Table of Content

1.0 Abstract

Stroke is a dangerous medical condition that affects the arteries that supply blood to the brain and grows worse if it is not treated right away. In fact, according to the World Health Organisation (WHO), stroke is one of the leading causes of mortality and disability in the globe. However, if a stroke is discovered sooner or treated right away, its severity may be decreased or even averted. In order to predict strokes, five machine learning models were developed in this paper: Logistic Regression, Naive Bayes, K-Nearest Neighbours, Decision Tree Classifier, and Random Forest Classifier. To attain the greatest model accuracy possible, several sampling methods were examined together with feature selection. The evaluation is conducted using a comparison of performance indicators including accuracy, precision, recall, and AUC value. The greatest accuracy was obtained using SMOTE data with Random Forest Classifier without feature selection with 95.00% accuracy, 0.94 precision, 0.96 Recall, and 0.99 AUC score. Oversampling dataset has been demonstrated to yield better results than undersampling dataset. This can be proven with the performance of models with RUS dataset. With just 78.00% accuracy, 0.76 precision, 0.82 recall, and 0.86 and 0.84 AUC scores, respectively, the Logistic Regression and Random Forest Classifier are the highest accuracy models employing RUS dataset.

2.0 Problem Statement

This research paper uses and reviews several Machine Learning methods to predict the likelihood of an individual getting a stroke. It is always better to hope for the best but prepare for the worst, taking early measurements can save an individual's precious life. More often than not, when a stroke patient finds out that they carry this disease, it is already too late. Therefore, when we can predict the probability of an individual getting a stroke, it allows them to take measurements and precautions. Thus, this research paper intends to find out what are the factors that contribute to getting a stroke as well as building a model to make predictions about the likelihood of getting a stroke

Research questions

- How does age affect the probability of getting a stroke?
- Is smoking a heavy factor in getting a stroke?
- How does having hypertension play a role in getting a stroke?
- What is the average glucose level of a patient that has a stroke?
- How does having heart diseases play a role in getting a stroke?
- Is gender a factor in getting a stroke?
- What is the average body mass index of a patient that has a stroke?
- Do work type and residence type contribute to getting a stroke?

Background and our motivation of choosing stroke prediction dataset

We are strongly aware that stroke is one of the highest lending causes of death globally. A stroke also known as brain attack happens when blood flow to the brain is interrupted or if a brain blood vessel breaks. Stroke is a serious medical emergency that can result in permanent brain damage, paralysis, or even death. The brain is the seat of our thoughts, feelings, and language, and it is responsible for all of our bodily functions. The brain regulates several bodily processes, including respiration and digestion.

To work properly, your brain needs oxygen. Your arteries provide oxygen-rich blood to all parts of your brain. Without oxygen, brain cells begin to die within minutes of a blockage in blood flow. When this happens, a stroke occurs (Mayo Clinic, 2022). There are two types of stroke namely Ischemic stroke and Hemorrhagic stroke (Centers for Disease Control and Prevention, 2022). When brain blood flow is interrupted due to the formation of clots or other debris, this is known as an ischemic stroke. When a brain vessel bursts or leaks, causing blood to spill out, this is known as a hemorrhagic stroke. The increased strain on brain cells caused by the bleeding is harmful. Stroke is the most common cause of permanent disability in the developed world. More than half of elderly stroke survivors have diminished mobility as a result of their disease.

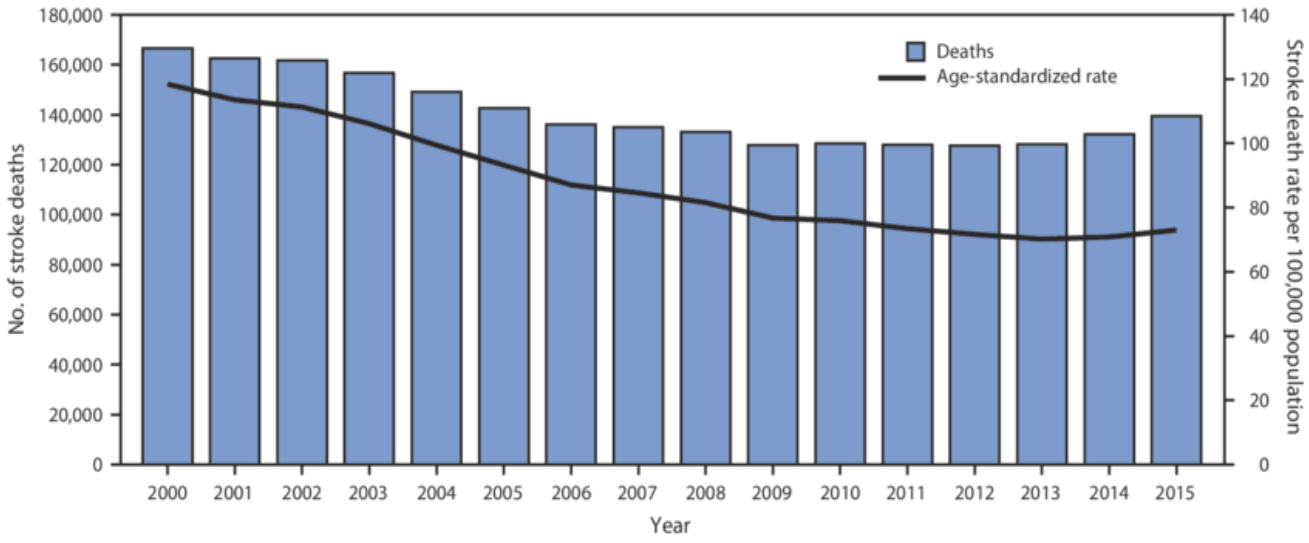


Figure 1: Chart of num of death between 2000 and 2015 (Xin Tong, 2017)

The chart above was a survey conducted between 2000 and 2015. As you can see, the annual death rate surpasses 100,000 every year in which the highest number of deaths was recorded in the year 2000 with a total number of 160,000. This has proven that stroke must be taken and dealt with seriously. Early precautions are crucial before it turns disastrous.

Our motivation for selecting this as our topic of choice is because we want to play a role in decreasing the number of deaths caused by stroke worldwide. There are a variety of common factors that contribute to getting a stroke which have been included in our dataset below such as

gender, age, smoking status, hypertension, heart disease and many more. We have carefully analysed the datasets so that we can predict whether an individual is likely to get a stroke based on his or her lifestyle. We can do so by building our prediction model utilising Machine Learning.

3.0 Literature Review

For the literature review, we will mainly focus on five research papers that are related to stroke prediction using machine learning to provide a better comparison.

Data Source

The research paper 'Stroke prediction through Data Science and Machine Learning' by Tavares, Jose-A [9] uses a dataset from Kaggle but they didn't specify which dataset they used. The dataset contains 5110 individuals with 10 features. Dev et al. [3] used the dataset of electronic health records released by McKinsey & Company. The dataset was retrieved from Kaggle and it contains the EHR records of 29072 patients. It has a total of 11 input attributes and 1 output feature. Alanazi et al. [2] used the National Health and Nutrition Examination Survey. They used the data from 2011 to 2015 which included 15,714 participants and 608 of them had strokes and to reduce the impact of imbalanced data, the final number of participants they include is 4186 which is after they reduce the number of participants who did not have strokes. For this dataset, it contains 21 attributes. Sailasya and Kumari [6] and Tazin et al. [10] used the same dataset which is the 'Stroke Prediction Dataset' from Kaggle, which includes 5110 rows and 12 columns.

Features

For four of the research papers, they have common features like age, gender, marital status, work type, residence type, hypertension, heart disease, avg glucose level, body mass index (BMI), smoking status and stroke. The features that are available in Alanazi et al. 's [2] paper are different compared to the others. The features they have are gender, age, and the numerical data of the blood test result which includes information like red cell distribution width, lymphocytes, red blood cell folate, segmented neutrophils, haemoglobin, red blood cell count and more.

Data Preprocessing

Data preprocessing is necessary before developing a model to get rid of noise and outliers that could cause the model to deviate from its intended training. At this point, all the obstacles to the model's optimal performance are removed. Next, the data from the selected dataset needs to be cleaned and formatted in preparation for model building. Tavares, Jose-A [9] decided to not

consider the marital status and BMI for model training after performing their feature selection with the help of the Pearson Correlation technique and backward elimination technique. However after Dev et al. [3] calculate the correlation between the features, it shows that none of the features are highly correlated with each other. This means each feature has an individual contribution to stroke prediction. For Alanazi et al. [2], they decided to analyse their data in three ways: without data resampling, with data imputation and with data resampling. When they applied data imputation, they replaced the missing values and deleted features that have more than 50% missing values. While Sailasya and Kumari [6] and Tazin et al. [10] used the same dataset, they both omitted the column 'id' since it won't help much in predicting the stroke, and they also fill in the missing value of BMI with the mean of the column data. Moreover, they perform label encoding which transfers strings into integers that the machine can understand. The dataset is also highly imbalanced wherein only 249 rows out of 5110 rows have strokes. The way they handled the imbalanced data are different. Sailasya and Kumari [6] used the method of undersampling which the majority class is undersampled to match the minority class. In this case, the dataset will have 249 rows. While for Tazin et al. [10], they used the SMOTE technique which will result in the dataset having 4861 rows.

Statistical Methods

For the statistical methods, we find out that most of the research used central measurements like mean, mode, and median of the features to handle the missing data in the dataset. Other than that, most of the studies implement Pearson's correlation coefficient when performing feature selection. If two features are highly correlated, one of them can be ignored since it does not provide any additional contribution to predicting stroke. Moving on, various kinds of graphs are utilised to visualise the data. Tavares, Jose-A [9] used a heatmap to visualise the correlation matrix between features. Other graphs like histograms, box plots, biplots, bar charts, and more are also plotted to provide better data analysis.

Machine Learning

The authors of the chosen research papers and articles utilise various different machine learning algorithms to process the data and output an interpretable set of results. The algorithms used were Bayes Net, Decision Tree, J48, K Nearest Neighbour (KNN), Logistic Regression, Naïve

Bayes, Neural Network, Random Forest, Support Vector Machine (SVM), Voting Classifier, XGBoost Classifier. The algorithms are able to identify the patterns of factors that may predict stroke, this allows them to be used to create machine learning models which produce predictive models. Tavares, Jose-A [9] utilised most of the listed algorithms for their data training and results, to be specific they used Decision Tree, KNN, Logistic Regression, Naïve Bayes, Neural Network, Random Forest, SVM, and XGBoost Classifier. On the other hand, Dev et al. [3] utilises Decision Tree, Neural Network, and Random Forest for its predictive model. Alanazi et al. [2] made use of four (4) algorithms, two (2) of which are not used by any other paper, namely Bayes Net and J48, while the common ones are Naïve Bayes and Random Forest Classifier. Sailasya and Kumari [6] make use of Decision Tree, Logistic Regression, Naïve Bayes, Random Forest and also SVM. Finally, Tazin et al. [10] use the following algorithms: Decision Tree, Logistic Regression and Random Forest.

Performance Metrics

It is essential to test the models to observe the performance as well as review the metrics used. For the research papers discussed for this review there are various algorithms, namely: Accuracy, Precision, F-Score, Miss Rate, Accuracy Variance, is the most important PPV and NPV, F1-score and also the Area under the curve (AUC).

Accuracy

The accuracy varies in the different researches, for Tavares, Jose-A [9] the best results were from the Random Forest Classifier with an accuracy of 92.3%. While the most inaccurate is Naïve Bayes with an accuracy of 79%. Meanwhile, Dev et al. [3] had the best accuracy as 80% from the Neural Network method. Alanazi et al. [2] have produced the results where the highest accuracy is from Random Forest Classifier with an accuracy of 96%. The best performance in the research of Sailasya and Kumari [6] found that Naïve Bayes Classification with accuracy of 82% and the most inaccurate was Decision Tree with the percentage being 66%. Finally, Tazin et al. [10] where the Random Forest algorithm resulted in the highest accuracy of 96%.

Conclusion

Reviewing the selected research papers has allowed us to get an understanding on the findings and discoveries that have already been made in our selected problem. We can use this knowledge to identify the similarities and differences between the various researches as well as to understand their different approaches in solving the problem. In this case it is identifying the methods and algorithms used to get the most accurate result. This understanding can be used to select the areas to focus on for our study and also to identify areas where we can make changes to improve the final results.

Critical Evaluation

Out of the five research papers reviewed the best accuracy was by Tazin et al. [10], which resulted in a 96% accuracy using a Random Forest algorithm. Comparatively the second highest in all of the researches was 94% accuracy also by Tazin et al. [10] but with a Decision Tree algorithm. On the other hand the most inaccurate result was 66% by Sailasya and Kumari [6] from when they used a Decision Tree algorithm. From this we can see that the same algorithm, Decision Tree, can produce the second best results and also the worst, simply based on how the model is designed.

4.0 Data

4.1 Brief Description of Data

The dataset for this research project was obtained from Kaggle. This multivariate dataset includes one target variable and has a total of 5110 records (rows) and 12 attributes (columns). The variables include avg glucose level," age," Residence type," work type," gender," "id," stroke". hypertension," smoking status," bmi," hypertension," ever married and " heart disease," The variable "stroke," which has a value of either "1" or "0," depending on whether the patient has a history of strokes or not, is the target variable for this dataset. The dataset is severely unbalanced since the prospect of '0' (no history of stroke) has 4861 entries, much outnumbering the possibility of '1' (has had a history of stroke) with just 249 rows in the stroke column. Pre-processing of data would need to be performed.

Attributes	Data Attribute Type	Type Of Data	Description
1. id	Integer	Numerical (Discrete)	Patients are identified by distinct integer values.
2. gender	String	Categorical (Nominal)	Patients' gender (Female, Male , Other)
3. age	Integer	Numerical (Discrete)	The patients' age
4. hypertension	Integer	Categorical (Nominal)	Hypertension is present in the patients. 1 - Patient has hypertension. 0 - Patient has no hypertension
5. heart_disease	Integer	Categorical (Nominal)	Patients' heart disease status

			1 - has heart disease 0 - has no heart disease
6. ever_married	String	Categorical (Nominal)	Marriage status of patients Yes - Married No - Not Married
7. work_type	String	Categorical (Nominal)	Patients' types of work (children, Govt_job, Never_worked, Private, Self-employed)
8. residence_type	String	Categorical (Nominal)	Housing type of patients (Urban, Rural)
9. avg_glucose_level	Float	Numerical (Continuous)	Value of the average glucose level of patients
10. bmi	Float	Numerical (Continuous)	Body mass index of patients
11. smoking_status	String	Categorical (Nominal)	Smoking status of patients (never smoked, formerly smoked , smokes , unknown)
12. stroke	Integer	Categorical (Nominal)	History of stroke 0 - Does not had stroke 1 - Had stroke

Table 1: Description of dataset

5.0 Cleaning Data

5.1 Data Preprocessing

Data preparation is crucial for ensuring that missing values and outliers in the dataset are dealt with, as well as dataset balance, in order to deliver better and more reliable results from the subsequent phases of model creation. Everything that affects and interrupts a model's development and causes it to work inefficiently should be addressed during this phase.A data cleaning technique is used to ensure that the dataset is clean and suitable for model building once the necessary dataset has been acquired. Since the column "id" has no bearing on the result of "stroke", our target variable, it is disregarded and eliminated from the dataset.

Handling Missing Values

Following that, the dataset's null and missing values are examined and identified as shown in Figure 2 below. It is shown that there are 201 null values in the ‘bmi’ attribute.

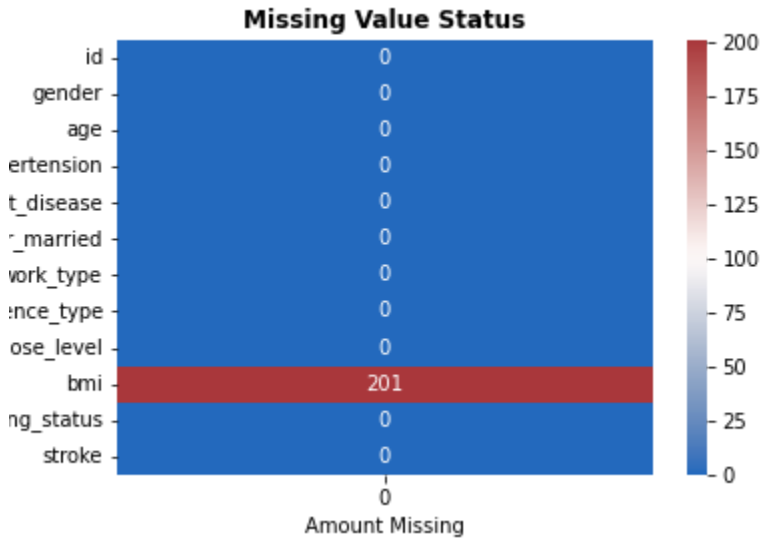


Figure 2: Missing value in the dataset

Any null or missing values found in the dataset is replaced with the mean value. Hence, the null value in ‘bmi’ is being replaced by its mean value. After replacing the null value, Figure 3 below shows that there is no more null value in the dataset.

Figure 3: Missing value in dataset after cleaning the data

Handling Outliers

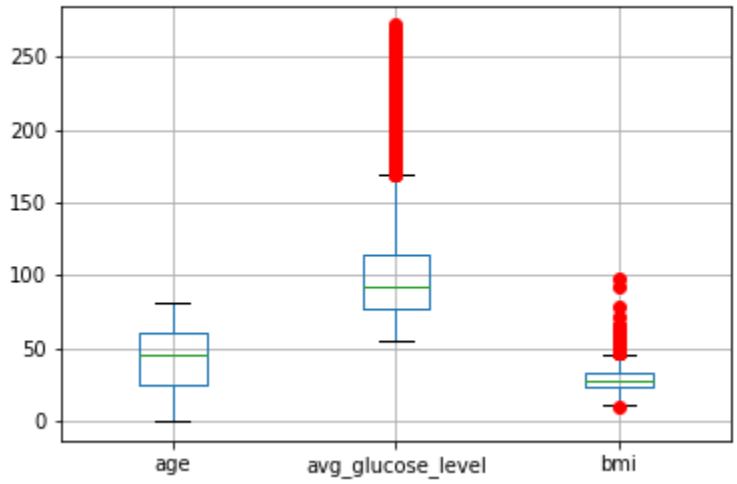


Figure 4: Box and whisker plot of the numerical attributes with outliers

The variables "avg_glucose_level" and "bmi," which are shown by red circles above and below the highest and lowest ranges of the variables, seemed to have a significant number of outliers, as illustrated in Figure 4. Outliers in the variables "avg_glucose_level" and "bmi" will be addressed by inserting the variable's median.

Finding the lowest and highest ranges and substituting out-of-range values with the median of that data column are done using the interquartile range (IQR) approach.

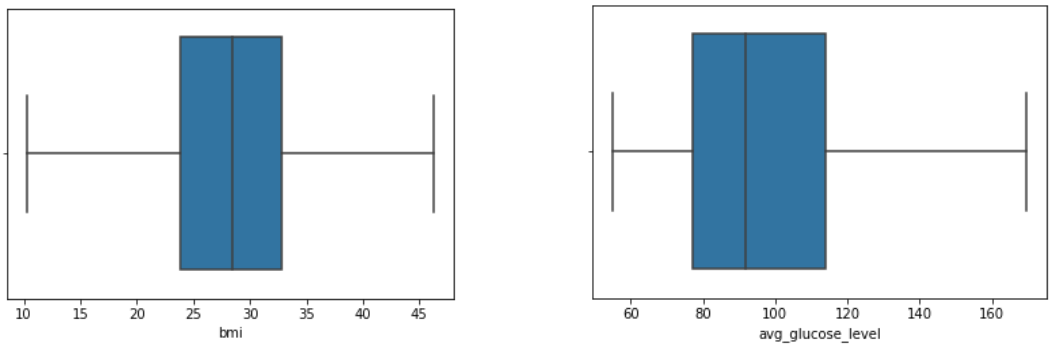


Figure 5: Box and whisker plot of ‘bmi’ and ‘avg_glucose_level’ after replacing outliers value with median

Clearing ‘Other’ category in ‘gender’ variable

Furthermore, the attribute 'gender' has been discovered to contain 1 entry of ‘other’ and is being eliminated.

5.2 Label Encoding

After the dataset has been cleaned, the Data Transformation process begins, which includes label encoding and dataset balancing.

```
gender        : {'Female': 0, 'Male': 1}
ever_married  : {'No': 0, 'Yes': 1}
work_type     : {'Govt_job': 0, 'Never_worked': 1, 'Private': 2, 'Self-employed': 3, 'children': 4}
residence_type: {'Rural': 0, 'Urban': 1}
smoking_status: {'Unknown': 0, 'formerly smoked': 1, 'never smoked': 2, 'smokes': 3}
```

Figure 6: Label Encoding

In order for the machine to recognise and comprehend categorical attributes with string data types, label encoding transforms them into integer values as shown in Figure 6 above. The conversion of attributes with string data types into integers is necessary since machines are often trained on numerical data. Five columns of attributes from the dataset are found to be in the data type of strings, namely 'gender,' 'ever_married,' 'work_type,' 'Residence_type,' and ‘smoking_status'. All strings were encoded to integer values through the process of label encoding, producing a workable dataset.

6.0 Data Visualisation and Statistical Analysis

After cleaning the data, the data is being visualised and analysed with statistical methods in this section.

6.1 Data Visualisation

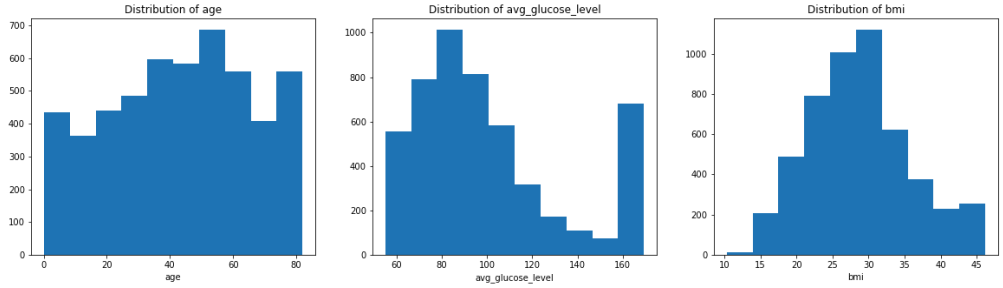


Figure 7: Distribution of numerical attributes

The distribution plots above plotted with numerical attributes can be used to visualise the skewness of the data after it is being processed. From these plots, it can be seen that there exists skewness in the data. The distribution of 'age' is slightly negatively skewed and the 'avg_glucose_level' and 'bmi' is positively skewed . This skewness is further proved by calculating its skewness.

```
age                 -0.137430
avg_glucose_level    0.935216
bmi                  0.431847
dtype: float64
```

Figure 8: Skewness of numerical attributes

Negative values of skewness indicate that the data is skewed to the left while positive value of skewness indicates that the data is skewed to the right. Based on Figure 8 above, we can prove that, 'age' is skewed to the left while 'avg_glucose_level' and 'bmi' is skewed to the right.

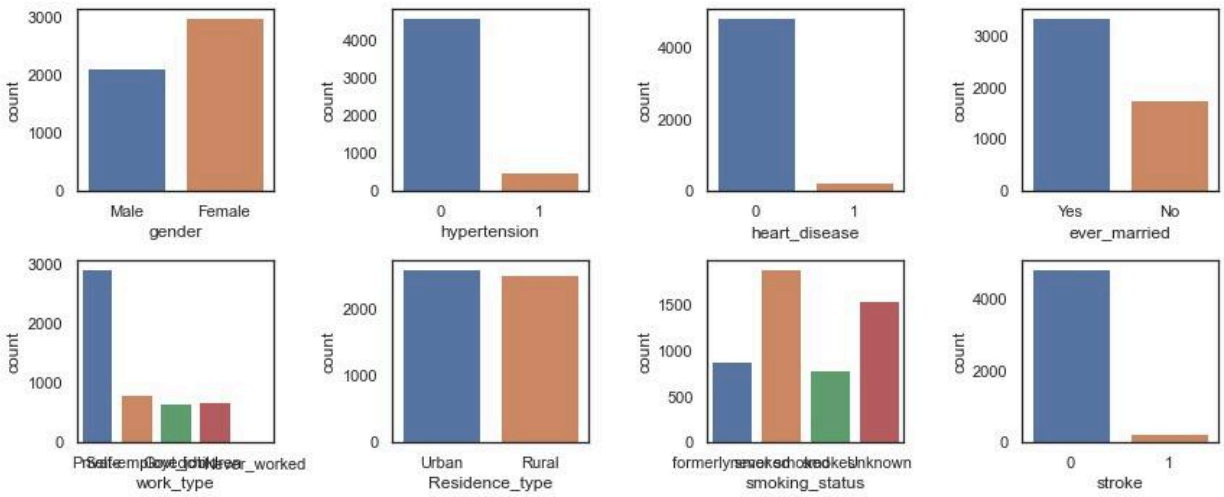


Figure 9: Distribution of categorical attributes

All the categorical data are plotted in a bar chart as shown in Figure 9 above except for ‘id’ as it was removed as stated in Section 3.1, Data Preprocessing. Based on these bar charts, it is shown that, except for gender, work_type, and smoking_status, the majority of the categorical variables are binary. The count of each categorical variable will also be compared in Section 7.0, EDA.

In addition, it is clear that most variables, including "stroke," exhibit some degree of class imbalance. A class imbalance would have an impact on the model's accuracy since it would highlight the dominant class while inaccurately portraying the minority class. Hence, it requires to be handled in Section 8.0, Class Imbalance.

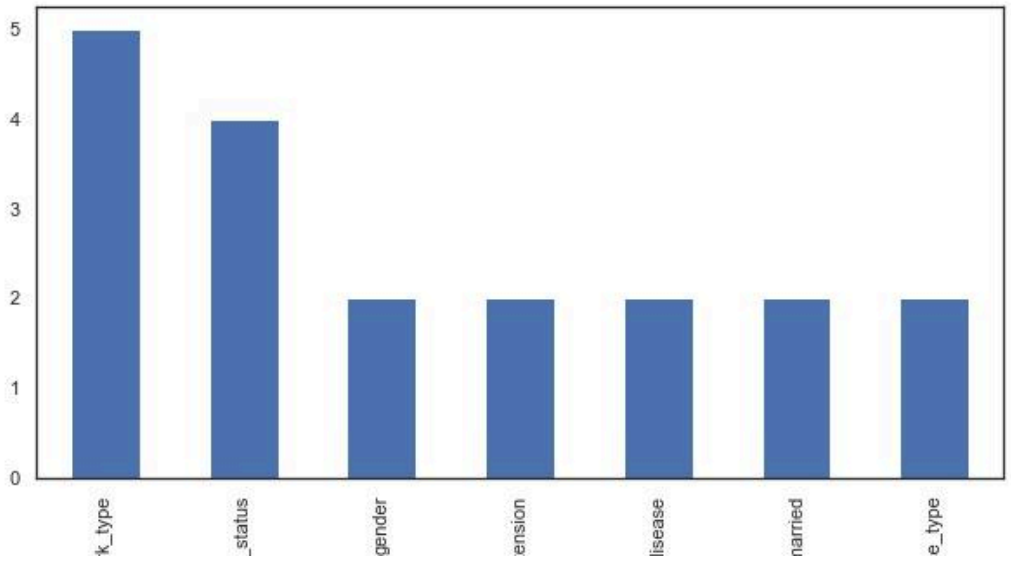


Figure 10: Number of categories in each categorical attribute

The plot above tells the number of categories each categorical variable has. Below shows a table of the total number of categories in each categorical variable.

Categorical Attribute	Number of Categories
work_type	5
smoking_status	4
gender	2
hypertension	2
heart_disease	2
ever_married	2
residence_type	2

Table 2: Number of categories in each categorical attribute

6.2 Statistical Analysis

The Table 3 below is a statistical breakdown of the numerical attributes measured, including the mean, first quartile, median, third quartile, minimum, and maximum values.

	age	avg_glucose_level	bmi
count	5109.000000	5109.000000	5109.000000
mean	43.229986	100.980861	28.720405
std	22.613575	33.198179	7.114056
min	0.080000	55.120000	10.310000
25%	25.000000	77.240000	23.800000
50%	45.000000	91.880000	28.400000
75%	61.000000	114.090000	32.800000
max	82.000000	169.300000	46.200000

Table 3: Statistical breakdown of numerical variable

The mean of age in the dataset reveals a 43-year-old average age. Moving on, the average blood sugar level of the dataset is 101, which is considered to be normal. The average BMI of the dataset is 28.72, which denotes overweight.

```
2    1892
0    1544
1     884
3     789
Name: smoking_status, dtype: int64
```

```
2    2924
3     819
4     687
0     657
1      22
Name: work_type, dtype: int64
```

```
0    2994
1    2115
Name: gender, dtype: int64
```

```
0    4833
1     276
Name: heart_disease, dtype: int64
```

```
1    2596
0    2513
Name: residence_type, dtype: int64
```

```
0    4611
1     498
Name: hypertension, dtype: int64
```

```
1    3353
0    1756
Name: ever_married, dtype: int64
```

Figure 11: Frequency table of categorical variable

Figure 11 above shows the frequency table of each categorical variable. It is summarised in the table below.

Categorical Attribute	Description	Count
gender	0: Female	2994
	1: Male	2115
work_type	0: Govt_job	657
	1: Never_worked	22
	2: Private	2924
	3: Self_employed	819
	4: Children	687
ever_married	0: No	1756
	1: Yes	3353
residence_type	0: Rural	2513
	1: Urban	2596
smoking_status	0: Unknown	1544

	1: Formerly smoked	884
	2: Never smoked	1892
	3: Smokes	789
hypertension	0: No	4611
	1: Yes	498
heart_disease	0: No	4833
	1: Yes	276

Table 4: Counts in each categories of categorical attributes

On the other hand, based on Figure 12, the mean age of each gender in the data is shown as below.

```
gender
0    43.757395
1    42.483385
Name: age, dtype: float64
```

Figure 12: Mean age of genders in the dataset

It is shown that the mean age of female in this dataset is 43.7 while the mean age of male in this dataset is 42.5

Furthermore, based on Figure 13, the mean age of individuals getting stroke in the dataset is shown as below.

```
stroke
0    41.974831
1    67.728193
Name: age, dtype: float64
```

Figure 13: Mean age of stroke in the dataset

It can be seen that the average age of getting a stroke is 68.

7.0 Exploratory Data Analysis (EDA)

EDA was completed after resampling, and features were selected based on the Pearson's Correlation heatmap and the Chi Square Test for categorical variables. An appropriate response should be given to every independent variable in the correlation heatmap that significantly correlates with another independent variable. Having a value > 0.5 indicates a strong correlation. When two or more independent variables exhibit strong intercorrelations, this is referred to as multicollinearity (Scott, n.d.). With this being said, multicollinearity may lead to wider confidence intervals, which can therefore provide probabilities that are less reliable when it comes to the effect of independent variables in a model. Hence, it is suggested to remove one of the variables involved in the multicollinearity.

Furthermore, p-values are also computed using Pearson's Correlation and Chi Square test. A p-value, also known as a probability value, is a numerical indicator of the likelihood that the null hypothesis is accepted (McLeod, n.d.). The null hypothesis may be rejected when the p-value is low (e.g., 0.05).

7.1 Bivariate Analysis

In this section, we will be analysing the categorical and numerical variables in the dataset with the target variable, 'stroke'. Figure 14 below shows the analysis between 'stroke' and each of the numerical variables in the dataset.

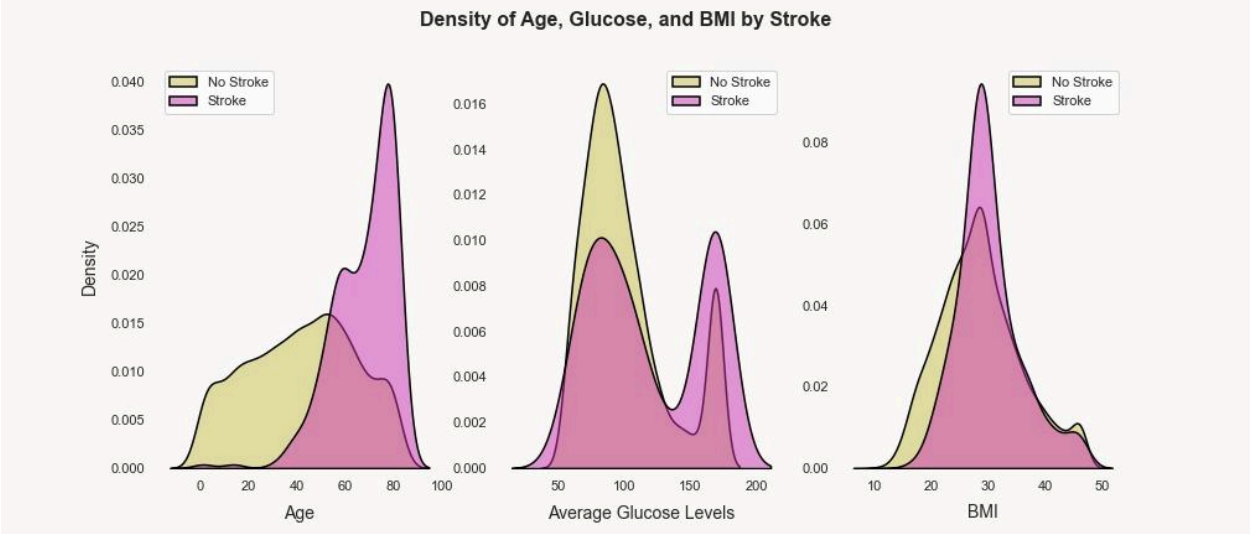


Figure 14: Analysis Between Stroke and Numerical Attributes

Stroke and Age

From Figure 14, it can be seen that the density of individuals who age above 50 suffered stroke more.

Stroke and Average Glucose Level

From Figure 14, it shows that the density of individuals who had glucose levels of less than 100 and more than 150 suffered stroke more.

Stroke and BMI

From Figure 14, it shows that overweight individuals have the highest density in suffering a stroke.

Stroke and Hypertension

```
A person with hypertension has a probability of 13.25 % get a stroke
A person without hypertension has a probability of  3.97 % get a stroke
```

Figure 15: Analysis between stroke and hypertension

As we have seen in Figure 15, stroke probability for those who have hypertension are quite different than for those who don't. Their percentage is 13.25% and 3.97% respectively. It means that individuals with hypertension are almost 3.3 times more likely to get stroke than the ones who don't have hypertension.

Stroke and Gender

```
A female person has a probability of 4.71 % get a stroke
A male person has a probability of 5.11 % get a stroke
```

Figure 16: Analysis between stroke and gender

Males are more likely to have a stroke than females, although the difference is rather minor.

Stroke and Heart Disease

```
A person with heart disease has a probability of 17.03 % get a stroke
A person without heart disease has a probability of 4.18 % get a stroke
```

Figure 17: Analysis between stroke and heart disease

As we've seen in Figure 17 above, individuals who have heart disease have a much higher risk of stroke than those without it. While those without heart disease have a 4.18% chance of having a stroke, those with heart disease have a 17.03% chance of experiencing a stroke. It indicates that those who have heart disease are 4.07 times more likely to get a stroke than those who do not.

Stroke and Married Status

```
A person married (or married before) has a probability of 6.56 % get a stroke

A person never married has a probability of 1.65 % get a stroke
```

Figure 18: Analysis between stroke and married status

Based on Figure 18, stroke risk differs significantly between those with and without a history of marriage. A person with a history of marriage has a 6.56% chance of having a stroke, whereas a person without a history of marriage has a 1.65% chance. It indicates that a person who is married (or has been married in the past) is 5.7 times more likely to get a stroke than someone who has never been married.

Stroke and Work Type

```
A person with private work type has a probability of 5.1 % get a stroke

Self-employed person has a probability of 7.94 % get a stroke

A person with a goverment job has a probability of 5.02 % get a stroke

A child has a probability of 0.29 % get a stroke

A person never worked has a probability of 0.0 % get a stroke
```

Figure 19: Analysis between stroke and type of work

Compared to other job types, self-employment increases the risk of stroke. Individuals who work in the public sector and the private sector approximately have an equal risk of stroke.

Stroke and Residence Type

```
A person, who lives in urban area, has a probability of 5.2 %  get a stroke

A person, who lives in rural area, has a probability of 4.54 % get a stroke
```

Figure 20: Analysis between stroke and residence type

As can be observed in Figure 20, there is little variation in an individual's residence types. An individual who lives in a rural region has a slightly higher chance of getting a stroke than someone who lives in an urban area. The difference is negligible, however.

Stroke and Smoking Status

```
A formerly smoked person has a probability of 7.92 % get a stroke

A person never smoked has a probability of 4.76 % get a stroke

A person smokes has a probability of 5.32 % get a stroke

A person whom smoking history is not known,has a probability of 3.04 % get a stroke
```

Figure 21: Analysis between stroke and smoking status

Depending on whether a person smokes, there are variations. An individual who has smoked in the past has a 1.66 times higher risk of having a stroke than someone who has never smoked. A smoker has a 1.11 times greater chance of having a stroke than someone who has never smoked. However, there is hardly any difference between smokers and non-smokers in terms of stroke risk.

7.2 Pearson's Correlation

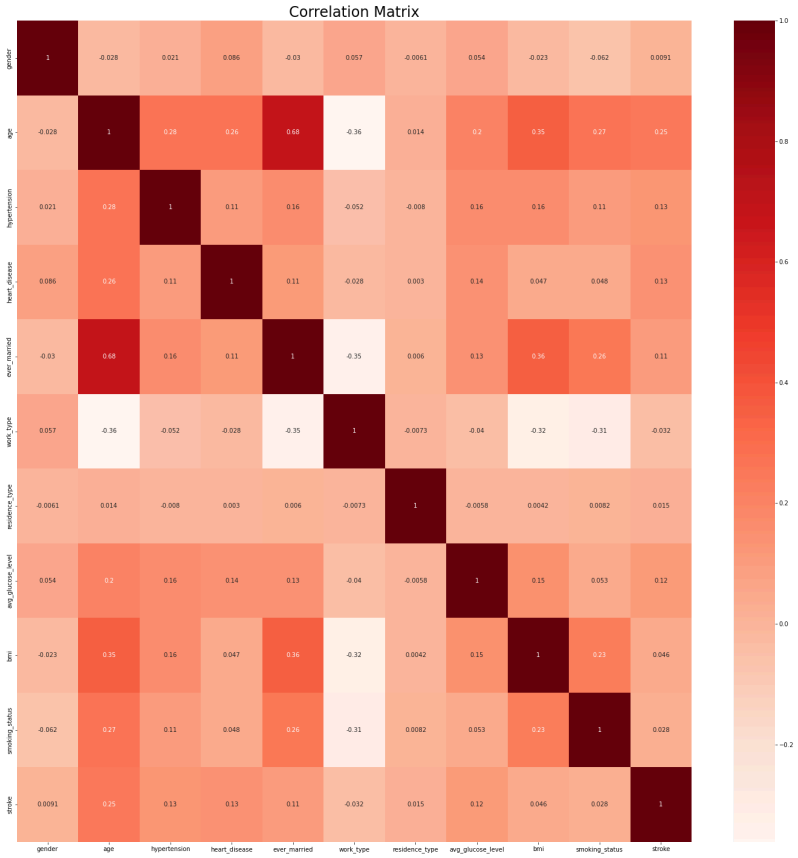


Figure 22: Pearson's Correlation Matrix

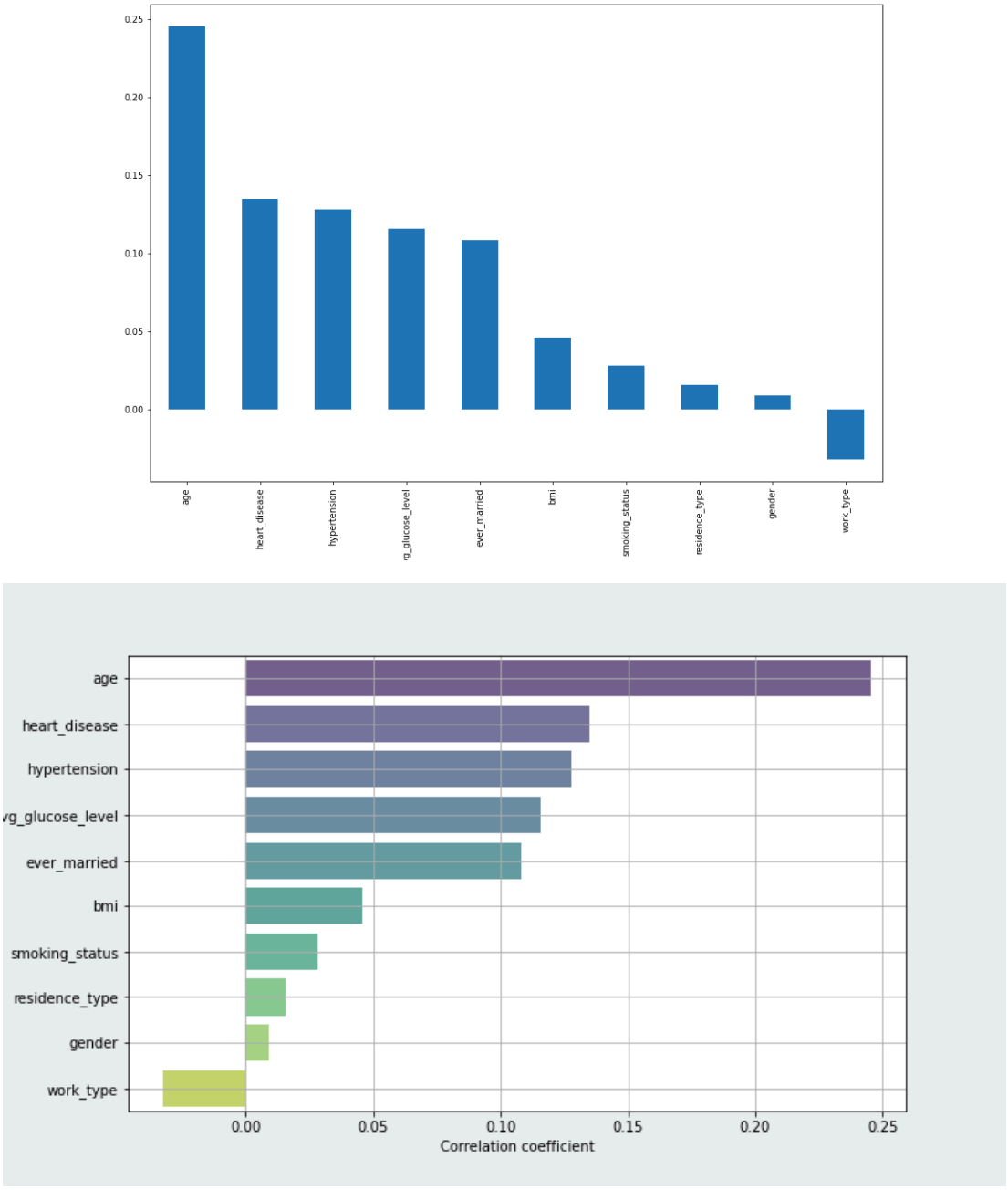


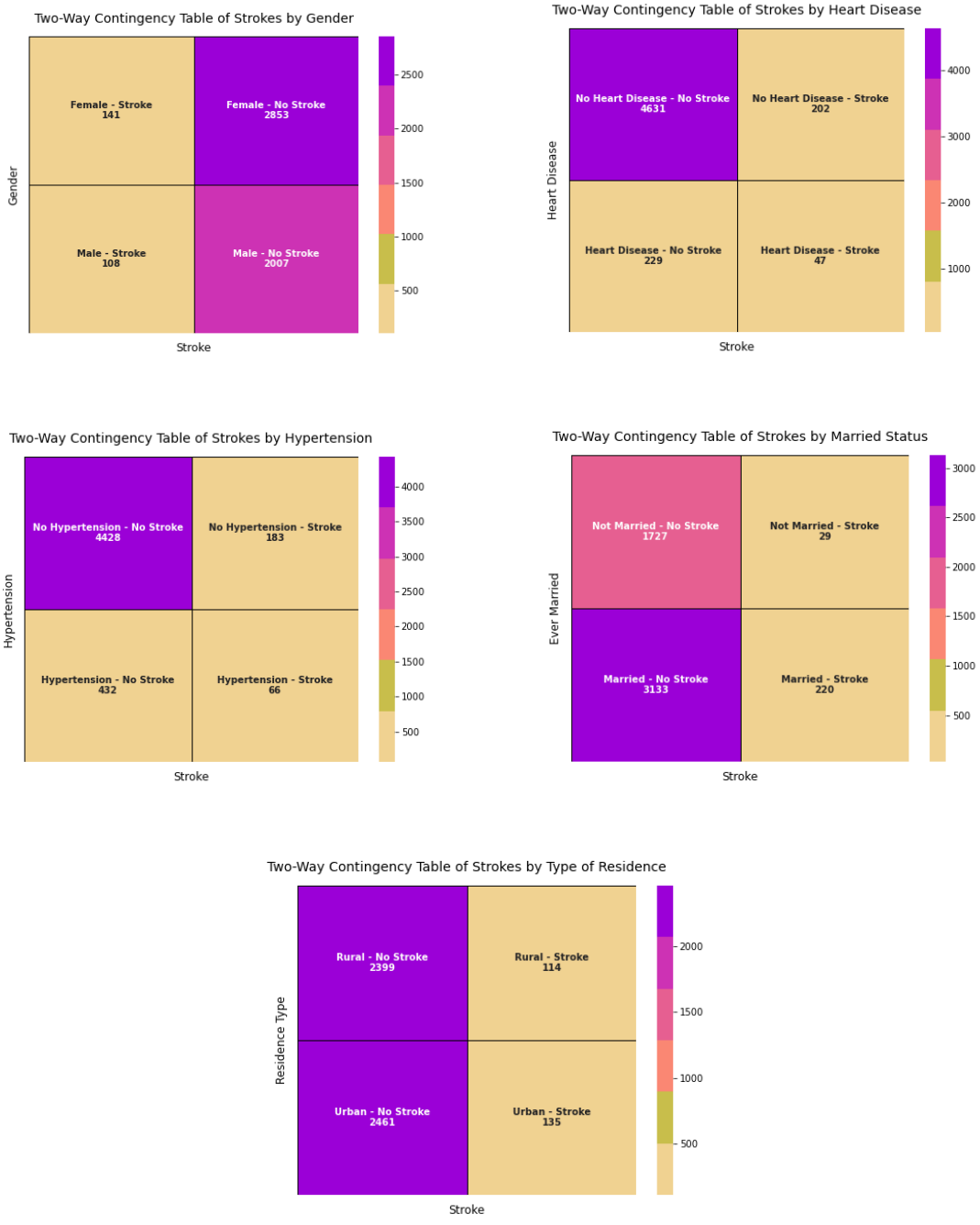
Figure 23: Correlation Scores of each variable compared to the target variable, ‘stroke’

As can be seen on Figures 22 and 23 above. With a score of 0.25, the attribute “age” is considered to be highly correlated to target variable “stroke’. The variables "hypertension," "heart_disease," "ever_married," and "avg_glucose_level" are also thought to be associated factors to the target variable, "stroke," according to Figures 22 and 23.

Other than that, in Figure 22, it shows that variables ‘age’ and ‘ever_married’ have multicollinearity. Due to its lower correlation with the target variable, ‘stroke’, when compared to the formal variable, ‘age’, and its inability to predict whether a person would get a stroke or not. The latter variable, ‘ever_married’ is eliminated.

7.3 Chi-square Test

In Figure 24 below, it shows the two way contingency table of categorical variables and the target variable, ‘stroke’.



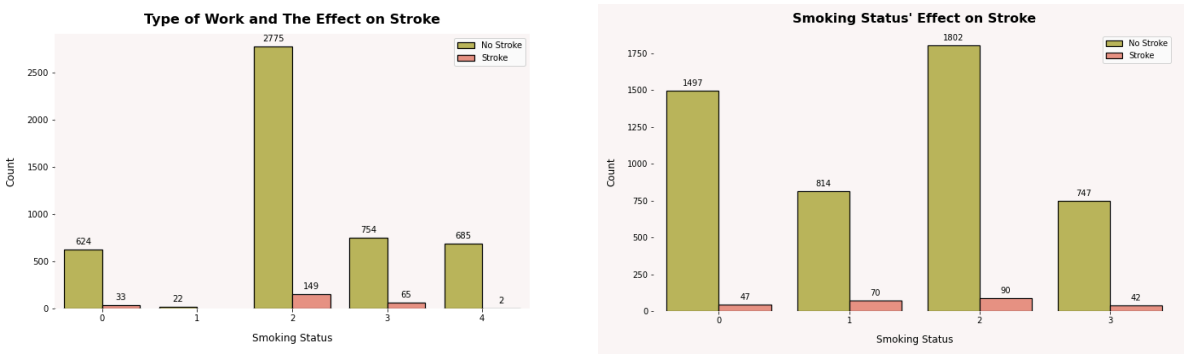


Figure 24: Contingency Table for Chi-square Test Analysis between Categorical and Target Variables

Target Variable		Categorical Variable	Count	Percentage
Stroke	1: Stroke	**Gender**		
		Male	108	2.11%
		Female	141	2.76%
	0: No Stroke	Male	2007	39.28%
		Female	2853	55.84%
Stroke	1: Stroke	**Heart Disease**		
		Has heart disease	47	0.92%
		No heart disease	202	3.95%
	0: No Stroke	Has heart disease	229	4.48%
		No heart disease	4361	85.36%
Stroke	1: Stroke	**Hypertension**		
		Has hypertension	66	1.29%
		No hypertension	183	3.58%

	0: No Stroke	Has hypertension	432	8.46%
		No hypertension	4428	86.67%
Stroke	1: Stroke	**Married Status**		
		Has married	220	4.31%
		Not married	29	0.57%
	0: No Stroke	Has married	3133	61.32%
		Not married	1727	33.80%
Stroke	1: Stroke	**Residence Type**		
		Urban	135	2.64%
		Rural	114	2.23%
	0: No Stroke	Urban	2461	48.17%
		Rural	2399	46.96%
Stroke	1: Stroke	**Work Type**		
		Government Job	33	0.65%
		Never Worked	0	0%
		Private	349	6.83%
		Self Employed	65	1.27%
		Children	2	0.04%
	0: No Stroke	Government Job	624	12.2%
		Never Worked	22	0.43%

		Private	2775	54.32%
		Self Employed	754	14.76%
		Children	685	13.4%
Stroke	1: Stroke	**Smoking Status**		
		Unknown	47	0.92%
		Formerly Smoked	70	1.39%
		Never Smoked	90	1.76%
		Smokes	42	0.82%
	0: No Stroke	Unknown	1497	29.30%
		Formerly Smoked	814	15.93%
		Never Smoked	1802	35.27%
		Smokes	747	14.62%

Table 5: Distribution of categorical attributes

Based on Table 5 above, we are able to analyse the distribution of the categorical attributes in the dataset.

Gender and Stroke

We realise that the dataset contains 2.76% of female individuals who have a stroke and 2.11% of male individuals who have stroke. There are more female individuals who have stroke compared to male individuals in this dataset.

Heart disease and Stroke

In the dataset, the count percentage of an individual with heart disease and stroke is 0.92% while the count percentage of an individual without heart disease but with stroke is 3.95%. It contains a significant 3.03% difference.

Hypertension and Stroke

Based on Table 5, we realised that the percentage count of individuals with hypertension and stroke is 1.29% while individuals without hypertension but has stroke is 3.58% in this dataset. The percentage difference is approximately 2 times.

Married Status and Stroke

In this relationship, the category that has the largest data is individuals who have ever married and have no stroke. It contains a percentage of 61.32% of the dataset in translation to 3133 counts.

Residence Type and Stroke

In this dataset, the percentage count difference between individuals with stroke who stayed at urban and rural is significantly small as they have a very similar percentage. Similarly, this applies to individuals without stroke who stayed in urban and rural areas as their percentage count difference in the dataset is significantly similar.

Work Type and Stroke

Based on Table 5 above, we could observe that the category that contains the largest count within this relationship in this dataset is individuals without stroke who worked in the private sector as it occupies 54.32% of the dataset. In this dataset, there appears a category that has 0% which is individuals who have never worked and has stroke.

Smoking Status and Stroke

The percentage count of individuals who smoke and have stroke in this dataset is 0.82% which is almost similar to individuals with unknown smoking status and have stroke with a percentage

count of 0.92%. The largest percentage count within this relationship is individuals who have never smoked and have no stroke with a percentage count of 35.27%.

Table 6 below contains the Chi-Square Test findings from the contingency tables shown in Figure 24.

	target	cat_feature	chi2	p-value	rounded p	alpha	H0	relation
0	stroke	gender	0.246877	6.192826e-01	0.619	0.05	Fail to reject	Independent
1	stroke	hypertension	75.418319	3.808401e-18	0.000	0.05	Rejected	Dependent
2	stroke	heart_disease	87.957296	6.688296e-21	0.000	0.05	Rejected	Dependent
3	stroke	ever_married	20.595671	5.672425e-06	0.000	0.05	Rejected	Dependent
4	stroke	work_type	2.927120	8.710275e-02	0.087	0.05	Fail to reject	Independent
5	stroke	residence_type	0.597112	4.396819e-01	0.440	0.05	Fail to reject	Independent
6	stroke	smoking_status	3.365543	6.657364e-02	0.067	0.05	Fail to reject	Independent
7	stroke	stroke	4860.000000	0.000000e+00	0.000	0.05	Rejected	Dependent

Table 6: Results of p-values of each categorical variable in Chi-square test

From the Chi Square Test, we obtained the ‘chi square value’, the ‘p-value’ and ‘rounded p value’ of the target variable and each categorical attribute. The corresponding p-values were calculated as shown above, and it was discovered that 'work_type', ‘gender’, ‘smoking_status’ and ‘residence_type' had no relevance in predicting stroke since the p-value are more than 0.05 and it fails to reject the null hypothesis as it is independent from the target variable, ‘stroke’.

7.4 Feature Selection

However, we decided not to reject ‘smoking_status’. This is because the p-value of the correlation shown in Pearson's correlation between ‘smoking_status’ and ‘stroke’ is < 0.05. This is shown in Figure 25 below.

```
Correlation between smoking_status and stroke is 0.028107558141114593
P-value of this correlation is 0.044541717547538245
```

Figure 25: Correlation coefficient and p-value between ‘smoking_status’ and ‘stroke’

Furthermore, ‘work_type’ is not rejected although it has a p-value of > 0.05 as shown in the Chi-square test in Table 6 above. However, according to Pearson's correlation analysis, the p-value of the correlation is as close as 0 as shown in Figure 26 below.

```
Correlation between work_type and stroke is -0.032323159385199966
P-value of this correlation is 0.020865493997985547
```

Figure 26: Correlation coefficient and p-value between ‘age’ and ‘stroke’

Therefore, based on the result of the Chi-square test, the variables ‘residence_type’ and ‘gender’ are rejected.

In conclusion, the final attributes chosen for the model development after feature selection includes “heart disease’,‘avg_glucose_level’,‘work_type’,‘age’,‘smoking_status’,’bmi’ and ‘hypertension’. The training set has a total of seven attributes.

8.0 Class Imbalance

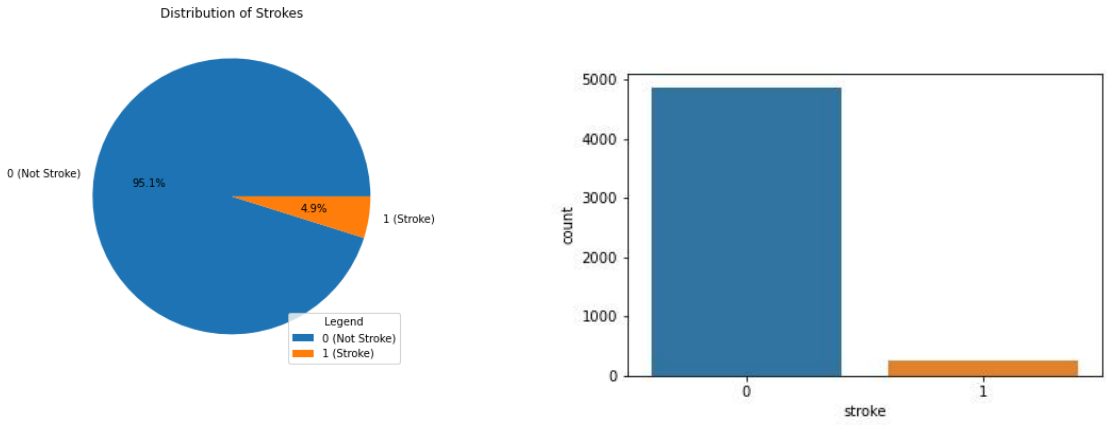


Figure 27: Distribution of stroke before balancing

The obtained dataset, which has 5110 rows, is very imbalanced since only 249 rows or 95.1% of the patient records contain a stroke, whereas 4861 rows or 4.9% have no entries that indicate a stroke. As seen on Figure 27, there exists a large class imbalance in our target variable ‘stroke’, where only 4.9% of our dataset indicates a positive under ‘stroke’. With such unbalanced data, it is not possible to train and test a machine learning model since the outcomes would be erroneous and ineffective predictions with poor accuracy and recall scores.

Because sampling methods might result in an increase or reduction in the dataset, this could have a significant impact on how accurate a model is. Undersampling and oversampling are the two main approaches for resampling a dataset to improve the class's balance. To deal with this issue, we did both undersampling, Random Under Sampling (RUS) and oversampling, Synthetic Minority Oversampling Technique (SMOTE) to reduce the imbalance.

In this paper, we will be using both RUS and SMOTE to compare the accuracy results of each machine learning model. The explanation of both sampling techniques are explained as below:

8.1 RUS

In order to implement Random Under-Sampling (RUS), a certain number of samples from the majority class are randomly removed from the training dataset. This results in a smaller proportion of training data instances belonging to the majority class in the modified dataset. This can be done repeatedly until the desired class distribution is reached, such as when there are the same amount of samples in each class. This approach comes in handy when there is a class imbalance even though there is enough number of samples in the minority class.

```
print('Original dataset shape:\n', y.value_counts())
print('Undersampled dataset shape:\n', y_rus.value_counts())
```

```
Original dataset shape:
 0    4860
1     249
Name: stroke, dtype: int64
Undersampled dataset shape:
 0    249
1    249
```

Figure 28: Original dataset and undersampled dataset

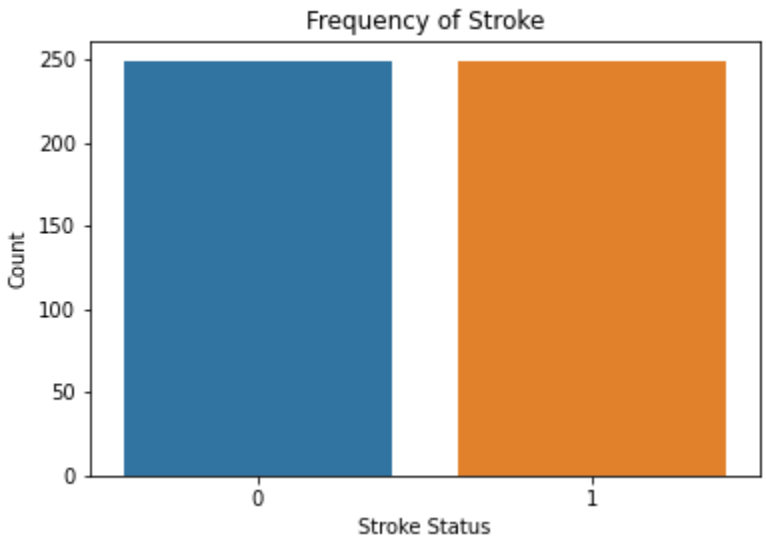


Figure 29: Histogram of the undersampled data

Based on the figures above, as you can see, the majority class(have no stroke) and the minority class(have stroke) are imbalanced. The majority class recorded a number of 4840 whereas the

minority class only recorded a number of 249. Therefore, RUS is used here to downsize the majority class to ensure both the classes are balanced.

8.2 SMOTE

Synthetic Minority Oversampling Technique (SMOTE), In the case of datasets with a class imbalance but a significant number of examples in the minority class, this method may be preferable. This is how the technique normally works, it first finds the dissimilarity between a sample and its nearest neighbour. Then, it will multiply a random number ranging from 0 to 1. Moving on, it incorporates this difference into the sample to produce a new synthetic example in the feature space. Finally, it will continue with the next number up to the user-defined number. SMOTE increases the number of cases in our dataset in a balanced way.

```
print('Original dataset shape:\n', y_smote.value_counts())
print('Oversampled dataset shape:\n', y_oversample.value_counts())
```

```
Original dataset shape:
 0    4860
1     249
Name: stroke, dtype: int64
Oversampled dataset shape:
 1    4860
0    4860
```

Figure 30: Original dataset and oversampled dataset

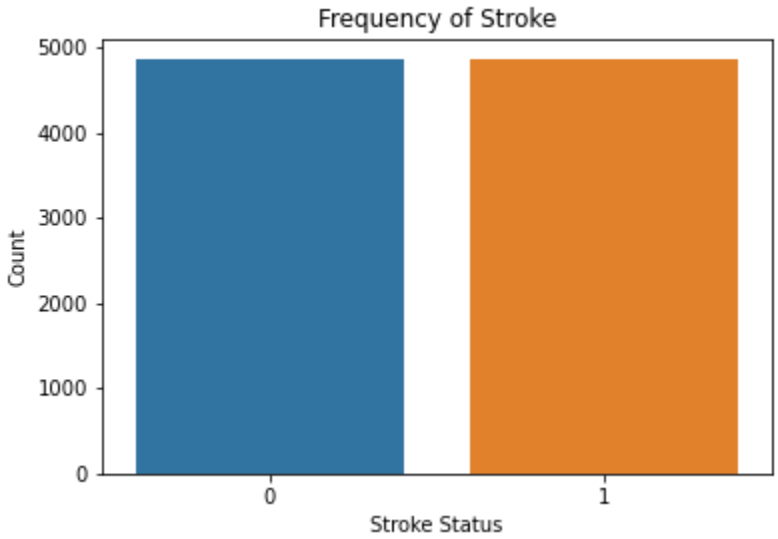


Figure 31: Histogram of the oversampled data using SMOTE

As for SMOTE, instead of downsizing the majority class, we upsized the minority class following the steps explained above. As you can see, originally, there was a total of 4860 in the majority class (no stroke) and only 249 in the minority class (stroke). Therefore, we performed SMOTE to upsize the minority class so that both the classes are balanced.

9.0 Model Development

Data that has been undersampled as well as oversampled is partitioned into training and testing with the ratio 80:20 respectively. We also used machine learning models such as Decision Tree Classifier (DTC), Random Forest Classifier (RFC), Naive Bayes Classifier (NBC), K-Nearest Neighbours (KNN), and Logistic Regression (LR) to help fit the data. We decided to lean more towards classification algorithms as the dataset's target variable, stroke, is categorical variable. Hence, our decision to use supervised classification ML algorithms.We decided to use 4 different types of train and model splits. Which is, RUS without feature selection, RUS with feature selection, SMOTE without feature selection, and finally SMOTE with feature selection.

9.1 Proposed Model Development

Random Forest Classifier (RFC)

The Random Forest Classifier, often known as RFC, is an example of an ensemble methods to machine learning. This is due to the fact that RFC is made up of many decision trees, each of which is trained independently using a random sample of data. It is a simple algorithm that can handle complex issues. Each decision tree casts a vote for one of the output classes while the trees are being trained. In comparison to any individual tree in the model, the combined predictions from the trees provide superior outcomes. Additionally, RFC chooses features and automatically handles missing data without changing hyperparameters.

K-Nearest Neighbour

For supervised classification tasks like the one we're working on, K-Nearest Neighbors (KNN) is a method that may be employed. It is a slow learner since it doesn't use training data to make new discoveries; instead, it maintains it and uses it to make classifications. Before allocating data items to a class, it compares their qualities to those of their nearest neighbours and the most similar measure. As a consequence, finding the right k parameter to get the best result is crucial. The best k parameter between 1 and 20 was determined through manual cross validation.

Decision Tree Classifier

The Decision Tree Classifier (DTC) is a supervised machine learning method with a tree topology. A decision tree is simple to build: it hierarchically separates data into subsets, which are then further split into smaller divisions or branches until they become "pure," which indicates that the features within the branch all belong to the same class. These groupings are referred to as "leaves."

Naive Bayes Classifier

The Naive Bayes Classifier is a probabilistic classifier that is based on Bayes' theorem. It is predicated on the idea that the existence of one feature in a class is unrelated to the presence of another feature in the same class.It is the possibility of an event happening given that another event has already occurred. For example, from a deck of 52 cards, two cards are pulled without replacement. Given that the first card drawn was also an ace, what is the likelihood that the second card will be an ace?

Logistic Regression

By computing the likelihood of each member of the set, Logistic Regression is used to classify items of a set into two categories (binary classification). We first establish a probability threshold. If the likelihood of a given element exceeds the probability threshold, we categorise that element in one of two groups. This is especially helpful in our case as we are faced with a binary classification which is predicting the event of a stroke.

9.2 Performance Metrics

K-Fold Validation

In order to measure the K-Fold Validation, we must first gauge how closely the model's predictions match the actual data in order to assess a model's performance on a dataset. The mean squared error (MSE) is the most typical metric for measuring this. The MSE will be less the more closely the model predictions match the observations. The use of only one testing set has the drawback that the MSE might vary greatly depending on which data were used in the training and testing sets. One method of avoiding this issue is to fit a model many times using various training and testing sets, then compute the MSE as the average of all the MSEs. We can prevent this problem by doing this.

The formula for MSE is the following.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where:

- y_i is the ith observed value.
- ŷi is the corresponding predicted value.
- n = the number of observations

To calculate the MSE, multiply the observed value by the predicted value and square the difference. Repeat for each observation. Then add all of the squared values together and divide by the number of observations.

K-fold cross-validation is a particular variation of the broad technique known as cross-validation. A dataset is initially randomly divided into k groups, or "folds," of approximately similar size. Next, decide which fold will be the holdout set. On the remaining

k-1 folds, fit the model. On the basis of the observations in the held-out fold, calculate the test MSE. Finally, perform this procedure k times, with a different set serving as the holdout set each time. The average of the k test MSEs is then calculated as the overall test MSE. In general, the test MSE's bias decreases as we increase the number of folds used in k-fold cross-validation, but the variance increases. In practice, we usually pick between five and ten folds and in this case, the parameter k is set to a factor of 5. This number of folds has been demonstrated to yield a reliable estimate of test MS and an appropriate balance between bias and variance.

Accuracy, Precision, Recall, ROC Curve & AUC

We must differentiate between two groups, stroke and non-stroke, with the non-stroke category including the majority of the data points. For instance, suppose the negative classes were far more abundant, as is typical in many datasets. Problems of this kind are examples of the extremely common circumstance in data science when accuracy is not a reliable sign of model efficacy. One may evaluate the effectiveness of the models using the performance metrics listed below.

Accuracy

The most basic and clear performance statistic is accuracy, which is calculated as the proportion of successfully predicted observations to all observations. One can believe that if a model achieves high accuracy, the model is reliable. Although accuracy is a useful metric, this metric can only be trusted when the values of false positives and false negatives are almost equal. Therefore, accuracy alone is not enough and other metrics have to be considered to evaluate the performance of the models.

Recall

We know intuitively that the stroke prediction problem isn't helped by categorising all data points as negative (non-stroke), and that we should instead concentrate on finding positive occurrences. Statistics refers to recall as a model's ability to discover all relevant instances within a data set. The total number of true positives divided by the total number of true positives plus false negatives yields recall.

False negatives are data points are data points that the model classifies as negative when they are really positive, while true positives are data points that the model labels as positive when they are actually positive (meaning they are correct). In the case of stroke predictions, true positives are people who were correctly identified as to have a stroke, and false negatives are people who the model incorrectly labels as not to have a stroke, but who were to have a stroke. In other words, recall can be defined as a model's capacity to locate every data point belonging to the class that interests us in a data collection.

Precision

The amount of true positives divided by the total of true positives and false positives is used to calculate precision. Precision is the percentage of data points in the relevant class that our model actually predicted were present. Recall is the capacity to find every relevant instance of a class in a data collection, whereas accuracy measures the proportion of those data points.

If we categorise everyone as having had a stroke, our recall increases to 1.0, indicating a perfect classifier. The measures we decide to maximise involve trade-offs, as is the case with most data science topics. When recall and accuracy are concerned, increasing recollection results in a loss in precision, and vice versa.

F1 Score

In some circumstances, we may be aware that maximising recall or accuracy at the expense of the other statistic is what we need. In our stroke prediction, for example, we would probably strive for a recall of close to 1.0 to identify all patients who truly have the condition, and we may tolerate a low precision in which we wrongly identify some patients as having the disease when they don't. However, we may combine accuracy and recall using the F1 score to get the best possible results. The harmonic mean of accuracy and recall is the F1 score, which accounts for both metrics. Harmonic mean is used as it penalises extreme numbers. A classifier with a simple average of 0.5 but an F1 score of 0 has a precision of 1.0 and a recall of 0.0. The F1 score, a specific illustration of the basic FX metric that can be changed to give greater weight to either recall or precision, equally weights both metrics. The F1 score is attempted to be

maximised in order to provide a classification model with the best possible recall and precision balance.

ROC Curve

We are able to graphically display recall and precision using a technique called a ROC curve. The ROC curve shows how the recall vs. accuracy relationship varies when the threshold for identifying a positive data point in our model is changed. When a data point passes the threshold, it is categorised as positive. We may establish a threshold in our stroke prediction model that, if met, will classify a patient as having the condition. Based on our selected dataset, each patient may get a score from zero to one according to the model. By altering the threshold, we try to achieve the right accuracy vs. recall balance. On the y-axis and x-axis, respectively, a ROC curve compares the proportion of true positives to false positives. The false positive rate (FPR) indicates the risk of a false alarm, while the true positive rate (TPR), is the recall rate.

Area Under the Curve (AUC)

AUC is a measurement that spans from zero to one, with a higher value indicating better classification performance, and it may be used to compute the whole ROC curve of a model. If the AUC value is 1.0, the classifier is able to discriminate between all positive and negative class points with accuracy; if the AUC value is 0 the classifier will treat and anticipate that all positive and negative class points are positive.Positive and negative class values may often be distinguished by the classifier when the AUC values fall between 0.5 and 1. Rather than False Negative (FN) or False Positive (FP) data, True Positive (TP) and True Negative (TN) statistics are found. When the AUC is equal to 0.5, the classifier cannot distinguish between positive and negative class points and will estimate a random or constant class for the whole set of data points.

9.3 Performance Analysis

	Model	Accuracy	K-Fold Validation	ROC Score	Precision	Recall	F1 Score
1	Random Forest Classifier	0.77	0.731266	0.818727	0.741379	0.843137	0.788991
2	Naive Bayes	0.77	0.703544	0.850740	0.833333	0.686275	0.752688
3	Logistic Regresstion	0.76	0.754051	0.842737	0.745455	0.803922	0.773585
4	K-Nearest Neighbours	0.75	0.721266	0.762705	0.709677	0.862745	0.778761
5	Decision Tree Classifier	0.62	0.663228	0.619448	0.622642	0.647059	0.634615

Table 7: Performance Analysis of Random Under Sampling (RUS) without Feature Selection

	Model	Accuracy	K-Fold Validation	ROC Score	Precision	Recall	F1 Score
1	Logistic Regresstion	0.78	0.758987	0.855942	0.763636	0.823529	0.792453
2	Random Forest Classifier	0.78	0.726076	0.832133	0.763636	0.823529	0.792453
3	K-Nearest Neighbours	0.73	0.721266	0.759904	0.700000	0.823529	0.756757
4	Naive Bayes	0.71	0.678354	0.852341	0.805556	0.568627	0.666667
5	Decision Tree Classifier	0.63	0.683323	0.629652	0.634615	0.647059	0.640777

Table 8: Performance Analysis of Random Under Sampling (RUS) with Feature Selection

	Model	Accuracy	K-Fold Validation	ROC Score	Precision	Recall	F1 Score
1	Random Forest Classifier	0.95	0.936086	0.989952	0.941473	0.957906	0.949618
2	Decision Tree Classifier	0.91	0.900721	0.910969	0.896142	0.930185	0.912846
3	K-Nearest Neighbours	0.89	0.882974	0.953680	0.837743	0.975359	0.901328
4	Logistic Regresstion	0.82	0.806327	0.894759	0.802941	0.840862	0.821464
5	Naive Bayes	0.79	0.789995	0.872469	0.755357	0.868583	0.808023

Table 9: Performance Analysis of Synthetic Minority Oversampling Technique (SMOTE) without Feature Selection

	Model	Accuracy	K-Fold Validation	ROC Score	Precision	Recall	F1 Score
1	Decision Tree Classifier	0.92	0.901233	0.919719	0.906561	0.936345	0.921212
2	K-Nearest Neighbours	0.89	0.880145	0.948380	0.838053	0.972279	0.900190
3	Random Forest Classifier	0.81	0.928883	0.984826	0.771788	0.881930	0.823191
4	Logistic Regresstion	0.80	0.789609	0.869944	0.772259	0.845996	0.807447
5	Naive Bayes	0.60	0.788710	0.864336	0.685185	0.379877	0.488771

Table 10: Performance Analysis of Synthetic Minority Oversampling Technique (SMOTE) with Feature Selection

The performance analysis of the various models and approaches are summarised from Tables 7 to 10. Based on the performance analysis table above, it is obvious that in this dataset, the oversampling technique, SMOTE, has a better overall accuracy compared to the undersampling technique, RUS.

Based on the performance analysis table, it shows that Random Forest Classifier (RFC) using SMOTE dataset without feature selection attained the greatest accuracy among all models and approaches applied, with a 95% accuracy rate and a 93.60% rate for K-Fold cross validation, as well as 0.94 precision and 0.96 recall. Additionally, it has an AUC value of 0.99 which is high.

In the performance analysis using RUS technique, the highest accuracy achieved is 78% for both Logistic Regression and Random Forest Classifier. Both of these dataset have been applied with feature selection. Both of the techniques have an average accuracy of 75.90% and 72.60% respectively after cross validation. On the other hand, the performance analysis using SMOTE technique, the highest accuracy is 95% for Random Forest Classifier without feature selection. It has an average accuracy of 93.60% after cross validation. We could observe a significant improvement in accuracy of these models from RUS to SMOTE. Hence, it has been established that sampling methods have a significant impact on model performance. It could be due to the volume of data since undersampling could result in overlap or implicit information loss.

The sample methods also have an impact on recall, which is crucial in medical situations and demonstrated significant improvement from RUS to SMOTE. KNN obtained the greatest recall of 0.86 in the undersampled dataset without feature selection, and the greatest recall of 0.98 in

the oversampled dataset without feature selection. Precision and F1 score also significantly improve after switching from RUS to SMOTE.

Furthermore, all of the models attained scores between 0.5 and 1, which, based on the obtained AUC values, indicates that they can distinguish between positive and negative class values. Majority of the modes achieve an AUC score of more than 0.8 except for Decision Tree Classifier which is employed in RUS sampling technique, which obtained a value of 0.62 without feature selection and 0.63 with feature selection. However, it is evident that when the Decision Tree Classifier model is being trained with SMOTE dataset, the AUC score increases significantly to 0.91 without feature selection and 0.92 with feature selection.

The observation that various models behave quite differently when employing various sampling approaches is an intriguing result. For instance, when applied with feature selection, Logistic Regression and Random Forest Classifier performed best with RUS, but Logistic Regression and Naive Bayes performed best with RUS when not applied with feature selection. With SMOTE without feature selection, Random Forest Classifier generated the best results, whereas with feature selection, Decision Tree Classifier produced the best results.

However, this investigation did not find any changes that were statistically significant when applied with feature selection. In fact, most of the models' accuracy has drastically decreased after being applied with feature selection. We observed a decrease of 0.8% in average in the accuracy for RUS dataset when being applied with feature selection and an average of 6.8% decrease in the accuracy for SMOTE dataset when being applied with feature selection. For instance, the largest decrease shown in the performance analysis table above is Naive Bayes trained on SMOTE dataset. When Naive Bayes model is being implemented on an oversampled dataset without feature selection, it has an accuracy of 79%. However, we observed a 19% decrease when the same model is being implemented on the oversampled dataset with feature selection as it only has an accuracy of 60%. Although most of the models' accuracy observed a decrease but there is an exception, for example, in Logistic Regression, Random Forest Classifier and Decision Tree Classifier that is being trained on RUS dataset with feature selection applied, it can be seen that there is an increase of 2%, 1% and 1% respectively when

compared with dataset without feature selection. On the other hand, after feature selection, all the models' performance on the SMOTE dataset was somewhat poorer.

Average decrease in accuracy for RUS dataset after feature selection

$$= \frac{(77-78)+(77-71)+(76-78)+(75-73)+(62-63)}{5}$$

$$= 0.8\%$$

Average decrease in accuracy for SMOTE dataset after feature selection

$$= \frac{(95-81)+(91-92)+(89-89)+(82-80)+(79-60)}{5}$$

$$= 6.8\%$$

Based on the performance analysis of our studies, we're able to conclude that sampling method plays an important role in affecting the performance of the machine learning algorithm, while feature selection, on the other hand does not show significant contribution to the performance, in fact, it decreases the performance of some models in some cases especially on oversampled dataset.

Based on Table 11, different authors used different machine learning models to predict stroke in patients. The table below shows the highest accuracy of the model that is being implemented by each reference. The Random Forest Classifier with SMOTE without feature selection is the best algorithm in this study, and it has the highest accuracy of 95%, putting the accuracy of this model close to other machine learning algorithms that are used by Eman M Alanazi et al. [2] and Tahia Tazin et al. [10], both of which used Random Forest Classifier and achieved the same accuracy of 96%. Additionally, the greatest accuracy found in this study is significantly greater than the Naive Bayes models used by Sailasya et al. [6] and Random Forest Classifier used by Tavares [9] with 82.0% and 92.3%, respectively.

However, comparing the models that are used on the RUS data in this paper with the models tabulated as below, even with the highest accuracy of 78% with the use of Logistic Regression and Random Forest Classifier on RUS with feature selection sample, it is clear that the

accuracy is still lower than the references below. The accuracy of 80% that is obtained by the Neural Network model that is being implemented by Dev et al. [3], has a significantly low accuracy if compared with the models that are being used on upsampled data in this paper.

Reference	Model	Accuracy
Tavares, Jose-A	Random Forest Classifier	92.3%
Dev et al.	Neural Network	80%
Eman M Alanazi et al.	Random Forest Classifier	96%
Sailasya et al.	Naive Bayes Classification	82%
Tahia Tazin et al.	Random Forest Classifier	96%

Table 11: Performance results of models by other authors

10.0 Limitation and Future Studies

10.1 Limitation

The main constraint of this research study turned out to be the inability to gather enough data due to limited amounts of datasets to provide better statistical and analytical outcomes. The use of various sampling approaches also presents a number of additional restrictions, such as undersampling, which led to the deletion of significant quantities of data and may have an impact on the accuracy and dependability of the models. On the other hand, excessive sampling of artificial data through the oversampling method, SMOTE, may result in an overfitting of the model to the artificial data. As a consequence of the data not being compared to real-world data, a false positive result might ensue. Finding the best algorithm for patient stroke prediction is also impossible with the restricted number of machine learning models being implemented. In actuality, the lack of implementation of deep learning models has led to less desirable outcomes.

10.2 Future Studies

A more balanced dataset which includes a broader dataset with real world data should be used to improve model development results in future studies. Extra caution should be also exercised when datasets are cleaned and processed, assuring that the dataset is fit to be used.To better decide the optimal model for stroke prediction, other machine learning models including Support Vector Machine, AdaBoost, XGBoost, ANN, and Weighted Voting should be applied to the dataset. Deep learning models are also encouraged as they are able to provide more accurate and trustworthy findings, be it statistical or analytical. It is with this in mind that the Artificial Neural Network (ANN) may be suggested to be tested on the dataset.

11.0 Conclusion

A stroke has to be treated straight away since it might be deadly. Stroke is a serious medical condition that may be triggered by any number of illnesses, including diabetes, high blood sugar, high cholesterol, and hypertension, to mention a few. The development of machine learning algorithms that can predict a patient's stroke risk in advance may assist to reduce that risk or perhaps fully avert a future stroke that is severe. This research study effectively predicts stroke in patients based on their health metrics using multiple machine learning algorithms while also experimenting with different sampling approaches and feature selection to improve model accuracy.

Based on literature review, it seems that Random forest classifiers have seen the most promising results so far. Similar results were observed in our case too. Based on our own modelling, our decision tree classifier with SMOTE without feature selection was able to outperform every other classifier with sampling techniques, boasting an accuracy of 94.6%. Of the sampling techniques, the SMOTE dataset seems to have a better effect on accuracy of the models. An example of this can be seen in Random Forest Classifiers Models, SMOTE sampling when applied, had an increased accuracy of 18% when compared to RUS sampling. Feature selection, on the other hand, has shown mixed effects on the accuracy of the models, and further study would be needed to determine its feasibility.

12.0 References

[1] *About Stroke | cdc.gov*. (2022, November 2). CDC. Retrieved November 23, 2022, from https://www.cdc.gov/stroke/about.htm

[2] Alanazi, E. M., Abdou, A., & Luo, J. (2021, December 2). Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models. *JMIR Formative Research*, *5*(12), 23440. 10.2196/23440

[3] Dev, S., Wang, H., & Nwosu, C. S. (2022, November). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytics*, *2*, 100032. 10.1016/j.health.2022.100032

[4] McLeod, S. (n.d.). *P-Value and Statistical Significance*. Simply Psychology. Retrieved November 18, 2022, from https://www.simplypsychology.org/p-value.html

[5] *Overview - - - Stroke*. (n.d.). NHS. Retrieved November 23, 2022, from https://www.nhs.uk/conditions/stroke/

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

[7] Scott, G. (n.d.). *Multicollinearity*. Investopedia. Retrieved November 18, 2022, from https://www.investopedia.com/terms/m/multicollinearity.asp

[8] *Stroke - Symptoms and causes*. (2022, January 20). Mayo Clinic. Retrieved November 23, 2022, from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

[9] Taveres, J.-A. (2021). Stroke prediction through Data Science and Machine Learning Algorithms. 10.13140/RG.2.2.33027.43040

[10] Tazin, T., Alam, M. N., Dola, N. N., Bourouis, M. S., & Bourouis, S. (2021, December 26). Stroke Disease Detection and Prediction Using Robust Learning Approaches. *Journal of Healthcare Engineering*, *2021*, 12. 10.1155/2021/7633381

[11] Tong, X. (2017). *FIGURE 1. Stroke deaths and age-standardized stroke death rate among...* ResearchGate. Retrieved November 23, 2022, from https://www.researchgate.net/figure/Stroke-deaths-and-age-standardized-stroke-death-rate-among-adults-aged-35-years-United_fig1_319568256

[12] *WHO EMRO | Stroke, Cerebrovascular accident | Health topics*. (n.d.). WHO Regional Office for the Eastern Mediterranean. Retrieved November 23, 2022, from http://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html