

College of Business,

Technology and   
Engineering

MSc Dissertation Report

**Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm**

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in **Big Data Analytics**

|  |  |
| --- | --- |
| Student Name | Oluwafemi Emmanuel Zachariah |
| Student ID | 31034108 |
| Supervisor | Olamilekan Shobayo |
| Date of Submission | September 8, 2022 |

STATEMENT 1:

This dissertation does NOT contain confidential material and thus   
can be made available to staff and students via the library.

**ABSTRACT**

Stroke is one of the major causes of death across the world, Stroke disease is caused when there is an obstruction in the free flow of blood to different region of the brain, it has been studied and documented in different research worldwide. Early detection of the both the modifiable and non-modifiable risk factor of stroke would help in the reducing the death rate incurred by this illness. Quite a few researchers have predicted the incidence of stroke disease from patients’ medical features using different machine learning algorithms. However, there have not been many studies that have predicted stroke disease with demographic and behavioural data. This work seeks to predict the incidence of stroke disease using Random Forest, Decision Tree and logistic regression machine learning algorithm to determine stroke best predictive algorithm out of the three algorithms. The data used for this study has 5,110 observations with 10 attributes which was gotten from Kaggle, it contains health records which were acquired from different hospitals of Bangladesh by a group of researchers for academic purpose. This study reveals the different stroke risk factors and other medical attributes that may trigger the incidence of stroke disease. Afterwards the result of this algorithms was evaluated using confusion matrix and ROC curve. Random forest gave the best performance with accuracy of 94.11% however logistic regression and decision tree have accuracy of 91.43% and 88.83% respectively. The result of this research recommended that stroke disease have several risk factors which are divided into modifiable and non-modifiable risk factors. Non-modifiable stroke risk factors cannot be changed but may be managed if detected at the early stage, the incidence of stroke disease increases with aging i.e., adults are mostly at risk of developing stroke disease, male gender has the highest stroke incidence but female experience higher incidence at older age. Body mass index, hypertension and heart disease are very strong modifiable risk factor of stroke disease.

**ACKKNOLEDGEMENT**

I would like to express my deepest gratitude to Almighty God for the success of this work. The project and the certification mean so much to me, nothing would have been achievable without the divine intervention of God. My journey to United Kingdom is not just by mistake, God made it happen.

My reserved appreciation goes to my supervisor, Mr Olamilekan Shobayo for his assistance, valuable comments, guidance and contribution during this dissertation. You are really God sent; I appreciate all your labour of love. It would be practically impossible without his support to achieve this project. Your constructive comments, patience, enthusiastic contribution and invaluable technical suggestion has immensely transformed my research skills. It has helped me to develop self-confidence and greatly influence my personal development as a data analyst. I am so much grateful for his commenting and correcting my views and evaluating my work as well as helping me develop new idea. I also appreciate other lecturers in the Department of data analytics for all their remarkable contribution in the completion of my programme in the University. May God bless you all.

I owe my parent Mr Morebise Zachariah and Mrs Olamide Zachariah unquantifiable gratitude for their immeasurable support both financially and morally. Words of mouth cannot describe your contribution to my success most especially in acquiring my master’s degree in the UK. My appreciation goes to my wife Deborah, the completion of this project would not be achievable without her encouragement and support. She has shown a deep level of unconditional love and provided me with courage and strong motivation to complete this research work.

I sincerely thank my other siblings Aanuoluwapo, Ogooluwa, Oluwaseun and Elizabeth for the prayers both day and night, I just want to let you know that your prayers have done wonders, I love you so much. Gratitude goes to the rest of my family members for their inspiration, calls and encouragement during the rigours of the research work. You all have contributed immensely towards my success.

I specially appreciate my bosom friend Adeniyi Bukola for his support so far. You’re truly a good friend, thank you for your help, advice and encouragement most especially to cheer me up at the point I needed him. My acknowledgement would not be complete if I fail to appreciate my course mates the likes of Debo, AY, Tosin, Bassey and Henry. You guys are amazing.

**TABLE OF CONTENTS**

Title Page……………………………………………………………………………I

Abstract.......................................................................................................................II

Acknowledgement…………………………………………………………………...III

Table of Content..........................................................................................................IV

**Chapter one**

* 1. Introduction­­­­ ………………………………………………………………………1
  2. Motivation of Research …………………………………………………………..2
  3. Research Questions, Aims and Objectives………………………………………..4
     1. Research Aims ………………………………………………………..4
     2. Research Questions……………………………………………………..…4
     3. Research Objectives……………………………………………………….5
  4. Deliverable………………………………………………………………………..5
  5. Contribution to knowledge………………………………………………………..5
  6. Structure of the Research………………………………………………………….6

**Chapter two**

2.0 Introduction……………………………………………………………………...…7

2.1 Overview of Stroke Disease………………………………………………….…….8

2.2 Diagnosis of Stroke Disease…………………………………………………...…...8

2.3 Stroke Risk Factors………………………………………………………………....8

2.3.1 Non-modifiable Risk factor of Stroke Disease……….…………………….…9

2.3.2 Modifiable Risk Factor of Stroke Disease………………………………….…11

2.3.2.1 Epidemiology of Stroke and Hypertension……………………………..…...12

2.3.2.2 Body Mass Index and Risk of Stroke……………………………………..…13

2.4 Review of Predictive Analysis in Healthcare using Different Algorithms….……...13

2.5 Summary……………………………………………………………………...……..18

**Chapter Three**

3.0 Introduction………………………………………………………………………19

3.1 Research Philosophy, Approach, Strategy and Methodology…………………....19

3.2 Tools and Techniques, Data Collection and Analysis………………..…………..20

3.2.1 Data Mining……………………………………………………………..…..20

3.2.2 Machine Learning………………………………………………………..….20

3.3 Theoretical Overview of Predictive Model…………………………………….....22

3.3.1 Logistic Regression………………………………………………………….23

3.3.2 Decision Tree………………………………………………………………..24

3.3.3 Random Forest……………………………………………………………….24

3.4 Data Description………………………………………………………………….25

3.5 Data Pre-processing……………………………………………………………....30

3.6 Research Ethics, Risk and Issues………………………………………………....33

3.6.1 Ethics and Legal issues………………………………………………………33

**Chapter** **four**

4.0 Introduction………………………………………………………………….……35

4.1 Exploratory Data Analysis of Demographic factors and Attributes in the Dataset.35

4.1.1 Chances of Getting Stroke Disease using the Gender Attribute……………..36

4.2 Machine Learning Algorithms…………………………………………………….43

4.3 Model Accuracy…………………………………………………………….……..45

**Chapter five**

5.0 Introduction……………………………………………………………….………..47

5.1 Result Findings………………………………………………………………….….47

5.1.1 Result of Variable correlation…………………………………………………47

5.2 Summary……………………………………………………………………………50

**Chapter six**

6.0 Introduction…………………………………………………………………………52

6.1 Conclusion…………………………………………………………………………..52

6.2 Recommendation……………………………………………………………………53

6.3 Limitation of the study………………………………………………………………53

References ………………………………………………………………………………55

Appendix…………………………………………………………………………………62

**Chapter One**

**Introduction**

**1.1 Introduction**

The study of stroke disease has been in existence for a long time and it has been documented in various studies worldwide. Warlow (1998) in his study of epidemiology of stroke disease identified mortality, incidence, prevalence, long term outcome and cost as the various facets of stroke burden. About 2-4% of healthcare cost is consumed by stroke treatment which however increases to about 4% of the healthcare cost in industrialized countries. Stroke is considered as the second most frequent cause of death worldwide which amounts to loss of five million lives on a yearly basis, the proportion of death caused by stroke disease ranges from 10-12% across the western countries with average of the victims around 65 years. It is considered as a medical emergency which requires urgent attention. To reduce damage done to the brain and other part of the body, early detection and proper management is therefore necessary to minimize stroke death rate worldwide. The different studies of stroke disease and the cost of treatment has prompted many researchers both in the medical field and in the IT space using machine learning algorithms to reduce its occurrence and prevent it prevalence among the populace.

Different machine learning algorithms have been used to predict stroke, for instance build a stroke prediction model using Artificial neural networks to detect stroke disease at the early stage. However, only few studies have carried out comparison between the different machine learning algorithms that has been used for the prediction of stroke to determine the most effective machine learning algorithms to predict stroke disease however there is a need to discover which of the algorithms have the highest level of accuracy to predict stroke disease. Many studies have used algorithms such as decision tree, KNN, logistic regression etc. but few studies have implemented random forest algorithm in the prediction of stoke disease using stoke set. Random forest has track record of high accuracy in the prediction of diseases in different medical research work therefore this work hereby seeks to investigate the prediction of stroke disease using three selected algorithms which are logistic regression, decision tree and random forest to identify the algorithm with the best performance.

This project focuses on the prevalence of stroke disease and seek to investigate the various risk factors of stroke disease both modifiable and non-modifiable risk factors, review literature on stroke disease and evaluate various predictive models used in predicting the occurrence of stroke disease. It specifically focuses on getting deep knowledge on demographic pattern of its occurrence and development of the disease to gain insightful knowledge on how it can be prevented or cured.

This work mainly focusses on building three machine learning algorithms for the prediction of stroke disease. The algorithms are Random Forest, Decision Tree and Logistic regression. The choice of this algorithms is due to their known level of accuracy among other machine learning algorithms which is exactly what is needed to for this project because it is medical related research henceforth, we need a reliable algorithm to build the required model. This project will however compare the level of performance of these algorithms and henceforth recommend the best predictive algorithm for stroke disease.

**1.2 Motivation of Research**

Across the globe, cardiovascular disease is one of the major causes of death. World health Organization reported in 2019 that about 17.9 million lost their lives on a yearly basis to this illness which contribute to 32% of death globally and Stroke is one of the major causes of cardiovascular disease in the world (Rochmah et al., 2021). WHO in 2001 emphasized that middle-income and low-income countries have a proportion of 85.5% of the total death rate of stroke disease worldwide, the body opined that the disability adjusted life years (DALYs) in both low-income and middle-income countries was approximately times seven of those lost in high income countries.

Chart, bar chart

Description automatically generated

*Fig 1: Estimate of death rate by WHO. (Sirsat, et al., 2020)*

In the research carried out by Mathers et al. (2006), it was emphasized that in every 20 adults who are age 14 and above, there is a likelihood of one out of them being affected by stroke disease although they argued that the death rate significantly differs in countries but insisted that low-income countries are likely to witness high death rate. (Rothwell et al., 2012). Stroke disease is a threat to human lives especially the old people within the age of 50 years and above, it is the major chronic disabilities, one of the causes of dementia which results to death in most of the cases if not properly handled. (Roger et al., 2011).

The world impact of stroke disease tends to be higher in developed countries like United Kingdom and United States. In the US for instance, 3% of the population of the adult has prevalence of stroke disease which is approximately 7 million people. Rochmah et al. (2021), among other authors established that stroke disease is identified as one principal cause of cardiovascular disease worldwide.

Furthermore, Centres for Disease Control and Prevention CDC, (2005) argued that under-developed country are at the risk of higher incidence of stroke disease and death rate due to the lack of adequate management and control system. However early detection and awareness of different risk factors and warning signs of this disease would minimize its occurrence among the populace and will be managed properly.

Application of data mining techniques and machine learning algorithm in the prediction and control of stroke disease would therefore relatively reveal new knowledge in the treatment of stroke disease especially to detect patients who are at risk of stroke disease based on some risk factors, help medical practitioners for decision making at the management level as well as generating relevant scientific hypotheses to enhance early detection of this disease so it can be properly managed.

Various models have been used in analysing stroke disease. Take for instance, Lumney et al. (2002) in their attempt to build stroke prediction model used cox proportional hazard model which is one of the most popular statistical methods in medical research. This model has been studied by other researchers and applied in predicting different diseases including stroke. Although the accuracy of the model depends on the relevance of the preselected features. (Khosla et al. 2010).

Qu et al. (2011) in their investigation in optimizing acute stroke outcome prediction models compared generalized regression neural networks (GRNN) and Logistic regression (LR). The outcome of the research showed that GRNN significantly performed better than LR model in predicting prognosis of acute stroke patients. They further suggested that GRNN is an optimal model for the prediction of acute stroke.

**1.3 Research Question, Aims & Objectives**

**1.3.1 Research Aims**

This project aim is to gain in-depth understanding of the various risk factors which cause stroke disease, build stroke prediction models using various risk factors from historical dataset to help individuals and medical practitioners in clinical decision-making by early detection and proper management of the disease.

**1.3.2 Research Questions**

* **What are the major risk factors for stroke disease?** This question will be answered by review different research work about stroke disease as well as the risk factors that has been identified.
* **How can we enhance stroke disease predictive model using these risk factors?** This question will be answered by building three machine learning models which are Random Forest, Logistic regression and Decision Tree.
* **What is the optimal statistical model for the prediction of stroke disease?** To answer this question, we explore the performance of the developed model, the model with the highest level of performance will be suggested as the optimal statistical model for stroke prediction.

**1.3.3 Research Objectives**

The project’s main objective can be summarized as follows:

* To investigate in detail the literature of both the theoretical and practical area based on previous studies of stroke disease.
* To generate insights on different medical attributes that could lead to stroke disease.
* To investigate range of predictive models used by stroke disease researchers as well as the risk factors associated with the disease.
* To use insights generated from objective 1 and 2 to develop predictive models appropriate to estimate risk of having stroke disease based on some modifiable and non-modifiable risk factors.
* Evaluate the performance of the developed models to identify the best out of the models.

**1.4 Deliverable**

A predictive model which appropriately make prediction whether an individual is at risk of developing stroke disease based on his/her modifiable and non-modifiable risk factor. This model will help medical practitioners in the early detection of the disease so it can be prevented or managed.

**1.5 Contribution to knowledge**

This research is to investigate both the risk modifiable and non-modifiable risk factor of stroke disease as well as building three statistical model which are Decision tree (DT), Logistic Regression (LR) and Random Forest using these risk factors as variables to predict the likelihood of the occurrence of stroke disease based on some hereditary and medical features of an individual. In addition, the performance of this models will be evaluated, and the best model will be chosen as the best predicting model.

There is no evidence of literature which documents the use of these three models concurrently to prediction of stroke disease henceforth, this research would be a source of reference to similar research in the future and to other professionals in the field of stroke disease all over the world especially in underdeveloped countries.

In addition, the study would further serve as point of reference and a source of literature to other researchers investigating the application of machine learning in the medical field.

At the end of this research, it is expected that the result of the analysis would be helpful in the early detection of stroke disease, as well as facilitating the selection of appropriate machine learning model and statistical techniques in analysing stroke datasets in other stroke related research. This project would be beneficial to the public in general as this result of this work would create awareness of the risk factors of stroke and this will reduce the prevalence of stroke in the society.

**1.6 Structure of the Research**

This research work is divided into five chapters, chapter one gives detailed introduction about the research topic, motivation, research question, research aims and objectives, deliverables and contribution to knowledge.

In chapter two, critical review of related literature is discussed, it explores the detailed review of the three machine learning algorithms used for the research work including the theoretical framework. Research methodology, the research approach for modelling such as decision tree, logistics regression and Random Forest is discussed in chapter three to investigate the research problem.

Chapter four discuss the outcome of the research with exploratory analysis of the variables and the result of the analysis and final recommendation is discussed in chapter five.

**Chapter two**

**Literature review**

1. **Introduction**

In this section, a review of literature is established to identify and justify this proposed project. This section focusses mainly the overview of stroke disease, the risk factors as well as a review of different existing predictive algorithms in healthcare. This will in turn generate insights in prevention and treatment of stroke disease. in addition, it will reveal how some of these factors can be avoided as much as possible. This chapter explain in detail the factors causing stroke disease, early detection of this factors would give an idea on how the disease can be treated.

**2.1 Overview of stroke disease**

Stroke disease is caused when there is an obstruction in the free flow of blood to different region of the brain. When this happens, cells in this area suffers from insufficient oxygen and the required nutrients to function properly which will result to death of the brain cells.

It may be explained as an abnormal function of some area in the body which may be due to some injury incurred in the brain, spinal cord, muscles or nerves of the body by a vascular cause. (Sacco et al., 2013). Research has shown that ischemic stroke is the most common type of stroke, which is likely caused by reduced blood flow, generally because of arterial occlusion. Venous infraction, which is due to blockage of blood vessels, veins is a rare type of ischemic stroke. However, the remaining proportion of 10-40% of stroke occurrence taking regional epidemiology into consideration are haemorrhagic stroke which is caused by cerebral aneurysm (Zhang et al., 2003)

Feigin et al., (2003) in their study established that the rate of stroke among the United Kingdom population ranges from 1.3/1000 which is a bit similar to that in Japan, Philippines stroke rate ranges from 1.7/1000, New Zealand ranges from 10.2/10000 which is very high. The occurrence of stroke disease is higher among the male gender compared to their female equals (Hollander et al., 2003) and recent study have shown that blacks are more likely to be diagnosed of stroke compared to the white population.

**2.2 Diagnosis of Stroke**

The major clinical feature of stroke is the unexpected body system malfunction which arises from the nerves, spinal cord or the function of the brain. The timing of this malfunction may be hidden if the victim suddenly awakes with symptoms of stroke disease or the victim is unable to communicate or probably do not have insights to determine the time of deficits. (Bruce and pooja, 2020).

Stroke symptoms which are under recognized according to (Bruce and pooja, 2020) are nausea, decreased level of consciousness, vomiting and vertigo are identified to be common in the setting of occlusions in the posterior circulation. (Arch et al., 2016).

Early detection and proper management of stroke would be needed to reduce further harm in the affected region of the brain as well as preventing further complications in the body system (Robert B, 2002).

When the populace is aware of the different risk factors of stroke disease, this would reduce its occurrence in the community. Vinereanu et al. (2017) revealed in their research that little percentage of participants of about 60-70 % were able to identify the risk factors of stroke disease however, Amelia et al. (2017) argued that stroke disease unlike other disease such as myocardial infection which happens due to large vessel atherosclerotic disease that deteriorate the function of coronary arteries. Stroke diseases occur in different ways therefore it is difficult to identify its risk factors.

**2.3 Stroke risk factors**

Risk factors of Stroke disease has been investigated by different researchers in the past, it is majorly categorized into two forms which are ischemic and haemorrhagic strokes. In the research of Adams et al. (2017), they break haemorrhagic stroke into intraparenchymal or subarachnoid and ischemic into etiologic subcategories namely large-artery atherosclerosis, cardio-embolism, small-vessel occlusion, stroke of other determined aetiology, and stroke of undetermined etiologic thought, all these to capture the causes of stroke disease.

However, the two types of strokes have similar risk factor and symptoms, so clinical symptoms and signs is not sufficient to differentiate between these two types of strokes. Haemorrhagic stroke occurs when there is one or more vessels in the brain bleeds or when bleeding into the outer and inner layer of the tissue which covers the brain, ischemic stroke is due to the occlusion of the arteries in the cerebral circulation by atherosclerotic plague.

In ischemic stroke for instance, the risk factors vary in the etiological categories according to (Tirschwell et al., 2004). One of the major risk factors of haemorrhagic is hypertension although it can result to atherosclerotic disease which will eventually lead to ischemic stroke. Adams et al. (2017) emphasis that hyperlipidaemia which is considered as a coronary risk factor is also a stroke disease risk factor.

Stroke Risk factors may be classified into modifiable (such as diet and conditions etc.) and non-modifiable risk (such as race, age etc.) factors. It may also be categorized into short-term risk or triggers (such as sepsis, infectious disease, stress), intermediate risk factors (such as hypertension and hyperlipidaemia) and long-term risk factors (e.g race and sex) (Amelia et al., 2017).

Race-ethnicity, age, sex and genetics has been identified as nonmodifiable risk factors of stroke disease.

**2.3.1 Non-Modifiable Risk Factors of Stroke Disease**

Roger et al. (2012) in their research on heart disease and stroke emphasised that stroke disease is a disease of aging henceforth the occurrence of stroke increases rapidly with aging. Kissela et al. (2012) opined that stroke disease in 2005 have 69.2 years average age of incidence, however recent studies argued that the mean age of incidence has now been increased to 20-54 years age group.

In childhood, study have shown that the incidence of stroke disease is higher in male compared to female and weak functional outcomes (Appelros et al. 2003)

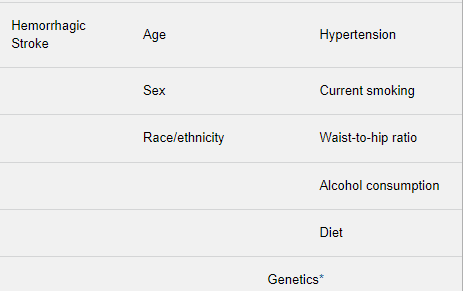
In middle age, women begin to witness high rate of occurrence of stroke disease and further increases as they attain menopause age and as they are losing females hormones. (Appelros et al. 2003). After middle age, stroke disease in women increases and studies have shown that higher stroke incidence have been reported over the years in elderly women above 85 years compared to elderly men. (Bots et al. 2017)

However, testosterone level is stable in life span of women compared to oestrogen level between 30 to 70 years (Holmegard et al. 2016), reduced level of testosterone has been attributed to the increase occurrence of stroke disease in men but no visible relationship was detected for testosterone level and risk of stroke in women. Holmegard et al. (2016) in their investigation on sex hormones and stroke disease have not detected any clear relationship between high or low level of testosterone and the occurrence of stroke in women. Furthermore, dehydroepiandrosterone which is adrenal hormone suitable for synthesis of testosterone and oestrogen was investigated and confirmed as one of the factors which increases the risk of stroke disease most especially in men.

In addition, Choudhury et al. (2015) also emphasis that stroke is associated with several risk factors which may either be modifiable or irreversible. He further explains that irreversible factors are age, hereditary, race, and gender and modifiable factors are, heart failure, diabetics, heart disease, hyperlipidaemia, hypertension and so on. Both ideas however are driven towards the same point

Xavier et al. (2010) carried out international research to investigate risk factors of stroke disease. The research comprises of 22 countries with a total of three thousand (3000) patients were used for the study, 2337 out of the overall sample were ischemic stroke patients and 637 were haemorrhagic patients. The table below gives the summary of their findings.

Graphical user interface, text, application, email

Description automatically generated

*Figure 1- Modifiable and non-modifiable risk factors of stroke by (Xavier et al., 2010).*

Among other factors, age is a major non-modifiable stroke risk factor, the incidence of stroke increases with respect to aging and approximately doubles every 10 years between when the individual approaches age 45-85 (Choudhury et al., 2015). However, Roger, et al. (2012) reported that incident of stroke doubles after 55 years.

**2.3.2 Modifiable Risk Factors of Stroke Disease**

Since the nonmodifiable risk factors cannot be controlled or prevented, attention should be focused on the modifiable risk factors of stroke disease as intervention strategies to curtail these factors would consequently lessen the incidence of the disease. Roger et al., (2012) argued that it is very crucial to identify and modify these risk factors at the early stage to avoid further complications on the victims. Modifiable risk factors may be divided into two namely behavioural and medical condition modifiable risk factors. Medical risk factors include high blood pressure, hypertension, diabetes Mellitus hyperlipidaemia e.t.c. while behavioural risk factors are smoking, substance abuse, alcohol intake, eating unhealthy diet e.t.c.

**2.3.2.1 Epidemiology of Stroke and Hypertension**

Chobannian et.al., (2003) among other authors argued that hypertension is one of the most crucial but modifiable risk factor of stroke disease, their study emphasis on the linear and direct relationship between blood pressure and stroke disease. it has been identified among all the modifiable risk factors as the cause of mortality and third threat to healthy life over the years (Ezzati et al., 2002). It was established in the study of (Kearney et al., 2005) that the 26% world adult population are hypertensive in 2000, this is expected to increase to 29% in 2025. However, the rate of prevalence has a low widespread in some geographical location such as Korea with 19%, some high spread has also been reported in countries like Germany (55.3%). The rate of incidence of hypertension is high among men in European countries and the United States however there may be more incidence in women in sub-Sahara Africa, turkey, Indian and quite a number in Latin America. (Kearney et al., 2005)

Kearney et al., (2005) reported that the incidence of stroke is more likely to increase in these regions as there would be increase in age in both genders. The study further emphasised that 2/3 of hypertensive people are from a developing country which may be due to high population and inadequate health care system.

**2.3.2.2 Body mass Index and Risk of stroke**

One of the well documented risk factors for heart disease is excess weight. (Rimm et al., 1995 ), Hubert et al., 1991) established in their research that excessive weight gain is associated with various risk factor which may indirectly lead to stroke disease including diabetes and hypertension. Research has also shown that excessive body mass index (BMI) would increase risk of stroke disease (Hubert et al., 1991) especially ischemic stroke. Rexrode et al., (1997) in their study examine the relationship between waist-hip ratio and abdominal obesity which may also be considered as risk factor of stroke disease.

The prevalence of stroke disease has slightly reduced over the years among old adults but increased among young people. (Medin et al., 3004), (Rosengren et al., 2013). No clear reason has been detected for rise in stroke incidence among young adults but it is suspected to have been from obesity epidemic (Falkstedt, et al., 2007). Abnormal body mass index is a risk factor of stroke disease (Falkstedt, et al., 2007), there is high association between body mass index in young adulthood and stroke risk in adult most especially in men.

**2.4 Review of Predictive Analysis in Healthcare Using different Algorithms**

Over the years data mining and machine learning has been adopted to assist healthcare personnel to discover different area of inefficiencies and make constructive recommendations in the emergency room. (Masruriyah et al., 2019).

Yen et al. (2011) in the quest to generate a predictive model for cerebrovascular disease used data mining techniques adopting classification algorithm such as decision tree, Bayesian classifier and back propagation neural networks. The algorithm used four hundred and ninety-three (493) samples and eight different variables (patient’s information) which includes blood test and diagnosis as well as physical examination results. This model used was evaluated using sensitivity and accuracy indicators. Decision tree had 92.29% sensitivity and 98.01% accuracy, Bayesian got 87.1% and 91.3% sensitivity and accuracy respectively while back propagation neural network gives 94.82% sensitivity and 97.87% accuracy. Decision trees was finally chosen as the classification algorithm in the prediction of cerebrovascular diseases due to its high performance and level of accuracy when compared to Bayesian classifier and neural networks. However, artificial neural network (ANN) was adopted by Shanthi et al. (2008) to predict Thrombo-embolic stroke disease generating 89% level of accuracy. In the overall research work, Artificial Neural Networks (ANN) exhibits good performance level for predicting stroke disease. In the same way, artificial neural network produces amazing prediction of heart attack in the outcome of research made by Soni et al. (2011), K-means clustering algorithm was adopted after the dataset has been pre-processed and the researchers used Neural Networks with back propagation for training. Fifteen attributes (variables) were used for the prediction with basic data mining techniques like clustering, Artificial Neural Networks etc. The outcome showed that the built algorithm is efficient to predict heart disease (Sudha et al., 2012). Decision tree was finally chosen as the best performing model although Bayesian classification has a similar accuracy (96.5%) to Decision tree with accuracy of 99.2% but no other predictive model like Neural Networks, K nearest Neighbour or other classification-based clustering will be more efficient (Masilamani, 2010).

Other several studies have adopted Artificial Neural Network to predict cardiovascular disease. In the research carried out by Celler et al. (1998), the application of Neural network to classify normal and abnormal Electrocardiogram waveform, the abnormal electrocardiogram had six disease condition. Waveforms were recognized by this classifier with 70% level of accuracy.

In the research carried out by Amini et al. (2013) for stroke incidence prediction, 807 records of both healthy and unhealthy subjects were collected using standard checklist with fifty stroke risk factors which include hyperlipidaemia, alcohol intake, diabetics status and previous experience of cardiovascular disease. K nearest neighbour and decision tree were used for the mining techniques and the result showed that k nearest neighbour had 94% while decision tree was 95%. However, Shanthi et al. (2008) proposed Artificial Neural Networks as a functional model to support existing diagnosed methods in the prediction of Thrombo-embolic stroke disease. The researchers investigated the application of backpropagation and Artificial Neural Networks and the overall outcome showed that Artificial Neural Networks significantly performed better than backpropagation with predictive accuracy of 89%.

In the same vein, Masruriyah et al., (2019) argued that hospitals who are in possession of large patients’ medical records including stroke still do not have appropriate analytics tools to extract information which make it difficult for paramedics, doctors, and physicians to gain adequate knowledge. The researchers therefore identified the need for predictive analytics for early detection of the types of strokes to prevent worsening of patients’ condition status. Artificial Neural Networks method was used for the prediction with 95.15% level of accuracy. In the study of Adams et al. (2016), machine learning was used for stroke prediction. The study established a classification model in the prediction of ischemic stroke using K Nearest Neighbour and decision tree, four hundred case study were collected from different hospitals. The outcome of the research however showed that decision tree performs better than k nearest neighbour with 89% level of accuracy hence the researchers recommended KNN to medical specialist for diagnosis of ischemic stroke patients.

In the research carried out by Govindarajan et al., (2020) using machine learning approach to categorize stroke disorder, used stroke dataset which comprises 507 stroke patients’ records, five machine learning approaches were used which are artificial neural networks (ANN), boosting and bagging, support vector machine, and random forest. Cheng et al., (2014) also used Artificial Neural Networks (ANN) in the study to estimate ischemic stroke prognosis, 82 patients’ records diagnosed with ischemic stroke were used for the analysis and the outcome of their work showed 95% level of accuracy.

In the study performed by Chin et al., (2017) to develop an automated system which will detect ischemic stroke at the early stage using CNN deep learning algorithm, the idea was to enter CT image of the brain, image processing will therefore be done with the aid of the developed system to remove the region which may not be possible for the stroke possible for the stroke area. After this has been achieved, the system will further select the patch image and use data augmentation method to increase the patched image. CNN due to its proven record to recognize ischemic stroke was used to train and test 256 patched images, the result showed that 90% accuracy was achieved and the authors recommend that the paper would effectively assist doctor diagnosis of stroke disease.

Singh et al. (2019) in the quest to develop stroke severity index used 3,577 acute ischemic stroke patients’ records, various machine learning models were adopted for their predictions, but linear regression was their final choice of model with predictive accuracy of 95%.

Cheon et at. (2019) investigated the incidence of stroke among the Korean population, for the sake of this research, medical service use and health behaviour data were used due to its easy accessibility compared to medical imaging data, Deep Neural Networks was used to detect the incidence of stroke disease using the acquired data with 15,099 observations. Principal component analysis (PCA) was implored to extract relevant variables from the dataset for stroke prediction. This machine learning method i.e., scaled/deep neural networks approach performed better as they got the area under the curve value of 83% henceforth it was recommended for doctors to early detect the incidence of stroke disease.

Monteiro et al. (2018) applied machine learning techniques to improve the prediction of functional outcome in ischemic stroke patients. This study is only applied to patients three months after their admission. The result of this research showed area under the curve value of 90%. Kansadub et al, (2015) developed stroke predictive model based on demographic data. Classification algorithms which are Naïve Baiyes, Decision tree and Neural networks were used to analyse and predict stroke using demographic data. The result showed that decision tree performed slightly better with 75% level of accuracy, Naïve bayes and Neural networks have 72 and 74 level of accuracy respectively which is very close. However, in the aspect of safety of life, Neural Networks was concluded as the most effective approach due to its high value of False positive rate (FP) and low false negative (FN).

Süt N et al. (2012) in quest to build predictive model for the prediction of mortality in stroke patients using multilayer perceptron with 6 layers. The dataset used for the study consist of 584 observations of stroke patients which was analysed using multilayer perceptron (MLP), neural networks. The different prognostic factors such as sex, age, hypertension e.t.c on motality in stroke were trained with the six Multilayer Perceptron Neural Networks models. Operative curve methods were implored to examine and compare the performance of the MLP using receiver operating characteristics curve (ROC) method. Quick propagation was chosen as the best algorithm for the prediction of mortality in stroke with predicting accuracy of 80.7%.

Rajinitanth et al. (2012) in their study to build a model which will detect ischemic stroke disease used support vector machine, Artificial Neural Networks (ANN) k- Nearest Neighbour (KNN) and Decision tree. Support vector machine (SVM) with predictive accuracy of 98% was chosen as the final predictive model while KNN, ANN and Decision tree have 97%, 96% and 92% classification accuracy respectively.

Colak et al. (2015) applied knowledge process to predict stroke disease. 297 patients record comprising of 130 sick and 197 healthy people from emergency medicine unit were collected, nine risk factors were used for the prediction however the selection of features was in accordance to Grammar’s V test. Multilayer perceptron, ANN and support vector machine approach were used to build the model.

Maier et al. (2015) in their study to develop classifier for ischemic stroke used nine classification methods namely Random decision forest, generalized linear model, convolutional neural networks etc. to develop the classifier. Thirty-seven (37) multiparametric ischemic stroke dataset was used for study where comparison was made between the nine classification methods. Random decision forest (RDFs) and convolutional neural networks (CNNs) gives the highest classification accuracy than other methods.

Adam et al. (2016) developed a classification model for ischemic stroke, only decision tree and k-nearest neighbour (KNN) was used for this study with 400 ischemic stroke patient’s dataset which were collected from different Sudanese hospitals. The result of their research showed that decision tree performs significantly better the KNN.

In the works of Chantamit-o-pas et al. (2018), Recurrent Neural Network (RNN) was investigated with Long Short-term Memory (LSTM) hidden unit. The main idea was to investigate the ability of Long Short-Term Memory (LSTM) to recognize patterns in multi-labelled classification of cerebrovascular system or stroke. The dataset used for this research were gotten from Department of Medical service in Thailand with 326,152 observations, three models were compared in this research which are Back propagation, Recurrent Neural Network and Long Short-Term Recurrent Neural Network 227 risk factors for stroke were used for calculating and making prediction. The outcome of this research showed that Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) with accuracy of 92.79% performed significantly better than Back propagation and Recurrent Neural Network (RNN) with 89.12% and 88.28% respectively. However, the precision and F1 measure scores are different from those of Back propagation and Recurrent Neural Network algorithm, therefore the accuracy rate of the algorithms depends on the size of the sample observation. The result of this work will be more reliable if there is a larger dataset for this prediction. However, the findings of the research carried out by (Masruriyah et al., 2019) to predict stroke disease generated 95.15% accuracy using Artificial Neural network Algorithm although he admitted that this model has disadvantage which is as a result of inadequacy of medical data about the history of patient illness and the type of drug taken for one month that can be used in discovering adequate treatment for the target patient.

**2.5 Summary**

Stroke is a heterogenous illness which occur as a result of various risk factors and it is classified into modifiable (e.g., hypertension, diet, smoking etc.) and non-modifiable factors (e.g sex, age ethnicity etc.), although both haemorrhagic and ischemic stroke have slightly different risk factors, but their risk factors are interwoven and almost the same. The systematic review of predictive analysis in healthcare using different algorithms have several important findings. Firstly, series of algorithms have been used in the healthcare sector to make different prediction, the various algorithms however have different level of accuracy which has been critically analysed in the review above. In summary, the overall algorithm discussed in this review include Artificial Neural network, Back Propagation, Bayesian Classifier, Decision Tree, Recurrent Neural Network, K-Nearest Neighbour and Long Short-Term Memory Recurrent Neural Network. It was observed that decision tree is one of the best predictive algorithms due to its level of accuracy and multiple record of high performance is different used cases. In several medical applications also, Artificial neural networks tested and confirmed as a useful tool to make predictions on several occasions

The research carried out by Shanthi et al. (2008) is similarly to this work, but his research only considered stroke disease caused by thrombus (blood clot). In addition, one algorithm (Artificial Neural Network) was used without considering other algorithms to compare which will significantly perform better. This work however seeks to use multiple algorithms in the prediction of stroke disease so that the algorithm with the highest level of accuracy would be selected as the best predictive model.

**Chapter Three**

**Research Methodology**

**3.0 Introduction**

This section describes the research philosophy, approach, methodology and strategy applied to answer the stated research questions in other to achieve the research objectives. The procedure for data collection, data pre-processing as well as other additional techniques is be explained.

Since the deliverable of this project is generating predictive models which appropriately predict the stroke disease, theoretical overview of predictive model and the proposed strategy for data analysis is discussed here in this chapter.

**3.1 Research Philosophy, Approach, Strategy and Methodology.**

Pragmatic research philosophy is applied in this project to address the research questions and to build the various models to fulfil the research objectives because it is a problem focused research. The objective of this project focuses on practical and consequential and not only the theoretical aspect of the problem i.e it combines the idea of positivism and interpretivism. Positivism approach would have as well be considered for this research, but positivism would only examine this problem from the objective point of view, and this will not be effective to achieve the research objective. Pragmatic approach enables a mixed approach i.e it allows the use of multiple methods and techniques hence it will be suitable for this research.

This project would employ **abductive approach** because it seeks to develop different models in which one out of the models will be considered as the best stroke predictive model from the overall level of performance.

The review of literature of similar research work will be done to have a background idea which is more of a **deductive approach.** Patient historical dataset will be used to build the models which is an open-source dataset, henceforth a **comparative** type of case study strategy would be adopted because the project seeks to compare different models to examine their level of performance to therefore seek which one is better.

**3.2 Tools and Techniques, Data collection and Analysis**

**3.2.1 Data mining**

Before the machine learning aspect of this research will be executed, the project seeks to reveal patterns in the dataset for insightful ideas about the behaviour of the different attributes in our dataset. The dataset consists of both quantitative and qualitative attributes therefore numbers of statistical graphs such as bar charts, histogram, scatter plots, box plots etc. which is be presented using tableau and python software.

The data mining aspect of this research will be executed using python programming language and tableau although there are numerous data analytics software which can be used as well but due to familiarity, number of available toolkits framework, and APIs, python is considered for this task.

Library such as scikit learn, NumPy, pandas, matplotlib and SciPy is extensively used for this section of the project because they are free and easily accessible tools for data analysis in python.

**3.2.2** **Machine Learning**

In a general perspective, there are two types of techniques used by machine learning which are supervised learning and unsupervised learning. Supervised machine learning trains the model with a proportion of the dataset and test model with the rest of the dataset so it can predict future output for a response to a new dataset. Supervised is divided into two classification and regression. Classification techniques is useful when making a discrete prediction while regression focusses on making continuous responses.

In the same vein, unsupervised machine learning exposes hidden patterns in the dataset and this type of machine learning majorly focus on clustering.

***Diagram

Description automatically generated***

***Fig 3.1****: Machine learning techniques for supervised and unsupervised learning (by Cleve Moler, 2022 )*

Classification techniques of supervised machine learning will be adopted to fulfil the research objective in this research. There are numbers of classification algorithms but three algorithms which are Random Forest, Decision Tree, and Logistic Regression where the algorithm which has the highest level of performance will be considered as the final choice of model. In this section, a proportion of the dataset trains the model/algorithm, and the other part will **test** the model. The beneficiary of this project as stated in the introduction section of this work would majorly be medical practitioners and individuals who have access to their medical records, they would detect early if a patient is at risk of stroke disease base on some risk factors so it can be controlled at the early stage, therefore there is a needs to evaluate if deliverable meets the needs of the beneficiaries.

Python programming language is one of the most preferred languages for machine learning although some developers still prefer using R programming, Java, and other language base on their level of familiarity. This section of the project will extensively adopt the use of python base on familiarity, ease of use as well as availability.

**3.3 Theoretical overview of predictive model**

Predictive analysis may be defined as analytic process which is used in forecasting an event which may happen in the future. This phenomenon uses techniques in statistics, machine learning, data mining and artificial intelligence to make accurate predictions of future events. (Elkan C, 2012).

In most cases, prediction model is usually applied in discriminating between two groups, Fisher’s discriminant function is a good demonstrator of such model (Fishers, 1936) which is also known as Linear Discriminant Analysis (LDA). Linear Discriminant Analysis refers to an approach to the method of discriminating between two groups. This process is implored to identify the variable that discriminate between two or more naturally occurring groups.

This method helps in making classification which involve decision as to whether an individual belong to a particular group or not based on his/her record. Take for instance, a patient may be categorized as being at risk of stroke or not based on his medical records.

In this study, three techniques which are logistics regression, decision tree and Random Forest will be used for stroke prediction. It is expected that the outcome of this project will be binary i.e., the outcome will be whether the patient will have stroke disease or not.

In general, the objective of logistic regression is to establish the relationship between the dependent variable and the independent variable. The dependent variable in this research are the predictive factors which are the various risk factors discussed in the review of literature. This model is suitable because the aim of this project is to detect whether a patient is likely to have stroke disease or not.

Decision tree may be explained as the categorization of complex data into simplified subset in other to identify the presence or absence of other components. It is useful in identifying both homogeneous and heterogenous data. Henceforth this will be useful in this piece of research as it will be helpful in breaking/splitting the patients into groups i.e., those with and without stroke disease.

**3.3.1 Logistic Regression**

Linear regression is the fundamental of logistic regression, linear regression may be described as a tool which predicts relationship between dependent variable say Y and one or more independent variables say Xi . and it enables us to see the model as well as the importance of the relationship. it is represented by the equation below:

**+** ε …………1

Where are constants values, are the independent variables and ε is the component error.

However logistic regression gives an estimate of the probability of an event occurring or not. It makes a prediction Y (dependent variable) from set of known independent variables which is not necessarily a numerical value i.e., it may be a categorical variable which and it gives room for adjustment of multiple predictors. This is one of the attributes which makes logistic regression useful for observational data analysis most especially when the potential bias resulting from the differences in the group being compared needs to be adjusted r It is represented by the equation:

…………… (eq 2)

………………(eq3)

P is the measure of probability that Y is independent on the dependent variable (X) and , are the coefficients of X.

Logistic regression on like linear regression do not only cover the linearity relationship between the dependent variable but also the non-linearity relationship which may either be categorical or continuous. Its final result is always in binary unlike linear regression which is given in a continuous form. Logistic regression does not assume linearity, its graph showed it has a sigmoid shape which accept linearity, near linearity and non-linearity event.

Chart, line chart

Description automatically generated

**Fig 3.2: Logistic regression functions**

**3.3.2 Decision tree**

Decision tree is one of the simplest form of multiple variable analysis, it is a supervised machine learning algorithm known for its unique way of splitting data into segment or branches.

([Barry D](https://learning.oreilly.com/search/?query=author%3A%22Barry%20de%20Ville%22&sort=relevance&highlight=true) and [Padraic N](https://learning.oreilly.com/search/?query=author%3A%22Padraic%20Neville%22&sort=relevance&highlight=true), 2013). An upturn decision tree is formed by the branches with connection from the top of the tree. In a decision tree diagram, the various field’s name i.e., the variable to be analysed are usually displayed together with the distribution of values which are in the object of analysis. The figure below demonstrated a sample of tree diagram.

**Diagram, schematic

Description automatically generated**

**Fig 3.3:** *Illustration of decision tree (Decision tree algorithm by* [*Nagesh Singh Chauhan*](https://www.kdnuggets.com/author/nagesh-chauhan)*, KDnuggets , 2022)*

Decision tree just like logistic regression accepts both continuous and categorical variables, the diagram above shows that all the field and field value is captured by the algorithm. Fields and column are used to create rules assigned to the nodes in the branch are observations from the dataset. Furthermore, decision tree classifies by sorting the dataset from the root nodes to the terminal notes. The nodes at the end are called leaves or terminal nodes. All nodes are mutually exclusive that is no two nodes can have the same observation from the parent data, the moment the decision rule have been generated, this rule can be applied to predict new nodes value for an entirely new data.

**3.3.3 Random Forest**

Random forest is generated from the combination of different tree classifiers, a random vector which is sampled independently from the input vector formed each tree classifier where classification of an input vector is as a result of vote unit cast by each tree to the most popular class. Random forest classifier uses sets of CARTs for classification and prediction. In the opinion of ([Breiman, 2001](https://www.sciencedirect.com/science/article/pii/S0924271616000265#b0050)), he emphasized that random forest is a classifier which combines tree-structured classifiers say {h(x, Ꝋk ), k = 1,...}, { Ꝋk } represents series of independent identical distributed random vectors and the most popular class receive a unit vote from each tree at input x ([Breiman, 2001](https://www.sciencedirect.com/science/article/pii/S0924271616000265#b0050)).

Random forest trees are created by using a bagging approach to draw subset of the training sample. It implies that similar sample can be selected in multiple times however they may be some sample which will not be selected all. Figure 3.4 shows the diagrammatic illustration of this process.

Diagram

Description automatically generated

***Fig 3.4 Training and***[***classification***](https://www.sciencedirect.com/topics/computer-science/classification)***phases of***[***Random Forest classifier***](https://www.sciencedirect.com/topics/computer-science/random-forest-classifier) ***(Belgiu, M., & Drăguţ, L. (2016))***

The in-bag samples i.e., the two-third of the sample are used to train the tree while the remaining portion which is the one-third of the also known as out-of-the-bag sample are used in internal cross-validation techniques to examine the performance of the random forest model. ([Breiman, 2001](https://www.sciencedirect.com/science/article/pii/S0924271616000265#b0050)).

**3.4 Data Description**

The dataset used for this research is gotten from Kaggle. It contains health records which were acquired from different hospitals of Bangladesh by a group of researchers for academic purpose.

Graphical user interface, table, Excel

Description automatically generated

***Fig 3.5: Screenshot of Stroke dataset***

The dataset contains health record of 5110 of patients with ten attributes which will be used extensively for analysis and prediction in this project. The attributes are age, gender, hypertension, work type, heart disease, average glucose level, body mass index (BMI), marital status, smoking status and stroke i.e., whether the patient previously had stroke or not.

Figure 3.6 describe the variables in the dataset. Out of the total population, the chart showed that over three thousand (3000) of the observations are female while the remaining population are females, none of the respondents were identified as other sex.

***Chart, bar chart

Description automatically generated***

***Fig 3.6 Distribution of stroke dataset by gender.***

From figure 3.7 below Over 3000 patients confirmed they have ever married in the record from the stroke dataset while others are not married. It is obvious we have higher records of married patients in the dataset.

***Chart, bar chart

Description automatically generated*Chart, pie chart

Description automatically generated**

***Figure 3.7: Distribution of stroke dataset by Marital status***

It also revealed that 79.9% of the population in the dataset confirmed they have at least married but 20.1% have never been married.

Figure 3.8 shows the smoking status the observation in the stoke dataset, the number non-smokers are higher in the dataset.

***Chart, bar chart

Description automatically generated***

***Figure 3.8: Distribution by smoking status***

In figure 3.9, it was observed that few record on the dataset have previously had stroke incidence.

**Chart, bar chart

Description automatically generated**

**Figure 3.9: Distribution of stroke dataset by previous incidence of stroke disease.**

figure 3.10 reveal that most of the patients whose record were in the dataset run a personal business.

Chart, bar chart, histogram

Description automatically generated

***Figure 3.10: Distribution by Work Type***

Figure 3.11 below shows that 50.8% of the population resides in Urban area while the remaining 49.2% lives in Rural area.

**Chart, pie chart

Description automatically generated**

**Fig 3.11 Distribution by Residence Type.**

**3.5 Data pre-processing**

For effective result and data quality, it is necessary to perform series of data quality checks on the dataset. Firstly, missing values in the dataset were replaced with the mean of other values and invalid character values were removed. The smoking status attribute have missing values were replaced with the group by age attribute. Proper checks were made to identify duplicate data. The dataset is being normalize and label encoding to a categorical data.

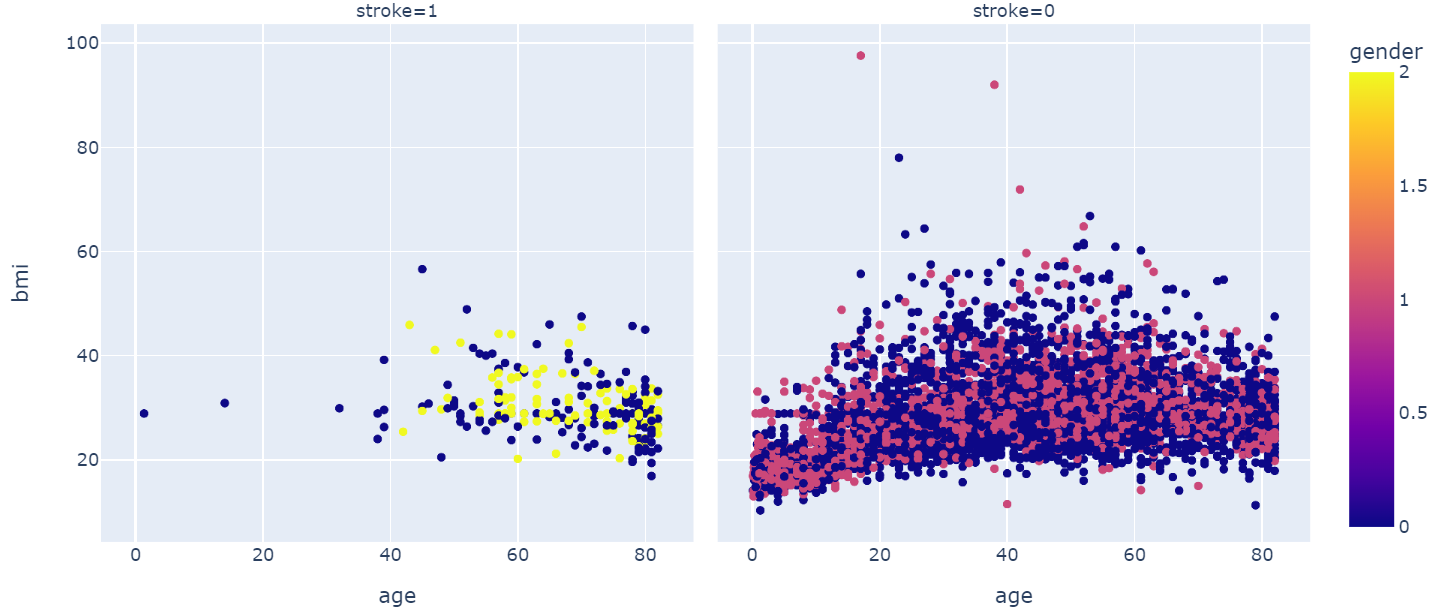


Fig 3.12: Scatter diagram of stroke disease

Obviously, the proportion of stroke incidence class is under-represented from the data description. There is an imbalance in the dataset as 95% of the dataset present no stroke, henceforth it would be difficult for machine learning algorithm to manage this kind of data. It is therefore important to oversample the minority class i.e., the stroke class. To achieve this, the examples in the minority class will be duplicated. Although, no new information is added to the model but new example will be synthesised from the existing examples. The two most popular oversampling techniques are Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic (ADASYN).

It is one of the most popular and easiest oversampling algorithms which was proposed by Chawla et al. (2002), it is the most widely used approach in synthesising new examples from the minority class. SMOTE uses K nearest neighbours’ algorithm to add synthetic minority class minority observation to the minority class observation. However, one of the major challenges of SMOTE approach is that it formulates synthetic observation arbitrarily using all the minority class observation. Figure 3.12 shows most of the “no stroke” points are not close to the edge line so the work used SMOTE instead of ADASYN for the oversampling of the minority class. The total observation in the dataset before SMOTE was 5110 and 9722 after SMOTE, now number of stroke incidence in the dataset is not the same with those who have never had stroke disease.

In addition, thorough outlier check was performed on the dataset to detect outliers in order to avoid over fitting. This process is called normalization, it detected and help in bringing the outlier to a value where all the mean is zero in the training set with the help of z-score. Z-score measure the distance between the mean of a point and the standard deviation. It’s the number of standard deviations away from the mean point. When the mean of a point gives a value above three, then it is different from the original point and as such the point may be an outlier.

Chart, box and whisker chart

Description automatically generated

**Fig 3.13: Box plot and Whisker**

Figure 3.13 shows the box plot for the numerical attributes in the dataset, the chart reveals observation which are numerically apart from the other dataset. This is known as check for outliers, the point located outside the whisker of the boxplot are the potential outliers in the dataset. BMI observations have a quite high observation and the average glucose level as well, the age attribute seems not to have outliers.

Furthermore, the dataset was divided into three groups i.e., train, test and validation 70% of the data was used to train the model, 15% was used for testing and the remaining 15% was used for validation.

Appropriate feature selection was carried out to select the number of input variable in the model to reduce the computational cost of modelling and to improve the overall performance of the model. This process trains a machine learning model with for each of the features and determine the model performance each feature.

Chart, bar chart

Description automatically generated

**Fig 3.13: Chart showing feature importance.**

**3.6 Research Ethics, Risk, and Issues**

**3.6.1 Ethics and legal issues**

The dataset for this project is from an open source as well as the programming software to be used and it is accessible to everyone, the researchers who uploaded it have taken consideration of some ethical issue in relation to the use of this dataset. However, a proper scrutiny has been made to examine the ethical issue with respect to the ethics checklist and there was no ethical issue. Furthermore, In the conduct of healthcare research, there are core ethical principles which must strictly be adhere to according to Helsinki Declaration (1964).

* Beneficence & Non-Malfeasance i.e., the research should prevent risk, harm, and hazard. The data used for this research will not in any form harm or cause either emotional, mental distress to the audience or public.
* Integrity: The outcome of this research will help in the early detection of stroke disease for proper control of the disease at the early stage, henceforth the outcome will contribute to knowledge which will benefit the public.
* Informed Consent: at the point of data collection, the group of researchers who collected the data ensured the patients are informed about the purpose of the data collection before it was conducted.
* Confidentiality/Anonymity: Adequate precautions have been taken such that details collected in this data have no link to the individual involved.

Find the complete ethics form in the appendix section.

**Chapter Four**

**Data Analysis**

**4.0 Introduction**

In this chapter, detailed explanation about the demographic factors is discussed with the aid of charts and diagram. Furthermore, data analysis of the sampled dataset which is basically in two perspectives. Firstly, exploratory analysis about the dataset is carried out to gain deep understanding of the patterns in the used datasets. In the other hand, we conducted the predictions on the dataset using the proposed algorithm i.e., logistic regression, decision tree and Random Forest.

This chapter also give a breakdown of the accuracies, precisions, recall, confusion matrix, ROC etc. of each of the algorithms in a tabular form and other relevant statistical charts which would further generate insights on this topic of discussion.

**4.1 Exploratory Data Analysis of Demographic factors and Attributes in the Dataset**

Figure 4.1 below shows that the mean age in the dataset is 43 years, however the maximum age is 82 years. About 25% of the patient’s age are below 25 years which is also similar to the % population of age 60 and above.

The glucose level has a mean value of 106 although, the recommended limit according to experts is 72-99mg/dL when fasting and 140mg/dL two hours after eating. The standard deviation shows that quite a number of the population deviates from this limit. Patients with BMI <33 are effectively represented but there is no clarity in the representation of people with BMI> 33.

Table

Description automatically generated

**Fig 4.1: Exploratory analysis of stroke dataset**

Figure 4.1 showed that patients diagnosed with heart disease in the dataset are 9.7% of the population and 8.4% suffered from heart disease. it further shows that only 4.87% of the entire population in the dataset have stroke disease, however the remaining 95.13% of the population have never had stroke disease. It would be quite difficult to run a machine learning algorithm directly on this dataset because these classes are under-represented compared to other binary counterparts.

Chart, pie chart

Description automatically generated

**Fig 4.2: Distribution by Age Group**

In addition to the discussion above, figure 4.2 also reveal the percentage representation of the age groups in the dataset. It is clear that from age 60 and above have the highest population in the dataset which is 25.5% of the population. Age 40-50 is just 14.5% of the entire population which is the smallest of all the age range.

**4.1.1 Chances of Getting Stroke Disease using the Gender Attribute**

This section examines the chances of occurrence of stroke disease using the gender attributes in the dataset. The chart generate in this section would display the attribute which would likely have higher level stroke incidence based on the information in the dataset.

Chart, bar chart

Description automatically generated

**Fig 4.3 Gender difference and Chance of Stroke Disease**

Figure 4.3 confirms that the male gender is most likely to be at risk of stroke disease. Although there is not clear justification because the charts show both are at risk with slight difference.

**Chart, bar chart

Description automatically generated**

**Fig 4.4 Chances of stroke disease in hypertensive patients**

Figure 4.4 shows the chances of incidence of stroke disease in patients with previous record of hypertension. The chart showed that hypertensive patients are more at risk of stroke disease.

Chart, bar chart

Description automatically generated

**Fig 4.5. Heart disease and incidence of stroke**

Figure 4.5 reveal the incidence of stroke in patients diagnosed previously with heart disease. It confirms that heart disease is one of the leading causes of stroke disease as argued different authors in the literature review.

Chart, bar chart

Description automatically generated

**Fig 4.6. Marriage and incidence of stroke disease**

Figure 4.6 shows that the chances of stroke disease is higher for people who have ever married, this difference may be due to the population of young people in the dataset whose chances of getting stroke disease is verry low. So, it will be ideal to plot the graph for different age groups in the dataset.

Chart, bar chart

Description automatically generated

**Fig 4.7. Chances of stroke disease for different age groups.**

Further analysis of the chances of stroke disease among the different age group showed that married people from age 0-36 have more chances of stroke disease most especially at age 30-40 years. However, there is high chances of stroke incidence among individuals who are not married yet except for 50-60 years and 80-90 years where married people have higher chances of the incidence of stroke.

Chart, bar chart

Description automatically generated

**Fig 4.8. Work type and incidence of stroke disease**

Figure 4.8 above shows that self-employed individual has the highest chances of stroke incidence. Those with working with private or government firms have almost equal chances of stroke incidence while children and those who have never gotten a job have an exceptionally low chances of stroke incidence.

Figure 4.9 below show that residence living in urban area have high chance of stroke incidence compared to rural area. Although the difference is not so high.

Chart, bar chart

Description automatically generated

**Fig 4.9 Chances of stroke disease and Residence type.**

Figure 4.10 reveals that individual who are former smokers are at the risk of stroke disease followed by the active smokers. Other categories such as never smoked and those with unknow status also have the tendencies of witnessing the incidence of stroke disease.

**Chart, bar chart

Description automatically generated**

**Fig 4.10. Chances of stroke disease with smoking status.**

Figure 4.11 below shows a progression in the incidence of stroke disease as aging individual age increases. Age 80-90 has the highest chances of stroke disease incidence followed by 70-80 age group. Age 0-10 and 10-20 have similar chances of stroke disease according to the charts. In conclusion the chart confirmed that incidence of stroke disease.

**Chart, bar chart

Description automatically generated**

**4.11. Age and chances of stroke disease.**

Figure 4.12 reveals the relationship between the various body mass index range and the chances of stroke incidence within this groups. It shows that BMI between 45-50 have the highest chances of stroke disease followed by 25-30. BMI within 15-20 has the lowest chances of stroke incidence. However other groups have almost similar chances of stroke incidence.

Chart, bar chart

Description automatically generated

**Fig 4.12 BMI and Stroke disease.**

Figure 4.13 shows that 30-60, 60-90 and 90-120 average glucose level have similar chances of stroke incidence, however the chances increase at 150-180 and 180-210. Average glucose level between 240-270 have the highest chances of stroke disease,

**Chart, bar chart

Description automatically generated**

**Fig 4.13: Chances of stroke disease and glucose level.**

**4.2 Machine learning Algorithm**

The dataset used for this research as discussed in chapter three was divided into training set which was 70% of the dataset, testing set 15% of the dataset while the remaining 15% was used for validation. The performances of these models are measured in relation to its effectiveness in identifying stroke patients.

False positive group consist of patients who have been wrongly classified by the model as stroke patient and patients who have been incorrectly classified as non-stroke patients constitute the false negative group. In the other hand, true positive group consist of patients who have been correctly identified as stroke patients while true negative group are the patients who have been correctly classified as non-stroke patients. The four groups explained above are usually displayed by a confusion matrix to reveal the classification accuracy of the different algorithms with the expression of their sizes compared to the overall data.

Receiver Operating Characteristics (ROC) is one of the most used measures which displays the performance of a binary classifier system in graphical form. It is a graphical curve which is generated by plotting True Positive Rate (TPR) and False Positive Rate (FPR) at different threshold. ROC demonstrate the level to which the model can distinguish between two classes. Recall or sensitivity which is also known as true positive rate (TPR) measure the ratio of the true positive values and the overall observation in the dataset. False Positive Rate (FPR) which is also known as fall-out is the measure of ratio of the True Positive and the total observation in the dataset. The figures below are various confusion matrix which shows the results and summary of actual dependent variable and the predicted ones for the three algorithms used. the point in the Receiver Operating Characteristics (ROC) is represented by each prediction result of a confusion matrix. The section will henceforth discuss the result of the research using both confusion matrix and the ROC curve. In the confusion matrix in the figures below, class “true” indicate stroke patients while “false” represent the non-stroke patients.

**Logistic Regression**

Figure 4.14 shows that 47.43% (692) of the dataset was correctly identified as non-stroke patients by logistic regression classifier and 44% (642) are rightly identified as stroke patients. In the other hand, the table also shows that only 2.6% (38) constitute the false positive group while 5.96% (87) are identified as the false negative group.

A picture containing chart

Description automatically generated

**Fig 4.14 Confusion matrix for logistic regression algorithm.**

**Decision Tree**

In figure 4.15 below, 44.83% (654) are identified as true negative, 44% (642) correctly represent the true positive group by decision tree classifier. However, the false positive and false negative are 5.12% (76) and 5.96% respectively.

A picture containing chart

Description automatically generated

**Fig 4.15 Confusion matrix for Decision Tree algorithm.**

**Random forest**

Random Forest classifier correctly identify 46.61% (680) as non-stroke patients and 47.50% (693) were truly identified as stroke patients. In the other hand, 3.43% (50) were incorrectly identified as non-stroke patients while 2.47% (36) were wrongly identified as stroke patients by Random Forest classifier.

**Chart

Description automatically generated with medium confidence**

**Fig 4.16 Confusion matrix for Random Forest Classifier.**

**4.3 Model Accuracy**

figure 4.17 shows that random forest performs significantly better among the other three algorithms, followed by logistic regression. Decision tree has the lowest level of accuracy.

Chart, box and whisker chart

Description automatically generated

**Fig 4.17 Accuracy of model on Validation set**

**Chapter Five**

**Discussion of Result findings and Summary.**

**5.0 Introduction**

In order to predict the incidence of stroke disease, chapter three and four have detailed analysis of the used dataset following the proposed strategy in chapter one using the three-machine learning algorithm which are logistic regression, decision tree and random forest. This section hereby discusses the result findings, show that the research questions have been answered. Chapter two section of this project discussed both the theoretical and practical aspect explicitly in the review of literature about existing research work on machine learning in the healthcare as well as the application of machine learning algorithms in the prediction of stroke disease and the various accuracies, recall and precision generated. Henceforth, that section of the project has answered the first research objective.

In addition, the review of literature has revealed the various medical attributes i.e., both modifiable and non-modifiable risk factors of stroke disease which were considered as the medical attributes of stroke disease. Figure 3.13 in chapter three also give an overview of the feature performance in the overall model as well as the feature performance.

**5.1 Result findings**

**5.1.1 Result of Variable correlation.**

The result in figure 5.1 It visualizes the relationship between the stroke features and other features in the dataset. The table shows that age and ever married are slightly correlated with correlation value of 0.68 and work type and unknown smoking type have slightly high correlation value of 0.51. However, Working\_type have a negative corelation with age (-0.63), ever married (0.54) and BMI (0.44). It is clear that from the correlation matrix below that none of the features has profoundly effect on the stroke feature.

**Chart, treemap chart

Description automatically generated**

**Fig 5.1 Correlation matrix for stroke Dataset**

The ROC curves in figure 4.18 demonstrates smooth plot for the three algorithms, a good TPR/FPR ratio is demonstrated with the sharp rise in the ROC. The line of discrimination that is the line above the straight line shows that the plot has a good predictive outcome. Figure 4.18 shows that the Random Forest have the highest predictive accuracy with value nearly close to 1. This will further be established in the table below.

Chart, line chart

Description automatically generated

**Fig 5.2 ROC curve for the three algorithms**

Additionally, figure 5.2 gives the respective Area Under Curve value of all the three algorithms, AUC represents the measure of classifier’s ability to discriminate between classes. When the AUC equals 1, it implies that the classifier would be can effectively distinguish all the negative and positive class point. On the other hand, if the AUC value equals zero, the classifier would expect both the positive and negative class point as positive. Generally, if 0.5<AUC< 1 we say that the classifier has a high chance of discriminate between negative value class and positive value class because the occurrence of true positive and true negative will be higher than false positive and false negative. However, the classifier will not be able to differentiate between positive and negative class when the Area Under Curve (AUC) is 0.5. Figure 5.1 above shows that Random Forest has the highest AUC value (0.979) followed by logistic regression with AUC value of (0.974) while decision tree has the lowest AUC value of (0.888). The three classifiers with their respective AUC value have high probability to differentiate between negative and positive class point.

**Choice of best Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Classification Report** | | | |
|  | Random Forest | Logistic Regression | Decision tree |
| Precision | 93.27% | 94.41% | 89.41% |
| Recall | 95.06% | 88.06% | 88.07% |
| F1 | 94.16% | 91.12% | 88.73% |
| Accuracy | 94.11% | 91.43% | 88.83% |

**Fig 5.3 Classification Report of the three Algorithms**

Figure 5.3 confirmed that random forest has the highest accuracy with 94.11%. logistic regression has 91.43% while decision tree has the lowest accuracy level with 88.83%. However, regression has a better precision strength when compared to Random Forest and Decision Tree but random forest perform significantly better than both Logistic regression and Decision tree in in its recall and sensitivity value. Random forest is however chosen as the best algorithm of the three used machine learning algorithms.

**5.2 Summary**

This proposed research work used three machine learning algorithms to predict the incidence of stroke disease using ten features considered as modifiable and non-modifiable risk factors. In the used dataset, two attributes which are age and gender were considered as the non-modifiable risk factors while the other factors namely, hypertension, heart disease, marital status, smoking status, body mass index, residence type and average glucose level were the modifiable risk factors. The dataset used for this project is imbalance as 95% of the observation are classified as “no stroke” patients, SMOTE was applied to correct this error in other to get effective prediction and the maximum level of accuracy. The original dataset has 5110 observations which later rise to 9722 when SMOTE was applied to oversample the minority class. In addition, the data was divided into three groups namely the test set, train sets and validation sets, appropriate feature selection process was made to enhance the overall performance of the model and to reduce the computational cost. Evaluation of the algorithm performance showed that random forest performs significantly better with 94.11% level of accuracy than the two other algorithms i.e decision tree and logistic regression which have 91.43% and 88.83% respectively. Random forest is therefore considered as the best predictive algorithm for the incidence of stroke

**Chapter six**

**Conclusion, Recommendation and Limitation of the study**

**6.1 Introduction**

In order to predict the incidence disease, chapter four and five discussed the data analysis section, result findings and summary of the work. This chapter however discussed the conclusion, the research limitation as well as final recommendation for future research.

**6.2 Conclusion**

This project has demonstrated prediction of stroke disease using both exploratory data analysis and predictive algorithms, the algorithms have been tested and validated using the test and validation test. Based on the research problem stated in chapter one, it can be argued that this research work has met the aims and objectives whereby individuals have adequate knowledge about stroke incidence, the causes i.e., the different medical attributes and how it can be prevented. Individuals are now informed about the factors which causes stroke disease, effective management of this factors will prevent the incidence of stroke disease.

The relationship between stroke disease and other disease have also been established in this research work. In future research work, more machine learning algorithms should be explored and combined with deep learning-based imaging such as magnetic resonance imaging, computerized tomography scan to enhance the performance of stroke prediction model.

**6.3 Limitation of the study**

This research work has some limitation just like every other research. Firstly, the feature of the used dataset i.e., the attributes is just ten (10), the non-modifiable features out of the ten attributes is only two which are age and gender. There are a lot of non-modifiable risk factors that can trigger the incidence of stroke disease which were not in the dataset. In future work, researcher should work more in collecting a more robust patients features to enhance higher level of accuracy. In addition, the dataset has 5110 observations only which was collected from random specialist hospital in Bangladesh, this isn’t enough to make generalization. Furthermore, only 5% of the patients in this dataset were stroke patients, future data collection procedures should correct this issue to generate a more accurate prediction.

This process can as well be improved in future research work, the data collection procedure should be extended to other countries to examine the incidence of stroke disease in different countries in order to make more insightful decision.

In further research, other machine learning algorithms should be implemented on a more robust dataset to see if a better level of accuracy could be generated.

**6.4 Recommendation from Research**

The outcome of this research work suggests that:

* Stroke disease have several risk factors which are divided into modifiable and non-modifiable risk factors.
* Non-modifiable stroke risk factors cannot be changed but may be managed if detected at the early stage
* The incidence of stroke disease increases with aging i.e., adults are mostly at risk of developing stroke disease.
* Male gender has the highest stroke incidence but female experience higher incidence at older age.
* Body mass index, hypertension and heart disease are very strong modifiable risk factor of stroke disease.
* Marital and smoking status contributes to the incidence of stroke disease,
* Random forest among the other two algorithms i.e., logistic regression and decision tree has the highest predictive accuracy.

**REFERENCES**

Abdull, M. A. S. (2011). Data mining techniques and breast cancer prediction: A case study of Libya. Sheffield Hallam University (United Kingdom).

Adams Jr, H. P., Bendixen, B. H., Kappelle, L. J., Biller, J., Love, B. B., Gordon, D. L., & Marsh 3rd, E. E. (1993). Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment.*Stroke, 24*(1), 35-41.

Adam SY, Yousif A, Bashir MB (2016) Classification of ischemic stroke using machine learning algorithms. Int J Comput Appl 149(10):26–31

Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., Norouzi, R., & Toghianfar, N. (2013). Prediction and control of stroke by data mining. International Journal of Preventive Medicine, 4(Suppl 2), S245.

Appelros, P., Nydevik, I., & Viitanen, M. (2003). Poor outcome after first-ever stroke: predictors for death, dependency, and recurrent stroke within the first year. Stroke, 34(1), 122–126. <https://doi.org/10.1161/01.str.0000047852.05842.3c>

Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, 24-31.

Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. Statistics in medicine, 24(11), 1713-1723.

Blankstein, R., Ward, R. P., Arnsdorf, M., Jones, B., Lou, Y. B., & Pine, M. (2005). Female gender is an independent predictor of operative mortality after coronary artery bypass graft surgery: contemporary analysis of 31 Midwestern hospitals. Circulation, 112(9\_supplement), I-323.

Bots, S. H., Peters, S., & Woodward, M. (2017). Sex differences in coronary heart disease and stroke mortality: a global assessment of the effect of ageing between 1980 and 2010. BMJ global health, 2(2), e000298. <https://doi.org/10.1136/bmjgh-2017-000298>

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Bulygin, K. V., Beeraka, N. M., Saitgareeva, A. R., Nikolenko, V. N., Gareev, I., Beylerli, O., ... & Aliev, G. (2020). Can miRNAs be considered as diagnostic and therapeutic molecules in ischemic stroke pathogenesis?—Current status. International Journal of Molecular Sciences, 21(18), 6728.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. Nature methods, 15(1), 5.

Centres for Disease Control and Prevention. (2009). Prevalence and most common causes of disability among adults--United States, 2005. MMWR: Morbidity and Mortality Weekly Report, 58(16), 421-426.

Chantamit-o-pas, P., & Goyal, M. (2018, July). Long short-term memory recurrent neural network for stroke prediction. In *International Conference on Machine Learning and Data Mining in Pattern Recognition* (pp. 312-323). Springer, Cham.

Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. International journal of environmental research and public health, 16(11), 1876.

Chin, C. L., Lin, B. J., Wu, G. R., Weng, T. C., Yang, C. S., Su, R. C., & Pan, Y. J. (2017, November). An automated early ischemic stroke detection system using CNN deep learning algorithm. In 2017 IEEE 8th International conference on awareness science and technology (iCAST) (pp. 368-372). IEEE.

Chobanian, A. V., Bakris, G. L., Black, H. R., Cushman, W. C., Green, L. A., Izzo Jr, J. L., ... & National High Blood Pressure Education Program Coordinating Committee. (2003). Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. hypertension, 42(6), 1206-1252.

Choudhury, M. J. H., Chowdhury, M. T. I., Nayeem, A., & Jahan, W. A. (2015). Modifiable and non-modifiable risk factors of stroke: A review update. Journal of National Institute of Neurosciences Bangladesh, 1(1), 22-26.

Cleve Moler (2022). How machine learning works. MathWorks. https://www.mathworks.com/discovery/machine-learning.html

Colak C, Karaman E, Turtay MG (2015) Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed 119(3):181–185

Donnan, G. A., Fisher, M., Macleod, M., & Davis, S. M. (2008). Stroke. The Lancet (British Edition), 371(9624), 1612-1623. 10.1016/S0140-6736(08)60694-7

Elkan, C. (2013). Predictive analytics and data mining (Vol. 600). San Diego: University of California.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1, 179-188.

Fedesoriano, (2020). Stroke Prediction Dataset, version 1. Retrieved June 1, 2022 from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/download?datasetVersionNumber=1

Ezzati, M., Rodgers, A., Vander Hoorn, S., Lopez, A.D., Lin, R.B., Murray, C.J.L. and Group Comparative Risk Assessment Collaborating, 2004. Distribution of major health risks: findings from the Global Burden of Disease study. PLoS medicine, 1(1), p.e27.

Falkstedt, D., Hemmingsson, T., Rasmussen, F., & Lundberg, I. (2007). Body mass index in late adolescence and its association with coronary heart disease and stroke in middle age among Swedish men. International journal of obesity, 31(5), 777-783.

Feigin, V. L., Lawes, C. M., Bennett, D. A., & Anderson, C. S. (2003). Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. The lancet neurology, 2(1), 43-53.

Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. Neural Computing and Applications, 32(3), 817-828.

Hubert, H. B., Feinleib, M., McNamara, P. M., & Castelli, W. P. (1983). Obesity as an independent risk factor for cardiovascular disease: a 26-year follow-up of participants in the Framingham Heart Study. Circulation, 67(5), 968-977.

Kansadub, T., Thammaboosadee, S., Kiattisin, S., & Jalayondeja, C. (2015, November). Stroke risk prediction model based on demographic data. In 2015 8th Biomedical Engineering International Conference (BMEiCON) (pp. 1-3). IEEE.

Kearney, P. M., Whelton, M., Reynolds, K., Muntner, P., Whelton, P. K., & He, J. (2005). Global burden of hypertension: analysis of worldwide data. The lancet, 365(9455), 217-223.

Khosla, A., Cao, Y., Lin, C. C. Y., Chiu, H. K., Hu, J., & Lee, H. (2010, July). An integrated machine learning approach to stroke prediction. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 183-192).

Lumley, T., Kronmal, R. A., Cushman, M., Manolio, T. A., & Goldstein, S. (2002). A stroke prediction score in the elderly: validation and Web-based application. *Journal of clinical epidemiology*, *55*(2), 129-136.

Masruriyah, A. F. N., Djatna, T., Hardhienata, M. K. D., Handayani, H. H., & Wahiddin, D. (2019). Predictive Analytics For Stroke Disease. Paper presented at the 2019 Fourth International Conference on Informatics and Computing (ICIC), 1-4.

M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun, & M. S. Kaiser. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. Paper presented at the - 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 1464-1469. 10.1109/ICECA49313.2020.9297525

Maier O, Schröder C, Forkert ND, Martinetz T, Handels H (2015) Classifiers for ischemic stroke lesion segmentation: a comparison study. PLoS ONE 10(12):e0145118

Mathers CD, Lopez AD, Murray CJL, et al. The burden of disease and mortality by condition: data, methods, and results for 2001. Global Burden of Disease and Risk Factors. New York: Oxford University Press; 2006: 45–240.

Medin, J., Nordlund, A., & Ekberg, K. (2004). Increasing stroke incidence in Sweden between 1989 and 2000 among persons aged 30 to 65 years: evidence from the Swedish Hospital Discharge Register. Stroke, 35(5), 1047-1051.

Monteiro, M., Fonseca, A. C., Freitas, A. T., e Melo, T. P., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(6), 1953-1959.

Ovbiagele, B., & Nguyen-Huynh, M. N. (2011a). Stroke epidemiology: advancing our understanding of disease mechanism and therapy. Neurotherapeutics, 8(3), 319-329.

Ovbiagele, B., & Nguyen-Huynh, M. N. (2011b). Stroke epidemiology: advancing our understanding of disease mechanism and therapy. Neurotherapeutics, 8(3), 319-329.

Qu, S., Zhou, M., Jiao, S., Zhang, Z., Xue, K., Long, J., Zha, F., Chen, Y., Li, J., & Yang, Q. (2022). Optimizing acute stroke outcome prediction models: Comparison of generalized regression neural networks and logistic regressions. PloS One, 17(5), e0267747.

Rajinikanth, V., & Satapathy, S. C. (2018). Segmentation of ischemic stroke lesion in brain MRI based on social group optimization and Fuzzy-Tsallis entropy. Arabian Journal for Science and Engineering, 43(8), 4365-4378.

Rexrode, K. M., Hennekens, C. H., Willett, W. C., Colditz, G. A., Stampfer, M. J., Rich-Edwards, J. W., ... & Manson, J. E. (1997). A prospective study of body mass index, weight change, and risk of stroke in women. Jama, 277(19), 1539-1545.

Rickham p. p. (1964). Human experimentation. code of ethics of the world medical association. declaration of helsinki. British medical journal, 2(5402), 177. <https://doi.org/10.1136/bmj.2.5402.177>

Rimm, E. B., Stampfer, M. J., Giovannucci, E., Ascherio, A., Spiegelman, D., Colditz, G. A., & Willett, W. C. (1995). Body size and fat distribution as predictors of coronary heart disease among middle-aged and older US men. American journal of epidemiology, 141(12), 1117-1127.

Robert B. (2002). What is a stroke. Mayo Clinic. Retrieved April 11 2022, from https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

Rochmah, T. N., Rahmawati, I. T., Dahlui, M., Budiarto, W., & Bilqis, N. (2021). Economic burden of stroke disease: a systematic review. International Journal of Environmental Research and Public Health, 18(14), 7552.

Roger VL, Go AS, Lloyd-Jones DM, et al. Heart disease and stroke statistics—2011 update: a report from the american heart association. Circulation 2011;123:e18–e209.

Roger, V. L., Go, A. S., Lloyd-Jones, D. M., Benjamin, E. J., Berry, J. D., Borden, W. B., Bravata, D. M., Dai, S., Ford, E. S., & Fox, C. S. (2012). Executive summary: heart disease and stroke statistics-2012 update a report from the American Heart Association. Circulation, 125(1), 188-197.

Rosengren, A., Giang, K. W., Lappas, G., Jern, C., Torén, K., & Björck, L. (2013). Twenty-four-year trends in the incidence of ischemic stroke in Sweden from 1987 to 2010. Stroke, 44(9), 2388-2393.

Rothwell PM, Coull AJ, Silver LE, et al. Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). Lancet 2005; 366: 1773–83.

Sacco, R. L., Kasner, S. E., Broderick, J. P., Caplan, L. R., Connors, J. J., Culebras, A., ... & Vinters, H. V. (2013). An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke, 44(7), 2064-2089.

Shanthi, D., Sahoo, G., & Saravanan, N. (2009). Designing an artificial neural network model for the prediction of thrombo-embolic stroke. International Journals of Biometric and Bioinformatics (IJBB), 3(1), 10-18.

Singh, M. S., Choudhary, P., & Thongam, K. (2019, September). A comparative analysis for various stroke prediction techniques. In International Conference on Computer Vision and Image Processing (pp. 98-106). Springer, Singapore.

Sirsat, M. S., Fermé, E., & Câmara, J. (2020). Machine learning for brain stroke: a review. Journal of Stroke and Cerebrovascular Diseases, 29(10), 105162

Süt N, Çelik Y (2012) Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turk J Med Sci 42(5):886–893

Lumley, R. A. Kronmal, M. Cushman, T. A. Manolio, and S. Goldstein (2002). A stroke prediction score in the elderly: Validation and web-based application. Journal of Clinical Epidemiology, 55(2):129–136, February 2002.

Tirschwell, D. L., Smith, N. L., Heckbert, S. R., Lemaitre, R. N., Longstreth, W. T., & Psaty, B. M. (2004). Association of cholesterol with stroke risk varies in stroke subtypes and patient subgroups. Neurology, 63(10), 1868-1875.

Vinereanu, D., Lopes, R. D., Mulder, H., Gersh, B. J., Hanna, M., de Barros E Silva, Pedro GM, Atar, D., Wallentin, L., Granger, C. B., & Alexander, J. H. (2017). Echocardiographic risk factors for stroke and outcomes in patients with atrial fibrillation anticoagulated with apixaban or warfarin. Stroke, 48(12), 3266-3273.

Warlow, C. P. (1998). Epidemiology of stroke. The Lancet, 352, S1-S4.

Xavier, D., Liu, L., Zhang, H., Chin, S. L., & Rao-Melacini, P. (2010). INTERSTROKE investigators: risk factors for ischaemic and intracerebral haemorrhagic stroke in 22 countries (the INTERSTROKE study): a case-control study. Lancet, 376(9735), 112-123.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. ACM Transactions on Intelligent Systems and Technology (TIST), 8(3), 1-19.

Zhang, L. F., Yang, J., Hong, Z., Yuan, G. G., Zhou, B. F., Zhao, L. C., ... & Wu, Y. F. (2003). Proportion of different subtypes of stroke in China. Stroke, 34(9), 2091-2096.

**Appendix**

**Research Proposal**

**1.0 INTRODUCTION AND JUSTIFICATION**

Across the globe, cardiovascular disease is one of the major causes of death. World health Organization reported in 2019 that about 17.9 million lost their lives on a yearly basis to this illness which contribute to 32% of death globally. Rochmah et al. (2021), among other authors established that stroke disease is identified as one principal cause of cardiovascular disease worldwide.

Old age disabilities in United States have been attributed to stroke disease (Ovbiagele et al. ,2011) and it has been considered as the 4th killer disease in the country and one of the major causes of death across the globe (Roger et al., 2012). Furthermore, Centres for Disease Control and Prevention CDC, (2005) argued that under-developed country are at the risk of higher incidence of stroke disease and death rate due to the lack of adequate management and control system. However early detection and awareness of different risk factors and warning signs of this disease would minimize its occurrence among the populace and will be managed properly.

Application of data mining techniques and machine learning algorithm in the prediction and control of stroke disease would therefore relatively reveal new knowledge in the treatment of stroke disease especially to detect patients who are at risk of stroke disease based on some risk factors, help medical practitioners for decision making at the management level as well as generating relevant scientific hypotheses to enhance early detection of this disease so it can be properly managed.

# 2.0 Research question, aims & objectives

## 2.1 Research Aims

This project aim is to gain in-depth understanding of the various risk factors which cause stroke disease, build stroke prediction models using various risk factors from historical dataset in order to help individuals and medical practitioners in clinical decision-making by early detection and proper management of the disease.

## 2.2 Research question

What are the major risk factors for stroke disease and how can we enhance stroke disease predictive model using these risk factors?

## 2.3 Research Objectives

The project’s main objective can be summarized as follows:

* To investigate into details the literature of both the theoretical and practical area based on previous studies of stroke disease.
* To generate insights on different medical attributes that could lead to stroke disease.
* To investigate range of predictive models used by stroke disease researchers as well as the risk factors associated with the disease.
* To use insights generated from objective 1 and 2 to develop predictive models appropriate to estimate risk of having stroke disease based on some modifiable and non-modifiable risk factors.
* Evaluate the performance of the developed models to identify the best out of the models.
* To evaluate if the outcome of the deliverable meets the needs of the beneficiary.

## 2.4 Deliverable

A predictive model which appropriately make prediction whether an individual is at risk of developing stroke disease based on his/her modifiable and non-modifiable risk factor. This model will help medical practitioners the in early detection of the disease so it can be prevented or managed.

# 3.0 Literature review

In this section, a review of literature is established to identify and justify this proposed project. This section focusses mainly the overview of stroke disease, the risk factors as well as a review of different existing predictive algorithms in healthcare.

**3.1 Overview of stroke disease**

Stroke disease is caused when there is an obstruction in the free flow of blood to different region of the brain. When this happens, cells in this area suffers from insufficient oxygen and the required nutrients to function properly which will result to death of the brain cells. Early detection and proper management of stroke would be needed to reduce further harm in the affected region of the brain as well as preventing further complications in the body system (Robert B, 2002).

When the populace is aware of the different risk factors of stroke disease, this would reduce its occurrence in the community. Vinereanu et al. (2017) revealed in their research that little percentage of participants of about 60-70 % were able to identify the risk factors of stroke disease however, Amelia et al. (2017) argued that stroke disease unlike other disease such as myocardial infection which happens due to large vessel atherosclerotic disease which deteriorate the function of coronary arteries. Stroke diseases occur in different ways therefore it is difficult to identify its risk factors.

**3.2 Stroke risk factors**

Risk factors of Stroke disease has been investigated by different researchers in the past, it is majorly categorized into two forms which are ischemic and haemorrhagic strokes. In the research of Adams et al. (2017), they break haemorrhagic stroke into intraparenchymal or subarachnoid and ischemic into etiologic subcategories namely large-artery atherosclerosis, cardio-embolism, small-vessel occlusion, stroke of other determined aetiology, and stroke of undetermined etiologic thought, all these to capture the causes of stroke disease. However, the two types of strokes have similar risk factor with little differences.

In ischemic stroke for instance, the risk factors vary in the etiological categories according to (Tirschwell et al., 2004). One of the major risk factors of haemorrhagic is hypertension although it can result to atherosclerotic disease which will eventually lead to ischemic stroke. Adams et al. (2017) emphasis that hyperlipidaemia which is considered as a coronary risk factor is also a stroke disease risk factor.

**2.2.1 Modifiable and Non-Modifiable Risk Factors of Stroke Disease**

Stroke Risk factors may be classified into modifiable (such as diet and conditions etc.) and non-modifiable risk (such as race, age etc.) factors. It may also be categorized into short-term risk or triggers (such as sepsis, infectious disease, stress), intermediate risk factors (such as hypertension and hyperlipidaemia) and long-term risk factors (e.g race and sex) (Amelia et al., 2017). Similarly, Choudhury et al. (2015) emphasis that stroke is associated with several risk factors which may either be modifiable or irreversible. He further explains that irreversible factors are age, hereditary, race, and gender and modifiable factors are, heart failure, diabetics, heart disease, hyperlipidaemia, hypertension and so on. Both ideas are driven towards the same point, there are some risk factors that can be manage or probably cured while some are inborn which implies no medical remedy can prevent the risk.

Xavier et al. (2010) carried out international research to investigate risk factors of stroke disease. The research comprises of 22 countries with a total of three thousand (3000) patients were used for the study, 2337 out of the overall sample were ischemic stroke patients and 637 were haemorrhagic patients. The table below gives the summary of their findings.

Graphical user interface, text, application, email

Description automatically generatedGraphical user interface, text, application, email

Description automatically generated

*Figure 1- Modifiable and non-modifiable risk factors of stroke by (Xavier et al., 2010).*

Among other factors, age is a major non-modifiable stroke risk factor, the incidence of stroke increases with respect to aging and approximately doubles every 10 years between when the individual approaches age 45-85 (Choudhury et al., 2015). However, Roger, et al. (2012) reported that incident of stroke doubles after 55 years.

**2.3 Review of Predictive Analysis of Stroke Disease Using Different Algorithms**

Over the years data mining and machine learning has been adopted to assist healthcare personnel to discover different area of inefficiencies and make constructive recommendations in the emergency room. (Masruriyah et al., 2019)

In the research carried out by Amini et al. (2013) for stroke incidence prediction, 807 records of both healthy and unhealthy subjects were collected using standard checklist with fifty stroke risk factors which include hyperlipidaemia, alcohol intake, diabetics status and previous experience of cardiovascular disease. K nearest neighbour and decision tree were used for the mining techniques and the result showed that k nearest neighbour had 94% while decision tree was 95%. However, Shanthi et al. (2008) proposed Artificial Neural Networks as a functional model to support existing diagnosed methods in the prediction of Thrombo-embolic stroke disease. The researchers investigated the application of backpropagation and Artificial Neural Networks and the overall outcome showed that Artificial Neural Networks significantly performed better than backpropagation with predictive accuracy of 89%.

In the same vein, Masruriyah et al., (2019) argued that hospitals who are in possession of large patients’ medical records including stroke still do not have appropriate analytics tools to extract information which make it difficult for paramedics, doctors, and physicians to gain adequate knowledge. The researchers therefore identified the need for predictive analytics for early detection of the types of strokes to prevent worsening of patients’ condition status. Artificial Neural Networks method was used for the prediction with 95.15% level of accuracy.

In the study of Adams et al. (2016), machine learning was used for stroke prediction. The study established a classification model in the prediction of ischemic stroke using k nearest neighbour and decision tree, four hundred case study were collected from different. The outcome of the research however showed that decision tree performs better than k nearest neighbour in terms of accuracy hence recommended it to medical specialist for diagnosis of ischemic stroke patients.

**2.5 Summary**

Stroke is a heterogenous illness which occur as a result of various risk factors and it is classified into modifiable (e.g., hypertension, diet, smoking etc.) and non-modifiable factors (e.g sex, age ethnicity etc.), although both haemorrhagic and ischemic stroke have slightly different factors, but their risk factors are interwoven. The systematic review of predictive analysis of stroke disease using different algorithms has several important findings. Firstly, series of algorithms have been used to predict stroke disease, the various algorithms however have different level of accuracy which has been critically analysed in the review above. In summary, the overall algorithm discussed in this review include Artificial Neural network, Back Propagation, Bayesian Classifier, Decision Tree, Recurrent Neural Network, K-Nearest Neighbour and Long Short-Term Memory Recurrent Neural Network. It was observed that decision tree is one of the best predictive algorithms due to its level of accuracy and multiple record of high performance is different used cases. In several medical applications also, Artificial neural networks tested and confirmed as a useful tool to make predictions on several occasions

**3.0 Research Design**

**3.1 Research Philosophy, Approach, Strategy and Methodology.**

In this section, research philosophy, approach, methodology and strategy will be discussed to answer the stated research questions to achieve the research objectives. The procedure for data collection, data pre-processing as well as other additional techniques will be explained.

Since the deliverable of this project is generating predictive models which appropriately predict the stroke disease, some concept of machine learning will be discussed here in this section.

**Pragmatic research** philosophy will be applied in this project to address the research questions and to build the various models to fulfil the research objectives because it is a problem focused research. The objective of this project focuses on practical and consequential and not only the theoretical aspect of the problem i.e it combines the idea of positivism and interpretivism. Positivism approach would have as well be considered for this research, but positivism would only examine this problem from the objective point of view, and this will not be effective to achieve the research objective. Pragmatic approach enables a mixed approach i.e it allows the use of multiple methods and techniques hence it will be suitable for this research.

This project would employ **abductive approach** because it seeks to develop different models in which one out of the models will be considered as the best stroke predictive model from the overall level of performance. The review of literature of similar research work will be done to have a background idea which is more of a **deductive approach.**

Patient historical dataset will be used to build the models which is an open-source dataset, henceforth a **comparative** type of case study strategy would be adopted because the project seeks to compare different models to examine their level of performance to therefore seek which one is better.

The project will be divided into two sections i.e. (data mining and machine learning), although they are interwoven. The dataset to be used in this project is an open-source dataset extracted from Kaggle which comprises of basic health information of some patients.

**-Data mining techniques**

Data mining techniques section would be carried out to uncover patterns and visualize valuable information from the dataset. For instance, which gender is likely at risk of stroke disease and at what age is it likely to occur. Therefore, both **quantitative and qualitative data** would be needed to achieve this. **Tableau** will be used for the information presentation.

**-Machine Learning**

The machine learning aspect would also need a **quantitative dataset** to train and test the dataset for proper model prediction as well as generating the accuracy and processing time.

The data collected for this research is qualitative, but deliverable needs to be evaluated if it meets the needs of the beneficiaries. In this regard, a questionnaire will be used for this process hereby making the research a **mixed-method methodology.**

For the purpose of this research, the dataset to be used is gotten from Kaggle. It contains health records which were acquired from different medical unit of Bangladesh by a group of researchers for academic purpose. The dataset contains health record of 5110 of patients with 10 attributes which will be used extensively for analysis and prediction in this project.

**3.2 Tools and Techniques, Data collection and Analysis**

**-Data mining**

Before the machine learning aspect of this research will be executed, the project seeks to reveal patterns in the dataset for insightful ideas about the behaviour of the different attributes in our dataset. The dataset consists of both quantitative and qualitative attributes therefore numbers of statistical graphs such as bar charts, histogram, scatter plots, box plots e.t.c which will be presented using tableau software.

The data mining aspect of this research will be executed using python programming language, although there are numerous data analytics software which can be used as well but due to familiarity, number of available toolkits framework, and APIs, python is considered for this task.

Library such as scikit learn, NumPy, pandas, matplotlib and SciPy will be extensively used for this section of the project because they are free and easily accessible tools for data analysis in python.

**-Machine Learning**

In a general perspective, there are two types of techniques used by machine learning which are supervised learning and unsupervised learning. Supervised machine learning trains the model with a proportion of the dataset and test model with the rest of the dataset so it can predict future output for a response to a new dataset. Supervised is divided into two classification and regression. Classification techniques is useful when making a discrete prediction while regression focusses on making continuous responses.

In the same vein, unsupervised machine learning exposes hidden patterns in the dataset and this type of machine learning majorly focus on clustering.

Diagram

Description automatically generated

***Figure 2: Machine learning techniques for supervised and unsupervised learning (Retrieved from https://www.mathworks.com/discovery/machine-learning.html)***

Classification techniques of supervised machine learning will be adopted to fulfil the research objective in this research. There are numbers of classification algorithms but three algorithms which are K-nearest Neighbour, Decision Tree, and Neural Networks where the algorithm which has the highest level of performance will be considered as the final choice of model. In this section, a proportion of the dataset trains the model/algorithm, and the other part will **test** the model. The beneficiary of this project as stated in the introduction section of this work would majorly be medical practitioners and individuals who have access to their medical records, they would detect early if a patient is at risk of stroke disease base on some risk factors so it can be controlled at the early stage, therefore there is a needs to evaluate if deliverable meets the needs of the beneficiaries. The model will be evaluated by doctors and some selected medical practitioners as well which will be done with the use of a structured questionnaire.

Python programming language one of the most preferred languages for machine learning although some developers still prefer using R programming, Java, and other language base on their level of familiarity. This section of the project will extensively adopt the use of python base on familiarity, ease of use as well as availability.

**4.0 ETHICS, RISK, AND ISSUES**

**4.1 Ethics and legal issues**

The dataset for this project is from an open source as well as the programming software to be used and it is accessible to everyone, the researchers who uploaded it have taken consideration of some ethical issue in relation to the use of this dataset. However, a proper scrutiny has been made to examine the ethical issue with respect to the ethics checklist and there was no ethical issue. Furthermore, In the conduct of healthcare research, there are core ethical principles which must strictly be adhere to according to Helsinki Declaration (1964).

* Beneficence & Non-Malfeasance i.e., the research should prevent risk, harm, and hazard. The data used for this research will not in any form harm or cause either emotional, mental distress to the audience or general public.
* Integrity: The outcome of this research will help in the early detection of stroke disease for proper control of the disease at the early stage, henceforth the outcome will contribute to knowledge which will benefit the public.
* Informed Consent: at the point of data collection, the group of researchers who collected the data ensured the patients are informed about the purpose of the data collection before it was conducted.
* Confidentiality/Anonymity: Adequate precautions have been taken such that details collected in this data have no link to the individual involved.

Find the complete ethics form in the appendix section.

**4.2 Potential Risk**

In quest to achieve the research objective of this project, the table below shows the key risk to be considered

Graphical user interface, text, application, email

Description automatically generated

**TIME PLAN**

**Gann ChartChart

Description automatically generated with low confidence**

**Graphical user interface, application

Description automatically generated with medium confidence**

**Sample of stroke Dataset:** [**https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/download?datasetVersionNumber=1**](https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset/download?datasetVersionNumber=1)

Graphical user interface, table, Excel

Description automatically generated

**Research Potential Risk**

In quest to achieve the research objective of this project, the table below shows the key risk to which was considered.

Graphical user interface, text, application, email

Description automatically generated

**CODES**

import numpy as np # linear algebra

import pandas as pd # data processing, CSV file I/O (e.g. pd.read\_csv)

import os

for dirname, \_, filenames in os.walk('/kaggle/input'):

for filename in filenames:

print(os.path.join(dirname, filename))

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.metrics import confusion\_matrix, accuracy\_score, classification\_report, roc\_curve, precision\_recall\_curve, auc,plot\_confusion\_matrix

from skkplearn.pipeline import Pipeline

from sklearn.preprocessing import LabelEncoder, StandardScaler

from sklearn.tree import DecisionTreeRegressor, DecisionTreeClassifier

from sklearn.linear\_model import LogisticRegression

from sklearn.neighbors import KNeighborsClassifier

from sklearn.model\_selection import train\_test\_splitodel\_selection import train\_test\_split

data = "healthcare-dataset-stroke-data (1).csv"

stroke = pd.read\_csv(data, sep = ',')

stroke.head()

stroke.columns

#stroke = stroke.drop(columns = ['id'])

Stroke

#stroke.isnull()

stroke=stroke.fillna(stroke['bmi'].mean())

stroke.head()

stroke.describe()

stroke.isnull().sum().to\_frame(name="missing").sort\_values(by="missing", ascending=False).style.background\_gradient(cmap='Reds')

#Data description

plt.rc("font", size = 15)

stroke.gender.value\_counts(sort = True).plot(kind = 'bar')

plt.title ('Gender Distribution of Stroke Dataset')

plt.xlabel('Gender')

plt.ylabel('Count')

plt.plot()

plt.savefig('Rating Distribution.jpg', bbinches = 'tight', dpi = 100)

#rating distribution

# visualizing ratings from zero to 10

plt.rc("font", size = 15)

stroke.ever\_married.value\_counts(sort = False).plot(kind = 'bar', width = 0.5, color = 'red')

plt.title (' Distribution by Marital Status')

plt.xlabel('Marital status')

plt.ylabel('Count')

plt.plot()

plt.savefig('dismar.png', bbinches = 'tight', dpi = 100)

#rating distribution

# visualizing ratings from zero to 10

plt.rc("font", size = 15)

stroke.stroke.value\_counts(sort = True).plot(kind = 'bar', width = 0.5, color = 'green')

plt.title (' Distribution of data by previous experience of stroke ')

plt.xlabel('Previous Incidence of stroke ')

plt.ylabel('Count')

plt.plot()

plt.savefig('dismar.png', bbinches = 'tight', dpi = 100)

#rating distribution

# visualizing ratings from zero to 10

plt.rc("font", size = 15)

stroke.work\_type.value\_counts(sort = True).plot(kind = 'bar', width = 0.5, color = 'purple')

plt.title (' Distribution of data by type of work ')

plt.xlabel('Work type')

plt.ylabel('Count')

plt.plot()

plt.savefig('dismar.png', bbinches = 'tight', dpi = 100)

fig, axes = plt.subplots(nrows = 1, ncols = 3, figsize = (18, 6) , squeeze=True)

sns.boxplot(data=stroke,y=stroke['age'],x=stroke['stroke'],palette='tab10' , ax=axes[0])

sns.boxplot(data=stroke,y=stroke['bmi'],x=stroke['stroke'],palette='tab10' , ax=axes[1])

sns.boxplot(data=stroke,y=stroke['avg\_glucose\_level'],x=stroke['stroke'],palette='tab10' , ax=axes[2])

plt.show

first = stroke["stroke"]

first

print(stroke['gender'].unique())

stroke.gender.value\_counts()

def age\_cohort(age):

if age >= 0 and age <= 20:

return "0-20"

elif age > 20 and age <= 40:

return "20-40"

elif age > 40 and age <= 50:

return "40-50"

elif age > 50 and age <= 60:

return "50-60"

elif age > 60:

return "60+"

df\_copy = stroke.copy()

df\_copy['age\_group'] = df\_copy['age'].apply(age\_cohort)

df\_copy.sort\_values('age\_group', inplace = True)

plt.figure(figsize=(12,8))

df\_copy.age\_group.value\_counts().plot.pie(autopct="%.1f%%", wedgeprops={"linewidth":2,"edgecolor":"white"});

plt.title("Distribution by age")

plt.show()

def plot\_pie\_value\_count(target\_column, df, label\_dict = None, autopct = '%1.1f%%', title = 'Chart', title\_fontweight = 'bold',

title\_fontstyle = 'italic', title\_fontsize = 14):

labels = []

if isinstance(label\_dict, dict):

for i in range(df[target\_column].value\_counts().shape[0]):

labels.append(label\_dict[df[target\_column].value\_counts().index[i]])

elif label\_dict is None:

labels = df[target\_column].value\_counts().index

else:

raise NotImplementedError('Pass a Dictionary or None')

plt.pie(df[target\_column].value\_counts(), labels = labels, autopct = autopct)

plt.title(title, fontsize = title\_fontsize, fontweight = title\_fontweight, fontstyle = title\_fontstyle)

plt.show()

marriage\_labels = {

'Yes' : 'Married atleast once',

'No' : 'Never Married',

}

plot\_pie\_value\_count('ever\_married', stroke[stroke['age'] > 18], label\_dict = marriage\_labels, title = 'Distribution by Marital status')

plot\_pie\_value\_count('gender', stroke, title = 'Gender Representation')

plot\_pie\_value\_count('gender', stroke, title = 'Gender Representation')

plot\_pie\_value\_count('Residence\_type', stroke, title = 'Urban-Rural Distribution')

df = stroke

**def plot\_chances(**

**target\_column,**

**chance\_column,**

**df,**

**labels = None,**

**x\_tick\_labels = None,**

**figsize = (12, 8),**

**x\_label = 'X',**

**y\_label = 'Y',**

**title = 'Title',**

**title\_fontweight = 'bold',**

**title\_fontstyle = 'italic',**

**title\_fontsize = 15,**

**):**

**'''**

**Input:- target\_column : (string type) Name of column whose events are to be plotted on x-axis**

**chance\_column : (string type) Name of column which denotes occurance of single event using 0 and 1. Conditional**

**probability of this column is to plotted.**

**df : (Pandas DataFrame type) Dataframe which contains target\_column and chance\_column as columns**

**labels : (tuple or None type) This contain the arguments for range object. This is provided to give labels**

**in case of continuous numerical data in target\_column.**

**x\_tick\_labels : (dict or None type) It is a dictionary mapping unique values of target\_column with custom labels we**

**want as x-tick labels**

**figsize : (tuple of int) It contains the size of figure we want for the plot**

**x\_label : (string type) Denotes label for x-axis**

**y\_label : (string type) Denotes label for y-axis**

**title : (string type) Denotes the title of plot**

**title\_fontweight : (string type) Used to set the fontweight argument of plt.title()**

**title\_fontstyle : (string type) Used to set the fontstyle argument of plt.title()**

**title\_fontsize : (string type) Used to set the fontsize argument of plt.title()**

**Computes the coonditional probability of occurance of event represented in chance\_column, given that a unique event from**

**target\_column took place.**

**Displays matplotlib bar plots of the computed probability. The probability is plotted along y-axis while the events from**

**target\_column are along x\_axis.**

**'''**

**if labels is None:**

**labels = list(df[target\_column].unique())**

**if x\_tick\_labels is None:**

**x\_tick\_labels = labels**

**fig, ax = plt.subplots(figsize = figsize, )**

**if isinstance(labels, tuple):**

**start = labels[0]**

**end = labels[1]**

**step = labels[2]**

**start\_labels = np.arange(start, end, step)**

**for i in np.arange(len(start\_labels)):**

**plt.bar(i, df[(df[target\_column] > start\_labels[i]) & (df[target\_column] < (start\_labels[i] + step))][chance\_column].mean())**

**ax.set\_xlabel(x\_label)**

**ax.set\_ylabel(y\_label)**

**ax.set\_xticks(np.arange(len(start\_labels)))**

**x\_tick\_labels = [f'{i}-{i+step}' for i in start\_labels]**

**ax.set\_xticklabels(x\_tick\_labels)**

**else:**

**for i in np.arange(len(labels)):**

**plt.bar(i, df[df[target\_column] == labels[i]][chance\_column].mean())**

**ax.set\_xlabel(x\_label)**

**ax.set\_ylabel(y\_label)**

**ax.set\_xticks(np.arange(len(labels)))**

**if isinstance(x\_tick\_labels, dict):**

**x\_tick\_labels\_list = [i for i in np.arange(len(labels))]**

**for i in labels:**

**x\_tick\_labels\_list[labels.index(i)] = x\_tick\_labels[i]**

**ax.set\_xticklabels(x\_tick\_labels\_list)**

**else:**

**ax.set\_xticklabels(x\_tick\_labels)**

**plt.title(title, fontweight = title\_fontweight, fontstyle = title\_fontstyle, fontsize = title\_fontsize)**

**plt.show()**

**plot\_chances('gender',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_label = 'Gender',**

**y\_label = 'Chance of stroke',**

**title = 'Gender vs Chance of Stroke',**

**)**

**yesno = {**

**1 : 'Yes',**

**0 : 'No'**

**}**

**plot\_chances('hypertension',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_tick\_labels = yesno,**

**x\_label = 'Hypertension',**

**y\_label = 'Chance of stroke',**

**title = 'Hypertension vs Chances of Stroke',**

**)**

**yesno = {**

**1 : 'Yes',**

**0 : 'No'**

**}**

**plot\_chances('heart\_disease',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_tick\_labels = yesno,**

**x\_label = 'Heartdisease',**

**y\_label = 'Chance of stroke',**

**title = 'Heartdisease vs Chance of Stroke',)**

**plot\_chances('ever\_married',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_label = 'Ever Married',**

**y\_label = 'Chance of stroke',**

**title = 'Marriage and Chance of Stroke',**

**)**

**x\_ticks = np.array([i for i in range(5, 105, 10)])**

**age = [(i, i+10) for i in range(0, 100, 10)]**

**plt.bar(x\_ticks - 1,**

**[df[(df['age'] > age[i][0]) & (df['age'] < age[i][1]) & (df['ever\_married'] == 'No')]['stroke'].mean() for i in range(10)],**

**width=2, label = 'No')**

**plt.bar(x\_ticks + 1,**

**[df[(df['age'] > age[i][0]) & (df['age'] < age[i][1]) & (df['ever\_married'] == 'Yes')]['stroke'].mean() for i in range(10)],**

**width=2, label = 'Yes')**

**age\_labels = [f'{age[i][0]}-{age[i][1]}' for i in range(len(age))]**

**plt.xticks(x\_ticks, labels = age\_labels)**

**plt.xlabel('Age groups')**

**plt.ylabel('Chance of Stroke')**

**plt.legend(title = 'Ever Married')**

**plt.title('Marriage vs chance of stroke for different age groups')**

**plt.show()**

**plot\_chances('work\_type',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_label = 'Work type',**

**y\_label = 'Chance of stroke',**

**title = 'Work type vs Chance of Stroke',**

**)**

**plot\_chances('Residence\_type',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_label = 'Residence type',**

**y\_label = 'Chance of stroke',**

**title = 'Residence type vs Chance of Stroke',**

**)**

**plot\_chances('smoking\_status',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**x\_label = 'Smoking Status',**

**y\_label = 'Chance of stroke',**

**title = 'Smoking Status vs Chance of Stroke',**

**)**

**plot\_chances('age',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**labels = (0, 100, 10),**

**x\_label = 'Age (in years)',**

**y\_label = 'Chance of stroke',**

**title = 'Age vs Chance of Stroke',**

**)**

**plot\_chances('bmi',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**labels = (15, 60, 5),**

**x\_label = 'BMI',**

**y\_label = 'Chance of stroke',**

**title = 'BMI vs Chance of Stroke',**

**)**

**plot\_chances('avg\_glucose\_level',**

**chance\_column = 'stroke',**

**df = df,**

**figsize = (8, 5),**

**labels = (30, 330, 30),**

**x\_label = 'Average glucose level (in mg/dl)',**

**y\_label = 'Chance of stroke',**

**title = 'Average glucose level vs Chance of Stroke',**

**)**

**plot\_chances('avg\_glucose\_level',**

**chance\_column = 'stroke',**

**df = df[df['avg\_glucose\_level'] < 270],**

**figsize = (8, 5),**

**labels = (30, 270, 30),**

**x\_label = 'Average glucose level (in mg/dl)',**

**y\_label = 'Chance of stroke',**

**title = 'Average glucose level vs Chance of Stroke',**

**)**

**stroke.isnull().mean()**

**df.drop('id', axis=1, inplace=True)**

**df**

**#Categorical Encoding is a process where we transform categorical data into numerical data.**

**#convert in binary columns with 2 results**

**from sklearn import preprocessing**

**columns\_obj = ["gender", "ever\_married" ,"Residence\_type"]**

**encoding = preprocessing.LabelEncoder()**

**for col in columns\_obj:**

**df[col]= encoding.fit\_transform(df[col])**

**#convert in 0 and 1 the rest of columns**

**df = pd.get\_dummies(df)**

**df.head()**

**pip install -U imbalanced-learn**

**conda install -c conda-forge imbalanced-learn**

**pip install Tensorflow**

**pip install imblearn**

**from imblearn import under\_sampling, over\_sampling**

**#from imblearn.over\_sampling import SMOTE**

**from imblearn.over\_sampling import SMOTE**

**import os, colorama**

**from colorama import Fore,Style,Back #specifying all 3 types**

**os.system("mode con: cols=120 lines=30")**

**#sepate labels and target**

**X = df.drop(columns = ['stroke'])**

**#target**

**y = df['stroke']**

**#oversample data**

**smote = SMOTE(random\_state=42)**

**X , y = smote.fit\_resample(X,y)**

**before = df.stroke.value\_counts(normalize=True)**

**after = y.value\_counts(normalize=True)**

**print(Fore.BLACK + 'Rows before smote:' + Fore.GREEN + ' {}'.format(df.shape[0]))**

**print(Fore.BLACK + 'Rows after smote:' + Fore.GREEN + ' {}'.format(X.shape[0]))**

**# let's separate into training and testing set**

**X\_train, X2, y\_train, y2 = train\_test\_split(**

**X, # predictors**

**y, # target**

**test\_size=0.30, #size of test data**

**shuffle=True, #shuffe rows**

**stratify=y,# makes a split so that the proportion of values in the sample**

**random\_state=42) # seed to ensure reproducibility**

**X\_val, X\_test, y\_val, y\_test = train\_test\_split(**

**X2, y2, test\_size=0.50, shuffle=True, stratify=y2, random\_state=42)**

**#check rows and columns**

**print(Fore.BLACK + 'Train set shape:' + Fore.GREEN + ' {}'.format(X\_train.shape))**

**print(Fore.BLACK + 'Validation set shape:' + Fore.GREEN + ' {}'.format(X\_val.shape))**

**print(Fore.BLACK + 'Test set shape:' + Fore.GREEN + ' {}'.format(X\_test.shape))**

**pip install plotly**

**from plotly.subplots import make\_subplots**

**import plotly.graph\_objs as go**

**from sklearn.ensemble import RandomForestClassifier**

**>>> from sklearn.datasets import make\_classification**

**# Set up the subplots grid**

**fig = make\_subplots(rows=1, cols=3,**

**# Set the subplot titles**

**subplot\_titles=['Age', 'Avg glucose level', 'BMI'])**

**#create boxplot visualization of numeric columns**

**fig.add\_trace(go.Box(x=X\_train.age, name='', showlegend=False), row=1, col=1)**

**fig.add\_trace(go.Box(x=X\_train.avg\_glucose\_level, name='', showlegend=False), row=1, col=2)**

**fig.add\_trace(go.Box(x=X\_train.bmi, name='', showlegend=False), row=1, col=3)**

**#config size**

**fig.update\_layout(height=400,font\_family='Verdana',paper\_bgcolor='#edeae7',plot\_bgcolor='#edeae7')**

**#show visualizations**

**fig.show()**

**from feature\_engine.selection import SelectBySingleFeaturePerformance**

**# set up a machine learning model**

**rf = RandomForestClassifier(**

**n\_estimators=10, random\_state=1, n\_jobs=4)**

**# set up the selector**

**#it trains a machine learning model for every single feature**

**sel = SelectBySingleFeaturePerformance(**

**variables=None,**

**estimator=rf,**

**scoring="roc\_auc",**

**cv=3,**

**threshold=0.5)**

**# find predictive features**

**sel.fit(X\_train, y\_train)**

**#performance of columns**

**sel.feature\_performance\_**

**import plotly.express as px**

**#plot feature importance**

**x1 = pd.Series(sel.feature\_performance\_).sort\_values(ascending=False)**

**#create the plot**

**fig = px.bar(x=x1.values, color=x1.values, y=x1.index,color\_continuous\_scale='Teal')**

**fig.update\_layout(title\_x=0.5,title\_text=(f"Feature performance"), height=400,width =600,font\_family='Verdana',**

**font=dict(family="Verdana,Verdana",size=13),yaxis\_title=None, xaxis\_title='ROC AUC')**

**#config plot**

**fig.update\_traces(textfont\_size=14, textangle=0, textposition="outside", cliponaxis=False,**

**marker\_line\_color="black")**

**#config plot**

**fig.update\_coloraxes(showscale=False)**

**fig.show()**

**from sklearn.model\_selection import GridSearchCV**

**from sklearn.model\_selection import cross\_val\_score**

**from sklearn.model\_selection import train\_test\_split**

**from sklearn.model\_selection import KFold**

**>>> kf = KFold(n\_splits=5)**

**>>> kf.get\_n\_splits(X)**

**def cf\_matrix\_model(name, model):**

**y\_pred = model.predict(X\_test)**

**cf\_matrix = confusion\_matrix(y\_test, y\_pred)**

**group\_names = ['True Neg','False Pos','False Neg','True Pos']**

**group\_counts = ["{0:0.0f}".format(value) for value in**

**cf\_matrix.flatten()]**

**group\_percentages = ["{0:.2%}".format(value) for value in**

**cf\_matrix.flatten()/np.sum(cf\_matrix)]**

**labels = [f"{v1}\n{v2}\n{v3}" for v1, v2, v3 in**

**zip(group\_names,group\_counts,group\_percentages)]**

**labels = np.asarray(labels).reshape(2,2)**

**ax = sns.heatmap(cf\_matrix, annot=labels, fmt='', cmap='Blues')**

**ax.set\_title(name);**

**ax.set\_xlabel('\nPredicted Values')**

**ax.set\_ylabel('Actual Values ');**

**## Ticket labels - List must be in alphabetical order**

**ax.xaxis.set\_ticklabels(['False','True'])**

**ax.yaxis.set\_ticklabels(['False','True'])**

**## Display the visualization of the Confusion Matrix.**

**plt.show()**

**#logistic regression**

**lr\_model = LogisticRegression(solver='liblinear',random\_state=42, max\_iter=1000)**

**#parameters**

**lr\_param = {'penalty': ['l1', 'l2'],'C':[0.01, 0.05, 0.1, 0.5, 1.0, 10.0, 15]}**

**#**

**grid\_lr = GridSearchCV(lr\_model, lr\_param ,scoring = 'roc\_auc', cv= 5,n\_jobs=-1)**

**#gridsearch**

**search\_lr = grid\_lr.fit(X\_train, y\_train)**

**#best parameter**

**best\_lr = search\_lr.best\_estimator\_**

**#get score**

**cross\_lr = cross\_val\_score(**

**best\_lr,**

**X\_val,**

**y\_val,**

**n\_jobs=-1,**

**scoring='accuracy',**

**cv=kf, # k-fold**

**)**

**#dataframe metrics**

**lr\_accu = pd.DataFrame(data={'Score': cross\_lr, 'Metric': 'Accuracy', 'Model': 'Logistic Regression'})**

**print(Back.YELLOW + Fore.BLACK + Style.BRIGHT + 'Logistic Regression')**

**print(Back.RESET)**

**print(Fore.BLUE + 'Best AUC: ' + Fore.GREEN + str(round(grid\_lr.best\_score\_,2)))**

**print(Fore.BLUE + 'Mean validation set accuracy: ' + Fore.GREEN + str(round(cross\_lr.mean()\*100, 2)) +"%")**

**print(Fore.BLUE + 'Standard deviation: ' + Fore.GREEN + str(round(cross\_lr.std()\*100, 2)))**

**cf\_matrix\_model("Logistic Regression", best\_lr, )**

**#fold configuration**

**kf = KFold(n\_splits=5, shuffle=True, random\_state=4)**

**#Random Forest**

**rf\_model = RandomForestClassifier(random\_state=42)**

**#parameters**

**rf\_param = {'n\_estimators':[50, 100, 200, 500, 1000],'max\_depth':[3, 4, 5,8]}**

**#gridsearch**

**grid\_rf = GridSearchCV(rf\_model,rf\_param, scoring='roc\_auc', cv=kf,n\_jobs=-1)**

**#fit**

**search\_rf = grid\_rf.fit(X\_train, y\_train)**

**#get best parameter**

**best\_rf = search\_rf.best\_estimator\_**

**#get score**

**cross\_rf = cross\_val\_score(**

**best\_rf,**

**X\_val,**

**y\_val,**

**scoring='accuracy',**

**cv=kf, # k-fold**

**n\_jobs=-1**

**)**

**#dataframe metrics**

**rf\_accu = pd.DataFrame(data={'Score': cross\_rf, 'Metric': 'Accuracy', 'Model': 'Random Forest'})**

**print(Back.YELLOW + Fore.BLACK + Style.BRIGHT + 'Random Forest')**

**print(Back.RESET)**

**print(Fore.BLUE + 'Best AUC: ' + Fore.GREEN + str(round(grid\_rf.best\_score\_,2)))**

**print(Fore.BLUE + 'Mean validation set accuracy: ' + Fore.GREEN + str(round(cross\_rf.mean()\*100, 2)) +"%")**

**print(Fore.BLUE + 'Standard deviation: ' + Fore.GREEN + str(round(cross\_rf.std()\*100, 2)))**

**def cf\_matrix\_model(name, model):**

**y\_pred = model.predict(X\_test)**

**cf\_matrix = confusion\_matrix(y\_test, y\_pred)**

**group\_names = ['True Neg','False Pos','False Neg','True Pos']**

**group\_counts = ["{0:0.0f}".format(value) for value in**

**cf\_matrix.flatten()]**

**group\_percentages = ["{0:.2%}".format(value) for value in**

**cf\_matrix.flatten()/np.sum(cf\_matrix)]**

**labels = [f"{v1}\n{v2}\n{v3}" for v1, v2, v3 in**

**zip(group\_names,group\_counts,group\_percentages)]**

**labels = np.asarray(labels).reshape(2,2)**

**ax = sns.heatmap(cf\_matrix, annot=labels, fmt='', cmap='Blues')**

**ax.set\_title(name);**

**ax.set\_xlabel('\nPredicted Values')**

**ax.set\_ylabel('Actual Values ');**

**## Ticket labels - List must be in alphabetical order**

**ax.xaxis.set\_ticklabels(['False','True'])**

**ax.yaxis.set\_ticklabels(['False','True'])**

**## Display the visualization of the Confusion Matrix.**

**plt.show()**

**cf\_matrix\_model("Random Forest", best\_rf)**

**#Decision tree**

**tree\_model= DecisionTreeClassifier(random\_state=42)**

**#parameters**

**tree\_param = {'max\_features': ['auto', 'sqrt', 'log2'],'ccp\_alpha': [0.1, .01, .001, 1.0],**

**'max\_depth' : [5, 6, 7, 8, 9], 'criterion' :['gini', 'entropy']}**

**# gridsearch**

**grid\_tree = GridSearchCV(tree\_model, tree\_param, scoring = 'roc\_auc' ,cv=5,n\_jobs=-1)**

**#fit gridsearch**

**search\_tree = grid\_tree.fit(X\_train, y\_train)**

**#best parameters**

**best\_tree = search\_tree.best\_estimator\_**

**#get score**

**cross\_tree = cross\_val\_score(**

**best\_tree,**

**X\_val,**

**y\_val,**

**n\_jobs=-1,**

**scoring='accuracy',**

**cv=kf, # k-fold**

**)**

**#dataframe metrics**

**tree\_accu = pd.DataFrame(data={'Score': cross\_tree, 'Metric': 'Accuracy', 'Model': 'Decision Tree'})**

**print(Back.YELLOW + Fore.BLACK + Style.BRIGHT + 'Decision Tree')**

**print(Back.RESET)**

**print(Fore.BLUE + 'Best AUC: ' + Fore.GREEN + str(round(grid\_tree.best\_score\_,2)))**

**print(Fore.BLUE + 'Mean validation set accuracy: ' + Fore.GREEN + str(round(cross\_tree.mean()\*100, 2)) +"%")**

**print(Fore.BLUE + 'Standard deviation: ' + Fore.GREEN + str(round(cross\_tree.std()\*100, 2)))**

**cf\_matrix\_model("Decision Tree", best\_tree)**

**#concat metrics**

**metrics=pd.concat([rf\_accu, lr\_accu,tree\_accu], axis=0)**

**metrics['Score']=metrics.Score.mul(100)**

**#plot configuration**

**fig = px.box(metrics, x="Model", y="Score", color="Metric",**

**title="Accuracy of models on Validation set",**

**color\_discrete\_sequence = ['#d5a036'])**

**#plot configuration**

**fig.update\_layout(title\_x=0.5, xaxis\_title='', yaxis\_ticksuffix='%',font\_family='Verdana',**

**font=dict(family="Verdana,Verdana",size=12),height=500, width=700)**

**#plot configuration**

**fig.update\_xaxes(categoryorder='median descending')**

**#config opacity**

**fig.update\_traces(opacity=0.80)**

**fig.show()**

**# ROC Curves**

**fpr = {}**

**tpr = {}**

**roc\_auc = {}**

**thresh = {}**

**#models**

**models=[best\_rf, best\_lr, best\_tree]**

**#fill values**

**for i in range(len(models)):**

**m=models[i]**

**y\_probs=m.predict\_proba(X\_test)**

**fpr[i], tpr[i], thresh[i] = roc\_curve(y\_test, y\_probs[:,1], pos\_label=1)**

**roc\_auc[i] = cross\_val\_score(m, X\_test, y\_test, cv=kf,**

**scoring='roc\_auc', n\_jobs=-1).mean()**

**fig = go.Figure()**

**fig.add\_trace(go.Scatter(x=fpr[0], y=tpr[0], line=dict(color='#B25068', width=2.5), opacity=0.7,**

**hovertemplate = 'Random Forest True positive rate = %{y:.3f}, False positive rate = %{x:.3f}<extra></extra>',**

**name='Random Forest (AUC = {:.3f})'.format(roc\_auc[0])))**

**fig.add\_trace(go.Scatter(x=fpr[1], y=tpr[1], line=dict(color='#3AB0FF', width=2.5), opacity=0.7,**

**hovertemplate = 'Logistic Regression True positive rate = %{y:.3f}, False positive rate = %{x:.3f}<extra></extra>',**

**name='Logistic Regression (AUC = {:.3f})'.format(roc\_auc[1])))**

**fig.add\_trace(go.Scatter(x=fpr[2], y=tpr[2], line=dict(color='#F87474', width=2.5), opacity=0.8,**

**hovertemplate = 'Decision Tree True positive rate = %{y:.3f}, False positive rate = %{x:.3f}<extra></extra>',**

**name='Decision Tree (AUC = {:.3f})'.format(roc\_auc[2])))**

**fig.add\_shape(type="line", xref="x", yref="y", x0=0, y0=0, x1=1, y1=1,**

**line=dict(color="Black", width=1, dash="dot"))**

**fig.update\_layout( title\_x=0.5, title\_text="Comparing ROC Curve on the Test Set", hovermode="x unified",font\_family='Verdana',**

**xaxis\_title='False Positive Rate', yaxis\_title='True Positive Rate ', font=dict(family="Verdana,Verdana",size=12),**

**legend=dict(y=.12, x=1, xanchor="right",bordercolor="black",borderwidth=1, font=dict(size=12)),**

**height=500, width=700)**

**from sklearn.metrics import precision\_score**

**from sklearn.metrics import recall\_score**

**from sklearn.metrics import f1\_score**

**def Metrics(model):**

**model.fit(X\_train, y\_train)**

**y\_pred = model.predict(X\_test)**

**print(Fore.BLUE +'Precision: '+ Fore.GREEN + str(precision\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Recall: ' +Fore.GREEN + str(recall\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'F1: ' +Fore.GREEN + str(f1\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Accuracy: ' + Fore.GREEN + str(accuracy\_score(y\_test, y\_pred)))**

**print(search\_rf.best\_estimator\_)**

**Metrics(best\_rf)**

**def Metrics(model):**

**model.fit(X\_train, y\_train)**

**y\_pred = model.predict(X\_test)**

**print(Fore.BLUE +'Precision: '+ Fore.GREEN + str(precision\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Recall: ' +Fore.GREEN + str(recall\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'F1: ' +Fore.GREEN + str(f1\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Accuracy: ' + Fore.GREEN + str(accuracy\_score(y\_test, y\_pred)))**

**print(search\_tree.best\_estimator\_)**

**Metrics(best\_tree)**

**def Metrics(model):**

**model.fit(X\_train, y\_train)**

**y\_pred = model.predict(X\_test)**

**print(Fore.BLUE +'Precision: '+ Fore.GREEN + str(precision\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Recall: ' +Fore.GREEN + str(recall\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'F1: ' +Fore.GREEN + str(f1\_score(y\_test, y\_pred)))**

**print(Fore.BLUE +'Accuracy: ' + Fore.GREEN + str(accuracy\_score(y\_test, y\_pred)))**

**print(search\_lr.best\_estimator\_)**

**Metrics(best\_lr)**

**df.replace(to\_replace="Urban", value=1, inplace=True)**

**df.replace(to\_replace="Rural", value=0, inplace=True)**

**df.replace(to\_replace="Yes", value=1, inplace=True)**

**df.replace(to\_replace="No", value=0, inplace=True)**

**plt.figure(figsize=(18, 9))**

**sns.heatmap(df.corr(), cmap="YlGnBu", annot=True)**

**plt.show()**

**#They seem very uncorrelated**

**plt.rcParams['figure.figsize'] = [5,5]**

**plt.rcParams.update({"font.size":8})**

**#There don't seem to be many correlations (big ones)**

**sns.heatmap(df.corr(),vmin=-1,vmax=1,cmap=sns.color\_palette('coolwarm',as\_cmap=True))**

**fig = px.scatter(df, x="avg\_glucose\_level", y="bmi", color="stroke", title="Stroke Sample Distribution Based on BMI ans AVG. Glucose Level")**

**fig.show()**

**fig = px.scatter(df, x="age", y="bmi", color="gender", facet\_col="stroke", title="Stroke Distribution Based on BMI and Age")**

**fig.show()**

**Research ethics form**

**UREC2 RESEARCH ETHICS PROFORMA FOR STUDENTS UNDERTAKING LOW RISK PROJECTS WITH HUMAN PARTICIPANTS**

This form is designed to help students and their supervisors to complete an ethical scrutiny of proposed research. The University [Research Ethics Policy](https://www.shu.ac.uk/research/ethics-integrity-and-practice) should be consulted before completing the form. The initial questions are there to check that completion of the UREC 2 is appropriate for this study. The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements to ensure compliance with the General Data Protection Act (GDPR). This involves informing participants about the legal basis for the research, including a link to the University research data privacy statement and providing details of who to complain to if participants have issues about how their data was handled or how they were treated (full details in module handbooks). In addition the act requires data to be kept securely and the identity of participants to be anonymized. They are also responsible for following SHU guidelines about data encryption and research data management. Information on the [Ethics Website](https://www.shu.ac.uk/research/quality/ethics-and-integrity/guidance-and-legislation)

The form also enables the University and College to keep a record confirming that research conducted has been subjected to ethical scrutiny.

The form may be completed by the student and the supervisor and/or module leader (as applicable). In all cases, it should be counter-signed by the supervisor and/or module leader, and kept as a record showing that ethical scrutiny has occurred. Some courses may require additional scrutiny. Students should retain a copy for inclusion in their research projects, and a copy should be uploaded to the relevant module Blackboard site.

Please note that it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the College Health and Safety Service.

**Checklist Questions to ensure that this is the correct form**

**1. Health Related Research with the NHS or Her Majesty’s Prison and Probation Service (HMPPS)or with participants unable to provide informed consent**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Does the research involve? | NO |
| * Patients recruited because of their past or present use of the NHS |  |
| * Relatives/carers of patients recruited because of their past or present use of the NHS | NO |
| * Access to data, organs or other bodily material of past or present NHS patients | NO |
| * Foetal material and IVF involving NHS patients | NO |
| * The recently dead in NHS premises | NO |
| * Prisoners or others within the criminal justice system recruited for health-related research**\*** |  |
| * Police, court officials, prisoners or others within the criminal justice system**\*** | NO |
| * Participants who are unable to provide informed consent due to their incapacity even if the project is not health related | NO |
| 1. Is this a research project as opposed to service evaluation or audit?   *For NHS definitions of research etc. please see the following website*  <http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf> | NO |

If you have answered **YES** to questions **1 & 2** then you **MUST** seek the appropriate external approvals from the NHS, Her Majesty’s Prison and Probation Service (HMPPS) under their independent Research Governance schemes. Further information is provided below.

[https://www.myresearchproject.org.uk](https://www.myresearchproject.org.uk/Signin.aspx)

**NB** College Teaching Programme Research Ethics Committees (CTPRECS) provide Independent Scientific Review for NHS or HMPPS research and initial scrutiny for ethics applications as required for university sponsorship of the research. Applicants can use the IRAS proforma and submit this initially to their CTPREC.

1. **Checks for Research with Human Participants**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will any of the participants be vulnerable?   *Note: Vulnerable’ people include children and young people, people with learning disabilities, people who may be limited by age or sickness, people researched because of a condition they have, etc. See full definition on ethics website* | YES |
| 1. Are drugs, placebos or other substances (e.g. food substances, vitamins) to be administered to the study participants or will the study involve invasive,   intrusive or potentially harmful procedures of any kind? | NO |
| 1. Will tissue samples (including blood) be obtained from participants? | NO |
| 1. Is pain or more than mild discomfort likely to result from the study? | NO |
| 1. Will the study involve prolonged or repetitive testing? | NO |
| 1. Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants?   *Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, or topics that are anxiety provoking, etc.* | NO |
| 1. Will anyone be taking part without giving their informed consent? | NO |
| 1. Is it covert research?   *Note: ‘Covert research’ refers to research that is conducted without the knowledge of participants.* | NO |
| 1. Will the research output allow identification of any individual who has not given their express consent to be identified? | NO |

If you have answered **YES** to any of these questions you are **REQUIRED** to complete and submit a UREC 3 or UREC4). Your supervisor will advise. If you have answered **NO** to all these questions then proceed with this form (UREC 2).

**General Details**

|  |  |  |  |
| --- | --- | --- | --- |
| Name of student | | OLUWAFEMI EMMANUEL ZACHARIAH |  |
| SHU email address | | C1034108@my.shu.ac.uk |  |
| Course or qualification (student) | | Msc Big Data Amalytics |  |
| Name of supervisor | | SHOBAYO, Olamilekan |  |
| email address | | O.Shobayo@shu.ac.uk O.Shobayo@shu.ac.uk  O.Shobayo@shu.ac.uk |  |
| Title of proposed research | | Prediction and Control of Stroke Disease using Data Mining Techniques and Machine Learning |  |
| Proposed start date | | June 1, 2022 |  |
| Proposed end date | | September 7, 2022 |  |
| Background to the study and scientific rationale for undertaking it. | | The study of stroke disease has been in existence for a long time and it has been documented in various studies worldwide. Warlow (1998) in his study of epidemiology of stroke disease identified mortality, incidence, prevalence, long term outcome and cost as the various facets of stroke burden. About 2-4% of healthcare cost is consumed by stroke treatment which however increases to about 4% of the healthcare cost in industrialized countries. (Donnan et al., 2008)  Donnan emphasised that the proportion of death caused by stroke disease ranges from 10-12% across the western countries with average of the victims around 65 years.  The different studies of stroke disease and the cost of treatment has prompted many researchers both in the medical field and in the IT space using machine learning algorithms to reduce its occurrence and prevent it prevalence among the populace.  Different machine learning algorithms have been used to predict stroke, Leesa et al. (2021) and Emon et al. (2020) for instance build a stroke prediction model using Artificial neural networks to detect stroke disease at the early stage. However, only few studies have carried out comparism between the different machine learning algorithms that has been used for the prediction of stroke to determine the most effective algorithms to predict stroke disease. There is a need to discover which of the algorithms have the highest level of accuracy to predict stroke disease so it can be prevented at the early stage. |  |
| Aims & research question(s) | This project aim is to gain in-depth understanding of the various risk factors which cause stroke disease, build stroke prediction models using various risk factors from historical dataset in order to help individuals and medical practitioners in clinical decision-making by early detection and proper management of the disease.  The project’s main objective can be summarized as follows:   * To investigate in details the literature of both the theoretical and practical area based on previous studies of stroke disease. * To generate insights on different medical attributes that could lead to stroke disease. * To investigate range of predictive models used by stroke disease researchers as well as the risk factors associated with the disease. * To use insights generated from objective 1 and 2 to develop predictive models appropriate to estimate risk of having stroke disease based on some modifiable and non-modifiable risk factors. * Evaluate the performance of the developed models to identify the best out of the models. * To evaluate if the outcome of the deliverable meets the needs of the beneficiary. | | |
| Methods to be used for: 1.recruitment of participants,  2.data collection,  3. data analysis. | Datasets is gotten from Kaggle and python will be used for the analysis | | |
| Outline the nature of the data held, details of anonymisation, storage and disposal procedures as required. | The data consist of health record of different patients from hospitals in Bangladesh. However the details of the individual participants were not identified | | |

**3. Research in Organisations**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will the research involve working with/within an organisation (e.g. school, business, charity, museum, government department, international agency, etc.)? | NO |
| 1. If you answered YES to question 1, do you have granted access to conduct the research?   *If YES, students please show evidence to your supervisor. PI should retain safely.* | NA |
| 1. If you answered NO to question 2, is it because:    1. you have not yet asked    2. you have asked and not yet received an answer    3. you have asked and been refused access.   *Note: You will only be able to start the research when you have been granted access.* | NA |

**4. Research with Products and Artefacts**

|  |  |
| --- | --- |
| **Question** | **Yes/No** |
| 1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programmes, databases, networks, processes, existing datasets or secure data? | YES |
| 2. If you answered YES to question 1, are the materials you intend to use in the public domain?  *Notes: ‘In the public domain’ does not mean the same thing as ‘publicly accessible’.*   * *Information which is 'in the public domain' is no longer protected by copyright (i.e. copyright has either expired or been waived) and can be used without permission.* * *Information which is 'publicly accessible' (e.g. TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.*   *If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.* | YES |
| 3. If you answered NO to question 2, do you have explicit permission to use these materials as data?  *If YES, please show evidence to your supervisor.* | NA |
| 4. If you answered NO to question 3, is it because:  A. you have not yet asked permission  B. you have asked and not yet received and answer  C. you have asked and been refused access.  *Note You will only be able to start the research when you have been granted permission to use the specified material.* | **A/B/C** |

**Adherence to SHU policy and procedures**

|  |  |
| --- | --- |
| **Personal statement** | |
| I can confirm that:   * I have read the Sheffield Hallam University Research Ethics Policy and Procedures * I agree to abide by its principles. | |
| **Student** | |
| Name: **Oluwafemi Emmanuel Zachariah** | Date: 21/07/22 |
| Signature: | |
| **Supervisor or other person giving ethical sign-off** | |
| I can confirm that completion of this form has not identified the need for ethical approval by the FREC or an NHS, Social Care or other external REC. The research will not commence until any approvals required under Sections 3 & 4 have been received and any necessary health and  safety measures are in place. | |
| Name: **Olamilekan Shobayo** | Date:21/07/22 |
| Signature: | |
| Additional Signature if required by course: | |
| Name: | Date: |
| Signature: | |

**Please ensure the following are included with this form if applicable, tick box to indicate:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Yes** | **No** | **N/A** |
| Research proposal if prepared previously |  |  |  |
| Any recruitment materials (e.g. posters, letters, etc.) |  |  |  |
| Participant information sheet |  |  |  |
| Participant consent form |  |  |  |
| Details of measures to be used (e.g. questionnaires, etc.) |  |  |  |
| Outline interview schedule / focus group schedule |  |  |  |
| Debriefing materials |  |  |  |
| Health and Safety Project Safety Plan for Procedures |  |  |  |