

A Comprehensive Analysis Of Risk Factors And Predictive Modelling For Stroke Incidence

Project Report submitted to the
SDM COLLEGE (Autonomous)



in partial fulfilment of the degree of

**MASTER OF SCIENCE
IN
STATISTICS**

by

Dhanya J D

Under the supervision of

Asst. Prof. Ms. Supriya Shivadasan Padmavati

Department of Post Graduate Studies in Statistics

SRI DHARMASTHALA MANJUNATHESHWARA

COLLEGE (Autonomous)

UJIRE - 574240

Karnataka, INDIA

January 2024

SRI DHARMASTHALA MANJUNATHESHWARA COLLEGE
(AUTONOMOUS)
UJIRE - 574240



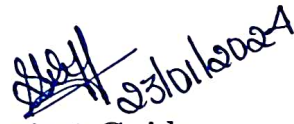
DEPARTMENT OF STATISTICS

CERTIFICATE

Certified that this is the bonafide record of project work done by Ms. Dhanya J D during the year 2024 as a part of her M.Sc (Statistics) third semester course work.

Reg. No.

2	2	4	0	0	7
---	---	---	---	---	---


Project Guide


Examiner

- 1.
- 2.



Place: Ujire

Date:


Head of the Department
Head of the Department
P.G. Studies in Statistics
SDM College (Autonomous)
Ujire - 574240, D.K. Karnataka

DECLARATION

I Dhanya J D , hearby declare that the matter embodied in this report entitled 'A Comprehensive Analysis Of Risk Factors And Predictive Modelling For Stroke Incidence' is a bonafide record of project work carried out by me under the guidance and supervision of Asst. Prof. Ms. SUPRIYA SHIVADASAN PADMA-VATHI, Department of Statistics, SDM College Ujire -574240, Karnataka, India. I further declare that no part of the work contained in the report has previously been formed the basis for the award of any Degree, Diploma, Associateship, Fellowship or any other similar title or recognition of any other university.

Place: Ujire

Dhanya J. D.
Signature

Date:

ACKNOWLEDGEMENTS

Firstly, I would like to thank our Principal **Dr. Kumara Hegde . B. A.** for providing the necessary facilities for the completion of this project work in our college.

*I would also thank our Dean **Dr. Vishwanath P** for his support. It is my privilege to thank our HOD **Dr. Savitha Kumari** for her suggestions and support.*

I am very grateful to my Research Supervisor, Asst. Prof. Ms. Supriya Shivadasan Padmavati, Department of Statistics, SDM College, Ujire, for her kind help and encouragement throughout my project work.

I gratefully acknowledge my teachers at the Department of Statistics, SDM College, Ujire, Asst. Prof. Ms. Shwetha Kumari and Asst. Prof. Mr. Pradeep K for their support during my project work.

I am also thankful to all my family members and friends for their constant encouragement and help in each step.

My sincere thanks also goes to the students of SDM College, Ujire, who have helped me directly or indirectly during my project work. Finally to all who helped me in many ways, I say, 'Thank You!'.

Contents

1 Chapter 1	7
Introduction	7
1.1 Motivation	9
1.2 Literature Review	9
1.3 Objectives	11
1.4 Scope of the study	11
2 Chapter 2	12
Methodology	12
2.1 Introduction	12
2.2 Sampling Strategy	12
2.2.1 Collection of data	12
2.3 Statistical Techniques Used	13
2.3.1 Bar Graph	13
2.3.2 Box Plot	13
2.3.3 Violin Plot	13
2.3.4 Column chart	13
2.3.5 Chi-square test for Independence of Attributes	14
2.3.6 t-test	14
2.3.7 Logistic Regression Model	14
2.3.8 Random forest classifier	15
3 Chapter 3	16
3.1 Results and Discussion	16
3.1.1 To examine patterns in stroke occurrence between genders.	16
3.1.2 To determine the most significant predictors of stroke.	17
3.1.3 To examine how different work types influence stroke incidence.	18
3.1.4 To investigate the connection between average glucose levels,BMI and stroke occurrence.	19
3.1.5 To determine if any residence type(rural or urban) contribute to stroke incidence.	21
3.1.6 To develop and train machine learning model capable of predicting the probability of stroke.	23
4 Chapter 4	24
4.1 Conclusion	24
4.2 Overall Conclusion	25

5	Chapter 5	26
5.1	Summary	26
6	Chapter 6	27
6.1	Bibliography	27
7	Chapter 7	28
7.1	Appendix	28

List of Tables

1	Frequency of Stroke Occurrence by Gender	16
2	Contingency table of stroke occurrence by gender	17
3	Table showing the coefficients	17
4	Table showing the frequency of stroke incidence for each work type	18
5	Frequency table of stroke occurrence by residence type	21
6	Contingency table of stroke occurrence by gender	22
7	Table of classification	23

List of Figures

1	Frequency of Stroke Occurrence by Gender	16
2	Frequency of stroke occurrence by each work type.	18
3	Distribution of average glucose level by stroke	19
4	Distribution of BMI by stroke	19
5	Violin plot of stroke incidence by residence type	21
6	Receiver operating characteristic curve	23

1 Chapter 1

Introduction

Strokes, often called "brain attacks" are serious health challenges that affect people and communities worldwide. They happen when something goes wrong with the blood flow to the brain, causing issues with movement, thinking, and overall life quality. Figuring out how often these strokes happen is what we call Stroke Incidence – it's like counting how many times these brain issues occur in a group of people over a certain time.

Why does Stroke Incidence matter? Well, it's like a helpful tool for doctors and researchers. It helps them understand how often these brain issues occur in a group of people over a specific time. But why does this matter? Well, it's not just about counting strokes; it's about finding patterns and trends. It's about uncovering the story behind the strokes – what makes them happen and how we can prevent them. This study into Stroke Incidence looks at different things that might affect strokes, like age, gender, health, lifestyle, where people live, and the kind of work they do.

Our daily choices also play a role. Imagine lifestyle choices as the decisions we make every day – like whether or not to smoke. Studies show that smoking can increase the likelihood of having a stroke. So, understanding these lifestyle factors is like recognizing habits that might be harmful to our health and finding ways to change them for the better. The place you call home can also impact Stroke Incidence. It's not the same everywhere; strokes might happen more often in cities or rural areas. This can be due to differences in healthcare access, the environment, and the way people live. Exploring these regional distinctions is like understanding the unique challenges different communities face when it comes to strokes. Some jobs may expose individuals to higher stress levels or require a lot of sitting. This can contribute to an increased risk of strokes. Studying different occupations is like recognizing patterns that help us figure out how work life might be linked to strokes. It's about making sure people stay healthy, even in their workplaces.

As we embark on this exploration of Stroke Incidence, our main goal is to answer the big questions – why and how do strokes occur? By studying the information, we aim to find out what things make strokes more likely and how we can stop them.

Looking ahead, we want to do more than just understand what happened in the past. We want to predict the future. This involves creating models that can forecast the likelihood of someone having a stroke based on various factors.

These models help doctors and researchers plan ahead, making sure they're ready to prevent strokes before they happen.

In the end, Stroke Incidence isn't just about counting strokes; it's a complex study that goes beyond numbers. It's about understanding the multitude of factors that contribute to strokes – age, health, lifestyle, location, and occupation. This exploration connects dots across demographics, health metrics, geography, and work life, weaving a narrative that goes beyond statistics. The ultimate aim is not just to quantify strokes but to illuminate the path toward a future where strokes are not only counted but significantly reduced through targeted and informed interventions, creating healthier communities for everyone.

1.1 Motivation

Starting This project, "Understanding and Preventing Strokes" is like beginning a mission to keep everyone in community safe and healthy. Strokes can be tough, affecting how people move and think. This project on a mission to find out why they happen and, most importantly, how to stop them.

Think of this project as a search for important information to create a guide that helps everyone live a healthier life. Main goal is to prevent strokes and help people make choices that keep them safe and happy. By learning more, I want to help both doctors and people like you make smart choices that lower the chances of having a stroke. It's like having a helpful tool that shows you how to live a healthier life.

This project isn't just about numbers and data; it's about making a real difference in people's lives. If this can prevent even one person from facing the challenges of a stroke, it's a big step toward creating a community where everyone can live better and healthier. That's what keeps me excited and motivated everyday – the idea that this work might really help make the world a safer and healthier place for everyone.

1.2 Literature Review

In 2021, Chun M, Clarke R, Cairns B J, Clifton D, Bennett D, Chen Y conducted a study on "Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults". The aims of this study were to compare Cox and ML models for prediction of risk of stroke in China at varying intervals of follow-up (ie, stroke within 9 years, 0–3 years, 3–6 years, 6–9 years) and to identify individuals for whom ML models might be superior to conventional Cox-based approaches for stroke risk prediction and develop and evaluate an ensemble model combining both approaches to identify individuals at high risk of stroke. The results highlight the potential value of expanding the use of ML in clinical practice.

In 2021, Qi Wang, Lulu Zhang, Yidan Li, Xiang Tang conducted a study on "Development of stroke predictive model in community-dwelling population: A longitudinal cohort study in Southeast China". This study aims to develop a highly accurate prediction model of stroke with a list of lifestyle behaviors and clinical characteristics to distinguish high-risk groups in the community-dwelling population. The predictive models could predict 2-year stroke with high accuracy. The models provided an effective tool for identifying high-risk groups and supplied guidance for improving prevention and treatment strategies in community-dwelling population.

In 2021, Eman M Alanazi, Aalaa Abdou and Jake Luo conducted a study on "Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms". The aim of this study was to apply computational methods using machine learning techniques to predict stroke from lab test data. They found that accurate and sensitive machine learning models can be created to predict stroke from lab test data. The predictive model, built using data from lab tests, was easy to use and had high accuracy.

In 2022, Qiu Y, Cheng S, Wu Y conducted a study on "Development of rapid and effective risk prediction models for stroke in the Chinese population". The purpose of this study was to use easily obtained and directly observable clinical features to establish predictive models to identify patients at increased risk of stroke. This work provides a rapid and accurate tool for stroke risk assessment, which can help to improve the efficiency of stroke screening medical services and the management of high-risk groups.

In 2022, Dritsas, Elias, and Maria Trigka conducted a study on "Stroke Risk Prediction with Machine Learning Techniques". In this research work, with the aid of machine learning, several models are developed and evaluated to design a robust framework for the long-term risk prediction of stroke occurrence. The main contribution of this study is a stacking method that achieves a high performance that is validated by various metrics, such as AUC, precision, recall, F-measure and accuracy. The experiment results showed that the stacking classification outperforms the other methods, with an AUC of 98.9%, F-measure, precision and recall of 97.4% and an accuracy of 98%.

In 2018, Seung Nam Min, Se Jin Park, Dong Joon Kim, Murali Subramaniyam, Kyung-Sun Lee Eur Neurol conducted a study on "Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea". They aimed to derive a model equation for developing a stroke pre-diagnosis algorithm with the potentially modifiable risk factors. They developed a logistic regression model based on information regarding several well-known modifiable risk factors. The developed model could correctly discriminate between normal subjects and stroke patients in 65% of cases.

1.3 Objectives

- To examine patterns in stroke occurrence between genders.
- To determine the most significant predictors of stroke.
- To examine how different work types influence stroke incidence.
- To investigate the connection between average glucose levels, BMI and stroke occurrence.
- To determine if any residence type(rural or urban) contribute to stroke incidence.
- To develop and train machine learning model capable of predicting the probability of stroke.

1.4 Scope of the study

This project aims to create a guide that assists people in making healthier choices, ultimately making our community a safer and healthier place. The project's scope is to provide valuable information that helps both individuals and doctors in lowering the chances of strokes, contributing to an overall healthier lifestyle. By focusing on prevention and creating awareness, the project aims to have a positive impact on people's lives, reducing the challenges associated with strokes and fostering a safer environment for everyone.

2 Chapter 2

Methodology

2.1 Introduction

In this chapter, the details of the collection of data on a stroke incidence is included. This chapter also includes the details of the statistical tools and methods used for the analysis of the data.

Section 2.2 gives the details about the strategy used in data collection and section 2.3 gives the details about the various statistical techniques used for the data analysis.

2.2 Sampling Strategy

2.2.1 Collection of data

This research is based on a secondary dataset from Kaggle. The number of participants was 4982, and all of the attributes are described as follows:

- **Age(years):** This feature refers to the age of the participants.
- **Gender:** This feature refers to the participants gender.
- **Hypertension:** This feature refers to whether this participant is hypertensive or not.
- **Heart disease:** This feature refers to whether this participant suffers from heart disease or not.
- **Ever married:** This feature represents the marital status of the participants.
- **Work type:** This feature represents the participant's work status.
- **Residence type:** This feature represents the participant's living status and has 2 categories(urban,rural).
- **Avg glucose level:** This feature captures the participants average glucose level.
- **BMI(Kg/m^2):** This feature captures the body mass of the participants.
- **Smoking status:** This feature captures the participant's smoking status.
- **Stroke:** This feature represents if the participant previously had a stroke or not.

2.3 Statistical Techniques Used

The statistical software like 'Python' is used for the analysis and interpretation of the data and excel is also used to analyze the data. The statistical techniques used to carry out the analysis are given as follows :

2.3.1 Bar Graph

A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a column chart.

It allows you to compare different sets of data among different groups easily. It instantly demonstrates this relationship using two axes, where the categories are on one axis and the various values are on the other. A bar graph can also illustrate important changes in data throughout a period of time.

2.3.2 Box Plot

A boxplot, also known as a box-and-whisker plot, is a graphical representation of the distribution of a dataset. It provides a visual summary of key statistical measures, including the median, quartiles, and potential outliers. The box encompasses the interquartile range (IQR), representing the middle 50% of the data, while the line inside the box indicates the median. Whiskers extend from the box to depict the data's range, and individual data points beyond the whiskers may be considered outliers. Boxplots are valuable for comparing datasets, identifying central tendencies, and detecting outliers, making them a widely used tool in exploratory data analysis.

2.3.3 Violin Plot

A violin plot is a sophisticated data visualization that combines features of a boxplot and a kernel density plot. Resembling the shape of a violin, the plot provides insights into the distribution and probability density of the data. The width of the violin at any point indicates the estimated probability density, offering a more nuanced view of the data's distribution. Often incorporating a box-and-whisker plot within, the violin plot visually captures the spread, median, and potential outliers. This type of plot is particularly useful for displaying the density of data across different regions of the distribution and is commonly employed in scenarios where traditional boxplots might oversimplify the data's complexity.

2.3.4 Column chart

A column chart, often referred to as a bar chart, is a visual representation of data using vertical or horizontal bars to showcase values within distinct cate-

gories. Each column's height or length corresponds to a specific value, making it easy to compare magnitudes across different groups. Typically employed to illustrate changes over time or compare quantities between various categories, column charts have two axes vertical and horizontal representing values and categories, respectively. The color-coded bars enhance differentiation between datasets, and the inclusion of titles and labels ensures clarity. This graphical tool is widely utilized in fields like business, statistics, and data analysis for its effectiveness in conveying information and facilitating a quick understanding of trends and comparisons.

2.3.5 Chi-square test for Independence of Attributes

The chi-squared (χ^2) test is a statistical technique used to determine if there is a significant association between two categorical variables. It is employed to assess whether observed data in a contingency table differs significantly from what would be expected if there were no association. The test calculates a chi-squared statistic, which measures the disparity between observed and expected frequencies. By comparing this statistic to a chi-squared distribution, it yields a p-value that indicates the likelihood of observing such an association by chance. A small p-value (typically less than 0.05) suggests a statistically significant relationship between the variables, while a larger p-value indicates a lack of association. Chi-squared tests are valuable tools for assessing relationships and conducting goodness-of-fit tests in statistical analysis. Python libraries like `scipy.stats` and `pandas` enable the execution of chi-squared tests.

2.3.6 t-test

The t-test is a statistical method used to evaluate whether there is a significant difference between the means of two groups. It calculates a t-statistic based on sample data, taking into account both the average difference and the variability within the groups. The t-test compares this statistic to a critical value from the t-distribution, helping researchers assess if the observed differences are likely due to chance or if they reflect a genuine distinction between the groups. A lower p-value resulting from the test indicates stronger evidence against the null hypothesis, suggesting a meaningful difference between the group means. T-tests are widely employed in scientific research, allowing investigators to make informed conclusions about the significance of observed differences in various scenarios, such as comparing the means of two treatment groups in an experiment.

2.3.7 Logistic Regression Model

Logistic regression is process of modelling the probability of a discrete outcome given an input variable. Logistic regression is one of the commonly used algorithms in machine learning for binary classification problems, which are problems

with two class values. Logistic regression can also estimate the probabilities of outcomes. Hence it can be used for classification by creating a model.

The assumption of logistic regression are :

- The dependent variable should be binary
- The observations should be independent of each other
- There should be little or no multicollinearity among the independent variables
- There must be a linear relationship between the independent variables and log odds

The logistic regression model is given by, $y = x'\beta + \epsilon$, the link function used here is the logit link, given by,

$$\eta = \ln\left(\frac{\pi}{1 - \pi}\right)$$

The odds are defined as the probability that a particular outcome is a case divided by the probability that it is a noncase. The logistic function is given by,

$$F(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

where $F(x)$ is the probability that the dependent variable equals a case, give some linear combination x of the predictors, β_0 is the intercept from the linear regression equation, $\beta_1 x$ is the regression coefficient multiplied by some value of the predictor, base e denotes the exponential function.

2.3.8 Random forest classifier

The Random Forest Classifier is an ensemble learning algorithm that builds a multitude of decision trees during training and merges their predictions to improve overall accuracy and robustness. Each tree in the forest is constructed independently, using a random subset of the training data and a random subset of features for each split. This randomness helps prevent overfitting and promotes diversity among the individual trees. The final prediction is made by averaging or taking a vote across all the trees, resulting in a robust and flexible model. Random Forests are widely used for both classification and regression tasks due to their ability to handle complex relationships in data, resist overfitting, and provide insights into feature importance.

3 Chapter 3

3.1 Results and Discussion

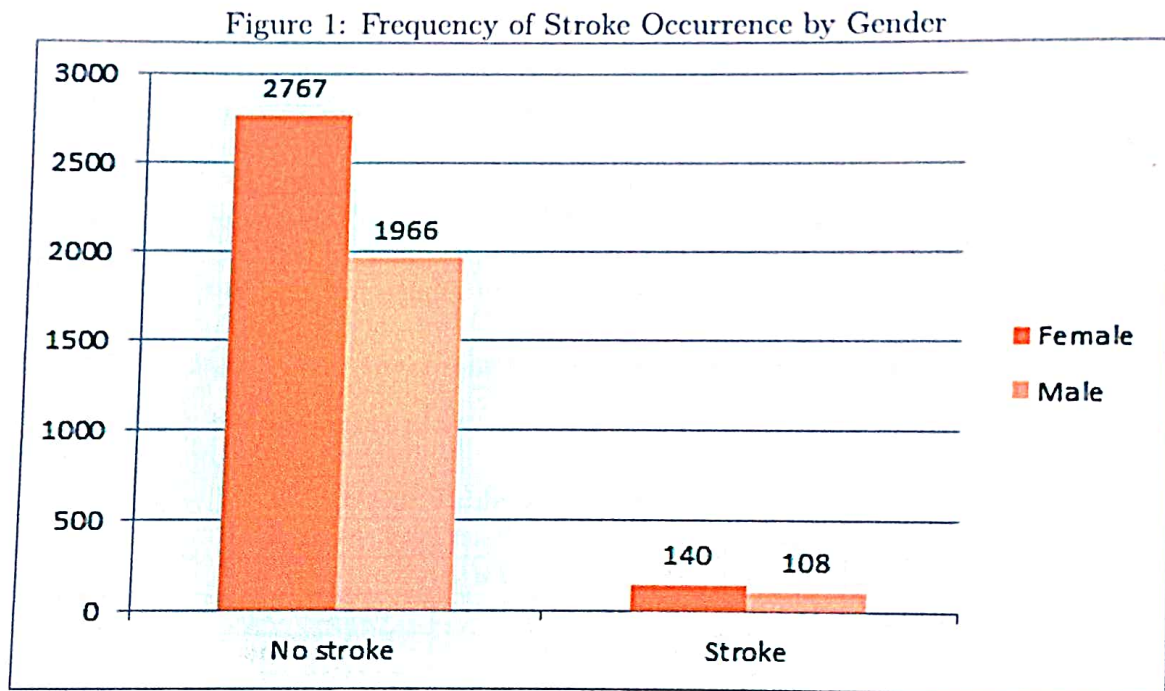
3.1.1 To examine patterns in stroke occurrence between genders.

The following table provides a detailed breakdown of the frequency of stroke occurrence by gender.

Table 1: Frequency of Stroke Occurrence by Gender

Gender	No stroke	Stroke
Female	2767	140
Male	1966	108

The following column chart shows the frequency of stroke occurrence by gender.



The column chart visually compares the frequency of stroke occurrence between males and females, with numerical values displayed on top of each bar. For females, 2767 individuals did not experience a stroke, and 140 individuals experienced a stroke. For males, 1966 individuals did not experience a stroke, and 108 individuals experienced a stroke. Both males and females show some instances of stroke occurrence, but further statistical analysis is needed to determine if there are significant differences.

Chi-Square test :

The hypothesis are as follows :

H_0 : There is no significant association between gender and stroke occurrence.

H_1 : There is a significant association between gender and stroke occurrence.

The following contingency table shows the distribution of stroke occurrences between genders.

Table 2: Contingency table of stroke occurrence by gender

Gender	No stroke	Stroke
Female	2767	140
Male	1966	108

- Chi-square value: 0.3135
- P-value: 0.5755
- Degrees of freedom: 1

The contingency table shows the distribution of stroke occurrences between genders. The chi-square test results in a chi-square value of 0.3135 and a p-value of 0.5755. With a p-value greater than 0.05, we do not reject the null hypothesis, so there is no significant association between gender and stroke occurrence. Therefore, based on the analysis, I conclude that there is no statistically significant association between gender and the occurrence of strokes.

3.1.2 To determine the most significant predictors of stroke.

Determining the predictors of stroke using logistic regression.

Table 3: Table showing the coefficients

Variables	coefficient	std err	z	p > z
Age	0.0690	0.005	13.149	0.000
Hypertension	0.3944	0.163	2.414	0.016
Heart disease	0.3226	0.188	1.717	0.086
Avg glucose level	0.0039	0.001	3.288	0.001
Bmi	0.0088	0.012	0.708	0.479

- Accuracy = 0.9402

The model indicates that age, hypertension, and average glucose level are statistically significant predictors of stroke. Individuals with hypertension tend to have a higher likelihood of stroke. The model may not find a statistically significant association between heart disease, BMI, and stroke.

3.1.3 To examine how different work types influence stroke incidence.

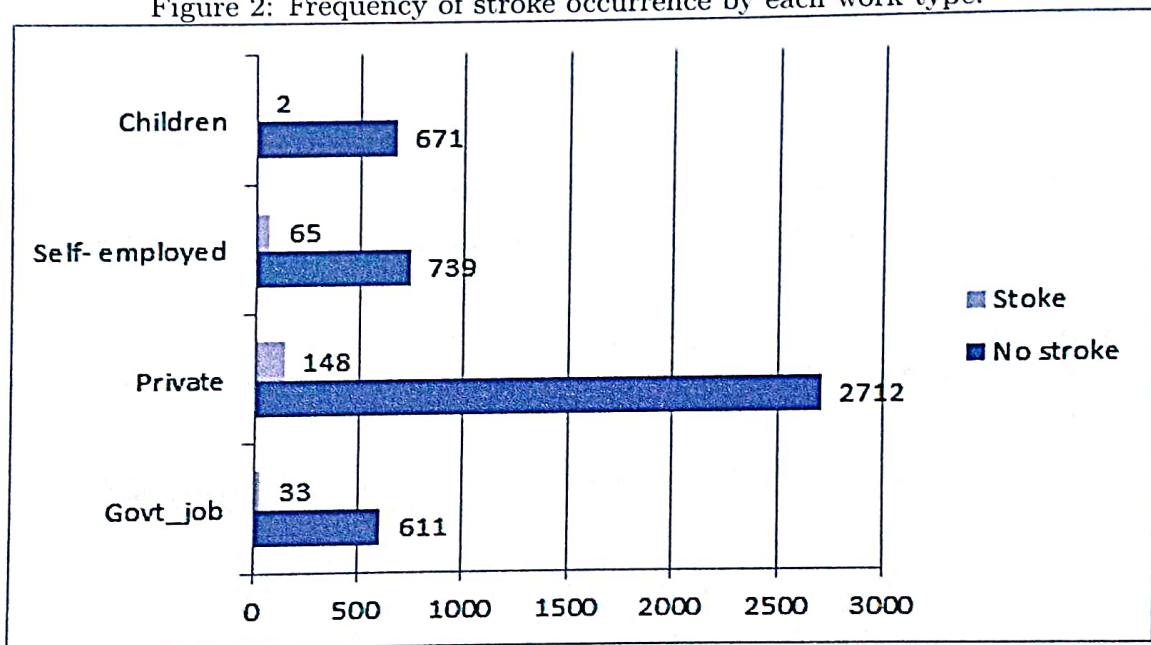
Analyzing how different work types influence stroke incidence using bar chart. The following table gives the frequency of stroke incidence for each work type.

Table 4: Table showing the frequency of stroke incidence for each work type

Work type	No stroke	stroke
Govt job	611	33
Private	2712	148
Self-employed	739	65
children	671	2

The following bar chart shows the frequency of stroke incidence for each work type.

Figure 2: Frequency of stroke occurrence by each work type.



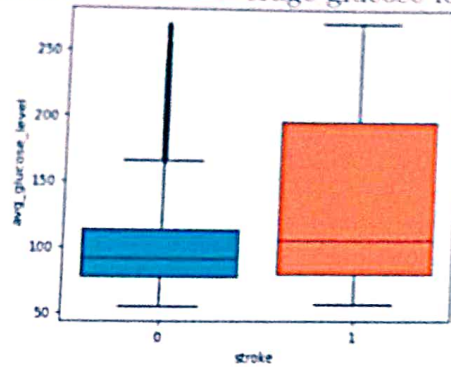
The bar chart visually represents the distribution of stroke incidents and non-stroke incidents for each work type. Among the work types, "Private" has the highest number of both stroke and non-stroke incidents, followed by "Self-employed", "Govt job" and "Children". The bars for non-stroke incidents are significantly higher than those for stroke incidents for all work types, indicating a lower incidence of strokes.

3.1.4 To investigate the connection between average glucose levels, BMI and stroke occurrence.

Analyzing the relationship between the average glucose level, BMI and stroke occurrences using boxplot.

The following boxplots shows the distribution of average glucose level by stroke.

Figure 3: Distribution of average glucose level by stroke

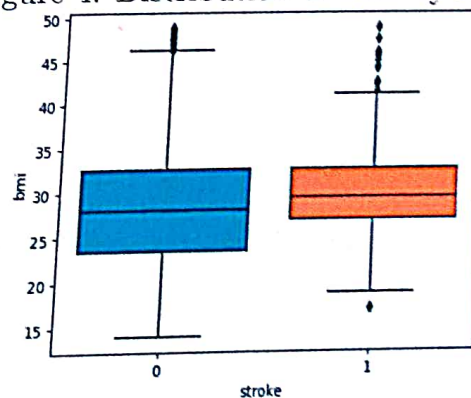


- For individuals without strokes, the median average glucose level is 91.45.
- For individuals with strokes, the median average glucose level is higher at 105.04.

The box plot for "avg glucose level" indicates that the distribution of average glucose levels tends to be higher for individuals who have had a stroke compared to those who have not.

The following boxplots shows the distribution of BMI by stroke.

Figure 4: Distribution of BMI by stroke



- For individuals without strokes, the median BMI is 28.00.
- For individuals with strokes, the median BMI is slightly higher at 29.45.

The box plot for "BMI" suggests that individuals with strokes tend to have a slightly higher median BMI compared to those without strokes.

There appears to be a relationship between average glucose levels, BMI, and the occurrence of strokes. Higher median values for average glucose levels and BMI in the stroke group suggest that these factors may be associated with an increased likelihood of experiencing a stroke. However, further statistical tests would be needed to confirm these observations.

T- test :

For average glucose level :

The hypothesis are as follows,

H_0 : There is no significant difference in average glucose levels between individuals with strokes and those without strokes.

H_1 : There is a significant difference in average glucose levels between individuals with strokes and those without strokes.

- T-Statistic: 9.49
- P-Value: 3.64e-21

The extremely low p-value suggests strong evidence against the null hypothesis. The positive t-statistic indicates that the average glucose levels for individuals with strokes are significantly higher than those without strokes. So, there is a significant difference in average glucose levels between individuals with and without strokes.

For BMI level :

The hypothesis are as follows,

H_0 : There is no significant difference in BMI between individuals with strokes and those without strokes.

H_1 : There is a significant difference in BMI between individuals with strokes and those without strokes.

- T-Statistic: 4.02
- P-Value: 5.82e-05

The very low p-value for BMI indicates strong evidence against the null hypothesis. The positive t-statistic implies that the BMI for individuals with strokes is significantly higher than for those without strokes. There is a significant difference in BMI between individuals with strokes and those without strokes.

Based on the investigation into the connection between average glucose levels, BMI, and stroke occurrence, the results of the t-tests provide strong evidence that both average glucose levels and BMI are significantly associated with stroke incidence. The higher average glucose levels and BMI observed in individuals with strokes suggest these factors may contribute to an increased risk of stroke.

3.1.5 To determine if any residence type(rural or urban) contribute to stroke incidence.

Analyzing if any residence type contribute to stroke incidence using violin plot.

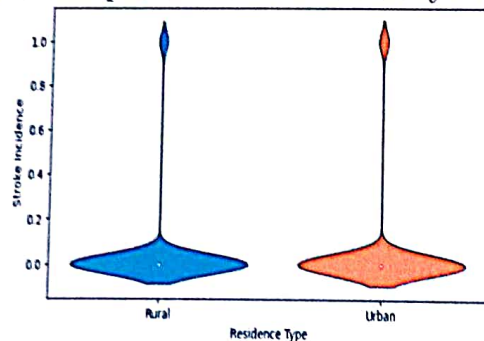
The following frequency table provide insights into the relationship between residence type (rural or urban) and stroke incidence.

Table 5: Frequency table of stroke occurrence by residence type

Residence type	No stroke	Stroke
Rural	2336	113
Urban	2397	135

The following violin plot displays the distribution of stroke incidence in both rural and urban areas.

Figure 5: Violin plot of stroke incidence by residence type



The wider sections of the violins represent areas of higher frequency in the dataset. In the plot, the distribution of stroke incidence in rural and urban areas appears similar, with a comparable spread of cases. The frequency table supports the observations from the violin plot. The total counts of stroke cases in rural and

urban areas are relatively close, with rural areas having 113 cases and urban areas having 135 cases.

Based on both the violin plot and the frequency table, it seems that there is no substantial difference in stroke incidence between rural and urban areas. Therefore, the analysis does not strongly indicate that residence type (rural or urban) contributes significantly to stroke incidence.

Analyzing if any residence type contribute to stroke incidence using chi-square test.

The hypothesis are as follows :

H_0 : There is no association between residence type (rural or urban) and stroke incidence.

H_1 : There is association between residence type (rural or urban) and stroke incidence.

The following contingency table shows the distribution of stroke incidence by residence type.

Table 6: Contingency table of stroke occurrence by gender

Residence type	No stroke	Stroke
Rural	2336	113
Urban	2397	135

- Chi-square value: 1.2078
- P-value: 0.2717
- Degrees of freedom: 1

Based on the Chi-square test results, The p-value is which is greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis. There is not enough statistical evidence to determine that residence type (rural or urban) significantly contributes to stroke incidence. The data does not support a significant association between residence type and the occurrence of strokes.

3.1.6 To develop and train machine learning model capable of predicting the probability of stroke.

- Model type : Random Forest Classifier

The following table shows the report of classification.

Table 7: Table of classification

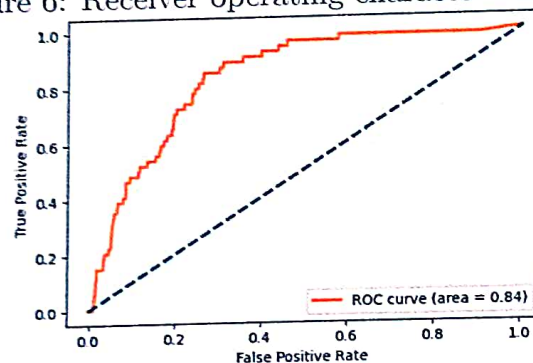
Classes	Precision	Recall	f1-score
No stroke	0.95	1.00	0.97
Stroke	0.94	1.00	0.96

Overall accuracy is the ratio of correctly predicted observations to the total observations. In this case:

Accuracy: 0.95 (95%)

The following Receiver Operating Characteristic (ROC) curve is plotted to visualize the trade-off between true positive rate and false positive rate.

Figure 6: Receiver operating characteristic curve



ROC curve measures the ability of the model to distinguish between classes. In this case:

Auccuracy : 0.835 (83.5%)

The model performs well in terms of accuracy, achieving an overall accuracy of 95%. However, when looking at class 1(stroke), the precision and recall are high (94% and 100%, respectively), suggesting good performance in identifying strokes. The AUC-ROC score of 0.835 indicates a good ability of the model to discriminate between positive and negative cases. The model shows promising results in identifying stroke cases.

4 Chapter 4

4.1 Conclusion

- The column chart visually compares the frequency of stroke occurrence between males and females. Both males and females show some instances of stroke occurrence, but further statistical analysis is needed to determine if there are significant differences.
- Based on the chi square test, concluded that there is no statistically significant association between gender and the occurrence of strokes.
- The logistic model indicates that age, hypertension, and average glucose level are statistically significant predictors of stroke. Individuals with hypertension tend to have a higher likelihood of stroke. The model may not find a statistically significant association between heart disease, BMI, and stroke.
- The bar chart visually represents the distribution of stroke incidents and non-stroke incidents for each work type. The bars for non-stroke incidents are significantly higher than those for stroke incidents for all work types, indicating a lower incidence of strokes.
- The box plot for "avg glucose level" indicates that the distribution of average glucose levels tends to be higher for individuals who have had a stroke compared to those who have not. The box plot for "BMI" suggests that individuals with strokes tend to have a slightly higher median BMI compared to those without strokes.
- Based on the investigation into the connection between average glucose levels, BMI, and stroke occurrence, the results of the t-tests provide strong evidence that both average glucose levels and BMI are significantly associated with stroke incidence.
- Based on both the violin plot and the frequency table, it seems that there is no substantial difference in stroke incidence between rural and urban areas.
- Based on the Chi-square test results, The p-value is which is greater than the typical significance level of 0.05. Therefore, we fail to reject the null hypothesis. There is not enough statistical evidence to determine that residence type (rural or urban) significantly contributes to stroke incidence.
- The AUC-ROC score of 0.835 indicates a good ability of the model to discriminate between positive and negative cases. The model shows promising results in identifying stroke cases.

4.2 Overall Conclusion

In conclusion, the project results indicate that gender and residence type do not exhibit a significant association with stroke occurrence. However, work type and its impact on strokes are noteworthy. The bar chart analysis reveals varying incident counts across different work types, with "Private" jobs showing the highest frequency of both stroke and non-stroke incidents. The Chi-square test results for residence type suggest no statistically significant contribution to stroke incidence.

Moreover, the predictive model demonstrates robust performance, particularly in identifying strokes, with age, hypertension, and average glucose level emerging as significant predictors. The importance of addressing modifiable risk factors, such as managing average glucose levels and maintaining a healthy BMI, is underscored. These findings collectively enhance our understanding of the multifaceted factors influencing stroke risk, providing valuable insights for preventive healthcare strategies.

5 Chapter 5

5.1 Summary

The project, titled "A Comprehensive Analysis Of Risk Factors And Predictive Modelling For Stroke Incidence," delved into the exploration of various factors contributing to the occurrence of strokes. By collecting data from Kaggle, the study examined gender, work type, residence type, and numerous health parameters, such as age, hypertension, and average glucose levels. Notably, the analysis uncovered that gender and residence type exhibited no significant association with stroke incidence. However, work type played a role, with "Private" jobs showing the highest frequency of incidents. The predictive model, incorporating key factors like age and average glucose levels, showcased robust performance in identifying stroke cases. These findings emphasize the importance of addressing modifiable risk factors for effective stroke prevention.

6 Chapter 6

6.1 Bibliography

- Chun M., Clarke R., Cairns B. J., Clifton D., Bennett D., Chen Y., et al.. (2021). *Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults*. J. Am. Med. Inform. Assoc. 28, 1719–1727.
- Qi Wang, Lulu Zhang, Yidan Li, Xiang Tang(2022).*Development of stroke predictive model in community-dwelling population: A longitudinal cohort study in Southeast China*.Front Aging Neurosci. 2022; 14: 1036215.
- Eman M Alanazi, Aalaa Abdou, Jake Luo (2021).*Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models*.JMIR Form Res. 2021 Dec; 5(12): e23440.
- Qiu Y, Cheng S, Wu Y, et al(2022).Development of rapid and effective risk prediction models for stroke in the Chinese population: a cross-sectional study.*BMJ Open* 2023;13:e068045. doi: 10.1136/bmjopen-2022-068045
- Dritsas, Elias, and Maria Trigka. 2022. "Stroke Risk Prediction with Machine Learning Techniques" *Sensors* 22, no. 13: 4670.
- Seung Nam Min; Se Jin Park; Dong Joon Kim; Murali Subramaniam; Kyung-Sun Lee Eur Neurol (2018).*Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study in Korea*.Volume 79, Issue(3-4): 214–220.
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8324240/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9813513/>
- <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8686476/>
- <https://bmjopen.bmj.com/content/13/3/e068045.citation-tools>
- <https://doi.org/10.3390/s22134670>
- <https://doi.org/10.1159/000488366>

7 Chapter 7

7.1 Appendix

Python codes used for data analysis.

Examining patterns in stroke occurrence by gender using column chart.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Create a table
table = pd.crosstab(data['gender'], data['stroke'],
                    margins=True, margins_name='Total')

# Create a bar chart
plt.figure(figsize=(10, 6))
ax = sns.countplot(x='gender', hue='stroke', data=data)

# Adding labels and title
plt.xlabel('Gender')
plt.ylabel('Frequency')
plt.title('Frequency of Stroke Occurrence by Gender')

# Adding legend
plt.legend(title='Stroke Occurrence', labels=['No Stroke', 'Stroke'])

# Adding numerical values on top of the bars
for p in ax.patches:
    ax.annotate(f'{p.get_height()}', (p.get_x() +
    p.get_width() / 2., p.get_height()),
    ha='center', va='center', xytext=(0, 10), textcoords='offset points')

# Show the table
print("\nTable - Frequency of Stroke Occurrence by Gender:")
print(table)

# Show the plot
plt.show()
```

```
# Conclusion
print("\nConclusion based on the Bar Chart and Table:")
print("The table provides a detailed breakdown of the
frequency of stroke occurrence by gender.")
print("The bar chart visually compares the frequency of stroke
occurrence between males and females, with
numerical values displayed on top of each bar.")
print("Both males and females show some
instances of stroke occurrence, but further
statistical analysis is needed to determine
if there are significant differences.")
```

Examining patterns in stroke occurrence by gender using chi-square test.

```
import pandas as pd
from scipy.stats import chi2_contingency
from tabulate import tabulate

# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Create a contingency table
contingency_table = pd.crosstab(data['gender'], data['stroke'])

# Print the contingency table
print("\nContingency Table:")
print(tabulate(contingency_table, headers='keys', tablefmt='pretty'))

# Perform the Chi-Square Test
chi2, p, _, _ = chi2_contingency(contingency_table)

# Display the results of the Chi-Square Test
print("\nChi-Square Test Results:")
print(f"Chi2 value: {chi2:.4f}")
print(f"P-value: {p:.4f}")

# Interpretation of results
if p < 0.05:
    print("\nThere is a significant association between
gender and stroke occurrence.")
else:
    print("\nThere is no significant association between
gender and stroke occurrence.")
```

Determining the most significant predictors of stroke using logistic regression.

```
import pandas as pd
import numpy as np
import statsmodels.api as sm

# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Convert 'stroke' to numeric (assuming '1' indicates stroke)
data['stroke'] = pd.to_numeric(data['stroke'], errors='coerce')

# Encode categorical variables
data_encoded = pd.get_dummies(data, columns=['gender', 'ever_married',
'work_type', 'Residence_type', 'smoking_status'])

# Fill missing values with mean
data_encoded.fillna(data_encoded.mean(), inplace=True)

# Add a constant term for the intercept
data_encoded['const'] = 1

# Define predictors
predictors = ['age', 'hypertension', 'heart_disease',
'avg_glucose_level', 'bmi', 'const']

# Create the logistic regression model
logit_model = sm.Logit(data_encoded['stroke'], data_encoded[predictors])

# Fit the model
result = logit_model.fit()

# Display the logistic regression results
print(result.summary())
```

Examining how different work types influence stroke incidence using bar chart.

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.stats import chi2_contingency
from tabulate import tabulate
```



```

# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Create a bar chart using seaborn
plt.figure(figsize=(10, 6))
sns.countplot(x='work_type', hue='stroke', data=data)
plt.title('Frequency of Stroke Incidence by Work Type')
plt.xlabel('Work Type')
plt.ylabel('Frequency')
plt.show()

# Interpretation
print("\nInterpretation:")
print("The bar chart visually represents the frequency of stroke incidence for each work type.")
print("Each bar is divided into two segments, corresponding to 'No Stroke' and 'Stroke' categories.")
print("The contingency table and Chi-Square Test results provide statistical evidence of the association between work type and stroke incidence.")
print("Further analysis of the chart shows the distribution of stroke cases across different work types.")

```

Investigate the connection between average glucose levels, BMI and stroke occurrence.

Using boxplot.

```

import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Create Box Plots
plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
sns.boxplot(x='stroke', y='avg_glucose_level', data=data)
plt.title('Distribution of avg_glucose_level by Stroke Occurrence')
plt.subplot(1, 2, 2)
sns.boxplot(x='stroke', y='bmi', data=data)
plt.title('Distribution of BMI by Stroke Occurrence')
plt.show()

```

Using t-test

```
import pandas as pd
from scipy.stats import ttest_ind

# Load the dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# T-Test
t_stat_glucose, p_value_glucose = ttest_ind(data[data['stroke'] == 1]
['avg_glucose_level'], data[data['stroke'] == 0]
['avg_glucose_level'])
t_stat_bmi, p_value_bmi = ttest_ind(data[data['stroke'] == 1]
['bmi'], data[data['stroke'] == 0]
['bmi'])

# T-Test Results
print("T-Test Results:")
print(f"T-Test for avg_glucose_level: T-statistic = {t_stat_glucose},
P-value = {p_value_glucose}")
print(f"T-Test for BMI: T-statistic = {t_stat_bmi}, P-value = {p_value_bmi}")

# Conclusion
print("\nConclusion:")
print("There is a significant difference in average glucose
levels between individuals with and without strokes.")
print("There is a significant difference in BMI between
individuals with and without strokes.")
```

Determining if any residence type(rural or urban) contribute to stroke incidence.

Using violin plot

```
import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Load your dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Filter data for rural and urban areas
rural_data = data[data['Residence_type'] == 'Rural']['stroke']
urban_data = data[data['Residence_type'] == 'Urban']['stroke']

# Create a DataFrame for plotting
```

```

plot_data = pd.DataFrame({'Rural': rural_data, 'Urban': urban_data})

# Create a violin plot
plt.figure(figsize=(8, 4))
sns.violinplot(x='variable', y='value', data=pd.melt(plot_data),
palette="muted")

# Add labels and title
plt.xlabel('Residence Type')
plt.ylabel('Stroke Incidence')
plt.title('Violin Plot of Stroke Incidence in Rural and Urban Areas')

# Show the plot
plt.show()

# Frequency table for stroke incidence in rural and urban areas
frequency_table = pd.crosstab(data['Residence_type'],
data['stroke'], margins=True, margins_name='Total')
print("\nFrequency Table:")
print(frequency_table)

    Using chi-square test

import pandas as pd
from scipy.stats import chi2_contingency
from tabulate import tabulate

# Load your dataset
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Create a contingency table
contingency_table = pd.crosstab(data['Residence_type'], data['stroke'])

# Print the contingency table in a tabular format
print("Contingency Table:")
print(tabulate(contingency_table, headers='keys', tablefmt='fancy_grid'))

# Perform Chi-square test
chi2, p_value, degrees_of_freedom, _ = chi2_contingency(contingency_table)

# Print the results
print("\nChi-square Test Results:")
print(f"Chi2: {chi2}")
print(f"Degrees of Freedom: {degrees_of_freedom}")

```



```

print(f"P-value: {p_value}")

# Interpretation
if p_value < 0.05:
    print("The Chi-square test suggests a significant association
between Residence_type and stroke incidence.")
else:
    print("There is not enough evidence to conclude a
significant association between Residence_type and stroke incidence.")

    Develop and train machine learning model capable of predicting the
probability of stroke.

import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix,
roc_auc_score, roc_curve
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline

# Load your dataset (replace 'your_dataset_path' with the actual path)
data = pd.read_csv("C:\\Users\\hp\\Downloads\\brain_stroke.csv")

# Assuming 'stroke' is the target variable and other columns are features
X = data.drop('stroke', axis=1)
y = data['stroke']

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

# Define a column transformer for one-hot encoding categorical variables
categorical_cols = ['gender', 'ever_married', 'work_type',
'Residence_type', 'smoking_status']
numeric_cols = ['age', 'hypertension', 'heart_disease',
'avglucose_level', 'bmi']

preprocessor = ColumnTransformer(
    transformers=[
        ('num', 'passthrough', numeric_cols),
        ('cat', OneHotEncoder(), categorical_cols)
    ]
)

```

```

))

# Build the pipeline with preprocessing and classifier
clf = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(n_estimators=100,
max_depth=10, random_state=42))
])

# Train the model
clf.fit(X_train, y_train)

# Make predictions on the test set
y_pred = clf.predict(X_test)

# Print classification report and confusion matrix
print("Classification Report:")
print(classification_report(y_test, y_pred))

print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

# Calculate AUC-ROC
y_prob = clf.predict_proba(X_test)[:, 1]
auc_roc = roc_auc_score(y_test, y_prob)
print(f"AUC-ROC: {auc_roc}")

# Plot ROC curve
fpr, tpr, _ = roc_curve(y_test, y_prob)

plt.figure(figsize=(6, 4))
plt.plot(fpr, tpr, color='darkorange', lw=2, label='ROC
curve (area = {:.2f})'.format(auc_roc))
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc="lower right")
plt.show()

```