**✅ Step 1: Initial Data Audit**

* Review dataset structure: check column names, types, and sample entries.
* Confirm expected types (e.g., numeric vs categorical).
* Check for duplicates (and remove if any exist).

**✅ Step 2: Handle Missing or Inconsistent Values**

* **Check for missing/null values** across all columns.
* **Handle inconsistencies**:
  + For smoking\_history, treat "No Info" as missing:
    - Option 1 (recommended): Replace with "unknown" (a valid category).
    - Option 2: Impute based on demographic or clinical similarity (age, gender, comorbidity).
* If any bmi, HbA1c\_level, or blood\_glucose\_level values are missing (none in this image), impute appropriately (mean/median by group).

**✅ Step 3: Encode Categorical Variables**

* **gender**: Convert to binary (e.g., Female = 0, Male = 1).
* **smoking\_history**: Apply one-hot encoding or ordinal encoding (if a natural order is assumed).
* **diabetes**: Already binary — no action needed.
* Optional: cast categorical features to category type for memory efficiency.

**✅ Step 4: Feature Engineering (Recommended)**

**4.1 Age Grouping**

* Create age bands:
  + <30
  + 30–60
  + >60

**4.2 BMI Categorization**

Based on WHO standards:

* Underweight: BMI < 18.5
* Normal: 18.5 ≤ BMI < 25
* Overweight: 25 ≤ BMI < 30
* Obese: BMI ≥ 30

**4.3 Comorbidity Flag**

* Combine hypertension and heart\_disease:
  + 1 if either condition is present, 0 otherwise.

**✅ Step 5: Final Dataset Checks**

* Confirm no missing values remain.
* Ensure correct data types after transformations.
* Verify all engineered and encoded features are correctly added.
* Save preprocessed dataset or proceed with train-test split.