**Classification Predictive Modelling of Stroke using Decision Tree Classifier**   
Individual Report

**Helly Hiteshkumar Thakar**

Student ID Number: **19155625**

**AI and Machine Learning Project 2022-23: CMP6202**

**BSc (Hons) Computer and Data Science**

Birmingham City University

Faculty of Computing Engineering & the Built Environment

School of Computing & Digital Technology

Table of Contents

[1 Report Introduction 3](#_Toc117525292)

[1.1 Dataset identification 3](#_Toc117525293)

[1.2 Supervised learning task identification 4](#_Toc117525294)

[1.3 Team Identification 4](#_Toc117525295)

[2 Exploratory Data Analysis 5](#_Toc117525296)

[2.1 Question(s) identification 5](#_Toc117525297)

[2.2 Splitting the dataset 5](#_Toc117525298)

[2.3 Exploratory Data Analysis process and results 7](#_Toc117525299)

[2.4 EDA conclusions 16](#_Toc117525300)

[3 Experimental Design 18](#_Toc117525301)

[3.1 Identification of your chosen supervised learning algorithm(s) 18](#_Toc117525302)

[3.2 Identification of appropriate evaluation techniques 18](#_Toc117525303)

[3.3 Data cleaning and Pre-processing transformations 21](#_Toc117525304)

[3.4 Limitations and Options 23](#_Toc117525305)

[4 Predictive Modelling / Model Development 23](#_Toc117525306)

[4.1 The predictive modelling process 23](#_Toc117525307)

[4.2 Evaluation results on “seen” data 25](#_Toc117525308)

[5 Evaluation and further modelling improvements 26](#_Toc117525309)

[5.1 Initial evaluation comparison 26](#_Toc117525310)

[6 Conclusion 26](#_Toc117525311)

[6.1 Summary of results 26](#_Toc117525312)

[6.2 Suggested further improvements to the model development process 26](#_Toc117525313)

[6.3 Reflection on Research Team 26](#_Toc117525314)

[6.4 Reflection on Individual Learning 27](#_Toc117525315)

[7 References 27](#_Toc117525316)

[8 Bibliography 28](#_Toc117525317)

# Report Introduction

A non-communicable illness called stroke is to blame for about 11% of all fatalities. Machine learning can be used to forecast the onset of a stroke thanks to advancements in medical technology. The algorithms used in machine learning are beneficial in supplying accurate analysis and producing accurate predictions. The majority of the prior efforts on stroke concern heart stroke prediction [[1](https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.116.308398)]. The research on brain stroke is rather limited. This report uses machine learning to forecast the likelihood of a brain stroke. The classification algorithm utilised is crucial to the methods adopted and outcomes realised. This model has a drawback because it was trained on tabular data rather than actual brain imagery through which we can understand the relationships between specific areas of the brain and how they are functioning and locate the areas of the brain that are affected by neurological disorders, as we can see here stroke. The decision tree machine learning classification technique is demonstrated in the study. It is possible to evaluate a few more machine learning algorithms using the findings of this research.

To continue with this challenge, a dataset from Kaggle is picked that has a variety of physiological characteristics as its properties. These characteristics are analysed subsequently and utilised in the final forecast. Then supervised learning task is identified and further data split is done into train, test and validate and then Exploratory Data Analysis is done on the training set and further the report specifies about the chosen supervised learning algorithm and then it talks about the evaluation techniques used. Afterwards, data cleaning and pre-processing is done to get more accurate result and then predictive modelling process is explained with results on train data( seen data).

Further, it explains the evaluation results on the validate data (initially seen data)of the model and comparison with the other team member’s results. Following it explains the improvements attempted on model further and then it discusses about the final model evaluation results.

At end the conclusion summarises existing results from the model development process and provides further enhancement recommendations and finally reflection on research team and individual learning.

## Dataset identification

Stroke, which occurs when an artery becomes clogged or a blood vessel breaks, causes the blood flow to a part of the brain to stop, is a potentially deadly medical condition (haemorrhagic stroke). This harm may affect not just how your body functions, but also how you think, feel, and communicate. The blood's delivery of nutrients and oxygen is essential for the brain's proper operation, just like it is for all other organs. If the blood flow is restricted or stopped, brain cells begin to deteriorate. A brain damage, a disability, or even death might occur from this. A balanced, healthy lifestyle that abstains from harmful practises like drinking and smoking and manages body mass index can reduce the risk of stroke. The dataset that was utilised in this study was discovered on Kaggle and was initially acquired from different medical centres in Bangladesh [[2](https://ieeexplore.ieee.org/document/9297525)]. It was collected by researchers from Jahangir University of Bangladesh for research on stroke. The patient details are collected based on their various health conditions, which is in the occurrence of stroke disease. The dataset contains data on 5110 individuals, which are of following types:

1) Categorical Features: gender, ever married, work type, Residence type, smoking status

2) Binary Numerical Features: hypertension, heart disease, stroke

3) Continuous Numerical Features: age, avg glucose level, bmi

Each of their attributes are listed below with its description:

Table

Description automatically generated

Table 1: Description of attributes and its values

So here our target variable is stroke, and all other attributes are our feature variable.

## Supervised learning task identification

A prediction or inference created in response to a problem or question and the present state of the data is known as a machine learning task. The classification task, for instance, classifies data. supervised machine learning is used to decide which of two classes (categories) a given item of data belongs to.

Here we are trying to build a predictive model which is capable of automatically predicting whether a given patient should be diagnosed with stroke. The selected target ground truth value is "stroke" and stroke is a nominal variable. Classification techniques are used to predict nominal values (categories).

Based on problem mentioned above we get the collection of following factors:

A) The selected target ground truth value is "stroke"

B) "stroke" is a nominal variable

C) Classification techniques are used to predict nominal values (categories)

Because A, B, C are all true, therefore, due to our target ground truth value "stroke" being a nominal variable we believe that classification would be the appropriate supervised learning task to build this predictive model.

## Team Identification

The table below lists the research team members as well as the initial model they created for the first iteration of supervised learning model development:

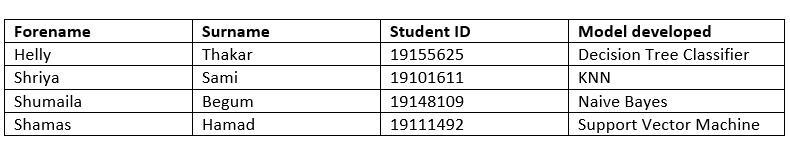


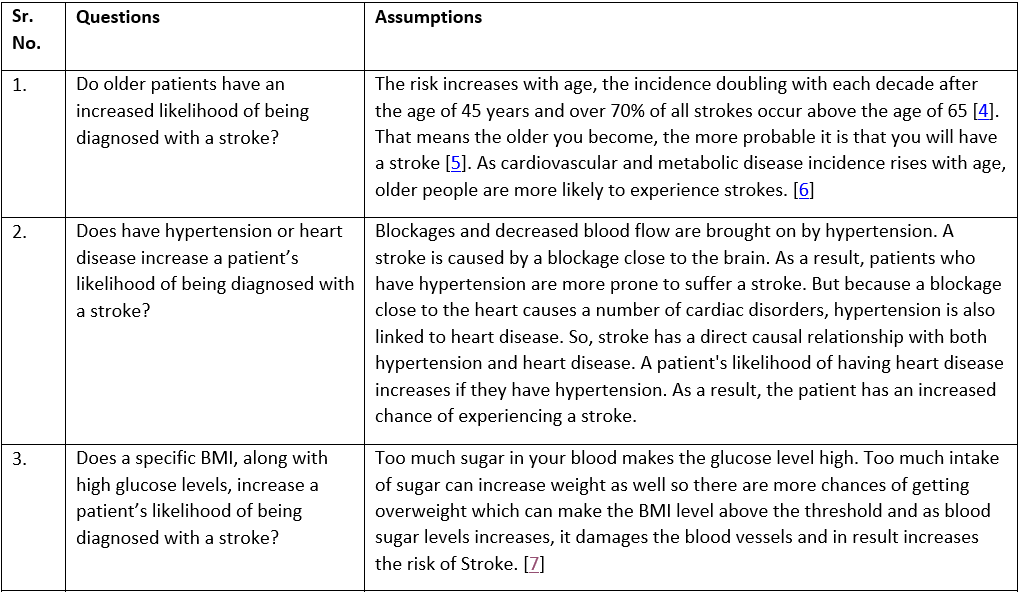
Table 2: Members of Research Team and the initial models developed by them

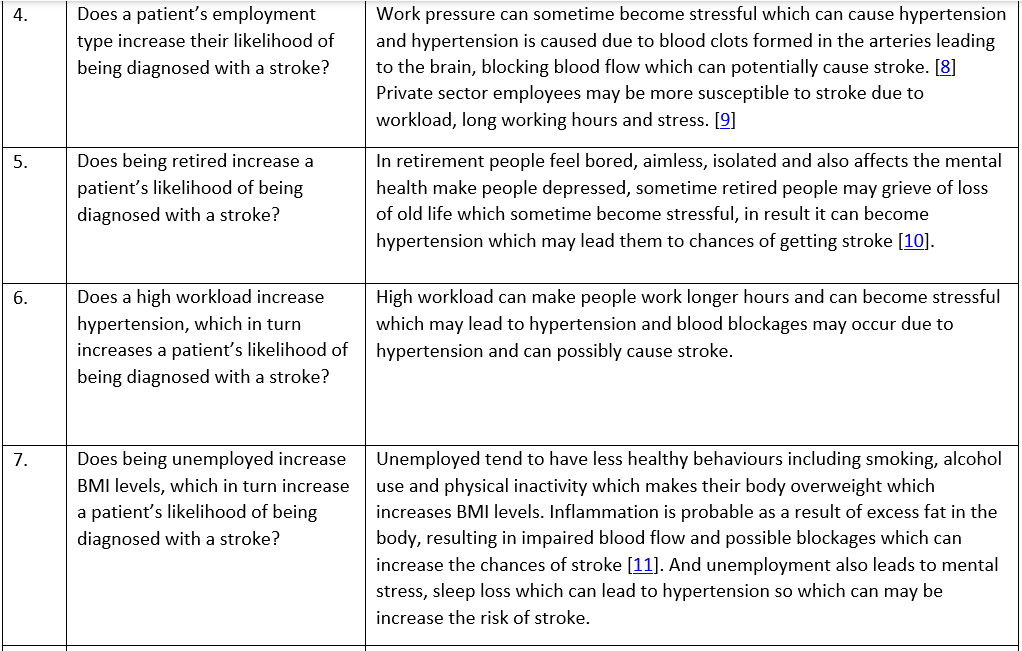
# Exploratory Data Analysis

As now we have identified our dataset and supervised learning task which are stroke and classification technique. Now, we will try to get some initial pattern from our data to further help to choose an algorithm and create our model and evaluate our result. So this section we will go through the Exploratory Data Analysis process by starting with asking questions to our data then splitting the dataset to get a train setting on which we can perform our exploratory data analysis.

## Questions identification

Having a well stated research question allows investigators to focus their study and work toward supporting or rejecting a certain hypothesis. Once the research question has been defined, the remaining components of the study can be identified logically. [[3](https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.15196#:~:text='%20Having%20a%20clearly%20defined%20research,in%20a%20logical%20man%2D%20ner.)]. And it allows us to get clear answer. So by general observation of the stroke dataset and literature search about stroke we will investigate on the following questions and assumptions to allow us to specifically confirm / reject our assumed answer at the conclusion of the EDA:





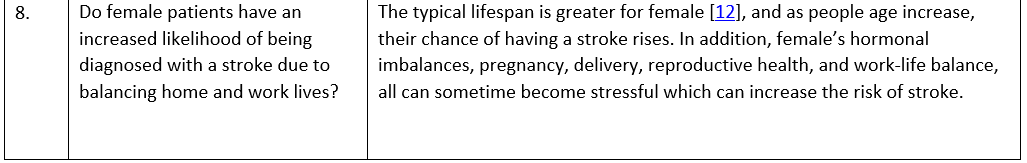


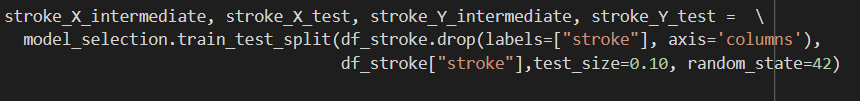
Table 3: Questions with assumptions on the dataset

## Splitting the dataset

The goal of dividing our data into training and validating and testing subsets is to be able to represent previously "seen", and currently “initially unseen” and "unseen" data throughout the model building process. As a result, almost all of your pre-processing (and even EDA) should be done on a sizable sample of previously "seen" data. This is to prevent any patterns / insights from accidentally leaking into the model and interfering with / biassing the evaluation results obtained on this “initially unseen” and "unseen" data at the end of our model development, thereby preventing data leakage [[13](https://datascience.stackexchange.com/questions/52282/is-it-right-to-impute-train-and-test-set)]. And the initially unseen evaluation can been used to improve model and get better evaluation results on “unseen” data.

So here we are splitting the data into three parts: train (“seen”), validate(“initially unseen”) and test (“unseen”).

To split the data we used the **model\_selection.train\_test\_split** method from the Scikit Learn package to get two subsets of data, one for intermediate and one for testing and dividing into the features (the X values) and the targets (the Y values). In the method, we have the **random\_state** value. The **random\_state** can be any positive number, (except 0). It is the parameter used to control the random number generator used [[14](https://scikit-learn.org/stable/glossary.html#term-random_state)]. We need to pick a number to be our **random\_state**. There are a few ways to split the dataset into such as splitting the data into 70:30 ratio or 80:20.

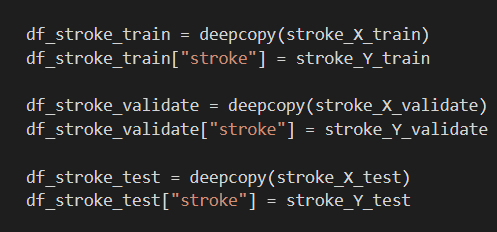




Code-snippet 1:

Code-snippet 1 : Train, Validate and Test Split

A straightforward way to do this is to use the **model\_selection.train\_test\_split** method which will first split the dataset into intermediate (90%), testing (10%). Then, use the **model\_selection.train\_test\_split** method again on the intermediate data only. This will split the intermediate dataset into training (80%) and validation (10%). So we are keeping the text size as 0.1 and here, we are defining the random state in our code because, if we don't, a new random value would be generated each time we run (execute) our code, changing the values of the train, test, and validation datasets.



Code-snippet 2: Generating deep copies of data frame

As well as we are saving our test data with the data frame for later and data frame with all our training data which are done as deep copy, which we will be used for EDA and model building. Then checking the shape of the train, test and validate data to ensure its properly divided:

Graphical user interface, application

Description automatically generated

Code-snippet 3: Checking the shape of the data frames after split

## Exploratory Data Analysis process and results

To prevent data leakage as discussed in 2.2 and avoid overfitting which is not getting good predictions on unseen data rather than seen, we will be doing our EDA solely on seen data which is training set while minimising the developing too many insights on analysing unseen data beforehand. We will start by visualising the potential relationships between feature and target variable using visualising libraries such as seaborn and matplotlib. We will performing univariate, bivariate, and multivariate analysis on our feature and target variables using count plot, histogram, and boxplot. More in depth about analysis and plots is discussed in the 7 Appendix.

### Univariate Analysis

Text

Description automatically generated

Code-snippet 4: Code to visualise count plot of target variable - Stroke

Chart, bar chart

Description automatically generated

Figure 1: Count of patients with and without stroke in the training data set

In Figure 1, less than 500 patients have experienced stroke in our training set, while more than 3500 have not. So 201 patients out of 4139 have a stroke, whereas 3938 do not. This means highly unbalanced data distribution.

Text

Description automatically generated

Code-snippet 5: Code to visualise count plot of feature variable - Gender

Chart, bar chart, treemap chart

Description automatically generated

Figure 2: Count of Female and Male patients in the training data set

Females outweigh males in our training data as we can see in Figure 2. Out of a total of 4139 patients, there are 2419 women and 1720 males.

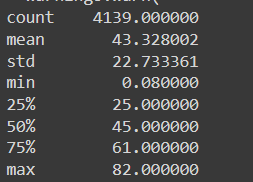


Text

Description automatically generated

Code-snippet 6 & 7: Code to visualise distribution of age variable

Output of code:



|  |  |
| --- | --- |
| Chart, box and whisker chart  Description automatically generated |  |

Figure 3 & 4: Age variable distribution

We can see for the Stroke dataset's Age attribute:

* Q1 (the 25th percentile) is about 25 years. So, 25% of all patients are less than or equal to 25 years
* Q2 (the median and 50th percentile) is about 45 years. So, 50% of all patients are less than or equal to 45 years
* Q3 (the 75th percentile) is about 61 years. So, 75% of all patients are less than or equal to 45 years

There for most of the patients falls in between the age of 25 years to 61 years in our training data. By keep in my mind above all point and the distribution is somewhat negatively skewed which slightly supports our first assumption that as age, the chances of getting a stroke increase because most of the patients are mid- adults or older and the skew also shows that.

Text

Description automatically generated

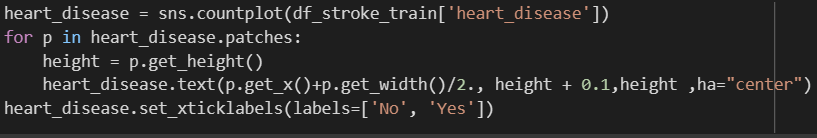
Code-snippet 8: Code to visualise count plot of feature variable - Hypertension

Chart, bar chart

Description automatically generated

Figure 5: Count of patients with hypertension and without in the training data set

There are much fewer people who do not have hypertension than those who do. Out of 4139 people in our training set, 3731 do not have hypertension and 408 do.



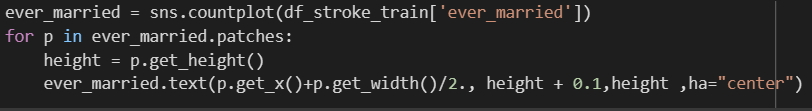
Code-snippet 9: Code to visualise count plot of feature variable – Heart disease

Chart, bar chart

Description automatically generated

Figure 6: Count of patients with heart disease and without in the training data set

Less patients have heart disease than who don’t. 219 patients have heart disease out of 4139 patients and 3920 patients have no heart disease.



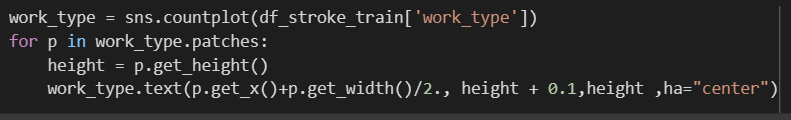
Code-snippet 10: Code to visualise count plot of feature variable – Ever married

Chart, bar chart

Description automatically generated

Figure 7: Count of patients who are married and not married in the training data set

The number of married people outnumbers the number of unmarried people (which makes sense given the distribution of age 25 to 65 in the Figure 3 & 4).



Code-snippet 11: Code to visualise count plot of feature variable – Work type

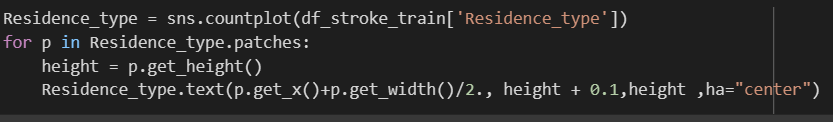
Chart, bar chart

Description automatically generated

Figure 8: Count of patients with scenario they work in, in the training data set

According to our training data most of our patients working in private sectors and the second most are self-employed and the patients who have children and have government job are of similar amount. There are least people who never worked in our training data.

In reference to figure 7, married people have responsibility in real world which shows they might be working in private sector so here we can see most of the patients are working in private.



Code-snippet 12: Code to visualise count plot of feature variable – Residence type

Chart, bar chart, treemap chart

Description automatically generated

Figure 9: Count of patients living in Rural or Urban Area

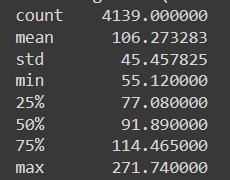
Most of the patients are living in urban which are 2129 patients compared to rural which are 2010 patients in our training dataset. As from figure 8, most of the people are working in private sector that show here in figure 9 that they might be living in urban area as most jobs are in city than villages in real world.

Text

Description automatically generated

Code-snippet 13 & 14: Code to visualise distribution of average glucose level variable

Output of code:



|  |  |
| --- | --- |
|  |  |

Figure 10 & 11: Distribution of average glucose level in blood of patients

We can see for the Stroke training dataset's average glucose level attribute:

* Q1 (the 25th percentile) is about 77 mg/dl. So, 25% of all patients have less than or equal to 77 mg/dl which is good.
* Q2 (the median and 50th percentile) is about 92 mg/dl. So, 50% of all patients are less than or equal to 92 mg/dl which is good.
* Q3 (the 75th percentile) is about 115 mg/dl. So, 75% of all patients are less than or equal to 115 mg/dl which is also good.

So most of the patients have glucose level between 77 to 115 mg/dl in our training data and the distribution is positive skewed. That indicates most of the people are not diabetic patients. And the outliers shows that there are few diabetic patients which are outside the "minimum" and "maximum" expected range of values which is more than 115 mg/dl. And the distribution is positive skewed.

Text

Description automatically generated

Code-snippet 15 & 16: Code to visualise distribution of bmi (body mass index) variable

Output of code:

A screenshot of a computer

Description automatically generated with low confidence

|  |  |
| --- | --- |
|  |  |

Figure 11 & 12: Distribution of BMI (Body Mass Index) of patients

We can see for the Stroke training dataset's average glucose level attribute:

* Q1 (the 25th percentile) is about 24. So, 25% of all patients have less than or equal to 24 which are good.
* Q2 (the median and 50th percentile) is about 28.30. So, 50% of all patients are less than or equal to 28.30 which are overweight.
* Q3 (the 75th percentile) is about 33. So, 75% of all patients are less than or equal to 33 which are also overweight.

So most of the patients have BMI between 24 to 33 in our training data and the normal BMI range is **18.5 to 24.9**. So it shows most of the patients are overweight in our training data. And the outliers shows that there are patients who are overweight which are outside the "minimum" and "maximum" expected range of values, who have more than 33 BMI which slightly support our 3rd assumption in 2.1 so they can be prone to get a stroke.

Text

Description automatically generated

Code-snippet 17: Code to visualise count plot of feature variable – Smoking status

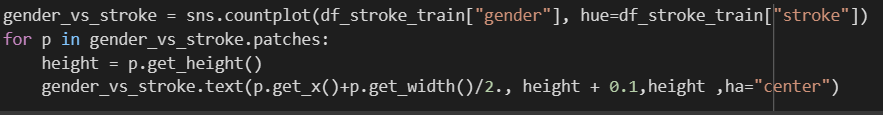
Chart, bar chart

Description automatically generated

Figure 13: Count of patient’s smoking status

In our data most of the stroke patients have Never Smoked which are 1508 out of 4139 and second highest count is of unknown so other than never smoked there are more patients whose smoking status is unknown in our training data. There might be patients who don’t who to share their smoking status or the value might be not filled or missed during the data collection.

### Bivariate Analysis



Code-snippet 18: Code to visualise the count plot of feature - gender with target - Stroke

Chart, bar chart

Description automatically generated

Figure 14: Count of Female and Male patients with and without Stroke

The overall number of female participants in this dataset is higher than the total number of male patients. There are more 35 females than the count of males who had stroke. Currently we are unable to find compelling evidence in our dataset that supports our 8th assumption in 2.1 fully and answers 8th question as well.



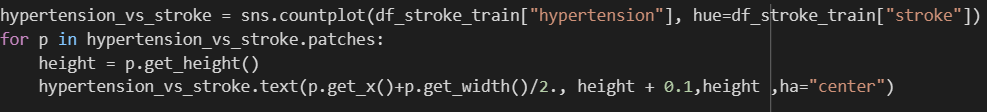
Code-snippet 19: Code to visualise the boxplot for distribution of target - stroke among feature - age

Chart, box and whisker chart

Description automatically generated

Figure 15: Distribution of patients having stroke among age

Despite the fact that the majority of patients did not have a stroke, it can occur at any age. The age range of participants who have had a stroke is 60 to 80 as we can see in figure 15. So this support our 1st assumption which we did in 2.1 and answers our 1st question that older patients have an increased likelihood of being diagnosed with a stroke. And as we can see, there are two outliers that indicate that two individuals between the age of 0 and 20 had a stroke, which may be accurate given that although many people think of stroke as an old disease, there is an incredibly significant chance that it will happen to a child during the perinatal era [[15](https://www.ninds.nih.gov/health-information/public-education/brain-basics/brain-basics-preventing-stroke#:~:text=Some%20ways%20that%20work%3A%20Maintain,Exercise%20more.)].



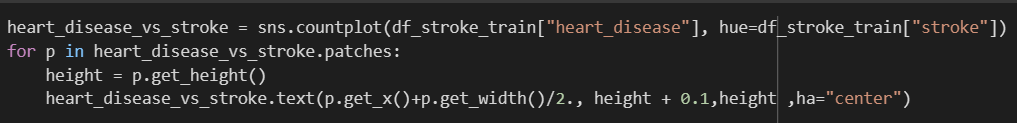
Code-snippet 20: Code to visualise the count plot of feature - hypertension with target - stroke

Chart, bar chart

Description automatically generated

Figure 16: Count of patients with and without hypertension vs with and without stroke

In this training dataset, there are significantly more people without hypertension than those who do. The patients who had hypertension also had stroke are less than patient who had stroke but did not had hypertension, so, here it answers our 2nd question in 2.1 and also does not support our 2nd assumption that patients who have hypertension likelihood of being diagnosed with a stroke because here the number of patients who had hypertension but does not had stroke are more than who had stroke had hypertension.



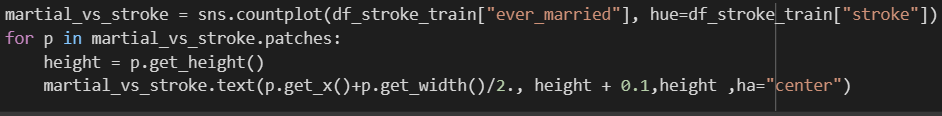
Code-snippet 21: Code to visualise the count plot of feature – heart disease with target - stroke

Chart, bar chart

Description automatically generated

Figure 17: Count of patients with and without heart disease vs with and without stroke

There are very few people who had heart disease than people who don’t. The patients who had heart disease also had stroke are less than patients who had stroke but who did not had heart disease, so, here it answers our 2nd question in 2.1 and also does not support our 2nd assumption that patients who have heart disease likelihood of being diagnosed with a stroke because here the number of patients who had heart disease but does not had stroke are more than who had stroke had heart disease.



Code-snippet 22: Code to visualise the count plot of feature – ever married with target - stroke

Chart, bar chart

Description automatically generated

Figure 18: Count of patients married and not married vs with and without stroke

Based on the above figure, it seems patient who are married had stroke are more than patients who are not married and had not had stroke. Married people have more responsibility and have to manage work-life-balance so they might have more stress which can be a reason for stroke.

Text

Description automatically generated

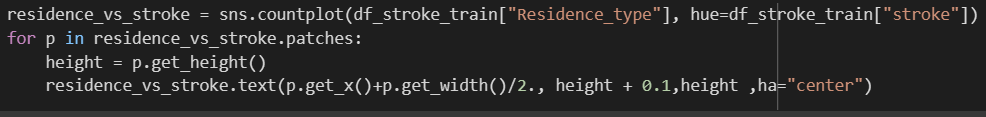
Code-snippet 23: Code to visualise the count plot of feature – work type with target - stroke

Chart, bar chart

Description automatically generated

Figure 19: Count of patient’s working scenario vs with and without stroke

Most stroke patients are either self-employed or employed by the government in addition to the private sector. The majority of stroke patients among them all work in the private sector which supports our 4th assumption in 2.1 that given their workload and potential for stress, private sector employees may be more susceptible to stroke and also answer the 4th question that patients employment type does increase their likelihood of getting stroke and here we can see patients who are working in private sector are more prone to get a stroke.



Code-snippet 24: Code to visualise the count plot of feature – residence type with target - stroke

Chart, bar chart

Description automatically generated

Figure 20: Count of patient’s dwelling type vs with and without stroke

From the above figure, we can see patients who are living in urban had stroke outnumbers patients who are living in rural had stroke. May be patients who are living in urban might have more stressful life, might be working in private sectors(refer. Fig. 19), might be married (refer. Fig. 18) and moved from rural to urban area due to job (refer. Fig. 19), which can be reasons for likelihood of getting a stroke.



Code-snippet 25:

Chart, box and whisker chart

Description automatically generated

Figure 21:

The boxplot demonstrates that individuals who have had a stroke have a higher average blood glucose level than individuals who have not.

Text

Description automatically generated

Code-snippet 26:

Chart, box and whisker chart

Description automatically generated

Figure 22:

According to the boxplot above, patients who have had a stroke had higher BMI which is between 29 to 33

Text

Description automatically generated

Code-snippet 27: Code to visualise the count plot of feature – smoking status with target - stroke

Chart, bar chart

Description automatically generated

Figure 23: Count of patient’s smoking status vs with and without stroke

We can observe that smoking habits do have an effect on the chance of having a stroke. Smoking contributes to atherosclerosis, a build-up of fatty substances in the carotid artery, the major artery supplying blood to the brain. Furthermore, smoking raises blood pressure. Smoking causes carbon monoxide, which lowers the quantity of oxygen in the brain's blood supply. Smoking also thickens the blood, which facilitates clotting.

### Multivariate Analysis

And then we performed multivariate analysis of relationship between the features and features variable by



Code-snippet 28:

A picture containing shoji, building

Description automatically generated

Figure 24:

A screenshot of a computer

Description automatically generated with medium confidence

The features may be found to have little correlation with one another. There is no multicollinearity issue as a result.

## EDA conclusions

The following conclusion may be derived from this dataset's exploratory analysis:

1. As people age, their chance of getting a stroke increases.

2. The risk of having a stroke is increased in people with higher average glucose levels.

3. Stroke risk is increased for people who have had hypertension in the past.

4. Stroke risk is increased in those who have had a cardiac condition in the past.

5. Patient/ people who are married, working in private sector, and living in urban area have likelihood of getting a stroke so variable like marital status, type of employment and dwelling are statistically significant.

6. People who has higher glucose level and higher BMI are prone to get a stroke.

6. Smoking status does affect stroke risk, although its importance is not as great as that of other key variables. People who smoke regularly at an increased chance of suffering a stroke.

# Experimental Design

## Identification of your chosen supervised learning algorithm(s)

The supervised learning algorithms chose here is Decision Tree Algorithm.

First off, a decision tree is a tree-based method in which each route leading from the root is characterised by a data separating sequence until a Boolean result is attained at the leaf node. It is a node-and-connection-containing hierarchical exemplification of knowledge relationships. Nodes indicate purposes when relations are used to categorise.

One of the effective techniques frequently employed in a variety of domains, including machine learning, image processing, and pattern recognition, are decision trees. A set of fundamental tests are efficiently and cogently united by decision trees, where each test compares a numerical property to a threshold value. The conceptual principles in the neural network of connections between nodes are significantly simpler to create than the numerical weights. DT is used mostly for grouping reasons. Additionally, a common categorization model in data mining is decision trees. Each tree is made up of nodes and branches. Each node represents a feature in a classification category, and each subset specifies a value the node may accept. Decision trees have several implementation domains due to their straightforward analysis and accuracy on various data formats.

The primary goal of the decision trees algorithm, which is a member of the family of supervised learning algorithms, is to build a training model that can be used to predict the class or value of target variables by learning decision rules inferred from the training data.

There are several benefits and drawback of decision trees:

Benefits

1) Easy to understand

2) Rapidly transformed into a set of production-related principles.

3) Has the ability to categorise both categorical and numerical results, but the characteristic produced must be categorical.

4) No a priori hypotheses are made while taking the quality of the outcomes into account.

Drawbacks:

1) If the best decision-making mechanism is inhibited, bad decisions may result.

2) The decision tree is fascinating since it has several tiers.

3) The complexity of the decision tree's computation may rise with additional training samples.

With the proper methods, these drawbacks may be readily solved. We only need to approach them mindfully and appropriately.

Finally, decision trees are appropriate because they let us choose the best feasible answer that is simple, requires little to no pre-processing, and is flexible when it comes to data.

## Identification of appropriate evaluation techniques

Metrics used to assess a model depend on the kind of machine learning activity it is used for. For the classification job, for instance, a predicted category's similarity to the actual category is used to gauge the model's performance. In order to assess our models, we may use a variety of indicators. The model's kind and strategy for execution have a significant impact on the metric chosen. Confusion matrix and F1-score was used here.

The confusion matrix divides predictions into a number of relevant categories, highlighting how one class may be mistaken for another:Table

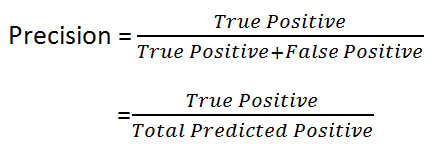
Description automatically generated

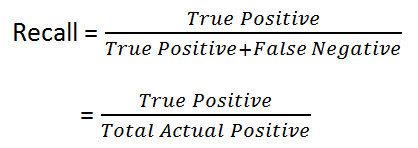
It provides details on the kinds of errors being produced by the classifier as well as the faults themselves. It exhibits the disarray and fuzziness of a classification model's predictions.

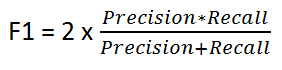
Graphical user interface, application

Description automatically generated

Here we can see how there are most true negative values and few true positive values with no false negative and no false positive value. In the above we can see that we appear to have no false positives or false negatives when predicting our training set - this seems reasonable, but we will need to be careful, just because our classifier performs well on data it has already seen during training that doesn't mean that it will perform well on unseen data

The proportion of correctly predicted positive classes to all items with positive predictions is called precision.  
  
  
  
This, intuitively, reveals the accuracy or precision of the positive predictions made by our model. When we think False Positives are more significant than False Negatives, precision is crucial.

Recall is defined as the proportion of accurately predicted positive classes to all positively categorised items:  
  
  
  
It measures the proportion of occasions when we were able to forecast the good outcomes (or recall). When we think False Negatives are more significant than False Positives, recall is crucial.

The F1-score is a single performance statistic that considers both recall and accuracy. It is also frequently referred to as the F-Measure. It is calculated by averaging the two measures harmonically:

The F1-score is a fantastic tool for evaluating how well different classifiers perform. It may be used to decide which model, out of many with different accuracy and/or recall levels, gives the "best" answers to the question at hand. Because of this, it is frequently applied in practise as a performance indicator to rank models.

Table, calendar

Description automatically generated

## Data cleaning and Pre-processing transformations

Data cleaning is the process of removing erroneous, damaged, badly formatted, duplicate, or missing data from a dataset. When combining several data sources, there are many possibilities for data to be duplicated or improperly categorised.

Missing values occur when certain participants or variables don't have data recorded for them. Numerous factors, such as improper data entry, device malfunctions, deleted files, and others, can result in data loss. In this case, we are imputing the missing data. Imputation is the process of replacing a missing value with a different value based on a reasonable assumption. You add more information to the dataset to make up for the missing value for a more complete dataset. We have a wide range of imputation methods at our disposal. The easiest method to impute missing data is to replace them with the mean or median value for that variable.

Since the distribution of the data values is symmetrical, there are few to no noticeable outliers, and the mean is the metric whose computation considers all of the values in the data set, we are using it. In this scenario, we address the missing numbers via mean imputing. First of all we are checking whether there are missing values:

A screenshot of a computer

Description automatically generated with low confidence

And as we can see 152 values are missing in bmi so now we have to fill that missing values so that we can train our model and get the accuracy.

We will start by defining the function and passing in the dataframe and column then we're using Sklearn's fit method to 'fit' the imputer on the specified column then we're using Sklearn's transform method to 'transform' the specified column and are returning the imputer object for future use. This imputer object holds the initial mean from the training set. Therefore, this mean can be used for the testing set. We can't create a mean from the testing set for two reasons: 1. the testing data must be kept unseen, so we can't use these values 2. we don't know how much data we have. In a real-world example, there will always be more data, hence a mean can't be devised at any one time. In the next code block, we're setting bmi\_imputer as the result of calling the function impute\_attribute with the parameters of dataframe and column - X\_train and bmi.

Text

Description automatically generated

After imputing we are again checking the data:

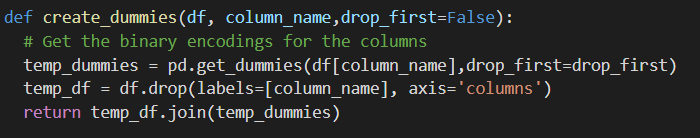
Text

Description automatically generated

And as we can see there are no missing values.

Then we need to encode our categorical data into forms that the ML algorithms can better understand. This is widespread practice and there are various tools available within ML libraries to allow us to do this, some of these methods can be found in the SKLearn Preprocessing library, for example OneHotEncoding.

So from pandas we are using the method get\_dummies to create dummy variable from categorical variable:



A dummy variable is a binary variable that indicates whether a separate categorical variable takes on a specific value.

A computer screen capture

Description automatically generated with medium confidence

As you can see above all categorical variables are converted into dummy variables.

## Limitations and Options

We don't have enough time to do the balancing of the data cause it's very unbalanced, so, we've got class imbalance but so it limits how accurate our model could be and we would have used our SMOTE or isolation forest or one of the other techniques to balance the data such as oversampling of the data.

# Predictive Modelling / Model Development

## The predictive modelling process

To increase the accuracy and efficiency of a machine learning model we used our pre-processed data. We started by training the model using features and target:

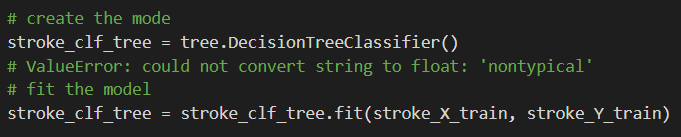
Text

Description automatically generated

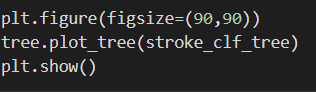
Here we are dropping the stroke from features because it is our target variable .

DecisionTreeClassifier is a class capable of performing multi-class classification on a dataset. As with other classifiers, DecisionTreeClassifier takes as input two samples: stroke\_X\_train holding the train samples, and stroke\_Y\_train of integer values, holding the class labels for the training samples then fit function is used to it function adjusts weights according to data values so that better accuracy can be achieved.

So we are creating and fitting the model:



Then we created a visualisation of the tree using the plot\_tree fuction from sklearn:



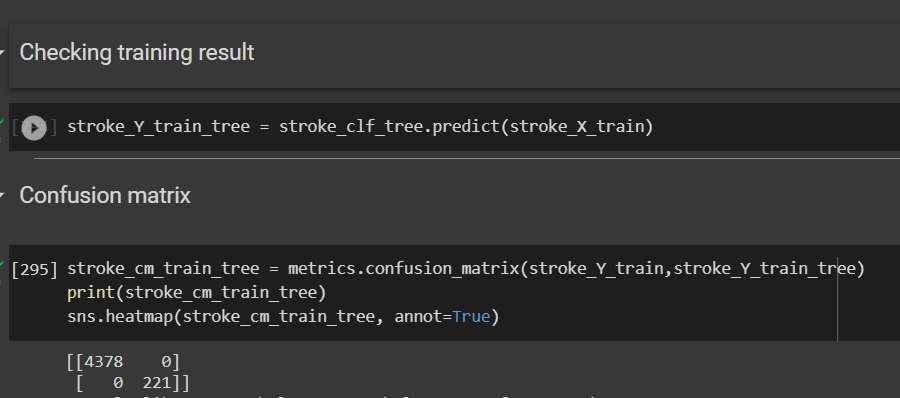
This is what we get

Background pattern

Description automatically generated with medium confidence

## Evaluation results on “seen” data

After being fitted, the model can then be used to predict the class of samples:



And then to get in dept evaluation we applied confusion\_matirix function to evaluate classification accuracy by computing the confusion matrix with each row corresponding to the true class as we can see from the above screen-capture. Here is what we got:

.Graphical user interface, application

Description automatically generated

In the above example we can see that we appear to have no false positives or false negatives when predicting our training set - this seems reasonable, but we will need to be careful, just because our classifier performs well on data it has already seen during training that doesn't mean that it will perform well on unseen data.

And then used the classification\_report function from sklearn which builds a text report showing the main classification metrics which gave the precision recall and f1-score:

A picture containing table

Description automatically generated

Here we notice that our decision tree classifier seems to perform good in terms of its training metric results.

# Evaluation and further modelling improvements

Show the results of model improvement by applying SMOTE analysis, data scaling, data normalisation, feature selection, hyperparameter optimization

Model might be overfitting so it’s giving 100% accuracy and then we performed EDA on testing dataset to see weather the number of records split for stroke dataset so are they people with stroke or not

## Initial evaluation comparison

## Further modelling improvements attempted

## Final Evaluation results

# Conclusion

## Summary of results

## Suggested further improvements to the model development process

We could have used gant chart in the future and I would have also used github, padlet.

For example, for those of you dealing with imbalanced classification problems, might stratification be useful for this type of data?

<https://code.likeagirl.io/good-train-test-split-an-approach-to-better-accuracy-91427584b614>

binning can be added to the histplot while doing the eda of age, glucose level amd bmi

## Reflection on Research Team

Firstly by working in research team we got the opportunity to gain experience from each other. We descriptively discussed the process of doing the research step by step and asking for guidance from each other and by being in different course one in computer science and other three in computer and data science, we were able to get guidance from wider range of experienced professionals when taking feedback and tackling problems as we were immediately discussing either writing descriptive about what the tutors suggested about tackling a particular problem or arranging a quick meeting so we can see everyone’s agreement. Overall there was effective communication, collaboration readiness and regular and open discussion as well as effective brainstorming.

#Did you planned out initial stages that you were working on effectively?

We individually planned out first stages by refereeing to pervious labs that were done in the lectures as well as researched more to get an idea of everything beforehand.

## Reflection on Individual Learning

We individually took advantage of support sessions available during the module to get feedback and support if we were stuck on any part of the coursework and to effectively draft the report from the feedback give to us after the session. Based on the support sessions we improved the code and removed unwanted stuff and did more EDA and tried to understand why we need certain methods to improve our result individually.

The insights gained through this module were how we can do better for world using data, creating model and save one life, one business.

The insights gained through the process of trying the course work were how I can create amazing model by refining the accuracy through number of iterations on data, algorithm, and parameters. And how model made through dataset of specific region can only predict things of that regions like here we are predicting stroke from the data collected from different medical centres in Bangladesh so it can predict stroke of that region only, it cannot predict stroke of China due to its climate, lifestyle etc.

I use the support available from tutors in the labs/additional sessions by

The step I found most difficult was …… because ……

The easiest step was model training because sklearn has easy code to just call the algorithm

This project has the potential to be spent and used in

# References

Sailasya, G. and Kumari, G.L.A. (2021). Analyzing the Performance of Stroke Prediction using ML Classification Algorithms. *International Journal of Advanced Computer Science and Applications*, [online] 12(6). doi:10.14569/ijacsa.2021.0120662.

‌

Emon, M.U., Keya, M.S., Meghla, T.I., Rahman, Md.M., Mamun, M.S.A. and Kaiser, M.S. (2020). *Performance Analysis of Machine Learning Approaches in Stroke Prediction*. [online] IEEE Xplore. doi:10.1109/ICECA49313.2020.9297525.

‌

# Bibliography

[ROC Curve with Visualization API — scikit-learn 1.2.0 documentation](https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_roc_curve_visualization_api.html)

*Urinary incontinence and pelvic organ prolapse in women: management*, <https://www.ncbi.nlm.nih.gov/books/NBK577759/bin/bm2.pdf>.

*How to Create Dummy Variables in Python with Pandas?*, <https://www.geeksforgeeks.org/how-to-create-dummy-variables-in-python-with-pandas/>.

*1.10. Decision Trees — scikit-learn 1.2.0 documentation*, <https://scikit-learn.org/stable/modules/tree.html>.

1. [Neuroscience for Kids - Imaging (washington.edu)](https://faculty.washington.edu/chudler/image.html#:~:text=Brain%20imaging%20methods%20allow%20neuroscientists,are%20affected%20by%20neurological%20disorders.)

2. Boehme, A.K., Esenwa, C. and Elkind, M.S.V. (2017). Stroke Risk Factors, Genetics, and Prevention. *Circulation research*, [online] 120(3), pp.472–495. doi:10.1161/CIRCRESAHA.116.308398.

3. [A good study starts with a clearly defined question - Lane - 2018 - BJOG: An International Journal of Obstetrics & Gynaecology - Wiley Online Library](https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1111/1471-0528.15196#:~:text='%20Having%20a%20clearly%20defined%20research,in%20a%20logical%20man%2D%20ner.)

4. [Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3006180/#:~:text=The%20risk%20increases%20with%20age,above%20the%20age%20of%2065.)

5. [Know Your Risk for Stroke | cdc.gov](https://www.cdc.gov/stroke/risk_factors.htm)

6. [Stroke prevention and management in older adults - PubMed (nih.gov)](https://pubmed.ncbi.nlm.nih.gov/16966926/)

7.

8. [Job-loss associated with weight gain - but diet and physical activity may not be the reason - Communicating Diet and Activity Research (cam.ac.uk)](https://www.cedar.iph.cam.ac.uk/2015/09/21/job-loss-associated-weight-gain/)

9. [Around the Globe, Women Outlive Men | PRB](https://www.prb.org/resources/around-the-globe-women-outlive-men/)

10. [High blood pressure dangers: Hypertension's effects on your body - Mayo Clinic](https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/high-blood-pressure/art-20045868)

11. [Perceptions of work stress causes and effective interventions in employees working in public, private and non-governmental organisations: a qualitative study - PMC (nih.gov)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5353523/)

12. [Adjusting to Retirement - HelpGuide.org](https://www.helpguide.org/articles/aging-issues/adjusting-to-retirement.htm)

13. [classification - Is it right to impute Train and Test set? - Data Science Stack Exchange](https://datascience.stackexchange.com/questions/52282/is-it-right-to-impute-train-and-test-set)

14. [Glossary of Common Terms and API Elements — scikit-learn 1.2.0 documentation](https://scikit-learn.org/stable/glossary.html#term-random_state)

‌

# Appendix

Countplot will be used to see the counts of observations of stroke in each categorical bin using bars.