M1 Stroke Risk Project — Full Study & Defense Guide

1) Objective of M1

Data preparation only: load, clean, explore, preprocess. No training yet to avoid leakage and respect ML workflow.

2) Import Libraries — Why

pandas (data), numpy (math), matplotlib/seaborn (EDA visuals), sklearn (splits + pipelines). Only what was needed.

3) Load Dataset

df = pd.read_csv()

We first inspect structure, types, and missing values to inform processing decisions.

4) Drop ID Column

ID has no predictive value and can create noise or leakage.

5) EDA Overview

.head(), .info(), .describe()

We checked distributions, dtype, and missing values.

Visualizations used:

• Histograms — detect skew, outliers

• Boxplots — identify numeric outliers

• Countplots — categorical frequency

• Correlation heatmap — numeric relations

Reason: understand data behavior before preprocessing.

6) Data Split (Train/Validation/Test)

Stratified split to preserve class imbalance (stroke cases are rare).

Split rationale:

Train ≈ 60%, Val ≈ 20%, Test ≈ 20%

7) Feature Categorization

Numeric: age, bmi, avg_glucose_level

Categorical: gender, ever_married, work_type, Residence_type

Ordinal: smoking_status (ordered by medical logic)

Binary: hypertension, heart_disease

Reason: preprocessing depends on type.

8) Imputation Strategy

Numeric → median (robust to outliers)

Categorical → mode

Ordinal → mode then encoded

9) Encoding Strategy

Categorical → OneHotEncoder (no imposed order)

Ordinal → OrdinalEncoder with medically correct order

Binary → passthrough

10) Scaling Decision

Scaler imported but not used yet — optional for M2 depending on model type.

11) Pipelines + ColumnTransformer

We built modular, reproducible preprocessing.

Ensures same transformations during inference.

Prevents leakage.

12) Fit + Transform Only on Train Set

X_train_prepared = fit_transform

X_val/test_prepared = transform only

13) Final Output

Clean, encoded dataset ready for modeling.

Professor Defense Key Points

• We avoided leakage

• Stratification preserves rare class ratio

• Median vs mean: medical skewness

• One-hot for nominal; ordinal for smoking

• Pipelines follow production ML practices

End of M1 — M2 will train and evaluate models.