**Modeling and Evaluation**

**1. Model Selection and Setup**

To predict the likelihood of in-hospital mortality (expired), two supervised machine learning models were employed:

* **Logistic Regression**: A baseline linear model widely used for binary classification due to its interpretability and simplicity.
* **Random Forest Classifier**: An ensemble-based non-linear model that aggregates decisions from multiple decision trees, often yielding improved accuracy and robustness.

The dataset was split into training (80%) and testing (20%) subsets, using stratified sampling to maintain the proportion of mortality cases in both sets. This approach helps generalize the model performance to unseen data.

**2. Model Training**

Both models were trained on the preprocessed feature matrix (X\_train) and target vector (y\_train). Hyperparameters for the logistic regression were kept default except increasing the iteration limit for convergence. The random forest model was configured with 100 trees (n\_estimators=100) and a fixed random seed for reproducibility.

**3. Performance Metrics**

Model performance was evaluated on the test set (X\_test, y\_test) using the following metrics:

* **Accuracy**: Overall correctness of predictions.
* **Precision**: Proportion of predicted positive cases that were actually positive.
* **Recall (Sensitivity)**: Proportion of actual positive cases correctly identified.
* **F1-Score**: Harmonic mean of precision and recall, balancing false positives and false negatives.
* **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**: Measures the model’s ability to discriminate between classes across all classification thresholds.

These metrics provide a comprehensive understanding of the classifier’s effectiveness, especially in medical settings where false negatives (missing mortality cases) can be critical.

**4. Results**

* **Logistic Regression** served as a baseline, providing interpretable coefficients but relatively modest predictive power.
* **Random Forest** outperformed logistic regression by capturing non-linear interactions and complex feature relationships, demonstrated by higher AUC and improved classification scores.

**5. Visualization**

* **Confusion Matrix**: Presented as a heatmap to visually assess the count of true positives, true negatives, false positives, and false negatives, facilitating error analysis.
* **ROC Curve**: Plotted for both models to illustrate the trade-off between true positive rate and false positive rate at various thresholds, highlighting the superior discriminative ability of the random forest.
* **Feature Importance**: Extracted from the random forest model to identify the top predictors of in-hospital mortality. Key features included interaction intensity, length of stay, and specific clinical event counts, suggesting these factors are critical indicators of patient outcome.

**6. Discussion**

The improved performance of the random forest model suggests that patient outcomes are influenced by complex, non-linear interactions among demographic, clinical, and procedural factors. The engineered features capturing patient care intensity and complexity contributed significantly to model accuracy, underscoring the importance of comprehensive feature engineering in clinical predictive modeling.

However, while the random forest offers better accuracy, logistic regression’s interpretability makes it valuable for clinical insights and decision-making transparency. Combining both approaches can enhance the applicability of predictive models in healthcare settings.

**7. Limitations and Future Work**

* The dataset, though rich, does not capture all aspects of patient health, such as detailed lab values, physiological time series, or clinician notes’ semantics.
* Future work could include temporal modeling using recurrent neural networks or transformers, incorporation of richer clinical data, and external validation on different ICU populations.