Predicting Mortality Using Healthcare Data

1st Sogand Soghrati Ghasbeh
PhD student at Industrial and Operations engineering
University of Michigan
Ann Arbor, USA
soghrati@umich.edu

***Abstract*—Accurate mortality prediction is a critical task in healthcare, enabling clinicians to identify high-risk patients and allocate resources effectively. In this study, we developed and evaluated machine learning models to predict patient mortality using clinical data. A Random Forest Classifier was employed as the best performing model, achieving an Area Under the Curve (AUC) of 0.84. Feature engineering, including the creation of composite variables, and synthetic oversampling techniques were utilized to improve model performance and handle class imbalance. Feature importance analysis identified key predictors, such as bicarbonate levels, anion gap, and leucocyte count, which align with clinical relevance. The results demonstrate the potential of machine learning models to enhance decision-making and improve patient outcomes in critical care settings.**

***Index Terms*—Mortality Prediction, Machine Learning, Feature Engineering, Healthcare Analytics**

I. Introduction

Mortality prediction is a critical challenge in modern healthcare, with significant implications for patient care, resource allocation, and decision-making. Hospitals often operate under constrained resources, such as intensive care unit (ICU) beds and medical staff, particularly during times of high demand, such as pandemics or emergencies. Accurate mortality prediction enables healthcare providers to prioritize high-risk patients, ensuring timely interventions and improving patient outcomes. Furthermore, identifying key factors contributing to mortality offers valuable clinical insights for guiding treatment strategies and preventive measures.

This project aimed to develop a robust predictive framework for mortality using a publicly available healthcare dataset. Key challenges, including missing data, class imbalance, and complex relationships among variables, were addressed through pre-processing techniques and machine learning methods. This project also emphasizes the importance of feature engineering to simplify the models while maintaining strong predictive performance. By deriving meaningful features, we demonstrate that fewer, more interpretable features can provide insights into patient outcomes while reducing complexity.

II. Literature Review

With significant advancements and developments of robust machine learning models, the prediction of mortality using healthcare data has been extensively studied. For example, [3] developed a real-time machine learning model for short-term mortality prediction in critical care, using ensemble methods such as Random Forests and Gradient Boosting to model complex relationships among clinical features. Similarly, [2] validated a mortality prediction model for older adults across diverse cohorts, emphasizing the importance of external validation to ensure the generalizability of findings.

Our project aims to enhance the models presented in existing literature by developing a new machine learning algorithm that surpasses their performance.

III. Method

A. Dataset Description

The dataset used in this project was obtained from the HuggingFace platform which is accessible via this link. It includes variables such as age, gender, pre-existing conditions, vital signs, and lab test results. The target variable is *mortality*, which indicates whether a patient survived or not.

B. Problem Formulation

The mortality prediction problem is formulated as a binary classification task. The input consists of patient features, including demographic and clinical data, while the output is the predicted probability of mortality. The primary objective is to maximize the predictive performance of the model, particularly in terms of the Area Under the Curve (AUC) and Accuracy metric.

C. Preprocessing

To prepare the dataset for modeling, several pre-processing steps were implemented; Missing values were imputed using statistical techniques, such as mean imputation for numerical variables and mode imputation for categorical variables. For more advanced models, new features were created as this project emphasizes the importance of feature engineering to simplify the model while maintaining strong predictive performance. Lastly, for the last model, as the dataset exhibited a significant class imbalance, oversampling techniques were applied to balance the classes.

D. Modeling Approach

The baseline model was a simple logistic regression to establish a reference performance. Advanced models, such as a logistic regression model after data pre-processing steps, and two Random Forest models with and without data balancing, were subsequently implemented to improve predictive accuracy. Model evaluation was conducted using the AUC and Accuracy metric, and feature importance analysis was

performed to identify key predictors. The AUC score, provides a single value summarizing the classifier's performance across all thresholds. It essentially measures how well the model distinguishes between positive and negative instances [1]. Typically, a model with an AUC above 0.8 is considered to perform well.

On the other hand Accuracy measures how accurate our model is in making predictions correctly. Mathematically, this metric is calculated as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

where TP stands for True positive, TN stands for True Negative, FP is False Positive, and FN shows False Negative values.

E. Tools

The project leveraged various Python libraries: **scikit-learn** for data preprocessing, model building, and evaluation, **imbalanced-learn** to handle class imbalance, **NumPy and Pandas** for data manipulation and exploratory analysis, **Matplotlib and Seaborn** for visualizations and data interpretation, and lastly, **SQLite3** for data processing and feature engineering.

RESULTS

The baseline model, a simple logistic regression model was developed first, achieving an Area Under the Curve (AUC) of **0.79**. To improve upon this, an analysis of feature correlations revealed a strong relationship between **Hematocrit** and **Red Blood Cell (RBC)**. These two variables are highly correlated because both measure aspects of red blood cells in the blood. To address this redundancy, feature engineering was applied to create a new feature representing the ratio of hematocrit to RBC count.

Additionally, instead of using individual complication indicators as separate features, a new feature was engineered to capture the total number of comorbidities. This approach reduces feature dimensionality while retaining critical information about patient conditions.

After implementing these enhancements, a new logistic regression model was fitted, followed by a Random Forest Classifier to compare their predictive power and accuracy. The results of these models, including their performance metrics, are presented in Table I.

Based on the results, after incorporating feature engineering, the logistic regression model showed improved performance, achieving an AUC of 0.80 and an accuracy of 0.86. The Random Forest (RF) classifier outperformed the baseline logistic regression with an AUC of 0.84 and an accuracy of 0.85. However, when the dataset was balanced using synthetic oversampling, the Random Forest model exhibited a slight drop in AUC to 0.81 while maintaining an accuracy of 0.85. These results highlight the importance of feature engineering and balancing techniques in enhancing model performance.

TABLE I
REGRESSION MODELS' METRICS

Model	*AUC*	*Accuracy*
Baseline LR[a]	0.79	0.85
Feature Engineered LR	0.80	0.86
RF[b]	0.84	0.85
Rf with balancing	0.81	0.85

[a]Logistic Regression, [b]Random Forest

The AUC plot of the best performing model is presented in Figure 1.

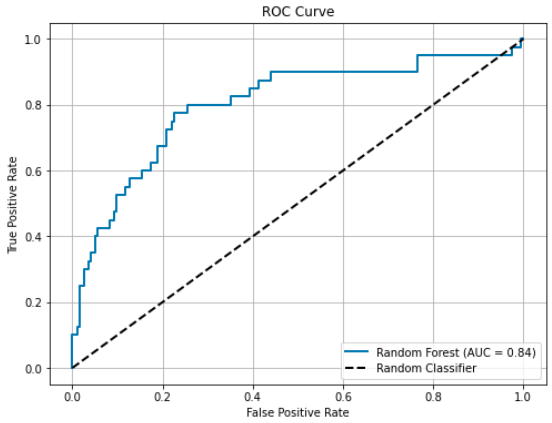

Fig. 1. Random Forest Model AUC

The feature importance for the best-performing model, is shown in Figure 2. The top features contributing to the model's predictive performance include Bicarbonate, Anion gap, and Leucocyte count, which align with clinical expectations as key indicators of patient health. Other significant features, such as Lymphocyte count, Urine output, and Lactic acid levels, further highlight the model's ability to capture complex relationships between variables and mortality risk.

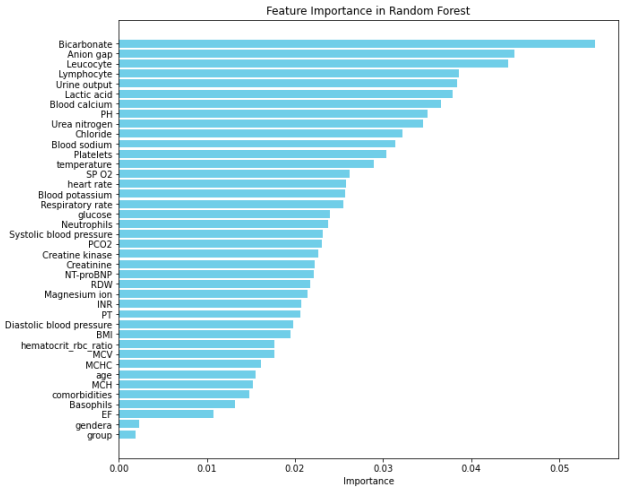

Fig. 2. Random Forest Model Feature Importance

Conclusion

This project demonstrates the utility of Logistic Regressions and Random Forest Classifiers in predicting patient mortality. The models achieved high AUC and Accuracy, underscoring their effectiveness in identifying high-risk patients. By leveraging feature importance analysis, the model provides insights into the factors most strongly associated with mortality, including Bicarbonate and Anion gap. By creating composite features, such as the hematocrit-to-RBC ratio and number of comorbidities, we demonstrated that it is possible to reduce the number of features without sacrificing performance. This approach not only simplifies the model but also aligns with clinical reasoning, making the predictions more meaningful and actionable.

Future work could focus on incorporating temporal trends in the data, such as changes in vital signs over time, or exploring ensemble techniques to improve predictive performance further. These extensions could enhance the model's utility in real-world healthcare decision-making.

References

[1] Bradley, AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern recognition. 1997 Jul 1;30(7):1145-59.

[2] Han, D. S. et al. Development and external validation of a mortality prediction model for older adults using two nationally representative cohorts. JAMA Internal Medicine, 2023; 183(3):211–220.

[3] Mao, Q. et al. Real-time machine learning model for short-term mortality prediction in critically ill patients. Critical Care, 2024; 28(1):122.