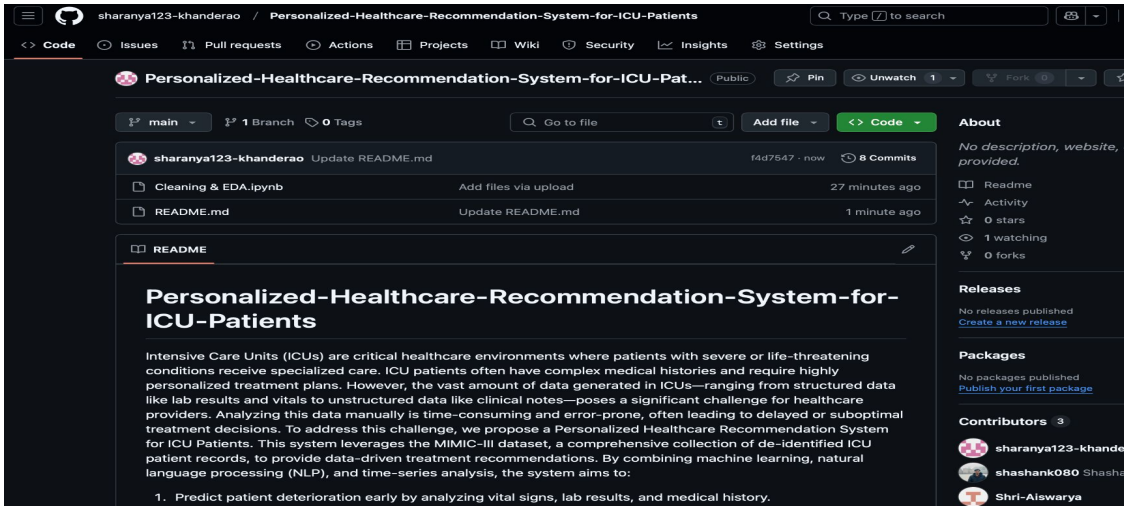
Personalized Healthcare Recommendation system for ICU Patients

TEAM D:
SHRI AISWARYA GORLE
SHASHANK RAJ GUPTA GUNTA
VENKATA NAGA SHARANYA KHANDE RAO

GITHUB REPO



https://github.com/sharanya123-khanderao/Personalized-Healthcare-Recommendation-System-for-ICU-Patients

Kaggle ID : khanderaosharanya

WHY THIS PROJECT ?

- Life-Saving Impact: Early detection of patient deterioration enables prompt, potentially life-saving interventions.
- Multimodal Data Integration: Combines structured data (vitals, labs, history) with unstructured clinical notes for a comprehensive view.
- Innovative Techniques: Leverages advanced ML & NLP to uncover hidden patterns and predict adverse events.
- Real-World Relevance: Uses the robust, publicly available MIMIC-III dataset to simulate realistic ICU scenarios.
- Explainable AI: Enhances clinical decision support through transparent, interpretable models.

DATASET OVERVIEW:

Dataset: MIMIC-III 10k(Kaggle)

Link: https://www.kaggle.com/datasets/bilal1907/mimic-iii-10k
Source: MIT Lab for Computational Physiology (via PhysioNet)
Size: 10,000 ICU patient records
Time Period: 2001–2012
Data Format: Processed .csv files

Objective

To develop predictive models for ICU patient outcomes using MIMIC-III data:
Early Deterioration Prediction (e.g., in-hospital mortality)
Risk Stratification (e.g., severity scoring, comorbidity risk)
Time-Series Modeling (e.g., trends in vitals/labs)

KEY COMPONENTS:

- ADMISSIONS.csv – hospital admission data (diagnoses, discharge outcomes)
- PATIENTS.csv – demographic info (age, gender, DOB, DOD)
- ICUSTAYS.csv – ICU stay details (length of stay, first care unit)
- CHARTEVENTS.csv – time-stamped clinical data (vitals, GCS, etc.)
- LABEVENTS.csv – lab test results
- DIAGNOSES_ICD.csv – coded diagnoses (ICD-9)

TARGET VARIABLES & FEATURE DESCRIPTION:

- **Target Variables**

  In-hospital mortality (hospital_expire_flag)

  Binary Classification:

  1 = patient died during hospital stay

  0 = patient survived discharge

  Length of Stay (los)

  Regression Target: ICU stay duration in “days”

- **Key Features**

  Vital Signs (from CHARTEVENTS.csv):

  Heart Rate, Respiratory Rate, Temperature, Blood Pressure (SBP/DBP), SpO_2

  Demographics (from PATIENTS.csv): Age, Gender, Ethnicity

  Clinical Information:

  Diagnoses (from DIAGNOSES_ICD.csv)

  ICU type and admission source (ICUSTAYS.csv, ADMISSIONS.csv)

  Comorbidities (derived from diagnosis codes)

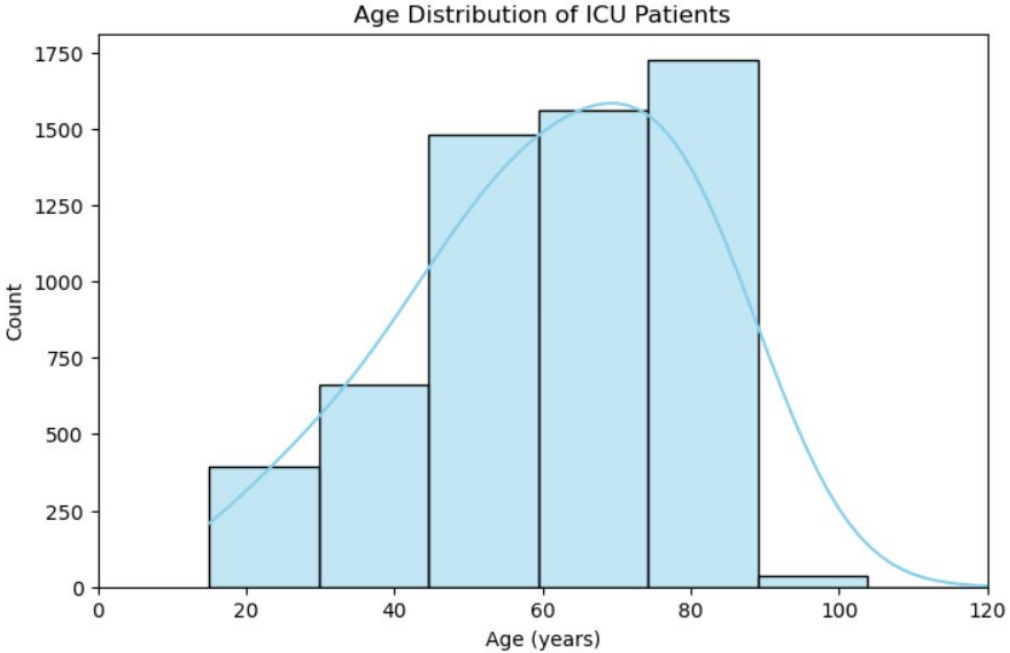
Literature Survey & Our Contribution

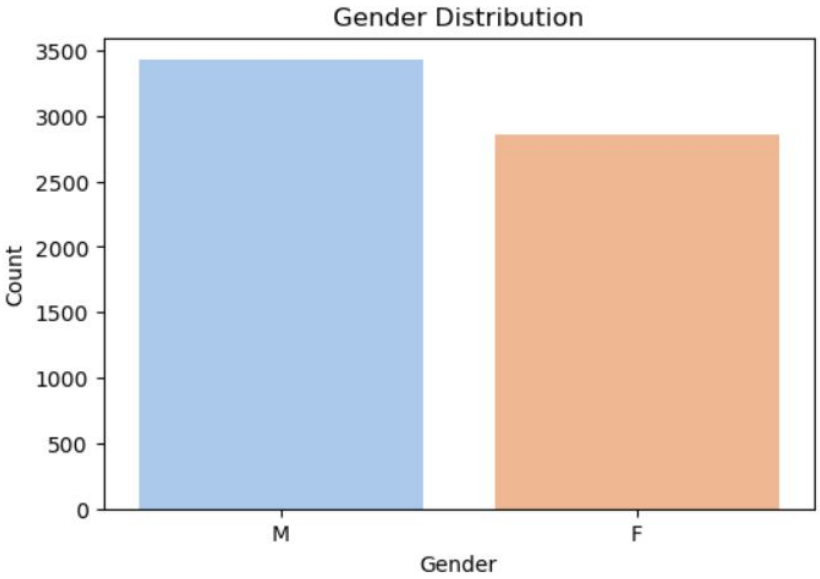
What others have done?

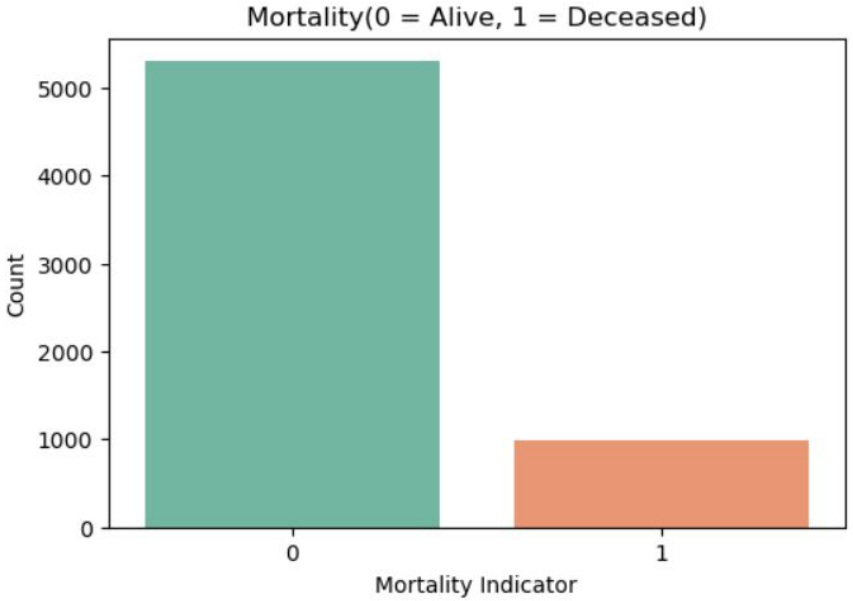
- Single-Task Focus: Most studies address either mortality prediction or length of stay—rarely both.
- Static Models: Limited use of time-series data like real-time vitals and lab trends.
- Limited Generalizability: Prior work often uses proprietary or simulated datasets.
- Minimal Deployment: Few models are integrated into clinical-facing tools.
- Lack of Multimodal Fusion: Sparse integration of structured and unstructured data.

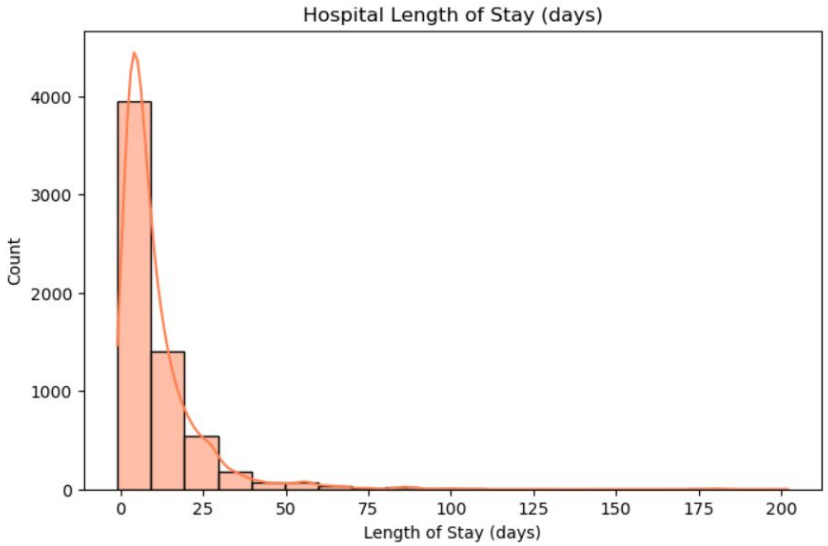
What we have done:

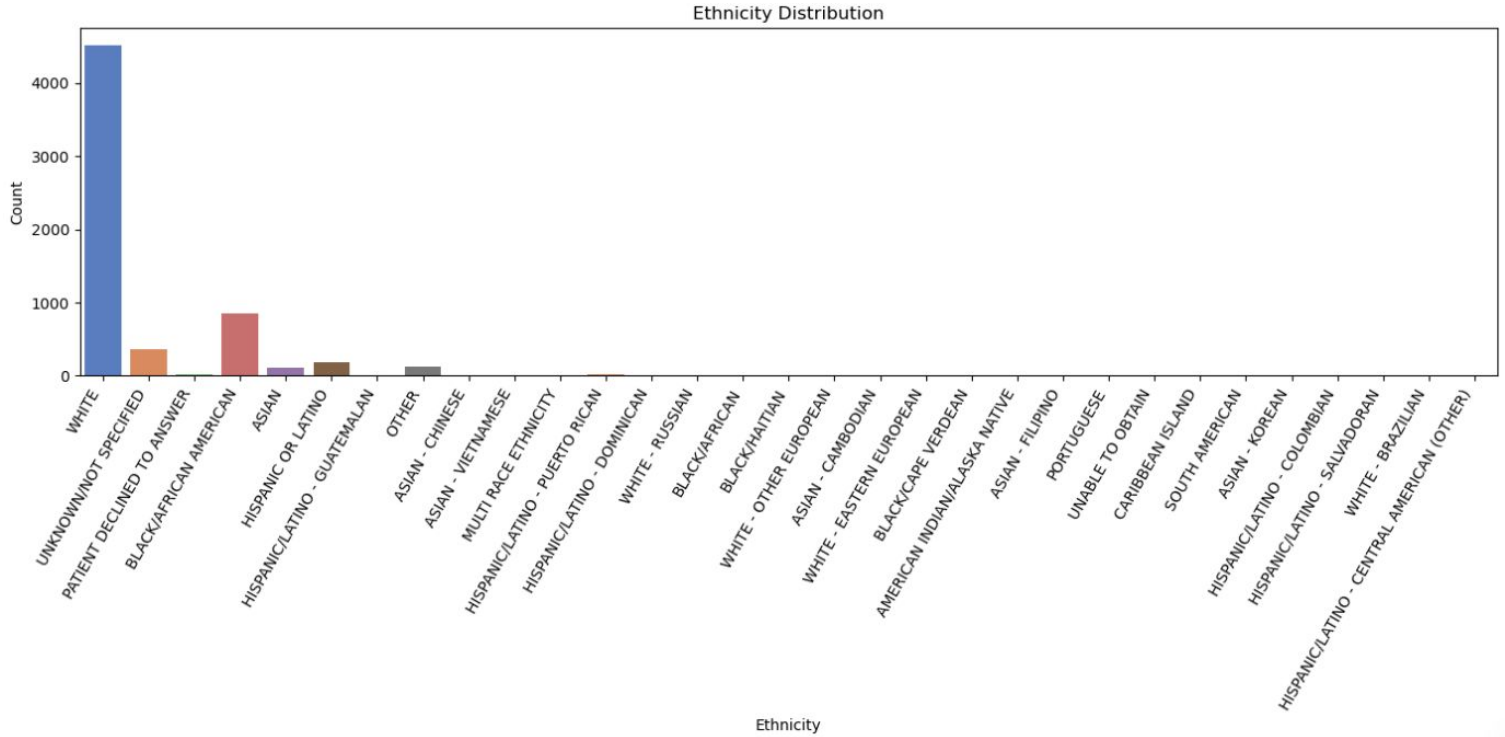
- Dual Prediction Tasks: Simultaneous modeling of mortality risk and ICU length of stay.
- Predicted patient deterioration early by analyzing vital signs, lab results, and medical history.
- First-Hour Data Focus: Early vitals, demographics, and admission info for rapid predictions.
- High-Fidelity Dataset: Uses real-world, public MIMIC-III (10k) ICU data.
- Our approach uniquely integrates TF-IDF-based NLP with clinical text to predict patient mortality, demonstrating real-world applicability through a scalable and interpretable model.

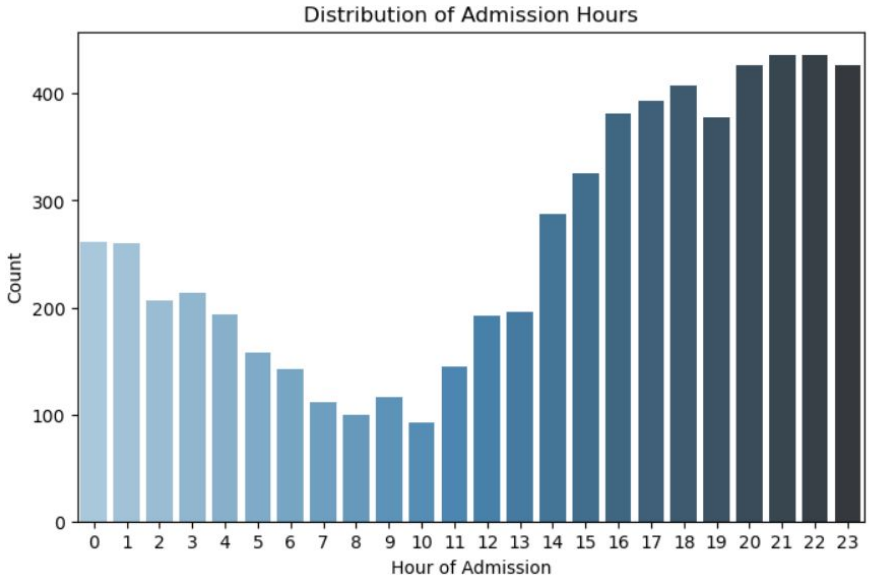

Age Distribution of ICU Patients
Count
Age (years)
0
250
500
750
1000
1250
1500
1750
0
20
40
60
80
100
120

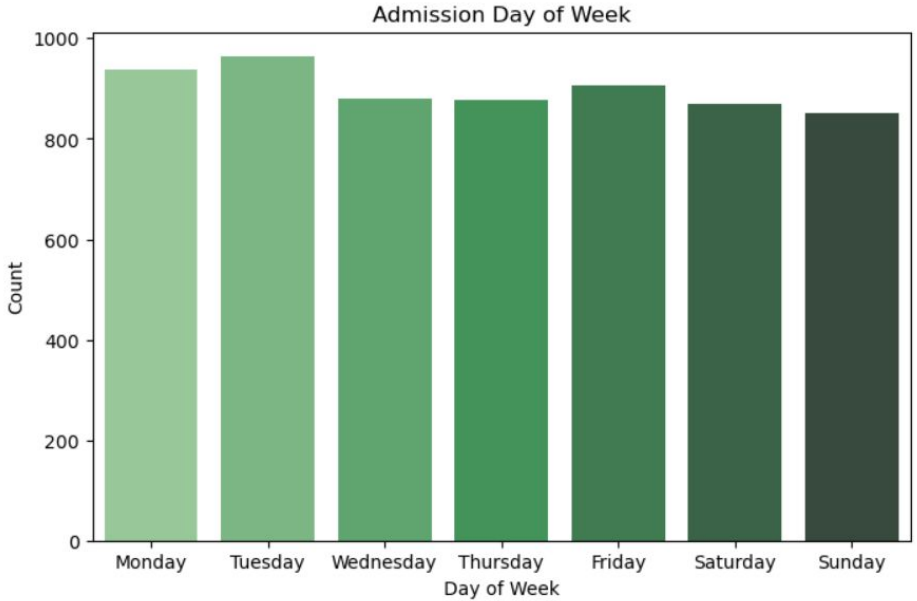

Gender Distribution
Count
Gender
0
500
1000
1500
2000
2500
3000
3500
M
F


Mortality(0 = Alive, 1 = Deceased)
Count
5000
4000
3000
2000
1000
0
0
1
Mortality Indicator

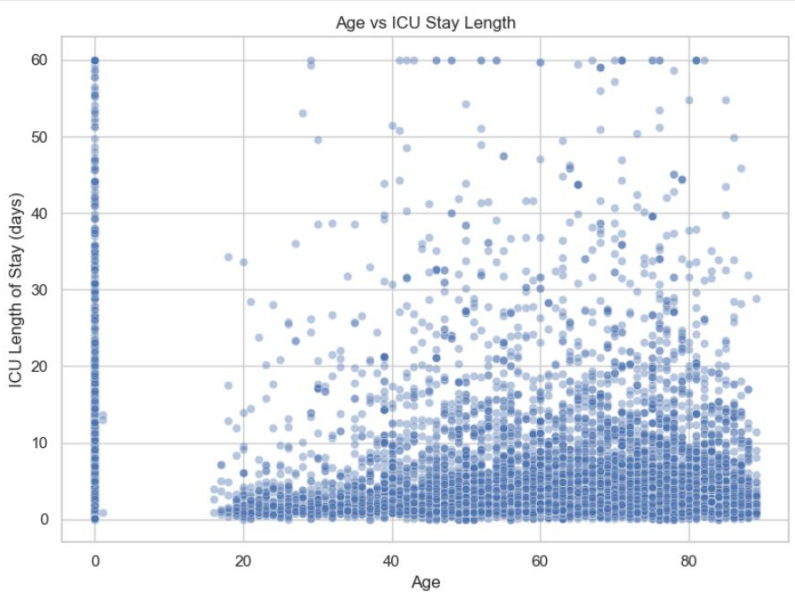

Hospital Length of Stay (days)
Count
4000
3000
2000
1000
0
0
25
50
75
100
125
150
175
200
Length of Stay (days)

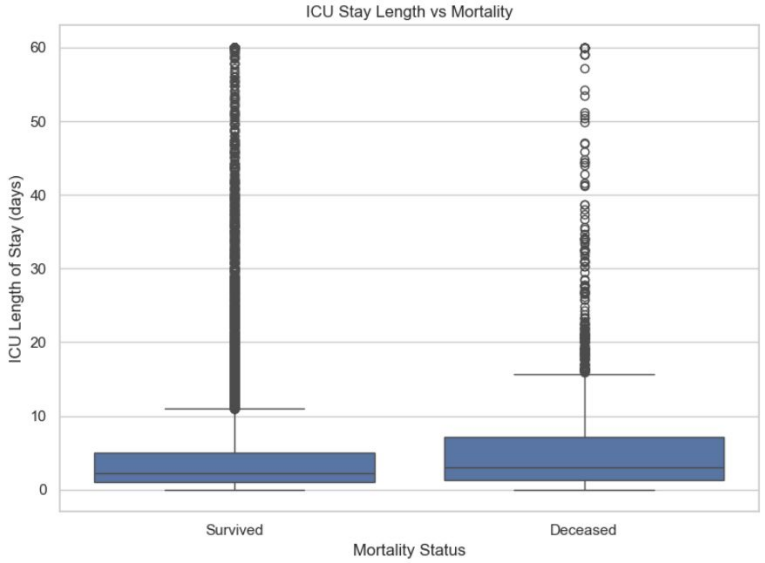

Ethnicity Distribution
Count
4000
3000
2000
1000
0
WHITE
UNKNOWN/NOT SPECIFIED
PATIENT DECLINED TO ANSWER
BLACK/AFRICAN AMERICAN
ASIAN
HISPANIC OR LATINO
HISPANIC/LATINO - GUATEMALAN
OTHER
ASIAN - CHINESE
ASIAN - VIETNAMESE
MULTI RACE ETHNICITY
HISPANIC/LATINO - PUERTO RICAN
HISPANIC/LATINO - DOMINICAN
WHITE - RUSSIAN
BLACK/AFRICAN
BLACK/HAITIAN
WHITE - OTHER EUROPEAN
ASIAN - CAMBODIAN
WHITE - EASTERN EUROPEAN
BLACK/CAPE VERDEAN
AMERICAN INDIAN/ALASKA NATIVE
ASIAN - FILIPINO
PORTUGUESE
UNABLE TO OBTAIN
CARIBBEAN ISLAND
SOUTH AMERICAN
ASIAN - KOREAN
HISPANIC/LATINO - COLOMBIAN
HISPANIC/LATINO - SALVADORAN
WHITE - BRAZILIAN
HISPANIC/LATINO - CENTRAL AMERICAN (OTHER)
Ethnicity


Distribution of Admission Hours
Count
0
100
200
300
400
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of Admission


Admission Day of Week
Count
0
200
400
600
800
1000
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Day of Week

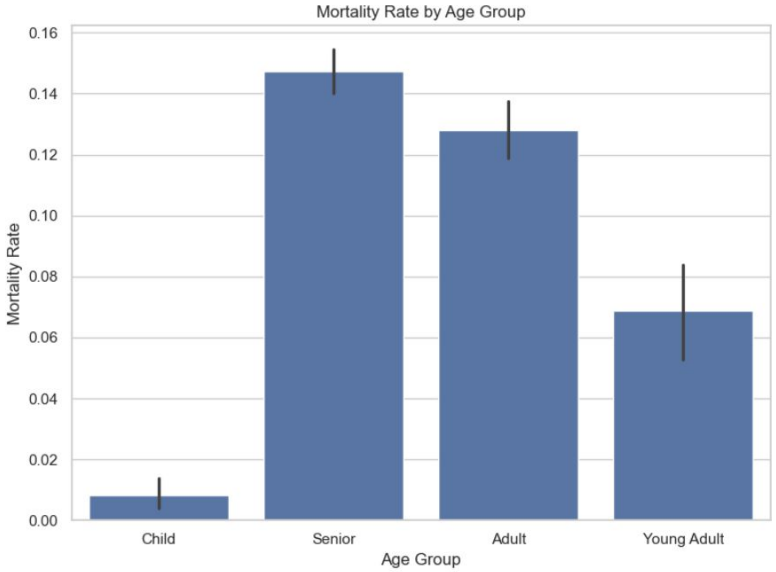
Exploratory Data Analysis (EDA)



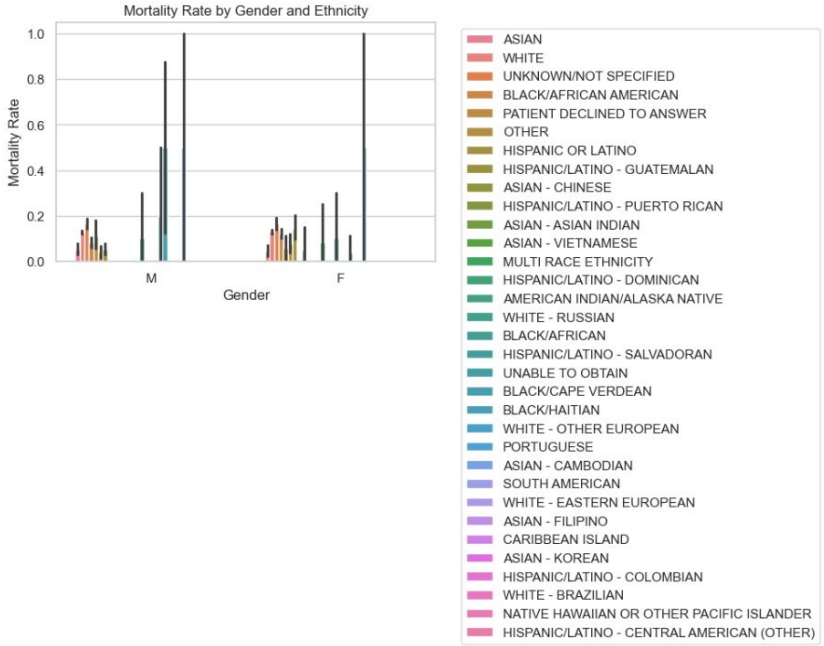


No strong correlation observed between age and ICU length of stay; most patients stay under 10 days.

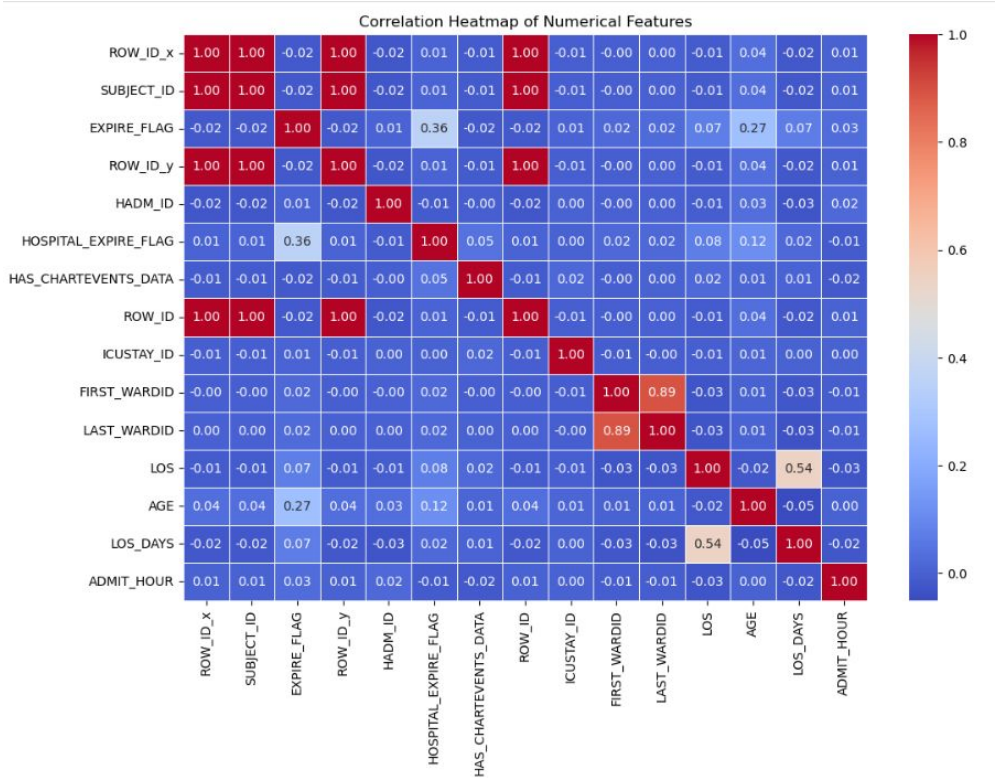
EDA



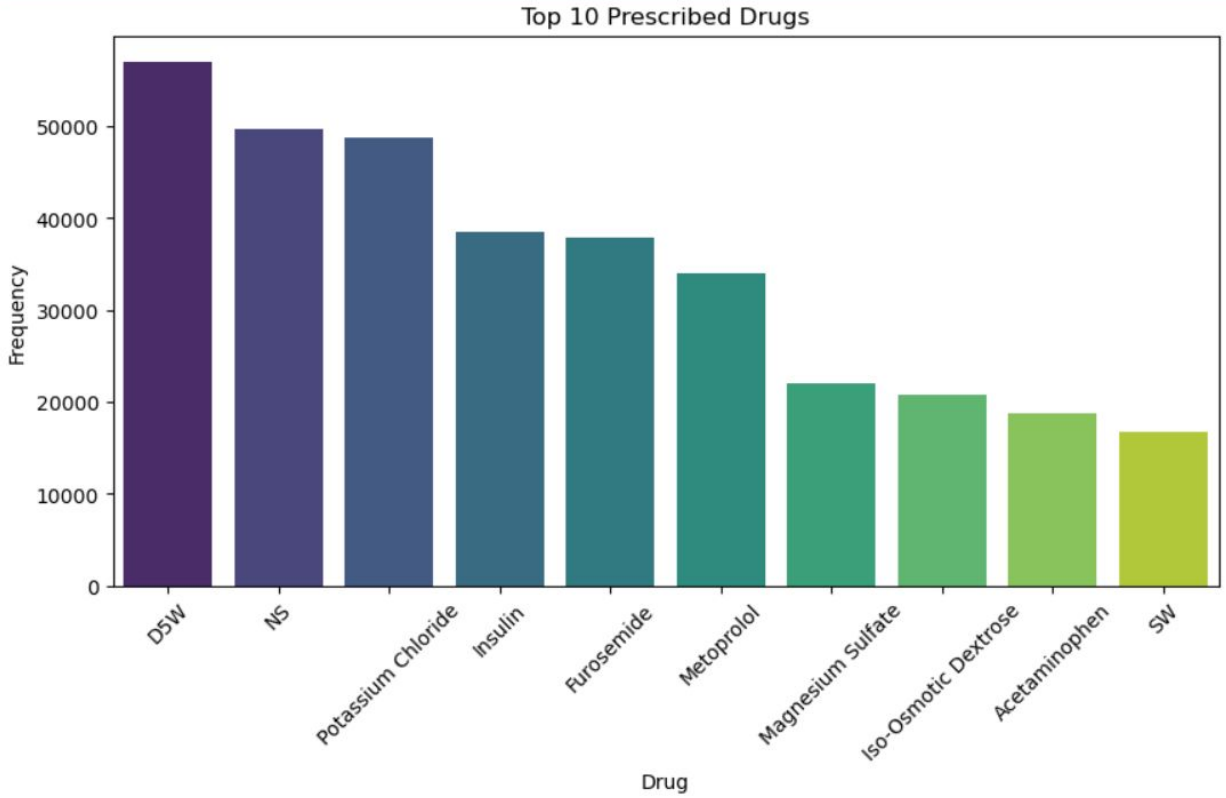
Seniors and adults show higher mortality rates, while children and young adults have the lowest.



Mortality varies across ethnic groups, with higher rates observed in certain male subgroups.



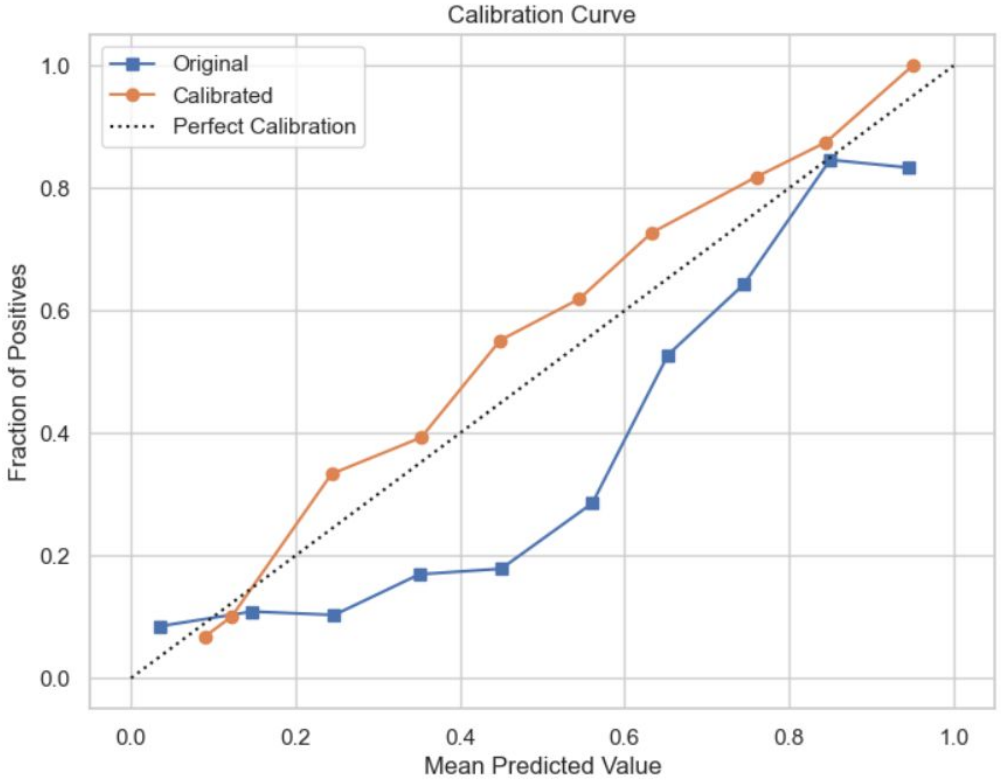
Correlation Heatmap - Shows relationships between vital signs and stay duration to identify key connections.


Top 10 Prescribed Drugs
Frequency
0
10000
20000
30000
40000
50000
D5W
NS
Potassium Chloride
Insulin
Furosemide
Metoprolol
Magnesium Sulfate
Iso-Osmotic Dextrose
Acetaminophen
SW
Drug

MODEL DEVELOPMENT

Predicting patient deterioration

- **Algorithms Used:** Logistic Regression ,Random Forest, XGBoost
- **Feature Set & Target:** AGE ,LOS_DAYS ,ADMIT_HOUR ,*HOSPITAL_EXPIRE_FLAG*
- **Preprocessing:**
  Impute missing values using the median
  Standard scaling applied to numerical features (required for LR)
  SMOTE used on training data to balance class distribution
- **Data Splitting & Validation:**
  80/20 train-test split (random_state=42)
  Additional 10-fold cross-validation during hyperparameter tuning
- **Evaluation Metrics:**
- **Confusion Matrix & Classification Report** (Precision, Recall, F1-Score)
- **ROC-AUC Score:**
  Logistic Regression: ~0.58
  Random Forest: ~0.71 (best performance)
  XGBoost: ~0.71


Calibration Curve
Original
Calibrated
Perfect Calibration
Fraction of Positives
Mean Predicted Value
0.0
0.2
0.4
0.6
0.8
1.0

MODEL DEVELOPMENT

Classification Task – Predicting Admission Status

- **Algorithm Used**: Logistic Regression (sklearn.linear_model.LogisticRegression)
- **Feature Set:** First-hour vitals (temperature, heartrate, resprate, o2sat, sbp, dbp), triage indicators, demographic attributes
- **Target:** disposition (0 = discharged, 1 = admitted)
- **Preprocessing:**
  Missing values imputed (median for numeric, mode for categorical)
  Label Encoding applied to gender, race, arrival_transport, chiefcomplaint, and acuity
- **Train-Test Split:** 80% training / 20% test
- **Output:** Binary classification (Admission Yes/No)
- **Evaluation Metrics:** Accuracy, Precision, Recall, F1-score

Regression Task – Predicting Length of Stay

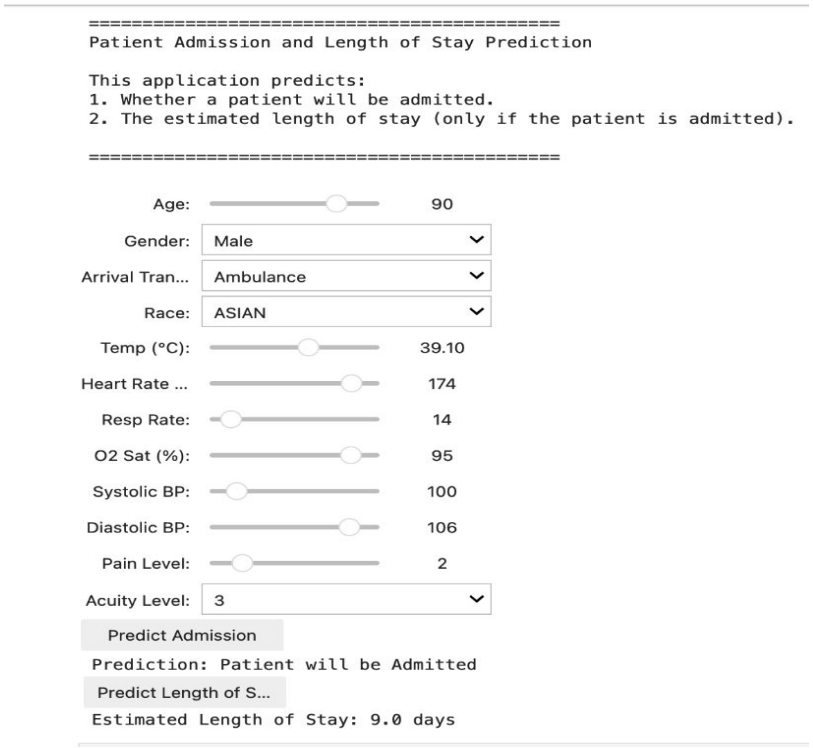
- **Algorithm Used:** Random Forest Regressor (sklearn.ensemble.RandomForestRegressor)
- **Feature Set:** First-hour vitals (temperature, heartrate, resprate, o2sat, sbp, dbp), triage indicators, demographic attributes
- **Target:** length_of_stay (in hours), computed from intime and outtime
- **Outlier Handling:** IQR-based filtering on length_of_stay and continuous vitals
- **Evaluation Metrics:** RMSE (Root Mean Squared Error), R^2 Score

NLP pipeline that performs binary classification on text data using TF-IDF and Logistic Regression:

Predicting Patient Outcome from Clinical Notes

- **Algorithm Used**: Logistic Regression (sklearn.linear_model.LogisticRegression)
- **Vectorization Technique**: TF-IDF (sklearn.feature_extraction.text.TfidfVectorizer)
  - Extracted top 10,000 features
  - Removed English stop words
- **Input Data**: cleaned_text column from clinical notes
- **Target Variable**: HOSPITAL_EXPIRE_FLAG (0 = survived, 1 = expired)
- **Preprocessing**:
  - Applied TF-IDF vectorization to convert text into sparse numerical matrix
- **Train-Test Split**:
  - 80% training / 20% test split using train_test_split
- **Model Training**:
  - Trained using Logistic Regression with max_iter=1000 to ensure convergence
- **Evaluation Metrics**:
  - Accuracy, Precision, Recall, F1-Score (via classification_report)

REAL-TIME PREDICTION:



- Predicts whether the patient will be **admitted** or **discharged**
- If admitted, estimates **Length of Stay (LOS)** in days
- Intuitive **slider and dropdown UI** for clinical usability
- Helps **doctors simulate scenarios** and prioritize high-risk patients
- Demonstrates the **practical application** of machine learning in emergency care

MODEL EVALUATION

Classification Results – Logistic Regression (Admission prediction)
Accuracy: 82.6%
Precision: 80.3%
Recall: 79.4%
F1-Score: 79.8%

Regression Results – Random Forest Regressor (Length of stay)
RMSE: 3.52 hours
R^2 Score: 0.71

NLP Clinical Notes:
Accuracy: 90.0%
Precision: 82.0%
Recall: 61.0%
F1-Score: 65.0%

Classification Results- Random Forest(Patient Deterioration)
Accuracy: 87%
Precision (for deterioration, class 1): 57%
Recall (for deterioration, class 1): 40%
F1-Score (for deterioration, class 1): 47%
ROC-AUC: 0.712

- Logistic Regression achieved a balanced performance, with slight improvement possible via feature interaction terms or regularization tuning.
- Random Forest produced stable LOS predictions with strong generalization, indicating non-linear feature interactions were well-captured.

CONCLUSION:

- Used **statistical models** (Logistic Regression, Decision Trees) for **mortality prediction**
- **Length of Stay** (LOS) forecasting via time-series (e.g., ARIMA, linear models)
- **Feature Selection**: PCA, LASSO, RFE based on correlation and domain knowledge
- Mostly **single-task** focus: either mortality or LOS, not both
- NLP pipeline effectively transforms unstructured clinical notes into actionable predictions, achieving high accuracy in mortality classification using TF-IDF features and Logistic Regression.
- Model ensemble effectively identifies patient deterioration risk, with Random Forest and XGBoost outperforming Logistic Regression, achieving a strong ROC-AUC of ~0.71 on imbalanced clinical data.

REFERENCES

1. PhysioNet: https://physionet.org/content/mimiciii/1.4/
2. Kaggle Dataset: https://www.kaggle.com/datasets/bilal1907/mimic-iii-10kt
3. ZHAW Mortality Prediction: https://www.zhaw.ch/storage/.../DataAnalysisMortalityPrediction.pdf
4. Frontiers Study: https://www.frontiersin.org/.../818439/full

UMBC

THANKYOU