PREDICTING HOSPITAL ADMISSION AT EMERGENCY DEPARTMENT TRIAGE USING MACHINE LEARNING

MA5755: Data Analysis & Visualization in R/Python/SQL

Submitted by-

Inderjeet(MA24M011)
Divyanshu Singh(MA24M008)
Devesh Pant(MA24M007)
Briti Sundar Ghatak(MA24M006)
Muskaan Jain(MA23C022)
Kadambari Pranav(MM24M014)

Under the guidance of

Prof. Neelesh S Upadhye

Department of Mathematics
Indian Institute of Technology Madras

Contents

Abstract

Emergency Departments (EDs) across the globe face persistent challenges such as overcrowding, long patient wait times, and limited medical resources, which can hinder timely care and patient outcomes. Efficient triaging and resource allocation are crucial to address these issues and improve operational workflow.

This project is focussed on developing Machine Learning models to predict the admission of the patient in the Emergency Department(ED) triage which will help in prioritizing patients and better use of hospital resources. A dataset of 5,60,486 patient records with 972 features was used for this project. The features were engineered to reduce the dimensionality of the problem to avoid learning from the noise. Models like Random forest, logistic regression, XGBoost classifiers were used for the prediction of patient admission in the Emergency Department(ED) triage. The results demonstrate the potential usage of machine learning models to help with the problem discussed initially.

1 Introduction

The Emergency Department (ED) is one of the most congested and high-pressure environments in a hospital setting. A significant proportion of patient admissions originate from the ED, placing immense strain on hospital resources and staff. Despite the fact that many ED visits result in patient discharge, delays in triage decisions contribute heavily to overcrowding. Efficiently predicting whether a patient will require hospital admission at the point of triage can play a critical role in streamlining patient flow and optimizing resource allocation.

This project aims to develop a machine learning-based predictive model to determine the likelihood of a patient being admitted to the hospital based on triage data and historical medical records extracted from Electronic Health Records (EHR). The dataset includes a variety of features such as prior patient disposition, historical vital signs, laboratory test results, and outpatient medication history.

To address this binary classification problem, several machine learning algorithms were employed, including Logistic Regression, Random Forest, and XGBoost. Information Gain was used as the primary metric for feature selection and dimensionality reduction, allowing the model to focus on the most informative attributes. By leveraging these techniques, the study identifies key predictors of hospital admission and demonstrates the potential of data-driven decision support tools in enhancing ED triage efficiency.

2 Dataset

2.1 Data Source

The data used for this project is taken from the Electronic Health Records (EHR) of the patients who visited the Yale New Haven Health system in the time frame of the years 2014-2017. These records are not publicly available due to the privacy policy and hence a de-identified and processed dataset of the patients was taken.

2.2 Data Description

The original dataset consisted of 560,486 patient records with 972 features.

Target variable: The target variable is the patient disposition (binary column). The target variable is imbalanced with only 28% admissions. The distribution is discussed in the Exploratory Data Analysis section.

Demographic variables: These variables include the age, gender, primary language, religion, ethnicity, insurance status, employment status, marital status of the patient.

Triage variables: Triage evaluation included variables routinely collected upon patient arrival, such as the presenting hospital name, arrival details (month, day, and 4-hour time bin), and mode of arrival. Clinical assessments at triage comprised the Emergency Severity Index (ESI) level assigned by the triage nurse and vital signs, including systolic and diastolic blood pressure, pulse rate, respiratory rate, oxygen saturation, use of an oxygen delivery device, and body temperature. These features collectively represent the triage variables used in the predictive modeling process.

Chief complaint variables: 200 unique chief complaints are considered like ankle injury, asthma, back pain, etc.

Hospital Usage statistic variables: These include the number of ED visits within one year, the number of admissions within one year, the disposition of the patient's previous ED visit, and the number of surgeries listed in the patient's record at the time of encounter.

Past medical history variables: 281 categories of issues in the Past medical history are recorded with a binary value of 1 if the patient has that issue or 0 otherwise.

Outpatient medication variables: 48 therapeutic categories of outpatient medication (e.g., cardiovascular, analgesics), with each category represented by a variable indicating the number of medications prescribed within that group.

Historical vitals: The maximum, minimum, median, and the last recorded value are the numeric values of the variables of each patient in a time frame of 1 year from the date of ED triage.

Historical labs: 150 frequent labs ordered by the previous EDs are chosen as two types of variables. If the result of the lab test is numeric, the maximum, minimum, median, and the last recorded value are included as the features. If the result of the lab test is categorical like positive/negative, then binary classification is used for the feature where 1 = Positive, 0 = Negative. The number of tests, the number of positives, and the last recorded value of each categorical lab were included as features.

Imaging and EKG counts: The number of previous orders of the following categories are chosen as the features/variables in the dataset. The categories are: EKG, X-Ray: Chest/other, echocardiogram, ultrasound, CT: Head/other, MRI, other imaging.

Category	**Number of Variables**
Response variable (Disposition)	1
Demographics	9
Triage evaluation	13
Chief complaint	200
Hospital usage statistic	4
Past medical history	281
Outpatient medications	48
Historical vitals	28
Historical labs	379
Imaging/EKG counts	9
Total	**972**

Table 1: Dataset categories and number of variables

2.3 Data Cleaning and Preprocessing

Handling Missing and Duplicate Values: Columns with more than 50% missing feature values were removed due to their limited utility. Patient records with missing values or duplicate entries were also excluded to maintain data integrity and consistency.

Encoding Categorical Variables: Categorical features with two unique values (e.g., gender, language, disposition) were converted to binary variables. Features with multiple categories (e.g., time of arrival, race, insurance status, marital status, and employment status) were grouped into meaningful bins where applicable and then one-hot encoded. Additionally, the column containing department names was label encoded due to its ordinal relevance.

Dimensional Reduction by Dropping Irrelevant Columns: Irrelevant features such as ethnicity and religion were dropped to simplify the model and reduce noise. Columns with over 95% zero values were also removed, as their limited presence across records makes them less likely to contribute meaningfully to the prediction task.

Addressing Multicollinearity: Multicollinearity was addressed in two stages. Initially, correlation analysis was performed on the numerical subset of the raw dataset, and one feature from each highly correlated pair (correlation > 0.9) was removed. After converting all categorical variables into numerical form, a second correlation check was performed on the fully numeric dataset to eliminate any additional multicollinear features introduced during encoding.

After preprocessing, we are left with 107319 records and 133 features.

2.4 Exploratory Data Analysis

Target variable class distribution: The data is imbalanced with 77693 discharge counts and 29626 admit counts. This resembles the real world data perfectly because many of the ED triage ends with the discharge of the patient.

Class distribution of the target variable is as follows:

Output	Counts	Percentage
Discharged	77693	72%
Admitted	29626	28%

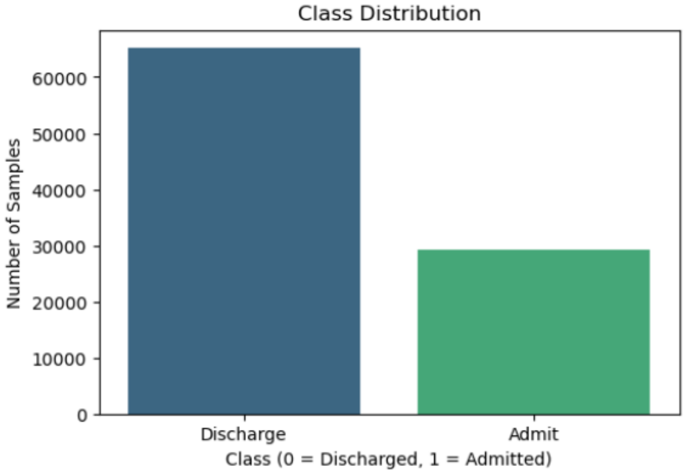

Figure 1: Class distribution of target variable

Emergency Severity Index:

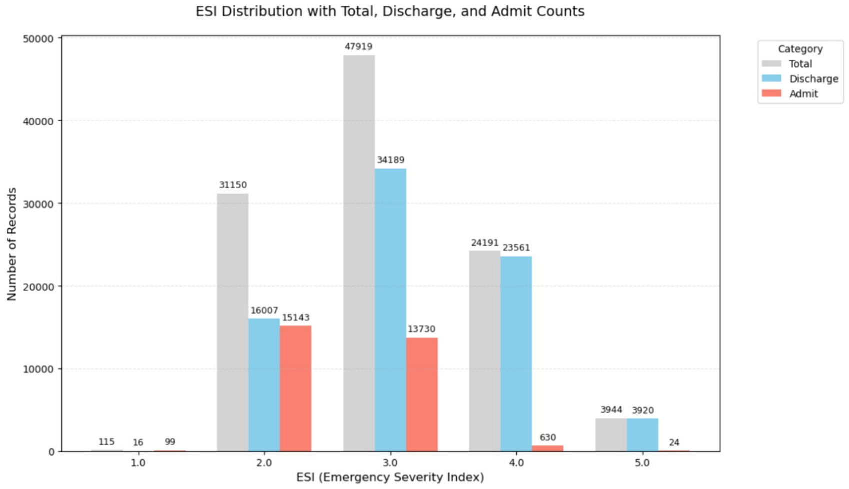

Figure 2: ESI distribution with the total, discharged and admitted counts

Observations: High-severity cases (ESI 1 & 2) have a high admission rate, especially ESI 2, which shows a large number of admissions relative to its total count. ESI 3 has the highest number of patients, but the majority are discharged, indicating a lower admission rate despite moderate severity. Low-severity cases (ESI 4 & 5) are rarely admitted, with most patients being discharged, reflecting their minimal need for hospitalization.

Cardiovascular medication counts:

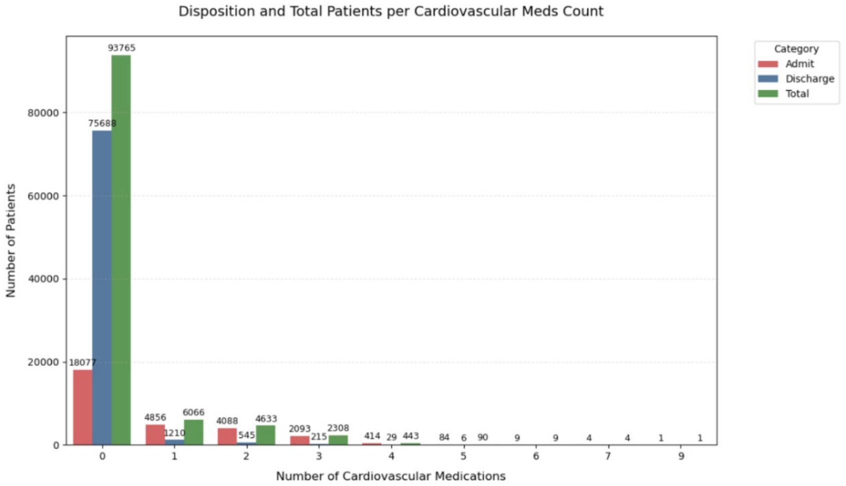

Figure 3: **Admitted, discharged and total counts distribution of cardiovascular medication bins**

Observations: Most patients $(93,765)$ are on 0 cardiovascular medications. Among those with 0 meds, discharges dominate. As the number of medications increases, the admission rate also increases. Patients on more than 3 meds have a very high admission proportion, indicating greater severity.

Gastrointestinal medication counts:

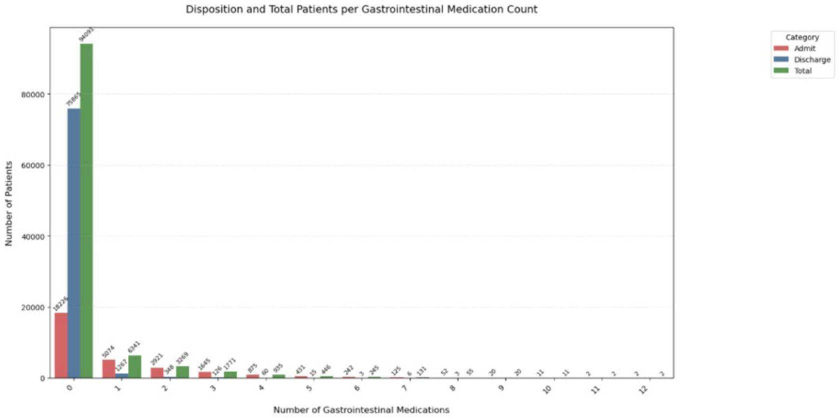

Figure 4: **Admitted, discharged and total counts distribution of gastrointestinal medication bins**

Observations: Most patients (94,027) are on 0 gastrointestinal medications. Among those with 0 meds, discharges dominate. As the number of GI medications increases, the admission rate also rises. Patients on 3 or more medications show a high admission proportion, suggesting higher severity or more complex conditions.

Age and admission:

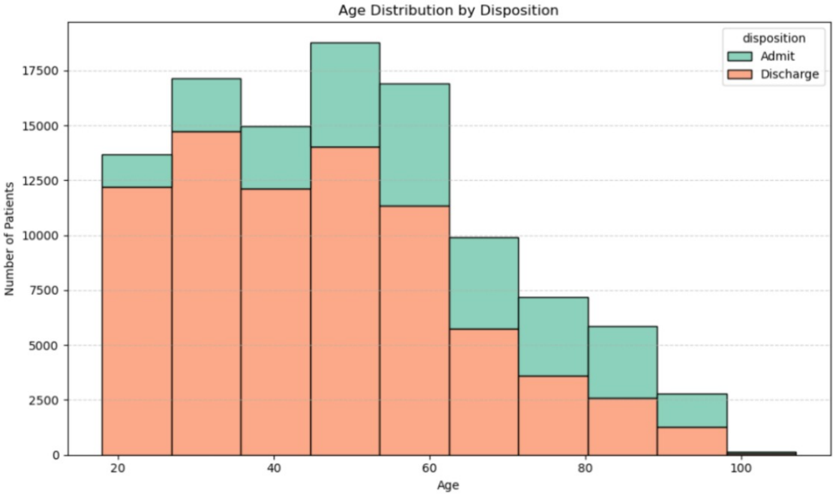

Figure 5: **Histogram distribution plot of age by disposition**

Observations: Most patients fall in the 30–50 age range. In younger age groups (below 60), discharges significantly outnumber admissions. However, as age increases beyond 60, the admission rate rises sharply. Patients above 80 show a high proportion of admissions, suggesting that age is a strong factor in patient disposition outcomes.

Patient Distribution by Total Medication Count:

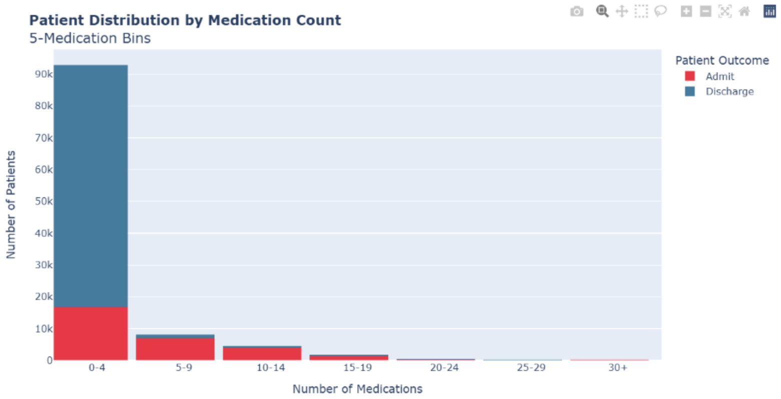

Figure 6: **Histogram distribution plot of Total medication by disposition**

Observations: The majority of patients (over 90,000) are on 0–4 medications, with discharges being dominant in this group. As medication count increases, patient count decreases sharply. However, admission rates rise proportionally in higher medication bins. Patients on 10+ medications are predominantly admitted, indicating that a higher number of medications may be associated with more severe conditions.

Total Medications and ESI:

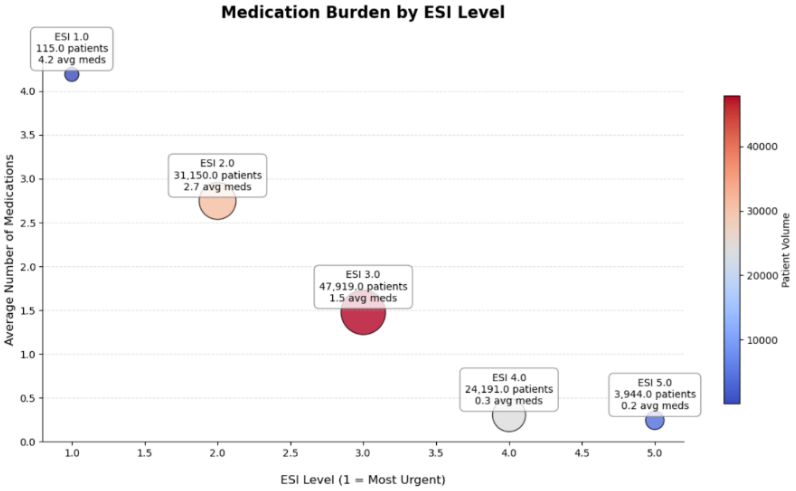

Figure 7: **Bubble Chart of Medication Burden by ESI (Emergency Severity Index) Level**

Observations: As ESI level increases (i.e., as urgency decreases), the average number of medications decreases significantly. ESI 1 patients (most urgent) have the highest medication burden, averaging 4.2 meds, while ESI 5 patients (least urgent) average just 0.2 meds. The highest patient volume is in ESI 3 (47,919 patients), with an average of 1.5 meds. This suggests that more urgent cases typically involve more complex treatment with a higher medication load. Additionally, Spearman's rho = -0.246 (p ¡ 0.0001) reveals a statistically significant negative monotonic relationship between age and ESI level, suggesting that as age increases, patients tend to present with more urgent conditions (lower ESI levels).

Age and ESI:

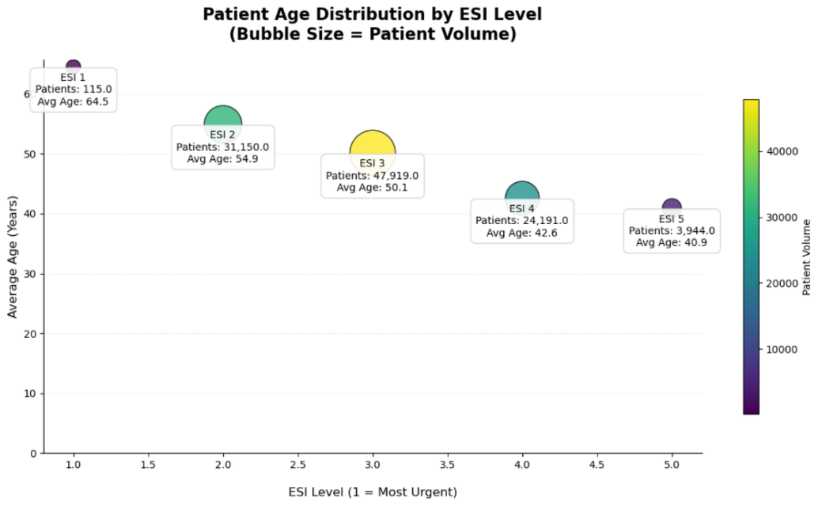

Figure 8: **Bubble Chart of Patient Age Distribution by ESI (Emergency Severity Index) Level**

Observations: As the ESI level increases (indicating less urgent conditions), the average age of patients decreases. For instance, patients in ESI 1 (most urgent) have the highest average age at 64.5 years, whereas ESI 5 patients average only 40.9 years. The patient volume peaks at ESI 3, with an average age of 50.1 years. Additionally, Spearman's rho = -0.243 (p ¡ 0.0001) reveals a statistically significant negative monotonic relationship between age and ESI level, suggesting that older patients tend to present with more urgent conditions (lower ESI levels).

Number of Surgeries count plot:

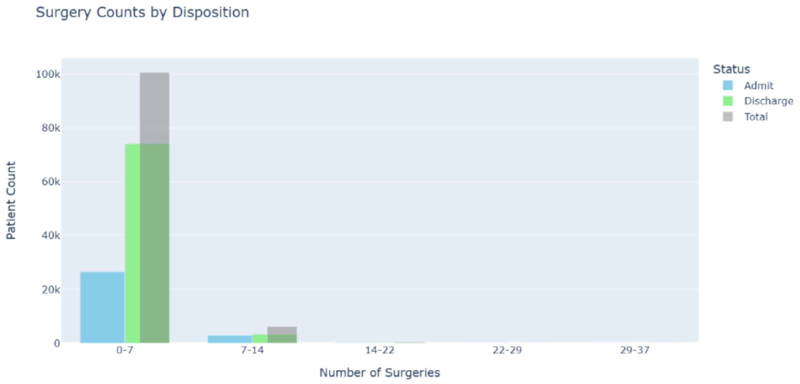

Figure 9: **Admitted, discharged and total counts distribution of number of surgeries**

Observations: As the number of surgeries increases, the patient count decreases significantly. For instance, patients with 0–7 surgeries represent the highest volume, totaling over 100,000, while those in the 7–14 range drop to a few thousand. Beyond 14 surgeries, the patient count becomes negligible. The majority of both admitted and discharged patients fall within the 0–7 surgeries group, indicating that most patients undergo a limited number of surgical procedures.

Arrival Mode plot:

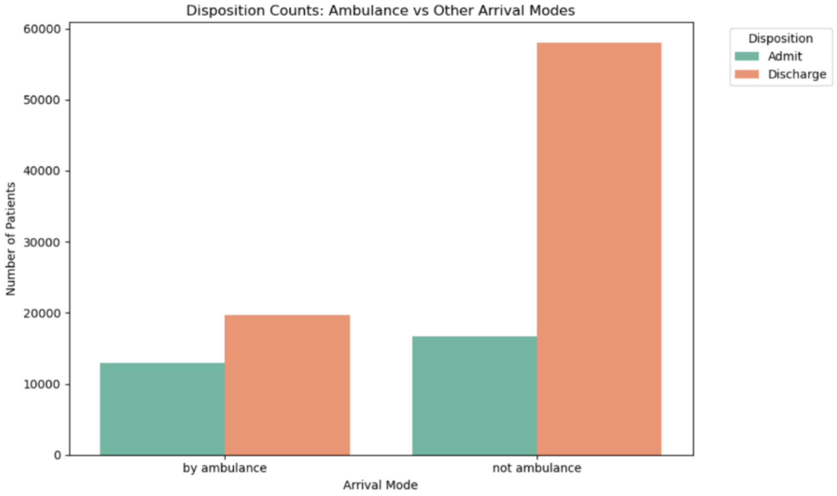

Figure 10: **Patient Disposition by Arrival Mode (Ambulance vs Other)**

Observations: Patients arriving by ambulance are more likely to be admitted compared to those arriving by other modes. However, the overall volume of patients arriving not by ambulance is much higher, with over 55,000 discharges compared to around 19,000 for ambulance arrivals. Admissions for both arrival modes are relatively similar, but discharges are significantly more common among non-ambulance arrivals, indicating that less severe cases tend to come via non-emergency transport.

3 Methodology

The methodology employed a comprehensive approach to develop predictive models for hospital admission decisions, beginning with extensive data preprocessing to ensure data quality and model robustness. Missing values were addressed by removing columns with over 50% missing data and eliminating any remaining null values along with duplicate entries. To reduce feature redundancy, highly correlated numeric variables were removed, while sparse features containing more than 95% zero values were excluded. Categorical variables were consolidated into clinically meaningful categories and appropriately encoded.

Exploratory data analysis revealed significant patterns and relationships within the dataset. The target variable exhibited a class imbalance with approximately 72% discharge cases versus 28% admissions. Key insights included the strong correlation between higher Emergency Severity Index (ESI) levels and increased admission rates, the predictive power of medication counts, and the association between advanced age (60+ years) with surgical history and higher admission likelihood. These findings informed subsequent feature engineering efforts, which created composite variables like total medication counts. The study employed three machine learning approaches. Ensemble methods - Random Forest and XGBoost - underwent hyperparameter optimization using RandomizedSearchCV while simultaneously performing feature selection to identify the most predictive variables. As a baseline comparison, Logistic Regression was implemented using standardized features with class weighting to address the inherent class imbalance. XGBoost established it as the optimal model for clinical deployment, demonstrating both high predictive accuracy and well-calibrated probability estimates.

The models were deployed via a Streamlit web application, enabling real-time prediction of hospital admission likelihood. The app accepts user inputs in comma separated file containing information such as patient demographics, vital signs, and medication counts, then displays predicted admission result. This end-to-end pipeline—spanning data cleaning, feature engineering, model training, and deployment—provides a scalable and clinically actionable solution for predicting hospital admissions.

4 Model Development and Training

The hospital admission prediction task was implemented as a binary classification problem using a structured modeling approach. The dataset was strategically divided into training

(80%) and test (20%) sets through stratified sampling to maintain the natural class distribution. Three modeling techniques were employed: Logistic Regression, Random Forest and XGBoost ensemble methods.

Implemented with class weighting and feature standardization, Logistic regression was trained on all features and evaluated using standard metrics including recall, F1 score, log loss, and precision-recall AUC. While Logistic Regression was fast and interpretable, its predictive performance was comparatively lower than ensemble models.

Both Random Forest and XGBoost were trained using the complete feature set, followed by hyperparameter optimization through RandomizedSearchCV. Subsequently, each ensemble method's built-in feature importance metrics were leveraged to identify and select the top 50 most predictive features. This enabled the training of streamlined model variants that maintained strong performance while improving computational efficiency.

Each model was finally evaluated on the held-out test set using recall, F1 score, log loss, and PR AUC. The ensemble models—especially XGBoost—offered superior predictive power, while Logistic Regression remained a valuable benchmark for interpretability and baseline comparison.

The predictive models were deployed in an interactive Streamlit web application that combines predictions from all three algorithms (Logistic Regression, Random Forest, and XGBoost) through majority voting to generate final admission recommendations. The app features an intuitive interface where clinicians can input patient demographics, clinical indicators, and medical history and receive real-time predictions.

5 Results

The evaluation revealed XGBoost as the top-performing model, achieving a PR AUC of 0.845 and log loss of 0.353 on the test set, demonstrating both high predictive accuracy and well-calibrated probabilities. Random Forest followed closely with a PR AUC of 0.83, while Logistic Regression provided a strong baseline (PR AUC: 0.820). All models maintained consistent performance between training and test sets, indicating robustness against overfitting. Feature importance analysis identified total medications, ESI level, previous disposition and age as the strongest predictors across all algorithms. Notably, using only the top 50 features preserved model performance, confirming these variables capture the most clinically relevant signals. These results validate the models' potential to support admission decisions while balancing accuracy and interpretability.

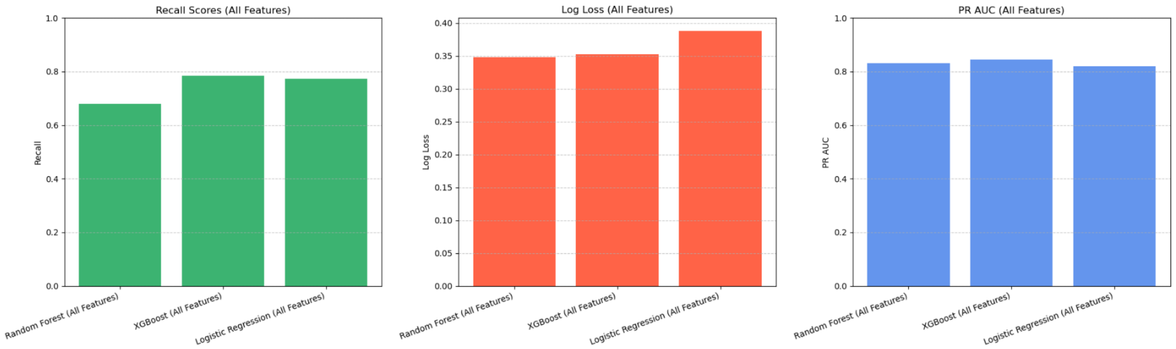

Figure 11: Model Performance Comparison

6 References

- https://doi.org/10.1371/journal.pone.0201016
- https://github.com/yaleemmlc/admissionprediction
- https://www.ynhhs.org/