VISVESVARAYA TECHNOLOGICAL UNIVERSITY

Jnana Sangama, Belgaum-590018

A PROJECT REPORT **(15CSP85)** ON

"Prediction of Stroke Using Machine Learning"

Submitted in Partial fulfillment of the Requirements for the Degree of

Bachelor of Engineering in Computer Science & Engineering

By

SHASHANK H N (1CR16CS155)

SRIKANTH S (1CR16CS165)

THEJAS A M (1CR16CS173)

KUNDER AKASH (1CR16CS074)

Under the Guidance of,

Prof. Kartheek G C R

Asst. Prof, Dept. of CSE

CELEBRATING 25 YEARS

CMRIT

CMR INSTITUTE OF TECHNOLOGY, BENGALURU.

ACCREDITED WITH A+ GRADE BY NAAC

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

CMR INSTITUTE OF TECHNOLOGY

#132, AECS LAYOUT, IT PARK ROAD, KUNDALAHALLI, BANGALORE-560037

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the project work entitled **"Prediction of Stroke Using Machine Learning"** carried out by **Shashank H N (1CR16CS155), Srikanth S (1CR16CS165), Thejas A M (1CR16CS173), Kunder Akash (1CR16CS074)**, bonafide students of CMR Institute of Technology, in partial fulfillment for the award of **Bachelor of Engineering** in Computer Science and Engineering of the Visveswaraiah Technological University, Belgaum during the year 2019-2020. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the Report deposited in the departmental library.

The project report has been approved as it satisfies the academic requirements in respect of Project work prescribed for the said Degree.

Prof. Kartheek G C R	**Dr. Prem Kumar Ramesh**	**Dr. Sanjay Jain**
Asst. Prof, Dept. of CSE	**Professor & Head**	**Principal**
Dept. of CSE, CMRIT	**Dept. of CSE, CMRIT**	**CMRIT**

DECLARATION

We, the students of Computer Science and Engineering, CMR Institute of Technology, Bangalore declare that the work entitled **" Prediction of Stroke Using Machine Learning "** has been successfully completed under the guidance of Prof. Kartheek G C R, Computer Science and Engineering Department, CMR Institute of technology, Bangalore. This dissertation work is submitted in partial fulfillment of the requirements for the award of Degree of Bachelor of Engineering in Computer Science and Engineering during the academic year 2019 - 2020. Further the matter embodied in the project report has not been submitted previously by anybody for the award of any degree or diploma to any university.

Place:

Date:

Team members:

SHASHANK H N (1CR16CS155) ____________________

SRIKANTH S (1CR16CS165) ____________________

THEJAS A M (1CR16CS173) ____________________

KUNDER AKASH (1CR16CS074) ____________________

ABSTRACT

Stroke is the second leading cause of death worldwide and remains an important health burden both for the individuals and for the national healthcare systems.

our project applies principles of machine learning over large existing data sets to effectively predict the stroke based on potentially modifiable risk factors.

Then it intended to develop the application to provide a personalized warning based on each user’s level of stroke risk and a lifestyle correction message about the stroke risk factors.

ACKNOWLEDGEMENT

I take this opportunity to express my sincere gratitude and respect to **CMR Institute of Technology, Bengaluru** for providing me a platform to pursue my studies and carry out my final year project

I have a great pleasure in expressing my deep sense of gratitude to **Dr. Sanjay Jain,** Principal, CMRIT, Bangalore, for his constant encouragement.

I would like to thank **Dr. Prem Kumar Ramesh,** Professor and Head, Department of Computer Science and Engineering, CMRIT, Bangalore, who has been a constant support and encouragement throughout the course of this project.

I consider it a privilege and honor to express my sincere gratitude to my guide **Kartheek G.C.R, Asst. Prof, Dept. of CSE,** Department of Computer Science and Engineering, for the valuable guidance throughout the tenure of this review.

I also extend my thanks to all the faculty of Computer Science and Engineering who directly or indirectly encouraged me.

Finally, I would like to thank my parents and friends for all their moral support they have given me during the completion of this work.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

PREAMBLE

1.1 Introduction

Stroke is the second leading cause of death worldwide and one of the most life-threatening diseases for persons above 65 years. It injures the brain like “heart attack” which injures the heart. Once a stroke disease occurs, it is not only cost huge medical care and permanent disability but can eventually lead to death. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage.

Stroke is a blood clot or bleed in the brain which can make permanent damage that has an effect on mobility, cognition, sight or communication. Stroke is considered as medical urgent situation and can cause long-term neurological damage, complications and often death. The majority of strokes are classified as ischemic embolic and Hemorrhagic. An ischemic embolic stroke happens when a blood clot forms away from the patient brain usually in the patient heart and travels through the patient bloodstream to lodge in narrower brain arteries. Hemorrhagic stroke is considered another type of brain stroke as it happens when an artery in the brain leaks blood or ruptures.

Strokes are sudden but many of the disease processes that precede them take a long time to develop. This is why age is the most clear-cut risk factor for stroke: the chance of blockage or breakage rises with every passing year, so – although it can strike at any age – stroke is much more likely the older we get.

The stroke risk factors included in the profile are age, systolic blood pressure, BMI, cholesterol, diabetes, smoking status and intensity, physical activity, alcohol drinking, past history (hypertension, coronary heart disease) and family history (stroke, coronary heart disease). On the basis of the risk factors in the profile, which can be readily determined on routine physical examination in a physician's office, stroke risk can be estimated. An individual's risk can be related to the average risk of stroke for persons of the same age and sex.

1.2 Problem Statement

Stroke is the second leading cause of death worldwide and remains an important health burden both for the individuals and for the national healthcare systems. Potentially modifiable risk factors for stroke include hypertension, cardiac disease, diabetes, and dysregulation of glucose metabolism, atrial fibrillation, and lifestyle factors.

Therefore, the goal of our project is to apply principles of machine learning over large existing data sets to effectively predict the stroke based on potentially modifiable risk factors. Then it intended to develop the application to provide a personalized warning based on each user's level of stroke risk and a lifestyle correction message about the stroke risk factors.

1.3 Relevance of the problem

Each year in the US, approximately 795,000 people have a new or recurrent stroke. In the US, every 40 seconds someone suffers a stroke, and every 4 minutes, someone dies of a stroke. Stroke is the 5th leading cause of death and the #1 cause of disability in the U.S. Most strokes are preventable, and if treated early, the likelihood of a good outcome after stroke can be significantly improved.

The most common type of stroke is an ischemic stroke, also called a cerebral infarction. A blood vessel supplying oxygen and nutrients to the brain becomes blocked, causing brain cell death. The damage is permanent because these cells cannot be replaced. However, the brain can adjust, so many patients improve, and some have no permanent disability. The other type of stroke is a cerebral hemorrhage, when a blood vessel in the brain ruptures, causing bleeding and damage to the brain tissue.

The strongest risk factor for both types of stroke is high blood pressure, also called hypertension. Other common risk factors for stroke include diabetes, an irregular heart rhythm called atrial fibrillation, high cholesterol, smoking, physical inactivity, a family history of stroke and chronic kidney disease. So, by predicting the chances of stroke based on these risk factors may save the lives of many.

1.4 Phase Description

Phase	Task	Description
Phase 1	Analysis	Analyzing the core of the IEEE paper and provide Literature review based on analysis.
Phase 2	Literature survey	Collect raw data and elaborate on literature surveys.
Phase 3	System analysis	Analyses the requirements of the project and lists the specific requirements needed.
Phase 4	Design	Object designing and Functional description
Phase 5	Implementation	Implement the code based on the object specification
Phase 6	Testing	Test the project according to Test Specification
Phase 7	Documentation	Prepare the document for this project with conclusion and future enhancement.

Table 1.6: Phase Description

1.5 Organization of the project report

The project report is organized as follows:

Chapter 2: **Literature Survey** - Gives a brief overview of the survey papers and the research sources that have been studied to establish a thorough understanding of the project under consideration.

Chapter 3: **Theoretical Background** - Establishes groundwork for the proposed project by giving a detailed analysis of the project topic, existing research relevant to the project, arguments in favor and against the existing solutions and finally explores the motivation behind the proposed solution.

Chapter 4: **System Requirement Specification** - Discusses in details about the different kinds of requirements needed to successfully complete the project.

Chapter 5: **System Analysis** - gives details about several analysis that are performed to facilitate taking decision of whether the project is feasible enough or not.

Chapter 6: **System Design** - Gives the design description of the project, conceptual and detailed design well supported with design diagrams.

Chapter 7: **Implementation** - Discusses the implementation details of the project and reasons the use of the programming language and development environment.

Chapter 8: **Testing** - Briefs the testing methods used for testing the different modules in the project.

Chapter 9: **Results and Performance Analysis** - Gives the snapshots and graphs of the proposed protocols.

Chapter 10: **Conclusion and Future Scope** - Gives the concluding remarks of the project, throwing light on its future aspects.

References: Lists the websites and references referred during the project work.

Chapter 2

LITERATURE SURVEY

2.1 Burden of Stroke in the World

Stroke is the second leading cause of death and leading cause of adult disability worldwide with 400-800 strokes per 100,000, 15 million new acute strokes every year, 28,500,000 disability adjusted life-years and 28-30-day case fatality ranging from 17% to 35%. The burden of stroke will likely worsen with stroke and heart disease related deaths projected to increase to five million in 2020, compared to three million in 1998. This will be a result of continuing health and demographic transition resulting in increase in vascular disease risk factors and population of the elderly. Developing countries account for 85% of the global deaths from stroke. The social and economic consequences of stroke are substantial. The cost of stroke for the year 2002 was estimated to be as high as $49.4 billion in the United States of America (USA), while costs after discharge were estimated to amount to 2.9 billion Euros in France.

2.2 Burden of Stroke in Africa

A systematic review of the existing literature to examine the burden and profile of stroke in the WHO African region reported an annual incidence rate of stroke of up to 316 per 100,000, a prevalence rate of 315 per 100,000 and a three-year fatality of up to 84% in Africa. Disabling stroke prevalence may be at least as high as in high-income areas. In 2002, model-based estimated age-adjusted stroke mortality rates ranged between 168 and 179 per 100,000 population for countries in the African region. Case–fatality data available from three hospital based urban stroke registers in South Africa (two South African and one from Zimbabwe) found 30 day case fatality ranging between 33 and 35%. Given the economic burden of stroke in the developed countries, a small fraction of such amounts can cause enormous economic damage to low income countries especially in SSA, given the younger age at which stroke occurs. A study done in Togo estimated direct cost per person of 936 Euros in only 17 days, about 170 times more than the average annual health spending of a Togolese.

2.3 Burden of Stroke in Uganda

The actual burden of stroke in Uganda is not known. According to WHO estimates for heart disease and stroke 2002, stroke was responsible for 11 per 1000 population (25,004,000) 4 disability adjusted life years and mortality of 11,043. Stroke is one of the common neurological diseases among patients admitted to the neurology ward at Mulago, Uganda's national referral hospital accounting for 21% of all neurological admissions. Unpublished research done at Mulago hospital, showed a 30-day case fatality of 43.8% among 133 patients admitted with stroke. The economic burden caused by stroke has not been explored in Uganda but given the very high dependent population (53%), high prevalence of HIV/AIDS, drug resistant TB and Malaria, the impact of stroke and other emerging non-communicable diseases on the resource limited economy is astronomical.

2.4 Paper Survey

In order to get required knowledge about various concepts related to the present analysis existing literature were studied. Some of the important conclusions were made through those are listed below.

1. **"*Computer Methods and Programs in Biomedicine*" - Jae–woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho –** The Purpose of this paper was Calculation of 10-year stroke prediction probability and classifying the user's individual probability of stroke into five categories.

2. **"*Probability of Stroke: A Risk Profile from the Framingham Study*" - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD -** In this paper, A health risk appraisal function has been developed for the prediction of stroke using the Framingham Study cohort.

3. **"*Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study*" - Min SN, Park SJ, Kim DJ, Subramaniyam M, Lee KS -** In this research, this paper aimed to derive a model equation for developing a stroke pre-diagnosis algorithm with the potentially modifiable risk factors.

4. "*Stroke prediction using artificial intelligence*"- **M. Sheetal Singh, Prakash Choudhary -** In this paper, Here, decision tree algorithm is used for feature selection process, principle component analysis algorithm is used for reducing the dimension and adopted back propagation neural network classification algorithm, to construct a classification model.

5. "*Medical software user interfaces, stroke MD application design (IEEE)*" **Elena Zamsa-**The article presents the design of an application interface for associated medical data visualization and management for neurologists in a stroke clustering and prediction system called Stroke MD.

6. **"*Focus on stroke: Predicting and preventing stroke*" Michael Regnier-** This paper focuses on cutting-edge prevention of stroke.

7. "*Effective Analysis and Predictive Model of Stroke Disease using Classification Methods*"**-A.Sudha, P.Gayathri, N.Jaisankar-** This paper, principle component analysis algorithm is used for reducing the dimensions and it determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not.

8. "*Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study*" **- Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal** - In this paper Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. This article aimed to develop and validate a set of deep learning algorithms for automated detection.

Chapter 3

THEORITICAL BACKGROUND

Theoretical background highlighting some topics related to project work. The description contains several topics which are worth to discuss and highlight some of their limitation that encourage going on finding solution as well as highlights some of their advantages for which reason these topics and their features are used in this project.

3.1 Causes of stroke and subtypes

Stroke or a cerebral vascular accident is the sudden death of brain cells due to inadequate blood flow and oxygen resulting from a blood clot occluding an artery in the brain or a blood vessel rupturing. When either of these things happens, brain cells begin to die and brain damage occurs. About two million brain cells die every minute during stroke with loss of abilities controlled by that area of the brain which include speech, movement and memory. A high number of dead brain cells are associated with increased risk of permanent brain damage, disability or death. Stroke is divided into two broad categories that define its pathophysiology: Ischemic stroke, Hemorrhagic stroke.

3.1.1 Ischemic stroke

Ischemic stroke is caused by blockage of arteries by blood clots or by the gradual build-up of plaque and other fatty deposits. These may originate from the affected vessel or may embolize from other intracranial and extracranial vessels or from the aortic arch. Majority of emboli originate from the heart as a result of valvular heart disease, arrhythmias, ischemic heart disease, bacterial and non-bacterial endocarditis and cardiomyopathies among others. Regarding the prevalence of ischemic stroke, it accounts for 85-90% of all strokes in the developed world with average age between 70-80 years. In Africa on the other hand, several studies have shown prevalence of ischemic stroke as low as 10-40%. An in-hospital study at Mulago hospital found 77.6% prevalence of ischemic stroke, mean age 62.2+ 16.5 years. A large multicenter case control study that involved 3000 stroke patients from 22 countries including low and middle income, showed 78% prevalence of ischemic stroke.

3.1.2 Hemorrhagic stroke

Hypertension is the most important risk factor for vessel rupture in the brain leading to haemorrhagic stroke. Other causes of rupture of vessel walls include cerebral aneurysms, arteriovenous malformations use of anticoagulants and vasculitis. Regarding the prevalence of haemorrhagic stroke, it accounts for 10-15% of all strokes in the developed world and is responsible for more than 30% of all stroke deaths. In Africa on the other hand, several studies have shown prevalence as high as 20-60%. In Pretoria South Africa, 32.8% of 116 stroke patients had haemorrhagic stroke on brain CT scan. Nakibuuka J et al 2012 found 22.4% prevalence of haemorrhagic stroke among 85 adult stroke patients at a Mulago hospital. The effects of a stroke depend on where the stroke occurs in the brain and how much the brain is damaged, but the clinical symptoms of stroke do not accurately predict its underlying cause or causes. Classic stroke symptoms include the acute onset of unilateral paralysis, loss of vision, speech impairment, memory loss, impaired reasoning ability, coma, or death.

3.2 Causes of mortality from stroke

Death from stroke is as a result of co-morbidities and/ or complications. Complications of stroke may arise at different time periods. The beginning of stroke symptoms and the first month following the stroke onset is the most critical period for survival with the highest number of fatalities in the first week. Complications of stroke include hyperglycemia, hypoglycemia, hypertension, hypotension, fever, infarct extension or rebleeding, cerebral oedema, herniation, coning, aspiration, aspiration pneumonia, urinary tract infection, cardiac dysrhythmia, deep venous thrombosis and pulmonary embolism among others. During the first week from stroke onset, death is usually due to transtentorial herniation and haemorrhage, with death due to haemorrhage happening within the first three days and death due to cerebral infarction usually occurring between the third to sixth day. One week after the onset of stroke, death is usually due to complications resulting from relative immobility such as pneumonia, sepsis and pulmonary embolism.

Different studies have found varied factors associated with stroke mortality in their setting. For example, the most common predictors of death from stroke for those aged more than 65 years of age reported by Mackay included previous stroke, atrial fibrillation and hypertension. Nigeria 6 reported a 12.6% 30-day case fatality of all strokes. Among patients with hemorrhagic stroke: fixed dilated pupil(s), a Glasgow coma score of less than 10 on admission, swallowing difficulties at admission, fever, lung infection, and no aspirin treatment were independent risk factors for a lethal outcome. Yikona J et al also observed that stroke severity, neurological deterioration during hospitalisation, non-use of antithrombolytics during hospital admission and lack of assessment by a stroke team were the most consistent predictors of case fatality at seven days, 30 days and one year after stroke. In Pretoria, South Africa, case fatality at 30 days was much higher, 22% for ischemic stroke, 58% for cerebral hemorrhagic stroke and hypertension was significantly associated with stroke. At Mulago hospital, 30 day case fatality of 43.8% was reported among 133 patients (mean age 65.8+ 15.8 years) with, fever > 37.50 (OR 2.81 (95%CI; 1.2-6.6) and impaired level of consciousness with a GCS <9 (OR0.13 95%CI; 0.005-0.35) significantly associated with increased mortality.

3.2.1 Complications of stroke

(a) Hyperglycemia/hypoglycemia

Hyperglycemia at the time of acute stroke is associated with poorer clinical outcomes, infarct progression and increased mortality especially in the first month of stroke, with non-diabetic patients more affected compared to diabetics. It is also associated with reduced functional recovery. Plasma glucose levels above eight mmol/L after acute stroke predicts a poor prognosis after correcting for age, stroke severity and stroke subtype and should be treated actively. The American Heart and Stroke Associations (AHA/ASA), The National Stroke foundation of Australia (NSF 2010), The National Institute for Clinical Excellence recommend cautious treatment of patients with glucose concentrations greater than 8-11mmol/L with subcutaneous insulin. Hypoglycemia on the other hand may cause focal neurological deficits that can be reversed by treatment.

(b) Fever

High body temperature within 24hrs from stroke onset is associated with poor outcome and large cerebral infarcts. High temperature seems to be a major determinant even for long-term mortality after stroke.

Hyperthermia acts through several mechanisms to worsen brain ischemia including: enhanced release of neurotransmitters, exaggerated oxygen radical production, more extensive blood brain barrier break down, increased numbers of potentially damaging ischemic depolarizations in the focal ischemic penumbra. Most common 7 causes include chest and urinary tract infections. Regular paracetamol and/or physical cooling measures are reportedly adequate.

(c) Blood pressure

Both hyper and hypotension in the first 24 hours after stroke are associated with poor outcome, and poor short and long-term prognosis. Hypertension may indicate oedema, hemorrhage, and increase in the risk of primary ICH or hypertensive encephalopathy. Many hypertensive patients also have pre-existing hypertension that may or may not have been treated prior to the stroke. According to AHA/ASA, for every ten-mmHg increase above 180 mmHg, the risk of neurological deterioration increases by 40% and the risk of poor outcome increases by 23%. A study in AHA/ASA, found that an elevated baseline mean arterial BP was not independently associated with poor outcomes, but elevations in mean BP over the first days after stroke were. They also found that in most patients, the BP spontaneously decreases over 4-10 days from stroke onset. BP changes may occur as a result of disturbed cardiovascular autonomic regulation, with changes in absolute BP levels and BP variability both possible.

A Cochrane review (65 Randomized Clinical Trials) concluded that insufficient data exists to evaluate BP lowering post-stroke. Recommended based practice is based on clinical experience and expert opinion with the AHA guidelines recommending starting or increasing anti-hypertensives in ischemic stroke if systolic blood pressure >220 mmHg or diastolic blood pressure >120mmHg, unless end-organ damage is due to high BP. Outside of organ dysfunction, the BP should be cautiously lowered by not more than 10-

20%, 15-25% in 24 hours (AHA>ASA). In acute ICH on the other hand, antihypertensive drugs including intravenously administered ones, can be used to maintain SBP<180mmHg (MAP 130mmHg). In the absence of intracranial pressure, neurosurgeons recommend 160/90mmHg (MAP 100mm Hg)

(d) Dysphagia

Dysphagia occurs in 27-55% of patients with new onset stroke. Only about 50% of those affected recovers normal swallowing by 6 months. It is associated with increased risk of complications such as aspiration, aspiration pneumonia, dehydration and malnutrition, hence early screening to prevent these complications. The National Stroke 8 Foundation recommends that all patients should be screened for swallowing deficits before being given food, drink or oral medications. Screening should be undertaken by personnel specifically trained in swallow screening and a failed bedside screen should be followed by a complete assessment by a speech pathologist prior to any oral ingestion.

3.3 Acute stroke management

There have been major advances in the treatment of acute stroke in recent years Fundamental to these advances have been; firstly, the use of thrombolysis in ischemic stroke; secondly protocol driven multi-disciplinary care in stroke units to improve survival, independence and quality of life; and thirdly development of national guidelines to assist in protocol development and standardization of care. When suspected of having a stroke or TIA, a rapid stroke screen is done to facilitate early referral to a stroke unit where available and also in eligible patients, being thrombolysed where possible.

A stroke unit is an area within a hospital where stroke patients are managed by a coordinated multidisciplinary team specialising in stroke management. There is overwhelming evidence (31 RCT) that stroke unit care significantly reduces death & disability after stroke compared with conventional care in general wards for all people with stroke OR 0.82, 95% CI 0.73-0.92. Stroke units in low resource settings have also been associated with positive patient outcomes. There are no stroke units in a hospital based epidemiological study on stroke reported that patients suspected to have stroke presented on average two days post ictal. It is recommended that in absence of stroke units, hospitals should adhere as closely as possible to the criteria of stroke unit care by use of the acute stroke care pathway. This pathway involves rapid assessment using validated tools, imaging using CT scan/MRI to confirm the diagnosis, routine investigations and added investigations for selected patients, thrombolysis, neuro and surgical interventions where possible, physiologic monitoring and management (GCS, vital signs), secondary prevention by addressing lifestyles modification, intervention to promote adherence to medications (antihypertensives, antiplatelet, anticoagulation, cholesterol lowering drugs, diabetes management) and lastly rehabilitation.

3.4 Risk factors of stroke

3.4.1 Traditional risk factors associated with stroke

Stroke can occur in anyone regardless of race, gender or age however the chances of having a stroke increase if an individual has certain risk factors that can cause a stroke. The best way to protect oneself and others is to understand personal risk and how to manage it. Studies have shown that 80% of strokes can be prevented in this way.

Stroke risk factors are divided into modifiable and non-modifiable. The modifiable risk factors are further subdivided into lifestyle risk factors or medical risk factors. Lifestyle risk factors which include smoking, alcohol use, physical inactivity and obesity can often be changed while medical risk factors such as high blood pressure, atrial fibrillation, diabetes mellitus and high cholesterol can usually be treated. A large multicenter (INTERSTROKE) case control study showed that there are ten factors that are associated with 90% of stroke risk and half of these are modifiable. Non-modifiable risk factors on the other hand though they cannot be controlled, they help to identify individuals at risk for stroke.

Types of Risk Factors

(1) Modifiable/ controllable risk factors Hypertension

(a) Hypertension

Hypertension is the force by which blood pushes against the arteries. If left untreated it can 10 weaken blood vessels and damage major organs such as the brain and lead to a stroke by accelerating atheroma and thrombus formation resulting into infarcts of large vessels, hyalinosis and fibrin deposition with resultant infarction of small vessel lacunes. Hyalinisation within the small cerebral vessels results in the formation of charcot bouchard micro aneurysms that result into intraparenchymal haemorrhage from perforating vessels. It also leads to rupture of diseased vessels and arterial venous malformations. In autopsy series, hypertension accounts for 40-50% of patients dying of non- traumatic haematomas.

Large prospective cohort studies and subsequent systematic reviews have shown hypertension to be the most powerful modifiable predictor of all strokes with estimated PAR ranging from 25-50%. There is also evidence to suggest that the PAR for hypertension may be influenced by ethnicity, stroke subtype, geographic location. Furthermore, a meta-analysis of 61 studies involving one million subjects (mostly Caucasian), reported a log-linear relationship between blood pressure and all stroke. As many as 73 million Americans have high blood pressure (one in every four adults) and 31.6% are not aware they have it. High blood pressure is still largely ignored as a public health problem in most developing countries despite a sharp rise in morbidity and mortality from diseases related to hypertension such as stroke. Also, hypertensive patients might not bother to visit health facilities to have their blood pressure taken as it is largely the asymptomatic.

Prevalence of hypertension is not known but isolated community-based studies report prevalence of 37.3-44% in rural and urban populations respectively. A hospital-based study reported 60% prevalence of hypertension among 85 adult stroke patients. Research has shown that less than 20% of hypertension occurs in isolation. Metabolically linked risk factors such as diabetes, obesity, dyslipidemias, all of which lead to amplification of stroke risk, commonly co-exist with hypertension. Excessive alcohol consumption, physical inactivity and a high salt diet can also lead to high blood pressure.

(b) Alcohol

Long-term heavy ingestion of alcohol (more than two drinks daily) is a recognized risk factor for all stroke (RR 1.6; 1.4-2.0), particularly for hemorrhagic stroke (RR 2.2; 1.5-3.2). This may be related to a decrease in HDL levels, reduced fibrinolysis, and increased platelet aggregation. Recent ingestion of large amounts of alcohol within 24 hours preceding stroke onset, which also includes non-habitual drinkers, has also been found to be a risk 11 factor in some studies.

(c) Obesity

Obesity and excessive weight put a strain on the entire circulatory system. It is strongly linked to cardiovascular diseases and type II diabetes mellitus (DM) through the promotion of insulin resistance and other associated physiological derangements, including dyslipidemia, elevated blood pressure and increased left ventricular mass. These lead to degenerative changes of vessel walls secondary to the process of atherosclerosis, fibrinoid necrosis of small arteries and arterioles with resultant cerebral infarction. In SSA, the past two decades have seen a dramatic increase in obesity as the region experiences what is called a 'double burden' of disease with malnutrition on the one hand and a growing prevalence of obesity on the other.

Malnourishment in the form of both starvation and overconsumption of cheap and fried foods is increasingly pandemic in Africa largely due to increasing urbanization as jobs have moved out of rural areas and into cities. While the rate of change in urban overweight/obesity did not significantly differ between the poor and the rich, it was substantially higher among the non-educated women than among their educated counterparts. In South Africa alone, 75% of the black population between the ages of 18 and 65 years and 50% of the white population are either overweight or obese (International Association for the Study of Obesity, IASO). Among certain tribes in Nigeria, women are traditionally fattened up before marriage to make them seem more attractive and healthier to their future husbands.

Heart Institute predicts that obesity-related heart disease will be the leading cause of death in sub-Saharan Africa by 2020. Despite the growing problem, many sub-Saharan Africans do not find the region's increasing waistline to be of concern because of social attitudes that make being overweight seem harmless, if not explicitly attractive, "index of affluence and power is linked 12 to one's size." HIV/AIDS is another unlikely factor in the acceptance of obesity. The virus, known as "slim disease" throughout Africa, is strongly associated with weight loss. So being fat is viewed as a "great thing because it means you don't have HIV". Some Africans purposely gain large amounts of weight to prove that they don't have the disease.

(d) Physical inactivity

Studies have shown a consistent association between physical activity and risk of ischemic and hemorrhagic stroke. The INTERSTROKE case control study (3000 stroke patients from 22 countries) showed regular physical activity (OR 0.69, 0.53-0.90) and a PAR of about 29% which was larger than was reported in INTERHEART for acute myocardial infarction (MI), (12%). People who exercise five or more times per week, have a reduced stroke risk. A meta-analysis of 31 observational studies confined to high income countries, reported that moderately intense physical activity reduced the risk of all strokes (RR 0.6; 0.5-0.9) with a comparable risk reduction for both ischemic and hemorrhagic stroke.

The prevalence of physical inactivity and sedentary behavior among adults is not known however a national survey carried out in 2003 as part of the WHO 24 countries Global School based Student Health Survey on physical inactivity and sedentary behavior among 4,218 school children (2,712 analyzed, mean age 14.9 years) showed more than 80% physical inactivity among both boys and girls, with 26.2% (95% CI 22.8-29.5) sedentary boys and 25.8% (95% CI 22.2-29.3) sedentary girls. A hospital-based study on the epidemiology of stroke among 85 stroke patients found the prevalence of physical inactivity of 40%.

(e) Diabetes mellitus

Diabetes accelerates atherosclerotic vessel disease. The resultant thrombosis, embolism of large arteries and lacunar disease, especially in the elderly is commonly associated with other risk factors. The incidence of type II DM which is known to increase the risk of developing heart disease and stroke has increased among black Africans in recent years. According to the International Association for the Study of Obesity (IASO), 8 to 12% of the black South African population has type II DM, compared with just 4% of the country's white population.

(f) Cigarette smoking

Nicotine contained in cigarettes raises BP which causes rupture of aneurysms and arterial venous malformations. Carbon monoxide produced during smoking reduces the amount of oxygen to the brain and cigarette smoke increases the amount of fibrinogen which increases the chance of clotting. In addition, increased serum proteolytic activity of smokers degrades the collagen of the plaque’s fibrous cap, which makes it susceptible to rupture. In a number of studies, smoking has been shown to be strongly associated with an increased risk for all strokes and subclinical carotid disease, with a graded linear association for the number of cigarettes smoked.

A study of high school students in Kampala (n= 2,789) and the rural tobacco growing town of Arua (n= 1528) found a current smoking prevalence of 5.3% in Kampala compared to 21.9% in Arua. The Global youth tobacco survey carried out in 2007 reported lifetime smoking prevalence rate of 16.6% among 13-15year old high school students. Case control studies have also found an association between environmental tobacco smoke and increased risk of stroke.

(g) Cardiac causes of stroke

Cardio-embolism is an important cause of ischemic stroke worldwide. In high income countries it accounts for 15 to 25% of all ischemic strokes. Limited data from low-income countries supports a similar frequency of cardio-embolism. Important causes of cardioembolism in high income countries include atrial fibrillation (AF), aortic arch disease and myocardial infarction. On the other hand, in low income countries, rheumatic heart disease in addition to factors mentioned above, remains prevalent and is a frequent cause of premature stroke in young patients in SSA. AF is the most important factor in the occurrence of cardio embolism with a four to seven-fold risk of systemic embolization when compared to comparable groups of patients without AF. The risk becomes greater with older patients.

The incidence of emboli varies from 9 to 27% in clinical reports and increases to 41% in necropsy studies. Fleming et al noted a 25% incidence of emboli among 500 14 patients with AF. O'Donell et al 2010 reported that AF was the most common cardiac source of thromboembolism in cases with ischemic stroke 203 (9%), with regional variation in prevalence: 86 (23%) in high-income countries, 14 (13%) in South America, 16 (7%) in Africa, 41 (6%) in India, and 46 (5%) in Southeast Asia. Cardiac aetiology was associated with an increased risk of ischemic stroke, but not intracerebral hemorrhagic stroke (OR 0·90, 0·52-1·56). Studies have identified three clinical predictors of subsequent thromboembolism: hypertension, recent onset of congestive heart failure (CHF) and previous history of thromboembolism.

(h) Dyslipidemias

Hyperlipidemia causes stroke secondary to thrombosis and embolism resulting from degenerative vessel changes due to the process of atherosclerosis. It is arguably the strongest risk factor for coronary heart disease (CHD) with an estimated PAR of over 50%. The relationship between cholesterol and stroke however is much less certain. For ischemic stroke, epidemiological evidence supporting an association has been inconsistent.

A meta-analysis of 45 prospective cohort studies reported no significant association between total cholesterol and ischemic stroke (RR 0.8; 0.6-1.1). However, a large multicenter case control study, found that increased concentration of total cholesterol was not associated with risk of ischemic stroke, but was associated with reduced risk of intracerebral hemorrhagic stroke. Additionally increased concentration of HDL cholesterol was associated with a reduced risk of ischemic stroke, and an increased risk of intracerebral hemorrhagic stroke. A hospital based found 31% prevalence of high total cholesterol (>200mg/dl) among 65 patients with ischemic stroke.

(2) Non-modifiable/uncontrollable stroke risk factors

(a) Age

Advanced age is associated with degenerative changes of the vessel wall resulting from hypertension alone or in connection with atherosclerosis, combined with hemodynamic action and natural weak points in the cerebral vessel wall leading to potential sites for rupture, thrombus formation and thromboembolism. Advanced age is a risk factor for both first and recurrent stroke, with doubling of stroke risk in each decade over age 55 years. There is a seven fold greater risk of dying from stroke than the general population, in 15 more than 65 years age group. Some risk factors tend to be more prevalent with advancing age such as hypertension and diabetes mellitus whereas others are acquired at a younger age like smoking and thus age can be considered a marker for the duration of exposure to a risk factor.

(b) Gender

Male mortality from stroke has been noted to exceed that of females in all age groups less than 70 years, with males having a 2.5 fold increased risk of stroke than females. Differences in frequency of key vascular risk factors have been implicated. Gender differences in risk factor profiles may predict differences in outcomes or responses to therapies therefore prevention strategies and public health efforts need to reflect these differences.

(c) Family history of stroke

This is an independent risk factor for ischemic stroke with onset before 70 years. The association is not only for large and small vessel disease but also for cryptogenic stroke. In a study of a cohort of men, the relative risk of stroke with positive paternal history was 2.4-fold and relative risk of stroke with maternal history of stroke was 1.4-fold.

(d) Previous history of stroke

Stroke recurrence after an initial stroke has varied widely from 3% to 22% at one year and 10% to 53% at five years in different studies.

3.4.2 Emerging Risk Factors

(a) HIV/AIDS

Clinical, radiological and post-mortem studies have found a strong association between HIV/AIDS, ischemic stroke and intracerebral haemorrhage. Vascular abnormalities, coagulation disorders and cardioembolism have been identified as the main causes of stroke among HIV\AIDS patients. HIV-associated dilated cardiomyopathy as a cause of cerebral infarction resulting from cardioembolism is well described in HIV infection.

A prospective analysis of stroke in 35 black South African HIV patients (mean age 32.1 years) reported ischemic stroke with coagulopathy as the commonest cause (60.7%) and, 16 protein S deficiency accounted for the majority (64%). The study also reported stroke to be the first manifestation of HIV infection in 20 out of 35 patients. All the HIV positive patients in this study presented with ischemic stroke.

(b) Psychosocial stress

This appears to be an important risk factor for stroke but studies are limited and the constructs of psychosocial stress are often imprecise. Globally, the PAR of implicated factors such as depression, perceived stress, social isolation and lack of social support for acute myocardial infarction was about 30%. A large multicenter case control study, reported a PAR of 4·6%, (2·1-9·6) for psychosocial stress and PAR 5·2%, (2·7-9·8) for depression with consistent estimates for ischemic and intracerebral hemorrhagic stroke. Depression was associated with an increased risk of all stroke and ischemic stroke, but not intracerebral haemorrhagic stroke (OR1·11, 0·82-1·52).

(c) Sickle cell disease

Sickle cell disease has recently been recognized as a problem of major public-health significance by the WHO with more than 70% of sufferers living in Africa. Stroke in various forms is the most common neurological complication known to occur in these patients. Vascular occlusions by sickled cells give rise to the pathological changes underlying the common neurological complications in this disease. These changes tend to be multiple and repetitive. They give rise to very varied and recurrent clinical manifestations of this complication. Nantulya et al, 1989, found 24.4% neurological complications among 25 children with SCA in Aga Khan Hospital in Nairobi, Kenya.

(d) Diseases of the vessel wall

Vasculitides such as polyarteritis nodosa, temporal arteritis; collagen vascular diseases like systemic lupus erythematosus, rheumatoid arthritis, syphilitic vasculitis, are known causes of stroke. There is activation of complement cascade by antigen-antibody complexes 17 lodging into gaps between endothelial vessels, with resultant production of lysozymes and destruction of vessel wall with haemorrhage and fibrinoid necrosis.

(e) Hypercoagulable states

Disseminated intravascular coagulation, haemoglobinopathies, thrombocytopenia, antiphospholipid antibody syndrome, hyperhomocysteinemia, deficiency of antithrombin III, Protein S and C and hyperfibrinogenemia lead to disturbance of normal properties of blood increasing the risk of cerebral infarction by formation of clots and cerebral hemorrhage. Antiphospholipid antibody syndrome and hyperhomocysteinemia associated with arterial thrombosis are particularly known to significantly increase risk of infarction Most epidemiological studies have reported a consistent and linear association between elevated levels of homocysteine and risk of ischemic stroke with meta-analyses reporting a 1.5 to 2 fold increase in risk.

3.5 Prevention of stroke

More than 70% of strokes are first events, thus making primary stroke prevention a particularly important aspect. Interventions should be targeted at behaviour modification, which however requires information about the baseline perceptions, knowledge and prevalence of risk factors in defined populations.

Summary

This chapter mainly concentrates on the basic theoretical background related to the topic of focus. It describes the survey work done with respect to the project, which includes the burden of stroke, causes of stroke and subtypes, causes of mortality from stroke and risk factors of stroke. Finally, it concludes by discussing various aspects for prevention of stroke. This survey work also finds out some of the major flaws associated with all these related topics, so that a suitable solution can be proposed in order to overcome these drawbacks.

Chapter 4

SYSTEM REQUIREMENT SPECIFICATION

Software requirement Specification is a fundamental document, which forms the foundation of the software development process. It not only lists the requirements of a system but also has a description of its major feature. An SRS is basically an organization's understanding (in writing) of a customer or potential client's system requirements and dependencies at a particular point in time (usually) prior to any actual design or development work. It's a two-way insurance policy that assures that both the client and the organization understand the other's requirements from that perspective at a given point in time.

The SRS also functions as a blueprint for completing a project with as little cost growth as possible. The SRS is often referred to as the "parent" document because all subsequent project management documents, such as design specifications, statements of work, software architecture specifications, testing and validation plans, and documentation plans, are related to it. It is important to note that an SRS contains functional and nonfunctional requirements only; it doesn't offer design suggestions, possible solutions to technology or business issues, or any other information other than what the development team understands the customer's system requirements to be.

4.1 Functional Requirement

Functional Requirement defines a function of a software system and how the system must behave when presented with specific inputs or conditions. These may include calculations, data manipulation and processing and other specific functionality. In this system following are the functional requirements: -

- Machine Learning Methodology
- Asset Visualization

4.1.1 Machine Learning Methodology

Using this methodology, the modeler can discover the “performance ceiling” for the data set before settling on a model. In many cases, a range of models will be equivalent in terms of performance so the practitioner can weigh the benefits of different methodologies.

Few methodologies used in our projects are:

- Decision Tree
- Naïve Bayes
- Artificial Neural Network

(a) Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

Decision tree is one of the important methods for handling high dimensional data. Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods.

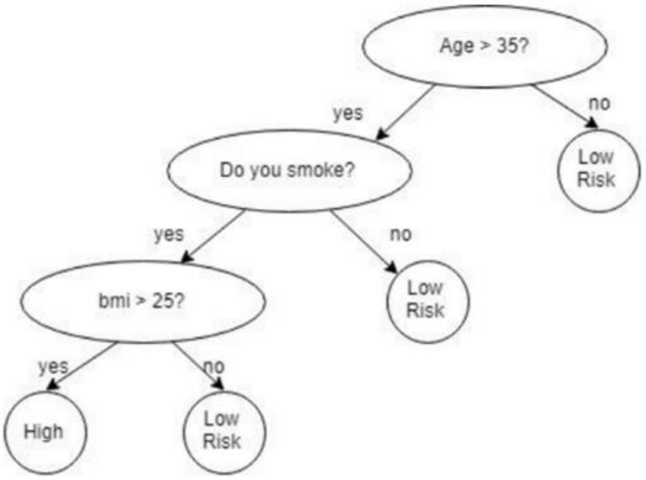


Fig 4(a): Decision tree

Tree based methods empower predictive models with high accuracy, stability and ease of interpretation. Unlike the linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand. Fig 4.1.1.1 represents the decision tree model for prediction of stroke diseases.

(b) Naïve Bayes

A Naïve Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem.

$$P(A\,|\,B) = \frac{P(B\,|\,A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of **A** happening, given that **B** has occurred. Hence, **B** is the evidence and **A** is the hypothesis. The assumption made here is that the predictors/features are independent. That is the presence of one particular feature does not affect the other. Hence it is called naïve.

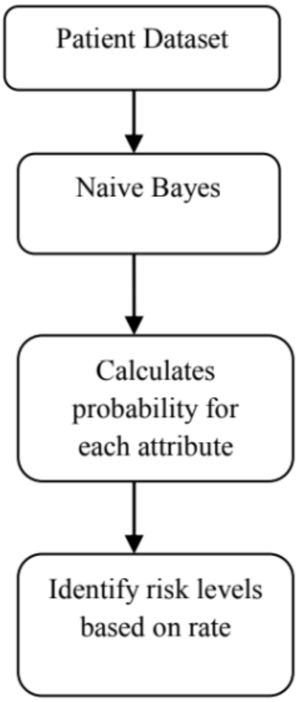

Patient Dataset
Naive Bayes
Calculates probability for each attribute
Identify risk levels based on rate

Fig 4.1(b): Bayesian classifier

Naïve Bayes algorithms are mostly used in sentiment analysis, spam filtering, recommendation systems etc. They are fast and easy to implement but their biggest disadvantage is that the requirement of predictors to be independent. In most of the real-life cases, the predictors are dependent, this hinders the performance of the classifier.

(c) Artificial Neural Network

Neural networks are a set of algorithms, modelled loosely after the human brain, that are designed to recognize patterns. They interpret sensory data through a kind of machine perception, labelling or clustering raw input. The patterns they recognize are numerical, contained in vectors, into which all real-world data, be it images, sound, text or time series, must be translated.

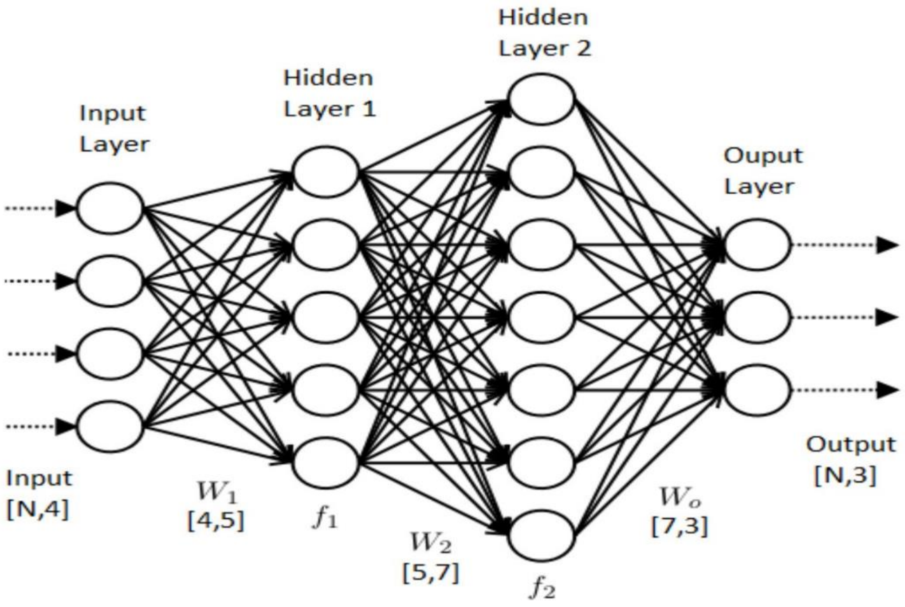


Fig 4.1(c): Artificial Neural Network

Neural networks help us cluster and classify. They help to group unlabeled data according to similarities among the example inputs, and they classify data when they have a labelled dataset to train on.

4.1.2 Asset Visualization

At a facility level, technicians accessing the user-interface will not be trained in Artificial Intelligence and Big Data. The key considerations when defining this requirement are the visualization of machine behavior and the ability to depict the health of machinery or the entire facility, and take specific action as a result.

4.2 Non-Functional Requirement

Non-functional requirements are the requirements which are not directly concerned with the specific function delivered by the system. They specify the criteria that can be used to judge the operation of a system rather than specific behaviors. They may relate to emergent system properties such as reliability, response time and store occupancy. Non-functional requirements arise through the user needs, because of budget constraints, organizational policies, the need for interoperability with other software and hardware systems or because of external factors such as: -

- Scalability
- Performance
- User Requirements
- Response time
- Maintainability
- Usability

4.2.1 Scalability

Analytics platform must be applicable to a machine or facility of any size. The solution must be able to add assets without a need for any incremental investment in hardware, software or dedicated labor hours.

4.2.2 Performance

The objective for an industrial analytics platform is to provide the production facility with accurate and timely data.

4.2.3 User Requirements

- There must be a user interface to configure the network.
- There must be an option for the user to select
- An option to view the performance parameters.
- The system should be user friendly, so that the client application is available at the system tray and user has to just click to select any options.

4.2.4 Response time

Response time is the elapsed time between an inquiry on a system and the response to that inquiry. Used as a measurement of system performance, response time may refer to service requests in a variety of technologies. Low response times may be critical to successful computing.

4.2.5 Maintainability

Maintainability is an important quality attribute and a difficult concept as it involves a number of measurements. Quality estimation means estimating maintainability of software. Maintainability is a set of attributes that bear on the effort needed to make specified modification.

4.2.6 Usability

One problem facing designers of interactive systems is catering to the wide range of users who will use a particular application. Understanding the user is critical to designing a usable interface. There are a number of ways of addressing this problem, including improved design methodologies using "intuitive" interface styles, adaptive interfaces, and better training and user support materials.

4.3 Data Requirements

Machine learning can enable new forms of predictive analytics and embed algorithm-driven intelligence into many software applications. However, none of that is possible without the right data, captured and processed the right way.

Machine learning algorithms consume and process large volumes of data to learn complex patterns about people, health, transactions, events, and so on. This intelligence is then incorporated into a predictive model. Comparisons to the model can reveal whether an entity is operating within acceptable parameters or is exhibiting an anomaly.

Machine learning is used to solve well bounded tasks such as classification and clustering. Note that a machine learning algorithm learns from so-called training data during development.

Figuring out what data are needed for a specific product or feature is the first and most important step in scoping data requirements. Machine learning models are nothing more than mathematical functions that take features as inputs, produce predictions as outputs, and learn how best to match predictions to patterns observed from the training data.

How much data do you need?

- In most cases, more data is better than less.
- If little or no data are available, transfer learning may help. In short, transfer learning allows you to take data and/or ML models from one task (e.g. classify dog breeds) and apply them to other tasks (e.g. classify cars). More on that in a future blog post.
- In cases where acquiring labeled data costs money (and/or time), define a goal of where you want to get to (in terms of model quality/performance) and a threshold of how much money/time you are willing/able to spend.
- At some point, more data will not help.

To illustrate these statements, here is a simplified graph that highlights the potential situations in which acquiring more data may or may not be beneficial.

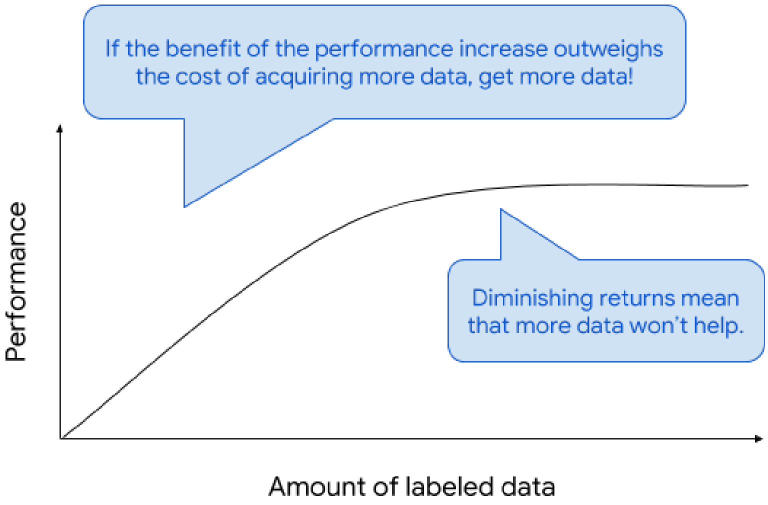


Fig 4.3(a): Graph of Performance vs Amount of Labeled data

The assumption is that most ML problems are on the steep part of this curve, i.e. acquiring more data will lead to better performance. However, in some cases, where a great deal of labeled training data already exists, there could be diminishing returns, i.e. training on more data doesn’t improve the model quality.

The following dataset is pre-defined in UCI Machine Learning Repository. The dataset has 73 features in which we selected those features which contribute most to our prediction output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features.

Summary

This chapter gives details of the functional requirements, non-functional requirements. Again, the non-functional requirements in turn contain scalability, performance, maintainability, user requirements etc.

Chapter 5

SYSTEM ANALYSIS AND DESIGN

5.1 System Architecture

A system architecture is the conceptual model that defines the structure, behavior, and more views of a system.

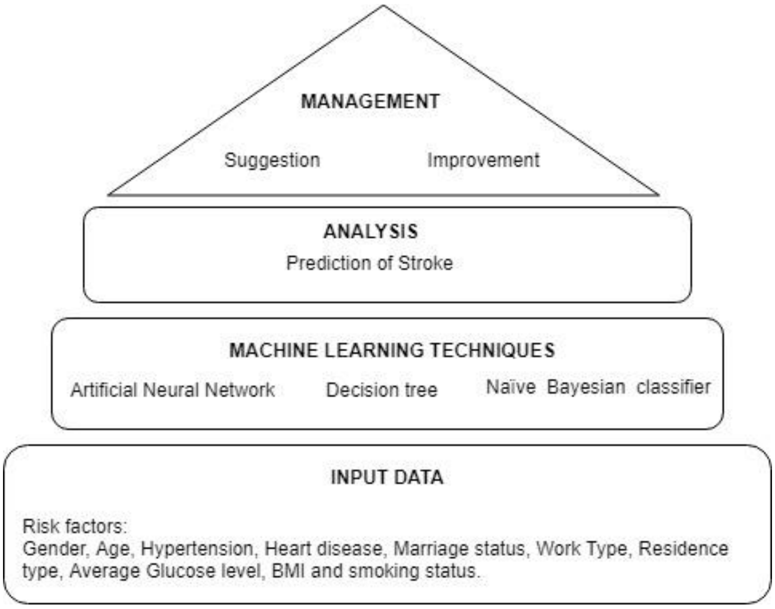

MANAGEMENT
Suggestion
Improvement
ANALYSIS
Prediction of Stroke
MACHINE LEARNING TECHNIQUES
Artificial Neural Network
Decision tree
Naïve Bayesian classifier
INPUT DATA
Risk factors:
Gender, Age, Hypertension, Heart disease, Marriage status, Work Type, Residence type, Average Glucose level, BMI and smoking status.

Fig 5.1: System Architecture

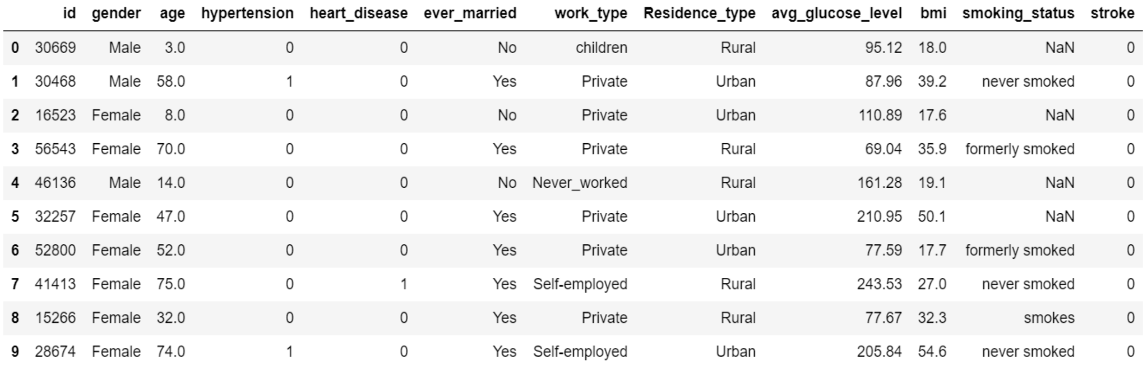
An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system. The overall logical structure of the project is divided into processing modules and a conceptual data structure is defined as Architectural Design.

Figure 5.1 shows the overall logical structure of the project with following modules:

- **Input data:** Risk factors like age, gender, hypertension, heart disease, BMI, Smoking status, Glucose level.
- **Machine Learning Techniques:** Artificial Neural Networks, Decision Tree, Naïve Bayes classifier.

- **Analysis:** Prediction and analysis of stroke whose performance is based on machine learning techniques.
- **Management:** Suggestion and improvement of stroke victims.

5.2: DATASET



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0
5	32257	Female	47.0	0	0	Yes	Private	Urban	210.95	50.1	NaN	0
6	52800	Female	52.0	0	0	Yes	Private	Urban	77.59	17.7	formerly smoked	0
7	41413	Female	75.0	0	1	Yes	Self-employed	Rural	243.53	27.0	never smoked	0
8	15266	Female	32.0	0	0	Yes	Private	Rural	77.67	32.3	smokes	0
9	28674	Female	74.0	1	0	Yes	Self-employed	Urban	205.84	54.6	never smoked	0

Fig 5.2: Stroke Dataset

Chapter 6

SYSTEM DESIGN

Overview

Design is a meaningful engineering representation of something that is to be built. It is the most crucial phase in the developments of a system. Software design is a process through which the requirements are translated into a representation of software. Design is a place where design is fostered in software Engineering. Based on the user requirements and the detailed analysis of the existing system, the new system must be designed. This is the phase of system designing. Design is the perfect way to accurately translate a customer’s requirement in the finished software product. Design creates a representation or model, provides details about software data structure, architecture, interfaces and components that are necessary to implement a system. The logical system design arrived at as a result of systems analysis is converted into physical system design.

6.1 System development methodology

System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress. The development method followed in this project is waterfall model.

6.1.1 Model phases

The waterfall model is a sequential software development process, in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of Requirement initiation, Analysis, Design, Implementation, Testing and maintenance.

Requirement Analysis: This phase is concerned about the collection of requirements of the system. This process involves generating document and requirement review.

System Design: Keeping the requirements in mind the system specifications are translated in to a software representation. In this phase the designer emphasizes on: - algorithm**,** data structure**,** software architecture etc.

Coding: In this phase programmer starts his coding in order to give a full sketch of product. In other words, system specifications are only converted in to machine readable compute code.

Implementation: The implementation phase involves the actual coding or programming of the software. The output of this phase is typically the library, executables, user manuals and additional software documentation

Testing: In this phase all programs (models) are integrated and tested to ensure that the complete system meets the software requirements. The testing is concerned with verification and validation.

Maintenance: The maintenance phase is the longest phase in which the software is updated to fulfill the changing customer need, adapt to accommodate change in the external environment, correct errors and oversights previously undetected in the testing phase, enhance the efficiency of the software.

6.1.2 Reason for choosing Agile Model as a development method

- Clear project objectives.
- Stable project requirements.
- Progress of the system is measurable.
- Strict sign-off requirements.
- Helps you to be the perfect.
- Logic of software development is clearly understood.
- Production of a formal specification
- Better resource allocation.
- Improves quality.

- Less human resources required as once one phase is finished those people can start working on to the next phase.

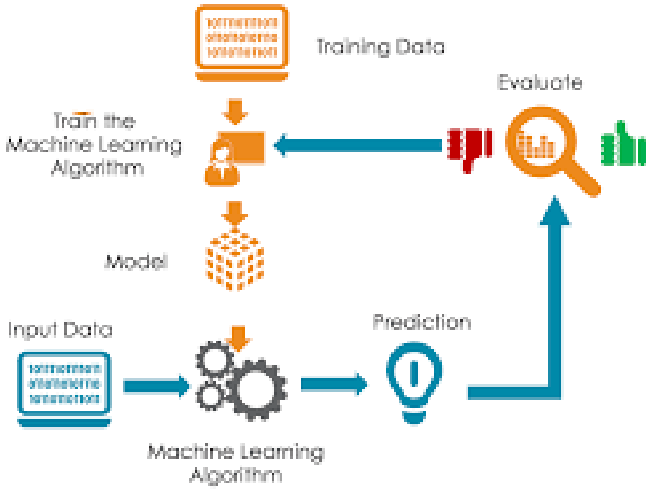

Training Data
Evaluate
Train the Machine Learning Algorithm
Model
Input Data
Prediction
Machine Learning Algorithm

Fig 6.1 (a): Agile model

6.2 Design Using UML

Designing UML diagram specifies, how the process within the system communicates along with how the objects with in the process collaborate using both static as well as dynamic UML diagrams since in this ever-changing world of Object-Oriented application development, it has been getting harder and harder to develop and manage high quality applications in reasonable amount of time. As a result of this challenge and the need for a universal object modeling language every one could use, the Unified Modeling Language (UML) is the Information industries version of blue print. It is a method for describing the systems architecture in detail. Easier to build or maintains system, and to ensure that the system will hold up to the requirement changes.

6.2.1 Data Flow Diagram

A data flow diagram (DFD) is graphic representation of the "flow" of data through an information system. A data flow diagram can also be used for the visualization of data processing (structured design). It is common practice for a designer to draw a context-level DFD first which shows the interaction between the system and outside entities. DFD’s show the flow of data from external entities into the system, how the data moves from one process to another, as well as its logical storage. There are only four symbols:

- Squares representing *external entities*, which are sources and destinations of information entering and leaving the system.
- Rounded rectangles representing *processes*, in other methodologies, may be called 'Activities', 'Actions', 'Procedures', 'Subsystems' etc. which take data as input, do processing to it, and output it.
- Arrows representing the *data flows*, which can either, be electronic data or physical items. It is impossible for data to flow from data store to data store except via a process, and external entities are not allowed to access data stores directly.
- The flat three-sided rectangle is representing data stores should both receive information for storing and provide it for further processing.

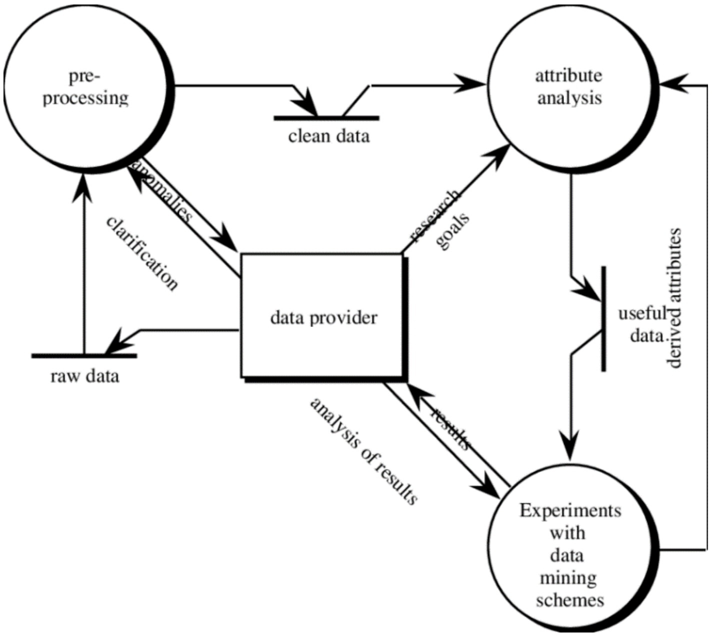


Fig 6.2 (a): Data flow diagram for a machine learning application

In the model presented in Figure 6.2.1, activity flows in a clockwise direction. In the pre-processing stage, the raw data is firstly represented as a single table, as required by the data mining algorithms. This table is translated into the ARFF (Attribute-Relation

File Format), an attribute/value table representation that includes header information on the attributes' data types.

The data may also require considerable 'cleansing', to remove outliers, handle missing values, detect erroneous values, and so forth. At this point the data provider (domain expert) and the data mining expert collaborate to transform the cleansed data into a form that will produce a readable, accurate data model when processed by a data mining algorithm. These two analysts may, for example, hypothesize that one or more attributes are irrelevant, and set aside these extraneous columns. Attributes may be manipulated mathematically, for example to convert all columns containing temperature measurements to a common scale, to normalize values in a given column, or to combine two or more columns into a single derived attribute.

One or more versions of the cleansed data are then processed by the data mining schemes. The domain expert determines which portions of the output are sufficiently novel or interesting to warrant further exploration, and which portions represent common knowledge for that field. The data mining expert interprets the algorithm's output and gives advice on further experiments that could be run with this data.

6.2.2 Workflow diagram

In the following section, we will have a look at some of the main steps of a typical machine learning task, and the diagram below should give us an intuitive understanding of how they are connected.

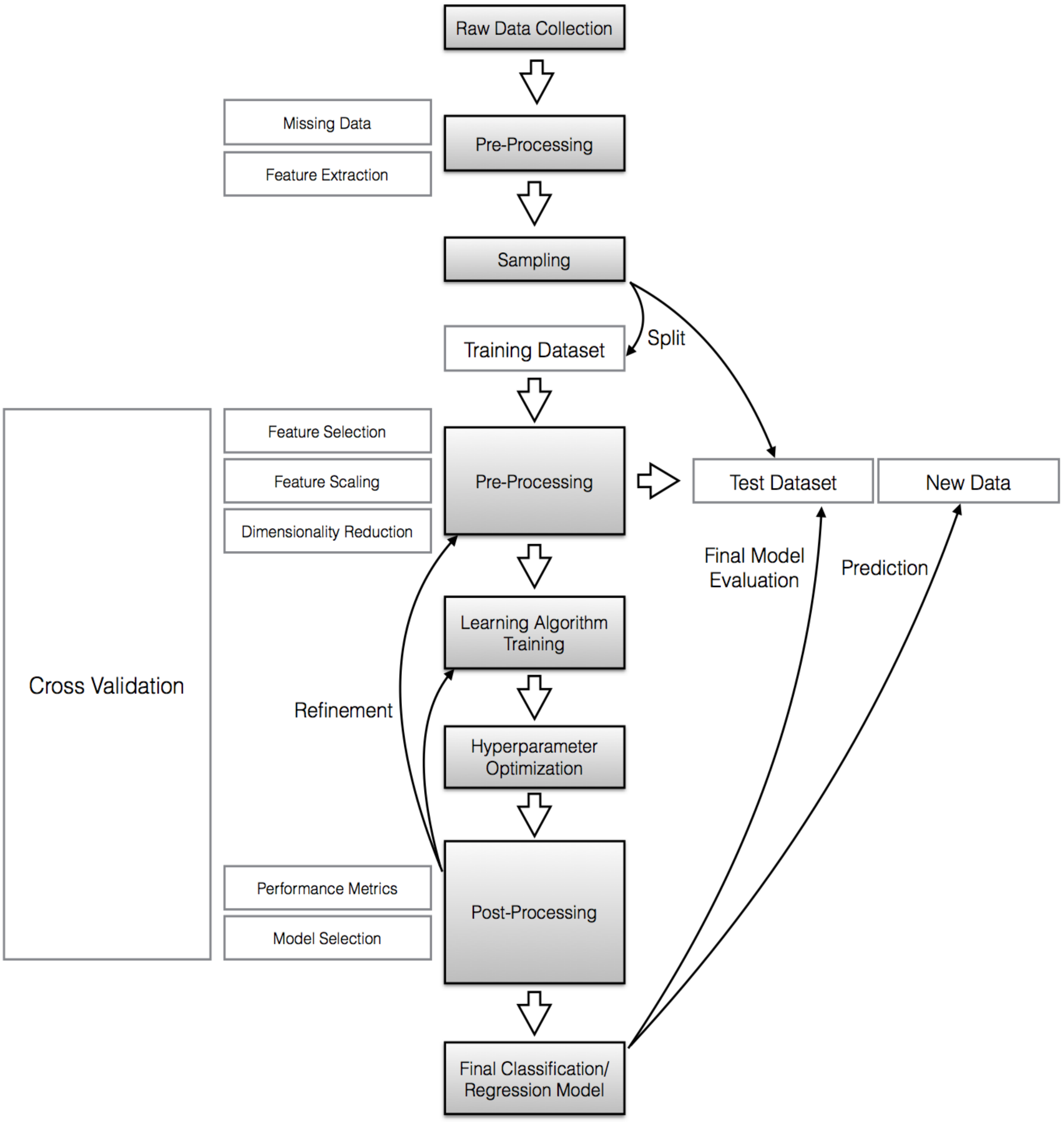


Fig 6.2 (c): Workflow diagram

Steps followed:

1: Raw data collection and feature extraction

When we'd download the UCI stroke dataset, we noticed that it is not already in "good shape", and it needs some preprocessing in order to obtain the dimensions of the sepals and petals in centimeters.

Occlusion of the leaves could be a problem that might lead to missing data: Many machine learning algorithms won't work correctly if data is missing in a dataset so that "ignoring" missing data might not be an option. If the sparsity (i.e., the number of empty cells in the dataset) is not too high, it is often recommended to remove either the samples rows that contain missing values, or the attribute columns for which data is missing. Another strategy for dealing with missing data would be imputation: Replacement of missing values using certain statistics rather than complete removal. For categorical data, the missing value can be interpolated from the most frequent category, and the sample average can be used to interpolate missing values for numerical attributes. In general, resubstituting of missing data by the overall sample mean.

2: Sampling

If we extracted certain features from our raw data, we would now randomly split our dataset into a training and a test dataset. The training dataset will be used to train the model, and the purpose of the test dataset is to evaluate the performance of the final model at the very end.

It is important that we use the test dataset only once in order to avoid overfitting when we compute the prediction-error metrics. Overfitting leads to classifiers that perform well on training data but do not generalize well so that the prediction-error on novel patterns is relatively high. Thus, techniques such as cross-validation are used in the model creation and refinement steps to evaluate the classification performance. An alternative strategy to re-use a test dataset for the model evaluation would be to create a third dataset, the so-called validation dataset.

3: Cross-Validation

Cross-validation is one of the most useful techniques to evaluate different combinations of feature selection, dimensionality reduction, and learning algorithms. There are multiple flavors of cross-validation, and the most common one would probably be k-fold cross-validation.

In k-fold cross-validation, the original training dataset is split into *k* different subsets (the so-called "folds") where 1 fold is retained as test set, and the other k-1 folds are used for training the model. E.g., if we set *k* equal to 4 (i.e., 4 folds), 3 different subsets of the original training set would be used to train the model, and the 4th fold would be used for evaluation. After 4 iteration, we can eventually calculate the average error rate (and standard deviation) of the model, which gives us an idea of how well our model generalizes.

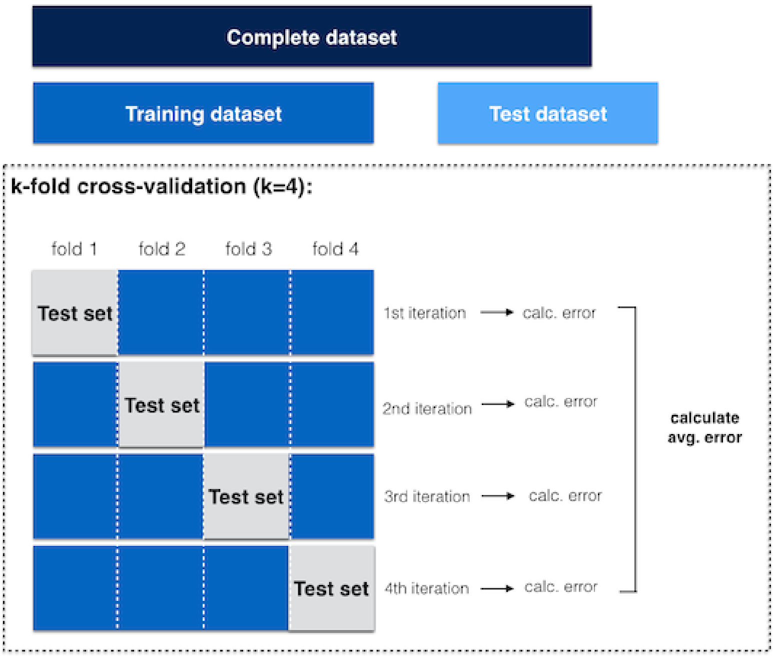

Complete dataset
Training dataset
Test dataset
k-fold cross-validation (k=4):
fold 1
fold 2
fold 3
fold 4
Test set
Test set
Test set
Test set
1st iteration → calc. error
2nd iteration → calc. error
3rd iteration → calc. error
4th iteration → calc. error
calculate avg. error

Fig 6.2 (d): k-fold cross validation

4: Normalization

Normalization and other feature scaling techniques are often mandatory in order to make comparisons between different attributes (e.g., to compute distances or similarities in cluster analysis), especially, if the attributes were measured on different scales (e.g., temperatures in Kelvin and Celsius); proper scaling of features is a requirement for most machine learning algorithms.

The term “normalization” is often used synonymous to “Min-Max scaling”: The scaling of attributes in a certain range, e.g., 0 to 1.

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Another common approach is the process of (z-score) “standardization” or “scaling to unit-variance”: Every sample is subtracted by the attribute’s mean and divided by the standard deviation so that the attribute will have the properties of a standard normal distribution (μ=0, σ=1).

$$z = \frac{x - \mu}{\sigma}$$

One important point that we have to keep in mind is that if we used any normalization or transformation technique on our training dataset, we'd have to use the same parameters on the test dataset and new unseen data.

5: Feature Selection and Dimensionality Reduction

Distinguishing between feature selection and dimensionality reduction might seem counter-intuitive at first, since feature selection will eventually lead (reduce dimensionality) to a smaller feature space.

In practice, the key difference between the terms “feature selection” and “dimensionality reduction” is that in feature selection, we keep the “original feature axis”, whereas dimensionality reduction usually involves a transformation technique.

The main purpose of those two approaches is to remove noise, increase computational efficiency by retaining only “useful” (discriminatory) information, and to avoid overfitting.

6: Learning algorithms and hyperparameter tuning

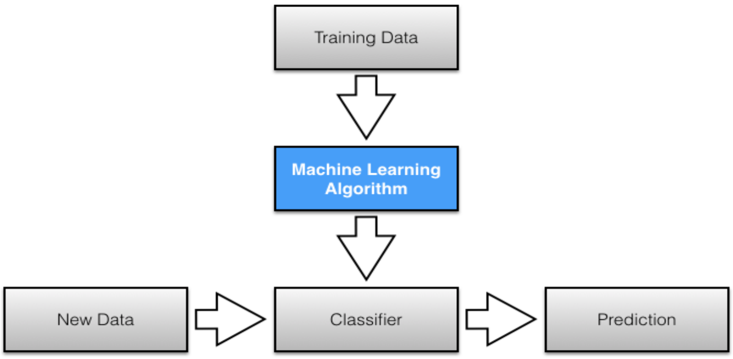


Fig 6.2 (e): Learning algorithms

There are an enormous number of different learning algorithms, and the details about the most popular ones are perfect topics for separate articles and applications. Here is just a very brief summary of four commonly used supervised learning algorithms:

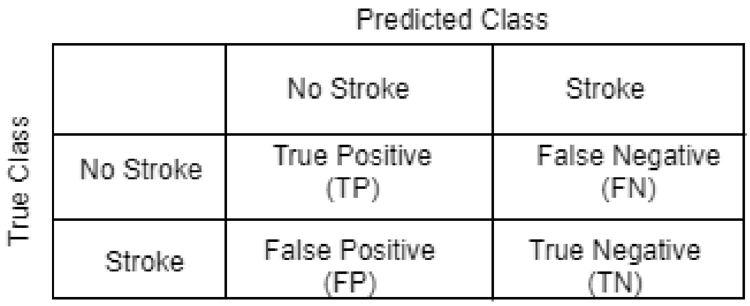
- **Bayes classifiers** are based on a statistical model (i.e., Bayes theorem: calculating posterior probabilities based on the prior probability and the so-called likelihood). A Naive Bayes classifier assumes that all attributes are conditionally independent, thereby, computing the likelihood is simplified to the product of the conditional probabilities of observing individual attributes given a particular class label.
- **Artificial Neural Networks (ANN)** are graph-like classifiers that mimic the structure of a human or animal "brain" where the interconnected nodes represent the neurons.
- **Decision tree classifiers** are tree like graphs, where nodes in the graph test certain conditions on a particular set of features, and branches split the decision towards the leaf nodes. Leaves represent lowest level in the graph and determine the class labels. Optimal tree is trained by minimizing Gini impurity, or maximizing information gain.

Hyperparameters are the parameters of a classifier or estimator that are not directly learned in the machine learning step from the training data but are optimized separately. The goals of hyperparameter optimization are to improve the performance of a classifier and to achieve good generalization of a learning algorithm. A popular method for hyperparameter optimization is Grid Search. Typically, Grid Search is implemented as an exhaustive search (in contrast to randomized parameter optimization) of candidate parameter values. After all possible parameter combination for a model are evaluated, the best combination will be retained.

7: Prediction-error metrics and model selection

A convenient tool for performance evaluation is the so-called confusion matrix, which is a square matrix that consists of columns and rows that list the number of instances as "actual class" vs. "predicted class" ratios.

A confusion matrix for a simple “Stroke vs. No Stroke” classification could look like:



Predicted Class

True Class

	No Stroke	Stroke
No Stroke	True Positive (TP)	False Negative (FN)
Stroke	False Positive (FP)	True Negative (TN)

Fig 6.2 (f): Confusion matrix

Often, the prediction “accuracy” or “error” is used to report classification performance. Accuracy is defined as the fraction of correct classifications out of the total number of samples; it is often used synonymous to specificity/precision although it is calculated differently. Accuracy is calculated as

$$\frac{TP + TN}{P + N}$$

where TP=True Positives, TN=True Negatives, P=Positives, N=Negatives.

Other indicators for classification performances are **Sensitivity**, **Specificity**, **Recall**, and **Precision**.

- Sensitivity (synonymous to recall) and precision are assessing the “True Positive Rate” for a binary classification problem: The probability to make a correct prediction for a “positive/true” case (e.g., in an attempt to predict a disease, the disease is correctly predicted for a patient who truly has this disease).

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

- Specificity describes the “True Negative Rate” for a binary classification problem: The probability to make a correct prediction for a “false/negative” case (e.g., in an attempt to predict a disease, no disease is predicted for a healthy patient).

$$Specificity = \frac{TN}{TN + FP}$$

In a typical supervised learning workflow, we would evaluate various different combinations of feature subspaces, learning algorithms, and hyperparameters before we select the model that has a satisfactory performance. As mentioned above, cross-validation is a good way for such an assessment in order to avoid overfitting to our training data.

Chapter 7

IMPLEMENTATION

The implementation phase of the project is where the detailed design is actually transformed into working code. Aim of the phase is to translate the design into a best possible solution in a suitable programming language. This chapter covers the implementation aspects of the project, giving details of the programming language and development environment used. It also gives an overview of the core modules of the project with their step by step flow.

The implementation stage requires the following tasks.

- Careful planning.
- Investigation of the system and constraints.
- Design of methods to achieve the changeover.
- Evaluation of the changeover method.
- Correct decisions regarding selection of the platform
- Appropriate selection of the language for application development

7.1 Pre – Processing

7.1.1 Clean the missing values both training and testing data

Data in real world are rarely clean and homogeneous. Typically, they tend to be incomplete, noisy, and inconsistent and it is an important task of a Data scientist to prepossess the data by filling missing values. It is important to be handled as they could lead to wrong prediction or classification for any given model being used.

Code to clean missing values:

```
train_data["bmi"] = train_data["bmi"].fillna(value=0)
```

7.1.2 Applying Label Encoder to convert object into integer

In ML models we are often required to convert the categorical i.e text features to its numeric representation. The two most common ways to do this is to use Label Encoder or OneHot Encoder. In our project we are using Label Encoder.

Code for applying Label Encoder:

```
str_data = train_data.select_dtypes(include=['object'])

str_data.columns

int_data = train_data.select_dtypes(include=['integer',"float"])

int_data.columns

label = LabelEncoder()

train_data['gender'] = label.fit_transform(train_data['gender'])

train_data['ever_married'] = label.fit_transform(train_data['ever_married'])

train_data['work_type'] = label.fit_transform(train_data['work_type'])

train_data['Residence_type'] = label.fit_transform(train_data['Residence_type'])

features = train_data[train_data['smoking_status'].notnull()]

features['smoking_status'] = label.fit_transform(features['smoking_status'])

features.head()
```

7.1.3 Balancing Dataset

Imbalanced class distribution is a scenario where the number of observations belonging to one class is significantly lower than those belonging to the other classes. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate. This happens because Machine Learning Algorithms are usually designed to improve accuracy by reducing the error. Thus, they do not take into account the class distribution / proportion or balance of classes. Here we solve such class imbalance problems using sampling techniques.

Code for balancing dataset:

```
from imblearn.over_sampling import RandomOverSampler,SMOTE
ros = RandomOverSampler(random_state=0)
X_res, y_res = ros.fit_sample(xtrain, ytrain)
print ('ROS Input Data Shape for Smoke Data: {}'.format(X_res.shape))
print ('ROS Output Data Shape for Smoke Data: {}'.format(y_res.shape))
```

7.1.4 Split the data into training and testing

The data set that we used is split into training data and test data. The training set contains a known output and the model learns on this data in order to be generalized to other data later on. We have the test dataset in order to test our model’s prediction on this subset.

Code for splitting data into train and test data:

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(X_res,y_res,test_size=0.2,random_state=41)
print(x_train.shape)
print(x_test.shape)
```

7.2 Modules in Implementation

This project mainly implements three algorithms. In this section code used for the implementing each of the algorithm is discussed. Here each algorithm is considered as an individual module. Therefore, the implementations of these three modules are described as follows:

- ➢ Decision Tree Model
- ➢ Naïve Bayes Model
- ➢ Artificial Neural Network Model

The scikit learn (python library) will help here to build a Machine learning model in Python. There are three types under scikit learn library:

- **Gaussian:** It is used in classification and it assumes that features follow a normal distribution.

- **Multinomial:** It is used for discrete counts. For example, let's say, we have a text classification problem. Here we can consider bernoulli trials which is one step further and instead of "word occurring in the document", we have "count how often word occurs in the document", you can think of it as "number of times outcome number x_i is observed over the n trials".

- **Bernoulli:** The binomial model is useful if your feature vectors are binary (i.e. zeros and ones). One application would be text classification with 'bag of words' model where the 1s & 0s are "word occurs in the document" and "word does not occur in the document" respectively.

Based on our data set, we chose Gaussian model to predict the output.

7.2.1 Building Decision Tree Model

If dataset consists of **"n"** attributes then deciding which attribute to place at the root or at different levels of the tree as internal nodes is a complicated step. By just randomly selecting any node to be the root can't solve the issue. If we follow a random approach, it may give us bad results with low accuracy.

For solving this attribute selection problem, researchers worked and devised some solutions. They suggested using some criterion like **information gain, gini index,** etc. These criterions will calculate values for every attribute. The values are sorted, and attributes are placed in the tree by following the order i.e, the attribute with a high value(in case of information gain) is placed at the root. In our project we are using information gain as criterion.

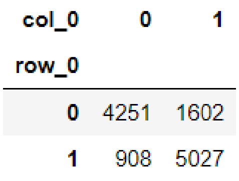
While using information Gain as a criterion, we assume attributes to be categorical. To measure the randomness or uncertainty of a random variable X is defined by **Entropy**.

Code for building decision tree model:

```
from sklearn.tree import DecisionTreeClassifier
dt_mod=DecisionTreeClassifier(criterion='entropy',max_depth=7,random_state=4)
dt_mod.fit(x_train,y_train)
y_pred=dt_mod.predict(x_test)
y_pred
ts_dt_score=dt_mod.score(x_test,y_test)
print("DTtest_score:",ts_dt_score)
tr_dt_score=dt_mod.score(x_train,y_train)
print("DTtrain_score:",tr_dt_score)
```

Code for Naïve Bayes Confusion matrix:

```
dt_conf_mtr=pd.crosstab(y_test,y_pred)
dt_conf_mtr
```



col_0	0	1
row_0		
0	4251	1602
1	908	5027

7.2.2 Building Naïve Bayes Model

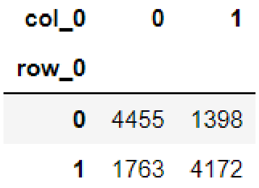
We use a Gaussian function to estimate the probability of a given attribute value, given the known mean and standard deviation for the attribute estimated from the training data. The result is the conditional probability of a given attribute value given a class value.

Code for building Naïve Bayes model:

```
from sklearn.naive_bayes import GaussianNB
model=GaussianNB()
model.fit(x_train,y_train)
predict=model.predict(x_test)
predict
test_score=model.score(x_test,y_test)
print("NBtest_score:",test_score)
train_score=model.score(x_train,y_train)
print("NBtrain_score:",train_score)
```

Code for Naïve Bayes Confusion matrix:

```
nb_conf_mtr=pd.crosstab(y_test,predict)
nb_conf_mtr
```



```
col_0     0     1
row_0
    0  4455  1398
    1  1763  4172
```

7.2.3 Building Artificial Neural Networks Model

```
from sklearn.neural_network import MLPClassifier
mlp_model=MLPClassifier(activation='relu',hidden_layer_sizes=(20,20),max_iter=1000,
                    batch_size=10,alpha=0.0001,learning_rate_init=0.001,
                    solver='adam',random_state=4)
mlp_model.fit(x_train,y_train)
mlp_predict= mlp_model.predict(x_test)
mlp_predict
ts_mlp_score=mlp_model.score(x_test,y_test)
print("NNtest_acore:",ts_mlp_score)
```

```
tr_mlp_score=mlp_model.score(x_train,y_train)

print("NNtrain_acore:",tr_mlp_score)
```

7.3 GUI for prediction

```
model_DT = pickle.load(open("model_DT.pkl", "rb"))

pred1 = model_DT.predict(np.array(sample).reshape(1, -1))

print("\n\nPredicted DT:", pred1)


model_NB = pickle.load(open("model_NB.pkl", "rb"))

pred2 = model_NB.predict(np.array(sample).reshape(1, -1))

print("\n\nPredicted NB:", pred2)

 model_MLP = pickle.load(open("model_MLP.pkl", "rb"))

pred3 = model_MLP.predict(np.array(sample).reshape(1, -1))

print("\n\nPredicted MLP:", pred3)


pred_lab = ["No Stroke", "Stroke"]


eml1 = Label(window)

eml1.grid(column=4, row=13)

eml1 = Label(window)

eml1.grid(column=4, row=14)


btn11_1 = Button(window, text="DT: "+pred_lab[pred1[0]], command=button_1, height=4, width=8)

btn11_1.grid(column=2, row=15)

btn11_2 = Button(window, text="NB: "+pred_lab[pred2[0]], command=button_1, height=4, width=8)

btn11_2.grid(column=4, row=15)

btn11_3 = Button(window, text="MLP:"+pred_lab[pred3[0]], command=button_1, height=4, width=8)
```

```
predict_btn = Button(window, text='Predict', command=predict, height=5, width=10)
window.mainloop()
```

Summary

The chapter discusses the implementation details of the different modules of the system and gives the step by step flow of each of them. Along with these, this chapter also highlights some of the important features of the platform and language used for implementation purpose.

Chapter 9

RESULTS AND PERFORMANCE EVALUATION

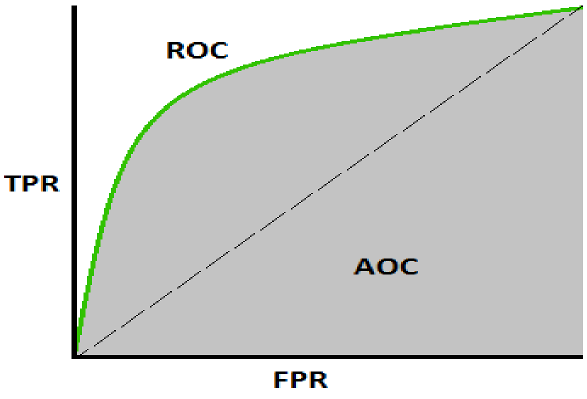
The following snapshots and graphs define the results or outputs that we will get after step by step execution of each proposed protocol for different values of time and speed.

9.1 Performance Analysis

In this section snapshot showing the performance of three algorithms proposed in this project i.e. Decision Tree, Naïve Bayes, Artificial Neural Network are compared.

AUC – ROC **(**Area Under The Curve - Receiver Operating Characteristics**) curve** is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.



$$\textbf{FPR} = 1 - \textbf{Specificity}$$

$$\textbf{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN} \qquad = \frac{FP}{TN + FP}$$

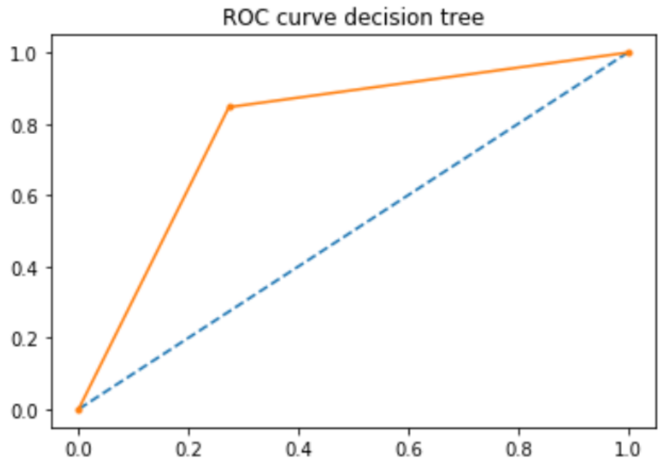


Fig 9.1(a): ROC of Decision Tree

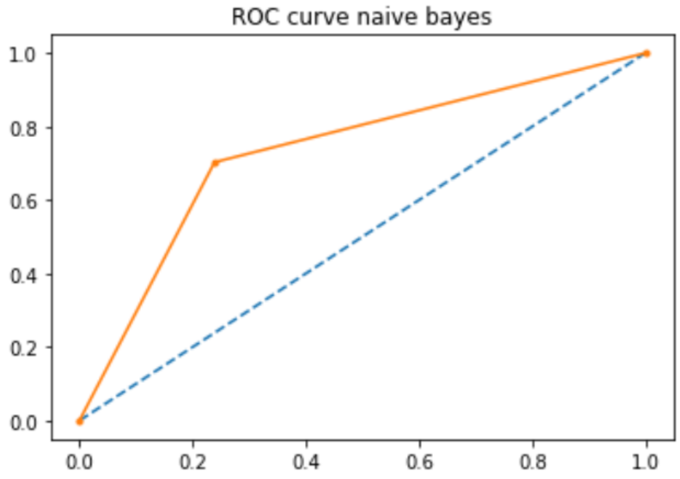


Fig 9.1(b): ROC of Naïve Bayes

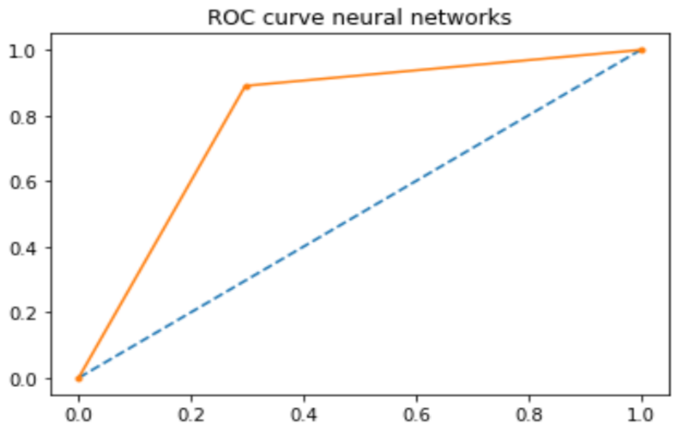


Fig 9.1(c): ROC of ANN

9.2 Graphs and Analysis

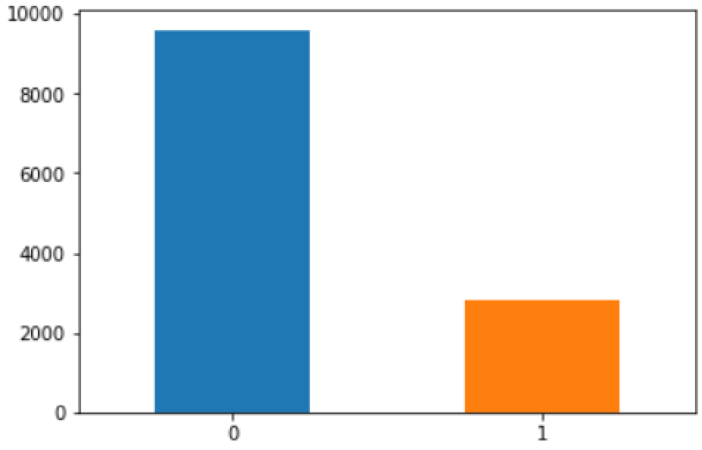


Fig 9.2 (a): Frequency of stroke

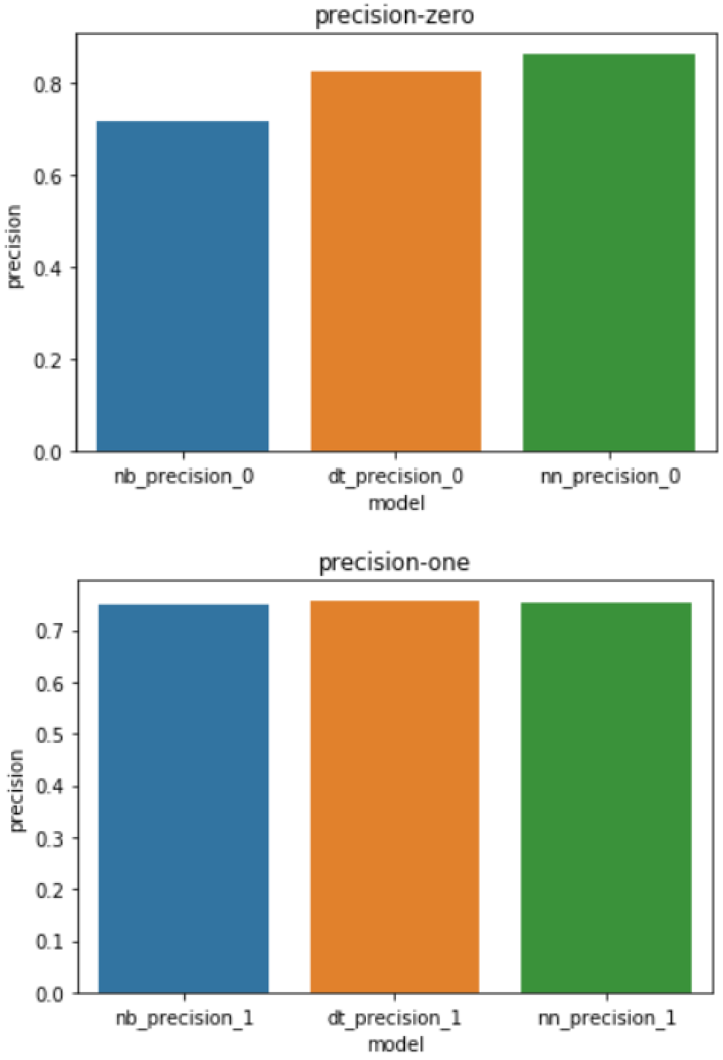


Fig 9.2 (b): precision-zero & precision-one

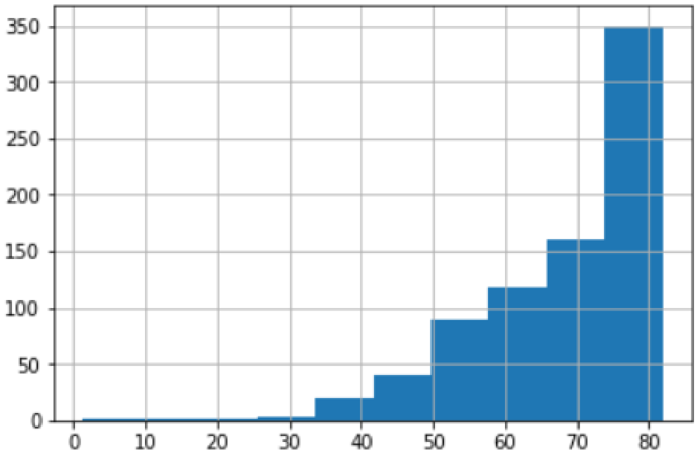


Fig 9.2 (c): Frequency of people with stroke on basis of age

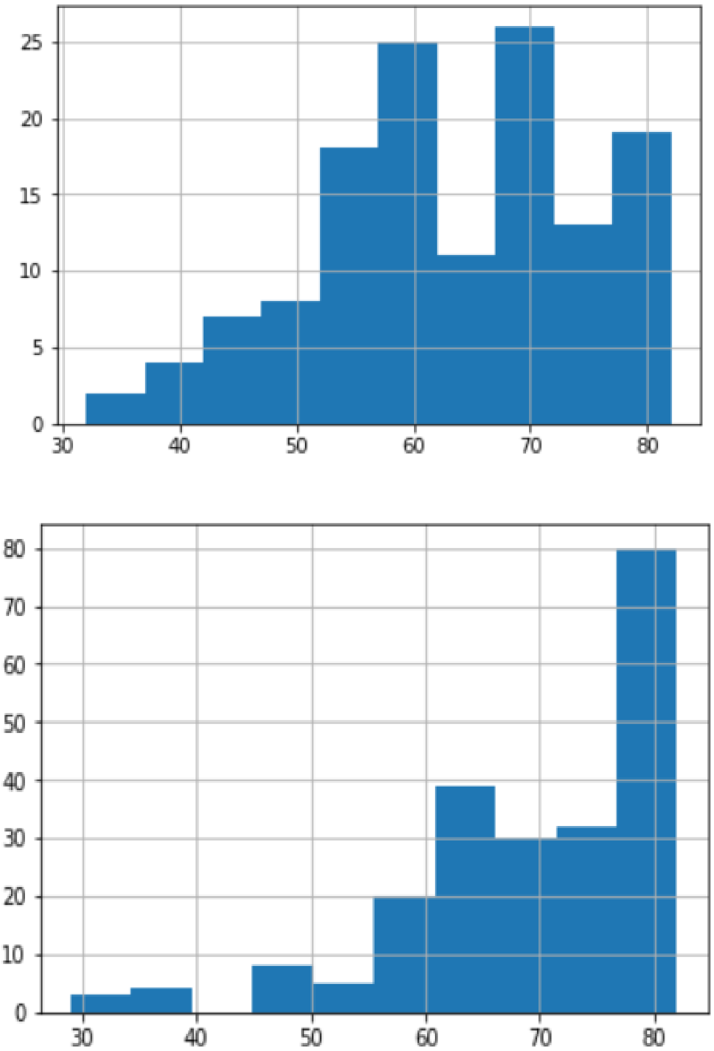
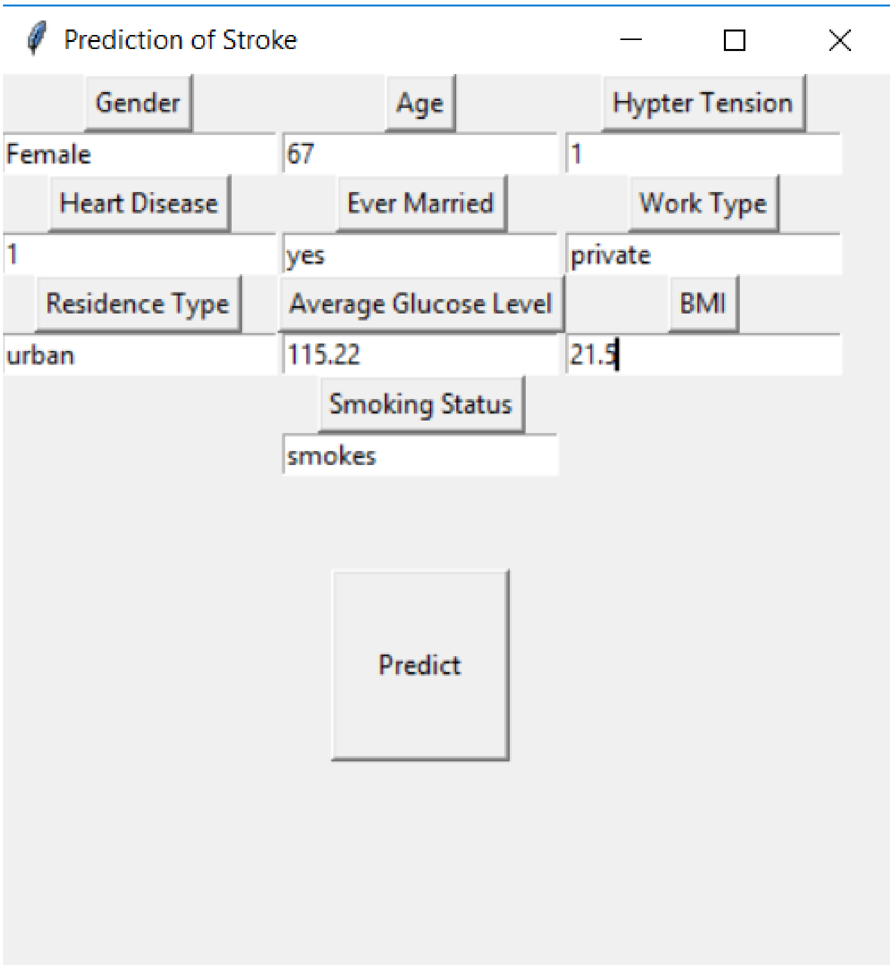


Fig 9.2 (d): Frequency of people who smokes & formerly smokes

9.3 GUI Snapshot



Gender	Age	Hypter Tension
Female	67	1
Heart Disease	Ever Married	Work Type
1	yes	private
Residence Type	Average Glucose Level	BMI
urban	115.22	21.5
	Smoking Status	
	smokes	

Fig 9.3: GUI Snapshot

Summary

This chapter gives a graphic view of the execution of the system. The output screens show each of the execution stages for each protocol. Based on the analysis the proposed protocol performs well in terms of energy consumption when compared to the existing AODV protocol.

Chapter 10

CONCLUSION & FUTURE SCOPE

10.1 Conclusion

Several assessments and prediction models, Decision Tree, Naive Bayes and Neural Network, showed acceptable accuracy in identifying stroke-prone patients. This project hence helps to predict the stroke risk using prediction model and provide personalized warning and the lifestyle correction message through a web application. By doing so, it urges medical users to strengthen the motivation of health management and induce changes in their health behaviors.

10.2 Future Scope

This project helps to predict the stroke risk using prediction model in older people and for people who are addicted to the risk factors as mentioned in the project.

In future, the same project can be extended to give the stroke percentage using the output of current project. This project can also be used to find the stroke probabilities in young people and underage people by collecting respective risk factor information's and doctors consulting.

REFERENCES

[1]. "*Computer Methods and Programs in the Biomedicine*" - Jae–woo Lee, Hyun-sun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyung-hee Cho

[2]. "*Probability of Stroke: A Risk Profile from the Framingham Study*" - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD

[3]. "*Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study*" - Min SN, Park SJ, Kim DJ, Subramaniyam M, Lee KS

[4]. "*Stroke prediction using artificial intelligence*"- M. Sheetal Singh, Prakash Choudhary

[5]. "*Medical software user interfaces, stroke MD application design (IEEE)*" - Elena Zamsa

[6]. "*Focus on stroke: Predicting and preventing stroke*" - Michael Regnier

[7]. "*Effective Analysis and Predictive Model of Stroke Disease using Classification Methods*" - A.Sudha, P.Gayathri, N.Jaisankar

[8]. "*Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study*" - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal

Websites

[1]. STROKE INFORMATION- https://en.wikipedia.org/wiki/Stroke

[2]. UNDERSTANDING SROKE - https://www.stroke.org/understand-stroke/what-is-stroke/

[3]. TYPES OF STROKE - https://www.medicalnewstoday.com/articles/7624.php

[4]. STROKE MANAGEMENT - https://jnis.bmj.com/content/10/4/358

[5]. CLASSIFICATION OF STROKE DISEASE USING MACHINE LEARNING - https://link.springer.com/article/10.1007%2Fs00521-019-04041-y