

SCHOOL OF COMPUTER SCIENCE

**CASE STUDY (Weightage 30%)
JAN 2025 SEMESTER**

MODULE NAME : Statistical Inference and Modelling
MODULE CODE : ITS66804
DUE DATE : Week 11
PLATFORM : MyTIMES


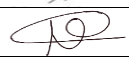



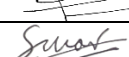
This paper consists of TEN (10) pages, inclusive of this page.

Group No: 19

Project Title:

STUDENT DECLARATION

- 1. I confirm that I am aware of the University's Regulation Governing Cheating in a University Test and Assignment and of the guidance issued by the School of Computing and IT concerning plagiarism and proper academic practice, and that the assessed work now submitted is in accordance with this regulation and guidance.*
- 2. I understand that, unless already agreed with the School of Computing and IT, assessed work may not be submitted that has previously been submitted, either in whole or in part, at this or any other institution.*
- 3. I recognise that should evidence emerge that my work fails to comply with either of the above declarations, then I may be liable to proceedings under Regulation*

No	Student Name	Student ID	Date	Signature	Score
1.	Safal Lohani	0369861			
2.	Nishchal Panta	0369948			
3.	Pragyan Shrestha	0370057			
4.	Rashik Thapa Magar	0370036			
5.	Runal Shrestha	0369972			
6.	Swornim Balla	0369952			

Marking Rubrics

Group Assignment Marking Rubrics					
Abstract (5 marks)	<p>5 marks A clear and concise abstract that gives the reader a clear idea of what the project is about and why it is interesting. The following components need to be included</p> <p>i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion</p>	<p>4 marks A clear abstract that gives the reader a clear idea of what the project is about. Four of the following components are included</p> <p>i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion</p>	<p>The abstract is difficult to read and/or is very vague and/or doesn't sell the project as well as it might have. Three of the following components are included</p> <p>i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion</p>	<p>2 marks Unable to read the abstract and/or is very vague and/or doesn't sell the project as well as it might have. Only two of the following components are included</p> <p>i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion</p>	<p>1 mark Unable to read the abstract. Only one of the following components is included</p> <p>i. Purpose and motivation of this research ii. Problem you are addressing iii. Methods and materials iv. Results v. Conclusion</p>
Introduction (10 marks)	<p>9-10 marks A readable write-up that explains what the problem is and why it is of interest. The</p>	<p>7-8 marks A readable write-up that explains what the problem is. Three of the following components</p>	<p>5-6 marks The write-up is difficult to read, somewhat vague, or doesn't make a really good</p>	<p>3-4 marks Unable to read the write-up and/or is very vague. Only one of the following</p>	<p>1-2 marks Unable to read the write-up. None of the following components are included.</p>

	following components need to be	are included. i. Problem ii. Negative	case for why the problem is of interest.	components are included. i. Problem	i. Problem ii. Negative impact of the
--	---------------------------------	---	--	--	--

	included i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	impact of the problem iii. Parties affected iv. Benefit of solving the problem	Two of the following components are included. i. Problem ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	ii. Negative impact of the problem iii. Parties affected iv. Benefit of solving the problem	problem iii. Parties affected iv. Benefit of solving the
--	---	--	--	---	--

Literature Review (20marks)	18-20 marks An outstanding overview, with an insightful analysis of prior work and a clear connection between prior work and the proposed method. The following components are given. (8 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	15- 17 marks A comprehensive overview of prior work that gives the reader a clear idea of what's out there and how the proposed method is different. Four of the following components are given. (6 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	10-14 marks A fairly good overview of prior work, and some connection is made to the proposed method. Three of the following components are given. (5 articles) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	5-9 marks An overview of several papers related to the proposed method, and some attempt is made to connect the prior work to the current method. Two of the following components are given. (4 marks) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review	1-4 marks An overview of several related papers, but not within a coherent conceptual frame- work. One of the following components are given. (2 marks) i. Introduction of the topic ii. Taxonomy Mapping iii. Paragraphs for each branch of the taxonomy tree iv. Conclusion v. Critical Review
Data (5marks)	5 marks The data are comprehensive	4 marks The data are fairly	3 marks The data are not	2 marks The explanations	1 mark The explanations

	<p>and clearly described. At least 6 of the following components are given.</p> <p>i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion</p>	<p>explained. At least 5 of the following components are given.</p> <p>i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion</p>	<p>comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given.</p> <p>i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion</p>	<p>are significantly flawed. At least 3 of the following components are given.</p> <p>i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion</p>	<p>are flawed. At least 2 of the following components are given.</p> <p>i. Source of the data ii. Description of the data and its context iii. Statistics of the data iv. Presentation, visualization and quantification of the data and images v. Conclusion</p>
Method (20 marks)	<p>17-20 marks The methods of analysis are comprehensive and clearly described. At least 6 of the following components are given.</p> <p>i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis methods iv. Statistical methods address the research v. Information</p>	<p>13-16 marks The methods of analysis are fairly explained. At least 5 of the following components are given.</p> <p>i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis methods iv. Statistical methods address the research objective v. Information</p>	<p>9-12 marks The methods of analysis are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given.</p> <p>i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis methods iv. Statistical methods</p>	<p>5-8 marks The methods of analysis are significantly flawed. At least 3 of the following components are given.</p> <p>i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis methods iv. Statistical methods</p>	<p>1-4 marks The methods of analysis are flawed. At least 2 of the following components are given.</p> <p>i. Explanatory data analysis ii. Statistical data analysis methods iii. Appropriate data analysis methods iv. Statistical methods address the research objective</p>

	objective v. Information on data analysis process vi. Clear relationship between methods	on data analysis process vi. Clear relationship between methods	address the research objective v. Information on data analysis process vi. Clear relationship between methods	address the research objective v. Information on data analysis process vi. Clear relationship between methods	v. Information on data analysis process vi. Clear relationship between methods
Result & Discussion (20 marks)	17-20 marks The results are comprehensive and clearly described. At least 6 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	13-16 marks The results are fairly explained. At least 5 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	9-12 marks The results are not comprehensive and/or there is a flaw in the explanation. At least 4 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the	5-8 marks The results are significantly flawed. At least 3 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question	1-4 marks The results are flawed. At least 2 of the following components are given. i. Subheadings are included and are clear and informative ii. Figures and tables are supported by text iii. Correct interpretation of the results iv. Results with tables and diagrams v. Additional insight to the content vi. Critical analysis of the results vii. Clearly addresses the research question

			research question		
Limitation and future Study (10 marks)	9-10 marks An insightful and correct analysis. The following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	7-8 marks A correct analysis that could be more complete and is not very insightful. One of the following components is missing. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	5-6 marks An incomplete or somewhat incorrect analysis. Two of the following components are missing. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	3-4 marks An incorrect analysis. One of the following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies	1-2 marks No analysis. None of the following components are given. i. Discussion addresses the major finding of the study ii. Results are interpreted with respect to outside sources iii. Identify the limitation or limitations iv. Explain these limitations in detail v. Propose a future direction for future studies
Conclusion (5 marks)	5 marks A clear and insightful summary of the paper, perhaps with interesting ideas for future work. The following components are given. i. Restate your research topic	4 marks A summary of the experiments is given, but the conclusion is a mere summary. The ideas for future work are not interesting. One of the following components is missing	3 marks A flawed conclusion. Two of the following components are missing. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics	2 marks An incorrect conclusion. Three of the following components are missing. i. Restate your research topic ii. Restate the objective iii. Summarize	1 marks No conclusion. One of the following components is given. i. Restate your research topic ii. Restate the objective iii. Summarize

	ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts	missing. i. Restate your research topic ii. Restate the objective iii. Summarize the main topics iv. Significance of results v. Conclude the thoughts	iv. Significance of results v. Conclude the thoughts	the main topics iv. Significance of results v. Conclude the thoughts	the main topics iv. Significance of results v. Conclude the thoughts
Format (5 marks)	5 marks A clear and correct formatting. The following components are given. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	4 marks A clear and correct formatting. One of the following components is missing. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	3 marks Two of the following components are missing. i. Number of pages 10 -5 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	2 marks Three of the following components are missing. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file	1 marks One of the following components is given. i. Number of pages 10 -15 ii. Use the correct template iii. Similarity index less than 20% iv. All the sections given in proper order v. Readable pdf file

Table of Contents

Marking Rubrics.....	2
Abstract.....	11
Introduction.....	12
Literature Review.....	14
Data.....	17
Methods.....	21
Results and discussion	24
Initial Exploratory Data Analysis.....	24
Data Cleaning.....	25
Further Exploratory Data Analysis	26
Methods Implementation	29
Data transformation	29
Model Deployment	31
Model Validation and Comparison	32
Critical analysis of the results	33
Addressing the Research Question	34
Limitations and Future Study.....	34
Conclusion	36
Appendixes	37
References.....	37

Table of figures

Figure 1: Data visualization of marital Status, gender, work type and residence.....	18
Figure 2: Data visualization of bmi, smoking status, hypertension, heart disease, glucose level and age	19
Figure 3:Stroke Histogram.....	20
Figure 4: Distribution of categorical data types.....	24
Figure 5:Histogram of Numerical Variables.....	25
Figure 6: Data Cleaning	26
Figure 7: Smoking Status Distribution after data cleaning	26
Figure 8: Boxplot of average Glucose Level and BMI.....	27
Figure 9:Stroke Rate by Hypertension, Heart Disease, Smoking Status, Marital Status, Work type	28
Figure 10: Correlation Heatmap	29
Figure 11: Label Encoding.....	29
Figure 12: Feature Engineering	30
Figure 13:Min_Max Scaling	30
Figure 14: Logistic Regression Implementation.....	31
Figure 15: Decision Tree Implementation	31
Figure 16: Random Forest Implementation	31
Figure 17:Confusion Matrix	32
Figure 18: Model performance Comparison.....	33

Abstract

Stroke is one of the leading reasons for mortality throughout the world. Survivors of stroke often suffer from long-term effects such as paralysis, vision loss, hearing issues, blood clots, coma, and many more. To reduce impacts of stroke early prediction is required. This study aims to employ various exploratory data analysis (EDA), statistical methods, and machine learning techniques on stroke prediction datasets to produce important insights and develop a predictive model by analyzing various health and lifestyle-related factors.

Stroke Prediction Dataset from Kaggle was used for this research. Various data preprocessing, visualization, statistical analysis, and Machine learning models, including logistic regression, decision tree, and random forest were applied and valuable insights and comparisons were generated. While valuable insights were discovered and high-accuracy models were prepared in this model, further work in the future is needed to get a better understanding of strokes and to produce a better stroke prediction model. The inclusion of personalized data is required for more practical healthcare applications.

This study highlights the importance of data-driven, statistical, and machine-learning techniques in the healthcare sector. This study also demonstrates how statistical methods and machine learning techniques can be used to accurately predict the occurrence of stroke. Future work in this field should focus on incorporating more personalized data into the dataset from a diverse population. Techniques like deep learning and neural networks should also be used in the future to produce a more practically applicable solution.

Introduction

Health is the most important asset for any human being. No number of resources can be useful to humankind without having good health. With advancements in technology data data-driven approaches and statistical modeling have become essential to identify or predict possible health issues and risks. With the use of these methods and techniques, healthcare professionals can predict possible risks and provide timely solutions and medications to try and eliminate the risks.

Stroke is a medical condition in which a part of the brain is deprived of blood flow, which causes damage to the brain tissue. Stroke can cause brain damage, physical disability, or even death. Stroke is one of the leading causes of mortality throughout the world, due to this there is a need for the usage of prevention strategies and regular research.

Problem Statement

Stroke is a serious medical condition that can lead to brain damage, physical disability, or even death. It causes about 6.5 million deaths every year. Survivors of stroke often suffer from long-term effects such as paralysis, vision loss, hearing issues, blood clots, coma, and many more. Due to the effects of stroke, a person or their family may face multiple financial, mental, and physical burdens. Delayed Diagnosis of stroke or lack of awareness can lead to complicated future issues. Despite numerous advancements stroke prevention remains challenging because various factors such as age, diabetes level, lifestyle habits, and blood pressure levels need to be considered while making such a critical prediction.

Negative Impact of the Problem

Strokes affect not just individuals but also families, the healthcare system, and the economy. A lot of people die from stroke every year. Even the survivors face various long-term conditions like mobility issues, hearing issues, vision loss, and many more. Financial burdens caused by treatment, recovery, rehabilitation, and caregiving process can place significant strain on families and the healthcare system. Delays in diagnosis and lack of awareness can result in preventable deaths and disabilities.

Parties Affected

Stroke affects various people. Particularly affects older adults, people who live an unhealthy lifestyle, and those individuals who have pre-existing health conditions. Strokes can also place emotional and financial burdens on the individual and family members. Stroke also affects the healthcare system as more resources need to be used to deal with the effects of stroke.

Benefits of Solving the Problem

Strokes are a serious medical condition. Timely identification and mitigation of stroke-related risk factors can change many lives for the better. Analysis of stroke-related data and the development of a statistical model for stroke prediction can help in timely intervention, reducing the occurrence of strokes and the severity of their effects. This study aims to analyze stroke-related datasets to discover patterns and use statistical methods to provide valuable stroke-related insights to the healthcare system and various stakeholders, which will lead to a better understanding of stroke awareness. This will ultimately result in enhanced prevention strategies, improved healthcare services, and reduced mortality rates.

Literature Review

Stroke is a medical condition in which a part of the brain is deprived of blood flow, which causes damage to the brain tissue. Stroke is one of the leading causes of mortality throughout the world. Even the survivors must suffer from brain damage, hearing issues, vision issues, and physical damage. In recent times, statistical modeling and data-driven approaches have enabled healthcare professionals to better understand strokes and identify the factors that can possibly lead to one. A robust fertilizer prediction model can help prevent many deaths and disabilities. Various research and studies have been conducted to better understand strokes and ways to detect strokes in an early stage.

Taxonomy Mapping

Many literatures were reviewed for the ideation of this project. The literature reviewed can be categorized into:

1. Understanding Stroke and its causes
2. Application of Machine Learning in Stroke prediction
3. Impact on healthcare system

Understanding stroke and its causes

"What is a Stroke?" Published by Stroke Recovery Association NSW provides an overview of stroke. This paper discusses and defines what a stroke is and classifies it into two different types: Ischemic stroke and Hemorrhagic stroke. This paper also highlights what Transient Ischemic Attack (TIA) is. This research underscores the necessity of public awareness and lifestyle modifications in stroke prevention. (What is a Stroke?, 2003)

"Stroke and Etiopathogenesis: What Is Known?" explores the impacts of genes on strokes. This paper dives deep into genetic factors including gene-specific mutations which can help in the assessment of stroke-related risk at a personal level. This study has also highlighted how incorporating genetic factors into machine learning models can lead to early detection of stroke. (Ciarambino, Crispino, Mastrolorenzo, Viceconti, & Giordano, 2022)

"Stroke: causes and clinical features" demonstrates how stroke can be related to lifestyle and environmental risk factors. This disease identifies conditions such as diabetes, hypertension, and

heart disease that can significantly increase the risk of getting a stroke. This study also suggests that lifestyle should be taken into account while making an assessment for stroke. (Murphy & Werring, 2020)

Application of Machine Learning in Stroke prediction

“A predictive analytics approach for stroke prediction using machine learning and neural networks” dives into the negative impact of stroke and provides insights into efforts to improve the diagnosis of stroke. This paper highlights which factors are most important in predicting strokes. This paper concludes that age, heart disease, average glucose level, and hypertension were the most important factors in detecting strokes. In this study, decision trees, neural networks, and random forests were used to analyze which gives a better understanding of stroke predictions. (Dev, et al., 2022)

“Performance Analysis of Machine Learning Approaches in Stroke Prediction” investigates the effectiveness of stroke predictions using various machine learning approaches. Hypertension, body mass index level, heart disease, average glucose level, smoking status, previous stroke, and age were considered while applying machine learning techniques. Using these attributes this paper applies 10 different models to predict strokes. Then every model was compared in this study. This study presents how some methods could achieve 97% accuracy which highlights the possibility of incorporating these methods into the healthcare system.(Emon, et al., 2020)

“An Integrated machine learning approach to Stroke Prediction” explores how machine learning methods such as SVMs, decision trees, and random forest perform in stroke classification. This study highlights how these models effectively predicted the occurrence of stroke. This model highlights the importance of the selection and interpretation of features in improving model efficiency. This paper highlights how machine learning principles can be incorporated into real-world health care systems. (Khosla, et al., 2010)

Impact on healthcare system

“Machine Learning in Stroke Medicine: Opportunities and Challenges for Risk Prediction and Prevention” examines how strokes impact the health care system. This paper highlights the opportunities and challenges of machine learning in stroke medicine. This study underscores how stroke-related cases account for a large financial burden on Individuals, families, and the healthcare system. This study suggests that the integration of machine learning techniques in healthcare can relieve these burdens by early detection and preventive measures. (Amann, 2021)

Conclusion

Reviewed literature highlights that significant progress has been made in understanding strokes using machine learning techniques. Prior studies have shown that the integration of machine learning techniques into healthcare systems can be impactful and can certainly help stroke prediction. Our project aims to produce a model that can handle class imbalance by utilizing models that can be visualized or interpreted such as decision trees and logistic regression. This work will provide insight into what factors are important while producing and analyzing stroke data and will try to incorporate a statistical model for early predictions and detection of stroke.

Critical review

The literature on stroke and stroke predictions highlights how statistical modeling and machine learning can be used to identify risk factors and predict possible strokes. While these models present high accuracy, they are often limited by class imbalances and lack real-world practicality. Most literature, while having high accuracy, is difficult to practically implement as the dataset they have used does not account for more personalized data such as genetic factors and lifestyle. While this project does not employ any personalized datasets, it would rather produce valuable exploratory data analytics and employ easily interpretable models. However, in the future, larger datasets and more personalized data, such as genetic factors, can be incorporated to produce more accurate and practically reliable models.

Data

Based on the review literature and taking our technical limitations into consideration, a dataset with a comprehensive collection of health-related and lifestyle features was needed. To fulfill these requirements, we have chosen **Stroke Prediction Dataset from Kaggle**.

Source of the data

Source: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Data Description and Context

This dataset consists of health-related information about individuals. This dataset includes 12 columns and 5110 rows. It includes 11 independent variables or features and one target variable. Features of this dataset include important health and lifestyle-related factors such as age, gender, hypertension, heart disease, average glucose level, BMI, and smoking status.

This project aims to analyze and apply statistical methods and machine learning to analyze any relations between features and stroke using EDA techniques and to produce a model that can accurately predict the probability of having a stroke. As this dataset contains relevant health and lifestyle information, it is suitable for this project.

Basic Data Statistics

Number of Rows: 5,110

Number of Columns: 12

Features	Target
id	stroke: 0 (no stroke), 1 (stroke)
gender: Male, Female, or Other	
age	
hypertension:0 (no hypertension), 1 (hypertension)	

heart_disease: 0 (no heart disease), 1 (heart disease)	
ever_married: Yes or No	
work_type: Type of work	
Residence_type: Urban or Rural	
avg_glucose_level: Average glucose leve	
bmi: Body Mass Index	
smoking_status: Smoking status (formerly smoked, never smoked, smokes, unknown)	

Data visualization

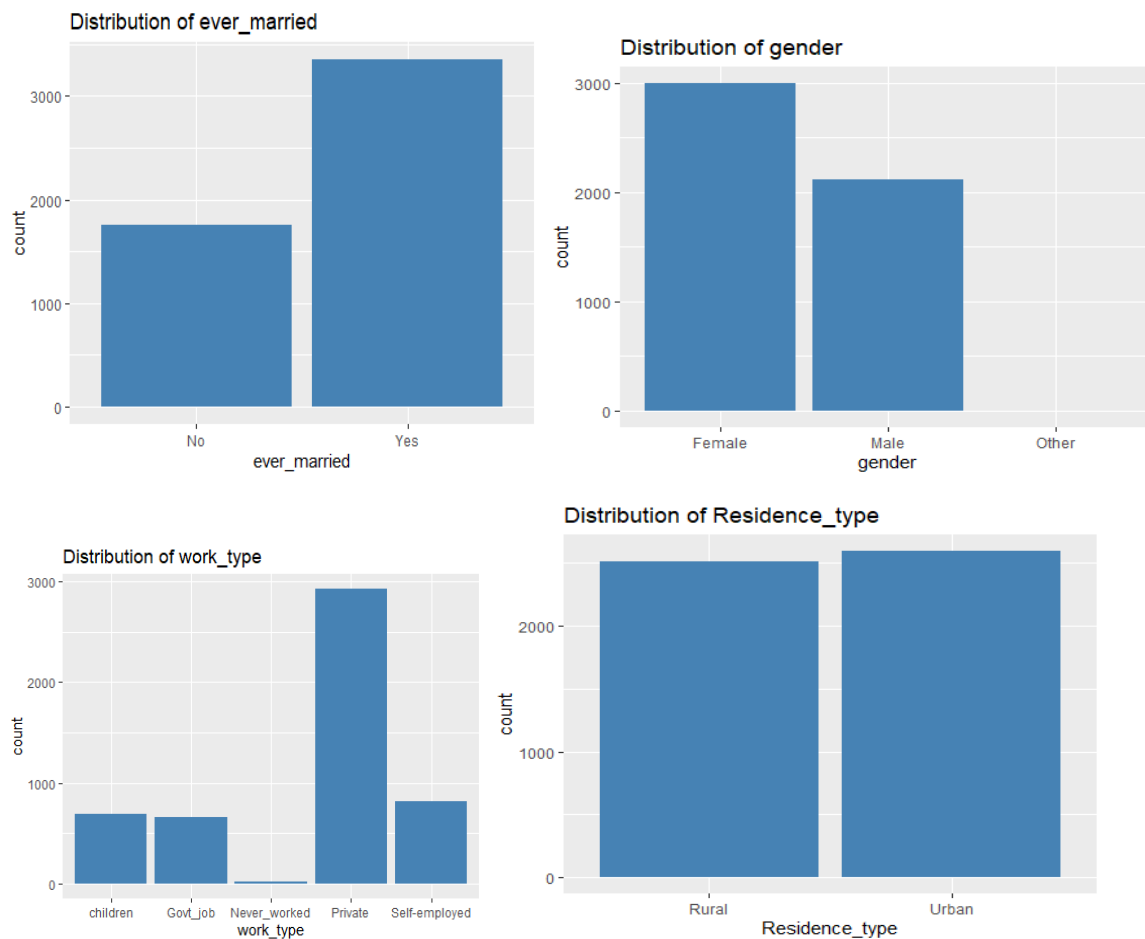


Figure 1: Data visualization of marital Status, gender, work type and residence

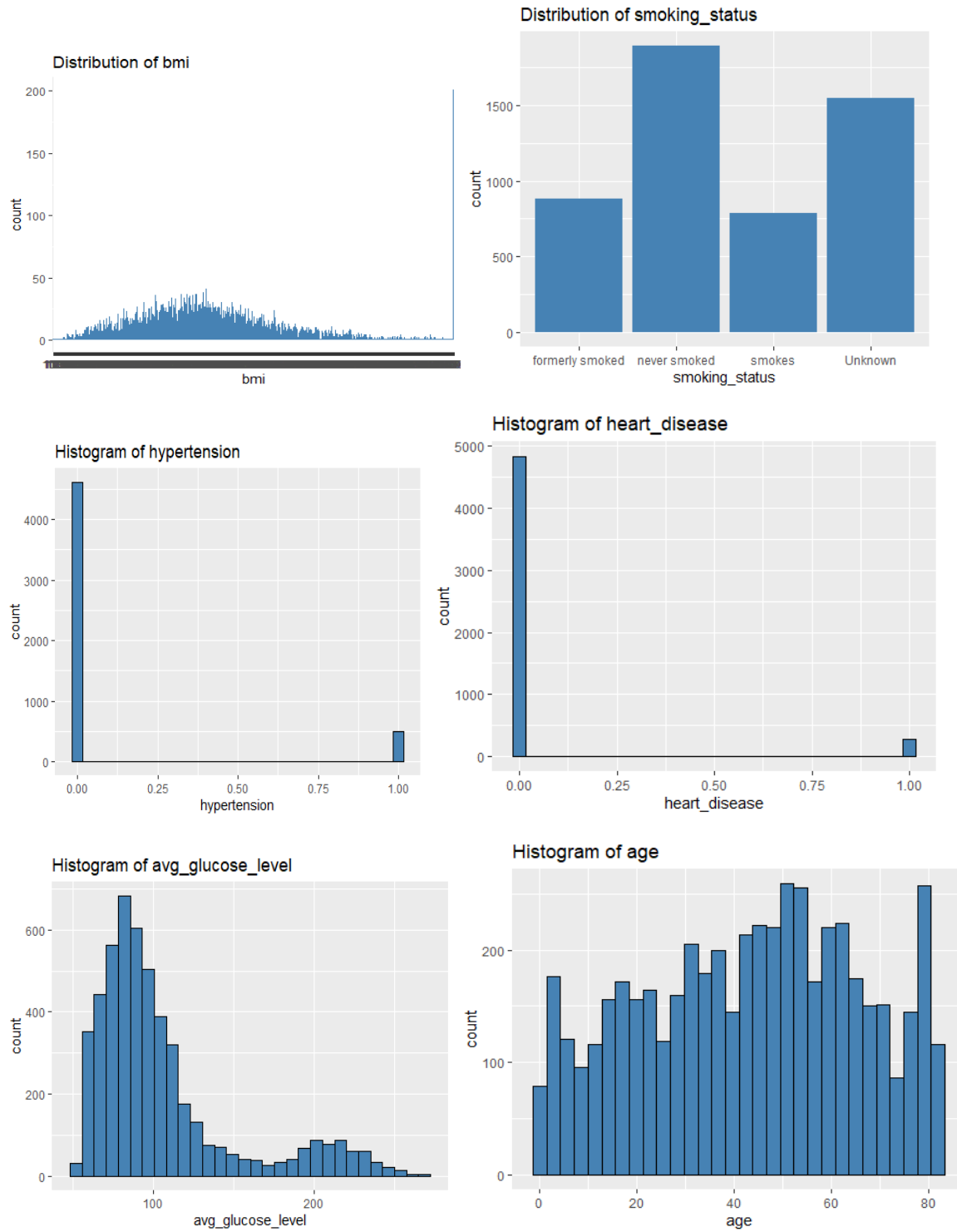


Figure 2: Data visualization of bmi, smoking status, hypertension, heart disease, glucose level and age

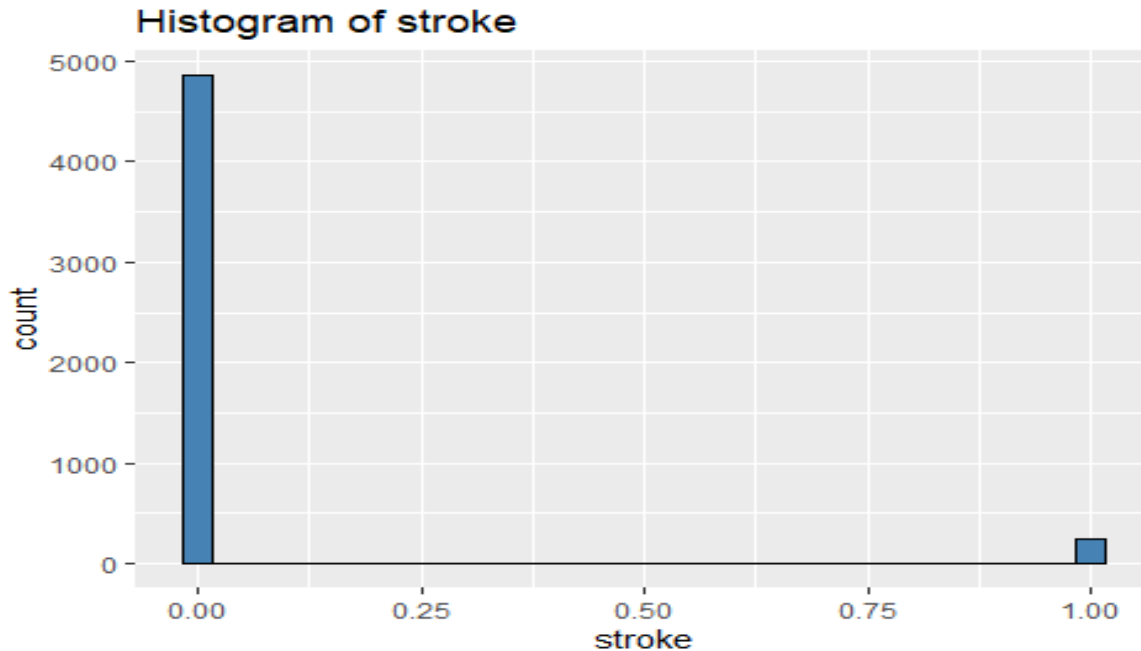


Figure 3:Stroke Histogram

Quantification

This dataset has only 249 instances of stroke and 4861 cases of no stroke which highlights a class imbalance in this dataset. The average age of this dataset is 43.2 years, which indicates that most of the people in this data are adults.

The Stroke Prediction Dataset is a well-structured dataset with important features regarding health-related factors, age, and lifestyle-related factors. Although its limitations of having less data and not having personalized data such as genetic information we can generate valuable insights by analyzing the data. By analyzing this dataset researchers can identify patterns and develop models that are critical for stroke prediction and prevention.

Methods

In this project, we will employ various Exploratory Data analysis techniques, data cleaning methods, high-level statistics, machine learning techniques, and statistical modeling methods to analyze and produce insights from this dataset using R. The methods we will use for this project are:

- Exploratory data analysis
- Statistical Data Analysis methods
- Data analysis and data cleaning
- Statistical Methods
- Information on data analysis process
- Relation between Method and Research Objectives

Exploratory Data Analysis

Exploratory Data Analysis is a process in which we analyze datasets by creating various visualizations and generating high-level statistical summaries. The primary goal of this step is to identify any patterns, correlations, and missing values and understand the distribution of the variables. This step provides a robust roadmap before applying statistical models.

For this project, we have used various libraries such as skimr, tidyverse, rpart, ggplot etc to create meaningful visualizations. Some of the visualizations that were created are:

- Bar plot for categorical data. This plot was made to visualize distribution of various categorical data.
- Histograms for numerical data. This plot visualizes the skewness of the dataset and also gives us an idea of the range of data we are dealing with.
- Distribution of smoking status after data cleaning.
- Smoking status by age group.
- Box plot of glucose and bmi level to detect and analyze outliers
- Point plot of glucose vs bmi colored by stroke
- Stroke rate by hypertension, heart disease, Smoking status, marital status, type of work
- Correlation heatmap between various numerics factors of this dataset

Using these graphs and figures we can find various correlations, patterns, missing data etc.

Statistical Data Analysis Methods

Statistical data analysis methods are used to describe, summarize, and interpret data. In this method, we will draw valuable conclusions from the data. In this study the following statistical techniques are employed:

- Descriptive statistics of the data set to summarize numerical and categorical data using metrics such as mean, median, standard deviation, and mode.
- Using EDA, we will derive insights that will help us to identify correlations, outliers, and other patterns.

Data Analysis, Cleaning and Preprocessing

Data will be analyzed and cleaned to make sure that further statistical methods can be employed in this step. This includes the following step:

- Missing or unknown values handling: We will handle any N/A and Unknown values by either imputing or removing them
- Feature selection: Unnecessary columns such as id will be removed
- Label encoding will be performed on the data to make sure it is compatible with further statistical and machine learning models.

Statistical Methods

We will be using statistical and machine learning methods to properly examine and identify relationships between a dependent variable and other independent variables. With application and validation of these methods, we will also get a better understanding of how accurate prediction can be made using stroke-related data. The following methods will be used:

- **Logistic regression:** Logistic regression is a method applied to that can classify or predict the probability of an outcome such as (yes/no) or (0/1). It is simple and easy to implement and interpret.
- **Decision Tree:** Decision tree is a machine learning technique that can be used for both classification and regression. In this model, there are root nodes, branches, internal nodes, and leaf nodes. This model is simple to interpret.

- **Random Forest:** Random forest is a technique in which multiple decision trees are combined. For classification, the output produced by random forest is the output that is selected by most of its trees. This model is a bit complex to implement and interpret compared to logistic regression and decision tree. This method produces better model robustness, and higher accuracy and mitigates overfitting issues.

Information on Data Analysis Process

The entire data analysis processes applied in this project are as follows:

- Data collection: Stroke Dataset was loaded to perform various EDA, visualization, and statistical methods.
- Exploratory Data Analysis (EDA): Various visualizations were created to understand the dataset better and to identify any patterns if possible.
- Data Cleaning & Preprocessing: Missing values and unknown values were handled by imputation and removal. Variables were encoded to make it easier to apply various statistical methods.
- Feature engineering: Unnecessary features were removed to make it easier to implement various models and methods.
- Model development: Logistic Regression, Decision Tree, and Random Forest were used to create models and analyze data.
- Model Evaluation: Models developed were analyzed and validated based on confusion matrix, accuracy, precision, recall, and F1-Score.

Relationship between Methods and Research Objectives

EDA and correlation analysis will help in identifying stroke related risk factors. This will also give us a better insight into how various factors such as age, hypertension, heart disease, bmi, glucose levels and other lifestyle factors impact the probability of getting a stroke. Statistical Methods and machine learning techniques trained on stroke prediction dataset will give us an insight of how accurately stroke can be predicted. These insights can be helpful in early diagnosis and prevention of stroke.

Results and discussion

Initial Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a process that provides a better understanding of datasets. Pattern identification, insight gathering, and visualization preparation are carried out in this step. We will be using R libraries such as tidyverse, ggplot, skimr, reshape2 will be used to create a visualization of this dataset.

Initially, we are going to create distribution graphs for categorical data types and histograms for numerical data types to understand data distribution.

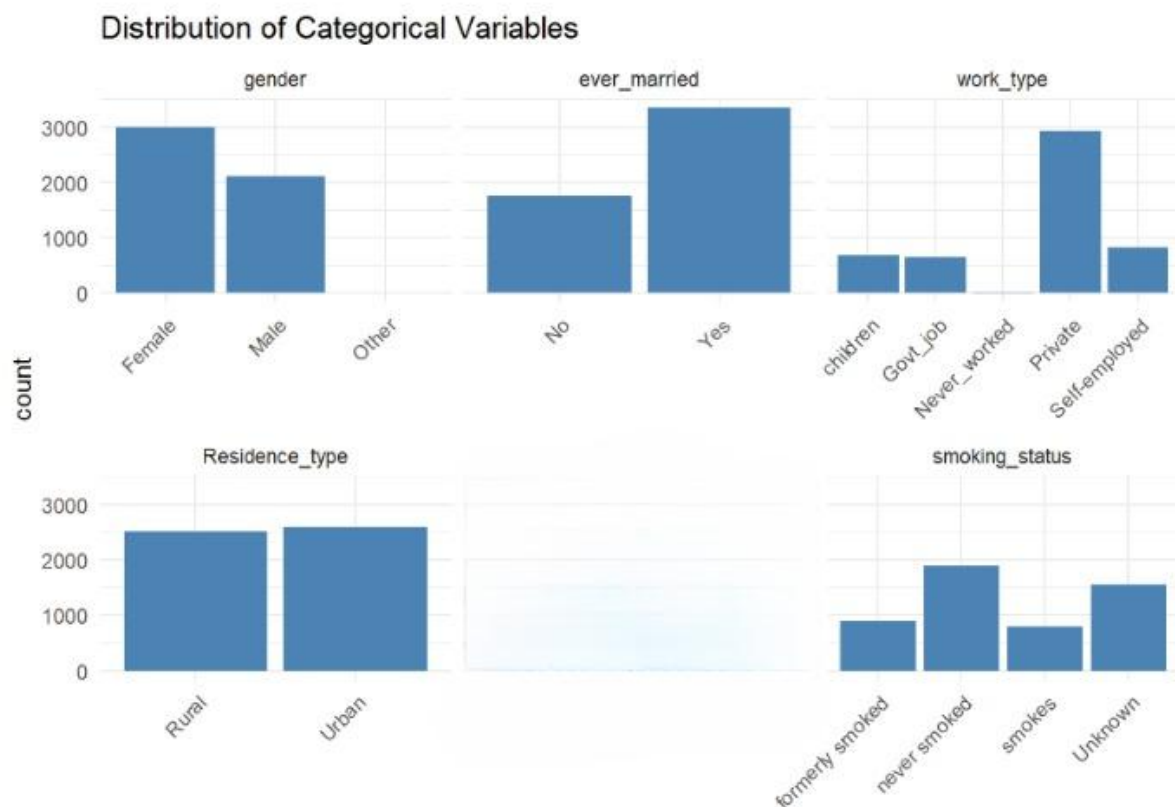


Figure 4: Distribution of categorical data types

Insight: Most of the individuals in this dataset are females. Most of them are also married and work in a private company. The type of residence is balanced. There is a high number of people whose smoking status is unknown which might require imputation to analyze the dataset better.

Histogram of Numerical Variables

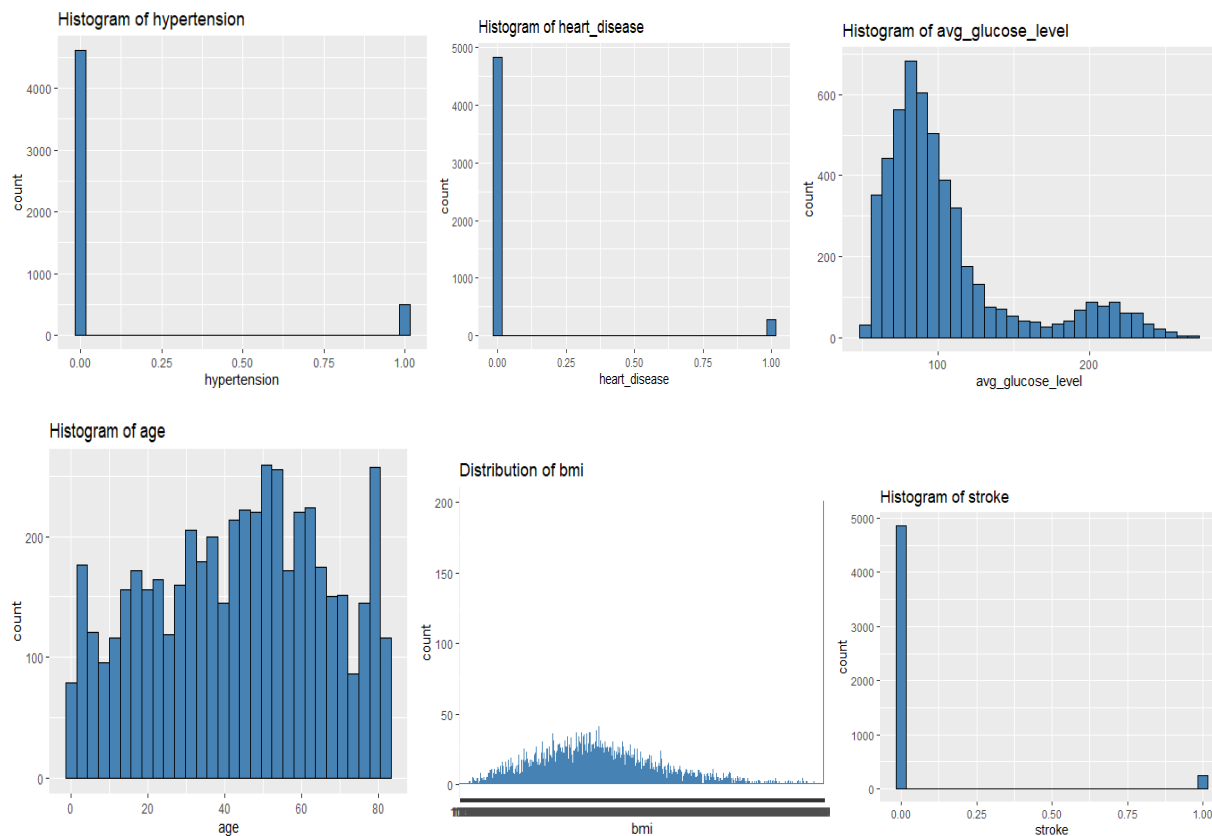


Figure 5: Histogram of Numerical Variables

Insight: This dataset includes people from various age groups but most of them are adults. Hypertension, heart disease, and stroke are denoted by 0 and 1 in. (1 = Yes, 0 = No) We can see there is a class imbalance in hypertension heart disease, and stroke. People with differing levels of glucose are included in this dataset.

Data Cleaning

Before Moving further into EDA, we will identify any N/A and unknown variables and either remove or impute them.

```
miss_scan_count(data = data, search = list("N/A", "Unknown"))
```

This code was used to Identify and list any N/A and Unknown existing in the dataset.

After running this code, we identified that there were **201 N/A values in BMI** and **1544 Unknown values** in smoking status.

201 N/A values in BMI were removed as they would have minimal impact on the data integrity. Bmi will also be converted to numeric datatypes. To treat **1544 Unknown Smoking status**, the most prevalent data in a specific age group were imputed. The following code was used to do so:

```
data_clean <- data[data$bmi != "N/A", ]
data_clean$bmi <- as.numeric(data_clean$bmi)

data_clean <- data_clean %>%
  mutate(age_group = cut(age, breaks = c(0, 20, 40, 60, 80, 100),
    labels = c("0-20", "20-40", "40-60", "60-80", "80-100"),
    include.lowest = TRUE))

# Find the most common smoking status for each age group
mode_by_age_group <- data_clean %>%
  filter(smoking_status != "Unknown") %>%
  group_by(age_group, smoking_status) %>%
  summarise(count = n(), .groups = "drop") %>%
  arrange(age_group, desc(count)) %>%
  group_by(age_group) %>%
  filter(row_number() == 1) %>% # Select the most common smoking status
  ungroup() %>%
  select(age_group, smoking_status)

# Merge mode information back into the dataset
data_clean <- data_clean %>%
  left_join(mode_by_age_group, by = "age_group", suffix = c("", "_mode")) %>%
  mutate(smoking_status = ifelse(smoking_status == "Unknown",
    smoking_status_mode, smoking_status)) %>%
  select(-smoking_status_mode)
```

Figure 6: Data Cleaning

Further Exploratory Data Analysis

Distribution of smoking status after data cleaning.

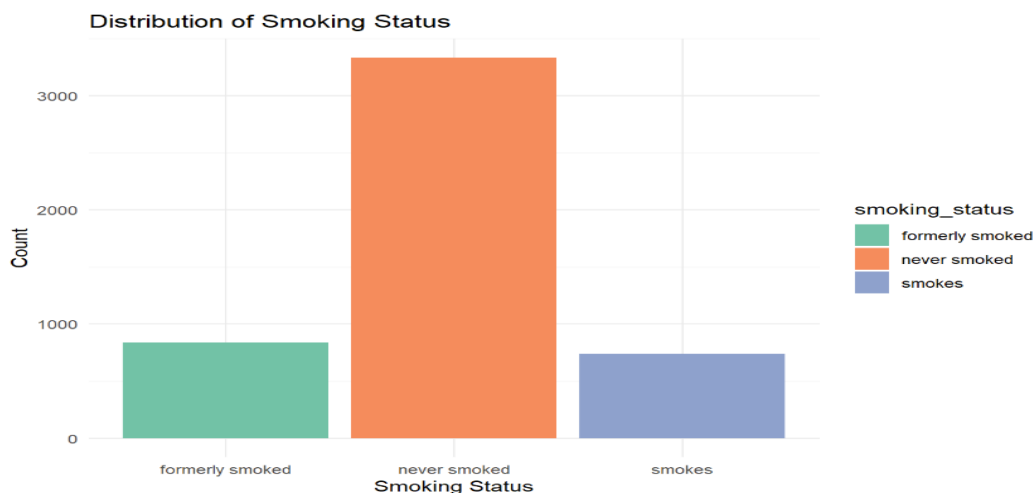


Figure 7: Smoking Status Distribution after data cleaning

Insight: After imputation the number of people who never smoked became very high.

Box Plot of Average Glucose Level And BMI

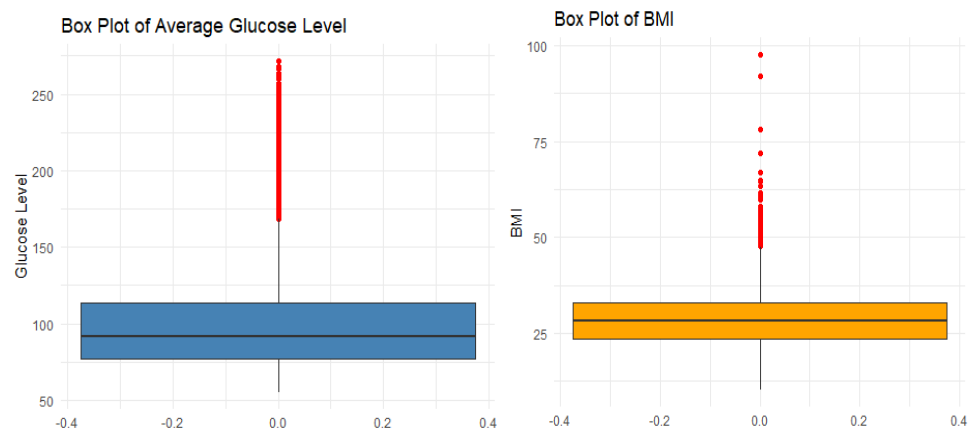
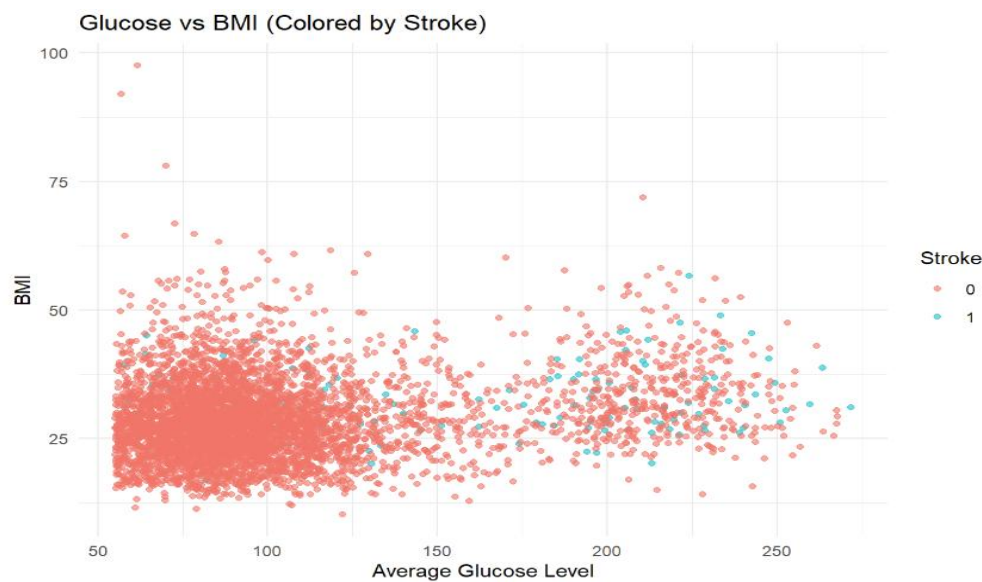


Figure 8: Boxplot of average Glucose Level and BMI

Insights: From this boxplot, we can see the spread of data on Glucose Level and BMI. From this plot, we can see that multiple outliers exist in Glucose Level and BMI. But these outliers do not require removal as they might represent real-world data as some people might have higher glucose levels and BMI.

Point plot of Average Glucose level and BMI



Insights: From this plot, we can see that individuals with higher BMI and Glucose levels are more likely to be affected by stroke. From this, we can see that there is a correlation between lifestyle and the probability of getting a stroke.

Stroke Rate by Hypertension, Heart Disease, Smoking Status, Marital Status, Work type

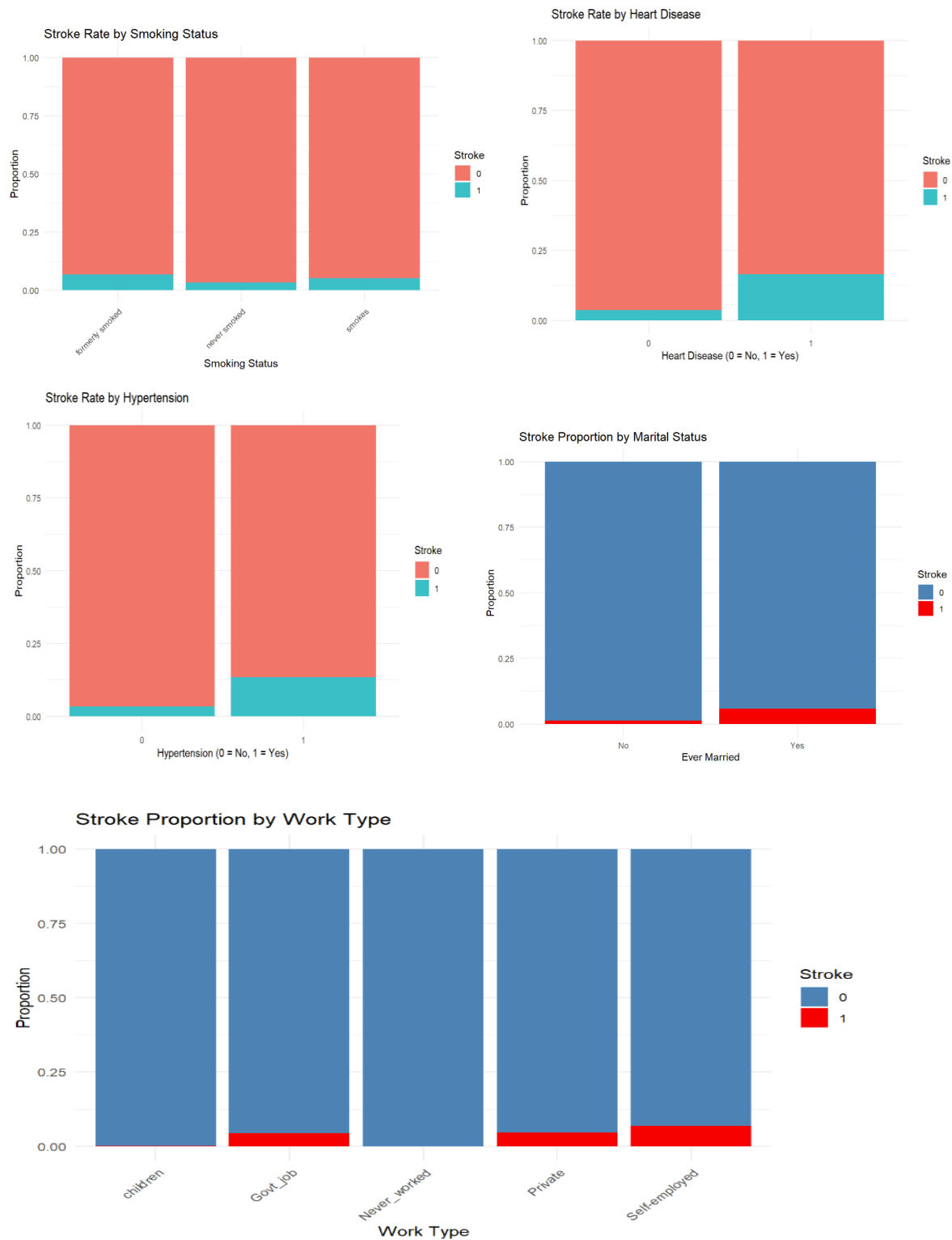


Figure 9: Stroke Rate by Hypertension, Heart Disease, Smoking Status, Marital Status, Work type

Insights: From these graphs we can extract the following insights:

- People who smoke or people who have formerly smoked are more likely to suffer from stroke.
- People who have suffered from heart disease are more likely to suffer from a stroke.
- People who suffer from Hypertension are more likely to be impacted from a stroke.
- People who are married are more likely to be impacted from a stroke. This could also be an indication that people of higher age group are more likely to get strokes.
- People who work are more likely to get stroke rather than people who have never worked. This could be an indication of a correlation between stress and stroke.

Correlation heatmap

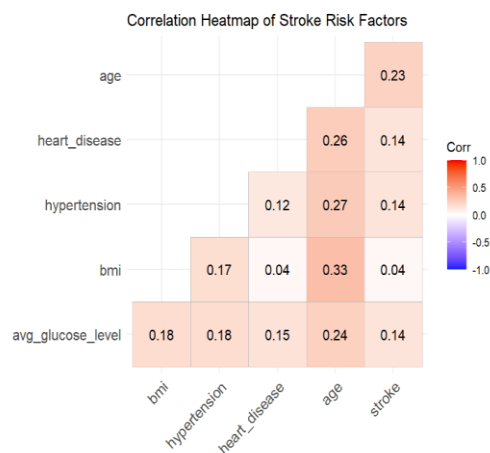


Figure 10: Correlation Heatmap

Insights: From this correlation heatmap we can see that there is a correlation between age, heart disease, hypertension, bmi and glucose level. We can also see that age has a major impact on stroke rather than other factors.

Methods Implementation

Data transformation

Before Implementing statistical methods and machine learning techniques we need to transform the data. The following steps are carried out:

1. **Label encoding:** Label encoding is a process in which all categorical data are converted into numerical data. This is done to make model implementation easier.

```
{r}
data_clean <- data_clean %>%
  mutate(across(all_of(categorical_vars), ~ as.numeric(factor(.)) - 1))
```

Figure 11: Label Encoding

2. **Feature engineering:** Feature Engineering is a process in which features that are not required in model development are either modified or eliminated. In this dataset **id** is not a defining factor for stroke. Due to this reason, we can eliminate this feature from the dataset. Also, we will be removing age groups from the dataset as we had created to perform imputation in smoke data.

```
```{r}
data_clean <- data_clean %>% select(-id)

...

```{r}
data_clean <- data_clean %>% select(-age_group)
```
```

Figure 12: Feature Engineering

3. **Min-Max Scaling:** Min-Max Scaling is done to convert all numerical data in the dataset into a range of 0 to 1. This step is carried out to make sure that all features of the dataset contribute equally when developing a model.

```
```{r}
# Apply Min-Max Scaling to numeric variables
numeric_vars <- names(data_clean)[sapply(data_clean, is.numeric)]
preProcess_min_max <- preProcess(data_clean[, numeric_vars], method = "range")
data_clean_scaled <- predict(preProcess_min_max, data_clean)
```
```

Figure 13: Min\_Max Scaling

#### 4. Data Splitting

The dataset was split into 70/30. 70% of the dataset will be used to train the model while 30% of the dataset will be used to test the model. This split ensures that the model is trained on a large chunk of the data while retaining a significant amount for testing.

```
Split data (70% train, 30% test)
set.seed(42)
trainIndex <- createDataPartition(data_clean_scaled$stroke, p = 0.7, list = FALSE)
train_data <- data_clean_scaled[trainIndex,]
test_data <- data_clean_scaled[-trainIndex,]

Convert stroke variable to factor
train_data$stroke <- factor(ifelse(train_data$stroke == 1, "Yes", "No"))
test_data$stroke <- factor(ifelse(test_data$stroke == 1, "Yes", "No"), levels = levels(train_data$stroke))
```

## Model Deployment

Now we are going to deploy Logistic regression, Decision Tree and Random Forest.

### 1. Logistic Regression

```
Logistic Regression Model
log_model <- glm(stroke ~ ., data = train_data, family = binomial)

Predictions
log_predictions_prob <- predict(log_model, test_data, type = "response")
log_predictions <- ifelse(log_predictions_prob > 0.5, "Yes", "No")
log_predictions <- factor(log_predictions, levels = levels(test_data$stroke))

Confusion Matrix
log_conf_matrix <- confusionMatrix(log_predictions, test_data$stroke)
print(log_conf_matrix)
```

Figure 14: Logistic Regression Implementation

### 2. Decision Tree

```
Decision Tree Model
dt_model <- rpart(stroke ~ ., data = train_data, method = "class")

Predictions
dt_predictions <- predict(dt_model, test_data, type = "class")
dt_predictions <- factor(dt_predictions, levels = levels(test_data$stroke))

Confusion Matrix
dt_conf_matrix <- confusionMatrix(dt_predictions, test_data$stroke)
print(dt_conf_matrix)
```

Figure 15: Decision Tree Implementation

### 3. Random Forest

```
##{r}
Random Forest Model
rf_model <- randomForest(stroke ~ ., data = train_data, ntree = 100, mtry = 3, importance = TRUE)

Predictions
rf_predictions <- predict(rf_model, test_data)
rf_predictions <- factor(rf_predictions, levels = levels(test_data$stroke))

Confusion Matrix
rf_conf_matrix <- confusionMatrix(rf_predictions, test_data$stroke)
print(rf_conf_matrix)
```

Figure 16: Random Forest Implementation

This is how we have implemented statistical and machine learning methods into our project.

## Model Validation and Comparison

Model Validation is a process in which we evaluate a model's performance to ensure that the models deployed are generating the results required for a certain task. We will be using Accuracy, F1 score, Precision, Recall, and Confusion Matrix to evaluate our models.

### Confusion Matrix

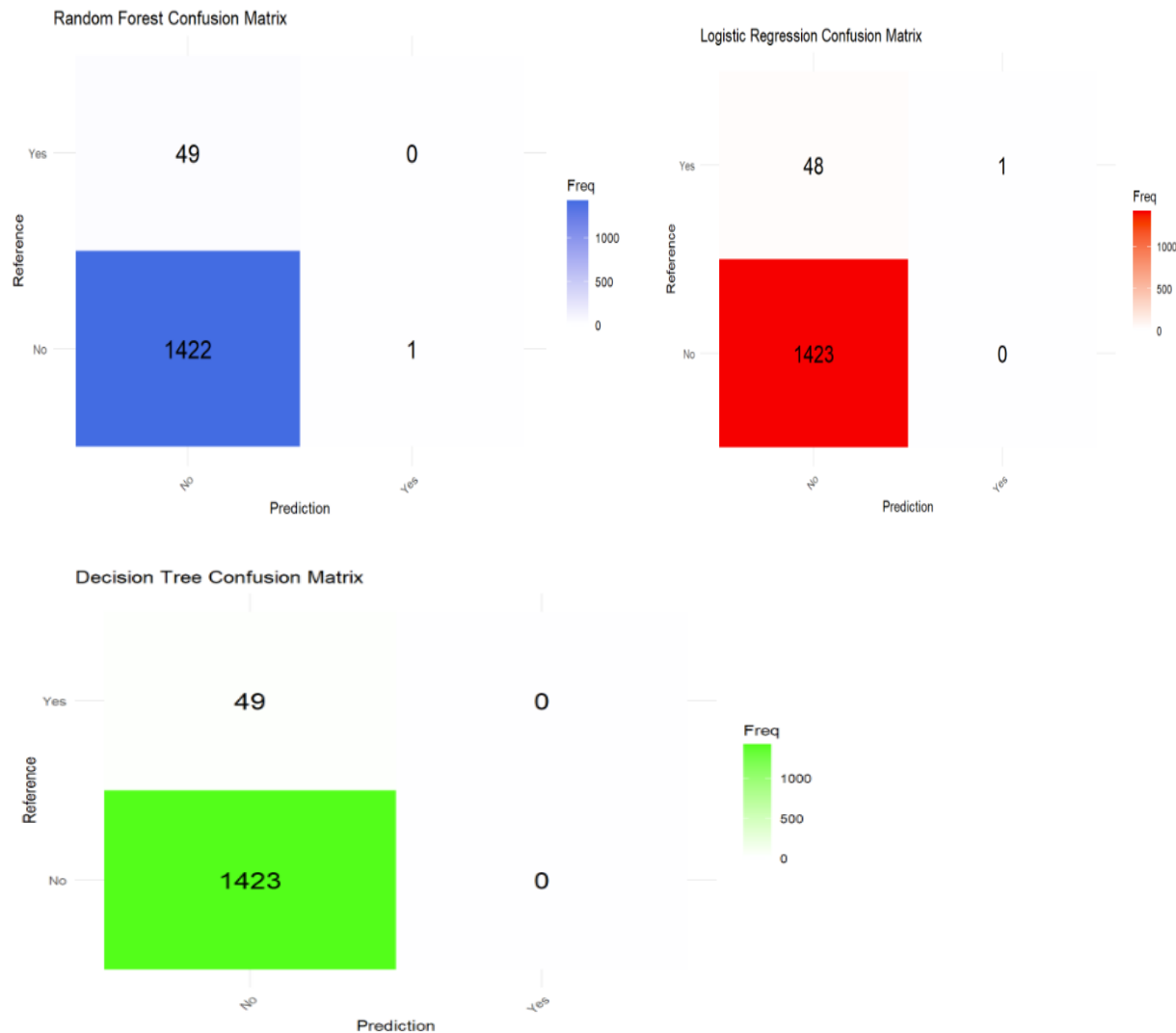


Figure 17: Confusion Matrix

Confusion matrix a detailed breakdown of the predictions of models. From these confusion matrixes, we can analyze that the models and methods implemented by us are able to accurately predict whether a person will have a stroke or not.



## Model Comparison using Accuracy, F1 score, Precision and Recall

Model Performance Comparison Table

| Model               | Accuracy | Precision | Recall | F1_Score |
|---------------------|----------|-----------|--------|----------|
| Random Forest       | 0.9660   | 0.9667    | 0.9993 | 0.9827   |
| Logistic Regression | 0.9674   | 0.9674    | 1.0000 | 0.9834   |
| Decision Tree       | 0.9667   | 0.9667    | 1.0000 | 0.9831   |

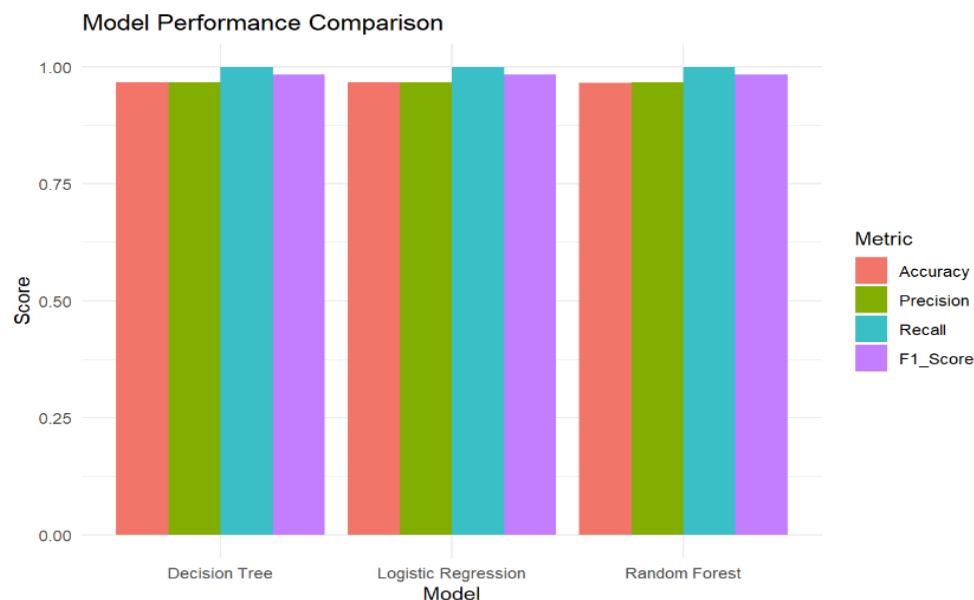


Figure 18: Model performance Comparison

From the Model performance comparison table and model performance bar graph it can be observed that all three models implemented by us i.e., Decision Tree, Logistic Regression, and Random Forest perform exceptionally well when applied to our dataset. All of them yield high accuracy, precision, recall, and F1 score. All of these models work well while predicting stroke according to our dataset.

## Critical analysis of the results

By analyzing various methods applied in this project the critical insights we discovered that hypertension and heart disease are the most influential factors in predicting stroke. Glucose levels also play a major role. The dataset contained fewer instances of stroke occurrence compared to

non-stroke cases, also there are heavy class imbalances in other features. This class imbalance may have affected model training making these models less useful for practical use.

## Addressing the Research Question

From the implementation of these methods, we discovered the correlation between **age, hypertension, heart disease, glucose levels**, and the probability of getting a stroke. This insight could be valuable while implementing preventive measures. As our models performed remarkably, more personalized datasets can be used to determine the best stroke prediction approach. If more personalized datasets are used and various statistical and machine learning methods are applied, a highly accurate stroke prediction system that can be practically applied might be possible.

## Limitations and Future Study

Despite our study producing promising results, there are several limitations that must be addressed. Limitations of our study are:

1. **Class Imbalance in the dataset:** Despite containing various important health related factors in the dataset, there was a major class imbalance as the instances occurrence of stroke was way more than that of not getting a stroke. There were imbalances in other features too. This issue makes the model less reliable for real life implementation despite its high accuracy.
2. **Lack of personalized data:** his dataset primarily relies on very generic data. Only some of the health and lifestyle related data is available in this dataset. Factors like genetic makeup, medical history and other lifestyle habits were not included in this dataset.
3. **Data Source and quality:** Our dataset was sourced from Kaggle. This dataset might not fully represent individuals from diverse backgrounds. Data quality, integrity and missing values might result in poor practical implementation of models.
4. **Limited Features:** While our data set included important features, the inclusion of other potential factors like stress level, physical activity etc. could have resulted in a more accurate and practically applicable model.

To address these limitations, future studies should try to integrate the following techniques and ideas:

- Future studies should try to gather and include more personalized data such as genetic makeup, family histories and other lifestyle factors. This will improve the accuracy and practicality of stroke prediction models in real life.
- Techniques like Synthetic Minority Over-sampling Technique (SMOTE) can be used in future work to balance the dataset and future improve the model's performance.
- Deep learning techniques such as neural networks and recurrent neural networks can be explored to improve the accuracy of stroke prediction even more.
- Implementation of models such as SHAP (SHapley Additive exPlanations) which can be easily visualized, can help healthcare professionals better understand, and trust recommendations produced by these models.
- Future work should be validated by implementing them in a real-world environment or a real-world healthcare system. Clinical trials and other examinations can be used to ensure that these models can be practically applied.

By addressing these limitations, researchers can develop more reliable and practically implementable models that can be integrated into real-world healthcare systems. This will ultimately help in early detection and development of stroke-related prevention strategies.

## Conclusion

Stroke is a critical medical condition that can lead to brain damage, physical disabilities, and even death. This project focuses on analyzing stroke-related data using various exploratory data analysis (EDA), machine learning, and statistical modeling techniques. Strokes have a severe impact on individuals, families, and the overall healthcare system. This project is aimed at discovering patterns and correlations that lead to a higher probability of getting a stroke. This project also aims to develop predictive models that can analyze various health and lifestyle-related indicators to assess stroke-related risks.

Various EDA, statistical analysis, statistical models, and machine learning techniques were implemented in this study to analyze stroke-related factors. Logistic regression, decision tree, and random forest models were implemented on the stroke prediction dataset. These models were compared and analyzed with all of them producing similar accuracy, precision, recall, and F1 scores. We also used a confusion matrix to visualize the accuracy of these models.

The findings of this project demonstrate the potential of data-driven approaches in impacting healthcare decision-making. Healthcare professionals are able to detect risk factors for stroke and initiate preventive measures for at-risk patients with the application of machine learning. This study also highlights the importance of the integration of predictive models in real-life healthcare systems. This study also highlighted the various limitations. One of the major issues faced in this project was class imbalance, as this imbalance can result in these models not being practical during real-life implementation. While this study was able to recognize key health and lifestyle related factors, more personalized information such as genetic makeup and other lifestyle factors can further increase model validity and real-life practicality.

In conclusion, this project uses data-driven approaches to identify patterns and produce predictive models to predict and reduce the risks of getting a stroke. Various EDA, statistical modeling, and machine learning techniques were implemented in this project but further advancements and improvements in machine learning and data collection techniques can produce an improved model which can be practically implemented.

## Appendixes

### References

- Amann, J. (2021). Machine Learning in Stroke Medicine: Opportunities and Challenges for Risk Prediction and Prevention. *Artificial Intelligence in Brain and Mental Health: Philosophical, Ethical & Policy Issues*.
- Ciarambino, T., Crispino, P., Mastrolorenzo, E., Viceconti, A., & Giordano, M. (2022). Stroke and Etiopathogenesis: What Is Known? *Genes*.
- Dev, S., Wang, H., Nwosu, C. S., Jain, N., Veeravalli, B., & John, D. (2022). A predictive analytics approach for stroke prediction using machine learning and neural networks. *Healthcare Analytic*.
- Emon, M. U., Keya, M. S., Meghla, T. I., Rahman, M. M., Mamun, M. S., & Kaiser, M. S. (2020). Performance Analysis of Machine Learning Approaches in Stroke Prediction. *4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1464-1469.
- Khosla, A., Cao, Y., Lin, C. C.-Y., Chiu, H.-K., Hu, J., & Lee, H. (2010). An integrated machine learning approach to stroke prediction. *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 183-192.
- Murphy, S. J., & Werring, D. J. (2020). Stroke: causes and clinical features. *Medicine*, 561-566.
- What is a Stroke? (2003).