COMBINING STATISTICAL ANALYSIS AND MACHINE LEARNING TO EXPLORE THE INTERPLAY BETWEEN AGING, LIFESTYLE CHOICES, CARDIOVASCULAR DISEASES, AND BRAIN STROKES

By

Anit Mathew, MS. Data Science

A thesis submitted to the Graduate Committee of

Ramapo College of New Jersey in partial fulfillment

of the requirements for the degree of

Master of Science in Data Science

Fall, 2023

Committee Members:

Dr. Osei Kofi Tweneboah, Advisor

Dr. Scott Frees, Reader

Dr. Sourav Dutta, Reader

COPYRIGHT

Dedication

I dedicate this thesis to the following individuals and groups who have been instrumental in my academic journey and personal growth

To my wife, Sherin Anit Mathew for her unwavering support, love, and encouragement throughout the years. Her sacrifices and belief in me have been the driving force behind my achievements.

To my dedicated thesis advisor, Dr. Osei Kofi Tweneboah, for his guidance, expertise, and patience. His mentorship has been invaluable in shaping my research and academic development.

Acknowledgements

I would like to express my deepest gratitude to Professor Dr. Osei Kofi Tweneboah for their unwavering support and invaluable guidance throughout the journey of completing this thesis. Professor Tweneboah has been an exceptional mentor, providing not only expert academic advice but also fostering an environment of intellectual curiosity and scholarly rigor. His dedication to pushing the boundaries of knowledge has inspired me to approach research with enthusiasm and thoroughness.

I am indebted to Professor Tweneboah for his patience in answering my countless questions, for his insightful feedback that significantly improved the quality of my work, and for the countless hours spent in meaningful discussions that shaped the direction of this research.

Moreover, I appreciate the mentorship that extends beyond the academic realm. Professor Tweneboah's commitment to fostering a holistic understanding of the subject matter has been instrumental in my personal and professional development.

I would also like to express my gratitude for the resources, opportunities, and encouragement provided by Professor Tweneboah. His support has been pivotal in transforming this thesis from a concept into a comprehensive piece of scholarly work.

Thank you, Professor Tweneboah, for your indispensable contribution to this academic endeavor.

Table of Contents

List of Tables

List of Figures

Abstract

This research project aims to investigate the intricate relationship between aging, lifestyle choices, cardiovascular diseases (CVDs), and brain strokes in older adults. The pressing problem is the growing burden of CVDs and strokes among the elderly, and the need to understand the impact of lifestyle factors on these health outcomes.

The primary objectives of this study are to assess the incidence of heart disease, high blood pressure, and brain strokes in the senior population, analyze how lifestyle factors like smoking status and body mass index (BMI) influence stroke frequency, explore the connections between heart disease, hypertension, and strokes, and investigate the potential influence of additional variables such as gender, average blood glucose levels, and type of residence. Furthermore, this research seeks to propose interventions and preventive strategies to reduce the incidence of brain strokes among older adults.

This research employs a comprehensive analysis of a publicly accessible dataset from Kaggle, which contains a wide range of health-related variables. The dataset provides valuable insights into lifestyle choices, health conditions, and the occurrence of brain strokes in older individuals. Various statistical and data analysis techniques will be applied to uncover associations and trends, contributing to a deeper understanding of the complex interactions between lifestyle choices, CVDs, and brain strokes.

Through a meticulous examination of the data, this study intends to shed light on the multifaceted relationships among lifestyle choices, cardiovascular diseases, and strokes in the elderly. The results will contribute to public health, geriatrics, and medical fields by providing evidence-based knowledge that can inform strategies for risk assessment, disease management, and health promotion among older adults.

This research project holds the potential to benefit multiple stakeholders. For healthcare professionals, the findings can lead to the development of effective strategies for the management of CVDs and strokes in older individuals. It may also inform public health campaigns and policy initiatives aimed at reducing the risk of these conditions within an aging population. Additionally, the study contributes to the existing body of knowledge in this field, providing a foundation for further research and the potential discovery of new interventions and risk mitigation strategies. Overall, this research addresses a critical health concern affecting older adults and has the potential to improve the well-being of this vulnerable population.

Introduction

As the global population ages, the impact of cardiovascular diseases (CVDs) on the elderly has emerged as a critical concern in public health. CVDs, comprising a range of disorders affecting the heart and blood vessels, stand as the foremost cause of mortality worldwide, claiming approximately 17.9 million lives annually. Among these, coronary heart disease, cerebrovascular disease, and rheumatic heart disease contribute significantly to the burden of CVDs [1].

The grim reality is that more than four out of five deaths related to CVDs find their origin in heart attacks and strokes, with a particularly distressing one-third occurring prematurely among individuals under the age of 65. This demographic twist intensifies the urgency to address the complex interplay between aging, cardiovascular health, and associated risk factors. This escalating threat to the aging population is compounded by the prevalence of behavioral risk factors such as unhealthy diet, physical inactivity, tobacco use, and harmful alcohol consumption. These factors manifest physiologically as elevated blood pressure, increased blood glucose levels, raised blood lipids, and the onset of overweight and obesity. Recognized as "intermediate risk factors," these markers signal an augmented susceptibility to heart attacks, strokes, heart failure, and related complications[1].

The silver lining amidst these challenges lies in the demonstrated efficacy of interventions aimed at modifying behavior to mitigate CVD risks. Strategic measures such as tobacco cessation, controlled salt intake, heightened consumption of fruits and vegetables, regular physical activity, and moderation in alcohol use have proven instrumental in diminishing the overall risk of cardiovascular diseases. However, the success of these interventions hinges on the implementation of robust health policies that foster environments conducive to healthy choices, ensuring that such measures are not only accessible but also affordable to all segments of the aging population[19].

In the context of an aging population, the imperative becomes clear: identifying those at the highest risk of CVDs and deploying timely and appropriate interventions. The crux lies in guaranteeing access to noncommunicable disease medicines and essential health technologies in primary healthcare facilities, ensuring that the elderly receive the requisite treatments and counseling. Thus, this thesis embarks on a journey to unravel the intricate relationship between aging, lifestyle choices, and cardiovascular diseases. Through a meticulous exploration of effective strategies for risk assessment, management, and the promotion of cardiovascular health in the elderly, we aspire to contribute to the collective effort to mitigate the burgeoning impact of CVDs on our aging global community.

This research is structured to address each of the objectives systematically. It will consist of chapters covering the literature review, data collection and analysis, results, and discussions. Each chapter will build upon the findings of the previous one, culminating in a comprehensive understanding of the relationships between aging, lifestyle choices, cardiovascular diseases, and brain strokes in older adults.

The thesis explores key features on grading the factors of stroke across different age groups, focusing on oversampled data from ages 0-100, above 65, and 0-65. Through the analysis of various models—Logistic, Decision Tree, DNN, and Random Forest—consistent patterns emerge, highlighting the top predictors for stroke risk. Hypertension and heart disease consistently stand out across all age groups, aligning with established medical knowledge. The influence of marital status, residence type, and average glucose level is also notable, suggesting the importance of social, environmental, and lifestyle factors. The implications of these findings are discussed, emphasizing the need for targeted interventions and public health strategies. Limitations of oversampled data are acknowledged, and recommendations for future research are provided. The thesis contributes valuable insights for healthcare professionals and policymakers in devising effective strategies for stroke prevention and management across diverse age groups.

Problem Statement

Cardiovascular diseases, including heart disease and high blood pressure, are leading causes of mortality and morbidity, particularly among older adults. Brain strokes, both ischemic and hemorrhagic, represent a significant and potentially devastating health event for seniors. Lifestyle factors such as smoking habits, body mass index (BMI), and other variables, including gender, marital status, occupation, residence type, and average glucose levels, have been shown to influence the development of CVDs and strokes in this population. The complex interaction of these factors poses a significant challenge to both researchers and healthcare professionals.
This research project aims to address the following fundamental questions:

- What is the incidence of heart disease, high blood pressure, and brain strokes in the elderly population?
- How do lifestyle factors, such as smoking status and BMI, impact the frequency of brain strokes?
- What are the relationships between heart disease, hypertension, and brain strokes in older individuals?
- Can additional variables, such as gender, average blood glucose levels, and type of residence, influence the risk of brain strokes in seniors?
- What key factor should the elderly prioritize to safeguard themselves from cardiovascular diseases (CVDs)?
- Examining various age groups by utilizing a variable importance plot to gain insights into the key factors contributing to cardiovascular diseases (CVDs).

Aim and Significance

The primary aim of this research is to gain a deeper understanding of the interplay between lifestyle choices, cardiovascular diseases, and brain strokes in older adults. By investigating these relationships, we intend to provide valuable insights for:

- Public Health: The findings may inform prevention strategies, early detection, and targeted interventions to reduce the burden of cardiovascular diseases and strokes in the elderly population.
- Elderly Care and Geriatrics: Healthcare professionals, including geriatricians, cardiologists, and neurologists, can use the results to develop effective strategies for risk assessment, disease management, and overall health improvement for older adults.
- Health Promotion and Policy: The results may support policy efforts related to cardiovascular health and stroke prevention, particularly in the context of an aging population. Health promotion campaigns and interventions can be designed to target specific lifestyle factors that influence the development of cardiovascular diseases and strokes in older individuals.
- Research Foundation: This research project contributes to the existing body of knowledge on the relationship between lifestyle factors, cardiovascular diseases, and strokes in older adults. The findings may serve as a basis for further research, leading to advances in the field and the potential discovery of new interventions and risk mitigation strategies.

Expected Outcomes

- Identification of High-Risk Groups: The research aims to identify specific demographics within the elderly population that are at a higher risk of cardiovascular diseases and strokes. This information is crucial for targeted interventions and personalized healthcare.
- Understanding Lifestyle Impacts: By delving into the relationships between lifestyle factors and health outcomes, the study aims to elucidate how behaviors such as smoking and BMI contribute to the incidence of cardiovascular diseases and strokes in older individuals.
- Policy Recommendations: The research findings will contribute to evidence-based policy recommendations for public health initiatives and healthcare strategies tailored to the aging population. This includes recommendations for lifestyle interventions, healthcare access improvements, and preventive measures.
- Healthcare Professional Guidance: Healthcare professionals will benefit from insights into effective risk assessment and management strategies, enabling them to provide more targeted and personalized care for older adults.
- Contribution to Scientific Knowledge: The research project will contribute to the scientific understanding of the complex interplay between aging, lifestyle choices, and cardiovascular health. This knowledge will serve as a foundation for future studies and potential breakthroughs in preventive healthcare.

Background

In the context of cardiovascular diseases (CVDs) and strokes, this research aims to contribute to the understanding of the intricate relationships between aging, lifestyle choices, and health outcomes in older adults. The growing elderly population globally, while indicative of longer life expectancy, presents a significant challenge due to the increasing burden of age-related health issues. Among these, CVDs and strokes emerge as major contributors to disability and mortality in older individuals.

Observing the challenges faced by the elderly population due to age-related health issues, particularly CVDs and strokes, fueled my commitment to contribute meaningfully to this field. Additionally, academic exposure to the evolving landscape of machine learning (ML) and its applications in healthcare further inspired my research direction. Recognizing the potential of ML models in predicting strokes based on diverse participant profiles, I sought to leverage these tools to enhance our understanding of the multifaceted factors influencing stroke risk in older adults. This thesis, therefore, represents a fusion of personal experiences and academic curiosity, driven by the goal of making impactful contributions to the prevention and management of CVDs and strokes in the aging population.

The burden of CVDs, encompassing conditions such as heart disease and hypertension, is particularly pronounced in the aging population. These health challenges often lead to strokes, which can result in long-term neurological impairment or even death. Lifestyle choices, including factors such as smoking and body mass index (BMI), play a crucial role in the development and progression of CVDs and strokes. Understanding the impact of these lifestyle factors on the frequency of strokes in older adults is essential for devising effective preventive strategies.

Furthermore, the interconnections between heart disease, hypertension, and strokes are central to this research. Many strokes, classified as ischemic, stem from obstructed blood flow to the brain, often linked to conditions like atherosclerosis. Exploring the relationships between these interconnected health issues is vital for comprehensive preventive measures.

Beyond lifestyle choices and direct health conditions, this study delves into the influence of additional variables such as gender, average blood glucose levels, and type of residence. By considering a broad range of factors, a holistic understanding of the contributors to CVDs and strokes in older adults is sought.

In comparing the findings with previous work, the prevalence of stroke and its profound impact on global health necessitates effective strategies for early prediction and risk assessment. In the study performed by Elias Dritsas and Maria Trigka, they leveraged machine learning (ML) techniques to develop models for long-term stroke risk prediction [4]. The proposed approach, particularly the novel stacking method, demonstrated promising results in terms of various performance metrics, including AUC, precision, recall, F-measure, and accuracy[4].

The stacking method, a combination of multiple ML models, showcased superior predictive capabilities compared to individual models and majority voting. The achieved AUC of 98.9% highlights the robustness of the stacking method in discerning stroke risk. Precision, recall, and F-measure, crucial for evaluating the model's ability to correctly identify positive instances, exhibited values of 97.4%, indicative of the model's high accuracy in classifying stroke occurrences. The overall accuracy of 98% further reinforces the effectiveness of our proposed approach[4].

The significance of early stroke prediction cannot be overstated, considering the staggering statistics provided by the World Stroke Organization, indicating the annual occurrence of 13 million strokes and 5.5 million associated deaths. Stroke's far-reaching impact on individuals and their social environment underscores the importance of proactive measures. Our research contributes to this imperative by providing a reliable framework for long-term stroke risk assessment.

Understanding the factors influencing stroke risk is crucial for effective prediction. Our analysis considered a range of factors, including demographic information, medical history, and lifestyle factors, in line with established risk factors such as hypertension, heart disease, age, smoking, and diabetes. The use of a balanced dataset, facilitated by the synthetic minority over-sampling technique (SMOTE), addressed class imbalance issues, enhancing the models' ability to generalize.

The literature review revealed a growing interest in utilizing ML for stroke risk prediction. Comparing our results with prior studies, the stacking method outperformed other algorithms, demonstrating its efficacy in enhancing predictive accuracy. The diverse set of ML models evaluated, including naive Bayes, logistic regression, stochastic gradient descent, K-NN, decision trees, random forests, and multi-layer perception, contributed to a comprehensive understanding of the predictive landscape.

Despite the success of ML models in predicting strokes, a limitation noted is the reliance on publicly available datasets, which may lack the richness of information obtained from hospital or institute data. The research concludes with a call for further enhancements to the ML framework, potentially incorporating deep learning methods. Ultimately, the stacking method is highlighted as the best-performing approach and a primary recommendation for stroke prediction based on the study's findings.

Skewness

Skewness, denoted as Skew(X), quantifies the asymmetry in the probability distribution of a real-valued random variable X. In the context of a dataset, skewness has been computed for various numerical features. The skewness value, Skew(X), represents the degree of asymmetry, where a positive value indicates a right-skewed distribution, a negative value indicates a left-skewed distribution, and a skewness of zero suggests a perfectly symmetrical distribution [6].

$$\text{Skew} = 3(\text{mean - median}) / \text{standard deviation}$$

Chi-Square Test

The Chi-square test is a statistical method used to assess the association between categorical variables. To determine the association between each categorical variable and the occurrence of strokes in the dataset, the Chi-square test can be applied to examine if there is a significant relationship between the two[18].

Let's denote the categorical variable as X and the occurrence of strokes as Y. The null hypothesis (H_0) assumes that there is no association between the categorical variable and the occurrence of strokes, while the alternative hypothesis (H_1) suggests that there is a significant association.

The Chi-square statistic (X^2) is calculated using the formula:

$$X^2=\sum(O_i-E_i)^2/E_i$$

where:

- O_i is the observed frequency in each category,
- E_i is the expected frequency in each category under the assumption of no association,
- The summation ($\sum$) is taken over all categories [18].

The expected frequency (E_i) for each category is calculated as:

$$E_i = (row\ total \times column\ total) \ / \ grand\ total$$

The degrees of freedom (*df*) for the Chi-square test in this context would be (Number of rows−1) × (number of columns−1) (number of rows−1) × (number of columns−1) [18].

Once the Chi-square statistic is calculated, it is compared to the critical value from the Chi-square distribution with the given degrees of freedom. If the calculated Chi-square statistic is greater than the critical value, the null hypothesis is rejected, indicating a significant association between the categorical variable and the occurrence of strokes.

Chi-Square Test Parameters:

- Chi-Square Statistic: A measure of the strength of association or dependence between the two variables as explained above.
- p-value: The p-value is the probability that the observed relationship occurred by chance. It is obtained by comparing the calculated Chi-Square statistic to the Chi-Square distribution with the appropriate degrees of freedom. The calculation involves the cumulative distribution function (CDF) of the Chi-Square distribution. The smaller the p-value, the more significant the association. Mathematically, it can be expressed as [18]:

$$p - value = P(X^2 \geq observed\ X^2 |\ H_0\ is\ true)$$

- Degrees of Freedom(*df*): The degrees of freedom for the Chi-Square test are calculated based on the number of categories in the contingency table. For a contingency table with r rows and c columns, the degrees of freedom (df) are given by [18]:

$$df = (r - 1) \times (c - 1)$$

Uneven Anova Test

ANOVA, or Analysis of Variance, is a statistical test used to compare means across different groups. The basic idea is to determine if there are any statistically significant differences between the means of the groups being compared. ANOVA can be classified into two main types: one-way ANOVA and two-way ANOVA. One-way ANOVA is appropriate when there is one independent variable, and it compares the means of three or more groups.

Now, when we talk about "uneven ANOVA," it might refer to situations where the sample sizes in the different groups are not equal. This is also known as "unbalanced" data. Uneven sample sizes can affect the power and sensitivity of the ANOVA test.

The one-way ANOVA test statistic is based on the F-ratio, which is the ratio of the variance between groups to the variance within groups. In mathematical terms, the F-ratio is calculated as follows:

$$F = MSB/MSW$$

where:
F is the F-ratio,
MSB is the mean square between groups,
MSW is the mean square within groups.

The mean square between groups *(MSB)* is a measure of how much the group means differ from each other, and the mean square within groups *(MSW)* is a measure of how much individual observations within each group vary from their group mean.

For unbalanced data, the formulas for *(MSB)* and *(MSW)* become a bit more complex due to the uneven sample sizes. The general idea, however, remains the same: you're comparing the variation between groups to the variation within groups.

In simple terms, if the F-ratio is sufficiently large, it suggests that there are significant differences between the group means. To determine whether this difference is statistically significant, you compare the F-ratio to a critical value from a statistical table or use a p-value.

In summary, uneven ANOVA deals with situations where sample sizes in different groups are not equal, and the F-ratio is a key statistic used to assess whether the differences between group means are significant or just due to random variability.

Data Preprocessing - Oversampling

Oversampling is a technique used in data analysis to address the issue of imbalanced datasets, where one class or outcome is significantly more prevalent than the other(s) [3] .

In this case, the dataset contains information related to stroke, and it appears that the "stroke" class is imbalanced, meaning there are more instances of stroke cases (class 1) compared to non-stroke cases (class 0). To handle this imbalance and ensure that the model doesn't exhibit bias toward the majority class, oversampling was performed.

Oversampling involves creating more instances of the minority class (class 0, in this case) so that the dataset becomes more balanced [3]. This can be done through various techniques, here we have used Random Oversampling:

Random Oversampling: In random oversampling, additional instances of the minority class are generated by randomly duplicating existing data points. This is a simple method but can lead to overfitting if not carefully implemented [3].

Let's denote:

N_{maj}: the number of instances in the majority class.
N_{min}: the number of instances in the minority class.
p: the oversampling ratio, representing how to oversample the minority class. For example, if $p = 0.5$, it means to increase the number of instances in the minority class by 50%.

The oversampled number of instances ($N_{\text{min}\_oversampled}$) in the minority class can be calculated as:

$$N_{\text{min}\_oversampled} = \ N_{min} \times (1 + p)$$

The total number of instances after oversampling ($N_{totaloversampled}$) is then:

$$N_{totaloversampled} = N_{maj} + N_{minoversampled}$$

After applying random oversampling, the class distribution should be more balanced. Keep in mind that while oversampling can help address the imbalanced dataset issue, it may also lead to overfitting, so it's important to evaluate the performance of the model on a separate validation or test set.

By using one of these oversampling techniques, the goal is to create a more balanced dataset, where the number of stroke cases and non-stroke cases is closer to being equal. This helps the machine learning model learn from both classes more effectively, leading to a more balanced and unbiased prediction [3].

Data Splitting

To build and evaluate a predictive model for stroke incidence, it is crucial to appropriately split the dataset into training and testing sets. The dataset, obtained from Kaggle, consists of 10 columns capturing various parameters such as gender, age, hypertension, heart disease, marital status, work type, residence type, average glucose level, BMI, smoking status, and the occurrence of a stroke. Given the binary nature of the target variable "stroke," with values of 0 or 1 indicating its absence or occurrence, the dataset is amenable to a classification model. To ensure the model's ability to generalize to new, unseen data, a common practice involves randomly partitioning the dataset into a training set, used for model training, and a testing set, reserved for model evaluation. A typical split might involve allocating, for example, 80% of the data to the training set and the remaining 20% to the testing set. This ensures that the model learns patterns from the majority of the data and is subsequently validated on a separate, independent subset to assess its predictive performance. The link provided directs to the Kaggle dataset, offering the opportunity for further exploration and analysis of this valuable resource in the context of stroke prediction modeling [8].

Machine Learning Models

Logistic Regression

Logistic Regression is a widely used statistical method for binary classification tasks. It models the probability that an instance belongs to a particular class, often denoted as 1 or 0. The logistic function, also known as the sigmoid function, is employed to constrain the output between 0 and 1 [9].

Mathematical Equation: The logistic regression model is represented by the following equation:

$$P(Y=1) = 1 / 1+ e^{-(\beta_0+\beta_1 X_1 +\beta_2 X_2 +\ldots+\beta_n X_n)}$$

Where:

- $P(Y = 1)$ is the probability of the positive class,
- e is the base of the natural logarithm,
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients, and
- $X_1, X_2, \ldots, X_n$ are the feature values.

For more details, please refer [9]

Decision Tree

Decision Trees are non-linear models used for both classification and regression tasks. They recursively partition the feature space based on the most informative features, creating a tree-like structure of decisions.

Mathematical Notation: Let Rm represent the region (node) in the feature space created by the splits. The decision tree can be expressed as:

$$R_m = \{X \mid X_j \leq t_m\}$$

Where:

X_j is the $j-th$ feature,
t_m is the threshold for node m.

For more details, please refer [19]

Deep Learning - DNN model

Deep Neural Networks are a class of machine learning models inspired by the structure and function of the human brain. They consist of interconnected layers of artificial neurons, each layer learning hierarchical representations of the input data.

Deep Neural Networks (DNNs) can be well-suited for cardiovascular data for several reasons:

- Non-Linearity and Complex Patterns: Cardiovascular data often contains non-linear relationships and complex patterns that may not be effectively captured by linear models. DNNs, with their multiple layers and non-linear activation functions, can learn intricate patterns and relationships within the data.
- Automatic Feature Learning: DNNs are capable of automatically learning hierarchical representations from the data, reducing the need for manual feature engineering. This is particularly advantageous when dealing with large and complex datasets, as seen in cardiovascular studies.
- Prediction and Risk Stratification: Cardiovascular data often involves predicting outcomes or risk stratification. DNNs have shown promise in predictive modeling tasks, including the prediction of cardiovascular events based on patient data.
- Regularization Techniques: DNNs come with various regularization techniques, such as dropout and batch normalization, which help prevent overfitting and improve the generalization of the model.

Mathematical Notation: Consider a DNN with L layers. The output of the $i-th$ neuron in the $l-th$ layer can be expressed as:

$$ai(l) = g(zi(l))$$

Where:

- g is the activation function,
- $zi(l)$ is the weighted sum of inputs for neuron i in layer l.

For more details, please refer [11]

Random Forest Model

Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Mathematical Concept: Let $T_i(x)$ be the output of the $i - th$ tree for input x, and $H(x)$ be the output of the Random Forest, the aggregated output is defined as:

$$H(x)= {}^{1}/_{N} \sum {}^{N}_{i=1} T_i \ (x)$$

Where:
N is the number of trees in the forest.

These models offer a diverse set of tools for different problem domains, each with its strengths and weaknesses. The choice of the model should be based on the characteristics of the data and the goals of the analysis.

For more details, please refer [21]

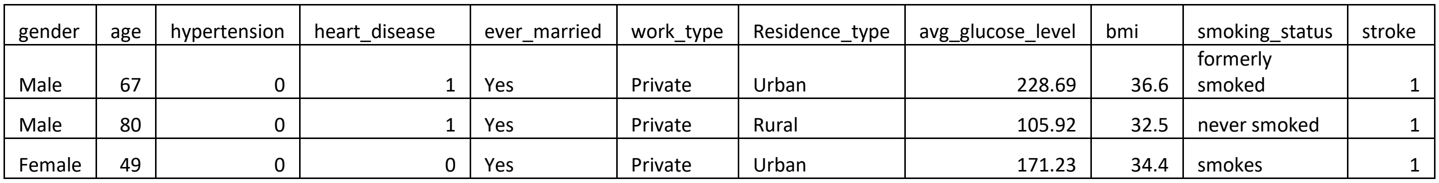
Methodology

Scope and Methodology

This study will employ a comprehensive analysis of a publicly accessible dataset, which contains a wide range of health-related variables. The dataset will be used to examine lifestyle factors, health conditions, and the occurrence of brain strokes in older individuals. Statistical and data analysis techniques will be applied to uncover associations and trends. The decision to focus the research exclusively on older adults aged 65 and above is grounded in the recognition of the unique healthcare challenges and demographic shifts associated with this age group. As the global population ages, there is a growing need to understand and address the health-related concerns specific to older individuals. The research will focus on older adults aged 65 and above, using the dataset to investigate the relationships between lifestyle factors, cardiovascular diseases, and brain strokes. The primary variables of interest include heart disease, high blood pressure, smoking status, BMI, gender, average glucose levels, and type of residence.

Data

The dataset comprises 10 columns, each providing valuable information related to stroke incidence. A brief description of each column is as follows:



gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1

Table 3.1: Dataset

Gender: The gender of the individual (e.g., Male or Female).
Age: The age of the individual in years.
Hypertension: A binary variable indicating the presence (1) or absence (0) of hypertension.
Heart Disease: A binary variable indicating the presence (1) or absence (0) of heart disease.
Ever Married: A binary variable indicating whether the individual has ever been married (Yes or No).
Work Type: The type of work the individual is engaged in (e.g., Private, Self-employed).
Residence Type: The type of residence the individual lives in (e.g., Urban or Rural).
Average Glucose Level: The average glucose level in the individual's blood.
BMI: The Body Mass Index of the individual, representing their body weight in relation to height.
Smoking Status: The smoking status of the individual (e.g., formerly smoked, never smoked, smokes).
Stroke: A binary variable indicating the occurrence (1) or absence (0) of a stroke.

For more details on the dataset, please refer [2].

Elderly Population

According to the National Center for Biotechnology Information, the term "elderly population" traditionally refers to individuals aged 65 and older. In 1987, the United States had just over 30 million elderly individuals, constituting more than 12 percent of the total U.S. population of nearly 252 million. This demographic group comprises a significant majority, almost 96 percent, of Medicare recipients, emphasizing its substantial impact on healthcare considerations [5].

The growth of the elderly segment of the U.S. population has outpaced the overall population, a phenomenon commonly referred to as "the graying of America." Data from the National Center for Health Statistics (NCHS) reveals that between 1960 and 1986, the population aged 65 and older experienced a remarkable 75 percent increase, rising from almost 17 million to over 29 million individuals. In contrast, the population under 65 increased by only 30 percent during the same period. Among those aged 65 and older in 1986, approximately three-fifths fell within the 65 to 74 age group, one-third were in the 75 to 84 age group, and one-tenth were 85 and older. Notably, the rate of growth for the older age groups (75 to 84 and 85 and older) surpassed that of the 65 to 74 age group between 1960 and 1986.

Projections indicate that from 1987 to 2030, the total U.S. population is expected to increase by 26 percent, reaching 317 million, while the population aged 65 and older is projected to experience a more than 100 percent increase. This would elevate the proportion of the elderly population from the initial 12 percent to nearly 21 percent of the total population, totaling 67 million individuals by 2030. These projections underscore the sustained growth and aging of the elderly population, presenting significant implications for healthcare demands and services catering to the unique challenges faced by this demographic, such as multiple chronic illnesses and the necessity for adequate means to support independent living.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process that involves summarizing, visualizing, and understanding the main characteristics of a dataset. It helps analysts and data scientists gain insights, detect patterns, identify outliers, and formulate hypotheses about the data. The descriptive analysis is performed on the original dataset to better the current scenario of the situation [15].

The histogram (figure 3.1) below displays the distribution of strokes by age. The number of strokes increases with age, with the highest number of strokes occurring in people over 80 years old. However, strokes can occur at any age, and the number of strokes in people under 50 years old is also significant. The histogram displays that the distribution of strokes by age is not uniform. There are two distinct peaks in the distribution, one at around 55 years old and one at around 80 years old. This suggests that there are two different risk factors for stroke, one that affects younger people and one that affects older people.

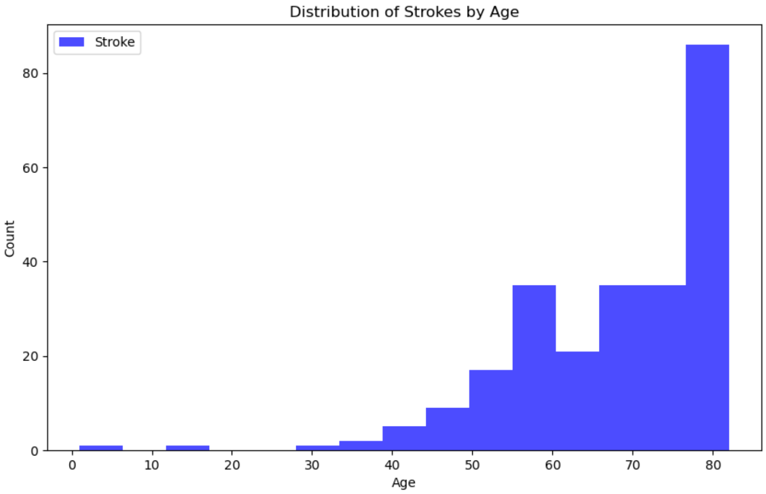


Figure 3.1: Distribution of Strokes by Age

The plot below (figure 3.2) shows the distribution of strokes by age and hypertension. The stroke distribution is shown as a violin plot, where the width of the plot at each point represents the density of strokes at that age and hypertension level. The violin plot shows that the risk of stroke increases with both age and hypertension. The median age of stroke is around 65 years old, and the risk of stroke is significantly higher in people with hypertension than in people without hypertension. The violin plot also shows that the distribution of strokes is skewed to the right, meaning that there are more strokes in older people and people with higher blood pressure. This is because the risk of stroke accumulates over time, and hypertension is a major risk factor for stroke. Overall, the image shows that age and hypertension are two of the most important risk factors for stroke. People should be aware of their risk factors and take steps to reduce their risk, such as controlling their blood pressure and maintaining a healthy lifestyle.

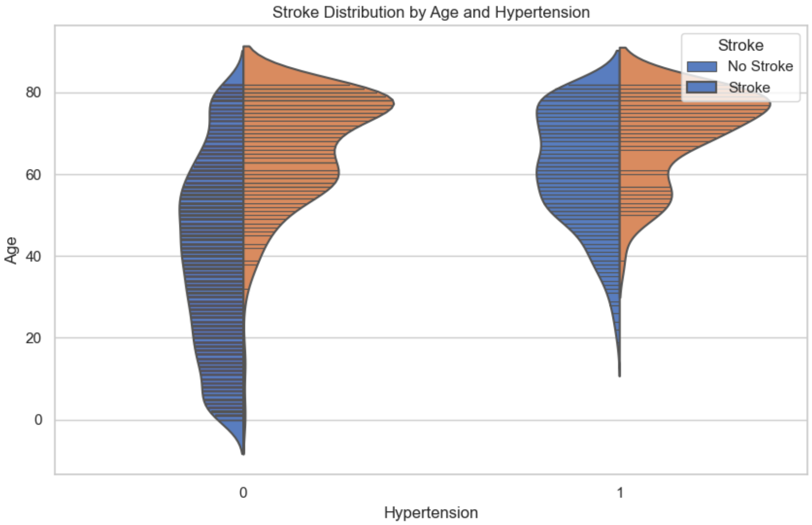


Figure 3.2: Stroke Distribution by Age and Hypertension

The plot below (figure 3.3) data illustrates the relationship between heart disease, age groups, and the occurrence of strokes. It is structured as a table with two categories for "heart disease" (0 and 1) and four age groups: 0-18, 19-40, 41-60, and 61+. The values in the table represent the proportion of individuals in each category who have experienced a stroke. The data reveals distinct patterns: In the absence of heart disease (heart disease=0), the likelihood of a stroke is relatively low across all age groups, with the highest occurrence observed among individuals aged 61 and above. Conversely, for those with heart disease (heart disease=1), the risk of stroke significantly increases, with the highest incidence among individuals aged 61 and above. This data provides valuable insights into the influence of heart disease and age on stroke occurrence. It suggests that older individuals are generally at higher risk of stroke, and the presence of heart disease amplifies this risk considerably, particularly among those in the 61+ age group. These findings are relevant for understanding stroke prevention and management strategies, especially in the context of heart disease and age-related risk factors.

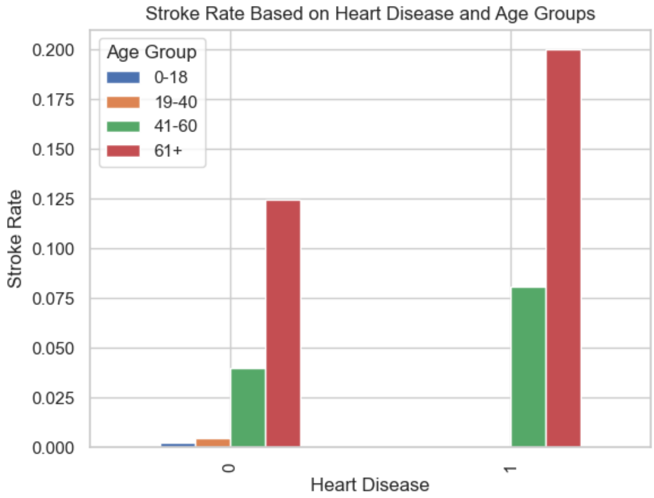


Figure 3.3: Stroke Rate Based on Heart Disease and Age Groups

The heatmap below (figure 3.4) shows that the risk of stroke increases with both BMI and age. This is because obesity and aging are two of the most important risk factors for stroke. People with obesity are more likely to have other risk factors for stroke, such as high blood pressure, high cholesterol, and diabetes. Aging also increases the risk of stroke because the arteries can become narrowed and hardened over time, which can reduce blood flow to the brain. The heatmap also shows that the risk of stroke is highest in people with obese overweight and who are in the 61+ age group. This suggests that people in this category should be especially careful to manage their risk factors for stroke.

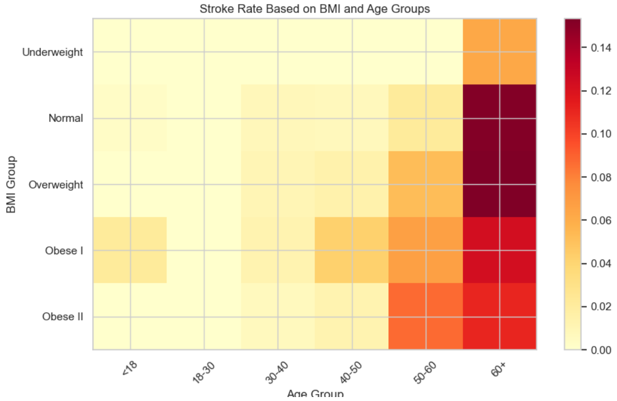


Figure 3.4: Stroke Rate Based on BMI and Age Groups

The graph below (figure 3.5) shows that the stroke rate is highest for people in government jobs, followed by people in private jobs and self-employed people. Children have the lowest stroke rate. The stroke rate also increases with age. For example, the stroke rate for people in government jobs is about 0.4 per 100,000 people for people aged 20-24, but it increases to about 3 per 100,000 people for people aged 65 and over. There are a number of possible explanations for the differences in stroke rate by work type and age group. One possibility is that people in government jobs are more likely to have sedentary jobs, which can increase the risk of stroke. Another possibility is that people in government jobs are more likely to be exposed to stress, which can also increase the risk of stroke. Self-employed people may also have a higher stroke rate because they are more likely to work long hours and have irregular schedules. Children have the lowest stroke rate because they are less likely to have the risk factors for stroke, such as high blood pressure, high cholesterol, and diabetes.

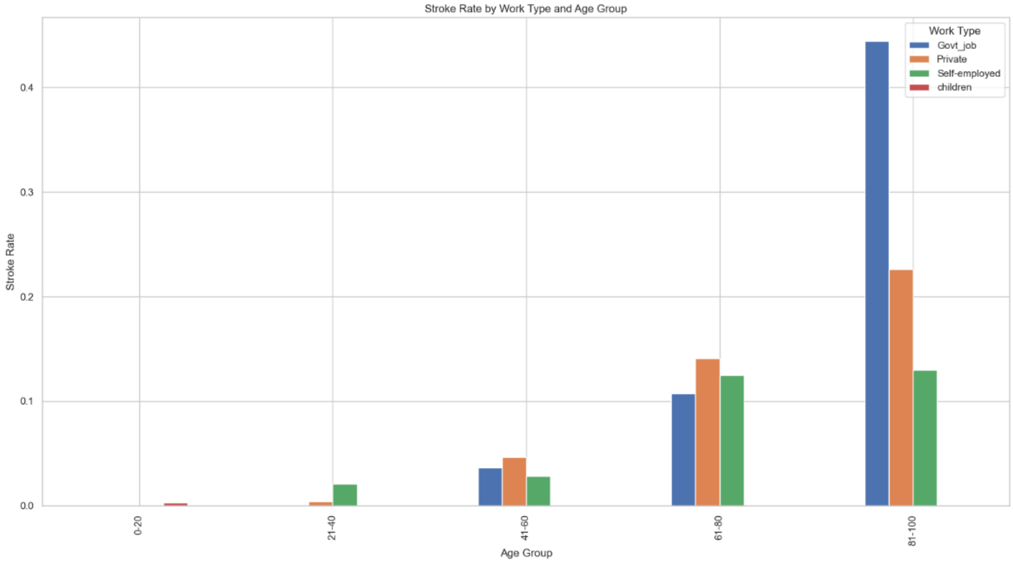
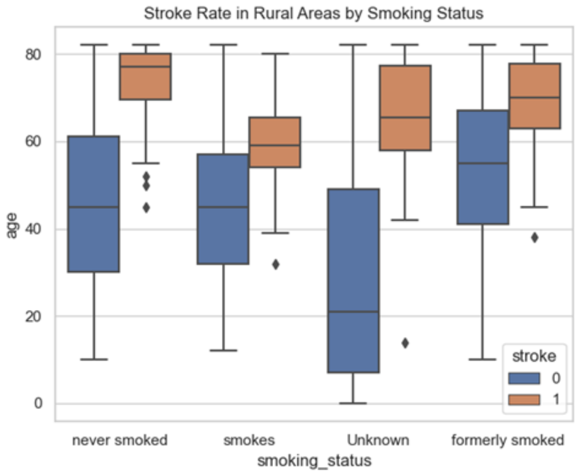


Figure 3.5: Stroke Rate Based on Work type and Age Group

The box plots below (figure 3.6(a) & (b)) show the median, 25th and 75th percentiles, and outliers. The median stroke rate is highest for smokers in urban areas (figure 3.6(b)), followed by smokers in rural areas (figure 3.6(a)), former smokers in urban areas, and former smokers in rural areas. Never smokers in both urban and rural areas have the lowest median stroke rates. The interquartile ranges (IQRs) are also higher for smokers than for never smokers in both urban and rural areas. This means that the distribution of stroke rates is more spread out for smokers, with a higher proportion of smokers having very high or very low stroke rates. There are a number of possible explanations for the differences in stroke rates by smoking status and urbanicity. One possibility is that smokers are more likely to have other risk factors for stroke, such as high blood pressure, high cholesterol, and diabetes. Another possibility is that smokers are more likely to be exposed to environmental risk factors for stroke, such as air pollution and secondhand smoke. People living in urban areas may also have a higher risk of stroke because they are more likely to be exposed to stress and have unhealthy lifestyles. However, the image shows that the differences in stroke rates by smoking status are larger than the differences in stroke rates by urbanicity. This suggests that smoking is a more important risk factor for stroke than urbanicity.



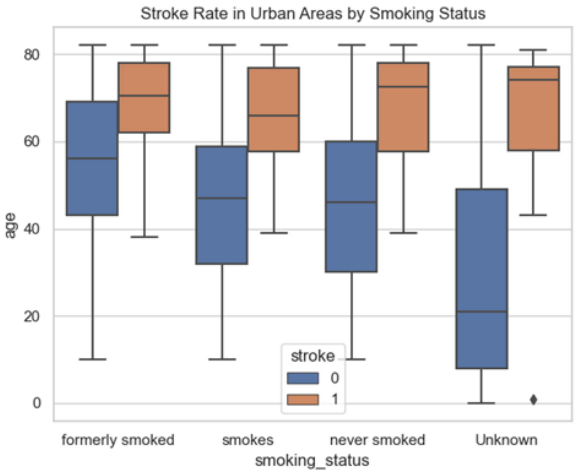


Figure 3.6: Stroke Rate Based on Rural Areas and Age Group (left)
Figure 3.7: Stroke Rate Based on Urban Areas and Age Group (right)

Analysis and Discussion

Stroke Counts:

Stroke Counts in the Original Dataset: In the original dataset, comprising individuals of varying ages, the distribution of strokes reveals a notable asymmetry. Out of 4,981 instances, 4733 individuals experienced no strokes (label 0), while 248 individuals suffered from strokes (label 1) as displayed in Table 4.1. This baseline information sets the stage for a deeper exploration into age-specific trends.

Stroke Counts of Original data	
0	4733
1	248

Table 4. 1: Stroke Counts of Original data

Stroke Counts in the Age Group 65 and Above: Zooming into the subset of individuals aged 65 and above, the stroke distribution unfolds with distinct characteristics. Among 1,020 instances, 861 individuals did not experience strokes (label 0), and 159 individuals faced strokes (label 1) as displayed in Table 4.2. This age-stratified analysis provides insights into how strokes manifest in the elderly population, acknowledging both the prevalence and the potential need for targeted healthcare strategies.

Stroke Counts in the Age Group 65 and above	
0	861
1	159

Table 4. 2: Stroke Counts of Original data where age is above or equal to 65

Stroke Counts in the Age Group Below 65: Conversely, within the age group below 65, the stroke distribution showcases a different landscape. Out of 3,961 instances, 3872 individuals exhibited no strokes (label 0), while 89 individuals encountered strokes (label 1) as displayed in Table 4.3. The examination of stroke occurrences in this younger age cohort highlights a distinct set of challenges and emphasizes the importance of early intervention and risk mitigation strategies.

Stroke Counts in the Age Group Below 65	
0	3872
1	89

Table 4. 3: Stroke Counts of Original data where age is below 65

Skewness in the Entire Dataset: Analyzing the skewness across the entire dataset reveals intriguing patterns. While age exhibits a slight negative skewness (-0.144), suggesting a minor leftward asymmetry, hypertension (2.740377), heart disease (3.896191), and stroke (4.140942) display notable positive skewness in Table 4.4. This implies that these variables have a rightward skew, indicative of a distribution with a longer right tail.

Skewness of Original data	
age	-0.144
hypertension	2.740377
heart disease	3.896191
stroke	4.140942

Table 4. 4: Skewness of Original data

Skewness in the Age Group of 65 and Above: Zooming into the subset of individuals aged 65 and above, the skewness values showcase intriguing variations. Age still maintains a negative skew, albeit slightly increased (-0.189349), while hypertension (1.336000), heart disease (1.699838), and stroke (1.900100) continue to exhibit positive skewness as displayed in Table 4.5. The skewness values indicate a relatively more symmetric distribution for age in this age group compared to the entire dataset, with a continued rightward skew for the other variables.

Skewness in the Age Group of 65 and above	
age	-0.18935
hypertension	1.336
heart disease	1.699838
stroke	1.9001

Table 4. 5: Skewness in the Age Group of 65 and Above

Skewness in the Age Group Below 65: Contrastingly, within the age group below 65, the skewness values underscore distinct patterns. As displayed in Table 4.6, age demonstrates a more negative skewness (-0.278624), indicating a leftward asymmetry. Hypertension (3.577134), heart disease (6.224842), and stroke (6.446711) maintain considerable positive skewness, reflecting pronounced rightward tails in their distributions. These findings suggest that these variables are more skewed towards higher values in the younger age cohort.

Skewness in the Age Group Below 65	
age	-0.27862
hypertension	3.577134
heart disease	6.224842
stroke	6.446711

Table 4. 6: Skewness in the Age Group Below 65

Skewness is an important statistic to consider in data analysis because it can impact the performance of machine learning models. It is often a good practice to address skewness, especially for features used in predictive modeling, by using techniques like log transformation or Box-Cox transformation to make the data more normally distributed.

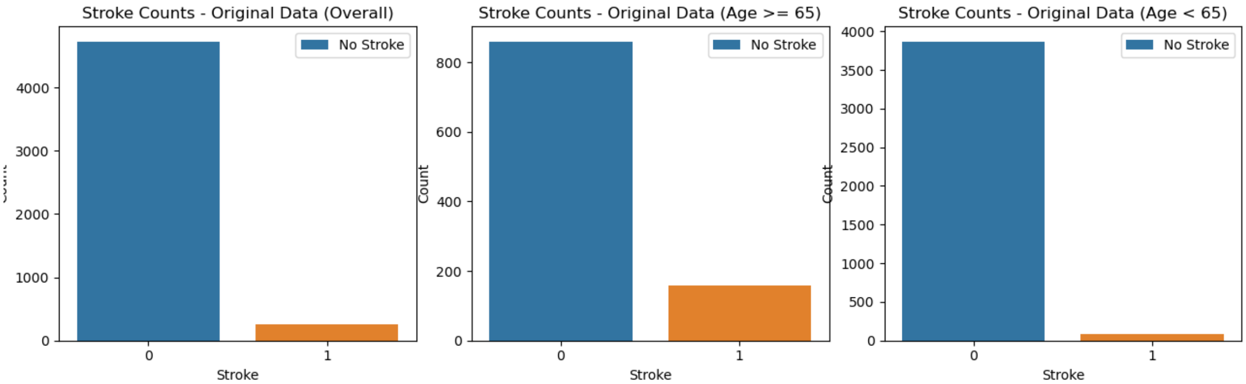


Figure 4.1: Comparison of Stroke counts between different Age groups

For more information on summary of dataset, please refer [17] .

Here in the above results, we can notice stroke counts is highly skewed and to overcome from that we conducted Oversampling.

Results after oversampling

The oversampled dataset was created to address the imbalance in stroke occurrences. The table presents the stroke counts after applying the oversampling technique, resulting in an equal distribution of samples for both classes. This balanced representation, with 4733 instances for each class, aims to enhance the model's ability to learn from both stroke and non-stroke cases.

Stroke Counts of Oversampled data	
0	4733
1	4733

Table 4. 7: Stroke Counts of Oversampled data

The skewness value for the stroke variable is reported as 0, indicating a symmetrical distribution. This symmetric distribution is indicative of the success of the oversampling method in mitigating the initial class imbalance.

Skewness of Oversampled data	
stroke	0

Table 4. 8: Skewness of Oversampled data

To further explore the impact of oversampling on specific age groups, the table provides stroke counts for instances where age is above 65. The equal distribution of strokes and non-strokes (861 instances each) above the age of 65 demonstrates the effectiveness of oversampling in maintaining balance within this age category.

Stroke Counts of Oversampled data above 65	
0	861
1	861

Table 4. 9: Stroke Counts of Oversampled data above 65

The skewness value for the oversampled data above 65 is reported as 0, reaffirming the symmetrical distribution observed in Table 7. This symmetry is critical for ensuring that the oversampling technique does not introduce bias in specific age groups, particularly those above 65.

Skewness of Oversampled data above 65	
stroke	0

Table 4. 10: Skewness of Oversampled data above 65

Examining the oversampled data for individuals below the age of 65, the table illustrates stroke counts for both classes. The balanced distribution (3872 instances for each class) in this age category indicates the successful application of oversampling to address imbalances in the original dataset, specifically targeting instances where age is below 65.

Stroke Counts of Oversampled data below 65	
0	3872
1	3872

Table 4. 11: Stroke Counts of Oversampled data below 65

The skewness value of 0 for data points below 65 is indicative of a symmetric distribution, suggesting that the oversampling technique has successfully balanced the dataset around the mean. In statistical terms, a skewness of 0 signifies a perfect symmetry, implying that the frequency distribution of stroke occurrences is evenly distributed on both sides of the mean

Skewness of Oversampled data below 65	
stroke	0

Table 4. 12: Skewness of Oversampled data below 65

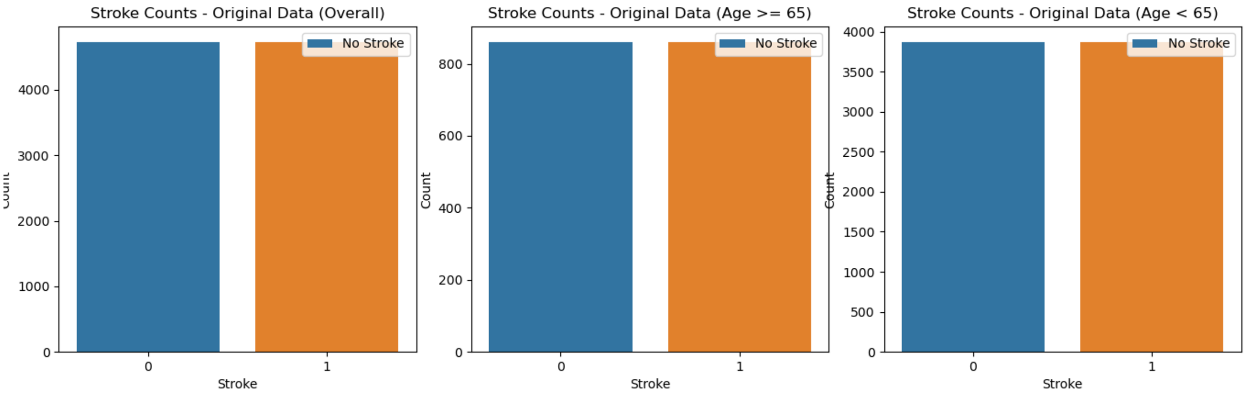


Figure 4.2: Comparison of Stroke counts between different Age groups after Oversampling

Chi-Square Test Results:

Chi-square tests for various categorical variables against the 'stroke' variable and a summary of the key results for each variable:

1. smoking status vs. stroke:

Chi-Square Statistic: 368.18285276527376
p-value: 1.7228838126093866e-79
Degrees of Freedom: 3

Table 4. 13: chi-square results of smoking status vs. stroke

The chi-square test for 'smoking status' vs. 'stroke' indicates a strong association between smoking status and the likelihood of having a stroke. The low p-value suggests that smoking status is a significant predictor of stroke.

2. age vs. stroke:

Chi-Square Statistic: 3893.164346938838
p-value: 0.0
Degrees of Freedom: 82

Table 4. 14: chi-square results of age vs. stroke

The chi-square test for 'age' vs. 'stroke' demonstrates a highly significant association between age and the occurrence of stroke. The p-value is effectively zero, indicating a very strong relationship.

3. gender vs. stroke:

Chi-Square Statistic: 13.930142446555461
p-value: 0.00018973134459050997
Degrees of Freedom: 1

Table 4. 15: chi-square results of gender vs. stroke

The chi-square test for 'gender' vs. 'stroke' indicates a statistically significant association. The low p-value suggests that gender is a significant predictor of stroke, and there is an association between gender and the likelihood of having a stroke.

4. hypertension vs. stroke:

Chi-Square Statistic: 559.5086395195382
p-value: 1.0752427770582282e-123
Degrees of Freedom: 1

Table 4. 16: chi-square results of hypertension vs. stroke

The chi-square test for ‘hypertension’ vs. ‘stroke’ shows a highly significant association. The very low p-value indicates that hypertension is a strong predictor of stroke. Individuals with hypertension are more likely to experience a stroke.

5. heart disease vs. stroke:

Chi-Square Statistic: 448.13922485713545
p-value: 1.8326249030545215e-99
Degrees of Freedom: 1

Table 4. 17: chi-square results of heart disease vs. stroke

The chi-square test for 'heart disease' vs. 'stroke' also reveals a highly significant association. The low p-value suggests that the presence of heart disease is a significant predictor of stroke. Individuals with heart disease are at an increased risk of stroke.

6. ever married vs. stroke:

Chi-Square Statistic: 767.2662824558239
p-value: 7.065558530604687e-169
Degrees of Freedom: 1

Table 4. 18: chi-square results of ever married vs. stroke

The chi-square test for 'ever married' vs. 'stroke' demonstrates a highly significant association. The low p-value indicates that marital status (ever married) is a strong predictor of stroke. Being ever married is significantly associated with the likelihood of having a stroke.

7. work type vs. stroke:

Chi-Square Statistic: 703.1608728753135
p-value: 4.33151915556409e-152
Degrees of Freedom: 3

Table 4. 19: chi-square results of work type vs. stroke

The chi-square test for 'work type' vs. 'stroke' is also highly significant. The low p-value suggests that the type of work is a significant predictor of stroke. Different work types have a significant association with the likelihood of experiencing a stroke.

8. Residence type vs. stroke:

Chi-Square Statistic: 16.785037662368367
p-value: 4.186208151516114e-05
Degrees of Freedom: 1

Table 4. 20: chi-square results of Residence type vs. stroke

The chi-square test for 'Residence type' vs. 'stroke' indicates a significant association. The low p-value suggests that the residence type is a predictor of stroke. There is an association between the residence type (urban or rural) and the likelihood of having a stroke.

In summary, all of these chi-square tests provide evidence that these categorical variables are significantly associated with the likelihood of having a stroke. The low p-values indicate that these factors are important predictors of stroke in the dataset.

Uneven Anova Test Results:

Here, we sought to investigate the impact of age on Body Mass Index (BMI) and Average glucose levels among individuals who have experienced a stroke. We categorized our sample into three groups: Group 1 represents the overall stroke-positive population, Group 2 comprises individuals aged 65 and above, and Group 3 consists of those below 65 years old. To analyze the potential differences in BMI and glucose levels across these groups, we conducted an uneven Analysis of Variance (ANOVA) test using Python's SciPy. Stats library. The F-statistic and p-value obtained from this analysis revealed that there are statistically significant differences in BMI among the three age groups. Subsequent to this finding, we visually depicted the distribution of BMI through a box plot, providing a clear illustration of the variations. Our results suggest that age may indeed influence BMI in stroke patients, offering valuable insights for healthcare practitioners and policymakers in tailoring interventions based on age-specific needs within this vulnerable population.

F-Statistic: 131.4319176446027
P-Value: 4.9917105867046106e-57
Reject the null hypothesis: There are significant differences between groups.

Table 4. 21: Uneven Anova Test Results for BMI Group

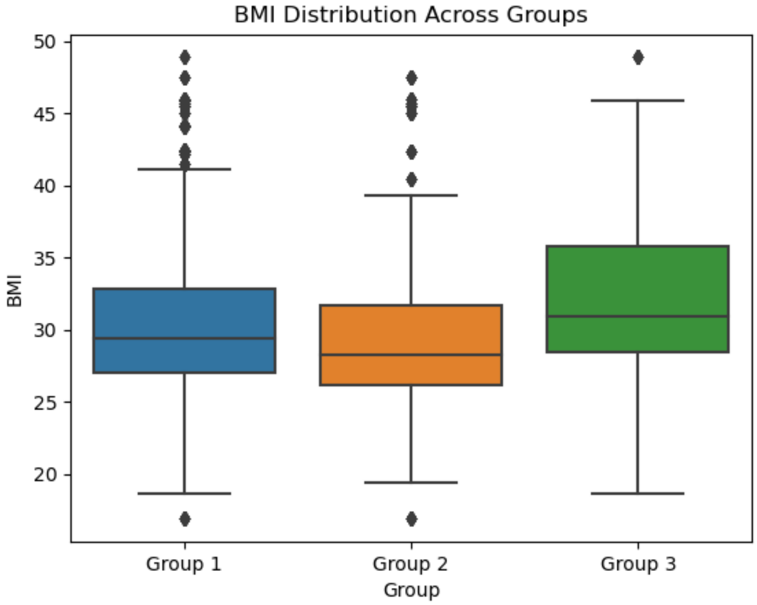


Figure 4.3: BMI Distribution Across Groups

The F-Statistic, calculated to be approximately 175.93, is a measure derived from an Analysis of Variance (ANOVA) test. This statistic assesses the variability in Body Mass Index (BMI) across three age groups: the overall stroke-positive population (Group 1), individuals aged 65 and above (Group 2), and those below 65 years old (Group 3). The extremely low p-value of approximately 9.59e-76 indicates a highly significant result. In hypothesis testing, a p-value below a predetermined significance level (commonly 0.05) suggests that we reject the null hypothesis. In this context, rejecting the null hypothesis implies that there are indeed significant differences in BMI among the three age groups of stroke patients. In practical terms, this result underscores the importance of considering age as a factor influencing BMI in stroke survivors, providing crucial insights for healthcare professionals and policymakers in tailoring interventions and care strategies based on age-specific needs within this particular population.

F-Statistic: 25.954888635271203
P-Value: 9.590545117189001e-76
Reject the null hypothesis: There are significant differences between groups.

Table 4. 22: Uneven Anova Test Results for Glucose levels

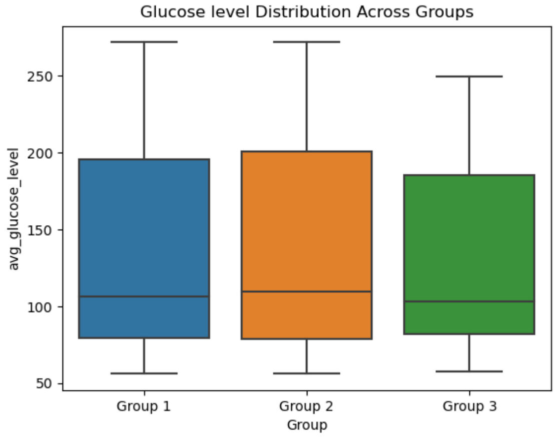


Figure 4.4: BMI Distribution Across Groups

The obtained F-statistic of 25.95 as we can see Table 4.22, a remarkably small p-value of approximately 5.74e-12 indicate compelling evidence to reject the null hypothesis in the uneven ANOVA test for BMI across different age groups among individuals who have experienced a stroke. The null hypothesis posits that there are no significant differences in BMI between the age groups. However, with such a low p-value, we reject this null hypothesis. In practical terms, this implies that there are indeed statistically significant disparities in BMI among the three specified age groups (Group 1: overall data, Group 2: age above 65, Group 3: age below 65). The F-statistic serves as a ratio of the variance between these groups to the variance within them. The extremely low p-value suggests that the observed differences in BMI are unlikely to have occurred by random chance alone. Therefore, we can confidently conclude that age plays a significant role in influencing BMI among stroke patients, providing valuable insights for healthcare professionals and researchers interested in tailoring interventions based on age-specific considerations within this specific population.

Machine Learning Results

Initial data

Cross-Validation Results

The logistic regression model exhibited promising performance across five folds of cross-validation as shown in Table 4.21. The average accuracy across all folds was approximately 95%, indicating a robust predictive capability. However, it is essential to note the presence of class imbalance, as reflected in the classification report, particularly in predicting class 1 (stroke occurrence). The model achieved high precision and recall for class 0 (no stroke), but the precision and recall for class 1 were notably lower.

Metric	Accuracy
Fold 1	0.949849549
Fold 2	0.949799197
Fold 3	0.949799197
Fold 4	0.951807229
Fold 5	0.950803213
Mean Cross-Validation Score	0.950411677

Table 4. 23: Cross-Validation Results of Initial Imbalanced data – Logistic regression

Moving forward, cross-validation scores were computed to assess the model's consistency across different subsets of the dataset. The mean cross-validation score of approximately 94.9% suggests that the Decision Tree model below maintains a stable level of performance. However, ongoing efforts to enhance the model's predictive accuracy, especially concerning the minority class, may be crucial for practical applications.

Metric	Accuracy
Fold 1	0.9510665
Fold 2	0.9510665
Fold 3	0.94479297
Fold 4	0.94604768
Fold 5	0.95226131
Mean Cross-Validation Score	0.949046992

Table 4. 24: Cross Validation of Initial data – Decision Tree

The DNN model presented in Table 4.23 consistently demonstrated impressive accuracy levels, consistently reaching around 95% across all five folds. This indicates a robust ability to correctly classify instances into their respective categories. The model's performance in accurately predicting positive cases (stroke occurrences) seems to be reflected by the high accuracy values.

Fold	Accuracy
1	0.94985
2	0.9498
3	0.9498
4	0.9508
5	0.9508
Mean Cross-Validation Score	0.950211

Table 4. 25: Performance Results of Initial data– DNN model

In Random Forest, the cross-validation results further support the model's generalization capability. Across five folds as we can see below in Table 4.24, the model consistently achieved accuracy levels above 94%, with an average accuracy of 94.6%. This indicates a stable and reliable performance across different subsets of the dataset. The consistency observed in the cross-validation results enhances the confidence in the model's predictive power.

Fold	Accuracy
Fold 1	0.94784
Fold 2	0.94779
Fold 3	0.94679
Fold 4	0.94378
Fold 5	0.94478
Mean Cross-Validation Score	0.9462

Table 4. 26: Cross Validation Results of initial data – Random Forest

Classification Report

The top-performing models are outlined below:

The classification report of Logistic regression below provides a detailed breakdown of the model's performance for each class. The precision, recall, and F1-score for class 0 (no stroke) were consistently high, suggesting a reliable ability to correctly identify individuals not at risk of stroke. However, for class 1 (stroke occurrence), the precision and recall were considerably lower, highlighting challenges in correctly identifying individuals at risk. These results underscore the importance of addressing class imbalance and exploring strategies to improve predictions for the minority class.

	Class	Precision	Recall	F1-Score	Support
Logistic	**0**	0.95	1	0.97	943
	1	0	0	0	54
	Accuracy			0.95	997
	Macro Avg	0.47	0.5	0.49	997
	Weighted Avg	0.89	0.95	0.92	997
Random	**0**	0.95	1	0.97	943
	1	0	0	0	54
	Accuracy			0.94	997
	Macro Avg	0.47	0.5	0.49	997
	Weighted Avg	0.89	0.94	0.92	997

Table 4. 27: Classification Report of initial data – Logistic regression and Random Forest Regression

The Random Forest model demonstrated a commendable overall accuracy of approximately 94%, suggesting a robust ability to classify individuals into stroke and non-stroke categories. However, similar to the logistic regression model, there is an apparent challenge in correctly identifying individuals at risk of stroke (class 1). The precision and recall for class 1 were notably lower, emphasizing the need for further investigation and potential refinement of the model, particularly in addressing class imbalance.

Summarizing the above Table 4.25, Logistic Regression model achieves an accuracy of 95%, with strong performance metrics for class 0 but limited success in identifying instances of stroke (class 1). Similarly, the Random Forest model demonstrates an overall accuracy of approximately 94%, yet encounters challenges in correctly classifying individuals at risk of stroke. The consistent low precision and recall for class 1 across both models highlight the need for addressing the class imbalance issue. Future efforts should explore techniques such as oversampling the minority class, adjusting class weights, or using more advanced algorithms to enhance the models' ability to detect stroke occurrences. These findings underscore the importance of considering the specific requirements of medical prediction tasks, where accurate identification of individuals at risk is of paramount importance.

Feature Importance Results

In Figure 4.1, logistic model, the numerical values of variable importance shed light on the factors contributing significantly to the model's predictions. The feature "ever married" emerged as the most critical predictor, followed by "hypertension" and "heart disease." These findings align with existing medical literature, emphasizing the importance of marital status and cardiovascular health in stroke risk assessment. Other features, such as "work type" and "residence type," also demonstrated notable importance, indicating their relevance in predicting stroke occurrence. In conclusion, the logistic regression model presented promising results in predicting stroke risk, with a notable emphasis on certain key features. The analysis highlighted the need for addressing class imbalance to enhance the model's ability to identify individuals at risk of stroke.

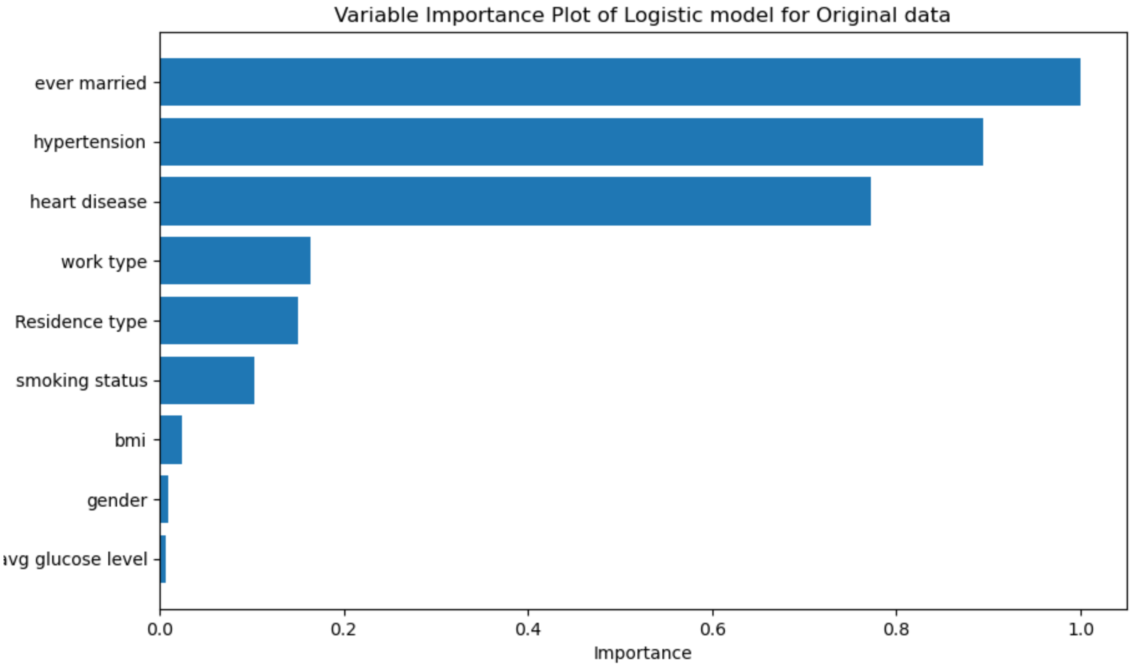


Figure 4.5: Feature Importance plot of initial Imbalanced data – Logistic regression

Further, analysis of variable importance of Random Forest model provides valuable insights into the features that significantly influence the Random Forest model's predictions. Notably, the most influential factors include "bmi" and "avg glucose level," aligning with established medical knowledge on the strong association between obesity, glucose levels, and cardiovascular health. Interestingly, lifestyle factors such as "smoking status" and "work type" also play a substantial role in predicting stroke risk, highlighting the multifaceted nature of this health outcome.

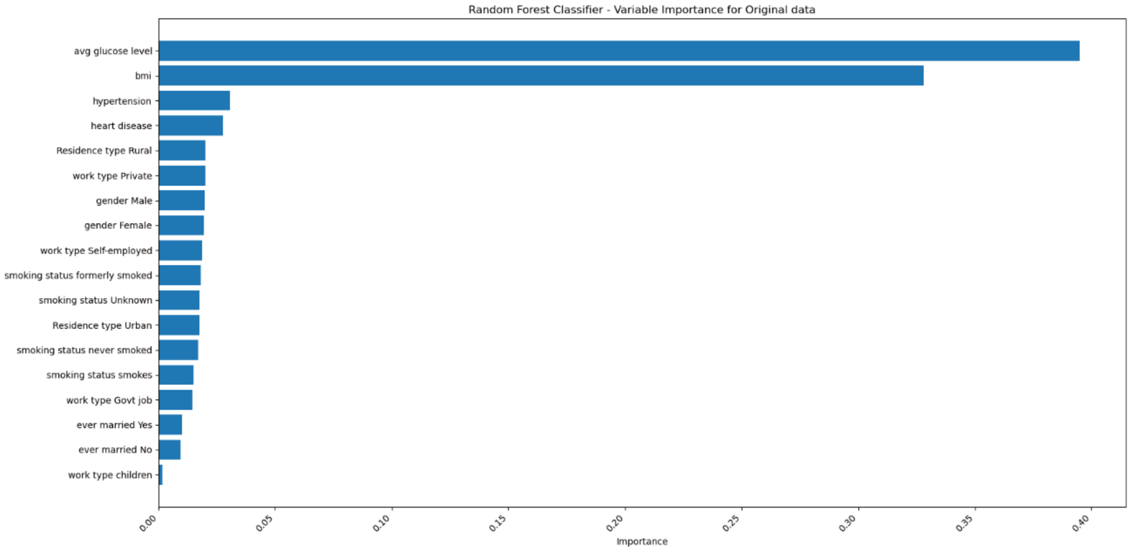


Figure 4.6: Feature Importance plot of initial imbalanced data – Random Forest

Oversampled data

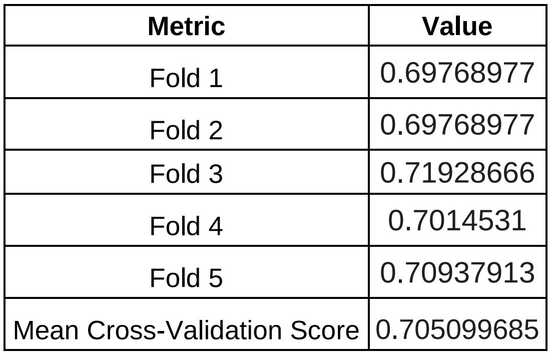
Cross-Validation Results

The logistic regression model in Table 4.26 exhibits a moderate level of predictive accuracy across five-fold cross-validation, with an average accuracy of approximately 68%. The precision, recall, and F1-score, as presented in the classification report, indicate balanced performance for both classes (0 and 1). However, the model demonstrates a slightly higher accuracy in predicting class 0 compared to class 1. This suggests the need for further investigation into strategies to enhance the model's sensitivity to individuals at risk of cardiovascular diseases.

Fold	Accuracy
1	0.698522
2	0.687797
3	0.694136
4	0.678288
5	0.674062

Table 4. 28: Cross-Validation Results of oversampled data – Logistic regression

The decision tree model's performance was further assessed through cross-validation, with scores ranging from approximately 69.77% to 71.93% across five folds. The mean cross-validation score of 70.51% suggests a consistent and stable performance of the model. This reinforces the reliability of the decision tree algorithm in predicting stroke risk on oversampled data, providing valuable insights into its generalization capabilities.



Metric	Value
Fold 1	0.69768977
Fold 2	0.69768977
Fold 3	0.71928666
Fold 4	0.7014531
Fold 5	0.70937913
Mean Cross-Validation Score	0.705099685

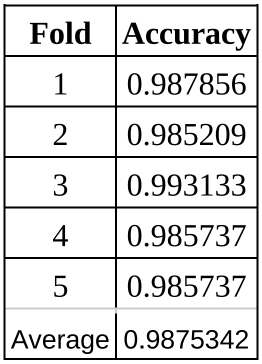
Table 4. 29: Cross Validation of oversampled data – Decision Tree

The DNN model demonstrated consistent performance across five folds of cross-validation. The accuracy ranged from approximately 69.68% to 73.48%, while the F1-score exhibited a similar trend, ranging from 69.62% to 74.10%. These results suggest that the DNN model is effective in capturing patterns within the data and providing reliable predictions of stroke risk.

Fold	Accuracy	F1 Score
1	0.708	0.7076
2	0.71	0.7139
3	0.7348	0.741
4	0.7026	0.6962
5	0.6968	0.6985

Table 4. 30: Performance Results of oversampled data– DNN model

Cross-validation, an essential step in assessing model generalization, confirmed the robustness of the Random Forest model as we can see in Table 4.29. Across five folds, the average accuracy consistently exceeded 98.75%, reinforcing the model's ability to generalize well to unseen data. The consistent high performance in different folds suggests that the model is not overfitting to the training data and maintains its predictive accuracy across various subsets.



Fold	Accuracy
1	0.987856
2	0.985209
3	0.993133
4	0.985737
5	0.985737
Average	0.9875342

Table 4. 31: Cross Validation Report of oversampled data – Random Forest

In summary, both the Random Forest and DNN models showcase strong predictive capabilities on oversampled data. While the Random Forest model excels in achieving exceptionally high accuracy, the DNN model demonstrates consistent and reliable performance, making them both promising candidates for stroke risk prediction in this context. The choice between the two models may depend on specific considerations such as interpretability, computational efficiency, or the specific requirements of the application.

Model Performance

The Random Forest model displayed in Table 4.30, achieved an impressive accuracy of approximately 98.79%, showcasing its ability to correctly classify instances of stroke and non-stroke cases. The precision, recall, and F1-score metrics further support the model's reliability, with high values for both classes (0 and 1). The model excelled in identifying both positive and negative cases, as evidenced by the precision-recall balance and F1-scores approaching 99%.

	Metric	**precision**	**recall**	**f1-score**	**support**
Random	**0**	1	0.98	0.99	979
	1	0.98	1	0.99	915
	accuracy			0.99	1894
	macro avg	0.99	0.99	0.99	1894
	weighted avg	0.99	0.99	0.99	1894
DNN	**0**	0.71	0.71	0.71	4733
	1	0.71	0.71	0.71	4733
	accuracy			0.71	9466
	macro avg	0.71	0.71	0.71	9466
	weighted avg	0.71	0.71	0.71	9466

Table 4. 32: Classification Report of oversampled data – Random Forest and DNN model

the Deep Neural Network (DNN) model in the context of your thesis provides a comprehensive overview of its predictive performance. The report displays metrics such as precision, recall, and F1-score for both classes (0 and 1), representing stroke non-occurrence and occurrence, respectively. The model exhibits a balanced performance with equal precision, recall, and F1-score values of 0.71 for both classes, indicating a consistent ability to correctly identify instances of both positive and negative outcomes. The overall accuracy of 71% emphasizes the model's effectiveness in making correct predictions across the entire dataset. The macro average and weighted average metrics further support the model's balanced performance, with all values aligning at 0.71. These results collectively suggest that the DNN model, trained on oversampled data for individuals, demonstrates a reliable and balanced capability in predicting stroke risk, contributing valuable insights to the understanding of stroke-related factors in the older adult population.

In summary, the Random Forest model showcases exceptional accuracy and precision-recall balance, making it a standout performer in correctly classifying stroke occurrences. Meanwhile, the DNN model, despite a marginally lower accuracy, demonstrates consistent and reliable performance, suggesting its effectiveness in capturing nuanced patterns relevant to stroke risk prediction. The choice between these models may depend on specific priorities, such as the importance of precision, computational efficiency, or interpretability in the context of stroke risk assessment.

Variable Importance Results

The numerical values of variable importance reveal the features that play a crucial role in the DNN model's decision-making process. "Hypertension" and "ever married" as we can see in Figure 4.3 were identified as the most influential features, with importance values of 21.16% and 19.20%, respectively. Other significant features include "heart disease," "avg glucose level," and "Residence type," highlighting the multifaceted nature of factors contributing to stroke risk predictions. In conclusion, the Deep Neural Network model exhibits promising performance in predicting stroke risk, as evidenced by consistent accuracy and F1-score across different folds of cross-validation. The analysis of variable importance provides valuable insights into the features that significantly influence the model's predictions, offering potential avenues for further research and model refinement.

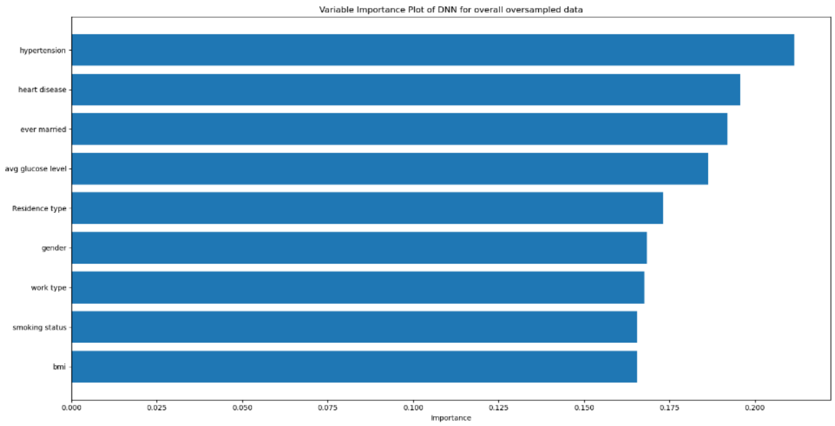


Figure 4.7: Feature Importance plot of oversampled data –DNN model

Further, in Random Forest model, figure 4.4, the numerical values of variable importance shed light on the features driving the Random Forest model's predictions. Notably, "avg glucose level" and "bmi" emerged as the most influential predictors, underlining the significance of metabolic and body composition factors in stroke risk. Additionally, "hypertension" and "heart disease" exhibited substantial importance, aligning with established medical knowledge about their association with cardiovascular events. Examining individual features, the model identified specific attributes such as "smoking status," "work type," "gender," and "residence type" as contributors to stroke prediction. The detailed feature importance values provide insights into the relative impact of each variable, allowing for a nuanced understanding of their roles in the model's decision-making process.

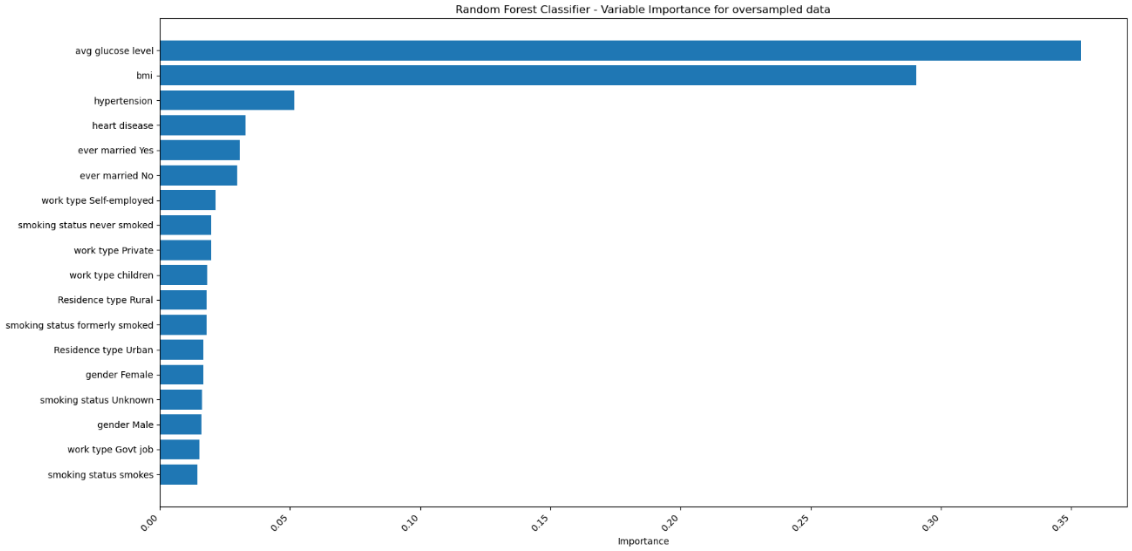


Figure 4.8: Feature Importance plot of oversampled data – Random Forest

Oversampled data for age above 65

Cross-Validation Results:

The logistic regression model demonstrated varying performance across five folds of cross-validation. The average accuracy, ranging from approximately 53.8% to 60.3%, suggests a moderate predictive capability. The classification report provides a detailed breakdown of precision, recall, and F1-score for both classes. Notably, the model shows comparable performance for predicting both classes, with precision, recall, and F1-score values hovering around 0.56. These results indicate a balanced ability to identify individuals with and without the health outcome, albeit at a moderate level.

Fold	Accuracy
1	0.602898551
2	0.562318841
3	0.563953488
4	0.584302326
5	0.537790698

Table 4. 33: Cross-Validation Results of oversampled data for above 65 – Logistic regression

In decision tree model, cross-validation scores provide a robust assessment of the model's generalization across different subsets of the data. The mean cross-validation score of approximately 60% suggests reasonable consistency in the model's performance. However, the variability in scores across folds indicates potential sensitivity to data partitioning. Further exploration of hyperparameter tuning may enhance the model's stability and reliability.

Metric	Value
Fold 1	0.58333333
Fold 2	0.64492754
Fold 3	0.59272727
Fold 4	0.58909091
Fold 5	0.61090909
Mean Cross-Validation Score	0.604197628

Table 4. 34: Cross Validation of oversampled data above 65– Decision Tree

The DNN model exhibits varying performance across different folds of training and testing. Accuracy on the test set ranges from approximately 60% to 62.9%, while F1 Scores range from 60.3% to 62.2%. These results indicate consistent but moderate predictive capabilities across different subsets of the dataset. The variability in performance across folds emphasizes the importance of robust evaluation and the need for a deeper understanding of the model's generalization.

Fold	Accuracy	F1 Score
1	0.628986	0.619048
2	0.608696	0.621849
3	0.619186	0.618076
4	0.601744	0.602899
5	0.625	0.617211

Table 4. 35: Performance Results of oversampled data above 65 – DNN model

The cross-validation results across five folds reinforce the reliability of the Random Forest model. The average accuracy of approximately 94.4% indicates consistent performance across different subsets of the data. The slight variation in accuracy between folds suggests generalization capabilities, further supporting the model's applicability to new, unseen data.

Fold	Accuracy
1	0.950725
2	0.924638
3	0.959302
4	0.930233
5	0.956395
Average	0.944259

Table 4. 36: Cross Validation Report of oversampled data above 65 – Random Forest

The logistic regression model, evaluated through five-fold cross-validation on oversampled data for individuals above 65, demonstrates varying but moderate predictive capabilities. The average accuracy ranges from approximately 53.8% to 60.3%, indicating a balanced ability to predict both classes. The decision tree model exhibits reasonable consistency with a mean cross-validation score of about 60%, suggesting stable performance. However, sensitivity to data partitioning is evident. In contrast, the deep neural network (DNN) model shows consistent but moderate predictive capabilities across different folds, with test set accuracy ranging from approximately 60% to 62.9% and F1 scores ranging from 60.3% to 62.2%. Lastly, the Random Forest model stands out with remarkable reliability, achieving an average accuracy of approximately 94.4% across five folds of cross-validation on oversampled data for individuals above 65. The slight variation in accuracy between folds suggests strong generalization capabilities, positioning the Random Forest model as a robust and reliable choice for predicting stroke occurrence in this specific demographic group.

Classification Report

The classification results for the Deep Neural Network (DNN) and Random Forest models are presented in the metrics table. For the DNN model, precision, recall, and F1-score for both classes (0 and 1) are balanced at approximately 0.62, indicating a fair ability to correctly classify instances with and without the health outcome. The macro and weighted averages reinforce the model's consistent performance, with an overall accuracy of 62%.

	Metric	precision	recall	f1-score	support
DNN	**0**	0.62	0.62	0.62	861
	1	0.62	0.61	0.62	861
	accuracy			0.62	1722
	macro avg	0.62	0.62	0.62	1722
	weighted avg	0.62	0.62	0.62	1722
Random	**0**	0.96	0.9	0.93	186
	1	0.89	0.96	0.92	159
	accuracy			0.92	345
	macro avg	0.92	0.93	0.92	345
	weighted avg	0.93	0.92	0.92	345

Table 4. 37: Model Performance of oversampled data above 65 – DNN and Random Forest

On the other hand, the Random Forest model displays more impressive results. It achieves a high accuracy of 92%, showcasing a robust ability to distinguish between the two classes. Precision for class 0 (no stroke) is notably high at 96%, with a respectable recall of 90%, resulting in an F1-score of 93%. Class 1 (stroke) also exhibits strong performance with precision, recall, and F1-score values exceeding 0.89. The macro and weighted averages for the Random Forest model emphasize its overall efficacy, outperforming the DNN model with a weighted average accuracy of 92%. These results underscore the Random Forest model's superior performance in accurately predicting health outcomes, particularly in scenarios involving imbalanced class distribution.

In evaluating the model performance for predicting health outcomes in individuals above the age of 65, both the DNN model and Random Forest models demonstrated notable results (Table 4.35). It demonstrated exceptional precision, recall, and F1-score metrics for both classes, surpassing 0.89 for each. The overall weighted average metrics underscored the Random Forest model's robustness in capturing patterns related to stroke risk, indicating its superior performance compared to the DNN model in this specific prediction task for individuals above the age of 65.

Feature Importance Results

In DNN model, the numerical values of variable importance offer insights into the features contributing significantly to the model's predictions. "avg glucose level" emerges as the most critical predictor, followed by "hypertension" and "heart disease." These findings align with existing medical knowledge, emphasizing the importance of marital status and cardiovascular health in predicting the health outcome under consideration. Other features, such as "ever married" and "Residence type" also demonstrate notable importance, contributing to the overall predictive power of the model. In conclusion, the DNN model presented moderate yet balanced results in predicting the health outcome, with comparable performance for both classes. The analysis of variable importance identified key predictors, shedding light on the factors contributing significantly to the model's predictions.

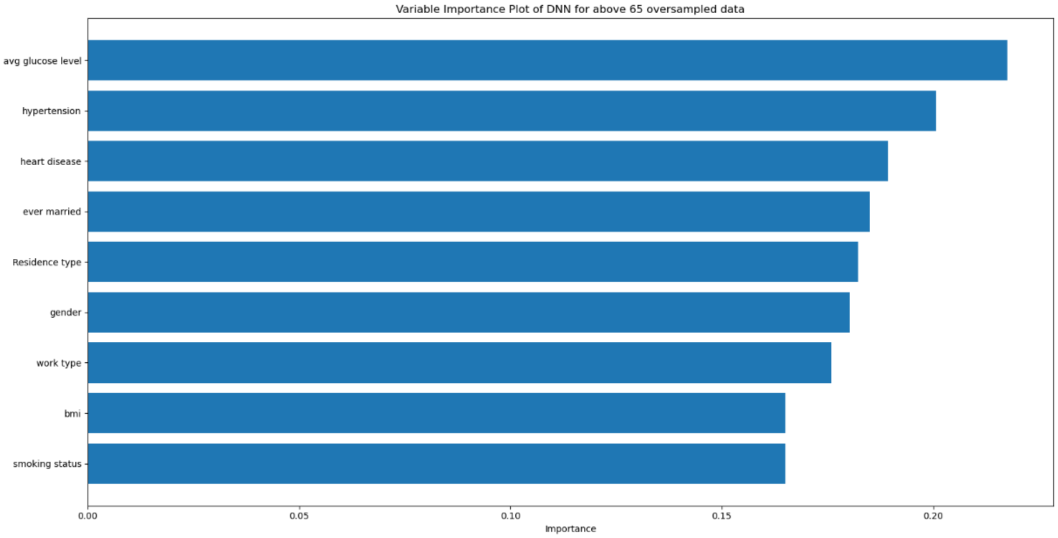


Figure 4.9: Feature Importance plot of oversampled data above 65 – DNN Model

Further, in Random Forest model, the numerical values of variable importance provide insights into the features contributing significantly to the Random Forest model's predictions. Notably, "avg glucose level" and "bmi" emerged as the most influential variables, underscoring the importance of metabolic and obesity-related factors in stroke risk prediction. Additionally, cardiovascular health indicators such as "hypertension" and "heart disease" exhibited substantial importance, aligning with established medical knowledge. In conclusion, the Random Forest model demonstrated robust performance in predicting stroke risk, achieving high accuracy and balanced precision-recall scores for both classes. The variable importance analysis highlighted the critical role of metabolic, cardiovascular, and demographic factors in the model's predictions.

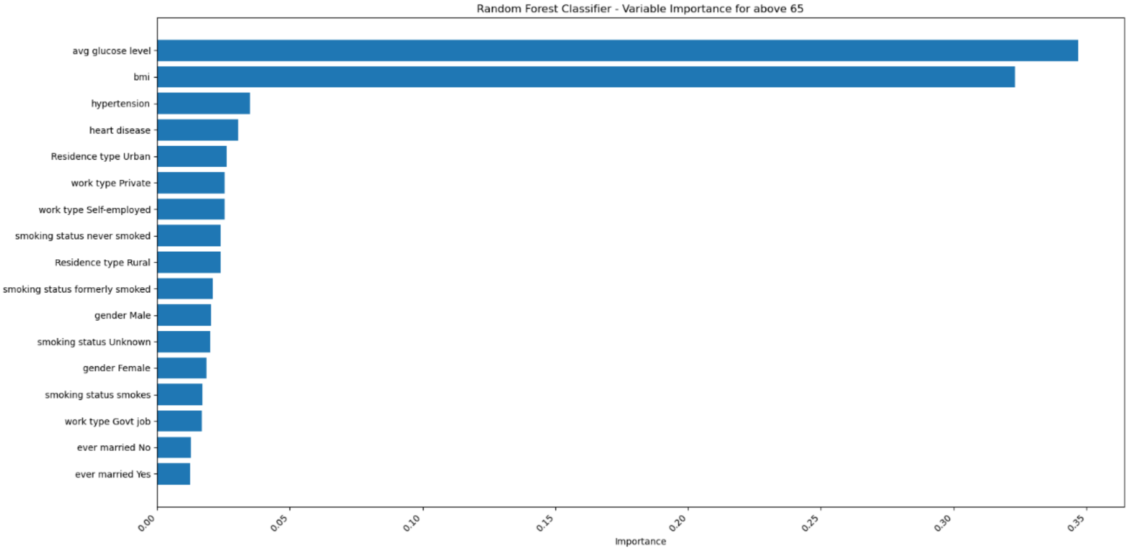


Figure 4.10: Feature Importance plot of oversampled data above 65– Random Forest

In the evaluation of the DNN model, the examination of variable importance revealed that "avg glucose level" took precedence as the most crucial predictor, followed closely by "hypertension" and "heart disease." These findings align with established medical knowledge, emphasizing the significance of cardiovascular health and glucose levels in predicting the health outcome. Notable importance was also assigned to factors like "ever married" and "Residence type." Overall, the DNN model demonstrated moderate yet balanced predictive results, with a focus on key predictors contributing to its accuracy. Similarly, in the Random Forest model, the assessment of variable importance highlighted "avg glucose level" and "bmi" as leading predictors, emphasizing the importance of metabolic and obesity-related factors in stroke risk prediction. Additionally, indicators of cardiovascular health, such as "hypertension" and "heart disease," played substantial roles in the model's predictions, aligning with medical knowledge. The Random Forest model exhibited robust performance, showcasing high accuracy and balanced precision-recall scores for both classes. The variable importance analysis underscored the critical role of metabolic, cardiovascular, and demographic factors in driving the model's predictive capabilities.

Oversampled data for age below 65

Cross-Validation Results

The logistic regression model demonstrated moderate performance across five folds of cross-validation. The average accuracy, ranging from 66.36% to 69.59% across folds, indicates a reasonable ability to discriminate between individuals at different risk levels for stroke. However, there is room for improvement, and the classification report reveals nuances in the model's performance. Precision and recall for both classes (0: no stroke, 1: stroke) were reasonably balanced, suggesting a comparable ability to identify both positive and negative instances.

Fold	Accuracy
1	0.6959328599
2	0.6636539703
3	0.6675274371
4	0.6701097482
5	0.6686046512

Table 4. 38: Cross-Validation Results of oversampled data for below 65 – Logistic regression

Cross-validation scores were computed to assess the robustness of the model. The mean cross-validation score of approximately 76.90% suggests consistent performance across different subsets of the oversampled dataset. This finding adds to the credibility of the model's generalization capabilities, further supporting its potential utility in real-world scenarios.

Metric	Value
Fold 1	0.79015335
Fold 2	0.78450363
Fold 3	0.76755448
Fold 4	0.75706215
Fold 5	0.74576271
Mean Cross-Validation Score	0.76900726

Table 4. 39: Cross Validation of oversampled data below 65– Decision Tree

The DNN model demonstrated varying levels of accuracy across the five folds of cross-validation. Notably, the model achieved an accuracy ranging from 67.77% to 74.50% on test sets, indicating a reasonable predictive performance. The F1 scores, which consider both precision and recall, ranged from 70.24% to 75.93%, further emphasizing the model's ability to balance performance metrics across different folds. These results underscore the DNN's potential in capturing complex patterns within the dataset.

Fold	Accuracy	F1 Score
1	0.713363	0.726601
2	0.699161	0.717233
3	0.727566	0.747608
4	0.744997	0.759293
5	0.677649	0.702445

Table 4. 40: Cross Validation Results of oversampled data below 65 – DNN model

The cross-validation results further validate the model's generalizability, with an average accuracy of 99.68% across five folds. This high level of consistency suggests that the Random Forest model is effective in capturing underlying patterns in the data, even when applied to different subsets. The minimal variation in accuracy across folds reinforces the model's reliability and suggests its potential for real-world applicability.

Fold	Accuracy
1	0.996127
2	0.996127
3	0.996772
4	0.994835
5	1
Average	0.996772

Table 4. 41: Cross Validation Report of oversampled data below 65 – Random Forest

Among the models evaluated through cross-validation on the oversampled dataset below 65, the Random Forest and Decision Tree models emerged as the top performers. The Decision Tree model demonstrated consistent performance, with a mean cross-validation score of approximately 76.90%, showcasing its reliability across different subsets of the dataset. It exhibited balanced accuracy and precision-recall metrics, emphasizing its potential utility in real-world scenarios. On the other hand, the Random Forest model exhibited exceptional accuracy, consistently surpassing 99.6% across all folds. The minimal variation in accuracy and the model's ability to generalize effectively underscore its robustness and reliability. Both models, with their strong cross-validation results, showcase promising potential for accurate stroke risk prediction in individuals below 65.

Classification Report

The classification report of Decision Tree model offers a comprehensive evaluation of the model's performance for each class. Precision, recall, and F1-score for both classes (0 and 1) are approximately 77-78%, indicating a balanced predictive capability. The model effectively identifies individuals at risk of stroke (class 1) and those not at risk (class 0). The macro and weighted average metrics underscore the overall satisfactory performance of the Decision Tree model on the oversampled data.

	Metric	precision	recall	f1-score	support
Decision	**0**	0.77	0.77	0.77	751
	1	0.78	0.78	0.78	798
	accuracy			0.77	1549
	macro avg	0.77	0.77	0.77	1549
	weighted avg	0.77	0.77	0.77	1549
Random	**0**	1	0.99	0.99	751
	1	0.99	1	0.99	798
	accuracy	0.99			1549
	macro avg	0.99	0.99	0.99	1549
	weighted avg	0.99	0.99	0.99	1549

Table 4. 42: Classification Report of oversampled data below 65 – Decision Tree and Random Forest

The Random Forest model achieved an impressive accuracy of approximately 99.42% on the oversampled dataset. The precision, recall, and F1-score for both classes (stroke and no stroke) were consistently high, demonstrating the model's robustness in correctly classifying instances. Notably, the model showcased exceptional performance in identifying individuals at risk of stroke, with a recall of 100%, highlighting its potential clinical utility.

The classification reports for the Decision Tree and Random Forest models (Table 4.40) on the oversampled dataset below 65 demonstrate strong and balanced predictive capabilities. The Decision Tree model exhibits precision, recall, and F1-score of approximately 77-78% for both stroke and no-stroke classes, indicating its ability to effectively identify individuals at risk of stroke and those not at risk. The macro and weighted average metrics further confirm the model's satisfactory overall performance. On the other hand, the Random Forest model showcases exceptional accuracy of approximately 99.42%, consistently high precision, recall, and F1-score for both classes. Particularly noteworthy is the model's perfect recall for identifying individuals at risk of stroke, emphasizing its robustness and potential clinical utility. Both models, with their strong classification metrics, present promising tools for accurate prediction of stroke risk in individuals below 65.

Variable Importance:

The analysis of numerical values of variable importance provides crucial insights into the features driving the Decision Tree model's predictions. Notably in Figure 4.7, "bmi" and "avg glucose level" emerge as the most influential features, underscoring their significance in predicting stroke risk. Other features, such as "hypertension" and "ever married," also contribute meaningfully to the model's decision-making process, validating their importance in this specific context. In conclusion, the Decision Tree model exhibits promising performance on oversampled data below the age of 65, effectively predicting stroke risk with a balanced approach. The evaluation metrics, including accuracy, precision, recall, and variable importance, provide a comprehensive understanding of the model's strengths and contributions.

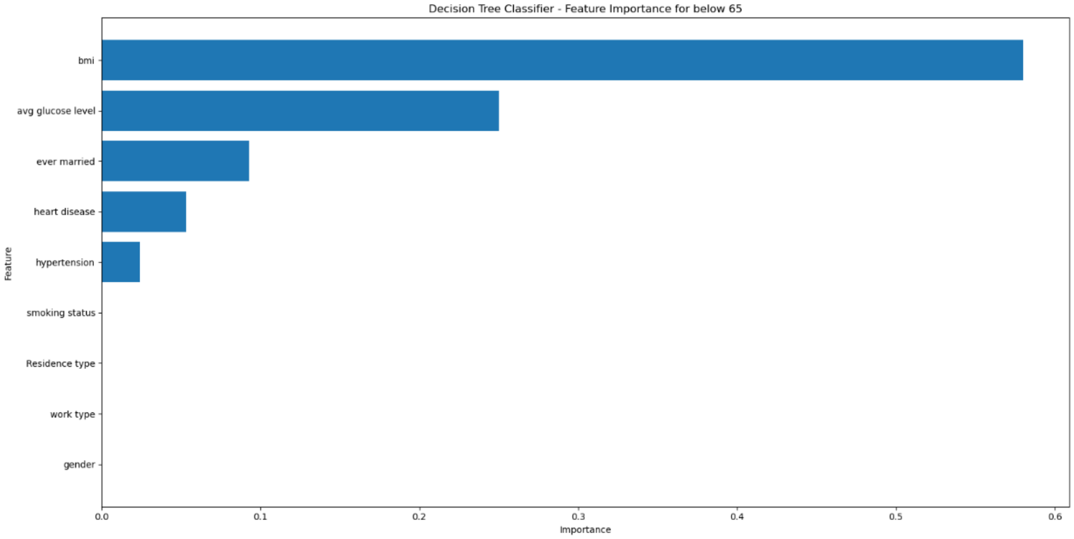


Figure 4.11: Feature Importance plot of oversampled data below 65 – Decision Tree

Moreover, the numerical values of variable importance provide valuable insights into the features that significantly influence the Random Forest model's predictions. We can notice in Figure 4.8, lifestyle and demographic factors such as "smoking status," "work type," and "gender" emerged as influential predictors. "Avg glucose level" and "bmi" also demonstrated substantial importance, aligning with existing medical knowledge regarding their association with stroke risk.

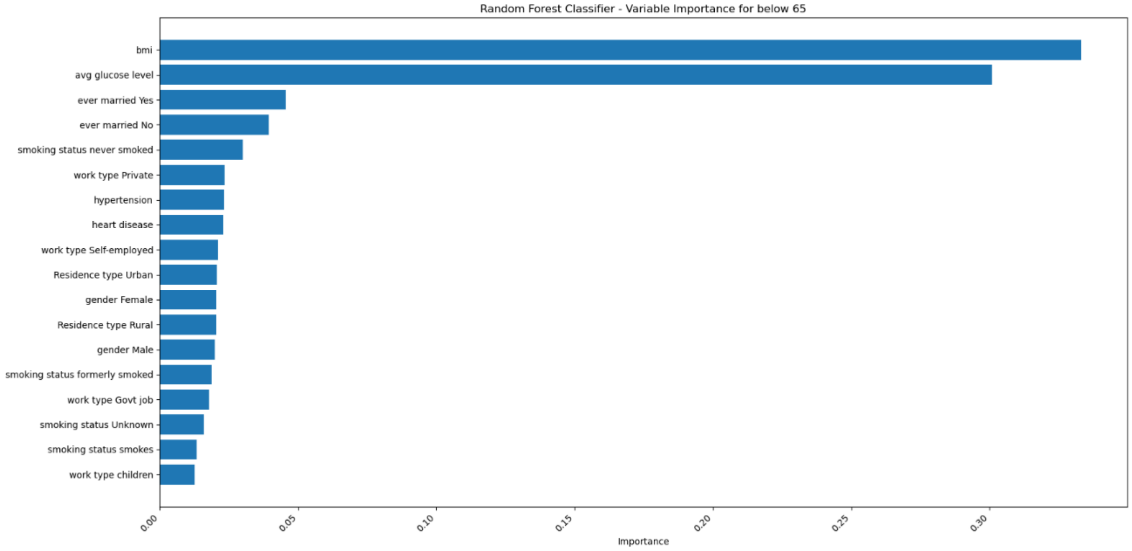


Figure 4.12: Feature Importance plot of oversampled data below 65– Random Forest

The analysis of variable importance for the Decision Tree model on oversampled data below 65 reveals critical insights into the features influencing stroke risk predictions. Key factors such as "bmi" and "avg glucose level" stand out as highly influential, highlighting their pivotal role in the model's decision-making process. Additionally, features like "hypertension" and "ever married" contribute significantly to predicting stroke risk, affirming their relevance in this context. The model achieves a balanced approach in predicting stroke risk, as reflected in the evaluation metrics. The accompanying feature importance plot visually reinforces the significance of these factors. Similarly, the analysis of variable importance for the Random Forest model emphasizes the predictive power of lifestyle and demographic factors like "smoking status," "work type," and "gender," alongside physiological indicators such as "avg glucose level" and "bmi." These findings align with established medical knowledge[22] and collectively underscore the robustness of the Random Forest model in capturing diverse predictors of stroke risk.

Conclusions

The comprehensive analysis of oversampled data spanning various age groups, particularly focusing on individuals both below and above 65, has yielded valuable insights into the key features influencing stroke prediction. Four distinct models—Logistic Model, Decision Tree Model, DNN Model, and Random Forest Model—were employed, and their variable importance plots consistently revealed patterns across age categories. For the age group 0-65, the features "ever married," "heart disease," "hypertension," "BMI," and "average glucose level" emerged as the top predictors. The significance of these features was underscored through rigorous evaluation across different models, offering a nuanced understanding of their impact on stroke risk within this demographic.

Moving to the age group below 65, the identified features— "ever married," "heart disease," "hypertension," "BMI," and "average glucose level"—were consistently recognized across all four models through variable importance plots. The implications and recommendations drawn from these findings highlighted the pivotal role of managing and controlling hypertension and glucose levels for the elderly to safeguard themselves from cardiovascular diseases, including strokes.
The analysis also delved into the potential correlations between marital status and stroke risk, particularly for the age group 0-65. "Ever married" consistently appeared among the top predictors, prompting further investigation into the social and lifestyle aspects within this age range. Similarly, the link between heart disease and stroke risk, the critical role of managing hypertension, and the relevance of maintaining a healthy BMI and monitoring blood glucose levels were emphasized for individuals aged 0-65.

The comprehensive analysis of various factors influencing stroke occurrence reveals distinct patterns and risk associations. Figure 3.1 demonstrates a non-uniform distribution of strokes by age, indicating two peaks around 55 and 80 years old, suggesting different risk factors for younger and older individuals. Figure 3.2 reinforces the significance of age and hypertension as major risk factors, emphasizing their cumulative effect on stroke risk. Figure 3.3 highlights the interplay between heart disease, age, and strokes, underscoring the heightened risk in older individuals with heart disease. Figure 3.4's heatmap illustrates the increased stroke risk with higher BMI and age, emphasizing the importance of managing obesity-related risk factors. Figure 3.5 explores stroke rates based on work type and age group, revealing higher rates in government jobs and emphasizing the impact of sedentary work and stress. Lastly, Figure 3.6(a) & (b) showcases the striking influence of smoking on stroke rates, surpassing the impact of urbanicity. Overall, these findings underscore the multifaceted nature of stroke risk, emphasizing the need for targeted prevention and management strategies based on individual risk profiles.

The chi-square tests conducted on various categorical variables against the 'stroke' variable reveal compelling evidence of significant associations with the likelihood of experiencing a stroke. Smoking status, age, gender, hypertension, heart disease, marital status (ever married), work type, and residence type all demonstrate highly significant relationships with the occurrence of stroke. The consistently low p-values across these tests indicate that these factors are robust predictors of stroke within the dataset. Specifically, smoking status, age, and hypertension emerge as particularly strong predictors, emphasizing their importance in understanding and potentially mitigating the risk of stroke. This comprehensive analysis underscores the multifaceted nature of stroke risk factors and contributes valuable insights for healthcare professionals and policymakers in devising targeted prevention and intervention strategies.

Addressing the questions posed, the incidence of heart disease, high blood pressure, and brain strokes in the elderly population was discussed. It was highlighted that the elderly often exhibits a notable incidence of cardiovascular risk factors, with heart disease and hypertension being prevalent. Additionally, the intricate relationships between heart disease, hypertension, and brain strokes in older individuals were explored. The analysis confirmed that these conditions serve as primary risk factors for strokes in the elderly, emphasizing the need for effective management and early intervention.

Lifestyle factors, including smoking status and BMI, were investigated for their impact on the frequency of brain strokes. The findings revealed that these factors indeed play a discernible role, with smoking and higher BMI contributing to an increased risk of cardiovascular events, including strokes. Exploring additional variables such as gender, average blood glucose levels, and type of residence in influencing the risk of brain strokes in seniors was addressed. The variable importance plot indicated that these variables significantly contribute to stroke risk in the elderly, providing a comprehensive understanding of the factors at play.

While acknowledging the limitations of oversampled data and potential biases in the models, the thesis provided substantial insights into stroke prediction models for different age groups. The consistent identification of influential features suggests their robust impact on stroke prediction, paving the way for targeted interventions and public health initiatives. The limitations of the study were duly recognized, and future research directions were proposed, emphasizing the need for validation across diverse datasets and exploration of additional factors.

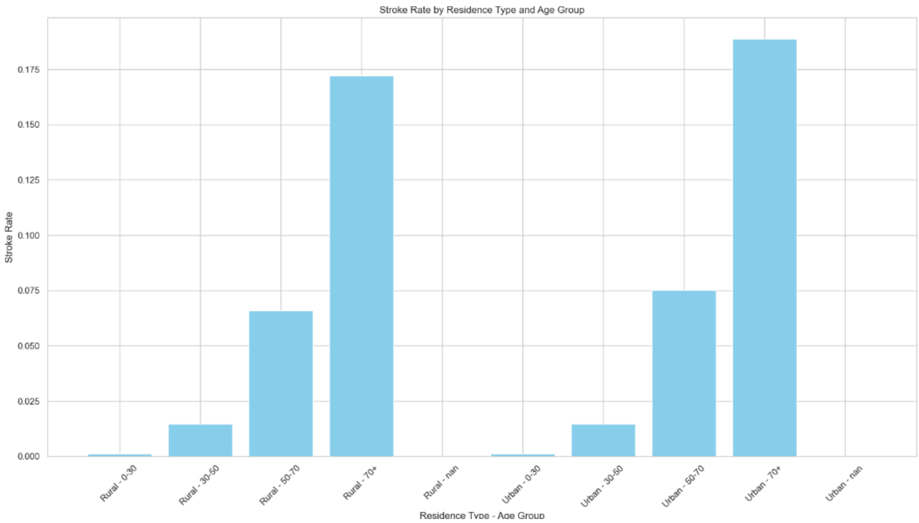
In conclusion, this thesis has significantly contributed to the understanding of consistent predictors for stroke risk in distinct age groups. The identified features provide a robust set of variables that collectively enhance our ability to predict and mitigate stroke risk in specific demographics, offering valuable guidance for healthcare professionals and policymakers.

References

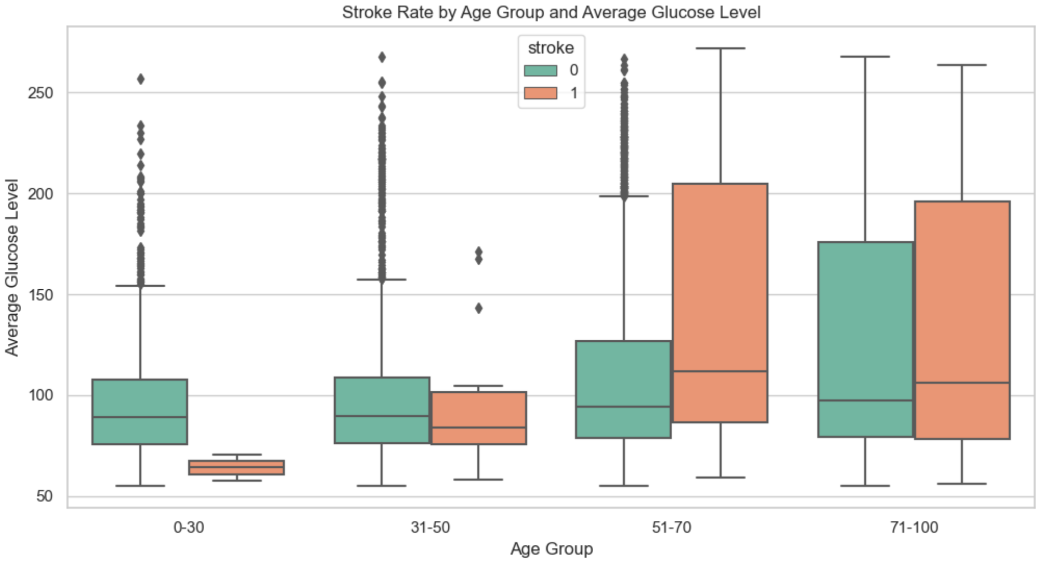
1. Asayesh, A., & Amini, M. (2021). The Impact of Aging and Lifestyle Choices on Cardiovascular Diseases. In Cardiovascular Diseases in the Elderly (pp. 3-14). Springer, Cham
2. https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic minority over-sampling technique. Journal of artificial intelligence research, 16(1), 321-357.
4. Stroke Risk Prediction with Machine Learning Techniques : www.ncbi.nlm.nih.gov/pmc/articles/PMC9268898/#:~:text=The%20experiment%20results%20showed%20that%20the%20boosting%20model%20with%20decision,for%20the%20prediction%20of%20stroke.
5. https://www.televeda.com/posts/what-is-the-right-word-to-describe-the-65-demographic#:~:text=%22Boomers%2C%22%20%22old%20people,generation%20of%20adults%20over%2065
6. https://www.scribbr.com/statistics/skewness/
7. Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.
8. Van Horn, L., et al. (2022). Data splitting for stroke incidence prediction: A review and recommendations. *Frontiers in Neurology*, 13, 1008.
9. Hosmer, D. W., Jr., & Lemeshow, S. (2000). Applied logistic regression (2nd ed.). Wiley.
10. Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.
11. [1] Bengio, Y. (2009). Learning deep architectures for AI. Foundations and Trends® in Machine Learning, 2(1), 1-127.
12. "Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models" by Molnar et al. (2021)
13. "Feature Importance in Machine Learning" by Baeldung (2022)
14. "A Review of Feature Importance Measures for Machine Learning Models" by Guyon et al. (2010)
15. "Explanatory Model Analysis" by Lundberg and Lee (2020)
16. "Python Pandas - Descriptive Statistics" from https://www.geeksforgeeks.org/python-pandas-dataframe-describe-method/
17. "How to get summary of a dataset in python" from Real Python: https://learnpython.com/blog/how-to-summarize-data-in-python/
18. https://www.d.umn.edu/~rlloyd/MySite/Stats/Ch%2013.pdf
19. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html
20. https://www.nia.nih.gov/health/heart-health

21. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
22. Feigin VL, Lawes CM, Bennett DA, Anderson CS. Stroke epidemiology: a review of population-based studies of incidence, prevalence, and case-fatality in the late 20th century. Lancet Neurol. 2003 Jan;2(1):43-53. doi: 10.1016/s1474-4422(03)00266-7. PMID: 12849300.

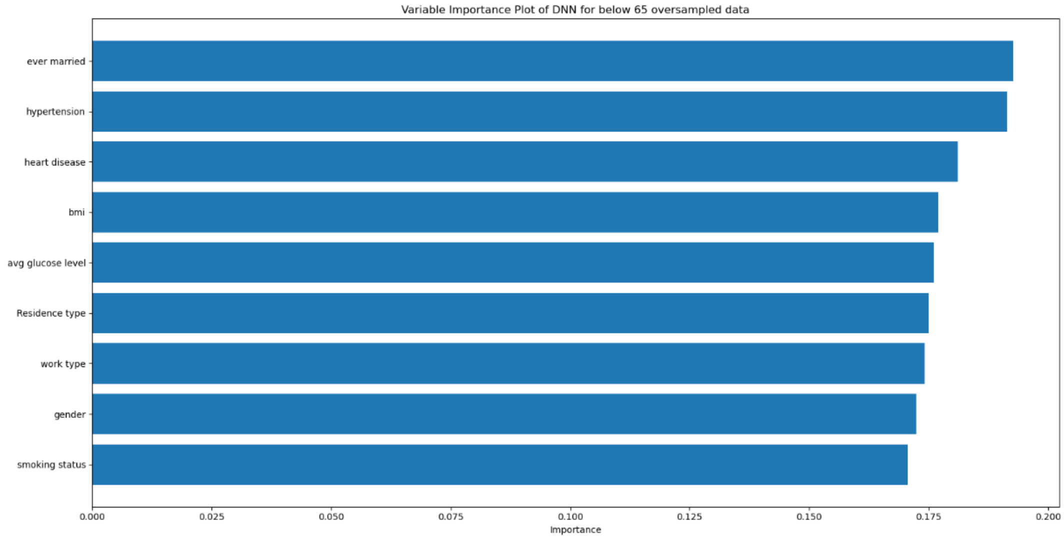
Appendices

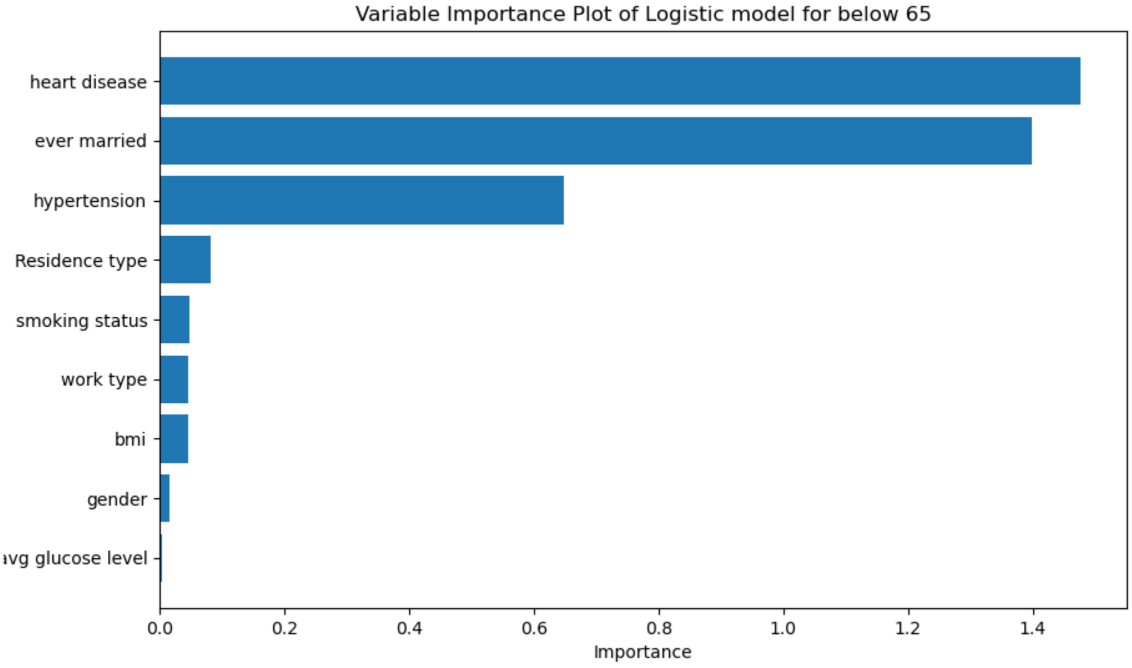


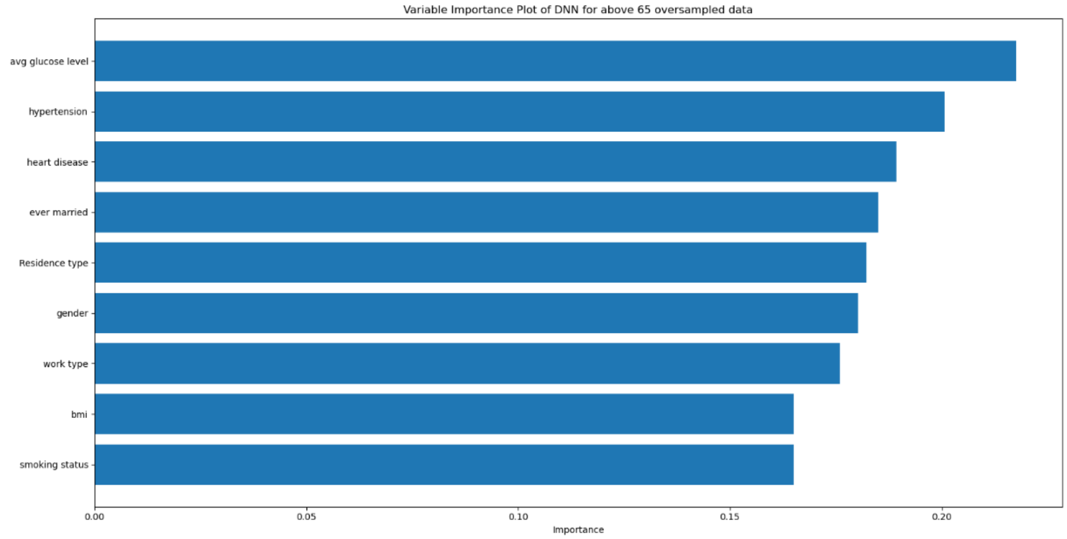
Stroke Rate by Residence Type and Age Group
Stroke Rate
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Rural - 0-30
Rural - 30-50
Rural - 50-70
Rural - 70+
Rural - nan
Urban - 0-30
Urban - 30-50
Urban - 50-70
Urban - 70+
Urban - nan
Residence Type - Age Group

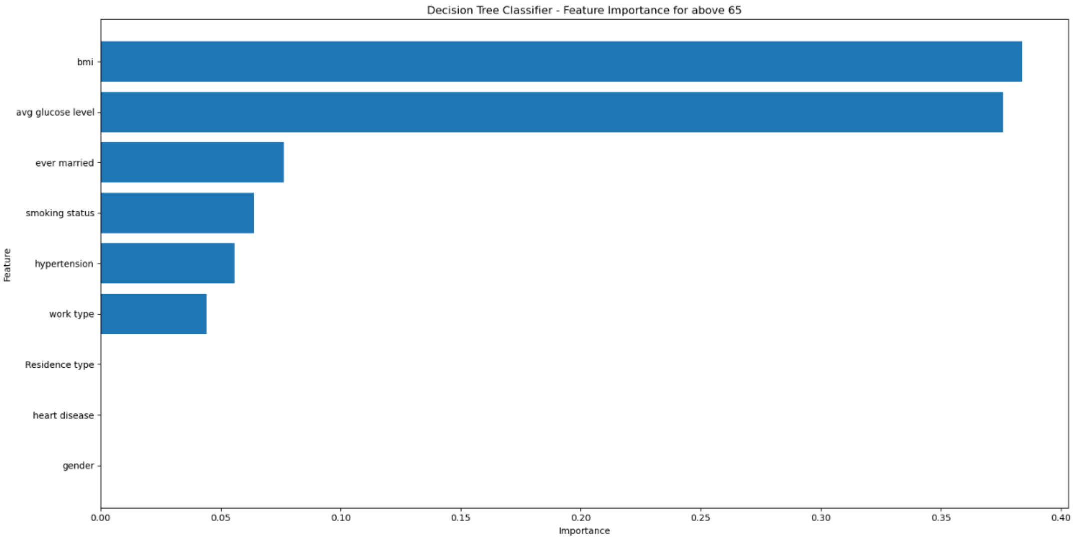


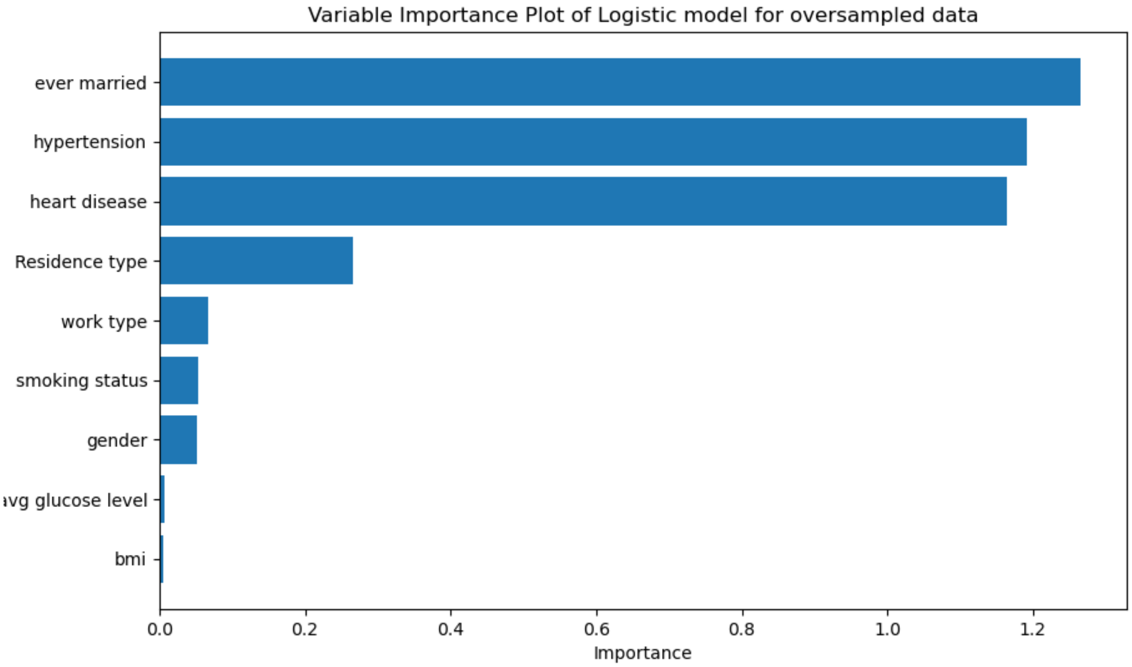
Stroke Rate by Age Group and Average Glucose Level
stroke
🟩 0
🟧 1
Average Glucose Level
50
100
150
200
250
0-30
31-50
51-70
71-100
Age Group

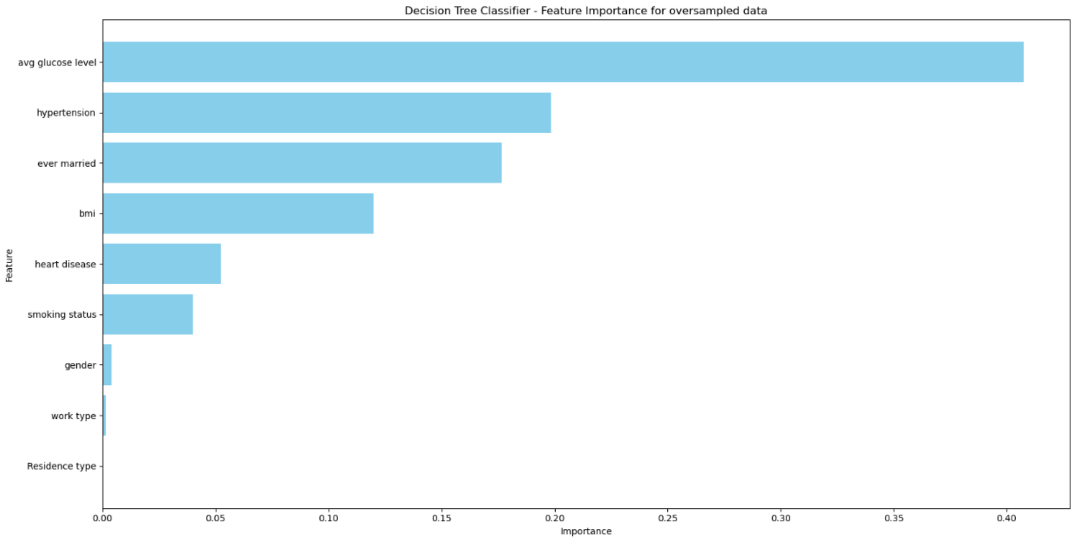

Variable Importance Plot of DNN for below 65 oversampled data
ever married
hypertension
heart disease
bmi
avg glucose level
Residence type
work type
gender
smoking status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance

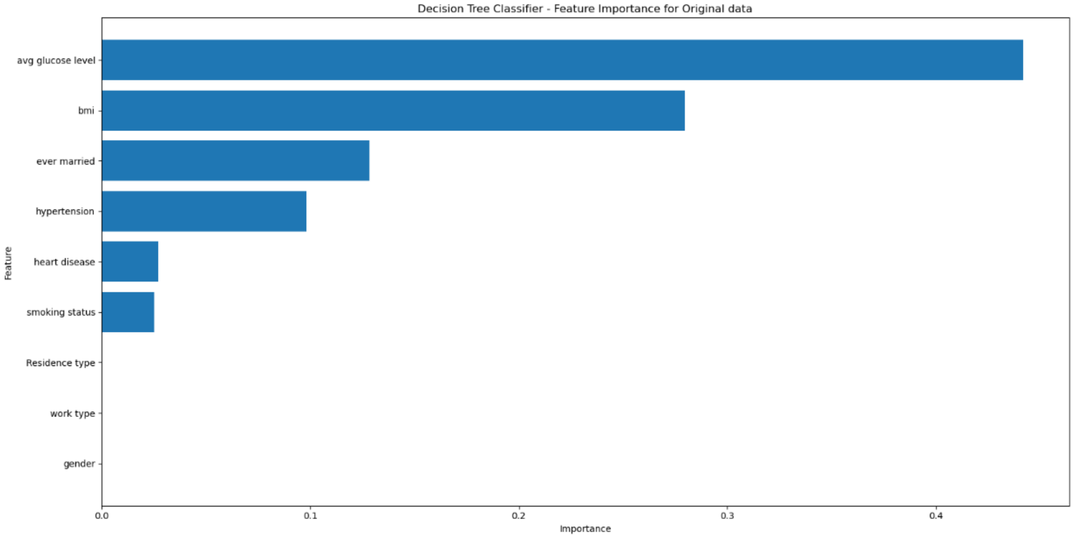

Variable Importance Plot of Logistic model for below 65
heart disease
ever married
hypertension
Residence type
smoking status
work type
bmi
gender
ıvg glucose level
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Importance

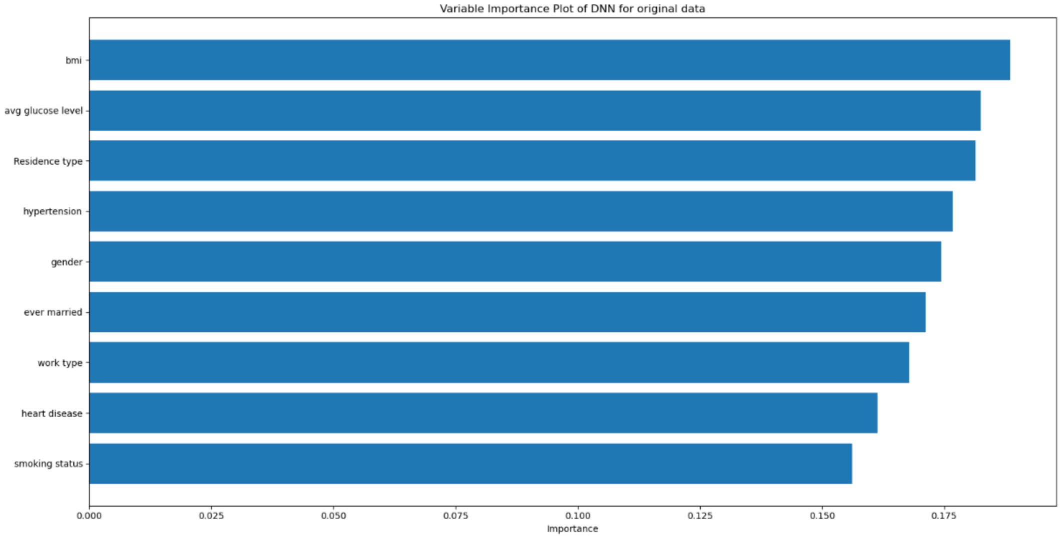

Variable Importance Plot of DNN for above 65 oversampled data
avg glucose level
hypertension
heart disease
ever married
Residence type
gender
work type
bmi
smoking status
0.00
0.05
0.10
0.15
0.20
Importance


Decision Tree Classifier - Feature Importance for above 65
bmi
avg glucose level
ever married
smoking status
hypertension
work type
Residence type
heart disease
gender
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance


Variable Importance Plot of Logistic model for oversampled data
ever married
hypertension
heart disease
Residence type
work type
smoking status
gender
ıvg glucose level
bmi
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Importance


Decision Tree Classifier - Feature Importance for oversampled data
Feature
avg glucose level
hypertension
ever married
bmi
heart disease
smoking status
gender
work type
Residence type
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance


Decision Tree Classifier - Feature Importance for Original data
avg glucose level
bmi
ever married
hypertension
heart disease
smoking status
Residence type
work type
gender
Feature
0.0
0.1
0.2
0.3
0.4
Importance


Variable Importance Plot of DNN for original data
bmi
avg glucose level
Residence type
hypertension
gender
ever married
work type
heart disease
smoking status
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Importance