Czech University of Life Sciences Prague

Faculty of Economics and Management

Department of Statistics

Bachelor Thesis

Predictive analysis of stroke case

Gao YuFeng

Declaration

I declare that I have worked on my bachelor thesis titled "Predictive analysis of stroke case" by myself and I have used only the sources mentioned at the end of the thesis. As the author of the bachelor thesis, I declare that the thesis does not break any copyrights.

In Prague on 09.03.2022 高域峰

Acknowledgement

I sincerely thank my supervisor Ing. Tomáš Hlavsa, Ph.D., for his guidance and assistance during my work on this thesis. I would also like to thank my parents for their support and trust in me all the time.

Predictive analysis of stroke cases

Abstract

The theoretical part of the thesis begins with our definition of stroke and the combined influencing factors that lead to it. The concept of big data is then explained and we introduce the basic concepts and principles of logistic regression algorithms for big data, before focusing on the development of statistical data processing and the basic methods of statistical and predictive analysis.

In the practical part of the work, we dealt with a large dataset of 5100 objects, a collection containing patient-related information placed on the kaggle online platform for educational purposes.

The practical work involved creating appropriate classification algorithms and using analytical tools to identify the factors that influence stroke. Based on these factors, logistic regression models and decision trees were constructed to predict strokes. Finally, the predictive power of the model was validated using the analytical tools and the application of new data.

Two data prediction models were developed using logistic regression and decision tree algorithms encapsulated in SAS studio to model data based on a dataset of stroke patients. The two data models show that four factors - age, average blood glucose level, hypertension and heart disease - are risk factors that influence stroke. Of these, age is the most important influencing factor. By comparing the two prediction models, a data model with a prediction accuracy of 84% was created, which can provide a reliable reference for stroke prediction.

Keywords: Big data, stroke, statistical analysis, predictive modeling

Prediktivní analýza případů mrtvice

Abstrakt

Teoretická část práce začíná naší definicí cévní mozkové příhody a kombinací ovlivňujících faktorů, které k ní vedou. Poté vysvětlujeme pojem big data a představujeme základní pojmy a principy algoritmů logistické regrese pro big data, načež se zaměřujeme na vývoj statistického zpracování dat a základní metody statistické a prediktivní analýzy.

V praktické části práce jsme se zabývali velkým datovým souborem o 5100 objektech, sbírkou obsahující informace týkající se pacientů umístěnou na online platformě kaggle pro výukové účely.

Praktická práce zahrnovala vytvoření vhodných klasifikačních algoritmů a použití analytických nástrojů k identifikaci faktorů, které ovlivňují cévní mozkovou příhodu. Na základě těchto faktorů byly zkonstruovány logistické regresní modely a rozhodovací stromy pro předpověď mozkových příhod. Nakonec byla ověřena prediktivní schopnost modelu pomocí analytických nástrojů a aplikace nových dat.

K modelování dat na základě souboru dat pacientů s cévní mozkovou příhodou byly vyvinuty dva modely predikce dat pomocí algoritmů logistické regrese a rozhodovacího stromu zapouzdřených ve studiu SAS. Tyto dva datové modely ukazují, že čtyři faktory - věk, průměrná hladina glukózy v krvi, hypertenze a srdeční onemocnění - jsou rizikovými faktory, které ovlivňují cévní mozkovou příhodu. Z nich je věk nejdůležitějším ovlivňujícím faktorem. Porovnáním obou predikčních modelů byl vytvořen datový model s přesností predikce 84 %, který může poskytnout spolehlivou referenci pro predikci mrtvice.

Klíčová slova: Big data, mrtvice, statistická analýza, prediktivní modelování

Table of content

List of pictures

List of tables

1 Introduction

In recent years, the generation and storage of data has increased geometrically due to the rapid development of technology. New data is being generated and stored every minute of every day. This growth also includes sources of data collection, such as popular web entries and search histories, information interactions on mobile devices and GPS satellite signals, to name but a few. Big Data is the term used to describe the vast amount of data that can affect our daily activities. With the enormous growth of data comes the challenge of getting the quality information we need from this vast amount of data. Big data analytics is the key to meeting this challenge. The Big Data market is relatively young but offers many possibilities for us to analyse and exploit this big data. Its applications and uses are also expanding over time. In the business sector, many companies are already using advanced data analytics to gain access to the vital information hidden in Big Data, enabling them to improve the management of their companies and gain a competitive edge in their operations.

There have also been amazing developments in the healthcare sector where, with the help of technology, we can identify trends in medical data through big data analytics. This analysis can help healthcare workers to make accurate predictions about any medical condition. This has led to improvements in medical conditions and reductions in the cost of treatment. It is fair to say that big data analytics has had a huge impact on the biomedical field. It has helped doctors to identify and control diseases at their early stages. Despite the new scientific advances in healthcare brought about by big data analytics in recent years, stroke remains a global public health problem and is one of the leading causes of death and disability in adults.

The focus of the paper is to use big data analysis to predict the occurrence of stroke and to identify the key factors required for stroke prediction and ultimately to determine which of the prediction models used is the best model for predicting stroke.

The paper is structured as follows. Section 3 describes the definition of stroke and the various factors that contribute to it through a literature review. It also explains the basic understanding of the definition of big data and data analysis and illustrates the challenges faced in big data analysis. Section 4 describes EDA analysis, correlation analysis and influencing factor importance analysis to screen for high risk factors affecting stroke and lists an introduction to various algorithms for data analysis. Section 5 details the data analysis algorithms used for predictive modelling and how they perform on the dataset. Finally, Section 6 concludes the paper and discusses future work.

2 Objectives and Methodology

2.1 Objectives

The huge amount of data generated every day is a requirement for big data characteristics. And in the medical field, human health testing and patient medical testing data is one of the big data generated every day in the world.

The main objective of this thesis is to identify risk factors influencing stroke.

While the above are the main objectives, there are a number of specific objectives to be achieved in this study:

1. to identify the key factors required to predict stroke.
2. to understand the correlations between the dependent and independent variables of stroke.
3. to use statistical predictive modelling techniques to predict stroke.
4. Comparing the performance of the models to arrive at the best model that is more accurate at predicting stroke

In the theoretical section, the basic concepts of big data are first introduced and a review of the literature on big data is presented. The concept of big data predictive analytics is presented and predictive analytics models are introduced. The theoretical part of the thesis also introduces the basic concepts of stroke and reviews the literature.

The practical part focuses on the preparation and analysis of the data and the construction of the predictive model. The basic information about the data set used in this study should first be presented and the data should be organised appropriately before proceeding with the analysis. The purpose of constructing the data model is to be able to identify and explain the important factors that influence stroke. The dataset used in the actual modelling is a publicly available predictive stroke dataset containing information on over 5,100 patients obtained from the Kaggle data site. In order to achieve the highest accuracy, different predictive modelling methods were used to provide the most satisfactory model.

2.2 Methodology

The methodology of the study was to first initially analyse the factors influencing stroke and screen out the influencing variables.

This was followed by a predictive analysis of stroke prediction data obtained from the kaggle data mining website. The dataset consisted of 5111 observations containing a dependent variable and 11 independent variables.

The specific methodology is described below.

- literature review of influencing factors
- Data pre-processing
- Data analysis
- Preparation of modelling data
- Building the model
- Evaluation

The first stage begins with a preliminary analysis of the influencing factors that would lead to stroke through a literature review to select the variables that have a greater impact on stroke.

The second stage will be the initial data collation of the original dataset and then obtaining organised, neat and high quality data. We can achieve data integration by recoding.

In the third stage of data analysis, interpretive data analysis is used and each variable after integration will be analysed independently through univariate statistical analysis. We also perform Spearman correlation coefficients and test for multicollinearity using tolerances and VIF. Through the second stage, we further screen and integrate the data, removing those variables that do not qualify.

The fourth stage, preparing the modelling data, aims to correct for previously identified multicollinearity problems and select eligible variables to prepare the data set for modelling.

The next stage, the modelling stage, involved the development of predictive models based on the previously collated dataset. Based on the literature review, two analytical methods, logistic regression and decision tree learning, fit the needs of this study and we decided to use these two models for modelling the data. Logistic regression is a classification algorithm that can be used to deal with binary classification as well as multivariate classification problems. It is a classical classification model commonly used in machine learning and data prediction. The coefficients derived from the logistic regression model are easy to understand and easy to interpret. Decision trees are a classical classification and regression algorithm in machine learning. Decision tree learning uses the CART method to construct a decision tree model of a binary tree with features selected on the basis of: the Gini index. This is because the CART algorithm is applicable to the question of whether the sample features are yes or no. In contrast to classification models such as neural networks, decision trees can be logically well interpreted and can handle both numeric and category types. The logistic regression model and the decision tree model in SAS analysis software were used to model the data and obtain predictive models respectively.

The final stage, evaluation, involved evaluating the previously developed forecasting models to assess whether their predictive power met the objectives of the study. During the evaluation phase, feedback is provided on whether the developed model meets the criteria and whether the prediction accuracy meets the criteria. In order to determine the best forecasting model, the following characteristics will be examined, the root mean square error (RMSE), the receiver operating characteristic (ROC) curve and the AUC curve.

3 Literature Review

3.1 Stroke

According to data released by (WHO, 2022), stroke has become the second leading cause of death in the world, accounting for 11 percent of all deaths worldwide, just five percent lower than the world's biggest killer, which is ischemic heart disease - 16% .

Traditionally, we think of stroke as a disease that belongs to the elderly. However, statistics show that the proportion of young people (under 20 years) and middle-aged people (20-64 years) affected by stroke has been increasing over the last 20 years (Tan, Ramazanu, Liaw, & Chua, 2022). Stroke rates are generally higher in people under 75 than in older people, particularly in lower income groups (Murphy & Werring, 2020). In addition, the incidence of stroke is increasing significantly in younger and middle-aged people and stroke should no longer be seen as a disease of the elderly. In other words, globally, the probability of stroke may have quietly shifted to younger people.

3.1.1 Definition

Strokes can be divided into two main categories: one caused by blockage of blood vessels (ischemic stroke) and one caused by hemorrhage (hemorrhagic stroke). Ischemia is due to an interruption of blood supply, while, hemorrhage is due to a ruptured or abnormal vascular structure in the brain. 80% of strokes are due to cerebral ischemia; the rest are due to hemorrhage.

Strokes occur when there is an interruption of blood flow to part of the brain. This kind of stroke, caused by interruption of blood flow, is an ischemic stroke and is the major cause of stroke in elderly patients (Sodeman & Sodeman, 2005).

Intracerebral hemorrhage (ICH) occurs as a result of bleeding from an arterial source directly into the brain substance. Although its relative frequency in patients with stroke is subject to geographic and racial variations, values between 5% and 10% are most commonly quoted (Kase, Mohr, & Caplan, 2004).

Both ischaemic and haemorrhagic strokes can cause abnormal brain function. Common stroke symptoms include inability to move some limbs, inability to speak, and inability to see beyond one side of the visual field. Patients are also affected by physical disability, cognitive impairment, etc (Tan, Ramazanu, Liaw, & Chua, 2022).

In addition, stroke is considered to be an acute event with a high mortality rate. However, if timely treatment is provided to the patient in a timely manner, the mortality rate can be reduced. The longer a patient receives treatment, the more irreparable damage is done to the brain tissue (Tan, Ramazanu, Liaw, & Chua, 2022). Mortality from stroke also tends to increase with age, and prevention is currently considered the best measure due to the ongoing lack of effective treatments.

According to the literature, we summarize stroke as, A series of conditions caused by a blockage (ischemia - reduced blood supply to the brain) or bleeding in the blood vessels supplying the brain, collectively known as a stroke, or we define it as the brain cell death is caused by ischemia in the brain.

3.1.2 Factors lead to stroke

Influencing factors. The causes of stroke are complex and varied, with hypertension being the most important risk factor for stroke. Other factors such as advanced age, hypertension, diabetes, hyperlipidaemia, overweight and obesity, high cholesterol, poor lifestyle habits (e.g. staying up late and smoking), lack of exercise, smoking and atrial fibrillation are all risk factors for stroke (Tan, Ramazanu, Liaw, & Chua, 2022). The following is a further description of the factors affecting stroke.

Classical risk factors for ischemic stroke onset, including ischemic heart disease, carotid artery disease, lipidemias, hypertension, obesity, tobacco use, atrial fibrillation, personal or family history of stroke, congenital heart defects, congestive heart failure, cardiac valve disorder, peripheral vascular disease (Raphael , Daniel, Santiago, Rose , & David, 2021).

Following a trial investigation, the results showed that high blood pressure, particularly high diastolic blood pressure, is a major risk factor for stroke. Abnormal glucose regulation, including diabetes, impaired fasting glucose, and impaired glucose tolerance, have been recognized as important risk factors for occurrence and recurrence of ischemic stroke in Europe and America. Abnormal lipid metabolism, obesity, and diabetes mellitus are known risk factors for ischemic vascular events. Heavy alcohol consumption increases the risk of stroke in men. There is a positive association between smoking and stroke risk, with the prevalence of stroke increasing with the intensity and duration of smoking. The prevalence of hypertension, diabetes, smoking, hyperlipidaemia and alcohol abuse as important risk factors for stroke is significantly influenced by lifestyle and nutrition (Jia, Liu, & Wang, 2011).

Stroke can modify risk factors, including high blood pressure, diabetes, dyslipidaemia, atrial fibrillation and other heart conditions, smoking, physical inactivity and obesity (Tatjana & Ralph, 2008).

Non-modifiable risk factors by (Stephen & David , 2020)**.**

Age: This is the most important factor in the risk of stroke. the incidence doubles every decade after the age of 55.

Gender: The risk of stroke is as high or higher in women than in men before the age of 65 because of the risk of pregnancy and use of oral contraceptives. At older ages, men have a slightly higher rate of stroke.

Modifiable risk factors by (Stephen & David , 2020)**.**

Hypertension: This is the most important modifiable risk factor for stroke overall. About half of all stroke victims have a history of hypertension. Even in those not defined as hypertensive, the higher the blood pressure, the higher the risk of stroke. This makes hypertension crucial for stroke prevention.

Diabetes: This is an independent risk factor for stroke. Stroke accounts for 20% of all deaths in people with diabetes.

Cardiac factors: cardioembolic infarction is the most severe ischaemic stroke subtype and has a high rate of disability and mortalit. The presence of atrial fibrillation increases with age. It accounts for 20-25% of strokes in patients >80 years of age.

Smoking: The risk of stroke is doubled by smoking. More than half of people who have a stroke have a history of smoking. Quitting smoking helps reduce stroke risk quickly

Hyperlipidaemia: The complex relationship between dyslipidemia and stroke. Elevated total cholesterol increases the risk of ischaemic stroke, while elevated high-density lipoprotein-cholesterol decreases the risk of ischaemic stroke. High blood lipid levels can cause blood vessels to become clogged, making it more likely to cause a stroke.

Alcohol consumption: Light alcohol consumption was associated with a lower risk of ischaemic stroke, while higher alcohol consumption was clearly associated with an increased risk of stroke. Alcohol consumption is linearly associated with the risk of ischaemic stroke. Elevated levels of alcohol in the blood vessels are associated with an increased risk of stroke.

Obesity: Much of the effect of body mass index on stroke risk is mediated by blood pressure and cholesterol. People who are physically active have a lower risk of stroke and lower overall stroke mortality compared to those who are inactive. Getting physical activity can help reduce the risk of stroke.

Therefore, 10 factors were initially identified in the dataset: Gender, age, hypertension, heart disease, marital status, type of work, type of residence, blood glucose level, body mass index and smoking status, are influential predictor variables for stroke.

3.2 Big data

Since the end of the 19th century and the beginning of the 21st century, science and technology have advanced rapidly, including the rapid development of information and Internet technology, which has also led to the rapid growth of digital data. The term "Big Data" was also born in the development of information and Internet technology.

The boundaries of Big Data are defined very vaguely, and generally for data that can no longer be processed by common means will be classified as Big Data. But Big Data is not that simple, but much more complex. Even many data files that do not require much storage space are still considered big data because of their complexity. At the same time, not all data files that require big data storage capacity are complex enough to be considered big data.

3.2.1 Definition

In the 1980s, the concept of "big data" appeared in The Third Wave by Toffler (1990), He described it as "the third wave of beautiful music," but does not define in further detail what Big Data was (Wang & Wang, 2021).

“The term ‘Big Data’, probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid 1990s, in which John Mashey figured prominently” (Diebold, 2012).

Whenever we say, "what is big data?" the first thing that comes to mind is definitely the sheer volume. (Laney, 2001) presents the three challenges of data management as volume, variety and velocity (or the "three Vs").

Big Data is defined by Gartner, Inc. as “Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making (Gartner IT Glossary, n.d.).”

“Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large data sets that are diverse, complex, and of a massive scale (Hashem, et al., 2015).”

In the literature show that big data has 3-5 characteristics; three of them: Volume, Velocity and Variety are common at all. Others Veracity and Value have not yet been agreed upon. Volume, Velocity, and Variety can be called the triple-V threshold. The Triple-V tipping threshold is the turning point where the datasets are transformed into big data.

3.2.2 Characteristics of Big data

Even without a uniform definition yet, big data have generally been considered to be characterized by the 5V, i.e., Volume, Variety, Velocity, Value and Veracity (Tang, Li, Du, Li, & Wu, 2022).

Volume refers to the size of the data. The size of Big Data is reported in terms of multi-terabytes and petabytes. 1024 terabytes in size is equal to a petabyte. As technology is updated, larger data sets can be collected in the future through more efficient storage technologies. From the other side for the type of data, or variety of data, which determines the meaning of "big" in it. For example, two datasets of the same volume data tables and videos may require different data management techniques (Gandomi & Haider, 2015).

Variety means the structural heterogeneity in a dataset. Technology updates have allowed different divisions of data types, structured, semi-structured and unstructured data.

- Structured data For example, relational databases and tabular data in Excel, structured data represents only 5% of the total amount of all data available (Cukier, 2010).
- Unstructured data Examples: text, images, audio, and video, which do not have a structure that can be analysed by machines (Gandomi & Haider, 2015).
- Semi-structured data is a type of data that falls between structured and unstructured data, the most typical example being a textual language called Extensible Markup Language (XML), which can be used to exchange data over the Internet (Gandomi & Haider, 2015).

Velocity means the efficiency of data generation per minute and second. With the advancement of technology, the use of devices such as smartphones and digital sensors has allowed for an incredible increase in the rate of data generation. High frequency data is starting to appear in a number of areas, most typically in the retail industry. The large amount of data generated by mobile service devices and mobile applications, such as geolocation, historical transaction history, etc., can optimize purchase recommendations and improve service levels for customers (Gandomi & Haider, 2015).

Value means Big Data has the characteristic of "low value density". Data in its raw form without any classification is generally useless, but by analysing the data will increase its original value and reach a higher value (Sivarajah, Kamal, Iran, & Weerakkody, 2017).

Veracity implies that there is an element of unreliability in the source of the data (Gandomi & Haider, 2015). For example, there is uncertainty about personal mood changes or personal emotions shared on Facebook social media. This involves human judgment itself, but the data itself has valuable information. Therefore by managing uncertain and imprecise data differently from normal data and analysing and utilizing it through specific analytic methods.

3.3 Models

Predictive analytics operates on big data, so that the key features of the models are said to be based on data derived from the algorithm itself rather than from the analyst's assumptions. In other words, the model is induced from the big data by the algorithm. The induction process can include identifying the variables in the model, defining the parameters of the model, the weights, coefficients in the model, or the complexity of the model. The coefficients or weights of the model and the shape of the model can be discovered by the algorithm.

Predictive modeling algorithms belong to the supervised learning group. This means that they try to find correlations between the input data and the target variables. The key goal of a predictive model is to discover the future values of the target variables based on the input data known today. The two main problems solved during supervised learning are classification and regression.

This chapter describes some of the statistical methods found in predictive analysis:

- Linear regression
- Decision tree
- logistic analysis
- Random forest
- Artificial Neural network

3.3.1 Linear regression

Regression analysis is a statistical technique which is usually used for the estimation of the relationship among different variables. In other words, we can use regression model to analyze the correlation among many variables having cause-effect relationships and based on our model to make a prediction for our data.

In the very beginning of the book named "Applied linear regression" by Weisberg (2005), regression is discussed with the fundamental graphical tool, a two-dimensional scatterplot. Based on the definition, the scatterplot of the response will be presented in the graph showing the relationship between dependent variable and independent variable.

The independent variables are also called as "predictors", and is shown as "Xi" in the graph; while the dependent variable is shown as "Y" in the scatterplot. Our aim is to understand how the value of Y change as X is varied over its range of possible values. Before we really get into the details of our sample, we can always first graphically look at our data to get a general picture of how our variables look like.

The following figure shows the very basic scatterplot of the linear regression model. As we can see in the figure that, the scatterplot is composed by a series of individual blue plots, and each plot represents corresponding value of “x” and “y”. If we want to know the prediction of a corresponding value of “y” based on a specific value of “x”, we need to understand the trend of the relationship between “x” and “y”, which is represented as a red straight line in the following graph. It is very clear to say that, the independent variable “x” has a positive relation with dependent variable “y” as the increase of x will cause the increase of y at the same time.

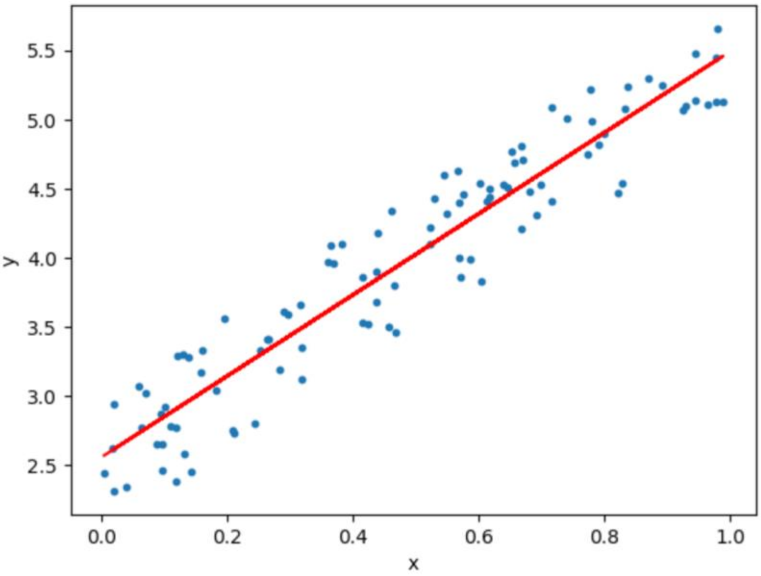


Figure 1 Linear regression (Agarwal, 2018).

The regression can be divided into two kinds, first one is using only one single independent variable, which is called “uni-variate regression”; and the other one is using more than one independent variable, which is called “multi-variate regression” (Tabachnick, Fidell, & Ullman, 2007). For practical reason, we commonly apply the later approach. However, we will introduce both the simple linear regression and the multivariate regression .

- Simple linear regression

The function of simple linear regression is expressed as follow:

$$E(Y|x) = \beta_0 + \beta_1 * x$$

Where E(), means the expected value, and we aim to find the possible value of Y when x is limited to some specific value. Beta zero β_0 is the intercept parameter, and beta one β_1 is the slope parameter. We usually call the parameter that can estimated the value in the model as “coefficient”.

The hypotheses of the simple linear regression is given as:

- $H_0: \beta_1 = 0$
- $H_1: \beta_1 \neq 0$

Ho is the null hypothesis, and if the null hypothesis is true, then we can say that based on our model, x has no effect on Y at all.

- Multivariate regression

The logic behind the multivariate regression is just like the simple linear regression one, we expect to find the best fit line of our model, but with more than one independent variable in the model (Abbott, 2014).

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i$$

Where Y is the dependent variable, and we have the X_i and different coefficient to explain Y. This function is actually quite similar to the simple linear regression, only adding more dependent variables in the model (Statistics Solutions, 2022).

Because we have more than one dependent variables in the model, the assumptions of multivariate regression is as follow:

- Linearity
- Normality
- No multi-collinearity
- No auto-correlation
- Homoscedasticity

3.3.2 Decision tree

Decision tree learning is widely used in data mining, statistics and machine learning. Decision tree learning is a decision model built using a tree structure based on the attributes of the input data to predict the value of a target attribute, and the output is a classification model. Each internal node in these trees represents an input variable respectively, the number of branches induced by the node is equal to the number of all possible values of that input variable, and the leaf nodes represent the final judgement outcome which is the value of the attribute. Decision tree learning can handle both continuous and categorical variables so it requires less effort in data preparation than other methods and is used here to predict the probability of a stroke.

According to the definition on The Economic Times, decision tree analysis is used to break down complex problems. Usually it involves making a tree-shaped diagram to chart out a course of action or a statistical probability analysis. We need to note here that the decision tree is not only used in finance and economic issues, but also in philosophy and machine learning problems. It is very useful to draw a final conclusion for the problems with many branches (The Economic Times, 2022).

There are five steps of decision tree analysis generally:

- Define the problem area for which decision making is necessary.
- Draw a decision tree with all possible solutions and their consequences.
- Input relevant variables with their respective probability values.
- Determine and allocate payoffs for each possible outcome.
- Calculate the Expected Monetary Value for every chance node in order to determine which solution is expected to provide the most value. Circles represent chance nodes in a tree diagram (Team Asana, 2021).

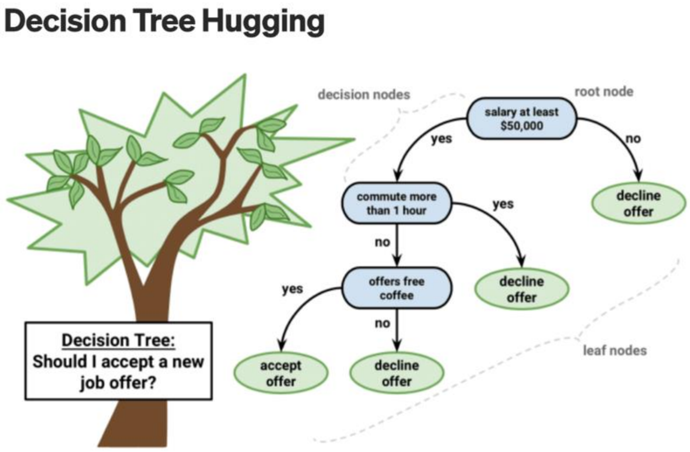


Figure 2 Decision tree (Gray, 2017).

For each of the decision tree algorithms described, the algorithmic steps are as follows (Abbott, 2014).

1. for each candidate input variable, evaluate the best way to split the data into two or more. Select the best way to split the data into subgroups defined by the splitting method.

2. Select one of the subgroups and repeat step 1 (this is the recursive part of the algorithm). Repeat for each subgroup.

3. Continue the segmentation until all records in the segmentation belong to the same target variable value, or apply another stop condition. The stopping condition can be a complex test of statistical significance or a simple minimum number of records.

CART analysis is a statistical method used to identify the explanatory variables that influence the target variable (Takeshi , et al., 2021). In the CART algorithm binary trees are constructed. The impurity measure is in the form of Gini index (Leszek , Maciej, Lena, & Piotr, 2014).

The CART algorithm, explains how to predict the value of an outcome variable based on other values. the output of CART is a decision tree where each fork is a fork of the predictor variable and each end node contains a prediction of the outcome variable.

We will now briefly describe the CART algorithm. The CART algorithm starts with a single node - the root. In the learning process, a specific subset Sq of the training data set S is processed in each created node. If all elements of the set Sq belong to the same class, the node is marked as a leaf and no segmentation is performed. Otherwise, the best attribute among the available attributes of the considered nodes is selected for segmentation according to the segmentation measurement function (Leszek , Maciej, Lena, & Piotr, 2014).

Any method has its advantages and its shortcomings, the decision tree model is not an exception as well. First we can talk about its strengths, it is very easy and direct to interpret the process of the whole decision-making. It is also quite valuable without requiring large amounts of hard data. Third, it really helps decision makers ascertain best, worst, and expected results for many kinds of scenarios. Last but not least, it can be combined with various decision models (Team Asana, 2021).

Decision tree learning algorithms are very efficient and scale well as the number of records or fields in the modeling data increase (Abbott, 2014).

Any extreme values are separated in small nodes and do not affect the classification problem (Tuffery, 2011).

On the other hand, there are some weaknesses of the decision trees model. First one we note here is that, even minor data changes could lead to major structure changes of our decision-making. Second that we mention is, information gain in decision trees can be biased. Third is uncertain values can lead to complex calculations and uncertain outcomes. Last but not least is, decision trees can often be relatively inaccurate (Team Asana, 2021).

3.3.3 Random forest

Random forests are a class of algorithms used to solve classification and regression problems, It is an integrated learning algorithm that is part of supervised learning. Commonly used for prediction, it is one of the most commonly used algorithms in machine learning. As ensemble methods, they grow several trees as base estimates and pool them together to make predictions. In order to obtain many different trees based on a single training set, the random forest procedure introduces randomness in the construction of the trees (Erwan , 2016). Random selection of data and random selection of features are the characteristics of the Random Forest algorithm.

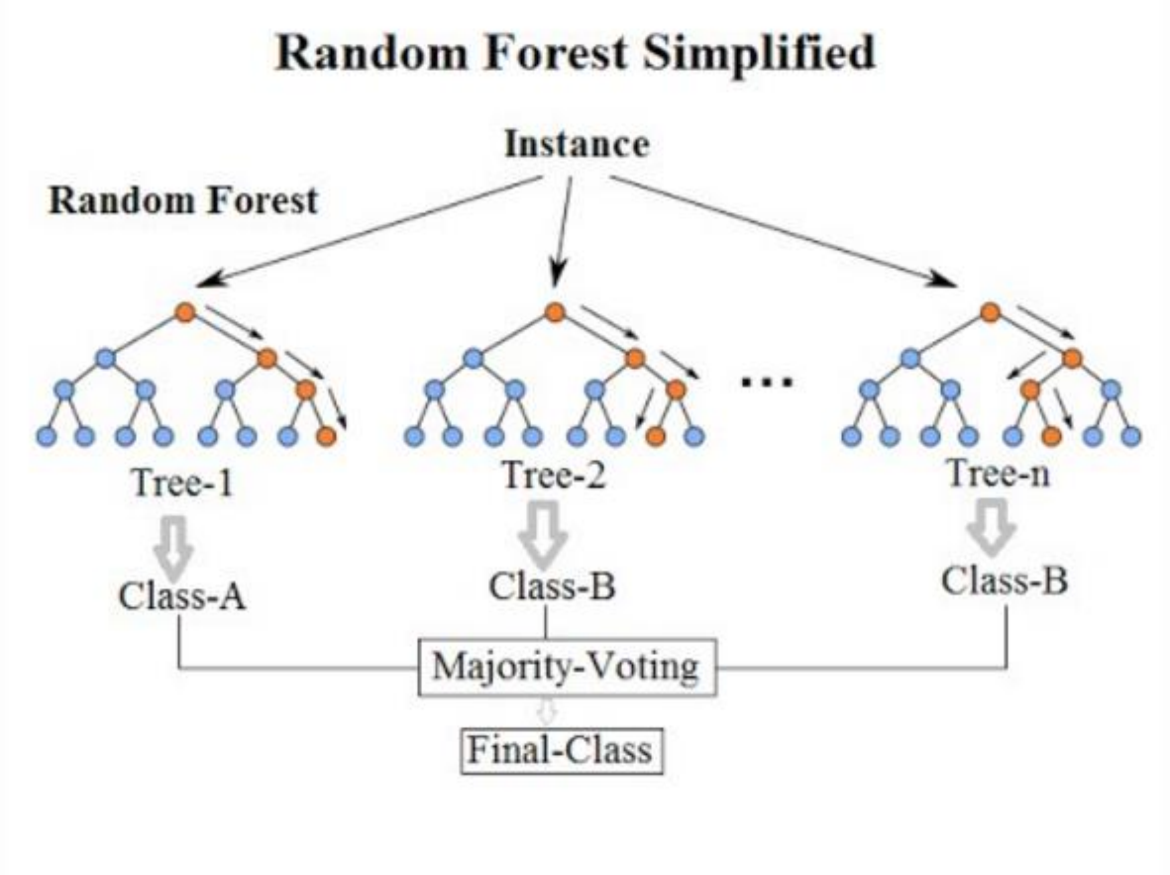


Figure 3 Random Forest Simplified (Venkata , 2020)

By an integrated classifier containing many randomly generated decision trees, several weak classifiers are combined to obtain a strong classifier with significantly superior classification performance.

The random forest algorithm requires a large variability among decision trees without any correlation with each other and no strong dependencies among the weak classifiers.

Random forest works by extracting multiple samples from the input large data, then the extracted samples are first trained with weak classifiers - decision trees, and finally these decision trees are combined together to produce the final classification or prediction result by voting. Random forests are not prone to overfitting and are somewhat resistant to noise.

Random Forest (RF) is an ensemble method based on decision (classification or regression) trees. Classification and regression trees (CART) are built by recursively partitioning the data according to covariates, so that the observations in the nodes become increasingly pure (in terms of outcomes) as the data move from the root to the terminal nodes. Each terminal node is usually summarized by the average outcome value of all the observations that last appeared there. There are two main differences between a tree in a random forest and a decision tree. As each tree is constructed in a random forest, at each internal node, only a random subset of covariates is evaluated for optimal partitioning. Trees in a random forest are usually fully grown and unpruned. The averaging process of a random forest greatly improves the accuracy of predictions over a single tree (Zhao, Su, Ge, & Fan, 2016).

For handling missing values, One of the main advantages of decision tree based methods is the ease to handle missing data. Unlike other modeling methods, it is not necessary to impute or delete any missing data before constructing trees in random forest (Zhao, Su, Ge, & Fan, 2016).

3.3.4 Artificial Neural Network

An artificial neural network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. ANNs are used for modelling non-linear problems and to predict the output values for given input parameters from their training values (M., 2010).

Neural networks are now the state of the art in machine learning for a variety of prediction tasks, including, but not limited to, facial recognition (1), robot navigation (2), and sentiment detection (3) (Emily , Andre , Cecile , & Mauricio, 2021).

Artificial neural network consists of a large number of nodes connected to each other. Each node represents a specific output function, called the activation function. Each connection between two nodes represents a weighted value for the signal passing through the connection, called the weight, which is equivalent to the memory of the artificial neural network. The output of the network varies depending on the connection, the weight and the activation function of the network.

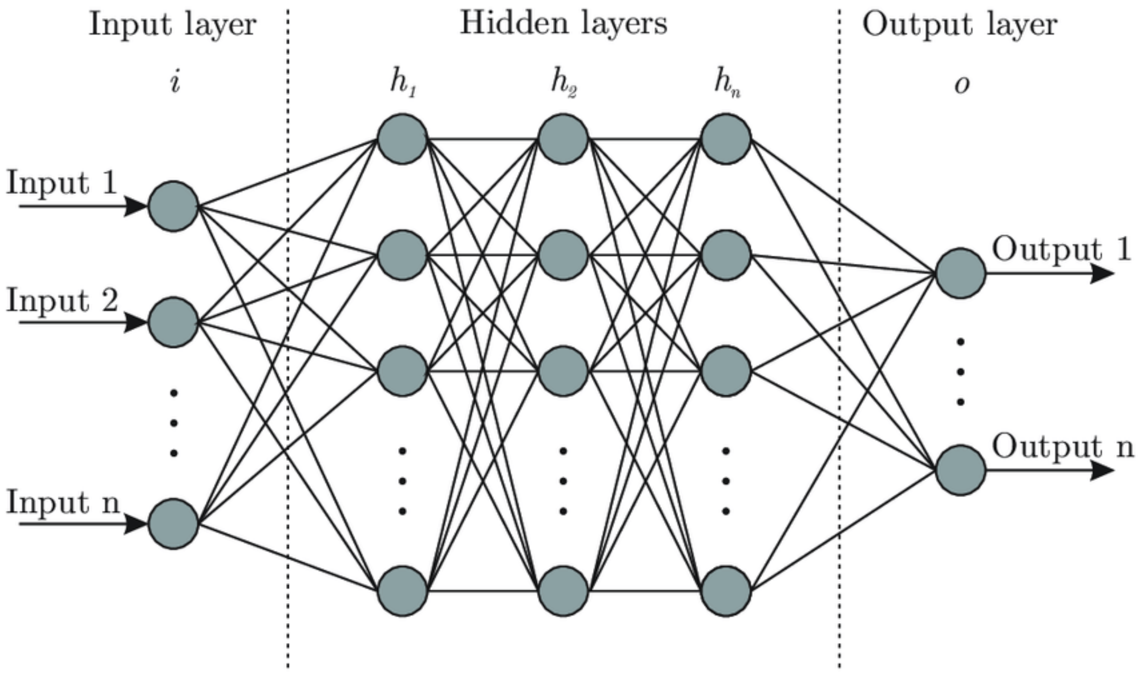


Figure 4 Artificial Neural Network (Facundo, Juan , & Víctor , 2018)

The ANN model is an extreme simplification of the human nervous system and consists of computational units similar to the neurons of the biological nervous system, known as artificial neurons. Primarily, the ANN model consists of three layers, namely input, hidden and output. Each neuron in the nth layer is interconnected with a neuron in the (n+1st) layer by some signal. Each connection is assigned a weight. After multiplying each input with the corresponding weight, the output can be calculated. The output is passed through the activation function to obtain the final ANN output (S.K. & Snehashish, 2021).

3.3.5 Logistic regression

Logistic regression is a generalized linear regression. The dependent variable of logistic regression can be either dichotomous or multi-categorical, but dichotomous is more commonly used. Therefore logistic regression is often used for partial linear classification of binary classification.

This logistic regression algorithm can be obtained by using the relationship between the large data set and the binary target variable. In logistic regression, the Odds ratio is used to determine the incidence rate or trigger probability. We generally think of it as the proportion of events that occur.

$$odds\,ratio = \frac{P(1)}{1-P(1)} = \frac{P(1)}{P(0)}$$

We can determine the parameters of the logistic regression by the maximum likelihood function

The degree of matching between the statistical model and the data sample is shown to be expressed by the likelihood function, and the maximum likelihood estimation is applied to find the parameters with the best match by the gradient descent method. When the parameters of the regression are determined, the odds ratio of logistic regression can be presented as follows: $Log\,odds\,ratio = \beta 0 + \beta 1 \cdot x1 + \cdots + \beta n \cdot xn$

The probability of an event occurring can be represented as a logistic curve to show. The shape of the function visible on Figure. Unlike linear regression, the shape of the logistic regression function is similar to the inverse "sin function". The logistic curve takes values in the interval [0;1], and generally we interpret the value of 0.5 as the decision boundary between the two classes. The probability of an event is represented by the following function:

$$\mathrm{P}(\mathrm{target} = 1) = \frac{1}{1 + e^{-(\beta 0 + \beta 1 \cdot x1 + \cdots + \beta n \cdot xn)}}$$

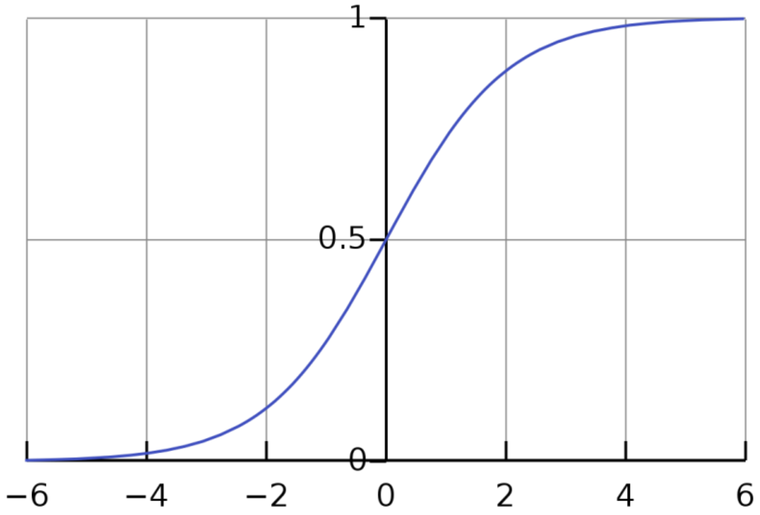


Figure 5 Logistic curve (Qef, 2008)

Since logistic regression is a number-based algorithm, all categorical variables need to be converted into a numerical form during data preparation.

3.3.6 Evaluation of model

When it comes to the evaluation of the logistic regression model, we can always use RMSE, precision and recall, ROC/AUC, cross validation etc. In the following paragraph, we will introduce some commonly used evaluation to check the performance of logistic regression model.

Confusion matrix and some relevant terms:

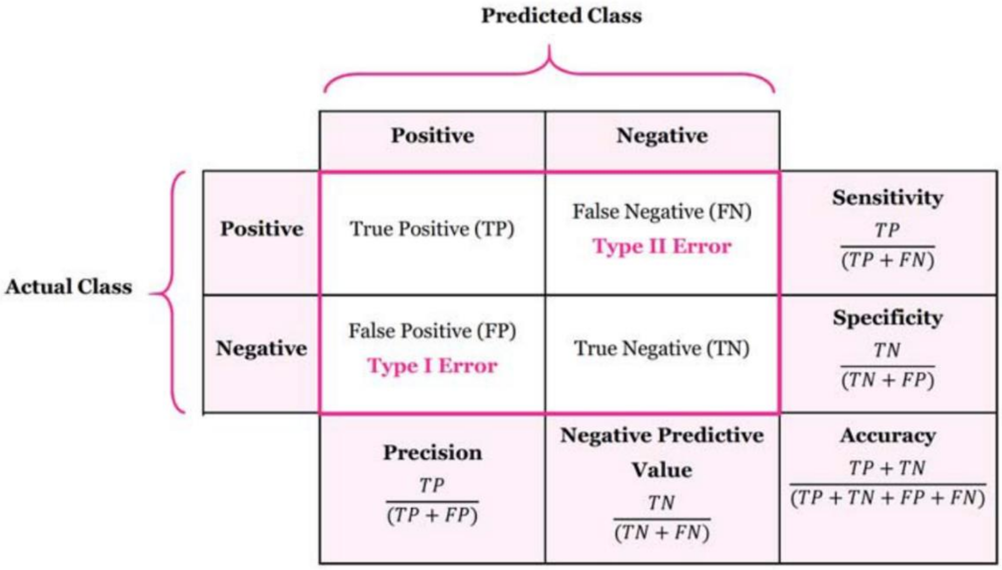
We will introduce the confusion matrix first in this part, which is quite useful for the further evaluation of the logistic regression model. The confusion matrix is a simple table that can be used to evaluate the performance of a classification model. It summarizes the count combinations of every predicted and actual class.

In the table, we have “positive/negative” which means the class is predicted as “positive/negative”, while “true/false” means whether our model detect or classify the data sample as “correct/incorrect” groups.

We need to stress that predicted class and actual class are two different concepts. For example, False Positive indicates that actual classification should be Negative, even though it is classified as positive by model. likewise, False Negative means actual classification should be Positive.

As shown in the following figure, we can see clearly there are four combinations of the results:

- True positive
- False negative
- True negative
- False positive



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Figure 6 Combination of results (An, 2020)

Based on the values of the above four situation, we can now calculate precision, sensitivity, specificity, accuracy and so on. For instance, the accuracy is the proportion of the total number of all the correct predictions.

$$accuracy = \frac{TP + TN}{(TP + TN + FP + FN)}$$

And we define the precision as the ratio of the total number of correctly classified positive classification and the total number of predicted positive classification. That’s why it is also known as positive predictive value.

$$\text{precision} = \frac{TP}{(TP + FP)}$$

Sensitivity is also called as “recall” or we just say “true positive rate”, is the proportion of the total number of actual positives that were identified correctly. Actual positives include situations of true positives and false negatives.

$$\text{sensitivity} = \frac{TP}{(TP + FN)}$$

As we probably all know that this is vital for the COVID-19 test. To be specific, if the COVID-19 test is 95% of sensitivity, this indicates that 95 out of 100 infected patients will be correctly diagnosed, and only 5 infected patients are off the radar based on test results.

Specificity should also be mentioned as the proportion of the total number of the actual negative that were identified correctly. The actual negative include situations of true negatives and false positives. It is also known as the true negative rate (TNR).

$$\text{Specificity} = \frac{TN}{(TN + FP)}$$

Nevertheless, we should be aware that there is a trade-off between recall and precision. It shows how the recall vs precision relationship changes as we vary the threshold for identifying a positive in our model. When we increase the recall rate by adjusting the classification threshold of a model, the precision rate is decreased.

After we introduce the confusion matrix and all relevant terms, we can get to know ROC/AUC. Generally, ROC means a curve plots the true positive rate on the y-axis versus the false positive rate on the x-axis. It shows the true positive and false positive rate for every probability threshold of a binary classifier.

In other words, the higher the result of ROC, the better is our model. And AUC is just "area under the curve", so it represents the similar meanings, with high value suggesting better performance of the model.

4 Practical Part

Introduction

This chapter introduces the research analysis as specified set out in the research methodology. The analysis of the data set is consistent with the objectives. The results of the analysis are presented in graphical form.

4.1 Overview of the Case Study

The dataset was sourced from Kaggle with health information of more than 5100 volunteers. Kaggle is the world's largest data science community with powerful tools and resources. After retrieving the Stroke Prediction Dataset from this site, data exploration and cleaning activities were performed. The data cleaning was started using EXCEL and SAS studio university version. From the given data, only 11 variables were selected and any duplicates were removed and their values were also defined before doing any analysis part. The purpose of analyzing this dataset is for academic purposes only and this data will not be used for any other purpose. It includes 5000 Health records of 5,000 patients with information about health indicators and lifestyle habits that may affect stroke. Before analyzing the data and its variables, it is important to understand the benefits of this study. Nowadays, stroke is the second leading cause of death worldwide and is a huge threat to health. Therefore, understanding and analyzing the factors that contribute to stroke and generating a predictive model to help medical practitioners and academics predict stroke from a set of data is of utmost importance to both groups.

The dataset was sourced from Kaggle, the world's largest data science community with powerful tools and resources, with health information from over 5,100 volunteers. After retrieving the stroke prediction dataset from this site, a data exploration and cleaning activity was undertaken. Data cleaning was initiated using EXCEL and SAS studio university edition. From the given data, only 11 variables were selected and any duplicates were removed and their values were also defined before any part of the analysis was performed. The purpose of analysing this dataset was for academic purposes only and this data will not be used for any other purpose.

It includes the health records of 5100 patients with information on health indicators and lifestyle habits that may affect stroke. Before analysing the data and its variables, it is important to understand the benefits of this study. Today, stroke is the second leading cause of death worldwide and poses a huge threat to health. Therefore, understanding and analysing the factors that contribute to stroke and generating a predictive model to help healthcare workers and academics predict stroke from a set of data is of utmost importance to both groups.

The variables in this dataset will be analysed in detail in the next sections of this specific study. The tool used to analyse the dataset was SAS Studio 9.4. SAS University Edition is an open source software where the data is run online and the results can be generated in SAS Studio, a computing software for analysing big data.

4.1.1 Data pre-processing

Attribute Information	Data type	Description
Id	Numerical	Unique identifier
gender	Categorical	"Male", "Female" or "Other"
age*	Categorical	age of the patient
hypertension*	Categorical	patient has hypertension or not
heart_disease*	Categorical	patient has a heart disease or not
ever_married	Categorical	"Yes" or "No"
work_type	Categorical	"Self-employed", "Private", "Never_worked","Govt_job" or "children"
residence_type	Categorical	"Urban" or "Rural"
avg_glucose_level	Categorical	average glucose level in blood
bmi*	Categorical	body mass index
smoking_status	Categorical	"formerly smoked", "ever smoked", "smokes" or "Unknown"
Stroke*	Categorical	patient had a stroke or not
Note: * see appendix for variable code value.		

Table 1 Category data, data type, Description
source: author's own work

Before we start EDA, the first and most important thing is to check for duplicates. If the result of checking duplicates is yes, we need to remove them by creating some indexes.

As seen in the figure above, there are no duplicates in the training set or even in the whole dataset because each patient's data is unique and cannot be duplicated.

Secondly, defining the types of variables present is a key part of our beginning so that we can understand them and their effect on the dependent variable and thus the statistical methods we will use. The categorical variables for this dataset are : Id、gender、age、hypertension 、 heart_disease 、 ever_married 、 work_type 、 residence_type 、 avg_glucose_level、bmi 、smoking_status and stroke.

4.2 Explanatory Data Analysis

The first phase of the study should provide an understanding of the dataset. The purpose of this phase is to provide as much information about the dataset as possible.

Explanatory data analysis is an activity that includes classification and numerical activity.

Therefore we need to have a comprehensive analysis and understanding of the variables and the effect of each variable on the dependent variable through EDA.

4.2.1 Univariate statistical analysis

First, we will use univariate statistical analysis to evaluate and analyze each variable separately.

The first variable to be described is the dependent variable 'y', which is the target variable stroke for the current data set. It is the one that contains information about whether the patient had a stroke or not.

Dependent variable: Stroke

```
proc sgplot data=WORK.STROKE_PRED;

     vbar stroke / fillattrs=(color=CXCAD5E5) datalabel
fillType=gradient
            stat=percent;
run;
```

The following table shows the number and percentage of people who had a stroke versus those who did not. It shows that there were 4,861 people who did not have a stroke and 249 people who had a stroke.

The visual distribution of the Y variable can be seen in Figure 5, where we can observe the percentage of people who have had a stroke versus those who have not had a stroke. We can see that this data set is very unbalanced. It shows that 95.1% of people have not had a stroke and the probability of having a stroke is 4.9%. From the graph by comparison we can get that in general the incidence of stroke is less than 5%.

Did the patient has a stroke?	Frequency	Percentage of total
Yes (1)	249	4.9%
No (0)	4861	95.1%

Table 2 Frequency statistics for variable strokes
Source: author's own work

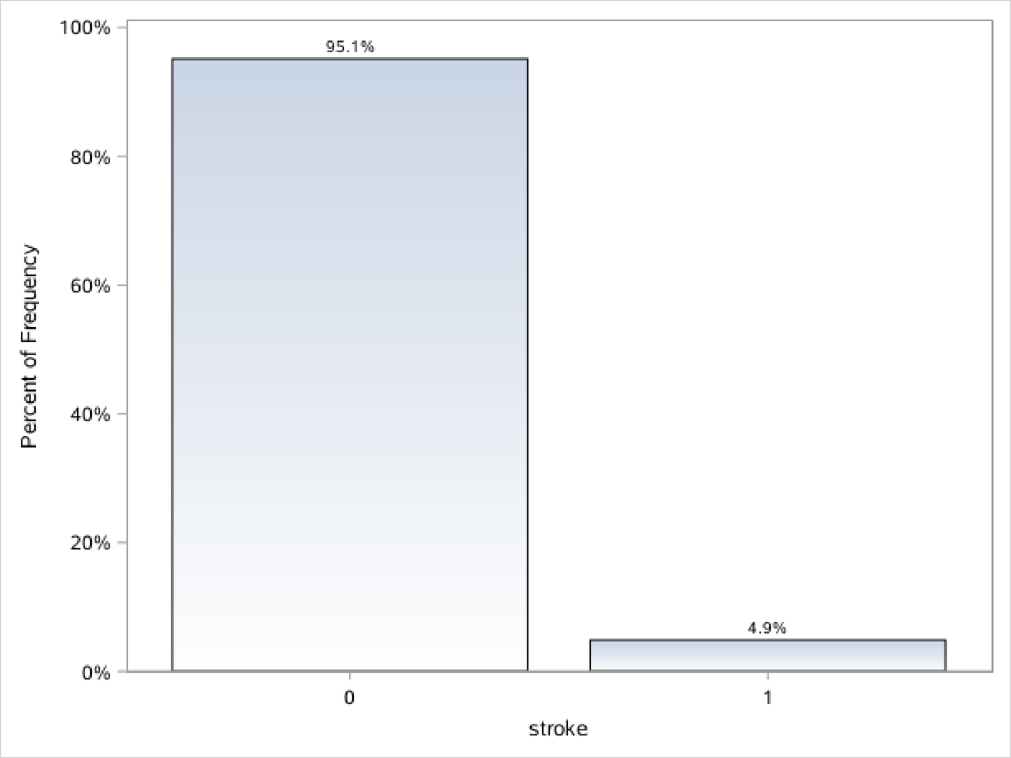
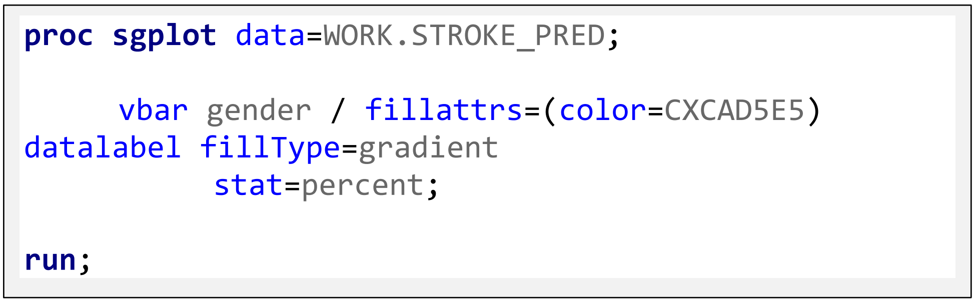


Figure 7 Distribution of dependent variable
source: author’s own work



```
proc sgplot data=WORK.STROKE_PRED;

     vbar gender / fillattrs=(color=CXCAD5E5) 
datalabel fillType=gradient
          stat=percent;

run;
```

Independent variable: Gender

As we can see from Figure 8, the data is slightly unevenly distributed, with 2,994 women accounting for 58.59% and 2,115 men accounting for 41.39% of the total. It is clear that the proportion of women is approximately ten percent higher than that of men.

Gender	Frequency	Percentage
Female	2994	58.59%
Male	2115	41.39%
Other	1	0.02%

Table 3 Frequency statistics for the variable gender
Source: author's own work

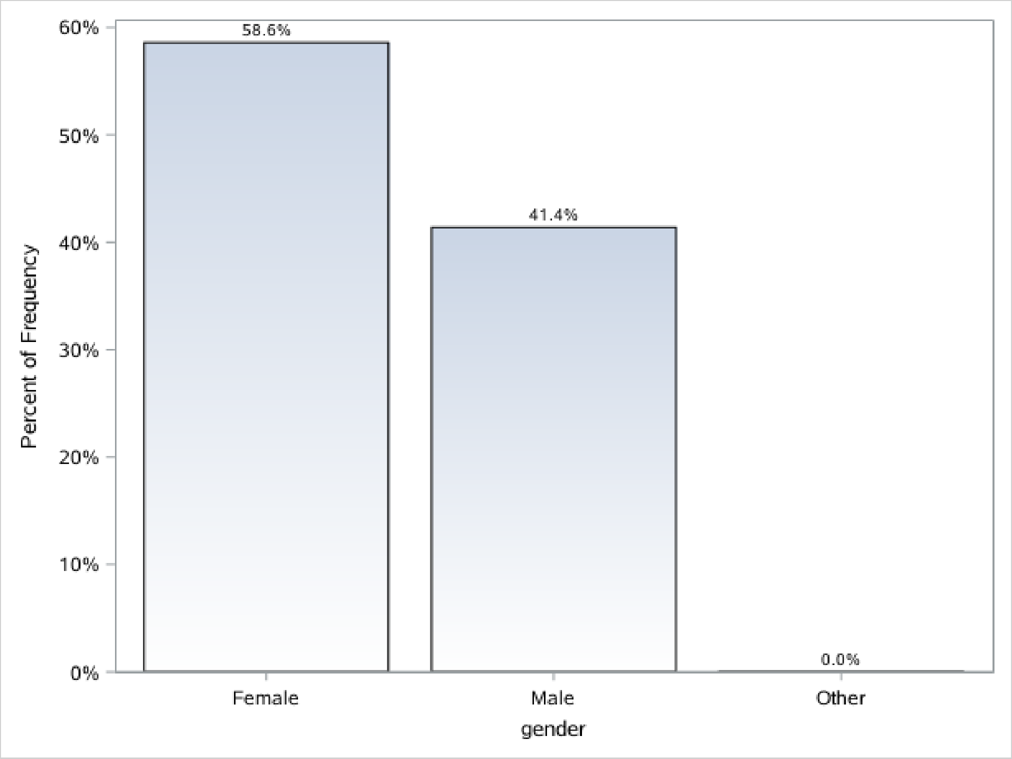
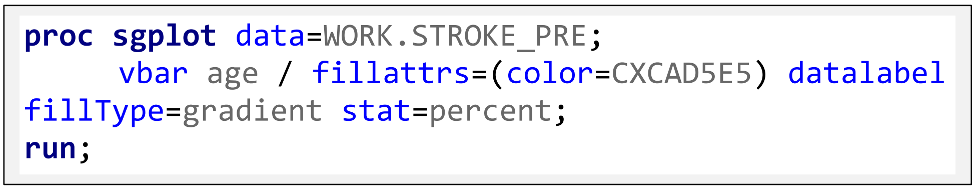


Figure 8 distribution of variable gender
source: author's own work

Independent variable: age



```
proc sgplot data=WORK.STROKE_PRE;
     vbar age / fillattrs=(color=CXCAD5E5) datalabel
fillType=gradient stat=percent;
run;
```

Here the ages are recoded as following table.

variables	Age (years)
1	0-10
2	11-20
3	21-30
4	31-40
5	41-50
6	51-60
7	61-70
8	71-80
9	81-90

Table 4 Recoded age
Source: author own work

This variable is represented by numerical values rather than actual values, each representing a different age group. As can be seen from the figure 7, out of the 9 age groups that the mean is 4.78, with 13.2% in group 4, 14.5% in group 5 and 16.1% in group 6. This indicates that of the nine age groups, the ages are concentrated in those in groups 4-6.

We can obtain that the age is concentrated among those who are between 31 and 60 years old.

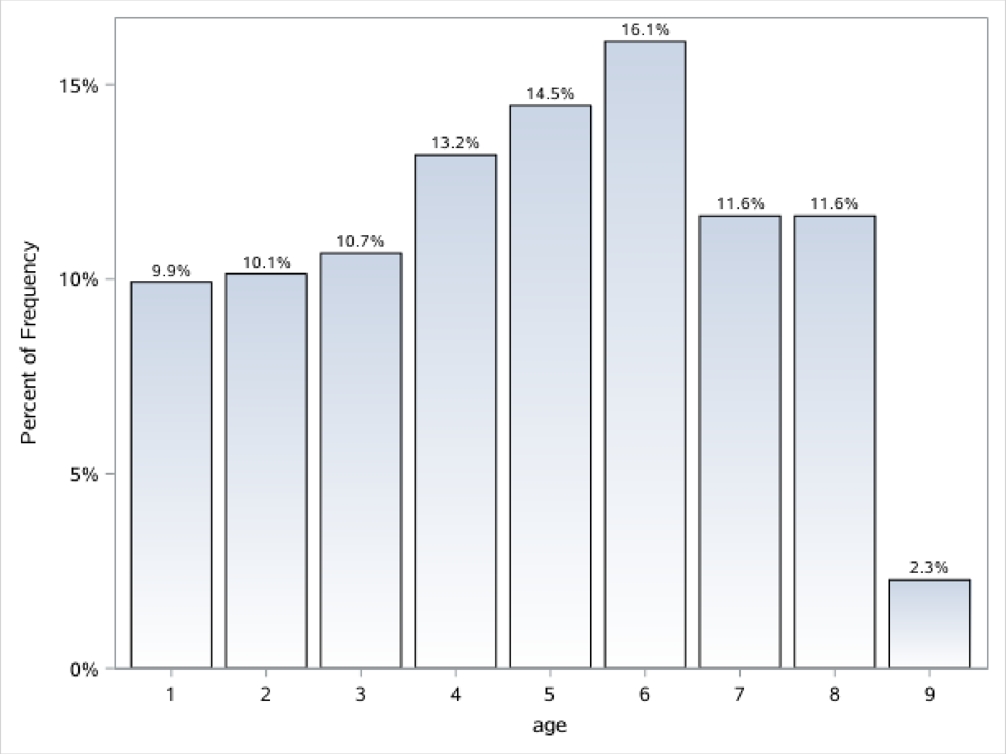
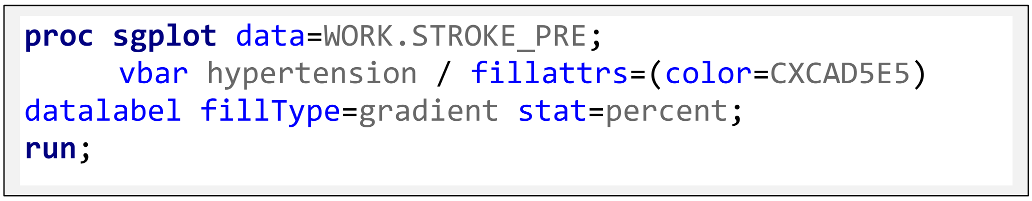


Figure 9 distribution of variable age
source: author's own work

Independent variable: hypertension



```
proc sgplot data=WORK.STROKE_PRE;
      vbar hypertension / fillattrs=(color=CXCAD5E5)
datalabel fillType=gradient stat=percent;
run;
```

The distribution of the variable hypertension data is to the right, and we can see a clear contrast in Figure 10, with 9.7% of people suffering from hypertension compared to 90.3% of people not suffering from hypertension.

Did the patient has a hypertension?	Frequency	Percentage of total
Yes (1)	498	9.7%
No (0)	4612	90.3%

Table 5 Frequency statistics for the variable hypertension
Source: author's own work

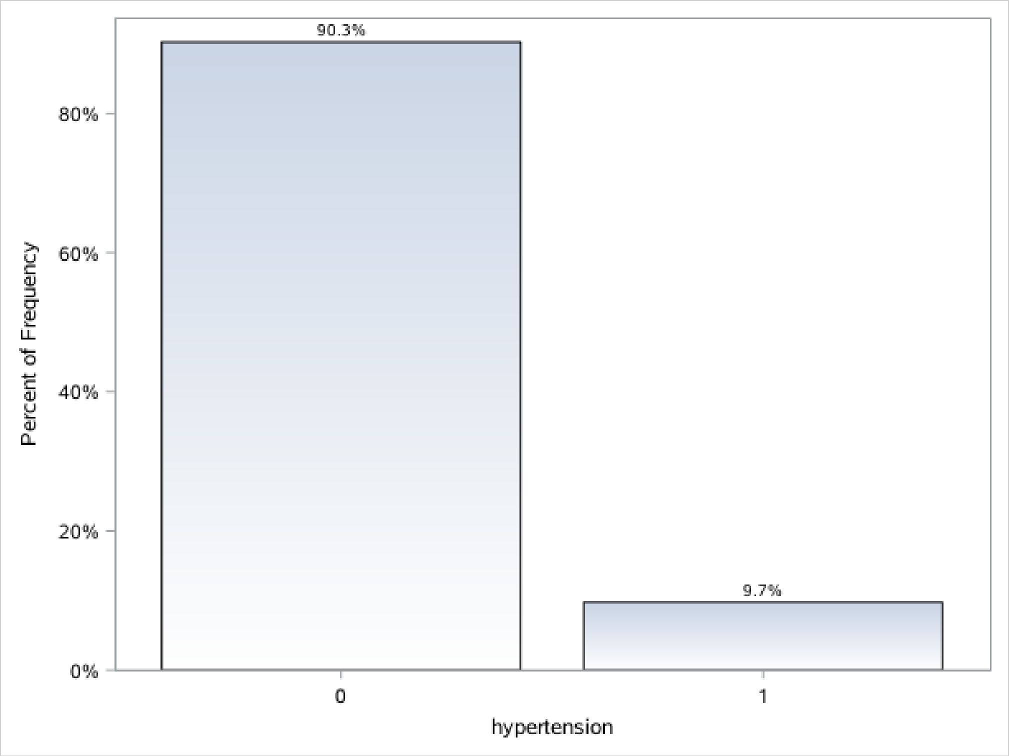
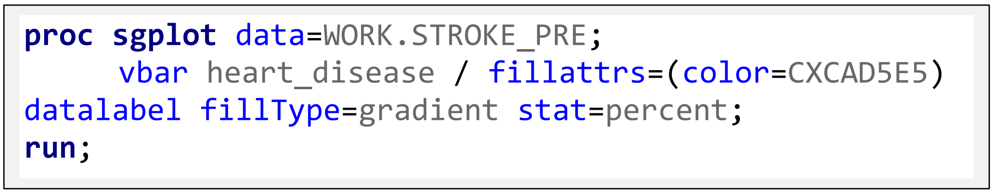


Figure 10 distribution of varibale hypertension
source: author's own work

Independent variable: heart_disease



```
proc sgplot data=WORK.STROKE_PRE;
     vbar heart_disease / fillattrs=(color=CXCAD5E5)
datalabel fillType=gradient stat=percent;
run;
```

As we can see from Figure 11, the variable heart disease is heavily distributed on the right side of the figure, with 276 people suffering from heart disease (5.4%), which is lower than those suffering from hypertension, and 4,834 people not suffering from heart disease (94.6%).

Did the patient has a heart disease?	Frequency	Percentage of total
Yes (1)	276	5.4%
No (0)	4834	94.6%

Table 6 Frequency statistics for the variable heart disease
Source: author's own work

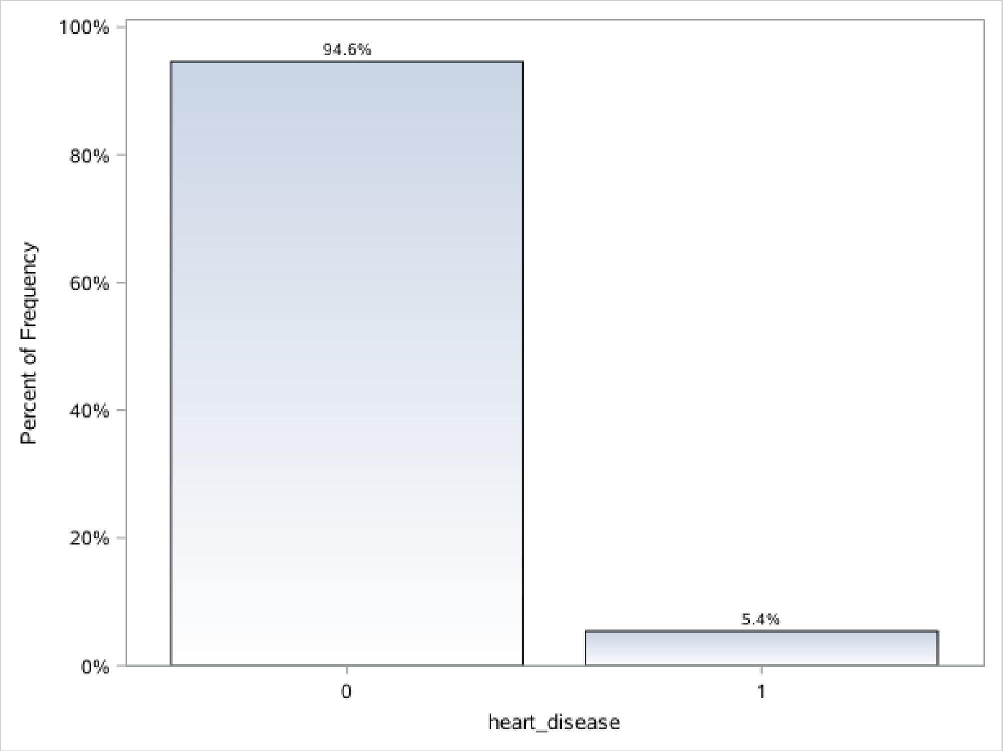
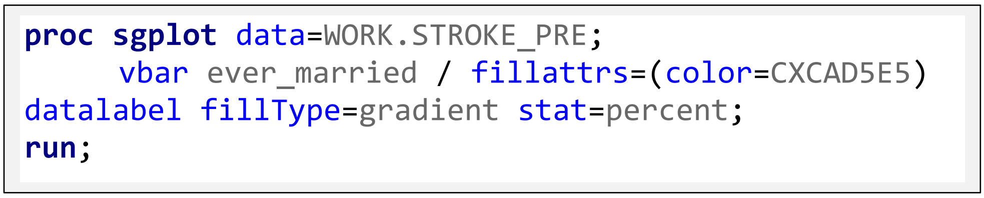


Figure 11 distribution of heart_disease
source: author's own work

Independent variable: ever_married



```
proc sgplot data=WORK.STROKE_PRE;
     vbar ever_married / fillattrs=(color=CXCAD5E5)
datalabel fillType=gradient stat=percent;
run;
```

By comparing the data in Figure 12, we can see that the distribution of data is slightly higher on the left side of the graph than on the right, with 3353 married people at 65.6%. The number of people who have never been married is at 1755, or 34.4%.

Has the patient ever married?	Frequency	Percentage of total
Yes (1)	3353	65.6%
No (0)	1757	34.4%

Table 7 Frequency statistics for the variable ever marry
Source: author's own work

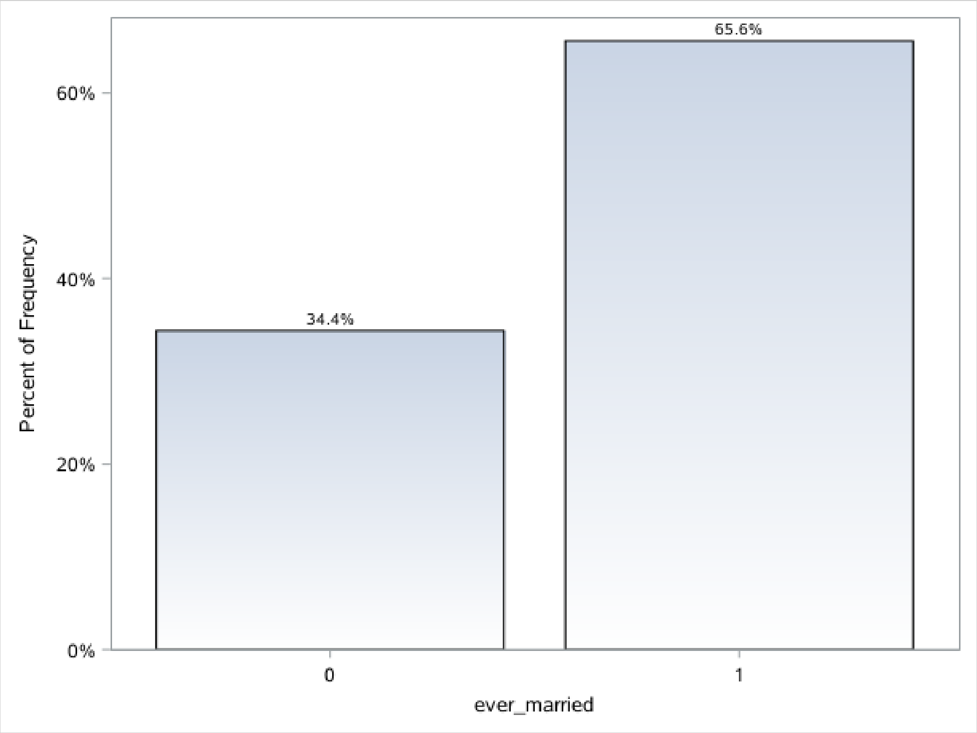


Figure 12 distribution of ever_married
source: author's own work

Independent variable: work_type

```
proc sgplot data=WORK.STROKE_PRE;
     vbar work_type / fillattrs=(color=CXCAD5E5)
datalabel fillType=gradient stat=percent;
run;
```

This variable gives the patient's work category, with four categorical values. From the chart below, As can be seen in Figure 13 below, the majority of people are self-employed or private business owners, with 3,744 people (73.3%). Only 0.4% 22 people had never worked before. Those who worked in government were 657 or 12.9% and children 687 or 13.4%, both relatively similar in number.

Category: Work_type	Frequency	Percentage of total
Self-employed or Private (1)	3744	73.3%
Never_worked (2)	22	0.4%
Govt_job (3)	657	12.9%
Children (4)	687	13.4%

Table 8 Frequency statistics for the variable work_type
Source: author's own work

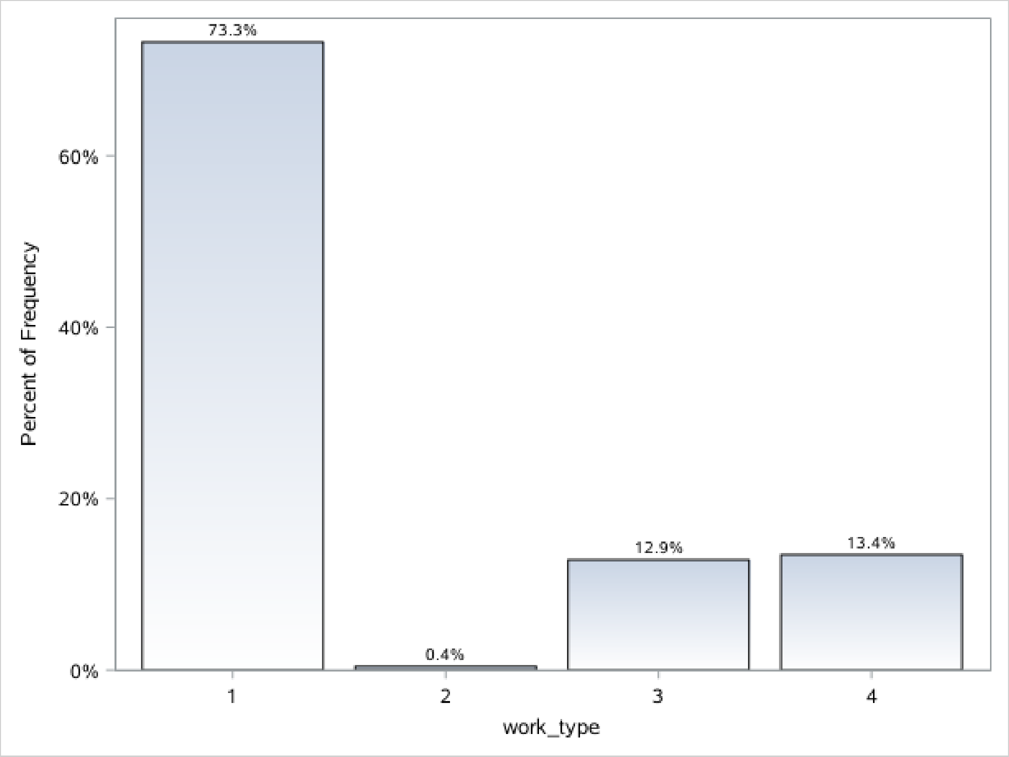
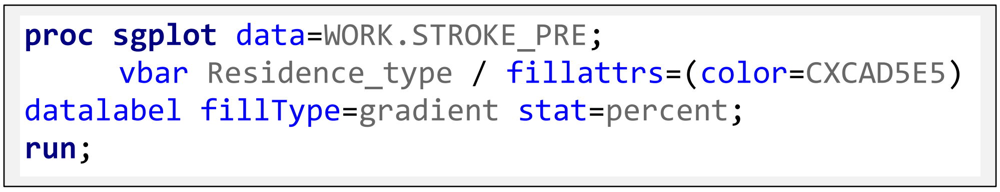


Figure 13 distribution of work_type
source: author's own work

Independent variable: Residence_type



```
proc sgplot data=WORK.STROKE_PRE;
     vbar Residence_type / fillattrs=(color=CXCAD5E5)
datalabel fillType=gradient stat=percent;
run;
```

This variable gives the type of residence category of the patient and has 2 categorical values. As can be seen from the graph below, 2514 of the population live in rural areas representing 49.2% of the total population and the other 2596 live in urban areas representing 50.8% of the total population. For the total number of people observed, the type of residence becomes evenly distributed.

Category: Residence_type	Frequency	Percentage of total
Rural (1)	2514	49.2%
Urban (0)	2596	50.8%

Table 9 Frequency statistics for the variable Residence_type
Source: author's own work

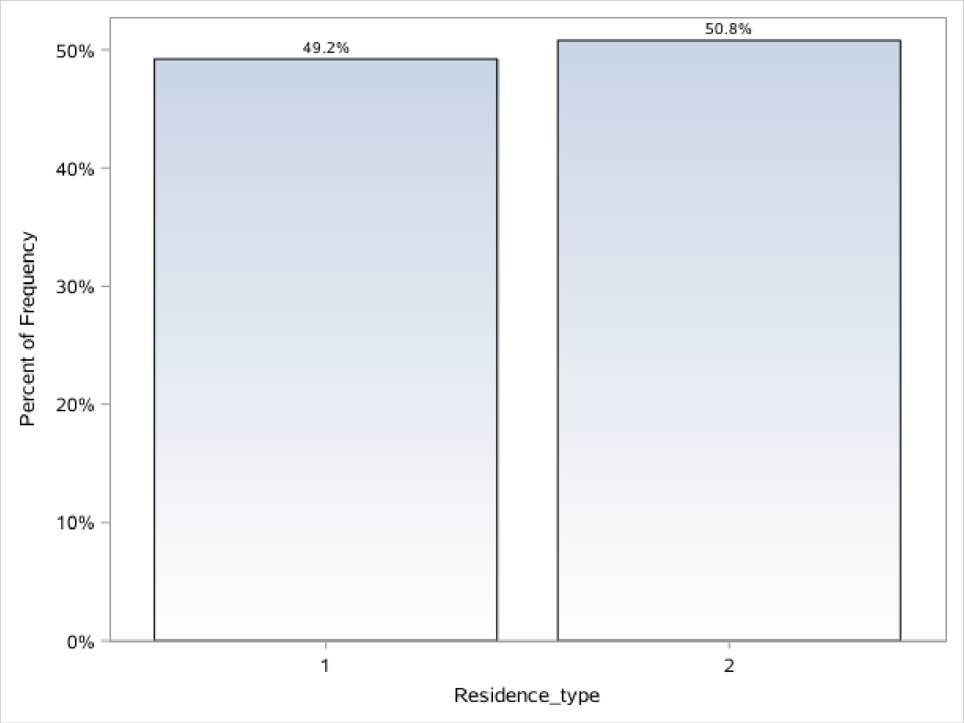
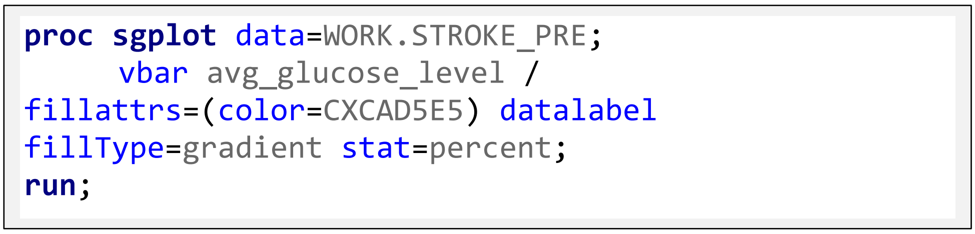


Figure 14 distribution of residence_type
source: author's own work

Independent variable: avg_glucose_level



```
proc sgplot data=WORK.STROKE_PRE;
     vbar avg_glucose_level /
fillattrs=(color=CXCAD5E5) datalabel
fillType=gradient stat=percent;
run;
```

The average blood glucose level, a variable that we can see from Figure 15 is mostly distributed to the left, with the two largest components being, 3189 people in the 50-100 range representing 62.4% of the total and 1196 people in the 101-150 range representing 23.4% of the total. The smallest number is 251-300 with only 22 people accounting for 0.4% of the total.

average glucose level in blood	Frequency	Percentage of total
50-100 (1)	3189	62.4%
101-150 (2)	1196	23.4%
151-200 (3)	306	6.0%
201-250 (4)	397	7.8%
251-300 (5)	22	0.4%

Table 10 Frequency statistics for the variable ave_glucose_level
Source: author's own work

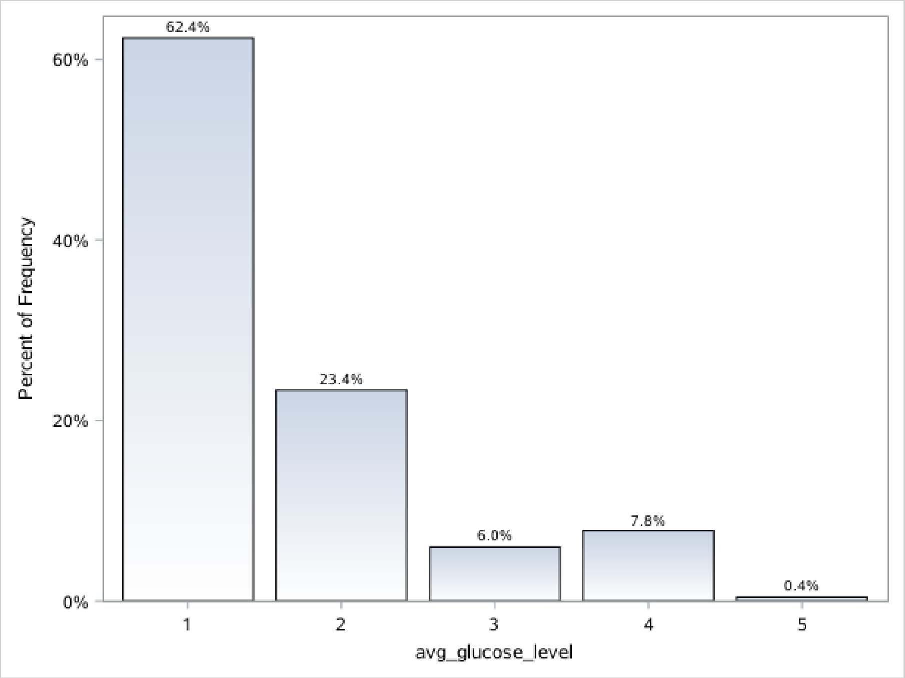
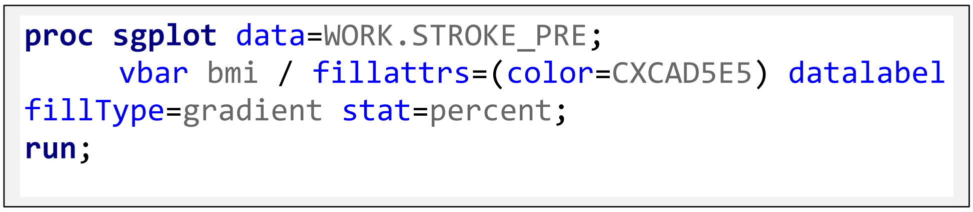


Figure 15 distribution of avg_glucose_level
source: author's own work

Independent variable: bmi



```
proc sgplot data=WORK.STROKE_PRE;
     vbar bmi / fillattrs=(color=CXCAD5E5) datalabel
fillType=gradient stat=percent;
run;
```

Body mass index is a variable that measures the proportionality of the body. This variable is mainly distributed on the left-hand side, as can be seen in Figure 16, where the vast majority of people are in the range of 21-40, with 3,808 people accounting for 74.5% of the total.

Body mass index	Frequency	Percentage of total
0-20 (1)	537	10.5%
21-40 (2)	3808	74.5%
41-60 (3)	336	6.6%
61-80 (4)	8	0.2%
81-100 (5)	2	0.05%
N/A (10)	419	8.2%

Table 11 Frequency statistics for the variable bmi
Source: author's own work

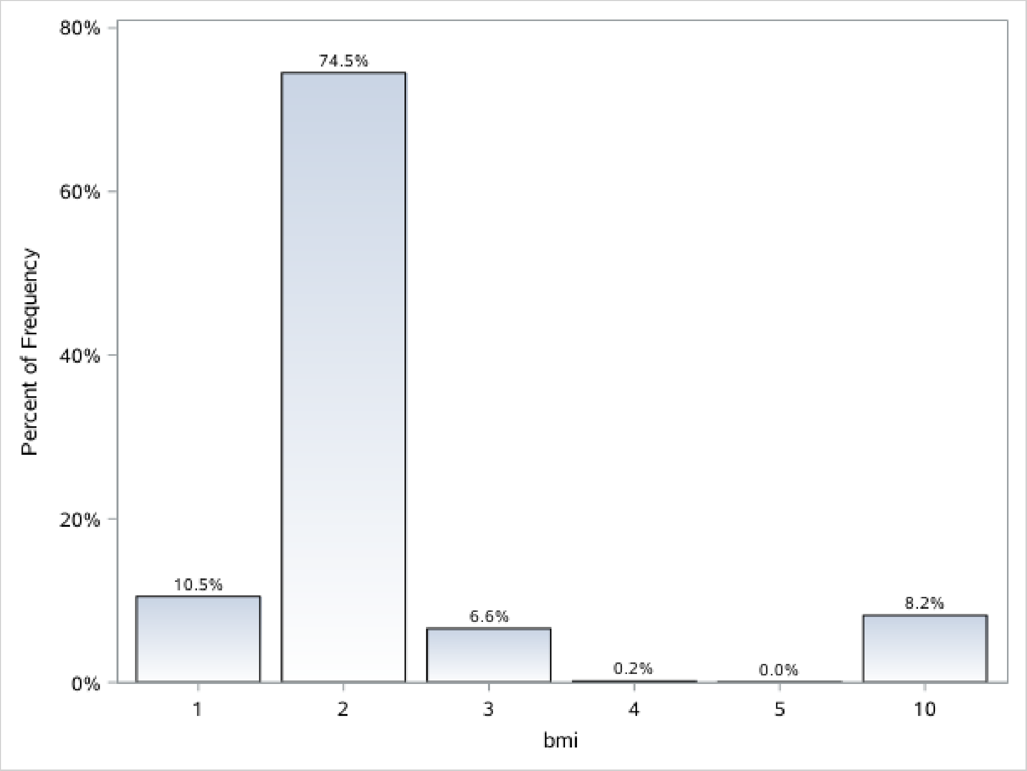


Figure 16 distribution of bmi
source: author's own work

Independent variable: smoking_status

```
proc sgplot data=WORK.STROKE_PRE;
      vbar smoking_status /
fillattrs=(color=CXCAD5E5) datalabel
fillType=gradient stat=percent;
run;
```

From Figure 17 we can see that the number of people who never smoked was 1892, or 37% of the total. The number of formerly smoked people is 885, or 17.3% of the total.

Smoking_status	Frequency	Percentage of total
formerly smoked (1)	885	17.3%
never smoked (2)	1892	37.0%
smokes (3)	789	15.4%
Unknow (10)	1544	30.2%

Table 12 Frequency statistics for the variable Smoking_status
Source: author's own work

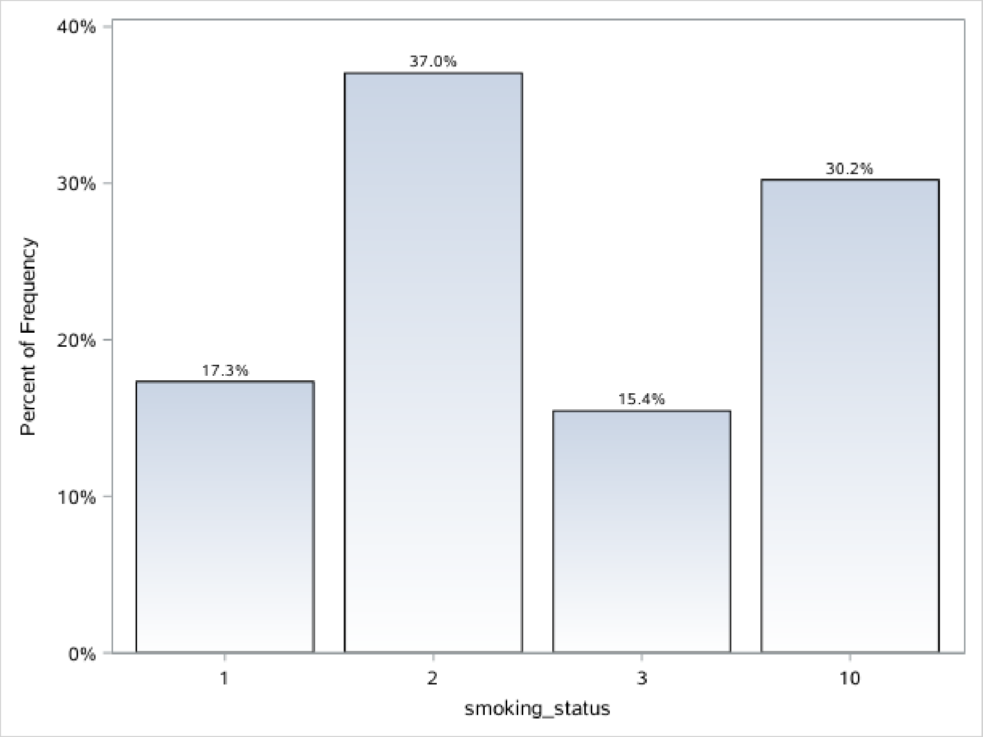


Figure 17 distribution of smoking_status
source: author's own work

4.3 Data preparation

After our initial analysis and understanding of the data through EDA, the next step will be data preparation. In the previous stage, we found that three variables present in the data (ID, gender and Residence_type) were not statistically significant for stroke, and we now prepared the data for modelling by removing these variables.

The variables to be used are listed in the following table:

Independent Variables	Classification code value
age	1="0-10" , 2="11-20" , 3="21-30", 4=”31-40”, 5="41-50" , 6="51-60" , 7="61-70" , 8="71-90" , 9="91-100"
hypertension	1=”patient has hypertension” , 2=” patient doesn’t have hypertension”
heart_disease	1=” Patient has heart disease” , 2=” Patient doesn’t have heart disease”
ever_married	1=”married” , 2=” not married ”
work_type	1=”Self-employed” , 1=”Private” , 2=”Never_worked” , 3=”Govt_job” , 4=”children”
avg_glucose_level	1=”50-100” , 2=”101-150” , 3=”151-200” , 4=“201-250” , 5=”251-300”

bmi	1=”0-20” , 2=”21-40” , 3=”41-60” , 4=”61-80” , 5=”81-100” , 6=”N/A”
smoking_status	1=”formerly smoked” , 2=”never smoked” , 3=”smokes” , 4=” Unknown”

Table 13 data description of the classification code
Source: author’s own work

Table 13 provides a description of the classification codes for these variables. The dataset used in this model was derived from the stroke prediction dataset we used earlier. The data were analysed by EDA using spearman correlation and the effects of correlation were excluded using VIF and tolerance testing, and finally we selected eight statistically significant variables to be used in modelling the data.

4.4 Building the model

4.4.1 Building a logistic regression model

We have prepared a data set that can be used to model the data for logistic regression in the previous step of preparing the data.

Now we can start building the regression model. The first major challenge we face is to split the dataset into two parts: the training set and the test set. The best way to divide the data set is to randomly select 25% of the data set as the test set and the remaining 75% as the training set. The training part of the dataset will be used to calculate the regression model parameters, while the test part will be used to verify the accuracy of the model parameters built. This is important to validate the model on data that has not been used for modelling.

The commands used to generate the training/test dataset in SAS Studio 9.4 are listed below.

```
proc surveyselect data=stroke_pred rate=0.75
out= stroke_pred_select outall
method=srs;
run;
data stroke_pred_train stroke_pred_test;
set stroke_pred_select;
if selected =1 then output stroke_pred_train;
else output stroke_pred_test;
run;
```

The total number of observations read was 5110 values, divided into two parts: the training set and the test set. The training set contains 3833 observations and the test set contains 1277 observations. The data had 9 variables, with stroke as the dependent variable and the other 8 variables as independent variables.

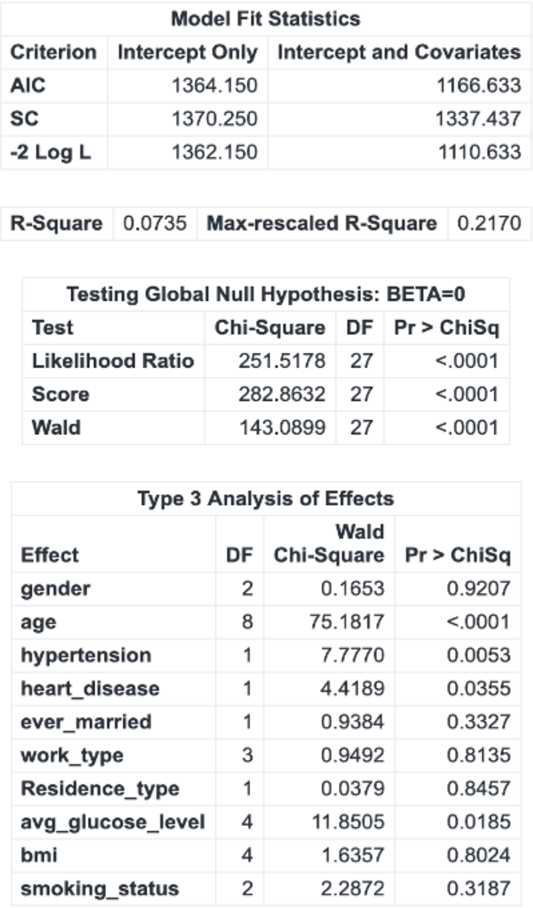
Variable	Train Dataset		Test Dataset	
	N	N miss	N	N miss
age	3833	0	1277	0

hypertension	3833	0	1277	0
heart_disease	3833	0	1277	0
ever_married	3833	0	1277	0
work_type	3833	0	1277	0
avg_glucose_level	3833	0	1277	0
bmi	3833	0	1277	0
smoking_status	3833	0	1277	0
stroke	3833	0	1277	0

Table 14 Train dataset and Test dataset.
source: author’s own work

A binary logistic regression model is developed as shown in Figure 18, The information is as follows:

Response Variable: stroke
Number of Response Levels 2
Model: binary logit
Optimization Technique: Fisher's scoring



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1364.150	1166.633
SC	1370.250	1337.437
-2 Log L	1362.150	1110.633

R-Square	0.0735	Max-rescaled R-Square	0.2170

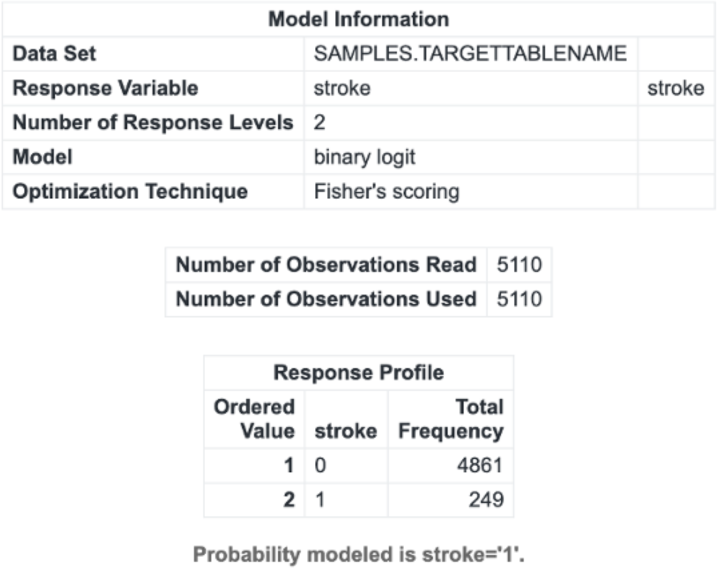
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	251.5178	27	<.0001
Score	282.8632	27	<.0001
Wald	143.0899	27	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
gender	2	0.1653	0.9207
age	8	75.1817	<.0001
hypertension	1	7.7770	0.0053
heart_disease	1	4.4189	0.0355
ever_married	1	0.9384	0.3327
work_type	3	0.9492	0.8135
Residence_type	1	0.0379	0.8457
avg_glucose_level	4	11.8505	0.0185
bmi	4	1.6357	0.8024
smoking_status	2	2.2872	0.3187

Figure 18 Binary logistic regression model-1
Source: author's own work

We can see that some of the independent variables in figure 18 have p-values above 0.05, which for the model as a whole means that the parameter is statistically insignificant. Parameters that are not significant should also be excluded from the model.

Therefore, we removed the variables gender, ever_married, work_type, Residence_type, bmi and smoking_status, Then we rebuilt the logistic regression model. Figures 19 and 20 show the basic and class level information of the reconstructed model.



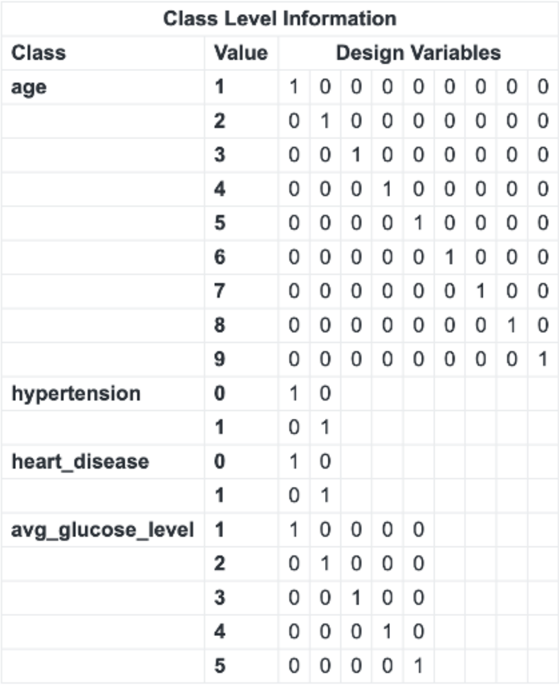
Model Information		
Data Set	SAMPLES.TARGETTABLENAME	
Response Variable	stroke	stroke
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	5110
Number of Observations Used	5110

Response Profile		
Ordered Value	**stroke**	**Total Frequency**
1	0	4861
2	1	249

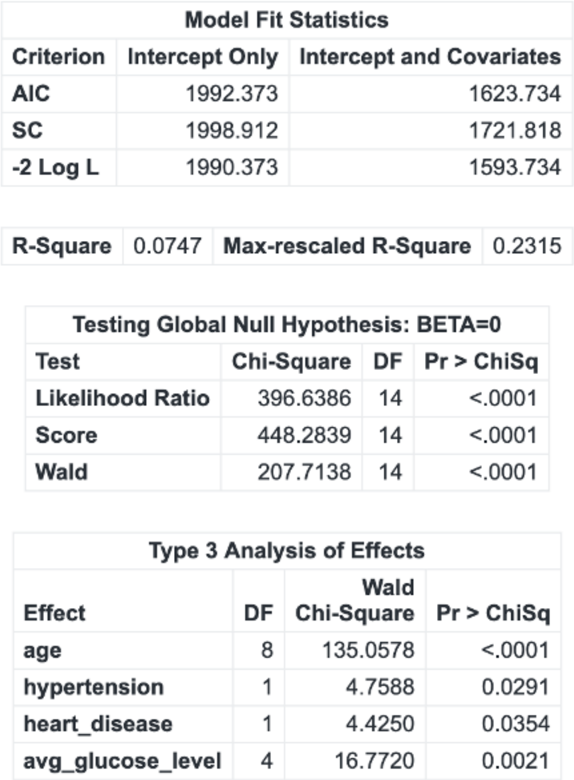
Probability modeled is stroke='1'.

Figure 19 Information about rebuilt binary logistic regression
Source: author's own work



Class	Value	Design Variables								
age	**1**	1	0	0	0	0	0	0	0	0
	2	0	1	0	0	0	0	0	0	0
	3	0	0	1	0	0	0	0	0	0
	4	0	0	0	1	0	0	0	0	0
	5	0	0	0	0	1	0	0	0	0
	6	0	0	0	0	0	1	0	0	0
	7	0	0	0	0	0	0	1	0	0
	8	0	0	0	0	0	0	0	1	0
	9	0	0	0	0	0	0	0	0	1
hypertension	**0**	1	0							
	1	0	1							
heart_disease	**0**	1	0							
	1	0	1							
avg_glucose_level	**1**	1	0	0	0	0				
	2	0	1	0	0	0				
	3	0	0	1	0	0				
	4	0	0	0	1	0				
	5	0	0	0	0	1				

Figure 20 class level information of BLR
Source: author's own work



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1992.373	1623.734
SC	1998.912	1721.818
-2 Log L	1990.373	1593.734

R-Square	0.0747	Max-rescaled R-Square	0.2315

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	396.6386	14	<.0001
Score	448.2839	14	<.0001
Wald	207.7138	14	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
age	8	135.0578	<.0001
hypertension	1	4.7588	0.0291
heart_disease	1	4.4250	0.0354
avg_glucose_level	4	16.7720	0.0021

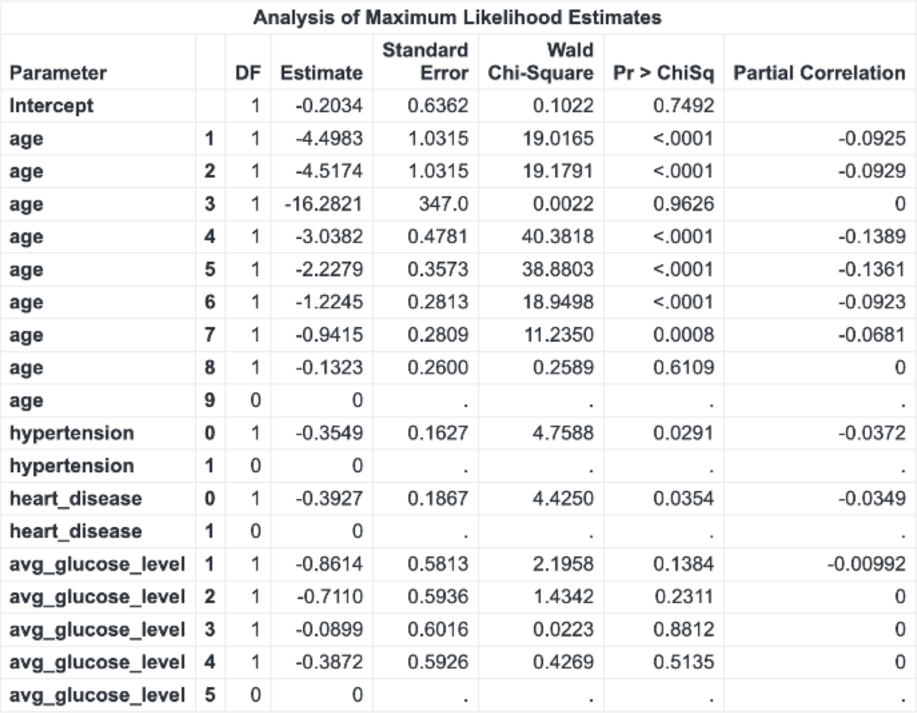
Figure 21 Model Fit Statistics and Analysis of Effects
Source: author's own work

From Figure 21 we can see that:

The middle section uses three tests (Likelihood Ratio Test, Score Test, Wald Test) to verify that the coefficients on all variables are zero at the same time. This shows that $\beta 1 = \beta 2 = \beta 3 = 0$ is valid or not. The results of all three tests are similar, with P-values less than 0.0001, indicating that the fitted model is significantly better than the model with only the constant term.

The p-values for age is <.0001, p-values for hypertension is 0.0291, p-values for heart-disease is 0.0354 and p-values for avg_glucose_level is 0.0021. Their p-values are all <0.05, We can say that all four independent variables are statistically significant at a 95% confidence level. Regression model tests showed that the independent variables age、hypertension、heart_disease and avg_glucose_level had a statistically significant effect on stroke.

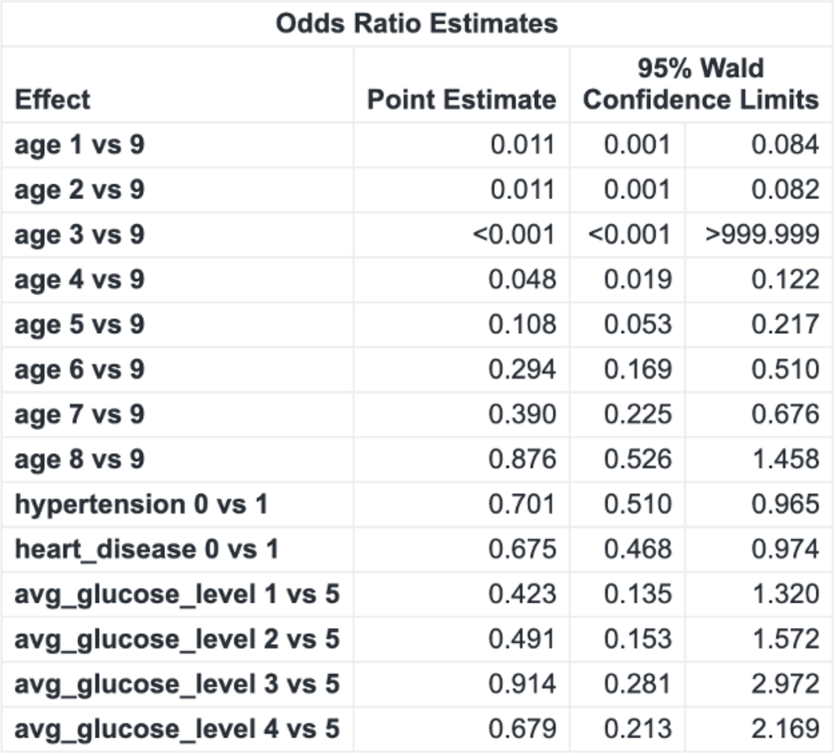
As shown in Figure 22 below, Through the estimation and subsequent validation of the significance of the parameters and the reconstruction of the model, we were able to find not only the variables that had a significant impact on the prediction of stroke, but also the values of the individual parameters.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Partial Correlation
Intercept		1	-0.2034	0.6362	0.1022	0.7492	
age	1	1	-4.4983	1.0315	19.0165	<.0001	-0.0925
age	2	1	-4.5174	1.0315	19.1791	<.0001	-0.0929
age	3	1	-16.2821	347.0	0.0022	0.9626	0
age	4	1	-3.0382	0.4781	40.3818	<.0001	-0.1389
age	5	1	-2.2279	0.3573	38.8803	<.0001	-0.1361
age	6	1	-1.2245	0.2813	18.9498	<.0001	-0.0923
age	7	1	-0.9415	0.2809	11.2350	0.0008	-0.0681
age	8	1	-0.1323	0.2600	0.2589	0.6109	0
age	9	0	0	.	.	.	.
hypertension	0	1	-0.3549	0.1627	4.7588	0.0291	-0.0372
hypertension	1	0	0	.	.	.	.
heart_disease	0	1	-0.3927	0.1867	4.4250	0.0354	-0.0349
heart_disease	1	0	0	.	.	.	.
avg_glucose_level	1	1	-0.8614	0.5813	2.1958	0.1384	-0.00992
avg_glucose_level	2	1	-0.7110	0.5936	1.4342	0.2311	0
avg_glucose_level	3	1	-0.0899	0.6016	0.0223	0.8812	0
avg_glucose_level	4	1	-0.3872	0.5926	0.4269	0.5135	0
avg_glucose_level	5	0	0	.	.	.	.

Figure 22 Analysis of Maximum Likelihood Estimates
Source: author’s own work



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age 1 vs 9	0.011	0.001	0.084
age 2 vs 9	0.011	0.001	0.082
age 3 vs 9	<0.001	<0.001	>999.999
age 4 vs 9	0.048	0.019	0.122
age 5 vs 9	0.108	0.053	0.217
age 6 vs 9	0.294	0.169	0.510
age 7 vs 9	0.390	0.225	0.676
age 8 vs 9	0.876	0.526	1.458
hypertension 0 vs 1	0.701	0.510	0.965
heart_disease 0 vs 1	0.675	0.468	0.974
avg_glucose_level 1 vs 5	0.423	0.135	1.320
avg_glucose_level 2 vs 5	0.491	0.153	1.572
avg_glucose_level 3 vs 5	0.914	0.281	2.972
avg_glucose_level 4 vs 5	0.679	0.213	2.169

Figure 23 Odd Ratio Estimates of BLR
Source: author’s own work

According to the two figures above 22 and 23, we can clearly see that:

When all other variables were held constant, the logarithm of the odds of having a stroke in age group 3 (21-30 years) relative to age group 9 (81-90 years) decreased by 16.2821. Accordingly, the odds ratio became $e^{-16.2821} =< 0.001$ times. And the logarithm of the probability of having a stroke in age group 8 (71-80 years) relative to the age group 9 (81-90 years) decreased by 0.1323. Accordingly, the odds ratio became $e^{-0.1323} = 0.876$ times. It can be seen that the older the patient is, the higher the probability of having a stroke. From this we get that age has a positive effect on the probability of stroke and that an increase in age leads to an increase in the probability of stroke.

When all other variables were held constant, the logarithm of the probability of having a stroke in those without hypertension relative to those with hypertension decreased by 0.3549. accordingly, the odds ratio became $e^{-0.3549} = 0.701$ times. We thus obtain that hypertension has a positive effect on the probability of stroke and that having hypertension leads to an increase in the probability of stroke.

When all other variables are held constant, the logarithm of the probability of having a stroke in a person without heart disease relative to a person with heart disease decreases by 0.3927. Correspondingly, the probability ratio becomes $e^{-0.3927} = 0.675$ times. As above, we can see that having heart disease has a positive effect on the probability of stroke, and that having heart disease leads to an increased probability of stroke.

When all other variables were held constant, the logarithm of the odds of having a stroke in average glucose level group 1 (50-100) relative to average glucose level group 5 (251-300) decreased by 0.8614. Accordingly, the odds ratio became $e^{-0.8614} = 0.423$ times. And the logarithm of the probability of having a stroke in average glucose level group 3 (151-200) relative to the average glucose level group 5 (251-300) decreased by 0.0899. Accordingly, the odds ratio became $e^{-0.0899} = 0.914$. It can be seen that the higher the average glucose level of the patient, the higher the probability of having a stroke. From this we get that the average glucose level has a positive effect on the probability of stroke and that an increase in the average glucose level leads to an increase in the probability of stroke.

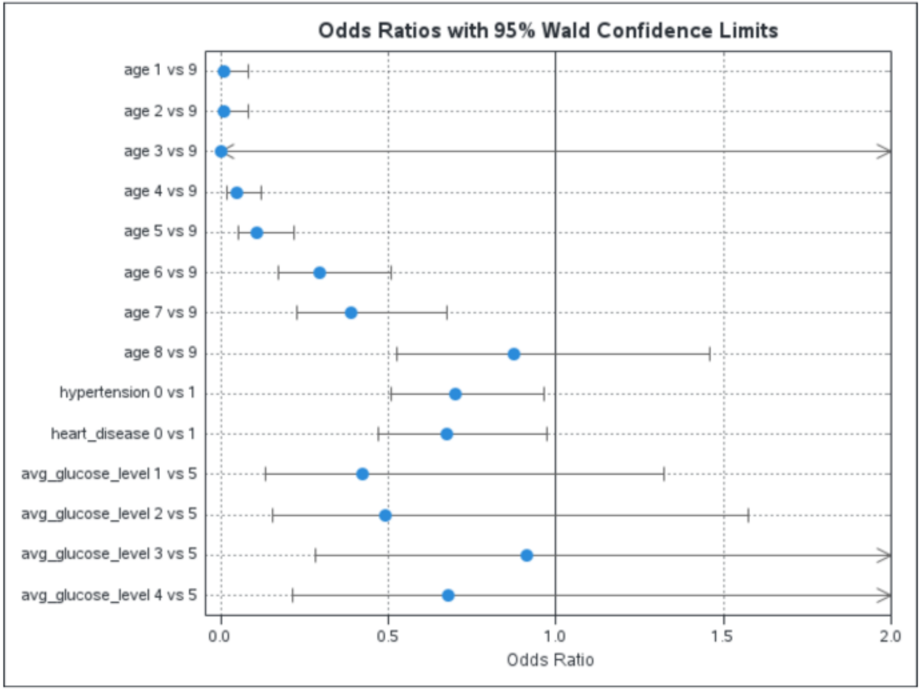
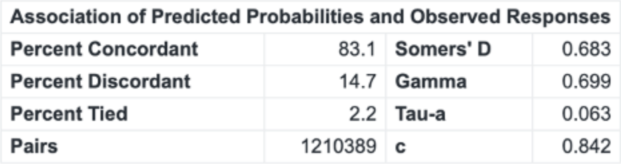


Figure 24 Odds Ratios with 95% Wald confidence Limits
Source: author’s own work



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	83.1	Somers' D	0.683
Percent Discordant	14.7	Gamma	0.699
Percent Tied	2.2	Tau-a	0.063
Pairs	1210389	c	0.842

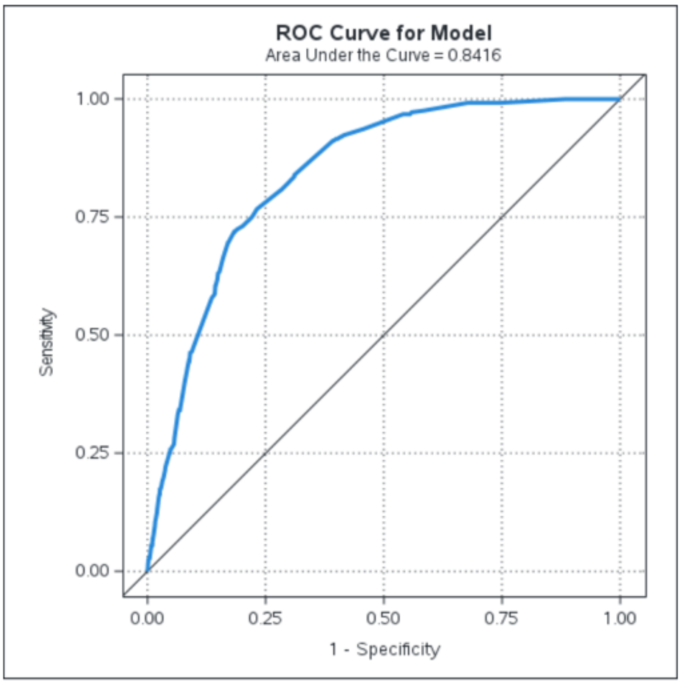


Figure 25 Predicted probability and ROC graphs
Source: author's own work

Figure 25 shows the correlation between predicted probability and observed response and the AUC image of the binary logistic regression, from the above figure Percent Concordan is 83.1%, while Percent Discordant is only 14.7%. c=0.842 indicates a model fit of 84.2%, which is the expected value we can accept.

We can clearly see the area under the ROC curve in the graph. auc = 0.8416

4.4.2 Build a decision tree model

There are two types of decision tree models that are used to construct them, namely CART and CHAID. CART stands for classification and regression trees where as CHAID represents Chi-Square automatic interaction detector.

Although CHAID can mine as much information as possible in the learning of the training sample set, its generated decision trees are relatively large in branching and size. The dichotomy of CART algorithm can simplify the size of decision trees and improve the efficiency of generating decision trees. Its operation is simpler. Of course, its performance is also very close to that of the entropy model.

- CHAID is a multinomial tree with slow operation speed, CART is a binary tree with fast operation speed.
- CART uses the Gini coefficient as an impurity measure for variables, reducing a large number of logarithmic operations.

CHAID is most frequently used for descriptive analysis whereas CART is frequently used in predictive analysis. In summary, the CART algorithm better fits the requirements of this study

Here we use a decision tree generated by the CART algorithm. Features are selected based on: the Gini index, the Gini coefficient represents the impurity of the model, the smaller the Gini coefficient, the lower the impurity, the better the features. the larger the Gini, the greater the uncertainty.

The CART algorithm uses a dichotomous recursive partitioning method. The generation of a decision tree is the process of recursively constructing a dichotomous decision tree. The algorithm always divides the current sample set into two sub-sample sets so that the resulting decision tree has only two branches per non-leaf node. The decision tree generated by the CART algorithm is therefore a binary tree with a simple structure. Therefore, the CART algorithm is suitable for problems where the sample features are yes or no.

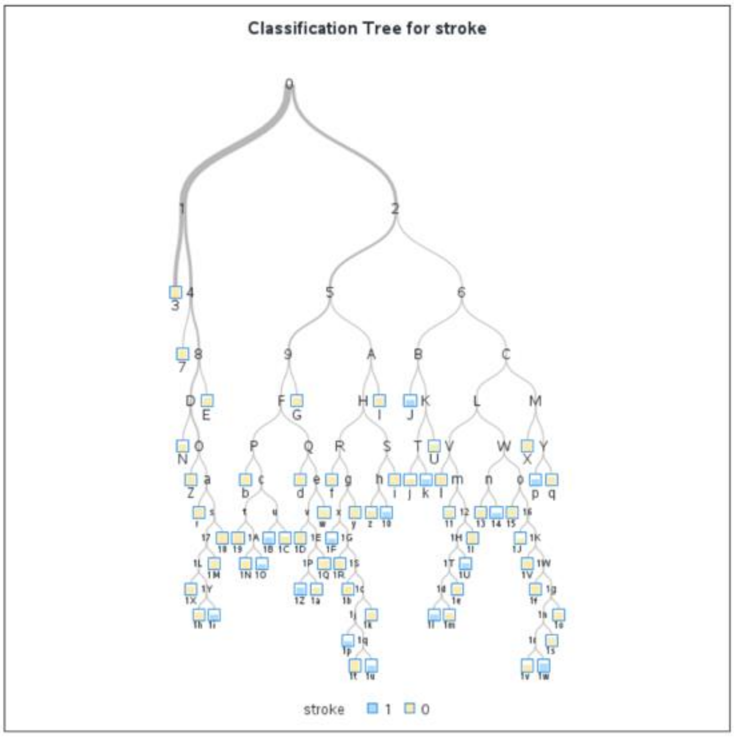


Figure 26 classification tree for stroke
Source: author's own work

Variable Importance				
Variable	Importance	Std Dev Importance	Relative Importance	Count
age	39.5845	0	1	15
avg_glucose_level	25.7277	0	0.6499	22
smoking_status	8.1609	0	0.2062	7
bmi	4.7117	0	0.119	3
Residence_type	3.1008	0	0.0783	4
work_type	3.0247	0	0.0764	3
gender	3.0189	0	0.0763	3
ever_married	2.3047	0	0.0582	2
heart_disease	0.6781	0	0.0171	1

Table 15 The importance of variables in a decision tree model
Source: author's own work

As we can see in Table 15 above, the most important variables for the decision tree model to predict stroke were stroke and mean blood glucose level, accounting for 39.5845% and 25.7277% of the total percentage, respectively. The least important variable was whether or not to change heart disease, with a percentage of 0.6781%. We can conclude that age and average blood glucose level are important influences on the risk of having a stroke.

Using the Relative Importance of the variables in Table 15, we can see that the factor that has the greatest impact on stroke is age, with the next second factor being avg_glucose_level. Therefore our starting segmentation threshold for Root Node 0 was chosen in age.

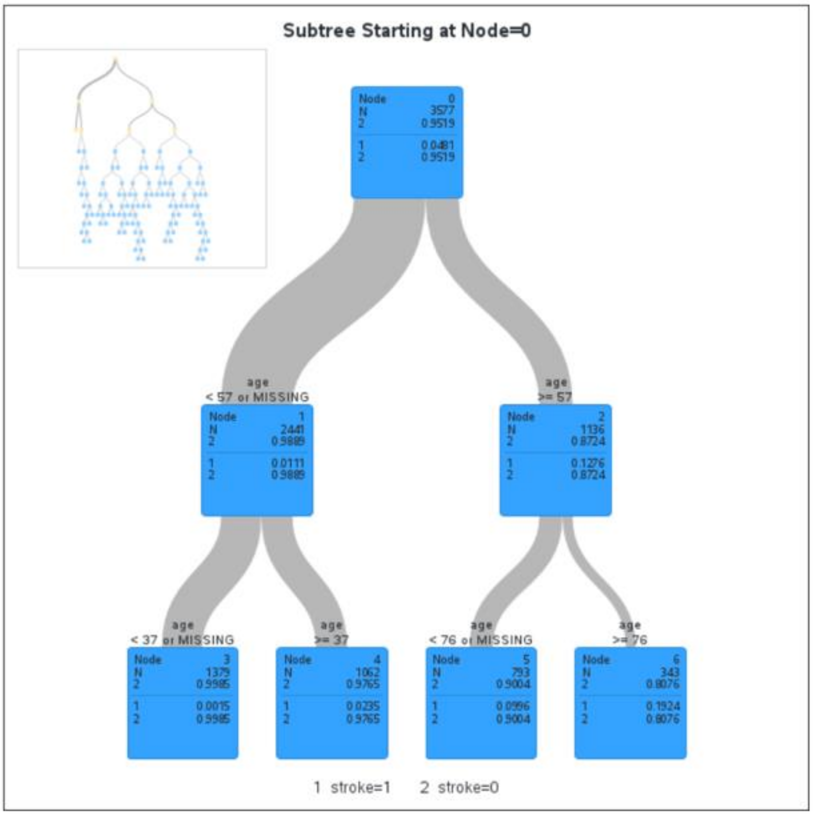


Figure 27 decision tree: subtree starting at node=0
Source: author's own work

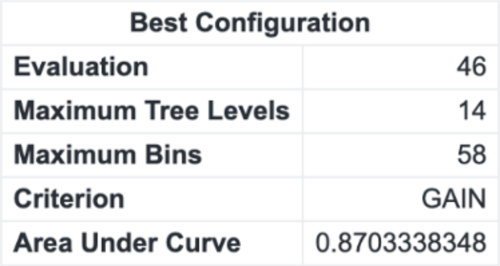
As shown in Figure 27 above, age=57 is taken as the splitting point from root Node 0 with a splitting threshold of age<57 , age>=57, and the following splitting points are taken as age=37 and age=76 with splitting thresholds of age<37, age>=37 and age<76, age>=76 respectively. The other age splitting points are age=7, age=17.

The segmentation thresholds for avg_glucose_level are avg_glucose_level <126.33, avg_glucose_level >=126.33

The segmentation threshold for hypertension is 1,0

The segmentation threshold for heart_disease is 1,0

The segmentation threshold for stroke is 1,0



Best Configuration	
Evaluation	46
Maximum Tree Levels	14
Maximum Bins	58
Criterion	GAIN
Area Under Curve	0.8703338348

Figure 28 best configuration of decision tree
Source: author's own work

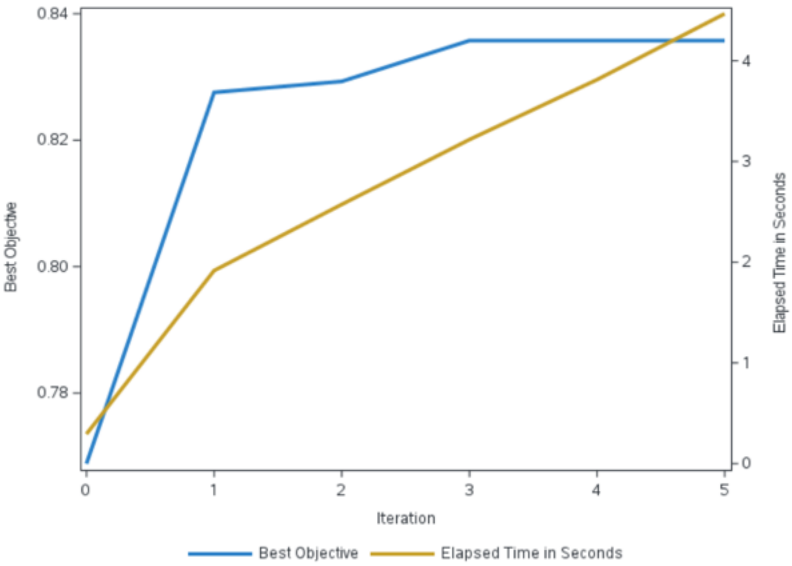


Figure 29 AUC curve – tree
Source: author's own work

Figure 29 shows the AUC of the decision tree, and from the above graph, both the accuracy and the ROC show that the model of the decision tree is higher, as it has the highest accuracy rate.

5 Results and Discussion

5.1 Models Evaluation

Comparing Figure 23 and Figure 27 In the logistic regression model, the AUC area under the curve equals 0.8416, which is slightly lower than the decision tree model's AUC area of 0.8703. This is 3% less accurate for prediction of large data, so we can clearly see that the decision tree model is slightly more accurate than the logistic regression model.

Given the overall specification and the purpose of the study, in its current form, the logistic regression results are better than the decision tree for warning of this activity by predicting stroke, albeit with slightly higher overall accuracy values. The tree model is characterised by more accurate predictions, whereas logistic regression, despite slightly lower predicted behaviour, provides a more detailed and accurate statistical analysis of risk factors. Given the acute and sudden nature of stroke in medicine and its high mortality rate, the logistic regression model seems more appropriate here, as predicting the probability of stroke in advance through big data, artificially interfering with and reducing the influence of risk factors, and reducing the probability of stroke occurring, is more beneficial than a generally accurate diagnosis of stroke in terms of reducing the financial loss to patients and reducing the medical burden on society.

Looking at the coefficients of the various influencing factors in the logistic regression, we were able to explain the effect of individual factors on the final outcome. We were therefore able to compare the results presented by our model with the characteristics of the individual factors mentioned in the theoretical section.

- Age
- Hypertension
- Heart disease
- Average blood glucose level

From the regression model we can see that age is a very important factor, which is also the most important factor in causing stroke as mentioned before. Thus as age increases, the chances of having a stroke also increase. In this case, the effect of the variables in the model is as expected.

Hypertension, a variable describing how high the body's blood pressure is. is proportional to the odds of having a stroke in the model; higher blood pressure means a greater risk of having a stroke, but it also means that the odds of having a stroke can be reduced by controlling blood pressure. is an artificially controlled variable and the results achieved in the model are consistent with expectations.

Heart disease, people with heart disease often have a variety of blood pressure problems and the fibrillation of the heart varies with age. Having heart disease often means that people are at high risk of having a stroke.

Finally, Average blood glucose level, this indicator means the level of blood glucose concentration in the blood and also represents the presence of diabetes. Stroke is responsible for 20% of deaths in people with diabetes. So an increase in blood glucose levels can also affect the probability of a stroke. This is in line with the expected findings.

6 Conclusion

Healthcare has always been one of the most pressing issues facing people's lives, and the impact of stroke is now expanding, and its effects on human life, including personal physical disability, property loss and overall workforce decline, prompted me to choose predicting stroke as the topic of my thesis. In turn, big data analytics provided the opportunity and challenge of building a predictive stroke model.

In many cities or countries with low population densities, it is simply not possible to generate predictive models without a large amount of data to support them. The data used in this thesis was obtained from the Kaggle data exploration website and was used for academic research purposes only.

This thesis, begins with a theoretical understanding of the terminology of stroke and the factors that influence it through a literature review, in addition to which the literature study provides a general discussion of the concept of big data, listing the structure and challenges of big data. This is followed by a list of various forms of data analysis models and their methods.

In four sections, this study compares models based in SAS studio through the algorithms of logistic regression and decision trees, and the test results show that decision trees perform slightly better than logistic regression models in terms of accuracy. The study showed that these algorithms were able to predict strokes with 84% accuracy. Both models demonstrated that age was the greatest risk factor for stroke and the second influence on stroke is average blood glucose levels. Hypertension and heart disease can also be influential factors for stroke.

This study could help provide useful information for stroke prediction in the health care system for manual intervention and early warning for people at high risk of stroke. For this thesis, there are some limitations and some factors (ethnicity, diet, bacteria, etc.) may have an impact on the probability of having a stroke, and proper data is needed for a more comprehensive analysis and comparison.

7 References

Kase, C. S., Mohr, J. P., & Caplan, L. R. (2004). *Stroke (Fourth Edition)* (Vols. Chapter 13 - Intracerebral Hemorrhage). ISBN 9780443066009, https://doi.org/10.1016/B0-44-306600-0/50017-1. (https://www.sciencedirect.com/science/article/pii/B0443066000500171): Churchill Livingstone.

Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst.* Prague: John Wiley \& Sons.

Agarwal, A. (2018, 10 5). *Linear Regression using Python*. Retrieved from Towards Data Science: https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2

An, J. (2020, 6 27). *How to Remember all these Classification Concepts forever*. Retrieved from medium: https://medium.com/swlh/how-to-remember-all-these-classification-concepts-forever-761c065be33

Cukier, K. (2010, 2 25). *Data, data everywhere: A special report on managing information*. Retrieved from The Economist: http://www.economist.com/node/15557443.

Diebold, F. X. (2012, November 26). A Personal Perspective on the Origin(s) and Development of 'Big Data': The Phenomenon, the Term, and the Discipline, Second Version. *PIER Working Paper*, pp. No. 13-003. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2202843

Emily , L., Andre , T., Cecile , V., & Mauricio, S. (2021). Toward the use of neural networks for influenza prediction at multiple spatial resolutions. *SCIENCE ADVANCES, 7*(25), DOI: 10.1126/sciadv.abb1237.

Erwan , S. (2016, 4). On the asymptotics of random forests. *Journal of Multivariate Analysis, 146*, pp. 72-83.

Facundo, B., Juan , G. M., & Víctor , F. D. (2018, 1 1). Prediction of wind pressure coefficients on building surfaces using artificial neural networks. *Energy and Buildings, 158*, pp. 1429-1441. Retrieved from https://www.researchgate.net/figure/Artificial-neural-network-architecture-ANN-i-h-1-h-2-h-n-o_fig1_321259051

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, Volume 35*, pp. 137-144.

Gartner IT Glossary. (n.d.). Retrieved from http://www.gartner.com/it-glossary/big-data/

Gray, C. (2017, 6 6). *Decision Tree Hugging*. Retrieved from Towards Data Science: https://towardsdatascience.com/decision-tree-hugging-b8851f853486

Hashem, I. A., Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, & Samee Ullah Khan. (2015). The rise of “big data” on cloud computing: Review and open research issues. *Information Systems, 47*, pp. 98-115.

Jia, Q., Liu, L., & Wang, Y. (2011, 9). Risk Factors and Prevention of Stroke in the Chinese Population. *Journal of Stroke and Cerebrovascular Diseases, 20*(5), pp. 395-400.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety.[J]. *META group research note, 6*(70), p. 1.

Leszek , R., Maciej, J., Lena, P., & Piotr, D. (2014, 5 10). The CART decision tree for mining data streams. *Information Sciences, 266*, pp. 1-15.

M., S. (2010). 5 - Use of artificial neural networks (ANNs) in colour measurement. In G. M.L. , *Colour Measurement* (pp. 125-146). ISBN 9781845695590, https://doi.org/10.1533/9780857090195.1.125. (https://www.sciencedirect.com/science/article/pii/B9781845695590500042): Woodhead Publishing.

Murphy, S. J., & Werring, D. J. (2020). Stroke: causes and clinical features. *Medicine, 48*(9), 561-566.

Qef. (2008, 7 1). *The logistic curve*. Retrieved from wikimedia: https://en.wikipedia.org/wiki/Sigmoid_function#/media/File:Logistic-curve.svg

Raphael , M., Daniel, K. M., Santiago, U. R., Rose , F., & David, A. J. (2021, 12). Classical Risk Factors for Ischemic Stroke are not Associated with Inpatient Post-Stroke Mortality in Sickle Cell Disease. *Journal of Stroke and Cerebrovascular Diseases, 30*(12), p. 106089.

S.K. , J., & Snehashish, C. (2021). Chapter 10 - Fuzzy eigenvalue problems of structural dynamics using ANN. In C. Snehashish, *New Paradigms in Computational Modeling and Its Applications* (pp. Pages 145-161). ISBN 9780128221334, https://doi.org/10.1016/B978-0-12-822133-4.00010-4. (https://www.sciencedirect.com/science/article/pii/B9780128221334000104): Academic Press.

Sivarajah, U., Kamal, M. M., Iran, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research, 70*, 263-286.

Sodeman, W. A., & Sodeman, T. C. (2005). *Ischemic Stroke: Patient and Caregiver's Guide, Instructions for Geriatric Patients (Third Edition)* (Vol. 2: Cerebrovascular Diseases). ISBN 9781416002031, https://doi.org/10.1016/B978-141600203-1.50010-X. (https://www.sciencedirect.com/science/article/pii/B978141600203150010X): Saunders.

Statistics Solutions. (2022). *Assumptions of Linear Regression*. Retrieved from statisticssolutions: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-linear-regression/

Stephen, M. J., & David , W. J. (2020, 9). Stroke: causes and clinical features. *Medicine, 48*(9), pp. 561-566.

Tabachnick, B., Fidell, L., & Ullman, J. (2007). *Using multivariate statistics* (Vol. 5). Boston: Boston, MA: pearson.

Takeshi , I., Yuji , I., Tetsuji, I., Naoki, I., Ryo , T., Haruki , T., . . . Osamu , A. (2021, 4). Decision Tree Algorithm Identifies Stroke Patients Likely Discharge Home After Rehabilitation Using Functional and Environmental Predictors. *Journal of Stroke and Cerebrovascular Diseases, 30*(4), p. 105636.

Tan, J., Ramazanu, S., Liaw, S. Y., & Chua, W. L. (2022). Effectiveness of Public Education Campaigns for Stroke Symptom Recognition and Response in Non-Elderly Adults: A Systematic Review and Meta-Analysis. *Journal of Stroke and Cerebrovascular Diseases, 31*(2), 106207.

Tang, L., Li, J., Du, H., Li, L., & Wu, J. (2022). Big Data in Forecasting Research: A Literature Review. *Big Data Research, 27*.

Tatjana, R., & Ralph, S. L. (2008, 11). Risk Factor Management to Prevent First Stroke. *Neurologic Clinics, 26*(4), pp. 1007-1045.

Team Asana. (2021, 12 6). *What is decision tree analysis? 5 steps to make better decisions*. Retrieved from asana: https://asana.com/zh-tw/resources/decision-tree-analysis

The Economic Times. (2022, 3 5). *Definition of 'Decision Tree Model'*. Retrieved from economictimes: https://economictimes.indiatimes.com/definition/decision-tree-model

Tuffery, S. (2011). *Data mining and statistics for decision making.* John Wiley \& Sons.

Venkata , J. (2020, 8 6). *TIBCO Community*. Retrieved from Random Forest Template for TIBCO Spotfire: https://community.tibco.com/wiki/random-forest-template-tibco-spotfire

Wang, B., & Wang, Y. (2021, 11). Big data in safety management: An overview. *Safety Science*, pp. Volume 143, 105414, ISSN 0925-7535, https://doi.org/10.1016/j.ssci.2021.105414.(https://www.sciencedirect.com/science/article/pii/S0925753521002587).

WHO. (2022, 12 9). *The top 10 causes of death*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

Zhao, P., Su, X., Ge, T., & Fan, J. (2016, 3). Propensity score and proximity matching using random forest. *Contemporary Clinical Trials, 47*, pp. 85-92.

8 Appendix

Hypertension: 0="have not hypertension" 1="have hypertension"	**Heart_disease :** 0=" have not heart disease" 1="have heart disease"	**Bmi:** 1="0-20" 2="21-40" 3="41-60" 4="61-80" 5="81-100"
Stroke: 0="have not stroke" 1="have stroke"	**Age:** 1="0-10" 2="11-20" 3="21-30" 4="31-40" 5="41-50" 6="51-60" 7="61-70" 8="71-80" 9="81-90"	**Work_type:** 1="Self-employed or Private" 2="Never_worked" 3="Govt_job" 4="Children"

Appendix 1. value coding for selected variables