

Computational Intelligence Project

2. Data Preprocessing & Exploratory Data Analysis (Eda)

Summary Report

Overview

The dataset used for this project is a stroke prediction dataset containing 5110 patient records and 12 features including age, hypertension, heart disease, BMI, glucose levels, marital status, working style, residency, and smoking habits. The target variable is Stroke (Yes/No).

Data Insights

- Most stroke cases occur in older age groups, especially above 50 years.
- Patients with hypertension and heart disease show a higher probability of stroke.
- Higher average glucose levels correlate with increased stroke risk.
- Lifestyle factors such as smoking also influence outcomes.
- Gender distribution showed similar stroke risk across males and females.
- The dataset was highly imbalanced – only around 5% patients had a stroke.

These insights helped determine which features are most significant for prediction.

Data Preprocessing Steps

Missing Values Handling

- Missing values in the **BMI** column were replaced using **median** imputation.

Categorical Data Encoding

- Converted categorical features (Gender, Work Type, etc.) into numerical format using **One-Hot Encoding**.

Feature Scaling

- Numerical columns (BMI, Avg Glucose Level) normalized using **Standard Scaling** to improve model performance.

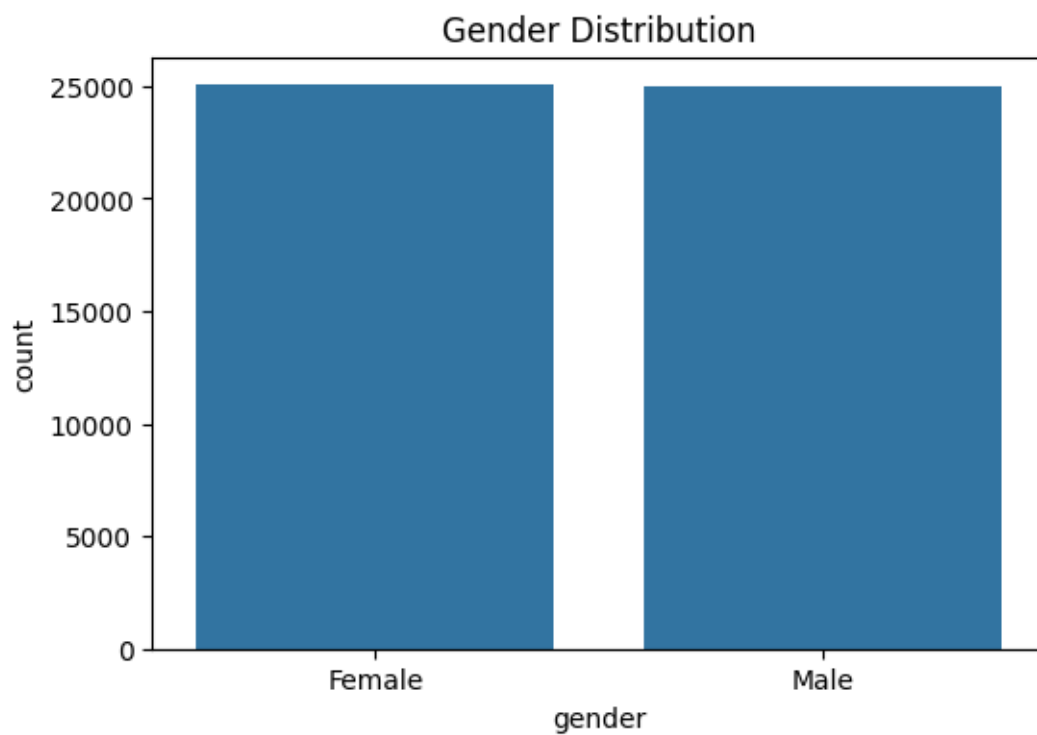
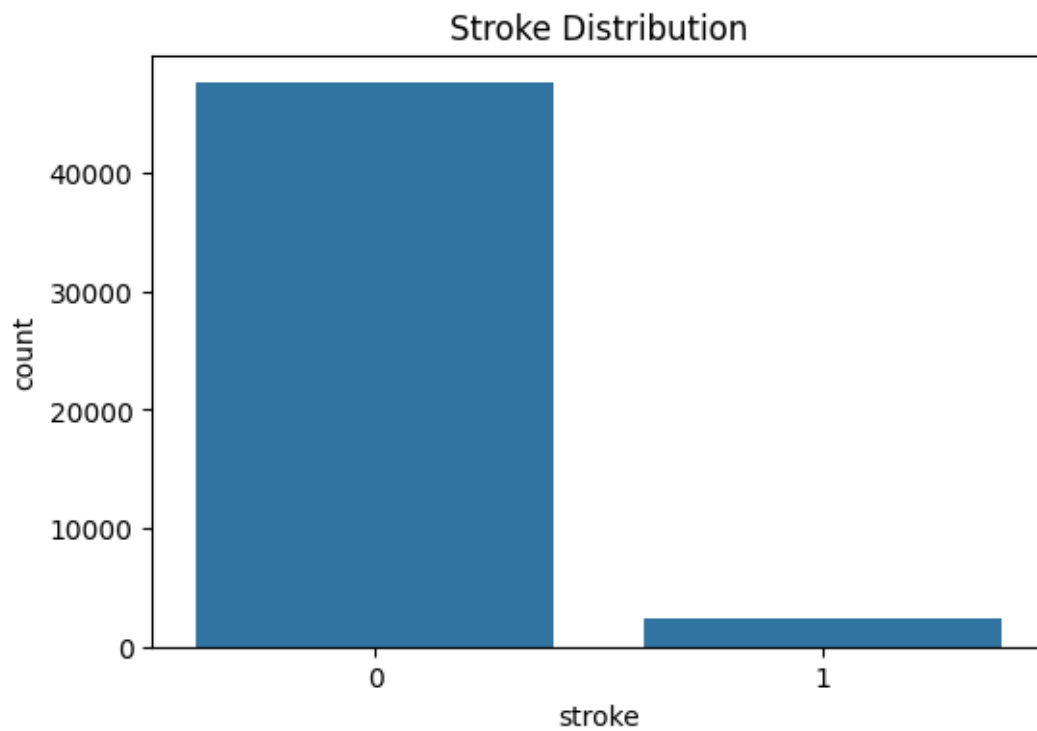
Class Imbalance Correction

- Applied **SMOTE oversampling** to balance the stroke and non-stroke classes for fair learning.

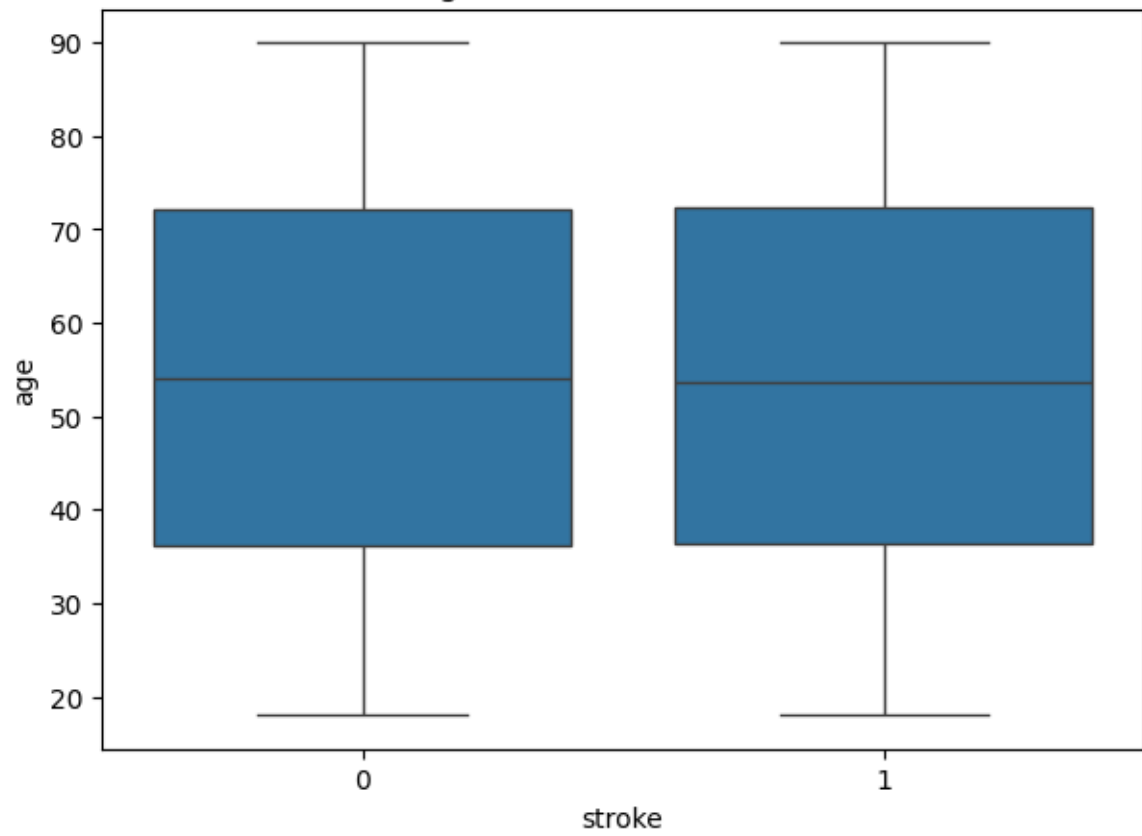
Train-Test Split

- Dataset was split into **80% training** and **20% testing** for unbiased model evaluation.

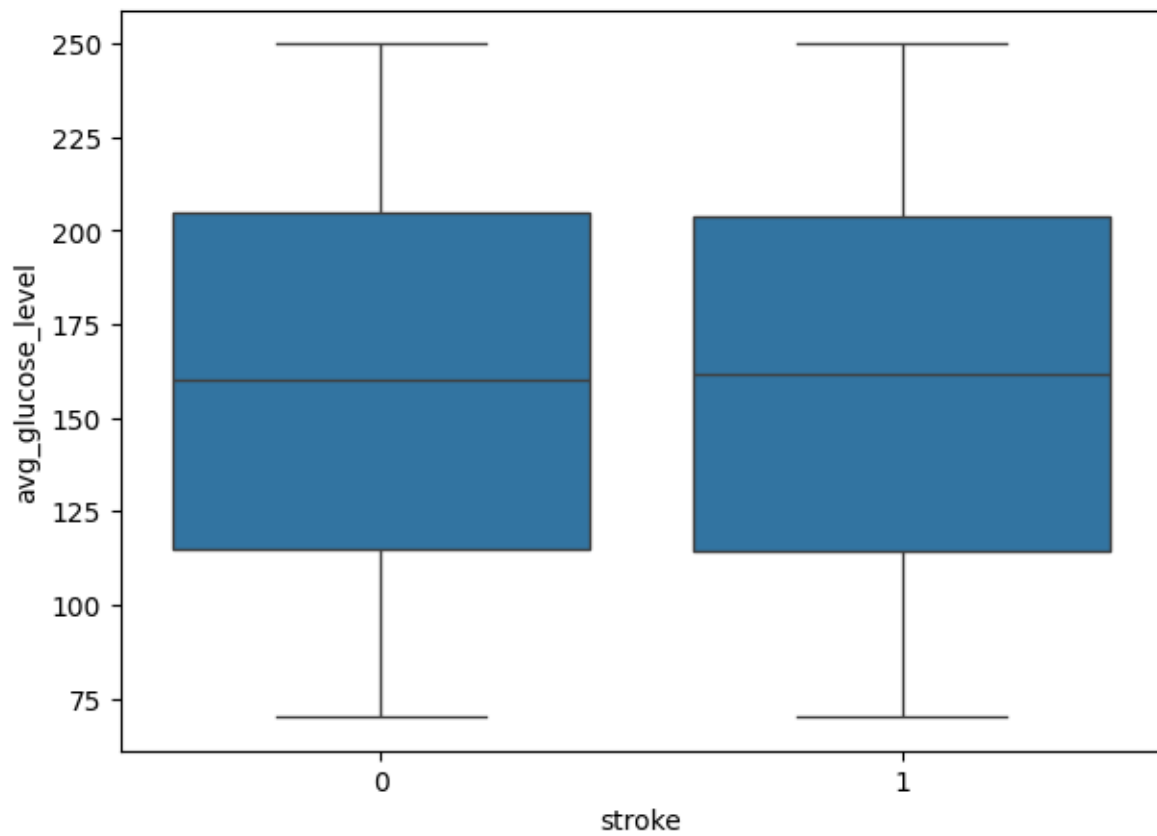
Visualizations

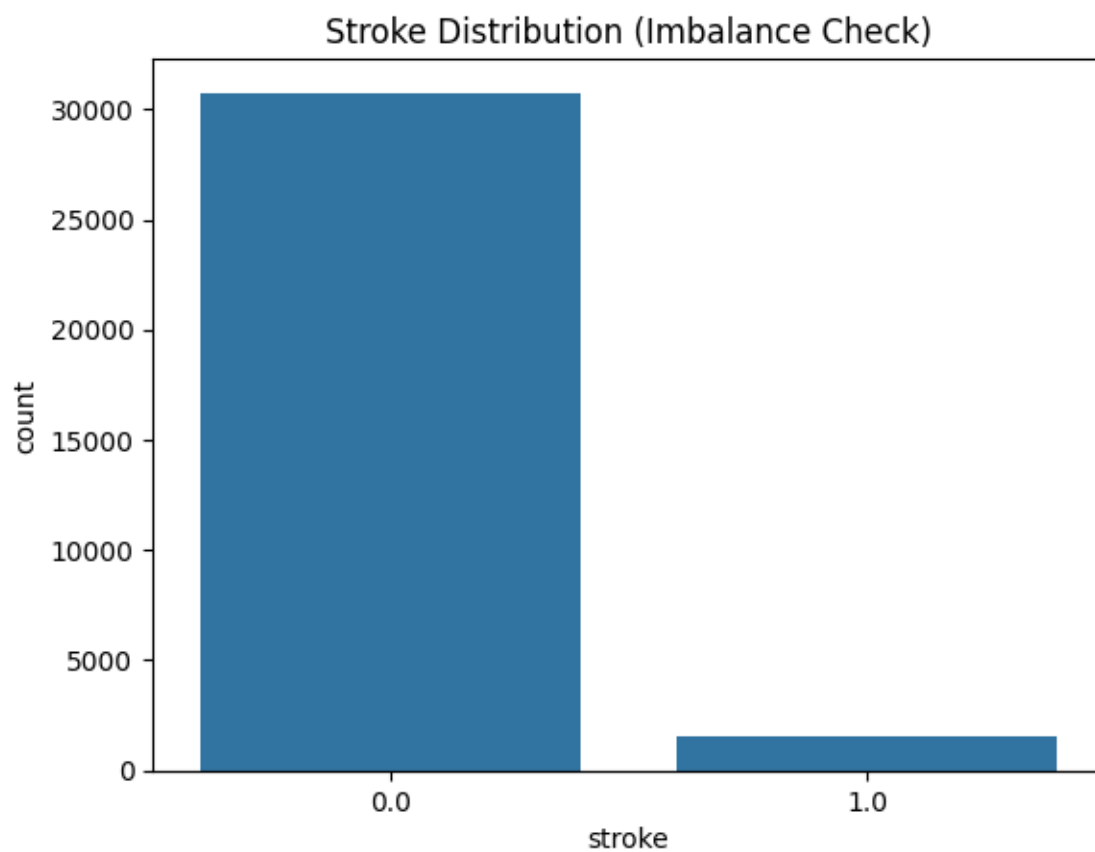
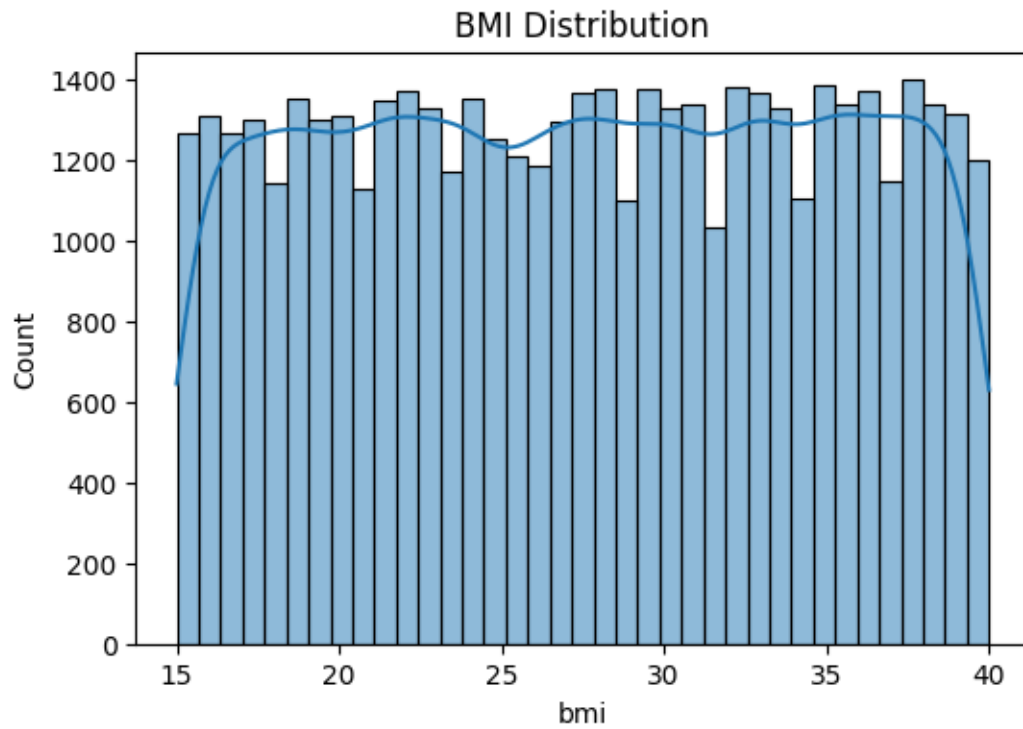


Age vs Stroke Occurrence



Glucose Level vs Stroke





Key Observations from EDA

- Age is the most critical factor for stroke.

- Strong correlation found between **glucose level, heart disease, and hypertension** with stroke.
- Smokers show slightly increased stroke likelihood.
- The data required careful balancing to avoid prediction bias.

Outcome of Preprocessing

The dataset is now:

- Clean, structured, and free of missing values
- Converted into machine-readable numeric form
- Balanced to improve model prediction capability

This ensures that the machine learning model can learn effectively and provide more accurate results.

Group no-3

Purva Nigade	202301070146
Aakanksha Sah	202301070148
Pratiksha Shinde	202301070155
Amruthavarshini Repalle	202301070158