Vietnam National University HCM
International University
School of Computer Science and Engineering

PROJECT REPORT

Introduction to Data Mining

Topic : Stroke Prediction

	Student Name	Student ID
1	Trần Đình Khôi Nguyên	ITITIU19166
2	Võ Vương Nhật Tân	ITITIU19205
3	Đặng Quang Vinh	ITITIU19247
4	Phan Ngọc Đông Minh	ITITIU20252
5	Hàng Huỳnh Công Thuận	ITITIU20021
6	Dương Trần Nhật Minh	ITDSIU20032

Table of Content

Introduction

Stroke is a leading cause of death and disability worldwide, which leads to a crucial need for improvement in stroke diagnosis and early signs detection. Early stroke prediction and intervention can significantly reduce the chance of irreversible damage or death. Thus, it is crucial to study the interdependency of factors that could relatively contribute to stroke. In this project, we applied data mining techniques to develop a stroke prediction model using a training dataset of over 5100 patients. We followed thoroughly the steps of data pre-processing, training, testing, fine-tuning and evaluating 2 classification algorithms, including Naïve Bayes and Logistic Regression, using 10-fold cross-validation technique.

Methodology

I. Exploratory Data Analysis

Before beginning the pre-processing step, it is first important to analyse and understand the dataset's characteristics. In this step, we looked at the data type, some descriptive statistics including missing values, mean, mode, median, correlation, …

1. Data type

We first let the program read the CSV file and use the info() function to understand the variable types of each column, and whether the columns held nominal or numerical values. We also checked for missing values and duplicated rows.

The dataset consists of 5110 instances, with 10 attributes, and 2 classes: stroke or no stroke. The attributes consists of numerical and nominal values. The attributes hypertension, heart_disease and stroke class in the original dataset are listed as a numerical value, although in fact, are categorical data. and should be converted to nominal values in the later preprocessing steps.

There are 201 missing values in the BMI column, and 1544 ‘Unknown’ in the smoking_status column and no duplicated values, suggesting that we should thoroughly examine the BMI and smoking_status column and its correlation to handle the missing values.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64  
 4   heart_disease      5110 non-null   int64  
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   int64
```

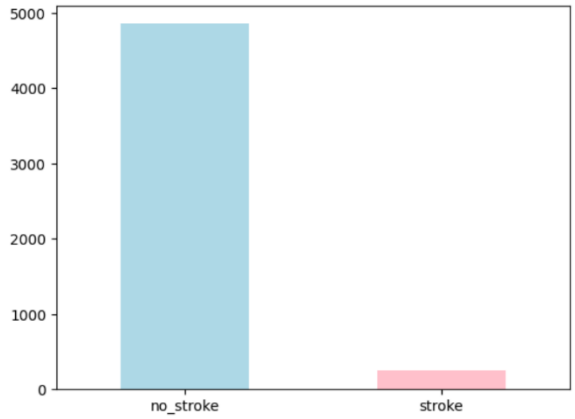
```
id :  Numerical
gender :  Nominal
age :  Numerical
hypertension :  Numerical
heart_disease :  Numerical
ever_married :  Nominal
work_type :  Nominal
Residence_type :  Nominal
avg_glucose_level :  Numerical
bmi :  Numerical
smoking_status :  Nominal
stroke :  Numerical
```

```
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      1544
stroke                 0
```

2. Statistical descriptor

The number of stroke vs non-stroke patients:

It is important to note that the number of no-stroke instances in this dataset accounts for 4861 out of 5110 íntances while stroke instances only account for 4.872% of the dataset. Thus, any model that always predicts no-stroke tested on any random sample of this dataset will likely achieve more than 95%. As the scope of our project is small, we will ignore this imbalance.

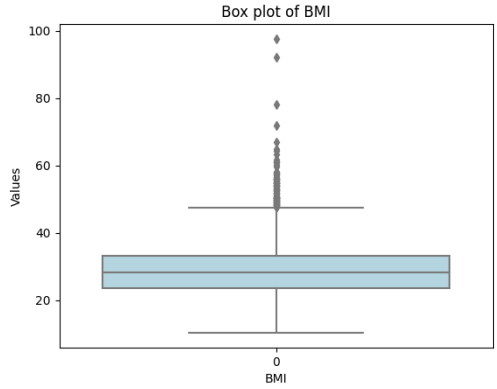


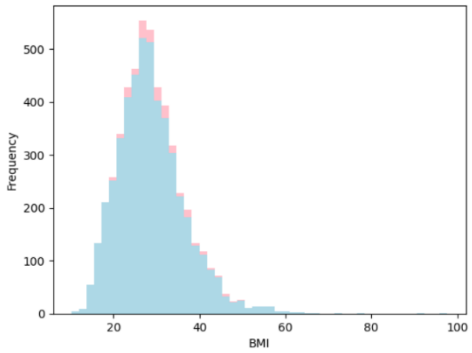
a. Univariate numerical variables analysis

The numerical data is summarized in the table below:

	age	avg_glucose_level	bmi
count	5110.000000	5110.000000	4909.000000
mean	43.226614	106.147677	28.893237
std	22.612647	45.283560	7.854067
min	0.080000	55.120000	10.300000
25%	25.000000	77.245000	23.500000
50%	45.000000	91.885000	28.100000
75%	61.000000	114.090000	33.100000
max	82.000000	271.740000	97.600000

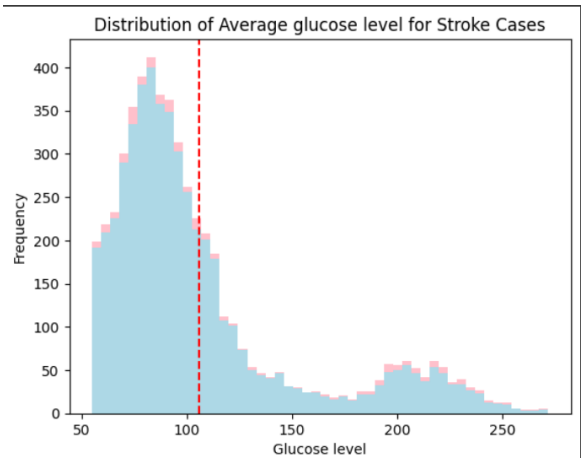
As we need to find the missing values of the BMI attribute, it is crucial to carefully analyse the box plot and distribution. BMI is the measure of body fat based on height and weight. In this dataset, BMI value ranged mostly from 15 to 45, with the mean of 28.89. The distribution of the BMI values appears to be slightly positively skewed with 110 outliers and extreme values on the scale's upper end. It can also be observed from the distribution that there is a correlation between higher BMI and the occurrence of stroke. Specifically, the data suggests that stroke are more likely to occur in patients with high BMI.





For average glucose level, the distribution shows a bimodel pattern, with the majority of the instances clustered at the first peak at 80 – 100mg/dL. The data is positively skewed with a mean of 106.147mg/dL and a standard deviation of 45.283mg/dL. The histogram does not reveal a clear relationship between glucose level and stroke occurences.

For age distribution, we can observe a normally distributed histogram ranging from 0 to 82 years. The histogram also implies a high frequency of stroke occurrence is correlated with higher age group.

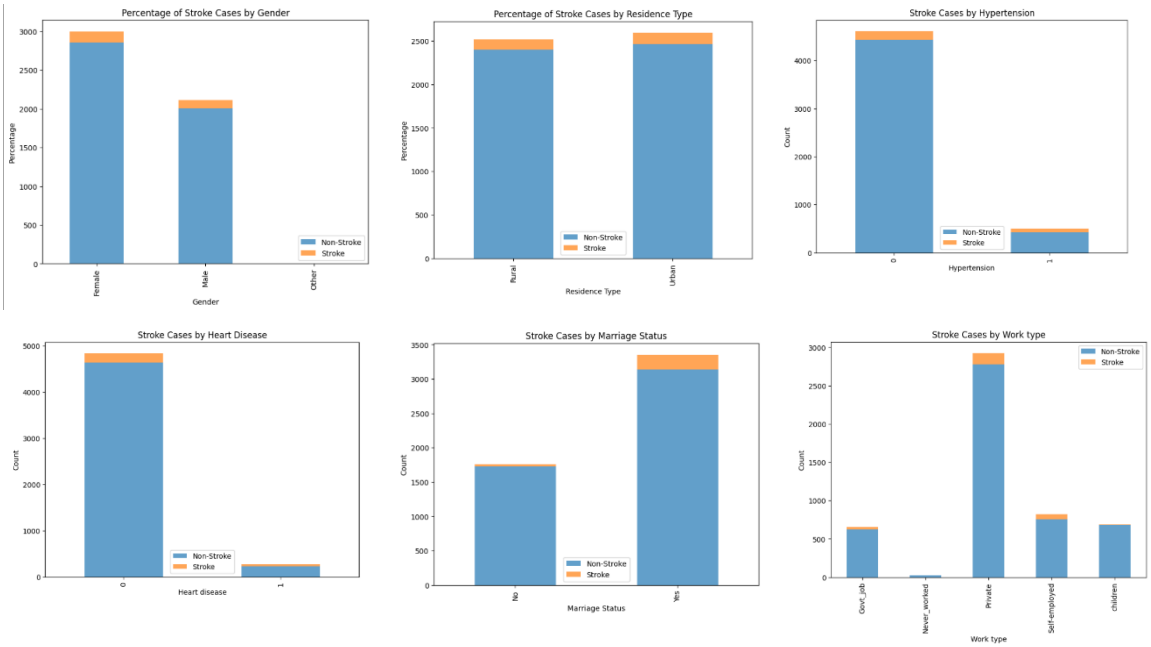


b. Univariate categorical variables analysis

Below are the plots that show certain features are linked with more stroke occurrences than others. Stroke occurs in 4.71% of females and 5.11% of males, suggesting the same risk for both genders. Similarly, the residential area does not seem to contribute significantly as a factor of stroke.

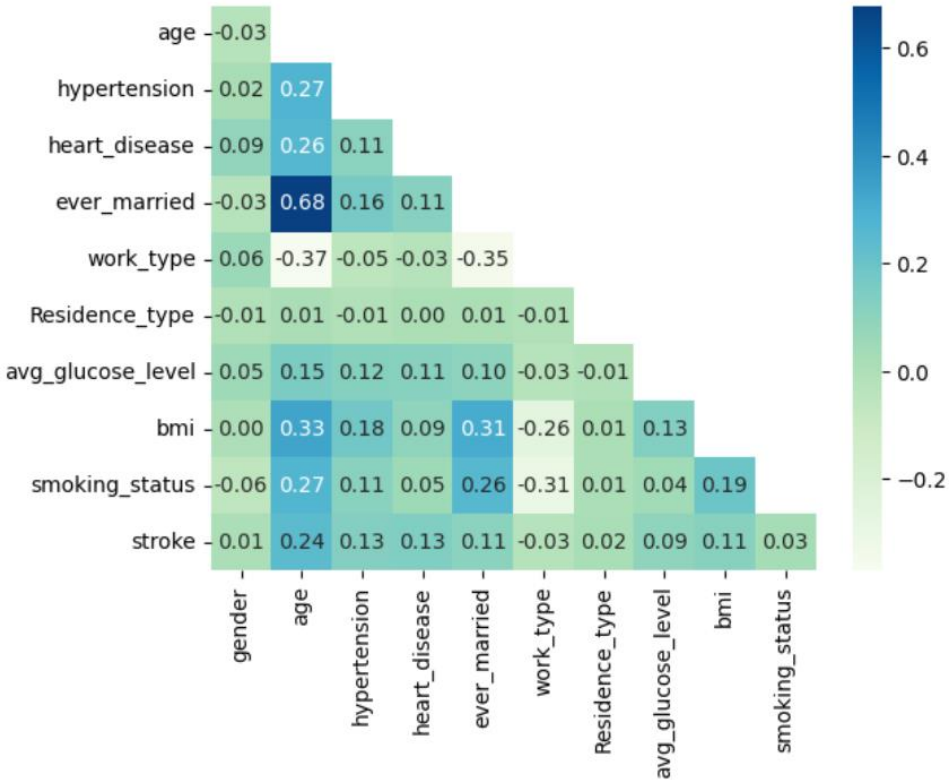
For marital status, the percentage of stroke occurrence is higher in the married group (6.56%) compared to the never-married group (1.65%). In terms of work type, children and people who have never worked have a low stroke occurrence of 0.29% and 0% respectively while the self-employed group has the highest stroke occurrence of 7.94%.

The risk of stroke for people with hypertension (13.45%) is comparatively high than the risk of stroke in the non-hypertension group. Similarly, the risk of stroke for people with heart disease is also a significant factor, as nearly 17% of the heart disease proportion experience stroke whereas only 4.18% of the people with no heart disease experience stroke. However, it is important to note that the occurrence of stroke is skewed towards those with hypertension and heart disease could be due to the imbalance of data.



c. Features correlation

The correlation matrix demonstrates a prominent positive correlation between stroke occurrence with age, hypertension, heart disease, marital status, and BMI. It is important to consider the correlation between the attributes for later fine-tuning processes.



II. Data Pre-processing

This process involved analysing the dataset for duplicates, missing values, and outliers. During exploratory data analysis, we found 201 missing values in the BMI column and 1554 'Unknown' instances in the smoking_status column.

Before replacing the missing BMI values, background research was conducted on BMI values. According to the CDC, the BMI of an average healthy person ranges from 18.5 to 24.9, while the mean of our dataset was 28.89. Further analysis of the dataset revealed 110 outlier instances with the smallest and largest outlier value of 47.9 and 97.6. These extreme values are highly unlikely to occur in real-life situations and could be a contributing factor to the positively skewed data. Thus, to handle the missing data, we replaced the 201 missing values with the median of the BMI distribution. The median was chosen as it is a measure of central tendency that is less affected by outliers.

Next, we considered the missing values of the smoking_status column. As the number of missing values accounted for nearly 30% of instances, we decided to keep 'Unknown' as a new category as imputing the data may introduce bias.

After handling the missing values, we split the data into training set and test set as well as converting them to .arff format in preparation for the models.

III. Implementation

After having cleaned the data, we built 2 models for stroke prediction: Naïve Bayes and Logistic Regression. We first loaded the .arff files of dataset to our models and initialized the training data instances. Both models were called from Weka Library API to and trained with the training set.

Naïve Bayes Classification algorithm is a probabilistic classifier which operates based on Bayes' Theorem. The algorithm assumes that all attributes are conditionally independent. We decided to use Naïve Bayes as it is computationally efficient and typically performs well with skewed data.

The Logistic Regression algorithm is a typical of binary classification algorithm which can describe the relationship between one dependent binary variable and one or more independent variables. As we have defined in the exploratory data analysis, the dataset is noisy and has many outliers, thus the logistic regression model is a suitable choice.

After training and testing our model, we found that the models were not performing particularly well. Therefore, we decided to reconduct the exploratory data analysis, in which we found some attributes were not highly correlated to the stroke classification, but rather, highly correlated among themselves. Therefore, we decided to fine-tune the algorithm by dropping 4 columns that were irrelevant including: gender, work type, residence type, and smoking status. The model results a better performance.

Evaluation

Both models are evaluated with 10-fold cross validation. The result is represented as follow:

	Naive Bayes	Logistic Regression
Accuracy	93.4%	95.8%
F-measure	97%	98%
RMSE	22%	19%
MAE	10%	7%

The Naive Bayes (NB) algorithm is a generative model, whereas Logistic Regression (LR) is a discriminative model. Typically, NB performs well on small datasets, while LR can achieve similar performance with regularization.

After applying both algorithms to the dataset at hand, it was found that Logistic Regression outperformed Naive Bayes in terms of accuracy and F-measure. However, it is noteworthy that the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) metrics of Naive Bayes were substantially lower than those of Logistic Regression.

Conclusion

In conclusion, we have successfully analyzed the stroke dataset, performed various preprocess steps in attempt to clean the data. We explored the correlation between the attributes and identified the most significant factors contributing to the occurrence of stroke being age, hypertension and heart disease. We have implemented 2 machine learning models to predict the stroke occurrences: Naïve Bayes and Logistic Regression. Both models was successful although the Logistic regression outperformed the Naïve Bayes in terms of accuracy and F-Measure.

However, we acknowledge that our dataset was imbalanced, with only 4.8% of the samples belonging to the positive class, leading to a bias in our models. For future work, it would be beneficial to gather a more balanced dataset with fewer outliers to improve the accuracy and reliability of our predictions.