

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

ANIT MATHEW

MS. DATA SCIENCE, FALL 2023

COMMITTEE MEMBERS:

DR. OSEI KOFI TWENEBOAH, ADVISOR

DR. SCOTT FREES, READER

DR. SOURAV DUTTA, READER

INTRODUCTION

- ▶ Global concern: Aging population and the impact of Cardiovascular Diseases (CVDs).
- ▶ CVDs as the leading cause of global mortality (17.9 million lives annually).
- ▶ Major contributors: Coronary heart disease, cerebrovascular disease, and rheumatic heart disease.
- ▶ Over 80% of CVD-related deaths result from heart attacks and strokes, with one-third occurring prematurely in individuals under 65.
- ▶ The thesis explores key features on grading the factors of stroke across different age groups, focusing on data from different age groups.

ELDERLY POPULATION

▶ **Definition of Elderly Population:**

- ▶ Individuals aged 65 and older.
- ▶ In 1987, the U.S. had over 30 million elderly individuals, constituting 12% of the total population.

▶ **Impact on Healthcare:**

- ▶ Almost 96% of Medicare recipients belong to the elderly population.
- ▶ Significant implications for healthcare considerations due to its substantial impact.

▶ **The Graying of America:**

- ▶ Between 1960 and 1986, a 75% increase in the elderly population (65 and older).
- ▶ Contrastingly, the population under 65 increased by only 30% during the same period.

▶ **Age Group Distribution (1986):**

- ▶ 65 to 74 age group: Three-fifths of the elderly population.
- ▶ 75 to 84 age group: One-third of the elderly population.
- ▶ 85 and older: One-tenth of the elderly population.

▶ **Future Projections (1987 to 2030):**

- ▶ Total U.S. population expected to increase by 26% (reaching 317 million).
- ▶ Elderly population projected to increase by more than 100%.
- ▶ Proportion of elderly population to rise from 12% to nearly 21% by 2030 (67 million individuals).

CHALLENGES AND URGENCY

- ▶ Aging population faces an escalating threat due to the complex interplay of aging, cardiovascular health, and risk factors.
- ▶ Prevalence of behavioral risk factors intensifies the urgency, leading to physiological markers associated with an elevated risk of heart-related complications.

PROBLEM STATEMENT

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 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?
- ▶ Can additional variables, such as gender, average blood glucose levels, and type of residence, influence the risk of brain strokes?
- ▶ What key factor should different age groups prioritize to safeguard themselves from 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 different age groups.

By investigating these relationships, we intend to provide valuable insights for:

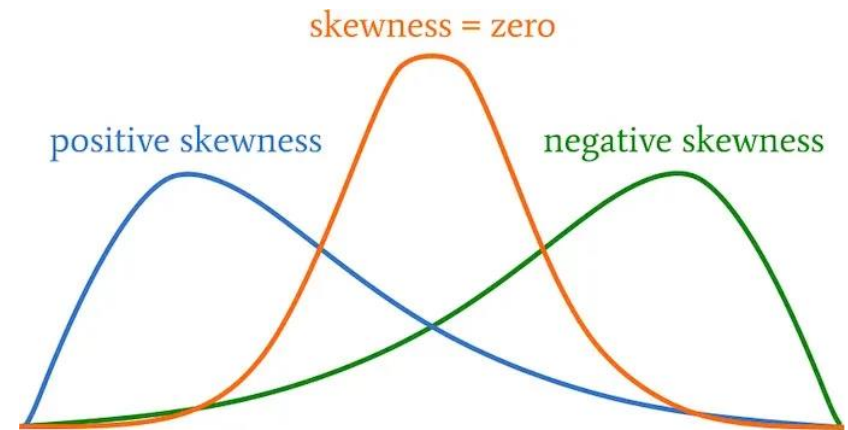
- ▶ Public Health
- ▶ Elderly Care and Geriatrics:
- ▶ Health Promotion and Policy
- ▶ Research Foundation

BACKGROUND

- ▶ The burden of CVDs, especially in the elderly, leads to strokes with long-term consequences. Lifestyle choices, such as smoking and body mass index, are crucial in CVD development, making it essential to understand their impact on stroke frequency.
- ▶ Drawing from a study by Elias Dritsas and Maria Trigka, the research employs ML techniques, particularly the stacking method, for long-term stroke risk prediction.
- ▶ The stacking method, combining multiple ML models, demonstrated superior predictive capabilities with an AUC of 98.9%, precision, recall, and F-measure values of 97.4%, and an overall accuracy of 98%.
- ▶ The significance of early stroke prediction is underscored by the staggering global statistics, and the research provides a reliable framework for long-term stroke risk assessment.
- ▶ Examining various age groups by utilizing a variable importance plot to gain insights into the key factors contributing to cardiovascular diseases (CVDs)


SKEWNESS

- ▶ Skewness is a measure of the asymmetry or lack of symmetry in a set of data. In simple terms, it tells us whether the data is more concentrated on one side.
- ▶ If the data is concentrated on the left side and the tail on the left is longer or fatter, it's considered negatively skewed.
- ▶ If the data is concentrated on the right side and the tail on the right is longer or fatter, it's considered positively skewed.
- ▶ In a perfectly symmetrical distribution, the skewness is zero. Skewness helps us understand the shape of the data distribution and can provide insights into its characteristics.
- ▶ In this thesis, we will be measuring skewness of stroke counts.



CHI- SQUARE TEST

- ▶ The Chi-Square Test is a statistical method that helps us figure out if there's a significant association between two categorical variables. In simpler terms, it helps us see if there's a relationship between two things that can be categorized (like yes/no, red/blue, etc.).
- ▶ Ex: Imagine you have data on whether people prefer tea or coffee and whether they are morning or night people. The Chi-Square Test would tell you if there's a connection between people's beverage preference and their preference for morning or night.
- ▶ In essence, it helps us determine if the differences in the distribution of categories are due to random chance or if there's a real relationship between the variables.

An illustration of a hand dropping several coins into the air. The coins are shown in a vertical sequence, suggesting motion. The hand is a simple line drawing with a blue cuff.

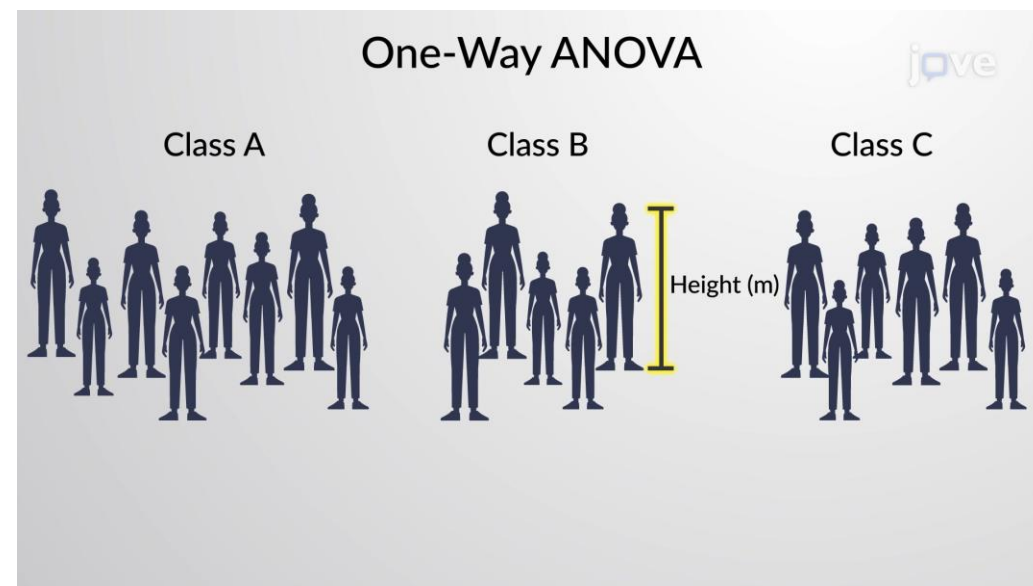
**Chi-Square (χ^2)
Statistic**

['kī-'skwer stə-'ti-stik]

A test that measures how a model compares to actual observed data.

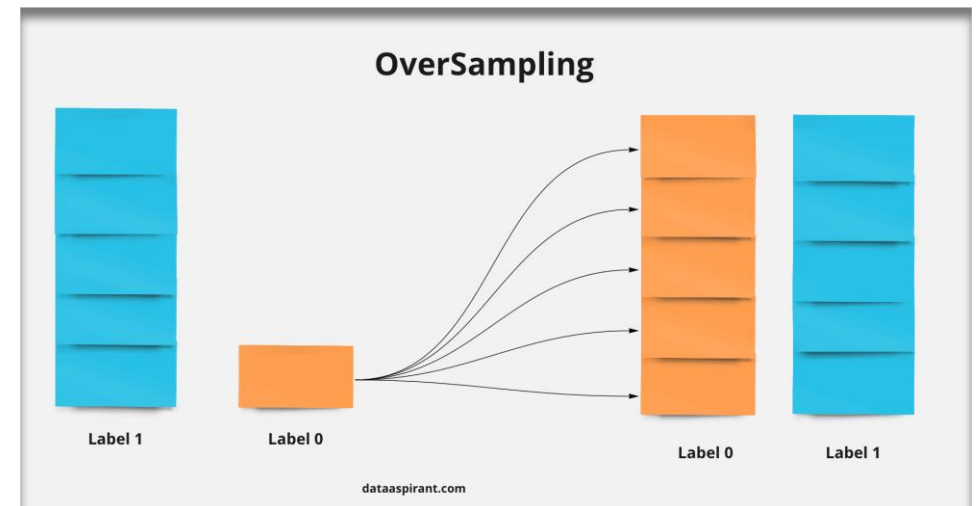
UNEVEN ANOVA TEST

- Uneven ANOVA, or one-way analysis of variance, is a statistical test used to compare the means of three or more groups to determine if there are significant differences among them. The term "uneven" here doesn't refer to the size of the groups but rather to the fact that the groups may have different sample sizes.



OVERSAMPLING

- ▶ Oversampling is like making sure everyone's voice is heard equally.
- ▶ Imagine you're trying to decide on a movie with your friends, but some friends speak softly, and their movie preferences might not be considered as much.
- ▶ In oversampling, we would give a microphone to those quieter friends more often so that their opinions have a fair chance, just like the louder ones.
- ▶ Similarly, in data, oversampling is a technique where we make sure that less common examples (like rare diseases in a health study) are represented more in the dataset, so the computer model can learn about them better and make fair predictions.

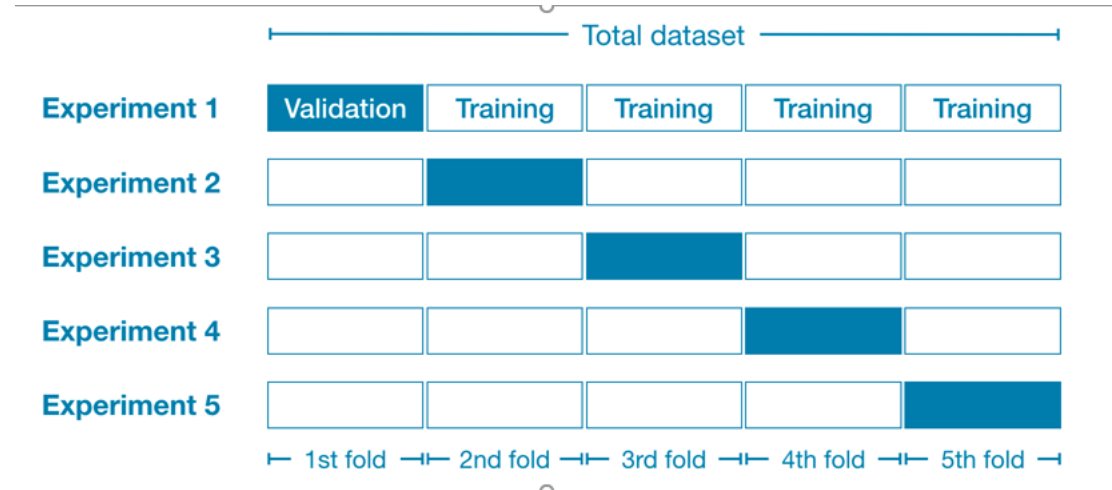


RANDOM OVERSAMPLING

- ▶ Random oversampling is like adding more copies of the minority group to make things fair.
- ▶ Imagine you have a group picture, but some friends are not as visible as others.
- ▶ Random oversampling is like taking more pictures with those less visible friends so that everyone gets an equal chance to be noticed.
- ▶ It helps balance the representation of different groups, especially when dealing with imbalanced data in things like predicting strokes or diseases.

CROSS VALIDATION

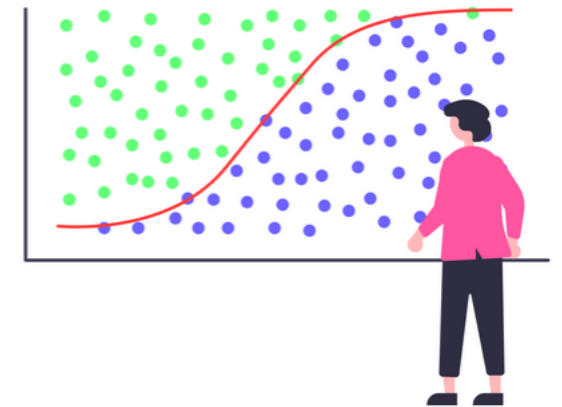
- ▶ Cross-validation is like a test that a student takes multiple times to ensure a fair assessment of their learning.
- ▶ Similarly, in machine learning, cross-validation is a technique used to evaluate how well a model will perform on new data.
- ▶ Instead of relying on a single split of data into training and testing sets, cross-validation involves dividing the data into multiple parts, training the model on different combinations, and testing it on the remaining data.
- ▶ This helps ensure that the model is robust and performs well in various scenarios, reducing the risk of overfitting or underfitting to a specific dataset.
- ▶ It's like giving the model different versions of the test to make sure it truly understands the material.



LOGISTIC REGRESSION

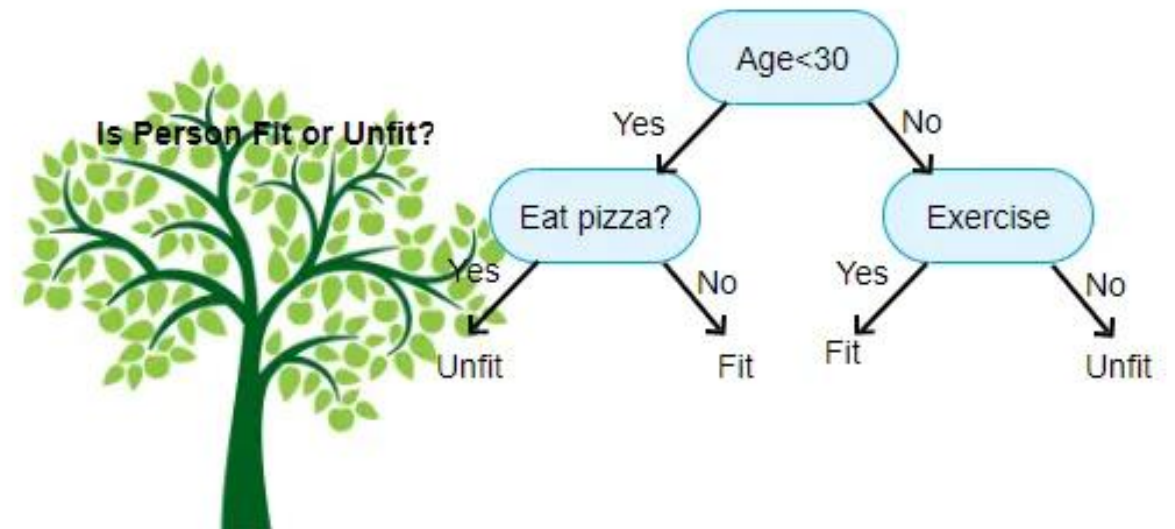
- ▶ In simple words, logistic regression is a statistical method used to predict the probability of an event happening. It's commonly used when the outcome you want to predict is binary, meaning it can have only two possible outcomes, like "yes" or "no," "success" or "failure," or "1" or "0."
- ▶ Imagine you're trying to predict whether a student will pass or fail an exam based on the number of hours they studied. Logistic regression takes into account the relationship between the hours of study and the likelihood of passing. Instead of giving a straight "yes" or "no" answer, it provides a probability score between 0 and 1. If the probability is closer to 1, it suggests a higher likelihood of passing; if closer to 0, it suggests a higher likelihood of failing.

WHAT IS
**LOGISTIC
REGRESSION?**



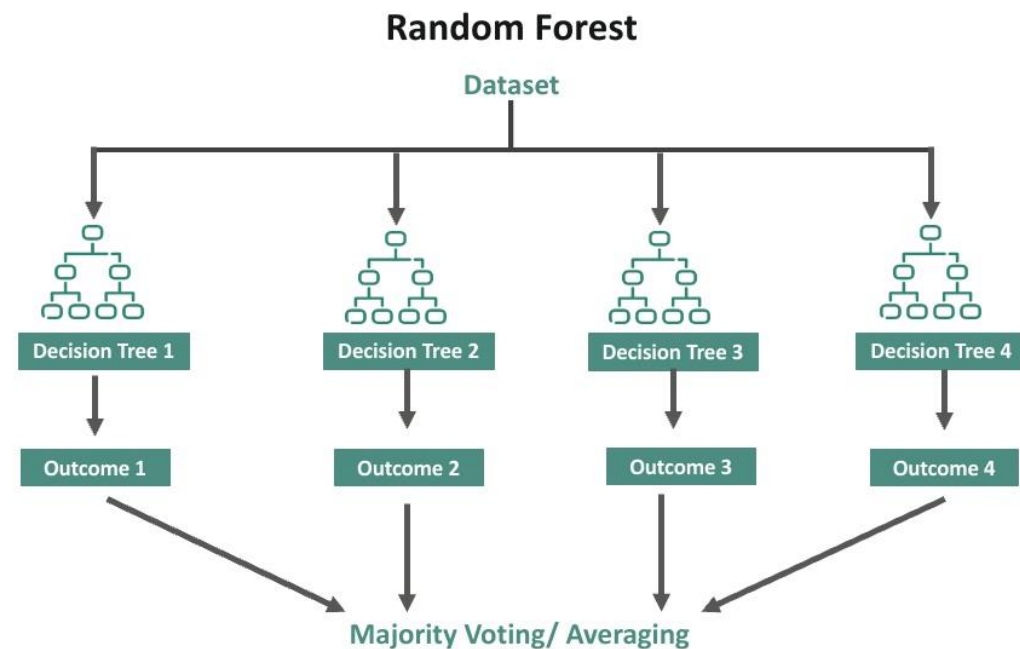
DECISION TREE

- ▶ A Decision Tree is like a smart flowchart that helps computers or people make decisions by asking a series of questions and narrowing down the options based on the answers.
- ▶ Imagine you're trying to decide what movie to watch. You start with a general question like, "Do I want to watch a comedy, drama, or action movie?" Depending on your answer, you move to more specific questions. If you chose comedy, you might ask, "Do I prefer classic comedies or recent releases?" Each question leads to more questions until you reach a specific movie choice.



RANDOM FOREST

- ▶ Imagine you have a big decision to make, like whether to go on a picnic or not. You might ask different friends for advice. Each friend has their own opinions and experiences, and they may consider different factors like weather, distance, or the availability of snacks.
- ▶ In the world of machine learning, Random Forest is like asking a bunch of friends (trees) for advice to make a decision (predict an outcome). Here's how it works:
 - ▶ **Build a Group of Friends (Trees):** Create a "forest" by training several decision trees. Each tree is like asking one friend for advice.
 - ▶ **Ask Each Friend (Tree) for Advice:** When you want to make a decision (predict an outcome), you ask each friend (tree) for their opinion. In machine learning, each tree makes its own prediction based on the features (like weather, distance, snacks) you provide.
 - ▶ **Count the Votes:** In Random Forest, the final decision is made by combining the opinions of all the friends (trees). The outcome with the most votes becomes the final prediction.



DEEP LEARNING – DNN MODEL

- ▶ Deep Learning, specifically Deep Neural Networks (DNN), can be understood as a computer system that learns to perform tasks by mimicking the human brain's way of processing information. Simply put, it's like training a computer to recognize patterns and make decisions on its own.
- ▶ Why the DNN model here?
 - ▶ **Feature Learning:** DNNs excel at feature learning, where the model identifies relevant features or patterns in the data during training. In stroke prediction, this is crucial for discerning subtle but significant indicators of risk, which may not be apparent through manual analysis.
 - ▶ **Handling Imbalanced Data:** Imbalanced datasets, where instances of strokes may be significantly fewer than non-stroke cases, are common in healthcare. DNNs, when appropriately configured, can address class imbalances, improving the model's ability to generalize and predict rare events like strokes.
 - ▶ **Large and Diverse Datasets:** DNNs thrive on large and diverse datasets. Stroke prediction datasets are often expansive, incorporating a wide range of variables. DNNs can effectively process and learn from such data, capturing nuances that may be challenging for simpler models.

SCOPE AND METHODOLOGY

- ▶ This study will thoroughly analyze a publicly available dataset encompassing a diverse set of health-related variables.
- ▶ The focus is on exploring the associations between lifestyle factors, health conditions, and the occurrence of brain strokes in individuals aged 65 and above.
- ▶ This exclusive focus on older adults recognizes this age group's unique healthcare challenges and demographic shifts.
- ▶ Given the global aging population, understanding and addressing the specific health concerns of older individuals is crucial.
- ▶ The research will employ statistical and data analysis techniques to unveil trends and associations, concentrating on variables such as 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
- ▶ The data is publicly accessible 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.

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

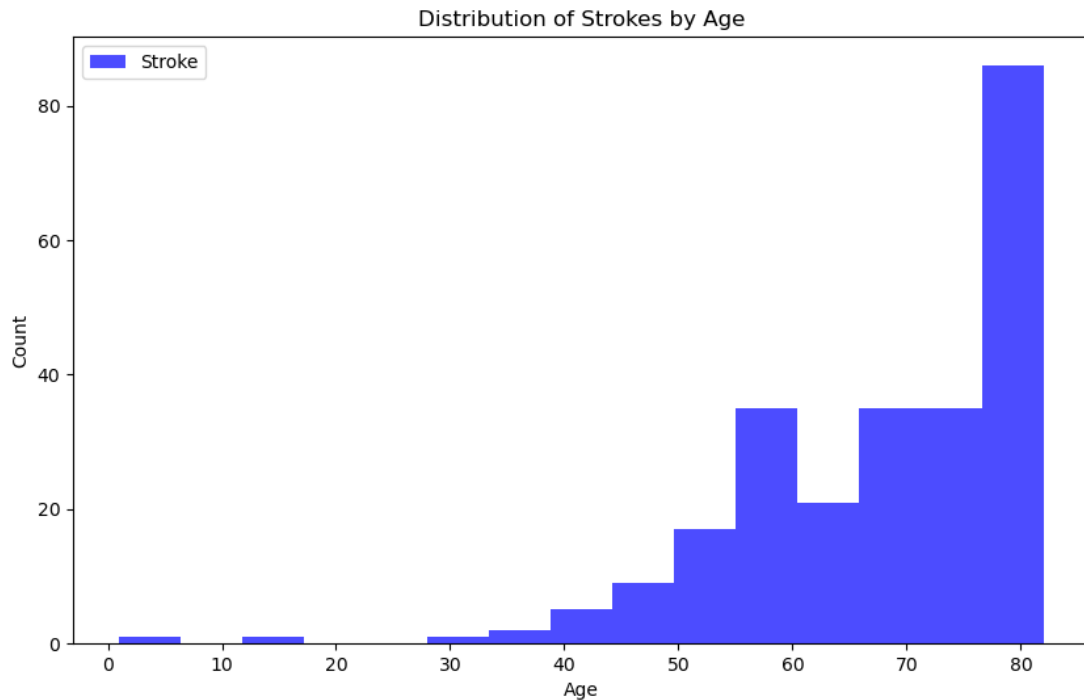
DATA

- ▶ 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

Exploratory Data Analysis (EDA)

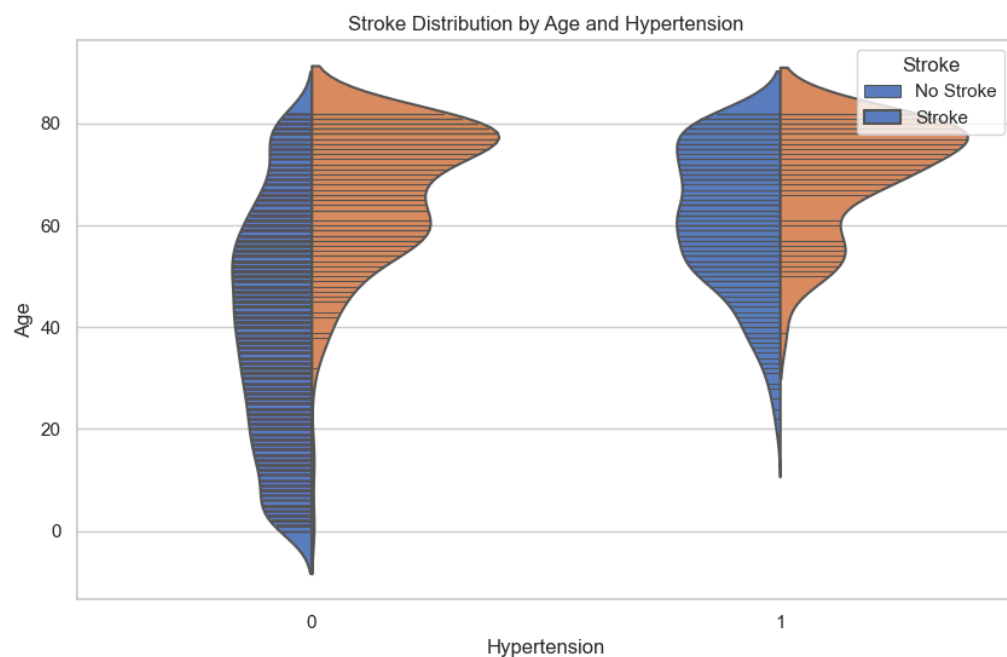
Exploratory Data Analysis (EDA) is a crucial initial step in the data analysis process that involves summarizing, visualizing, and understanding the main characteristics of a dataset.

Exploratory Data Analysis (EDA)



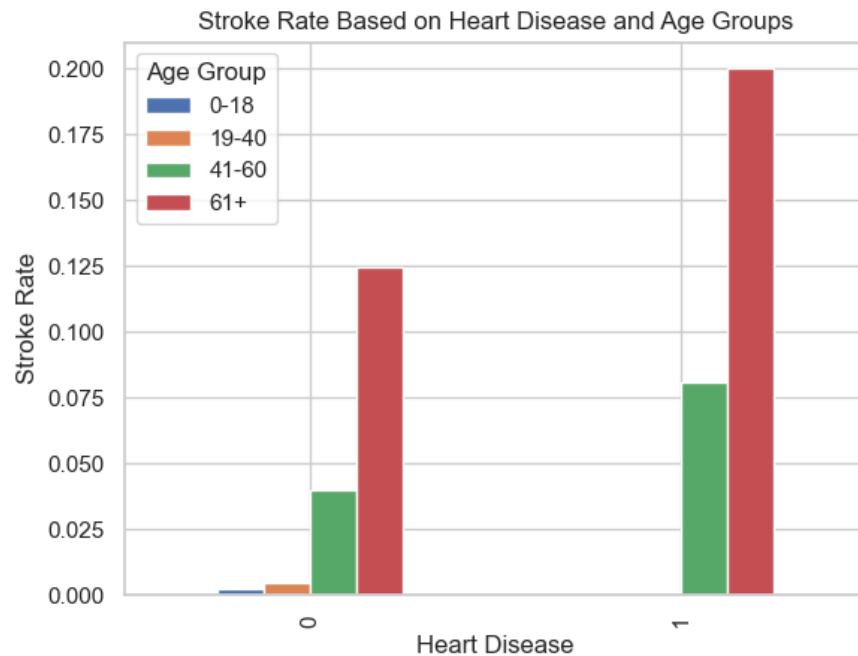
- Stroke incidence increases with age, peaking in individuals over 80 years old.
- Strokes can occur at any age, with a significant number occurring in people under 50.
- The distribution of strokes by age is not uniform, showing distinct peaks at around 55 and 80 years old.
- This suggests two different risk factors for strokes: one affecting younger individuals and another affecting older individuals.

Exploratory Data Analysis (EDA)



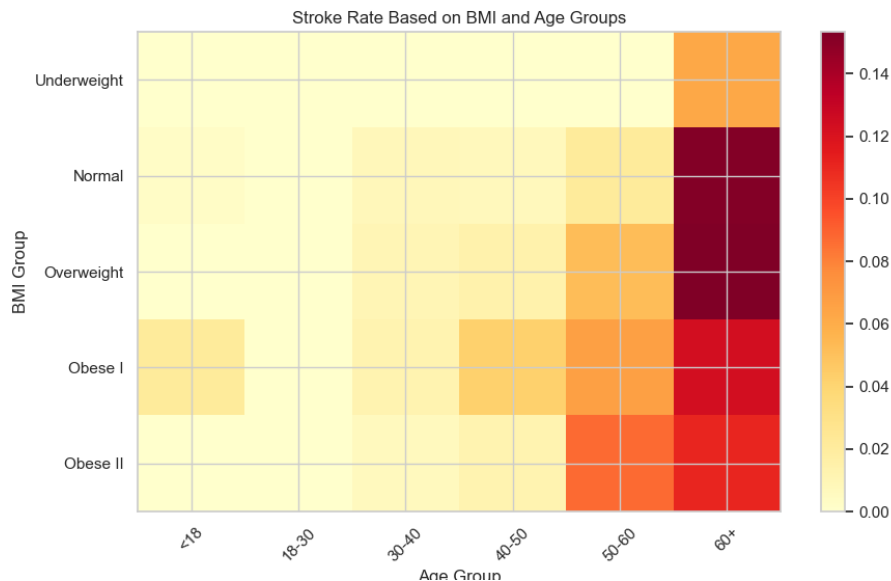
- ▶ Risk of stroke increases with age and hypertension.
- ▶ Median age of stroke is around 65 years.
- ▶ Higher risk in individuals with hypertension.
- ▶ Distribution of strokes is right-skewed, indicating more strokes in older age and higher blood pressure.
- ▶ Risk accumulates over time.
- ▶ Hypertension is a significant stroke risk factor.

Exploratory Data Analysis (EDA)



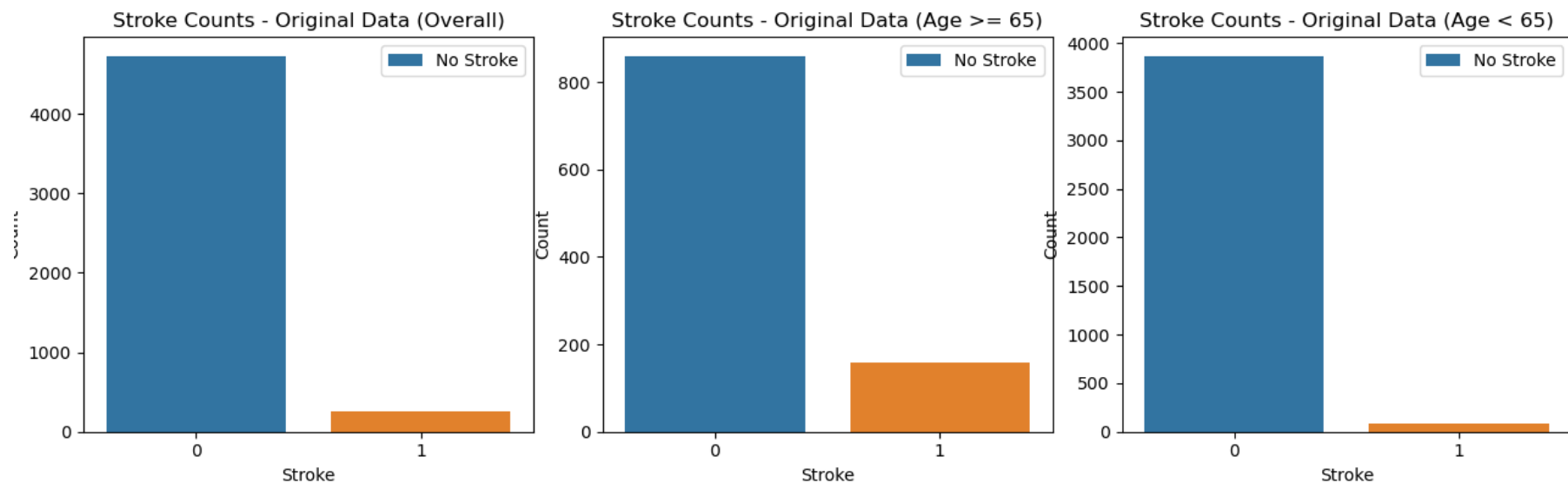
- ▶ For individuals without heart disease (heart disease=0), the likelihood of stroke is generally low across all age groups, with the highest occurrence in the 61+ age group.
- ▶ Conversely, for individuals with heart disease (heart disease=1), the risk of stroke significantly increases across all age groups, reaching the highest incidence among those aged 61 and above.
- ▶ Older individuals, particularly those in the 61+ age group, are generally at a higher risk of stroke.
- ▶ The presence of heart disease amplifies the risk of stroke, especially among individuals in the 61+ age group.

Exploratory Data Analysis (EDA)



- ▶ The heatmap indicates a correlation between increased risk of stroke and higher BMI (Body Mass Index) as well as advancing age.
- ▶ Obesity is identified as a significant risk factor for stroke, with individuals carrying excess weight being more prone to other stroke-related risk factors such as high blood pressure, high cholesterol, and diabetes.
- ▶ Advancing age contributes to an elevated risk of stroke due to the natural aging process causing arteries to narrow and harden, potentially leading to reduced blood flow to the brain.
- ▶ The heatmap highlights that the highest risk of stroke is observed in individuals who are both in the obese overweight category and aged 61 or above.

STROKE COUNTS & SKEWNESS



Stroke Counts of Original data	
0	4733 – 95%
1	248 – 5%

Stroke Counts in the Age Group 65 and above	
0	861 – 84.4%
1	159 – 15.6%

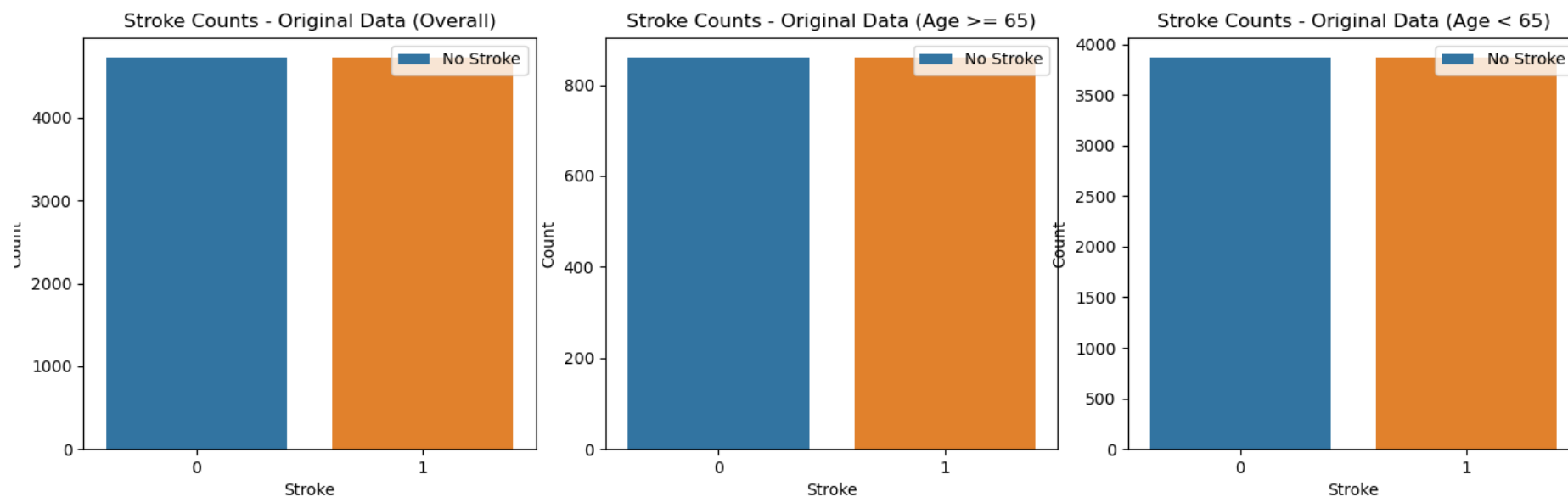
Stroke Counts in the Age Group Below 65	
0	3872 – 97.7%
1	89 – 2.3%

Skewness of Original data	
stroke	4.140942

Skewness in the Age Group of 65 and above	
stroke	1.9001

Skewness in the Age Group Below 65	
stroke	6.446711

RESULTS AFTER OVERSAMPLING



Stroke Counts of Oversampled data	
0	4733
1	4733

Skewness of Oversampled data	
stroke	0

Stroke Counts of Oversampled data above 65	
0	861
1	861

Skewness of Oversampled data above 65	
stroke	0

Stroke Counts of Oversampled data below 65	
0	3872
1	3872

Skewness of Oversampled data below 65	
stroke	0

UNEVEN ANOVA TEST RESULTS

Group 1 = Overall data with all age groups

Group 2 = Age above 65

Group 3 = Age below 65

BMI group

F-Statistic: 131.4319176446027

P-Value: 4.9917105867046106e-57

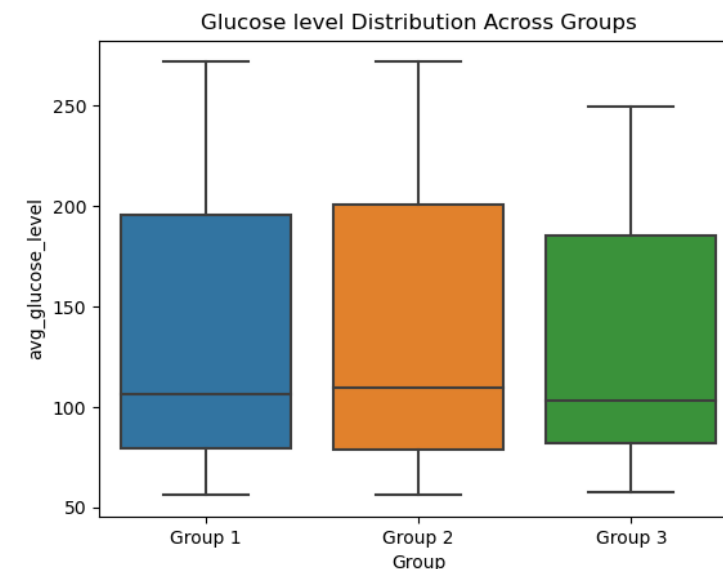
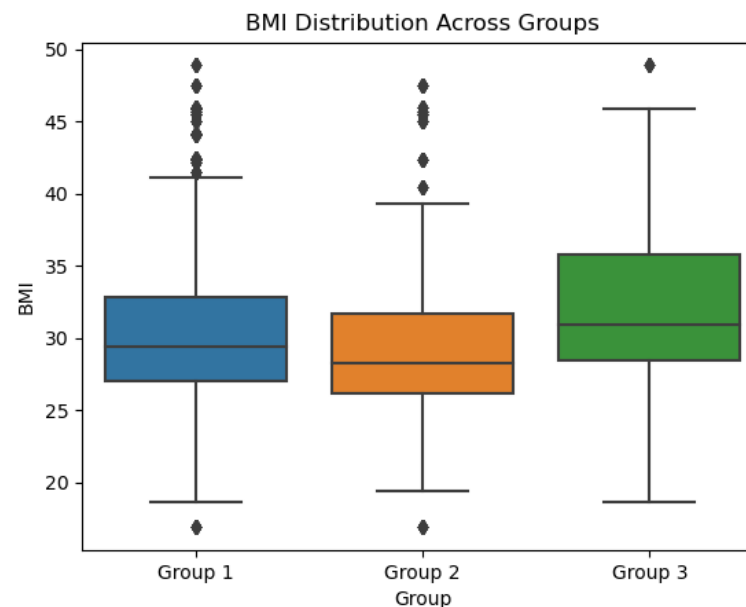
Reject the null hypothesis: There are significant differences between groups.

Average Glucose level group

F-Statistic: 22.25982386306607

P-Value: 2.2664737765469657e-10

Reject the null hypothesis: There are significant differences between groups.



UNEVEN ANOVA TEST RESULTS

- With a small p-value and a high F-statistic, we reject the null hypothesis.
- Therefore, there is statistical evidence to suggest that there are significant differences in the 'BMI' and 'avg glucose level' values among the three age groups (age < 65 , age ≥ 65 , and the overall group with stroke).

CHI-SQUARE TEST RESULTS

1. smoking status vs. stroke:

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

2. gender vs. stroke:

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

7. Residence type vs. stroke:

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

3. hypertension vs. stroke:

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

4. ever married vs. stroke:

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

5. heart disease vs. stroke:

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

6. work type vs. stroke:

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

CHI-SQUARE TEST RESULTS

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.

MACHINE LEARNING RESULTS – INITIAL DATA

Cross-Validation Results

Mean Cross-Validation Score	Accuracy
Logistic	0.950411677
Decision	0.949046992
DNN	0.950211
Random	0.9462

MACHINE LEARNING RESULTS – INITIAL DATA

Cross-Validation Results

► Logistic Regression:

- Strong performance with an average accuracy of approximately 95%.
- Demonstrated robust predictive capabilities in five-fold cross-validation.
- Challenges due to class imbalance, especially in predicting strokes.
- High precision and recall for the majority class (no stroke), but lower metrics for the minority class (stroke).

► Decision Tree:

- Stable performance with a mean cross-validation score of around 94.9%.
- Consistently demonstrated consistent and reliable predictive capabilities.

► Deep Neural Network (DNN):

- Achieved impressive accuracy levels of about 95% consistently.
- Reliable ability to classify instances accurately.
- No specific mention of addressing class imbalance.

► Random Forest:

- Demonstrated generalization capability.
- Consistently achieved accuracy levels above 94% across five folds.
- Reinforces confidence in its predictive power.

MACHINE LEARNING RESULTS – INITIAL DATA

Classification Report – Best two

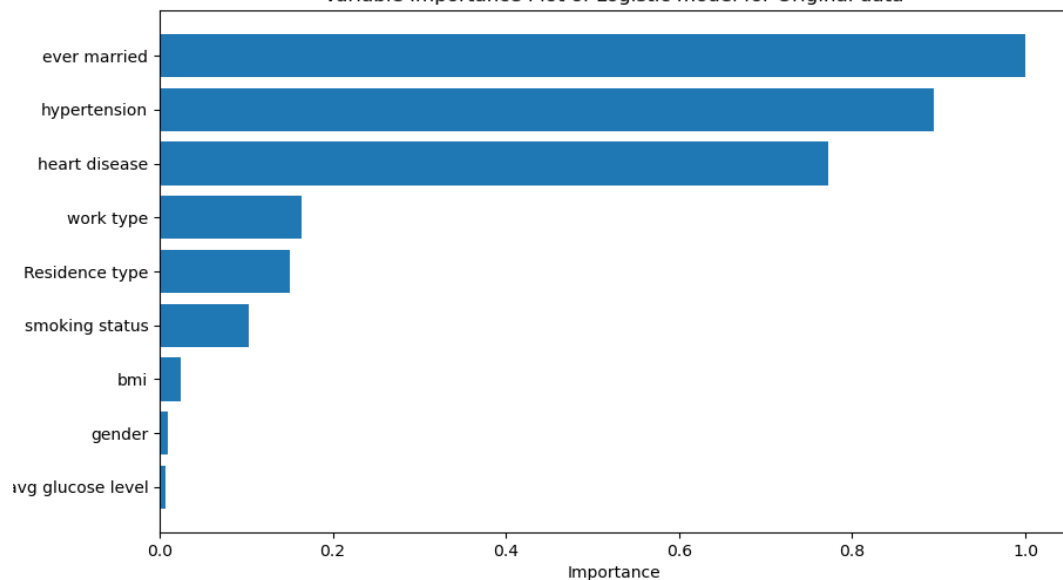
	Class	Precision	Recall	F1-Score	Support
Logistic	0	0.95	1	0.97	943
	1	0	0	0	54
	Accuracy			0.95	997
	Macro Avg	0.47	0.5	0.49	997
	Weighted Avg	0.89	0.95	0.92	997
Random	0	0.95	1	0.97	943
	1	0	0	0	54
	Accuracy			0.94	997
	Macro Avg	0.47	0.5	0.49	997
	Weighted Avg	0.89	0.94	0.92	997

- ▶ Logistic Regression model achieves 95% accuracy across five folds.
- ▶ Strong performance metrics for class 0 (no stroke), but limited success in identifying instances of stroke (class 1).
- ▶ Random Forest model demonstrates approximately 94% overall accuracy but faces challenges in classifying individuals at risk of stroke.
- ▶ Consistent low precision and recall for class 1 in both models indicate issues with identifying stroke occurrences.

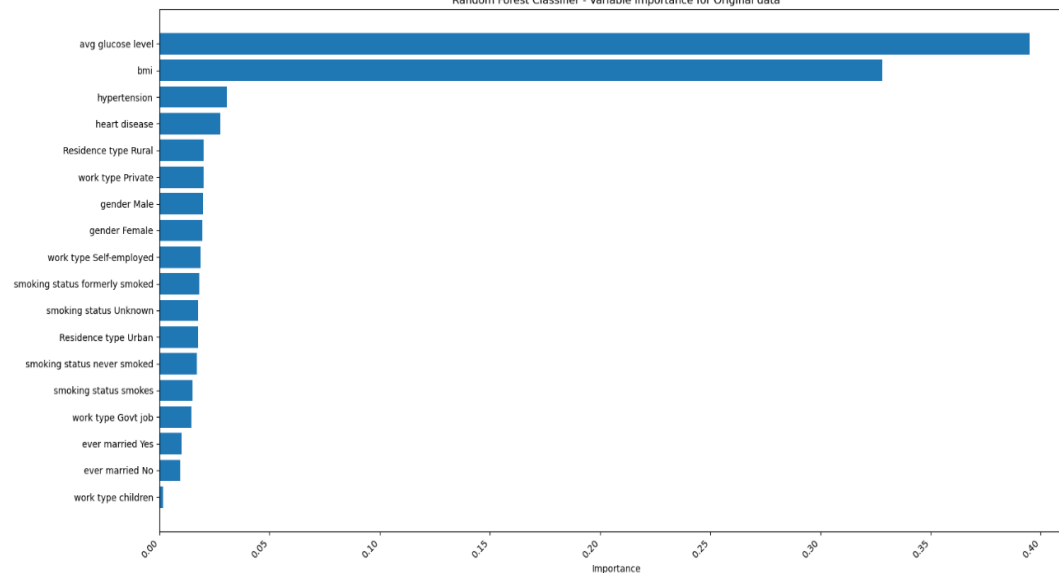
MACHINE LEARNING RESULTS – INITIAL DATA

Feature Importance plot– Best two

Variable Importance Plot of Logistic model for Original data



Random Forest Classifier - Variable Importance for Original data



MACHINE LEARNING RESULTS – INITIAL DATA

Feature Importance plot– Best two

► Logistic Model

- "ever married" emerged as the most critical predictor for stroke risk.
- "Hypertension" and "heart disease" followed as significant contributors, aligning with medical literature.
- Features like "work type" and "residence type" also showed notable importance in predicting strokes.

► Random Forest Model:

- Variable importance analysis revealed "bmi" and "avg glucose level" as the most influential factors.
- These findings align with established medical knowledge on the link between obesity, glucose levels, and cardiovascular health.
- Lifestyle factors such as "smoking status" and "work type" also played a substantial role in predicting stroke risk.

MACHINE LEARNING RESULTS – OVERSAMPLED DATA

Cross-Validation Results

Mean Cross-Validation Score	Accuracy
Logistic	0.6865
Decision	0.7050
DNN	0.7104
Random	0.9875

MACHINE LEARNING RESULTS – OVERSAMPLED DATA

Cross-Validation Results

- ▶ Logistic regression model: Average accuracy of 68% across five-fold cross-validation.
 - ▶ Balanced precision, recall, and F1-score for both classes (0 and 1).
 - ▶ Slightly higher accuracy in predicting class 0, indicating a need to improve sensitivity for cardiovascular disease risk.
- ▶ Decision tree model: Cross-validation scores ranged from approximately 69.77% to 71.93%.
 - ▶ Mean cross-validation score of 70.51% indicates consistent and stable performance.
 - ▶ Reliable predictions on oversampled data, showcasing generalization capabilities.
- ▶ DNN model: Consistent performance with accuracy ranging from approximately 69.68% to 73.48%.
 - ▶ F1-score ranging from 69.62% to 74.10%, suggesting effective pattern capture and reliable stroke risk predictions.
- ▶ Random Forest model: Exceptional average accuracy exceeding 98.75% across five folds.
 - ▶ Demonstrates robust generalization to unseen data, indicating high performance and reliability.
- ▶ Overall: Random Forest and DNN models showcase strong predictive capabilities on oversampled data.
 - ▶ Random Forest excels in exceptionally high accuracy, while DNN demonstrates consistent and reliable performance.
 - ▶ The choice between the two may depend on considerations such as interpretability, computational efficiency, or specific application requirements.

MACHINE LEARNING RESULTS – OVERSAMPLED DATA

Classification Report – Best two

	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

► Random Forest Model:

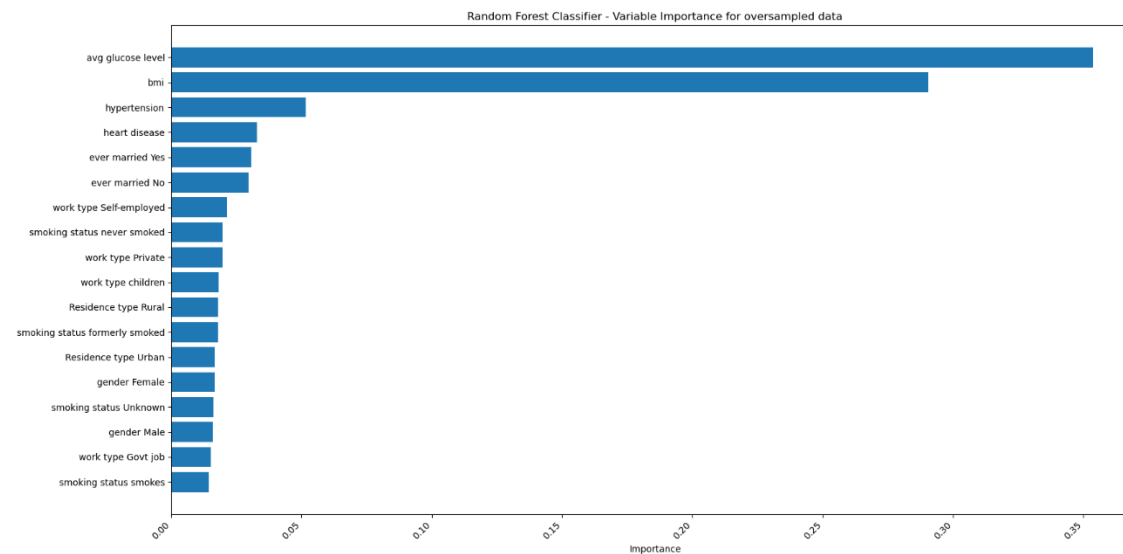
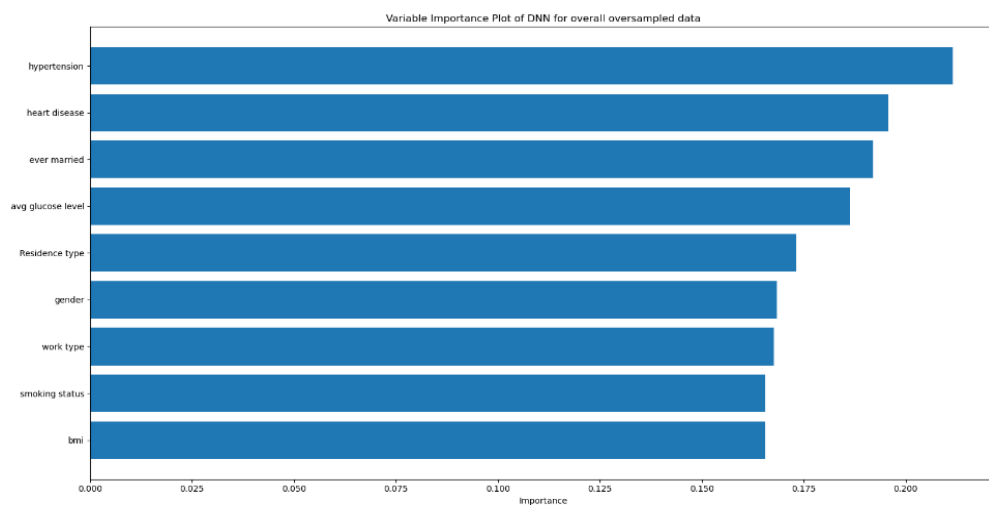
- Accuracy: 98.79%
- Precision, recall, and F1-scores for both classes (0 and 1) are high.
- Excellent balance in identifying both positive and negative cases, with precision-recall approaching 99%.

► Deep Neural Network (DNN) Model:

- Balanced performance with equal precision, recall, and F1-score values of 0.71 for both classes.
- Overall accuracy: 71%.
- Macro average and weighted average metrics support balanced performance, all values aligning at 0.71.

MACHINE LEARNING RESULTS – OVERSAMPLED DATA

Feature Importance plot– Best two



MACHINE LEARNING RESULTS – OVERSAMPLED DATA

Feature Importance plot– Best two

► DNN Model Variable Importance:

- Key features influencing the DNN model include "Hypertension" (21.16%) and "Ever Married" (19.20%).
- Other notable contributors are "Heart Disease," "Avg Glucose Level," and "Residence Type."

► Random Forest Model Variable Importance:

- In the Random Forest model, "Avg Glucose Level" and "BMI" emerge as the most influential predictors.
- "Hypertension" and "Heart Disease" show substantial importance, aligning with established medical knowledge.
- Specific attributes such as "Smoking Status," "Work Type," "Gender," and "Residence Type" contribute to stroke prediction.

MACHINE LEARNING RESULTS – ABOVE 65

Cross-Validation Results

Mean Cross-Validation Score	Accuracy
Logistic	0.57
Decision	0.6041
DNN	0.6167
Random	0.9442

MACHINE LEARNING RESULTS – ABOVE 65

Cross-Validation Results

- ▶ Logistic regression model exhibited varying but moderate performance (53.8% to 60.3%) across five folds, indicating a balanced ability to predict both classes.
- ▶ Decision tree model showed reasonable consistency (mean cross-validation score of about 60%) but displayed sensitivity to data partitioning, suggesting potential for improvement through hyperparameter tuning.
- ▶ DNN model demonstrated consistent but moderate predictive capabilities (60% to 62.9% accuracy,) across different folds, highlighting the need for robust evaluation and understanding of generalization.
- ▶ Random Forest model proved reliable with an average accuracy of approximately 94.4% across five folds, indicating strong generalization capabilities and making it a robust choice for predicting stroke occurrence in individuals above 65.
- ▶ Logistic regression model's evaluation on oversampled data for individuals above 65 revealed varying but moderate predictive capabilities, emphasizing a balanced ability to predict both classes.

MACHINE LEARNING RESULTS – ABOVE 65

Classification Report – Best two

	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

► Deep Neural Network (DNN) Model:

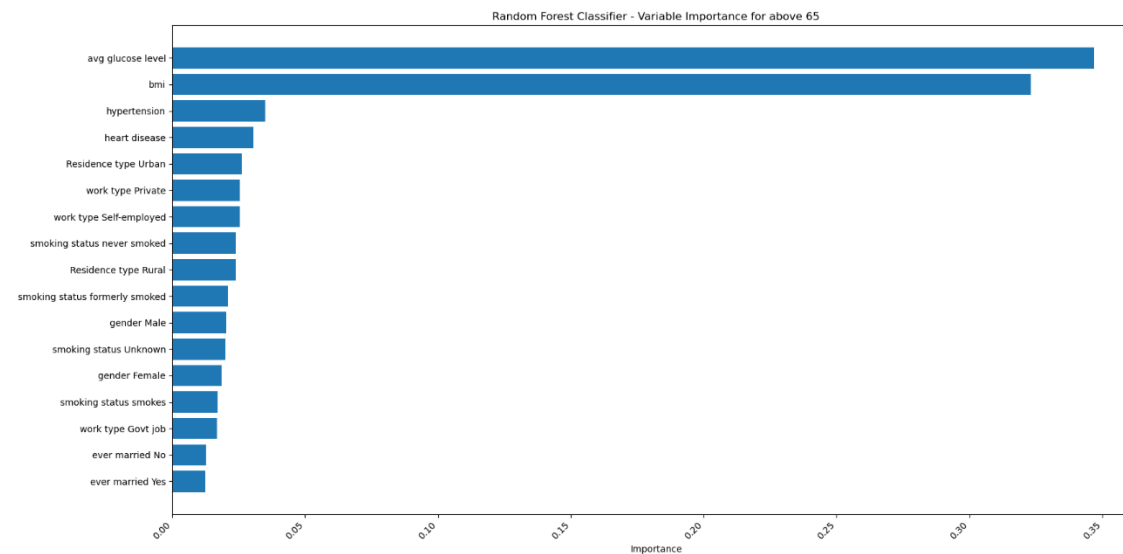
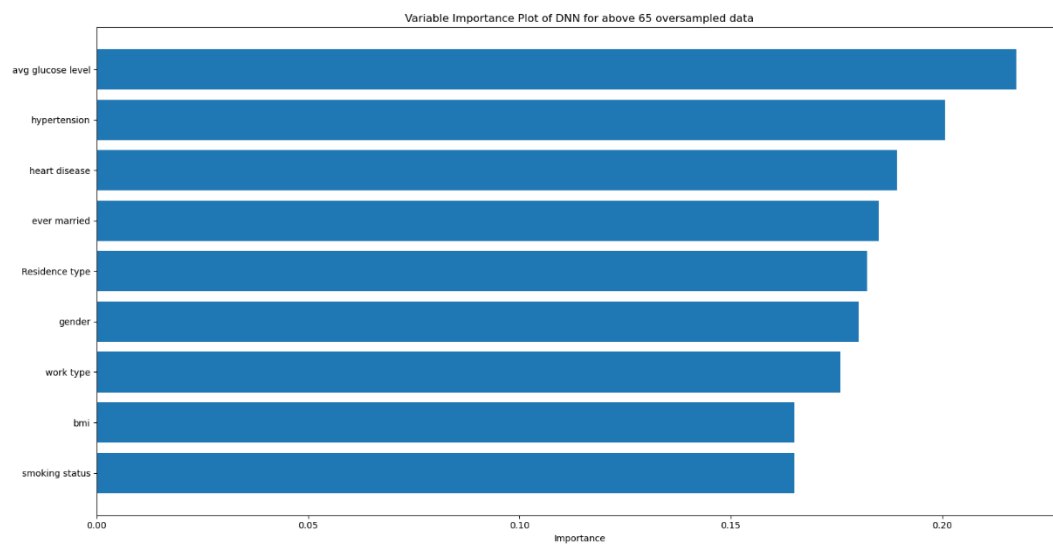
- Precision, recall, and F1-score for both classes (0 and 1) are balanced at approximately 0.62, indicating fair classification ability.
- Macro and weighted averages support consistent performance, yielding an overall accuracy of 62%.

► Random Forest Model:

- Achieves high accuracy of 92%, demonstrating a robust ability to distinguish between classes.
- Class 0 (no stroke) displays notably high precision at 96%, with a respectable recall of 90% and an F1-score of 93%.
- Class 1 (stroke) also exhibits strong performance with precision, recall, and F1-score values exceeding 0.89.
- Macro and weighted averages emphasize overall efficacy, surpassing the DNN model with a weighted average accuracy of 92%.

MACHINE LEARNING RESULTS – ABOVE 65

Feature Importance plot– Best two



MACHINE LEARNING RESULTS – ABOVE 65

Feature Importance plot– Best two

▶ DNN Model Insights:

- ▶ "Avg glucose level," "hypertension," and "heart disease." are key predictors.
- ▶ Aligns with established medical knowledge, emphasizing marital status and cardiovascular health.
- ▶ Other Important Features are "Ever married" and "Residence type."

▶ Random Forest Model Insights:

- ▶ "Avg glucose level" and "bmi" highlighted as most influential.
- ▶ Emphasizes metabolic and obesity-related factors, along with cardiovascular health.
- ▶ Key Predictors are "Hypertension" and "heart disease."

MACHINE LEARNING RESULTS – BELOW 65

Cross-Validation Results

Mean Cross-Validation Score	Accuracy
Logistic	0.6732
Decision	0.7690
DNN	0.7125
Random	0.9967

MACHINE LEARNING RESULTS – BELOW 65

Cross-Validation Results

- ▶ Logistic regression model exhibited moderate performance (66.36% to 69.59% accuracy) in predicting stroke risk across five folds.
- ▶ Classification report indicated reasonably balanced precision and recall for both stroke and no-stroke classes.
- ▶ Cross-validation scores for logistic regression averaged approximately 76.90%, highlighting consistent performance on oversampled data below 65.
- ▶ Decision Tree model demonstrated stability with a mean cross-validation score of around 76.90%, emphasizing reliable performance across different dataset subsets.
- ▶ DNN model showed varying accuracy (67.77% to 74.50%) across folds, with F1 scores ranging from 70.24% to 75.93%, showcasing its ability to balance performance metrics.
- ▶ Random Forest model consistently achieved high accuracy (99.68% on average) across five folds, indicating robust generalization capabilities.
- ▶ Decision Tree and Random Forest emerged as top performers, with balanced metrics and high accuracy, showcasing promising potential for accurate stroke risk prediction in individuals below 65.

MACHINE LEARNING RESULTS – BELOW 65

Classification Report – Best two

	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

► Decision Tree Model Performance

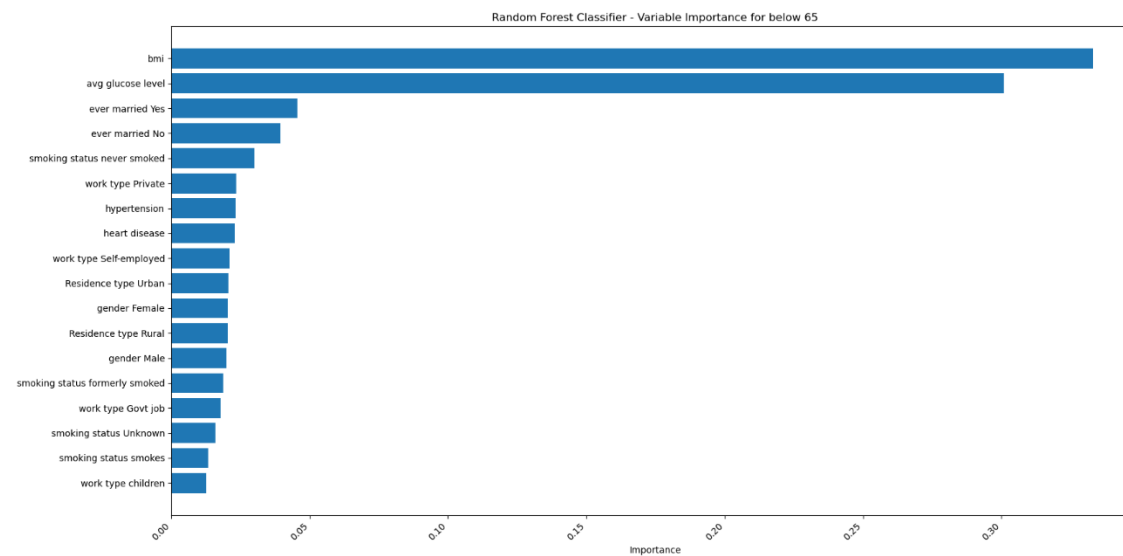
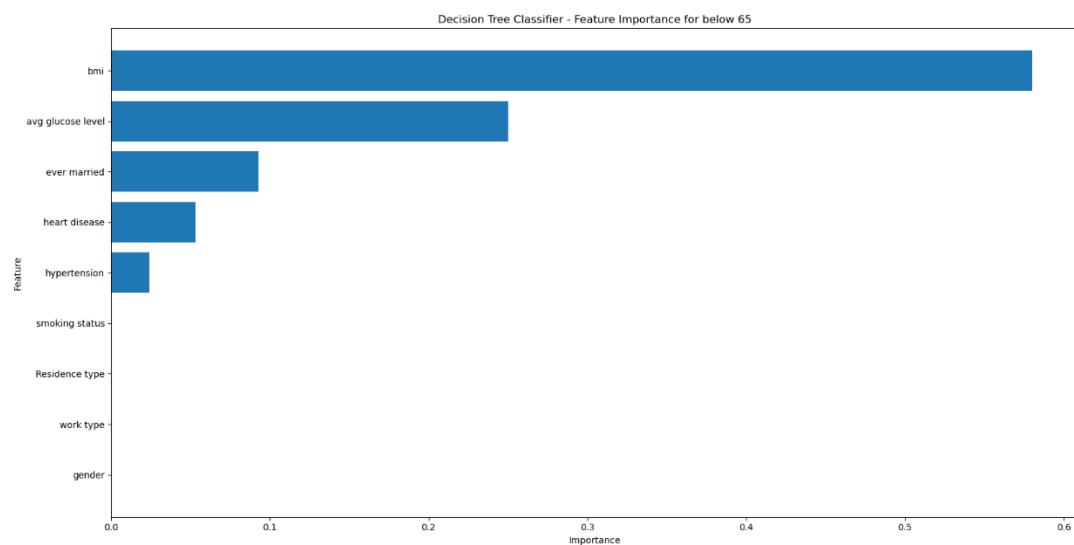
- Precision, recall, and F1-score for classes 0 and 1: approximately 77-78%.
- Balanced predictive capability, effectively identifying stroke risk and no-stroke cases.
- Macro and weighted average metrics confirm overall satisfactory performance on oversampled data.

► Random Forest Model Performance

- Impressive accuracy of approximately 99.42% on the oversampled dataset.
- Consistently high precision, recall, and F1-score for both stroke and no-stroke classes.
- Exceptional recall of 100% for identifying individuals at risk of stroke, highlighting clinical potential.

MACHINE LEARNING RESULTS – BELOW 65

Feature Importance plot– Best two



MACHINE LEARNING RESULTS – ABOVE 65

Feature Importance plot– Best two

► Decision Tree Model Insights:

- Analysis of variable importance highlights "bmi" and "avg glucose level" as most influential features.
- "Hypertension" and "ever married" also contribute meaningfully to stroke risk prediction.
- Promising performance on oversampled data below age 65, demonstrating a balanced approach.
- Evaluation metrics (accuracy, precision, recall) offer a comprehensive understanding of strengths.

► Random Forest Model Insights:

- Feature importance analysis reveals lifestyle and demographic factors like "smoking status," "work type," and "gender" as influential predictors.
- "Avg glucose level" and "bmi" also hold substantial importance in predicting stroke risk.
- The model demonstrates robustness in capturing diverse predictors of stroke risk.
- Aligns with established medical knowledge regarding the association of certain factors with stroke risk.

DISCUSSION

Age Group 0-100:

- ▶ Consistently across various models and datasets, the predictors "avg glucose level," "hypertension," "heart disease," "ever married," and "BMI" emerge as crucial factors in predicting stroke risk.
- ▶ Emphasizing the need for a comprehensive approach that includes monitoring glucose levels, managing hypertension and heart health, exploring social determinants such as marital status, and addressing obesity in stroke prevention strategies.
- ▶ Chi-square tests confirmed significant associations of smoking status, age, gender, and other variables with stroke occurrence.
- ▶ Random Forest consistently highlights "avg glucose level" and "hypertension."

DISCUSSION

Age Group 0-65:

- ▶ Four models consistently identified "ever married," "heart disease," "hypertension," "BMI," and "average glucose level" as top predictors.
- ▶ Managing hypertension and glucose levels emerged as crucial recommendations for cardiovascular health.
- ▶ The analysis explored potential correlations between marital status and stroke risk, prompting further investigation into social and lifestyle aspects.
- ▶ Heart disease and hypertension were underscored as critical factors, reinforcing the need for their effective management.
- ▶ Chi-square tests confirmed significant associations of smoking status, age, gender, and other variables with stroke occurrence.
- ▶ The research provides valuable insights for healthcare professionals and policymakers to devise effective prevention and intervention strategies tailored to the age group 0-65.

DISCUSSION

Age Group Above 65:

- ▶ Consistent identification of influential features, including heart disease, hypertension, smoking, and BMI, in stroke prediction models for individuals above 65.
- ▶ Incidence of heart disease, high blood pressure, and their correlation with strokes confirmed their prevalence in the elderly population.
- ▶ The analysis revealed lifestyle factors, such as smoking and higher BMI, contributing to increased cardiovascular risk in seniors.
- ▶ Exploration of additional variables like gender, average blood glucose levels, and residence type provided a comprehensive understanding of factors influencing stroke risk.
- ▶ The study's insights can inform healthcare professionals and policymakers in developing effective strategies for stroke prevention and management in the elderly population.

CONCLUSION

- ▶ The analysis explores correlations between marital status and stroke risk, highlighting "ever married" as a top predictor for the age group 0-65.
- ▶ Significant risk factors include age, hypertension, heart disease, BMI, and smoking, with age demonstrating non-uniform distribution peaks at 55 and 80 years.
- ▶ Chi-square tests confirm significant associations of smoking status, age, gender, hypertension, heart disease, marital status, work type, and residence type with stroke occurrence.
- ▶ Heart disease and hypertension are primary risk factors for strokes in the elderly.
- ▶ Lifestyle factors (smoking, BMI) impact stroke frequency, with smoking and higher BMI contributing to increased risk.
- ▶ Additional variables (gender, average blood glucose levels, type of residence) significantly contribute to stroke risk in older people.

FUTURE WORK

▶ **Future Work Recommendations:**

▶ **Validation Across Diverse Datasets:**

- ▶ Conduct validation studies using datasets from various sources to assess the generalizability of the identified predictors across different populations.

▶ **Exploration of Additional Factors:**

- ▶ Extend the analysis by incorporating additional variables that might contribute to stroke risk, considering socio-economic, genetic, and environmental factors.

▶ **Longitudinal Studies:**

- ▶ Engage in longitudinal studies to track changes in identified predictors over time, providing a dynamic understanding of stroke risk factors.

▶ **Machine Learning Algorithm Enhancement:**

- ▶ Explore the refinement and enhancement of machine learning algorithms to improve prediction accuracy and account for evolving patterns in stroke risk.

▶ **Implementation of Intervention Strategies:**

- ▶ Collaborate with healthcare professionals to design and implement targeted intervention strategies based on identified predictors, aiming to prevent strokes and enhance public health outcomes.

REFERENCE

- ▶ 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
- ▶ <https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset>
- ▶ 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.
- ▶ 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.
- ▶ <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>
- ▶ <https://www.scribbr.com/statistics/skewness/>
- ▶ **Agresti, A. (2013). Categorical data analysis (3rd ed.). Hoboken, NJ: Wiley.**
- ▶ Van Horn, L., et al. (2022). Data splitting for stroke incidence prediction: A review and recommendations. *Frontiers in Neurology*, 13, 1008.
- ▶ Hosmer, D. W., Jr., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). Wiley.
- ▶ Breiman, L., Friedman, J. H., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- ▶ [1] Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning*, 2(1), 1-127.
- ▶ **"Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models"** by Molnar et al. (2021)
- ▶ **"Feature Importance in Machine Learning"** by Baeldung (2022)
- ▶ **"A Review of Feature Importance Measures for Machine Learning Models"** by Guyon et al. (2010)
- ▶ **"Explanatory Model Analysis"** by Lundberg and Lee (2020)
- ▶ **"Python Pandas - Descriptive Statistics"** from <https://www.geeksforgeeks.org/python-pandas-dataframe-describe-method/>
- ▶ "How to get summary of a dataset in python" from Real Python: <https://learnpython.com/blog/how-to-summarize-data-in-python/>
- ▶ <https://www.d.umn.edu/~rlloyd/MySite/Stats/Ch%2013.pdf>
- ▶ <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
- ▶ <https://www.nia.nih.gov/health/heart-health>
- ▶ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html>
- ▶ 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.

QUESTIONS?



