COMS- W4995 AML PROJECT
PROJECT DELIVERABLE #3 - FINAL REPORT DELIVERABLE
GROUP 3

Atharva Kulkarni (ak5070), Divya Tadimeti (dt2760),
Xiaoyi Zhou (xz3126), Yolanda Zhu (xz3013), Yu Yen Liu (cl4202)

1. Introduction

Our project focuses on developing and evaluating machine-learning models capable of predicting in-hospital mortality for heart failure (HF) patients in the Intensive Care Unit (ICU). Given the severeness of underlying diseases in HF patients and the limitation in medical understanding regarding the survival outcomes of ICU-admitted HF patients, this initiative has great significance. By gaining a deeper insight into the variables affecting these outcomes, our models aim to help physicians in making more informed decisions about patient prognosis. This predictive approach is important, especially considering the complexity involved in treating heart failure, where even intensive medical treatments don't always offer clear insights into patient survival during hospitalization.

In our data analytical process, we initially identified and dropped features that had more than 10% missing values. We assessed the correlations among various features and eliminated those that were highly correlated to streamline our dataset subsequently. Before using machine learning models, we divided the dataset into two parts: a development set and a test set. Our approach involved experimenting with tree-based and boosting models to fit our data and generate predictions. Ultimately, it was observed that the boosting models outperformed the tree-based models in terms of prediction accuracy.

2. Data Description

The dataset contains 3 kinds of variables: Demographic characteristics, vital signs, and laboratory variables. The target variable is Outcome (in-hospital mortality), defined as the vital status at the time of hospital discharge in survivors and non-survivors: 0 stands for Alive, and 1 stands for Death. Among 48 features in the raw dataset, 10 of them are categorical variables, the left are numerical variables. This dataset is highly imbalanced since the 'Alive' category takes the vast majority of the target as seen in Fig. 1, so in the subsequent analysis, we tried to utilize the oversampling and SMOTE techniques to handle this issue.

For missing values within the dataset, we found that all features have less than 25% missing values as seen in Fig. 2 but for the robustness of our analysis, we decided to drop all features with more than 10% missing values. Then we removed features exhibiting high correlations with others, specifically those with correlation coefficients exceeding 0.7. This led to the drop of RBC, MCV, and INR from our dataset. The rationale behind this decision is likely rooted in the intrinsic medical relationships these features share.

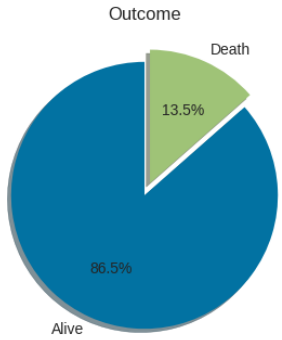

Fig. 1: Ratio of the target value (Outcome)

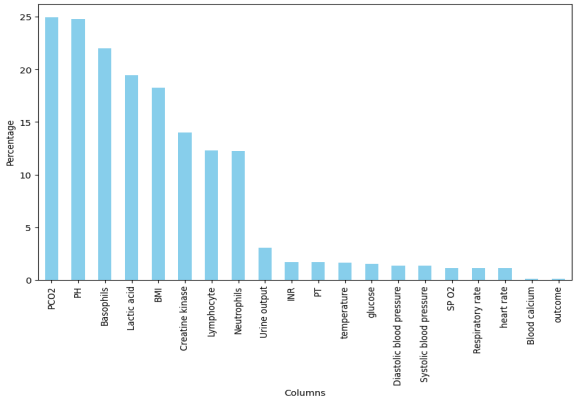

Fig. 2: Percentage of missing values per column

3. Methodology

We aim to find the model with the best performance, our methodology firstly, involves making sure that our imbalanced dataset is ready to train and test our models. We use multiple techniques to oversample the data namely random oversampling, SMOTE. Secondly, we implement the decision tree, random forest, Gradient Boosting, and XGBoost and also explore PCA to check if it helps us achieve our goal.

In this project, we initially assessed the performance of a Default Decision Tree (DT) model on an imbalanced dataset, yielding a mean AUC of 0.5426 and a mean Average Precision of 0.2153 using 5-fold Cross Validation. Addressing the class imbalance through Random Oversampling significantly improved the DT classifier's performance, resulting in a mean AUC of 0.7852 and mean Average Precision of 0.4607. Further enhancement was achieved by employing Synthetic Minority Oversampling Technique (SMOTE) on the development dataset, yielding a shape of 1862 instances with positive labels (845) and negative labels (1017). The DT classifier improved, with a mean AUC of 0.8711 and mean Average Precision of 0.8411.

Subsequently, we chose the SMOTE dataset for further analysis, where the DT model exhibited promising results with an accuracy of 0.90, precision of 0.90, recall of 0.28, and an F1-Score of 0.43. Moving on to ensemble methods, Random Forests showcased accuracy of 0.907, a perfect precision, and AUC score in 0.656 performance with feature importance attributed to Hypertensive, Anion Gap, and Atrial Fibrillation. We had to perform hyperparameter tuning since the model was overfitting. Similarly, Gradient Boosting and XGBoost demonstrated strong results, with key features such as Hypertensive, Atrial Fibrillation, and Gendera for Gradient Boosting, and Depression, Renal Failure, and Urine Output for XGBoost. For XGBoost, we performed hyperparameter tuning and observed minor fluctuations in the scores which led to perfect scores. We think that this fluctuation is perhaps caused by relatively small dataset size

We also tried exploration of Principal Component Analysis (PCA) involving standard scaling and stratified splitting, retaining 95% of the variance. However, the application of PCA yielded mixed results for DT, Random Forest, Gradient Boosting, and XGBoost classifiers. Notably, the Gradient Boosting model displayed a decrease in performance, indicating that the introduction of PCA may not universally enhance the performance of tree-based models. Basically, we compare and analyze various oversampling techniques, particularly SMOTE, in addressing class imbalance and enhancing performance. Ensemble learning methods, namely Random Forests, Gradient Boosting, and XGBoost, exhibited robust performance on the SMOTE dataset. The introduction of Principal Component Analysis, while beneficial in some cases, did not necessarily improve the performance of tree-based models, emphasizing the importance of considering dataset-specific characteristics when applying dimensionality reduction techniques.

Model Name	Accuracy	Precision	Recall	F1-Score	ROC- AUC
Decision Tree	0.90	0.90	0.28	0.48	0.64
Random Forest	0.91	1.00	0.31	0.48	0.66
Gradient Boost	0.928	0.941	0.500	0.653	0.75
XGBoost	1.00	1.00	1.00	1.00	1.00

Table 1: Model performance comparison.

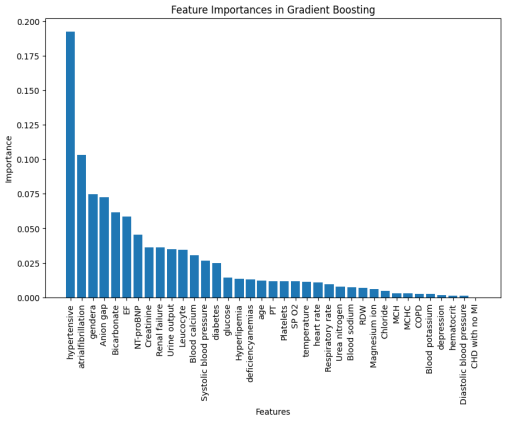

Fig. 3: Feature Importance plot for Gradient Boosting

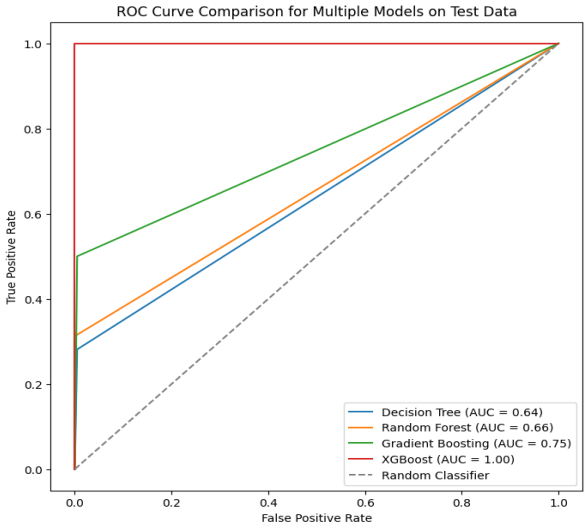

Fig. 4: ROC Curve for Model Comparison

4. Conclusion

In conclusion, our project focused on developing machine-learning models for predicting in-hospital mortality in heart failure patients within the ICU. The initial challenges of handling a highly imbalanced dataset were successfully addressed through oversampling techniques such as random oversampling and SMOTE. Our analysis revealed that boosting models, specifically Gradient Boosting and XGBoost, outperformed other tree-based models in terms of prediction accuracy on the SMOTE dataset. Furthermore, feature selection played a crucial role in model performance, with Gradient Boosting and XGBoost highlighting the significance of variables like Hypertensive, Atrial Fibrillation and Gendera as seen in the Fig.3.The exploration of Principal Component Analysis (PCA) yielded mixed results, indicating that its application may not universally enhance the performance of tree-based models.

We compare the ROC curve for models based on SMOTE data as seen in Fig.4. We can say that factors such as Hypertensive and Atrial Fibrillation are really important and should be looked after if we have to really save the patient in the ICU. To conclude, our work highlights the importance of approaches in addressing class imbalance and the need for careful consideration of feature selection techniques. Through ensemble learning methods in our analysis, particularly on the SMOTE dataset, we have provided valuable insights for future applications in predicting outcomes for heart failure patients in the ICU.