Table of Contents

1. Introduction

1.1 Problem Statement

Since stroke has become the third disease led to death in Malaysia, and the people suffered into stroke in Malaysia keep increasing since 2016 until now, we tend to find out what are the reasons causing people suffered into stroke keep increasing in recent years. We would like to use the data research to predict how stroke happens, using different method to figure out which method will increase the accuracy on stroke prediction in an individual.

1.2 Research Question

The framework of this research will be established on the core theme of exploring the potential factors triggering strokes through quantifying historical data. This report begins by examining the range of age that experiences stroke the most. Followed by investigating if gender or smoking influences the risk of having a stroke. Subsequently, proving that people with high BMI have a higher risk of getting a stroke. Finally, explore if the living environment is one of the factors causing a stroke. Different algorithm will be used to predict the risk of having stroke, the one with highest accuracy will be used.

1.3 What is Stroke?

Stroke which also known as brain attack. Usually occurs when something obstructs the brain's blood flow or when a brain blood artery ruptured. Part of the brain either die or suffers harm in both scenarios. When there is something block the blood flaw will cause the brain lack of oxygen and nutrients which lead the brain cells start to die within minutes. A stroke may result in permanent brain damage, chronic disability, or even fatality. It can be divided into 2 types which is ischemic stroke and hemorrhagic stroke. Mild weakness, paralysis, or numbness on one side of the body or face can all be symptoms of a stroke. Other symptoms include abrupt weakness, a strong headache, vision problems, trouble speaking or comprehending speech, and difficulty seeing.

1.4 Stroke in Malaysia

There were 47,911 incident cases, 19,928 fatalities, 443,995 prevalent cases, and 512,726 DALYs lost due to stroke in Malaysia in 2019 (Tan & Venketasubramanian, *Stroke burden in Malaysia* 2022). The DALY is used to evaluate the total burden of illness. However, Malaysia has a lower age and sex-standardized stroke mortality and DALYs compared to many other countries in Southeast Asia (*Indicator metadata registry details* 2022). Multiple national health and morbidity studies from 2006 showed that risk factors including diabetes, hyperlipidemia, and obesity are becoming more and more common. These risk factors have been linked to an increase in stroke incidence in people under the age of 65, with the age group of 35 to 39 years showing the highest rise which 53.3% in males and 50.4% in women respectively. (Tan & Venketasubramanian, *Stroke burden in Malaysia* 2022) Based on the latest report given by WHO in 2020, the death caused by stroke in Malaysia has reached to 21592 or 12.85% of total deaths. The death rate of stroke in Malaysia is 81.65 per 100,000 which in the 92 in the world and stroke is in the third leading cause of death within Malaysia. Diagram below shows trends in the incidence of inpatient stroke for males (A-D) and women (EH) in Malaysia between the years of 2008 and 2016.

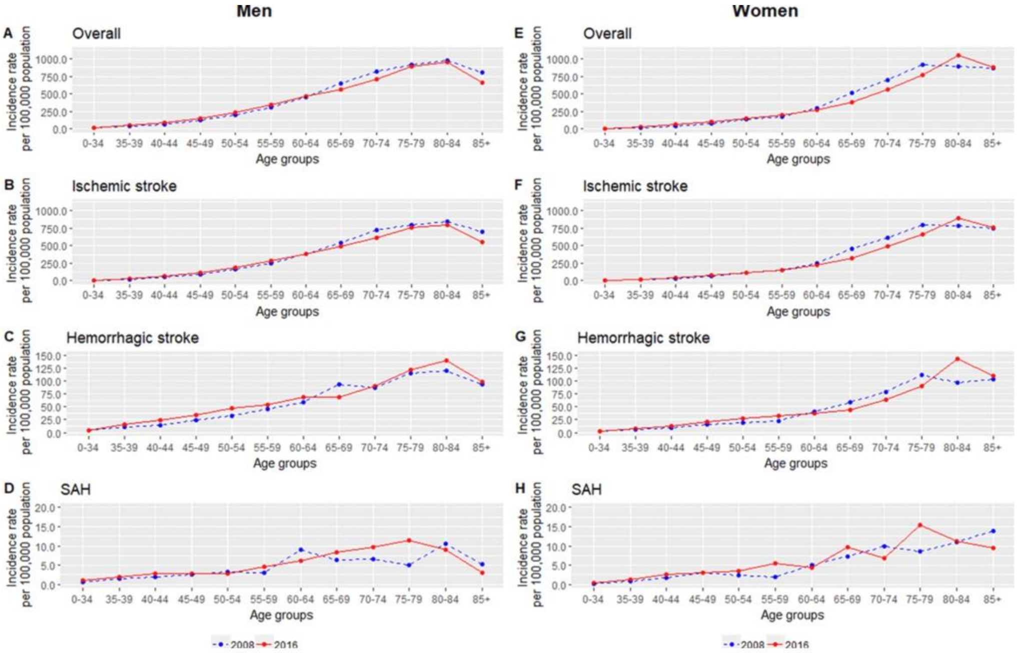


Figure 1. Trends in the incidence of inpatient stroke for males (A-D) and women (E-H) in Malaysia between the years of 2008 and 2016.

2. Related Work

2.1 Attributes

Based on our research, the common key attributes that will be used are gender, daily activities and BMI. There are many factors that could cause stroke, such as age (if someone is over 55 years of age, they are clearly more likely to be affected, although stroke is described at any age, even in children), gender, BMI, family history, blood pressure, lipid profile, blood sugar, ever married, smoking status and many more. One of the research papers was using, age, gender, hypertension, heart disease, ever married work type, residence type, avg glucose level, BMI, smoking status and stroke, these are all the attributes that were used to perform the training as well as testing. *(Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Sensors 2022)* In addition to this, there is a research paper used some clinical outcome as their attributes namely acute kidney injury, vascular complications, acute stroke, blood transfusion. The paper also used some extra patients characteristic such as prior stroke, chronic lung disease, coagulopathy, diabetes controlled, obesity and so on. *(Cardiovascular Revascularization Medicine, Jul 2022).* Researchers have studied multiple methodologies such as case-control studies, case series, prospective cohort studies in the past few decades and identified non-modifiable risk markers (male gender, older age and genetic factor) and modifiable markers (hypertension, smoking status and diabetes mellitus). *(Omae, Stroke risk factors and stroke prevention 1992)*

2.2 Machine Learning Algorithms Used

After reading all the research papers, there are some few common machine learning algorithms that were being used for their dataset namely naïve bayes, random forest, logistic regression, linear regression, k-nearest neighbors and decision tree. All the research paper used multiple machine learning algorithms in order to come out with a comparable and accurate result for the prediction of stroke. Besides, one of the research papers showed that among all the algorithms used, Stacking algorithms has the highest accuracy which is 0.98 (98%). *(Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Sensors 2022)(**Figure 2. Average Performance of ML models**)*

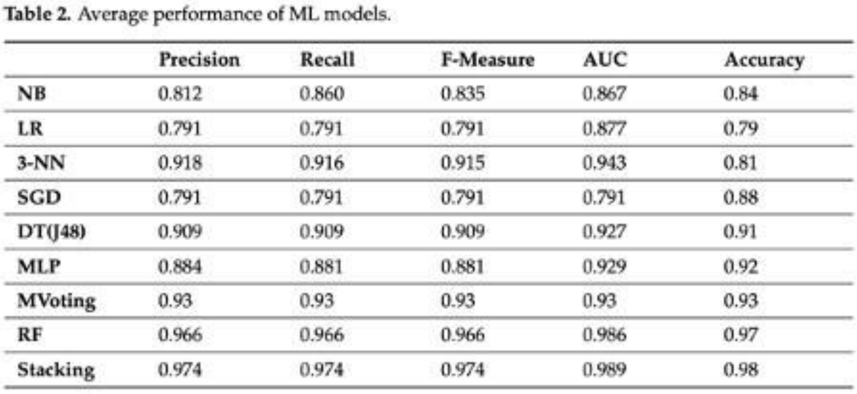


Table 2. Average performance of ML models.

	Precision	Recall	F-Measure	AUC	Accuracy
NB	0.812	0.860	0.835	0.867	0.84
LR	0.791	0.791	0.791	0.877	0.79
3-NN	0.918	0.916	0.915	0.943	0.81
SGD	0.791	0.791	0.791	0.791	0.88
DT(J48)	0.909	0.909	0.909	0.927	0.91
MLP	0.884	0.881	0.881	0.929	0.92
MVoting	0.93	0.93	0.93	0.93	0.93
RF	0.966	0.966	0.966	0.986	0.97
Stacking	0.974	0.974	0.974	0.989	0.98

Figure 2 Average Performance of ML models

2.3 Performance Measurement

Performance measurement can be applied to each machine learning algorithms and comes out with a result that let the user able to determine which algorithms is more suitable and prior to the dataset compared to other algorithms. There are several kinds of performance measurement such as, precision, recall, f-measure, AUC and accuracy which were used in a research paper. *(Dritsas, E.; Trigka, M. Stroke Risk Prediction with Machine Learning Techniques. Sensors 2022)* One of the article mentioned, mean absolute error, mean squared error, root mean squared error, r-squared these are all under regression metrics. Some other performance metrics such as, accuracy, confusion matrix, precision and recall, f1-score and au-roc are under classification metrics. There are two main type performance measurement metric (mentioned above) that could be used for our algorithm to measure the performance for each algorithms. (*Bajaj, Performance metrics in machine learning [complete guide] 2022)*

2.4 Analysis Tool

Among the journal articles reviewed, two chose R tools, either R Language or R Studio, as analysis tool for developing their prediction model (Dev, et al., 2022; Liu, et al., 2022). R is the programming language which performs the statistical computing and graphics while R Studio acts as an IDE to provide increased functionality. Besides R Language, Liu et al. (2022) used other analysis tools such as H2O Driverless AI and Python. H2O Driverless AI is an AI platform which automates complex data science and machine learning process such as feature engineering, model validation, tuning, selection and deployment which improves both efficiency and accuracy (H2O.ai, n.d.). On the other hand, in the study of Biswas et al. (2022) and Bansal et al. (2022), analysis of data was done using analysis tools for Python, with the former choosing Jupyter Notebook, a web-based interactive computational environment as the IDE. Jupyter Notebook supports various programming languages besides Python such as Julia, Scala and R. Besides that, MATLAB R2020b, a high-level programming language commonly used by engineers and scientists, was utilized by Imura et al. (2021) in their study to compare the performance metrics of different supervised machine learning algorithms in predicting stroke patient's discharge possibility. In another research paper, two analysis tools namely Weka and Stata were used. The former was used for implementation of Random Tree, Linear Regression and K-Nearest Neighbor while the latter was used for general statistical analysis such as finding the mean, standard deviation, median, counts, percentages and running Chi-squared tests to compare categorical variables.

2.5 Limitations

While machine learning gives systems the capability to learn without explicit programming, each machine learning model built, even those with desirable performance metrics, still has its limitations when applied in real life. Similarly, limitations of several journal articles were identified to aid in training the machine learning model in this report. Firstly, most of the journal articles are the outcome of retrospective studies. Retrospective study is when preexisting data are analyzed, such as data from existing medical records of patients or data from events that have taken place (Li, et al., 2022; Biswas, et al., 2022; Nwosu, et al., 2019). As the researchers are not involved in the collection of these data, the data may be subjected to misclassification or recall bias. In addition, in the journal article by Li et al. (2022), only the LASSO regression model, was used to screen the radiomic features of diffusion-weighted imaging (DWI) to predict the outcome after acute ischemic stroke patients receive a mechanical thrombectomy. Hence, the limitation of the study would be the lack of training and testing of other feature screening methods. The next limitation is the usage of crosssectional data, which are data collected by observing many subjects during a fixed period of time (Liu, et al., 2022). Hence, the resulting machine learning model cannot be used to infer the causality or treatment decision. In addition, the feature deletion and observation deletion process in the study by Liu, et al. (2022), also introduce bias into the journal article's results. As an example, deletion of observation with above 30% missing values or deletion of informative features with small sample size.

2.6 Critical Review

In all of the journal articles analyzed, feature selection was carried out. Some studies opt to use existing feature selection algorithm such as the LASSO regression model for radiomics characteristics feature extraction (Li, et al., 2022) and BoostARoota to exclude redundant features (Liu, et al., 2022). In the journal article by Liu et al. (2022), the BoostARoosta algorithms was ran thirty times with different random seed to identify robust features for their analysis. Alternatively, Sung, et al. (2015) in their research used a three-step feature selection process. First, any features that presents in less than 1% or more than 99% of patient claims are dropped. Second, correlation-based feature selection (CFS) was used. The study by Nwosu et al. (2019) has a similar approach for their feature selection process, which is excluding one of two features that are highly correlated. An excellent feature subset consists of features that are highly correlated to the outcome and not correlated with each other (Sung, et al., 2015). The third step in Sung et al. (2015) study is to obtain expert opinion on the feature subset selected. In summary, feature selection is essential in ensuring the machine learning models that we train in this study recognizes the right patterns in the data and is able to make prediction with high accuracy. In addition, majority of the research papers implements multiple machine learning algorithms. The performance metrics of different machine learning algorithms are then compared to identify the best model.

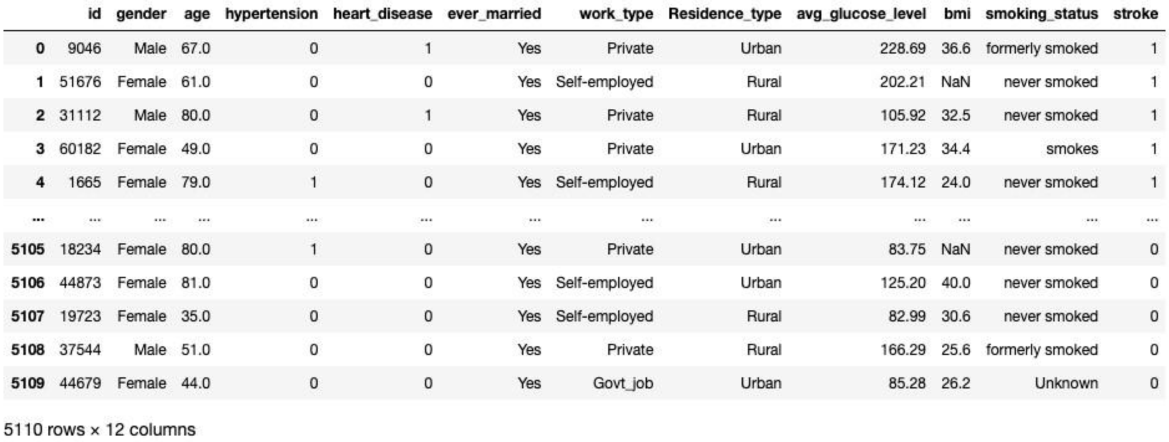
2.7 Conclusion

Referring to all the information summarized from the various research papers, several key aspects were recognized and will be implemented in the design of our machine learning model. Firstly, it has been identified that common features that are used to train stroke prediction machine learning model consists of age, gender, BMI, blood pressure, blood sugar, smoking status and more. Other less common but influential attributes can also be included to develop a model with higher accuracy but having more attributes does not necessarily mean a better machine learning model. Next, it can be summarized that multiple machine learning algorithms should be implemented to train the stroke prediction machine learning model in our study. Comparison of the performance metrics of the models developed using different machine learning algorithm will aid in determining the best models. In addition, the analyzation tool chosen for our study is Jupyter Notebook which allows cell by cell running, promoting effective communication of ideas and results. In addition, there are built-in data analysis libraries such as pandas and numpy, also data visualization libraries such as matplotlib and Seaborn. As for the limitations, steps will be taken to avoid excess observation deletion such as by populating Null data with the calculated median of the feature.

3. Data and Method

3.1 Data

The dataset used in this study is a stroke prediction dataset obtained from Kaggle. It is open to be used by the public. It contains 5110 instances and 12 attributes. These attributes and the unique values they contain are included in Figure 3 below.



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

Figure 3. Data and attributes used

The table below shows the attributes and theirs unique values (if any).

Table 1. Attributes and unique values (if any)

Attribute	Unique Values	Remarks
id		
gender	1. Male 2. Female 3. Other	
age	0.08 (< 1 month) to 82 years	
hypertension	0 or 1	0: if the patient does not have hypertension 1: if the patient has hypertension
heart_disease	0 or 1	0: if the patient does not have hypertension 1: if the patient has hypertension
ever_married	'Yes' or 'No'	-
work_type	1. Private 2. Self-employed 3. Children 4. Govt_job 5. Never_worked	-
Residence_type	1. Urban 2. Rural	-
avg_glucose_level	55.12 to 271.74	Average glucose level in the blood
bmi	10.3 to 97.6	

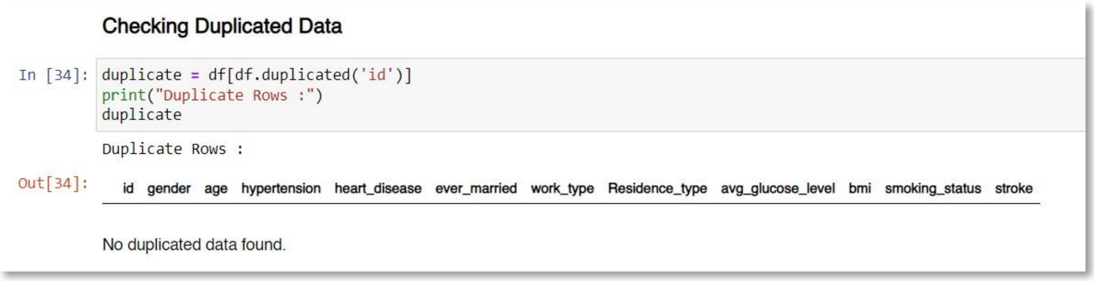
smoking_status	1. Never smoked 2. Unknown 3. Formerly smoked 4. Smokes	
stroke	0 or 1	0: if the patient did not have a stroke 1: if the patient had a stroke

3.2 Method

3.2.1 Data Pre-processing

Data Cleaning

To ensure the quality of the data used and the results obtained from different predictive models, several attempts have been made to clean the data such as identifying duplicates, null values and outliers. The original dataset consists of 5110 rows of data with 12 columns, including the target.



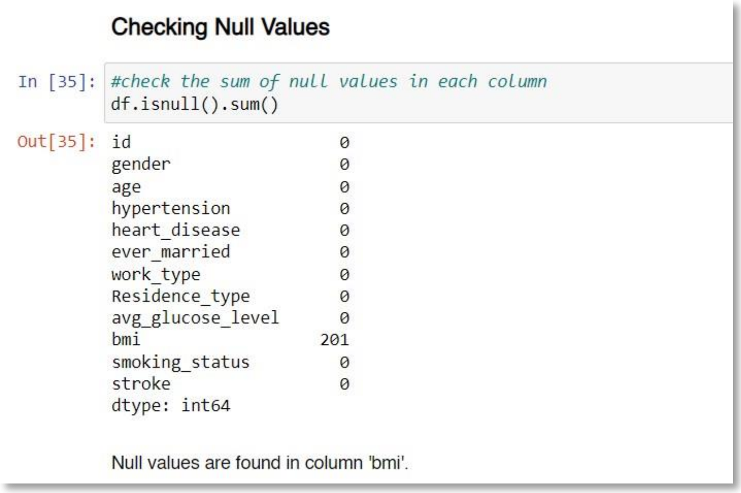
Checking Duplicated Data

```
In [34]: duplicate = df[df.duplicated('id')]
         print("Duplicate Rows :")
         duplicate

         Duplicate Rows :

Out[34]:   id  gender  age  hypertension  heart_disease  ever_married  work_type  Residence_type  avg_glucose_level  bmi  smoking_status  stroke

         No duplicated data found.
```

Figure 4: Identifying duplicates by searching for duplicated patient ‘id’

In the figure above, no duplicated ‘id’ was found. Hence, no rows are dropped in this step.

Missing values were found in the ‘bmi’ column. The null values were replaced with the median of the ‘bmi’ feature using the SimpleImputer class from the scikit-learn module which allows replacing missing values using the statistics of columns which these missing values are located. With that, all null values in the dataset have been handled.



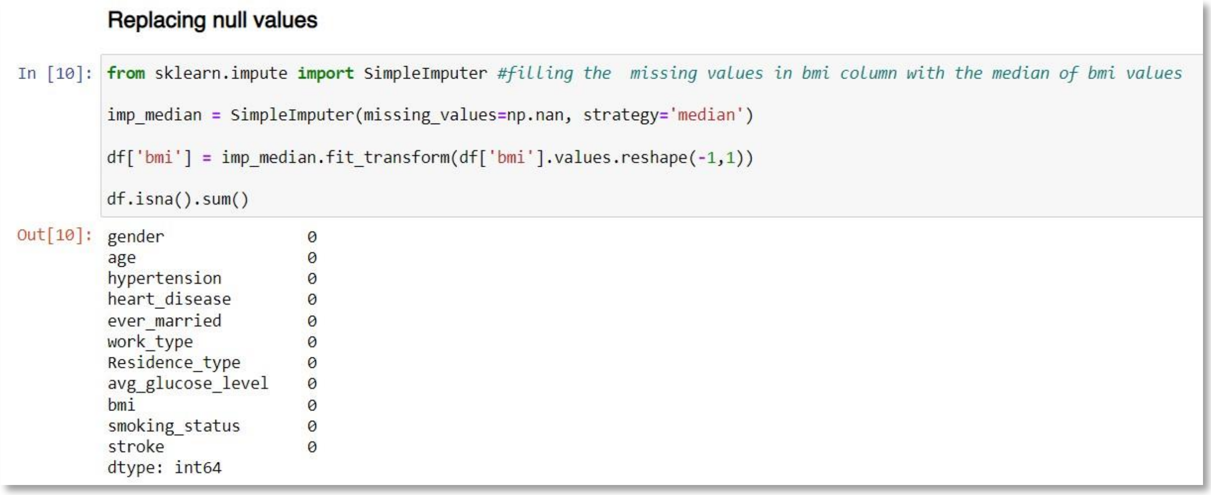
Checking Null Values

In [35]:
```
#check the sum of null values in each column
df.isnull().sum()
```

Out[35]:
```
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke                 0
dtype: int64
```

Null values are found in column 'bmi'.

Figure 5: A total of 201 null values were identified in the 'bmi' column.



Replacing null values

In [10]:
```
from sklearn.impute import SimpleImputer #filling the  missing values in bmi column with the median of bmi values

imp_median = SimpleImputer(missing_values=np.nan, strategy='median')

df['bmi'] = imp_median.fit_transform(df['bmi'].values.reshape(-1,1))

df.isna().sum()
```

Out[10]:
```
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                  0
smoking_status       0
stroke               0
dtype: int64
```

Figure 6: Sum of 'bmi' null values is 0 after imputation with median of 'bmi' column.

To identify outliers, box and whisker plot of independent variables (features) against the dependent variable (target) were plotted. Excluding categorical data, outliers were observed in age, bmi and average glucose levels.

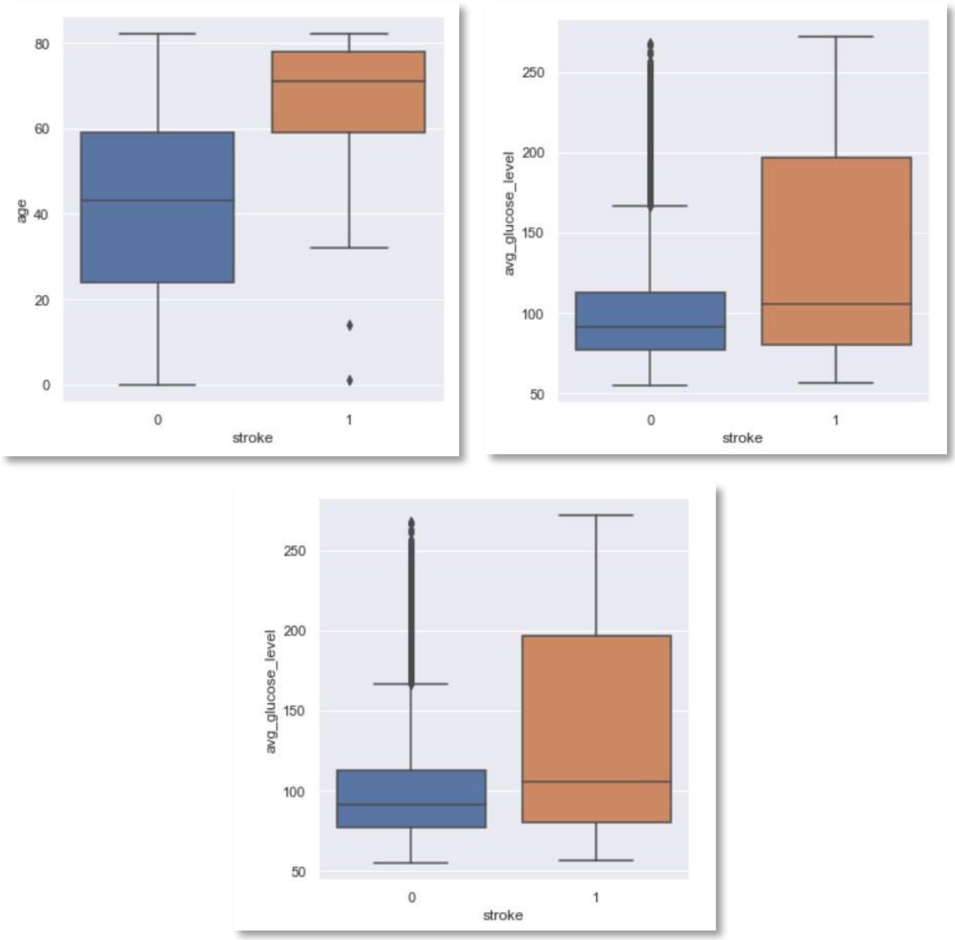


Figure 7: Outliers in each feature can be observed on the box plot.

For the 'age' feature, index of the columns containing the outliers (age is less than 20 and has stroke) was dropped. For the 'avg_glucose_level' feature, index of rows with average glucose level above 220 and does not have stroke was dropped. Lastly for the 'stroke' feature, rows which satisfies the below condition were dropped:

i. bmi is greater than 50 and does not have stroke, ii.
bmi is greater than 40 and has stroke, iii. bmi is
less than 19 and has stroke.

After dropping the rows mentioned above and the target column, 'stroke', the dataset consists of 4838 rows and 11 columns.

Data Encoding

In machine learning, encoding is known as the process of replacing categorical values with unique numeric values in the range of 0 and the number of classes minus 1. Encoding is necessary in this study as most machine learning algorithms used can only be trained using numerical values. Label-Encoder was used to encode the categorical features in the dataset such as 'gender', 'ever_married', 'work_type', 'Residence_type' and 'smoking_status'.

Data Balancing

A dataset is said to be imbalanced when the target class has a skewed distribution of observations. Two techniques that can be used to handle imbalanced data is oversampling or undersampling. Oversampling is the act of creating synthetic observations in the minority class by duplicating observations (Brownlee, 2020). In contrast, undersampling is the act of maintaining data in the minority class and decreasing the size of the majority class (2U, 2022).

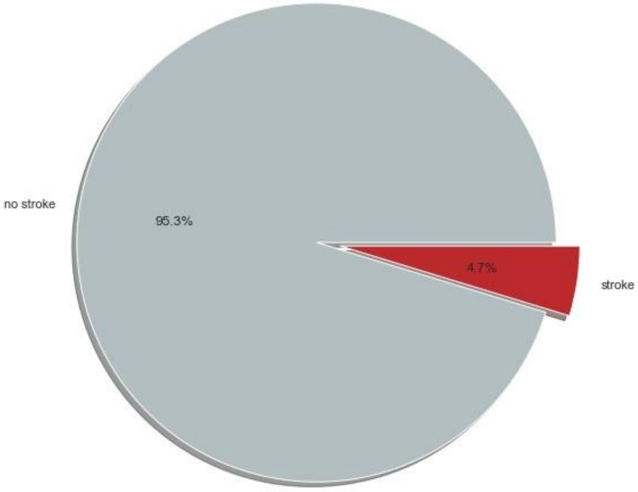


Figure 8: The pie chart shows that the dataset is imbalance, as 95.3% of the data are no stroke data with only 4.7% has stroke.

In this study, the dataset used is imbalance with 'no stroke' data as the majority class. As the size of the minority class, 'stroke', is only 228, undersampling will produce a small balanced dataset which is inadequate as the representation of true distribution of data in the population. This may then lead to a less robust machine learning model which produces inaccurate results. Hence, the oversampling technique was implemented using the Synthetic Minority Oversampling Technique, also known as SMOTE. This technique was first proposed in 2002 by Nitesh Chawla, et al. Firstly, from the minority class, a random instance, *x*, is chosen. SMOTE then utilizes k-nearest neighbors algorithm to identify the *x*'s k-nearest neighbors. One of the k-nearest neighbors is then randomly chosen and a synthetic instance, *y*, in created. Other synthetic instances created will be the convex combination of the two instances, *x* and *y*.

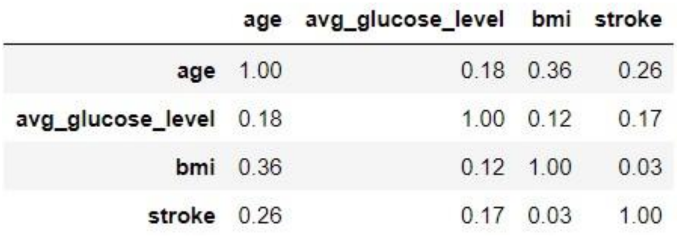
Data Splitting

As this is a retrospective study which uses existing medical data to train different predictive models, the dataset is split into two subsets, training and validation. The method train_test_split was used to split the data, with the size of the validation subset equals to 25%. The training subset will be fitted into different machine learning algorithms to create trained models. Finally, the trained models will be passed the validation subset to predict. The performance of the different trained models in predicting the target is then compared.

Besides using the conventional train_test_split method, K-Fold Cross Validation is also implemented. The disadvantages of the train_test_split method such as only producing one training subset and one validation subset, as well as fluctuating accuracy when using different random_state value can be overcome using K-Fold Cross Validation (MLTut, 2020). It is a method which splits a dataset into K number of segments and each segment will be used as the training validation subset one by one, while the other segments form the training subset.

3.2.2 Exploratory Data Analysis (EDA)

Correlation



	age	avg_glucose_level	bmi	stroke
age	1.00	0.18	0.36	0.26
avg_glucose_level	0.18	1.00	0.12	0.17
bmi	0.36	0.12	1.00	0.03
stroke	0.26	0.17	0.03	1.00

Figure 9*: *Correlation table

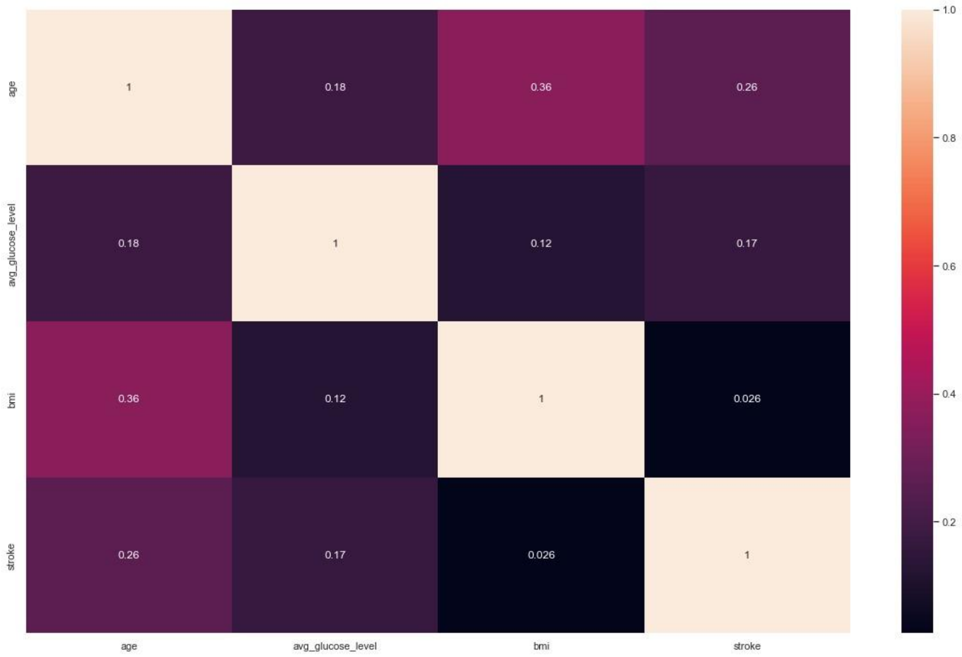
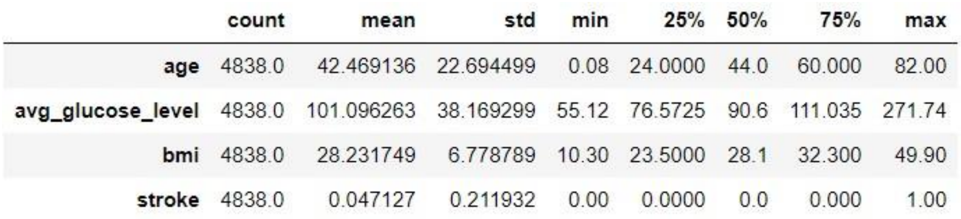


Figure 10: Heatmap

From figure 9 and 10, the correlation matrix we can see the variables 'bmi' and 'age' are highly correlated to each other since they have the correlation of 0.36. The variables 'stroke' and 'bmi' are least correlated to each other since they have a correlation of 0.026. Furthermore, the rest of the variables are related to each other but not considered as highly correlated.

Measure Of Central Tendency and Dispersion



	count	mean	std	min	25%	50%	75%	max
age	4838.0	42.469136	22.694499	0.08	24.0000	44.0	60.000	82.00
avg_glucose_level	4838.0	101.096263	38.169299	55.12	76.5725	90.6	111.035	271.74
bmi	4838.0	28.231749	6.778789	10.30	23.5000	28.1	32.300	49.90
stroke	4838.0	0.047127	0.211932	0.00	0.0000	0.0	0.000	1.00

Figure 11: Table for measure of central tendency and dispersion

Figure 11 shows the measure dispersion and the measure of central tendency of the numerical variables. We can observe the mean, standard deviation, minimum, first quartile range (25%), median (50%) and third quartile range (75%) and maximum for all the numerical data.

Data Visualisation

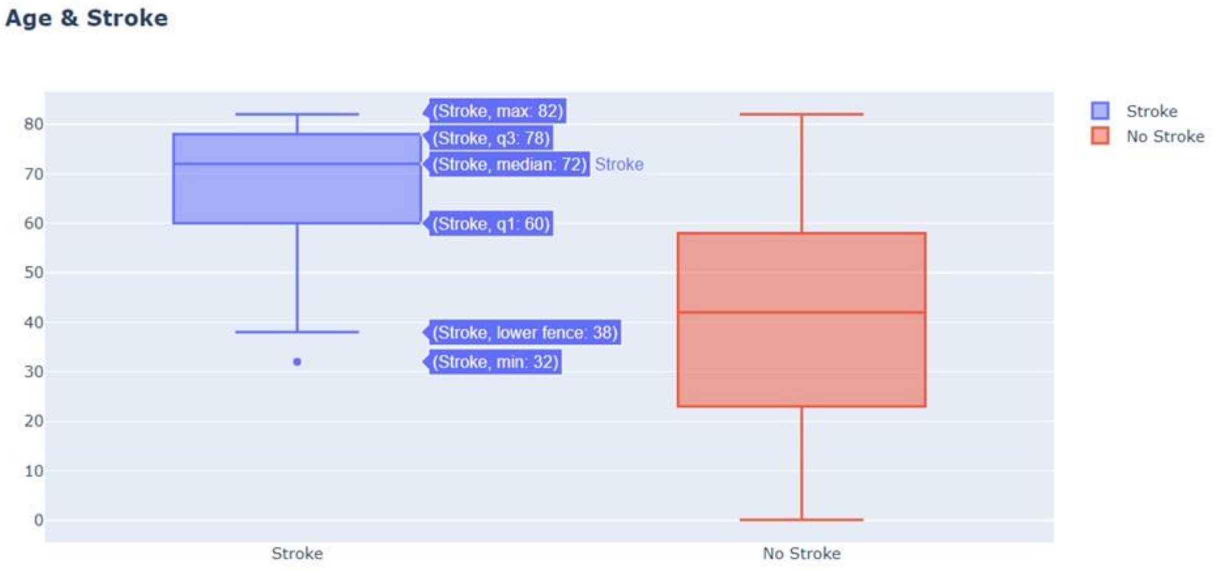


Figure 12: Box plot for 'age' column for stroke and no stroke data

Figure above shows the box plots for age and stroke. We can observe that people with higher age have higher risk of getting stroke. The majority of patients that have stroke are aged around 60 to 78 years old.

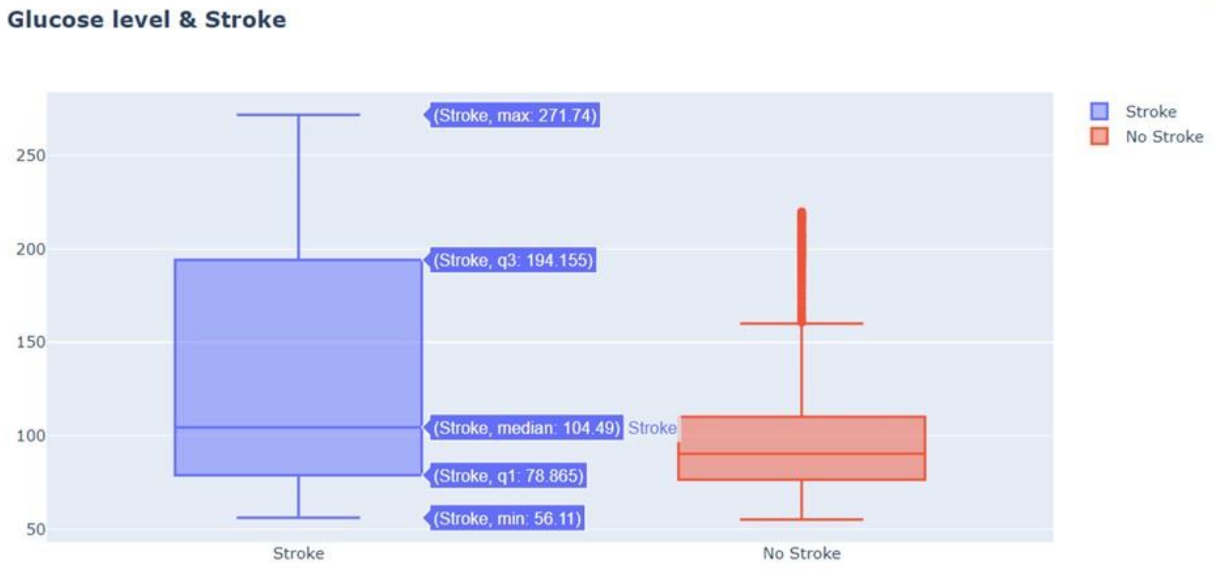


Figure 13: Box plot for 'avg_glucose_level' column for stroke and no stroke data

Figure 13 shows the box plot of average glucose level and stroke. Stroke occurred in a patient whose average post-meal blood glucose level was greater than 150 mg/dL (milligrammes per decilitre). Patients with a value of greater than 200 mg/dL have diabetes. If the patient's level fell between 140 and 199 mg/dL, the patient will be considered as pre-diabetes. Diabetes is one of the risk factors for stroke, and people with prediabetes are more likely to get a stroke.

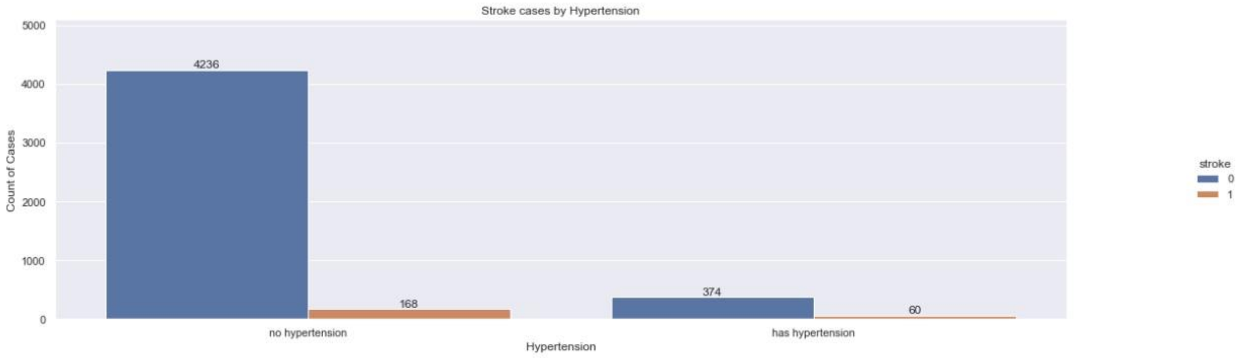


Figure 14: Box plot for 'hypertension' feature for stroke and no stroke data

We can see that there are 60 people with hypertension has stroke and 168 people without hypertension has stroke. We can observe that people with hypertension have lower risk of getting stroke while people without hypertension have higher risk of getting stroke.

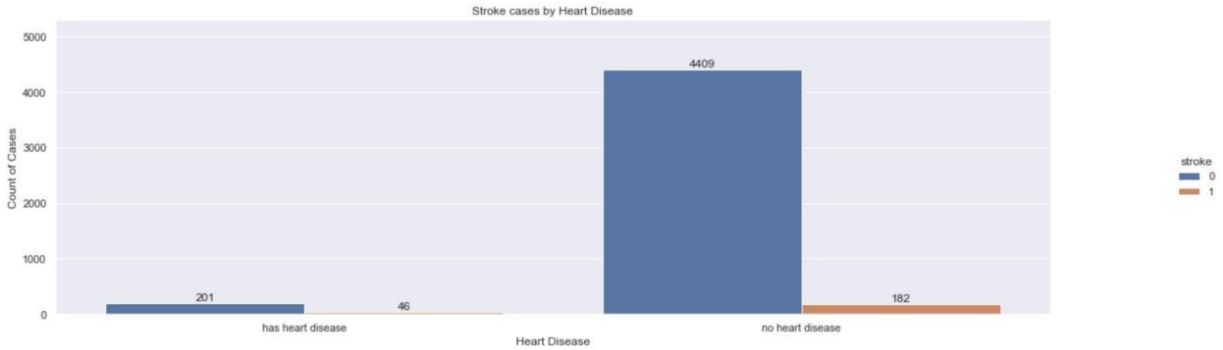


Figure 15: Box plot for 'heart_disease' feature for stroke and no stroke data

There are 186 people without heart disease has stroke and 46 people with heart disease has stroke. Thus, people without heart disease have a higher risk of getting stroke.

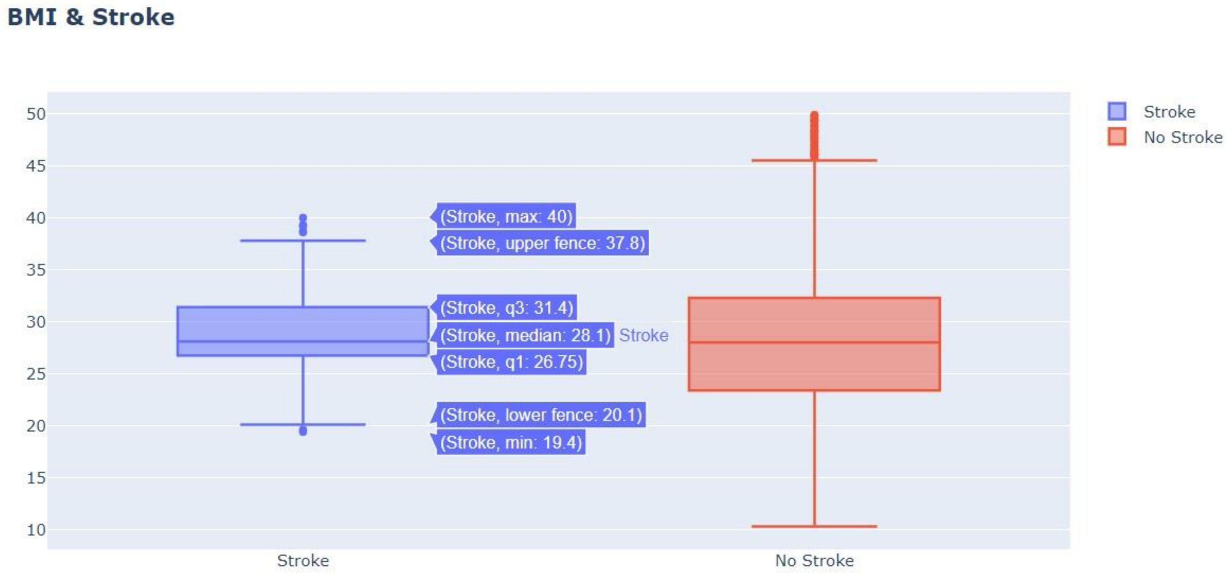


Figure 16: BMI and Stroke

Figure 16 shows the box plot of the column bmi and stroke. The majority of the patients that had BMI between 26 and 32 were the highest to suffer a stroke. Thus, higher BMI does not increase the risk of stroke.

3.2.3 Machine Learning Classification Algorithms

Throughout this report, we have used different kind of machine learning classification algorithms to undergo the prediction, the algorithm used is explained in below.

1. Support Vector Machine (SVM): Defined as a supervised machine learning algorithm that is utilized for classification and regression. It is considered a reliable classification algorithm due to its efficiency even with a limited set of data. It is commonly referred to as regression concerns, but it is more appropriate to label it as categorization.
2. K-Nearest Neighbors (KNN): Described as a non-parametric supervised learning classifier that adopt proximity to classify or predict how a single data point will be categorized.

3. Random Forest: A widely used machine learning technique that integrates the output of many decision trees to arrive at a single conclusion. Its widespread use is motivated by its adaptability and usability since it can solve classification and regression issues.
4. Decision Tree: It is similar to a flowchart in which every single internal node represents the test of an attribute, the test result is represented on the branch, and the leaf node which is also known as the terminal node represents the class label.

Since KNN and SVM are based on Euclidean distances, they are unable to be used directly to handle the categorical variables in a dataset, only numerical variables can be applied directly in these two algorithms. However, both numerical data and categorical data is able to apply on random forest and decision tree. Throughout the report, we will convert all the categorical data into numerical data by using LabelEncoder() from sklearn library based on the original datasets. Table 2 below shows the detail information of each algorithm used.

Table 2. Detail Information of Each Algorithm Used

Predictive Model	Data Type			Trained on Dataset		
	Numerical	Categorical	Original	Numerical Variables Only	Categorical Variables Only	Discretization
SVM	√	X	√	√	X	X
KNN	√	X	√	√	X	X
Random Forest	√	√	√	√	X	X
Decision Tree	√	√	√	√	X	X

3.2.4 Performance Measurements

Accuracy

Accuracy metric is one of the simplest classification metrics that most of the machine learning algorithms can use it to come out with the performance result intuitively. Accuracy can be formulated as number of correct predictions divide by the total number of predictions. Accuracy is suitable for the data that are approximately balanced. An example is our dataset, stroke prediction. Before data being balanced, our data is highly unbalanced, over 90% belongs to one class, and another 10% belongs to another, this will lead our accuracy becomes low, in other word, the performance of algorithms is bad.

Precision

Precision of a machine learning algorithm model describes how many detected items are truly relevant it is calculated by dividing the true positives by overall positive. The true and false positive indicates that how many positive predictions that are true and false, same with the true and false negative predictions, how many negative predictions that are true and false. The formula for precision is TP divides by TP + FP.

Recall

Recall is similar to precision; it is one of the metrics that quantifies the number of true positive predictions made out of all positive predictions that could have been made. Precision only comments on the true positives predictions out of all positive's predictions. However, recall provides an indication of missed positive predictions.

F-Measure

Since precision and recall is complement each other, which mean 100% recall will have 0% precision and vice versa. F-Measure comes in to tackle this issue. It will maximize both precision and recall.

Cross Validation

A statistical method that is used to estimate the performance of the machine learning algorithms' model. It is used to protect against overfitting in a predictive model, especially in a case where the amount of data may be limited. In a simple explanation, cross validation is dividing the dataset into partitions or is called folds, each fold will contain the same percentage of the total data. Next, is create the models for each folds, for instance, model 1 will use fold 1 as testing data and the rest folds will be the training data. Each model will have its own accuracy, using all the performance of each model, coming out with the average accuracy/performance of this algorithms.

GridSearchCV

The objective of grid search is to identify the best hyperparameter values to obtain the best prediction results from out model. Grid search calculates the performance for each combination of all the supplied hyperparameters and their values. Next, it will choose the ideal value for the hyperparameters. The processing and consuming time are based on the amount of hyperparameters that involved.

4. Result and Discussion

4.1 Result

4.1.1 Support Vector Machine

Table 3. SVM with the best results after implementing GridSearch

Support Vector Machine	Accuracy	Recall	Precision	F1-Score
Default results	0.91	0.91	0.92	0.91
After cross validation	0.91	0.91	0.92	0.91
After GridSearch	0.97	0.98	0.96	0.97

The table above shows that SVM obtained the best results after implementing GridSearch, which finds the best parameters to fit the model with. It was found that the best parameters were C=10, gamma=1, probability=True. Grid search also includes cross validation to test the model with the entire dataset.

4.1.2 K-Nearest Neighbors

Table 4. K-Nearest Neighbors with the best results after implementing GridSearch

K-Nearest Neighbors	Accuracy	Recall	Precision	F1-Score
Default results	0.93	0.99	0.89	0.94
After cross validation	0.85	0.92	0.81	0.86
After GridSearch	0.98	1.0	0.96	0.98

The table above shows that KNN obtained the best results after implementing GridSearch, which finds the best parameters to fit the model with. It was found that the best parameters were n_neighbors=1.

4.1.3 Random Forest

Table 5. Random Forest with the best results after implementing GridSearch

Random Forest	Accuracy	Recall	Precision	F1-Score
Default results	0.97	0.96	0.99	0.97
After cross validation	0.97	0.96	0.99	0.97
After GridSearch	0.97	0.95	0.99	0.97

The table above shows that random forest obtained the best results after implementing cross validation, which trains and tests the data with different sections of the dataset each time. This means that the model can be tested multiple times with different test data.

4.1.4 Decision Tree

Table 6. Decision Tree with the best results after implementing GridSearch

Decision Tree	Accuracy	Recall	Precision	F1-Score
Default results	0.95	0.94	0.95	0.95
After cross validation	0.95	0.95	0.94	0.95
After GridSearch	0.95	0.95	0.95	0.95

The table above shows that decision tree obtained the best results after implementing GridSearch, which finds the best parameters to fit the model with. It was found that the best parameters were max_depth=12.

4.2 Discussion

Overall, models obtained the best results when GridSearch is implemented. The model with the best result is K-Nearest Neighbours (KNN), as it obtained a better recall score of 1.0, as compared with the 0.95 recall score obtained by random forest, which is the next best model overall. Recall is the performance metric that ultimately differentiates the results obtained by these two models as recall is prioritized in this study. This is due to the importance of false negatives (FN) preceding the importance of false positives (FP) when predicting stroke (Krish Naik, 2020). On the confusion matrix, FN is when the actual value is 1, but predicted value is 0. In the context of predicting stroke, it means the situations when a person actually has stroke, but the model predicts that they don't. This leads to disastrous results as the patient would not have the correct information that they could have stroke. Hence, they would not have taken the steps to improve their health, receive treatment and prevent the stroke from occurring.

In accuracy, KNN also performed best. However, due to the dataset being imbalanced, other performance metrics must be included to provide more insights to the performance of each model. KNN also performed best for recall. In precision, random forest performed best. Precision focuses on FP, which in this study is when a person actually does not have stroke but is predicted to have stroke by the model. Giving a patient this false prediction could result in excessive worry that could be prevented. While this error is important to identify, FN is significantly more so. Lastly, KNN also produced the best F1-score. This metric is used to consider when recall and precision are equally as important, that is, when FN and FP should both be considered (Krish Naik, 2020).

5. Limitations and Future Study

In this study, we have faced a few limitations that can be improved to conduct a better future study. The first limitation is limited classification models. We only used four classification models for our study which are SVM, KNN, Random Forest and Decision Tree. More classification models can be used for future studies to analyse, better understand and compare the data to determine whether there is any better classification model that can provide more precision than the classification model that we used in this study. Besides, the uneven distribution of the dataset is also one of the limitations. For example, the dataset for this study had substantially fewer men than women. Thus, in order to further test and validate the performance of the prediction model a more balanced dataset can be used for further study. Other than that, using retrospective studies is also a limitation of this study. This is because we do have not enough time to collect data ourselves. Retrospective studies have an inferior level of evidence, and they might not have complete data compared with prospective studies. Hence, for future study, using prospective studies would be better since it has more complete data than retrospective studies.

6. Conclusion

To sum up, we have used four different machine learning algorithms which are Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree as well as Random Forest to develop a more accurate prediction model of the prediction of stroke using the datasets we found in this report. Random forest is the most accurate model to predict the risk of having a stroke which contains an accuracy of 97%.

Based on the result, we can find out who is the majority of the victim suffers from a stroke. Females who gets into stroke are slightly higher than males. Besides, people in the age range of 60 to 80 or with a glucose level higher than 104 are easier to suffer from stroke. Also, people who stay in an urban area have a higher risk to get a stroke compared to people who live in a rural area. Last but not least, those BMI between 27 to 31 or who have a smoking habit have a higher risk to get into a stroke.

In short, we should use different classification models if we want to do further predictions in the future to compare the results so that we are able to find out the best fit model among them which can provide the highest accuracy and best result.

7. References

1. 2U, 2022. *What Is Undersampling?*. [Online] Available at: https://www.mastersindatascience.org/learning/statistics-datascience/undersampling/
2. Bajaj, A. (2022) Performance metrics in machine learning [complete guide], neptune.ai. Available at: https://neptune.ai/blog/performance-metrics-in-machine-learningcomplete-guide (Accessed: November 23, 2022).
3. Brownlee, J. (2020) How to calculate precision, recall, and F-measure for imbalanced classification, MachineLearningMastery.com. Available at: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalancedclassification/ (Accessed: November 23, 2022).
4. DecisionTree hyper parameter optimization using grid search - (no date) ProjectPro. Available at: https://www.projectpro.io/recipes/optimize-hyper-parameters-ofdecisiontree-model-using-grid-search-in-python (Accessed: November 23, 2022).
5. Dritsas, E. and Trigka, M. (2022) “Stroke risk prediction with machine learning techniques,” Sensors, 22(13), p. 4670. Available at: https://doi.org/10.3390/s22134670.
6. Federico Soriano Palacios. (n.d.). Stroke Prediction Dataset (Version 1) [Dataset]. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
7. H2O.ai, n.d. Overview — Using Driverless AI 1.10.3.1 documentation. [Online] Available at: https://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html
8. King, S.B. (2022) “‘states' rights’ for healthcare?,” Cardiovascular Revascularization Medicine [Preprint]. Available at: https://doi.org/10.1016/j.carrev.2022.09.008.
9. Indicator metadata registry details (2022) World Health Organization. World Health Organization. Available at: https://www.who.int/data/gho/indicator-metadataregistry/imr-details/158 (Accessed: November 23, 2022).
10. Li, Y. et al., 2022. Combining machine learning with radiomics features in predicting outcomes after mechanical thrombectomy in patients with acute ischemic stroke. Computer Methods and Programs in Biomedicine, Volume 225.
11. Liu, J. et al., 2022. Machine learning algorithms identify demographics, dietary features, and blood biomarkers associated with stroke records. Journal of the Neurological Sciences, Volume 4440.
12. MLTut, 2020. K Fold Cross-Validation in Machine Learning? How does K Fold Work?. [Online] Available at: https://www.mltut.com/k-fold-cross-validation-in-machine-learning-howdoes-k-fold-work/
13. N;, T.K.S.V. (2022) Stroke burden in Malaysia, Cerebrovascular diseases extra. U.S. National Library of Medicine. Available at: https://pubmed.ncbi.nlm.nih.gov/35325896/ (Accessed: November 23, 2022).
14. Omae, T. (1992) “Stroke risk factors and stroke prevention,” Journal of Stroke and Cerebrovascular Diseases, 2(1), pp. 45–46. Available at: https://doi.org/10.1016/s10523057(10)80035-7.

15. Precision versus recall - essential metrics in machine learning (2022) Graphite Note. Available at: https://graphite-note.com/precision-versus-recall-machine-learning# (Accessed: November 23, 2022).

16. Sung, S.-F.et al., 2015. Developing a stroke severity index based on administrative data was feasible using data mining techniques. *Journal of Clinical Epidemiology,* 68(11), pp. 1292-1300.

17. Krish Naik. (2020, January 27). *Tutorial 34- Performance Metrics For Classification Problem In Machine Learning- Part1* [Video]. YouTube. https://www.youtube.com/watch?v=aWAnNHXIKww