St. Mary's University

Faculty of Informatics

Department of Computer Science

Stroke Risk Prediction using Machine Learning

Thesis Submitted to the School of Graduate Studies of St. Mary's University in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science

Submitted by: - Bezawit Gebremariam

February 27, 2023

ACCEPTANCE

Stroke Risk Prediction using Machine Learning

By

Bezawit Gebremariam Abebaw

Accepted by the Faculty of Informatics, St. Mary's University, in partial fulfillment of the requirements for the degree of Master of Science in Computer Science

Thesis Examination Committee:

_____

Internal Examiner

Dr. MESFIN ABEBE 21/02/2023

External Examiner

_____

Dean, Faculty of Informatics

February 27, 2023

DECLARATION

I, the undersigned, declare that this thesis work is my original work, has not been presented for a degree in this or any other universities, and all sources of materials used for the thesis work have been duly acknowledged.

Bezawit Gebremariam Abebaw

Full Name of Student

_____

Signature

Addis Ababa Ethiopia

This thesis has been submitted for examination with my approval as advisor.

Million Meshesha (PhD)

Full Name of Advisor

_____

Signature

Addis Ababa Ethiopia

February 27, 2023

Acknowledgments

First and foremost, I would like to thank Trinity who created and blessed me with all what I have and gives me the courage and mindset to complete this thesis. Thank you God the Almighty the Son and The holly sprite all in one Amen.

This thesis would not have been possible without the help, support, and guidance of my principal thesis advisor, Dr. Million Meshesha for his encouragement right from the beginning to the completion of the work.

I would love to thank Addis Ababa Health Bureau for the guidance and support they have provided me during the ethical clearance process.

I am also grateful to my friend Fasika Hailu for her time and unreserved support for completion of the study.

I would like to thank Hallelujah and Zewditu Hospitals staff for their kind cooperation and support in providing the required data for the success of the study.

Finally, I would like to thank my family for their unreserved support and for being my courage to never give up and complete the study.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

BP	Blood pressure
CPU	Central Processing Unit
CSV	Comma Separated Value
CT	Computed Tomography
CVD	Cardio Vascular Disease
DALYs	Disability Adjusted Life Years Lost
EMR	Electronic Medical Record
FBS	Fasting Blood Sugar
GBD	Global Burden of Disease
IDEs	Integrated Development Environments
K-NN	K- Nearest Neighbors
LMIC	Low and Middle Income Countries
ML	Machine Learning
NCD	Non Communicable Disease
RAM	Random Access Memory
RBC	Red Blood Cell
RF	Random Forest
SBP	Systolic Blood Pressure
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
WEKA	Aikato Environment for Knowledge Analysis
WHO	World Health Organization

ABSTRACT

Stroke occurs due to an interruption of supply in oxygen, blood and other nutrients. Identifying and treating stroke is time consuming and expensive specially, in developing countries like Ethiopia. Prediction of stroke risk will help to recognize, detect and treat the disease at early stage and this will reduce (disability, death and cost) that occur from stroke. By addressing the problem at early stage individuals can control their life style and medical status, government can prepare healthcare strategy towards the solution. This will save life, reduce disability and the amount of investment the government dedicate for the disease. By utilizing ML techniques, it is possible to anticipate the onset of stroke with the development of technology in medical sector. ML is a science of feeding computers data and information inorder to make them learn then improve the learning through time. An ideal stroke risk assessment tool that takes into account different risk factors, widely applicable and acceptable does not exist. Stroke has different risk factors including non clinical risk factors like genetic, life style, living area of individuals.

In this study, three machine learning algorithm models are developed for stroke risk prediction. Demographic and diagnosis data from Hallelujah and Zewditu hospitals is used to analyze and come up with stroke risk prediction models. After the business understanding and data understanding phases, data preparation task is done to clean the data from inconsistency, duplication and error then the data becomes ready for the experiment. For predictive model construction, machine learning algorithms such as Logistic Regression, SVM, and Random Forest (RF) Decision Tree with Anaconda python programming was used to conduct all the experiments. Confusion Matrix is used to test the performance of the models. Based on the research findings, the Random Forest (RF) Decision classifier produced an accuracy of 99.3%, SVM an accuracy of 96.63% and Logistic Regression an accuracy of 94%. Therefore, the Random Forest (RF) Decision Tree classifier is proposed for constructing stroke risk prediction model. Based on the proposed optimal model in this study, we recommend future research to integrate the stroke risk prediction model with Health Information System and to use different attributes on addiction of patients' towards Cigarette smoking, drug use, alcohol consumption, which are not included in this study.

Keywords: Stroke Risk; Stroke Risk Prediction; Machine learning; Logistic Regression; SVM; Random Forest

Chapter One

Introduction

1.1 Background

Stroke is Rapidly developing clinical signs of focal and at times global disturbance of cerebral function which is lasting more than 24 hours or leading to death with no apparent cause than that of vascular origin" [1]. According to the World Health Organization (WHO), it is the third leading cause of death in the world.

Stroke occurs as a result of ischemia or haemorrhage. Ischaemic strokes accounts for about 80% of strokes worldwide and may be about 60-80% in Africa and it is caused by thrombosis or embolism resulting in loss of blood supply to part of the brain [2].

About 10-20% of strokes worldwide are caused by haemorrhage. This percentage can reach as high as 40% in Africa may be due to untreated or poorly controlled hypertension in those who take medication. Hemorrhagic stroke happens due to sudden release of blood into the brain commonly due to high blood pressure [2] .

Approximately 85% of death from stroke occurs in low to middle income countries and the burden of the disease from stroke has been increasing in sub-Saharan Africa [1]. The continent of Africa is disproportionately affected by stroke due to population growth, poor and under-developed healthcare systems, unchecked industrialization, and the increased adoption of Western diets [3]. And these have significant impact to increase major risk factors of stroke like hypertension, diabetes and obesity.

Stroke will impact in daily-life such as memories, movement, vision, speech, and literal ability. Identifying stroke is tedious and time-consuming for medical practitioners and Period of treatment in stroke patients depends on symptom and damage of organs. It is beneficial if machine learning method used to predict stroke disease to reduce amount of risk patients have before the initial disease. And the prediction result is helpful for prevention and early treatment. Applying machine

learning techniques to the health-risk assessment problem will be a possible approach to have more accurate predictions.

Different industries utilized Machine-learning technologies and predictive analytics for decades [4]. The healthcare sector has begun adopting these technologies and different papers published over the years and have tried to develop to predict risks of various disease , the healthcare sector adopting these technologies for a variety of applications such as staffing prediction, chronic disease management and population health risk assessment [4]. Therefore, applying machine learning techniques to stroke-risk assessment problem is a possible approach to have prediction of stroke risk in patients.

The most common barriers that makes healthcare access limited in Ethiopia are popularity of traditional healers and transport problem [3]. According to [3] in South Africa 85% of patients believed that their symptoms were not serious or they self-resolve and due to this they delay in medical attention. Patients believed that their symptoms were not serious or they self-resolve and due to this they delays in medical attention. In Ethiopia, stroke is the most common neurological condition in patients admitted to hospitals and is associated with significant morbidity and mortality [3]. Care and treatment of stroke is poor in Ethiopia so that the prediction result will help patients to understand about their major risk factors and follow-up their medical condition. It has been estimated that 68% of adult Ethiopians in the country's capital city, Addis Ababa, have one or more of the following CVD risk factors: daily smoking, regular khat chewing, binge drinking, obesity, abdominal obesity, physical inactivity, or high blood pressure [3].

Therefore, applying machine learning technique for stroke risk prediction will help to reduce the impact and treat the disease early. Systems to make predictions on the possibility that a patient might need medical help in the near feature will also help medical practitioners to prevent and detect stroke early [5].

1.2 Statement of the Problem.

Stroke impacts memories, movement, vision, speech, and literal ability. Identifying stroke is boring and time-consuming for medical practitioners. In Ethiopia, stroke is the most common neurological condition in patients admitted to hospitals and is associated with significant morbidity and mortality [3]. According to a study [3] in South Africa 85% of stroke patients delays in

medical attention because almost half of them believed that their symptoms were not serious or they self-resolve. So that identifying risk factors of stroke and predict stroke risk of individuals enables patients to understand their situation and make them alert for any symptoms and follow up their medical condition closely.

System for predicting stroke risk from demographic and diagnosis data of patients is very important to show how much a patient exposed to the disease. This help to detect and treat stroke early, which is emerging in developing countries like Ethiopia. This also helps to reduce the investment of the government by preventing and treating stroke at early stage.

There are existing studies regarding risk prediction in stroke patients [6] [7] [8] [9] . However, researchers concentrated on identifying major risk factors affecting stroke in Ethiopian context. As per the researcher knowledge there is no study conducted in Ethiopian context to predict stroke risk. Therefore, the aim of this study is to develop and identify stroke risk prediction model using machine learning algorithm. Since stroke risk factors depend on lifestyle, genetic and living area of individual, we cannot use the result of prediction made for foreign countries. Hence, there is a need to identify risk factors and construct a model for predicting stroke.

Therefore, it is necessary to create risk prediction model for stroke using diagnosis data along with demographic data.

1.3 Research Questions

This study tries to answer the following research questions:

- What are the attributes to be used for determining stroke risk?
- Which classification algorithm is suitable for accurate prediction of stroke risk?
- To what extent the proposed predictive model works?
- What are the three major risk factors for stroke risk occurrence?

1.4 Objective of the Study

This research has the following general and specific objectives.

1.4.1 General Objective

The **general objective** of the study is to develop a predictive model that determines stroke risk.

1.4.2 Specific Objectives

The following are specific objectives that are addressed to achieve the general objective of the study:

- To review related literature to identify suitable methods and techniques.
- To capture and prepare data for model construction.
- To identify attributes or variables that are used to predict stroke risk of individuals.
- To select appropriate machine learning algorithms to build stroke risk predictive model.
- To evaluate the predictive model for its relevance and accuracy.
- To identify major risk factors from the attributes used for stroke prediction.

1.5 Scope and Limitation of the Study

The main focus of this research is investigating different classification algorithms to design and develop stroke risk prediction model that can predict individual risk to have stroke in the near feature. Different classification algorithms are investigated and suitable techniques used. This is followed by creating classification models on training data sets and test to check the performance of the system.

Since there is no organized patient data for training and testing, in this research limited number of datasets is used.

One of the attributes selected during data collection task is addiction of patients' towards smoking cigarette but we cannot find history of smoking on stroke patients records/cards. Due to the fact

that smoking cigarette status of patients was not found on patients' record/card, smoking attribute is dropped from the data set.

1.6 Significance of the Study

This study has significant contribution for early detection and treatment of stroke. Overall this study will decrease the mortality rate and damage caused by stroke. In addition, it identifies risk factors that help medical practitioners to control and treat the disease early and individuals to change and follow up their medical status. It also provides significance contribution by leading government future plan to prevent the disease and to change lifestyle of the society.

Finally, the study will contribute for the knowledge building of constructing model for stroke risk prediction from local medical data, from built models, which model better suit for stroke prediction using the selected attributes. Which attributes are major risk factors for stroke risk prediction in patients and it will also contribute for opening research ideas for new researchers to implement other algorithms, using different dataset in different hospitals located in different parts of Ethiopia with additional attributes about patients' addiction and lifestyle.

1.7. Methodology of the study

This study follows experimental research. Experimental research is a research conducted with a scientific approach using two sets of variables. The first sets of attributes acts as a constant, which is used to measure the differences in class attribute.
Towards conducting an extensive experiment, we undertake the following three steps: data collection and preparation, implementation, evaluation.

1.7.1 Data collection and preparation

Before proceeding to data collection, thesis proposal has been submitted to Addis Ababa Health Bureau, and based on the guidance, support of the office this thesis has got Ethical Clearance from the office, which is mandatory to collect data and conduct health researches. The office provided recommendation letter for data collection for the two hospitals, based on which the data collection has been done.

All the data for this work is collected from Hallelujah Hospital and Zewditu Hospital, in Addis Ababa, Ethiopia. The data that is collected from Hallelujah hospital is extracted from Electronic Health Record of the hospital, and the data from Zewditu hospital is manually collected from patient cards. Demographic data like age, weight, height, sex, geographical location, life style (alcohol consumption, smoking, Drug use) diagnosis data like cholesterol level, heartbeat, Glucose level, RBC, blood pressure and stroke status which is labeled as stroke or non-stroke.

To assure the relevance of the dataset, the data attributes compared with different standard datasets in the field and advice from professional neurologists from both hospitals has been considered in order to get the context specific relevant data for accurate prediction. Once the data was collected, it has been preprocessed to keep the quality of the data for constructing model for stroke risk prediction.

1.7.2 Development tools

The Development tool that is used in this research is Anaconda program using Python. Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.), that aims to simplify package management and deployment[10]. Installing Anaconda gives access to different environments that allow to code in either Python or R. These environments, platforms or apps called Integrated Development Environments (IDEs) which greatly ease the development of code.

Visualization and interact with data through visual drill-down capabilities and dashboards is the best way to manage modern and huge dataset[11]. And Anaconda Python has better visualization that allows users to gain insight into the data. Python is an object-oriented, interpreted, mid-level programming language that is easy to learn and use while being versatile enough to tackle a variety of tasks [12]. Python gives an opportunity to code to any individual with its cross platform compatibility (Mac, Windows, Linux, Ubuntu) and it is free to use.

1.7.3 Evaluation Method

There are various methods to measure the accuracy of models: Lift Charts & Gain Charts, ROC Curve, Gini coefficient, Cross Validation, Root Mean Squared Error and Confusion Matrix. Using the right metric can have more influence on your model performance than the algorithm you use [13]. In this study confusion matrix method is used to test performance of the models. Confusion Matrix shows accuracy, true positive, false positive, Sensitivity & specificity of the model. Additionally, it has a table showing the number of predictions for each class compared to the number of instances that actually belong to each class that helps to get an overview of the types of mistakes the algorithm made.

1.8 Organization of the Study

This study is organized into five chapters.

Chapter One is an introduction part which covers background, statements of the problem, Objective of the study, scope and limitations of the study, significance of the study and description about the method of the research.

Chapter Two is a review of literature and related works. Moreover, the review of literatures, books, journals, publications and researches on stroke, machine learning and researches that has been done on stroke risk prediction using machine learning algorithms.

Chapter Three present Data preparation tasks including business understanding, data understanding, data preprocessing and data conversion steps.

Chapter Four presents the proposed architecture and how the data is prepared before the experiment by explaining Data preprocessing tasks. In addition, it presents how the training and testing dataset split, the predictive models experimented using the proposed algorithms, comparison of the results of the models, identification of major risk factor and finally discussion of the result is presented.

Finally, Chapter Five presents conclusion and recommendations of future works.

Chapter Two

Literature Review

This chapter includes an overview of literatures that are used as guides for this work. The overview covers the concept of stroke, machine learning, machine learning algorithms and few related works that have been done for stroke risk prediction and stroke in general.

2.1 Overview of Stroke

Stroke is an abnormality of vascular system in the brain affecting neurologically such as muscle weakness, numbness, and probably mortality.
According to the World Health Organization (WHO), Stroke is "a global disturbance of cerebral function lasting more than 24 hours or leading to death without apparent cause that of vascular origin"

Stroke can be classified into two types as ischemic stroke and hemorrhagic strokes. Stroke is caused by interruption of blood supply to the brain cells, which damages and results brain cells death. Blood flow interruption may be caused by either a clot in the blood vessel or rupture in a blood vessel. Stroke caused due to a clot in the blood vessel is called Ischemic stroke [7] and the type of stroke due to a rupture of blood vessel is Hemorrhagic stroke.
Different risk factors increase the risk of stroke in general. Lifestyle risk factors include diet, cigarette smoking habits, overweight and obesity, physical inactivity, alcohol consumption, family and genetic factors, age, sex, drug use, race, oral contraceptive use, geographic location, season, climate and socioeconomic factors whereas medical conditions include Atrial fibrillation, Blood pressure[2].

It has been estimated that 68% of adult Ethiopians in the country's capital city, Addis Ababa, have one or more of the following CVD risk factors: daily smoking, regular khat chewing, binge drinking, obesity, abdominal obesity, physical inactivity, or high blood pressure [2].

2.1.1 Haemorrhagic stroke

There are two types of haemorrhagic stroke: one resulting from intracerebral haemorrhage secondary to hypertension, cerebral amyloid angiopathy, or degenerative arterial disease; and the other secondary to subarachnoid haemorrhages caused by rupture of an aneurysm [14]. The major risk factors for haemorrhagic stroke are advanced age, heavy alcohol consumption and hypertension. And cocaine abuse is an important cause of cerebral haemorrhage in young people. Focal neurological symptoms, vomiting, drowsiness, stiff neck and seizures are uncommon but headache may be present.Large haemorrhages may cause stupor or coma. Most sub-arachnoid haemorrhages appear suddenly with intense headache, vomiting and neurological deficit and altered consciousness may occur in about 50% of patients. Occasionally, number of prodromal neurological symptoms occurs before a haemorrhage from an enlarging aneurysm causing pressure on the surrounding tissue or as a result of a leak of blood into the subarachnoid space ("Warning Leaks") symptoms such as paralysis of a limb, difficulty in speaking, visual impairment and sudden unexplained headache.

2.1.2 Ischaemic stroke

Neurological symptoms and signs of an ischaemic stroke usually appear suddenly, but less frequently, they occur in a progressive manner that's why it is also called stroke-in-progress [14]. The signs of Ischaemic stroke depends on the location of the occlusion and the extent of the collateral flow but the typical presentation is sudden onset of hemiparesis in an older person. Atherosclerotic ischaemic stroke occurs without warning in more than 80% of the cases and it is more common in the older people. Ischemic stroke accounts for around 80% of all strokes and its incidence is accelerating in developing countries due to unhealthy lifestyles [7].

2.2 Machine Learning

Machine Learning is the process of getting computers learn by feeding data and information in the form of real-world interactions and observation, and then improve their learning over time [15]. As it is scientific and mature modeling method Machine Learning (ML) is attraction more attention than traditional modeling approaches such as Cox Proportional hazard model [16]. The two widely adopted machine learning methods are Supervised and Unsupervised learning but machine learning has other models called semi supervised and reinforcement learning too.

2.2.1 Supervised learning

In Supervised Learning, algorithms learn from labeled data. After understanding the data, the algorithm determines which label should be given to new data based on pattern and associating the patterns to the unlabeled new data [9]. Algorithms trained using labeled data which means the algorithm receives inputs with their corresponding outputs, and the algorithm find a method to determine how to arrive at those inputs and outputs. The algorithm makes predictions and corrected by operator this process is going to be repeated until the algorithm reaches high level of performance. It is used for applications where historical data predicts likely future of events. The supervised learning can be used with methods such as prediction, regression and classification.

2.2.2 Unsupervised learning

Algorithm learns without predefined historical data. It structure and sort the data or learns according to certain characteristics of the data set. The goal is to find some structure within the data by exploring it and the algorithm analyze and organize the data in a way that describes its structure. The unsupervised learning technique works on transactional data very well. It can be used with methods clustering and dimension reduction.

2.2.3 Semi supervised learning

Semi-supervised learning is a combination of supervised and unsupervised learning techniques and used for the same applications as supervised learning. But semi supervised learning uses both

labeled and unlabeled data. It uses a large amount of unlabeled data with a small amount of labeled data because unlabeled data is less expensive. By using the data it learns to label unlabeled data. The semi supervised learning can be used with methods such as regression, classification and prediction.

2.2.4 Reinforcement learning

The algorithm is provided with a set of actions, parameters and end values [17].In reinforcement learning the algorithm learns from trial and error by identifying which actions yield greatest values. By learning from past experiences it begins to adapt new approach based on the situation and the goal in reinforcement learning is to learn the best policy to achieve the best result.

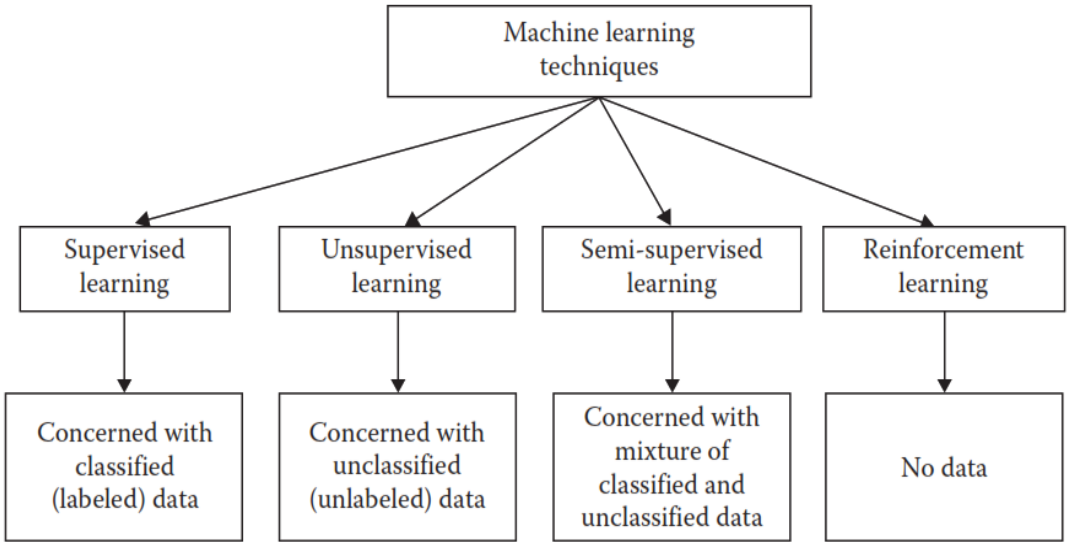

Figure 2.1: Different machine learning techniques and their required data [18]

2.3 Machine Learning Algorithms

Machine Learning Algorithm is a type of method that used data to create models automatically [19]. These algorithms can be used for prediction, classification or clustering tasks. Over the period of time many techniques and methodologies were developed for machine learning tasks[20]. In this study three machine learning algorithms namely Logistic Regression, Support Vector Machine (SVM) and RF Decision Tree has been used.

2.3.1 Logistic Regression

Logistic regression use previous data provided to estimate a probability of an event occurring in the feature [17]. Logistic regression is classification algorithm used when dependent or target variable has two values (0/1, yes/no, true/false) based on the given of independent variables. In many medical data classification tasks Logistic regression and artificial neural networks are the models of choice [21].

There are two models of logistic regression, the first one is binary logistic regression and the second one is multinomial logistic regression. **Binary logistic regression** used when the dependent variable is dichotomous and the independent variables are either continuous or categorical, when the dependent variables are either continuous or categorical. The relationship between one nominal dependent variable and one or more independent variables is explained by **Multinomial regression**. A multinomial logistic regression can be employed when the dependent variable is comprised of more than two categories and dichotomous [22].

A simple logistic function:

$$Y = \frac{e^x}{1 + e^x} = \frac{1}{1 + e^{-x}}$$

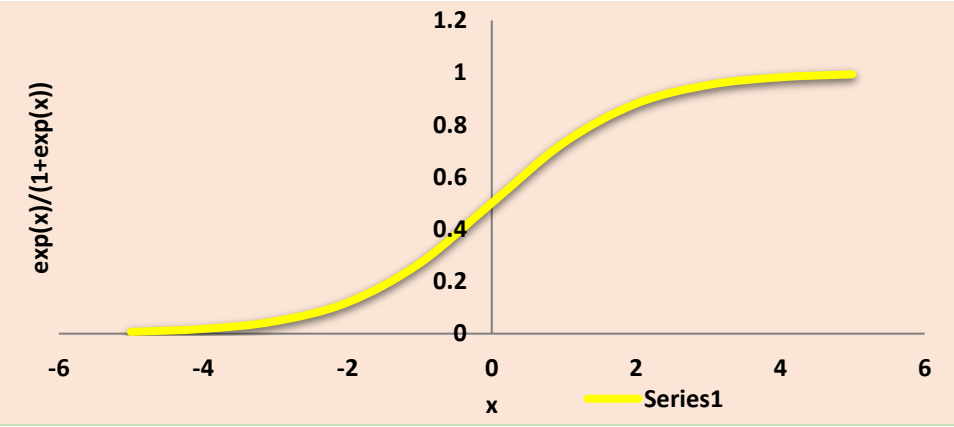

Figure 2.2: Graph of logistic curve where a = 0 and b = 1.

To provide flexibility, the logistic function can be extended to the form:

$$Y = \frac{e^{a+\beta x}}{1 + e^{a+\beta x}} = \frac{1}{1 + e^{-(a+\beta x)}}$$

Where **a** and **b** determine the logistic intercept and slope.

There are different advantages of logistic regression: it is easily extendable to multiple classes (multinomial regression), quick to train, very fast at classifying unknown records, resistant to overfitting and it has good accuracy for many simple datasets. Currently, logistic regression and artificial neural networks are the most widely used models in biomedicine, 28,500 for logistic regression, 1100 for decision trees, 8500 for neural networks, 1300 for k-nearest neighbors, and 100 for support vector machines when measured by number of publications indexed in MEDLINE, which is the National Library of Medicine's premier bibliographic database that contains more than 29 million references to journal articles [21].

2.3.2 Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional (N—the number of features) space that distinctly classifies the data points [23]. Hyperplanes are decision boundaries that help to classify the data points. The dimension of the hyperplane depends up on the number of input features, when the number of input features is 2 the hyperplane is a line, when the number is 3 the hyperplane is two-dimensional plane but the dimension of the hyperplane become difficult when the number of input features exceeds 3. From many possible hyperplanes to separate classes, the objective is to find the hyperplane that has the maximum margin. Future data points can be classified with more confidence and by maximizing the margin distance which provides some reinforcement [23].

For classification problems, SVM tries to find a maximal margin hyperplane that separates two classes[24]. There are many possible hyperplanes that could be chosen to separate two classes of data points. The main objective is to find a hyperplane that separate the two data points with maximum margin distance. Many researchers have used support vector machine (SVM) in neuromuscular disorder diagnosis [7].

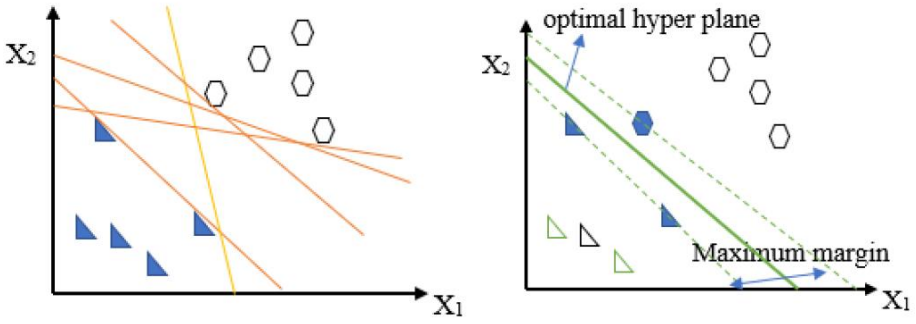

Figure 2.3 Graph of SVM hyperplanes that separate two classes

2.3.3 Decision tree

Inductive learning is a well-known data mining approach to acquire knowledge automatically, where decision tree is one kind of inductive learning algorithms that offers an efficient and practical method for generalizing classification rules from previous concrete cases that already solved by domain experts. Its goal is to discover knowledge that comprehensible to users and with a high predictive accuracy [25]. Decision Tree is a popular method for data mining in creating model initiated by finding the best attribute related with class is a root node, after that, finding next attribute for branches [8].

Decision trees have the advantage that they are not black-box models, but can easily be expressed as rules and in many application domains, this advantage weighs more heavily than the drawbacks, so that these models are widely used in medicine[21].

Decision tree builds classification or regression models in the form of a tree structure, it incrementally develop decision tree associated with the data while it breaks down a data set into smaller and smaller subsets at the same time [26].

The final result is a tree with decision nodes and leaf nodes where a decision node has two or more branches while a leaf node represents a classification or decision. The topmost decision node in a tree, which is best predictor is called root node.

There are two types of decision tree based on the type of target: Categorical Variable Decision Tree which has categorical target variable and Continuous Variable Decision Tree: which has continuous target variable [27].

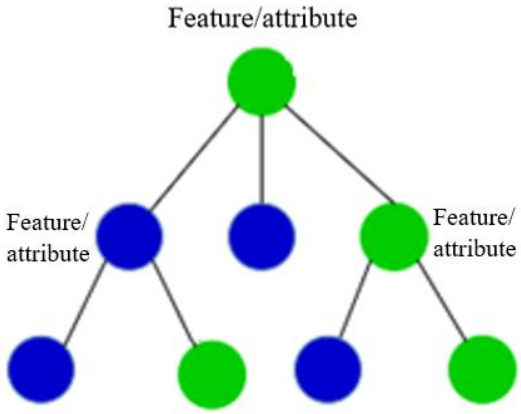

Figure 2.4: Decision Tree

Based on a set of features or attributes present in the data a decision tree makes a series of decisions. The features/attributes and conditions can change based on the data and complexity of the problem but the overall idea remains the same.

Decision trees have different advantages:

- Easy to Understand the output from decision tree is very easy to understand and it doesn't need any statistical knowledge to relate hypothesis[27].
- Useful in Data exploration: Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables [27].
- Less data cleaning required: Decision tree requires less data cleaning and it is not influenced by missing values to a fair degree.
- Data type is not a constraint: Decision tree can handle both categorical and numerical variables.
- Non-Prametric Method: Decision tree Decision tree has no assumption about space distribution and classifier structure as a result it is called a non-parametric method [27].

Disadvantages of Decision trees:

- Over fitting: it is one of the most practical difficulty this problem gets solved by setting constraints on model parameters and pruning.
- Not fit for continuous variables: Decision tree looses information when it categorizes variables in different categories when it works on continuous numeric variables.

2.3.3.1 Random Forest

Random forest is a collection of decision trees and there are a lot of differences in their behavior. Random Forest has a significant performance improvement when it is compared to single tree classifier like C4.5 [28]. Bagging is a principle used by Random forest. Bagging also known as bootstrap aggregation is ensemble technique used by random forest, which chooses a random sample from the data set. Hence each model is generated from the samples (bootstrap samples) provided by the original data with replacement known as row sampling, it is a step of row sampling with replacement called bootstrap. Each model is trained independently and generates results. Aggregation is the process that involves combining all results and generating output based on majority voting.

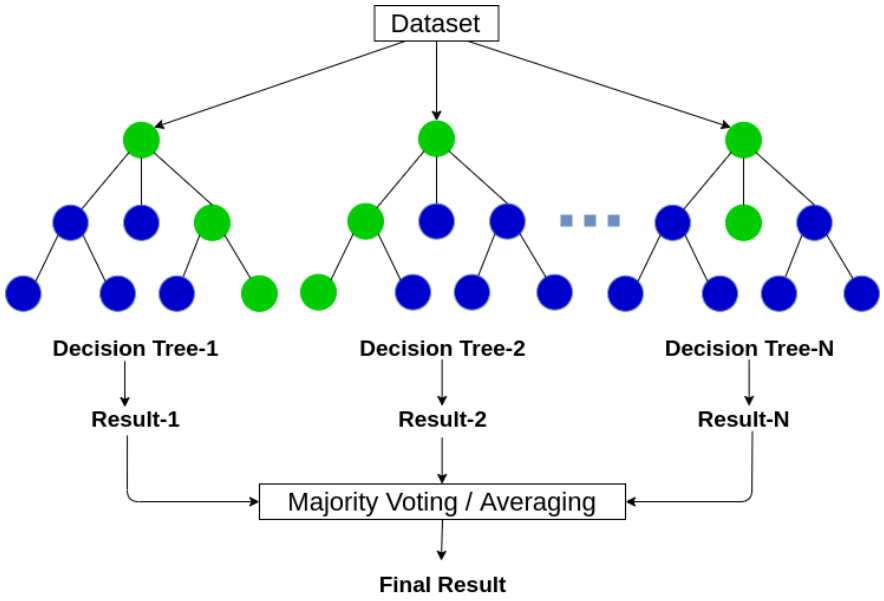

Figure 2.5: Random Forest Decision Tree

RF Decision Tree is a forest of **randomly created decision trees** and each node in the decision tree works on a random subset of features to calculate the output. The random forest then combines the output of individual decision trees to generate the final output by using majority voting.

Important Features of Random Forest

- **Diversity-** Each tree uses different attributes while making an individual tree not all attributes are used in a single tree.
- **Immune to the curse of dimensionality-** Each tree does not consider all the features so that the feature space is reduced.
- **Parallelization-** We can make full use of the processor to build random forests by creating each tree independently out of different data and attributes.
- **Train-Test split-** In a random forest we don't have to segregate the data for trail and test as there will always be 30% of the data which is not seen by the decision tree.
- **Stability-** The result is based on majority voting or averaging so stability arises.

Decision trees	Random Forest
1. Decision trees normally suffer from the problem of overfitting if it's allowed to grow without any control.	1. Random forests are created from subsets of data and the final output is based on average or majority ranking and hence the problem of overfitting is taken care of.
2. A single decision tree is faster in computation.	2. It is comparatively slower but hyperparameters can be used to make the model faster.
3. When a data set with features is taken as input by a decision tree it will formulate some set of rules to do prediction.	3. Random forest randomly selects observations, builds a decision tree and the average result is taken. It doesn't use any set of formulas.

Table 2.1: The difference Between Decision Tree & Random Forest [28]

Random forests use hype parameters either to make the model faster or to enhance predictive power and performance of the models. If the trees are diverse and acceptable, random forests are much more successful compared to decision trees. In the cases of large numbers of instances, Random Forest achieves increased classification performance and yields results that are accurate and precise compared to single decision tree. It also overcome the missing values problem and over-fitting problem generated due to missing values in the datasets. Therefore, if one has to choose from tree based classifiers set for classification problems, Random Forest is recommended for a variety of classification problems with confidence [28].

2.4 Related works

Teerapat [8] used demographic data of patients to predict stroke disease and it was found that Decision Tree was more accurate, compared with other two machine learning classification algorithms, such as Naïve Bays, and Neural Network , but in the aspect of safety of life, Neural Network was the best method by accuracy. Decision Tree was proved the best accuracy and lowest FP, by which low FP rate means high accuracy in predicting the patients were stroke but really were non-stroke, while FN predicts incorrect but patient was actually stroke. Decision tree gives an accuracy of 0.75, Naïve Bayes-0.72 and Neural Network-0.74 respectively. FN is the worst case because it might lead to mortality since the patient is really stroke but the prediction is opposite.

In this study, the best method is Decision tree with 0.75 accuracy but in aspects of life, Neural Network is the best model with highest in FP and lowest in FN value. Finally, they have proposed that further studies should use the diagnosis data together with demographic data of patients.

Jeena R S [7] objective of their study was to develop a machine learning based approach to predict the possibility of stroke in people having the symptoms or risk factors of stroke. The dataset for their study is taken from International Stroke trial Database. Database includes patient information, patient history, hospital details, risk factors and symptoms. 350 samples were taken and they have used (Age, Sex, Walking symptoms, Atrial Fibrillation, Face deficit, Arm / Hand deficit, Leg/ Foot deficit, Infarct visible on CT, Dyphasia, Hemianopia, Visuospatial disorder and Cerebellar signs) attributes .

Their study uses support vector machine with different kernel functions and have been applied on 300 training samples and tested with 50 samples. Polynomial, quadratic, radial basis and linear functions are applied and all give different accuracy. The study use the parameters sensitivity, specificity, accuracy, precision and F1 score to evaluate the performance of various kernel functions of SVM classifier. SVM has been implemented in MATLAB and with different kernel functions. The appropriate choice of kernel function for detection of stroke has been investigated and different accuracy level is achieved in each kernel function: linear-91%, Quadratic, 81%, RBF-59% and Polynomial-87%. From the functions accuracy of the linear kernel function is a better choice due to better accuracy level. Linear function provides a greater accuracy of 91 % and they have proposed it to be used as a standard function in SVM for stroke prediction.

Their work can aid physicians to plan for better medication and can provide the patient with early diagnosis of stroke .This work demonstrates the predictive power of SVM with a small set of input parameters. They have proposed to extend the method to large database by considering more input attributes so as to improve the system performance.

E. Dritsas [29] have developed several models and evaluated to design a robust framework for the long- term risk prediction of stroke occurrence. And they have used Naïve Bayed, Random Forest, Logistic Regression, K-Nearest Neighbors, Stochastic Gradient Descent, Decision Tree, Multilayer Perception, Majority Voting and Stacking Machine learning algorithms. The study used a dataset from Kaggle. From the dataset, they have focused on participants who are over 18 years old. The number of participants was 3254 with ten attributes that are used as input to ML models and one target class. The attributes used are: Age (years) , Gender , Hypertension, Heart_disease , Ever married , Work type, (private 65.02%, self-employed 19.21%, govt_job 15.67% and never_worked 0.1%), Residence ,urban , rural), Avg glucose level, BMI , Smoking Status (smoke, never smoked and formerly smoked) and target class Stroke.

WEKA 3.8.6 environment is used to evaluate the Machine Learning models performance. The other two models are outperformed by stacking classification which achieved F-measure, precision and recall of 97.4% , AUC of 98.9% and an accuracy of 98%. Similarly, RF and majority voting classifiers achieved high values. Focusing on the AUC metric, the stacking and RF models have

approximately similar discrimination abilities, which show that, with a high probability of 98.9% and 98.6%, respectively, both models can successfully identify the stroke from the non-stroke instances. Besides stacking and RF, the K-NN model is the next one, with an essentially high AUC equal to 94.3%. And finally they proposed future purpose of the study is to enhance the ML framework via the employment of deep learning methods and to use brain CT scans and to evaluate the predictive ability of deep learning as a challenging but promising direction.

Le Zheng [9] objective of their study is to propose and validate a risk model predictive of stroke in future I year targeting at patients in Maine. They have used a dataset derived from Electronic Medical Records (EMR) and clinical notes provided by Health Information Exchange (HIE). Natural language processing (NLP) techniques were applied to collect clinical histories from notes. A retrospective cohort of 180,196 patients and a prospective cohort of 347,504 patients were constructed for model development and validation, respectively. Initially, around 30,000 demographic and clinical features were extracted from EMR and notes. They have used Natural language processing (NLP) techniques to collect clinical histories from notes. And 26 features having significant p-values «0.05) by multivariate analysis were selected for model development. These features included 19 diagnoses, 2 prescription medications, counts of emergency department visits, costs and chronic diseases, and 5 NLP features. Retrospective cohort was randomized into training, calibration, and blind-testing sub cohorts, with the ratio of patients with stroke to those without stroke maintained at the same level in each sub cohort. A logistic regression model was built with the training sub cohort. The output of the model was a risk score (ranging between 0 and 1) describing the probability of having stroke in future 1 year. They choose threshold so that both positive predictive value (PPV) and sensitivity reached at acceptable levels. The model was then validated with the blind-testing sub-cohort, and tested on the prospective cohort. The c-statistics for the retrospective and prospective predictions were 0.892 and 0.887, respectively. At a PPV of 0.262, the model correctly identified 4l.0% (3,028 of 7,387) of prospective patients who had stroke in future 1 year. Such prospective performance highlights the effectiveness of the model in identifying population having stroke over a large, independent cohort. They have also tested performance of the Framingham study model with their prospective cohort. A c-statistics of 0.836 was achieved. Furthermore, their model was applied to patients across all age, all payor, and all disease groups, while the Framingham model was applied to patients with age 54+.

They have proposed Implementation of their model onto a real-time monitoring platform of statewide population can provide healthcare providers with early warnings of health status of population, which benefits timely administration of population with chronic conditions.

Sultan et al [1] have conducted a cross-sectional study in the adult emergency center (EC) of an urban university hospital in Addis Ababa, Ethiopia, from August 2015 to January 2016. They conducted a cross-sectional study in the adult emergency center (EC) of an urban university hospital in Addis Ababa, Ethiopia, from August 2015 to January 2016. The study site is a tertiary referral hospital with neurological and neurosurgical expertise. The investigator screened all patients presenting to the adult EC with stroke-like symptoms or altered level of consciousness for eligibility at the triage. All patients with CT confirmed diagnosis of acute stroke were also included. The researchers excluded patients with post-traumatic neurologic deficits or new stroke symptoms in the context of central nervous system lesions due to malignancy or infection.

They have reviewed emergency center clinical records for documented physical examination findings and results of laboratory investigations and radiographic examinations. The researchers sample size calculation of 114 was based on a similar study conducted by Chalachew et al. [69]. They have performed analysis using SPSS version 20 and used descriptive statistics as well as multivariable logistic regression models to evaluate associations between stroke types and stroke risk factors, and between delayed presentation and clinical indicators.

In this study, 104 patients are included from those total patients 58 or 56% are male and the minimum age is 53 years. Patient demographics age, sex educational status, occupation and mode of arrival are summarized and from clinical patient data Hypertension, cardiac illness, diabetes, ischaemic heart disease and hypertensive heart, previous history of stroke, HIV, Smoking, TIA and family history of stroke attributes are used. The most common stroke risk factor is hypertension with 49%, the second risk factor is cardiac illness with 20% and the third is diabetes with 11% importance. Evaluating predictors of delayed presentation in logistic regression, no individual predictor found or no statistically signification combination predictors found (omnibus likelihood ratio test p=0.91). Different symptoms are presented including loss of consciousness (48%, n=50), facial palsy (43%, n=45), hemiparesis (47%, n=39), hemiplegia (29%, n=28) and

aphasia (33%, n=19). The systolic blood pressure(SBP) of 25 patients is >160mmHg and the median SBP at triage was 140mmHg.

There were high degree of haemorrhagic stroke patients than ischaemic stroke (n-58; 56%) Vs (n=46; 44%). Patients with ischaemic strokes are more likely to have cardiac disease (p=.005). Patients with hypertension were significantly more likely to have haemorrhagic stroke ($p < .001$) and presented with loss of consciousness (p=.01). Cardiac illness with odds ratio of 3.00, p=0.01, 95% confidence interval (1.37, 11.6) is found to be the only risk factor related with ischaemic stroke in multivariable logistic regression model.

In this study, the researchers identified the stroke types, risk factors, and clinical presentations of stroke victims presenting to an urban university emergency center in Ethiopia. While ischaemic strokes account for almost 90% of strokes in the United States, they found that over half of stroke victims in our setting suffered from haemorrhagic strokes. A prior study of stroke in Ethiopia by Abebe et al. found a higher rate of ischaemic stroke, but several other studies in sub-Saharan Africa have found that haemorrhagic strokes are more common [1][2][3] [31][32]. Hypertension, cardiac disease and diabetes were the most common risk factors presenting both types of stroke, which conforms to the results of other studies in LMICs [8,16]. In addition, from patients with one known risk factor, most of them are not receiving appropriate treatment that can reduce the risk of stroke and almost half of the patients had one or more identified stroke risk factors. In general, the major factors that play an important role in shaping patients outcome are system related factors are system and patient related factors [17].

Reliance on natural healers, support systems at home and financial strain, proximity to health care facilities and knowledge of stroke symptoms significance are considered patient related factors.

Different researches has been done on stroke risk prediction and this research use clinical data set along with demographic data. As recommended : Further studies should use the diagnosis together with demographic data of patients in prediction of the results by which, better accuracy or getting new rules can be expected [8].

Researchers have been done on stroke risk prediction by using different dataset, attributes, models and datasets from different location or geographical area. Kansadub et al.[8] Suggested that further

researches can be done by using clinical data along with demographic data of patients these will improve accuracy or enable to drive new rules. A study by Rolon-Mérette et al. [8] explore the predictive power of SVM with small data set but it is recommended to use another models to have better accuracy or get rules. Chao and Wong [23] used clinical data with demographic data to predict stroke risk of individuals across all age and group and the have used logistic regression model for stroke risk prediction. Having different models and compare the results between models will enable to get new rules or better accuracy. In addition, this research used dataset with eleven attributes that are believed high risk factors for stroke in Ethiopian context by consulting neurologist on the field and reviewing health researches in Ethiopia. Most the currently used CVD/stroke prediction algorithms are based on the Framingham study of a primarily white population of North America, which may not be accurate enough for other racial/ethnic groups [33] .

In this study three machine learning models using SVM, Decision Tree and Logistic Regression has been built. Reason for using these three algorithms SVM because of its , Many researchers have used support vector machine (SVM) in neuromuscular disorder diagnosis [7]. Logistic regression because it is easy to implement and Logistic regression and artificial neural networks are the models of choice in many medical data classification tasks [16][21]. Decision Trees has used because Decision trees are very easy to interpret[34]. Decision trees have the advantage that they are not black-box models, but can easily be expressed as rules and in many application domains, this advantage weighs more heavily than the drawbacks, so that these models are widely used in medicine[21].

2.4.1 Summary of Related Works

STROKE RISK PREDICTION STUDIES				
Author (year)	Problem	Approach	Result	Gap
Teerapat [8] (2016)	Used demographic data of patients to predict stroke disease.	Classification	Decision tree gives an accuracy of 0.75, Naïve Bayes-0.72 and Neural Network-0.74	Use the diagnosis data together with demographic data of patients
Jeena R S [7] (2016)	Predict the possibility of stroke in people having the symptoms or risk factors of stroke	Classification	SVM with kernel function linear-91%, Quadratic, 81%, RBF-59% and Polynomial-87% accuracy.	Use large data set
E. Dritsas [29] (2021)	Developing a model for a robust framework for the long- term risk prediction of stroke occurrence	Classification	Stacking classification achieves accuracy of 98%	Enhance accuracy employing deep learning methods and to use brain CT scans
LOCAL STUDIES ABOUT STROKE				
Author (year)	Problem	Approach	Result	Gap

Sultan et al., (2017)	Epidemiology of stroke patients in Tikur Anbessa Specialized Hospital: Emphasizing clinical characteristics of Hemorrhagic Stroke Patients	assess the epidemiology of stroke patients seen in Black Lion Hospital	46.1% of the patients reviewed had hemorrhagic stroke. Hypertension appeared to be predominant, 103/139 (78.3%) male with a higher hemorrhagic stroke rate than females (62% vs. 38%)	Single centered facility based, patients sicker & milder symptoms selected may not represent all types of patients, unable to determine socioeconomic status of patients.
Alemayehu, (2013)	Assessment of Stoke PatientsOccurrence of Unusually High Number of Haemorrhagic Stroke Casesin TASH	assess the clinical characteristics and risk factors among patients presented with stoke	Among the total 55.3% had hemorrhagic and 35.1% had ischemic stroke. 63 (55%) were females. hypertension and Diabetes Mellitus with 69.3 % and 14.9% identified as major risk factors.	large sample size with measurement of the lipid profile of patients needed.
Fekadu et al., (2019)	Risk factors, clinical presentations and predictors of stroke among adult patients admitted to stroke unit of Jimma university medical center	prospective observational study	51.7% of patients had ischemic while 48.3% had hemorrhagic stroke. The most common risk factor hypertension (75.9%) followed by family history (33.6%),	Sample size was small, single center focused and all patients are from southwest part of the country.

			alcohol intake (22.4%), smoking (17.2%) and heart failure (17.2%). males comprised 62.9%.	
Abate, (2021)	The burden of stroke and modifiable risk factors in Ethiopia:	A systemic review and meta-analysis	hemorrhagic and ischemic stroke were 46.42% and 51.40%. Modifiable risk factor of hypertension, alcohol consumption and dyslipidemia were 49%, 24.96%, and 20.99% . 90% of stroke patients had one or more modifiable risk factors.	developing and testing a conceptual model that can use accessibility to screening, treatment or sociocultural aspects.

2.5 Model Evaluation

This section presents the evaluation method that is used to evaluate the performance of the model Stroke Risk Prediction. In order to check the performance of a classification based machine learning model, the confusion matrix is deployed [35]. It is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known[36]. The study used Confusion matrix as evaluation methods because Confusion matrix is a summarized table of the number of correct and incorrect predictions yielded by a classifier (or a classification model) for binary classification tasks [35].

2.5.1 Confusion Matrix

The confusion matrix is a tool for predictive analysis in machine learning [35]. It is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known [36].

Also we can say Confusion matrix is a summarized table of the number of correct and incorrect predictions yielded by a classifier (or a classification model) for binary classification tasks [35].

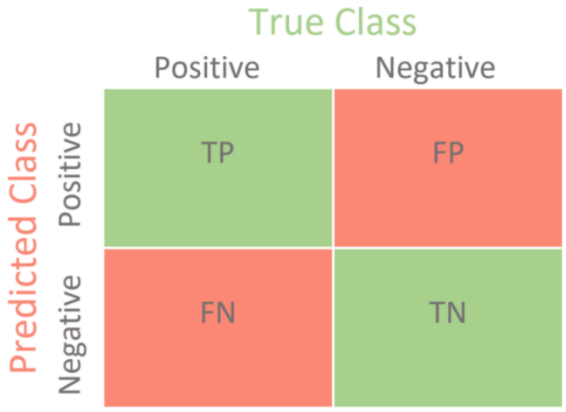

Figure 2.6 Confusion Matrix [35]

Confusion Matrix has four dimensions

- True Positives (TP) − When both actual value & predicted value of data point is 1.
- True Negatives (TN) − When both actual value & predicted value of data point is 0.
- False Positives (FP) also known as Type 1 error− When actual value of data point is 0 & predicted value of data point is 1.
- False Negatives (FN) also known as Type 2 error− When actual value of data point is 1 & predicted value of data point is 0

Accuracy

Accuracy is a measure for how many correct predictions your model made for the complete test dataset [35]. It is a ratio of correctly classified patients (TP+TN) to the total number of patients (TP+TN+FP+FN).

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}$$

Mis classification :

The Misclassification Rate is a performance metric that tells you the fraction of the predictions that were wrong, without distinguishing between positive and negative predictions.

$$\text{Mis classification} = \frac{FP+FN}{TP+TN+FP+FN}$$

Precision

Precision is about how many correctly predicted cases turned out to be positive. And this determine whether the model is reliable or not.

Precision is a useful metric in case where False Positive is a higher concern than False Negative [35].

$$\text{Precision}=\frac{TP}{TP+TN}$$

Recall (sensitivity)

Recall describes that how many of actual positive cases are predicted correctly by the model and it is a useful metric in cases where False Negative trumps False Positive.

$$\text{Recall}= \frac{TP}{TP+FN}$$

A higher recall means that most of the positive cases (TP+FN) will be labeled as positive (TP) and this will lead to a higher number of FP measurement and a lower overall accuracy [35]. And a low recall means a high number of FN which means it should be positive but labeled as negative. This means the model have more certainty if positive case found and this is to be a true positive.

F1-Score

F1-Score is a combined idea about Precision and Recall and it reaches to maximum when Precision is equal to Recall.

$$\text{F1-Score} = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}}$$

The interpretability of F1-Score is poor which means we use in combination with other evaluation metrics that gives us a complete picture otherwise we don't know whether the classifier is maximizing precision or recall.

False Positive rate

False positive rate is a measure for how many results get predicted as positive out of all the negative cases [60].

$$\text{FPR} = \frac{FP}{TN + FP}$$

Chapter Three

Data Preparation

3.1. Overview

In this phase after the sources are completely identified, proper selection, constructing, formatting and cleansing the data involves raising the quality of records to the desired level considering the analysis techniques selected. Data coding, data cleaning, attribute selection, and data transformation steps undertaken in this stage. Finally, the data set becomes ready for the modeling techniques implemented.

3.2 Business understanding

Understanding the business objective requirement from the business perspectives is crucial to convert the data into a new knowledge by applying data mining techniques. So that we have gone in-depth and remark important points that will help us to understand, define and analyze the problem to be addressed in a best way.

3.2.1 Stroke Overview

The most recent Global Burden of Disease (GBD) 2019 stroke burden estimates1 showed that stroke remains the second leading cause of death and the third leading cause of death and disability combined (as expressed by disability-adjusted life-years lost—DALYs) in the world [37]. According to World Stroke Organization (WSO) the estimated global cost of stroke is over US$891 billion and with the bulk of the global stroke burden lower income and lower–middle income countries (LMIC) takes the highest burden (86.0% of deaths and 89.0% of DALYs). Incidence and mortality of stroke differ between countries, geographical regions, and ethnic groups [31].

Healthcare providers, system leaders, governments and the general population need an effort towards implementing widely available and effective prevention strategies, increasing raising awareness, educating population and individuals about their risk factors [37]. The disease by large

can be prevented by making simple changes in the way people live their lives or simply by changing our lifestyle[33]. Ethiopia faces the unenviable threat of a triple burden of disease: infectious diseases, Non-Communicable Diseases (NCDs), and injuries [31]. The age-adjusted death rate of stroke in Ethiopia is 89.82 per 100 000 of the population and The magnitude of stroke-related deaths in the country is 6.23%. Besides, 90% of the burden of stroke is attributable to modifiable risk factors as per previous reports.

Metabolic factors (high blood pressure, obesity, fasting plasma glucose, cardiac disorder, and total cholesterol) accounted for 72% of stroke DALYs, and behavioral factors (smoking, poor diet, and physical inactivity) accounted for 66% [31]. In Ethiopia, a nationwide comprehensive study on stroke burden and its risk factors are lacking and more studies need to be done nationwide as the same time in each region.

Features of stroke:

- Weakness or sudden numbness leg, arm or in the face, especially on one side of the body
- Sudden confusion, trouble speaking, or difficulty understanding speech
- Sudden trouble seeing in one or both eyes
- Dizziness, loss of balance, lack of coordination, sudden trouble in walking.
- Impairment or loss of consciousness
- Arm, leg or face transient weakness, numbness or paralysis of typically on one side of body
- Difficulty in understanding others or garbled speech or transient slurred speech.
- Double vision or transient blindness in one or both eyes.
- Amaurosisfugax (curtain like appearance in front eye)
- Transient dizziness or loss of balance or coordination

3.2.2 Risk Factor of Stroke

For prevention, it is important to identify risk factor for stroke [33]. Some recognized risk factors of stroke are:

1. *Well documented Modifiable Risk Factors:*

Hypertension

Diabetes Mellitus

Dyslipidemia

Obesity and Body fat distribution

Physical inactivity

Tobacco use

Structured cardiac diseases

Atrial fibrillation

Sickle cell disease

Carotid stenosis

Excessive Alcohol consumption

Unhealthy diet and nutrition

2. *Less – well documented Modifiable Risk Factors:*

Migraine

Metabolic Syndrome

Drug Abuse

Obstructive Sleep apnea

Hyperhomocysteinemia

Hypercoagulability

Elevated Lp (a)

Inflammation and Infection

3. *Non-modifiable risk factors:*

Genetic factors

Increasing age

Low birth weight

Race/ethnicity

Low socio-economic status

Male gender

Stroke prevention aims at either by controlling various risk factors that increase the chance of having a stroke or reducing the likelihood of having a stroke by reducing the chances of developing risk factors.

3.2.3 Importance of Stroke Prediction

Stroke is a medical disorder in which the blood arteries in the brain are ruptured, causing damage to the brain [38]. Stroke symptoms develop when the supply of blood, oxygen and other nutrients to the brain interrupted. Prediction of stroke risk will help to recognize, detect and treat the disease in early stage and this will reduce the impact (disability, death and control) of stroke. By addressing the problem at early stage individuals can control their lifestyle and medical status and government can prepare healthcare strategy towards the solution and this will save life and reduce the amount of investment the government dedicates for the disease. Specially in developing countries like Ethiopia identifying and treating stroke is time consuming and expensive. With the development of technology in the medical sector, it is now possible to anticipate the onset of a stroke by utilizing ML techniques [38].

Early recognition and detection of the various warning signs and risk factors can help to reduce the impact of the stroke in general.

Different Machine Learning models have been developed to predict stroke and in this study three models have been built using different machine learning algorithms namely Logistic Regression, Support Vector Machine (SVM) and RF.

An ideal stroke risk assessment tool that is simple, widely applicable and accepted, and takes into account the effects of multiple risk factors does not exist [33]. And on this study demographic and diagnosis data from three different hospitals in Addis will be used to analyze and come up with stroke risk prediction models.

Methods of primary Stroke Prevention are the following [33]:

- Mass (population-wide) strategy
- High Risk Strategy

And the gaps that are identified in primary stroke/ cardiovascular disease (CVD) prevention are Lack of awareness, Under usage of population-wide strategies, False reassurance of low risk , Management of blood pressure , Cost Barrier and **Lack of local stroke /CVD prediction algorithms**: According to [33]. Most the currently used CVD/stroke prediction algorithms based on the Framingham study of a primarily white population of North America, which may not be accurate enough for other racial/ethnic groups.

3.3 Data Understanding

After understanding the business where the problem is located, the next step is to understand and analyze the data that is available for analysis and model building.

3.2.1 Data Source Description

For this study, the data collected from selected two hospitals in Addis Ababa, Ethiopia.
All the data for the study collected from Hallelujah Hospital and Zewditu Hospitals in Addis Ababa, Ethiopia. Demographic & clinical data of patients collected from both hospitals. The data consists 9373 records with twelve attributes with one attribute stroke or not have stroke.

The data collected from Hallelujah hospital is extracted from the health management system which consists over 113,000 patient records until the date 01/12/2014. The data collected from Zewditu hospital is taken from physical patient cards until the date 05/13/2014.
From the attributes selected to use for prediction of stroke smoking status of patient is not registered on patient record /card so that the rest eleven attributes listed below are used for further processing.
The attributes description and categorical values explored for the study described in the below table:

S.No	Attributes	Description
1	Age	Age of the patient
2	Weight	Weight of the patient in kg
3	Height	Height of the patient in meter
4	BMI	Body Mass Index of the patient
5	Sex	Male/ Female
6	Cholesterol level	Blood Cholesterol level of the patient
7	Blood Pressure	Blood pressure of the patient
8	Pulse Rate	Heart Beat Rate of the patient
9	RBC	Red Blood Cell count of the patient
10	FBS	Fasting Blood Sugar of the patient
11	Location	(Where the patient lives or come from)
12	Label	(Stroke or non-Stroke)

Table 3.1: Attributes and description

3.4 Data Preprocessing

Data available for mining is raw data and may be in different formats as it comes from different sources, it may consist of noisy data, irrelevant attributes, missing data etc. Data needs to be preprocessed before applying any kind of data mining algorithm [31].We have applied preprocessing tasks on the data.

3.4.1 Data Integration

Data integration involves removing inconsistencies in attribute names or attribute value between data sets of different sources when the data to be mine comes from different sources.

In this study, the demographic data of patients and diagnosis (clinical) data of patients collected from two hospitals. So integration of the data is used to combine two datasets from different data sources by using the same names of attributes. The data collected from Hallelujah hospital is in excel format and the data collected from Zewditu hospital manually so the integration mainly focus encoding all data from Zewditu Hospital to Microsoft excel file. The attribute height value in both hospitals use both in meter and in centimeter so all values of the attribute converted to meter.

3.4.2 Data Cleaning

Data cleaning involve detecting and correcting errors in the data, filling in missing values, etc.

In this study the data set has been checked from errors and missing values. And to clean the dataset correction of different values has been made based on the source document and also there are missing values that apparently leave blank. And missing values in this dataset has been found **180** in number and it has been decided to drop all records with missing values. And the remaining original dataset has been passed for further processing.

3.4.3 Data Field Selection

Many irrelevant attributes may be present in the data to be mined so that irrelevant attributes need to be removed. And also many mining algorithms don't perform well with large amounts of attributes and that is why feature selection technique is needed to be applied before any data mining algorithm applied. The main objectives of feature selection are to improve the model and avoid overfitting. At this stage the attribute Smoking status values are unknown so that the attribute dropped since it will not have an impact on the outcome of the data.

3.4.4 Method of Data Quality Assurance

The datasets were checked for completeness and correctness of the required attributes and integrity before analysis and prediction. The study explored different non-machine learning studies on stroke risk and stroke risk prediction in addition consultation from neurological surgeons at both hospitals considered. This helps the research to ensure the required attributes are complete and adequate for prediction of Stroke risk.

3.4.5 Data Discretization

Discretization needs to be applied when the data mining algorithm cannot cope with continuous attributes [31]. This step takes only few discrete values to transform continuous attribute into a categorical attribute.

The attribute blood pressure has two values systolic BP and Diastolic BP. Systolic BP used for this study.

3.5. Data conversion

After all the necessary data preprocessing activities taken place as described in the earlier sections, the preprocessed data is loaded to anaconda python application for model building. The dataset is converted to *csv* format that is suitable for python anaconda.

	Age	Weight	Height	BMI	Sex	Cholestr	Blood Pre	Pulse Rate	RBC	FBS	Smoking	Location	Label	Label
2	70	82	1.67	29.40227	MALE	206	140	76	4.75	95	unknown	Afar Zone 01 woreda: Asayta	0	Non Stroke
3	62	69	1.53	29.47584	MALE	208	150	93	5.48	61	unknown	Addis Ababa Bole woreda: 7	0	Non Stroke
4	36	80	1.62	30.48316	MALE	191	110	71	5.55	62	unknown	Addis Ababa Bole woreda: 10	0	Non Stroke
5	76	74	1.57	30.0215	MALE	168	143	101	4.48	113	unknown	Addis Ababa Yeka woreda: 12	0	Non Stroke
6	44	112	1.52	48.47645	FEMALE	184	150	90	6.14	101	unknown	Addis Ababa Addis Ketema woreda: 3	0	Non Stroke
7	38	58.7	1.81	17.91765	FEMALE	179	109	77	5.35	109	unknown	Addis Ababa Kirkos woreda: 9	0	Non Stroke
8	45	81	1.68	28.69898	FEMALE	79	145	80	5.79	99	unknown	Addis Ababa Kirkos woreda: 2	0	Non Stroke
9	85	75	1.76	24.21229	MALE	177	170	73	5.34	89	unknown	Addis Ababa Nifas Silk-Lafto woreda:	0	Non Stroke
10	85	63	1.71	21.54509	MALE	149	130	75	4.55	131	unknown	Addis Ababa Arada woreda: 7	0	Non Stroke
11	40	77.4	1.76	24.98709	MALE	201	149	94	5.88	113	unknown	Addis Ababa Kirkos woreda: 10	0	Non Stroke
12	45	69	1.58	27.6398	FEMALE	234		78	6.34	289	unknown	Addis Ababa Bole woreda: 8	0	Non Stroke
13	47	57	1.56	23.42209	MALE	173	180	78	6.82	102	unknown	Somali Warder woreda: Bokh	1	Stroke
14	66	111	1.75	36.2449	FEMALE	178	100	95	6.94	79	unknown	Addis Ababa Bole woreda: 3	1	Stroke
15	65	80	1.58	32.04615	FEMALE	100	130	64	5.45	105	unknown	Addis Ababa Akaki Kaliti woreda: 1	1	Stroke
16	47	98	1.65	35.99633	MALE	122.5	180	100	5.65	80	unknown	Somali Warder woreda: Bokh	1	Stroke
17	97	64	1.54	26.986	MALE	123	140	69	5.53	84	unknown	Addis Ababa Kolfe Keranio woreda: 1	1	Stroke
18	56	80	1.51	35.08618	MALE	116	144	84	4.73	69	unknown	Addis Ababa Arada woreda: 8	1	Stroke
19	79	73	1.71	24.96495	MALE	251	220	117	5.14	104	unknown	Addis Ababa Bole woreda: 8	1	Stroke
20	66	55	1.62	20.95717	FEMALE	150	100	77	4.97	103	unknown	Addis Ababa Bole woreda: 3	1	Stroke
21	79	92.4	1.56	37.96844	MALE	206	130	106	4.85	104	unknown	Addis Ababa Akaki Kaliti woreda: 4	1	Stroke
22	79	69	1.64	25.65437	FEMALE	288	140	111	5.73	78	unknown	Addis Ababa Arada woreda: 2	1	Stroke
23	65	67	1.59	26.50212	FEMALE	130	130	72	6.29	80	unknown	Addis Ababa Akaki Kaliti woreda: 1	1	Stroke

Table 3.2: Sample dataset in csv

Chapter Four

Experiment and Result Discussion

4.1 Overview

This chapter discusses about how models are constructed and experiment is carried out based on the proposed framework. In addition this experiment and evaluation shows the realization of the architecture presented in chapter three and presents the selected machine learning models Decision Tree, SVM and Logistic regression.

The experiment and interpretation of the results presented then performance of each models evaluated. This experiment and analysis is carried out, trained and tested with anaconda python program on Hp Elitebook 840 G6 laptop, Processor- Intel(R) Core(TM) i5-8265U, CPU-1.60GHZ, 8GB RAM and 250GB hard disk. The operating system used on the hardware is Microsoft Windows 10, 64 bit operating system.

All the preprocessing activities performed on the dataset as described in chapter three and some of the major tasks performed are presented here. This section present the realization of the framework architecture and focuses on presenting summary of the major experiments made in the process of arriving at the optimal model to achieve the objective set in chapter One.

Series of Experiments are conducted based on which algorithms prediction model with varying accuracies, sizes and precisions are obtained. This section also presents several activities done during development including evaluating the models built, selecting the best model and providing explanation of the selected model.

4.2 The Proposed Architecture

The proposed architecture in this research is illustrated in Figure 4.1.The proposed architecture shows the steps followed in the Prediction of stroke risk of individuals from the dataset obtained.

In the proposed architecture, patient data taken as input and stored in the dataset and this dataset contains patients with stroke case and patients without stroke case. To validate the proposed model, we use patient dataset.

The architecture passes through different stages after business understanding that has described on chapter three. After understanding the business from the business perspective, the first task is obtaining ethical clearance from AA Health Bureau after that, data collection is taken place on both Hallelujah and Zewditu Hospitals. The data collected from the two hospitals is merged into one data set, then the dataset (patient data) gone through preprocessing tasks to clean attributes in the dataset. And then the dataset is splitted into test dataset which is 30% of the data set which is going to be used to evaluate the performance of the system and training data set 70%of the dataset used to train the model. After splitting the dataset the training dataset (70%) goes through classification models namely Decision tree, SVM and Logistic regression to train the model and obtain the trained model. The test dataset (30%) directly pass to the stroke prediction model obtained. Finally, the Stroke Risk Prediction Model predicts Stroke Risk or Not.

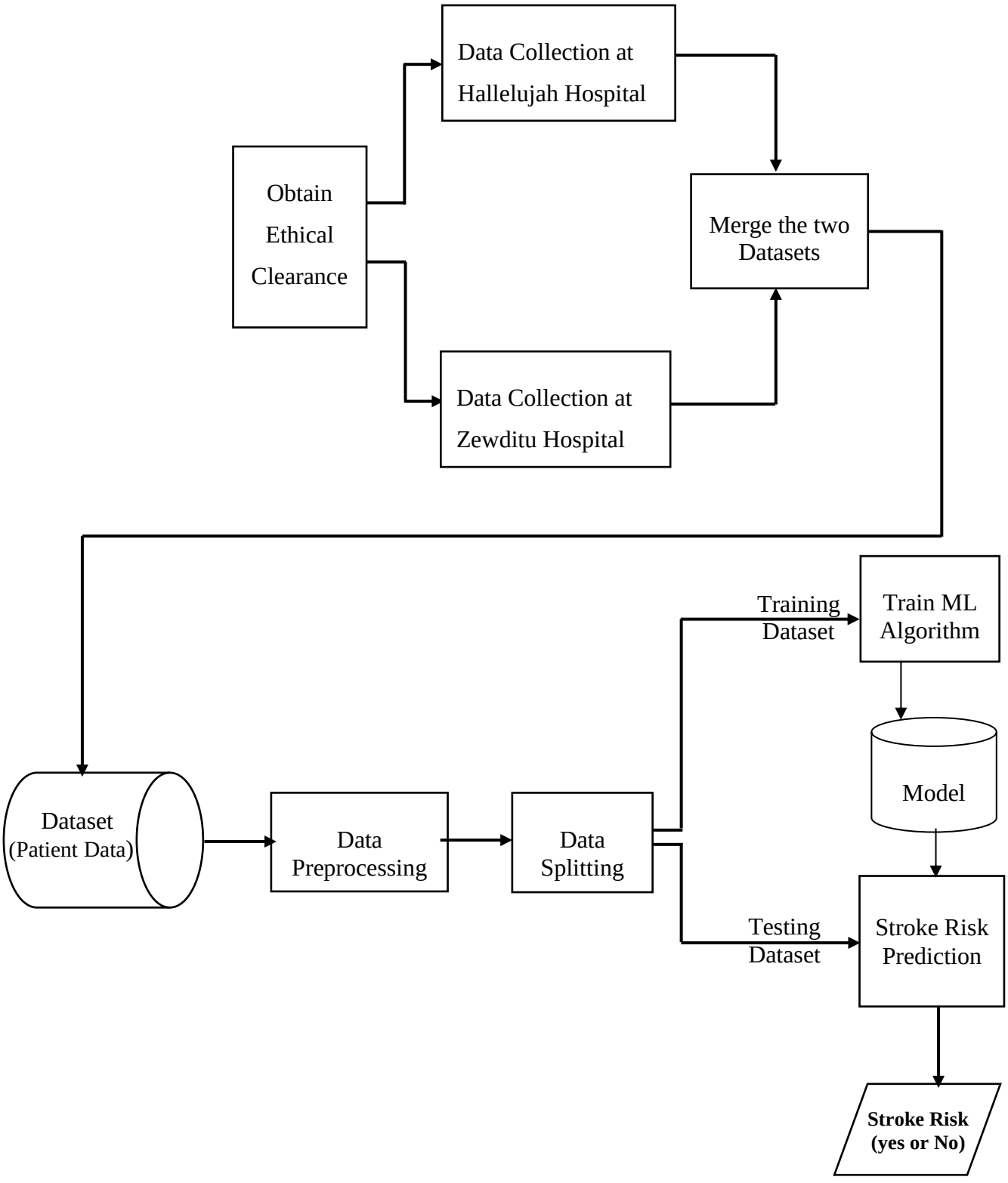

Figure 4.1: The proposed architecture for Stroke Risk Prediction Model

4.3 Dataset for Experiment

For this study, the data collected from Hallelujah Hospital and Zewditu Hospital. The data from Hallelujah and Zewditu hospitals has both patients' demographic and clinical data with 9373 records. The data contained 9373 patient records with 12 attributes one of the attributes is dependent field representing either the patient has Stroke or not.

4.3.1 Data preprocessing

The collected raw data that contains lists of integers/numbers and texts/categorical values cannot feed directly into a machine learning model. This collected raw dataset contains categorical features such as sex, location and label. These listed features are not appropriate input for the any learning algorithm. Due to these reasons, data preprocessing techniques was applied that include data cleaning, data normalization, data balancing, and data numericalization that map symbolic-valued/non-numerical to numeric.

In data preprocessing, cleaning the data and making it suitable for a machine learning model tasks were performed, which also increases the accuracy and efficiency of a machine learning model. As mentioned above, the raw dataset attributes have both numeric and non-numeric values. Hence, non-numeric attributes translated to numeric values by using encoding/ numericalization technique. Then, the data normalization was applied only to numeric attributes, and data balancing was applied to overcome the problem of the imbalanced dataset which leads to biases in machine learning.

Python libraries are need for performing data processing tasks. Numpy, pandas, matplotlip, scikit-learn and some python functions and classes are imported to carried out experiments using the following python code.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import QuantileTransformer
```

After importing python libraries, then import the collected patient data in Comma Separated Value (CSV) files format using read_csv () python function.

```
#Loading the collected patient data
dataset = pd.read_csv('Patient_dat_final.csv', low_memory=False)
```

4.3.1.1 Data Cleaning

Data cleaning is the most important task in data preprocessing for preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. In our case, 180 missing data point values are existed in dataset such missing values are listed below Figure 4.2.

```
#number of missing data points
dataset.isnull().sum()
```

```
Age                0
Weight            11
Height             0
BMI                0
Sex                0
Cholestrol        18
Blood Pressure   114
Pulse Rate         0
RBC                4
FBS               33
Location           0
Label              0
```

Figure 4.2. List of missing data point values

There are many fillings missing value techniques such as filling missing data with a value using imputation, average, mode and median, and deleting columns with missing data and deleting rows with missing data.

Instead of using filling missing data values on the attributes that mentioned above in Figure.4.2. Deleting rows with missing data technique was applied to remove the 180 missing data points from our dataset. The original total dataset was 9373 data rows, after removing 164 missing data rows, 9209 data rows remained in the database. The remained data rows were used for further data processing.

```
#drop missing value of all rows
data_final = dataset.dropna(how='any',axis=0)
data_final
```

4.3.1.2 Data Numericalization

As known, machine learning algorithms take as input only numerical values to work correctly. So, we applied feature conversion to non-numeric features to transform their unified format. There are 9 numeric features and 2 non-numeric features in our collected dataset. The input value of machine learning algorithms should be a numeric matrix. Therefore, the non-numeric features, such as 'Sex', and 'Location' features are converted into numeric form. For example, the feature 'Location' has Ten types of attributes, 'Addis Ababa', 'SNNP', 'Dire Dawa', 'Oromiya', 'Amhara', 'Benshangule','Gambela', 'Somali', 'Tigray' and 'Afar', and its numeric values are encoded as binary vectors using dummies value approach as follows:

```
#encoding using Using dummies values approach
datlocation_trans = pd.get_dummies(datlocation,columns=['Location'],prefix=['Location'])
datlocation_trans
```

Location_Addis Aba	Location_Afar	Location_Amhara	Location_Beneshangul	Location_Dire Dawa	Location_Gambela	Location_Oromiya	Location_SNNP	Location_Som	Location_Tigray
1	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	1

Table 4.1 Numeric values of location attributes

The datasets contain stroke, and non-stroke classes label. We have used Label Encoder encoding technique to transform categorical features to numeric. The non-stroke class is assigned as label 0 and stroke class is labeled as 1.

```
from sklearn.preprocessing import LabelEncoder
label_encoder1 = LabelEncoder()
dlable= label_encoder1.fit_transform(Y)
la=['Label']
```

4.3.1.3 Data Normalization

Normalization is a technique often applied as part of data preparation for machine learning. Data normalization is a data preprocessing task and one of the first to be performed during intellectual analysis, particularly in the case of tabular data[39]. This feature is important since the scale used for the values for each variable might be different. The best practice is to normalize the data and transform all the values to a common scale.

The collected dataset features have various data values in different ranges/scales. So, the collected datasets didn't give appropriate results and the learning algorithm couldn't work efficiently. Hence, a data normalization technique is applied for transforming dataset features in a common range to avoid larger numeric feature values dominance over smaller numeric feature values. Quantile transform and min-max scaling are used to transform adjusted features value distribution to normal distribution and reduce the negative effect of marginal values respectively. Then, all un-normalized feature values in the datasets were rescaled to the range of predefined lower and upper bounds (0 to 1) by using min-max scaling. This method was used to map the range of our

dataset R to the range of R` in the specified range of [min_vlaue, max_vlaue] by using the following formula.

$$R` = (maxtarget - mintarget) \times [\frac{(R - min\_vlaue)}{max\_value - min\_value}] + mintarget$$

Where min_vlaue and max_value represent the maximum and minimum value of the dataset R within the range of values that assign as min_value= 0 and max_value = 1.

The dataset has 12 input variables, one output variable, and 9209 rows of data points. Figure 4.3 depicted the data distribution, minimum and maximum values of each input variable. Some of the dataset input variables have a skewed distribution, different minimum, and maximum values.

```
#check the distribution of patient raw data using histogram the data has normal distrbution or not normal
import matplotlib.pyplot as plt
num_column = ['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',
       'Pulse Rate', 'RBC', 'FBS']
hitd_dat[num_column].hist(figsize=(8,6),grid=True,color='Brown');
plt.tight_layout()
plt.show()
```

Therefore, quantile transformation and min_max scaler s were applied on each input variables to change skewed distribution to a normal probability distribution and normalize the input variables to improve the modeling performance.

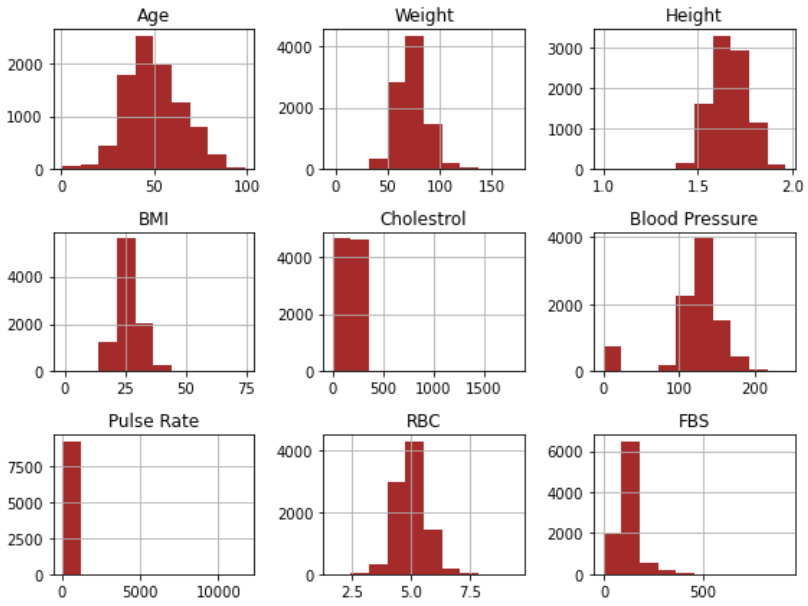

Figure 4.3 Numerical input variables for the patient dataset

We applied the quantile transform using the QuantileTransformer class and set the output_distribution argument to “normal”.

```
#applying data transformation method by using QuantileTransformer
from sklearn.preprocessing import QuantileTransformer
qt = QuantileTransformer(output_distribution= 'normal')
dataset_qrt=qt.fit_transform(hitd_dat)
dataset_qrt=pd.DataFrame(dataset_qrt)
dataset_qrt.columns = ['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',
       'Pulse Rate', 'RBC', 'FBS']
dataset_qrt
```

Figure 4.4 shows that the shape of the histograms for each variable looks like Gaussian distribution as compared to the raw data shown in Figure 4.3.

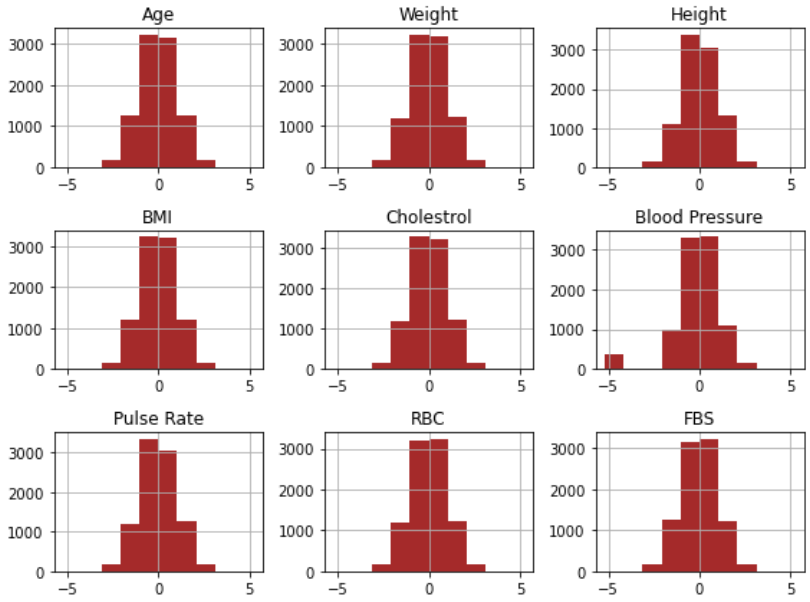

Figure 4.4. Normal quantile transformed numerical input variables for patient dataset

The MinMaxScaler was applied directly on the normalized quantile transformeddataset to normalize the input variables. We used the default configuration and its scale values to the range of 0 and 1, and then a normalized version of our dataset was created. Figure 4.5 shows that the distributions have been adjusted similarly as in Figure 4.4 and don't look like their original raw distributions as can be seen in Figure 4.3. but the minimum and maximum values for each input variable are now 0.0 and 1.0 respectively.

```
# applying minmax scaller with default parameter
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
scaled = scaler.fit_transform(dataset_qrt)
scaled=pd.DataFrame(scaled)
scaled.columns= ['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',
       'Pulse Rate', 'RBC', 'FBS']
scaled
```

```
# view the distribution of data after applyed Minmax scaller
import matplotlib.pyplot as plt
scaled.hist(figsize=(8, 6),grid=True,color='Brown');
plt.tight_layout()
plt.show()
```

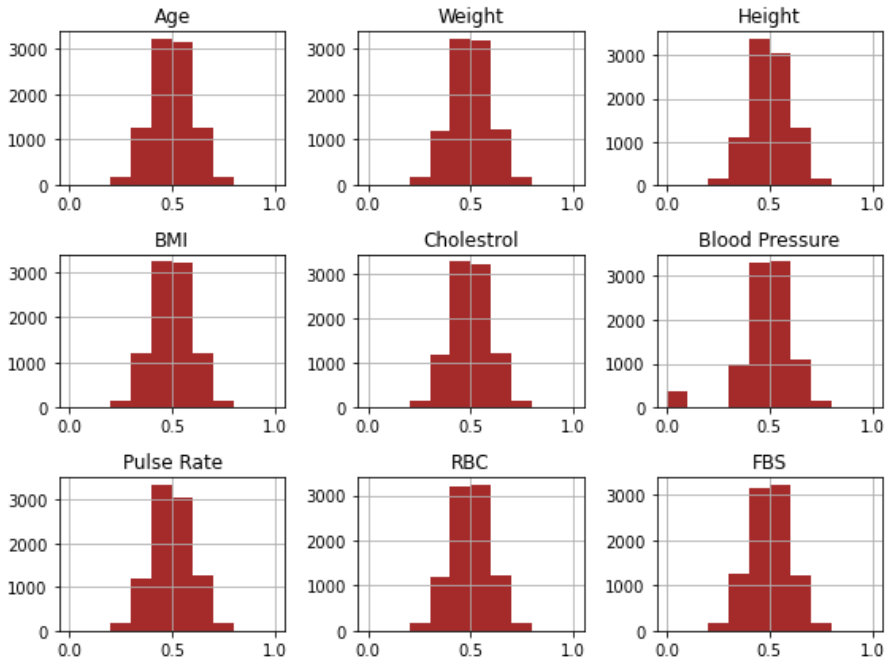

Figure 4.5. MinMaxScaler scaled input variables for the patient dataset

4.3.1.4 Data Balancing

The problem of imbalanced datasets leads to biases in machine learning. An imbalanced dataset occurs when there is a significantly unequal representation of one class than the others. It is necessary to balance the data before training a classifier [40]. The imbalance in our dataset can be a representative of improper class distribution or/and the existence of errors. To solve this problem, researchers suggest three methods such as sampling-based method, ensemble-based method, and cost-sensitive using class weight computing method [41]. The cost-sensitive using class weight computing method has been used. As it can be seen in Figure 4.6, our dataset samples are unevenly distributed. For example, the ‘‘non-stroke’ data accounts for more than 91 percent of the data.

However, the “stroke” data account for merely 9 percent. This leads to a training result bias towards larger data categories.

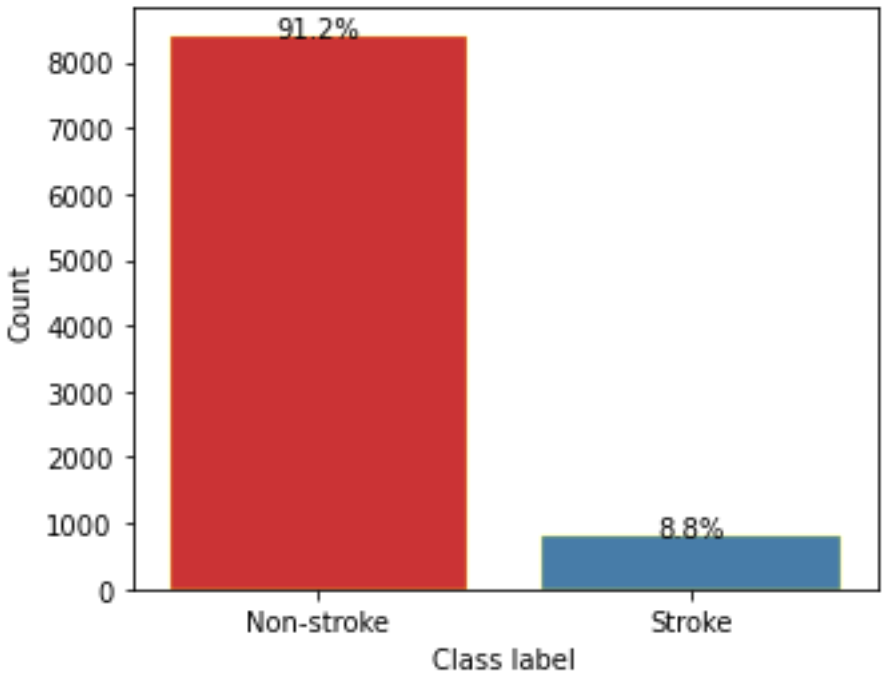

Figure 4.6. Dataset class distribution

4.4 Training and Test dataset

The dataset contains 12 features (9 continuous or discrete numerical attributes and 2 categorical or symbolical attributes) as shown in Table 4.1.

Class	No of training Sample	No of testing sample	Total sample
Non- stroke	5876	2523	8399
Stroke	570	240	810
Total number of samples = 9209			
Total training sample		6446	
Total testing sample		2763	

Table 4.2: Training and testing dataset samples

In this study, before the data was used to train the Logistic Regression, SVM and RF (random forest) Decision tree, we divided the preprocessed dataset into two parts. The first one, X, is the part of the dataset without a labeling feature, and the other one, Y, is the part of the dataset that makes this learning algorithm to be supervised learning using the following python code.

```
#separate depenendent and independent variables from dataset
X = data_patient_scalled.iloc[:,0:-1]
Y= data_patient_scalled.iloc[:,12]
```

Then, X and Y are split into two parts; X_train, X_test, Y_train and Y_test. 70% of the data train parts are used in the training section. The rest of the data, which is 30%, is used for testing to evaluate the performance of the model.

```
from sklearn.model_selection import train_test_split
```

```
X_train, X_test,Y_train,Y_test = train_test_split(X,Y,test_size=0.3)
```

After applying train_test_split method using above python code and running it, the training set contains 6446 instances and the testing set comprises 2763 instances. But, as it can be seen in Figure 4.6, our dataset samples are unevenly distributed. For example, the ‘‘non-stroke data account for more than 90 percent of data. However, the “stroke” data account for merely 9 percent. This leads to a training result bias towards larger data categories. To solve this problem, a cost-sensitive using class weight computing method has been used. The training and testing dataset samples are shown in Table 4.1.

4.5 Predictive Modeling

Predictive modeling is the process of developing a mathematical tool or model that generates an accurate prediction [42]. Traditionally, humans analyzed the data, but the volume of data surpasses their ability to make sense of it, which made them automate systems that can learn from the data and the changes in data to adapt to the shifting data landscape [43].

Predictive modeling works by analyzing historical and current data and generating a model to help predict future outcome/behavior. In predictive modeling data is collected and preprocessed then model is formulated, predictions are made, and the model is validated (or revised) as additional data becomes available.

4.5.1 Logistic Regression Model

In this subsection, stroke risk prediction model was build using Logistic Regression algorithm.

Logistic regression used to predict of descriptive variables or order variables based on a set of independent variables as if some of them are continuous variables), the other section being discrete variables (descriptive, order) [3]. Specifically, binary logistic regression was applied i.e. the dependent variable only takes two values (0,1). Logistic regression Python libraries with other libraries that mentioned on data preprocessing section are used for building the model.

```
########################################### Applying Logistic regression #
from sklearn.linear_model import LogisticRegression
```

Experiment 1

After importing required python libraries, create the model using logistic regression function. Fit the model on the 6446 instances of training data and 10 predictors variable with 10 dummies location variables and one dependent variable (label) using fit () function. Perform prediction on the 2763 instance of testing data using predict () function. The first experiment was conducted using the default logistic regression parameter with calculated class weight. This class_weight parameter value is calculated using a cost-sensitive weight computing method the same as RF algorithm class_weight.

```
##################################### Applying Logistic regression ############################
from sklearn.linear_model import LogisticRegression
#Create a Model and Train It
lr = LogisticRegression(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1,
                        class_weight={0: 0.5002575210372604, 1: 5.65438596491228}, random_state=None, solver='lbfgs',max_iter=100, verbose=0,
                        warm_start=False, n_jobs=None, l1_ratio=None)
```

```
lr.fit(X_train,Y_train)
#predict the model with testing data
pred = lr.predict(X_test)
```

Evaluating Model Performance

In this subsection, the performance evaluation and analysis of the logistic regression-based stroke risk prediction model is explained. This built prediction model performance evaluated using Accuracy, confusion matrix and classification report.

Accuracy: the performance of logistic regression accuracy is 93.0%, This accuracy considered as very good accuracy result.

```
# import evalution metrics
import sklearn.metrics as skm
accir= metrics.accuracy_score(pred,Y_test).round(2) * 100
print('Accuracy =', accir)
```

```
Accuracy = 93.0
```

Confusion matrix:

Confusion matrix measures prediction accuracy of the model in tabular form. It contains the number of correct and incorrect predictions summed up class-wise. A confusion_matrix() function was used to create confusion matrix and it provide the actual and predicted outputs as the arguments in the function.

```
#Ploting the confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test,pred)
def plot_confusion_matrix(cm, classes, normalized=True, cmap='bone'):
    plt.figure(figsize=[4, 3])
    norm_cm = cm
    if normalized:
        norm_cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        sns.heatmap(norm_cm, annot=cm, fmt='g', xticklabels=classes, yticklabels=classes, cmap=cmap)

plot_confusion_matrix(cm, ['Non-Stroke', 'Stroke'])
```

After running the above confusion matrix code, the following tabular form shown below consists of 4 different combinations of actual and predicted values.

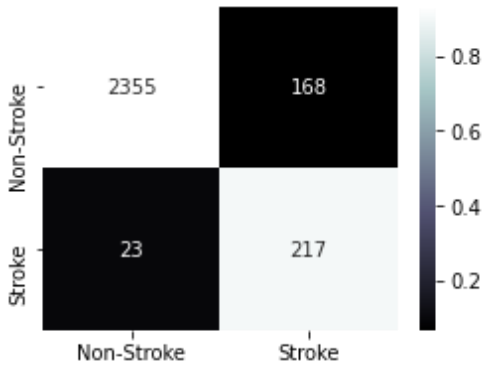

Table 4.3 Confusion matrix result L1

From 2763 total test data instances it has been split into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that in the first-row non-stroke class 2355 instances are correctly predict to the class non-stroke. However, 168 instance of non-stroke class are incorrectly predicted into stroke classes. For the second row, 217 are stroke instances are correctly predicted to stroke class, however, 23 instances are incorrectly predicted to the class of non-stroke.

Classification Report:

The classification report includes accuracy, precision, F1-score and recall.

```
import sklearn.metrics as skm
print(classification_report(Y_test, pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.99      0.93      0.96      2523
      Stroke       0.56      0.90      0.69       240

    accuracy                           0.93      2763
   macro avg       0.78      0.92      0.83      2763
weighted avg       0.95      0.93      0.94      2763
```

According to the above classification report, precision is express how accurate our model is or how our model makes a prediction accurately. In our case, the logistic regression model predicted 56% of patients are going to risk from stroke. So, this prediction model is 56% accurate when it says that a sample is stroke.

Recall evaluation metrics measure the model ability to detect positive sample. There are patients who have stroke in the test set and the Logistic Regression predication model can identify it 90 % of stroke patient.

F1-score measure the overall performance of the model. Our prediction model predicted correctly 69 % of stroke patients are predicted correctly.

Improve Logistic Regression prediction model performance

In order to improve the model by setting different parameters with values that differ from the default parameters value. For instance, solver, regularization strength c, and penalty (regularization) are parameters of logistic regression. In the previous experiment, we used these above-listed parameters and the remaining parameters with default values including weight was calculated using the cost-sensitive using class weight method. 5 experiments were conducted to improve the performance of the model by changing the aforementioned parameters value different from the default. The final assigned values of these parameters that perform the model performance are given bellow.

Experiment 2:

```
#exprement five
from sklearn.linear_model import LogisticRegression
#Create a Model and Train It
lr5 = LogisticRegression(C= 0.1, solver='liblinear',penalty='l2', class_weight={0: 0.5002575210372604, 1: 5.65438596491228})
lr5.fit(X_train,Y_train)
#predict the model with testing data
pred5 = lr5.predict(X_test)
```

Evaluating Model Performance

Accuracy is improved from 93 % to 94 % as shown below.

```
# import evalution metrics
import sklearn.metrics as skm
accir5= metrics.accuracy_score(pred5,Y_test).round(2) * 100
print('Accuracy =', accir5)
```

```
Accuracy = 94.0
```

Confusion matrix: is expressed using tabular form like table. The following confusion matrix consists of 4 different combinations of actual and predicted values.

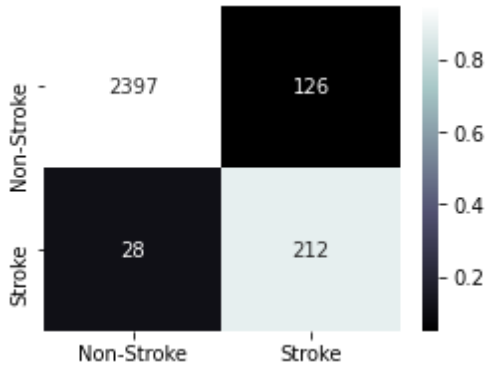

Table 4.4 Confusion matrix result L2

We have 2763 total test data instances it has been randomly split into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that

in the first-row non-stroke class 2397 instances are correctly predict to the class non-stroke. However, 126 instance of non-stroke class are incorrectly predicted into the stroke classes. For the second row, 212 are stroke instances are correctly predicted to stroke class, however, 28 instances are incorrectly predicted to the class of non-stroke.

The following classification result including precision, recall and f1-score was better as compared to the result that gained in previous experiments. When we looking the precision result, 60% of patients are going to risk from stroke. Thus, this prediction model is 60% accurate when it says that a sample is stroke.

As a result, Logistic Regression with C =0.1, solver= 'liblinear', penality= 'l2' and class_weight parameter value was calculated using cost sensitive weight computing method achieves 63% precision, 88% recall and 73% f1-score. 88 % of stroke patient accurately identify as stroke and 73% of stroke patient sample accurately predicted.

```
import sklearn.metrics as skm
print(classification_report(Y_test, pred5, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.99      0.95      0.97      2523
      Stroke       0.63      0.88      0.73       240

    accuracy                           0.94      2763
   macro avg       0.81      0.92      0.85      2763
weighted avg       0.96      0.94      0.95      2763
```

4.5.2 Support Vector Machine (SVM) Model

SVM algorithm was used for building stroke risk prediction model. This algorithm performs a classification by making optimal multidimensional hyperplane to discriminate two classes by maximizing the margin between two data groups [44]. This algorithm used kernel function to transform the input space into a multidimensional space. In real-world situations, some data points

in the two classes are difficult to be separated. SVM solves this kind of problem perfectly by using parameters such as:

- C is the regularization parameter that minimized the misclassifications and maximization of margin. It also controls over-fitting of the model by specifying tolerance for misclassification.
- kernel functions including linear, polynomial, sigmoid, and radial basis functions (RBF) to add more dimensions to the low dimensional space, as a result that two classes could be separable in the high dimensional space.
- Gamma controls the degree of nonlinearity of the model.

In this research, these aforementioned SVM parameters were required to build an optimal SVM prediction model. Sklearn.SVM Python libraries with other libraries that mentioned on data preprocessing section were used for building the model.

```
#importing svm library frm sklear to conduct experment1
from sklearn import svm
```

Experiment 1:

The 9209 preprocessed data was used for building SVM prediction model. Fit the model on the 6446 instances of training data and 10 predictor variables with 10 dummies location variables and one dependent variable (label) using the fit () function. Perform prediction on the 2763 instances of testing data using predict () function. The first experiment was conducted using the default SVM parameter with calculated class weight. This class_weight parameter value was calculated using a cost-sensitive weight computing method the same as RF and logistic regression algorithms class_weight.

```
# class weight calculated by using cost sensitive method
weights= {0: 0.5002575210372604, 1: 5.65438596491228}
```

```
# using defaultparamer values with calculated weight
rbf1 = svm.SVC(kernel='rbf', C=1,gamma='scale', class_weight=weights)
rbf1.fit(X_train, Y_train)
```

```
SVC(C=1, class_weight={0: 0.5002575210372604, 1: 5.65438596491228})
```

```
rbf_pred = rbf1.predict(X_test)
```

Evaluating Model Performance of experiment 1:

In this subsection, the SVM-based stroke risk prediction model performance evaluation and analysis are explained. This built prediction model performance was evaluated using Accuracy, confusion matrix and classification report. Testing data sets were used to assess the performance of the build models.

Accuracy: the performance of SVM accuracy is 94.61%, This accuracy considered as very good accuracy result.

```
from sklearn.metrics import accuracy_score
import sklearn.metrics as skm
rbf_accuracy = accuracy_score(Y_test, rbf_pred)
print('Accuracy (Radial Kernel): ', "%.2f" % (rbf_accuracy*100))
```

```
Accuracy (Radial Kernel):  94.61
```

Confusion Matrix:

Confusion matrix measures prediction accuracy of the model in tabular form. It contains the number of correct and incorrect predictions summed up class-wise. A confusion_matrix() function was used to create confusion matrix and it provide the actual and predicted outputs as the arguments in the function. To generate confusion matrix by using the following python confusion matrix function.

```
#Ploting the confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test,rbf_pred)
def plot_confusion_matrix(cm, classes, normalized=True, cmap='bone'):
    plt.figure(figsize=[4, 3])
    norm_cm = cm
    if normalized:
        norm_cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        sns.heatmap(norm_cm, annot=cm, fmt='g', xticklabels=classes, yticklabels=classes, cmap=cmap)

plot_confusion_matrix(cm, ['Non-Stroke', 'Stroke'])
```

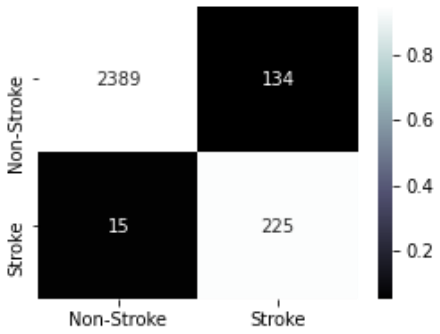

Table 4.5 Confusion matrix result S1

The tabular form is shown above consists of 4 different combinations of actual and predicted values. We have 2763 total test data instances it has been randomly split into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that in the first-row non-stroke class 2389 instances are correctly predict to the class non-stroke. However, 134 instance of non-stroke class are incorrectly predicted into the stroke classes. For the second row, 225 are stroke instances are correctly predicted to stroke class, however, 15 instances are incorrectly predicted to the class of non-stroke.

Classification report:

The classification report comprises accuracy, precision, F1-score and recall. In order to generate classification report evaluation result of the build model, we used the following classification report function.

```
import sklearn.metrics as skm
print(classification_report(Y_test, rbf_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.99      0.95      0.97      2523
      Stroke       0.63      0.94      0.75       240

    accuracy                           0.95      2763
   macro avg       0.81      0.94      0.86      2763
weighted avg       0.96      0.95      0.95      2763
```

The performance evaluation result as shown above classification report, precision is express how accurate our model is or how our model makes a prediction accurately. In our case, the SVM model predicted 99% of patients are going to risk from stroke. So, this prediction model is 99% accurate when it says that a sample is stroke.

Recall evaluation metrics measure the model ability to detect positive sample. There are patients who have stroke in the test set and the SVM model can identify it 63 % of stroke patient.

F1-score measure the overall performance of the model. Our prediction model predicted correctly 75% of stroke patients are predicted correctly.

The performance evaluation results described above used default parameters including the radial basis kernel function as the kernel, C=1 as the penalty/regularization, and gamma='scale' as the nonlinearity controller. In these evaluation results, recall and f1-score were not achieved well. So, model improvement is necessary by tuning the value of the parameter.

Improve SVM prediction model performance

The previously built SVM model needs improvement. To increase the performance evaluation results of the model, we need to find the optimal parameter by setting parameters with different values. In this research, radial base, polynomial and linear functions were used. Moreover, the regularization C and gamma parameters were used by assigning a variety of their values. As we discussed before, in the previous experiment, SVM parameters with default values and calculated

class weight were used. Extensive experiments were conducted to improve the performance of the model by using the value of the aforementioned parameter different from the default. Then, we will describe the following three experiments from the most performed experiments.

Experiment 2:

This experiment conducted using radial base kernel function, C =0.1 and gamma = 0.1 parameters with value.

```
# by chanching the value of c regularization and gamma parametrs value
rbf2 = svm.SVC(kernel='rbf', C=0.1,gamma=0.1, class_weight=weights)#.fit(X_train, Y_train)
rbf2.fit(X_train, Y_train)
```

```
SVC(C=0.1, class_weight={0: 0.5002575210372604, 1: 5.65438596491228}, gamma=0.1)
```

```
rbf2_pred = rbf2.predict(X_test)
```

Evaluating model performance Experiment 2:

To evaluate this experiment, we used the same evaluation metrics just as others.

Accuracy: It achieves better accuracy as compared to previous.

```
from sklearn.metrics import accuracy_score
import sklearn.metrics as skm
rbf2_accuracy = accuracy_score(Y_test, rbf2_pred)
print('Accuracy (Radial Kernel): ', "%.2f" % (rbf2_accuracy*100))
```

```
Accuracy (Radial Kernel):  96.45
```

Confusion matrix:

The following tabular form shown below consists of 4 different combinations of actual and predicted values.

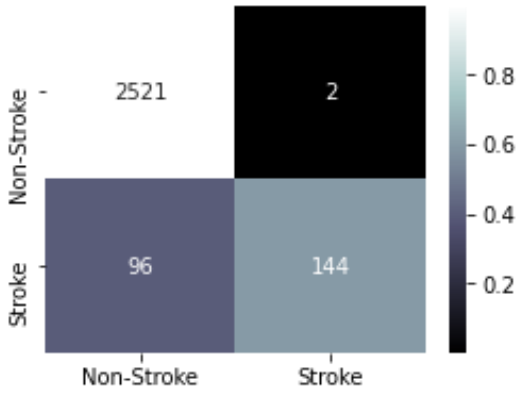

Table 4.6: Confusion matrix result S2

From 2763 total test data instances it has been randomly divided into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that in the first-row non-stroke class 2521 instances are correctly predict to the class non-stroke. However, only 2 instance of non-stroke class are incorrectly predicted into stroke classes. For the second row, 144 instances are stroke instances are correctly predicted to stroke class, however, 96 instances are incorrectly predicted to the class of non-stroke.

Classification report:

The following result comprises precision, recall and f1-score. In this experiment, 99% of precision, 60 % of recall and 75 % of recall were achieved.

```
import sklearn.metrics as skm
print(classification_report(Y_test, rbf2_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.96      1.00      0.98      2523
      Stroke       0.99      0.60      0.75       240

    accuracy                           0.96      2763
   macro avg       0.97      0.80      0.86      2763
weighted avg       0.97      0.96      0.96      2763
```

Experiment 3:

This experiment conducted using polynomial kernel function, C =0.1 , gamma = 0.2 and degree = 2 parameters with value.

```
# by chanching the value of c regularization, gamma and degree parametrs value
poly2 = svm.SVC(kernel='poly', degree=2, C=0.1,gamma=0.2, class_weight=weights)#.fit(X_train, Y_train)
poly2.fit(X_train, Y_train)
```

```
SVC(C=0.1, class_weight={0: 0.5002575210372604, 1: 5.65438596491228}, degree=2,
    gamma=0.2, kernel='poly')
```

```
poly2_pred = poly2.predict(X_test)
```

Evaluating Model performance of Experiment 3:

Accuracy, confusion matrix and classification report evaluation metrics are used to assess this experiment.

Accuracy: 96.63% of accuracy was achieved. This accuracy is better as compared to experiment 1 and the previous experiment that used default parameter values.

```
from sklearn.metrics import accuracy_score
import sklearn.metrics as skm
poly2_accuracy = accuracy_score(Y_test, poly2_pred)
print('Accuracy (poly Kernel): ', "%.2f" % (poly2_accuracy*100))
```

```
Accuracy (poly Kernel):  96.63
```

Confusion Matrix:

The following confusion matrix result consists of 4 different combinations of actual and predicted values.

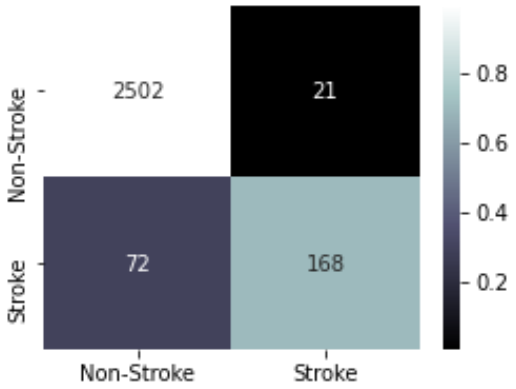

Table 4.7: Confusion matrix result S3

From 2763 total test data instances it has been randomly divided into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that in the first-row non-stroke class 2502 instances are correctly predict to the class non-stroke. However, only 21 instance of non-stroke class are incorrectly predicted into stroke classes. For the second row, 168 instances are stroke instances are correctly predicted to stroke class, however, 72 instances are incorrectly predicted to the class of non-stroke.

Classification Report:

In experiment 3, 89% of precision, 70 % of recall and 78 % of recall were achieved using polynomial kernel function with degree = 2.

```
import sklearn.metrics as skm
print(classification_report(Y_test, poly2_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.97      0.99      0.98      2523
      Stroke       0.89      0.70      0.78       240

    accuracy                           0.97      2763
   macro avg       0.93      0.85      0.88      2763
weighted avg       0.96      0.97      0.96      2763
```

Experiment 4:

This experiment was conducted using linear kernel function, C =0.03 and gamma = 0.1 parameters with value through the training set, 0.7 size data, and testing set 0.3 size data.

```
# by changing the C ,and gamma parametrs values
lrn2 = svm.SVC(kernel='linear', C=0.03,gamma=0.1, class_weight=weights)#.fit(X_train, Y_train)
lrn2.fit(X_train, Y_train)
```

```
SVC(C=0.03, class_weight={0: 0.5002575210372604, 1: 5.65438596491228},
    gamma=0.1, kernel='linear')
```

```
lnr2_pred = lrn2.predict(X_test)
```

Evaluation Model Performance of Experiment 4:

In Experiment 4, the taste data and evaluation metrics we used were similar to the taste data and evaluation metrics of the previous experiments.

Accuracy: When using linear kernel function with C =0.01, and gamma =0.1 The testing accuracy of the model is 96.56%.

```
from sklearn.metrics import accuracy_score
import sklearn.metrics as skm
lnr2_accuracy = accuracy_score(Y_test, lnr2_pred)
print('Accuracy (linear Kernel): ', "%.2f" % (lnr2_accuracy*100))
```

```
Accuracy (linear Kernel):  96.56
```

Confusion Matrix:

2763 total test data instances it has been randomly divided into 2523 instances that are non-stroke and 240 instances are stroke. From the bellow confusion matrix result, we can observe that in the first-row non-stroke class 2503 instances are correctly predicted to the class non-stroke. However, only 20 instances of non-stroke classes are incorrectly predicted into stroke classes. For the second row, 165 instances are stroke instances are correctly predicted to the stroke class, however, 75 instances are incorrectly predicted to the class of non-stroke.

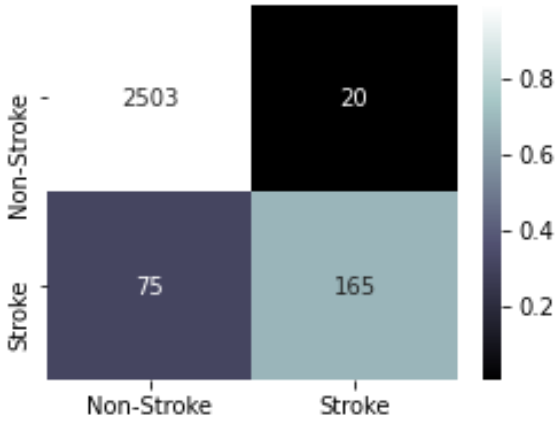

Table 4.8: Confusion matrix result S4

Classification Report:

When using linear kernel function, C=0.03, and gamma= 0.1 parameter values. 89% of precision, 69 % of recall, and 78 % of recall were achieved for correctly identifying stroke.

```
import sklearn.metrics as skm
print(classification_report(Y_test, lnr2_pred, target_names=target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       0.97      0.99      0.98      2523
      Stroke       0.89      0.69      0.78       240

    accuracy                           0.97      2763
   macro avg       0.93      0.84      0.88      2763
weighted avg       0.96      0.97      0.96      2763
```

When comparing these experiments to each other, the evaluation result of experiment 3 was better as compared to experiment 1,2 and experiment 4.

4.5.3 Random Forest (RF) Decision Tree Model

This subsection deals with how the random forest decision tree prediction model is built using tuning hyper parameters. The tuning hyper parameters with values python codes are as follows.

```
from sklearn.ensemble import RandomForestClassifier

modelrfc = RandomForestClassifier(n_estimators = 15, max_features= 'sqrt', bootstrap=True,criterion = 'entropy',
                                  random_state = 60,min_samples_leaf=70,min_samples_split=480, max_depth = 5,
                                  class_weight= {0: 0.49200687174422303, 1: 5.595486111111111}
)
modelrfc.fit(X_train, Y_train)
```

In this experiment a Prediction model building is done using RF Decision tree algorithm. Cost sensitive weight computing method used to calculate dictionary format class_weight parameter value. The following python function code was applied to calculate class_weight parameter values.

```
#weight assign to traing instance
def CreateBalancedSampleWeights2(Y_train, largest_class_weight_coef):
    classes = np.unique(Y_train, axis = 0)
    classes.sort()
    class_samples = np.bincount(Y_train)
    total_samples = class_samples.sum()
    n_classes = len(class_samples)
    weights = total_samples / (n_classes * class_samples * 1.0)
    class_weight_dict = {key : value for (key, value) in zip(classes, weights)}
    class_weight_dict
    class_weight_dict[classes[0]] = class_weight_dict[classes[0]] * largest_class_weight_coef
    print(class_weight_dict)
    sample_weights = [class_weight_dict[y] for y in Y_train]
    return sample_weights
```

Experiment

The preprocessed dataset used for conducted experiment. This dataset consists of 9209 data rows and 11 independent attributes with 10 dummies attribute and one dependent attribute/class (which is Label) and the remaining features are independent attributes.

Evaluating Model Performance

In this subsection, the performance evaluation and analysis of the machine learning based stroke risk prediction model using RF are explained. To evaluate the classification performance of random forest (RF) models, we select accuracy, classification report includes (precision, recall and f1-score), and confusion matrix.

```
#Predict the response for test dataset
y_predrf = modelrfc.predict(X_test)
accrf = metrics.accuracy_score(y_predrf,Y_test).round(3)*100
print("Accuracy = ", accrf)
```

```
Accuracy =  99.3
```

After running the above python code, the RF achieved 99.3% accuracy.

Confusion Matrix:

the following confusion matrix was found after using the following python code.

```
#Ploting the confusion matrix
from sklearn.metrics import classification_report, confusion_matrix, f1_score,auc,recall_score,precision_score
cm = confusion_matrix(Y_test,y_predrf)
def plot_confusion_matrix(cm, classes, normalized=True, cmap='bone'):
    plt.figure(figsize=[4, 3])
    norm_cm = cm
    if normalized:
        norm_cm = cm.astype('float') / cm.sum(axis=1)[:, np.newaxis]
        sns.heatmap(norm_cm, annot=cm, fmt='g', xticklabels=classes, yticklabels=classes, cmap=cmap)

plot_confusion_matrix(cm, ['Non-Stroke', 'Stroke'])
```

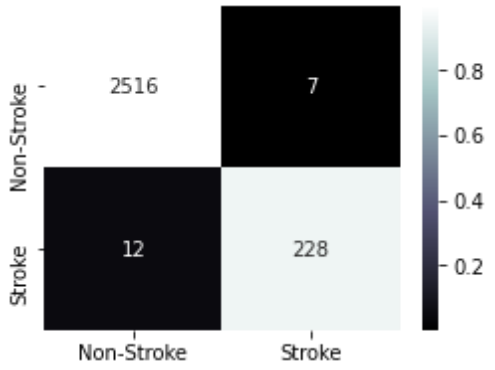

Table 4.9: Confusion matrix result D1

From 2763 total test data instances it has been randomly split into 2523 instances are non-stroke and 240 instances are stroke. From the above confusion matrix result, we can observe that in the first-row non-stroke class 2516 instances are correctly predict to the class non-stroke. However, only 8 instance of non-stroke class are incorrectly predicted into stroke classes. For the second

row, 228 are stroke instances are correctly predicted to stroke class, however, 12 instances are incorrectly predicted to the class of non-stroke

Classification report:

It includes f1-score, precision, and recall. We find the following classification report using the following python code.

```
import sklearn.metrics as skm
print(classification_report(Y_test, y_predrf, target_names = target_names))
```

```
              precision    recall  f1-score   support

  Non-stroke       1.00      1.00      1.00      2523
      Stroke       0.97      0.95      0.96       240

    accuracy                           0.99      2763
   macro avg       0.98      0.97      0.98      2763
weighted avg       0.99      0.99      0.99      2763
```

Based on the above classification report, precision is express how accurate our model is or how our model makes a prediction accurately. The RF model predicted 97% of patients are going to risk from stroke. So, this prediction model is 97% accurate when it says that a sample is stroke.

Recall evaluation metrics measure the model ability to detect positive sample. There are patients who have stroke in the test set and the RF model can identify it 95 % of stroke patient.

F1-score measure the overall performance of the model. Our RF prediction model predicted correctly 96 % of stroke patients are predicted correctly.

4.7 Comparing classification algorithms

Comparative performance analysis of the three machine learning algorithms was conducted with each other. The results are based on the result of the evaluation metrics such as accuracy, precision, recall, F1-score, and FAR. As per the analysis of the accuracy, precision, recall, and F1-score the

RF prediction models performed better compared to logistic regression and SVM models as shown in Figures 4.7 – 4.10.

The accuracy of the proposed stroke risk prediction model by using the three classifiers including RF, SVM, and logistic regression is shown in Figure 4.7. It can be seen from Figure 4.7 that the accuracy of the RF higher than the accuracy of the SVM and logistic regression, but SVM is relatively achieving a better accuracy result as compared to logistic regression algorithm. Based on the extensive experimental results, the RF algorithms outperformed SVM and logistic regression algorithms.

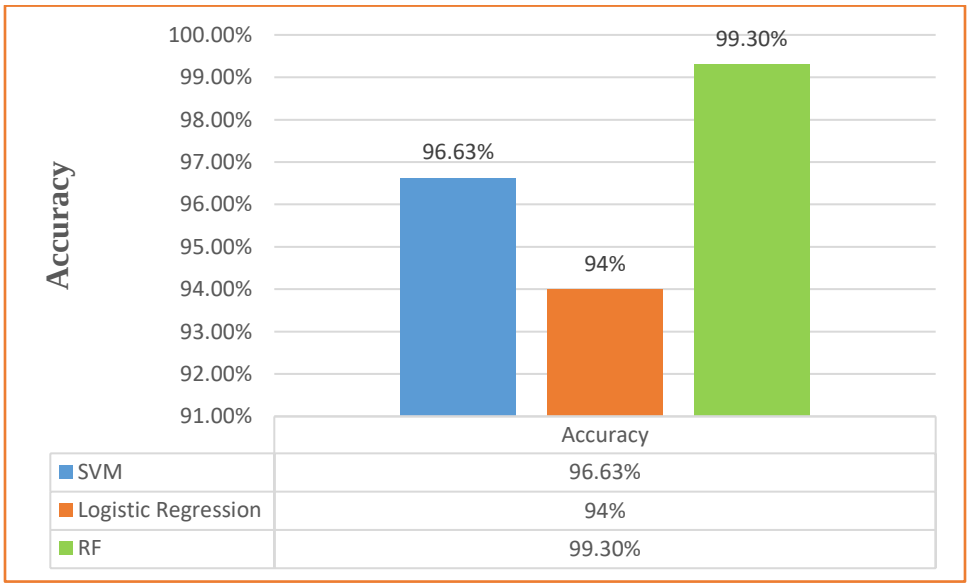

Figure 4.7: Accuracy of the three classifiers

The precision of the three machine learning models on the testing dataset is shown in Figure 4.8. When tune parameters values, we can see that the precision of RF is the highest while that of logistic regression is the lowest in the case of stroke class. In the case of non-stroke class, RF gives 100% precision as compared to the other models. However, the precision of logistic regression is lower when identifying stroke class.

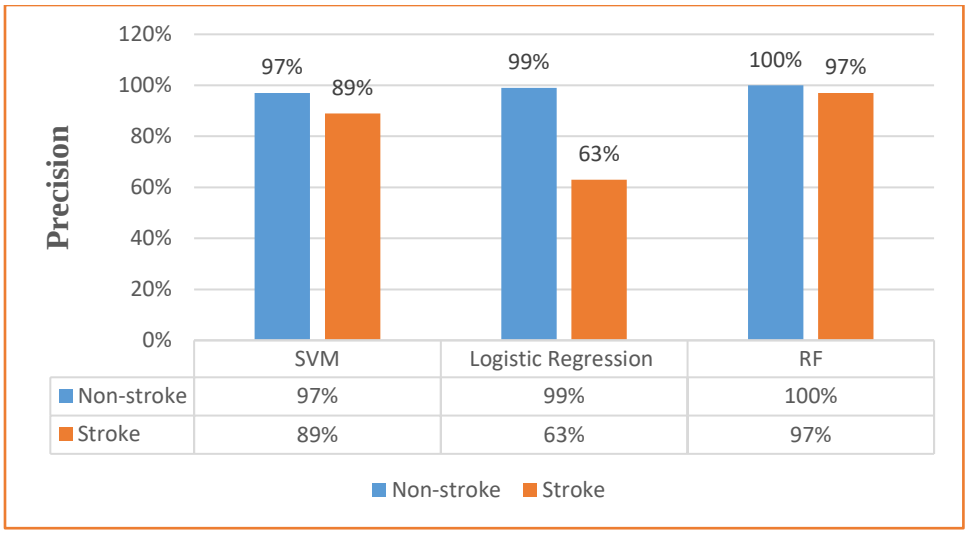

Figure 4.8: The precision of the three classifiers in both classes

Figure 4.9 shows the recall of the three-machine learning including (RF, SVM and logistic regression) for using the test dataset. We can easily conclude that the RF model perform better than the other models with both classes. But, SVM achieved higher recall compared to the logistic regression model for the identification of the non-stroke class.

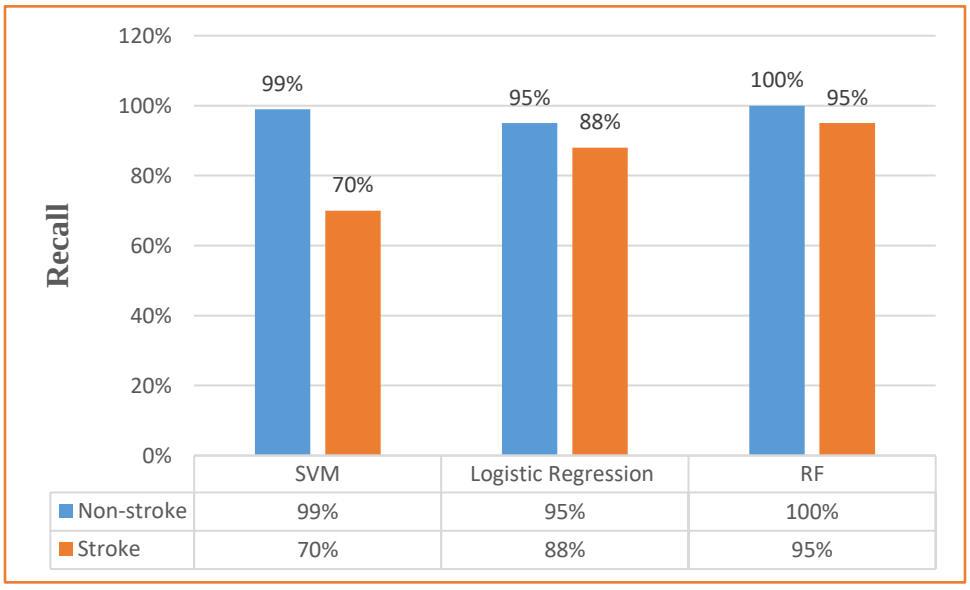

Figure 4.9: Recall of the three classifiers in both classes

Figure 4.10 shows the F1-score performance of the three classifiers such as RF, SVM, and logistic regression. Among them, RF gets the highest F1-score in both classes than the other models.

However, the F1 scores of the SVM are higher than the F1 scores of the logistic regression algorithm.

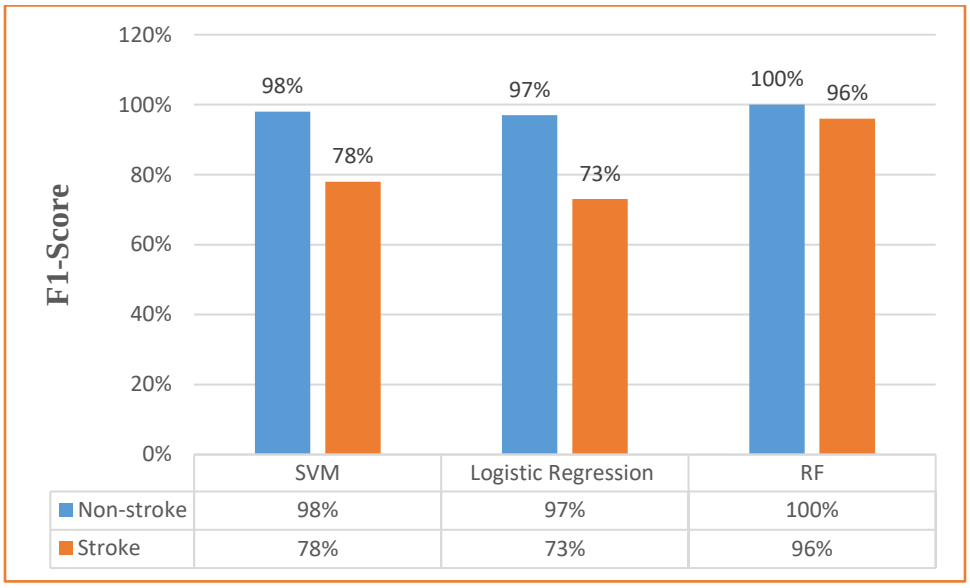

Figure 4.10: F1 Score of the three classifiers in both classes

In general, based on its performance result random forest (RF) decision tree is recommended for stroke risk prediction.

4.6 Identifying risk factor of stoke using RF

In order to identify the most important factors affecting stroke risk using the RF method and to conduct all tests for this method, we used feature importance. The World Health Organization states that the most important factors that help to determine the patient's exposure to stroke are (weight, BMI, Blood pressure, RBC, FBS, Pulse rate, cholesterol, BMI and others) [45]. These factors have been taken into account in this research to determine whether the patients were exposed to stroke or not. To determine the most influenced features/ factors, the feature importance method was applied. In RF, the importance of the features was calculated using Gini approach. In this research, the dataset contains 11 features/factors, now we identify which features/factors are the most impacted on predicting the class /target variable, stroke risk. Then, the feature importance variable was used to calculate the feature importance scores of each feature using the following python code.

```
importances = modelrfc.feature_importances_
# Sort the feature importance in descending order
sorted_indices = np.argsort(importances)[::-1]
stroke_labels = data_patient_scalled.columns#[0:]
for f in range(X_train.shape[1]):
    print("%2d) %-*s %f" % (f + 1, 30,
                            stroke_labels[sorted_indices[f]],
                            importances[sorted_indices[f]]))
```

After running the above python code, we obtained the following importance score for each feature in decreasing order.

	Features	Importance score
6	Blood Pressure	0.784455
0	Age	0.076958
7	Pulse Rate	0.050950
9	FBS	0.046988
4	BMI	0.012614
8	RBC	0.006876
1	Weight	0.005295
5	Cholestrol	0.004400
3	Height	0.004156
2	Sex	0.002639
18	Location_Somali	0.002367
10	Location_Addis Ababa	0.002253
16	Location_Oromiya	0.000050
12	Location_Amhara	0.000000
13	Location_Beneshangul	0.000000
14	Location_Dire Dawa	0.000000
15	Location_Gambela	0.000000
17	Location_SNNP	0.000000
11	Location_Afar	0.000000
19	Location_Tigray	0.000000

Table 4.10 : Importance of risk factors for stroke risk prediction

Next, visualize the above feature scores with matplotlib python code.

```
import matplotlib.pyplot as plt
f, ax = plt.subplots(figsize=(8, 6))
plt.title('Feature Importance')
plt.bar(range(X_train.shape[1]), importances[sorted_indices], align='center')
plt.xticks(range(X_train.shape[1]), X_train.columns[sorted_indices], rotation=90)
plt.tight_layout()
plt.show()
```

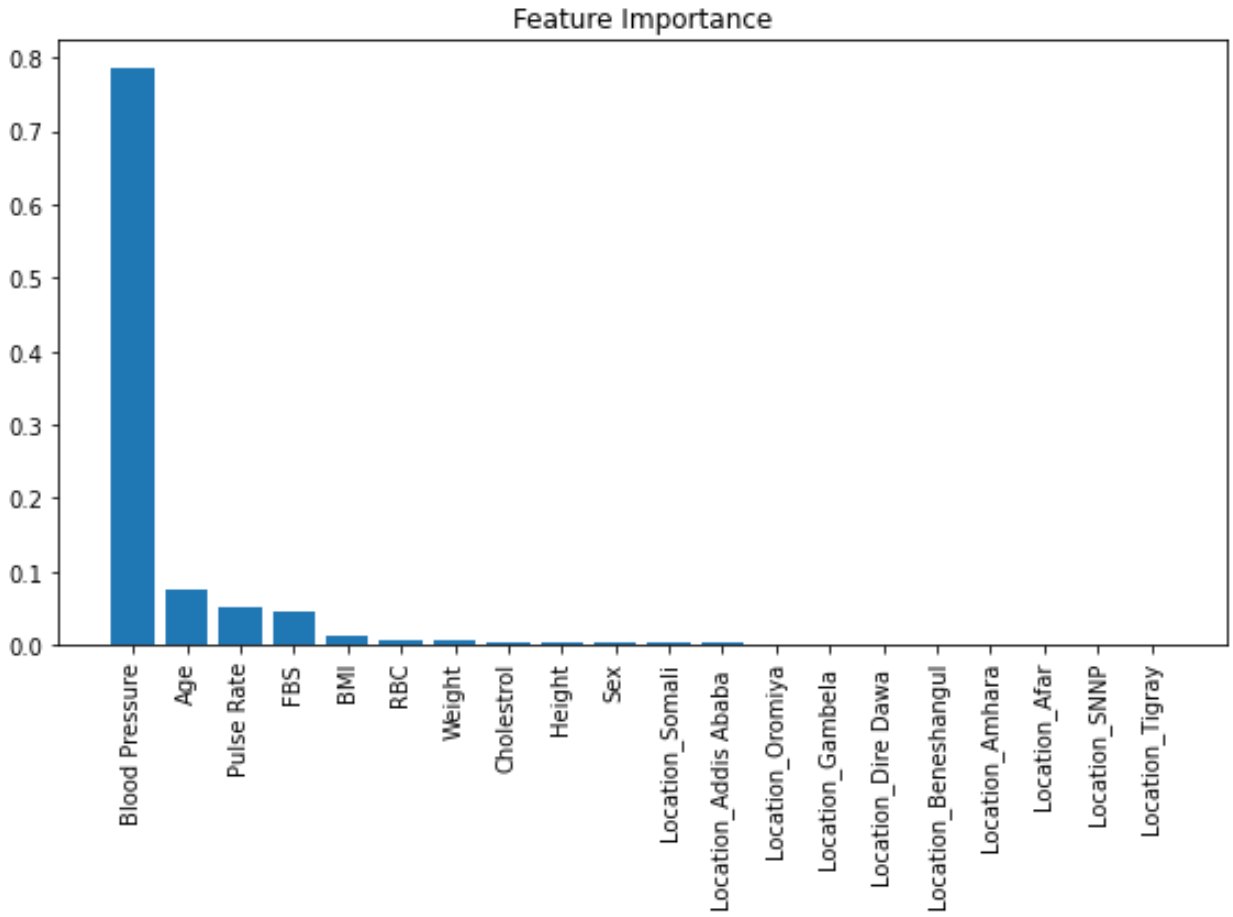

Figure 4.11. Shows stroke patient dataset features with feature importance scores plotted using the RF algorithm.

From the above Figure 4.11 and feature importance score depicted in table 4.10, we would be able to know that, out of 11 factors/features, blood pressure has 78.45% importance score for stroke

risk prediction. The next three factors such as age, pulse rate, and BMI has 18% importance. Locations, except Somalia, Addis Ababa and Oromiya have no importance for stroke risk prediction.

4.8 Discussion of result

The main objective of this research is to construct an optimal model that predicts stroke risk.

In this research, all experiments are performed with 6446 (70%) samples for training and 2763 (30%) samples for testing samples. All experiments, including logistic regression, SVM, RF experiments, are performed separately on the same training and test data.

The SVM with an accuracy of 96.63% for the testing data and takes the second position after the RF model and logistic regression classifier algorithm achieved an accuracy of 94% on the same test set data. Furthermore, as can be seen from the recall results, the RF (100%, 95%), SVM (99%, 70%), and logistic regression (95%, 88%) for non-stroke and stroke classes respectively. RF model has achieved better of both classes But, SVM and logistic regression outperform recall results only in the case of non-stroke. Hence based on experimental result RF is recommended for stroke risk prediction.

The major risk factor identified in this study, out of 11 factors/features is high blood pressure and it has 78.45% importance score for stroke risk prediction. The next three factors such as age, pulse rate, and BMI has 18% importance. This shows that this study has the same result related to identifying major risk factor for stroke prediction. Studies show that the most common risk factor associated with stroke in Ethiopia context is high blood pressure [3] [31][32] [45]. The second risk factor in this study is age where other studies [3]identifies age as one of the major factors.

The third risk factor in this study is Pulse Rate where other studies [1] [3] [31][32] [45] didn’t use this attribute, but we found out that it has correlation with stroke risk.

Chapter Five

Conclusion and Recommendation

5.1. Overview

Community based preventive strategy and patient education with early detection stroke and strict control will be an effective way of preventing stroke. Additionally, public education of symptoms of stroke and the need for immediate evaluation when noticing these symptoms could have an impact on the overall stroke management.

Statistics of World Health Organization (WHO) stated that stroke is the third leading cause of mortality in females and males. Identifying stroke is tedious and time-consuming for medical practitioners so that Machine Learning is needed for predicting stroke[1].

Hospitals collect demographic data, diagnosis data, etc, from patients that are very huge and complex. Machine learning technic helps to extract knowledge from these data, which is important for understanding risk factors and early detection and prevention.

Machine learning algorithms can be applied to dramatically reduce the burden on the health sector and stroke prediction models needs to be done by including data from every corner of the country to come up with general model that can be used for prevention, detection and treatment of stroke.

This chapter summarizes the entire findings of the study. It is divided into two main sections the conclusion and recommendations of the future works.

5.2 Conclusions

The necessary data for the experiment obtained from the Hallelujah Hospital on excel sheets from health management system and the second data source is from Zewditu Hospital, which was extracted from patient cards manually a total dataset of 9373. Then the necessary preprocessing activities applied on the dataset after which 9373 data was prepared for the experiment.

Classification and prediction model building was experimented with Decision Tree, SVM and Logistic Regression algorithms. And the tools were used to simulate all the experiment is Anaconda using Python programming. The confusion matrix used and then accuracy and sensitivity were calculated.

From the results of the experiments, we concluded that the Machine Learning algorithms Techniques can be effectively applied for stroke risk prediction by developing predictive models with acceptable level of accuracy in a given attributes.

Therefore, hospitals in Ethiopia can apply Machine Learning Models to detect, identify and manage patient risk for stroke. Hospitals in Ethiopia can apply by adopting Machine Learning Models to detect patients' possible risk for stroke. This will help physicians to easily identify the risks and manage the risk factors before it results to stroke case this will reduce the amount of time spend on stroke assessment and the amount investment for stroke.

However, due to the quality and size of dataset used, an increased size in the dataset with increase in amount and diversity of attributes involving different hospitals in all regions of the country needs to be done. This could have an impact for an inclusive model for stroke risk prediction. It could have enabled the research to make use of larger data with more attributes than those used in this study to help address other areas of problem in the health industry of the country. All Hospitals need to open their doors for researchers in order to get more data in quality and size. The model building experiment conducted based on the collected datasets from Hallelujah & Zewditu hospitals with just 9373, dataset but it is the researcher's belief that it would have resulted an inclusive model if it includes other Hospital dataset in different regions of Ethiopia with big data size. If other dataset with more attributes that are not used in this study were also utilized the result of the model can be improved. The next section presents the recommendations drawn based on the result of the study.

The result of this study shows that, the research questions formulated at the beginning are answered as shown below.

The first research question is to identify attributes or variables that are used to predict stroke risk of individuals. Accordingly, in this study blood pressure, Age, Weight, Height, BMI (Body Mass Index), Sex, Cholesterol, Pulse Rate, RBC (Red Blood Cell Count), FBS (Fasting Blood Sugar), location status attributes are found to be effective, relevant and important predictors for the target

class Stroke for predicting stroke risk of individuals.

Once the data is collected and prepared, appropriate machine learning algorithms, such as RF Decision Tree, SVM and Logistic Regression algorithms are selected to build the predictive model. Experimental result finally shows that RF Decision Tree is found the most suitable for constructing an optimal model for Stroke Risk Prediction. Based on the selected RF model, blood pressure is the most critical factor for stroke risk prediction, which is followed by age, pulse rate and BMI attributes.

On the other side, because of lack of information on smoking cigarette status of patients we are unable to investigate its effect on stroke risk prediction.

5.3 Recommendations

Even though the investigation undertaken is mainly for academic purpose, it will have important contribution for health practitioner and for other researchers interested in similar area. Although the results of this study are inspiring, there are problem areas that need further investigation for future work to attain better inclusive model and also bring it to an operational level. Therefore, the researcher forwards the following issues as a future research direction based on this study: -

- Since there is no organized medical data available, we have dropped one of the selected attributes smoking because information about the status of patients' smoking not found. Feature studies can use drug addiction, smoking and alcohol consumption attributes together with the attributes used in this study.
- In this study, demographic and clinical data of patients is used, to make sure that the prediction result correct feature researches can use Brain CT/MRI image to train and detect stroke.
- In this study an attempt made to construct an optimal model to make it applicable in the given problem area so that, there is a need to connect the model to health information systems or knowledge base system to make it usable.

- Since, stroke risk differs from the living area, genetic makeup and lifestyle of the patients, there is a need to collect patient data for stroke prediction from different parts of the country.
- There are many unstructured medical data so there is a need to automatically extract important information to convert to structured data for constructing stroke predictive model in the feature.

Reference:

[1] M. Sultan, F. Debebe, A. Azazh, and G. W. Hassen, “Epidemiology of stroke patients in Tikur Anbessa Specialized Hospital: Emphasizing clinical characteristics of Hemorrhagic Stroke Patients,” *Ethiop. J. Heal. Dev.*, vol. 31, no. 1, pp. 13–17, 2017.

[2] B. Fantahun, M. G. Ministry, R. H. Disease, and A. View, “Ethiopian National Guideline on Major NCDs Guidelines on Clinical and Programmatic Management of Major Non Communicable Diseases,” no. August, 2016, doi: 10.13140/RG.2.2.24757.06889.

[3] A. Zewdie *et al.*, “Prospective assessment of patients with stroke in Tikur Anbessa Specialised Hospital, Addis Ababa, Ethiopia,” *African J. Emerg. Med.*, vol. 8, no. 1, pp. 21–24, 2018, doi: 10.1016/j.afjem.2017.11.001.

[4] D. Deshmukh and A. More, “International Journal of Innovative Research in Computer and Communication Engineering Applying Big Data in Higher Education,” 2017, doi: 10.15680/IJIRCCE.2017.

[5] Y. Cao, Yu & Chiu, Hsu-Kuang & Khosla, Aditya & Chiung, Cliff & Lin, “CS229 Project: A Machine Learning Approach to Stroke Risk Prediction,” no. 1, 2009.

[6] H. McHeick, H. Nasser, M. Dbouk, and A. Nasser, “Stroke Prediction Context-Aware Health Care System,” *Proc. - 2016 IEEE 1st Int. Conf. Connect. Heal. Appl. Syst. Eng. Technol. CHASE 2016*, pp. 30–35, 2016, doi: 10.1109/CHASE.2016.49.

[7] J. R. S, A. Professor, T. Dept.of ECE, College of Engineering, D. S. K. A. Professor, T. Rajiv Gandhi Institute of Development Studies Vellayambalam, and II, “Stroke Prediction Using SVM,” pp. 600–602, 2016.

[8] T. Kansadub, S. Thammaboosadee, S. Kiattisin, and C. Jalayondeja, “Stroke risk prediction model based on demographic data,” *BMEiCON 2015 - 8th Biomed. Eng. Int. Conf.*, pp. 3–5, 2016, doi: 10.1109/BMEiCON.2015.7399556.

[9] L. Zheng *et al.*, “Risk prediction of stroke: A prospective statewide study on patients in Maine,” *Proc. - 2015 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2015*, pp. 853–855, 2015,

doi: 10.1109/BIBM.2015.7359796.

[10] S. S. Prakash and M. Nivethitha Devi, “GUI BASED HEART STROKE PREDICTION USING MACHINE LEARNING ALGORITHMS,” *Int. J. Innov. Res. Electr. Electron. Instrum. Control Eng.*, vol. 9, no. 4, pp. 2321–5526, 2021, doi: 10.17148/IJIREEICE.2021.9434.

[11] A. Al-Khoder and H. Harmouch, “Evaluating four of the most popular Open Source and Free Data Mining Tools,” *Int. J. Acad. Sci. Res.*, vol. 3, no. 1, pp. 2272–6446, 2015, [Online]. Available: www.ijasrjournal.org

[12] D. Rolon-Mérette, M. Ross, T. Rolon-Mérette, and K. Church, “Introduction to Anaconda and Python: Installation and setup,” *Quant. Methods Psychol.*, vol. 16, no. 5, pp. S3–S11, 2020, doi: 10.20982/tqmp.16.5.s003.

[13] 2016 Sandeepraut,July, “How to evaluate Data Science models - Data Science Central,” *Data Science Central*. https://www.datasciencecentral.com/profiles/blogs/how-to-evaluate-data-science-models

[14] R. Wittenauer and L. Smith, “Ischaemic and Haemorrhagic Stroke,” *Prior. Med. Eur. World "A Public Heal. Approach to Innov. Pap. 6.6 Ischaem. Haemorrh. Stroke*, no. December, p. 46, 2012.

[15] D. Faggella, “What is Machine Learning Emerj.” https://emerj.com/ai-glossary-terms/what-is-machine-learning/ (accessed Jul. 21, 2019).

[16] C. Gofton and J. George, *Pathophysiology, diagnosis and management*, vol. 50, no. 10. 2021. doi: 10.31128/AJGP-05-21-5974.

[17] K. Wakefield, “A guide to machine learning algorithms and their applications,” *Sas Uk*, 2019. https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

[18] M. Mohammed, M. B. Khan, and E. B. M. Bashie, *Machine learning: Algorithms and applications*, no. December. 2016. doi: 10.1201/9781315371658.

[19] “Machine learning algorithms explained InfoWorld.” https://www.infoworld.com/article/3394399/machine-learning-algorithms-explained.html (accessed Jul. 22, 2019).

[20] Vikramaditya Jakkula, S. of EECS, and Washington State University, “Tutorial on Support Vector Machine,” *Appl. Comput. Math.*, vol. 6, no. 4, p. 13, 2016, doi: 10.11648/j.acm.s.2017060401.11.

[21] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: A methodology review,” *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, 2002, doi: 10.1016/S1532-0464(03)00034-0.

[22] G. (University of V. Yunyong, “A Cloud Computing Based Platform for Geographically Distributed Health Data Mining Supervisory Committee A Cloud Computing Based Platform for Geographically Distributed Health Data Mining,” 2013.

[23] G. Rohith, “Support Vector Machine — Introduction to Machine Learning Algorithms.” p. 1, 2018. [Online]. Available: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[24] M. A. R. Khalid, M. Alwaqdani, and M. A. H. Farquad, “Comparative Analysis of Support Vector Machine: Employing Various Optimization Algorithms,” *Proc. - 2015 14th Int. Conf. Inf. Technol. ICIT 2015*, no. 1, pp. 171–174, 2016, doi: 10.1109/ICIT.2015.52.

[25] S. Chao and F. Wong, “An incremental decision tree learning methodology regarding attributes in medical data mining,” *Proc. 2009 Int. Conf. Mach. Learn. Cybern.*, vol. 3, no. July, pp. 1694–1699, 2009, doi: 10.1109/ICMLC.2009.5212333.

[26] C. Sehra, “Decision Trees Explained Easily - Chirag Sehra - Medium,” *Medium.Com*. p. 1, 2018.

[27] NPTEL and A. VIDHYA, “What is a Decision Tree ? How does it work.” pp. 1–20, 2016.

[28] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random forests and decision trees,” *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[29] E. Dritsas and M. Trigka, “Stroke Risk Prediction with Machine Learning Techniques,” *Sensors*, vol. 22, no. 13, 2022, doi: 10.3390/s22134670.

[30] S. G. Abdissa, “Guidelines on Clinical and Programmatic Management of Major Non Communicable Diseases Ethiopian National Guideline on Major NCDs 2016,” no. December, 2018, doi: 10.13140/RG.2.2.24757.06889.

[31] T. W. Abate, B. Zeleke, A. Genanew, and B. W. Abate, “The burden of stroke and modifiable risk factors in Ethiopia: A systemic review and meta-analysis,” *PLoS One*, vol. 16, no. November, pp. 1–19, 2021, doi: 10.1371/journal.pone.0259244.

[32] C. M. Alemayehu, “Assessment of Stoke Patients: Occurrence of Unusually High Number of Haemorrhagic Stroke Casesin Tikur Anbessa Specialized Hospital, Addis Ababa, Ethiopia,” *Clin. Med. Res.*, vol. 2, no. 5, p. 94, 2013, doi: 10.11648/j.cmr.20130205.11.

[33] Manaj Jhalani, “Guidelines for prevention and management of stroke,” *Natl. Program. Prev. Control Cancer, Diabetes, Cardiovasc. Dis. Stroke (NPCDCS), Gov. India Guidel. Prev. B. C. V, Silva, D. A. De, Macleod, M. R., Coutts, S. B., Schwamm, L. H., Davis, S. M., D*, no. 61, pp. 1–16, 2019.

[34] R. S. and C. O’Neil, *[Rachel_Schutt,_Cathy_O’Neil]_Doing_Data_Science*, 1st ed. O’Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472., 2014.

[35] Z. Karimi, “Confusion Matrix,”, *research gate*, no. October, pp. 0–4, 2021.

[36] “Confusion matrix,” *Dr.P.K.Chaurasia*, Mahatema Gandi Central University, 2019

[37] V. L. Feigin *et al.*, “World Stroke Organization (WSO): Global Stroke Fact Sheet 2022,” *Int. J. Stroke*, vol. 17, no. 1, pp. 18–29, 2022, doi: 10.1177/17474930211065917.

[38] T. Tazin, M. N. Alam, N. N. Dola, M. S. Bari, S. Bourouis, and M. Monirujjaman Khan, “Stroke Disease Detection and Prediction Using Robust Learning Approaches,” *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/7633381.

[39] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, “A Two-Step Data

Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, vol. 10, no. 11, pp. 1–18, 2022, doi: 10.3390/math10111942.

[40] S. Birla, K. Kohli, and A. Dutta, "Machine Learning on imbalanced data in Credit Risk," *7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016*, 2016, doi: 10.1109/IEMCON.2016.7746326.

[41] L. Liu, X. Wu, S. Li, Y. Li, S. Tan, and Y. Bai, "Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, pp. 1–16, 2022, doi: 10.1186/s12911-022-01821-w.

[42] M. Kuhn and K. Johnson, *Applied predictive modeling*. 2013. doi: 10.1007/978-1-4614-6849-3.

[43] P. Verhagen, "Predictive Modeling," *Encycl. Archaeol. Sci.*, pp. 1–3, 2018, doi: 10.1002/9781119188230.saseas0475.

[44] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.*, vol. 10, no. 1, 2010, doi: 10.1186/1472-6947-10-16.

[45] G. Fekadu, L. Chelkeba, and A. Kebede, "Risk factors, clinical presentations and predictors of stroke among adult patients admitted to stroke unit of Jimma university medical center, south west Ethiopia: prospective observational study," *BMC Neurol.*, vol. 19, no. 1, pp. 1–11, 2019, doi: 10.1186/s12883-019-1409-0.

Appendix I:

Sample codes

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.preprocessing import LabelEncoder

from sklearn.preprocessing import OneHotEncoder

from sklearn.preprocessing import MinMaxScaler

from sklearn.preprocessing import QuantileTransformer


#Loading the collected patient data

dataset = pd.read_csv('Patient_dat_final.csv', low_memory=False)

dataset.info()

dataset.describe()

Dataset

#number of missing data points

dataset.isnull().sum()


#checking missing values using isnull()

dataset.isnull()


#drop missing value of all rows
```

```
data_final = dataset.dropna(how='any',axis=0)
data_final
data_final.shape
data_final.isnull().sum()

#data_final.to_csv("data_final_stroke.csv", index=False)
# after removing the 162 missing values rows
dataset_final= pd.read_csv('data_final_stroke.csv',low_memory=False)
datagen=dataset_final.loc[:,['Sex']]
datagen

#translate catigorical data point to numerical using LabelEncoder
from sklearn.preprocessing import LabelEncoder
label_encoder = LabelEncoder()
sexenc= label_encoder.fit_transform(datagen)
la=['Sex']
sexenc= pd.DataFrame(sexenc,columns=la)
sexenc
datlocation=dataset_final.loc[:,['Location']]
datlocation
datlocation['Location'].unique()
array(['Addis Ababa ', 'SNNP', 'Amhara', 'Dire Dawa ', 'Oromiya',
       'Somali', 'Afar', 'Tigray', 'Gambela ', 'Beneshangul'], dtype=object)
```

```
#encoding using Using dummies values approach
datlocation_trans = pd.get_dummies(datlocation,columns=['Location'],prefix=['Location'])
datlocation_trans
datlocation['Location'].unique()
array(['Addis Ababa ', 'SNNP', 'Amhara', 'Dire Dawa ', 'Oromiya',
       'Somali', 'Afar', 'Tigray', 'Gambela ', 'Beneshangul'],  dtype=object)


#datlocation_trans.to_csv("datlocation_trans.csv", index=False)
Y=dataset_final.loc[:,[ 'Label']]
import seaborn as sns
from matplotlib import pyplot
plt.figure(figsize=(5,4))
g =sns.countplot(Y['Label'], palette='Set1', lw=0.3,  ec=['Yellow'])
for p in g.patches:
    w,h,= p.get_width(),p.get_height()
    x,y = p.get_xy()
    g.text(x+w/2, y+h , '{:.1f}%'.format(100*h/9209),horizontalalignment='center')
plt.xlabel('Class label')
plt.ylabel('Count')
pyplot.show()
from sklearn.preprocessing import LabelEncoder
label_encoder1 = LabelEncoder()
```

```
dlable= label_encoder1.fit_transform(Y)
la=['Label']
y = column_or_1d(y, warn=True)
Y= pd.DataFrame(dlable,columns=la)
Y
dataset_final.drop(['Sex','Label','Location'],axis=1, inplace=True)
hitd_dat = dataset_final
dataset_final
dataset_final = pd.concat([dataset_final,sexenc,datlocation_trans],axis=1)
dataset_final
dataset_final.columns
Index(['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',
       'Pulse Rate', 'RBC', 'FBS', 'Sex', 'Location_Addis Ababa ',
       'Location_Afar', 'Location_Amhara', 'Location_Beneshangul',
       'Location_Dire Dawa ', 'Location_Gambela ', 'Location_Oromiya',
       'Location_SNNP', 'Location_Somali', 'Location_Tigray'], dtype='object')


#creating new columns to rearange the colums position
new_cols= ['Age', 'Weight', 'Sex', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',
       'Pulse Rate', 'RBC', 'FBS',
       'Location_Addis Ababa ', 'Location_Afar', 'Location_Amhara',
       'Location_Beneshangul', 'Location_Dire Dawa ', 'Location_Gambela ',
       'Location_Oromiya', 'Location_SNNP', 'Location_Somali', 'Location_Tigray']
```

```
dataset_final=dataset_final[new_cols]

dataset_final

dataset_final= pd.concat([dataset_final,Y],axis=1)

dataset_final

#dataset_final.to_csv("dataset_spatient_after_encode.csv", index=False)


#check the distribution of patient raw data using histogram the data has normal distrbution or not normal

import matplotlib.pyplot as plt

num_column = ['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure', 'Pulse Rate', 'RBC', 'FBS']

hitd_dat[num_column].hist(figsize=(8,6),grid=True,color='Brown');

plt.tight_layout()

plt.show()


#applying data transformation method by using QuantileTransformer

from sklearn.preprocessing import QuantileTransformer

qt = QuantileTransformer(output_distribution= 'normal')

dataset_qrt=qt.fit_transform(hitd_dat)

dataset_qrt=pd.DataFrame(dataset_qrt)

dataset_qrt.columns = ['Age', 'Weight', 'Height', 'BMI', 'Cholestrol', 'Blood Pressure',

    'Pulse Rate', 'RBC', 'FBS']

dataset_qrt
```