STROKE PATIENT HEALTHCARE

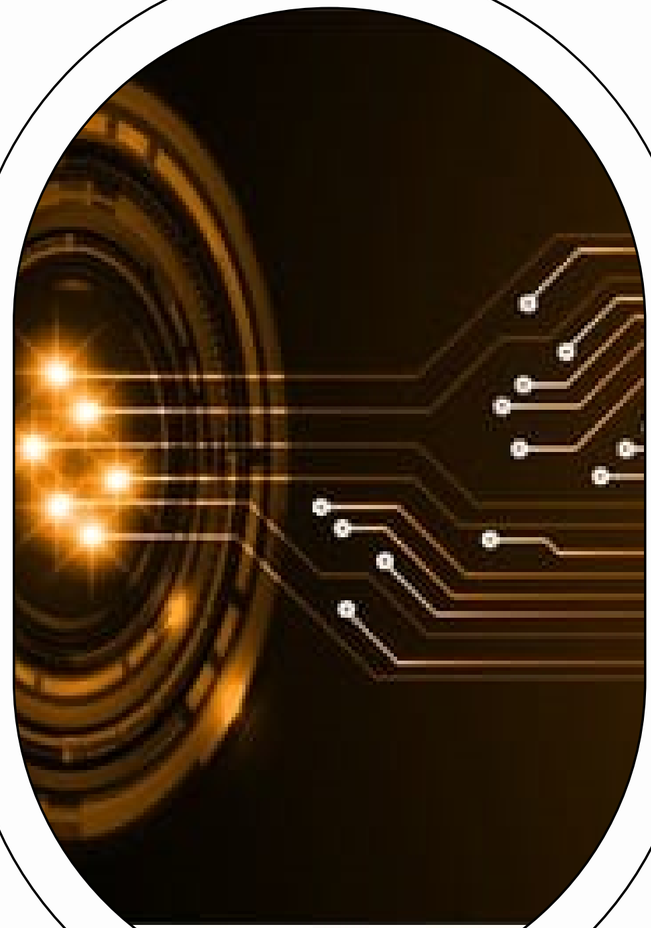
Using Deep Learning

CONTENTS

Abstract

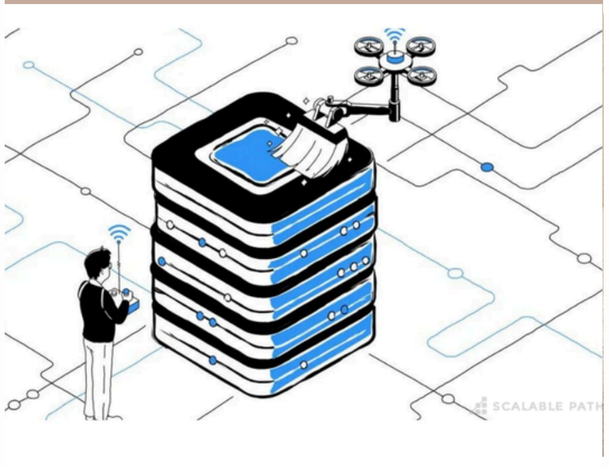
This project leverages deep learning techniques to develop a predictive model for stroke risk assessment based on healthcare data. By analyzing key patient parameters such as age, hypertension, heart disease, and lifestyle factors, the model identifies patterns and correlations indicative of stroke risk.

The project involves multiple stages, including data collection, preprocessing, visualization, and the application of advanced encoding methods to handle categorical data. Key metrics such as precision, recall, and F1-score were used to assess model performance.
The results demonstrate the potential of deep learning to assist healthcare providers in making data-driven decisions for early stroke detection.



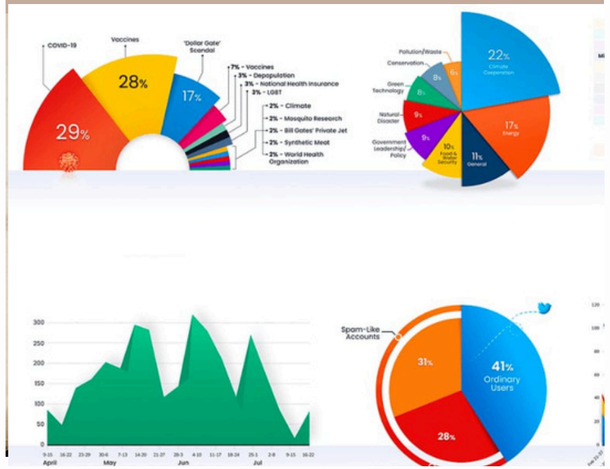
Project Overview

MILESTONE 1



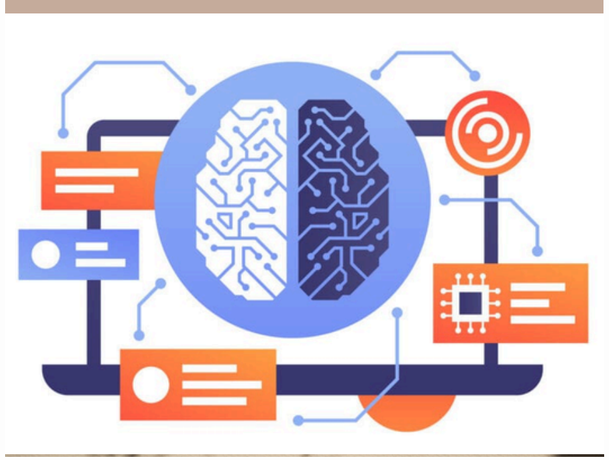
Data Exploration and Pre Processing.

MILESTONE 2



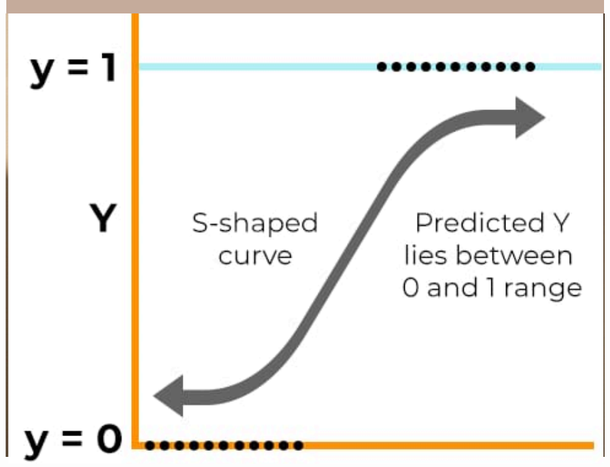
Data Visualization and Data Encoding

MILESTONE 3



Machine Learning Models

MILESTONE 4



High Accuracy Model

MILESTONE 1

Data Exploration & Pre processing

ANALYSIS	DEFINITION	OBSERVATION
df.shape()	Returns a tuple representing the number of rows and columns in the dataset.	The dataset contains 5110 rows and 12 columns, indicating sufficient data for analysis and modeling.
df.info()	Displays a concise summary of the DataFrame, including non-null counts and data types.	Identified bmi as the only column with missing values. Other columns are complete, with data types categorized as int, float, or object.
df.describe()	Provides statistical summary (mean, std, min, max, etc.) for numerical columns.	Key numerical features like age, avg_glucose_level, and bmi were analyzed. The values show a wide range, with notable outliers in glucose levels and BMI.

ANALYSIS	DEFINITION	OBSERVATION
df.describe (include=object)	Provides statistical summary for categorical columns (e.g., count, unique values, top category,frequency).	The dataset has higher representation of females (2,994 occurrences) and married individuals (3,353 occurrences).Private employment is the most common work type, urban areas are slightly more prevalent (2,596 occurrences), and the majority of participants have never smoked(1,892 occurrences).
df.smoking_status.unique()	Displays unique values in the smoking_status column.	The unique values are: ['formerly smoked', 'never smoked', 'smokes','Unknown']. These categories will need encoding for model compatibility.
df.isnull().sum()	Counts the number of missing values for each column.	Counts the number of missing valuesfor each column.

MILESTONE 2

Data Visualization & Data Encoding

VISUALIZATION TECHNIQUES

Bar Charts and Line Graphs: Compare disease incidence or trends over time.

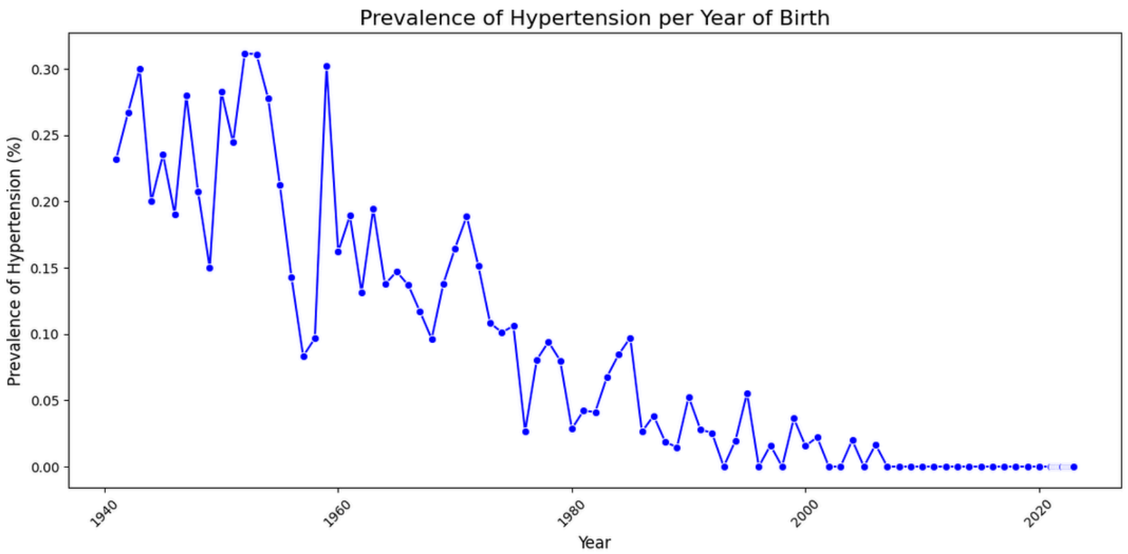
Heatmaps: Show geographical distribution of diseases or resource allocation.

Pie Charts: Display proportional breakdowns, such as causes of hospital admissions.

Scatter Plots: Highlight correlations, e.g., between cholesterol levels and stroke risk.

Histograms: Examine distributions, such as patient age or blood pressure levels.

Prevalence rates of hypertension, heart disease, stroke, and other health conditions changed across different age groups or genders



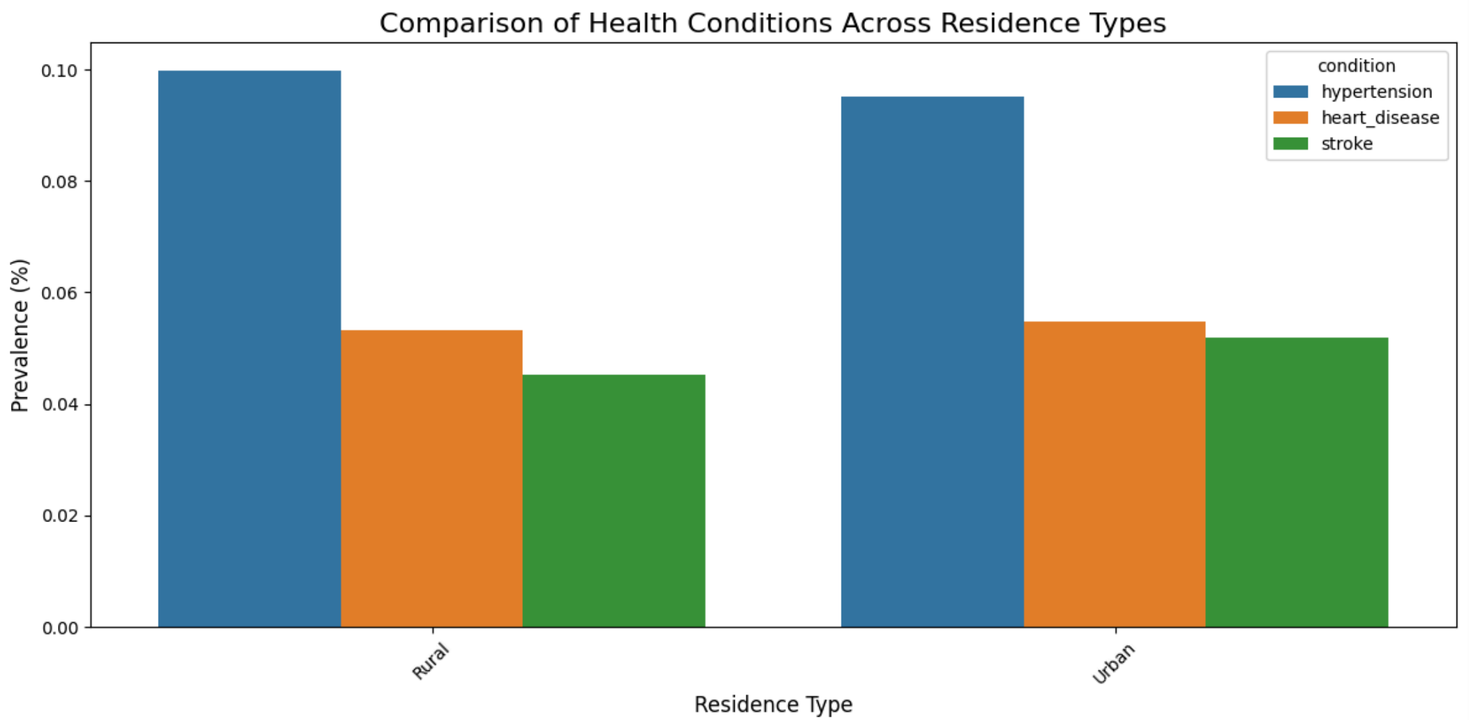
OBSERVATIONS

1.The graph shows the prevalence of hypertension over different years of birth.

2.People born earlier (e.g., pre-1990) show higher rates of hypertension, likely due to age and long-term exposure to lifestyle factors.

3.By limiting the x-axis range from 1990 to 2021, we focus on recent cohorts, helping to zoom into changes in health patterns in the current population.

Comparison of the prevalence of health conditions (e.g., hypertension, heart disease, stroke) across different region

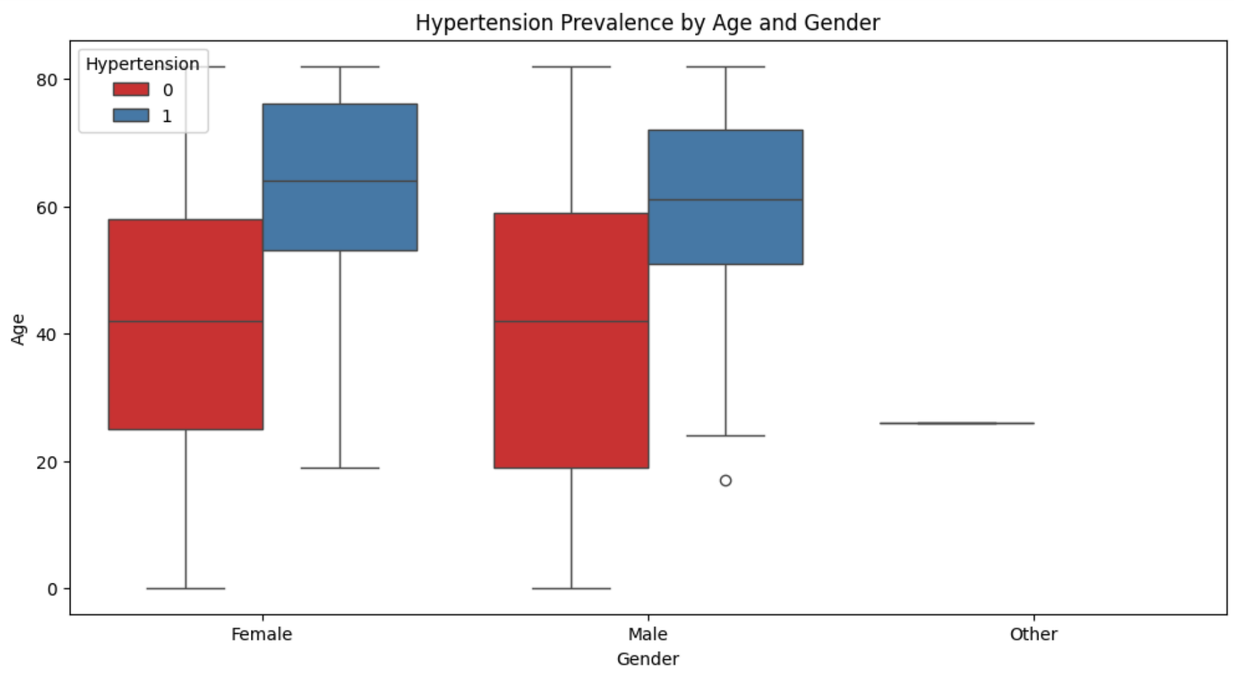


OBSERVATIONS

1. *Urban areas may show higher prevalence of heart disease, which is often linked to sedentary lifestyles, unhealthy diets, and higher pollution levels. The availability of healthcare services in urban areas might lead to better diagnosis, but lifestyle factors could contribute to higher rates.*

2. *Rural areas may show higher prevalence of stroke and hypertension. Limited access to healthcare, fewer health awareness programs, and lifestyle factors (such as poor diet and less physical activity) may lead to these higher rates. Lack of early diagnosis and treatment may contribute to the higher prevalence of stroke in rural populations.*

Prevalence of health conditions (hypertension, heart disease, stroke) vary based on age or BMI, and how does it differ between males and females in the healthcare dataset



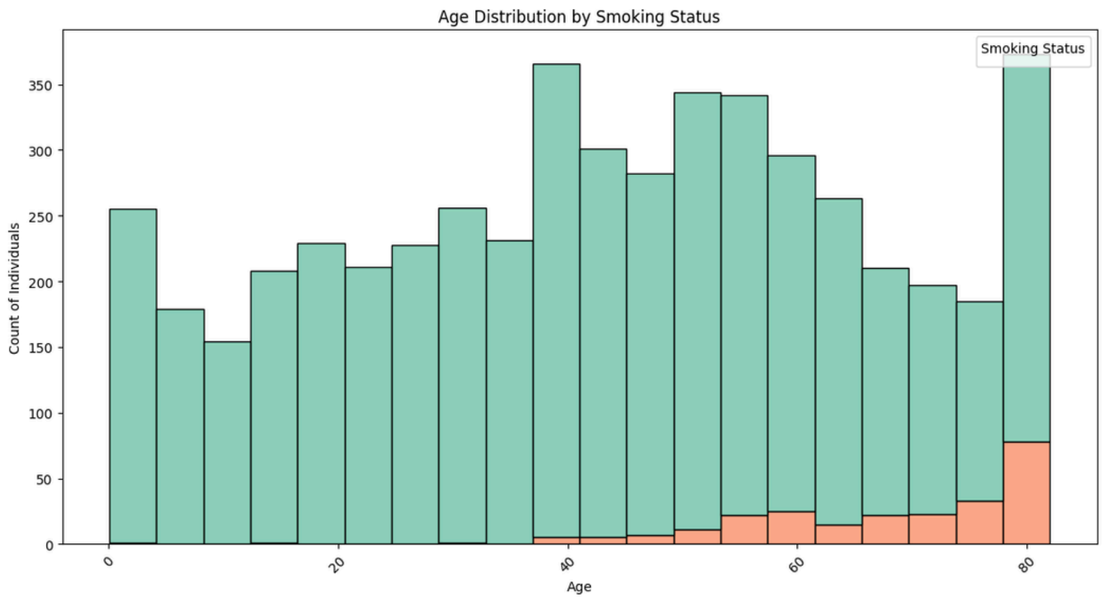
OBSERVATIONS

1. ***The female group (hue = "gender") shows a higher concentration of individuals with hypertension in the older age ranges. The box for hypertensive women likely starts rising as age increases (post-menopausal women are more prone to hypertension).***

2. ***For non-hypertensive women, the age distribution could show a broader range, with many younger females having no hypertension.***

3. ***The median age for hypertensive females may be slightly lower than for hypertensive males, suggesting that hypertension might develop earlier in females (possibly due to hormonal factors or menopause)***

How does the prevalence of hypertension vary across different age groups and smoking status in the healthcare dataset



OBSERVATIONS

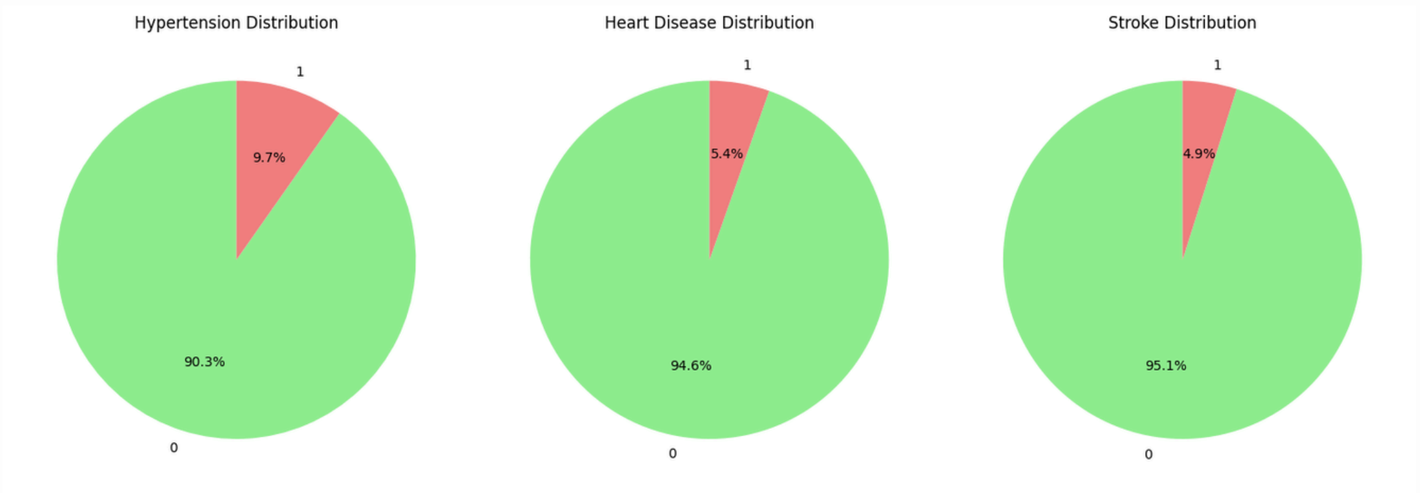
1. *As we move to older age groups (e.g., 60-69, 70-79, and 80+), the number of hypertensive individuals increases, indicating that age is a strong risk factor for developing hypertension.*

2. *The younger age groups (e.g., 18-29, 30-39) show much lower counts of hypertension, which is expected since hypertension tends to be less common in younger populations.*

3. *Among all age groups, smokers and formerly smoked individuals tend to have a higher count of hypertensive individuals compared to those who never smoked. This suggests that smoking is a significant contributor to the development of hypertension.*

4. *In younger age groups (18-29), although the overall number of hypertensive individuals is low, smokers still show a higher count compared to non-smokers, indicating that smoking may increase the risk of hypertension even at a younger age.*

Growing focus on treating certain health conditions (e.g., hypertension, heart disease, stroke) over the years,



OBSERVATIONS

1.Prevalence of Hypertension: The pie chart will show what proportion of individuals in the dataset are hypertensive (those with hypertension = 1) versus those who are not (those with hypertension = 0).

2.If the portion with hypertension is large, it indicates that a significant number of individuals in the dataset are suffering from high blood pressure, which is a common health condition, especially with advancing age.

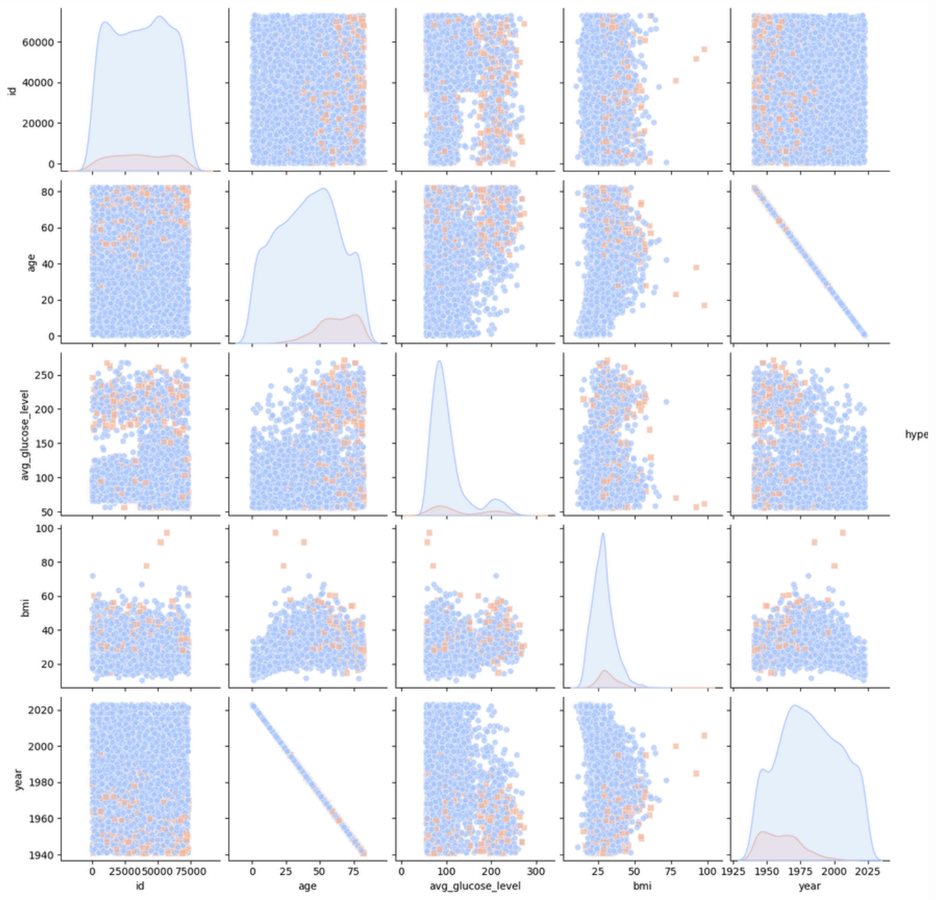
3.If the portion without hypertension is large, this suggests that most individuals in the dataset are either normotensive (normal blood pressure) or not diagnosed with hypertension.

Features age, BMI, and average glucose level distributed across individuals with and without hypertension, and are there any noticeable correlations between these variables that may suggest risk factors for hypertension

OBSERVATIONS

1.*In the scatter plots between age and other features (like bmi and avg_glucose_level), you may notice that individuals with hypertension tend to be older. Specifically, hypertensive individuals could be more concentrated in the higher age ranges, suggesting that age may be a risk factor for hypertension.*

2.*The KDE plot on the diagonal for age might show a higher density of older individuals among those with hypertension, implying a correlation between age and the likelihood of having hypertension.*



Data Encoding

Data encoding is the process of converting categorical data into numerical representations so it can be used effectively by machine learning models.

Residence_type	0-Rural 1-Urban
Work_type	work_type_Never_worked work_type_Private work_type_Self-employed
Smoking_status	smoking_status_never smoked smoking_status_formerly smoked smoking_status_smokes

MILESTONE 3

Machine Learning Models

Machine Learning (ML) models are algorithms that enable computers to learn patterns from data and make predictions or decisions without being explicitly programmed. Common models are :

- Linear Regression
- Ridge Regression
- Logistic Regression
- Lasso Regression

Pre processing & Data Preparation for Regression Models

Handling Missing Values

Encoding Categorical Variables

Feature Selection

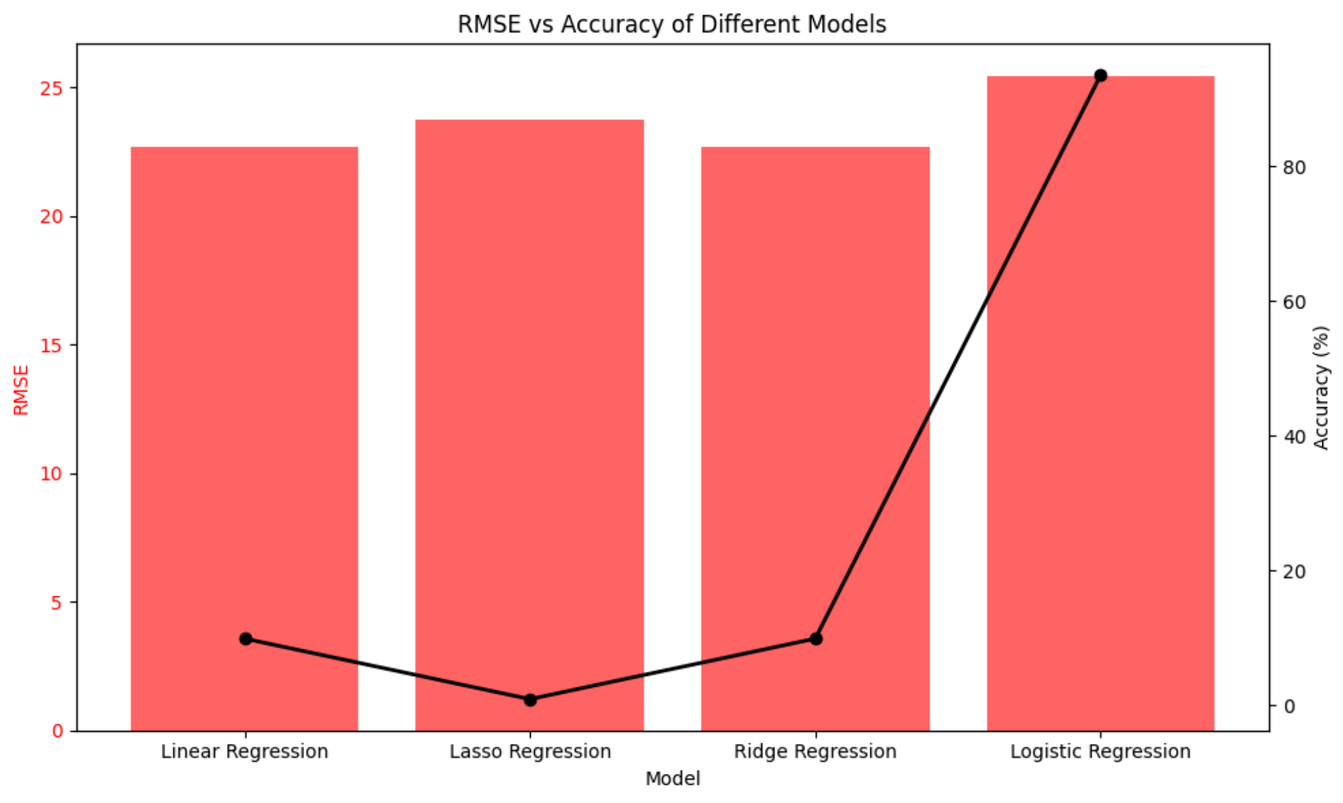
Standardization

Train-Test Split

Model Performance : Accuracy & RMSE

MODELS	ACCURACY	RMSE
LOGISTIC REGRESSION	**93.54**	**25.41**
LINEAR REGRESSION	**0.09**	**22.65**
RIDGE REGRESSION	**0.09**	**22.65**
LASSO REGRESSION	**0.009**	**23.76**

Visualize Models



OBSERVATIONS

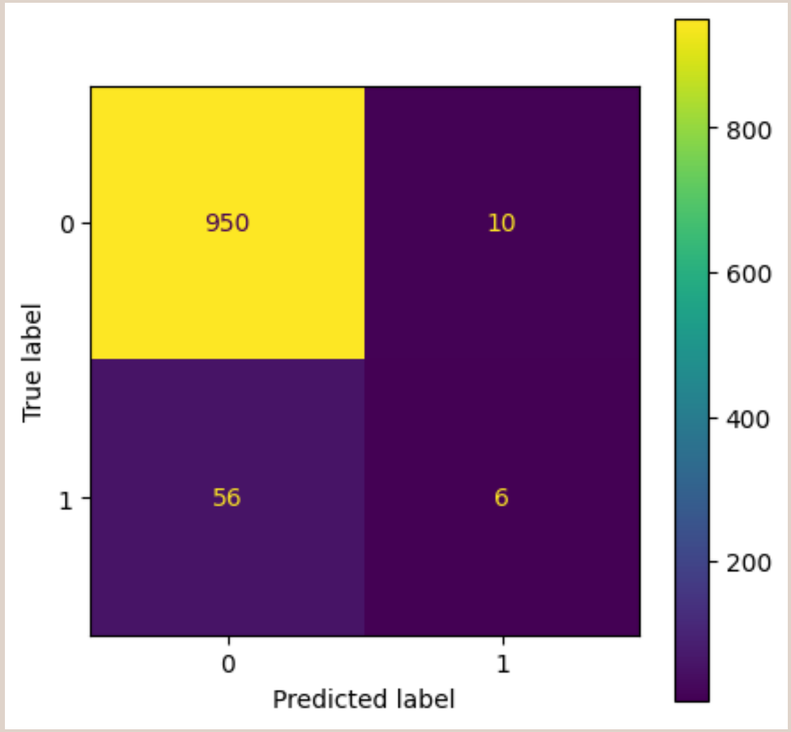
Linear Regression will likely strike a balance with decent accuracy and low RMSE, making it suitable for regression tasks.

Lasso and Ridge Regression might show slightly higher RMSE but are useful when overfitting is a concern.

Logistic Regression's high accuracy reflects its suitability for classification tasks, though its RMSE is misleading because it isn't designed for continuous output.

MILESTONE 4

Precision, Recall, F1 Score, and Accuracy



Precision (P): Measures how many predicted "Stroke" cases are correct.

P= TP/[TP+FP]

R**ecall (R):** Measures how many actual "Stroke" cases are correctly identified.

R=TP/[TP+FN]

F1 Score: Harmonic mean of Precision and Recall, providing a balance between both.

F1 Score= 2*(P x R/P+R)

Accuracy: Overall correctness of the predictions.

Accuracy=(TP+TN)/(TP+TN+FP+FN)

OBSERVATION OF LOGISTIC REGRESSION MODEL

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
LOGISTIC REGRESSION	**93.54**	**25.41**	**0.096**	**0.153**

High accuracy but very low recall and F1 scores, indicating poor minority class detection. Choosing a logistic regression model based on its high accuracy score is a logical decision.

Is accuracy a factor for model performance?

Accuracy alone is not a reliable metric for evaluating model performance, especially in cases of class imbalance. While high accuracy values (e.g., Logistic Regression: 96.13%, Linear Regression: 96.02%) might seem impressive, they can be misleading when the minority class is poorly identified.

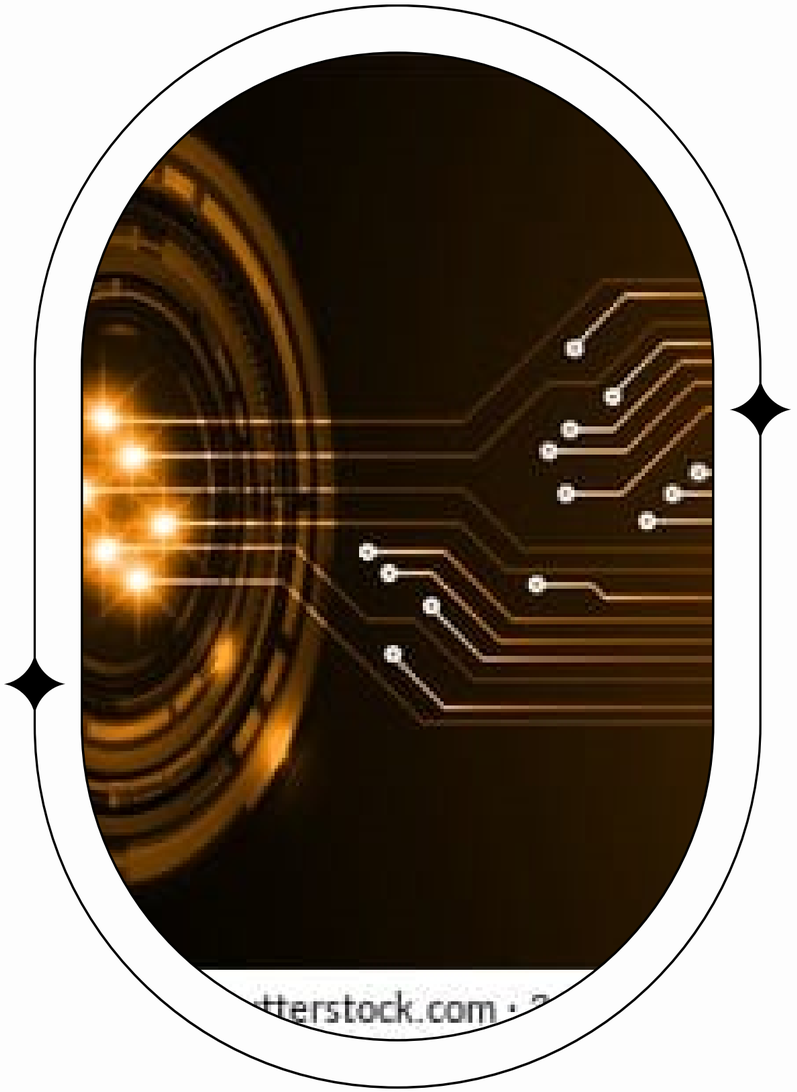
Is the dataset biased?

The data set is the stroke patient's dataset

- ***Class 0 (No Stroke): 95.74%***
- ***Class 1 (Stroke): 4.26%***

This severe imbalance skews the model toward predicting the majority class, leading to poor detection of the minority class.

- ***Steps to mitigate bias:***
  - ***Changing random state***
  - ***Threshold tuning***
  - ***Adjusting class weights***
  - ***Resampling***



Logistic Model Evaluation

Strengths:

● High Recall : A recall of 0.9677 means that the model correctly identifies 96.77% of the actual positive instances (e.g., correctly identifying people who have had a stroke)

● Good Accuracy : Performs well overall in correct predictions.

Weaknesses:

● Low Precision : A precision of 0.375 means that only 37.5% of the instances predicted as positive (e.g., stroke in this case) are correct.
There are a high number of False Positives (FP) relative to True Positives (TP).

● low F1 Score : the low precision (37.5%) severely affects the F1 score. This indicates that the model is identifying many true positive cases (high recall), but it is also making a significant number of false positives (low precision), which drags down the F1 score.

Conclusion:

Ridge Regression excels in recall but requires improvement in precision to balance performance

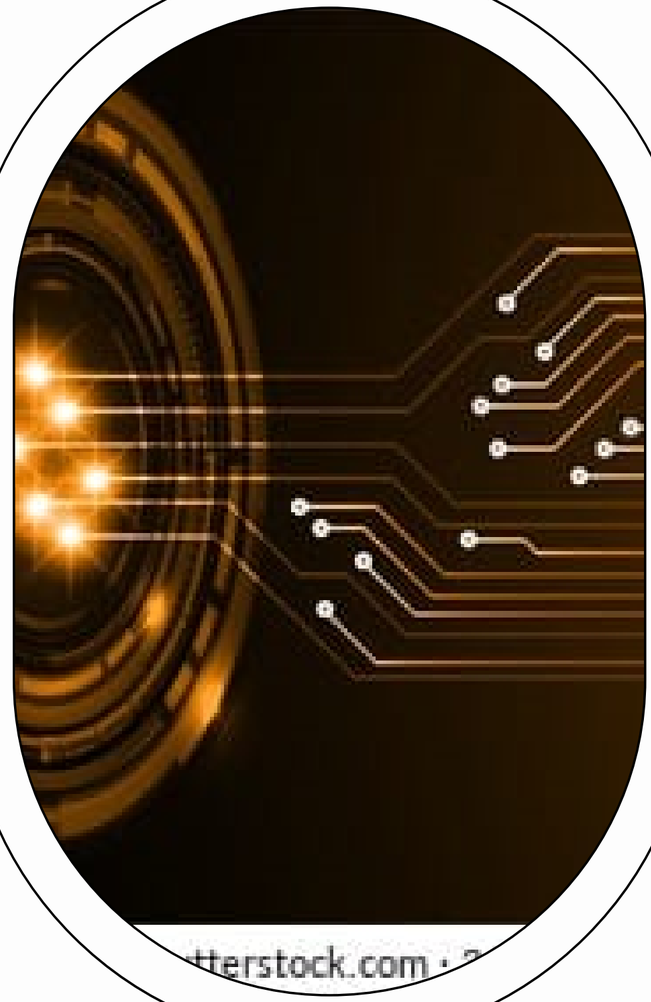
CONCLUSION

Class Imbalance: *Dataset is heavily skewed (95% "No Stroke", 5% "Stroke"), causing bias towards the majority class.*

Impact on Performance*: High accuracy but poor detection of strokes due to the imbalance.*

Evaluation Metrics*: Metrics like recall and F1-score are more reliable than accuracy for evaluating minority class detection.*

Analysis Conducted*: Data preprocessing, visualizations, and modeling plays significant role in*



THANK YOU

Regards
Harshita Agrawal