**Medication Recommendation Model**

**Project Goal and Objective:**

This project demonstrates the step-by-step process of finding the best model that can be utilized to recommend a medication based on condition, symptom and other demographic information for a patient. The dataset used to perform this analysis is sourced from <https://www.kaggle.com/datasets/asjad99/mimiciii/data>, the data is further grouped and synthetic data is added to create a reasonable pool of records to analyze and train a model.

MIMIC-III (Medical Information Mart for Intensive Care) is a freely accessible database developed by the MIT Lab for Computational Physiology. It contains detailed information about over 60,000 ICU admissions to Beth Israel Deaconess Medical Center between 2001 and 2012. For this analysis, the data is derived from Patient, Prescriptions Admissions, ICD and Drug code datasets.

By building and implementing a predictive model of medication recommendation in an EHR system will enable providers to make informed medication decisions and will prove to be a useful tool to study medication effectiveness and additionally will help minimize hospital and provider visits.

**Methodology**

**Data Loading**

The dataset is loaded into a Pandas DataFrame. The CSV file contains the following:

Features related to patient demographics, symptoms, conditions A target variable indicating the recommended medication

The first few rows are analyzed to understand its structure and content.

**Input**

Columns Description  
PatientID - Incremental value masked  
Age - Age  
Gender - Gender M/F  
BMI - Body Mass Index  
Weight\_kg - Weight  
Height\_cm - Height  
Chronic\_Conditions - Existing condition  
Symptoms - Symptoms  
Diagnosis - Diagnosis code  
Recommended\_Medication - Medication mapped  
NDC - National Drug Code  
Dosage - Medication Dosage  
Duration - Duration for Medication  
Treatmen\_Effectiveness - Effectiveness of Medication  
Adverse\_Reactions - Any allergic reactions or side-effects  
Recovery\_Time\_Days - Recovery time average  
16 Columns 8 Numeric and 8 Categorical

**Exploratory Data Analysis**

EDA is a crucial step to understand the dataset and identify potential issues. Here, we:

Check for missing values in each column using value counts.  
Summarize numerical features using descriptive statistics and histograms  
Summarize categorical features  
Identify outliers for numeric datatypes  
Find correlation between features using heatmap  
Visualize categorical feature distribution against top 10 values of target variable (Recommended\_Medication)  
Visualize the top 10 distribution of the target variable (Recommended\_Medication)

**Data Preprocessing and Feature Engineering**

Preprocessing is essential to prepare the data for model training. The steps include:

Separating features (X) and the target variable (y). Identifying categorical and numerical columns. Creating pipelines for:  
\* Numerical Features: Imputing missing values and scaling.  
\* Ordinal Features: Imputing missing values and ordinal encoding.  
\* Categorical Features: Imputing missing values and one-hot encoding  
\* Text Features: Separately use CountVectorizer to convert text into bag of words  
Combining these transformations into a unified preprocessing pipeline.

**Split dataset into Train and Test**

Split features and target variable into train and test datasets for model training and ensure target classes with atleast 2 samples are present in both datasets for stratification. Since the target label is multilabel, encode target variable using MultiLabelBinarizer Evaluate distribution of target variable between train and test datasets

**Baseline Performance Analysis using MultiOutputClassifier DummyClassifier and LogisticRegression**

The target variable is a multi-label target and shows imbalance through these scores

* Baseline Accuracy: 0.025672075563090337
* Baseline F1 Score: 0.06671387839819111  
  Confusion matrix analysis for top 10 medications: Overall, the model demonstrates high sensitivity and predicts top 10 medications [Metformin,Azithromycin], with many true positives. However, the relatively equal number of false positives and false negatives indicate that the model is not biased and maintains a level between precision and recall.

Plot feature distribution after preprocessing

**Train a simple LogisticRegression Model:**

The Logistic Regression model shows outstanding performance for most classes, with perfect precision, recall, and F1-scores (all 1.00) for most classes. This indicates that the model accurately classifies most of the samples in the dataset.  
However, the model struggles with a few classes with low (e.g., Class 6, Class 10, Class 14, and Class 15), precision, recall, and F1-score are 0.00. Furthermore, classes with low support (e.g., Class 7, Class 9, Class 21) show imbalanced metrics with precision and recall both being around 0.50, which indicates that the model has difficulty distinguishing these smaller classes accurately. The macro average metrics (precision: 0.71, recall: 0.68, F1: 0.69) reflect this issue, suggesting that while the model is performing well in larger classes, there is room for improvement in handling imbalanced or minority classes.

**Conclusion**

The feature coefficients contributing to medication recommendation are: Chronic\_condition, Dosage and Symptoms