Milestone 1 Report: Stroke Prediction Dataset

# Dataset Overview

Dataset Source: healthcare\_stroke\_dataset.csv

Objective: Understand patterns related to stroke occurrence based on demographic and health-related attributes.

## General Information

- Number of Rows: [Fill based on df.shape]

- Number of Columns: [Fill based on df.shape]

- Missing Values: Present in bmi and smoking\_status

- Duplicates: [Fill from df.duplicated().sum()]

## Data Types Summary

- Numerical Columns: age, avg\_glucose\_level, bmi

- Binary Columns: hypertension, heart\_disease, stroke

- Categorical Columns: gender, ever\_married, work\_type, Residence\_type, smoking\_status

# Data Quality Issues

|  |  |  |
| --- | --- | --- |
| Issue | Column(s) | Resolution |
| Missing values | bmi | Filled with median value |
| Unknown category placeholder | smoking\_status | Replaced '-' with 'unknown' |
| Duplicate rows | [Fill if any] | [Handled or Not Needed] |

# Exploratory Data Analysis (EDA)

## Summary Statistics

Descriptive statistics for numerical columns showed variation in age, glucose levels, and BMI. No obvious data entry errors were found (e.g., negative values).

## Outlier Detection (Boxplots)

- age: Normally distributed with slight right skew.

- avg\_glucose\_level: Contains potential outliers (high glucose values).

- bmi: Mild outliers present; handled by median imputation.

## Categorical Distribution

Countplots for all categorical features showed:

- Class imbalance in stroke (very few positive cases).

- Diverse categories in work\_type and smoking\_status.

# Data Cleaning Summary

- Missing bmi values -> Imputed with median

- Placeholder smoking\_status = '-' -> Replaced with 'unknown'

- One-hot encoding -> Applied to all categorical variables

- Binary columns conversion -> Converted to int

- Dropped irrelevant columns -> id, date

Cleaned dataset exported as: preprocessed\_stroke\_data.csv

# Deliverables Summary

|  |  |
| --- | --- |
| Deliverable | Description |
| Dataset Exploration Report | Summary of the data, distributions, and quality issues |
| EDA Notebook | Includes visualizations: heatmaps, boxplots, countplots, and summary stats |
| Cleaned Dataset | Saved as preprocessed\_stroke\_data.csv for modeling and analysis |