Zhaocheng Yang
Instructor: Andras Zsom
Brown University | Data Science Institute
Dec 15, 2024
GitHub: https://github.com/Barcayzc/Project1030

Predicting Sepsis Survival Using Clinical Data

1. Introduction

Sepsis is a life-threatening condition caused by the body's exaggerated immune response to infection, often resulting in organ failure and death. According to the World Health Organization (WHO), sepsis affects over 30 million people annually, leading to approximately 6 million deaths worldwide [1]. In the United States, it accounts for over $24 billion in healthcare costs each year [2]. Despite medical advancements, the complexity and heterogeneity of sepsis make its diagnosis and management particularly challenging [3][4].

Timely risk assessment is critical, as delayed intervention can result in death within hours. While laboratory tests provide valuable insights, their results often arrive too late for immediate clinical decision-making. Machine learning offers a promising solution by enabling rapid and accurate survival prediction using a small set of readily available medical features. This project aims to predict whether a hospitalized patient with sepsis is likely to survive or die within approximately nine days, using three features—age, sex, and the number of sepsis-related episodes—through a binary classification framework involving four machine learning algorithms.

Dataset

The dataset, sourced from the UC Irvine Machine Learning Repository [5], contains 110,204 hospital admissions in Norway from 2011 to 2012. These admissions include patients diagnosed with infections, systemic inflammatory response syndrome (SIRS), sepsis, or septic shock. The data, derived from the Norwegian Patient Registry and Statistics Norway, is robust and reliable for survival prediction studies.

Previous Work

Previous studies using this dataset applied five machine learning classifiers—linear regression, support vector machines (SVM with linear and radial kernels), gradient boosting, and naïve Bayes—alongside the ROSE oversampling technique to handle class imbalance [5]. However, unlike earlier work that focused on predicting the majority class (survived), this study prioritizes the clinically critical minority class (deceased). Predicting the minority class in imbalanced datasets presents greater challenges, which may impact performance but reduce false negatives, which is believed more important in this scenario.

2. Exploratory Data Analysis

Our exploratory data analysis (EDA) aimed to uncover the distribution and relationships of the three key features—age, sex, and septic episode number—with patient survival outcomes. Below is a concise summary of the key findings.

Feature	Explanation	Measurement	Range
Age	The age of the patient at the hospital stay	Years	[0, 100]
Episode number	The number of septic episodes experienced by the patient	Integer	[1, 5]
Sex	0: male; 1: female	Binary	0,1
[Target] Survival	0: Deceased; 1: Alive	Boolean	0,1

<Table 1. Meanings, measurement units, and intervals of each feature of the dataset>

Class Imbalance in the Target Variable

Figure 1 reveals a significant class imbalance in the target variable, Hospital_Outcome. Specifically, 92.6% of patients survived (class 1), while only 7.4% were deceased (class 0). This imbalance highlights the need for careful handling of the minority class (class 0) during data preparation to ensure reliable predictive performance.

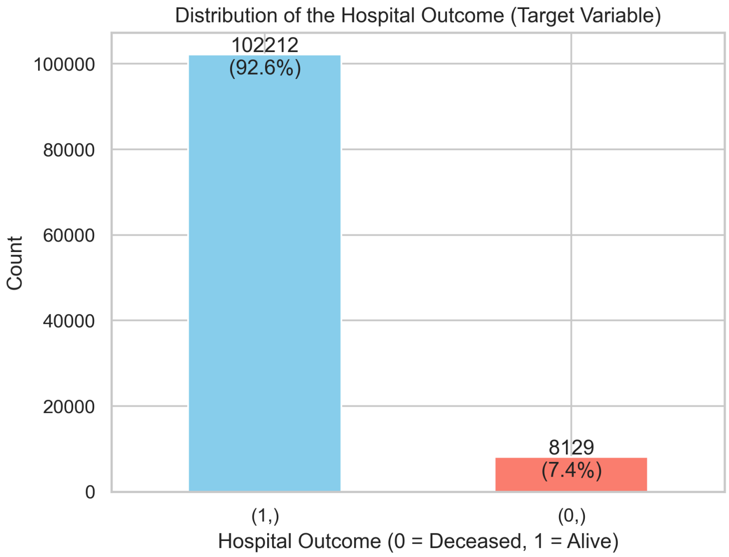


Figure 1. The target variable 'Hospital Outcome' shows a highly imbalanced distribution. Class (1) dominates with 92.6% (102,212 instances), while class (0) represents only 7.4% (8,129 instances). Addressing this imbalance is critical for model performance.

Age and Survival Outcomes

As shown in Figure 2, survival outcomes vary significantly by age:

Young patients (0–10 years): Exceptionally high survival rates with minimal mortality.

Middle-aged patients (30–60 years): Survival remains high, with a slight increase in mortality.

Older patients (≥60 years): Mortality rises sharply, peaking in the 70–85 age group.

This trend aligns with clinical expectations, underscoring the increased vulnerability of older patients to sepsis-related mortality.

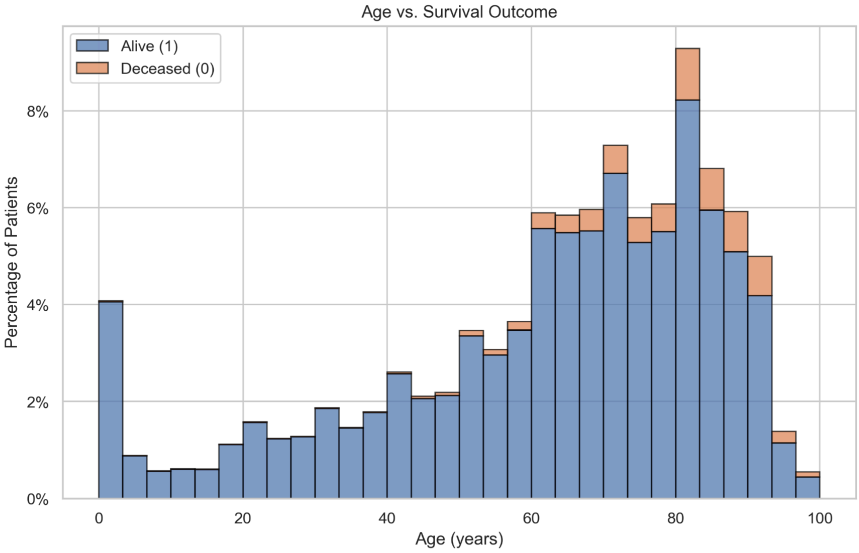


Figure 2. Distribution of patient age and survival outcomes. Patients aged 0-10 years exhibit very high survival rates, whereas mortality increases significantly among patients aged 60+ years, particularly peaking around the 70-85 year range. This trend highlights the relationship between age and hospital outcomes, emphasizing the vulnerability of older patients.

Gender and Survival Outcomes

Figure 3 shows the relationship between gender and survival.

Males: 92.1% survival, 7.9% mortality; Females: 93.2% survival, 6.8% mortality.

While females exhibit slightly better survival rates, the overall impact of gender appears minimal, suggesting it may not be a strong predictor of outcomes.

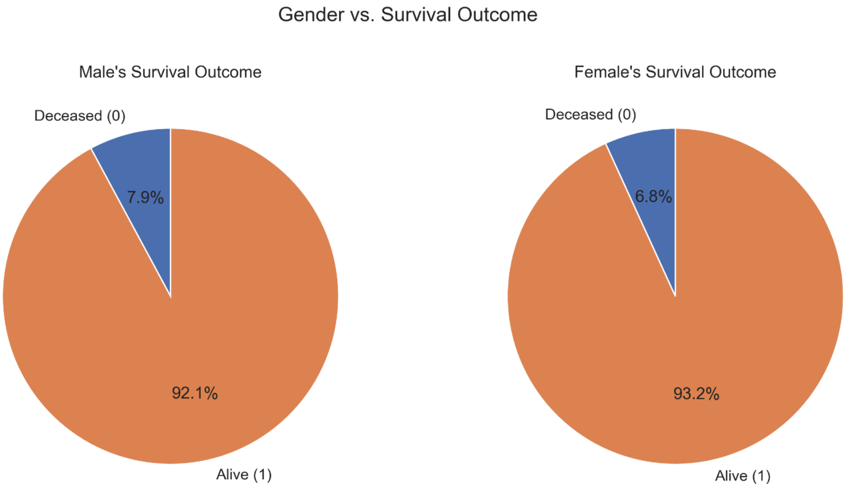


Figure 3. Gender-based survival outcomes show a slightly higher survival rate for females (93.2%) compared to males (92.1%), with both groups having a small percentage of deceased patients.

Episode Number and Survival Outcomes

Figure 4 visualizes survival across five septic episodes.

Despite slight variations between each, survival rates consistently exceed 91% across all episodes, indicating a limited but observable relationship between episode number and survival outcomes.

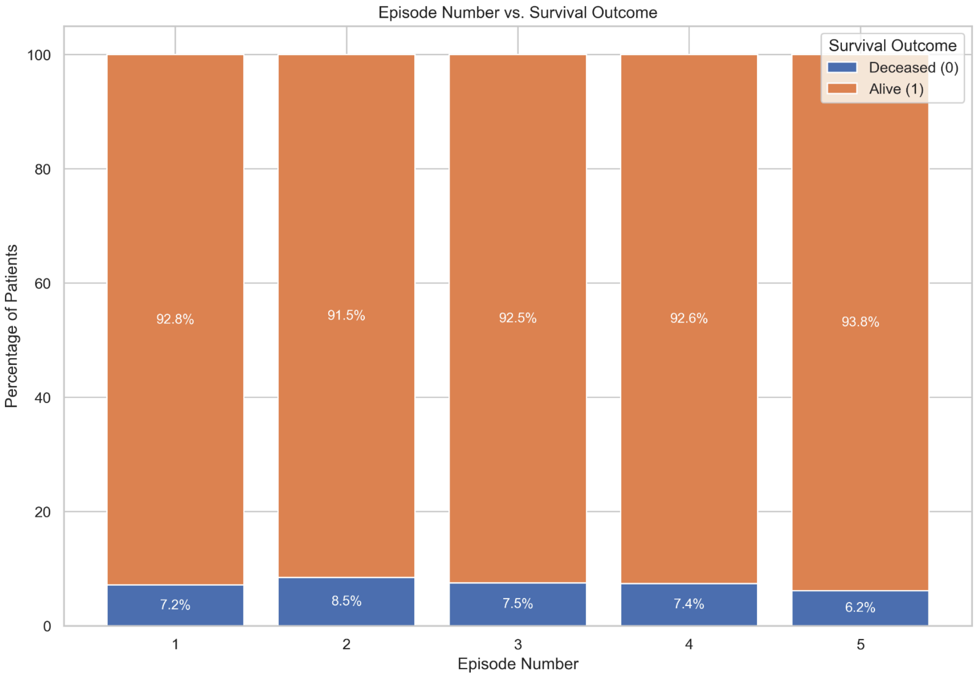


Figure 4. Survival outcomes across episode numbers reveal slight variations, with Episode 5 showing the lowest mortality (6.2%) and Episode 2 the highest (8.5%), while survival rates remain consistently above 91%.

Correlation Analysis

Figure 5 presents the correlation matrix for numerical features and the target variable:

Age: Weak negative correlation with survival (-0.17), indicating older patients are at higher mortality risk.

Sex and Episode Number: Near-zero correlations with survival and each other, suggesting limited linear relationships.

These findings suggest that non-linear interactions may better capture the relationships between features and survival, motivating the use of non-linear machine learning models.

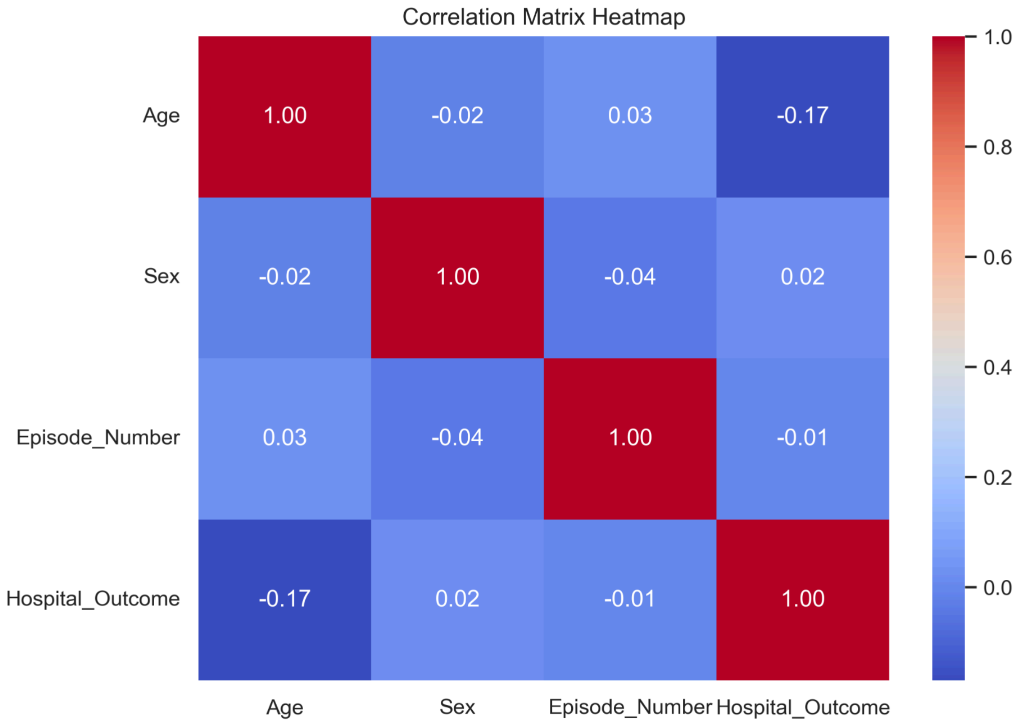


FIgure 5. Correlation matrix heatmap showing weak relationships among features and target variable, with Age exhibiting a slight negative correlation (-0.17) with Hospital Outcome.

3. Methods

Data Preprocessing

A preprocessing pipeline was designed using ColumnTransformer to prepare the dataset for machine learning:

1. Ordinal Features: Episode_Number was encoded using OrdinalEncoder to convert categorical values into numerical format.
2. Scaling: Age was normalized using MinMaxScaler to scale values between 0 and 1.
3. Passthrough: Sex was already numeric and passed through without transformation.

Train-Test Splitting and Cross-Validation

The dataset was split into 80% training and 20% testing using stratified sampling to maintain the class distribution (92.6% alive, 7.4% deceased). To enhance robustness in this highly imbalanced dataset, 4-fold KFold cross-validation was applied during training, repeated across five random states to reduce variability from data splits.

Machine Learning Pipeline

An automated pipeline streamlined the entire process:

1. Preprocessing: Applied the pipeline described above.
2. Feature Engineering: Polynomial features (degree = 2, interaction only) were generated using PolynomialFeatures to capture potential non-linear relationships, excluding bias terms to avoid redundancy.
3. Scaling: Applied StandardScaler globally after feature engineering to ensure consistent scaling across models.
4. Model Training: Preprocessed data was fed into the respective machine learning algorithms.

Models and Hyperparameter Tuning

We evaluated five machine learning models (Table 2), encompassing both linear and non-linear methods. GridSearchCV was used to optimize hyperparameters by performing exhaustive searches over predefined grids tailored to each model.

Evaluation Metric

Due to the significant class imbalance, the Area Under the Precision-Recall Curve (AUC-PR) was chosen as the primary evaluation metric. Unlike AUC-ROC, AUC-PR focuses on the minority class (class 0: deceased), making it more appropriate for imbalanced datasets.
A custom scoring function used precision_recall_curve to compute precision and recall, followed by the auc function to calculate the area under the curve.
The minority class (class 0) was explicitly set as the positive class (pos_label=0) throughout the pipeline.

Handling Class Imbalance

To address class imbalance, models such as Logistic Regression, Random Forest, and SVC were configured with class_weight='balanced', which adjusts weights inversely proportional to class frequencies. The scale_pos_weight parameter in XGBoost was set to the ratio of majority to minority class instances to balance the loss function effectively.

ML Model	Hyperparameters
Logistic Regression	'penalty': ['elasticnet’], 'C': [1e-2, 1e-1, 1e0, 1e1], 'l1_ratio': [0, 0.25, 0.5, 0.75, 1]
Random Forest Classifier	max_depth: [2, 3, 4, 5, 6], max_features: [0.7, 0.75, 0.8, 0.85, 0.9]
Kneighbors Classifier	n_neighbors: [1e1,1e2,1e3], metric: ['euclidean', 'manhattan'], weights: ['uniform']
XGBoost Classifier	reg_alpha: [1e0, 1e1, 1e2], reg_lambda: [1e-2, 1e-1, 1e0, 1e1], max_depth: [1,3,5,7]
Support Vector Classifier	gamma: [1e-1, 1e0, 1e1], C: [1e-1, 1e0, 1e1]

<Table 2. ML Models and their Corresponding Hyperparameters

4. Results

Model Performance and Comparison to Baseline

The Random Forest Classifier achieved the highest mean AUC-PR score of 0.2159, outperforming other models (KNeighborsClassifier, LogisticRegression, XGBClassifier, and SVC), which hovered around 0.14 (Figure 6). However, no model surpassed the baseline AUC-PR score of 0.537.

The Precision-Recall curve for the Random Forest Classifier (Figure 7) shows an average precision (AP) of 0.11, demonstrating limited success in identifying the minority class due to extreme class imbalance.

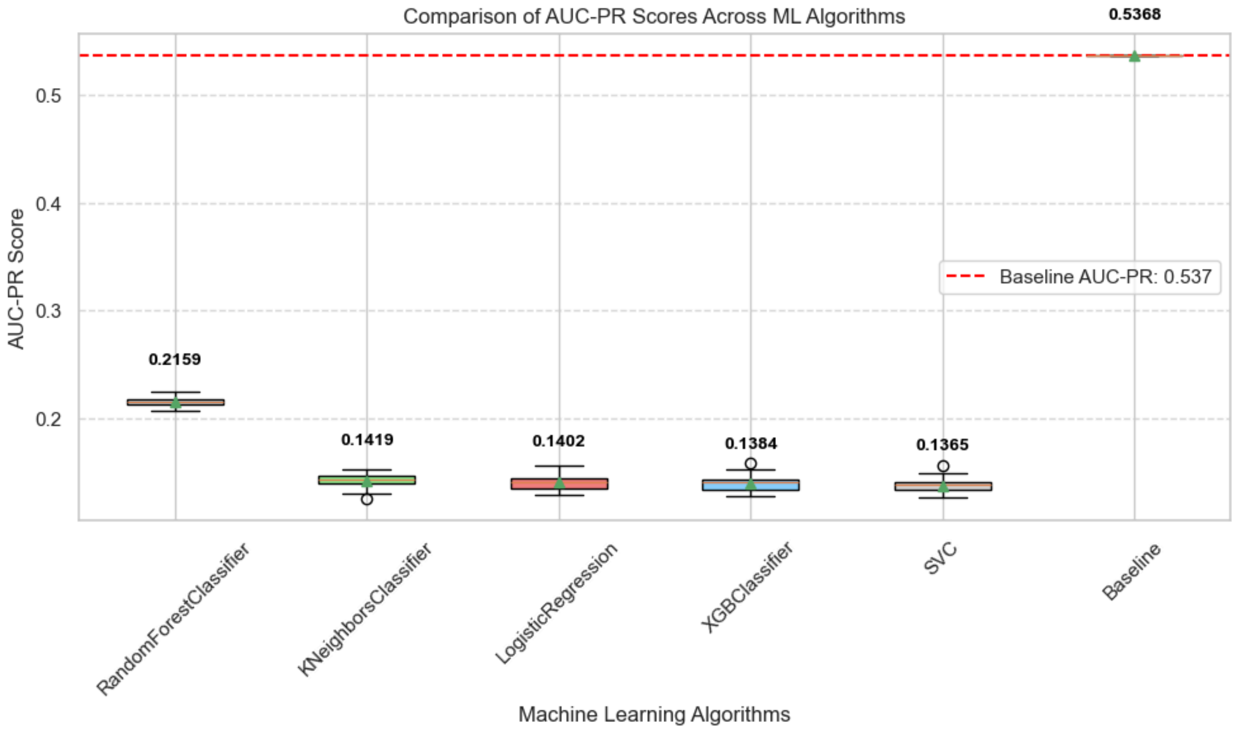


Figure 6. This boxplot compares the AUC-PR scores for multiple machine learning algorithms, including RandomForestClassifier, KNeighborsClassifier, LogisticRegression, XGBClassifier, and SVC.

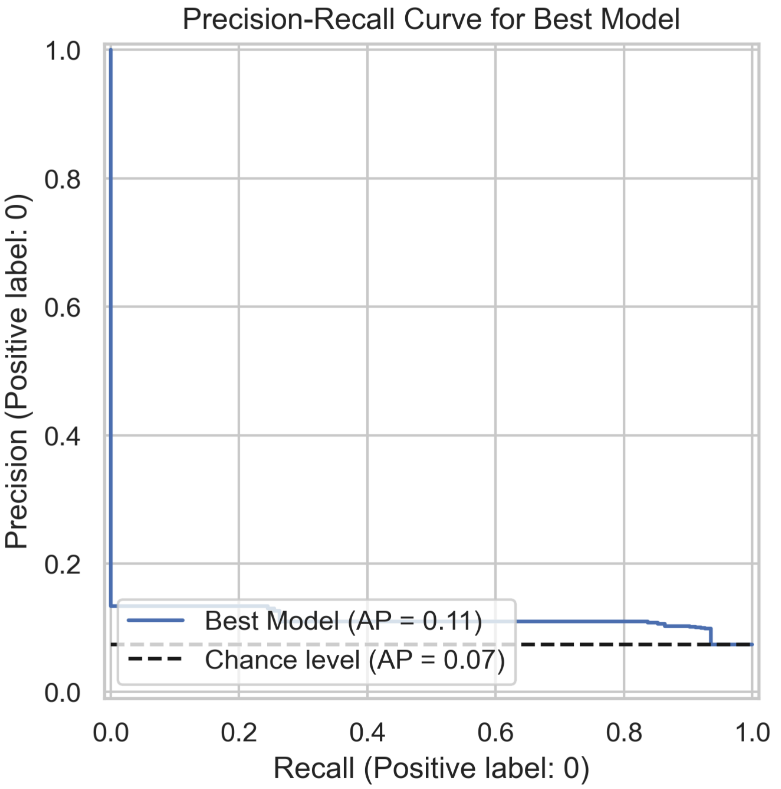


Figure 7. The Precision-Recall (PR) curve shows the performance of the best model (RandomForestClassifier) when predicting the minority class (deceased).

Confusion Matrix Analysis

The confusion matrix (Figure 8) highlights the model's bias toward predicting the majority class:

Class 0 (deceased): The model correctly predicted 1,482 cases but misclassified 144 as Class 1.

Class 1 (alive): While 7,365 cases were correctly identified, 13,078 were incorrectly predicted as Class 0.

These results underscore the challenge of minority class detection in highly imbalanced datasets.

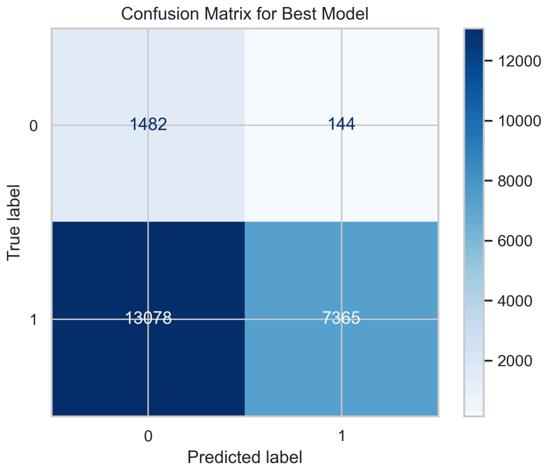


Figure 8. This confusion matrix visualizes the classification results for the best-performing model.

Feature Importance Analysis

Global Feature Importance:

1. Gini Importance (Figure 9): Identified Sex as the most significant feature, followed by Episode_Number, with Age showing minimal importance.
2. Permutation Importance (Figure 10): Contrastingly, Age emerged as the most influential feature, with Sex and Episode Number contributing minimally.
3. SHAP Values (Figure 11): SHAP analysis confirmed Age as the strongest predictor, with higher values increasing the likelihood of class 0 predictions. Sex and Episode_Number showed moderate but inconsistent effects.

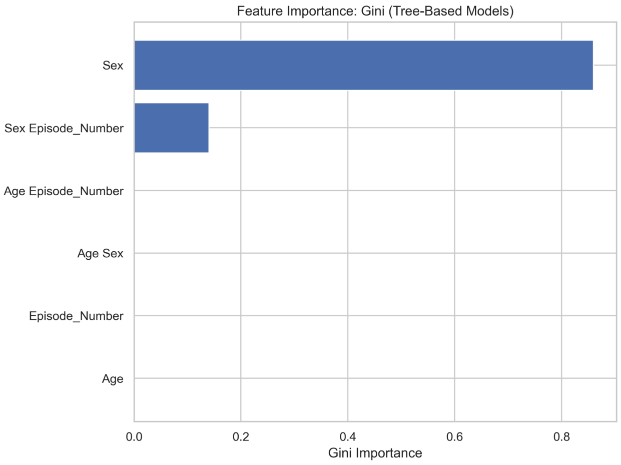


Figure 9. Gini-based feature importance highlights Sex as the dominant feature, with minimal contributions from other features.

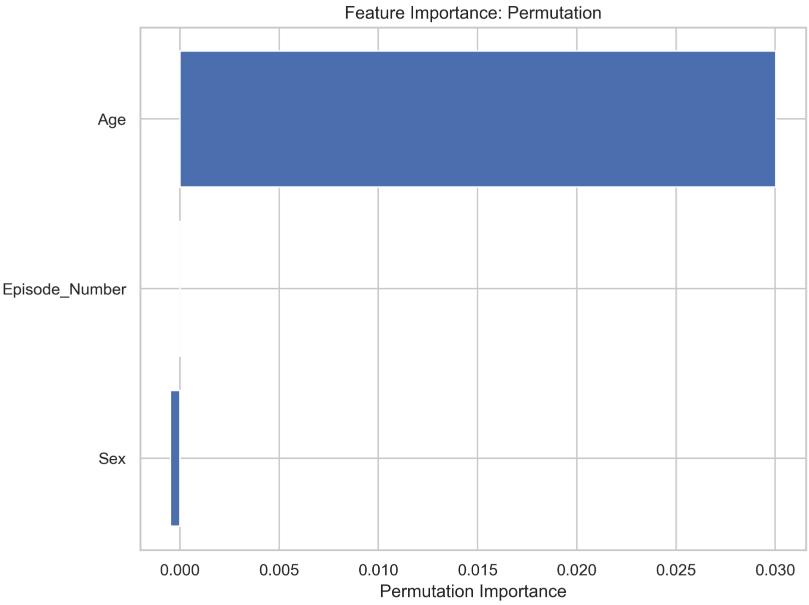


Figure 10. Permutation-based importance identifies Age as the most significant feature

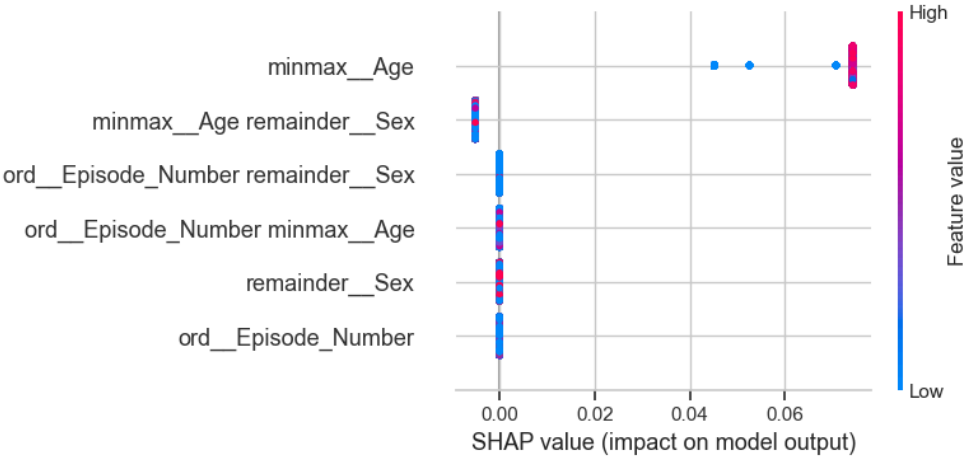


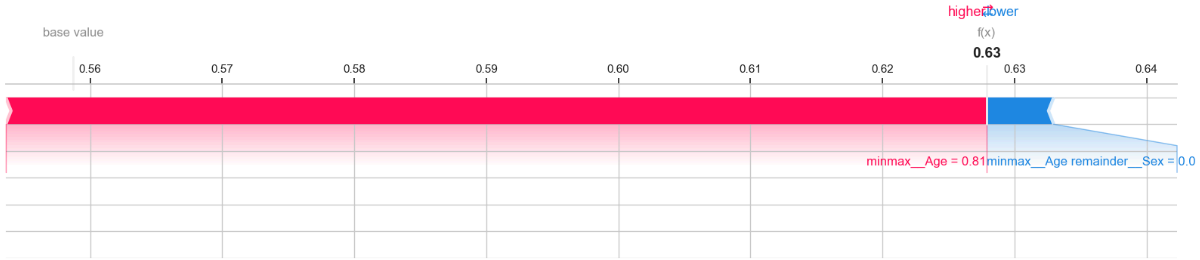
Figure 11. SHAP summary plot highlights Age as the most impactful feature, with interactions involving Sex and Episode_Number showing moderate contributions.

Local Feature Importance:

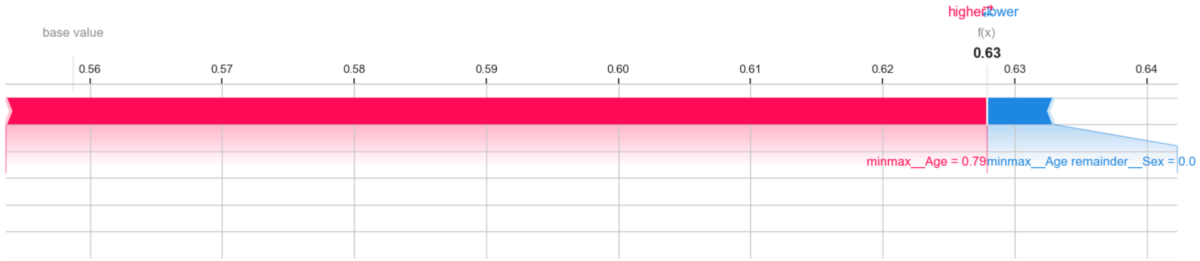
SHAP force plots for two sample instances (indices 10000 and 20000) revealed Age as the dominant feature driving predictions. At index 10000, for example, Age (minmax__Age = 0.81) pushed the prediction upward, while Sex and Episode_Number had negligible effects. Despite slight differences in feature values, predictions for both instances converged to 0.63, indicating similar probabilities.

Summary of Findings

The analysis highlights Age as the most critical predictor both globally and locally, consistent with clinical expectations of higher mortality risks for older patients. However, the models struggled to outperform the baseline, reflecting the difficulty of predicting the minority class in an imbalanced dataset. While Sex showed high importance in Gini-based metrics, its lower impact in permutation and SHAP analyses suggests that Gini importance reflects its role in decision tree splits, whereas permutation and SHAP better capture its actual predictive power.



<Figure 11. Shap value for index 10000>



<Figure 12. Shap value for index 20000>

5. Outlook

1. Oversampling (ROSE):

To address the class imbalance, I plan to implement the Random Over-Sampling Examples (ROSE) technique. ROSE generates synthetic data points for the minority class by interpolating within the feature space, creating a more balanced dataset. This approach is expected to enhance model performance, particularly for predicting the "Deceased" class, by improving sensitivity and ensuring equitable predictions.

2. Advanced Models:

Future work will explore advanced machine learning models to improve accuracy and robustness. These include linear and radial SVM for their strong decision boundary capabilities, gradient boosting for iterative error minimization, and Naïve Bayes, which is well-suited for imbalanced data. Additionally, neural networks can be applied to capture complex, non-linear relationships, offering the potential for significant performance gains.

3. Threshold Tuning:

I plan to incorporate threshold tuning to optimize the classification decision boundary. By adjusting the threshold based on precision-recall trade-offs or other evaluation metrics, the model can prioritize recall for the "Deceased" class while maintaining acceptable precision. This strategy ensures the model aligns with real-world goals, such as minimizing false negatives for critical outcomes.

References

1. World Health Organization. WHO: Improving the Prevention, Diagnosis and Clinical Management of Sepsis. https://www.who.int/sepsis/en/. Accessed 23 February 2020.
2. Rudd, K. E. et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. Lancet 395, 200–211 (2020).
3. Leligdowicz, A. & Matthay, M. A. Heterogeneity in sepsis: new biological evidence with clinical applications. Crit. Care 23, 80 (2019).
4. Arnold, C. News feature: the quest to solve sepsis. Proc. Nat. Acad. Sci. 115, 3988–3991 (2018).
5. Chicco, D. & Jurman, G. (2020). Sepsis Survival Minimal Clinical Records [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C53C8N.