**Data Preprocessing and Feature Engineering**

**1. Data Loading and Initial Inspection**

The dataset was imported using Python’s pandas library:

import pandas as pd

df = pd.read\_csv('mimic3c.csv')

An initial inspection (df.head(), df.info()) confirmed the structure, data types, and completeness of the dataset.

**2. Handling Missing Values**

The dataset had no significant missing values in the selected features, which simplified preprocessing. Features with missing data, if any, would have been handled via imputation or exclusion depending on their relevance and proportion of missingness. In this case, no imputation was necessary.

**3. Feature Selection and Cleaning**

Non-informative or redundant columns, such as unique identifiers (hadm\_id), free-text fields (AdmitDiagnosis, AdmitProcedure), and the original target (ExpiredHospital), were removed or transformed for modeling purposes.

**4. Feature Engineering**

To improve the predictive power and capture relevant clinical insights, new features were derived based on domain knowledge and exploratory data analysis:

* **Interaction Intensity per Day** (interactions\_per\_day):  
  This feature captures the average number of clinical interactions (lab tests, medication administrations, notes, etc.) normalized by length of stay, reflecting patient complexity and care intensity.
* **Lab-to-Medication Ratio** (lab\_rx\_ratio):  
  Calculated as the ratio of lab tests ordered to medications administered, this feature provides a proxy for diagnostic effort versus treatment intensity.
* **Procedure-to-Diagnosis Ratio** (proc\_diag\_ratio):  
  Reflects the balance between the number of procedures performed and diagnoses assigned, potentially indicating intervention aggressiveness.
* **Chart Event Density** (chart\_density):  
  Number of charted events per day, representing monitoring frequency.
* **Total Unique Event Types** (total\_unique\_events):  
  A count of distinct clinical event categories with any recorded activity, serving as a proxy for case complexity.
* **High Intensity Flag** (high\_intensity\_flag):  
  A binary indicator identifying admissions with both a high number of clinical notes and procedures, suggestive of severe or complex cases.

**5. Categorical Variable Encoding**

Categorical variables such as gender, admit\_type, insurance, ethnicity, and others were transformed using **one-hot encoding**. This technique converts categorical levels into binary indicator variables, enabling models to process them effectively without assuming any ordinal relationship.

**6. Skewness and Normalization**

Several numeric features, including total interaction counts and medication counts, exhibited skewed distributions. These were log-transformed using a natural logarithm with an offset (log1p) to reduce skewness, stabilize variance, and improve model convergence.

**7. Final Dataset Preparation**

The final dataset for modeling comprised:

* Cleaned numeric features, including both original and engineered variables.
* One-hot encoded categorical features.
* A binary target variable (expired), indicating hospital mortality.

This comprehensive preprocessing pipeline ensures the dataset is well-suited for predictive modeling tasks.