**Dataset Description and Context**

**What is the MIMIC-III Clinical Classifier Dataset?**

The dataset used in this project, mimic3c.csv, is a **curated and preprocessed version of the MIMIC-III (Medical Information Mart for Intensive Care) dataset**, originally compiled by the MIT Lab for Computational Physiology. This version has been made publicly available via Kaggle by Dr. Scarlat [@drscarlat](https://www.kaggle.com/datasets/drscarlat/mimic3c) and serves as a **condensed, structured tabular format** for modeling clinical outcomes of ICU patients.

MIMIC-III is a comprehensive, de-identified dataset that contains information about **over 40,000 critical care patients** admitted to the Beth Israel Deaconess Medical Center in Boston between 2001 and 2012. It is widely used in clinical research and machine learning applications in healthcare.

**Why This Dataset?**

This curated version of MIMIC-III was selected for the following reasons:

* **Publicly Available and De-identified**: Enables ethical, reproducible research without requiring credentialed access.
* **Clean and Pre-Aggregated**: The dataset has already aggregated many raw-level events into higher-level features (e.g., number of labs, medications, procedures), making it directly usable for predictive modeling.
* **Outcome-Oriented**: It includes a clear binary outcome (ExpiredHospital), which denotes in-hospital mortality, aligning perfectly with the project’s objective of predicting ICU patient outcomes.
* **Rich Feature Set**: Contains a mix of demographic, clinical, administrative, and procedural variables, enabling multifactorial analysis.

**How the Dataset is Used**

The dataset is loaded using Python and the pandas library:

import pandas as pd

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

This command imports the CSV file into a structured **DataFrame**, allowing for easy manipulation, exploration, and modeling using Python-based data science tools.

**Key Columns in the Dataset**

Here are the most relevant attributes from mimic3c.csv:

|  |  |
| --- | --- |
| **Feature** | **Description** |
| hadm\_id | Hospital admission ID (unique identifier for each ICU stay) |
| gender, age | Patient demographics |
| LOSdays | Length of stay in days |
| admit\_type, admit\_location | Type and origin of hospital admission |
| insurance, religion, marital\_status, ethnicity | Social and financial information |
| AdmitDiagnosis, AdmitProcedure | Diagnosis and procedure on admission (free text) |
| NumLabs, NumRx, NumNotes, etc. | Count of different clinical events during admission |
| ExpiredHospital | Binary target: 1 = patient died in hospital, 0 = survived |
| TotalNumInteract | Total count of recorded interactions (combined) |
| LOSgroupNum | Binned length of stay categories (discrete class) |

**Target Variable**

* **ExpiredHospital**: This is the main target variable for prediction. It is binary:
  + 1 = patient expired during the hospital stay
  + 0 = patient was discharged alive

This variable is converted to an int type and stored as a new feature called expired for modeling purposes.

**Ethical Use**

Although the original MIMIC-III data requires credentialing and data use agreements, this curated version is **ethically sourced and anonymized** for public use on Kaggle. Researchers are still expected to handle the data responsibly and in accordance with privacy-preserving standards.