**Final Project Report - HealthCare Analytics**

**Group Number: 1**

**Group members:**

**Quang Nguyen (0794134)**

**Sai Susmitha Vanarasa (0793957)**

**Mounika Loka (0787475)**

**Aleena Alex Almeida (0788193)**

## **Title**

The primary objective of this report is to predict the rate of hospital mortality in intensive care units (ICU) for heart failure by validating the data and building a model. In addition, we seek to understand the critical aspects that contribute to a patient's survival and to investigate its link with some major risk factors by applying machine learning methodologies.

## **Introduction**

Patients’ mortality prediction models have long been developed to objectively assess the severity of the patients and discuss it with the medical care team to accomplish collaborative treatment. Furthermore, an accurate mortality risk assessment of patients at the time of hospitalization is required for estimating the scale of required medical resources based on the severity of the patient.

While several mortality prediction systems focused on adult critical care unit admissions, the model of death prediction for overall hospitalization was limited. To determine the scale of required medical resources according to the patient’s severity at the time of hospitalization, the mortality prediction model that covers the overall hospitalized population is required.

Machine learning applications that use nonlinear feature extraction in the clinical setting have been found to improve prediction ability and applying these techniques to this problem can result in the generation of an accurate prediction model for hospital mortality prediction. As a result, in this study, we used laboratory data to generate machine learning models of hospital mortality prediction with nonlinear feature extraction, compared them to a logistic regression model that uses linear feature extraction, and attempted to generate an accurate prediction model of hospital mortality at the time of hospitalization.

## **Related Work**

The predictors of in-hospital mortality for HF patients admitted to intensive care units (ICU) remain unknown. The goal of this study was to create and verify a prediction model for all-cause in-hospital mortality in HF patients admitted to the ICU.

Patients who met the inclusion criteria were found in the MIMIC-III database and randomly assigned to derivation or validation groups. In the derivation sample, XGBoost and LASSO

regression models were used to test for independent risk factors for in-hospital mortality. To create prediction models, multivariable logistic regression analysis was employed. The C-index, calibration plot, and decision curve analysis were used to examine the discrimination, calibration, and clinical usefulness of the predictive model. Following a pairwise comparison, the highest performing model was picked to create a nomogram based on the regression coefficients.

In-hospital mortality was 13.52 percent among the 1,177 hospitalizations. The XGBoost, LASSO regression, and GWTG-HF risk score models performed well in both groups. The XGBoost and LASSO regression models also performed well in terms of calibration. In a pairwise comparison, the prediction effectiveness of the XGBoost and LASSO regression models was higher than that of the GWTG-HF risk score model (P0.05). The XGBoost model was selected as our final model because of its more concise and wider net benefit threshold probability range, and it was presented as a nomogram.

## **Methods**

In machine learning, it is generally understood that no single algorithm is superior to the others. Below we present the general idea on how each of these supervised machine-learning algorithms work on the dataset and any assumptions we make in each case.

After finalizing the dataset, our main objective was to understand the dataset clearly,

Perform Exploratory Data Analysis (EDA).

Downloaded the Data set (.csv format) to a local folder, we used Pandas Data frame to stored data. We imported some of the basic libraries during analysis like Pandas, Matplotlib, NumPy, Xgboost, Imblearn, Graphviz. Used NLP algorithms as classification models, and they include Logistic Regression, KNN, SVM, Random Forest, XGBoost.

Table

Description automatically generated

In Data Cleaning, the null values are removed since the health test scores are too important to be replaced by mean or any estimated method.

It is generally known in machine learning that no single algorithm is superior to the others. We present a rough overview of how each of these supervised machine-learning methods works on the dataset, as well as any assumptions we make in each case, below. Our main goal after finishing the dataset was to clearly understand it and perform exploratory data analysis (EDA).

In EDA, we have answered the below questions:

* Which age group is most in the hospital?

Chart, histogram

Description automatically generated

The age group of 89-90 are most in the hospital.

* Which age group of patients dies more in the hospital?

Chart, sunburst chart

Description automatically generated

About 15% of patients died in the hospital and remaining are alive.

* Which gender is the most prevalent in the hospital?

Chart, sunburst chart

Description automatically generated

Gender male is the most prevalent in the hospital.

* What is the rate of non-survived patients with diabetes?

Chart, sunburst chart

Description automatically generated

There is 54.2% rate of non-survived patients with diabetes.

* What is the rate of non-survived patients with hypertension?

Chart, sunburst chart

Description automatically generated

The rate of 29.7% of patients are non – survived with hyper tension.

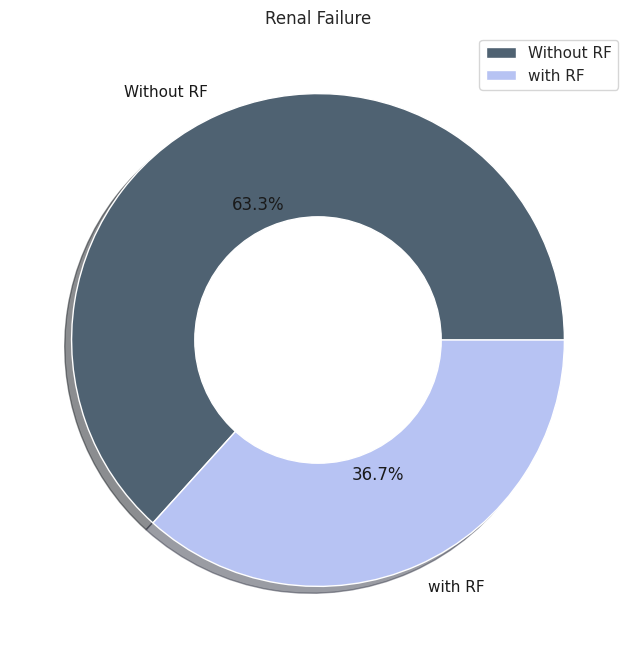
* How many patients in the hospital have depression?

Chart, sunburst chart

Description automatically generated

About 86.9% of patients in the hospital have depression.

* How many patients Alive in the hospital they are with renal failure?

Chart

Description automatically generated

About 36.7% of patients with Renal failure are alive in the hospital.

* How many patients Alive in the hospital they are with Anemia?

Chart, pie chart

Description automatically generated

By using a machine learning random forest classifier we can see the different Survivals Based on a few important features from the above plots.

We created a heatmap of the correlation matrix to analyze the relationships between the various features and the target variable.

Correlation matrix:

Chart, treemap chart

Description automatically generated

Confusion matrix:

Graphical user interface, text, application, chat or text message

Description automatically generatedChart, treemap chart

Description automatically generated

Weight:

Chart, table

Description automatically generated

There were different methods to find the feature importance’s like select k best, Chi2, select from model, rfe, rfecv & etc... for this study we have used feature importance’s from XGBoostClassifier model.

As per the above mentioned 18 variables were explaining 76% of the variation, out of these 18 variables, the first 9 variables explain 56% of the variation.

Data Processing:

Below is the heat map with null values.

Graphical user interface

Description automatically generated with medium confidence

* The original shape of dataset is (1177, 51)
* Drop Group column because it has no meaning
* We decided to drop all rows with nulls
  + df = df.dropna(how='any',axis=0)
* After dropping, the shape of dataset is (428, 50)

## **Methods and Modelling:**

**XGBoost before SMOTE:**

XGBoost is an open-source Python library that provides a gradient boosting framework. It helps in producing a highly efficient, flexible, and portable model. When it comes to predictions, XGBoost outperforms the other algorithms or machine learning frameworks. This is due to its accuracy and enhanced performance.

Text

Description automatically generated

Table

Description automatically generated

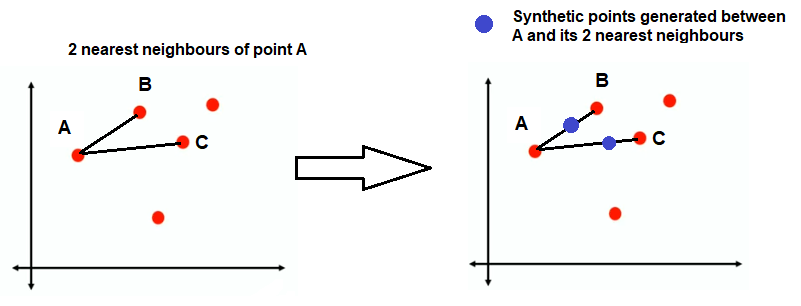
**SMOTE(Synthetic Minority Oversampling) Technique:**

There is a data imbalance issue. The percentage of live patients is always greater than the percentage of died patients. Due to this imbalance, the model tends to predict patients to be alive. Solution is to balance data between majority class(live) and minority (died) by using SMOTE technique.

Graphical user interface, text, application, chat or text message

Description automatically generated

* After dropping nulls and before SMOTE, the shape is (428, 50)
* After SMOTE, the shape is (726, 50)



**All Modeling Techniques after SMOTE:**

**KNN:**

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. In KNN, no model learning is required, and all work is done when a prediction is required. As a result, KNN is frequently referred to as a lazy learning algorithm.

**Table

Description automatically generated**

**SVM:**

SVM is a supervised learning method that analyzes data and categorizes it into one of 2 categories. A linear discriminative classifier seeks to construct a straight line between two sets of data to create a classification model. It simply finds a line or curve or a manifold that divides the classes.

**Table

Description automatically generated**

**Logistic Regression:**

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables.

Table

Description automatically generated

**Random Forest:**

Random Forest is a powerful and versatile supervised machine learning algorithm that grows and combines multiple decision trees to create a “forest.” It can be used for both classification and regression problems in R and Python.

Table

Description automatically generated

**XGBoost:**

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Table

Description automatically generated

## **Results:**

We did not mean to use numerous models, but after receiving an 81 percent on logistic regression, we wanted to investigate and try alternative models to obtain accuracy higher than what we received in the first instance.

|  |  |  |
| --- | --- | --- |
| **Machine Learning Model** | **Precision** | **Recall** |
| KNN | 0.77 | 0.77 |
| SVM | 0.81 | 0.81 |
| Logistic Regression | 0.81 | 0.81 |
| Random Forest | 0.93 | 0.93 |
| XGBoost | 0.95 | 0.95 |

However, by identifying the more important benefits and removing a few unnecessary features from the EDA, we were able to get 95 percent accuracy with no overfitting when using the XGBoost Model.

**ROC Curve:**

ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases.

Chart, line chart

Description automatically generated

## **Conclusion:**

To predict the Outcome, we started with five predictive models: the Logistic Regression Model, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), XGBoost, and Random Forest. We considered XGBoost which performed better than other models. Using predictive models to understand and act on the patient's chances of survival.

Analytical tools are used to understand a patient's journey from point of entry to point of exit. Knowing which health services were used allows the hospital to make data-driven decisions. A systematic process for reviewing cases to understand the activities that were taken and what actions could have been taken.

## **Contributions**

Sai Susmitha Vanarasa has pre-processed the data and cleaned the data in which we have outliers and there were no null values but there we have duplicate values.

Quang Nguyen has used feature engineering to get accurate data which are the main features in the data set and used ML models to get an accurate result for feature engineering.

Mounika Loka Analyzed the results after applying each model.

Allena Almeida has done Testing and evaluated the data on all the models to differentiate the features which contribute to variation in this project.

## **Deepnotes Link:**

Below is the link for deep notes where we have done our project work:

<https://deepnote.com/workspace/sebastian-nguyen-617e-55f7cc8b-d2d5-47ab-9c49-25884d40738d/project/Healthcare-Analytics-119f37b3-fe4b-4780-9c44-374fbed5aa5d/notebook/Data%20Analytics-20d3ff960cb94cb0b2ebe9bf255cd552>

## **References:**

Below is the link for dataset:

* <https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction>