Report: Sepsis Survival Prediction Model

This study addresses the critical need for a rapid, interpretable predictive model for patient survival in the context of a highly imbalanced Sepsis Survival dataset. A Logistic Regression (LR) model was strategically employed, leveraging only minimal clinical records—Age, Sex, and Number of Prior Sepsis Episodes—to ensure immediate clinical utility. By incorporating the *class_weight='balanced*' parameter, the model achieved a robust Area Under the Precision-Recall Curve (AP) of 0.956. Given the severe class imbalance (91% Alive vs. 9% Dead), this high AP score is most likely attributed to the model's strong performance on the majority class ('Alive'). The model remains highly interpretable, confirming Age($\beta=-0.8731$) as the most critical negative predictor of survival, valuable for rapid triage. However, the modest Balanced Accuracy (0.629) and Weighted F1-Score (0.689) confirm that performance on the clinically essential minority class ('Dead') requires improvement. Future work must focus on data augmentation with additional clinical markers and the exploration of more complex, non-linear models to boost overall accuracy and ensure higher recall for high-risk patients.

Context & Motivation

The primary objective of this project is to develop and validate an interpretable machine learning model—specifically, a Logistic Regression model—capable of predicting a negative patient outcome (30-day mortality) in a healthcare support application.

Given the time-sensitive and critical nature of Sepsis, where every hour of delayed treatment increases mortality risk, the model serves as an early alert system. By using minimal clinical records, the application aims to flag patients at the highest risk of mortality. This early warning enables healthcare providers to prioritize interventions, allocate critical resources, and apply life-saving protocols more efficiently, ultimately supporting evidence-based clinical decision-making and improving patient survival rates.

Problem Definition

Sepsis is a life-threatening condition defined as organ dysfunction caused by a dysregulated host response to infection. It is a global health crisis characterized by high incidence, severe morbidity, and high mortality.

Expected Social Impact

SDG 3: Good Health and Well-being

The project directly addresses the global imperative to improve health outcomes, particularly by tackling one of the most common causes of preventable death in hospital settings.

- SDG 3.2: By 2030, end preventable deaths of newborns and children under 5 years of age.
- Sepsis is a major killer of children; models that increase early detection can directly reduce pediatric mortality.SDG 3.4: By 2030, reduce by one third premature mortality from non-communicable diseases through prevention and treatment.
- While Sepsis is an infectious process, reducing its mortality contributes directly to the overall reduction of premature, avoidable deaths in the healthcare system.SDG 3.d:Strengthen the capacity of all countries... for early warning, risk reduction and management of national and global health risks.

The predictive model acts as an AI-powered early warning system, enhancing the capacity of healthcare providers to manage Sepsis risk.

Dataset Description

This dataset is designed for binary classification, with the primary goal of predicting the survival outcome of patients diagnosed with or at risk of sepsis using a minimal set of clinical features that are easy to collect quickly.
Source: UCI Machine Learning Repository (Originally published in Scientific Reports by Chicco and Jurman, 2020).

Avaliable here:
https://archive.ics.uci.edu/dataset/827/sepsis+survival+minimal+clinical+records

Key Features and Design Principles:

- hospital_outcome_1alive_0dead: Target
  - The survival status of the patient after approximately 9 days of admission.
- age_years: Feature
  - Age of the patient
- sex_0male_1female: Feature
  - Gender of the patient
- episode_number: Feature
  - Number of prior septic episodes the patient has experienced.

Data Quality and Distribution Analysis

1. Data Aggregation and Volume

The project utilizes data derived from three separate cohorts, which were aggregated to form a comprehensive dataset for modeling.

- Combined Instance Count: After concatenating the three cohorts, the final dataset contains 129391 patient instances. This large volume of data provides a solid foundation for training a robust predictive model.

2. Missing Data Integrity

A critical step in the data preparation phase confirmed that the final working dataset is complete:

- Finding: No missing data points (NaNs or Nulls) were found across the four primary clinical features (Age, Sex, Number of Prior Sepsis Episodes, and the derived interaction term).
- Implication: No need for imputation techniques (like mean or median substitution), thereby preventing the introduction of artificial variance or bias into the model's training process.

3. Duplicated Data Strategy

The decision regarding duplicate rows was strategic, directly addressing the nature of the data:

- Finding: Duplicate instances were intentionally retained in the dataset.
- Rationale: The raw data often contains multiple records for probably different patients but with the same features. By not dropping these records, the analysis preserves the natural distribution and real-world frequency of the clinical phenomena being studied. This approach ensures the model is trained on data that accurately represents the heterogeneity and recurrence observed in a typical clinical environment.

4. Data Distribution and Imbalance

The most significant structural feature of the target variable is its extreme imbalance:

- Target Variable: outcome (0 = Dead, 1 = Alive).
- Imbalance: The dataset exhibits a severe class imbalance, with the target distribution being approximately:
  - Class 1 (Alive): ≈91% of instances (Majority Class)
  - Class 0 (Dead): ≈9% of instances (Minority Class)

- Implication for Modeling: This imbalance makes the prediction of the minority class ('Dead') significantly challenging. Standard accuracy metrics would be misleading (e.g., a model predicting everyone 'Alive' would have 91% accuracy). This is the core justification for the strategic model choices:
    - Using class_weight='balanced' in the models to prioritize the minority class.
    - Prioritizing the Recall over Accuracy for evaluations.

Advanced Preprocessing and Model Evaluation Strategy

1. Data Balancing Technique: Tomek Links

To address the severe 91%/9% class imbalance, an advanced under-sampling technique, Tomek Links and SMOTE, was employed.

- Those methods refined the class boundaries, making the majority and minority classes more distinct and improving the quality of the data structure without drastically reducing the total number of majority instances and increasing minority instances.

2. Comprehensive Model Comparison

A critical component of the analysis involved a comparison of model performance across different data states to validate the final model choice:

- Model Spectrum: A wide array of classification models were tested, including both simple linear models (Logistic Regression) and complex non-linear ensemble models (e.g., Random Forest, XGBoost).
- Data States Evaluated: Each model was evaluated under two distinct data conditions:
    1. Imbalanced Data: Using the original 91%/9% distribution.
    2. Balanced Data: Using the data refined by SMOTE/Tomek.

3. Justification for Final Model Selection

To select the final model, several linear and non-linear classifiers were evaluated across both the original imbalanced data and the data cleaned using the SMOTE/Tomek Links technique.

Crucially, Logistic Regression and Decision Tree with balanced data, as well as Gradient Boosting (both states), SVM (balanced), and Random Forest (balanced) were unable to correctly make predictions for the critical minority class (Class 0, 'Dead'), rendering them unusable for high-risk triage.

The metrics below are for the models that successfully provided meaningful predictions across both classes:

Model	Data State	W. F1-Score	Balanced Accuracy	Primary Limitation / Status
KNN	Balanced (Tomek)	0.8611	0.5155	DISCARDED: Fails to predict minority class (Class 0), showing extreme bias.
Logistic Regression	Imbalanced*	0.6892	0.6293	FINAL CHOICE: Best balance of interpretability and honest prediction.
Random Forest	Imbalanced	0.6064	0.6334	Non-linear complexity without significant performance gain over LR.
Decision Tree	Imbalanced	0.5935	0.6311	Low F1-Score, similar Balanced Accuracy to LR.
Other Models (SVM, XGBoost)	All States	(Failure)	(Failure)	Unable to correctly predict the critical minority class (Dead).

∗ *Logistic Regression was run on imbalanced data but utilized the key hyperparameter class_weight='balanced'.*

The final selection of the Logistic Regression model is based on a rigorous analysis of classification integrity, prioritizing Balanced Accuracy and Clinical Interpretability over misleading aggregate metrics.

1. Rejection of KNN: The KNN model, despite achieving the highest Weighted F1-Score (0.861), was rejected because its extremely low Balanced Accuracy of 0.5155 confirms that it failed to learn the minority class ('Dead') and was

essentially predicting the majority class ('Alive') almost exclusively. In a clinical context, a model that fails to identify mortality risk is useless and dangerous.

2. Superior Classification Integrity: Logistic Regression (LR) achieved the highest Balanced Accuracy (0.6293) among all viable models. This indicates that LR provided the most honest and reliable performance across both the majority and minority classes, a critical requirement for a high-consequence prediction task.
3. Preservation of Interpretability: LR successfully balances this superior classification integrity with the project's primary strategic goal: Maximal Clinical Interpretability. The model provides clear, quantifiable coefficients (Odds Ratios) that explain the risk, a feature lacking entirely in the high-performing non-linear alternatives (like Random Forest).

Clinical Findings: Feature Importance

The model's coefficients provide direct, actionable intelligence on patient risk:

Feature	Coefficient (β)	Odds Ratio (exp(β))	Clinical Interpretation
Age	-0.8731	0.4177	Most Significant Negative Predictor. Age is the strongest factor correlated with decreased odds of survival, aligning with established academic literature on reduced physiological reserve in older patients.
Female	+0.2306	1.2594	Positive Survival Factor. Female patients have a moderate increase in the odds of survival compared to male patients, it could be biased because of imbalance between number of male and female data.
n_prior_sepsis	-0.2135	0.8078	Moderate Negative Factor. A history of multiple septic episodes reduces the odds of survival, indicating increased frailty and cumulative organ damage.

Recommendations for Future Development

To elevate the model from a high-quality triage tool to a comprehensive predictive system, strategic investment is required:

1. Data Augmentation: New data collection is necessary, focusing on introducing additional minimal clinical data, such as vital signs or basic lab values (e.g., lactate). These features are readily available and known to significantly improve sepsis prediction.
2. Addressing Distribution and Interaction Variance: Introducing new variables will change the data distribution and feature interactions, making possible more sophisticated non-linear modeling techniques(e.g., Gradient Boosting or Neural Networks) to predict caught those interactions. These models can better capture the complex, varying relationships between an expanded set of clinical risk factors, which is essential to improve recall and overall accuracy in diverse patient populations.