A STUDY ON PREDICTIVE MODELLING OF STROKE RISK FACTORS USING HEALTH AND LIFESTYLE INDICATORS

Dissertation Report submitted in partial fulfilment of the requirements for the award of the degree of

MASTER OF BUSINESS ADMINISTRATION

of

RV INSTITUTE OF MANAGEMENT

Autonomous Institution Affiliated to Bengaluru City University

By

B SHASHANK

REG NO: P18FW21M0096

Under the guidance of

Internal Guide

Dr. Vinay K S

Assistant Professor

RV Institute of Management

Rashtreeya Sikshana Samithi Trust

RV INSTITUTE OF MANAGEMENT

Autonomous Institution Affiliated to Bengaluru City University

CA 17, 36TH cross, 26th main, 4th T Block,

Jayanagar, Bangalore-560041

2023

DECLARATION BY THE STUDENT

I hereby declare that "***Predictive Modelling of Stroke Risk Factors Using Health and Lifestyle Indicators***" is the result of the project work carried out by me under the guidance of ***Dr. Vinay K.S.*** in partial fulfillment for the award of Master's Degree in Business Administration by RV Institute of Management, Autonomous Institution Affiliated to Bengaluru City University.

I also declare that this project is the outcome of my own efforts and that it has not been submitted to any other university or Institute for the award of any other degree or Diploma or Certificate.

Name: B SHASHANK

Register Number: P18FW21M0096 **Signature**

Place: Bengaluru

Date:

ACKNOWLEDGMENT

I would like to express my deepest gratitude to all those who have contributed to the successful completion of my final semester college dissertation on the predictive modelling of stroke risk factors using health and lifestyle indicators.

Express my sincere thanks to our director sir **Dr. PURUSHOTTAM BUNG** for his inspiration and support at various stages of the dissertation project.

I extend my heartfelt thanks to my Mentor and Internal guide **Asst Prof Dr. Vinay K.S.** for his invaluable guidance, expertise, and continuous encouragement throughout the research process. Their insightful feedback and constructive criticism have been instrumental in shaping the direction and quality of this study.

Last, but not the least, thanks my family and friends for their inputs to improve the project.

B SHASHANK

Reg. No: P18FW21M0096

GUIDE CERTIFICATE

This is to certify that Mr. **B SHASHANK** of RV Institute of Management, an Autonomous Institution Affiliated to Bengaluru City University, has undertaken a Master Thesis entitled "***Predictive Modelling of Stroke Risk Factors Based on Health and Lifestyle Indicators***" under my Guidance and it has not been submitted to any other University or Institute for the award of any other degree or Diploma or Certificate. His Conduct and work is Original, and Good.

Name: Dr. Vinay K.S.

Date: **Signature**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1:

INTRODUCTION

1.1 Brain an Overview:

The brain is a miraculous three-pound organ that governs all bodily functions, analyses information from the outside world, and embodies the essence of the mind and soul. The brain controls many things, including intelligence, creativity, emotion, and memory. The brain, which is protected within the skull, is made up of the cerebrum, cerebellum, and brainstem. The brain collects information from our five senses: sight, smell, touch, taste, and hearing - frequently all at once. It puts the messages together in a way that makes sense to us and can store the information in our memory. The brain regulates our thoughts, memory, and speech, as well as the movement of our arms and legs and the function of many organs in our body. The brain and spinal cord make up the central nervous system (CNS). The peripheral nervous system (PNS) consists of spinal nerves that branch from the spinal cord and cranial nerves that branch from the brain.

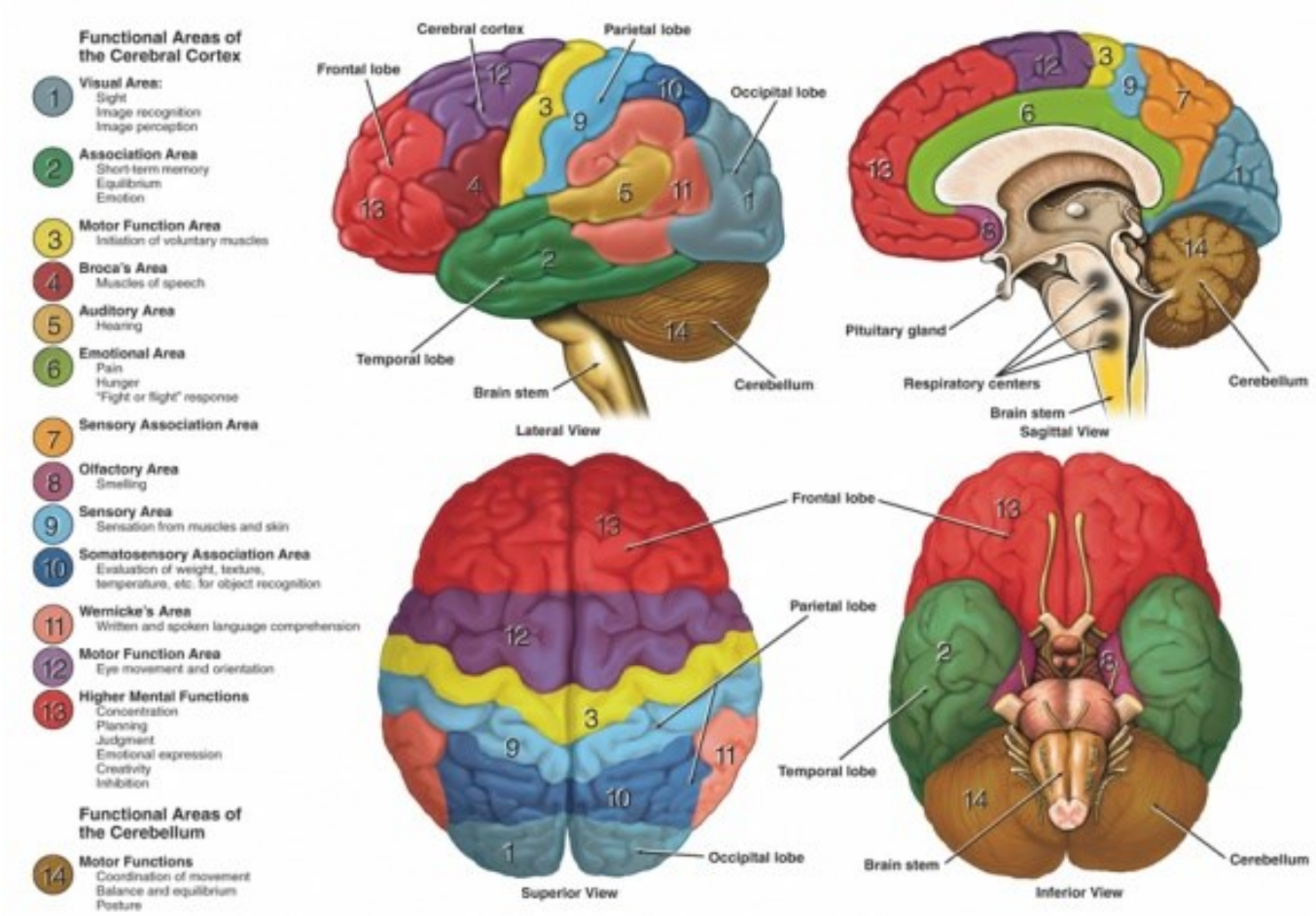


Fig 1.1

1.2 Structure of Brain:

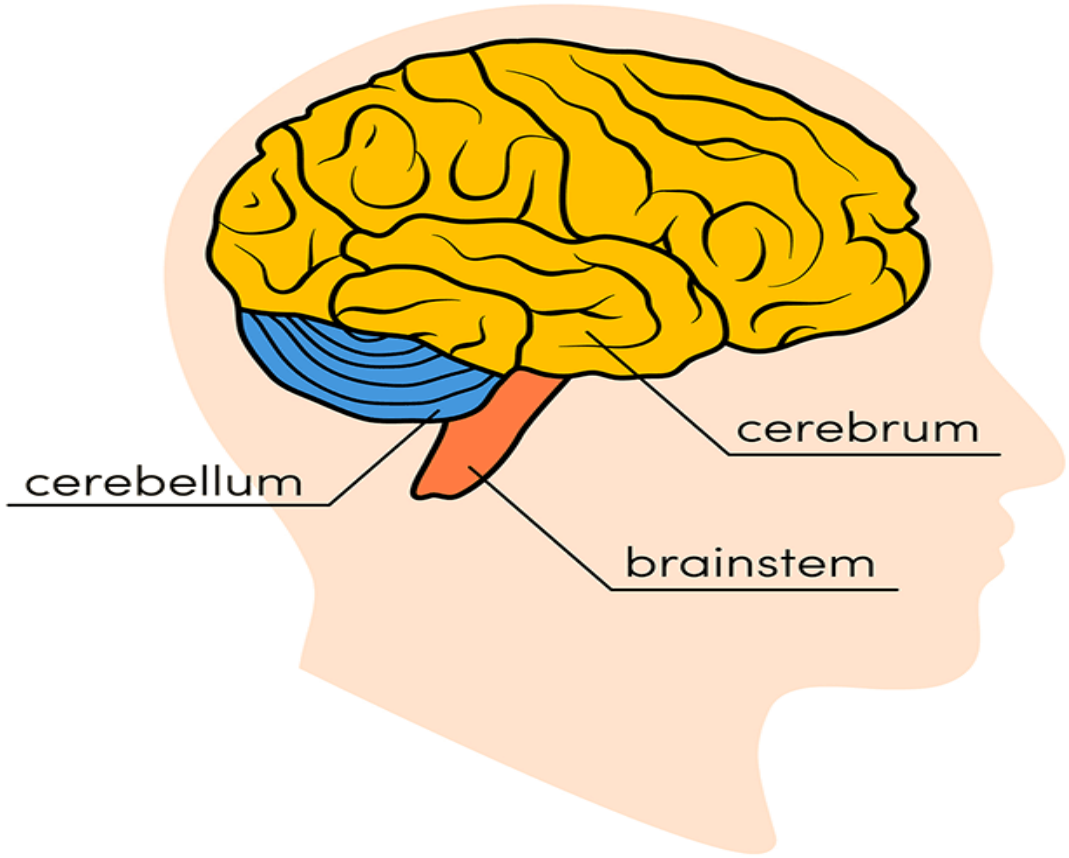


Fig 1.2

1.2.1 Cerebrum:

The cerebrum is the biggest region of the brain, consisting of right and left hemispheres. Higher functions include understanding touch, vision, and hearing, as well as speaking, reasoning, emotions, learning, and fine motor control. The cerebrum (front of the brain) consists of grey matter (the cerebral cortex) and white matter in its centre. The cerebrum, the main region of the brain, initiates and coordinates movement as well as regulates temperature. Other parts of the cerebrum are responsible for communication, judgment thought and reasoning, problem-solving, emotions, and learning. Other functions concern vision, hearing, touch, and other senses.

Cerebral Cortex is Latin for "bark," and it refers to the cerebrum's outer grey matter coating. Because of its folds, the cortex has a huge surface area and accounts for almost half of the

weight of the brain. The cerebral cortex is split into two halves called hemispheres. It has ridges (gyri) and folds (sulci) on it. The two halves come together at a huge, deep sulcus (the interhemispheric fissure, also known as the medial longitudinal fissure) that extends from the front to the rear of the head. The left half of the brain controls the left side of the body, whereas the right half controls the right side. The two parts communicate with one another via the corpus callosum, a huge, C-shaped mass of white matter and nerve connections. The corpus callosum is located in the cerebrum's center.

1.2.2 Brainstem:

The brainstem (middle of the brain) connects the cerebrum to the spinal cord. The midbrain, pons, and medulla are all part of the brainstem. The brainstem serves as a relay center for the cerebrum and cerebellum to the spinal cord. Breathing, heart rate, body temperature, wake and sleep cycles, digestion, sneezing, coughing, vomiting, and swallowing are all autonomic functions.

Midbrain- The middle part of the brain (or mesencephalon) is an intricate structure that contains a variety of neuron clusters (nuclei and colliculi), neuronal pathways, and other structures. These traits aid in a variety of activities, ranging from hearing and movement to calculating responses and detecting environmental changes. The midbrain also contains the substantia nigra, a Parkinson's disease-affected area rich in dopamine neurons and part of the basal ganglia, which allows movement and coordination.

Pons- Four of the 12 cranial nerves originate in the pons, allowing for a variety of tasks such as tear generation, chewing, blinking, concentrating vision, balancing, hearing, and facial expression. The pons, named after the Latin word for "bridge," connects the midbrain to the medulla.

Medulla- The medulla is located at the bottom of the brainstem and connects the brain to the spinal cord. The medulla is critical for survival. The medulla regulates various body operations, including heart rhythm, respiration, blood flow, and oxygen and carbon dioxide levels. The medulla is responsible for reflexive actions such as sneezing, vomiting, coughing, and swallowing.

The spinal cord runs from the bottom of the medulla to the bottom of the skull through a big hole. The spinal cord, which is supported by the vertebrae, transports messages between the brain and the rest of the body.

1.2.3 Cerebellum:

The cerebellum, sometimes known as the "little brain," is a fist-sized part of the brain found in the rear of the head, beneath the temporal as well as occipital and above the region known as the brainstem. The cerebellum is placed beneath the cerebrum. Its role is to coordinate muscular movements, maintain posture, and maintain balance like the cortex of the brain, has two hemispheres. The outside section includes neurons, and the interior area interacts with the cortex of the brain. Its purpose is to coordinate the voluntary contractions of muscles and to keep posture, balance, and equilibrium. New research is looking into the cerebellum's roles in thought, emotions, and social behavior, as well as its likely connection in addiction, autism, and schizophrenia.

1.3 Overview of Stroke:

A stroke, sometimes known as a "brain attack," is a medical emergency that requires prompt attention and a thorough understanding. It is a prominent cause of disability and mortality worldwide, making it critical for both healthcare professionals and general people to be aware of and knowledgeable about this condition. A stroke occurs when the blood supply to the brain is disrupted, either owing to a blockage in the blood vessels (ischemic stroke) or a rupture of the blood vessels (hemorrhagic stroke). A stroke can have serious effects because the damaged portion of the brain may not receive enough oxygen and nutrients, resulting in cell damage or death. Stroke symptoms can appear abruptly and include trouble speaking, weakness or numbness on one side of the body, and vision impairment. When blood flow to the brain is disrupted or there is sudden bleeding in the brain, a stroke can result. There are two types of strokes. An ischemic stroke occurs when blood flow to the brain is disrupted due to which the brain receives no oxygen or nutrients from the blood. Within minutes after being deprived of oxygen and nutrients, brain cells begin to die. A hemorrhagic stroke occurs as a result of sudden bleeding in the brain. The leaky blood puts pressure on brain cells, damaging them. Strokes are produced by clogged blood vessels (ischemic), with the remainder caused by internal bleeding (hemorrhagic).

1.4 Types of Strokes:

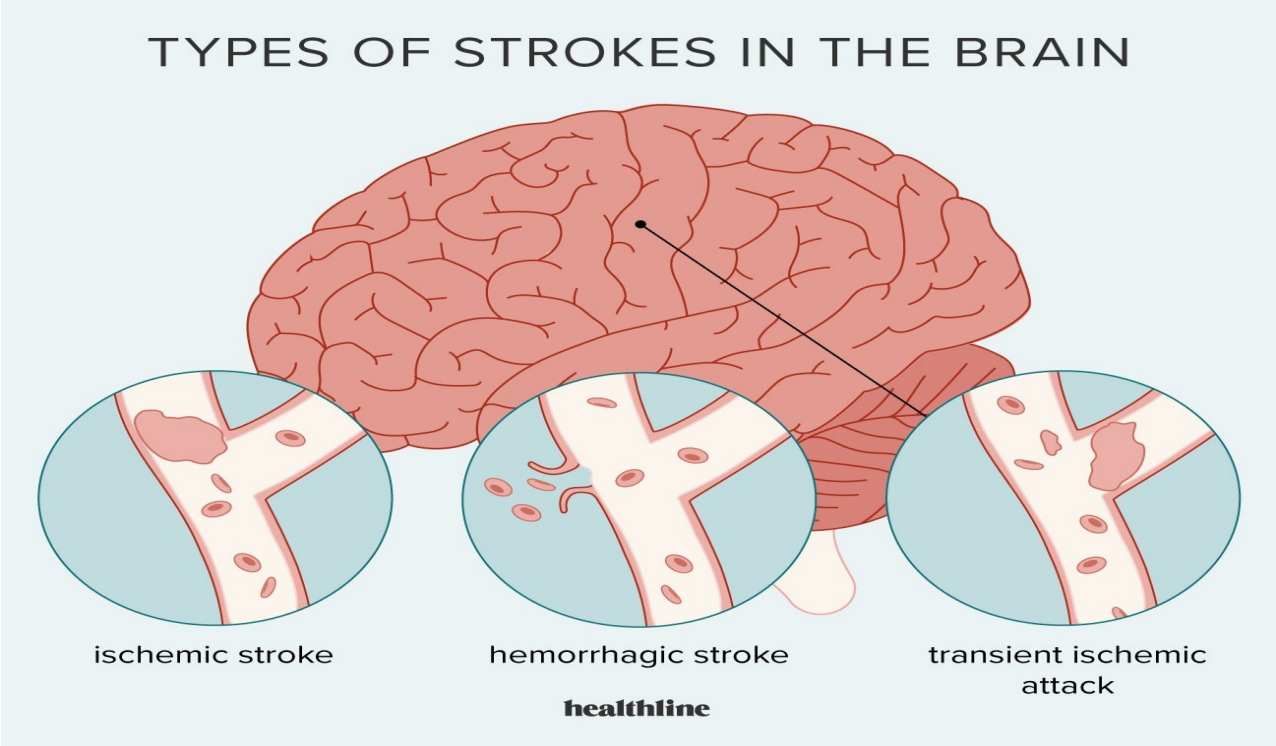


Fig 1.3

1.4.1 Transient ischemic attack (TIA) Stroke:

The transient ischemic attack (TIA), sometimes known as a "ministroke," occurs when a portion of the brain loses blood flow for a brief period of time. This results in stroke-like symptoms that normally go away within 24 hours. The fundamental difference between a stroke and a TIA is that the signs of a TIA often resolve themselves within a brief amount of time (a few hours to 24 hours). A stroke, on the contrary, can cause long-term symptoms and problems. This is because the obstruction in the circulatory vessel normally dissolves itself after a TIA.

1.4.2 Ischemic Stroke:

During a stroke that is ischemic, the blood supply to a portion of the brain is reduced, causing tissues in that area to malfunction. There are four possible explanations for this:

- A thrombotic (obstruction of a vessel in the body by a locally formed blood clot)
- Embolism (obstruction owing to a foreign embolus from elsewhere within the body),
- Systemic hypoperfusion (a broad reduction in blood supply, such as in shock)
- Thrombosis of the cerebral venous sinus.

Cryptogenic stroke (idiopathic) is defined as a stroke with no clear cause; this accounts for 30-40% of all ischemic stroke cases. There are several classification methods in acute ischemic stroke. The Oxford Community Stroke Project classification (OCSP, additionally referred to as the Bamford or Oxford classification) is based primarily on the initial symptoms; the stroke incident is categorized as total anterior circulation infarct (TACI), incomplete anterior circulation infarct (PACI), and lacunar infarct (LACI), or posterior circulation infarct (POCI) based on the extent of the symptoms. These four variables determine the severity of the ischemic stroke, the afflicted area of the brain, the underlying etiology, and the prognosis. The TOAST classification is based on clinical signs and symptoms as well as the findings of additional investigations; on this premise, A stroke can be caused by thrombosis or embolism caused by atherosclerosis of a big artery, an embolism that starts in the heart, full blockage of a tiny blood channel, another recognized cause, or an unclear reason (two probable causes, no cause discovered, or insufficient study). Cocaine and methamphetamine users are at a significant risk of having an ischemic stroke.

1.4.3 Hemorrhagic Stroke:

Hemorrhagic stroke is classified into two types:

Intracerebral hemorrhage is caused by either intraparenchymal hemorrhage (blood clot within the neural tissue) or intraventricular hemorrhage (coagulation within the brain's ventricular system).

Subarachnoid hemorrhage is blood loss that takes place beyond the brain tissue but within the skull, specifically between the membrane known as the arachnoid mater and the pia mater (the fragile deepest layer of the three meninges that protect the brain).

The above two primary kinds of hemorrhagic stroke are also two different forms of intracranial hemorrhage, which is the buildup of blood anyplace within the cranial vault; but the other forms of intracranial hemorrhage such as epidural hemorrhage (bleeding between the skull and the dura mater, that is the thick outermost layer of the meninges that surround the brain) and subdural hemorrhaging (bleeding within the subdural space), are not considered "hemorrhagic stroke"

Hemorrhagic stroke can arise as a result of blood vessel changes in the brain, including cerebral amyloidosis angiopathy, cerebral arteriovenous malformation, or an intracranial aneurysm, which can result in intraparenchymal or subarachnoid hemorrhage. Hemorrhagic stroke

typically results in specific symptoms (for example, subarachnoid hemorrhage creates a strong headache described as a thunderclap pain) or reveals evidence of a preexisting head injury, in addition to damage to the nervous system.

1.5 Symptoms of Stroke:

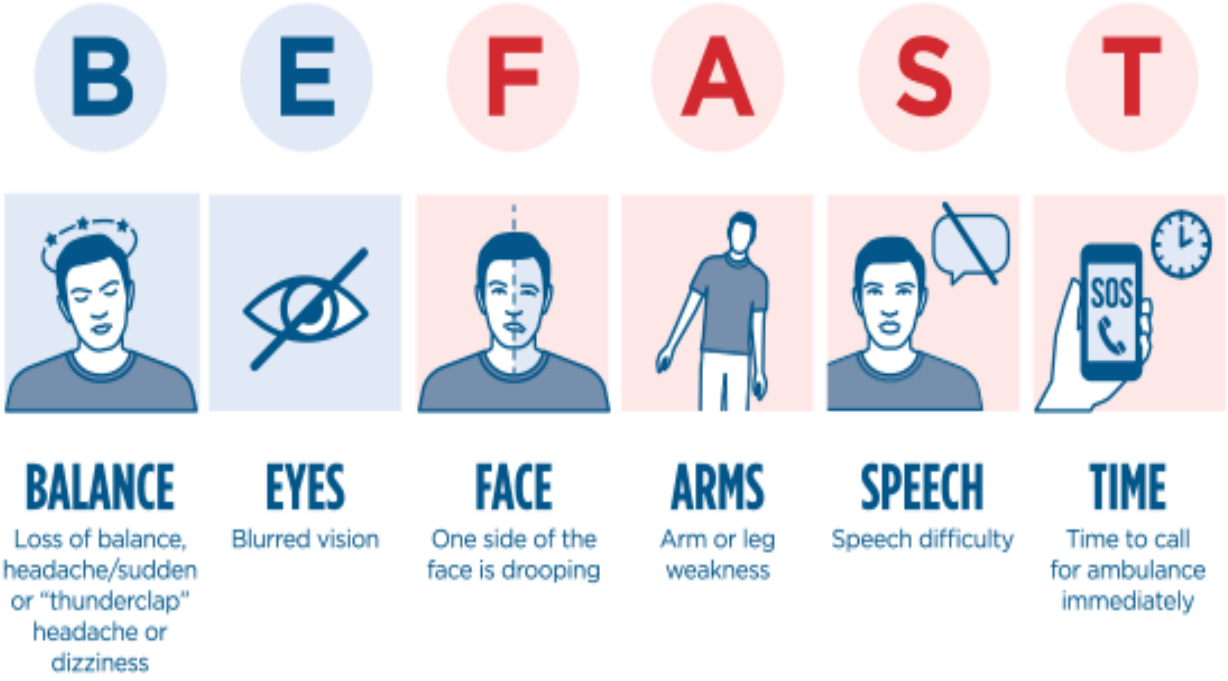


Fig 1.4

A stroke, a medical emergency caused by a sudden disruption in the flow of blood to the brain, manifests in a variety of ways, each of which indicates the probable severity and nature of the episode. These signs, typically abbreviated as FAST (Facial drooping, Arm weakness, Speech difficulty, and Time to call emergency services), serve as an important guide for both patients and bystanders in recognizing the seriousness of the situation.

Facial drooping is sometimes one of the first and most obvious indicators of a stroke. It happens usually when one side of the human face gets numb or weak, resulting in visible sagging or drooping, which is most obvious around the eyes and mouth. This imbalance in how one looks is an unmistakable sign that something is wrong neurologically and should demand rapid attention.

The second element of the FAST acronym is arm weakness, which is defined as an unexpected decrease of strength or weakness in one arm. This can show difficulty in elevating the arm or

a sense of weight, as well as a general absence of coordination. Individuals suffering from a stroke may find it difficult to execute fundamental duties with the affected hand, and the imbalance in energy between the two limbs is obvious.

The third factor, speech difficulties, emphasizes the communication difficulties that frequently accompany a stroke. Speech impairment can range between slurred or garbled phrases to an inability to convey concepts completely. Furthermore, individuals may struggle to interpret language that is spoken or written, complicating their capacity to communicate their needs or understand information. Speech-related symptoms reflect the brain's inability to manage the complex processes that contribute to language production and processing.

The FAST acronym's final component, Time to reach out emergency services, emphasizes the crucial need of quick action. In stroke situations, time is of crucial essence, and a delay in obtaining medical assistance can have a substantial impact on the possibility for recovery. The critical nature of the issue needs an instant contact to emergency services, allowing for a quick response that can limit the extent severe brain damage.

It is important to note, however, that the indications of a stroke might extend above the FAST signs, embracing a broader range of manifestations. Sensory problems, such as sudden tingling or numbness in the encounter, arms, as well as legs, may develop, indicating that there is involvement of certain sensory regions of the brain. Coordination and balance problems might also arise, resulting in dizziness, difficulties walking, or a loss of equilibrium. These muscular and sensory symptoms highlight the numerous ways in which a cerebrovascular accident can affect the complex brain networks that govern movement and physical awareness.

In conclusion, stroke symptoms include a wide range of manifestations, from the instantly noticeable face drooping, limb weakness, and speech problems to more subtle mental, sensory, and emotional abnormalities. Understanding the scope of these symptoms is critical for both patients and healthcare providers because it allows for early detection, timely intervention, and the commencement of proper post-stroke treatment. The multifarious character of stroke symptoms emphasizes the intricate nature of the brain's operations and the tremendous impact that disturbances in blood flow may have on many intellectual, sensory, including motor processes. As neurology and stroke treatment research progresses, a better knowledge of the intricacies of symptoms of stroke will lead to improved prevention, management, and rehabilitation measures, thereby enhancing outcomes for people suffering from this vital medical condition.

1.6 Factors Affecting the Stroke:

A variety of variables impact the occurrence and outcome of strokes, ranging from individual health behaviours to bigger societal and environmental influences. Understanding these multiple characteristics is critical for both preventative initiatives and effective stroke therapy.

Hypertension, often known as high blood pressure, is one of the most significant risk factors for stroke. Elevated blood pressure stresses the arteries, making them more vulnerable to injury or occlusion. Uncontrolled hypertension raises the risk of both ischemic and hemorrhagic strokes, highlighting the significance of regular arterial pressure monitoring and treatment as a fundamental preventative intervention.

Smoking, in addition to hypertension, is a modifiable risk factor with significant consequences for stroke risk. Tobacco smoke contains substances that not only damage blood vessels but also lead to the development of blood clots. As a result, smoking cessation becomes an essential component of stroke prevention, providing both immediate and long-term benefits.

Dietary habits have a critical role in stroke risk, with poor nutrition contributing to illnesses such as obesity, diabetes, and high cholesterol—all of which are established risk factors for strokes. A diet heavy in fatty and trans fats, sodium chloride, and cholesterol can cause plaque accumulation in arteries, reducing blood flow to the brain. A diet high in fruits, vegetables, whole grains, and lean meats, on the other hand, can improve cardiovascular health and lower the risk of stroke. A different approach lifestyle factor that interacts with nutrition to influence stroke risk is physical inactivity. Regular exercise not only aids in weight loss but also improves cardiovascular health lowers blood pressure, and lowers cholesterol levels. Sedentary habits, which are common in modern culture, contribute to overweight and other health problems, increasing the risk of stroke. Physical exercise, even if it is as simple as walking briskly, can be a strong preventative tool.

Diabetes, a metabolic condition characterised by increased glucose levels, is a major risk factor for stroke. Diabetes can harm blood vessels and raise the risk of atherosclerosis, a disease in which arteries constrict and stiffen. Diabetes treatment, including medication, lifestyle adjustments, and regular monitoring, is critical for lowering the risk of strokes and other coronary artery disease. Furthermore, atrial fibrillation, an abnormal heart rhythm, increases the risk of stroke. Blood can collect in the chambers of the heart with this disease, creating clots

that can go to the brain's surface and trigger a stroke. Anticoagulants are frequently recommended to treat atrial fibrillation and lower the risk of developing clots.

Genetic and genetic variables also have a role with stroke risk. Family history might reveal an individual's proclivity for specific ailments such as diabetes, high blood pressure, and cardiovascular disease, all of which increase the risk of stroke. While hereditary elements are unchangeable, being aware of familial trends enables proactive health administration and preventative interventions.

Gender and age are two more non-modifiable risk factors for stroke. Stroke risk increases significantly with age, with the possibility increasing significantly beyond the median age of 55. This is due to the cumulative impact of different risk variables, as well as changes to blood vessels and cardiovascular health with age. Men are more likely than women to have a stroke, although the risk increases with age, especially during pregnancy and the postmenopausal stage. Endocrine modifications, as well as pregnancy-related disorders, can all contribute to an increased risk throughout these specific life phases.

Disparities in stroke rates are also influenced by socioeconomic and healthcare access variables. Access to preventative healthcare may be difficult for those with lower socioeconomic levels, which might increase the incidence of risk factors and result in less effective management of preexisting diseases. Education levels have a strong correlation with health literacy, which affects people's capacity to comprehend and put preventative measures into practice. Environmental variables may also have an indirect influence on the risk of stroke by causing cardiovascular health problems, such as pollution in the air and exposure to chemicals.

A further factor in the variations in stroke rates is ethnicity and locale. African Americans, for example, have a greater risk of stroke than other racial groupings. Genetic predisposition, greater rates of high blood pressure, and socioeconomic variables all contribute to this increased risk. Additionally, residents in the "stroke belt" in America, which is an area with a greater frequency of stroke risk factors, have higher stroke rates, demonstrating the influence that environmental and local determinants.

Chronic stress, sadness, and social isolation are among the psychosocial variables that affect stroke risk. Depression may affect medication adherence and good lifestyle choices, while chronic stress can lead to bp and other risk factors for cardiovascular disease. Social exclusion

and a lack of assistance from others are linked to worse health outcomes, highlighting the link between mental and physical health in stroke care and prevention.

The interaction of these several variables highlights the complexity of the aetiology of stroke and calls for a comprehensive strategy for prevention and management. Stroke preventive public health interventions should focus on larger economic, social, and environmental causes in addition to individual behaviours. Additionally, healthcare programmes must take into account the wide range of hazards and customise therapies to meet the unique requirements of people and communities. Healthcare professionals, politicians, and citizens may together cooperate to lessen the burden of this serious medical illness by creating a thorough awareness of the variables causing stroke. A comprehensive strategy for reducing the risk of stroke and enhancing general cardiovascular health includes prevention, early identification, and efficient care.

1.7 Predictive Modeling:

Using data to predict future events and trends, predictive modeling, a sophisticated computer approach, has become a potent tool in a variety of fields. Predictive modeling is fundamentally the process of analyzing previous data, seeing trends, and making educated predictions about upcoming events or behaviors using algorithms based on statistics and machine learning approaches. This strategy is cross-disciplinary and has uses in business, medicine, marketing, environmental research, and more. Predictive modeling excels in bringing to light hidden insights, streamlining decision-making, and enhancing strategic planning.

Finance Industry:

Predictive modelling is essential to risk evaluation, portfolio administration, and fraud detection in finance. Financial institutions develop models that anticipate market trends, analyze investment risks, and optimize portfolio allocations using historical data on markets, economic indicators, and consumer behavior patterns. These models help in generating data-driven decisions, improving the efficiency and efficacy of financial strategies. They frequently use sophisticated algorithms like artificial neural networks and decision trees to make choices. Additionally, predictive modelling analyses transactional data to spot unusual trends in the field of fraud detection, allowing for the quick response and loss minimization.

Healthcare Industry:

Predictive modelling helps with personalized medication, illness forecasting, and resource management in the healthcare industry. Healthcare experts may create models to forecast particular health risks and customize treatment strategies by analyzing patient data, such as medical records, family histories, and lifestyle variables. Anticipating disease outbreaks and allocating resources most effectively during public health emergencies are two more applications of predictive modeling. Predictive models, for instance, were used to foresee the spread of the COVID-19 pandemic, allocate medical resources wisely, and direct public health measures.

Marketing Industry:

Predictive modelling in marketing changes how companies approach client interaction, revenue forecasts, and campaign optimization. Businesses may create models that forecast future purchase trends by analyzing client behavior, preferences, and past purchasing information. These models, which are frequently driven by algorithms that use machine learning, improve targeted marketing campaigns and allow businesses to provide customers with tailored experiences. Furthermore, sales forecasting models use previous sales information and outside variables to estimate future demand, enabling more precise inventory administration and manufacturing planning.

1.8 Effectiveness of Predictive Analytics:

Predictive modeling's success depends on the caliber and quantity of the data at hand. By giving users access to enormous datasets, big data technologies have dramatically improved the capacity of predictive modeling. Big data includes vast and complicated datasets that are difficult for typical data processing systems to manage. Big data's scalability benefits predictive models by making it possible to analyze enormous datasets in real-time. Across several sectors, the combination of large-scale data analytics and predictive modeling has enabled more precise forecasts, deeper insights, and improved decision-making.

Predictive modelling relies heavily on machine learning algorithms since they enable learning from data, pattern adaptation, and improvement over time. A prominent technique in predictive modeling is supervised learning, which includes teaching an algorithm on a labelled dataset while the computer discovers the connections between the input characteristics and the associated results. Once educated, the model is able to predict outcomes from fresh, unforeseen data. On the other hand, unsupervised learning is used when the data does not contain outcomes

that have been labelled. Unsupervised learning techniques like clustering algorithms may find patterns and put data points into groups based on similarities, exposing underlying dataset structures.

There are difficulties with interpretability, prejudice, and ethical issues in predictive modelling. Understanding the reasoning behind forecasts is harder as models get more complicated. Predictive models must be transparent and easy to understand, especially in industries like banking and healthcare where judgements can have far-reaching effects. Predictive models can reinforce data bias, whether it results from historical injustices or problems with sampling, producing unfair or biased results. When using sensitive data in predictive modelling, ethical issues also come into play, posing concerns about privacy, permission, and the ethical use of technology.

A crucial phase in the predictive modeling process is model validation, which makes sure the model works effectively with fresh, untested data. A typical problem in predictive modeling is overfitting, which happens when a model understands the training information too well and captures noise and unimportant patterns. Poor generalization to new data may be the outcome of this. To reduce overfitting and determine a model's resilience, regularization methods, and cross-validation are used.

With improvements in the field of artificial intelligence, and machine learning, and the incorporation of many data sources, the probable future of predictive modeling seems bright. Deep learning models, which are based on the design and operation of neural networks that exist in the human brain, are excellent at deducing complex patterns from data. These models have shown astounding success in natural language processing, picture recognition, and other challenging tasks. Predictive modeling that incorporates deep learning brings up new possibilities for handling high-dimensional data and addressing complex challenges.

The predictive modeling is a disruptive force that is changing how businesses use data to make decisions. Predictive modeling offers an effective lens using which to foresee future patterns and behaviors in a variety of fields, including economics, healthcare, marketing, and environmental research. Predictive modelling has become a dynamic and developing subject as a result of the combination of massive amounts of data, and machine learning, including advanced analytics, with persistent issues relating to ability to be understood, bias, and moral considerations. The future for predictive modeling promises the potential of more precise forecasts, better model interpretability, and expanded accessibility via user-friendly platforms

as technology develops. The next wave of advances in predictive modeling will probably be driven by the interaction between human knowledge and computing power, opening up new opportunities and insights in a variety of fields.

1.9 Importance of Predictive Modelling in Healthcare Industry:

With its important insights for physicians, administrators, and politicians, predictive modelling in health care analytics has established itself as a cornerstone of contemporary medical procedures. Predictive modelling is important in healthcare because it may use data to forecast probable health outcomes, allocate resources more effectively, and provide better patient care. This data-driven methodology not only improves decision-making but also helps to deliver healthcare more effectively and produce better patient results.

1.9.1 Advantages:

Predictive modeling in healthcare has the potential to revolutionize patient care via early intervention and individualized treatment plans, which is one of its main benefits. Predictive models can identify people who are at risk of acquiring particular health disorders by analyzing large datasets containing patient demographics, medical data, and clinical outcomes. Predictive modeling, for instance, can be used to estimate a patient's risk of developing diabetes, cardiovascular disease, or even readmission to the hospital. With this foresight, healthcare professionals may take proactive action to reduce risks and stop the course of illnesses by providing focused therapies, lifestyle changes, or carefully monitored care.

Another significant benefit of predictive modeling in the healthcare industry is the optimization of resource allocation. Healthcare organizations and hospitals face the problem of efficiently allocating resources, from controlling bed capacity to staff scheduling and inventory optimization. To effectively estimate demand, predictive algorithms can examine past statistics, patient admission trends, and seasonal fluctuations. This makes it possible for healthcare institutions to distribute resources efficiently, resulting in shorter wait times, improved patient flow, and increased operational effectiveness. Effective resource management also results in cost savings and improved use of healthcare resources.

Predictive modelling also aids in the development of evidence-based healthcare decisions. To guide treatment plans, intervention methods, and health policy choices, clinicians and administrators may depend on data-driven insights. By ensuring that medical choices are based on the most up-to-date knowledge, this evidence-based approach improves the standard of care

and patient outcomes. For instance, given on a patient's medical history, genetic makeup, and reactions to past therapies, predictive models can help determine the best pharmaceutical regimen for that patient.

Predictive modelling is essential for identifying at-risk groups and designing treatments to enhance overall community health in the field of population health management. Predictive models can identify populations or demographic groups with increased susceptibility to certain health conditions by analysing demographic data, socioeconomic characteristics, and health behaviours. In order to address particular health issues among such groups, this information enables healthcare professionals and policymakers to put into place specific actions, preventive therapies, and public health campaigns.

1.9.2 Disadvantages:

Predictive modelling in healthcare analytics has a number of benefits, but it also has certain drawbacks. The possibility of prejudice in the data utilized for instruction prediction models is one of the main worries. Inaccurate or biased results may result from the prediction model if past information contains biases based on race, gender, socioeconomic position, or other characteristics. Predictive modelling in the field of healthcare needs thorough evaluation of data quality, openness in model construction, and continual monitoring to spot and correct possible biases.

Another difficulty is understanding and interpreting the results of prediction models. It may be difficult for healthcare practitioners, particularly those with a background is data science, to comprehend the complicated inner workings of sophisticated algorithms. To promote confidence among healthcare practitioners and patients, it is essential to provide openness and interpretability in prediction models. Predictive modelling in healthcare has challenges due to inherent "black box" nature that surrounds certain powerful machine learning techniques.

Given the sensitivity of patient information, data security and privacy issues are of utmost importance in healthcare analytics. Predictive modelling depends on large datasets, which frequently include electronic medical records and other private health data. It is crucial to safeguard patient privacy and adhere to laws like the Healthcare Insurance Portability and Accountability Act (HIPAA). Implementing predictive modelling in healthcare continues to face obstacles, including protecting against unintentional access, data breaches, and maintaining secure data sharing practices.

1.10 Future Scope and Conclusion:

Healthcare analytics may change in the future thanks to the combination of predictive modeling with other cutting-edge technologies like the use of AI including the Internet of Things (IoT). Predictive models may become even more precise and timely with the help of wearable technology, and ongoing monitoring, including real-time data streams. For instance, smartphones that monitor patients remotely can produce continuous data streams that allow predictive modeling to identify minute changes in medical conditions, enabling swift intervention and preventive interventions.

In conclusion, predictive modelling in healthcare analytics is crucial because it provides a wealth of benefits for population health management, early intervention, resource optimisation, and evidence-based decision-making. A more effective and efficient way to offer healthcare is to be able to predict health outcomes, identify groups that are at risk, and allocate resources optimally. But issues like interpretability, data privacy, bias in the data, and the danger of over-relying on models call for careful thought and continual attempts to solve these issues. Predictive modelling in healthcare presents the potential of increased model interpretability, more precise predictions, and improved interaction with new technology for the improvement of patient care along with overall healthcare results.

CHAPTER 2:

REVIEW OF LITERATURE AND RESEARCH METHODOLOGY

2.1 Review of Literature:

In the study conducted by **Rahman, Senjut et.al., (2023)** on the Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. Several Classification techniques were compared under the study. The objective was to study the effectiveness of these algorithms to predict the occurrence of the stroke. The study found that the Random Forest Classifier had an accuracy of 99% compared to all other ML classifiers. When using the chosen features as input, the three-layer deep neural network (4-layer ANN) technique gave a greater accuracy of 92.39%. The results of the study demonstrated that machine learning methods performed better than deep neural networks. The study also concluded that severity of stroke can be lessened upon the early detection of the symptoms.

A study on Deep Learning and Machine Learning for Early Detection of Stroke and Hemorrhage was conducted by **Al-Mekhlafi, Zeyad Ghaleb et.al., (2022)** The effectiveness of machine learning, deep learning, and a hybrid approach combining deep learning and machine learning on the Magnetic Resonance Imaging (MRI) dataset for cerebral hemorrhage was evaluated using a dataset encompassing medical, physiological, and environmental assessments for stroke. The high-dimensional dataset was represented in a low-dimensional data space using the t-distributed Stochastic Neighbour Embedding technique. In the meanwhile, the irrelevant features were eliminated using the Recursive Feature Elimination method (RFE), which was used to order the features according to importance and their link to the target feature. Among the algorithms, the Random Forest method performed the best, with an overall accuracy of 99%. The Precision, Recall, and F1 scores for this algorithm's classification of stroke cases were 98%, 100%, and 99%, respectively. The MRI image dataset from the second dataset was assessed using a hybrid AlexNet+SVM model and AlexNet model. The hybrid model AlexNet+SVM outperformed the AlexNet model in accuracy, sensitivity, specificity, and Area Under the Curve (AUC) metrics, respectively, reaching values of 99.9%, 100%, 99.80%, and 99.86%.

Biswas, Nitish er.al., (2022) in A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach concluded that for effective early treatment and to lower the mortality rate, a precise stroke prediction is required. In order to better effectively detect stroke using unbalanced data, this study suggests a machine learning technique. This study balanced the data by using the random oversampling (ROS) approach.

In this work, we analyze eleven classifiers: Support Vector Machine, Random Forest, K-nearest Neighbors, Decision Tree, Naive Bayes, Vote Classifier, AdaBoost, Progressive Boosting, Multi-Layer Perception, which is and Nearest Centroid. Before employing the oversampling approach to balance the data, ten classifiers produce findings that are more accurate than 90%, while four classifiers produce results that were more accurate than 96%. According to the results, the support vector machine has the greatest accuracy, with a 99.99% F1-measure, 99.99% recall, and 99.99% precision values. With a 0.001% inaccuracy, Random Forest has the second-highest accuracy at 99.87%. Additionally, the most realistic model is used to build both a user-friendly online app and a mobile application that is easy to use.

Risk factor identification for stroke prognosis using machine-learning algorithms was the study conducted by **Ahammad, Tanvir et. al., (2022)** in which they found Several studies have centred on predicting strokes in patients. On the publicly accessible datasets, the majority had a respectable accuracy ratio of approximately 90%. A requirement for research is the ability to significantly improve classifier quality by combining many pre-processing jobs. Researchers should also identify the main causes of stroke disease and utilize sophisticated classifiers to predict the chance of stroke. On the basis of a publicly available clinical dataset, this study proposes an improved method for determining possible risk factors and forecasting the incidence of stroke. The approach takes into account and fills up considerable gaps left by earlier research. To identify the existence of stroke, it uses 10 classification models, including powerful boosting classifiers. Overall, the primary accomplishment of their work was to outperform state-of-the-art methods for stroke dataset by achieving a greater percentage of stroke prognosis (97% efficiency using boosting classifiers). In order to best predict patients' strokes in the actual world, doctors can employ gradient as well as ensemble boosting-tree-based models. Additionally, this analysis demonstrates that the most important risk variables are age, heart disease, blood sugar, hypertension, and marital status. In addition, the remaining qualities are crucial to getting the optimum results.

Dev, Soumyabrata et.at., (2022) in their paper on A predictive analytics approach for stroke prediction using machine learning and neural networks found that the most critical risk variables for diagnosing stroke in patients are age, heart disease, average blood glucose level, and hypertension. Furthermore, compared to employing all available input characteristics and other benchmarking methodologies, a perceptron neural network constructed using these four parameters has the greatest rate of accuracy and lowest miss rate. We describe our findings using a balanced dataset produced using sub-sampling approaches because the dataset is

significantly unbalanced with regard to the incidence of stroke. Age, heart disease, the average level of glucose in the blood, and hypertension were the most important risk factors for identifying stroke in patients. Furthermore, a perceptron artificial neural network built using these four variables has the highest rate of accuracy and the lowest miss rate when compared to using all available input features and other benchmarking approaches.

In a study by **Debnath, Sourav et.al., (2021)** on Predictive Analysis for Risk of Stroke Using Machine Learning Techniques found that analysis of several warning indicators can help predict stroke. In this project, we use a variety of machine learning methods to build a process of stroke risk prediction using our dataset. The first task in this study is assumed to be data imputation, feature selection, and data preparation. As the feature is used to simulate training and testing, several physical characteristics like heart disease, age, BMI, gender, hypertension, etc. are taken into consideration. To forecast the risk probability of stroke, this study implements the AdaBoost classifier, artificial neural network (ANN), decision tree classification algorithm, k-nearest neighbour (KNN) classifier, random forest, stochastic descent with gradients (SDG), support vector machine (SVM), and XGBoost classifier. Following that, the voting classifier is applied to these eight conventional classifiers. We discovered 98% accuracy in identifying the risk factor for stroke after comparing several machine learning algorithms using voting classifiers, which is superior than previous models.

Emon, Minhaz Uddin et.al., (2020) conducted a study on Performance Analysis of Machine Learning Approaches in Stroke Prediction in which they proposed that with the incidence of hypertension, heart disease, average glucose level, body mass index level, smoking status, prior stroke, and age, early prediction of stroke disorders using various machine learning algorithms is possible. In order to get the best accuracy, the output of the basic classifiers was then combined using the weighted voting method. And in this investigation, the weighted voting classifier outperforms the base classifiers, with an accuracy of 97%. This model provides the most accurate stroke prediction. The weighted voting classifier's area under the curve value is similarly high. The weighted classifier has the lowest rates of false positives and false negatives when compared to the competition. As a result, weighted voting is virtually the ideal classifier for predicting strokes, allowing both patients and doctors to prescribe treatments and identify probable strokes early on.

In their study **Wang, Wenjuan et.al., (2020)** This systematic review's objective is to locate and evaluate the reporting of and development of ML models for predicting outcomes following

stroke. The application of ML to forecast stroke outcomes is growing. Few, nevertheless, adhered to the fundamental reporting requirements for clinical prediction tools, and none made their models accessible in a form that allowed for usage or evaluation. Before it can be seriously considered for practice, significant changes in ML research conduct and documentation are required. Mortality and functional results were the consequences of stroke that were most commonly predicted. Random forests, support vector algorithms, decision tree models, and neural networks were the most frequently utilized machine learning techniques. With a standard deviation of 22 predictors taken into account, the overall population size was 475 on average.

Dev, Soumyabrata et.al., (2020) conducted a study on Identifying Stroke Indicators Using Rough Sets in which they found thet Globally, a variety of data mining approaches have been applied to the accurate prediction of the incidence of stroke according to the risk variables linked to the patients' electronic health records (EHRs). To improve prediction accuracy, it is necessary to remove the majority of the thousands of characteristics that EHRs frequently contain since they are repetitive or useless. The use of feature-selection techniques can enhance the model's predictive power and facilitate efficient handling of the archived input characteristics. For the purpose of detecting strokes, the numerous characteristics in EHR records are methodically analyzed. We provide a unique rough-set-based method for assessing the significance of different EHR information in stroke detection. They found that the most important factors for diagnosing stroke in patients were age, average blood sugar levels, heart disease, and hypertension. We also compared the suggested method to other well-known feature-selection methods. When rating the significance of each characteristic in identifying stroke, we achieved the best result.

Nwosu, Chidozie Shamrock et.al., (2019) in the study on Predicting Stroke from Electronic Health Records concluded that numerous risk factors have been linked in studies to the start of stroke in a person. Using these characteristics and data mining techniques on patient medical information, it has been possible to forecast the occurrence of stroke. But there hasn't been much research on how various stroke risk factors interact using electronic health records. In this study, we examine patient electronic health data to determine how risk variables affect stroke prediction. Additionally, we compare the effectiveness of cutting-edge artificial intelligence algorithms to forecast stroke using electronic medical data.

Hung, Chen Ying et.al., (2019) conducted a study on Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health

record database. The objective of the study was to attain great modelling power, deep neural networks (DNNs) are artificial intelligence algorithms. It is necessary to evaluate and validate the use of DNNs with large-scale data to estimate the risk of stroke. This work uses a sizable electronic health record collection to use a DNN for developing a stroke predicting model. Techniques and outcomes We performed a retrospective population-based investigation using the Taiwan National Health Insurance Research Database. The database was split into two testing sets (each with a 10% sample size) and one developmental dataset for training the models (using about 70% of all patients to train models and 10% for parameter tuning). The model's performance was comparable to those of other medical risk assessment scores. Conclusions One may create a highly effective model for ischemic stroke risk assessment by using a DNN algorithm to this sizable electronic health record set. To find out whether or not a similar DNN-based IDSS may enhance clinical practice, more investigation is required.

Min, Seung Nam et.al., (2019) Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea found that hypertension, heart illness, diabetes, dysregulation of blood sugar metabolism, atrial fibrillation, and unhealthy habits are some of the risk factors for stroke that may be controlled. In order to create a stroke pre-diagnosis algorithm with the theoretically controllable risk variables, we set out to build a model equation. Methods: For model development, we employed logistic regression and data from the Korea National Health Insurance Service (NHIS) database. We looked over 500,000 participants' NHIS records. 367 individuals with strokes were the subject of the regression analysis. In 65% of instances, the proposed model was able to distinguish between healthy participants and stroke patients. The model used in this work can be used in a clinical context to predict the likelihood of having a stroke within a year and enhance stroke preventive tactics for high-risk patients. The method used to create the algorithm for preventing strokes can be utilized to create analogous models for the early identification of other illnesses.

In paper by **Jeena, R. S et.al., (2017)** on stroke prediction using Support Vector Machine found that early stroke identification is crucial for prompt prevention and therapy. According to research, measurements taken from several risk indicators might provide useful data for stroke prediction. This study looks at the many physiological markers that are considered to be risk factors for the stroke prediction. Support Vector Machine (SVM) training and testing on data taken from the International Stroke Trial database proved successful. In this study, we built SVM using several kernel functions and discovered that the linear kernel provided 90% accuracy.

Singh, M. Sheetal et.al., (2017) in their paper on Stroke prediction using artificial intelligence concludes that comparing the effectiveness of prediction data mining systems has been the focus of several studies aimed at predicting different illnesses. On the cardiovascular wellness Study (CHS) dataset, we evaluate several approaches for predicting stroke with our algorithm in this article. Here, the principal component analysis method is used to reduce the dimension, the algorithm for decision trees is used to pick the features, and back propagation neural networks classification technique is used to build a classification model. Our study offers the best prediction model for the stroke illness with 97.7% accuracy after analyzing and comparing classification efficiency with other approaches and variation models.

In the paper titled Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review by **Goldstein, Benjamin A. et.al., (2017)** found that various studies often used a wide variety of predictors and had relatively high sample sizes (median sample size 1/426 100). Most frequently employed validation methods (n1/494 of 107) and stated model coefficients for repeatability (n1/483) were both used. Studies used median 1 out of 427 predictor variables, however they seldom used longitudinal data (n1/437), limiting their ability to fully exploit the range of EHR data. Only 26 of the research (n1/450) used cross-site validation, and fewer than fifty percent of the investigations were multicenter. Many studies didn't adequately address EHR data biases such data that is missing or loss to follow-up. Mortality (0.84), clinical forecasting (0.83), hospitalization (0.71), and service utilization (0.71) had the highest average c-statistics for various outcomes. Conclusions: For clinical risk prediction, EHR data provide both potential and obstacles.

Yahiya, Selma et.al., (2018) in their study on Classification of Ischemic Stroke using Machine Learning Algorithms states that the two categories of stroke shared a lot in common, making it challenging to precisely categorized the instances using medical techniques. Additionally, there are no definite distinctions between these categories. This research examined and analyzed recent data on ischemic stroke categorization. Additionally, the study used the k closest neighbors' method and decision tree algorithm to construct a model of classification for ischemic stroke. 400 cases from several Sudanese hospitals make up the dataset upon which the categorization model is built. Medical experts may categorize and identify ischemic stroke patients using the decision tree algorithm's findings. The study also showed that some characteristics may be utilized directly to identify the kind of ischemic stroke. These findings aid medical professionals in the categorization of ischemic strokes. The majority of ischemic stroke cases, according to the findings, are thrombotic ischemic strokes.

Luk, James K.H et.al., (2016) in their study Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects mainly focused on variable of age and found that patients of three age groups—65, 65 and 80, and 80 years—had significant disparities in their clinical traits and stroke sequelae. The older age groups had lower overall FIM scores both at admission and release. There was no discernible difference in the FIM score changes across these age groups. Age was not a reliable indicator of a successful outcome. In all patients and in all age groups, FIM at admission was an independent predictor of a successful outcome (discharge FIM 90). Predicting a positive prognosis for all patients was employment prior to stroke. Home ownership was a predictor for both the general population and the group between 65 and 80 years old prior to stroke. In the group of 80 years, the duration of stay indicated a favorable outcome. Conclusions: Predictors of a successful result after stroke rehabilitation include functional status at admission, employment, and property ownership prior to stroke, but not age per se. Age should not be a factor in denying stroke patients comprehensive therapy because older patients exhibit equivalent progress throughout rehabilitation.

Bin Emdad, Forhan et.al., (2016) conducted a study Towards Interpretable Multimodal Predictive Models for Early Mortality Prediction of Hemorrhagic Stroke Patients. The results of the study were that clinicians are searching for data-driven decision support tools as a result of the rising fatality rate from stroke over the previous eight years. Deep-learning-based prediction models with extremely fine EHR (electronic health record) data have recently demonstrated improved potential for predicting health outcomes. However, there hasn't been much research into the application of EHR-based deep neural network models for stroke with hemorrhage outcome prediction. This study suggests a group deep learning architecture to forecast early death in patients receiving intensive care unit care for hemorrhagic stroke. The suggested ensemble model outperformed the fusion model and other standard models (logistic regression, random forest decision tree, and XGBoost) with an accuracy of 83%. Additionally, we interpreted the ensemble model using SHAP values to pinpoint key characteristics for the prediction.

Saleh, Hager et.al., (2019) conducted a study on Stroke Prediction using Distributed Machine Learning Based on Apache Spark. The results of their study found one of the leading causes of severe, long-term disability and mortality worldwide is stroke. On the Health Care Dataset Stroke, we compare several distributed machine learning techniques in this work for predicting strokes. Apache Spark, a big data platform, is used to carry out this study. One of the most widely used big data systems, Apache Spark contains an MLlib library to manage huge data.

Spark and MLlib work together as an API to enable machine learning algorithms. Building the stroke prediction model required the use of four different machine learning classification algorithms: Decision Tree, Random Forest, Support Vector Machine Classifier, and Logistic Regression. The application of cross-validation and hyperparameter adjustment with machine learning techniques improved the outcomes. Performance metrics for machine learning models were calculated using precision, recall, accuracy, and F1-measure. According to the findings, Random Forest Classifier had the highest accuracy, at 90%.

Teerapat Kansadub et.al., (2016) in their study Stroke Risk Prediction Model based on Demographic Data used navie byes, decision tree and neural network to study the stroke risk. They found that stroke is currently the third most common cause of death across all life spans. The stroke was responsible for 255,307 incidents of death between 1994 and 2013, according to data from the Bureau of the National Economic and Social Development Board (NESDB). The length of therapy for stroke patients depends on their symptoms and any organ damage. The ability to forecast stroke illness through data analysis techniques like data mining would appear to be advantageous in lowering the number of individuals at risk before the condition ever manifests. The three classification algorithms utilized in this study—Decision Tree, Naive Bayes, and Neural Network—are model-based, outperform conventional statistics, and have a suitable model for identification. The demographic data of patients is the range of data usage. Before modelling, this work's initialization steps included attribute selection, grouping, and resampling. Indicators for evaluation in this study included area under the ROC curve (AUC) and accuracy. Naive Bayes had the best in AUC, and decision trees are the most accurate. The diagnosis of the patients should be a part of future study.

2.2 Statement of Problem:

- What are the most important health and lifestyle factors related to stroke risk?
- How can we create an accurate and dependable predictive model?
- How might the prediction model be used to focus interventions on persons at high risk of stroke?

2.3 Need for the study:

After heart disease, stroke is the second-leading cause of mortality worldwide. The World Health Organization predicts that the number of stroke-related fatalities will increase to a total of 12.2 million by 2030 from the estimated 9.1 million in 2020. The cause may be the result of

different concomitant medical illnesses, lifestyle choices, or hereditary factors. Understanding and evaluating the condition and associated risk factors becomes crucial. Utilizing current methods for analysis of information and predictive modelling can therefore aid in the development of stroke prevention strategies that will eventually enhance both the health and wellbeing of the individual and the community.

In the fields of healthcare and public health delivery, the research of predictive modeling of stroke hazards utilizing health and lifestyle variables is of utmost significance. Stroke is a common term for a "brain assault" and is one of the main causes of disability and death globally. Through the use of sophisticated modeling tools, it is now possible to forecast and comprehend the risk factors for stroke, offering a revolutionary new way to practice preventative treatment, personalize medication, and allocate healthcare resources.

Health factors such as High blood pressure, sometimes known as hypertension, are one of the most important health markers. It is generally known that hypertension raises the risk of both ischemic as well as hemorrhagic strokes. In order to identify those who have an increased risk of stroke, predictive modeling can analyze previous blood pressure data while taking patterns and swings into consideration. With the use of this knowledge, healthcare professionals may apply targeted treatments, such as managing medications and lifestyle changes, to reduce the risk of stroke and regulate blood pressure.

The modelling of stroke risk that takes into account lifestyle factors is essential. Significantly increasing the risk of stroke include unhealthy lifestyle decisions including smoking, being sedentary, and having bad eating habits. Predictive algorithms can examine previous lifestyle data to find trends linked to a higher risk of stroke. For instance, a person may be identified as having a greater risk of having a stroke if they have a history of cigarette smoking, engage in little physical exercise, and consume a diet heavy in saturated fats.

These insights enable medical professionals to use targeted counseling and intervention techniques to encourage better lives and lower the risk of stroke.

2.4 Scope of the Study:

Predictive modelling of stroke risk factors combining health and lifestyle variables has enormous promise, with enormous potential for revolutionary advances in preventive healthcare, personalised treatment, and population health management. The use of cutting-edge predictive modelling approaches offers a chance to revolutionise healthcare delivery, risk

assessment, and intervention design as societies struggle with an increasing incidence of stroke-related mortality and morbidity.

This study's scope is broad and includes a variety of elements that together help us understand and control stroke risk factors more thoroughly. The recognition and evaluation of many health markers that are essential to determining stroke risk is one of the main areas of focus. Predictive models rely heavily on health indicators as inputs, including but not restricted to cholesterol levels, blood pressure, type 2 diabetes, and cardiovascular health measures. These models may identify fine distinctions and connections by diving into previous health knowledge and patterns, offering a clearer understanding of the intricate interaction of factors influencing stroke risk.

By adding behavioural and environmental factors, the incorporation of way of life indicators into prediction models broadens the reach. Significantly increasing the risk of stroke include unhealthy lifestyle decisions including smoking, being sedentary, and having bad eating habits. Predictive modelling can find trends in lifestyle information, spot high-risk behaviours, and calculate how much of an influence they have on stroke risk. This information is crucial for creating targeted treatments and public health initiatives that attempt to change unhealthy lifestyle habits and reduce stroke risk on a larger scale.

The inclusion of socioeconomic variables broadens the scope further by acknowledging the complex relationship between social variables of wellness and stroke risk. People with lower socioeconomic position may have more difficulty getting healthcare, deal with greater levels of long-lasting stress, and adopt good lifestyle habits. Predictive modelling may identify susceptible groups by including socioeconomic characteristics, allowing policymakers and healthcare professionals to create interventions that address the root causes of social determinants and lessen health inequalities.

In order to improve the precision and application of predictive models, the scope also covers the investigation of cutting-edge technologies and approaches. Deep learning and other sophisticated algorithms for machine learning have the capacity to find nuanced patterns in large datasets. The potential for creating more reliable and accurate prediction models is increased by investigating the abilities of these mathematical techniques in combination with conventional statistical methods.

2.5 Objectives of the Study:

- Determine the most essential health and lifestyle factors linked to stroke risk.
- Create a predictive algorithm that recognizes persons who are at high risk of having a stroke.
- Test the predictive model using test data.
- Compare the accuracy of various ML algorithms in predicting the stroke.

2.6 Variables Under the Study:

2.6.1 Gender :

Gender appears as a key variable influencing stroke risk in the research of predictive modeling of stroke risk factors utilizing health and lifestyle indicators. Despite the fact that anybody can have a stroke, there are clear disparities among men and women in terms of frequency, risk factors, and results. Different patterns of stroke risk are influenced by gender-specific variances in health indicators, including hormonal impacts and genetic predispositions. Because men and women may have different risk profiles, adding gender as a variable to prediction models improves their accuracy and applicability. This emphasizes the significance of gender diversity in the creation of prediction models for the evaluation of stroke risk, as this will help to provide more individualized and successful preventative measures for a variety of groups. Gender being a categorical variable is coded as 1 and 2 where 1 represents male and 2 represents female.

2.6.2 Age:

In the research of predictive modelling of stroke risk factors utilising health and lifestyle indicators, the variable of age is a key element. The risk of stroke increases dramatically with age, and age is a reliable predictor of this risk. Numerous health and lifestyle changes brought on by ageing result in changes in risk factors. Given that key indicators of health, such as cholesterol levels, blood pressure, and the incidence of chronic illnesses, have a tendency to change over time, predictive models must take these age-related dynamics into consideration. Age-dependent differences are also seen for lifestyle variables including smoking, food, and physical activity. Predictive modelling that includes age as a variable improves risk assessments and allows for a deeper knowledge of the complex interactions between age, health, as well as lifestyle in affecting stroke outcomes.

2.6.3 Hypertension:

In the predictive modelling of stroke risk factors utilising health and lifestyle indicators, the variable of hypertension is crucial. High blood pressure, often known as hypertension, is one

of the most important and controllable indicators of risk for stroke. Predictive models must carefully take into account how blood pressure changes over time, utilising past data to spot patterns and trends. Increased blood pressure adds to blood vessel deterioration, which raises the risk of ischemic and strokes with haemorrhage. An intricate knowledge of how hypertensive interacts with various other medical conditions and lifestyle factors is necessary for effective prediction modelling for stroke risk. Given the fundamental role that hypertension plays, prediction algorithms may now classify people according to the blood pressure profiles, allowing healthcare practitioners to customise therapies and preventative measures. In addition to improving the precision of stroke risk assessments, the inclusion of hypertensive as an indicator in predictive modelling emphasises the crucial role that blood pressure control plays in reducing the total burden of stroke.

2.6.4 Heart disease:

In the predictive modelling of stroke risk factors utilising health and lifestyle indicators, the variable of heart disease acquires substantial relevance. As the cardiovascular system's health is closely related to the outcomes of strokes, those with pre-existing heart issues are at an increased risk of having a stroke. Since each of the several heart disease symptoms, such as heart failure, coronary artery disease, and arrhythmias, adds distinct dimensions to stroke risk, predictive models must take them into consideration. Heart disease frequently indicates an undiagnosed vascular susceptibility whereby impaired blood flow and structural deviations might facilitate both ischemic and haemorrhagic strokes. For a risk assessment to be correct, it is essential to comprehend how heart disease along with various health markers interact. By include coronary artery disease as a variable, predictive models are better equipped to identify subtle trends and adjust risk forecasts for people with certain cardiovascular profiles. This increases the accuracy of stroke risk evaluations while also highlighting the need for comprehensive cardiovascular treatment as a key component in reducing stroke risk. Recognising the heart disease variable in predictive modelling is a crucial first step towards individualised and targeted therapies for those who are at higher risk of stroke in the larger context of preventive healthcare.

2.6.5 Marriage:

The indicator of marital status offers a fascinating dimension to the investigation of cerebrovascular risk factors prediction modeling by combining lifestyle and health indicators. Research indicates that marital status may be an important feature in prediction models, despite

the fact that the association involving marriage and risk for stroke is complicated and nuanced. Married people may benefit from societal assistance, companionship, and shared obligations in terms of their health. Predictive models must take into account how these psychosocial variables may affect health outcomes, such as the chance of stroke. On the other hand, marriage-related pressures and lifestyle decisions might also have an influence on health. Individuals who are single, divorced, married, or widowed may have different risk profiles, depending on a variety of variables such as emotional support, financial security, and the possible impact of significant life events. Predictive modeling that takes marital status into account improves risk assessments and emphasizes the significance of comprehending the social including relational circumstances that affect general health and, possibly, stroke risk. It emphasizes the necessity for a comprehensive strategy that takes into account the interdependence of lifestyle, social, and health variables in predicting and avoiding unfavorable health effects like stroke.

2.6.6 Work type:

In the research of predictive modelling of stroke risk factors using health and lifestyle indicators, the variable of employment type emerges as a major contributor. Workplace environments and workplace-related factors can have a significant impact on a person's health and lifestyle decisions, which can affect stroke risk. Predictive models need to take into consideration the variety of work situations, taking into account the possibility of occupational stress, sedentary behaviour, and exposure to environmental elements that may increase the risk of stroke. People who work in stressful jobs or have unpredictable schedules, for instance, maybe more at risk. Additionally, the nature of the job may have an impact on eating choices, physical activity levels, and access to preventative healthcare. Recognition of job type as a variable improves the accuracy of prediction models and enables a more detailed understanding of the complex interactions between occupational characteristics, and health, including stroke risk. This method emphasizes the significance of taking into account a person's daily activities and routine in a larger perspective when assessing their stroke risk, ultimately resulting in focused and successful preventative interventions in a variety of professional contexts. The work type also includes the nature of work that may be private, full-time, part-time or unemployed.

2.6.7 Residence:

Because urban and rural contexts present different possibilities and difficulties for health and well-being, predictive models must take these distinctions into account. Urban dwellers may experience greater amounts of pollution, more availability to bad food alternatives, and more stresses, all of which may raise their risk of stroke. On the other hand, residents in rural locations could struggle with restricted access to healthcare services and adopt various occupational and lifestyle habits that can raise their risk of stroke. Researchers can find trends unique to rural and urban residents by integrating home type as an element in predictive modelling, enabling targeted treatments. This method takes into account the contextual variations that affect socioeconomic conditions, environmental exposures, and health behaviours in various residential contexts. Not only does the addition of housing type in prediction models improve the precision of stroke risk estimates, but also emphasises the need of taking environmental determinants of health into account. In order to develop more effective preventive measures, it is important to address the intricate relationships among residence type, good health, and lifestyle, taking into account the fact that different treatments may be required to address the particular problems that come with living in an urban or rural area.

2.6.8 Sugar/Glucose Level:

In the predictive modelling of stroke risk factors utilising health and lifestyle indicators, the variable of average blood glucose level takes vital relevance. Elevated glucose levels are a known risk factor for stroke and are frequently a sign of diseases like diabetes or problems with glucose metabolism. Because chronic hyperglycaemia increases the likelihood of atherosclerosis, inflammation, and vascular damage—all of which are risk factors for stroke—predictive models must carefully take into account the importance of average blood glucose levels in determining stroke risk. Predictive modelling becomes much more difficult as a result of lifestyle factors, such as food and exercise habits, which directly affect blood glucose levels. Furthermore, a diagnosis of diabetes or changes in blood glucose levels may interact with additional medical conditions, such as high blood pressure and cholesterol, increasing the risk of stroke in general. Researchers can identify complex patterns in glycaemic management and its relationship to stroke risk by using average blood sugar levels as an indicator in prediction models. This highlights the connection between metabolism health and cardiovascular results as well as improving the accuracy of stroke risk assessments. The need of thorough, multidimensional risk assessments is highlighted by the realisation that average glucose levels are a crucial variable in predictive modelling, enabling the development of treatments and preventative measures that are specifically targeted at people with certain glycaemic profiles.

2.6.9 BMI:

In the predictive modelling of stroke risk factors combining health and lifestyle indicators, the parameter of Body Mass Index (BMI) takes a vital role. BMI, which calculates body fat from a person's height and weight, is a useful proxy for determining adiposity overall obesity. A higher BMI has repeatedly been associated with a higher risk of cerebrovascular accident, including ischemic and hemorrhagic. BMI must be thoroughly included in predictive models as a continuous variable, taking into account its impact on numerous indicators of health and lifestyle variables. Obesity frequently occurs in conjunction with comorbid conditions such as diabetes, cardiovascular disease, and dyslipidaemia, which all of which are involved in the pathogenesis of strokes. The development of risk calculations based on BMI recognises the complexity of being overweight as a factor in health. This strategy highlights the need for focused treatments while also improving the accuracy of estimates of stroke risk. The fact that BMI is a crucial component in stroke prevention highlights the need of tackling overweight as a controllable risk factor, directing personalised and successful solutions for people with different body compositions. Such predictive modelling helps to elucidate the complex relationship between BMI, good health, and stroke risk in the larger context of public health.

2.6.10 Smoking:

In predictive modelling of stroke risk factors utilising health and lifestyle indicators, the variable of smoking plays a crucial role. It is generally known that smoking increases the risk of stroke and has negative effects on the circulatory system. To accurately predict outcomes, predictive models must carefully take into account smoking patterns as an unpredictable variable, taking into account both the short- and long-term effects on vascular health. Smoking raises the risk of both strokes that are ischemic or hemorrhagic by promoting atherosclerosis, causing more blood clots to develop, and raising blood pressure. Because they affect the severity of risk, lifestyle factors like the quantity and degree of smoking complicate predictive modelling. Smoking being a crucial factor highlights how crucial quitting tobacco is for preventing strokes. This method informs personalised treatment plans for people with different smoking histories, resulting in more successful public health programmes designed to lessen the general burden of stroke. Predictive modelling that includes smoking as a variable offers a thorough understanding of the complex connection that exists between tobacco consumption, lifestyle variables, and stroke risk, guiding proactive actions for individuals and population-level health.

2.6.11 Stroke Occurrence:

The incidence of a stroke itself is, without a doubt, the variable that is dependent in the investigation of predictive modelling of stroke risk factors. This crucial event is the culmination of the intricate interactions between several health and lifestyle factors, making it the focus of prediction models. A stroke, which is either ischemic or hemorrhagic, is a serious medical condition that has a high morbidity and fatality rate. The prediction models are made to examine a wide range of independent factors, including diabetes, high blood pressure, cigarette smoking, BMI, among others, and determine how they collectively affect the chance of having a stroke. The models serve as an important tool for risk identification by attempting to identify correlations, trends, and thresholds of risk that indicate an increased likelihood of stroke. Predictive models provide an anticipatory approach to stroke prevention by examining historical data, lifestyle habits, and health markers. In order to identify patients who are more at risk and perform targeted treatments, the models serve as essential decision-support tools for medical professionals. These treatments might include everything from tailored health education programmes to pharmaceutical management and lifestyle changes. Predictive modeling's incorporation of the dependent variable, which is "stroke" emphasises the seriousness of the projected health result and draws attention to the possibility of early action and risk reduction. The predictive models essentially operate as a link between the numerous risk factors and the actual, life-altering stroke occurrence, enabling a more educated and proactive approach towards stroke prevention and management of healthcare.

2.7 Hypothesis:

1) Ho: Occurrence of Stroke is not dependent on Health and Lifestyle Factors

   H1: Occurrence of Stroke is dependent on Health and Lifestyle Factors

2) Ho: Predictive models cannot accurately predict the possibility of the occurrence of stroke

   H1: Predictive models can accurately predict the possibility of the occurrence of stroke.

3) H0: All predictive models have the same level of accuracy of prediction.
   H1: All predictive models do not have the same level of accuracy of prediction.

2.8 Research Methodology:

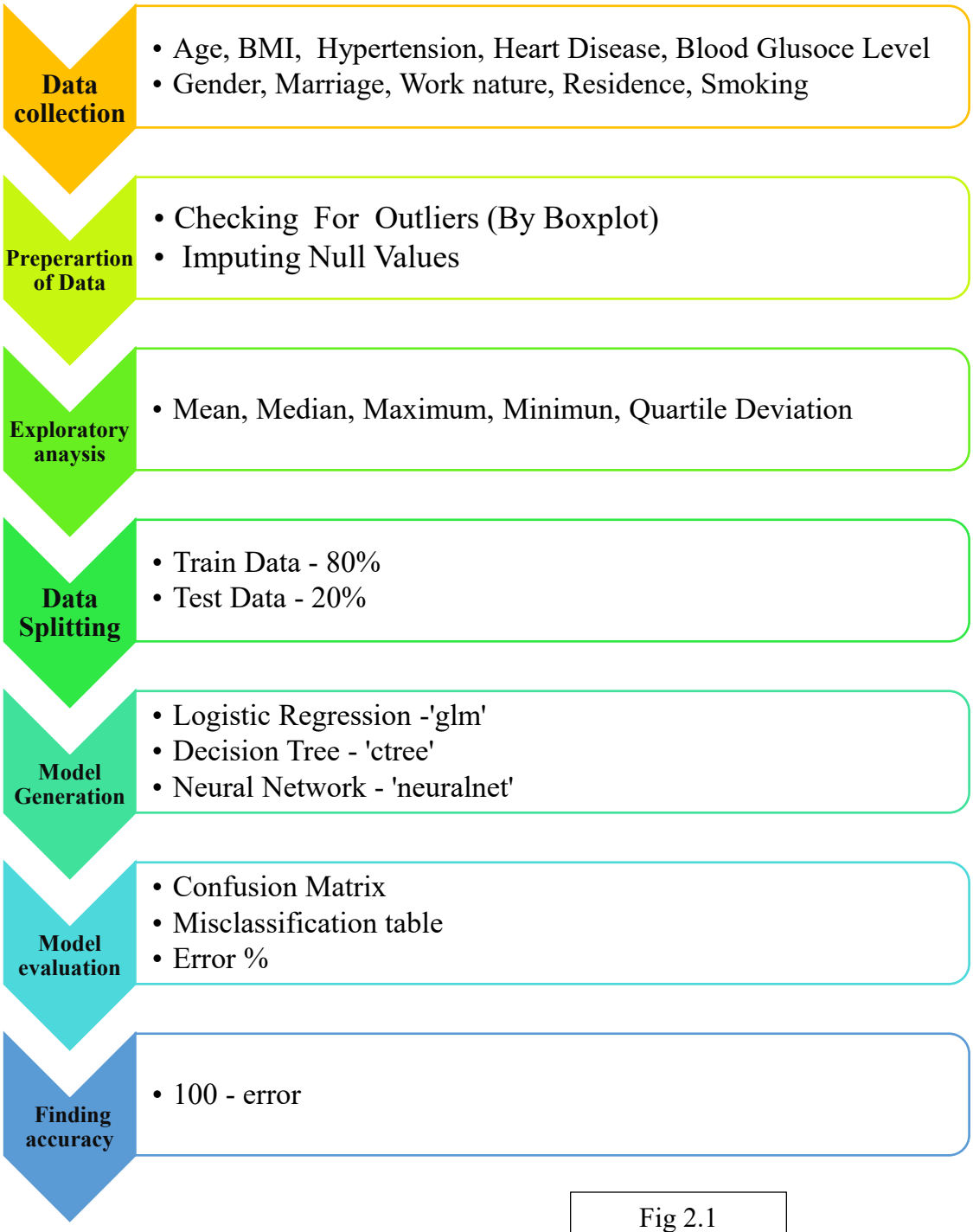


Fig 2.1

2.8.1 Data Collection:

The data is collected through various sources. These Sources include government websites, health organizations, and journals. The present dataset contains 4981 data points with independent variables including Gender, Age, Hypertension, Heart disease, Marriage, Nature

of Work, Residence, Sugar/Glucose Level, BMI, and Smoking. While Stroke is the dependent variable whose occurrence depends on the above factors.

2.8.2 Preparation of data for Analysis:

The data is prepared by checking for the null values. Some of the null values were removed completely while other null values were imputed using the mean.

The dataset is also checked for the presence of outliers if any. The method used for checking the presence of outliers is through the boxplots.

2.8.3 Exploratory Analysis:

A basic statistical analysis of the data was conducted. The statistical analysis gives a good overall view of the data and helps understand the data in a better way. The current analysis included the use of measures of central tendency such as minimum value in the data, maximum value in the data, mean, median, mode, first quartile, mid quartiles, and third quartiles.

2.8.4 Data Splitting:

The whole dataset is split into train data and test data. The splitting is done in the ratio of 80 to 20 i.e., 80% of the data is train data and 20% of the data is test data.

The train data is used for training the model for the prediction.

The test dataset is used for calculating the accuracy of the prediction done by the model.

2.8.5 Model Generation:

The model is generated using the packages and functions.

In the case of logistic regression, the ‘glm’ function is used to generate the logistic model.

In the decision tree initially, the following packages were installed ‘party’, ‘dplyr’, ‘rpart’, ‘party.kit’, ‘rpart.plot’. the function of ‘ctree’ is used for the generation of the decision tree model.

In neural networks the ‘Neuralnet’ library is imported first later on the ‘neuralnet’ function is used for the model generation.

2.8.6 Model Evaluation:

The generated model is evaluated based on the confusion matrix, error rate, and the accuracy of its prediction. Based on these above parameters the model is evaluated.

2.8.7 Finding the Accuracy of the model:

The accuracy of the model is determined based on the error percentage. This error is calculated using the confusion matrix that is generated.

2.9 Sources for Data Collection:

- World Health Organization.
- Health Ministry website.
- Journals based on Stroke and Cerebrovascular diseases.

2.10 Limitations of Study:

While showing great potential, the research of predictive modeling of stroke danger factors utilizing health and lifestyle variables has several drawbacks. For a thorough grasp of the difficulties that academics, healthcare practitioners, and policymakers could run into in this dynamic sector, acknowledging these limits is essential. These restrictions include a range of topics, including data-related difficulties and ethical issues, and they highlight how difficult it is to forecast and reduce the risk of stroke.

1) **Quality and Accessibility of Data:**

Predictive modelling of stroke risk variables has a number of major drawbacks, one of which is the quality and accessibility of data. Historic medical and lifestyle data are a major source of information for predictive models, which they use to find trends and produce precise risk assessments. Healthcare data, however, frequently exhibits variances in data quality, missing values, and incomplete records. Healthcare systems' disparate data-gathering methods can create biases and reduce the precision of predictive models, in addition to making it difficult to integrate data from many sources. To create strong and trustworthy models, it is imperative to address these data-related issues.

2) **Bias of prediction models:**

The possibility for bias in prediction models, which may come from a variety of sources, is another important constraint. Models may unintentionally reinforce racial, gender, socioeconomic status, and other prejudices if the previous information used to train them

contains such biases. The model could overestimate the danger for particular demographic groups, for instance, if such groups have traditionally had less access to healthcare. The careful evaluation of data representation, computational fairness, and constant monitoring to find and correct any biases are necessary for bias mitigation.

3) Interpretability:

The use of mathematical models for prediction in healthcare, especially the investigation of stroke risk factors, presents a difficulty in terms of interpretability. It might be difficult to comprehend the reasoning behind forecasts when using certain sophisticated machine learning algorithms since they work as complicated "black boxes." Predictive models may not be well-accepted by patients and healthcare professionals because of their lack of interpretability. An continuing topic of study and development in the subject is ensuring that models are both accurate and understandable.

4) Ethical issues:

There is a significant restriction due to the ethical issues regarding the usage of private healthcare information. To preserve patient privacy, healthcare data, which frequently contains personally identifiable information, must be subject to strict controls. Although following laws like the Health Insurance Portability and Accountability Act (HIPAA) is essential, finding the right balance between data privacy and the usefulness of predictive models is still difficult. The two most important stages in traversing the ethical terrain of predictive modelling are obtaining the patients' informed permission and putting in place strong security measures.

5) Generalization:

The generalizability of prediction models is another drawback. The generalizability of models developed using data from particular populations or healthcare settings may be poor. Predictive models may not be applicable in all situations due to the variety of patient demographics, medical procedures, and socioeconomic circumstances. In order to increase the generalizability of models, it is crucial to ensure that they are verified across a variety of populations.

6) Changes in Health indicators:

Changing health and lifestyle markers throughout time and temporal dynamics add another level of complexity. Predictive models frequently make the assumption that the connections between risk variables and results don't change over time. However, over longer time periods,

changes in population demographics, lifestyle patterns, or healthcare practises may have an influence on the validity of models. To take into account temporal changes and maintain the continuing precision of forecasts, continuous model validation and recalibration is required.

7) Complexity of lifestyle Indicators:

Predictive modelling is difficult since lifestyle indicators are changing. Diet, exercise, and smoking patterns are examples of lifestyle variables that might alter over time. If there have been recent changes, predictive models based on past lifestyle data could not adequately reflect current behaviours. The requirement for real-time, current lifestyle data becomes clear, and the incorporation of such data into forecasting techniques necessitates sophisticated monitoring systems and data gathering techniques.

CHAPTER 3:

PROFILE OF THE PREDICTIVE MODELS SELECTED FOR THE STUDY

3.1 Correlation:

A statistical metric known as correlation assesses the level of relationship between two variables. It reveals information on the nature and depth of their relationship. It is usual to represent the linear relationship among two continuous variables using the coefficient of Pearson correlation (r).

3.1.1 Equation of Correlation:

The following is the formula to determine the Pearson correlation coefficient:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2\ ]}}$$

Equation 3.1

The correlation coefficient (r) is a number that goes from -1 to 1. A value of 1 denotes a perfect linear relationship (positive or negative), -1 denotes a perfect linear relationship (positive or negative), and 0 denotes no linear relationship.

The number of paired data points in the sample, or n (Number of Data Points).

The total of the product of each pair of corresponding values from the two variables is known as ∑"xy" (total of the Product of Corresponding Values).

∑x (Sum of Variable X): The total of all variable X values found in the dataset.

∑y (Sum of Variable Y): The dataset's total count of the values for the variable Y.

$\sum x^2$ (Sum of Squares of Variable X): The total of variable X's squared values.

$\sum y^2$ (Sum of Squares of Variable Y): The total of variable Y's squared values.

3.1.2 Testing of hypotheses:

To assess if the measured correlation coefficient is substantially different from zero, indicating a meaningful link between the variables, hypothesis testing is frequently employed in correlation analysis. The other possibility (H1) often proposes a considerable connection, while the null hypothesis (H0) frequently asserts no link between the variables.

To calculate a t-statistic for the purpose of hypothesis testing, use the following formula.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Equation 3.2

Here, n is the number of data points and r is the observed correlation coefficient. Degrees of freedom (df) for this test are (n-2), which reflects the estimation of two variables (the slopes of the regression lines) during the calculation of the correlation coefficient.

3.1.3 The Correlation Coefficient's Interpretation:

Positive Correlation (r > 0): A positive correlation means that as one variable rises, the other variable also tends to rise. The strength of the positive association increases as the correlation coefficient approaches 1.

Negative Correlation (r < 0): A negative correlation shows that the tendency is for the other variable to drop as the first increases. The negative correlation is larger the closer the correlation coefficient is near -1.

A correlation value of 0 indicates that there is no association between the variables. It's crucial to keep in mind, though, that the Pearson correlation value may not account for all correlations.

3.1.4 Correlation Strength:

It is common practice to evaluate the connection's strength using the correlation coefficient's absolute value, or r.

- $r \leq 0.3$: Weak or non-existent correlation
- Moderate correlation, $0.3 \leq r < 0.7$
- Strong correlation, $r \geq 0.7$

3.1.5 Types of Correlation:

Positive Correlation: As an illustration, if we examine the connection between the quantity of study time and test performance, we may discover a positive correlation. Exam results often improve as study time increases.

Negative Correlation: Example, the amount of time spent sitting down and physical fitness may have a negative relationship with regard to health. The level of physical fitness tends to decline as sedentary behavior rises.

No Correlation: For instance, if we looked at the relationship between shoe size and academic achievement, we might not find a discernible connection. These factors are probably unrelated.

Correlation with a curvature: The relationship between variables can occasionally take the form of a curve rather than a straight line. Other techniques, such polynomial regression, could be more suitable for finding such associations than the Pearson correlation coefficient.

False Correlation: When two variables seem to be connected but really have a coincidental relationship, this is known as a false correlation. To prevent reading correlations incorrectly, it is crucial to take into account any confounding variables and the context of the data.

Analysis of correlations offers important insights into the connections between variables. Quantifying the strength and direction of linear correlations is the Pearson correlation coefficient. The correlation coefficients' accompanying hypothesis testing aids in determining the statistical significance of observed connections. Understanding the type of correlation helps us understand the nature of the link between the variables under examination. Interpreting the correlation coefficient includes taking into account both its magnitude and direction. Correlation does not always imply causation; hence it is important to be cautious when inferring causal relationships from correlations.

3.2 Logistic Regression:

A common statistical technique for simulating the connection between an outcome variable with one or more distinct variables is logistic regression. When the variable in question is categorical and represents two alternative outcomes, such as "success" and "failure," it is very helpful. Logistic regression is made to manage the probabilistic characteristics of binary results, unlike the use of linear regression, which is ideal for continuous outcomes.

The Logistic Regression Model's Formula: The logistic function, sometimes called the sigmoid function, serves as the foundation for the logistic regression model.

It is defined as: $$P(Y=1)=\frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_n X_n}}$$

Equation 3.3

Terms of Equation:

P(Y=1) is the likelihood that an event, such as success or the fulfilment of a condition, will occur.

The natural logarithm's base is e.

The intercept term is 0.

The coefficients for the independent variables X1, X2,..., and Xn are β1, β2,..., and βn.

When modeling binary outcomes, the logistic function is essential since it limits the predicted probability to values between 0 and 1.

3.2.1 Logistic regression coefficient interpretation:

The logistic regression coefficients (β0, β1, ..., βn) offer useful information about the correlation between the independent variables and the outcome's log chances. Each coefficient's sign shows the effect's direction:

An increase in X causes an increase in the log probabilities of the occurrence Y=1 if βi is positive.

If βi is negative, a rise in Xi causes the log chances of the occurrence Y=1 to drop. The magnitude of the coefficient (i) is a measure of how big the influence is. The odds ratio ($e^{\beta i}$), which shows how a one-unit change in the predictor impacts the likelihood of the event occurring, offers a more understandable measure of effect magnitude.

3.2.2 Considerations and Assumptions for Logistic Regression:

Linearity of Log Odds:

Logistic regression presupposes a linear connection between the variables that are not dependent and the outcome's log chances. Exploratory analysis of data and model diagnostics can verify this assumption.

Independence of Observation: The independence of observations is important since they should not depend on one another. Adjustments could be required if there is dependency, such as with time-series data.

Absence of Multicollinearity: There shouldn't be much correlation between the independent variables. The coefficient estimations may become unstable due to multicollinearity.

Absence of Outliers: Outliers can adversely affect the estimate of coefficients and the performance of models.

Binary Outcome: Logistic regression is intended for binary outcomes. Extensions such as logistic regression and multinomial logistic regression are utilized for outcomes with several categories.

Big Sample Size: With a big sample size, logistic regression works well. A general guideline is to propose a minimum of 10 occurrences for each predictor variable to provide reliable estimates.

3.2.3 Uses for Logistic Regression:

Because it can handle binary outcomes in a variety of ways, logistic regression has applications in many different domains. Examples of typical uses include:

Healthcare: Predicting a patient's chance of having a specific medical disease based on laboratory findings and patient features is known as medical research.

Marketing: Predicting a customer's propensity to buy based on past purchasing patterns and demographic data.

Credit Scoring: The evaluation of a customer's likelihood of loan default based on their financial & and credit history

Epidemiology is the study of the risk factors involved in the development of illnesses or other health consequences.

Social sciences: Examining variables that affect binary outcomes, such as voting patterns, work situations, or level of education.

3.2.4 Extensions and Challenges:

Data Unbalance: Logistic regression can be sensitive to datasets that have one result that is significantly more common than the other. To solve this problem, methods like over- or under sampling might be used.

Including interaction terms can let you model intricate connections between predictor variables.

Non-Linear Relationships: Polynomial regression, logistic regression, or other models that are not linear may be taken into account for non-linear relationships.

Regularisation Methods: To avoid overfitting when there are several predictors, regularisation methods like Lasso or Ridge regression can be used.

As a fundamental tool in the statistical toolbox, logistic regression provides a reliable approach for modelling binary outcomes. Many fields employ it because of its adaptability, interpretability, and simplicity of usage. Practitioners may use logistic regression to acquire important insights into the interactions between variables and make educated predictions in scenarios with binary outcomes by knowing the logistic function, key terms, and coefficient interpretation.

3.3 Decision Tree:

Decision Trees are helpful in both regression and classification applications, decision trees are a potent and adaptable machine learning approach. They are well-liked because of their ease of use, readability, and capacity for both category and numerical data. Recursively dividing the data in subsets according to the values of the input characteristics, decision trees make choices at each node until they reach the node with the leaf that displays the expected result.

3.3.1 Decision Tree's Basic Structure:

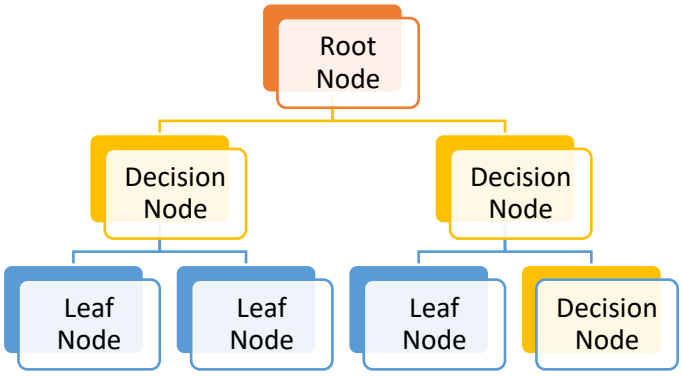


Fig 3.1

There are nodes, subdivisions, and leaves in a decision tree. Nodes stand in for points of decision depending on the values of characteristics, branches for potential outcomes and leaves for the final predictions.

3.3.2 Components of a decision tree:

The topmost node, or root node, represents the feature that divides the data most effectively depending on the selected criterion. It is where the decision-making process begins.

Nodes that come after the root nodes and serve as decision points are known as internal nodes. They divided the data into sets depending on the values of the features.

Branches: The linking edges between nodes that stand in for various decision-related outcomes.

The decision tree leaves (terminal nodes), where the ultimate predictions or categorization are formed. Each leaf is associated with a certain class or value in numbers.

3.3.3 Algorithms for Decision Trees:

The most prominent decision tree-building algorithms are ID3, C4.5, CART (Classification and Regression Trees), and Random Forests. These algorithms use several techniques to choose the optimum splits and build decision trees.

3.3.4 Building and Pruning Decision Trees:

Top-down construction is the method used to build decision trees. The algorithm chooses the most effective feature to divide the data into subsets, builds child nodes, and repeats the process for each subset until a stopping requirement is satisfied. Stopping criterion might be a maximum depth, a minimal sample count per leaf, or a minimal threshold for information gain.

In order to reduce overfitting, pruning is a post-processing procedure. It entails trimming branches from validation data that have limited predictive value. Pruning keeps the tree from becoming overly reliant on the training data, enhancing its ability to generalise to new data.

Handling Numerical and Categorical data:

By dividing categorical data into distinct categories, decision-making structures naturally handle categorical data. They choose the best thresholds to produce binary splits for numerical data. To account for numerical aspects, different algorithms employ a variety of strategies, such as binaries splits or multi-way splits.

3.3.5 Decision trees provide the following benefits:

Easy interpretation, which enables consumers to grasp the decision-making process intuitively.

Decision trees are resistant to abnormalities since they don't make any assumptions about the distribution of the data.

Decision trees can naturally handle both categorical and numerical data without the requirement for preprocessing.

Automatic Feature Selection: The tree may automatically cut off features that don't provide much prediction value.

Decision trees are able to simulate non-linear connections between features and the goal variable.

3.3.6 Decision trees' drawbacks:

Decision tree models are vulnerable to overfitting, which might capture noise within the training set of data. To solve this problem, pruning and establishing suitable hyperparameters are essential.

Stability: Minor modifications to the data may result in alternative tree architectures and have an impact on stability.

Biased Towards Dominant Classes: Decision trees frequently exhibit bias when predicting the dominant class in unbalanced datasets.

Top-down, greedy approaches may or may not produce the globally optimum tree.

3.3.7 Applications of Decision Tree:

Applications for decision trees include categorization tasks like identifying spam, credit scoring, or medical diagnosis.

Decision trees are appropriate for jobs like predicting property prices because they can forecast numerical results.

Ranking the value of features using decision trees can help with feature selection in complicated datasets.

Decision trees are helpful for anomaly detection because they can spot odd patterns in data.

Decision trees are useful tools for a variety of applications because they provide an understandable and straightforward approach to machine learning. In order to use decision trees efficiently, it is essential to comprehend the splitting criteria, building procedure, and potential difficulties. Although they have drawbacks, pruning strategies and ensemble approaches like Random Forests can improve their performance. The popularity of decision trees in machine learning is largely due to their versatility in handling both numerical data and categorical data, automated feature selection, and interpretability.

3.4 Neural Networks:

In the fields of artificial intelligence and machine learning, neural networks serve as a fundamental idea. These computational models, which draw their inspiration from the structure and operation of the human brain, are skilled in extracting intricate patterns and interpretations from data. Due to its capacity to handle a wide range of tasks, including audio and picture recognition, natural language processing, and more, neural networks have become increasingly popular. It is crucial to examine neural networks' design, parts, and underlying mathematical formulas in order to comprehend their complexities.

3.4.1 Architecture:

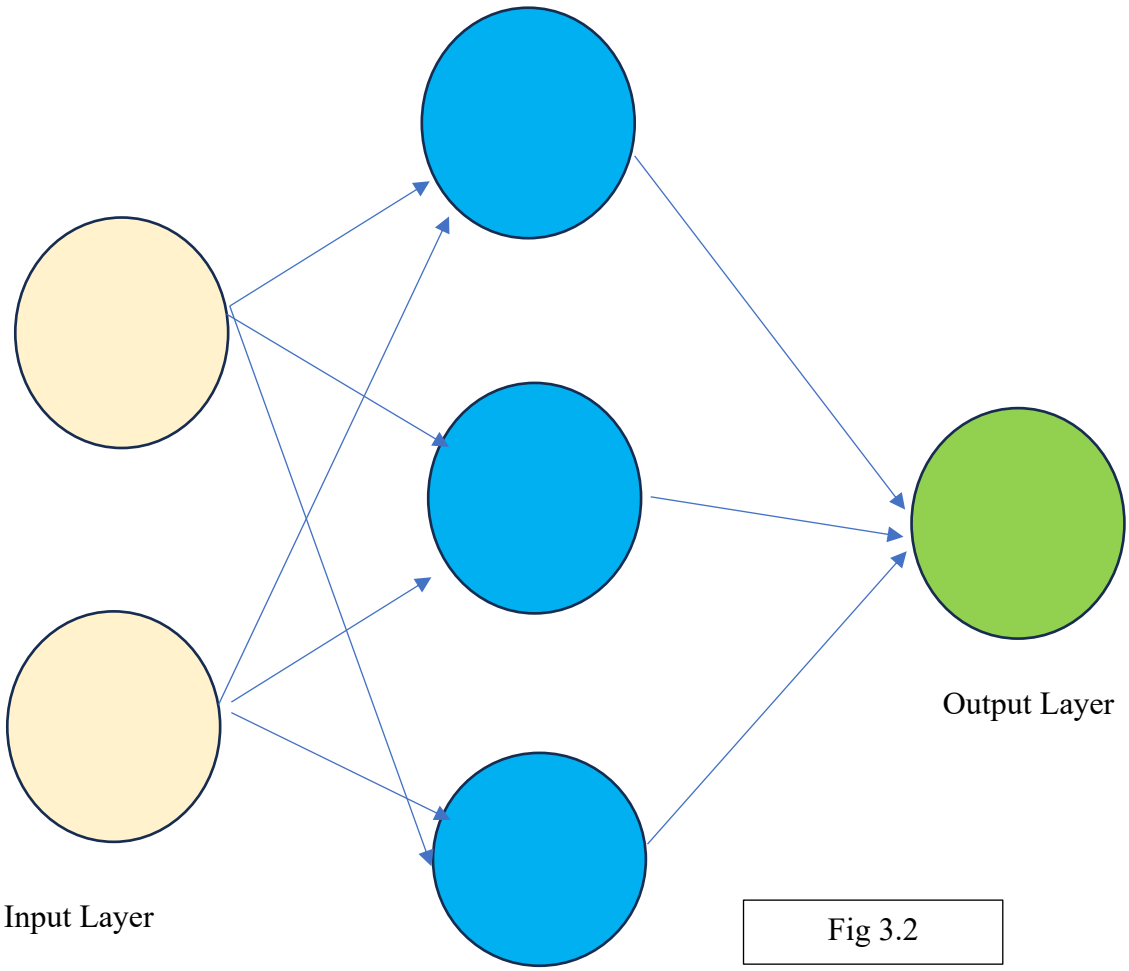


Fig 3.2

A linked layer of nodes, also known as neurons or perceptron’s, make up a neural network's basic structure. There are three different sorts of these layers:

Nodes that reflect the model's input features make up the input layer. Each node in the input layer correlates to a characteristic in the information set, and the total number of nodes depends on how dimensional the input data are.

Hidden Layers: Neural networks may include one or more hidden layers in between the layers of input and output. Every node in a layer that is hidden is linked to every other node in the layer above and below it. Hyperparameters that affect the network's ability to recognise complicated relationships in the data include the number of layers that are concealed and the number of nodes in each layer.

The output layer generates the predictions or classifications made by the model. Based on the kind of task—binary categorization, multi-class classification, regression, etc.—determines the number of nodes in the output layer.

3.4.2 Multilayer Perceptron model of Neural Network:

The neuron or perceptron is the basic building block of a neural network. The manner that each neuron in the human brain functions is an inspiration for how it operates. A neuron takes in incoming signals, interprets them, and then generates an output. The weighted sum of a perceptron's inputs, through an activation function, determines the output y mathematically:

$$y = f(\sum_{i=1}^{n} w_i\, x_i + b)$$

Equation 3.4

- xi is the input feature
- wi stands for the weights connected to each input,
- b is the bias term
- f is the activation function.

The network's capacity to simulate complicated interactions is greatly influenced by the selection of the activation function. The sigmoid, tangent hyperbolic (tanh), rectified linear unit (ReLU), and softmax are typical activation functions.

3.4.3 Neural network training:

Finding the best combination of weights and biases to train a neural network entail minimising the loss function. The essential steps involved in this procedure are as follows:

1) Initialization: Randomly initialise the biases and weights. For the model to avoid becoming stuck in local minima, proper initialization is essential.
2) Pass the input across the network to get predictions via forward propagation.
3) Calculate Loss: Determine the difference in value between the forecasts and the actual values.
4) Using backpropagation, you may calculate gradients and adjust weights and biases.
5) Up until the model converges, repeat: Iterate through steps 2-4 for several epochs.

3.4.4 Neural network types:

Feedforward Neural Networks (FNN): Information goes from the input layer to the output layer in a single direction in a neural network called a feedforward neural network (FNN).

Convolutional Neural Networks (CNN): Created for image processing, CNNs employ convolutional layers to record spatial feature hierarchies.

Recurrent neural networks (RNNs): Designed to handle sequential data, RNNs include connections that create cycles, enabling them to keep track of prior inputs.

Long Short-Term Memory (LSTM): A kind of RNN with specialized memory cells that lessen the issue of disappearing gradients and improve the capacity to detect long-term dependence.

Generative Adversarial Networks (GAN): The creation of realistic data is made possible by generative adversarial networks (GAN), which are made up of a discriminator network and a generator network that have undergone adversarial training.

3.4.5 Disadvantages or Challenges to Neural Networks:

Vanishing Gradient Problem: When backpropagating in deep networks, gradients may become extremely tiny, making it difficult to update the weights. Techniques like batch normalisation and the usage of certain activation functions (like ReLU) take care of this problem.

Deep neural network training may be computationally intensive and requires strong hardware, such as GPUs or TPUs.

Interpretability: Deep architecture neural networks, in particular, are frequently regarded as "black box" models, making it difficult to understand how they make decisions.

Neural networks sometimes need vast volumes of labelled data for training, which limits their application in situations with fewer data.

3.4.6 Applications of Neural Networks:

CNNs are excellent at applications like object identification, facial recognition, and picture categorization.

Text creation, sentiment analysis, and language translation are all common applications of recurrent neural networks and transformers in natural language processing (NLP).

voice Recognition: Systems for voice recognition use neural networks, particularly recurrent designs.

Neural networks help in illness diagnosis, disease prediction, and medical picture analysis.

Neural networks are essential to the development of autonomous cars because they make tasks like object identification and navigation possible.

In machine learning, neural networks offer a breakthrough paradigm that enables computers to learn and make judgements that are similar to human cognition. The levels of complexity of neural networks present both difficulties and previously unheard-of possibilities, ranging from the basic architecture to the sophisticated mathematics underlying forward and backward propagation. Addressing issues with interpretability, computational effectiveness, and data needs is still essential as the area develops. With its capacity to identify patterns in enormous datasets, neural networks have changed the face of artificial intelligence by powering systems for everything from voice and picture recognition to autonomous systems and medical diagnostics. We are getting closer to creating intelligent systems that can learn and adapt thanks to the ongoing development of neural network topologies and approaches.

CHAPTER 4:

DATA ANALYSIS AND INTERPRETATION

4.1 Correlation:

Indicators of the degree and direction of a linear relationship among two variables include correlation coefficients. It can be anything between -1 and 1, where -1 denotes a flawless negative correlation, 1 indicates an ideal positive correlation, and 0 denotes no connection at all.

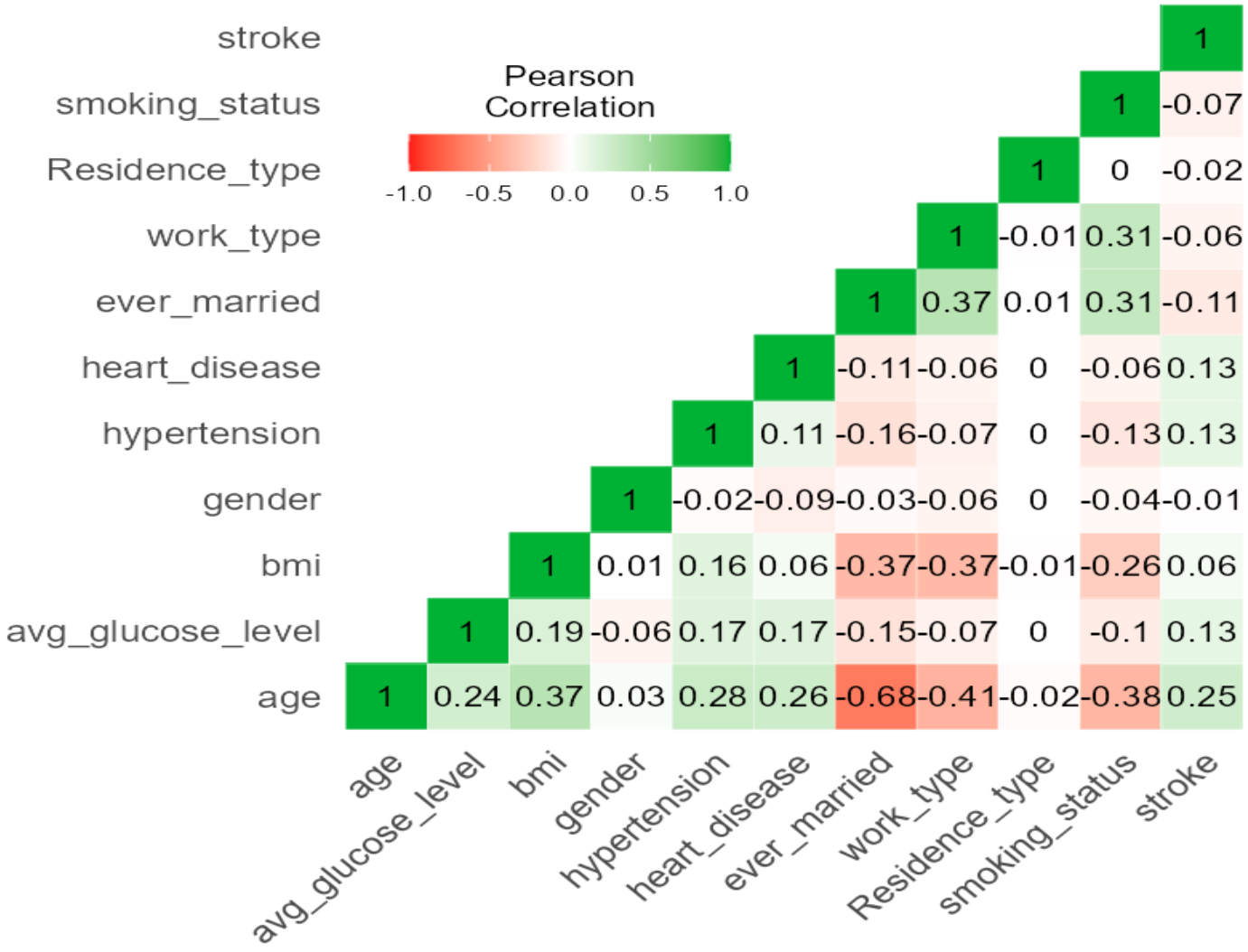


Fig 4.1

4.1.1 Output Analysis:

Age, smoking status, cardiac disease, average blood glucose level, hypertension, and BMI are all positively connected with stroke risk, according to the correlation matrix.

Accordingly, those who have greater values for these factors are at an increased risk of suffering a stroke.

Additionally, the correlation matrix demonstrates that the variables "ever married," "residence type," and "job type" have a negative connection with stroke risk.

Higher scores for these factors indicate a decreased risk of stroke in individuals.

Stroke and Age:

Age and stroke have a correlation value of 0.25. This indicates that the two variables have a somewhat positive association. This is most likely brought on by the fact that ageing raises the risk of stroke.

Stroke and Smoking status:

The link between smoking status and stroke is 0.37. This indicates that the two variables have a somewhat positive association. Smoking destroys blood vessels and raises the risk of stroke; therefore, this is probably why.

Smoke and Heart disease:

The connection between heart disease and stroke is 0.28. This indicates that the two variables have a somewhat positive association. This is probably because persons with heart problems are more prone to get blood clots in their veins, which can result in strokes.

Stroke and Blood sugar level:

The average glucose level and the connection between stroke and that number is 0.19. This indicates that the two variables have a sluggishly positive connection. High levels of glucose in the blood can damage the blood vessels and raise the risk of stroke, which is probably why this is the case.

Stroke and hypertension:

The connection between hypertension and stroke is 0.26. This indicates that the two variables have a somewhat positive association. This is most likely caused by hypertension, which damages the blood vessels and raises the possibility of stroke.

Stroke and BMI:

The stroke and BMI correlation value is 0.26. This indicates that the two variables have a somewhat positive association. This is probably because individuals with higher BMIs are more prone to acquire elevated cholesterol levels, high blood pressure, and diabetes, all of that constitute stroke risk factors.

Stroke and ever marriage:

Stroke and marriage history have a -0.37 association value. This indicates that the two variables have a somewhat negative association. This is probably because married people often experience better health results than single persons.

Stroke and residence type:

The stroke-residence type correlation coefficient is -0.11. This indicates that the two variables have a slender negative connection. This is probably because people who reside in rural locations frequently have less accessibility to healthcare and can be more inclined to participate in unhealthy habits like smoking and overeating.

Stroke and work type:

Stroke and job type have a correlation value of -0.06. This indicates that the two variables have a very little negative association. This is probably because those who work full-time might not have as much time for exercise and good eating.

Numerous factors are connected with the risk of stroke, as the correlation matrix demonstrates. Age, history of smoking, heart disease, blood pressure, glucose levels on average, BMI, previous marriages, kind of habitation, and type of employment are some of these factors.

Noting that correlation does not imply causation is crucial. A correlation between two variables does not always imply a causal relationship. The correlation matrix does, however, offer crucial details regarding the connections between various variables.

People who are at increased risk for stroke can make efforts to minimize their risk by adopting healthier choices, such as stopping smoking, eating a balanced diet, exercising on a regular basis, and keeping a healthy weight.

4.2 Correlation Heatmap between stroke and Health factors:

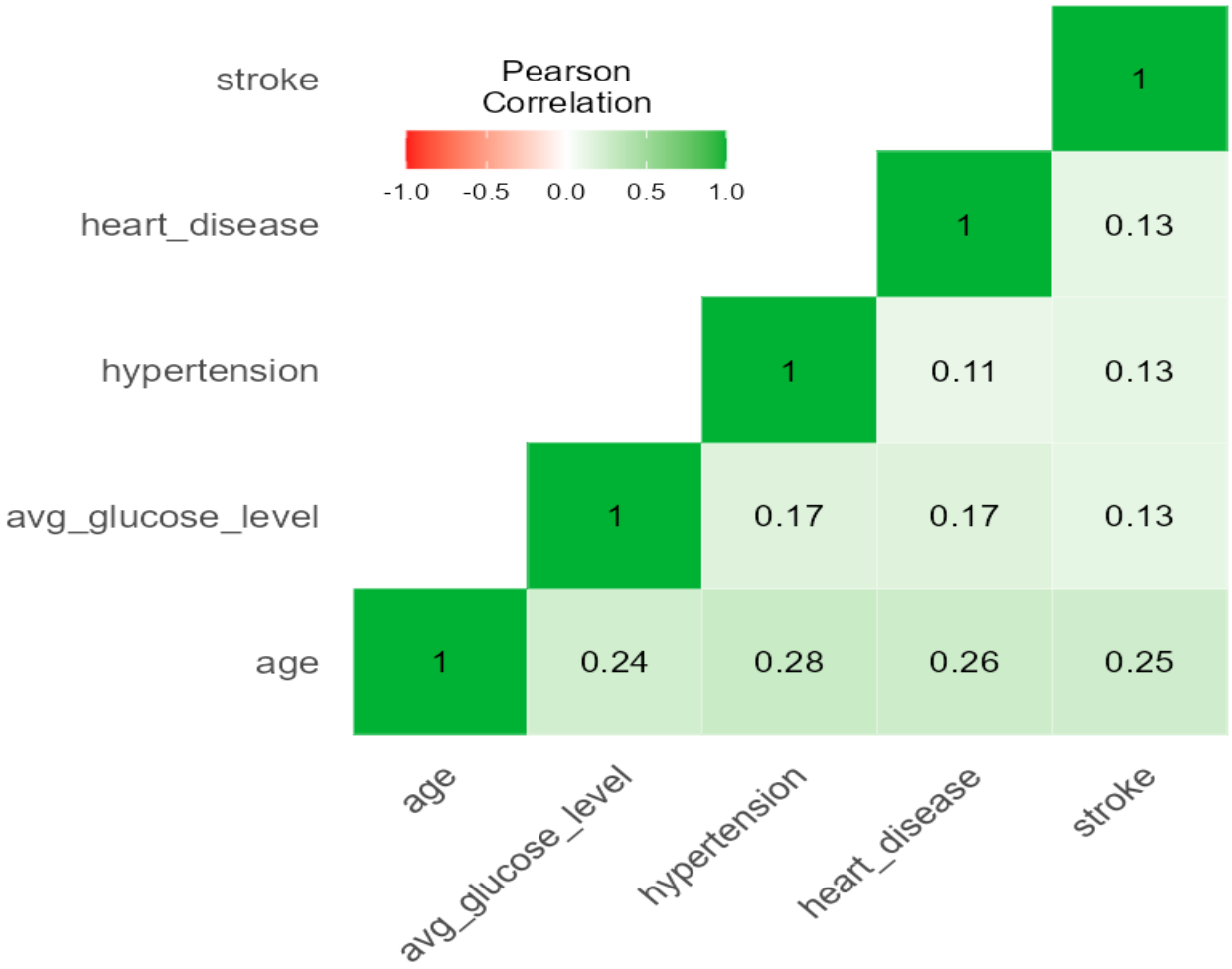


Fig 4.2

The above correlation heatmap shows the relation between stroke and health factors. It can be seen from the heatmap that all the factors relating to health such as age, glucose level, hypertension, and heart disease are positively correlated with the stroke. This indicates that these factors contribute more positively to the risk of stroke.

4.3 Logistic Regression Code:

#Importing the dataset

dat1<-data

#Exploratory Analysis

```
summary(dat1)

View(dat1)

str(dat1)

unique(dat1$State)

#Checking for null values

sum(is.na(dat1))

#Splitting the data into test and train

set.seed(100)

datsamp<-sample(1:nrow(dat1), size = 0.8*nrow(dat1))

train<-dat1[datsamp,]

test<-dat1[-datsamp,]

str(train)

View(train)

#Building Model

form=stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Residence
_type+avg_glucose_level+bmi+smoking_status

mod1<-glm(formula = form, data = train, family = "binomial")

summary(mod1)

#Generating the Predicted output of the model based on training data

x<-predict(mod1)

#Viewing the top 10 values

head(x)

head(train$stroke)

#To find the probability

xnew<-ifelse(x>0.5,1,0)
```

#Creating Confision matrix

```
cm<-table(xnew, train$stroke)
cm
```

#Calculating the error

```
err<-1-sum(diag(cm))/sum(cm)
err
```

#testing the data

```
mod2<-glm(formula = form, data=test, family = "binomial")
summary(mod2)
```

#Getting the Predicted values

```
xtest<-predict(mod2)
```

#Changing to discrete values

```
xtestnew<-ifelse(xtest>0.5,1,0)
str(xtestnew)
```

 # Confusion Maxtrix

```
cmt<-table(xtestnew, test$stroke)
cmt
```

 # Finding the Error

```
errt<-1-sum(diag(cmt))/sum(cmt)
errt
```

4.3.1 Output Analysis and Interpretation:

```
R 4.3.1 · ~/
> #Importing the dataset
> dat1<-Stroke.Data.Coded
>
> #Exploratory analysis
> summary(dat1)
     gender           age          hypertension     heart_disease
 Min.   :1.000   Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
 1st Qu.:1.000   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
 Median :2.000   Median :45.00   Median :0.00000   Median :0.00000
 Mean   :1.584   Mean   :43.42   Mean   :0.09617   Mean   :0.05521
 3rd Qu.:2.000   3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :2.000   Max.   :82.00   Max.   :1.00000   Max.   :1.00000
  ever_married     work_type     Residence_type  avg_glucose_level
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 55.12
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 77.23
 Median :1.000   Median :1.000   Median :1.000   Median : 91.85
 Mean   :1.341   Mean   :1.825   Mean   :1.492   Mean   :105.94
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:113.86
 Max.   :2.000   Max.   :4.000   Max.   :2.000   Max.   :271.74
      bmi        smoking_status      stroke
 Min.   :14.0   Min.   :1.000   Min.   :0.00000
 1st Qu.:23.7   1st Qu.:2.000   1st Qu.:0.00000
 Median :28.1   Median :2.000   Median :0.00000
 Mean   :28.5   Mean   :2.584   Mean   :0.04979
 3rd Qu.:32.6   3rd Qu.:4.000   3rd Qu.:0.00000
 Max.   :48.9   Max.   :4.000   Max.   :1.00000
> View(dat1)
```

Logistic regression is used to predict the outcome based on the probability. The output of logistic regression is in the form of binary i.e., either 0 or 1.

Initially, we import the dataset into the R environment. An exploratory analysis is conducted. The summary statistics help us understand the data using the measures of central tendency. The average age of the person in the data is around 43 years. The maximum age of the respondent was 82 years while the minimum age was 20. The median age of the population is 45 years, i.e., 50% of the respondents are below 45 years of age and 50% of the respondents are above the age of 45.

The average glucose level of the respondents was 105.94. showing most of the respondents were on the prediabetic stage. The median was found to be 91.85. The maximum value reported was 271.74 while the minimum value reported was 55.12.

BMI stands for body mass index. It is the ratio calculated by the height and weight of the person. From the dataset, we can see that the average BMI of the respondents was 28.5

Indicating most respondents were either overweight or obese. The median value of the BMI is 28.1. The least BMI was 14 while the maximum value was 48.9. bmi: A person's average BMI in the sample is 28.5.

Gender: Males make up 58% of the population in the dataset.

ever_married: Sixty-six percent of the individuals in the dataset are married.

Work_type: The bulk of the dataset's participants (60%) are full-time employees.

smoking_status: 74% of the participants in the sample do not smoke.

hypertension: Ninety-one percent of the population in the sample does not have hypertension.

heart disease: Heart disease is not present in 95% of the population in the sample.

```
> #Splitting the dataset into train and test
> set.seed(100)
> datsamp<-sample(1:nrow(dat1), size = 0.8*nrow(dat1))
> train<-dat1[datsamp,]
> test<-dat1[-datsamp,]
> str(train)
'data.frame':   3984 obs. of  11 variables:
 $ gender           : int  1 2 2 1 2 2 2 2 2 2 ...
 $ age              : num  79 1.8 69 31 45 44 63 30 53 65 ...
 $ hypertension     : int  0 0 0 0 0 0 1 0 0 0 ...
 $ heart_disease    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ever_married     : int  2 2 1 1 1 1 1 1 1 1 ...
 $ work_type        : int  2 4 2 2 3 1 2 3 2 1 ...
 $ Residence_type   : int  1 1 1 2 1 2 1 1 2 1 ...
 $ avg_glucose_level: num  128.7 58.3 79.7 61.1 80 ...
 $ bmi              : num  31 16.5 25 26.5 41.4 21.3 37.7 34 34.3 19.8 ...
 $ smoking_status   : int  4 4 2 2 2 2 2 2 1 1 ...
 $ stroke           : int  0 0 0 0 0 0 0 0 0 0 ...
> View(train)
>
> #Building Model
> form<-stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Resi
dence_type+avg_glucose_level+bmi+smoking_status
> mod1<-glm(formula = form, data = train, family = "binomial")
> summary(mod1)
```

A total of 4981 observations were collected under 11 variables. These variables were a combination of lifestyle factors and also the health indicators. Some of these were categorical in nature while the variables of bmi, avg glucose level and age were continuous. The data was later on checked for null values using the is.na function. The sum(is.na) is found to be 0 indicating that the dataset was complete without the presence of any missing or null values.

The dataset is later divided into test and train data. The train data consists of 80% of the data and is used to generate and train the logistic regression model. The test data contains 20% of

the whole dataset. This is used to test the efficiency and accuracy of the generated model to predict the outcome.

The Logistic regression model is generated using the glm function. The dependent variable for the model is the occurrence of stroke while rest of the variables are independent.

```
Call:
glm(formula = form, family = "binomial", data = train)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.788221   0.810497  -9.609  < 2e-16 ***
age                0.070471   0.005807  12.135  < 2e-16 ***
gender            -0.106396   0.154743  -0.688  0.49173
hypertension       0.320904   0.183772   1.746  0.08077 .
heart_disease      0.191843   0.215346   0.891  0.37300
ever_married       0.210103   0.246187   0.853  0.39342
work_type         -0.044819   0.104588  -0.429  0.66826
Residence_type    -0.034344   0.152562  -0.225  0.82189
avg_glucose_level  0.003523   0.001332   2.645  0.00816 **
bmi                0.012128   0.013678   0.887  0.37527
smoking_status     0.020148   0.072673   0.277  0.78159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1598.2  on 3983  degrees of freedom
Residual deviance: 1286.4  on 3973  degrees of freedom
AIC: 1308.4

Number of Fisher Scoring iterations: 7
```

```
> #Generating the Predicted output of model based on training data
> x<-predict(mod1)
> #Viewing the top 10 values
> head(x)
     3786       503      3430      3696      4090      3052
-1.121161 -7.181630 -2.428130 -5.081306 -3.964369 -4.201575
> head(train$stroke)
[1] 0 0 0 0 0 0
> #To find the probability
> xnew<-ifelse(x>0.5,1,0)
> #Creating Confusion matrix
> cm<-table(xnew,train$stroke)
> cm

xnew    0    1
   0 3782  202
> err<-1-sum(diag(cm))/sum(cm)
> err
[1] 0.05070281
> #testing the data
> mod2<-glm(formula = form, data=test,family = "binomial")
> summary(mod2)
```

The Confusion matrix also known as misclassification table has correctly classified 3782 observations with an error percentage of only 5%. This makes the training model efficient enough to be tested and predicted with the test data.

```
Call:
glm(formula = form, family = "binomial", data = test)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.074355   1.712541  -4.131 3.61e-05 ***
age                0.069723   0.013356   5.220 1.79e-07 ***
gender             0.246146   0.344831   0.714   0.4753
hypertension       0.733846   0.382159   1.920   0.0548 .
heart_disease      0.785267   0.413759   1.898   0.0577 .
ever_married       0.131241   0.508644   0.258   0.7964
work_type         -0.374538   0.245341  -1.527   0.1269
Residence_type    -0.403402   0.335756  -1.201   0.2296
avg_glucose_level  0.005267   0.002974   1.771   0.0765 .
bmi               -0.012634   0.030613  -0.413   0.6798
smoking_status     0.027567   0.167492   0.165   0.8693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 372.85  on 996  degrees of freedom
Residual deviance: 279.70  on 986  degrees of freedom
AIC: 301.7

Number of Fisher Scoring iterations: 8
```

```
> #Confusion Matrix
> cmt<-table(xtestnew, test$stroke)
> cmt

xtestnew   0   1
       0 943  54
> #Error
> errt<-1-sum(diag(cmt))/sum(cmt)
> errt
[1] 0.05416249
```

Upon testing with the test data, the new confusion matrix is obtained. The new misclassification table shows that 943 values have been correctly classified by the model with an error percentage of 5.4%. This score shows Logistic regression model is efficient enough to be used to predict the outcome or occurrence of stroke based on these input variables depicting the health and lifestyle factors.

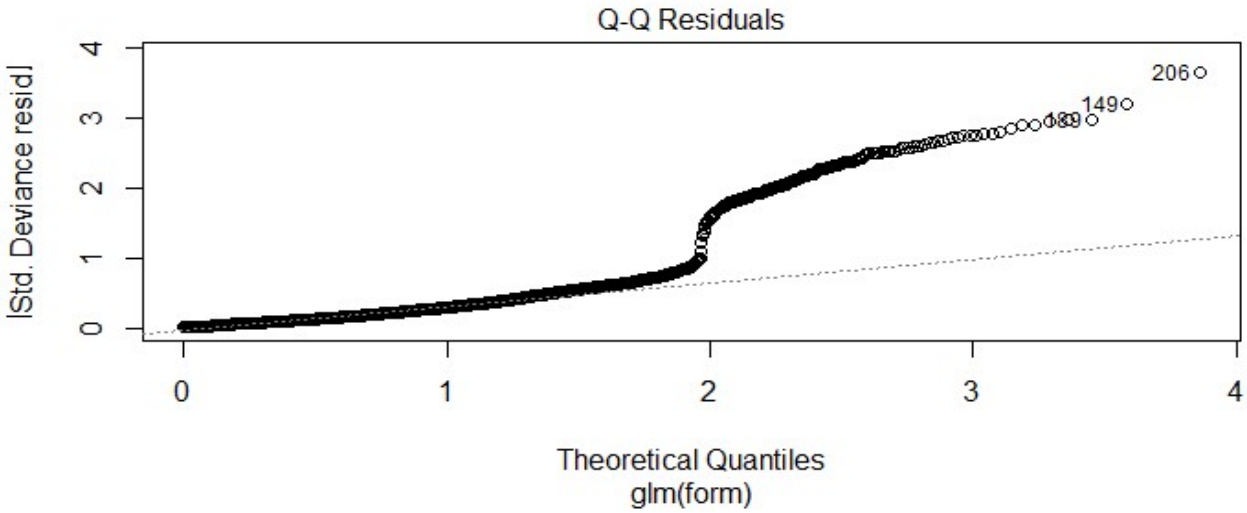


Fig 4.3

The graph demonstrates that as the total amount of hypothetical quantities rises, so does the projected likelihood of stroke. So, those who have a higher concentration of these contributing factors were more likely to experience a stroke. The graphic also demonstrates a nonlinear link between the theoretical values and the anticipated stroke probability. This indicates that the risk of stroke does not grow steadily with the presence of new risk factors. For instance, the risk of stroke may increase more from possessing two risk factors to acquiring three risk factors than from having only one risk factor to having two risk factors. Overall, the logistic regression plot indicates that a person's risk of stroke increases with the number of risk variables they have. For each added risk factor, the risk of stroke does not always rise linearly. It's crucial to keep in mind that the aforementioned logistical regression plot is only a straightforward model. It does not account for all the elements that may raise a person's risk of having a stroke. It can, however, give a broad notion of how these variables may affect stroke risk.

4.4 Decision Tree Code:

#Importing the dataset

```
dat1<-data
```

#Exploratory Analysis

```
summary(dat1)
```

```
View(dat1)

str(dat1)

unique(dat1$State)
```

#Checking For Null Values

```
sum(is.na(dat1))
```

#Splitting the data into Train and Test

```
set.seed(100)

datsamp<-sample(1:nrow(dat1), size = 0.8*nrow(dat1))

train<-dat1[datsamp,]

test<-dat1[-datsamp,]

str(train)

View(train)
```

#Installing the packages

```
install.packages("party")

library(party)

install.packages("party.kit")

library(partykit)

install.packages("dplyr")

library(dplyr)

install.packages("rpart")

library(rpart)

install.packages("rpart.plot")

library(rpart.plot)
```

#Building Model

```
form=stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Residence
_type+avg_glucose_level+bmi+smoking_status

mod1<-ctree(form, data = train)

summary(mod1)

plot(mod1)

#Fancy plot

fp<-rpart(form,data = train, method = 'class')

rpart.plot(fp, extra = 5)

#Generating the Predicted output of model based on training data

x<-predict(mod1)

#Viewing the top 10 values

head(x)

head(train$stroke)

#To find the probability

xnew<-ifelse(x>0.5,1,0)

#Creating Confusion matrix

cm<-table(xnew,train$stroke)

cm

#Error

err<-1-sum(diag(cm))/sum(cm)

err

#testing the data

mod2<-glm(formula = form, data=test,family = "binomial")

summary(mod2)

plot(mod2)
```

```
#Fancy plot

fp2<-rpart(form,data = test, method = 'class')

rpart.plot(fp2, extra = 5)

#Getting the Predicted values

xtest<-predict(mod2)

#Changing to discrete values

xtestnew<-ifelse(xtest>0.5,1,0)

str(xtestnew)

#Confusion Matrix

cmt<-table(xtestnew, test$stroke)

#Error Calculation

errt<-1-sum(diag(cmt))/sum(cmt)

errt
```

4.4.1 Output Analysis:

Decision trees are among the prominent machine learning algorithms used for both regression as well as classification. The decision trees are extremely helpful in cases where the person needs to decide at every next step considering various number of possibilities.

Initially, we imported the dataset into the R environment. The dataset is checked for null values and is imputed if found. The dataset is prepared for the analysis. The summary statistics used to conduct the exploratory analysis. This method uses the measures of central tendency such as mean, quartile deviations, mode, median etc., to understand and analyze the dataset.

```
> #Importing the data
> dat1<-Stroke.Data.Coded
> #Exploratory Analysis
> summary(dat1)
     gender           age          hypertension     heart_disease
 Min.   :1.000   Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
 1st Qu.:1.000   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
 Median :2.000   Median :45.00   Median :0.00000   Median :0.00000
 Mean   :1.584   Mean   :43.42   Mean   :0.09617   Mean   :0.05521
 3rd Qu.:2.000   3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :2.000   Max.   :82.00   Max.   :1.00000   Max.   :1.00000
  ever_married     work_type     Residence_type  avg_glucose_level
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 55.12
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 77.23
 Median :1.000   Median :1.000   Median :1.000   Median : 91.85
 Mean   :1.341   Mean   :1.825   Mean   :1.492   Mean   :105.94
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:113.86
 Max.   :2.000   Max.   :4.000   Max.   :2.000   Max.   :271.74
      bmi        smoking_status      stroke
 Min.   :14.0   Min.   :1.000   Min.   :0.00000
 1st Qu.:23.7   1st Qu.:2.000   1st Qu.:0.00000
 Median :28.1   Median :2.000   Median :0.00000
 Mean   :28.5   Mean   :2.584   Mean   :0.04979
 3rd Qu.:32.6   3rd Qu.:4.000   3rd Qu.:0.00000
 Max.   :48.9   Max.   :4.000   Max.   :1.00000
> View(dat1)
```

```
> str(dat1)
'data.frame':   4981 obs. of  11 variables:
 $ gender           : int  1 1 2 2 1 1 2 2 2 2 ...
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : int  0 0 0 1 0 1 0 0 1 0 ...
 $ heart_disease    : int  1 1 0 0 0 1 0 0 0 1 ...
 $ ever_married     : int  1 1 1 1 1 1 2 1 1 1 ...
 $ work_type        : int  1 1 1 2 1 1 1 1 1 3 ...
 $ Residence_type   : int  1 2 1 2 1 2 1 1 2 2 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status   : int  1 2 3 2 1 2 2 4 2 3 ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
> #Checking for null values
> sum(is.na(dat1))
[1] 0
```

```
> #Splitting the data into train and test
> set.seed(100)
> datsamp<-sample(1:nrow(dat1), size = 0.8*nrow(dat1))
> train<-dat1[datsamp,]
> test<-dat1[-datsamp,]
> str(train)
'data.frame':   3984 obs. of  11 variables:
 $ gender           : int  1 2 2 1 2 2 2 2 2 2 ...
 $ age              : num  79 1.8 69 31 45 44 63 30 53 65 ...
 $ hypertension     : int  0 0 0 0 0 0 1 0 0 0 ...
 $ heart_disease    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ever_married     : int  2 2 1 1 1 1 1 1 1 1 ...
 $ work_type        : int  2 4 2 2 3 1 2 3 2 1 ...
 $ Residence_type   : int  1 1 1 2 1 2 1 1 2 1 ...
 $ avg_glucose_level: num  128.7 58.3 79.7 61.1 80 ...
 $ bmi              : num  31 16.5 25 26.5 41.4 21.3 37.7 34 34.3 19.8 ...
 $ smoking_status   : int  4 4 2 2 2 2 2 2 1 1 ...
 $ stroke           : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
> #Building Model
> form<-stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Resi
dence_type+avg_glucose_level+bmi+smoking_status
> mod1<-ctree(form, data = train)
> summary(mod1)
  Length Class      Mode
1 7      constparty list
2 5      constparty list
3 1      constparty list
4 3      constparty list
5 1      constparty list
6 1      constparty list
7 1      constparty list
> plot(mod1)
> #Fancy plot
> fp<-rpart(form,data = train, method = 'class')
> rpart.plot(fp, extra = 5)
> #Generating the Predicted output of model based on training data
> x<-predict(mod1)
> #Viewing the top 10 values
> head(x)
        7         3         7         3         3         3
0.1741742 0.0046490 0.1741742 0.0046490 0.0046490 0.0046490
> head(train$stroke)
[1] 0 0 0 0 0 0

> #To find the probability
> xnew<-ifelse(x>0.5,1,0)
> #Creating Confusion matrix
> cm<-table(xnew,train$stroke)
> cm

xnew    0    1
   0 3790  194
> #Error
> err<-1-sum(diag(cm))/sum(cm)
> err
[1] 0.04869478
```

The misclassification table or the confusion matrix shows that the model generated using the train dataset was able to classify 3790 entries. This made the model perform with the efficiency of 95.2%. This make model ready o be used to predict the outcome based on the testdata.

```
> #testing the data
> mod2<-ctree(formula = form, data=test)
> summary(mod2)
  Length Class      Mode
1 5      constparty list
2 3      constparty list
3 1      constparty list
4 1      constparty list
5 1      constparty list
> plot(mod2)
> #Fancy plot
> fp2<-rpart(form,data = test, method = 'class')
> rpart.plot(fp2, extra = 5)
> #Getting the Predicted values
> xtest<-predict(mod2)
> #Changing to discrete values
> xtestnew<-ifelse(xtest>0.5,1,0)
> str(xtestnew)
 Named num [1:997] 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "names")= chr [1:997] "5" "5" "5" "5" ...
```

```
> #Getting the Predicted values
> xtest<-predict(mod2)
> #Changing to discrete values
> xtestnew<-ifelse(xtest>0.5,1,0)
> str(xtestnew)
 Named num [1:997] 0 0 0 0 0 0 0 0 0 0 ...
 - attr(*, "names")= chr [1:997] "15" "22" "24" "27" ...
> cmt<-table(xtestnew, test$stroke)
> errt<-1-sum(diag(cmt))/sum(cmt)
> errt
[1] 0.04613842
```

Upon predicting the model using the test dataset it can be found that the error percentage is only 4.6%. This indicates that the decision tree model is highly efficient enough to predict the outcome or possibility of occurrence of stroke pertaining to the person's demographic and health factors.

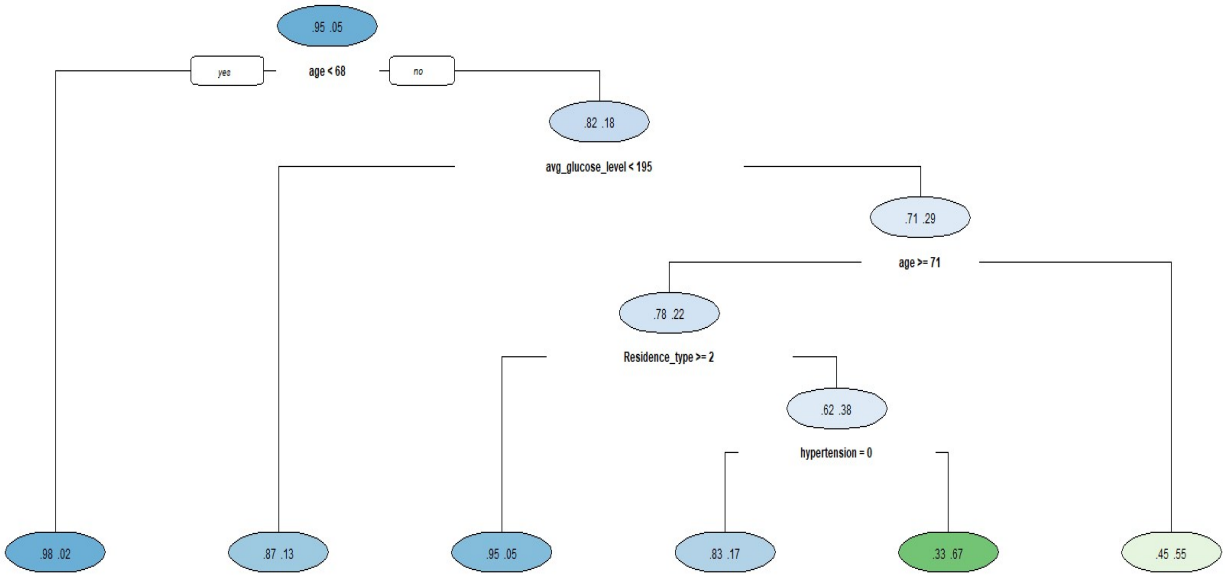


Fig 4.4

The root node, defined is "age 68," is where it all begins. As a result, the tree's initial inquiry is if the individual in question is under the age of 68. The tree shifts to the "avg_glucose level 195" child node on the left if the patient's age is less than 68. As a result, the tree will now

inquire as to whether the individual's average blood glucose level is less than 195 mg/dL. The tree shifts to the leftmost child node, "Residence_type>2," if the patient's average blood glucose level is less than 195. Accordingly, the tree's subsequent inquiry is whether the individual in need resides in a rural region (Residence_type>2). For individuals who satisfy the requirements for each leaf node, the probabilities at the bottom of every node in the leaf show the anticipated likelihood of stroke. For those who are under 68 and who have an average blood glucose level under 195, the chance of having a stroke is 95.05 percent. In general, the decision tree plot indicates that individuals who age, have greater mean glucose levels, and reside in rural regions are more likely to experience a stroke.

4.5 Neural Network Code:

#Importing the dataset

```
data<-Stroke.Data.Coded

data

View(data)

str(data)
```

#Checking for Null Values

```
is.na(data)

sum(is.na(data))/prod(dim(data))
```

#Splitting the data into train and test

```
set.seed(222)

datasplit<-sample(2, nrow(data), replace = TRUE, prob = c(0.8,0.2))

traindata<-data[datasplit==1,]

str(traindata)

testdata<-data[datasplit==2,]

str(testdata)
```

#Installing the packages

```
Install.packages("neural net")

library(neuralnet)

#Generating the model

form=stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Residence
_type+avg_glucose_level+bmi+smoking_status

model<-neuralnet(formula = form, data = traindata, hidden = 5 ,err.fct = 'ce',linear.output =
FALSE)

plot(model)

model$result.matrix

#Testing the Model on Train data

predict1<-compute(model, traindata)

head(predict1$net.result)

z<-predict1$net.result

pred2<-ifelse(z>0.5,1,0)

head(pred2)

#Confusion matrix

a<-table(pred2,traindata$stroke)

a

#Calculating the error

error1<-1-sum(diag(a))/sum(a)

error1

# Testing the Model on Test data

predict<-compute(model, testdata)

head(predict$net.result)

x<-predict$net.result
```

```
pred<-ifelse(x>0.5,1,0)

head(pred)
```

#Confusion Matrix Creation on Test Data

```
y<-table(pred,testdata$stroke)

y
```

#Calculating the error

```
error<-1-sum(diag(y))/sum(y)

error
```

4.5.1 Output Analysis:

Neural networks are machine learning algorithms that are inspired by the structure of the human brain. The neural networks are very much capable of understanding and mapping complex datasets. In this analysis, we have used a multilayer perceptron model. This model includes 3 layers viz., the input layer, the hidden layer, and the output layer. The input layer is used to feed the values to the model. The analysis and calculations occur within the hidden layer and the output layer is responsible for generating the output.

```
> #Importing the data
> data<-Stroke.Data.Coded
> View(data)
> str(data)
'data.frame':   4981 obs. of  11 variables:
 $ gender           : int  1 1 2 2 1 1 2 2 2 2 ...
 $ age              : num  67 80 49 79 81 74 69 78 81 61 ...
 $ hypertension     : int  0 0 0 1 0 1 0 0 1 0 ...
 $ heart_disease    : int  1 1 0 0 0 1 0 0 0 1 ...
 $ ever_married     : int  1 1 1 1 1 1 2 1 1 1 ...
 $ work_type        : int  1 1 1 2 1 1 1 1 1 3 ...
 $ Residence_type   : int  1 2 1 2 1 2 1 1 2 2 ...
 $ avg_glucose_level: num  229 106 171 174 186 ...
 $ bmi              : num  36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
 $ smoking_status   : int  1 2 3 2 1 2 2 4 2 3 ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
>
```

```
> summary(data)
     gender           age          hypertension     heart_disease
 Min.   :1.000   Min.   : 0.08   Min.   :0.00000   Min.   :0.00000
 1st Qu.:1.000   1st Qu.:25.00   1st Qu.:0.00000   1st Qu.:0.00000
 Median :2.000   Median :45.00   Median :0.00000   Median :0.00000
 Mean   :1.584   Mean   :43.42   Mean   :0.09617   Mean   :0.05521
 3rd Qu.:2.000   3rd Qu.:61.00   3rd Qu.:0.00000   3rd Qu.:0.00000
 Max.   :2.000   Max.   :82.00   Max.   :1.00000   Max.   :1.00000
  ever_married     work_type     Residence_type  avg_glucose_level
 Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 55.12
 1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.: 77.23
 Median :1.000   Median :1.000   Median :1.000   Median : 91.85
 Mean   :1.341   Mean   :1.825   Mean   :1.492   Mean   :105.94
 3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:113.86
 Max.   :2.000   Max.   :4.000   Max.   :2.000   Max.   :271.74
      bmi        smoking_status      stroke
 Min.   :14.0   Min.   :1.000   Min.   :0.00000
 1st Qu.:23.7   1st Qu.:2.000   1st Qu.:0.00000
 Median :28.1   Median :2.000   Median :0.00000
 Mean   :28.5   Mean   :2.584   Mean   :0.04979
 3rd Qu.:32.6   3rd Qu.:4.000   3rd Qu.:0.00000
 Max.   :48.9   Max.   :4.000   Max.   :1.00000
>
>
> #Checking for null values
> sum(is.na(data))/prod(dim(data))
[1] 0
```

```
> #Splitting the data into train and test
> set.seed(222)
> datasplit<-sample(2,nrow(data),replace = TRUE,prob = c(0.8,0.2))
> traindata<-data[datasplit==1,]
> str(traindata)
'data.frame':   4005 obs. of  11 variables:
 $ gender           : int  1 2 2 2 2 2 2 2 2 2 ...
 $ age              : num  80 49 79 69 78 81 61 54 79 50 ...
 $ hypertension     : int  0 0 1 0 0 1 0 0 0 1 ...
 $ heart_disease    : int  1 0 0 0 0 0 1 0 1 0 ...
 $ ever_married     : int  1 1 1 2 1 1 1 1 1 1 ...
 $ work_type        : int  1 1 2 1 1 1 3 1 1 2 ...
 $ Residence_type   : int  2 1 2 1 1 2 2 1 1 2 ...
 $ avg_glucose_level: num  105.9 171.2 174.1 94.4 58.6 ...
 $ bmi              : num  32.5 34.4 24 22.8 24.2 29.7 36.8 27.3 28.2 30.9 ...
 $ smoking_status   : int  2 3 2 2 4 2 3 3 2 2 ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
> testdata<-data[datasplit==2,]
> str(testdata)
'data.frame':   976 obs. of  11 variables:
 $ gender           : int  1 1 1 1 1 1 2 1 2 1 ...
 $ age              : num  67 81 74 75 71 42 82 54 39 73 ...
 $ hypertension     : int  0 0 1 1 0 0 1 0 1 1 ...
 $ heart_disease    : int  1 0 1 0 0 0 1 0 0 0 ...
 $ ever_married     : int  1 1 1 1 1 1 2 1 1 1 ...
 $ work_type        : int  1 1 1 1 1 1 1 1 1 2 ...
 $ Residence_type   : int  1 1 2 1 1 2 2 1 2 1 ...
 $ avg_glucose_level: num  228.7 186.2 70.1 221.3 102.9 ...
 $ bmi              : num  36.6 29 27.4 25.8 27.2 25.4 26.5 28.5 39.2 32.8 ...
 $ smoking_status   : int  1 1 2 3 1 4 1 2 3 2 ...
 $ stroke           : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
> #Model generation
> form<-stroke~age+gender+hypertension+heart_disease+ever_married+work_type+Resi
dence_type+avg_glucose_level+bmi+smoking_status
> model<-neuralnet(formula = form,data = traindata, hidden = 5, err.fct = 'ce',
linear.output = FALSE)
```

```
> # Testing the Model on Train data
> predict1<-compute(model,traindata)
> head(predict1$net.result)
         [,1]
2 0.08600437
3 0.08600437
4 0.08600437
7 0.08600437
8 0.08600437
9 0.08600437
> z<-predict1$net.result
> pred2<-ifelse(z>0.5,1,0)
> head(pred2)
  [,1]
2    0
3    0
4    0
7    0
8    0
9    0
```

```
> #Confusuion matrix
> a<-table(pred2,traindata$stroke)
> a

pred2    0    1
    0 3802  203
>
> #Calculation of error
> error1<-1-sum(diag(a))/sum(a)
> error1
[1] 0.05068664
```

The confusion matrix or the misclassification table shows that the neural network model was able to classify correctly around 3802 of entries. The error percentage is found to be 5% making the model highly efficient enough to be reliable for predicting based on the test dataset.

```
> # Testing the Model on Test data
> predict<-compute(model,testdata)
> head(predict$net.result)
            [,1]
1  0.0860043712
5  0.0860043712
6  0.0860043712
15 0.0860043712
21 0.0860043712
26 0.0005760471
> x<-predict$net.result
> pred<-ifelse(x>0.5,1,0)
> head(pred)
   [,1]
1     0
5     0
6     0
15    0
21    0
26    0
```

```
> #Confusion Matrix Creation on test data
> y<-table(pred,testdata$stroke)
> y

pred    0    1
   0 1408   75
> #Calculation of error
> error<-1-sum(diag(y))/sum(y)
> error
[1] 0.05057316
```

The model developed using the training dataset is now efficient enough to predict the outcome based on the test dataset. It can be seen from the confusion matrix that the model was able to classify 1408 entries correctly with a very minimal error of 5%. This means the neural network model was 95% efficient every time while predicting the outcome of possibility of occurrence of stroke based on the health and lifestyle factors of the person.

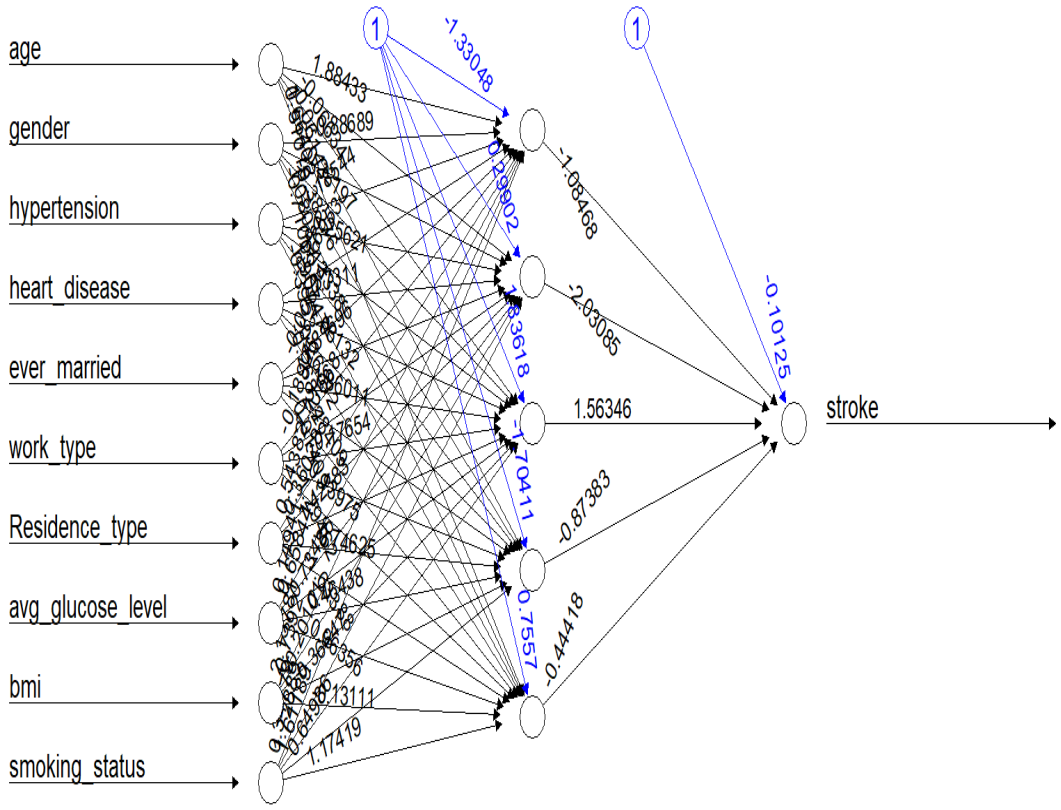


Fig 4.5

The above image displays the weights connected to each neural network variable. The weights show how significant each characteristic is in predicting the probability of having a stroke. The variable is more significant the higher the weight. For instance, the weights for age, high blood pressure, and coronary artery disease are all favourable, indicating a positive correlation

between these factors and the risk of stroke. Smoking raises the likelihood of stroke since the weight for being a smoker is also positive. The average blood sugar level and BMI have negative weights, indicating a negative correlation between these factors and the risk of stroke. This is probably because both high blood sugar and a high body mass index can damage vessels for blood and raise the possibility of stroke.

CHAPTER 5:

FINDINGS, SUGGESTIONS, AND CONCLUSION

5.1 Findings:

There was a total of 10 variables in the study. Viz., Age, Gender, BMI, Marriage, Employment, Residence, Smoking habit, Hypertension, Heart disease, and Blood Sugar level. These above variables can be categorized as Health factors and lifestyle factors. The health factors included BMI of the person, existence of any heart disease, the average blood sugar level of the person, Hypertension or Blood pressure, and Age.

These are the health factors that contribute to the possibility of the occurrence of stroke in a person. The correlation test conducted on these factors with respect to the occurrence of stroke resulted in a positive correlation. This shows that these factors were directly related to the dependent variable. Hence, we can find based on these factors that people at risk of these are more likely to develop a stroke than those with normal health conditions.

Apart from these health factors there are lifestyle factors too contribute to the person developing a stroke. There are various factors relating to the lifestyle of the person. The factors included in the study are the Gender, Residence of the person whether urban or rural. Marriage factor i.e., if the person is married or not, nature of employment such as full-time, part-time, or unemployed, Smoking habit of the person such as regular smoker, occasional smoker, or non-smoker. It was seen that these factors directly or indirectly affect the person's health thus leading to the risk of developing a stroke. E.g., consider the lifestyle factor of smoking habit. A person who smokes regularly is more at risk of experiencing a stroke due to the influence of tobacco.

Thus, encompassing all these health and lifestyle factors a predictive model was built using the R programming tool.

Logistic regression is a machine learning algorithm that is used for classification. The output of logistic regression is in binary form such as high or low, yes or no. The logistic regression model was able to classify with an accuracy of 94.6 % and an error rate of 5.4%.

The second machine learning algorithm under study was the decision tree. The decision tree is helpful in both classification and regression tasks. The unique nature of a decision tree is that it needs to make a decision at each step thus it takes into consideration various situations. Hence making it good and capable of predicting a reliable output. The decision tree predicted the outcome with an accuracy 95.4% of and an error rate of 4.6%.

Neural networks are the branch of machine learning models. These are developed to depict and work similar to a human brain. There are various classes of neural networks and the one under study here is multi-layer perceptron. The neural network develops its efficiency with time and learns to predict with better accuracy. In this study the neural network was able to predict the outcome with an accuracy of 95% and an error rate of 5%.

Thus, it can be concluded based on the error rate that logistic regression, decision tree, and neural network are all highly efficient and reliable enough to predict the outcome of possibility of occurrence of stroke based on the health and lifestyle factors.

5.2 Suggestions:

We can greatly improve preventative measures and further our understanding of this serious medical illness by starting research of predictive modelling for the likelihood of stroke based on lifestyle and health factors.

It is necessary to have a holistic strategy that takes into account many aspects of lifestyle and health in order to successfully dive into this topic. First off, including a wide range of health markers would give a full picture of a person's cardiovascular health, including physiological factors like cholesterol levels, blood pressure, and the body mass index (BMI).

These measurements have been recognised as key stroke risk indicators, and their incorporation in the prediction model would increase its precision.

Additionally, lifestyle variables are crucial in determining stroke risk, thus it is crucial to include them in the study. The state of your heart depends significantly on factors including eating habits, degree of exercise, smoking habits, and consumption of alcohol.

A thorough analysis of these lifestyle components will not only help create a predictive model that is more precise but also provide important information about risk factors that may be changed. This knowledge can then direct the creation of specialised preventative measures and therapies.

Additionally, it is crucial to take the time component of lifestyle influences into account. A dynamic view of lifestyle behaviours' effects on stroke risk might be possible with longitudinal data that records changes in these habits through time.

For instance, monitoring changes in eating habits, exercise routines, or attempts to quit smoking over a period of years helps clarify the overall implications of modifications to

lifestyle on cardiovascular health. By taking into account how risk factors change over time, such periodic considerations can considerably improve the model's prediction powers.

It is possible to improve the study's predicted accuracy by using sophisticated data analytics methods, such as the use of machine learning algorithms. In managing large datasets and seeing nuanced patterns that could elude standard statistical techniques, artificial intelligence models have demonstrated extraordinary effectiveness.

By utilising these methods, stroke risk might be better understood in its many nuances by revealing subtle connections between different health and lifestyle factors.

A varied and representative population of samples would be advantageous for the investigation. The findings would be more broadly applicable if they were inclusive of people from all age groups, genders, races, and socioeconomic situations.

A well-stratified sample enables the development of a prediction model that is more reliable and adaptable and can accommodate the population's various demographic traits.

The study can offer priceless insights under stroke prevention by adopting a holistic strategy, incorporating a variety of health and lifestyle variables, taking temporal dynamics into account, using modern analytics tools, and following ethical standards.

The ultimate objective is to create a reliable prediction model which not only identifies those who are at increased risk but also guides focused therapies and equips them to make wise lifestyle decisions to maintain cardiovascular health.

5.3 Conclusion:

As a result, the research on predictive modelling of stroke risk based on lifestyle and health characteristics provides significant insights that go far beyond the purview of scholarly investigation. The extensive investigation of physiological markers and lifestyle choices has shed light on the intricate interaction of factors affecting a person's vulnerability to stroke. A thorough picture of cardiovascular health has been made possible by the integration of several health indicators, such as cholesterol levels, blood pressure, BMI, and lifestyle factors including food, exercise, smoking, and alcohol use. These results highlight the complex interplay between changeable and non-changeable risk variables, opening the door for focused treatments that might alter public health policies.

We now have a more sophisticated understanding of stroke risk thanks to the study's temporal component. The study emphasises the cumulative effect of decisions on cardiovascular health by taking into account the change of lifestyle variables over time. This temporal perspective not only improves the model's ability to predict outcomes, but it also offers a complex account of how lifestyle choices affect people over the long run. This temporal knowledge will become increasingly important as we move forward in creating personalised preventative measures that take into account the ever-changing character of risk variables.

In terms of stroke risk prediction, the use of sophisticated analytics, in particular algorithms for machine learning, has become a shining example of innovation. These models open up new horizons for study and application because of their capacity to recognise nuanced patterns in large, complicated datasets. The work goes beyond the limitations of conventional statistical methods by utilising machine learning, which reveals subtle relationships and nonlinear linkages that help create a more complex prediction model.

One asset that strengthens the study's external reliability is the sample population's variety. The inclusion of people from a variety of different ages, genders, ethnic backgrounds, and socioeconomic statuses guarantees that the prediction model is both reliable and appropriate to a wide range of the population. This variety is consistent with the fact that stroke risk transcends demographic lines. The study's findings are thus applicable to creating treatments that speak to the various needs of various populations.

This study's prediction model is a powerful tool for preventative healthcare, not just a piece of academic literature. It has the capacity to recognise those who are at higher risk, enabling targeted therapies that can delay the beginning of strokes. Giving people this knowledge encourages a paradigm change in favour of proactive and knowledgeable healthcare decision-making.

The work on predictive modelling of risk for stroke, driven by lifestyle and health variables, essentially sits at the crossroads of serious scientific investigation and real-world social consequences. It outlines the future course for evidence-based healthcare solutions that are also keenly sensitive to the unique dynamics of risk. As we accept the learnings from the current research, we set out on a path to a day when strokes are not only anticipated but actively avoided, changing the face of cardiac wellness for future generations.

REFERENCES

Al-Mekhlafi Z, S. E. (2022). *Deep Learning and Machine Learning for Early Detection of Stroke and Haemorrhage.*

Bin Emdad F, T. S. (n.d.). *Towards Interpretable Multimodal Predictive Models for Early Mortality Prediction of Hemorrhagic Stroke Patients.*

Biswas N, U. K. (2022). *A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach.*

Dev S, W. H. (2022). *A predictive analytics approach for stroke prediction using machine learning and neural networks.*

Emon M, K. M. (2020). *Performance Analysis of Machine Learning Approaches in Stroke Prediction.*

Goldstein B, N. A. (2017). *Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review.*

Hung C, L. C. (2019). *Development of an intelligent decision support system for ischemic stroke risk assessment in a population-based electronic health record database.*

Islam R, D. S. (2021). *Predictive Analysis for Risk of Stroke Using Machine Learning Techniques.*

Jeena R, K. S. (2017). *Stroke prediction using SVM.*

Luk J, C. R. (2006). *Does age predict outcome in stroke rehabilitation? A study of 878 Chinese subjects.*

M, Y. S. (2016). *Classification of Ischemic Stroke using Machine Learning Algorithms.*

Min S, P. S. (2018). *Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea.*

Nwosu C, D. S. (2019). *Predicting Stroke from Electronic Health Records.*

Pathan M, J. Z. (2020). *Identifying Stroke Indicators Using Rough Sets.*

Rahman S, H. M. (2023). *Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques.*

Saleh H, A.-E. G. (2019). *Stroke Prediction using Distributed Machine Learning Based on Apache Spark.*

Singh M, C. P. (2017). *Stroke prediction using artificial intelligence.*

T, A. (2022). *RISK FACTOR IDENTIFICATION FOR STROKE PROGNOSIS USING MACHINE-LEARNING ALGORITHMS.*

Wang W, K. M. (2020). *A systematic review of machine learning models for predicting outcomes of stroke with structured data.*

Yahiya S, Y. A. (2022). *Classification of Ischemic Stroke using Machine Learning Algorithms.*

https://www.healthline.com/health/stroke

https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

https://www.mayoclinic.org/diseases-conditions/stroke/symptoms-causes/syc-20350113

https://my.clevelandclinic.org/health/diseases/5601-stroke

https://www.cdc.gov/stroke/about.htm

https://main.mohfw.gov.in/sites/default/files/Fin%20-%20Stroke%20guidelines_0.pdf

https://www.fda.gov/consumers/minority-health-and-health-equity-resources/stroke

https://en.wikipedia.org/wiki/Stroke#:~:text=Stroke%20is%20a%20medical%20condition,brain%20to%20stop%20functioning%20properly.

https://www.nhlbi.nih.gov/health/stroke#:~:text=A%20stroke%20can%20occur%20when,and%20nutrients%20from%20the%20blood.

Link to Dataset:

Link to Stroke Dataset

ANNEXURE

A: PLAGIARISM REPORT

DrillBit

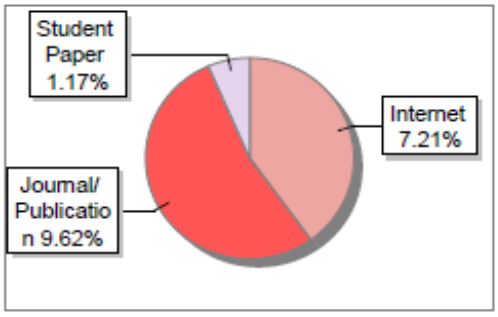
The Report is Generated by DrillBit Plagiarism Detection Software

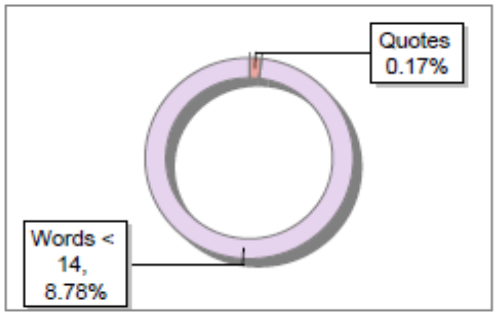
Submission Information

Author Name	B SHASHANK
Title	Predictive Modeling of Stroke Risk Factors usin..
Paper/Submission ID	1028458
Submitted by	nnsreddy.rvim@rvei.edu.in
Submission Date	2023-10-16 13:53:02
Total Pages	78
Document type	Dissertation

Result Information

Similarity 18 %





Exclude Information

Quotes	Not Excluded
References/Bibliography	Not Excluded
Sources: Less than 14 Words Similarity	Not Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

A Unique QR Code use to View/Download/Share Pdf File

DrillBit

DrillBit Similarity Report

18	207	B
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

	LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
1		Submitted to Visvesvaraya Technological University, Belagavi	1	Student Paper
2		ijcsmc.com	1	Publication
3		qdoc.tips	<1	Internet Data
4		www.mdpi.com	<1	Internet Data
5		Thesis Submitted to Shodhganga Repository	<1	Publication
6		Clustering technique-based least square support vector machine for EEG by Siuly-211	<1	Publication
7		Identifying Stroke Indicators Using Rough Sets by Pathan-2020	<1	Publication
8		A data mining system for providing analytical information on brain tum by Santos-213	<1	Publication
9		Distributed Computing and Internet Technology 17th International Conference, IC by Digant-2021	<1	Publication
10		www.dx.doi.org	<1	Publication
11		research.assaf.org.za	<1	Publication
12		docplayer.net	<1	Internet Data
13		Thesis Submitted to Shodhganga, shodhganga.inflibnet.ac.in	<1	Publication

14	coek.info	<1	Internet Data
15	liba.edu	<1	Publication
16	www.doaj.org	<1	Publication
17	www.dx.doi.org	<1	Publication
18	www.ijcaonline.org	<1	Publication
19	www.msdmanuals.com	<1	Internet Data
20	www.tobaccoinduceddiseases.org	<1	Publication
21	academicjournals.org	<1	Internet Data
22	index-of.es	<1	Publication
23	arxiv.org	<1	Publication
24	docplayer.net	<1	Internet Data
25	encord.com	<1	Internet Data
26	etd.aau.edu.et	<1	Publication
27	ajet.org.au	<1	Publication
28	biomedcentral.com	<1	Internet Data
29	Calibration and validation of multiple regression models for stormwater quality by Mourad-2005	<1	Publication
30	mdpi.com	<1	Internet Data
31	njamhaa.org	<1	Internet Data
32	worldwidescience.org	<1	Internet Data