PREDICTING ICU ADMISSION USING LOGISTIC REGRESSION AND MACHINE LEARNING ON ELECTRONIC HEALTH RECORDS

1. Background and Problem Statement

Early identification of patients at high risk of ICU admission is essential for optimizing resource allocation and improving clinical outcomes. Electronic Health Records (EHR) contain a rich set of variables such as vital signs, lab results, and comorbidities that can be used to build predictive models. This project leverages both classical statistical modeling and machine learning to predict ICU admissions in hospitalized patients, helping hospitals prioritize critical care interventions.

2. Objectives

- To develop predictive models for ICU admission using EHR data
- To compare logistic regression with machine learning methods such as random forest
- To evaluate model performance using appropriate classification metrics
- To generate interpretable insights for use by clinicians and healthcare administrators

3. Methodology

Data Source:

- Synthetic dataset modeled on ICU patient data from public sources such as MIMIC-III
- Variables: age, gender, vitals (heart rate, BP, respiratory rate), lab results (WBC, creatinine, hemoglobin), comorbidities (diabetes, COPD), ICU admission flag (binary outcome)

Software:

- **Python** using pandas, scikit-learn, matplotlib, seaborn

Steps:

1. **Data Cleaning and Preparation:**
    - Imputation of missing values using median strategy
    - Feature scaling (Min-Max normalization)
    - Encoding categorical variables

- Train-test split (80-20)

2. **Modeling:**
    - **Model 1:** Logistic Regression (baseline model)
    - **Model 2:** Random Forest Classifier
    - Evaluation metrics: Accuracy, Precision, Recall, F1-Score, ROC-AUC
    - Feature importance analysis for interpretability
3. **Model Tuning:**
    - Grid search with cross-validation (for random forest)
    - Regularization tuning (for logistic regression)

4. Results

- **Logistic Regression:**
    - Accuracy: 0.78 | AUC: 0.81
    - Significant predictors: WBC count ($p < 0.01$), heart rate ($p < 0.05$), creatinine ($p < 0.05$)
- **Random Forest:**
    - Accuracy: 0.85 | AUC: 0.89
    - Top features: WBC, respiratory rate, creatinine, age
    - Feature importance clearly highlighted actionable clinical factors
- ROC curves indicated improved sensitivity of the machine learning model without sacrificing specificity

5. Interpretation and Insights

- Machine learning models such as random forest offer superior performance over traditional logistic regression in classifying ICU admission risk
- However, logistic regression provides clearer interpretability and is still clinically relevant
- High WBC, abnormal vitals, and impaired kidney function (creatinine) are key indicators of ICU need

6. Limitations

- Synthetic data may not capture real-world noise and interdependencies
- Model lacks time-series input (e.g., vitals trend over hours)
- Feature engineering was limited to basic transformations

7. Future Work

- Use time-series modeling (e.g., LSTM, survival trees) for temporal ICU risk estimation
- Incorporate real-time EHR streaming data for dynamic prediction
- Extend analysis to multi-label outcomes (e.g., ICU + ventilation requirement)

8. Relevance to Stakeholders

Academic:

- Ideal for coursework in health informatics, applied ML, or predictive analytics
- Offers practical grounding in data preprocessing, classification models, and clinical use cases

Corporate/Hospital/Startup:

- Valuable for healthcare AI firms building triage support tools
- Supports hospital IT systems in implementing real-time ICU risk alerts
- Can guide medical device vendors and analytics consultants working with hospital data