize accuracy. That would not be very useful since we want to maximize specificity (in this analysis, a patient not having stroke is the positive class), so in essence, we want to predict that a patient has risk of getting stroke where that is actually the case as much as possible. To work around this, I used the Synthetic Minority Oversampling Technique to generate a new data set with a better balance of stroke victims and those who did not have stroke. In this new data, 51.2195122% of the patients were stroke victims which would allow us to build a more useful model. Like earlier stated, this problem is a classification problem. Also, we are more concerned with the accuracy of the model than it's interpretability. So I modeled with K-Nearest Neighbours and Ranfom Forest algorithms, both classification algorithms that give high accuracy.

TEST AND TRAIN DATA

After getting the "smoted" data set, I partitioned it into test and train data before evaluating the models ## K-Nearest Neighbours K-Nearest Neighbours (KNN) is a machine learning model that can be used for classification or regression problems, KNN makes classification predictions in the test set based on the classes of the closest k(tuning parameter that can be set) entry's in the training set. How close entries are is determined by their predictor values.

For my implementation, I tuned KNN between 5 and 50 and got a best fit of 20 as seen in the plot below.



Also, I obtained accuracy of 0.8030338 and a specificity of 0.7062201. While the accuracy is decent albeit not great, the specificity is even worse meaning that we will be unable to properly detect well enough patients with risk of having a stroke. We can see the confusion matrix below

	Actually 0	Actually 1
Predicted 0	1476	230
Predicted 1	614	1965

Random Forest

The random forest is a machine learning model that constructs a set of decision trees using data from the train set and uses those decision trees to make final predictions on the test set. The tuning parameter here is mtry which is the number of random variables from the data that would be used to construct each decision tree because of the long run time associated with random forest, I did not carry out any tuning of mtry. The mtry value selected by R was 6

The random forest model gave better accuracy than knn 0.9003501 and an even better specificity of as high as 0.8779904. We can see the confusion matrix below.

	Actually 0	Actually 1
Predicted 0	1835	172
Predicted 1	255	2023

Based on performance, random forest would be the model to use for solving this problem

CONCLUSION

In conclusion, we were able to build a model using random forest that allowed us to predict a patients risk factor to stroke with an overall accuracy of 0.9003501 and with the chances of predicting when a patient is actually at risk of stroke of 0.8779904. While these values are good they could potentially be improved by tuning the random forest model over a wide range of values and by looking at other factors that impact the occurence of stroke in patients for example: genetics (history of stroke in the patients family), previous occurence of stroke with the patient and the patients race.

REFERENCES

World Health Organisation: https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day#:~:text=Globally%2C%20stroke%20is%20the%20second,tobacco%20use%20and%20alcohol%20abuse.

Long-Term Survival and Causes of Death After Stroke: https://www.ahajournals.org/doi/10.1161/hs0901.094253

The CANHEART Stroke Study: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.118.004973#:~:text=Stroke%20mortality%20is%20higher%20in,rather%20than%20stroke%20case%20fatality.