Diagnosing Data: Predicting Hospital Patient Mortality

Seoyon Kwon
College of Engineering and Computer Science
University of Central Florida

Michael Owens
College of Engineering and Computer Science
University of Central Florida

Gehan Velivitiya
College of Engineering and Computer Science
University of Central Florida

Sandra Ann Mathew
College of Engineering and Computer Science
University of Central Florida

Michael Cruz
College of Engineering and Computer Science
University of Central Florida

Ngirimana Stephanie De Vasantha Nyiramurehe
College of Engineering and Computer Science
University of Central Florida

***Abstract*—In Intensive Care Units (ICU), the hospital staff's understanding of the severity of admitted patients' conditions is crucial and critical to their proper receipt of care. Patient biometric data, such as demographic data, medical history, and lab analyses, can provide valuable insight into the level and imperativeness of immediate attention. The hope is that, in promptly identifying and addressing dire situations, more lives can be saved.**

***To prevent death, one must be aware of its potentiality. Leveraging their collective knowledge and technical skills, the authors implemented data mining techniques to investigate this dilemma of the healthcare domain. The main objective was to develop a highly accurate predictive model for forecasting the mortality of hospital patients. Using various Machine Learning methods, several predictive models were constructed. Patient data was input into computational algorithms, including Decision Trees, Clustering algorithms, and General Linear Models (GLM), to produce a determination of fatality. The results were extensively tested for practicality and accuracy.**

It is the belief of the authors that these predictive models can lead to improvements in healthcare practices. The research highlights the potential of data-driven solutions in enhancing patient outcomes and overall healthcare efficiency.

Keywords— *healthcare, hospital mortality, data mining, machine learning, medical treatment, intensive care, predictive modeling*

I. Introduction

In a clinical care setting, the Intensive Care Unit (ICU) is known for having one of the highest in-hospital mortality rates due to the critical nature of its patients. Monitoring the health status of patients in the ICU is crucial for their survival, and accurately assessing their level of health can significantly impact the level of care they receive.

Our final project aims to predict the mortality of patients admitted to the ICU using data from the first 24 hours after admission. The concept of predicting patient severity in the ICU is not new. The Acute Physiology and Chronic Health Evaluation (APACHE) scoring system was first introduced in 1981 by George Washington University and has since become a widely used scale for assessing acute illness. Over the years, several versions of APACHE have been developed, with the latest being APACHE IV, released in 2006. APACHE utilizes physiologic measurements, demographics, and previous health conditions from the first 24 hours to assess the severity of acute illness. While APACHE is an invaluable tool for quick severity assessment using simple body metrics, it does not provide predictions on patient mortality; instead, it aids clinical practitioners in assessing patients.

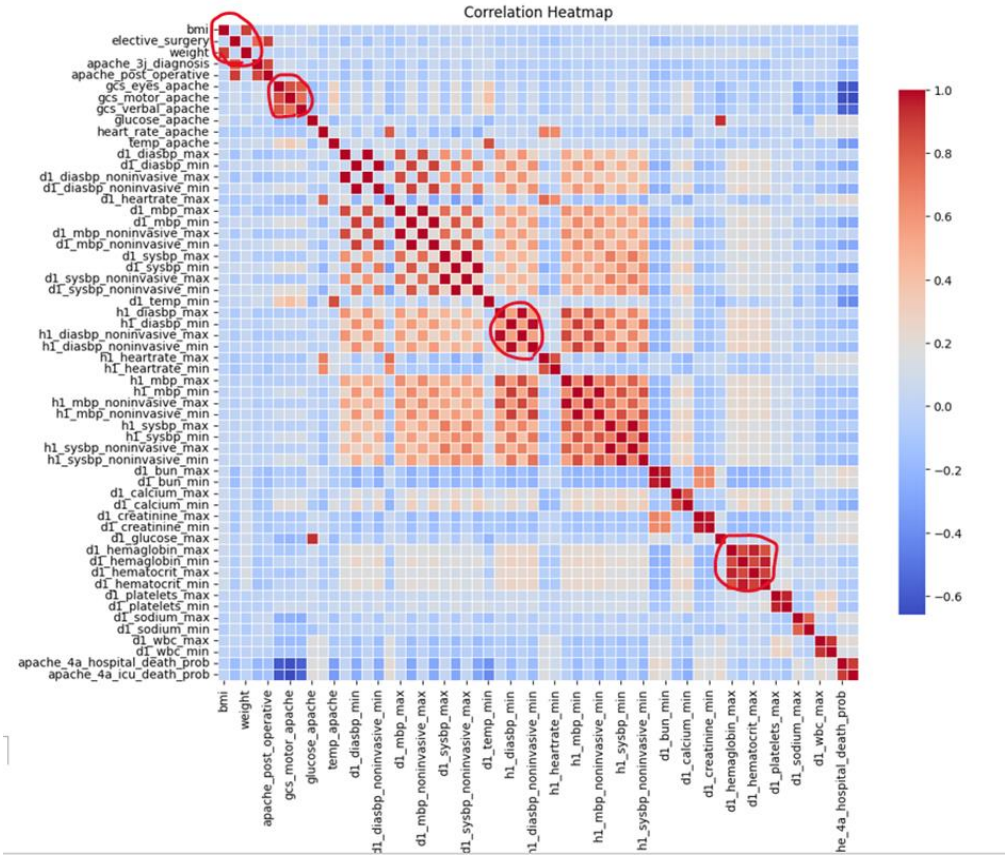
As data analytics students, our challenge is to predict patient mortality based on 24-hour vital metrics, similar to APACHE, in an attempt to explore the predictive potential of the data. By doing so, we hope to provide a broader perspective for assessing patients in clinical settings and discover the possibilities of leveraging data for mortality prediction.

II. Exploratory Data Analysis

The hospital critical care dataset used in this study was sourced from Kaggle, an online platform for sharing datasets. The original data stems from MIT's GOSSIS (Global Open Source Severity of Illness Score) database, designed to predict hospital mortality among patients admitted to the Intensive Care Unit (ICU). Data was collected from diverse hospitals across Argentina, Australia, New Zealand, Bangladesh, India, Nepal, Sri Lanka, Brazil, and the United States.

A. Overview

The dataset encompasses over 130,000 ICU visits worldwide, offering valuable insights such as patient demographics, comorbidity status, laboratory results, and medical measurements taken within the first 24 hours of ICU admission. With a total of 186 columns, it comprises 16 binary, 12 character, and 158 numeric columns. In terms of rows, it contains 91,713 records, and the prediction target variable is the binary outcome of hospital death. The variables included some patient demographic information such as patient id, race, gender, hospital admission types, Body mass index, pre-clinical histories, etc. Vital variables such as diastolic and systolic blood



(**figure 1**: Correlation Heatmap filtered for correlation coefficient +- 0.7)

pressure, body temperature, hear rate, white blood cell count, respiratory measure, oxygen, and arterial pH levels are calculated.

During the data collection phase, the dataset's high quality and reliability, characterized by minimal noise and consistent measurements was noteworthy. The comprehensive data dictionary and variable descriptions facilitate efficient understanding and interpretation for this project. Moreover, GOSSIS ensures data privacy by anonymizing the dataset, safeguarding sensitive and discriminatory information to prevent identification of individuals.

B. General Variable Conventions

An important thing to keep in mind when examining the data, were some of the conventions that are often used in the naming of the variables. For instance, a prefix of "h1_" or "d1_" refers to a max or min measurement that was taken during the first hour or over the day 1 (first 24 hours). The prefix "gcs_" refers to the Glasgow Coma Scale [1][2][3]. This refers to a scale developed by two neurosurgeons at the University of Glasgow in Scotland in the year 1974. It serves as a tool to relay an assessment of a patient's level of consciousness in cases of traumatic brain injuries and other various conditions. The GCS still remains as an important component in patient care and clinical decision making to this day.

For a suffix of "_apache" it most often refers to the fact that the measurement reported is the measurement taken within the first 24 hours that results or produces the highest APACHE III score[4]. This fits with our inclination that it is frequently better to arrive at a false positive in the medical field than a false-negative (the higher APACHE III score relating to a greater risk of mortality – and in our case the target variable of hospital_death being predicted as 1).

C. EDA – Pre-Processing

Due to the high dimensionality of the dataset, instead of scrutinizing each individual variable, this project focused on areas that raised concern. EDA (Exploratory Data analysis) revealed that some variables having as high as 92% null values. To maintain analysis precision and consider imputation costs, variables with more than 15% null values were dropped. Histogram was used to examine the distribution of variables, and it concluded standardizing variable is essential. Outliers which lack of any logical explanations, such as negative age, are also removed but any other outliers are kept. Thus, the binary target variable is highly imbalanced, only 9% of the target has a value of 1. Over and under sampling was used to balance out the binary target variable.

An observation was made regarding certain variables appearing to be closely related. **Figure 1** presents a correlation heatmap, indicating strong correlations (correlation coefficient greater than +- 0.7) among specific variables. It is expected to observe such high correlations, especially between variables like minimum and maximum heart rate and blood pressure.

Although many variables show positive correlations, variables like eye, motion, and verbal scores exhibit a strong negative correlation with the Apache death probability. To avoid target leakage, we removed all death probabilities from the previous Apache model, as they were highly correlated with the target variable. Additionally, irrelevant ID variables, such as patient and hospital IDs, were also excluded. Regarding categorical variables, gender and race did not demonstrate significant relationships with the target variable. However, they were retained in the model to capture any potentially valuable information.

We combined categorical labels which represent less than 5% of data size. Afterwards we performed releveling to make sure the problem base classes were selected

1) Additional Preprocessing with PCA

After the initial preprocessing, the resulting dataset was heavily imbalanced with only around 8% of positive (e.g., resulting in death) cases. To address this, both over and under-sampling was performed to balance the dataset. The newly modified dataset was almost perfectly balanced with 26,369 negative and 26,387 positive.

The next bottleneck faced was the relatively high dimensionality of the dataset. Even after initial preprocessing, the medical records still contained 98 features to account for. To accommodate this, Principal Component Analysis (PCA) was performed on the numerical targets. We chose to retain the first seven Principal Components (PC) which explains 50% of cumulative variance. Illustrated in **Table 1**, it can be observed that by including three additional PC, the Area Under the Curve (AUC) improves by only 0.0060. This is a comparatively small improvement in cumulative variance. Thus, it was decided that the 50% gain in accuracy was sufficient.

Cumulative Variance	No. of Principal Components (PC)	AUC of GLM
0.5	7	0.8547
0.6	10	0.8578
0.7	15	0.8606
0.8	21	0.8634

Table 1. Gain in accuracy vs number of PC's.

After performing PCA, the original numerical variables were replaced with the seven new PC. Manual binarization was then performed on the categorical features. At this point the dataset was determined to be ready for model fitting and evaluation.

2) Feature Selection

Feature selection presented one of the largest challenges with this dataset. We were required to try many different routes as we whittled down the numerous and various predictor variables. Our main approach to this was easily split into two camps. The first was through gaining common-sense and by utilizing critical thinking of the subject matter itself. We had to conduct research on the topics and information relative to the fields of study covered in our data (as none of us were medical professionals). We were able to find a few academic papers that helped us gain a deeper understanding of the applicable areas of medicine and health [5][6][7]. We looked to studies and predictive models of the past in order to help us better hone our own model and move more confidently through choosing which predictive data was/wasn't most useful [8][9].

The second prong of our feature-selection approach was going through using more concrete and statistical method. Our investigation included a box of tools in combination consisting of stepwise, forward selection, backward elimination, ridge regression, lasso regression, elastic net, and more. As medicine is an extremely complex subject area, we felt that accuracy should be more important than interpretability for this report (meaning a tendency to use backward deletion, and using AIC as our penalty). To discuss further thoughts on this, we initially deliberated on whether it would be more advantageous to have a false positive or a false negative for the prediction of someone's survival [10][11]. More on this will be discussed in the "future work" section of this report.

D. PCA– Looking at the Biplot

1) The Loadings

The PCA Bi-Plot served to allow a worthwhile further look into our data as part of our Principal Component Analysis, although it came with certain limitations and short-comings. As the initial plot provided a view that was overly cluttered, we decided to first look at a listing/readout of the first two principal components. This was done twice; first by ordering the loadings of PC1 from least to greatest, and then again doing the same with PC2. The intention was to make it easier to identify certain trends or groupings of similarity among the variables (admittedly a difficult task due to the large nature of our dataset and number of variables).

Cumulative Variance	No. of Principal Components (PC)	AUC of GLM
0.5	7	0.8547
0.6	10	0.8578
0.7	15	0.8606
0.8	21	0.8634

Table 2.The PC Loadings.

For PC1, a very large majority of the most negative loadings were all associated in some way with vitals related to blood pressure (these also had the largest absolute values). For example, these were variables like "h1_mbp_noninvasive_min," "h1_sysbp_noninvasive_min,"

"h1_mbp_max," "d1_mbp_noninvasive_min," and others. This was actually quite interesting and will be revisited later when discussing some of the implications for risk of death. On the positive end of the loading spectrum, there was a slight similarity in that a few of the variables grouped together were lab results for various concentrations of chemicals with the serum or plasma of the patient (for example, "d1_creatinine_max," "d1_calcium_max," "d1_bun_min" – here "bun" stands for "blood urea nitrogen"). However, of these, there were fewer variables with extremely similar loading scores.

When looking at the loading scores for PC2, the groupings of variables conveyed a different story. Here we saw quite a different breakdown in the variables in that the most negative values were all measurements of the Glasgow Coma Scale (GCS) that resulted in the highest APACHE III score – specifically "gcs_eyes_apache" "gcs_verbal_apache," and "gcs_motor_apache." One the positive end, the values were oddly enough a bit similar to the highly negative values of PC1, but in the opposite direction. The more useful takeaway here was the tight coupling and large magnitude of "gcs" variables. This shows that PC2 was capturing something related to this underlying structure of the data.

In looking back at the visuals, themselves, it seemed necessary to split the images up into the loading vectors, and the observations for the sake of clarity. This was done so below in Fig 2. It is clear to see the trend of large magnitudes to the left of the plot as all dealing in some way with blood pressure (invasive/noninvasive, both min and max for diastolic, systolic, and mean). It can also be seen (especially through the process of adding the variables individually, one-by-one) that the three "gcs" vectors are in a tight group heading downward and slightly to the left on the biplot. With Fig 3 the observations can be seen as the black dots, while the labels for the variables here are restricted to direction without magnitude for the sake of clarity.

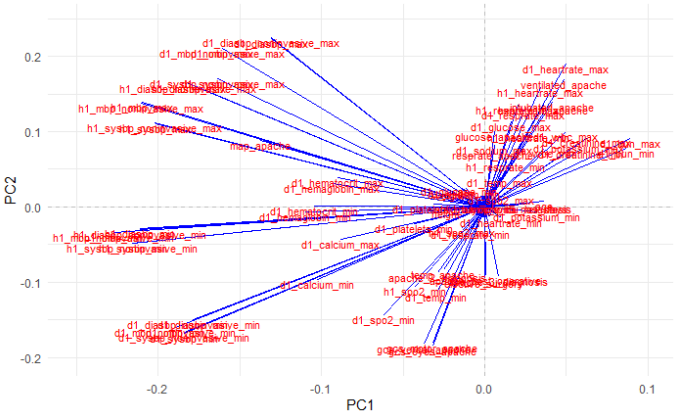


Fig 2. The PCA Biplot without Observations.

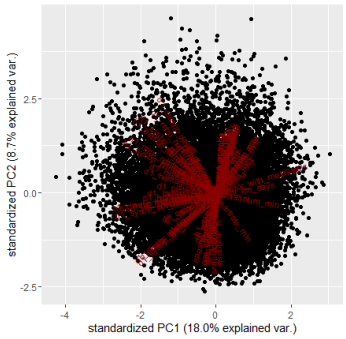


Fig 3. The PCA Biplot Showing Observation Distribution Black.

After seeing the nature of the biplot, and gaining the information discussed above, we wanted to investigate how the factor of the "hospital_death" target variable would be distributed on this biplot (recall that we had removed this target variable from the PCA with the intention of using it as dimensionality reduction, to then be used in a supervised learning model). To this end, we decided to overlay the biplot observations with the corresponding hospital_death attribute (shown below in Fig. 4 with blue where the patient survived, and red where the patient died). Also, it should be noted that we use the over/under-sampled, balanced data here. We double-checked that the resulting plot was identical to the original data since the sampling simply produces synthetic replicas of the data points. This was in fact the case.

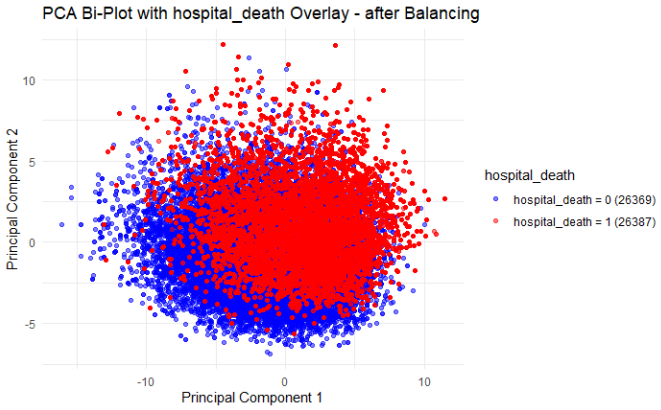


Fig 4. The PCA Biplot with ***hospital_death*** Overlay

The resulting graph gives quite a significant insight into the nature of our data and how it relates to the driving forces identified in our principle components. Notice that, in a general sense, the surviving patients tend to be located further to the bottom and further to the left, while the patients who died congregate more to the top and to the right. Obviously, this is not a rule by any means, just a trend to notice (again PC1 and PC2 only account for 27% of the total variance). What is interesting is that this aligned somewhat with how we were able to describe some of the driving forces behind principal component 1 and 2; with blood pressure and gcs variables. This paints a picture that

reinforces the importance of why blood-pressure is monitored so carefully in the ICU, in addition to why the Glasgow Coma Scale is such an important factor in determining the survivability of a patient.

It is quite impressive that PCA allows us to capture such information from a very high-dimensional dataset. By reducing the dimensions and looking at just PC1 and PC2, we certainly were able to gain some insight into possible major trends that drove the variability within the data. It is worth mentioning that the cumulative explained variance of both principal components 1 and 2, is only about 27%. Even so, as we will see later in this report, PC2 and PC1 provide some of the very highest in predictive power.

III. Building the Models

A. *Classification Models*

Classification is a supervised learning problem. It can predict a target label using the provided predictors. Different classification models are suitable based on the number of labels in the target variable. In our scenario we will be doing a binary classification with a suite of algorithms such as logistic GLM, trees etc.

In this use case, the classification model was designed to predict the mortality of the ICU patients. The variable of "*hospital_death*" was chosen as a binary target variable, with "0" representing death and "1" representing non-deaths.

1) Generalized Linear Model

The first attempt was to fit a Generalized Linear Model (GLM) with binomial family and logit link. The data was partitioned with stratification, allotting 75% to the training set and 25% to the test set. Running this model on the optimized dataset achieved a test accuracy of 0.76 at 0.5 cutoff with test AUC of 0.8547. Additionally, the sensitivity of the model was slightly higher than the specificity. The assessment of these results was that the model had achieved nominal improvement at predicting patient death. The GLM model proved to be achieving its intended purpose.

Building upon this, and in pursuit of further improvements, backward feature elimination was performed. This task was performed with the intention of reducing the Akaike Information Criterion (AIC), demonstrating a better balance between bias and variance. Despite these efforts, the improvements were infinitesimal—resulting in a new AUC of 0.8548 (an improvement of only 0.0001). Nevertheless, while arguably negligible, the improvements to the model were kept. This GLM had t, which we decided to move ahead with out of the two GLMs.

Feature	Estimate	Interpretation
Aids1	-0.6310	The odds of dying for having aids is $e^{-0.6310} < 1$ times of that in baseline level (i.e., aids0)
Leukemia1	0.233079	The odds of dying for having leukemia a is $e^{0.228490} > 1$ times of that in baseline level (i.e., leukemia0)
PC2	0.486650	A unit increase in PC1 results in $e^{0.486650} > 1$ increase the odds of dying increases.

Table 3. Interpretation of few variable coefficients of GLM given others fixed.

Looking at [Table 3] we can see deduce that a patient having aids has a lower odds of dying compared to one not having, which is an interesting observation. On the other hand a person having leukemial relatively has a high chance of mortality than one not having.

2) Decision Trees

Decision trees are non-parametric supervised machine learning models that use a flowchart-like structure to split the data. It is a hierarchical model and is used in the decision-making process as it helps with the visualization of how the data gets divided based on certain conditions and what can be the outcomes as you go down the tree. Thus, it clearly indicates the cause and effect relationship in the data. The major advantages of using decision trees in big data is that it helps in identifying the important predictors and the interaction between them easily with the help of visualization techniques. The most important predictor in the dataset will be the first variable that splits the data in a decision tree. These advantages were the major motivation to run a tree-based analysis in our project.

The most influential predictor of the tree was "PC2," [see Appendix] representing the first split in the branching of the tree. The use of a Decision Tree model was implemented to explore additional avenues of data mining, The initial Decision Tree was built with a max depth of seven and a Complexity Parameter (CP) of 0.0005. This resulted in a tree containing 47 splits and 48 terminals nodes..

The resulting test AUC of the initial tree was 0.8139—a relatively good performance. However, the factors of balance and interpretability needed to be taken in consideration. To produce a more balanced model, the decision tree was pruned at the CP value equal to one standard deviation from the cross-validation error. Now, the tree contained only 37 splits and 38 terminal nodes, making it more easily interpretable. Yet, pruning the tree also resulted in a decrease in the test AUC, bringing it down to 0.8068. It was decided that the minimal sacrifice in accuracy was a prudent trade-off for the comparative gains in balance and simplicity. Thus, the pruned tree was ultimately decided to be the final variant of the model.

3) Ensemble Model

The final approach was to use ensemble models. First, a random forest was fit with 50 trees and assigned a max_depth of seven. This yielded a test AUC of 0.9926. Unfortunately, a score this high is a strong indication of overfitting. Upon observation of the variable importance plot of this model, the most

influential predictor was identified as "PC2." Conversely, "apache_3j_bodysystem.Gynocological" was found to have 0 importance to the model and was, thus, never used for splitting. From [**Fig 6]** we can observe that the log odds of target(yhat) class 0(i.e., non-deaths) goes down (blue line) showing that the odds of patient dying increases. This is the same interpretation we observed through the GLM model for PC2 previously.

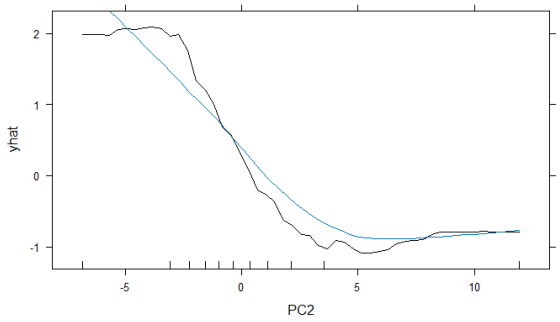


Fig 6. Partial Dependency Plot for PC2

The partial dependency plot is used to help explain and understand the impact of a given variable on the eventual prediction of a model. Here we looked at the impact of the variable with the highest predictive power; Principal Component 2. On the y axis, the value of "y hat" can be seen. This graph is developed by holding all other variables constant except for the selected PC2. This is what allows us to see its direct impact.

There is a definite trend that can be ascertained by noticing that as the PC2 value increases, the value of yhat decreases. The shape of the relationship is not quite linear (more of an s curve), however the nature of the relationship remains. Recalling the insights gained from our PCA, we can see that the gcs values and blood pressure are the driving forces behind this principal component, we were able to surmise that it is a combination of these factors that must play a very major role in determining the patient's survivability.

Finally, we fitted a boosted tree with max_depth of seven which gave us a test AUC of 0.9337, which was a promising result.

We then investigated the variable importance plot of the random forest. As stated earlier in the report, PC2 has the highest importance and therefore possesses the greatest amount of predictive power. Following, it intuitively makes sense that the other additional 6 chosen principal components occupy the next ranks of importance here. Interestingly, the gender being Male is the first non-PC variable that makes an appearance. From a medical perspective, this implies that gender has a great deal to do with mortality rates in the ICU. It should also be pointed out that only one of the variables (Apache_3j_bodysystem.Gynecological). This means that all but one variable played at least some role of importance in the decision trees.

(v.) Evaluations

Model	Test AUC
GLM Reduced	0.8548
1sd Decision Tree	0.8068
Random Forest	0.9926
XGB Tree	0.9337

Table 4. AUC scores of the final selected models.

Based on the results presented in [Table 4] the conclusion was to select the boosted model with 0.9337 AUC as our final classification model which had a higher specificity than sensitivity at 0.5 cutoff as well making it a reliable model to predict the hospital death.

*B. **Regression Models***

In simple terms, regression a mathematical tool used to ascertain the affect that independent variables have on one or more dependent variables. In the field of machine learning, the implementation of regression models can take many forms. Different model types, such as Linear Regression, Polynomial Regression, Ridge Regression, all have different methodologies and broad usage. However, some methods work better than others for different use cases. Situational context must be considered when determining how best to use this tool.

Mortality, as a binary variable, is not ideal for regression. Therefore, to demonstrate the capabilities of regression, the target variable of h1_heartrate_min was selected. Analysis was performed to determine how other biometrics would impact the minimum heart rate of patients.

1) Decision Tree Regression

The analysis was carried out in the platform python with the help of DecisionTreeRegressor() from the scikit learn library. For the fitting process, the imbalanced data was oversampled to ensure that the data was balanced. As the decision tree requires only a little data pre-processing, steps such as scaling or outlier elimination were not deemed necessary for fitting the tree. The variables in the dataset related to heartbeat such as 'h1_heartrate_max', 'heart_rate_apache' and 'd1_heartrate_max' were removed to prevent target leakage resulting in using 115 variables for fitting the model. There were attempts made to fit the model on the features generated using principal component analysis (PCA) which resulted in an inferior performance of the model with Mean squared error (MSE) of 261.355 and R-squared score of 0.43.

The selected regression decision tree model was fitted on the oversampled without the features generated using PCA and the categorical variables were encoded as well . The hyperparameter tuning was carried out using GridsearchCV and the selected tree with the best parameters gave the highest R squared score of 0.44 with MSE of 255.95. The developed tree was used to identify the most important predictors in the data with respect to the target variable as well and the most important predictor in the base decision tree is 'ventilated_apache'

2) Ensemble Model - Regression

The ensemble model was implemented using theXGBoost regressor. The hyperparameters were tuned with different values with the learning rate being one of them, The best performing model was fitted with a learning rate of 0.1 and the number of runs the XGBoost will try to learn was 200 which is tuned using n_estimators. The model produced an R squared value of 0.54 and MSE of 210.218. The most important predictor in the boosted tree is 'h1_resprate_min'.

Attempts were also made to fit a random forest by they became computationally heavy on the system and the results were never produced. Thus the XGBoost model was the best model in tree-based analysis for regression with the highest R squared and lowest MSE.

3) Elastic Net Regression

To investigate the relationship between the other biometric features and minimum heart rate, an Elastic Net Regression model was built. To do this, the data was first preprocessed and cleaned, similar to the classification models. Columns with more than 15% missing values or columns that contained exclusively only zeroes were removed. Random oversampling was performed to assist in balancing out the dataset. Additionally, some features were combined to make the dataset more comprehensive.

At this point the dataset was ready to be split. However, as a regression model, the categorical variables needed to be eliminated. Using one-hot encoding, the categorical variables were turned into numerical variables. With all variables in numerical form, the data was able to be scaled. It was now ready to be fit to a training model. An elastic net training model was prepared. The elastic net makes use of a combination of L1 and L2 penalties to improve balance between reduction of bias and variance.

To gauge the efficacy of the model, it was compared to a naïve linear regression model. The results were counter-intuitive as the base model performed better than the elastic net model. Even with a respectable R^2 score of 0.77, the elastic model fell short of the 0.81 produced by the naïve model. It is suspected this is due to the high dimensionality and multicollinearity produced by the vastly large medical dataset. The same was true of MSE yielding 103.33 again the base of 85.22. Likewise, the model RMSE of 7.15 was trumped by base 5.92.

4) Poisson GLM Regression

Further analysis was performed by implementing a Poisson GLM Regression model. Again, the data was preprocessed and cleaned, and oversampling was implemented. Categories that contained less than 5% observations were combined into an "Other" category. This model, however, utilized PCA to reduce dimensionality.

The dataset was fit to the Poisson GLM was deployed and, unlike the elastic net model, the Poisson model performed better than its base model. The resulting test yield an MSE of 183.56, RMSE of 13.55, and an R^2 value of 0.57. This was a highly significant improvement from the base model, which scored an MSE of 424.13, RMSE of 20.59, and an R^2 value of -4.10e-06. However, while the Poisson Regression model showed some capability, it was determined to be less effective then desirable at making accurate minimum heart rate predictions.

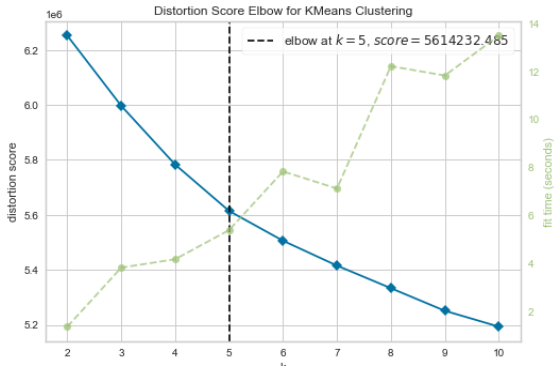
IV. Clustering

In our analysis, we leveraged the power of clustering algorithms, specifically K-means and hierarchical clustering, to address the challenge of dimensionality reduction, particularly in the context of large datasets with high feature counts. The curse of dimensionality can hinder data exploration and analysis, making clustering an effective approach to create a more manageable and concise representation of the data. With clustering, we organized data points into meaningful groups based on similarities, thereby forming a more compact representation of the dataset. By grouping similar data points together, we reduced the complexity of subsequent analyses, enabling us to gain valuable insights from the data efficiently. Both K-means and hierarchical clustering played essential roles in this dimensionality reduction process.

*A. **K-Means***

The application of K-means clustering to this dataset aimed to unveil hidden structures and patterns within the data. By grouping similar data points into clusters based on their similarity, valuable insights into natural groupings and segments present in the data were gained. The clustering process involved partitioning the dataset into K clusters, with the objective of minimizing the within-cluster sum of squares (inertia) – the sum of squared distances from each data point to its assigned cluster's centroid.

Determining the appropriate number of clusters, K, was of utmost importance in the K-means clustering process. To achieve this, both visualization and performance metrics were employed, enabling the identification of the optimal K value that balanced clustering quality and simplicity. Notably, the elbow plot, depicted in **Fig. 7**, pinpointed the "elbow" point where the inertia started to level off, signifying the optimal value of K as 5.



(Fig 7: Elbow method)

Furthermore, the silhouette score was utilized as a quantitative measure of clustering quality. Scores close to +1 indicated well-clustered data points close to their own cluster's centroid, while scores close to 0 suggested data points near decision boundaries between clusters, and scores close to -1 implied potential misclassifications. In this case, K=2 was chosen as it exhibited the highest silhouette score among the various K values, as evident in [see Appendix]

Subsequently, the generated clusters were harnessed as a form of feature engineering. By including the cluster labels as additional features, the aim was to explore the potential improvement of predictive model performance by capturing underlying patterns and dependencies within the data.

After performing the K-means clustering to reveal hidden structures and patterns within the dataset, we further investigated the impact of the obtained clusters by employing logistic regression. The objective was to assess how well the cluster labels could be utilized as additional features in predictive modeling. As a result of this analysis, the logistic regression model achieved an impressive accuracy of 0.91. This high accuracy indicates that the clusters indeed provided valuable information and could effectively contribute to the predictive power of the model. The successful integration of the cluster labels as features demonstrated their relevance in capturing underlying patterns and dependencies within the data, thus enhancing the overall performance of the predictive model. The results further underscored the significance of the K-means clustering approach as a valuable preprocessing step in extracting meaningful insights and improving the efficacy of subsequent machine learning task.

B. Hierarchial Clustering

Hierarchical clustering is a suitable choice for smaller datasets and scenarios where the number of clusters is not predetermined [12]. However, for our current project, we encountered challenges due to the large size of the dataset, making Hierarchical clustering computationally expensive and less optimal for our purposes. To address this issue, we used subset of dataset and reduced its dimensionality by utilizing the average of viral metrics, rather than all the minimum and maximum variables. Additionally, we standardized the data before running the algorithm with complete linkage. This approach yielded a silhouette score of 0.62 with the selection of 2 clusters. In an attempt to improve the results, we employed PCA (Principal Component Analysis) as a dimensionality as another reduction technique. However, this step resulted in a slightly lower silhouette score of 0.60 compared to the initial model. Both models, nevertheless, exhibited a Sum of Explained Variance Ratio of 0.74. Considering our goal of achieving high accuracy, it appears that Hierarchical clustering may not be the most optimal approach for this specific project. Other clustering methods or classification algorithms may be better suited to fulfill our objectives. Thus, since both K-means and Hierarchical clustering is unsupervised classification model, we won't be able to get meaningful result.

C. SVM (Support Vector Machine)

SVM is a powerful algorithm known for its ability to efficiently perform non-linear classification tasks using the kernel trick. By mapping inputs into high-dimensional feature space, SVM can effectively handle high-dimensional and complex classification datasets, making it suitable for tasks like text recognition or gene classification. In this project, we encountered the challenge of selecting the appropriate kernel for SVM, and we experimented with both linear and RBF (Radial Basis Function) kernels.For the RBF kernel, we determined the optimal gamma value using the formula 3/k, where k represents the number of features in our analysis [13]. This step was crucial to fine-tune the SVM model and achieve improved accuracy. In our experiments, SVM demonstrated impressive performance, achieving an AUC(Area Under Curve) score of 0.95. SVM outperformed other clustering methods, highlighting the its effectiveness of in tackling complex classification tasks.

V. Future Work

In this report, we acknowledge the uncertainty surrounding the impact of a positive or negative prognosis (or the possibility of false positives or negatives) on patients. The decision seems to have varying implications depending on the specific situation. One perspective suggests that for patients with a low survival chance, a positive prognosis could lead to faster triage and more allocated resources to improve their chances of survival. Conversely, in a crisis scenario, healthcare professionals may prioritize patients with higher chances of survival. These two viewpoints appear contradictory, and we recognize the need to establish threshold criteria that delineate upper and lower bounds for binary predictions. Such criteria might create a "window" of survival chance rather than a single border, which can also influence the selection of relevant features. As we move forward, we will thoroughly evaluate these ideas and analyze more data to address these critical questions.Additionally, exploring alternative algorithms could significantly enhance the analysis and prediction of this dataset. Deep learning models, particularly deep neural networks, offer the exceptional ability to automatically extract meaningful representations (features) directly from raw data. This advantage proves especially beneficial when dealing with high-dimensional and complex datasets, as these models can reveal intricate hierarchical patterns that might not be immediately apparent in the original data. Moreover, deep neural networks excel at capturing non-linear relationships between variables, enabling them to tackle complex tasks beyond the capabilities of simple linear models. With multiple hidden layers, these models proficiently model highly non-linear functions, effectively handling intricate patterns inherent in the data. Another crucial benefit of deep learning is its capability to handle categorical variables. By learning embeddings or representations for categorical features, deep learning models can capture underlying relationships between categories, leading to more accurate predictions.

VI. Conclusion

In conclusion, our final project focused on predicting the mortality of patients admitted to the Intensive Care Unit (ICU) using data from the first 24 hours after admission. The dataset,

sourced from MIT's GOSSIS database, provided valuable insights into patient demographics, comorbidity status, laboratory results, and vital measurements. Our data analytics exploration began with exploratory data analysis (EDA), where we carefully selected relevant variables, handled missing values, and identified correlations among features.For feature selection, we employed both domain knowledge and statistical methods, seeking to retain important predictors while removing noise and irrelevant variables. Our analysis revealed that principal component 2 (PC2) carried the highest predictive power, indicating the significance of vital metrics and blood pressure in determining patient outcomes. Regression analysis using decision trees and XGBoost not only validated the significance of PC2 but also achieved remarkable R-squared scores and MSE values. While an attempt was made to conduct a random forest analysis, its computational complexity presented challenges during the process. Clustering methods, namely K-means and hierarchical clustering, was attempted to reveal underlying groups of data. Both clustering methods proved less optimal due to the dataset's large size and problem context. In our classification analysis, Support Vector Machine (SVM) emerged as the most powerful algorithm, outperforming other methods with an impressive AUC score of 0.95. SVM's ability to handle non-linear classification tasks showcased its efficacy in dealing with complex medical data.

Our findings demonstrate the value of data mining techniques and machine learning in predicting ICU patient mortality. By leveraging data-driven insights, we gain a deeper understanding of patient severity and can offer more targeted and efficient care.

VII. Executive Summary

Understanding the severity of patients in Intensive Care Units (ICUs) is crucial for healthcare professionals to provide efficient care. Accurately predicting patient mortality can offer valuable clinical insights. Leveraging the extensive GOSSIS dataset, which includes over 91,000 observations from hospitals worldwide, this project achieves a remarkable 95% accuracy in classifying patient mortality. After an exhaustive evaluation of various data mining methods, Support Vector Machine (SVM) emerges as the most effective model for predicting patient mortality. This project conclude the significance of using SVM as a powerful tool for making critical healthcare assessment.

References

[1] Teasdale, G, and B Jennett. “Assessment of coma and impaired consciousness. A practical scale.” Lancet (London, England) vol. 2,7872 (1974): 81-4. doi:10.1016/s0140-6736(74)91639-0

[2] Jain, Shobhit. and Lindsay M. Iverson. “Glasgow Coma Scale.” StatPearls, StatPearls Publishing, 21 June 2022.

[3] Braine, Mary E, and Neal Cook. “The Glasgow Coma Scale and evidence-informed practice: a critical review of where we are and where we need to be.” Journal of clinical nursing vol. 26,1-2 (2017): 280-293. doi:10.1111/jocn.13390

[4] Knaus, W A et al. “The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults.” Chest vol. 100,6 (1991): 1619-36. doi:10.1378/chest.100.6.1619

[5] Chmielewski, Nicholas DNP, RN, CEN, CENP, NEA-BC, FAEN; Moretz, Jason MHA, BSN, RN, CEN, CTRN. ESI Triage Distribution in U.S. Emergency Departments. Advanced Emergency Nursing Journal 44(1):p 46-53, January/March 2022. | DOI: 10.1097/TME.0000000000000390

[6] Litvak, Eugene et al. “How Hospitals Can Save Lives and Themselves: Lessons on Patient Flow From the COVID-19 Pandemic.” Annals of surgery vol. 274,1 (2021): 37-39. doi:10.1097/SLA.0000000000004871

[7] Heng, H., Jazayeri, D., Shaw, L. et al. Hospital falls prevention with patient education: a scoping review. BMC Geriatr 20, 140 (2020). https://doi.org/10.1186/s12877-020-01515-w

[8] Raita, Y., Goto, T., Faridi, M.K. et al. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 23, 64 (2019). https://doi.org/10.1186/s13054-019-2351-7

[9] S. Cohen, N. Dagan, N. Cohen-Inger, D. Ofer and L. Rokach, "ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models," in IEEE Access, vol. 9, pp. 91584-91592, 2021, doi: 10.1109/ACCESS.2021.3091622.

[10] Kappen, T.H., van Klei, W.A., van Wolfswinkel, L. et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagn Progn Res 2, 11 (2018). https://doi.org/10.1186/s41512-018-0033-6

[11] Croft, P., Altman, D.G., Deeks, J.J. et al. The science of clinical practice: disease diagnosis or patient prognosis? Evidence about “what is likely to happen” should shape clinical practice. BMC Med 13, 20 (2015). https://doi.org/10.1186/s12916-014-0265-4

[12] Das, V. K. (1970, October 8). K-means clustering vs hierarchical clustering. Global Tech Council. https://www.globaltechcouncil.org/clustering/k-means-clustering-vs-hierarchical-clustering/

[13] IBM. (n.d.). SVM Node Expert Options. SVM node expert options. https://www.ibm.com/docs/en/spss-modeler/18.2.2?topic=node-svm-expert-options

[14] Knaus, W A et al. “APACHE-acute physiology and chronic health evaluation: a physiologically based classification system.” Critical care medicine vol. 9,8 (1981): 591-7. doi:10.1097/00003246-198108000-00008

[APPENDIX]

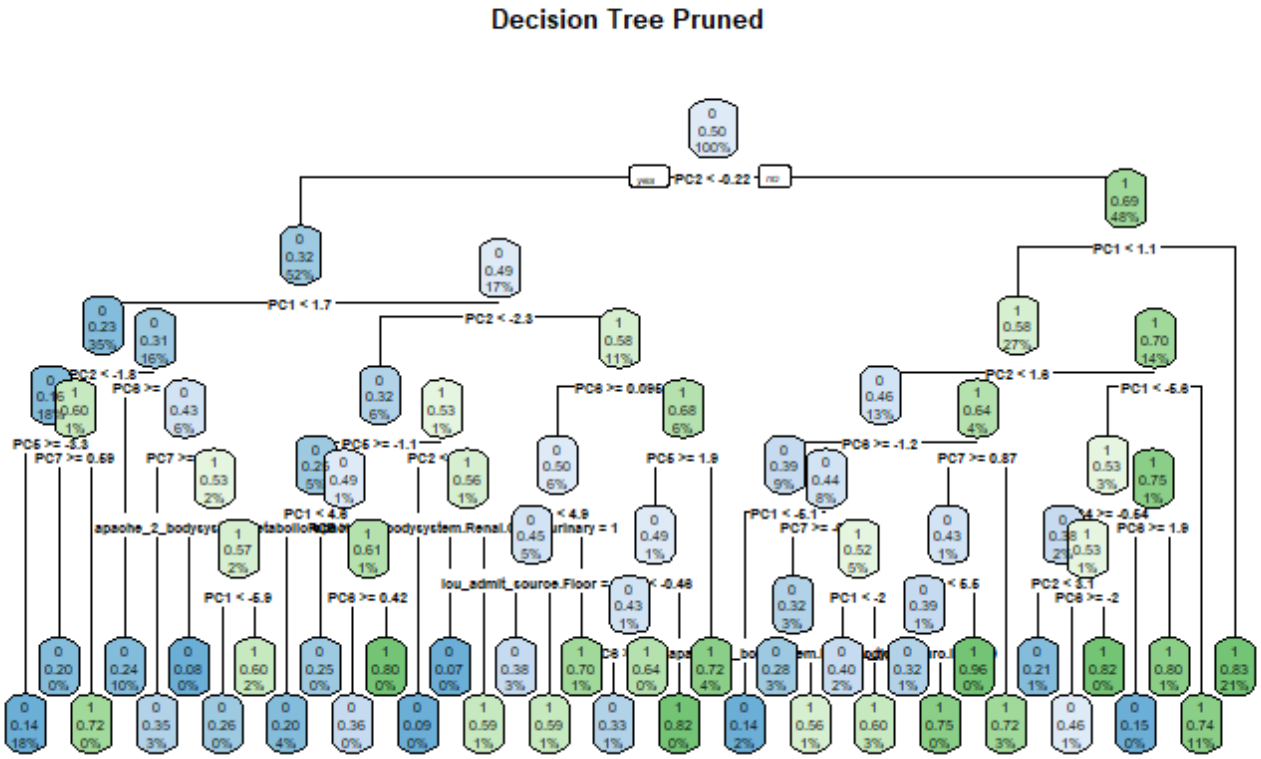


Fig A1. Decision tree pruned

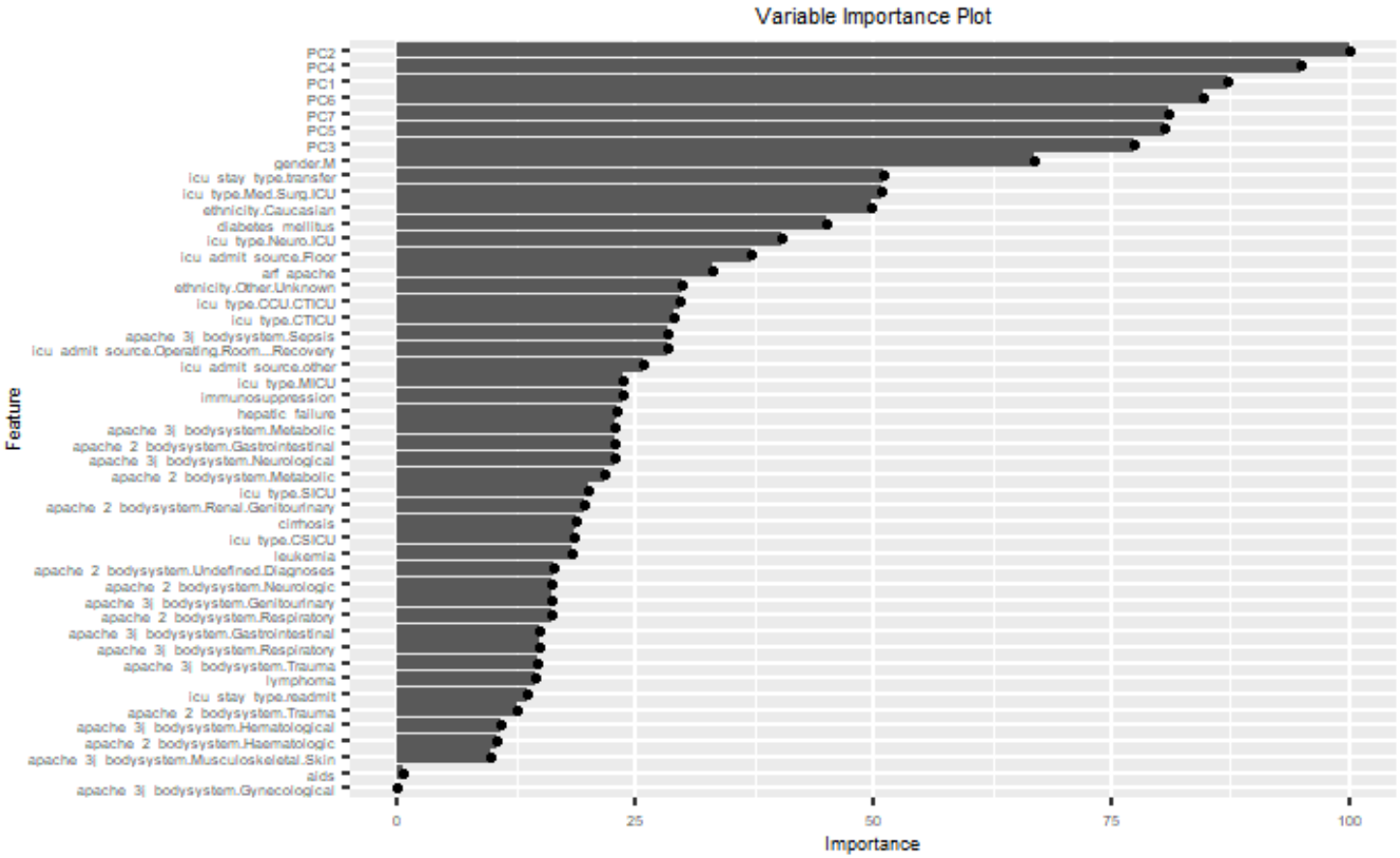


Fig A2. Variable importance