**TEMASEK POLYTECHNIC**

**SCHOOL OF INFORMATICS & IT**

**AY2022/2023 APRIL SEMESTER**

**DIPLOMA IN APPLIED ARTIFICIAL INTELLIGENCE**

**MACHINE LEARNING FOR DEVELOPERS (CAI2C08)**

Project Report

|  |  |
| --- | --- |
| Full Name | Admin No. |
| Zhang Haorui | 2101806D |

Shape

Description automatically generated with medium confidence

# **Introduction**

The topic selected is neuro health. According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

The dataset is Stroke Prediction and is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

**Data Dictionary**

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever\_married: "No" or "Yes"

7) work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"

8) Residence\_type: "Rural" or "Urban"

9) avg\_glucose\_level: average glucose level in blood

10) bmi: body mass index

11) smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"\*

12) stroke: 1 if the patient had a stroke or 0 if not

*\*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient.*

# **Data Exploration and Pre-processing of data**

## **Data Exploration**

### Insights 1

Chart, pie chart

Description automatically generated

From the pie chart above, 96% of participants did not have encountered stroke before while only 4% of participants encountered stroke.

As such, participants who did not encounter stroke are well represented. However, participants who had stroke are not well represented as it only represents 4% of the overall participants in the data. In this case, of 5110 participants involved in the dataset, only around 204 participants had stroke.

In conclusion, with no good representation of participants who had stroke, it may lead to biased data as the small range of participants who had stroke may not accurately let the machine fully train and familiarise participants who had stroke, as the machine did not train enough data rows that has participants who had stroke.

### Insights 2

Chart, box and whisker chart

Description automatically generated

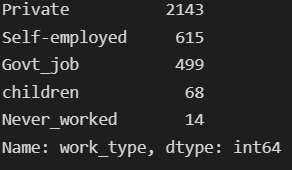
From the box plot provided above, the BMI range of participants involved in the dataset ranges from BMI of around 10.3 to around 97.6.

However, there are many outlier values where BMI is larger than 47.5. There might be a possibility that the participants could have mistaken their BMI for their weight.

As it could be a mistake on the recording side of the data, the columns of bmi where BMI is larger than 47.5 will be removed to prevent the machine learning model to train too much extreme values.

### Insights 3

Chart

Description automatically generated with low confidence

As seen from the histogram provided above, more than 2000 participants’ occupations are Private in the work\_type column, meaning that most participants work in the private sector.

However, on the other end, only 14 participants’ occupations are Never\_worked in the work\_type column, meaning that very few participants did not work before. It is not well represented to combine columns of Private, Self-employed, Govt-job, children, Never\_worked into Worked (3325 rows), and compare it with Never\_worked (14 rows) as there are too few data for the machine learning models to train the data.

As such, one-hot encoding of the work\_type into individual columns, and then performing feature selection will help out in finding which type of work\_type has a significant impact on the testing column.

## **Data Pre-processing**

### Missing Values

A picture containing text

Description automatically generated

As shown above, the bmi column has 201 missing values in the dataset.

As these values are missing completely at random due to errors caused by machines or humans, the rows with the missing BMI values will be deleted from the dataset.

Imputing the missing values with the mean or median of the BMI values will be biased as there are assumptions made about the missing values. Thus, the removal of missing values will be fair and not biased.

### Removal of Outliers

Chart, box and whisker chart

Description automatically generated

As seen from the box plot above and mentioned earlier, there are many outliers seen in the bmi column in the dataset.

Since there are many outliers in the dataset, the rows containing outlier bmi values will be removed so that no assumptions are made on the dataset, which may result in biased data.

Any BMI value higher than the equation q3+1.5\*iqr will be removed as it is considered a mild outlier.

After removing the missing values and outliers, there are 4799 rows left in the dataset.

### Binning

Graphical user interface, application

Description automatically generated with medium confidence(Before) Graphical user interface, application

Description automatically generated(After)

Since there are only 2 unique values in the is\_married column (yes or no), binning can be performed in this column.

As most machine learning models can only read numerical attributes, the categorical data in the is\_married column can be replaced with a Boolean data type of true or false value only.

This simplifies the data and allows the machine learning algorithm to understand the true/false values which represent 1 and 0 respectively.

### One-hot Encoding

#### **Residence\_type & gender columns**

For the columns in Residence\_type and gender, there are 2 unique values found in the dataset. Residence\_type contains rural and urban attributes and gender contains male and female attributes.

Binning will not be considered for these two columns as Boolean datatype does not bring in any meaning to these attributes. As such, numerical representative datatype will be considered instead of Boolean representative datatype.

In the Residence\_type column,

* 0 represents Rural residency,
* 1 represents Urban residency.

Graphical user interface, application

Description automatically generated(Before) Graphical user interface, application

Description automatically generated(After)

In the gender column,

* 0 represents Female,
* 1 represents Male.

A black screen with white text

Description automatically generated with low confidence(Before) Graphical user interface, application

Description automatically generated(After)

#### **Smoking\_status, work\_type columns**

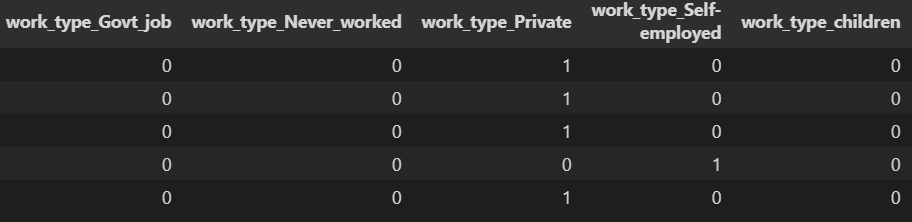
For the columns in smoking\_status, work\_type, one-hot encoding is useful for replacing categorical data with simple numeric data. This allows the machine learning models to better understand the data.

However, as there are more than 2 unique values in these two columns, representing numerical attributes will not work in these columns.

Many numerical attributes will let the machine learning model think that the order of those numbers is significant as the model assumes that the bigger numbers are more important than the smaller numbers.

As such, the proposed method is to create new columns for the different unique values. Adding the columns require more data rows, if not, the dataset may encounter the curse of dimensionality. However, in this case, there are 17 columns after creating the new columns and hence the bare minimum required rows would be 170 rows. The dataset has 3339 rows. Moreover, feature selection will be performed later to see which columns have a significant impact on the predicting column (stroke).

Graphical user interface, text, application

Description automatically generated(Before) (After)

Graphical user interface, application

Description automatically generated(Before) Graphical user interface, application, Teams

Description automatically generated(After)

### 

### Data Sampling

|  |
| --- |
| A picture containing text, shoji, clipart, tiled  Description automatically generatedChart, bar chart  Description automatically generated(After) |

A picture containing calendar

Description automatically generated (Before)

As seen in the stroke, hypertension and heart\_disease columns, these data columns have a contrasting difference in the rows of data between the 0 and 1 values.

Participants who have (a) stroke, (b) hypertension, (c) heart disease are very few compared to those participants who do not have the medical condition or illness in (a), (b), (c).

This means participants who had stroke, hypertension and heart disease are not well represented in the raw data and the machine learning model may not be able to capture such data quickly due to the very limited rows of such data.

This results in inaccurate model learning as there are not enough rows of data for the machine model to train for. As a result, the model will not be able to classify well which features will determine if a participant has stroke, or whether hypertension or heart disease are determining factors in causing a participant to get stroke.

Thus, random over-sampling may help in such data that are not well represented, where it randomly duplicates minority classes to ensure that data is more balanced with participants who have (a), (b), (c).

Text

Description automatically generated

As such, X will represent the other features to train and test in the dataset and y will represent the stroke feature to predict.

### Correlation Metrics

Table

Description automatically generated (Part 1)Table

Description automatically generated(Continue. Part 2)

The features in the dataset are not strongly correlated to each other.

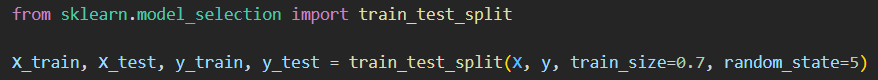
The column that is most correlated to the stroke column will be age (approx. 0.24), which means that the relationship between age and stroke column is the strongest among the other columns. This is further proven in performing the feature selection.

# **Methods and Improvements**

The column selected for prediction is the stroke column, which contains numeric data (1/0). As the predicting column (stroke) only contains two numerical unique values, machine learning classification models will work well for this dataset.

Regression models will be ignored as it only works for predicting column that has continuous numerical data.

## **Machine Learning Models**



The dataset is split into 70% train and 30% test and returns 4 variables calling the train\_test\_split (random\_state=5) function. Random\_state=5 allows us to all have the same reproducible effect and to ensure that the train/test values are the same.

* **Decision Trees**

Default values used. Random state is set to 1.

* **Logistic Regression**
* Default values used. Random state is set to 1.
* **Random Forest Classifier**

Default values used. Random state is set to 1.

* **K-Nearest Neighbour Classifier**

Default values used. Random state is set to 1.

* **Ensemble Learning – Gradient Boosting Classifier**

Default values used. Random state is set to 1.

The model trained is Random Forest Classifier as it has the best accuracy score of 0.99 compared to the other models. No overfitting of models occurred as the training set accuracy score (1) is very close to the testing set accuracy score (approx. 0.99). With such extremely low variance, this means that the model trained and can predict the dataset very well.

Text

Description automatically generated

## **Feature Selection**

Since there are a few columns in smoking and occupation, feature selection is necessary to find out if which column holds more importance than the other columns in predicting the test column (stroke).

Text

Description automatically generated

As seen from the result of feature importance above, the feature, age, has the most prominent impact in predicting whether a participant has stroke, while the feature, work\_type\_Never\_worked, has no impact on model prediction.

With that said, I will select the top 9 columns that have a significant impact on the predicting value.

The selected columns are

* age,
* avg\_glucose\_level,
* bmi,
* hypertension,
* residence\_type,
* heart\_disease,
* gender\_code,
* is\_married,
* smoking\_staus\_never smoked.

# **Results and Analysis**

Random Forest Classifier is used in the end as this model has the best accuracy score as compared to the other models.

As such, Random Forest Classifier is the best model for accurately predicting whether a person has stroke, based on the features in the X train variable.

Moreover, this is also proven by the mean absolute error of the training (0.15) and testing (0.13) sets, which shows a minor difference of approximately 0.02 between the two sets.

# **Conclusion**

In conclusion, the raw dataset provided is skewed towards participants who did not have stroke, hypertension, heart disease. Hence, sampling is required to overcome skewness and ensure that the dataset is evenly distributed.

When first run without sampling, the model did not capture the minority data rows completely and even took the minority data rows as outliers.

When running the Random Forest Classifier model, I was shocked at its training set accuracy, which is 1. For a moment, I thought that the model is overfitted.

However, after seeing the difference between the training set (1) and the testing set accuracy (approx. 0.99) of 0.01, it was concluded that the model is not overfitted as the testing set accuracy score is extremely close to the training set accuracy score. Hence, the variance between these two sets is very low, meaning that the model is not overfitted.

Moreover, the model’s F1 score is very good at approximately 0.99, meaning that the dataset is balanced and the model is good at predicting stroke.

Lastly, the use of grid search was not necessary since the training and testing set accuracy scores for Random Forest Classifier are very good at 1 and 0.99 respectively, meaning that the training of the dataset is well trained and that the model can accurately predict the predict column, which is stroke.

**References**

<https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/>

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

<https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/>

<https://towardsdatascience.com/one-common-misconception-about-random-forest-and-overfitting-47cae2e2c23b>