An overview of data pre-processing, visualisation and modelling

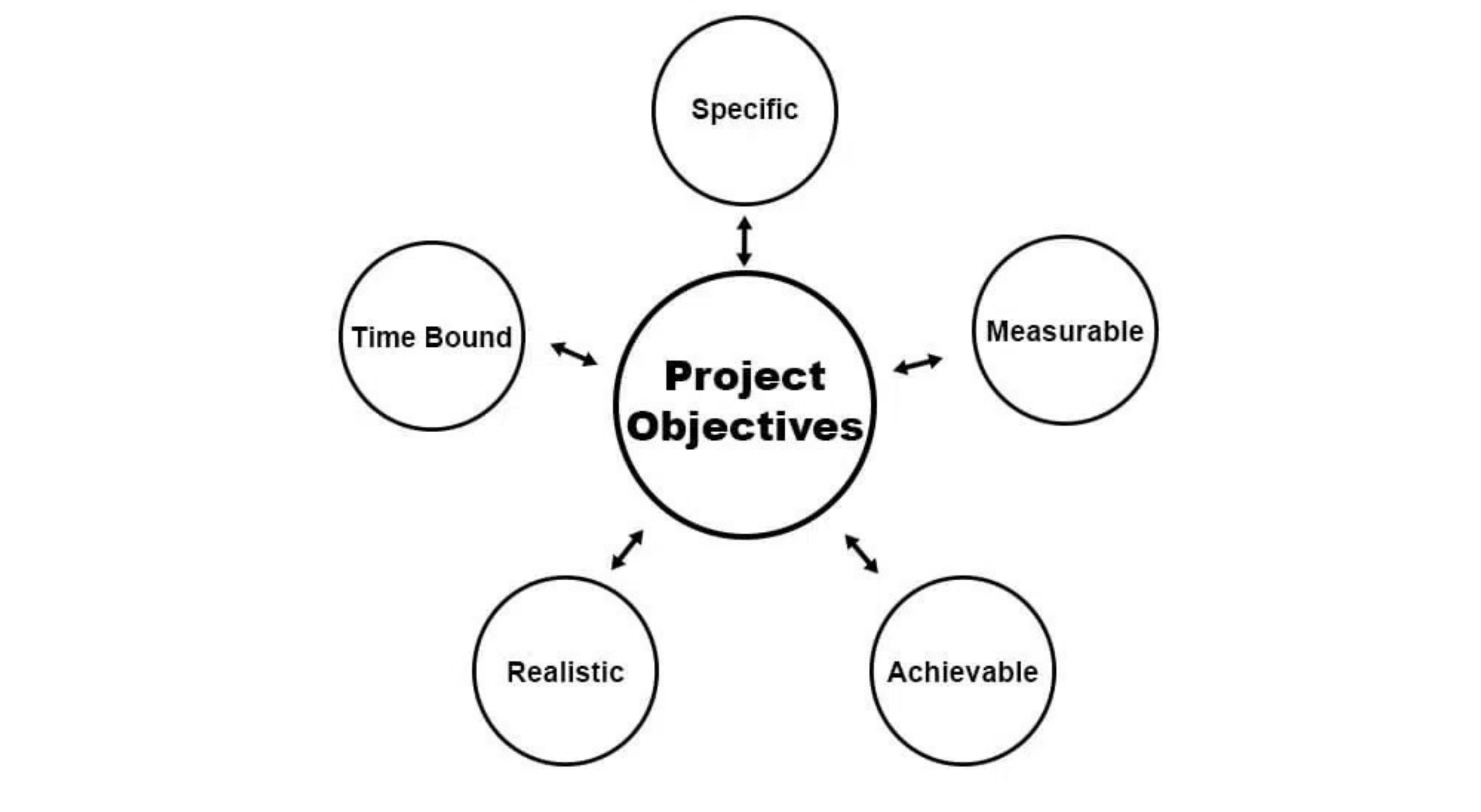
Stroke Patient Healthcare using Deep Learning

Contents

Objective	Provide a concise summary of the project goals and objectives
Project Overview	Outline the structure and scope of the project
Milestone 1	Includes data collection, cleaning, and preprocessing steps
Milestone 2	Data visualization phase retrieving patterns and trends.
Milestone 3	Data encoding techniques and building models to data
Milestone 4	Tuning the models and evaluating various scores
Model Final Evaluation	Present an assessment of the final model’s performance, including strengths and limitations
Final Insights	Summarize the key conclusions drawn from the project and its outcomes

Objective of the project

The goal of this project is to study healthcare data to find patterns and build a model that can predict health risks, like strokes. It uses data analysis, charts, and machine learning to give helpful insights and accurate predictions.



Project Overview

It follows four steps for progression from data exploration to model building

1 Data Pre-Processing

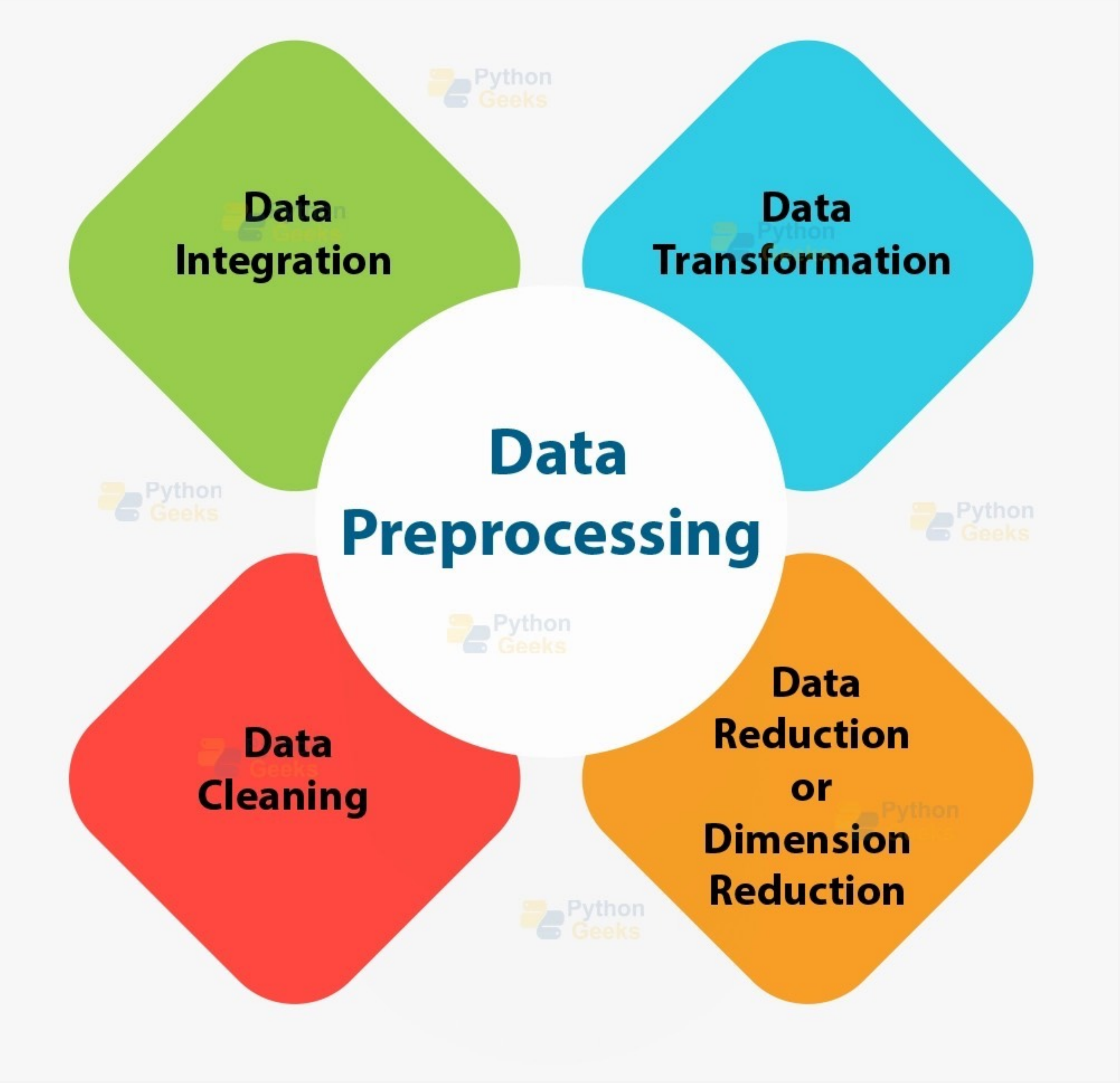
2 Data Visualisation

3 Data Encoding

4 ML models

1

Data Pre-Processing



Dataset Characteristics

- The dataset contains 5110 rows and 12 columns

- We have columns like "gender", "ever_married", "work_type", "residence_type" and "smoking_status" contain string values, which are represented using the "object" datatype in this data frame

- The columns like "age", "avg_glucose_level", "bmi" to be of float datatype

- The columns like "id", "heart_disease", "hypertension", "stroke" to be of int datatype.

Exploratory Analysis Summary

- df.describe(): Key numerical features like age, glucose levels, and BMI are shown in wide range.

- df.describe(include=object): The dataset consists of a majority of females, with 2,994 female participants compared to 2,116 males in terms of gender.

- df.isnull().sum(): The bmi column has 201 missing values, rest of columns have 0 missing values.

- df.ever_married.unique(): array([‘Yes’,’No’],dtype=object), refers to two unique attributes present in ever_married column which are ‘yes’ and ‘no’.

2

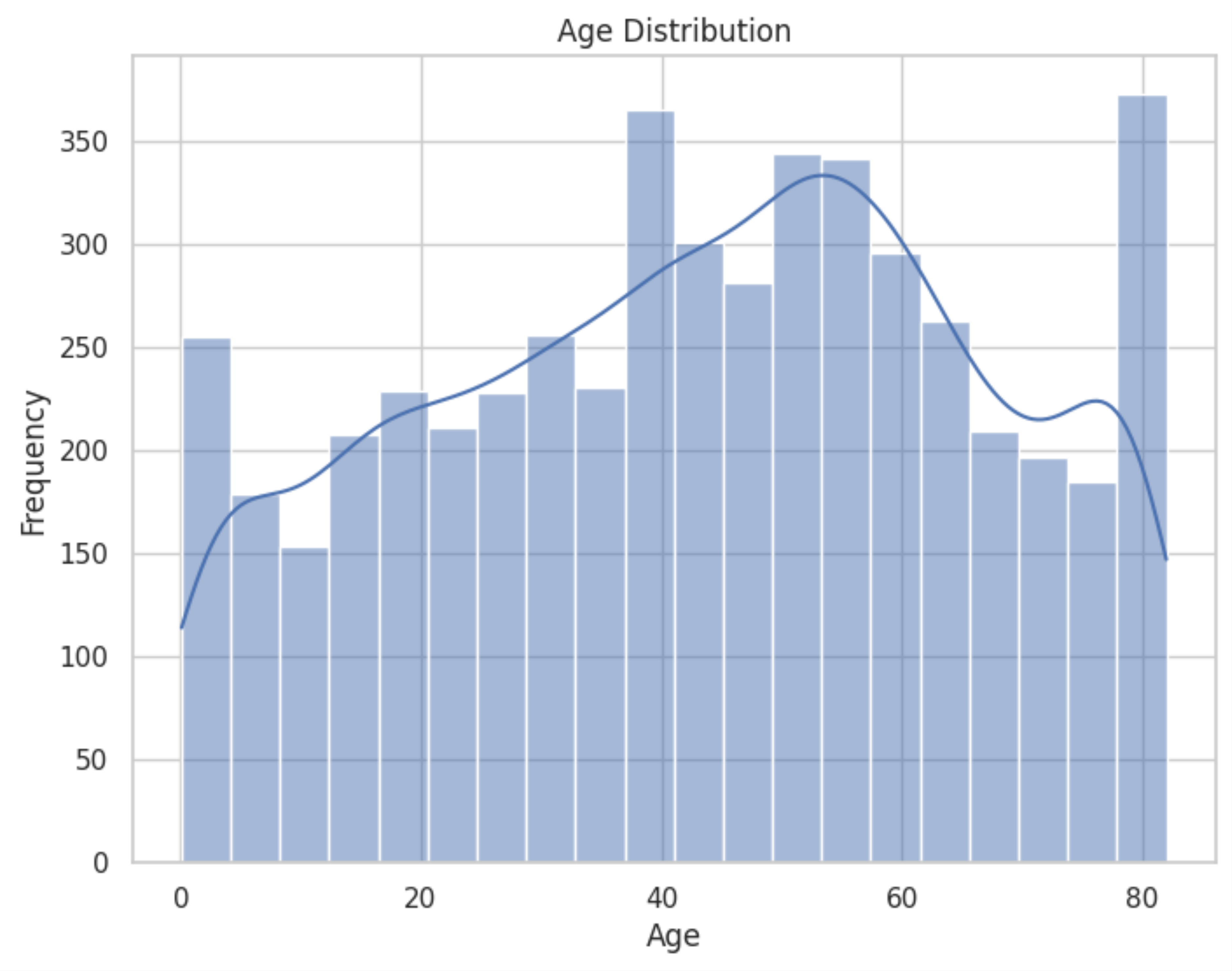
Data Visualisation

Data visualization is the practice of representing data through visuals like charts, graphs, and plots to make complex information easier to understand and analyze. It helps uncover patterns, trends, and outliers in the data, making it a vital tool for data analysis, decision-making, and effectively communicating insights.

Histogram for Age Distribution

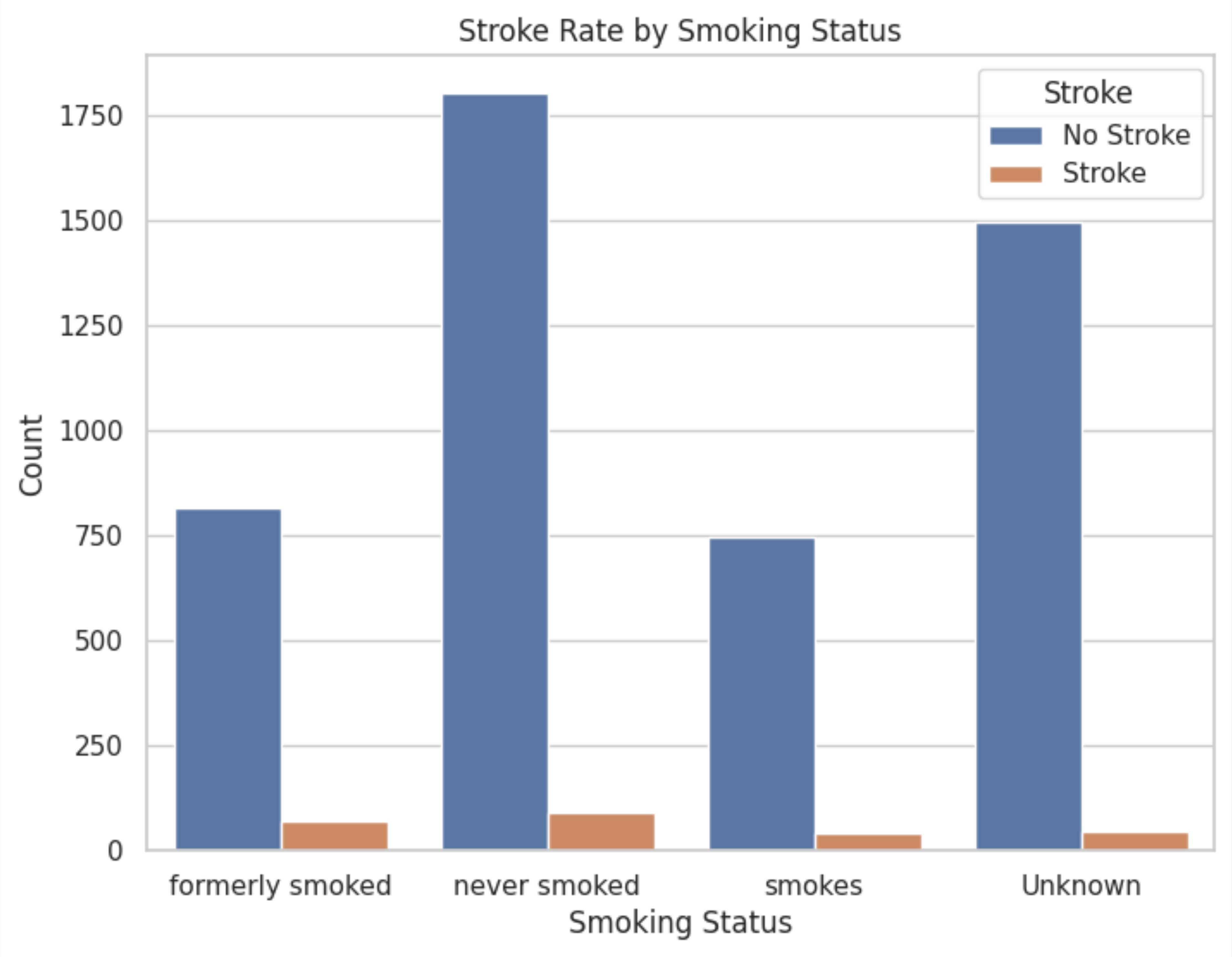
- Most people in the dataset are adults, with a high number of individuals aged between 40 and 80.

- The dataset includes more middle-aged and elderly individuals, which may influence the analysis if age is a significant factor in stroke risk.



Bar char of stroke rates by smoking status

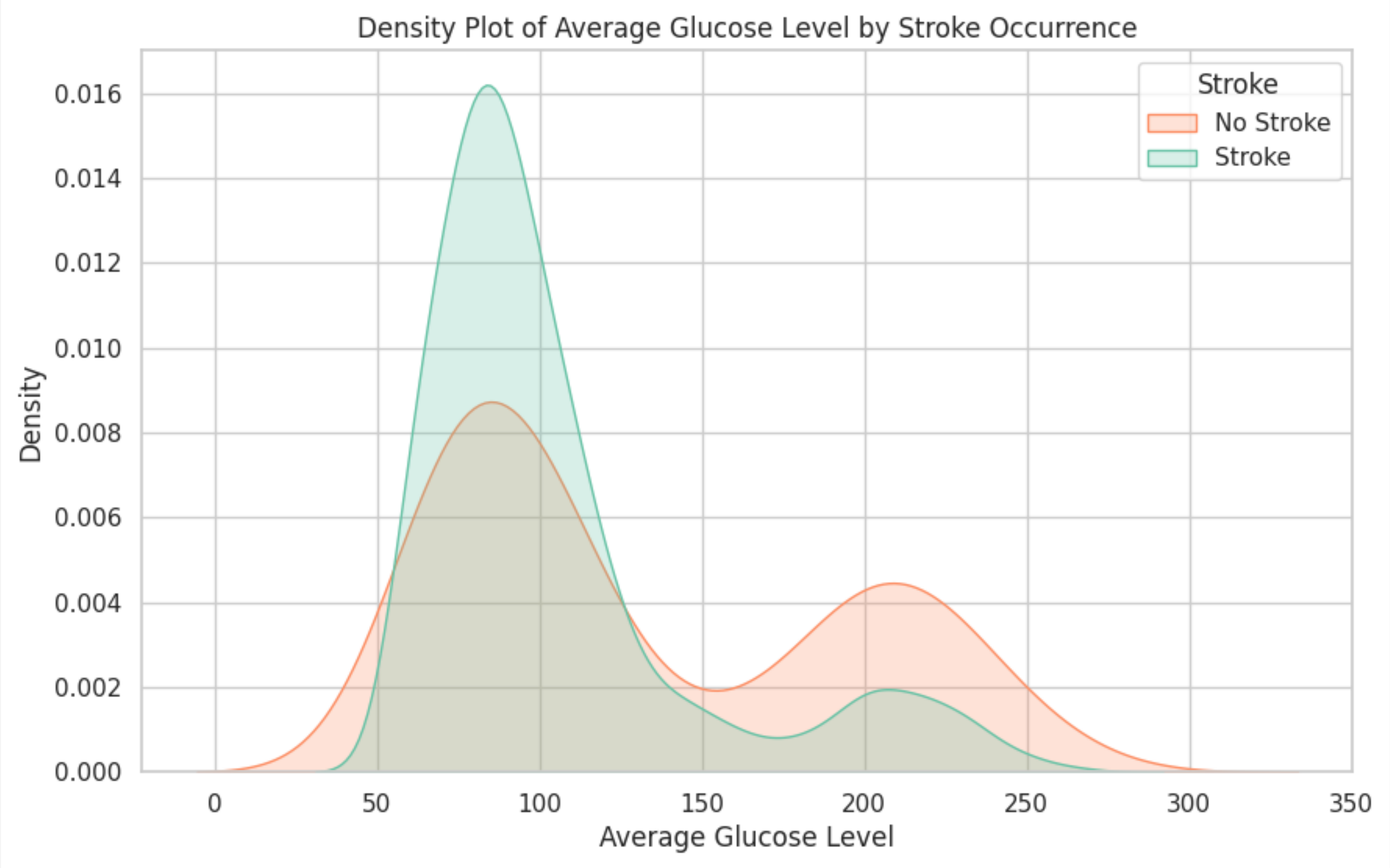
- Individuals who are current or former smokers seem to have a higher stroke rate compared to those who have never smoked.
- Smoking could be a contributing factor to stroke risk, supporting existing health advice that encourages quitting smoking for better health.



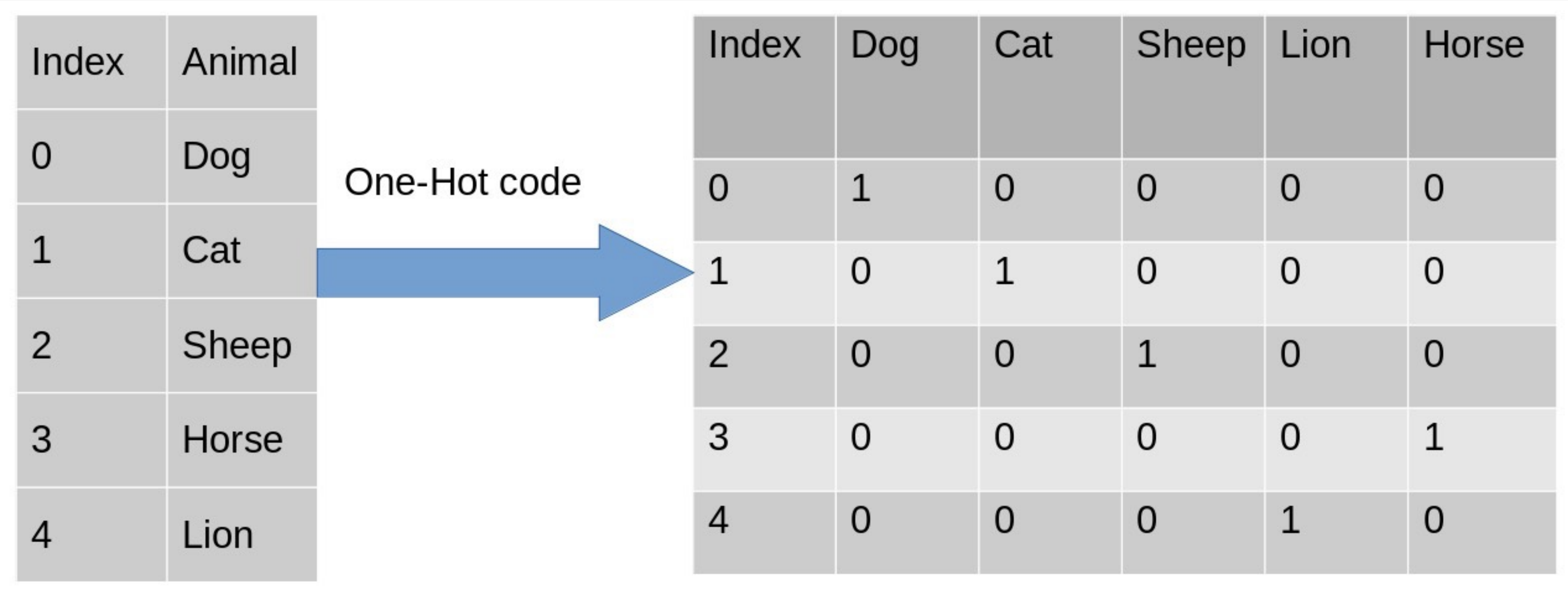
Density plot of avg_glucose_level grouped by stroke occurrence

- **Glucose Levels in Stroke and No-Stroke Groups:** People without a stroke (orange) mostly have lower glucose levels, around 70-100. People with a stroke (green) tend to have slightly higher glucose levels, peaking around 100, with some even reaching 200-250.

- **Higher Glucose Levels and Stroke**: The stroke group has more cases with high glucose levels (above 150). This suggests that higher glucose might be linked to a higher risk of stroke.



3 Data Encoding



Index	Animal
0	Dog
1	Cat
2	Sheep
3	Horse
4	Lion

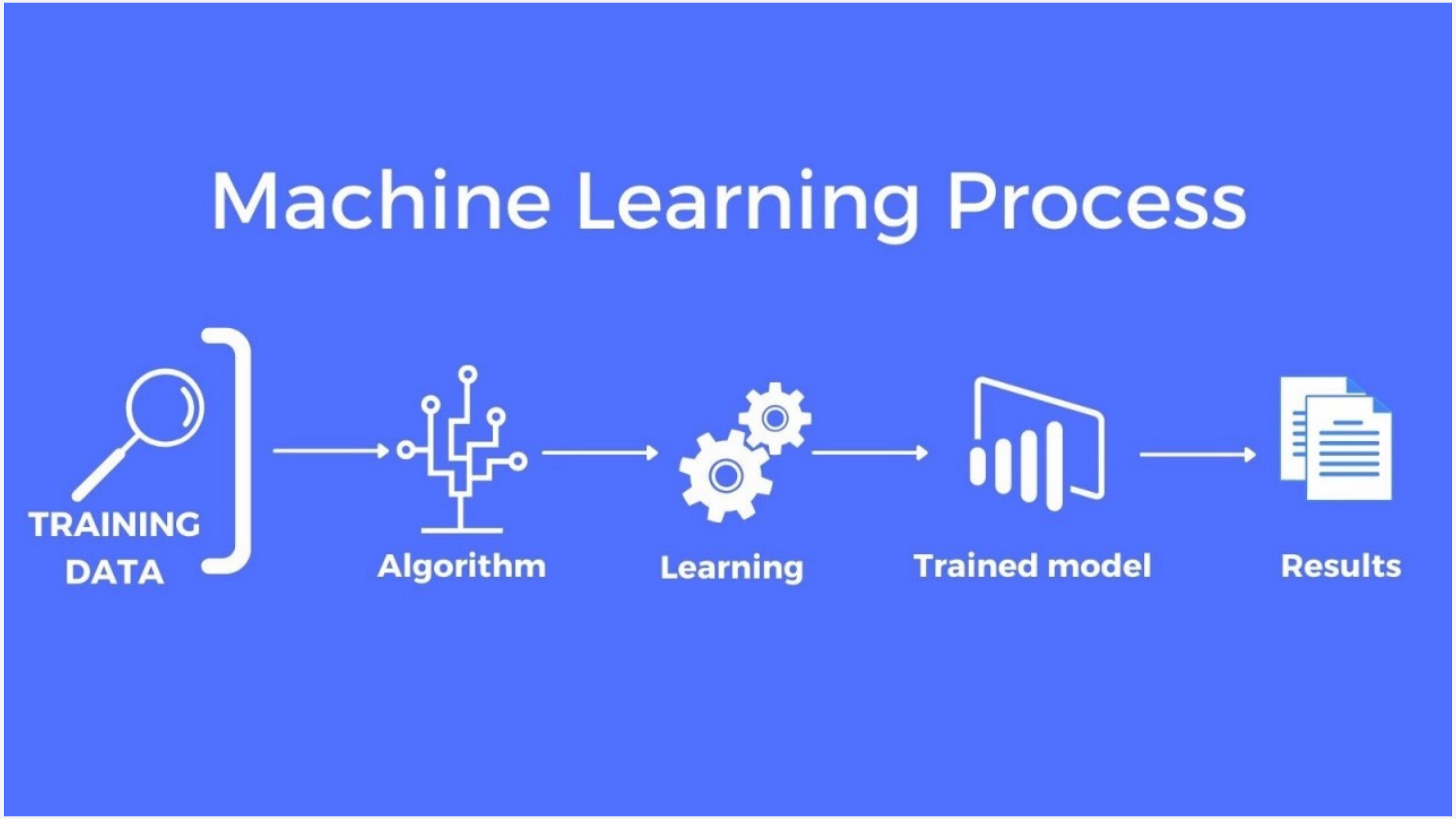
One-Hot code

Index	Dog	Cat	Sheep	Lion	Horse
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	0	1
4	0	0	0	1	0

Data encoding is the process of converting categorical data into numerical formats so that machine learning algorithms can interpret it.

- Residence_type column is converted into 0-Rural and 1-Urban
- Work_type column is converted into work_type_Govt_job, work_type_Never_worked, work_typed_Private, work_type_self-employed
- Smoking_status column is converted into smoking_status_Unknown, smoking_status_formerlysmoked, smoking_status_neversmoked, smoking_status_smoked.
- Gender column is converted to 0-Female, 1-Male

4 Machine Learning Models



Machine Learning

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

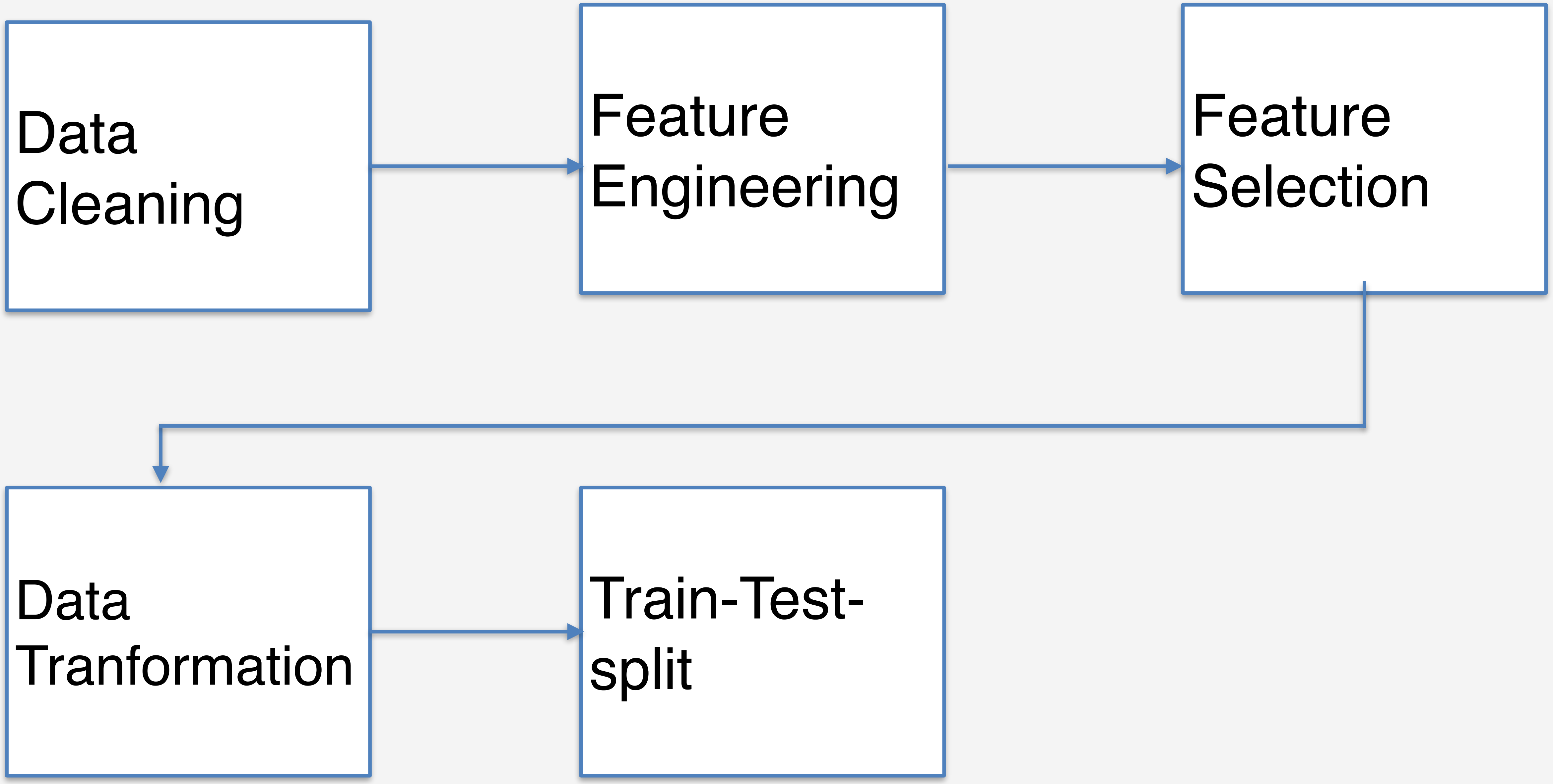
Linear Regression

Lasso Regression

Ridge Regression

Logistic Regression

Preprocessing and Data Preparation for learning models



Model Performance Summary

Linear Regression

It has R^2 score of 22.75 and accuracy of 9.1

Lasso Regression

It has R^2 score of 23.75 and accuracy of 0.94

Model Performance Summary

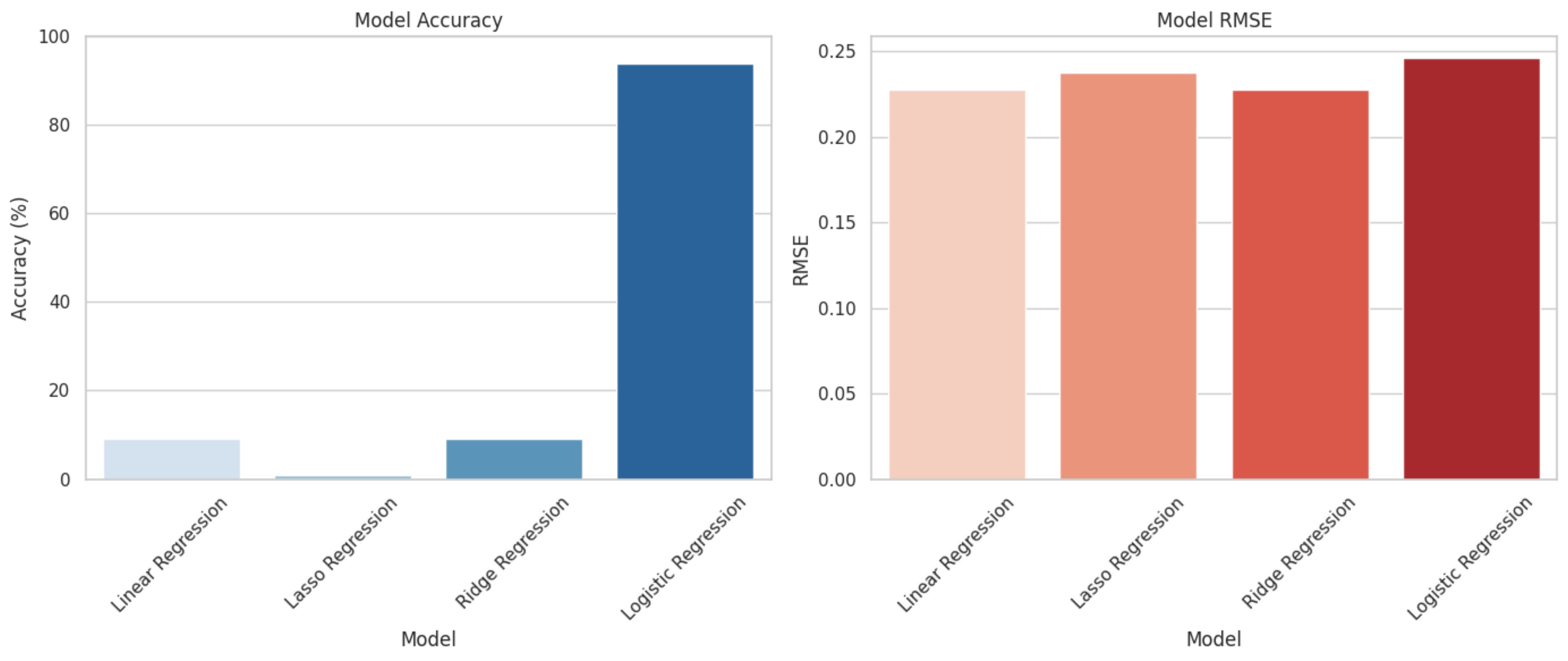
Ridge Regression

It has R^2 score of 22.7594 and accuracy of 9.1002

Logistic Regression

It has R^2 score of 24.63 and accuracy of 93.9

Visualising Different Models



Observations:

Different Models for Different Tasks:

- Logistic Regression is designed for classification problems, where you predict categories (like "yes" or "no"), such as whether someone had a stroke or not.
- Linear Regression, Ridge, and Lasso are designed for predicting continuous values (like age or temperature) and are better suited for regression tasks.

Why Logistic Regression is Better for Classification:

- Logistic Regression works well on binary classification tasks (two outcomes), so it gives high accuracy for problems like predicting stroke vs. no stroke.

Precision, Recall, F1 score & Accuracy

Confusion Matrix

Precision(P)

Measures how many predicted “Stroke” cases are correct.

$$Precision = \frac{TP}{TP + FP}$$

Recall(R)

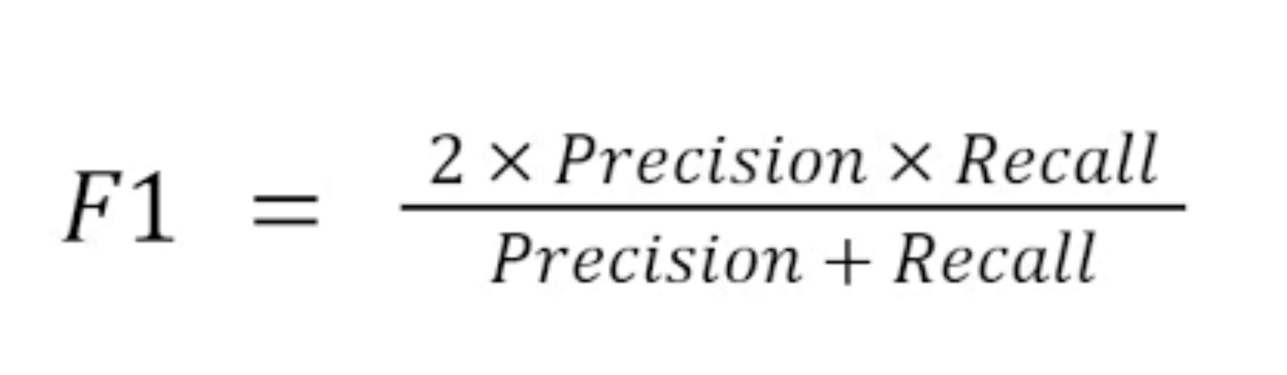
Measures how many actual “Stroke” cases are correctly identified

$$Recall = \frac{TP}{TP + FN}$$

Confusion Matrix

F1 Score

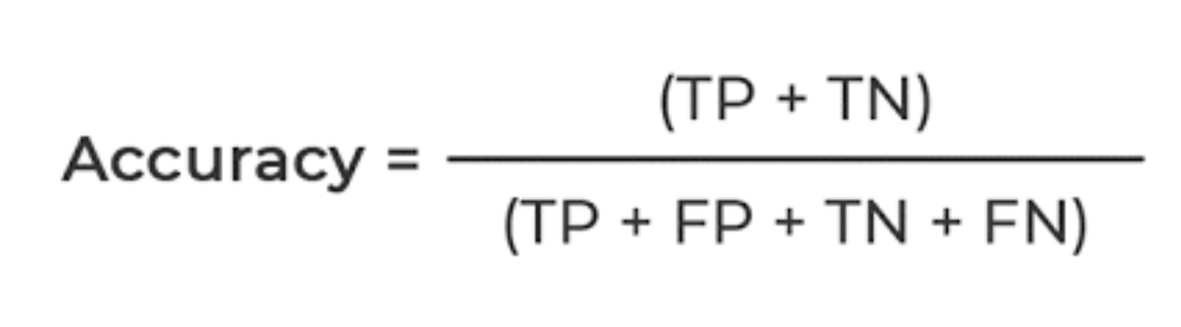
Harmonic mean of Precision and Recall, providing a balance between both



$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Accuracy

Overall correctness of the predictions.

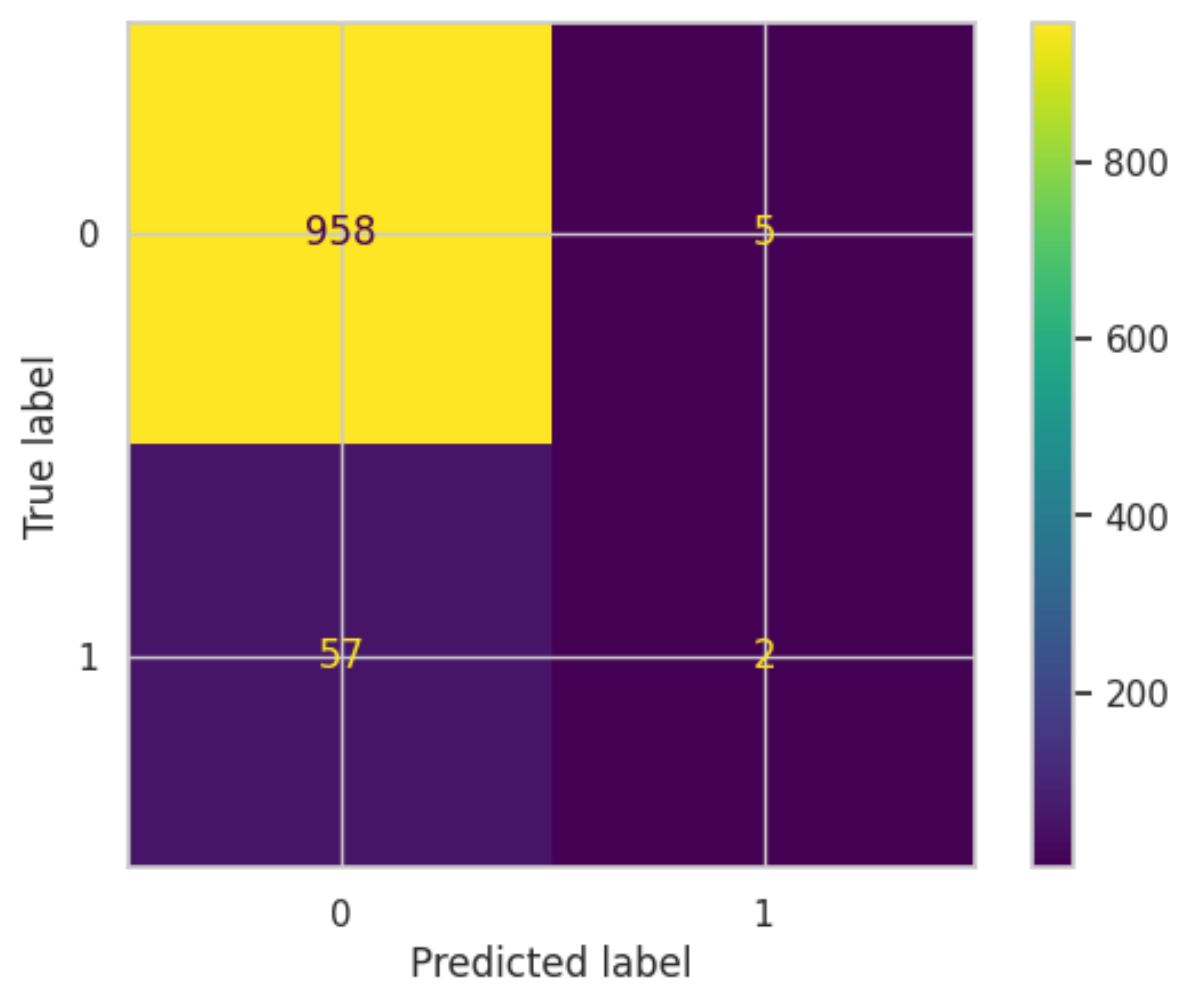


$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{TN} + \text{FN})}$$

Logistic Regression Evaluations

- Precision of 0.28
- Recall of 0.033
- F1 score of 0.060
- Accuracy of 93.93

High accuracy but very low recall and F1 scores, indicating poor minority class detection.



Is the Dataset Biased or not?

Bias in Dataset

Class Imbalance:

- The dataset has a high bias towards non-stroke cases, which could lead to models being overly confident in predicting the majority class.

Data Gaps:

- Missing bmi values could introduce bias if not handled properly.

Is the Dataset Biased or not?

Tackling Bias

Handling Class Imbalance:

- **Undersampling:** Randomly remove samples from the majority class to balance the dataset.

Addressing Missing Data:

- Use imputation strategies (e.g., mean, median, or regression imputation) to fill missing bmi values.

Evaluation Metrics:

- Use metrics like F1-score, precision-recall curves, to assess the model's performance

Final Insights

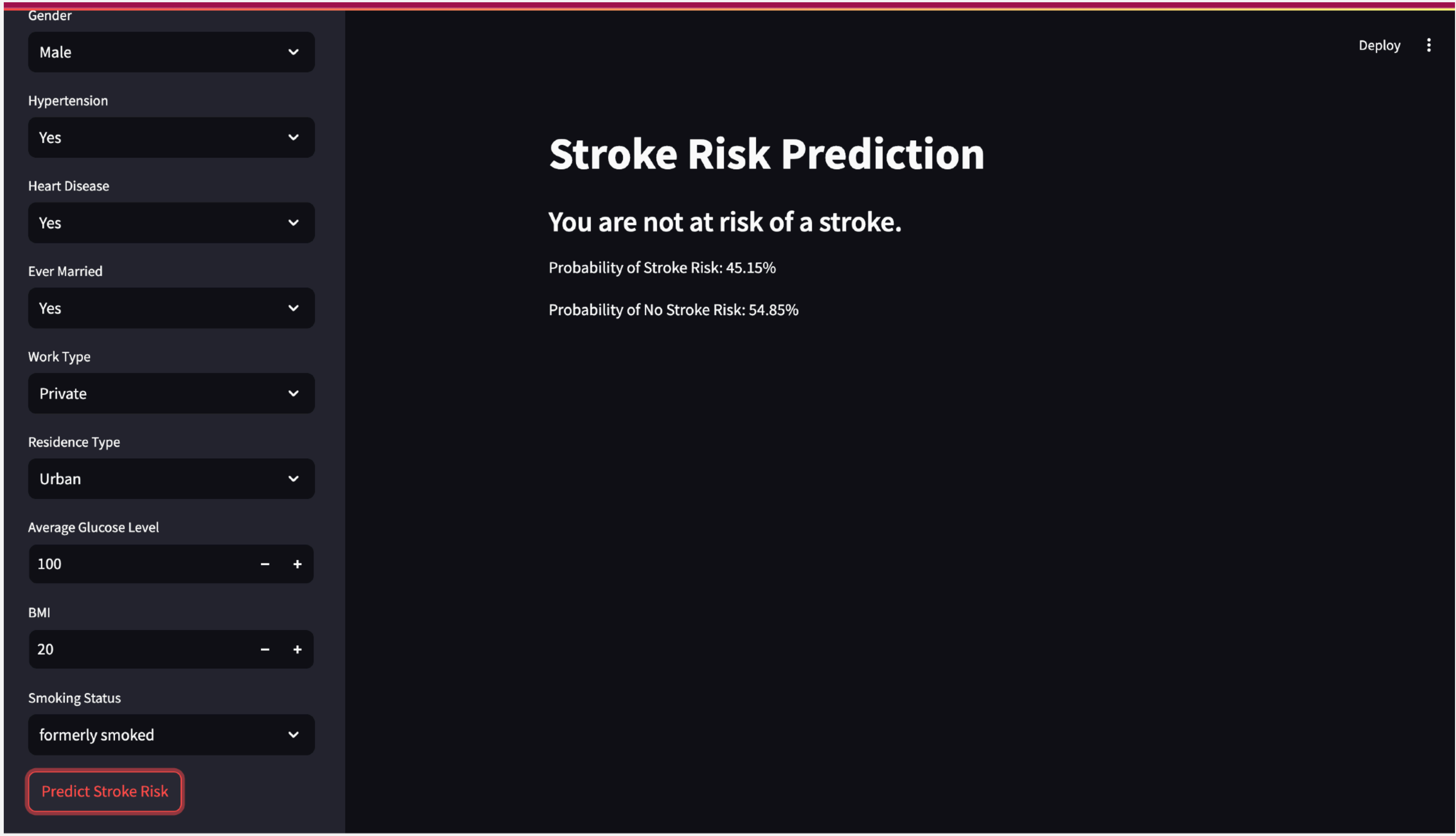
Class Imbalance: Most cases in the dataset are "No Stroke" (95%), with only 5% being "Stroke," causing bias towards the majority class.

Impact on Performance: The model shows high accuracy but struggles to detect strokes effectively due to this imbalance.

Better Metrics: Metrics like recall and F1-score are more useful than accuracy for evaluating stroke detection.

Steps Taken: Data cleaning, visualizations, and modeling helped improve the model's performance.

STREAMLIT DESIGN



Gender: Male

Hypertension: Yes

Heart Disease: Yes

Ever Married: Yes

Work Type: Private

Residence Type: Urban

Average Glucose Level: 100

BMI: 20

Smoking Status: formerly smoked

Stroke Risk Prediction

You are not at risk of a stroke.

Probability of Stroke Risk: 45.15%

Probability of No Stroke Risk: 54.85%

THANK YOU

Tulasiram Nimmagadda