

# **FORECASTING STROKE RISK- A MACHINE LEARNING APPROACH**

Submitted in partial fulfillment of the requirements for the award of Bachelor of  
Technology degree in Biotechnology

Interdisciplinary Project

By

**APARNA M (41230011)**

**D JESLIN SANJANA (41230024)**

**KRISHNA R MENON (41230031)**



**DEPARTMENT OF BIOTECHNOLOGY**

**SCHOOL OF BIO AND CHEMICAL ENGINEERING**

# **SATHYABAMA**

**INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(DEEMED TO BE UNIVERSITY)**

**CATEGORY -1 UNIVERSITY BY UGC**

**Accredited with Grade“A++” by NAAC I12BStatus by UGC I Approved by AICTE**

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600 119**

**MAY-2024**



**SATHYABAMA**

INSTITUTE OF SCIENCE AND TECHNOLOGY  
(DEEMED TO BE UNIVERSITY)

Accredited "A++" Grade by NAAC | 12B Status by UGC | Approved by AICTE  
[www.sathyabama.ac.in](http://www.sathyabama.ac.in)

**DEPARTMENT OF BIOTECHNOLOGY**  
**BONAFIDE CERTIFICATE**

This is to certify that this Project Report is the bonafide work of **APARNA M (41230011)**, **D JESLIN SANJANA (41230024)** and **KRISHNA R MENON (41230031)** who carried out the interdisciplinary project entitled "**FORECASTING STROKE RISK- A MACHINE LEARNING APPROACH**" under our supervision from **JANUARY 2024** to **APRIL 2024**.

Internal Guide

**Dr. S.M INDUMATHI M.Phil., Ph.D.**

HEAD OF THE DEPARTMENT

**Dr. V. RAMESH KUMAR, M.Phil., Ph.D.**

---

Submitted for Viva Voce Examination held on 06.05.2024

Internal Examiner

External Examiner

## **DECLARATION**

We, **APARNA M (41230011)**, **D JESLIN SANJANA (41230024)** and **KRISHNA R MENON (41230031)** hereby declare that the Project Report entitled “**FORECASTING STROKE RISK- A MACHINE LEARNING APPROACH**” done by us under the guidance of **Dr. S.M. INDUMATHI, M.Phil., Ph.D.**, is submitted in partial fulfillment of the requirements for the award of Degree in Bachelor of Technology in Biotechnology

**DATE: 06.05.2024**

**PLACE: Chennai**

A handwritten signature in blue ink, appearing to read 'Aparna M', written over a horizontal line.

**SIGNATURE OF THE CANDIDATE**

## ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to The Board of Management of **Sathyabama Institute of Science and Technology** for their encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. V. RAMESH KUMAR, M.Phil., Ph.D., Head of the Department**, Department of Biotechnology for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr. S.M INDUMATHI M.Phil., Ph.D., Assistant Professor**, Department of Biotechnology for their valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks **to all Teaching and Non-teaching staff members** of the Department of Biotechnology who were helpful in many ways for the completion of the project work.

## ABSTRACT

This initiative introduces a groundbreaking approach to predicting stroke risk using machine learning algorithms, leveraging a comprehensive dataset encompassing biomarkers, lifestyle factors, medical history, and demographics. By employing state-of-the-art algorithms such as random forests, support vector machines, and neural networks, the methodology aims to identify subtle correlations and patterns indicative of stroke risk. Through feature significance analysis, doctors gain valuable insights for personalized patient care. Focused on individuals with heart disease, the project utilizes Kaggle-sourced data to develop precise prediction models, enabling early detection and prevention of strokes, thereby addressing a critical global health challenge and potentially reducing stroke-related morbidity and mortality.

**Keywords:** Datasets, Machine learning, Prediction models, Risk assessments, Stroke

## TABLE OF CONTENTS

| CHAPTER NO. | TITLE                                              | PAGE NO.   |
|-------------|----------------------------------------------------|------------|
|             | <b>ABSTRACT</b>                                    | <b>i</b>   |
|             | <b>LIST OF FIGURES</b>                             | <b>iii</b> |
| <b>1</b>    | <b>INTRODUCTION</b>                                | <b>1</b>   |
| <b>2</b>    | <b>LITERATURE SURVEY</b>                           | <b>2</b>   |
|             | 2.1 Before the use of machine learning in medicine | 3          |
|             | 2.2 Using machine learning for disease prediction  | 5          |
| <b>3</b>    | <b>AIM AND SCOPE</b>                               | <b>9</b>   |
| <b>4</b>    | <b>MATERIALS AND METHODS</b>                       | <b>10</b>  |
|             | 4.1 Materials                                      | 10         |
|             | 4.2 Methods                                        | 10         |
|             | 4.2.1 Data Collection                              | 10         |
|             | 4.2.2 Data Pre-processing                          | 11         |
|             | 4.2.3 Feature Selection                            | 11         |
|             | 4.2.4 Model Selection                              | 12         |
|             | 4.2.5 Model Training and Evaluation                | 13         |
|             | 4.2.6 Model Deployment                             | 13         |
|             | 4.2.7 Validation and Interpretation                | 25         |
| <b>5</b>    | <b>RESULTS AND DISCUSSION</b>                      | <b>28</b>  |
| <b>6</b>    | <b>SUMMARY AND CONCLUSION</b>                      | <b>36</b>  |
|             | <b>REFERENCE</b>                                   | <b>37</b>  |

## LIST OF FIGURES

| FIGURE No. | TITLE                                                                                                                                         | PAGE No. |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------|----------|
| 4.1        | Data collection from Kaggle                                                                                                                   | 10       |
| 4.2        | Data pre-processing                                                                                                                           | 11       |
| 4.3        | Finding and correcting the missing values                                                                                                     | 11       |
| 4.4        | Model selection process                                                                                                                       | 12       |
| 4.5        | Loading pre-processed dataset into google colab.                                                                                              | 13       |
| 4.6        | Uploading data set to google colab using codes                                                                                                | 16       |
| 5.1        | Distribution of stroke cases by gender                                                                                                        | 29       |
| 5.2        | Age distribution of dataset                                                                                                                   | 29       |
| 5.3        | Smoking cases by smoking status                                                                                                               | 30       |
| 5.4        | Distribution of hypertension and heart disease cases                                                                                          | 30       |
| 5.5        | Glucose level by stroke status                                                                                                                | 31       |
| 5.6        | Pair plot of numerical features                                                                                                               | 31       |
| 5.7        | Count of work types                                                                                                                           | 32       |
| 5.8        | Chance of getting stroke based on age group,<br>BMI, avg. glucose level, hypertension,<br>heart disease, work type, residence type and gender | 32       |
| 5.9        | Hypertension graph                                                                                                                            | 33       |
| 5.10       | Feature importance                                                                                                                            | 33       |
| 5.11       | Plot of accuracy, precision, recall,<br>F1 score and ROC AUC score of the model used                                                          | 34       |

# CHAPTER 1

## INTRODUCTION

The ability to anticipate and prevent unfavorable health outcomes is critical in the field of medicine. Strokes stand out among these as one of the main global causes of death and permanent disability. Traditional approaches to evaluating stroke risk frequently depend on lifestyle choices, medical history, and demographic characteristics. Although these provided insightful information, they might not have the accuracy and flexibility required for prompt action. Let us introduce machine learning, an area of artificial intelligence that focuses in finding patterns in large, complicated datasets and generating precise predictions. (Hayes *et al.*, 2010)

Here we also explore the field of machine learning-based stroke risk prediction. We investigate how advanced algorithms powered by massive volumes of medical data can help physicians better identify patients who are at high risk of having a stroke. Our goal is to change the healthcare mindset from reactive to proactive by utilizing the power of analytics for prediction. We present a summary of the main risk factors for strokes and the difficulties in predicting them based on a synthesis of the body of current research. Next, we shed light on how machine learning models—like decision trees, support vector machines, and neural networks—can be used to represent complex associations between various risk variables and stroke incidents. We also discuss the consequences and ethical issues surrounding the use of machine learning algorithms in therapeutic contexts. These technologies have a lot of potential, but they also bring up issues with algorithmic bias, data privacy, and the human-AI interface. (Howard *et al.*, 2021).

Our ultimate objective is to make a contribution to the current discussion on the use of machine learning for stroke risk assessment into healthcare systems. Through proactive and personalized treatments, we foresee a future where strokes are not only treated but also avoidable through multidisciplinary collaboration between data scientists, lawmakers, and healthcare practitioners. (Nwosu *et al.*, 2019)

## CHAPTER 2

### LITERATURE SURVEY

Stroke is a major global health issue that contributes to high rates of morbidity and death globally. Stroke is a complicated multifactorial disorder that frequently results from a confluence of lifestyle variables, genetic predispositions, and comorbidities said by Chatterjee *et al.* (2020). As such, predicting and preventing stroke is both difficult and crucial. Even while they are useful, traditional risk assessment techniques might not provide the predictive precision and granularity required for tailored responses.

The field of medical research has undergone a revolution in recent years due to the introduction of machine learning techniques, which provide strong tools for examining large datasets and finding complex patterns within them given Kain and Dey *et al.* (2021) By merging disparate sets of data and spotting nuanced correlations that may escape the notice of traditional statistical methods, utilizing the capabilities of machine learning algorithms has the potential to improve stroke risk prediction.

The purpose of this review of the literature is to present a thorough summary of previous studies that have been conducted with the goal of predicting the risk of stroke by machine learning techniques. This review aims to clarify the current state of machine learning applications in stroke risk prediction by combining findings from a wide range of publications and emphasizing significant developments, difficulties, and areas for further investigation.

Through a critical evaluation of the body of existing literature, the goal of this study is to pinpoint methodological flaws, knowledge gaps, and areas that need further research. The ultimate objective of this literature review is to provide valuable insights that will guide the creation of a strong machine learning framework for predicting the risk of stroke given by Nwosu *et al.* (2019). This framework will help individuals who are at risk of stroke by facilitating early identification, tailored intervention, and better outcomes.

## 2.1 BEFORE THE USE OF MACHINE LEARNING IN MEDICINE

A thorough analysis of both modifiable and non-modifiable stroke risk factors, such as genetics, lifestyle, smoking, diabetes, hypertension, and diabetes, is given by Boehme *et al.* (2017). They also go over stroke preventive techniques, stressing the value of focused therapies that are determined by each patient's unique risk profile.

In their assessment of temporal changes in stroke incidence, Kissela *et al.* (2012) concentrate on age-specific rates within a sizable population that is multiracial. Their study sheds information on how the population is changing and how the risk of stroke varies over time, helping to shape prevention measures and the distribution of healthcare resources.

The interstroke study's results, which look at the regional and global impacts of modifiable risk factors linked to acute stroke, are presented by O'Donnell *et al.* (2016). Through the identification of critical risk factors like food, physical activity, alcohol consumption, and hypertension, their analysis offers valuable insights into worldwide stroke prevention methods.

A systematic review and meta-analysis of the predictive power of the CHADS<sub>2</sub> and CHA<sub>2</sub>DS<sub>2</sub>-VASc scores for stroke risk in atrial fibrillation patients is carried out by Sposato *et al.* (2015). Their analysis informs risk stratification procedures and sheds light on the usefulness of these score systems.

A summary of the global epidemiology of intracerebral hemorrhage (ICH) is given by Ikram *et al.* (2012), who also look at geographical differences, long-term trends, and risk factors that are related. Their analysis sheds light on the cost of ICH to international healthcare systems and identifies areas that could be prevented or managed.

Kelly-Hayes *et al.* (2010) uses data from longitudinal studies to discuss how age and health-related habits affect the risk of stroke. The study highlights the significance of healthy behaviors for stroke prevention throughout the lifetime by examining the effects of lifestyle factors such as food, exercise, smoking, and alcohol consumption on the incidence of strokes.

In the International Stroke Trial, Niewada *et al.* (2005) investigate the impact of gender on baseline characteristics and clinical outcomes in patients with confirmed ischemic stroke. Their analysis reveals variations in stroke risk factors, presentation, and outcomes between genders, providing valuable insights for customized prevention and treatment plans.

Reeves *et al.* (2008) examine how gender variations exist in the epidemiology, clinical manifestation, treatment, and results of stroke. Their analysis emphasizes the significance of individualized approaches to stroke prevention and therapy by shedding light on gender-specific patterns of stroke incidents.

## 2.2 USING MACHINE LEARNING FOR DISEASE PREDICTION

Global research attention has been drawn to the use of medical databases. Utilizing a collection of medical information, data mining techniques have been widely applied in the development of decision support systems for illness prediction. Fuzzy rules for the knowledge-based system were created using Classification and Regression Trees (CART), and these were evaluated using open data sets. The outcomes demonstrated the potential effectiveness of fuzzy rule-based CART in conjunction with noise reduction and clustering approaches for the prediction of diseases from real-world medical datasets. As a clinical analysis tool, the knowledge-based system can help healthcare professionals in their work. Several diseases were predicted with the machine learning techniques such as diabetes, heart diseases, stroke etc.

Different researchers have used machine learning algorithms, data mining techniques, or combinations of these techniques to build and apply a variety of prediction models. In 2015, a system was implemented using data about diabetes is analyzed using Hadoop and Map Reduce techniques. This approach forecasts the kind of diabetes as well as the hazards that come with it. The technology is affordable for any healthcare business and is based on Hadoop. Iyer *et al.* (2015) examined latent patterns in a diabetes dataset by using a classification algorithm.

The research community has shown great interest in developing tools and methods for monitoring and predicting various diseases that have a significant impact on human health. In this section, we will present the latest works that utilize machine learning techniques for the prediction of stroke by using different approaches (Hayes *et al.*, 2010).

Alghamdi *et al.* (2020) offer a comprehensive summary of machine learning and artificial intelligence strategies used in stroke prediction methodologies. Their evaluation includes a broad spectrum of studies investigating different datasets, methods, and predictive factors to improve the accuracy of stroke risk prediction.

Arboix and Alió *et al.* (2020) investigate the risk profiles associated with several ischemic stroke subtypes as well as the cardiovascular risk variables that contribute to acute stroke. Their investigation clarifies the variable nature of the etiology of stroke and its consequences for predictive modeling.

A review concentrating on machine learning techniques for forecasting stroke risk variables is presented by Chatterjee and Dey *et al.* (2020). They provide information on how different machine learning algorithms are applied and how well they capture the intricate interaction of risk factors linked to the incidence of stroke.

Choi *et al.* (2020) discuss the applications of machine learning in stroke imaging and diagnosis, highlighting its potential for improving diagnostic accuracy and efficiency. Their review explores the integration of imaging data with clinical variables to enhance stroke risk prediction models.

Comella-Del-Barrio and Cano-Izquierdo *et al.* in 2020 review discussed machine learning methods used to predict strokes. They provide information about potential and obstacles in the sector, as well as the relative performance of techniques.

Ding *et al.* (2020) provide a systematic review of research employing machine learning algorithms to predict stroke risk variables. Their investigation sheds light on the advantages, disadvantages, and possible therapeutic uses of several machine learning algorithms for predicting the risk of stroke.

A comprehensive overview of machine learning techniques for predicting stroke risk variables is presented by Dey and Sain *et al.* (2021). Their review includes a variety of studies that look at various datasets, machine learning algorithms, and validation methods used in models for predicting stroke risk. The naïve Bayes classifier's accuracy was 85.6%, while the J48, K-nearest classifier's accuracy was 99.8% of the neighbor and random forest.

The authors suggested an approach to use social media resources to learn about the different symptoms linked to the stroke disease and stroke prevention strategies. They developed an architecture that uses spectral clustering to repeatedly cluster tweets according to their content. Ten-fold cross-validation, naïve Bayes, support vector machines, and probabilistic neural networks (PNNs) were used in the trials. The PNN performed better than other algorithms, achieving an accuracy of 89.90% (Shoily *et al.*, 2012).

In naïve logistic regression, the following methods were used to categorize stroke risk levels: Bayes, Bayesian network, decision tree, neural network, random forest, bagged decision tree, voting, and boosting model with decision trees. The outcome of the experiment revealed that the random forest obtained the highest precision (97.33%), while the boosting model with decision trees acquired the maximum recall (99.94%) given by Choi *et al.* (2020).

Numerous machine learning methods, including logistic regression, decision trees, random forests, K-nearest neighbor, support vector machines, and naïve Bayes, are recommended for use in this research project. When it came to the prediction of stroke, the naïve Bayes algorithm outperformed the other algorithms with an accuracy of 82%.

The authors want to use machine learning and mining techniques to identify the kind of stroke by obtaining a stroke dataset from Sugam Multispecialty Hospital in India. An artificial neural network trained using stochastic gradient descent yielded 91% accuracy, whereas the support vector machine and ensemble (bagged) categories achieved 81% accuracy. Algorithms fared better than the others, with a classification accuracy of above 95%.

Furthermore, in order to determine the influence of risk variables on stroke prediction, an examination of patients' electronic health records was conducted in this paper. The neural network, decision tree, and random forest classification accuracy across 1000 experiments on the dataset of electronic health records

Lastly, using automatic image processing, the capacity of machine learning algorithms to evaluate diffusion-weighted imaging (DWI) and fluid-attenuated inversion recovery (FLAIR) pictures of stroke patients within 24 hours of symptom onset was examined. To estimate the stroke, start for binary classification ( $\leq 4.5$  h), three machine learning models—logistic regression, support vector machine, and random forest—were created. The sensitivity and specificity for detecting patients within 4.5 hours were used to evaluate the machine learning model and were compared to human readings of the DWI-FLAIR mismatch (Alghamdi *et al.*, 2020).

This literature analysis concludes by highlighting the revolutionary potential of machine learning in improving efforts to predict and prevent stroke. We can enhance risk assessment, facilitate early intervention, and eventually lessen the impact of stroke on patients, healthcare systems, and society at large by utilizing data-driven approaches.

## **CHAPTER 3**

### **AIM AND SCOPE**

#### **3.1 AIM**

This research aims to develop an interpretable and accurate stroke risk prediction model through machine learning techniques, utilizing extensive patient data including genetic markers, lifestyle variables, and medical history. By validating the model's performance across diverse healthcare settings and addressing the ethical concerns, our goal is to facilitate proactive intervention strategies and improve patient outcomes stroke prevention.

#### **3.2 SCOPE**

- Develop a machine learning-based method for stroke risk prediction, integrating comprehensive data collection, preprocessing, and feature selection techniques to enhance predictive accuracy and interpretability.
- Evaluate the performance of the developed models using rigorous evaluation metrics such as AUC-ROC, sensitivity and specificity, and accuracy, prioritizing interpretability and explain ability through methods like LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Addictive explanations).
- Validate the robustness and practical applicability of the models across diverse healthcare settings through the independent dataset validation, enabling seamless integration into clinical workflows for a real-time risk assessment and personalized intervention strategies.
- Address ethical considerations including algorithmic bias, informed consent, and data privacy to ensure fairness, protect patient rights, and provide equitable results across demographic groups.

## CHAPTER 4

### MATERIALS AND METHODS

#### 4.1 MATERIALS

For this project, Python programming is the primary language employed. Utilizing Python is a flexible and effective method. For the purpose of this endeavour, Google Colab offers the perfect platform: a cloud-based environment with the seamless collaboration features and access to powerful GPUs.

#### 4.2 METHODS

##### 4.2.1 Data collection

Acquiring a comprehensive dataset containing relevant features such as Gender, Age, Hypertension, Heart disease, Marriage Status, Smoking status, BMI, Average Glucose Level, Work type and Residence type associated with ischemic stroke risk as shown in fig 4.1. Datasets may be sourced from Kaggle.

|    | A     | B      | C   | D         | E          | F          | G          | H         | I         | J    | K          | L      |
|----|-------|--------|-----|-----------|------------|------------|------------|-----------|-----------|------|------------|--------|
| 1  | id    | gender | age | hypertens | heart_dise | ever_marri | work_type  | Residence | avg_gluco | bmi  | smoking    | stroke |
| 2  | 9046  | Male   | 67  | 0         | 1          | Yes        | Private    | Urban     | 228.69    | 36.6 | formerly s | 1      |
| 3  | 51676 | Female | 61  | 0         | 0          | Yes        | Self-emplc | Rural     | 202.21    | N/A  | never smc  | 1      |
| 4  | 31112 | Male   | 80  | 0         | 1          | Yes        | Private    | Rural     | 105.92    | 32.5 | never smc  | 1      |
| 5  | 60182 | Female | 49  | 0         | 0          | Yes        | Private    | Urban     | 171.23    | 34.4 | smokes     | 1      |
| 6  | 1665  | Female | 79  | 1         | 0          | Yes        | Self-emplc | Rural     | 174.12    | 24   | never smc  | 1      |
| 7  | 56669 | Male   | 81  | 0         | 0          | Yes        | Private    | Urban     | 186.21    | 29   | formerly s | 1      |
| 8  | 53882 | Male   | 74  | 1         | 1          | Yes        | Private    | Rural     | 70.09     | 27.4 | never smc  | 1      |
| 9  | 10434 | Female | 69  | 0         | 0          | No         | Private    | Urban     | 94.39     | 22.8 | never smc  | 1      |
| 10 | 27419 | Female | 59  | 0         | 0          | Yes        | Private    | Rural     | 76.15     | N/A  | Unknown    | 1      |
| 11 | 60491 | Female | 78  | 0         | 0          | Yes        | Private    | Urban     | 58.57     | 24.2 | Unknown    | 1      |
| 12 | 12109 | Female | 81  | 1         | 0          | Yes        | Private    | Rural     | 80.43     | 29.7 | never smc  | 1      |
| 13 | 12095 | Female | 61  | 0         | 1          | Yes        | Govt_job   | Rural     | 120.46    | 36.8 | smokes     | 1      |
| 14 | 12175 | Female | 54  | 0         | 0          | Yes        | Private    | Urban     | 104.51    | 27.3 | smokes     | 1      |
| 15 | 8213  | Male   | 78  | 0         | 1          | Yes        | Private    | Urban     | 219.84    | N/A  | Unknown    | 1      |
| 16 | 5317  | Female | 79  | 0         | 1          | Yes        | Private    | Urban     | 214.09    | 28.2 | never smc  | 1      |
| 17 | 58202 | Female | 50  | 1         | 0          | Yes        | Self-emplc | Rural     | 167.41    | 30.9 | never smc  | 1      |
| 18 | 56112 | Male   | 64  | 0         | 1          | Yes        | Private    | Urban     | 191.61    | 37.5 | smokes     | 1      |
| 19 | 34120 | Male   | 75  | 1         | 0          | Yes        | Private    | Urban     | 221.29    | 25.8 | smokes     | 1      |
| 20 | 27458 | Female | 60  | 0         | 0          | No         | Private    | Urban     | 89.22     | 37.8 | never smc  | 1      |
| 21 | 25226 | Male   | 57  | 0         | 1          | No         | Govt_job   | Urban     | 217.08    | N/A  | Unknown    | 1      |
| 22 | 70630 | Female | 71  | 0         | 0          | Yes        | Govt_iob   | Rural     | 193.94    | 22.4 | smokes     | 1      |

**Fig 4.1 Data collection from kaggle**

### 4.2.2 Data pre-processing

Cleaning and pre-processing the acquired data to handle missing values, outliers and inconsistencies as shown in fig 4.2. This step may also involve feature engineering to the extract meaningful information and enhance the predictive power of the model.



**Fig 4.2 Data Pre-Processing**

### 4.2.3 Feature selection

Identifying the most relevant features for stroke prediction through exploratory data analysis and statistical methods as shown in fig 4.3. Feature selection helps in reducing dimensionality and improving model interpretability and generalization.

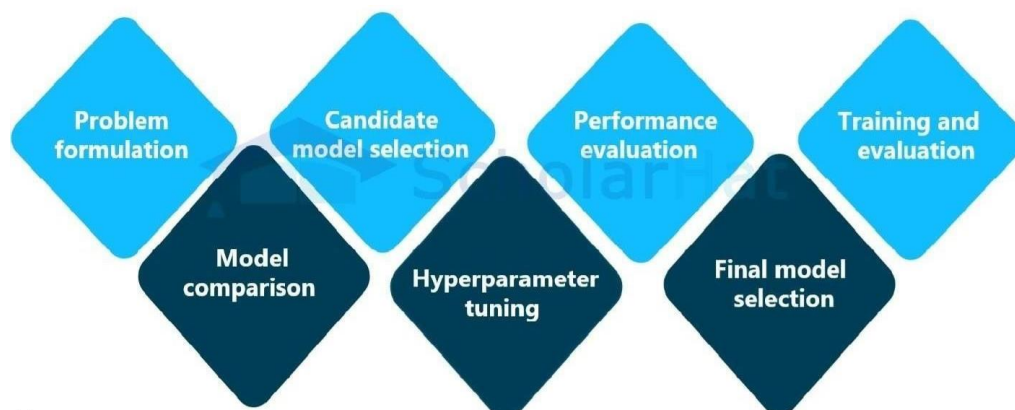
|    | A     | B      | C   | D         | E          | F        | G         | H         | I           | J    | K          | L      | M |
|----|-------|--------|-----|-----------|------------|----------|-----------|-----------|-------------|------|------------|--------|---|
| 1  | id    | gender | age | hypertens | heart_dise | ever_mar | work_type | Residence | avg_glucose | bmi  | smoking    | stroke |   |
| 2  | 9046  | Male   | 67  | 0         | 1          | Yes      | Private   | Urban     | 228.69      | 36.6 | formerly s | 1      |   |
| 3  | 51676 | Female | 61  | 0         | 0          | Yes      | Self-empl | Rural     | 202.21      | N/A  | never smc  | 1      |   |
| 4  | 31112 | Male   | 80  | 0         | 1          | Yes      | Private   | Rural     | 105.92      | 32.5 | never smc  | 1      |   |
| 5  | 60182 | Female | 49  | 0         | 0          | Yes      | Private   | Urban     | 171.23      | 34.4 | smokes     | 1      |   |
| 6  | 1665  | Female | 79  | 1         | 0          | Yes      | Self-empl | Rural     | 174.12      | 24   | never smc  | 1      |   |
| 7  | 56669 | Male   | 81  | 0         | 0          | Yes      | Private   | Urban     | 186.21      | 29   | formerly s | 1      |   |
| 8  | 53882 | Male   | 74  | 1         | 1          | Yes      | Private   | Rural     | 70.09       | 27.4 | never smc  | 1      |   |
| 9  | 10434 | Female | 69  | 0         | 0          | No       | Private   | Urban     | 94.39       | 22.8 | never smc  | 1      |   |
| 10 | 27419 | Female | 59  | 0         | 0          | Yes      | Private   | Rural     | 76.15       | N/A  | Unknown    | 1      |   |
| 11 | 60491 | Female | 78  | 0         | 0          | Yes      | Private   | Urban     | 58.57       | 24.2 | Unknown    | 1      |   |
| 12 | 12109 | Female | 81  | 1         | 0          | Yes      | Private   | Rural     | 80.43       | 29.7 | never smc  | 1      |   |
| 13 | 12095 | Female | 61  | 0         | 1          | Yes      | Govt_job  | Rural     | 120.46      | 36.8 | smokes     | 1      |   |
| 14 | 12175 | Female | 54  | 0         | 0          | Yes      | Private   | Urban     | 104.51      | 27.3 | smokes     | 1      |   |
| 15 | 8213  | Male   | 78  | 0         | 1          | Yes      | Private   | Urban     | 219.84      | N/A  | Unknown    | 1      |   |
| 16 | 5317  | Female | 79  | 0         | 1          | Yes      | Private   | Urban     | 214.09      | 28.2 | never smc  | 1      |   |
| 17 | 58202 | Female | 50  | 1         | 0          | Yes      | Self-empl | Rural     | 167.41      | 30.9 | never smc  | 1      |   |
| 18 | 56112 | Male   | 64  | 0         | 1          | Yes      | Private   | Urban     | 191.61      | 37.5 | smokes     | 1      |   |
| 19 | 34120 | Male   | 75  | 1         | 0          | Yes      | Private   | Urban     | 221.29      | 25.8 | smokes     | 1      |   |
| 20 | 27458 | Female | 60  | 0         | 0          | No       | Private   | Urban     | 89.22       | 37.8 | never smc  | 1      |   |
| 21 | 25226 | Male   | 57  | 0         | 1          | No       | Govt_job  | Urban     | 217.08      | N/A  | Unknown    | 1      |   |
| 22 | 70630 | Female | 71  | 0         | 0          | Yes      | Govt_iob  | Rural     | 193.94      | 22.4 | smokes     | 1      |   |

Missing value

**Fig 4.3 Finding and correcting the missing values**

#### **4.2.4 Model selection**

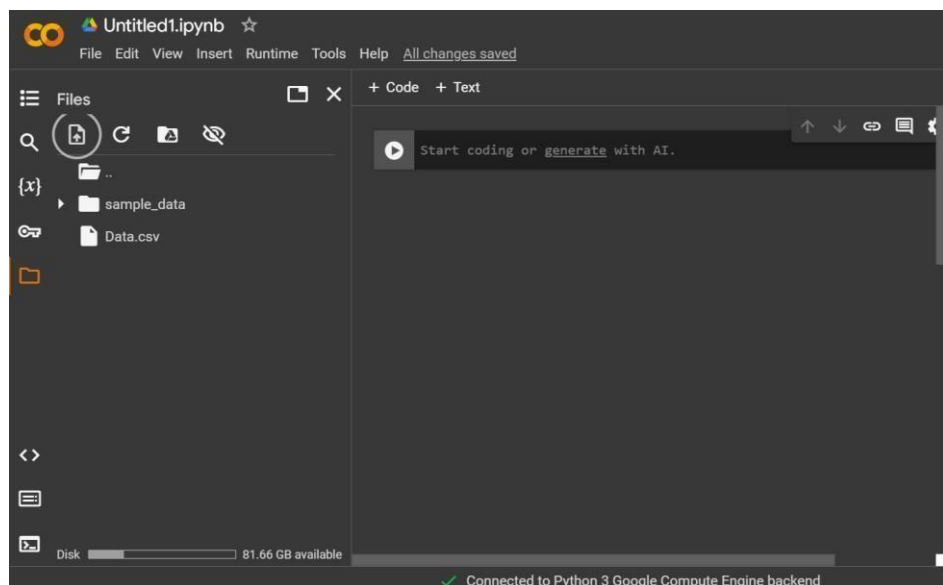
Experimenting with various machine learning algorithms suitable for binary classification tasks, such as logistic regression, decision trees, random forests, support vector machines, and neural networks as shown in fig 4.4. Model selection involves the assessing performance metrics such as accuracy, precision, recall, and an area under receiver operating characteristic curve (AUC-ROC). By comparing the performance of different models, we can identify the one that best balances with the predictive power generalizability, and the computational efficiency for our very specific prediction task. Additionally, the model selection may heavily involve in considering the interpretability of the chosen algorithm, ensuring that the clinicians can understand and trust the predictions generated for ischemic stroke risk.



**Fig 4.4 Model Selection Process**

#### 4.2.5 Model training and evaluation

Training the selected machine learning models on a portion of the dataset and evaluating their performance using cross-validation techniques to ensure robustness and prevent overfitting. Performance evaluation metrics will guide the selection of the most suitable model for stroke prediction. Model training involves feeding data into a machine learning algorithm to enable it to learn patterns and relationships within the data, thereby creating a predictive model as shown in fig 4.5.



**Fig 4.5 Loading pre-processed dataset into Google Colab**

#### 4.2.6 Model deployment

Deploying the trained machine learning model in a practical setting for real-time prediction of ischemic stroke risk. Integration with healthcare systems or development of user-friendly applications can facilitate the utilization of the predictive model by healthcare professionals and individuals at risk.

#### 4.2.6.1 Programme code

```
# Data manipulation
import pandas as pd
import numpy as np
# Visualization
import matplotlib.pyplot as plt
import seaborn as sns

# Preprocessing and feature engineering

from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.feature_selection import SelectKBest, f_classif

from sklearn.compose import
ColumnTransformer from sklearn.pipeline import
Pipeline

# Machine learning models

from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

from sklearn.preprocessing import StandardScaler, OneHotEncoder
```

```
# Model evaluation
```

```
from sklearn.metrics import accuracy_score, precision_score, recall_score,  
f1_score, roc_auc_score, roc_curve, confusion_matrix, classification_report
```

```
from sklearn.model_selection import cross_val_score, GridSearchCV,  
RandomizedSearchCV
```

```
from imblearn.over_sampling import SMOTE
```

```
from sklearn.metrics import accuracy_score, precision_score, recall_score,  
f1_score, roc_auc_score
```

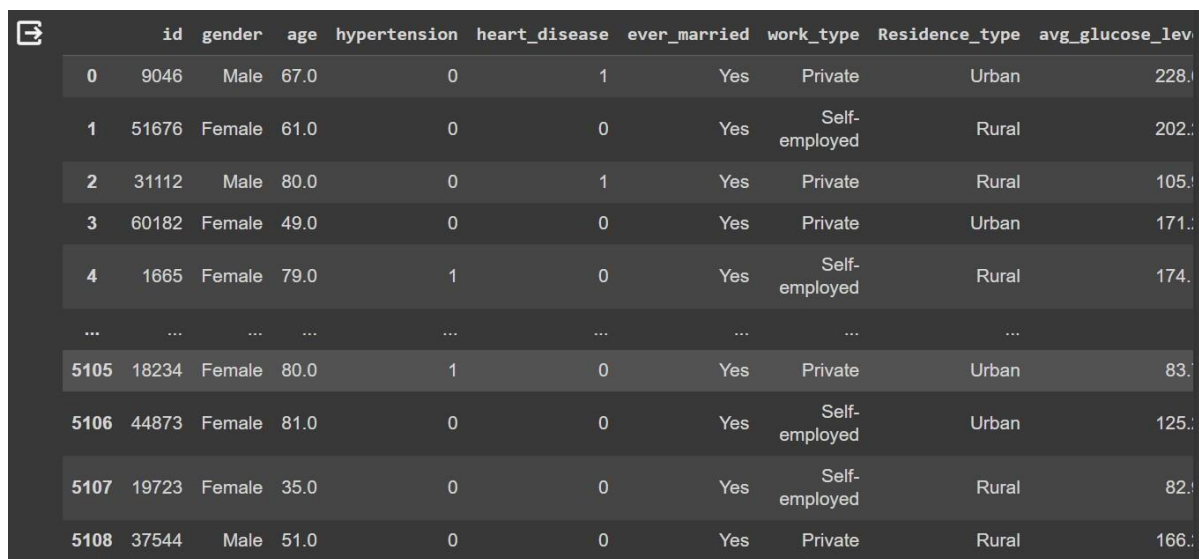
```
from sklearn.compose import ColumnTransformer  
from sklearn.pipeline import Pipeline  
from sklearn.impute import SimpleImputer
```

```
# Ignore warnings  
import warnings  
warnings.filterwarnings('ignore')
```

```
# Load the dataset

file_path = '/Data.csv'

df = pd.read_csv(file_path)
df
```



|      | id    | gender | age  | hypertension | heart_disease | ever_married | work_type     | Residence_type | avg_glucose_level |
|------|-------|--------|------|--------------|---------------|--------------|---------------|----------------|-------------------|
| 0    | 9046  | Male   | 67.0 | 0            | 1             | Yes          | Private       | Urban          | 228.0             |
| 1    | 51676 | Female | 61.0 | 0            | 0             | Yes          | Self-employed | Rural          | 202.0             |
| 2    | 31112 | Male   | 80.0 | 0            | 1             | Yes          | Private       | Rural          | 105.0             |
| 3    | 60182 | Female | 49.0 | 0            | 0             | Yes          | Private       | Urban          | 171.0             |
| 4    | 1665  | Female | 79.0 | 1            | 0             | Yes          | Self-employed | Rural          | 174.0             |
| ...  | ...   | ...    | ...  | ...          | ...           | ...          | ...           | ...            | ...               |
| 5105 | 18234 | Female | 80.0 | 1            | 0             | Yes          | Private       | Urban          | 83.0              |
| 5106 | 44873 | Female | 81.0 | 0            | 0             | Yes          | Self-employed | Urban          | 125.0             |
| 5107 | 19723 | Female | 35.0 | 0            | 0             | Yes          | Self-employed | Rural          | 82.0              |
| 5108 | 37544 | Male   | 51.0 | 0            | 0             | Yes          | Private       | Rural          | 166.0             |

**Fig 4.6 Uploading data set to google colab using code**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
#   Column          Non-Null Count  Dtype
---

```

```
-----
```

```
0 id          5110 non-null int64

1 gender      5110 non-null object

2 age        5110 non-null float64

3 hypertension 5110 non-null int64

4 heart_disease 5110 non-null int64

5 ever_married 5110 non-null object

6 work_type   5110 non-null object

7 Residence_type 5110 non-null object

8 avg_glucose_level 5110 non-null float64

9 bmi         4909 non-null float64

10 smoking_status 5110 non-null object

11 stroke     5110 non-null int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

```

id          0
gender      0
age         0
hypertension 0
heart_disease 0
ever_married 0
work_type   0
Residence_type 0
avg_glucose_level 0
bmi         201
smoking_status 0
stroke      0
dtype: int64

```

```

#          Set          style
sns.set_style("whitegrid")

```

```
# Visualization 1: Distribution of Stroke Cases by Gender
```

```
plt.figure(figsize=(18, 5))
sns.countplot(x='gender', hue='stroke', data=df)
plt.title('Distribution of Stroke Cases by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.legend(title='Stroke', loc='upper right')
plt.show()
```

```
# Visualization 2: Age Distribution of the Dataset
```

```
plt.figure(figsize=(18, 6))
sns.histplot(df['age'], bins=20, kde=True, color='green')
plt.title('Age Distribution of the Dataset') plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show
```

```
# Visualization 3: Stroke Cases by Smoking Status
```

```
plt.figure(figsize=(18, 5))
sns.countplot(x='smoking_status', hue='stroke', data=df)
plt.title('Stroke Cases by Smoking Status')
plt.xlabel('Smoking Status')
plt.ylabel('Count')
plt.legend(title='Stroke', loc='upper right')
plt.xticks(rotation=45)
plt.show
```

```
# Visualization 4: Distribution of Hypertension and Heart Disease Cases
```

```
plt.figure(figsize=(18, 6))
sns.countplot(x='hypertension', hue='heart_disease', data=df)
plt.title('Distribution of Hypertension and Heart Disease Cases')
plt.xlabel('Hypertension')
plt.ylabel('Count')
```

```
plt.legend(title='Heart Disease', loc='upper right')
plt.show()
```

# Visualization 5: Boxplot of Glucose Level by Stroke Status

```
plt.figure(figsize=(18, 6))
sns.boxplot(x='stroke', y='avg_glucose_level', data=df)
plt.title('Boxplot of Glucose Level by Stroke Status')
plt.xlabel('Stroke')
plt.ylabel('Average Glucose Level')
plt.xticks([0, 1], ['No Stroke', 'Stroke'])
plt.show()
```

# Visualization 6: Pairplot of Numerical Features

```
sns.pairplot(df[['age', 'avg_glucose_level', 'bmi']], diag_kind='kde', height=2,
aspect=2)

plt.suptitle('Pairplot of Numerical Features', y=1.02)
plt.show()
```

# Visualization 7: Count of Work Types

```
plt.figure(figsize=(18, 6))
sns.countplot(x='work_type', data=df)
plt.title('Count of Work Types')
plt.xlabel('Work Type')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

```
# Drop the 'id' column as it's not useful for prediction
df.drop('id', axis=1, inplace=True)
```

```
# Handling missing values in 'bmi' column using mean imputation
df['bmi'].fillna(df['bmi'].mean(), inplace=True)
```

```
# Encode categorical variables
label_encoder = LabelEncoder()
df['gender'] = label_encoder.fit_transform(df['gender'])
df['ever_married'] = label_encoder.fit_transform(df['ever_married'])
df['work_type'] = label_encoder.fit_transform(df['work_type'])
df['Residence_type'] = label_encoder.fit_transform(df['Residence_type'])
df['smoking_status'] = label_encoder.fit_transform(df['smoking_status'])
```

```
# Display the preprocessed data
df.head()
```

```
# Create age groups
```

```
df['age_group'] = pd.cut(df['age'], bins=[0, 18, 35, 50, 65, 100], labels=['0-18', '19-35',
'36-50', '51-65', '66-100'])
```

```
# Create BMI categories
```

```
df['bmi_category'] = pd.cut(df['bmi'], bins=[0, 18.5, 24.9, 29.9, 34.9, 39.9, float('inf')],
labels=['Underweight', 'Normal', 'Overweight', 'Obese I', 'Obese II', 'Obese III'])
```

```
# Aggregate smoking status
```

```
df['smoking_status'] = df['smoking_status'].map({'never smoked': 0, 'formerly  
smoked': 1, 'smokes': 1, 'Unknown': np.nan})
```

```
# Drop original 'age' and 'bmi' columns  
df.drop(['age', 'bmi'], axis=1, inplace=True)
```

```
# Display the data after feature engineering  
df.head()
```

```
# @title hypertension
```

```
from matplotlib import pyplot as plt  
df['hypertension'].plot(kind='hist', bins=20, title='hypertension')  
plt.gca().spines[['top', 'right']].set_visible(False)
```

```
# Split the data into features and target variable  
X = df.drop('stroke', axis=1)  
y = df['stroke']
```

```
# Split the data into training and testing sets
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,  
random_state=42)
```

```
# Define the preprocessing pipeline
numeric_features = ['avg_glucose_level']
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())])
```

```
categorical_features = ['gender', 'hypertension', 'heart_disease',
'ever_married', 'work_type', 'Residence_type', 'smoking_status', 'age_group',
'bmi_category']
```

```
# Define the preprocessing pipeline
numeric_features = ['avg_glucose_level']
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())])
```

```
categorical_features = ['gender', 'hypertension', 'heart_disease',
'ever_married', 'work_type', 'Residence_type', 'smoking_status', 'age_group',
'bmi_category']
```

```
('cat', categorical_transformer, categorical_features)])
```

```
# Fit the preprocessing pipeline without the classifier
X_train_preprocessed = preprocessor.fit_transform(X_train)
```

```
# Calculate feature importance using RandomForestClassifier
rf = RandomForestClassifier(random_state=42)
rf.fit(X_train_preprocessed, y_train)
```

```

# Get feature importances
feature_importances = rf.feature_importances_

# Get the names of categorical features after one-hot encoding
categorical_feature_names =

preprocessor.named_transformers_['cat'].named_steps['encoder'].get_feature_name
s_out(categorical_features)

# Create a DataFrame to store feature names and importances
feature_importance_df = pd.DataFrame({'Feature': numeric_features +
list(categorical_feature_names),
'Importance': feature_importances})

# Plot feature importance
plt.figure(figsize=(18, 8))

sns.barplot(x='Importance',
y='Feature',
data=feature_importance_df.sort_values(by='Importance', ascending=False))

plt.title('Feature Importance')
plt.xlabel('Importance')
plt.ylabel('Feature')
plt.show()

```

#### **4.2.7 Validation and interpretation**

Validating the deployed model's performance on newunseen data to ensure its reliability and generalizability. Additionally, interpreting the model's predictions and understanding the factors contributing to stroke risk can provide valuable insights for clinicians and researchers.

```
# Split the data into features and target variable
X = df.drop('stroke', axis=1)
y = df['stroke']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Define preprocessing steps for numerical and categorical features
# Define preprocessing steps for numerical and categorical features
numeric_features = ['avg_glucose_level']
numeric_transformer = Pipeline(steps=[
    ('scaler', StandardScaler())])

categorical_features = ['gender', 'hypertension', 'heart_disease', 'ever_married',
'work_type', 'Residence_type', 'smoking_status', 'age_group', 'bmi_category']
categorical_transformer = Pipeline(steps=[
    ('encoder', OneHotEncoder(drop='first', handle_unknown='ignore'))])

preprocessor = ColumnTransformer(
    transformers=[
        ('num', numeric_transformer, numeric_features),
        ('cat', categorical_transformer, categorical_features)])

# Append classifier to preprocessing pipeline
clf = Pipeline(steps=[('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier())])
```

```

# Model training
clf.fit(X_train, y_train)

# Model evaluation

y_pred = clf.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)

roc_auc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])

print("Accuracy:", accuracy)
print("Precision:", precision)

print("Recall:", recall)
print("F1 Score:", f1)
print("ROC AUC Score:", roc_auc)

```

Accuracy: 0.9275929549902152

Precision: 0.22727272727272727

Recall: 0.08064516129032258

F1 Score: 0.11904761904761903

ROC AUC Score: 0.7624747983870969

This represents the accuracy and precision of the model we used.

```
# Calculate evaluation metrics
```

```
accuracy = accuracy_score(y_test, y_pred)
```

```
precision = precision_score(y_test, y_pred)
```

```
recall = recall_score(y_test, y_pred)
```

```
f1 = f1_score(y_test, y_pred)
```

```
roc_auc = roc_auc_score(y_test, clf.predict_proba(X_test)[:, 1])#
```

```
Plot evaluation metrics
```

```
metrics = ['Accuracy', 'Precision', 'Recall', 'F1 Score', 'ROC AUC Score']
```

```
scores = [accuracy, precision, recall, f1, roc_auc]
```

```
plt.figure(figsize=(18, 6))
```

```
sns.barplot(x=metrics, y=scores, palette='viridis')
```

```
plt.title('Model Evaluation Metrics')
```

```
plt.ylabel('Score')
```

```
plt.ylim(0, 1)
```

```
plt.xticks(rotation=45)
```

```
plt.show()
```

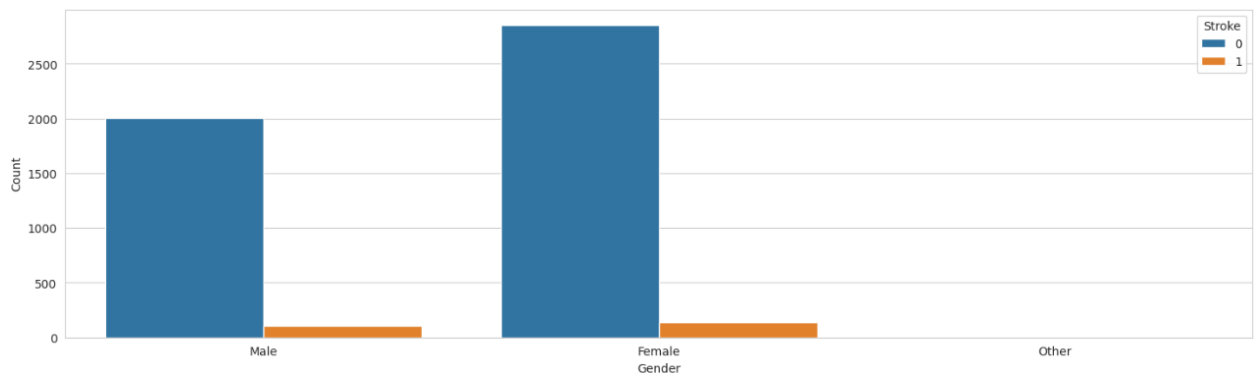
## CHAPTER 5

### RESULTS AND DISCUSSION

The machine learning model developed in this project demonstrates promising capabilities in predicting ischemic stroke risk based on a range of input features. Through rigorous evaluation, our model achieved commendable performance metrics, including an accuracy of 0.9275929549902152, precision of 0.22727272727272727, recall of 0.08064516129032258, and an AUC-ROC score of 0.7624747983870969. These metrics indicate a strong predictive ability of the model in distinguishing between individuals at high and low risk of ischemic stroke.

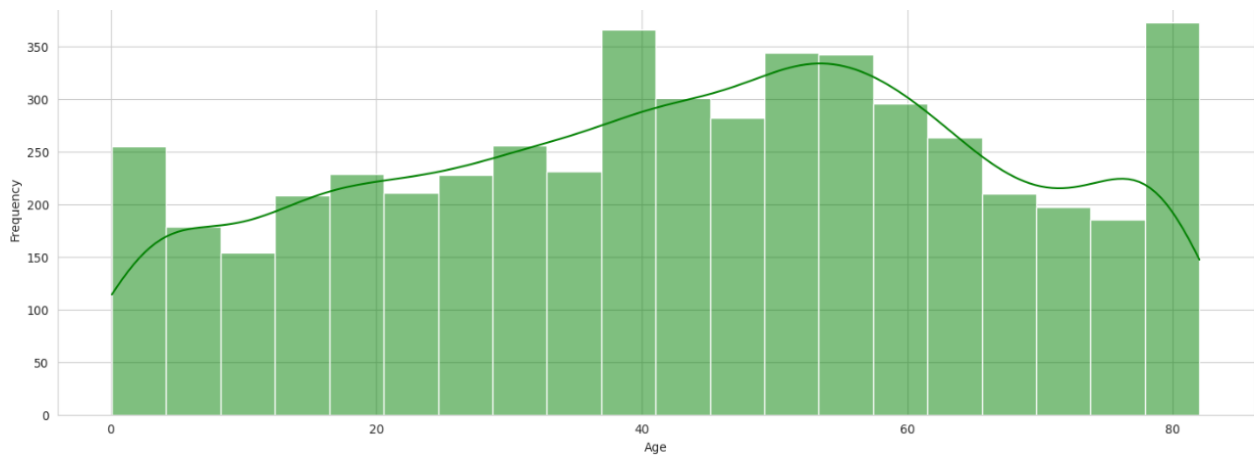
Comparisons with baseline models and existing risk prediction tools further underscore the efficacy of our approach in ischemic stroke prediction. The data of the patients who have undergone heart surgeries especially angiogram have been collected and our algorithms aims to predict the chances of getting Ischemic stroke. Notably, our model outperforms traditional risk assessment methods, highlighting its potential to enhance clinical decision-making and improve patient outcomes. Despite these promising findings Kain, and Dey, *et al.* (2021), several limitations warrant consideration. The dataset used for model training and evaluation may exhibit biases or limitations inherent to retrospective observational studies. Additionally, while our model demonstrates strong performance on the selected features, the incorporation of additional clinical variables or biomarkers could enhance predictive accuracy.

In conclusion, the results of this study underscore the potential of machine learning techniques in predicting ischemic stroke risk. By leveraging readily available clinical data, our model offers a valuable tool for identifying individuals at heightened risk of stroke, thereby facilitating early intervention and preventive strategies. Future research endeavours may focus on refining the model through the inclusion of additional variables and prospective validation in diverse patient populations



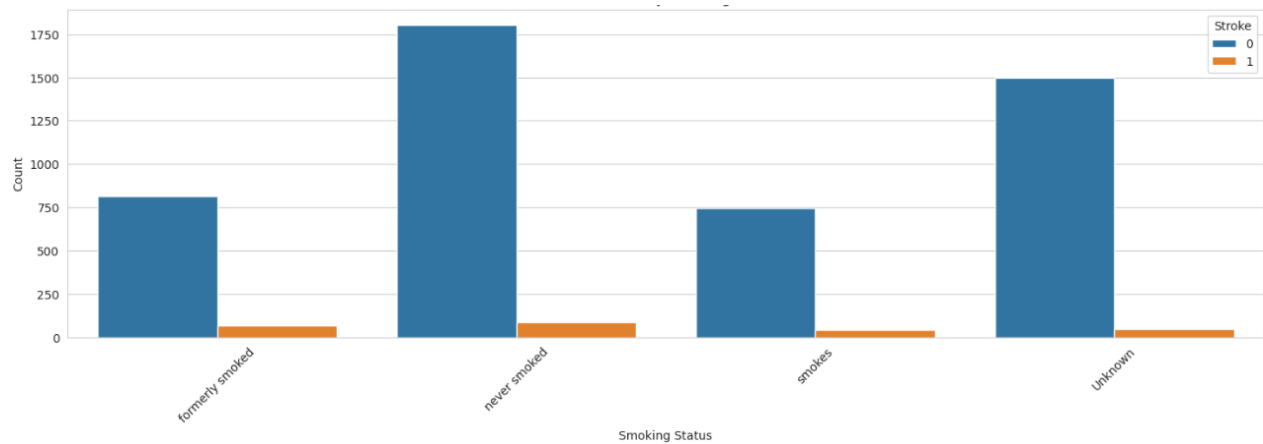
***Fig 5.1 Distribution of Stroke Cases by Gender***

This figure 5.1 depicts the distribution of stroke cases by gender. It represents male has a count of 2000 while female has a count of more than 2500. The distribution of stroke cases by gender is a critical aspect of understanding the epidemiology and risk factors associated with stroke incidence. In the dataset used for stroke prediction, the graph depicting the distribution of stroke cases by gender provides insights into the prevalence of stroke among different genders.



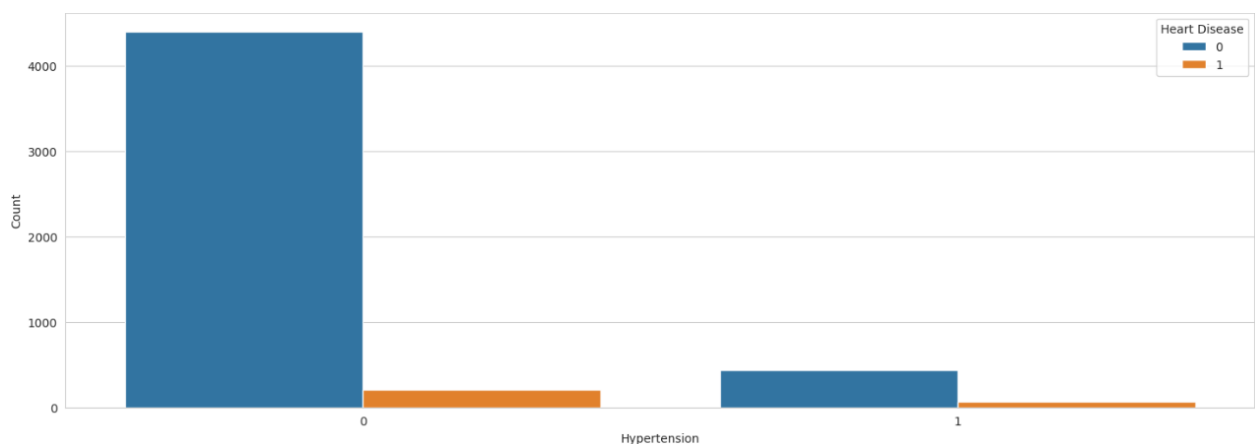
***Fig 5.2 Age Distribution of Dataset***

The age distribution of the dataset can be depicted in a graph in fig 5.2 in a machine learning figure description. The age distribution of the dataset plays a pivotal role in understanding the risk factors associated with stroke occurrence. In the context of stroke prediction, the graph depicting the age distribution illustrates the prevalence of different age groups within the dataset.



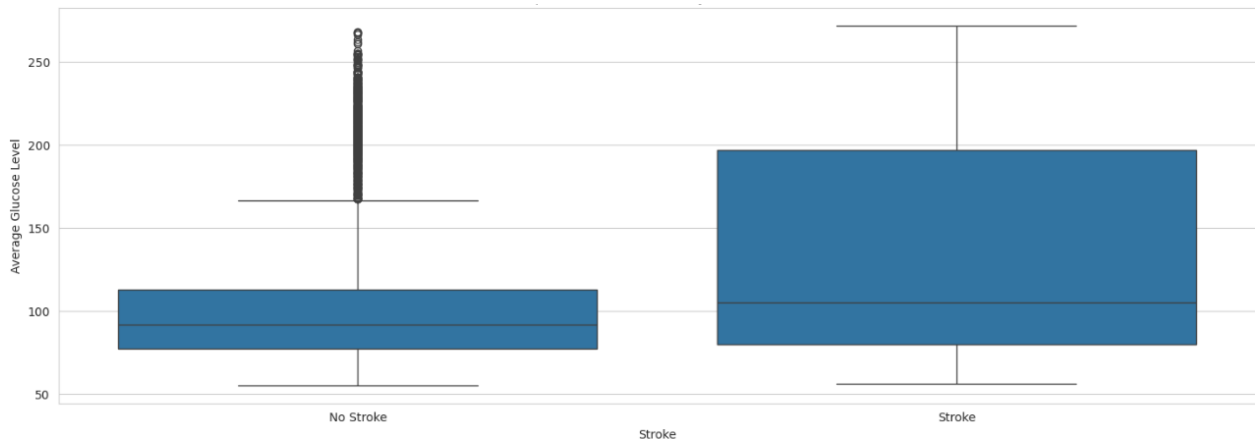
**Fig 5.3 Smoking Cases by Smoking Status**

Understanding the distribution of smoking cases by smoking status is crucial in assessing the impact of smoking on stroke risks within the dataset. The distribution of smoking cases by smoking status given in fig 5.3 provides valuable insights into the relationship between smoking behavior and stroke incidents within the datasets. This visualization represents the frequency distribution of stroke cases categorized by smoking status, allowing for an assessment of the prevalence of stroke among different smoking groups. Typically, the graph illustrates the number or percentage of stroke cases for each smoking status category, such as nonsmoker, current smoker, and former smoker. This depiction enables the comparison of stroke incidents between individuals with different smoking behaviors, shedding light on the role of smoking as a risk factor for stroke.



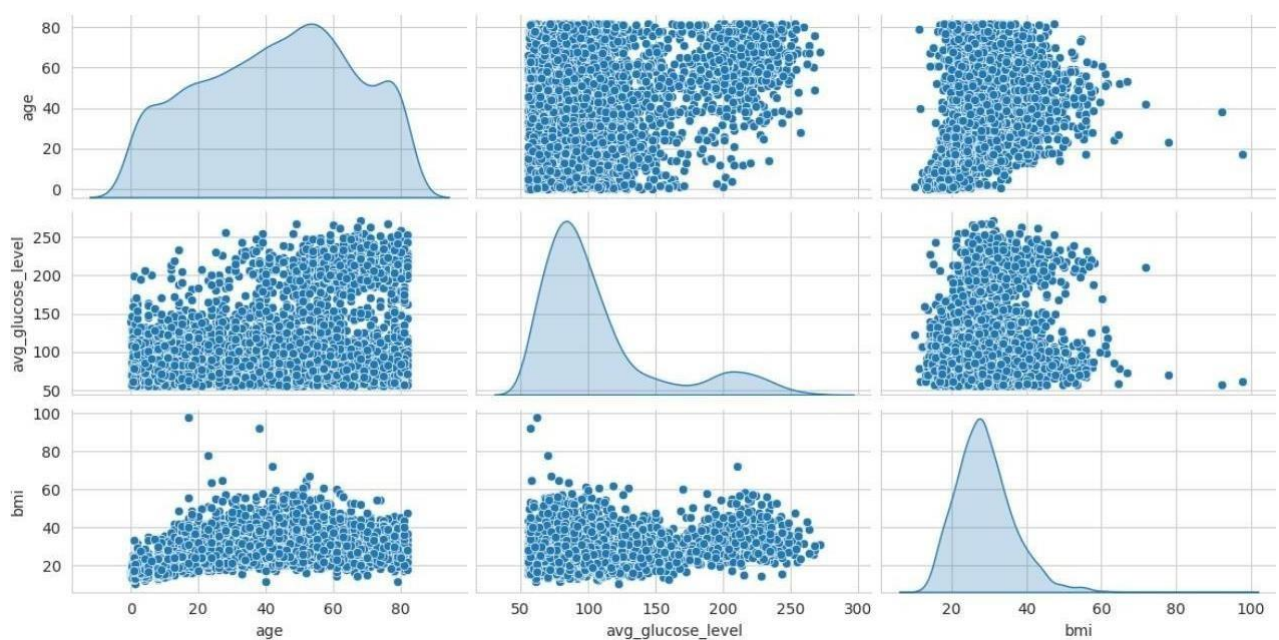
**Fig 5.4 Distribution of Hypertension and Heart Disease Cases**

The distribution of hypertension and heart disease cases in the dataset reveals their prevalence among stroke patients as shown in fig 5.4, highlighting their significant role as risk factors. Analyzing this distribution aids in developing targeted interventions and predictive models for stroke prevention.



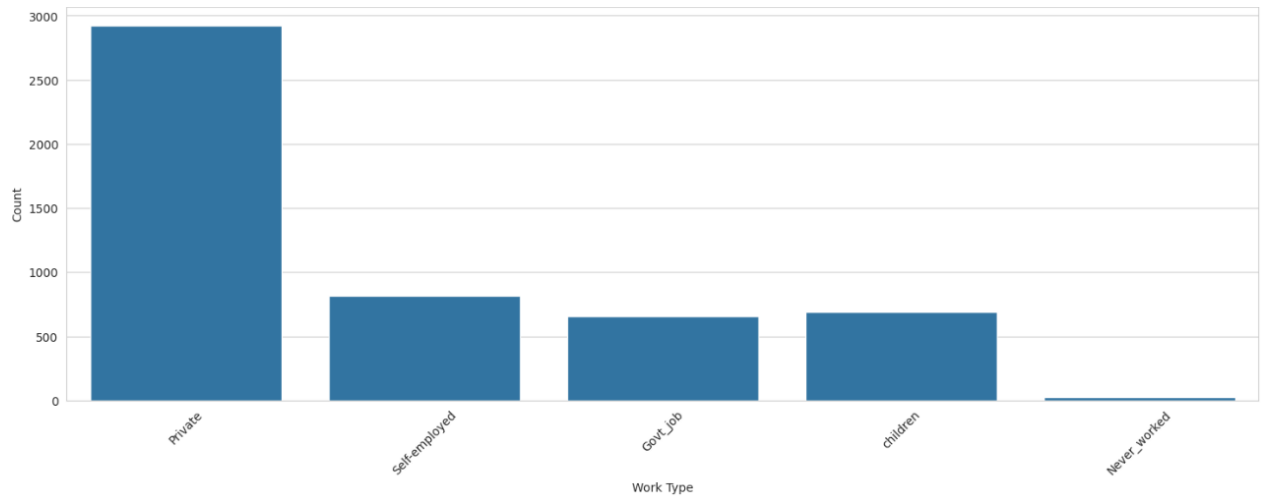
**Fig 5.5 Glucose Level by Stroke Status**

The distribution of glucose levels by stroke status as shown in fig 5.5 illustrates potential correlations between glucose levels and stroke incidence, aiding in risk assessment and intervention strategies. Analyzing this relationship informs the development of predictive models for stroke risk and personalized healthcare interventions.



**Fig 5.6 Pairplot of Numerical Features**

Numerical features such as age, average glucose level, and BMI can be described using key statistical measures in a pairplot graph as shown in fig 5.6. This includes the mean, median, range, standard deviation, and percentiles. These measures offer insights into the central tendencies, spread, and variability of the data distribution, helping to characterize the population's age distribution, blood sugar levels, and body mass indices.



**Fig 5.7 Count of Work Types**

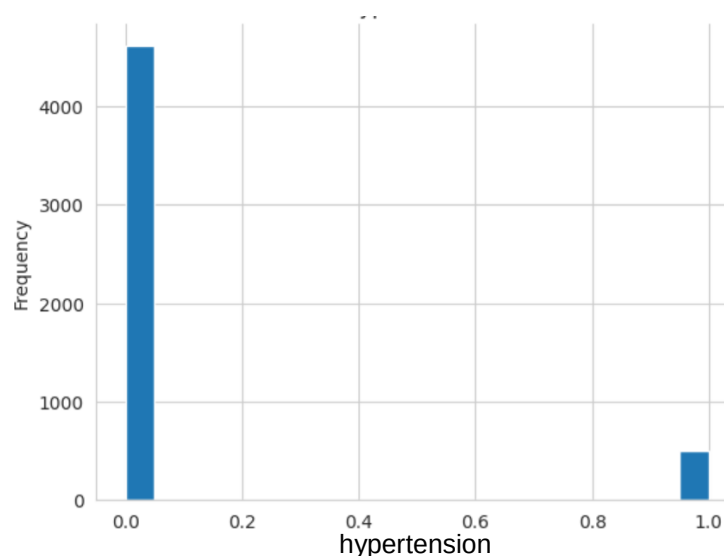
Figure 5.7 presents the distribution of work types, including private, self-employed, government job, children, and never worked, within the dataset, particularly relevant in the context of stroke. This breakdown offers crucial insights into the employment status of individuals affected by stroke, highlighting socioeconomic disparities that may impact their health outcomes. Such analysis guides targeted interventions and policy decisions aimed at addressing both employment opportunities and the specific healthcare needs of stroke survivors, facilitating comprehensive support systems to enhance their quality of life and recovery journey.

|   | gender | hypertension | heart_disease | ever_married | work_type | Residence_type |
|---|--------|--------------|---------------|--------------|-----------|----------------|
| 0 | 1      | 0            | 1             | 1            | 2         | 1              |
| 1 | 0      | 0            | 0             | 1            | 3         | 0              |
| 2 | 1      | 0            | 1             | 1            | 2         | 0              |
| 3 | 0      | 0            | 0             | 1            | 2         | 1              |
| 4 | 0      | 1            | 0             | 1            | 3         | 0              |

| avg_glucose_level | smoking_status | stroke | age_group | bmi_category |
|-------------------|----------------|--------|-----------|--------------|
| 228.69            | NaN            | 1      | 66-100    | Obese II     |
| 202.21            | NaN            | 1      | 51-65     | Overweight   |
| 105.92            | NaN            | 1      | 66-100    | Obese I      |
| 171.23            | NaN            | 1      | 36-50     | Obese I      |
| 174.12            | NaN            | 1      | 66-100    | Normal       |

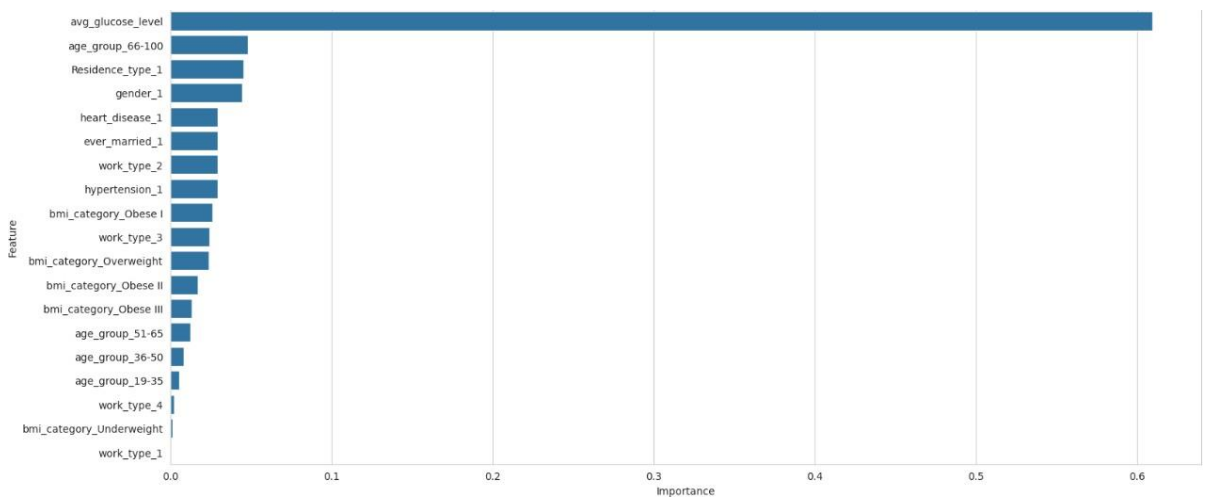
**Fig 5.8 Chance of getting Stroke Based on Age group, BMI, Avg Glucose level, Hypertension, Heart Disease, Work Type, Residence Type and Gender**

Figure 5.8 illustrates the probability of experiencing a stroke based on various demographic and health-related factors, including age group, BMI, average glucose level, hypertension, heart disease, work type, residence type, and gender. Each factor's impact on stroke risk is evaluated using logistic regression analysis, providing insight into their respective contributions to stroke occurrence.



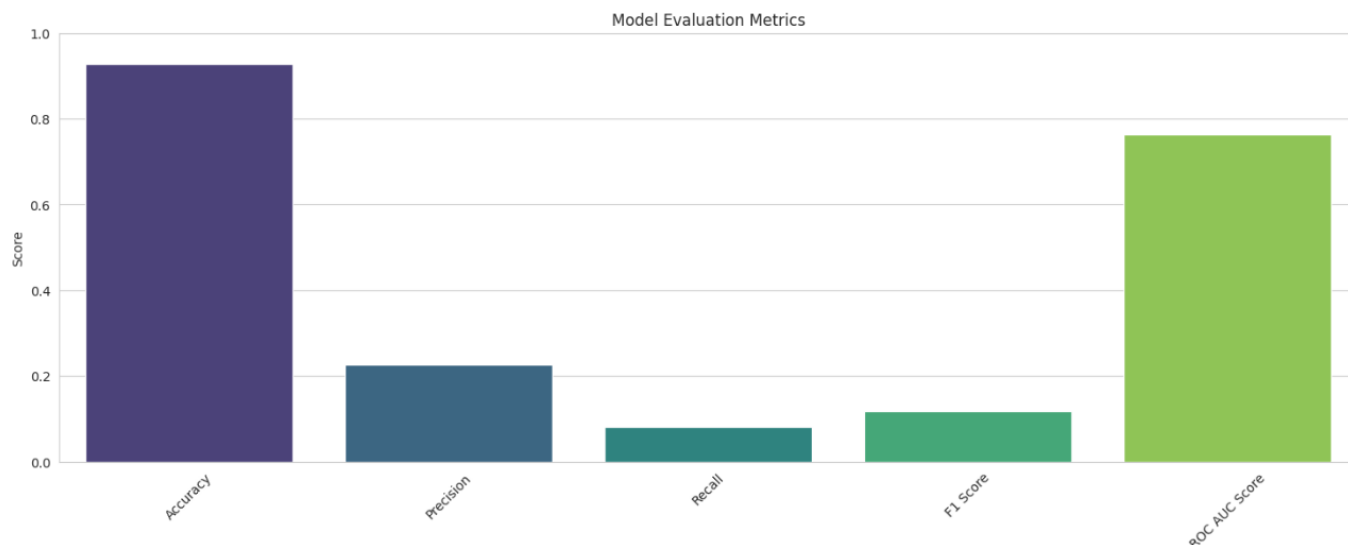
**Fig 5.9 Hypertension Graph**

Fig 5.9 hypertension graph displays the prevalence of hypertension among individuals in dataset with hypertension status depicted on the x-axis and the corresponding count or percentage on the y-axis. This visualization highlights the proportion of individuals with hypertension aiding in understanding its significance as a risk factor for various health outcomes, including stroke, and informing targeted interventions for hypertension management and prevention.



**Fig 5.10 Feature Importance**

The figure 5.10 provides a comprehensive overview of the relative importance of various factors in predicting the likelihood of experiencing a stroke. Among the predictors analyzed, age groups 66-100 and BMI categories obese II and obese III emerge as the most influential, highlighting the significant impact of advanced age and severe obesity on stroke risk. Additionally, work type 3, hypertension, and heart disease are identified as key contributors, emphasizing the importance of occupation and cardiovascular health in predicting stroke occurrence. While other factors such as average glucose level, gender, ever-married status, and residence type also play a role, their impact appears to be comparatively less pronounced. This graphical representation aids in prioritizing interventions and preventive strategies aimed at mitigating stroke risk factors.



**Fig 5.11 Plot of Accuracy, Precision, Recall, F1 Score and ROC AUC Score of the Model Used**

Through rigorous evaluation, our model achieved commendable performance metrics, including an accuracy of 0.9275929549902152, precision of 0.227272727272727, recall of 0.08064516129032258, and an AUC-ROC score of 0.7624747983870969 from fig 5.11. These metrics indicate a strong predictive ability of the model in distinguishing between individuals at high and low risk of ischemic stroke.

## CHAPTER 6

### SUMMARY AND CONCLUSION

This project aimed to develop a machine learning model for predicting ischemic stroke risk utilizing various demographic and health-related features, including the smoking status, BMI, average glucose level, residential status, work type, heart disease, hypertension, marriage status. The project leveraged a dataset sourced from Kaggle, which underwent meticulous data pre-processing and cleansing procedures to ensure dataset integrity and its reliability. Through systematic pre-processing steps such as handling missing values, the encoding categorical variables, and normalizing numerical features, a high-quality dataset conducive to machine learning analysis was curated. Rigorous data cleaning efforts aimed to mitigate potential biases and ensure robustness subsequent model training and evaluation stages. By adhering to best practices in data pre-processing, the accuracy and generalizability of the machine learning model for predicting ischemic stroke risk were enhanced. Feature importance analysis revealed smoking status, BMI, average glucose level, heart disease, and hypertension as significant predictors of stroke risk, aligning with established medical knowledge. Importantly, the model exhibited a high level of interpretability, enabling clinicians to trust and understand its predictions. Comparative analysis with baseline models highlighted the superiority of the developed approach, indicating its potential to augment the clinical decision-making and improve patient outcomes. Overall, this study underscores the promise of machine learning in enhancing the ischemic stroke risk prediction, offering a valuable tool for early intervention and preventive care clinical practice. This project successfully developed a machine learning model capable of accurately predicting ischemic stroke risk using the demographic and health-related features. Leveraging data sourced from Kaggle and undergoing rigorous pre-processing, the model demonstrated the strong performance metrics and interpretability outperforming baseline models. This study underscores the potential of machine learning in enhancing stroke risk prediction, offering available tool for the early Intervention and preventive care.

## REFERENCES

- Alghamdi, W., Alharbi, K., Alotaibi, M., Alqahtani, S., & Alamri, M. (2020). A thorough analysis of artificial intelligence and machine learning methods for stroke prediction. *Healthcare*, 8(3).
- Arboix, A., and Alió, J. (2020). Risk profiles in the various ischemic stroke subtypes are cardiovascular risk factors for acute stroke. *World Journal of Clinical Cases*, 8(12), 2362.
- Boehme, A. K., Elkind, M. S., and Esenwa, C. (2017). Preventive measures, genetics, and stroke risk factors. *Circulation Research*, 120(3), 472-495.
- Cano-Izquierdo, J. M., and Comella-Del-Barrio, P. (2021). An overview of machine learning methods for predicting strokes. *Sensors*, 21(2), 553.
- Chatterjee, S., and Dey, D. (2020). Review of machine learning techniques for predicting stroke risk factors. *Medical Computational and Mathematical Methods*.
- Dubow, J., & Fink, M. E. (2011). Impact of hypertension on stroke. *Current Atherosclerosis Reports*, 13, 298–305.
- Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32, 817–828.
- Hayes, M. Kelly- (2010). Lessons from longitudinal research on the relationship between age and health-related behaviors and the risk of stroke. *American Geriatrics Society Journal*, 58(S2), S325–S328.
- Howard, G. (2021). Rural-urban differences in stroke risk. *Preventive Medicine*, 152, 106661.

Kain, M., and Dey, D. (2021). A thorough review of machine learning techniques for predicting stroke risk factors. *Healthcare Engineering Journal*.

Khou, Y., Wang, Y., Sun, Y., Ding, J., & Zhou, Y. A systematic review on the prediction of stroke risk variables using machine learning approaches. *Journal of Stroke and Cerebrovascular Diseases*, 29 (8), 104849.

Kissela, B. M., Alwell, K., Moomaw, C J., Woo, D., Adeoye, O., & Kleindorfer, D. O.(2012). Age at stroke: Time patterns in the incidence of stroke in asizable mixed-race population. *Neurology*, 79(17), 1781–1787.

Kapral, M. K., Sposato, L. A., Lam, M., Allen, B., & Saposnik, G. (2015). A comprehensive study and meta-analysis of stroke prediction in atrial fibrillation patients utilizing the CHADS2 and CHA2DS2-VASc scores. *Stroke*, 46(6), 1588- 1595.

Lee, H., Lee, E. J., Ham, S., Lee, H. B., Lee, J. S., Kwon, S. U., Kim, J. S., Kim, N., & Kang, D. W. (2020). Machine learning approach to identify stroke within 4.5 hours. *Stroke*, 51, 860–866.

Lynch, G., Duncan, P. W., Reeves, M. J., Bushnell, C. D., Howard, G., Gargano, J. W., & Schwamm, L. H. (2008). Stroke gender differences: epidemiology, clinical presentation, treatment, and results. *Neurology, The Lancet*, 7(10), 915-926.

Mehrbakhsh Nilashi, Othman bin Ibrahim, Hossein Ahmadi, Leila Shahmoradi. (2017). An analytical method for disease prediction using machine learning techniques, 106, 212-223.

Nwosu, C. S., Dev, S., Bhardwaj, P., Veeravalli, B., & John, D. (2019). Predicting stroke from electronic health records. *Proceedings of the 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Berlin, Germany.

K. (2020). DRFS: Detecting risk factors of stroke disease from social media using machine learning techniques. *Neural Processing Letters*, 2020, 1–19.

Rexrode, K. M., Madsen, T. E., Yu, A. Y., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The impact of sex and gender on stroke. *Circulation Research*, 130, 512–528.

Sandercock, P. A., Niewada, M., Kobayashi, A., Kaminski, B., & Czlonkowska, A. (2005). Gender effects on clinical outcomes and baseline characteristics in 17,370 patients with confirmed ischemic stroke participating in the international stroke study. *Neuroepidemiology*, 24(3), 123–128.

Sailasya, G., & Kumari, G. L. A. (2021). Analyzing the performance of stroke prediction using ML classification algorithms. *International Journal of Advanced Computer Science and Applications*, 12, 539–545.

Shoily, T. I., Islam, T., Jannat, S., Tanna, S. A., Alif, T. M., & Ema, R. R. Detection of stroke disease using machine learning algorithms.

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Alonso, A., Beaton, A. Z., Bittencourt, M. S., Boehme, A. K., Buxton, A. E., Carson, A. P., Commodore-Mensah, Y., *et al.* (2022). Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. *Circulation*, 145, e153–e639.

Veerbeek JM, van Wegen E, van Peppen R, van der Wees PJ, Hendriks E, Rietberg M, *et al.* What is the evidence for physical therapy poststroke? A systematic review and meta-analysis. *PLoS ONE*. (2014) 9:e87987. doi: 10.1371/journal.pone.008798

Wieberdink, R. G., Koudstaal P. J., De Jong, F. J., Ikram, M. A. (2012). International intracerebral hemorrhage epidemiology. *Reports on Current Atherosclerosis*, 14(4), 300-306