

School of Computing, Engineering and Digital Technologies
Department of Computing and Games
Teesside University
Middlesbrough TS1 3BA

Brain Stroke Risk Prediction; Strategies for improving early brain stroke prediction and diagnosis using machine learning

An analysis, design and implementation report for the development of an application to Early
Brain Stroke Risk Prediction

Submitted in partial requirements for the degree of *MSc. Data Science*

Date: 10 January 2024

Athira Reghunath

W9561613

Supervisor: Glen Hopkinson

Abstract:

In the world of healthcare, think of machine learning as a superhero, a powerful tool that acts quickly and reliably to predict outcomes. It's like having a personal guide for each patient, customizing the care they receive. Now, let's talk about something crucial: strokes, a major cause of adult deaths and disabilities worldwide. Most strokes happen when an unexpected obstacle pops up in the pathways of both the brain and the heart. Imagine if we could spot various warning signs early on it is like putting on a superhero cape to minimize the impact of a stroke before it becomes a big problem.

This research aims to create a machine-learning methodology for the early prediction of stroke diseases by utilizing features like gender, hypertension, body mass index, heart disease, average glucose level, smoking status, marital status, employment type, and residential area. The study involves training five distinct classifiers, namely Logistics Regression, Decision Tree Classifier, K-Nearest Neighbor Classifier, Random Forest, and Support Vector Machine, using these extensive attributes. Additionally, the proposed approach attained an accuracy level of 93.93%.

For simple assessment, an online application has been created to inform patients about their risk of stroke. The patient can enter their details through the app and submission will lead to the stroke risk diagnosis page. This application works with the help of the machine learning algorithms used during the data modelling.

When reviewing the literature on stroke prediction, numerous studies were identified across various databases, but only a few proved suitable for further investigation. In this research used distinct machine-learning techniques compared to previous studies to achieve satisfactory outcomes. In summary, Support Vector Machine (SVM) and Random Forests emerged as effective methods across all categories. This research underscores the significance of employing diverse machine-learning techniques in the realm of brain stroke prediction.

Acknowledgement

The dissertation titled “Brain Stroke Risk Prediction; Strategies for Improving Early Brain Stroke Prediction and Diagnosis Using Machine Learning” would not have been successfully finished without the constant assistance of those around me. I would like to convey my appreciation to my family for their steadfast support throughout the entire course, particularly for providing me with mental encouragement to complete this dissertation. Additionally, I want to express my admiration for the support of my friends.

Most importantly, I want to extend my thanks to my research supervisor for his invaluable guidance and support throughout the entire dissertation process. Without his assistance, this subject would not have evolved into a research project. Lastly, but equally significant, I want to express my appreciation to everyone for their assistance, advice, and contributions to this project.

Table of Contents

Abstract.....	1
Acknowledgement.....	2
Table of Contents.....	3
List of Figures.....	4
1. Introduction	5
1.1 Research Background.....	5
1.2 Research Objectives.....	5
1.3 Research Questions.....	6
1.4 Research Structure.....	7
2. Literature Review	8
2.1 Historical Developments in Brain Stroke Prediction.....	8
2.2 Review of the Existing Literature on Brain Stroke Prediction.....	10
2.3 Techniques that are Going to Use.....	10
3. Methodology/Proposed System	11
3.1 Data Collection	11
3.2 Data Preprocessing	13
3.3 Exploratory Data Analysis	13
3.4 Web Application.....	21
4. Experimentation	24
4.1 Predictive Machine Learning Models.....	24
4.1.1 Logistic Regression	25
4.1.2 Decision Tree.....	26
4.1.3 Random Forest.....	26
4.1.4 K-Nearest Neighbors	27
4.1.5 Standard Vector Machine	28
4.1 Comparisson Between Models	28
5. Conclusion and Recommendations.....	30
5.1 Conclusion	30
5.2 Recommendations	30
5.3 Limitations.....	31
5.3 Future Research	31
6. References.....	32
7. Appendices	34

List of Figures

Fig 1: Distribution of Stroke	14
Fig 2: Distribution of Age.....	14
Fig 3: Distribution of Average Glucose Level	15
Fig 4: Distribution of Histogram of Body Mass Index	15
Fig 5: Pair plot of Age, Average Glucose Level and BMI	16
Fig 6: Count plot for Gender, Hypertension and Heart Disease	17
Fig 7: Count plot for Marital Status, Work Type and Residence Type	17
Fig 8: Count plot for Smoking Status and Stroke	18
Fig 9: Gender, Hypertension, Heart Disease, Marital Status, Job Type, Residence Type, Smoking Status Distributions with and without Stroke	19
Fig 10: Correlation Matrix Heat Map	20
Fig 11: Correlation with Stroke	21
Fig 12: Stroke Risk Prediction Home Page	22
Fig 13: Stroke Risk Prediction Home with Values	22
Fig 14: No Stroke Risk Diagnosed Page	23
Fig 15: Stroke Risk Diagnosed Page	23
Fig 16: Confusion Matrix Logistic Regression	25
Fig 17: Confusion Matrix Decision Tree	26
Fig 18: Confusion Matrix Random Forest	27
Fig 19: Confusion Matrix KNN	27
Fig 20: Confusion Matrix SVM	28
Fig 21: Model Comparison Bar Chart	29

1. Introduction

1.1 Research Background

A cerebrovascular accident, commonly known as a stroke, occurs when there is a disruption or reduction in blood flow to specific regions of the brain. This interruption in the supply of nutrients and oxygen results in the commencement of cell death, marking it as a severe medical emergency. Swift and immediate treatment is imperative to mitigate additional harm to the impacted brain region and forestall potential complications in other bodily areas. The World Health Organization reports a global incidence of fifteen million stroke cases annually, with one individual succumbing to a stroke approximately every four to five minutes.

There exist two categories of stroke: ischemic and hemorrhagic. An ischemic stroke is predominantly triggered by the blockage of a blood artery, although there are additional, less frequent causes. On the other hand, a hemorrhagic stroke is the result of bleeding directly into the brain or the space between the brain's membranes. The prevalence of this medical condition is increasing in developing nations such as China, where the occurrence of strokes is notably elevated. In the United States, there is an emerging pattern of enduring disability attributed to strokes, and the mortality rate from strokes in these countries has surged to a level ten times higher than it was in the last five decades.

The anticipated increase in the global population is expected to contribute to a rise in both the mortality rate and the number of individuals affected by this disease. However, early detection and timely treatment can play a crucial role in mitigating the mortality rate. Traditional methods, such as the Cox proportional hazard model, encounter limitations in effectively predicting strokes, particularly when dealing with high-dimensional data. In such situations, machine learning becomes a pivotal tool to enhance accuracy and efficiency in stroke prediction at a more cost-effective rate. Over the years, various machine learning classifiers like Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression have been applied in the medical field. These classifiers leverage patterns within large, imbalanced datasets to conduct precise analyses and provide reliable predictions.

This research showcases the efficacy of stroke prediction by employing strategies to balance data, utilizing machine learning algorithms, and incorporating diverse risk factors within an imbalanced dataset. The findings reveal that adopting this methodology achieves the utmost accuracy in predicting strokes.

1.2 Research Objectives

The research objectives for stroke risk prediction may include:

1. Development of Predictive Models:
 - Create and evaluate machine learning models to predict the risk of stroke.

- Explore the effectiveness of different algorithms such as Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression in stroke risk prediction.
2. Data Analysis and Feature Selection:
 - Conduct a comprehensive analysis of relevant data, identifying key risk factors associated with stroke.
 - Implement feature selection techniques to identify the most influential variables for accurate prediction.
 3. Comparison with Traditional Methods:
 - Compare the performance of machine learning models with traditional methods, such as the Cox proportional hazard model, in stroke risk prediction.
 - Assess the advantages and limitations of each approach to inform practical applications in clinical settings.
 4. Ethical Considerations:
 - Address ethical considerations in stroke risk prediction, including patient privacy and consent.
 - Evaluate and mitigate potential biases in the data and models to ensure fair and equitable predictions.
 5. Clinical Implementation and User-Friendly Interfaces:
 - Explore ways to integrate predictive models into clinical practice.
 - Develop user-friendly interfaces for healthcare professionals to easily interpret and utilize the predictions in real-world scenarios.
 6. Cost-Benefit Analysis:
 - Conduct a cost-benefit analysis of implementing stroke risk prediction models in healthcare settings.
 - Assess the economic feasibility and potential long-term benefits of integrating predictive tools into routine clinical practices.

1.3 Research Questions

- What machine learning models are most effective in early diagnosis of brain strokes, and how do they contribute to risk reduction?
- In what ways can machine learning methods be employed to assist physicians in the early detection and diagnosis of brain strokes, and what are the key advantages of such assistance?
- To what extent can the integration of machine learning in stroke prediction contribute to the reduction of mortality rates associated with brain strokes?
- How can machine learning models be optimized to enhance early diagnosis, particularly focusing on reducing false positives and false negatives in predicting brain strokes?

- What ethical considerations need to be addressed in the deployment of machine learning for stroke prediction, especially concerning patient privacy and informed consent?
- How can develop a simple risk prediction application for the easy diagnosis of stroke risk?

1.4 Research Structure

Below is a suggested research structure for a study on stroke prediction using machine learning:

1. Introduction:

- Background and context of stroke as a critical health issue.
- Importance of early diagnosis and prediction in reducing the risk and impact of strokes.
- Significance of implementing machine learning in stroke prediction.
- Clear statement of research objectives and questions.

2. Literature Review:

- Overview of existing literature on stroke epidemiology, risk factors, and consequences.
- Review of traditional methods for stroke prediction (e.g., cox proportional hazard model).
- Examination of prior studies using machine learning for stroke prediction.
- Identification of gaps in current research and the need for advanced predictive models.

3. Methodology/Proposed System:

- Overview of the envisioned machine learning framework designed for stroke prediction.
- Rationale behind selecting specific machine learning algorithms (e.g., “Random Forest, Support Vector Machine, Decision Tree, Logistic Regression”)
- Inclusion of diverse risk factors in shaping the predictive model.
- Implementation of data balancing methodologies to address imbalances in datasets.
- Ethical implications and the incorporation of user-friendly interfaces for risk assessment.

4. Experimentation:

- Data collection: Description of the dataset used, including details on the source, size, and characteristics.
- Preprocessing: Handling missing data, normalization, and feature selection.
- Implementation: Detailed steps for training and testing machine learning models.
- Evaluation metrics: Clarification of the standards utilized to evaluate the effectiveness of the models.
- Results and analysis: Presentation and interpretation of experimental findings.

5. Conclusion:

- Summary of key findings from the experimentation.
- Discussion of the implications of the results within the realm of predicting strokes.

- Constraints of the investigation and areas for future research.
- Overall conclusion on the effectiveness and feasibility of using machine learning for stroke prediction.

6. References:

- Comprehensive list of all the sources cited throughout the research.

7. Appendices:

- Appendices can consist of figures, raw data, computer programs etc

2. Literature Review

Stroke remains a major global health concern, necessitating effective risk prediction models to enable early detection and intervention. In recent times, there has been a growing trend among researchers to utilise machine learning methods to improve the precision and effectiveness of stroke risk assessment.

2.1 Historical Developments in Brain Stroke Prediction

Stroke, a major global cause of illness and death, has seen a rapid advancement in prediction techniques. Over the years, advancements in medical research, technology, and data analysis have shaped the landscape of brain stroke prediction. This literature review explores the historical developments that have paved the way for contemporary approaches to stroke risk prediction.

This study extensively reviewed over 20 research papers and articles to investigate the historical evolution within this field. This section provides concise summaries of some of these papers and their content.

The paper published in IEEE (2022) “A Machine Learning Approach to Detect the Brain Stroke Disease” by Akter et al. is concluding the paper as Accurately identifying the risk of brain stroke could substantially influence long-term mortality rates for individuals, irrespective of their social or cultural background. Achieving this goal necessitates early detection. Although machine learning has been used to predict brain strokes in a number of researches, this paper takes a similar tack while using a more creative tactic and using a larger dataset for model training. The used dataset includes observations from 5110 patients and includes 12 brain stroke-related variables. Three classifiers can be trained and tested on the dataset with greater adaptability when image processing techniques are used. In particular, the Random Forest classifier performs exceptionally well, with a far better accuracy of 95.30% when compared to other classifiers. Subsequent efforts will focus on applying a variety of feature selection techniques to generalise the model and strengthen its ability to handle datasets with large amounts of missing data, which will ultimately increase accuracy.

Ashrafuzzaman et al. (2022): “Prediction of Stroke Disease Using Deep CNN Based Approach’ (Journal of Advances in Information Technology):” Ashrafuzzaman, Saha, and Nur focus on the application of deep convolutional neural networks (CNNs) for stroke prediction. In order to determine the risk of having a stroke, this study looked at a number of individual characteristics. A healthcare

dataset that was downloaded from Kaggle was used for the inquiry. A CNN model was created to predict the likelihood that an individual will suffer a stroke, and various categorization methods were used to examine the dataset. The effectiveness of the model was then evaluated. According to the experimental results, the suggested model outperforms some of the current models and reaches a notable 95.5 percent accuracy. Early stroke diagnosis in patients may be possible using this paradigm. Moreover, the examination of several traits related to stroke revealed a recurring pattern among those attributes linked to an increased risk of stroke development.

Biswas et al. (2022): “A Comparative Analysis of Machine Learning Classifiers for Stroke Prediction” (Healthcare Analytics): Biswas and collaborators perform a comparative analysis of machine learning classifiers, aiming to identify the most effective algorithm for stroke prediction. This paper employs diverse machine learning classification methods to identify and forecast stroke outcomes. The proposed methodology consists of five main steps, which include loading stroke datasets, pre-processing data, hyperparameter tuning, classifier assessment, and cross-validation. As can be seen from the results, the Random Forest classifier and Support Vector Machine achieve the highest accuracy rates 99.85% and 99.99%, respectively. In addition, an intuitive mobile application and website have been created to improve the display of results. In a next study, deep learning and image processing methods will be used in an effort to get even more accurate outcomes. We believe that this research will advance the effectiveness of treatments for this illness.

Dritsas and Trigka (2022): “Stroke Risk Prediction with Machine Learning Techniques” (Sensors): Dritsas and Trigka explore various machine-learning techniques for stroke risk prediction. Their research probably explores how sensor data may be applied to make predictions that are more accurate, and their work took a risk but ended up with a successful outcome by analysing picture data from brain CT scans to evaluate how well deep learning models predict the likelihood of strokes.

JalajaJayalakshmi et al. (2021): “Analysis and Prediction of Stroke using Machine Learning Algorithms” (IEEE): Jalaja Jayalakshmi, Geetha, and Ijaz analyze and predict strokes using machine learning algorithms. Data mining techniques were utilised in this study to determine the risk factors linked to the start of strokes in patients. Four different machine learning classification techniques were used to evaluate the Stroke Prediction dataset. The AdaBoost and J48 classifiers performed better than the other two and showed similar likelihoods of predicting strokes among patients. Remarkably, 95.7% True Positive (TP) rate was attained by both classifiers. The analysis of the dataset also shows that those who identify as self-employed, married, and male had a higher risk of stroke.

Some of the major literature reviews are described above, when going through all these papers we can see that most of the studies happened after the 2020s. The reviewed literature highlights the diverse methodologies and approaches employed in stroke risk prediction. Researchers are investigating how to improve the precision, speed, and usability of stroke prediction models through the application of deep learning approaches, machine learning algorithms, and cloud deployment viewpoints. These predictive models are continuously improved by comparative analysis and performance reviews.

The integration of machine learning in stroke risk prediction is a rapidly evolving field. The studies reviewed demonstrate a broad spectrum of approaches, methodologies, and perspectives, indicating the multidimensional nature of research in stroke detection. This literature review sets the stage for a comprehensive master’s level project, emphasizing the potential for innovation and the improvement of stroke prediction models.

2.2 Review of the Existing Literature on Brain Stroke Prediction

This review critically examines existing literature on brain stroke prediction, encompassing diverse methodologies, data sources, and technological advancements. From a historical perspective, early efforts in stroke prediction were primarily clinical, relying on risk factor assessments. With the advent of neuroimaging technologies, studies incorporated CT and MRI scans to enhance diagnostic accuracy. Statistical models, such as the Framingham Stroke Risk Profile, played a crucial role in quantifying risk in the latter part of the 20th century.

The integration of machine learning (ML) techniques marked a paradigm shift, enabling researchers to analyze vast datasets for predictive insights. Numerous studies have applied ML algorithms, including decision trees, support vector machines, and neural networks, to enhance stroke prediction accuracy. These approaches leverage diverse data sources, ranging from demographic information to genetic and imaging data. Recent literature emphasizes the effectiveness of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in image-based stroke prediction. These advanced techniques demonstrate a capacity to extract intricate patterns from medical imaging data, offering promising avenues for improved accuracy.

As the field progresses, ethical considerations surrounding patient privacy, data security, and informed consent have gained prominence. Researchers and healthcare practitioners are tasked with striking a balance between the potential benefits of predictive models and safeguarding individual rights.

The existing literature on brain stroke prediction reflects a dynamic evolution from traditional clinical assessments to advanced data-driven approaches. While substantial progress has been made, there is a continual need for innovative solutions that balance predictive accuracy, ethical considerations, and real-world applicability. As researchers navigate these challenges, the collective insights from existing literature provide a solid foundation for shaping the future of brain stroke prediction.

2.3 Techniques that are Going to Use

The techniques employed in brain stroke prediction research vary and often include a combination of methodologies to enhance accuracy and reliability. Some of the prominent techniques used in this domain are:

Machine Learning Algorithms:

- **Decision Trees:** Tree-based models like decision trees and random forests are commonly utilized for their interpretability and ability to handle complex datasets.
- **Support Vector Machines (SVM):** SVMs are effective in classifying data into different categories, making them valuable for stroke risk prediction tasks.
- **Logistic Regression:** Logistic regression is a statistical method frequently applied for binary classification problems, including stroke prediction.
- **K-Nearest Neighbor:** The majority class or mean of the k-nearest data points in the feature space is the basis of a machine learning technique used for classification and regression applications.

Deep Learning Techniques:

- Convolutional Neural Networks (CNNs): CNNs are good at image-based tasks, they can be used to analyse medical imaging data, like MRI and CT scan results, to predict strokes.
- Recurrent Neural Networks (RNNs): RNNs are useful for capturing temporal dependencies in patient data over time because of their proficiency processing sequential data.

Data Pre-processing and Feature Selection:

- Image Processing: Techniques for enhancing and analyzing medical images, such as filtering and segmentation, contribute to extracting relevant features for stroke prediction.
- Dimensionality Reduction: The dimensionality of datasets can be decreased by techniques like Principal Component Analysis (PCA) and feature selection algorithms, which increase computational efficiency.

Predictive Analytics and Statistical Models:

- Statistical Risk Models: Traditional statistical models, including logistic regression and Cox proportional hazards models, are often employed to analyze the relationships between various risk factors and stroke occurrence.

Wearable Technology and Continuous Monitoring:

- Physiological Sensors: Wearable devices equipped with physiological sensors, such as heart rate monitors and accelerometers, enable continuous monitoring of vital signs for early detection of stroke risk.

Cloud Deployment and Web Applications:

- Cloud Computing: Utilizing cloud infrastructure enhances the scalability and accessibility of stroke prediction models.
- Web Application: Developing user-friendly web applications allows for easy representation and interpretation of results, promoting wider adoption.

The selection of techniques often depends on the specific goals of the research, available data, and the desired level of interpretability and generalizability of the predictive models.

3. Methodology/Proposed System

3.1 Data Collection

This project comprises analyzing a healthcare dataset to predict the likelihood of a stroke in individuals. The dataset is taken from Kaggle and it includes several patient health and lifestyle parameters such as “age”, “hypertension”, “heart disease”, “marital status”, “work type”, “residence type”, “average glucose level”, “BMI”, “smoking status”, and “gender”. Each record relates to a specific patient, and the features include a variety of risk variables associated with the development of a stroke.

There are 5110 rows and 12 columns in the dataset, the columns and the column descriptions are,

Id: It's a unique identifier

gender: “Male”, “Female” and “Other” are the values. This characteristic denotes the patient's gender, influencing stroke risk through biological disparities and gender-specific lifestyle patterns.

Age: This critical stroke prediction factor, which indicates the patient's age, emphasises that the risk of stroke increases with age. The World Health Organisation states that after the age of fifty-five, the risk of stroke doubles each decade.

Hypertension: 0 if the patient has no hypertension and 1 for the patient having hypertension. Hypertension significantly raises stroke risk by causing damage to blood vessels, making them susceptible to blockage or rupture.

Heart_disease: If the patient has no cardiac problems, the answer is 0. 1 if the patient has a cardiac condition. Patients with heart conditions face an elevated risk of stroke, as these ailments may result in clot formation in the heart that can migrate to the brain.

Ever_married: "Yes" for married and "No" for unmarried. This attribute reveals the marital status of the patient. Although not a direct factor influencing stroke risk, marital status could be linked to lifestyle aspects that impact the likelihood of a stroke. For instance, married individuals might display varying stress levels, physical activity routines, or dietary habits in comparison to their unmarried counterparts.

Work_type: The values "children", "Govt_jov", "Never_worked", "Private" or "Self-employed". Provide details about the patient's profession; this categorical characteristic could be associated with elevated stress levels, influencing the risk of stroke.

Residence_type: The values are "Rural" or "Urban". This characteristic indicates whether the individual lives in an urban or rural setting. The place of residence could be linked to the risk of stroke, considering factors such as accessibility to healthcare, air quality, and lifestyle choices.

Avg_glucose_level: Representing the average glucose level in the patient's blood, elevated levels can damage blood vessels, heightening the risk of stroke.

Bmi: The patient's body mass index, or BMI, is determined by dividing their weight in kilogrammes by the square of their height in metres. Obesity, or having a high body mass index (BMI), increases the risk of stroke and can cause or exacerbate illnesses like high blood pressure, hyperglycemia, and cardiac disease.

Smoking_status: "formerly smoked," "never smoked," "smoke," or "Unknown" are the available values. This categorical characteristic indicates if the patient currently smokes, has smoked in the past, or has never smoked. By causing damage to blood arteries, raising blood pressure, and reducing the amount of oxygen reaching the brain, smoking raises the risk of stroke.

Stroke: This column gives information about the person who is having a stroke risk and who is not. 1 if the patient had a stroke or 0 if not.

3.2 Data Preprocessing

Preprocessing, also known as data preparation, is the act of turning unprocessed data into a format that can be accessed. This is an important step in data analysis because we cannot work with raw data. It is necessary to assess the data quality prior to utilising machine learning techniques. By following these methods, we can clean the data.

- Handle missing values: Depending on the conditions, columns such as 'bmi' may contain missing values that need to be filled up or the rows eliminated.
- Handle duplicates: We must guarantee that the dataset does not contain any duplicated rows. If there are duplicates, we must select whether to keep one, keep all, or average the duplicates, based on the data and research topic.
- Handle outliers: Some columns, such as 'avg_glucose_level' and 'bmi,' are numerical and may contain outliers that skew the analysis. We should decide how to handle things, such as deleting them or transforming them to lessen their influence.
- Convert categorical variables to data types that are appropriate: Gender, ever_married, work_type, residence_type, and smoking_status are all categorical columns. We may need to change these to dummy variables depending on the analysis.
- Handle invalid values: For example, 'age' should not be negative, 'hypertension' and 'heart_disease' should be 0 or 1, 'gender' should be 'Male' or 'Female', and so on.

The 'id' attribute is a random number assigned for patient identification and does not influence stroke prediction, making it negligible. The remaining attributes provide essential patient information, comprising three continuous variables (age, average glucose level, and body mass index) and seven categorical factors ("gender, hypertension, heart_disease, ever_married, work_type, residence_type, and smoking_status").

In this dataset, the patients having stroke data is very less so we are not going to delete the null values. Instead of deleting null values replace it with the mean or median.

After the data cleaning process, the analytical dataset comprises 11 attributes with 5110 instances, including 208 instances representing data from patients with a history of stroke.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is like taking a first look at data to understand it better. It means checking and drawing pictures of the data to see what's important, find patterns, and see how things are connected. The idea is to learn about the data and help researchers or analysts decide what to do next. EDA involves doing things like cleaning up the data, making summaries, and creating charts and graphs to show what the data looks like. It's like a roadmap that helps figure out where to go next in studying the data.

Target Variable

The target variable is the main outcome or response being predicted in a statistical or machine learning model. In this dataset, Stroke is the target variable, And the distribution of strokes among the whole data is plotted in the below count plot:

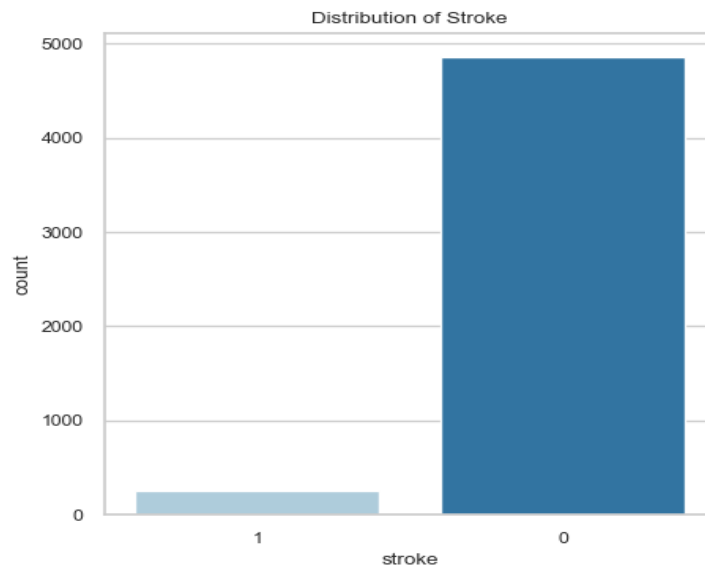


Fig 1: Distribution of Stroke

From Fig 1 it is very clear that the people having a risk of stroke is negligible compared to the people having no risk of stroke. However, when we are taking a huge data there will be a greater number of people having a stroke risk. The purpose of this project is to find out the age, gender and health conditions like heart disease, hypertension, body mass index and average glucose level are how affected by the risk of stroke and find out the way for an early prediction. The next step is to analyse the distribution of age, glucose level and BMI which means the numerical data in the dataset.

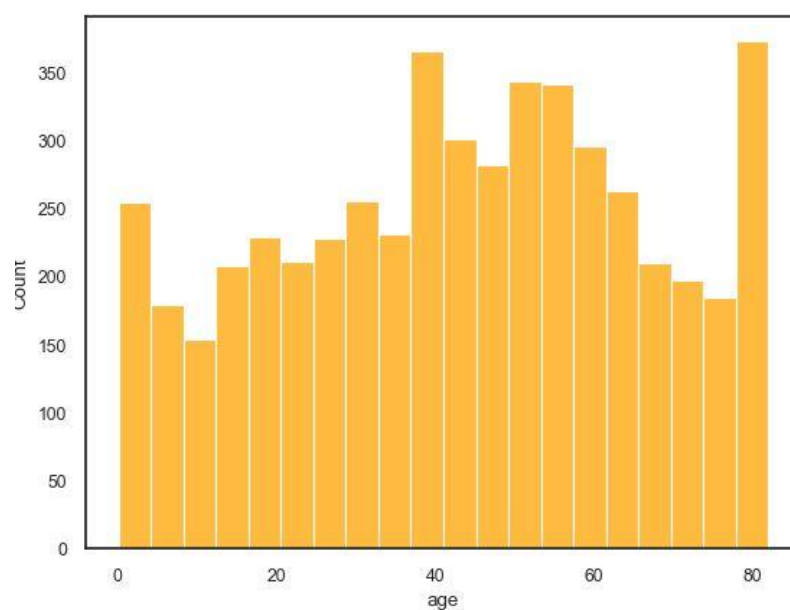


Fig 2: Distribution of Age

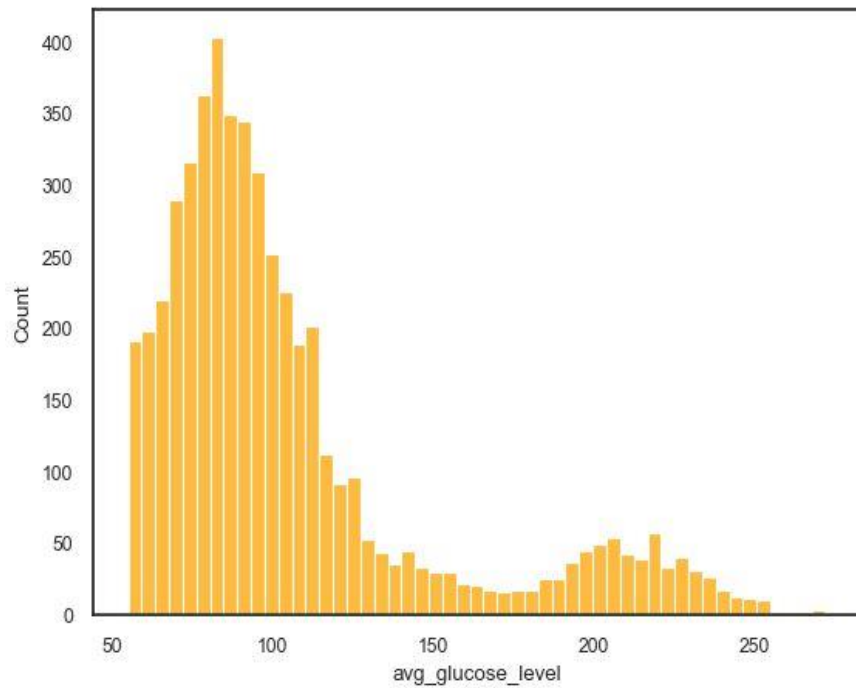


Fig 3: Distribution of Average Glucose Level

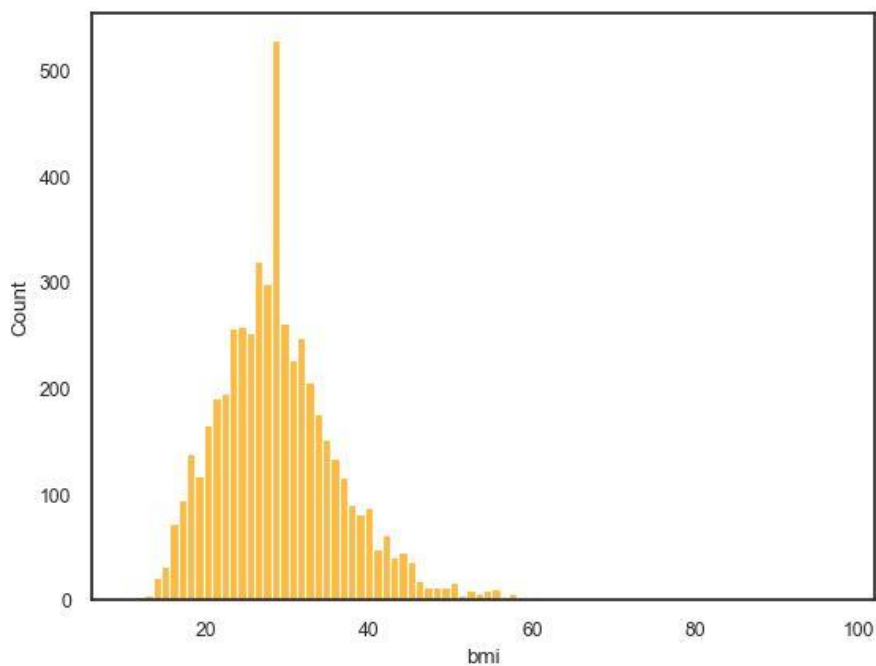


Fig 4: Distribution of Body Mass Index

The above figures Fig2, Fig 3 and Fig 4 depict the distribution of age, average glucose level, and body mass index across the entire dataset. It's evident that individuals above their 80s are predominantly represented in the dataset. Regarding average glucose level, a significant number falls within the range of 50 to 120, and the body mass index hovers around 50 for most individuals.

In the subsequent figure, Fig 5, a pairplot demonstrates the relationships between age, average glucose level, body mass index, and the occurrence of stroke among these.

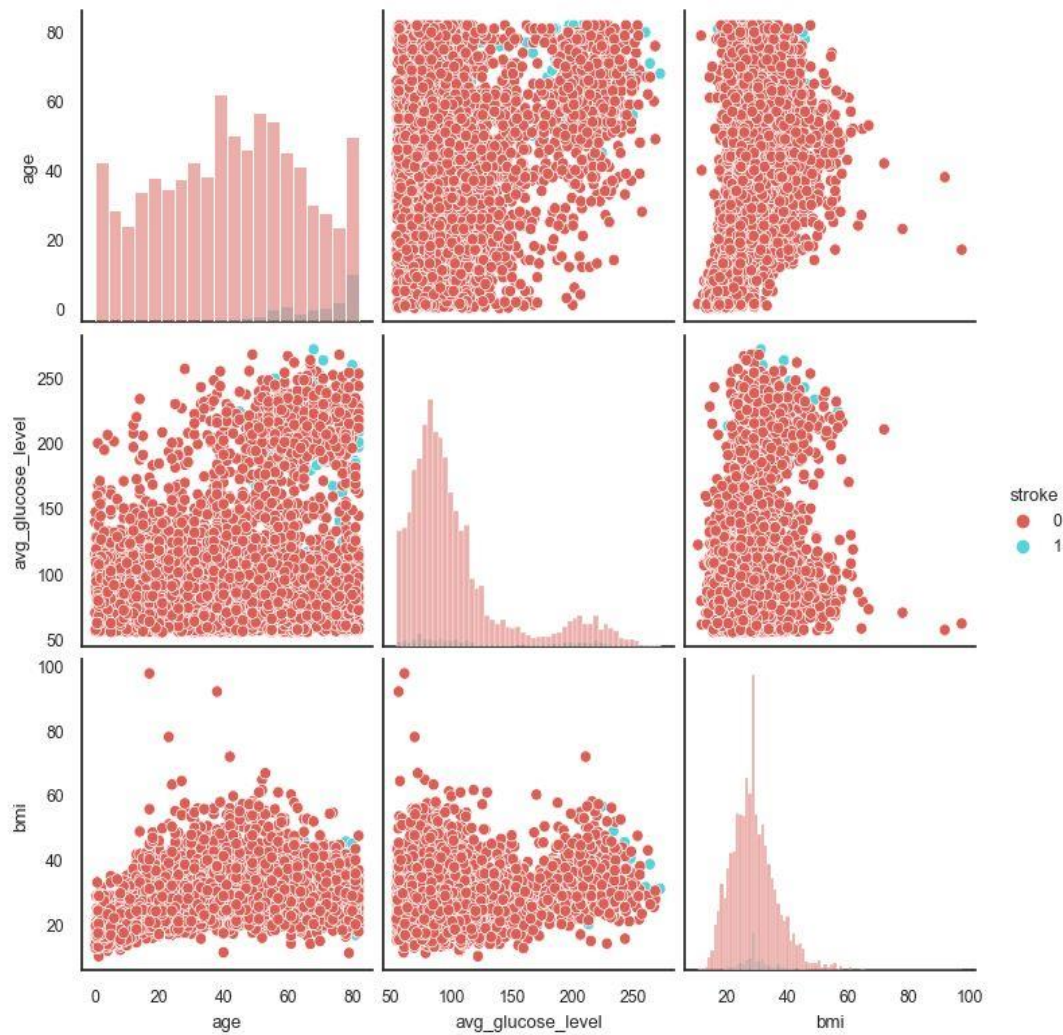


Fig 5: Pair plot of Age, Average Glucose Level and BMI

In the depicted figure Fig 5, the presence of red signifies an elevated risk of stroke, while blue indicates the absence of such risk. This pairplot analysis provides a comprehensive visual exploration of the interrelationships and distributions of age, avg_glucose_level, and BMI, categorized by stroke status. Examining the pairwise comparisons, certain trends become evident. In the age vs avg_glucose_level plot, although there isn't a distinct correlation between age and average glucose level, stroke patients (depicted in orange) exhibit an inclination towards older age and elevated glucose levels. Similarly, in the age vs BMI plot, no clear relationship emerges between age and BMI, yet stroke patients tend to be older, with BMI showing minimal divergence between stroke and non-stroke patients. The avg_glucose_level vs BMI plot reveals no discernible relationship between average glucose level and BMI; however, stroke patients consistently demonstrate higher glucose levels, irrespective of their BMI. The diagonal plots detailing variable distributions for stroke and non-stroke patients affirm the earlier univariate analysis, reinforcing that stroke patients typically skew older and have elevated glucose levels, while their BMI distribution aligns closely with non-stroke patients.

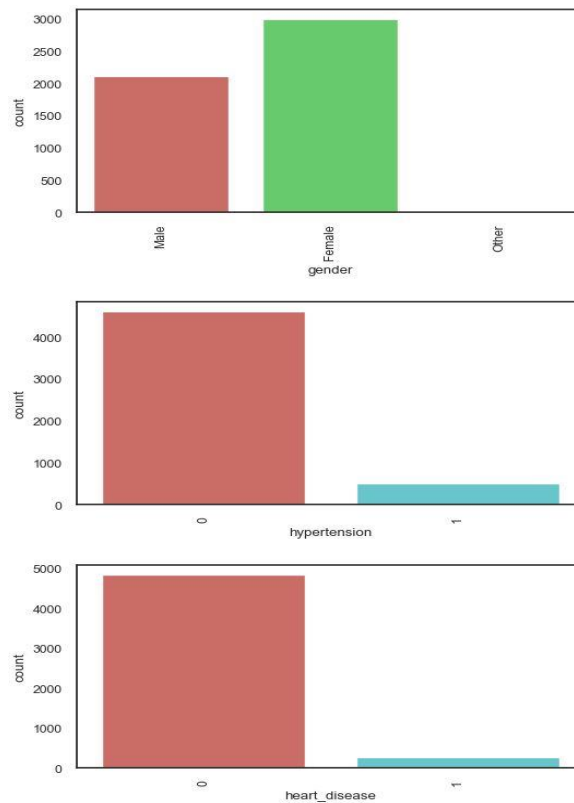


Fig 6: Countplot for Gender, Hypertension and Heart Disease

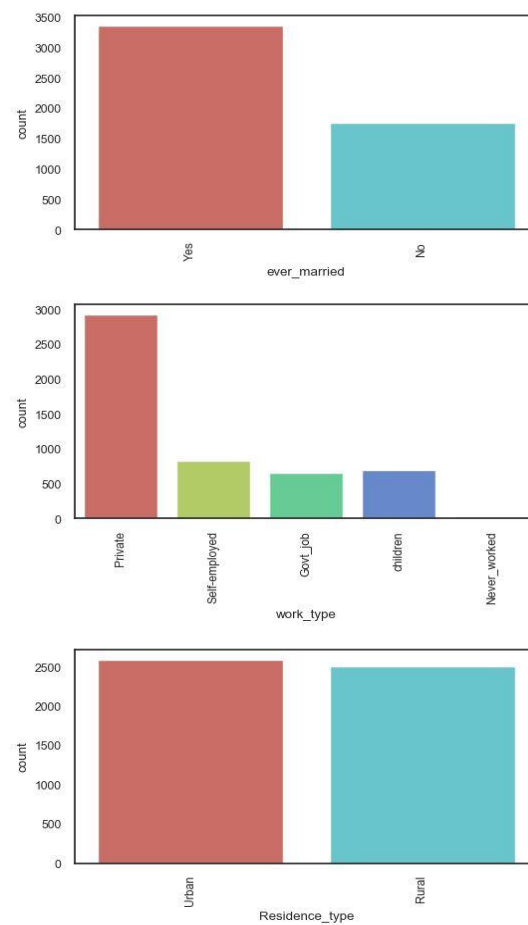


Fig 7: Count plot for Marital Status, Work Type and Residence Type

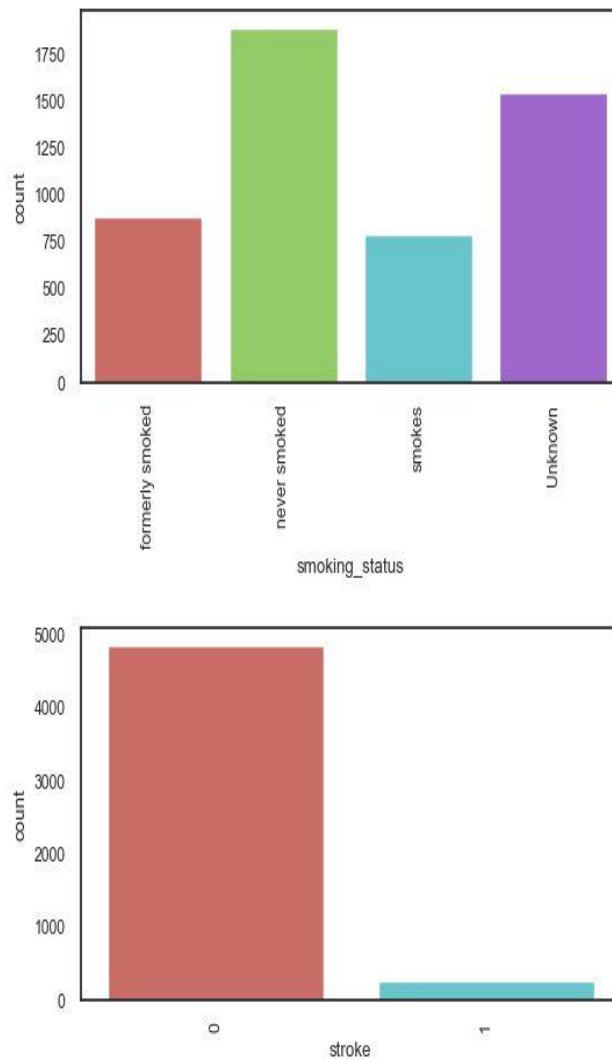


Fig 8: Count plot for Smoking Status and Stroke

The above figures Fig:6, Fig 7, and Fig 8 provide visual representations of the distribution of categorical variables, including gender, hypertension, heart disease, marital status, work type, residence type, and smoking status. The key observations derived from this analysis are as follows: the dataset contains a higher number of female patients compared to male patients, with a small fraction identifying as Other. Regarding hypertension, the majority of patients do not have this condition. Similar patterns are observed for heart disease, indicating that the majority of individuals do not have heart disease. Analyzing marital status reveals that most patients are married. In terms of work type, the dataset shows a predominant representation in the Private work category, with substantial numbers in Self-employed and children categories. Conversely, Govt_job and Never_worked categories have fewer patients. The distribution of residence type is almost equal between patients in urban and rural areas. A significant observation from the graphs is that the majority of patients have never smoked. Categories like formerly smoked and smokes exhibit lower numbers, while a substantial portion falls under the Unknown smoking status category.

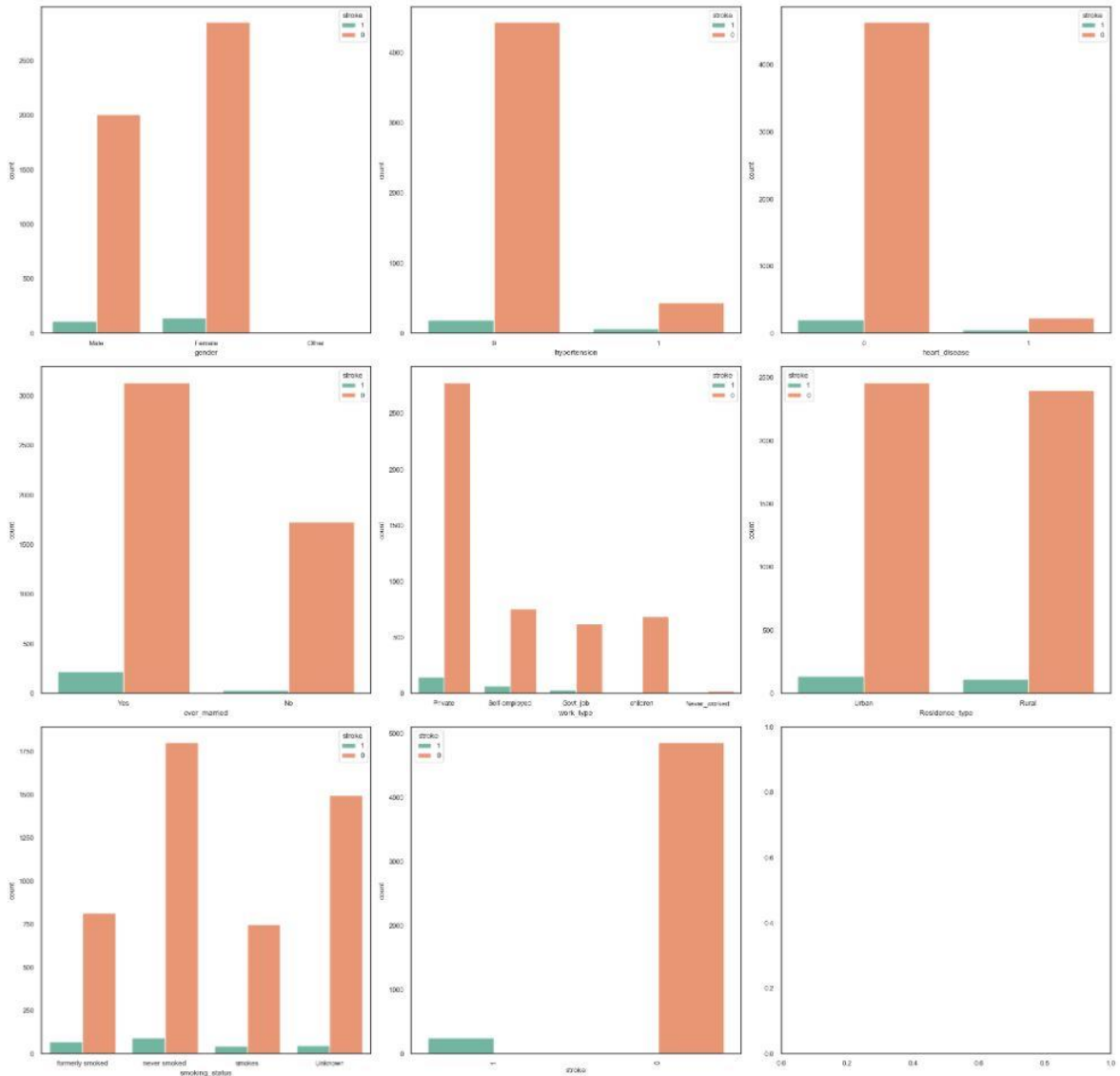


Fig 9: Gender, Hypertension, Heart Disease, Marital Status, Job Type, Residence Type, Smoking Status Distributions with and without Stroke

The above figure Fig 9. is a bivariate analysis of gender, hypertension, heart disease, marriage status, job type, residence type, smoking status etc. with stroke. From this analysis, it is observed that both males and females exhibit a similar proportion of stroke cases, with males having a slightly higher incidence. The Other category shows no instances of stroke, but this might be attributed to the small sample size within this category. Patients with hypertension or heart disease demonstrate a higher proportion of stroke cases compared to those without these conditions. Interestingly, individuals who have been married show a higher proportion of stroke cases than those who have not experienced marriage. In terms of work type, patients in self-employment or private jobs have a higher proportion of stroke cases compared to other occupational categories. The residence type does not significantly impact stroke occurrence, as the proportion is nearly equal for both urban and rural residents. Notably, patients who formerly smoked or currently smoke exhibit a higher proportion of stroke cases than those who have never smoked. The 'Unknown' smoking status category indicates a lower stroke proportion.

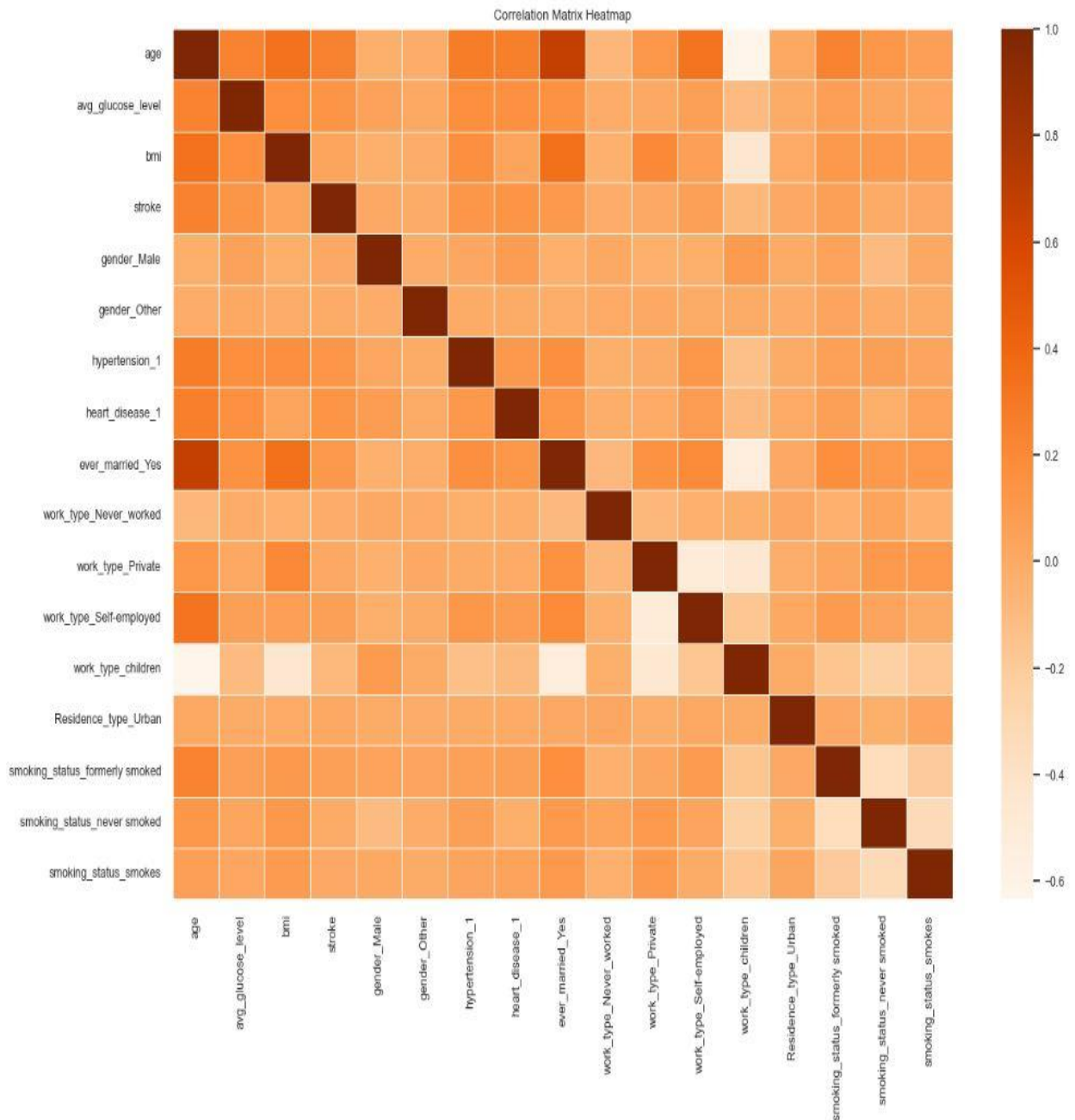


Fig 10: Correlation Matrix Heat Map

The above figure Fig 10. is heatmap which presents a visual representation of the correlation between every pair of features within the dataset. Each cell's color indicates the correlation coefficient between the respective pair of variables: a darker color signifies a stronger correlation.

The following figure Fig 11. showing the correlation with stroke. This heatmap orders the features according to their correlation with the target variable, stroke.

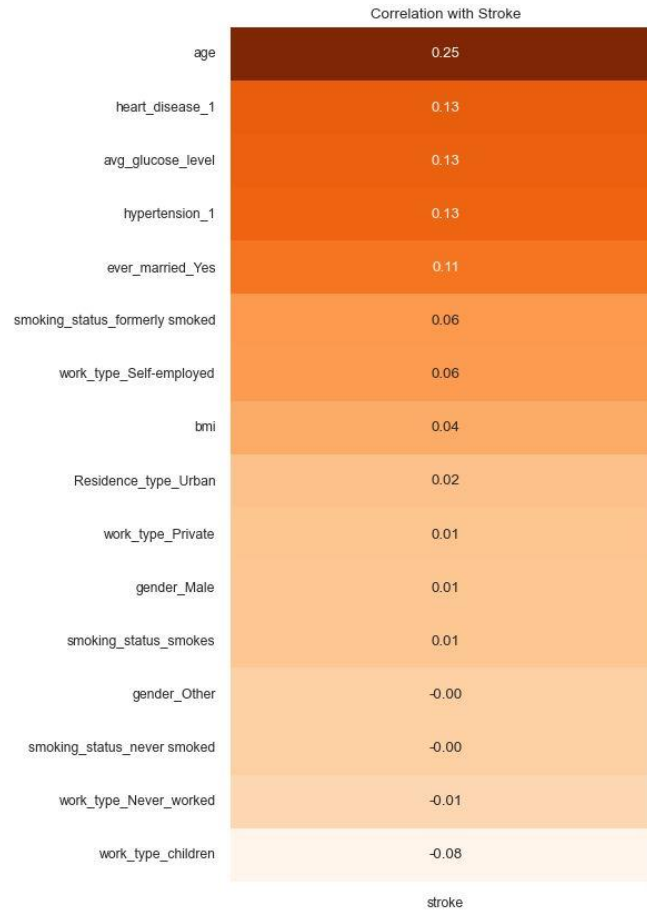


Fig 11: Correlation with Stroke

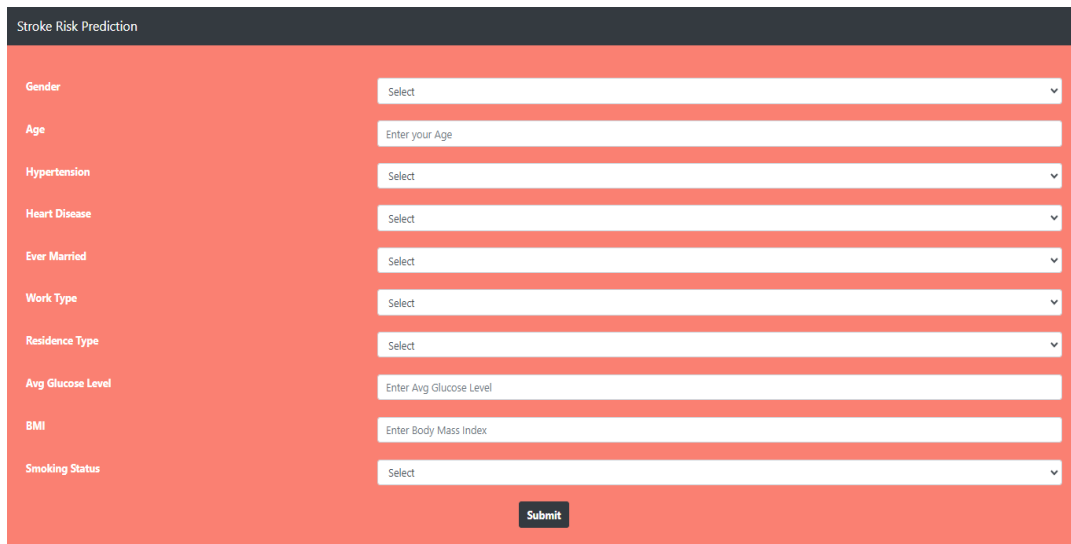
The above heatmap Fig 11. analysis reveals several noteworthy observations. The age feature exhibits the highest positive correlation with the stroke target variable, indicating that older individuals might be more susceptible to strokes, aligning with established medical insights. Additionally, hypertension and heart_disease features also display positive correlations, suggesting that individuals with hypertension or heart disease could be at a higher risk of experiencing a stroke. Positive correlations are observed for avg_glucose_level and ever_married_Yes, implying that elevated average glucose levels and being married might be linked to an increased likelihood of stroke. Moreover, work_type_Self-employed and gender_Male features show positive correlations, indicating that self-employed individuals and males may have a higher probability of experiencing a stroke. Conversely, features such as work_type_children and Residence_type_Urban exhibit negative correlations with stroke, implying that these factors might be associated with a lower likelihood of stroke occurrence.

3.4 Implementation of Web Application

This web application is an advanced tool meant to give people information about their possible risk of having a stroke. Personalised risk evaluations based on many health and lifestyle characteristics are intended to be provided by this application, which makes use of sophisticated machine learning algorithms and extensive data analysis. Since stroke is one of the world's leading causes of long-term disability and death, early detection and proactive risk management are vital. This web application serves as a user-friendly and accessible platform,

allowing individuals to input relevant information and receive valuable predictions regarding their stroke risk. Through a user-friendly interface and the predictive models, strive to contribute to the prevention and early intervention of strokes, ultimately promoting better healthcare outcomes and wellbeing. This web app is created by using python language and the tool used for python coding is PyCharm.

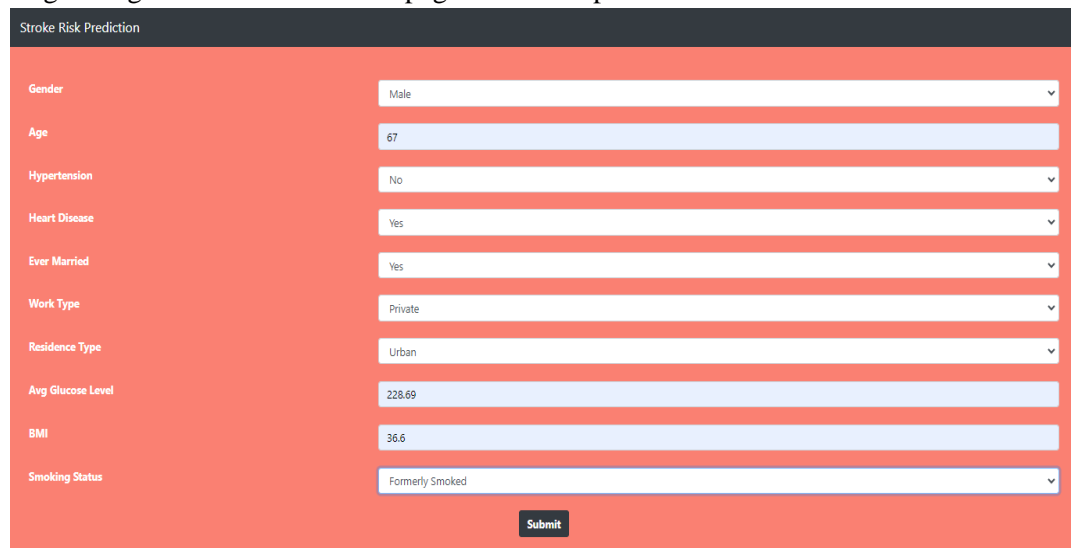
The below figure Fig 12. Is the home page of the Stroke Risk Prediction application. The user can enter their required inputs using the textboxes and some inputs can select from list boxes also.



The screenshot shows the 'Stroke Risk Prediction' application interface. It features a dark blue header with the title 'Stroke Risk Prediction'. Below the header, there is a list of input fields on the left and corresponding input boxes on the right. The input fields are: Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg Glucose Level, BMI, and Smoking Status. The input boxes are: a dropdown menu for Gender (showing 'Select'), a text box for Age (showing 'Enter your Age'), a dropdown menu for Hypertension (showing 'Select'), a dropdown menu for Heart Disease (showing 'Select'), a dropdown menu for Ever Married (showing 'Select'), a dropdown menu for Work Type (showing 'Select'), a dropdown menu for Residence Type (showing 'Select'), a text box for Avg Glucose Level (showing 'Enter Avg Glucose Level'), a text box for BMI (showing 'Enter Body Mass Index'), and a dropdown menu for Smoking Status (showing 'Select'). A 'Submit' button is located at the bottom right of the form.

Fig 12: Stroke Risk Prediction Home Page

Below figure Fig 13. Is the same home page with the input values in the fields.



The screenshot shows the 'Stroke Risk Prediction' application interface with input values entered. The input fields are: Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type, Residence Type, Avg Glucose Level, BMI, and Smoking Status. The input boxes are: a dropdown menu for Gender (showing 'Male'), a text box for Age (showing '67'), a dropdown menu for Hypertension (showing 'No'), a dropdown menu for Heart Disease (showing 'Yes'), a dropdown menu for Ever Married (showing 'Yes'), a dropdown menu for Work Type (showing 'Private'), a dropdown menu for Residence Type (showing 'Urban'), a text box for Avg Glucose Level (showing '228.69'), a text box for BMI (showing '36.6'), and a dropdown menu for Smoking Status (showing 'Formerly Smoked'). A 'Submit' button is located at the bottom right of the form.

Fig 13: Stroke Risk Prediction Home with Values

After entering all the required inputs, the user can submit the form by clicking the Submit button on the bottom of the screen.

If the person who is having no stroke risk will be navigating to the below page Fig 14. with the green background and a message will display like “You have been diagnosed with no Stroke Risk. Congratulations”.

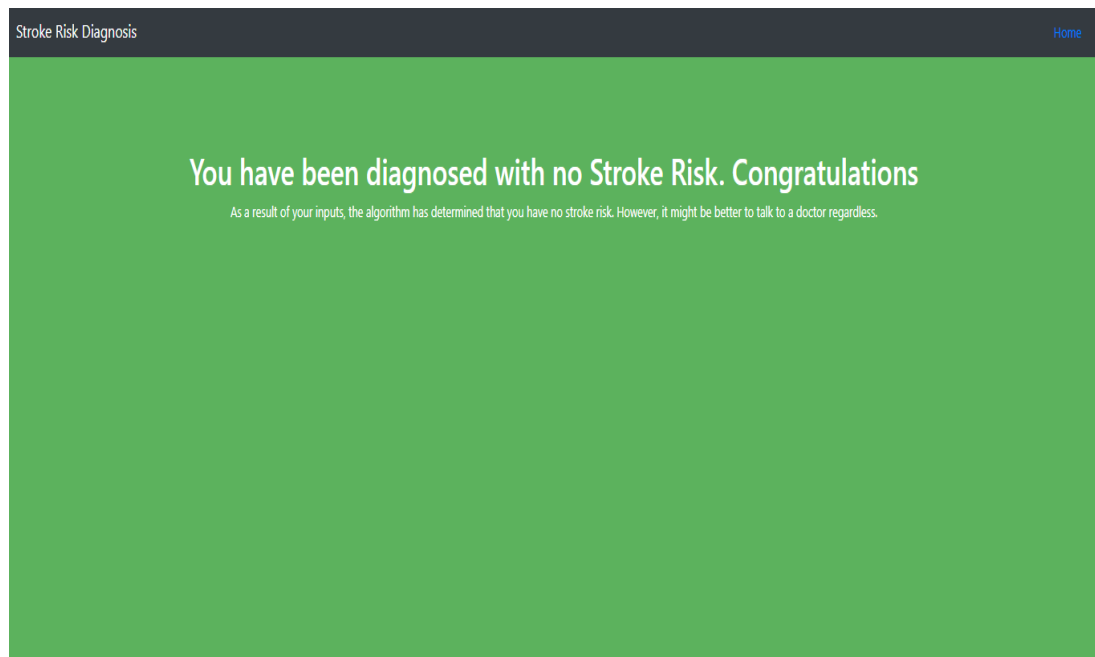


Fig 14: No Stroke Risk Diagnosed Page

If the person who is having a stroke risk will be navigating to the below page Fig 15. with the red background and a message will display like “You have been diagnosed with Stroke Risk. Please consult a doctor”.

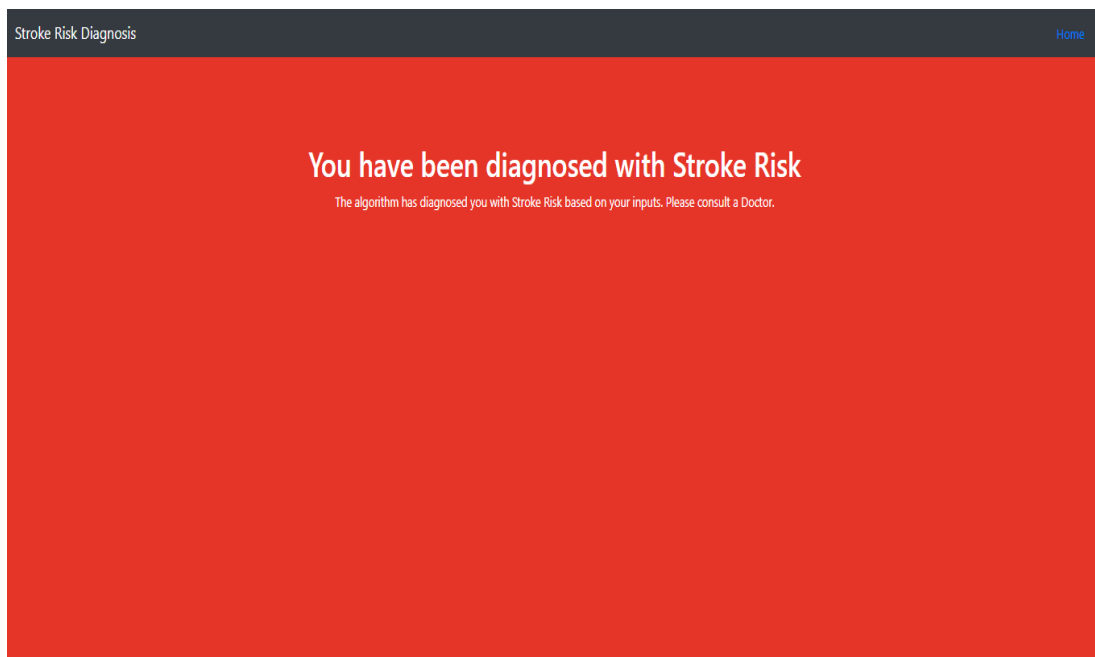


Fig 15: Stroke Risk Diagnosed Page

4. Experimentation

In a research paper, experimentation refers to the systematic and controlled investigation conducted to test hypotheses, answer specific research questions, or explore the relationships between variables. It involves a carefully designed and structured process where researchers manipulate certain factors, observe outcomes, and analyze the collected data to draw meaningful conclusions.

In this research paper, five machine learning models are used for a fast and easy prediction of the brain stroke risk. So, in this part of the research report, explaining the machine learning methods used to predict the stroke risk.

4.1 Predictive Machine Learning Models

Predictive machine learning models are computational algorithms designed to analyze data and make predictions about future outcomes or trends. These models leverage statistical patterns and relationships within historical data to learn and generate predictions for new, unseen data. The primary objective is to identify and understand patterns in the data that can be used to forecast outcomes, classify data into specific categories, or estimate numerical values.

Key characteristics and types of predictive machine learning models include:

Supervised Learning Models: In these models, the algorithm learns the link between input attributes and associated output labels by training it on labelled data. Neural networks, support vector machines, and linear regression are a few examples.

Unsupervised Learning Models: In unsupervised learning, structures and patterns are found in the data by training models on unlabeled data. Unsupervised learning models include clustering algorithms like k-means and dimensionality reduction methods like principal component analysis (PCA).

Regression Models: Regression models are used for predicting numerical values or quantities. Linear regression, polynomial regression, and decision tree regression are common regression techniques.

Classification Models: Classification models are designed to categorize data into predefined classes or groups. Examples include logistic regression, decision trees, random forests, and support vector machines.

Time Series Models: These models are specialized for predicting future values based on past time-ordered data. Autoregressive Integrated Moving Average (ARIMA), exponential smoothing methods, and recurrent neural networks (RNNs) are used for time series prediction.

Ensemble Models: To increase performance overall, ensemble approaches aggregate the predictions of several base models. Popular ensemble approaches include Gradient Boosting and Random Forest.

Deep Learning Models: Deep learning models, often based on neural networks, are particularly effective for complex tasks, such as image recognition, natural language processing, and speech recognition.

Predictive machine learning models find applications in various fields, including finance, healthcare, marketing, and manufacturing. They play a crucial role in decision-making processes by providing insights, identifying trends, and aiding in risk assessment. The success of these models depends on the quality and relevance of the data used for training and their ability to generalize well to new, unseen data.

This research is used five different machine learning models for the predictive analysis of stroke risk. The models used for the prediction are, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbors Classifier and Advanced Vector machine.

4.1.1 Logistic Regression

The logistic regression model for stroke risk prediction is a statistical tool used to assess the likelihood of an individual experiencing a stroke based on certain input features. This model operates under the assumption that the relationship between these features and the probability of stroke follows a logistic function. By analyzing historical data where the occurrence of strokes is known, the model learns to quantify the impact of various factors on stroke risk. The logistic regression approach is valued for its simplicity and interpretability, offering insights into which features contribute to increased or decreased odds of stroke. This predictive model aids in understanding and identifying key risk factors, facilitating informed decision-making in healthcare and enabling targeted preventive measures for individuals at higher risk of stroke.

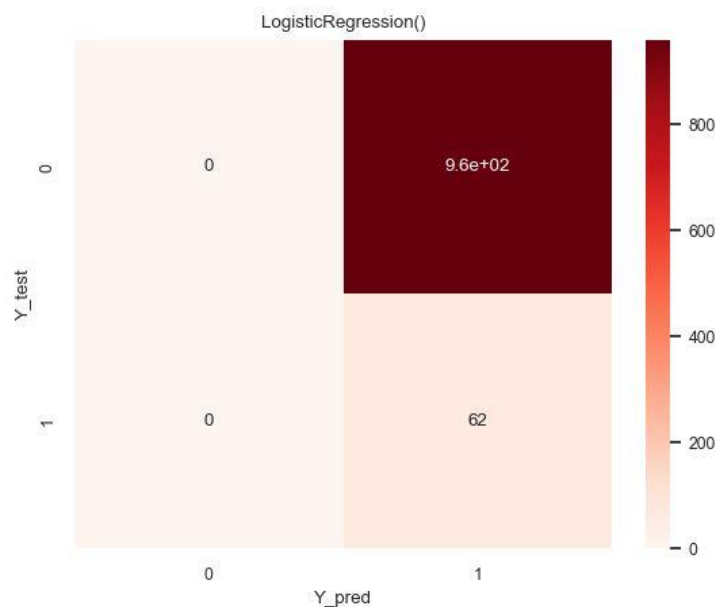


Fig 16: Confusion Matrix Logistic Regression

Here in this research, Logistic Regression gives 93.83% of accuracy score.

4.1.2 Decision Tree

The Decision Tree classifier for stroke risk prediction is a machine learning algorithm that employs a tree-like structure to make predictions regarding an individual's likelihood of experiencing a stroke. This model breaks down the data into a series of binary decisions based on different features, ultimately leading to a prediction at the leaf nodes of the tree. Each decision is made by selecting the feature that best separates the data into subsets with distinct stroke outcomes. The simplicity and interpretability of Decision Trees make them advantageous for understanding the contributing factors to stroke risk. By following the branches of the tree, healthcare professionals can identify key determinants and their thresholds, facilitating a straightforward assessment of an individual's risk of stroke and informing targeted preventive measures.

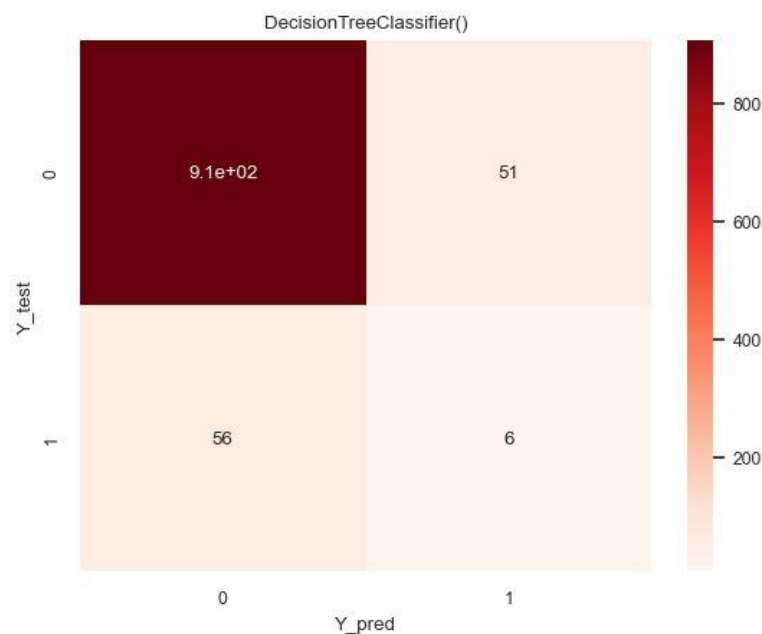


Fig 17: Confusion Matrix Decision Tree

Here in this research, Decision Tree Classifier gives 89.53% of accuracy score.

4.1.3 Random Forest

A powerful machine learning technique called the Random Forest classifier for stroke risk prediction constructs many decision trees and then combines their predictions to improve accuracy and resilience. A random subset of the data and features are used to train each decision tree in the Random Forest, resulting in a wide range of predictions. The total of each tree's individual forecasts is used to determine the final prediction. By using this method, the likelihood of overfitting is decreased and the model's capacity to manage intricate relationships within the data is enhanced. When it comes to predicting stroke risk, Random Forests are an effective tool for examining a range of variables and how they interact, giving a thorough and accurate estimate of a person's chance of having a stroke depending on their specific traits.

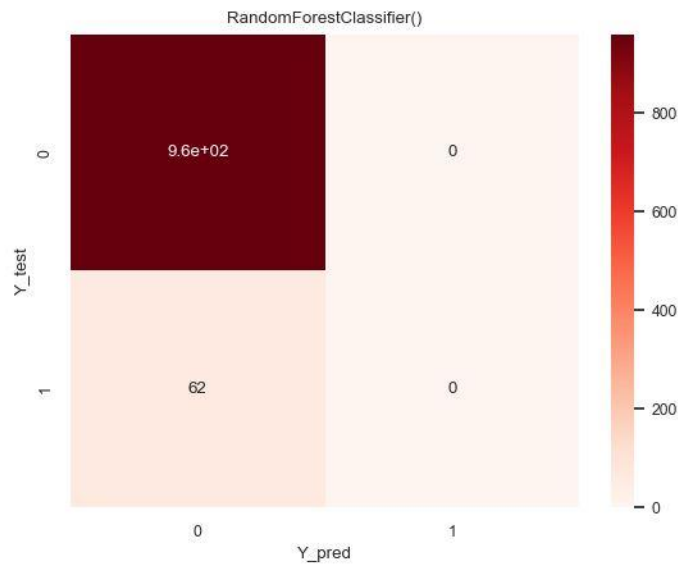


Fig 18: Confusion Matrix Random Forest

Here in this research, Random Forest Classifier gives an accuracy score 93.73%.

4.1.4 K-Nearest Neighbors Classifier

The K-Nearest Neighbors (KNN) classifier for stroke risk prediction is a straightforward yet effective machine learning algorithm. According to this method, a new data point's classification is based on the majority class of its K nearest neighbors in the feature space. The model's sensitivity to local fluctuations depends on the value of K. With regard to stroke risk, KNN uses feature similarity between individuals to provide predictions. For instance, if a person's attributes closely resemble those of others who have experienced a stroke, the model would classify them as at higher risk. The simplicity and adaptability of KNN make it a valuable tool for healthcare practitioners seeking a quick and intuitive method for personalized stroke risk assessment, especially when considering local patterns and individualized factors.

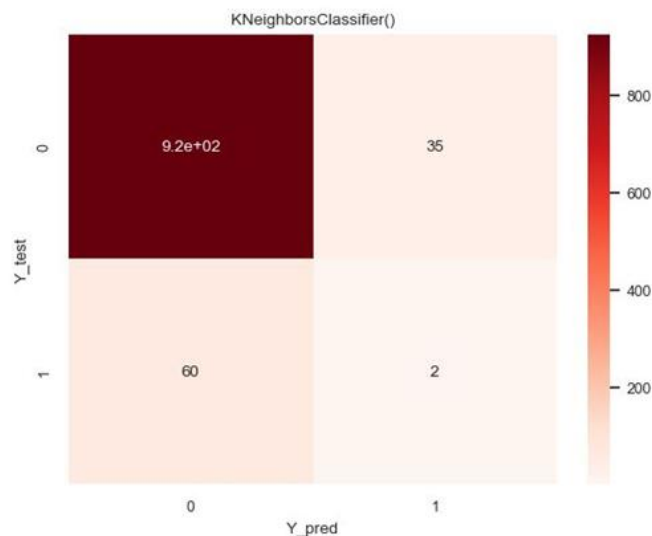


Fig 19: Confusion Matrix KNN

Here in this research, K-Nearest Neighbors Classifier gives an accuracy score 93.44%.

4.1.1 Support Vector Machine

An effective machine learning technique that is great at determining the best decision limits is the Support Vector Machine (SVM), which is used to forecast the risk of stroke. SVM maximises the margin between the two classes by locating the hyperplane that best divides people with and without stroke in the context of stroke risk assessment. When it comes to managing intricate datasets and identifying non-linear relationships, this algorithm excels. SVM helps identify people as either at risk of a stroke or not by taking into account their attributes to establish where they are in the feature space. Because of its adaptability, SVM is a useful tool in the healthcare industry. It can produce precise predictions based on a variety of parameters, which in turn helps medical personnel identify patients who can benefit from specific preventive measures.

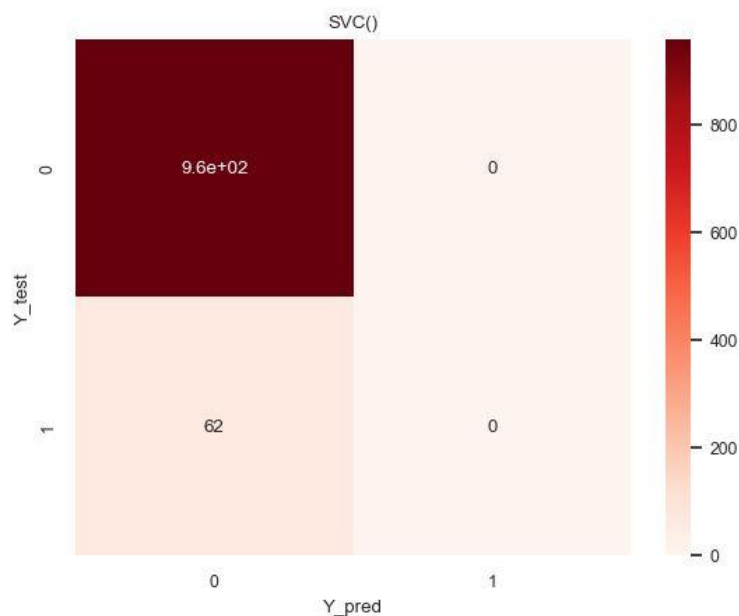


Fig 20: Confusion Matrix SVM

Here in this research, Support Vector Machine gives an accuracy score 93.93%.

4.2 Comparison Between Models

In the realm of machine learning for stroke risk prediction, the comparison between various models stands as a pivotal exploration to determine their effectiveness in providing accurate predictions. This comparative analysis delves into the performance metrics achieved by different models during cross-validation.

This comparative analysis aims to expose the advantages and disadvantages of each model, offering insightful information about each's capabilities and directing the choice of the best method for accurate and reliable stroke risk estimations.

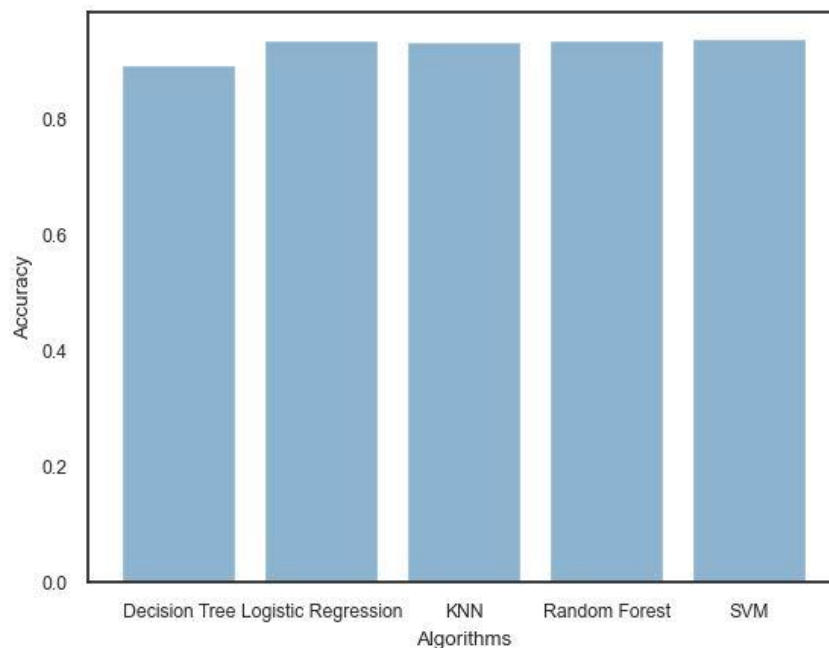


Fig 21: Model Comparison Bar Chart

Figure showing above Fig 21. Is a bar chart which showing the accuracy comparison of five different machine learning models used to predict the risk of brain stroke. Logistic Regression exhibited a commendable accuracy of 93.83%, showcasing its ability to effectively discern patterns and relationships within the dataset. Decision Tree, while slightly less accurate at 89.53%, demonstrated a reliable capacity for predicting stroke outcomes. Random Forest, with an accuracy of 93.73%, emerged as a robust ensemble learning method, excelling in capturing complex data relationships. K-Nearest Neighbors (KNN) exhibited strong performance with an accuracy of 93.44%, particularly effective in identifying localized patterns in the data. Support Vector Machine (SVM) stood out with the highest accuracy at 93.93%, indicating its proficiency in defining optimal decision boundaries. These cross-validation results provide a comprehensive evaluation of the models, offering insights into their respective strengths and highlighting their suitability for accurate stroke risk prediction in diverse scenarios.

From the analysis of all the five machine learning models, the highest accurate one is the Standard Vector Machine with a 93.93% of accuracy score. The second one is the Logistic Regression with an accuracy 93.83%. Random Forest, KNN and Decision tree are the consecutive places with 93.73%, 93.44% and 89.53% respectively.

5. Conclusion and Recommendations

5.1 Conclusion

In conclusion, the study on brain stroke risk prediction using machine learning has significantly advanced our understanding of proactive healthcare strategies. By harnessing the power of advanced algorithms and analyzing diverse patient data, the research underscores the potential for early detection and prevention of strokes. The findings emphasize the importance of a holistic approach to risk assessment, incorporating various demographic, lifestyle, and physiological factors. The successful application of machine learning models in identifying critical risk factors signifies a transformative shift towards personalized medicine. As we navigate the promising landscape of early brain stroke prediction, it is imperative to address challenges such as data diversity, ethical considerations, and real-world validation. This research provides a foundation for future endeavors, guiding healthcare practices towards more accurate, timely, and patient-centric stroke prevention strategies.

The major contributions are;

- This study attains an accuracy of 93.93%, surpassing the previous results achieved by other researchers in this specific area.
- In this research, five classifiers and various machine learning techniques, such as label encoding, outlier removal, and cross-validation, are used to achieve the optimal outcome.
- A web page and a web application are created based on this research, capable of accurately calculating results using real-time inputs.
- Among the five classifiers utilized, Support Vector Machine and Random Forest exhibit the highest accuracies, reaching 93.93% and 93.73%, respectively

5.2 Recommendations

To enhance brain stroke risk prediction, several recommendations emerge from this study. First and foremost, there is a critical need to expand and diversify datasets, incorporating a comprehensive range of demographic, genetic, and lifestyle factors. This expansion will not only fortify machine learning models but also ensure their applicability across diverse patient populations. Additionally, the integration of real-time physiological monitoring through wearable devices stands out as a promising avenue, providing dynamic updates for a more nuanced risk assessment. Interdisciplinary collaboration between healthcare professionals, data scientists, and technology experts is essential to bridge the gap between technical robustness and clinical relevance. Ethical considerations should remain at the forefront, necessitating the establishment and adherence to guidelines that safeguard patient privacy. These recommendations collectively form a roadmap towards refining and implementing effective brain stroke risk prediction strategies, contributing to proactive and personalized healthcare interventions.

5.3 Limitations

The application of machine learning in brain stroke risk prediction, while promising, is not without its limitations. One significant constraint lies in the dependence on available datasets, which may be limited in size and diversity. The accuracy and generalizability of machine learning models heavily rely on the quality and representativeness of the training data. Additionally, ethical concerns surrounding data privacy and security pose challenges, particularly when dealing with sensitive health information. There is a potential risk of biases in the data, which can lead to skewed predictions and affect the model's performance, especially if certain demographic or socioeconomic groups are underrepresented. Furthermore, the interpretability of machine learning models remains a challenge, making it difficult for healthcare professionals to trust and understand the reasoning behind specific predictions. Real-world validation and integration of these models into clinical settings are also hindered by the need for robust validation studies and the potential reluctance of healthcare practitioners to adopt novel technologies. These limitations highlight the complexities and considerations that must be addressed for the effective implementation of machine learning in brain stroke risk prediction.

5.4 Future Research

Future studies on machine learning-based brain stroke risk prediction have enormous potential to further knowledge and enhance preventative measures. Using cutting-edge imaging methods to capture complex brain patterns and structural connectivity, such as diffusion tensor imaging (DTI) or functional magnetic resonance imaging (fMRI), is one line of inquiry. This could improve the algorithms' ability to predict outcomes by adding precise neuroimaging data.

Longitudinal studies are also required to evaluate the long-term efficacy of machine learning models and monitor changes in risk factors over time. A more dynamic and individualised risk assessment can be achieved by having a better understanding of how risk factors change over time.

The incorporation of genetic data into machine learning models represents another promising direction. Identifying genetic markers associated with stroke risk may provide valuable insights into underlying genetic predispositions, enabling a more comprehensive understanding of individual susceptibility.

Exploring ensemble learning approaches, where multiple models are combined, could also be beneficial. Ensemble methods can potentially enhance model robustness and generalizability by leveraging the strengths of different algorithms.

To address the interpretability challenge, future research should focus on developing explainable AI techniques specific to brain stroke risk prediction models. Transparent models are crucial for gaining the trust of healthcare professionals and ensuring the effective implementation of these predictive tools in clinical settings.

Lastly, collaborative efforts between researchers, healthcare practitioners, and regulatory bodies are essential. Establishing standardized protocols for model validation, ethical considerations, and data privacy will pave the way for the responsible and widespread adoption of machine learning in brain stroke risk prediction. Future studies should aim to bridge the gap between research findings and practical implementation, ultimately contributing to improved patient outcomes and stroke prevention strategies.

6. References

- Akter, B. et al. (2022) 'A Machine Learning Approach to Detect the Brain Stroke Disease', IEEE
- Ashrafuzzaman, M., Saha, S. and Nur, K. (2022) 'Prediction of Stroke Disease Using Deep CNN Based Approach', Journal of Advances in Information Technology
- Biswas, N. et al. (2022) 'A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach', Healthcare Analytics (New York, N.Y)
- Dritsas, E. and Trigka, M. (2022) 'Stroke Risk Prediction with Machine Learning Techniques', Sensors (Basel, Switzerland)
- JalajaJayalakshmi, V., Geetha, V. and Ijaz, M.M. (2021) 'Analysis and Prediction of Stroke using Machine Learning Algorithms', IEEE
- Krishna, V. et al. (2021) 'Early Detection of Brain Stroke using Machine Learning Techniques', IEEE
- Kumari, R. and Garg, H. (2023) 'Interpretation and Analysis of Machine Learning Models for Brain Stroke Prediction', IEEE
- Nikita and Parashar, G. (2023) 'Brain Stroke Detection and Prediction Using Machine Learning Approach: A Cloud Deployment Perspective', IEEE
- Revathy, G. et al. (2023) 'Early Prediction of Stroke using Machine Learning', IEEE
- Shoily, T.I. et al. (2019) 'Detection of Stroke Disease using Machine Learning Algorithms', IEEE
- Srivastav, S., Guleria, K. and Sharma, S. (2023) 'Machine Learning Models for Early Brain Stroke Prediction: A Performance Analogy'
- Minhaz Uddin Emon; Maria Sultana Keya; Tamara Islam Meghla; Md. Mahfujur Rahman; M Shamim Al Mamun; M Shamim Kaiser. (2020) 'Performance Analysis of Machine Learning Approaches in Stroke Prediction' IEEE
- Tahia Tazin ,Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis , and Mohammad Monirujjaman Khan . (2021) 'Stroke Disease Detection and Prediction Using Robust Learning Approaches' Hindawi Journal of Healthcare Engineering
- Nugroho Sinung Adi; Richas Farhany; Rafidah Ghina; Herlina Napitupulu . (2021) 'Stroke Risk Prediction Model Using Machine Learning' IEEE
- Redwanul Islam; Sourav Debnath; Torikul Islam Palash. (2021) 'Predictive Analysis for Risk of Stroke Using Machine Learning Techniques' IEEE

Chrischell Lucas; Kathrina Clarisse Padrique; Mariah Christa Lansangan; Maria Jena Isabel Gusi; Marian Lubag.(2022) 'Machine Learning on Stroke Risk Prediction Systems as Complementary Technology for Neurologists: A Critical Review' IEEE

Adi, N.S. et al. (2021) 'Stroke Risk Prediction Model Using Machine Learning', IEEE

Dritsas, E. and Trigka, M. (2022) 'Stroke Risk Prediction with Machine Learning Techniques', Sensors (Basel, Switzerland)

Kurlekar, R. et al. (2023a) 'Stroke Risk Prediction Using Deep Neural Networks: Empowering Healthcare Services for Early Identification and Prevention', IEEE

Mostafa, S.A., Elzanfaly, D.S. and Yakoub, A.E. (2022) 'A Machine Learning Ensemble Classifier for Prediction of Brain Strokes', International Journal of Advanced Computer Science & Applications

Müller, Andreas C; Guido, Sarah. (2016) 'Introduction to machine learning with Python: a guide for data scientists', Beijing: O'Reilly

8. Appendices

```
File Edit Selection View Go Run Terminal Help
Restricted Mode is intended for safe code browsing. Trust this window to enable all features. Manage Learn More

app.py x
C:\Users\USER\Desktop>W5561613Athira_Beghath_Reseach_Artifacts>Web Application>app.py
1 from flask import Flask, render_template, request
2 import joblib
3 import os
4 import numpy as np
5 import pickle
6
7 app = Flask(__name__)
8 @app.route("/")
9 def index():
10     return render_template("home.html")
11
12 @app.route("/result", methods=['POST', 'GET'])
13 def result():
14     gender = int(request.form['gender'])
15     age = int(request.form['age'])
16     hypertension = int(request.form['hypertension'])
17     heart_disease = int(request.form['heart_disease'])
18     ever_married = int(request.form['ever_married'])
19     work_type = int(request.form['work_type'])
20     Residence_type = int(request.form['Residence_type'])
21     avg_glucose_level = float(request.form['avg_glucose_level'])
22     bmi = float(request.form['bmi'])
23     smoking_status = int(request.form['smoking_status'])
24
25     x = np.array([gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type,
26                  avg_glucose_level, bmi, smoking_status]).reshape(1, -1)
27
28     scaler_path = os.path.join('F:/Athira/Stroke_Prediction', 'models/scaler.pkl')
29     scaler = None
30     with open(scaler_path, 'rb') as scaler_file:
31         scaler = pickle.load(scaler_file)
32
33     x = scaler.transform(x)
34
35     model_path = os.path.join('F:/Athira/Stroke_Prediction', 'models/dt.sav')
36     dt = joblib.load(model_path)
37
38     y_pred = dt.predict(x)
39
40     if y_pred == 0:
41         return render_template("nstroke.html")
42     else:
43         return render_template("stroke.html")
44
45 if __name__ == "__main__":
46     app.run(debug=True, port=7384)
47
48
```

Fig 1: Web Application Python Code

```
File Edit View Navigate Code VCS Help Stroke_Prediction [F:/Athira/Stroke_Prediction] - home.html
app.py home.html
1 <!doctype html>
2 <html lang="en">
3
4 <head>
5     <meta charset="utf-8">
6     <meta name="viewport" content="width=device-width, initial-scale=1, shrink-to-fit=no">
7     <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/4.0.0/css/bootstrap.min.css" integrity="sha384-Iq384400WAA=O88XFPygyR44W1V7H90E233xnfCzLSA41GfW4W/dA155Xm" crossorigin="anonymous">
8     <title>Home</title>
9 </head>
10
11 <nav class="navbar navbar-dark bg-dark navbar-fixed-top">
12     <a class="navbar-brand" href="#">Stroke Risk Prediction</a>
13 </nav>
14
15 <body>
16     <div class="container-fluid">
17         <div class="col-lg-12">
18             <form action="{{ url_for('result') }}" method="post">
19                 <br>
20                 <br>
21                 <div class="form-group">
22                     <div class="row">
23                         <div class="col-lg-4 col-md-4">
24                             <label for="gender"><b>Gender</b></label>
25                         </div>
26                         <div class="col-lg-8 col-md-8">
27                             <select style="width: 100%;" class="form-control" name="gender" id="gender">
28                                 <option selected="selected">Select</option>
29                                 <option value="1">Male</option>
30                                 <option value="2">Female</option>
31                             </select>
32                         </div>
33                     </div>
34                 </div>
35                 <br>
36                 <div class="row">
37                     <div class="col-lg-4 col-md-4">
38                         <label for="age"><b>Age</b></label>
39                     </div>
40                     <div class="col-lg-8 col-md-8">
41                         <input type="text" class="form-control" id="age" name="age" placeholder="Enter your Age">
42                     </div>
43                 </div>
44                 <br>
45                 <div class="row">
46                     <div class="col-lg-4 col-md-4">
47                         <label for="hypertension"><b>Hypertension</b></label>
48                     </div>
49                     <div class="col-lg-8 col-md-8">
50                         <select style="width: 100%;" class="form-control" name="hypertension" id="hypertension">
51                             <option selected="selected">Select</option>
52                             <option value="1">Yes</option>
53                             <option value="0">No</option>
54                         </select>
55                     </div>
56                 </div>
57                 <br>
58                 <div class="row">
59
60
```

Fig 2: Web Application Home Python Code

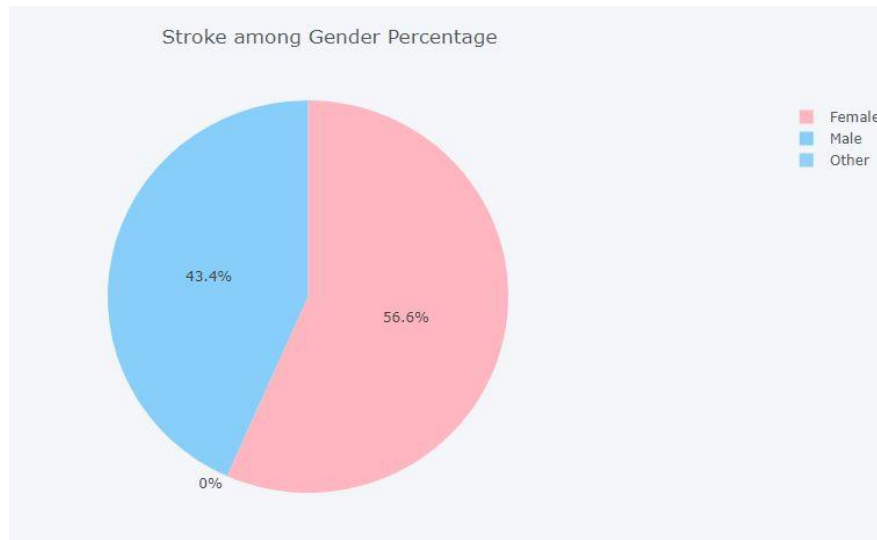


Fig 3: Percentage of Stroke among Gender

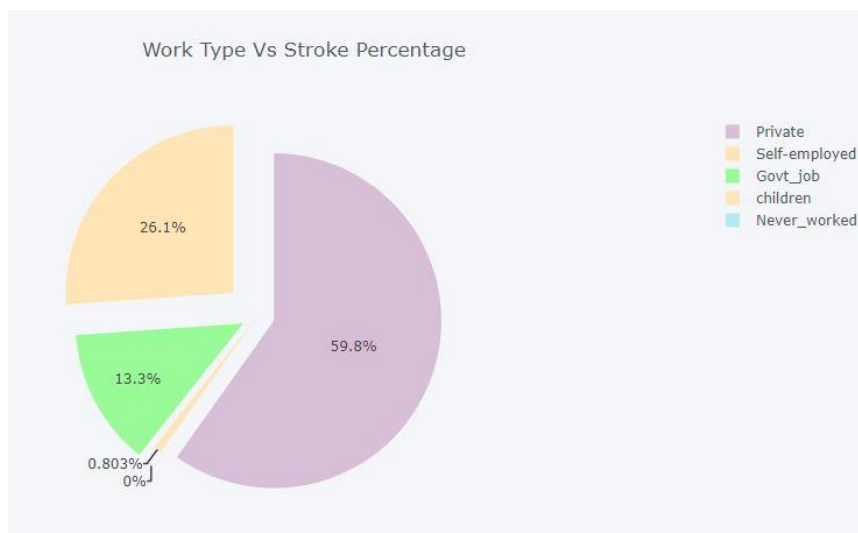


Fig 4: Percentage of Stroke among Work Type

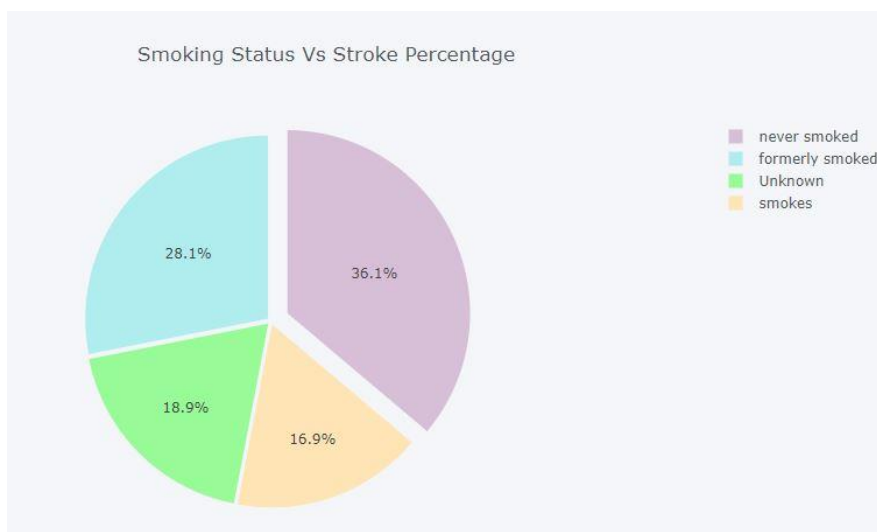


Fig 5: Percentage of Stroke among Smoking Statu

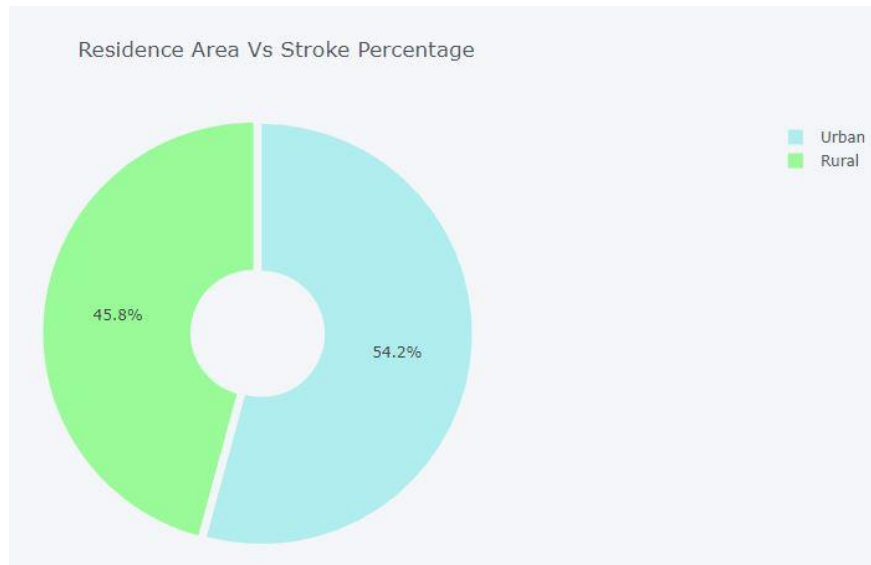


Fig 6: Percentage of Stroke among Residence Area

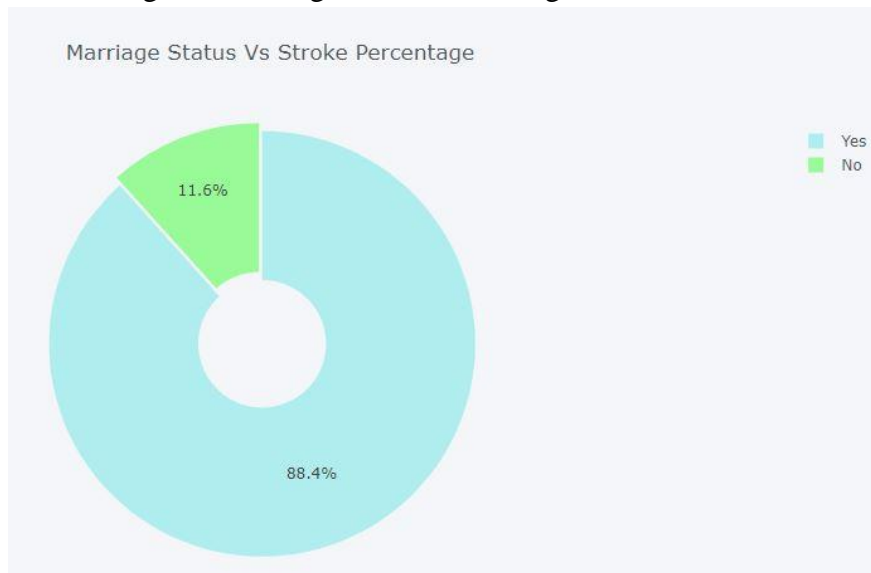


Fig 7: Percentage of Stroke among Marriage Status