Early Mortality Prediction in the Intensive Care Unit:
A Machine Learning Approach

Selam Mequanint and Ossama Mahmoud

Abstract

Early mortality prediction of hospitalized patients is important for assessing the severity of illness and deciding the appropriate treatment and intervention required. Several severity scoring models have been developed over the past decades, but still, early mortality prediction for intensive care unit patients remains a challenge. This study aims to investigate the use of machine learning in predicting early mortality using information collected during the first 6 hours of intensive care unit (ICU) admission. MIMIC-III ('Medical Information Mart for Intensive Care') data was used for this study. A total of 6,380 complete records were included for the model development. Multiple supervised classification algorithms including Decision Trees, Random Forest, Support Vector Machine (both linear and Radial-bases function (for non-linear)), Logistic Regression, and Neural Nets were explored to develop a model that best predicts patient-specific early mortality in ICU. Neural Nets exhibited the highest area-under-curve followed by Support Vector Machine with Radial-basis function (*RBF)* compared with the rest of the models considred for predicting ICU mortality in the first 6 hrs of admission.

Introduction

The intensive care unit (ICU) is a special department in a hospital where severely ill patients receive treatment. The ICU is one of the most expensive care because it requires high staff to patient ratio for intensive patient monitoring and complex treatment. The average daily cost of an ICU stay per patient is estimated to be three times more than the average cost of a day's stay in a general ward(Information (CIHI) 2016). The extra costs are attributed to ICU stays being more resource-intensive and involving many healthcare professionals, equipment and medication. Critical illness is associated with high mortality. Intensive care clinicians can factor an estimate of patient mortality when making decisions about the health of a patient. These decisions can allow physicians to allocate more resources to the more ill patients, ensure they are adequately monitored, and assess the possibility of moving patients to a comfort care option, palliative care.

Several models that predict risk for ICU mortality have been developed in the past (Le Gall et al. 2005, Lemeshow et al. 1993, and J. R. Le Gall, Lemeshow, and Saulnier 1993), but they are limited by technical and practical considerations, often use summary data from a full day of a patient's ICU stay which was manually documented by trained personnel. Critical care units are equipped with an advanced medical monitoring device, and the growing use of electronic charting that lends a great opportunity for machine learning approaches to predict mortality. Machine learning algorithms that use routinely captured ICU data to predict mortality within a few hours of ICU admission is a promising strategy to make a significant improvement in ICU outcomes, and cost savings in the healthcare system.

This study aims to investigate the use of machine learning in predicting mortality early in the ICU. The goal of the study is to provide clinicians with timely information that can enhance their understanding of a patient criticality and act as a flag for poor outcomes. In this study, we define early as the first 6 hours of admission. This time interval was reached based on a literature review and in consultation with clinicians. The 6 hours window was selected as a balance between having

too short windows with too many missing data points (low-frequency clinical data) and too large windows too to make any meaningful prediction.

Data Source

MIMIC-III (Johnson et al. 2016) ('Medical Information Mart for Intensive Care') data, an openly available dataset developed for Computational Physiology, comprising de-identified health data associated with critical care patients, was used for this project. MIMIC-III is a large, single-center database comprising information relating to patients admitted to critical care units at the Beth Israel Deaconess Medical Center in Boston, Massachusetts hospital. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. Data accessed for this study contains clinical information associated with a total of 45,183 adult patients (aged 17 years or above) admitted to critical care units between 2001 and 2012. Table 1 presents selected demographic and admission information of the available dataset.

Demographic	Mean (SD)	Range
Age	63(16.35)	17-89
Sex (%) F:M	42:58	
ICU length of stay *(days)*	4.41(6.46)	0.07-173.0
Hospital length of stay *(days)*	11.37(13.22)	-0.84-294.66
Hospital mortality (%)	11.37%	
Classes (Deceased: Discharged)	N = 45,338	5390: 39948

Table 1: Patient demographics and admission/discharge information

Experiment

Feature selection: Similar studies (Awad et al. 2017, Silva et al. 2012, Ramon et al. 2007, and Pirracchio et al. 2015) that used the MIMIC III database were reviewed, and clinical experts were consulted before selecting potential predictors to ensure that all clinically important variables are not missed. The average, minimum and maximum value of the selected features in the first 6 hours of ICU admission were computed and included as predictive features. Whether the patient died in the hospital after the ICU stay, or hospital mortality was used as a response variable.

Although the sample size is quite large, feature selection is important to eliminate highly correlated values. As such, a correlation matrix for the variables was constructed, and Figure 1 shows that there were highly correlated variables in the dataset. The heat map (Figure 1), that uses colour intesity to indicate the level of correlation between each feature (lighter-negatively related; darker colour shows positively related), was used to identify redundantly and even repeated lab tests. Bar graphs were used to see if there is any difference between the response classes regarding vital and clinical tests measure in the first 6 hours. Visible mean differences were observed in clinical measures between the two class (Appendix C-D). A total of 138 features were retained for further analysis.

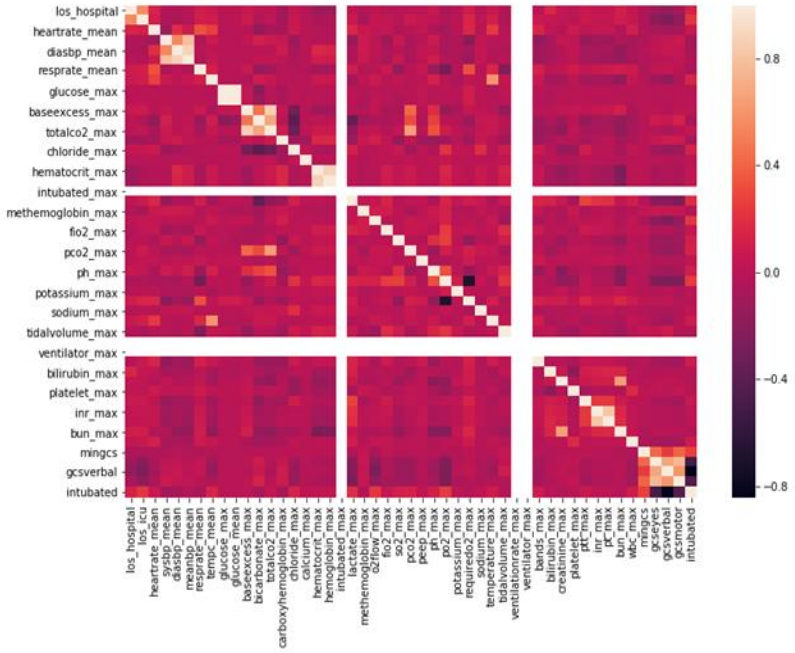


Figure 1. Correlation matrix of all variables considerate in study

Missing values: Data were assessed for missing values, and percentage of missing values for each feature by the response category (Deceased: Discharged) was calculated. Not all lab tests are measured for all patients within the first 6 hours of admission. Therefore there will be expected data missing for each patient which can be considered normal. However, if an individual patient record has multiple entries missing, it may be explained that this is because they were regarded as being less sick than others, so they were not prioritized for tests. Similarly, the patient may have died before all test can be done. Identifying these two scenarios is not easy with the data available.

Missing values were handled by setting up threshold at two levels: features and raw levels**.** Features with more than 60% missing and are clinically insignificant were dropped. After the removal of features with a large number of missing values, 90 features were left in the dataset. The features dropped during this process are mostly clinical measure that does not get measured in the first 6 hours such as urine analysis and stool test. Eleven of the 90 features are admission attributes such as length of hospital and ICU stay, time of admission, severity scores. Also, records with at most 10% missing values were kept in the dataset that resulted in a total of 21,448 records. Missing values in the feature matrix were imputed with the column means.

Balancing the data: After missing values were addressed, the data showed that only about 18% of patients died in the hospital. This creates imbalanced classes as the ratio between our two classes, died in hospital to discharged from the hospital, is 18:82. The under-sampling technique was used to balance the data. Although up-sampling is widely used, our dataset is quite large and creating more of the minority (deceased) class was not necessary. Thus, prototype selection algorithms (Lemaitre, Nogueira, and Aridas 2016) that allows to randomly select an equal number of records from the majority (discharged) class was used to balance the data. After under-sampling was performed, the total number of records retained was 6,380 with a class ratio of 50:50.

Scaling and standardization: Continuous values such as age, weight, and laboratory measures were scaled to a range between 0 and 1, as follows:

$$x_s = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

Where x_s is a scaled value, x_i is the original value, and min(x) and max(x) are the minimum and maximum values of the particular feature. Scaling prevents variables from over dominating in the supervised learning algorithms.

Data splitting: After data preparation complete by addressing missing value, imbalance data and normalization, a total of 6,380 records and 79 features were left and used for model development. Appendix lists the features used for the model development. The final dataset was split into a 20/80 split; 20% for testing and 80% for training sets.

Learning Algorithms

The study response variable, mortality, is a binary variable. As such, we used a set of supervised classification algorithms to develop different predictive models. These included Logistic Regression, Support Vector Machine (SVM), Decision Tree and Random Forest. To find optimal hyperparameters for each model 5-fold cross-validation was used. Finally, each model was tested with the test set, and performance metrics such as accuracy, precision, recall, F-score and AUC-ROC were calculated. All analyses were performed using python 3.7 and the sklearn library.

Logistic Regression (LR): LR is a statistical modelling technique for a dataset with a binary response and one or more variables that determine the response. The goal of LR is to find the best fitting model to describe the relationship between the binary outcome variable and a set of independent (predictor or explanatory) variables. LR was applied to the project data to predict mortality. Designed specifically for binary classification, LR is very well suited for our dataset.

Decision Tree: Decision trees are a non-parametric learning algorithm. The algorithm uses a tree-like structure of decisions to splits the data up. Each decision attempts to minimize the entropy of the split data. The purity of each result is maximized to minimize the entropy of the result of a decision. Then when classifying new samples, the sample is passed through the decision tree and follows a certain path. The final ratio of labels in the leaf, or last node in the path, determine the classification of the sample. (Visualization in Appendix B)

Random Forest: The random forest algorithm is an extension of the decision tree algorithm. The algorithm builds a sequence of trees each with a different random seed. Then when running new samples through the algorithm, the new sample is run through each tree created in the random forest, and the classification is determined by a majority vote of all the classification results of the induvial trees in the random forest. This technique is useful as it prevents overfitting.

Support Vector Machine (SVM): A support vector machine is a learning algorithm that attempts to maximize the decision boundary between training data. The decision boundary depends on a few samples of the data closest to the boundary known as support vectors. Then when the classifier needs to classify new samples, it merely sees on which side of the classifier the new sample is on and classifies it as such. The advantage of the SVM is that it is swift to analyze new samples. We have explored both linear and non-linear SVMs, allowing us to produce both linear and non-linear decision boundary.

Neural Net (NN): Neural nets are a classification algorithm that utilizes a sequence of layers of nodes each with variable weights connecting one layer with the next. Based on the training data

the weights are updated to ensure the best classification of the training data. The neural net used in this problem had two hidden layers.

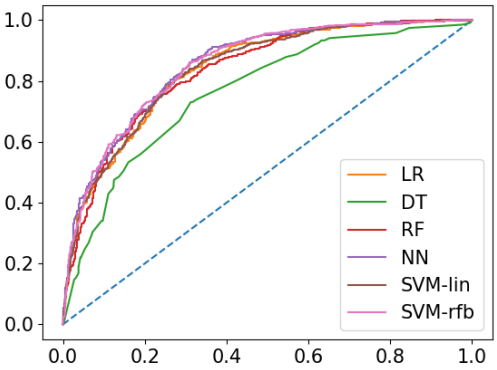
Results

The commonly used performance measures such as sensitivity, specificity and area under the receiver operator characteristic (AUC-ROC) curves were used to measure the performance of the tested models. The area under the ROC curves ranges from a perfect score of 1.0 to a score of 0.5 indicate random classification. In the ROC curve plot, the curve closer to the vertical axis is the most accurate predictive model.

The best cross-validated predictor of early ICU mortality according to AUC-ROC was the Neural Nets model as shown in Table 2. The AUC-ROC obtained from the six classifiers ranges from 0.762 to 0.858, which is not very wide. The Neural Nets provided the best accuracy out of the six different learning algorithms tested. It is of note that all six performed relatively well (Figure 2). Also, Neural Nets provided the highest (81%) recall rate (the proportion of death identified by the model out of the total death), which is an important metrics for our study because the cost of missing death undetected is high. False negatives are also particularly import for our project, as a false negative implies that the predictor predicted that a patient would die when in fact, he survived the ICU stay. A costly mistake, as you can imagine. The neural net was still the best classifier with the lowest rate of false negatives at 23%.

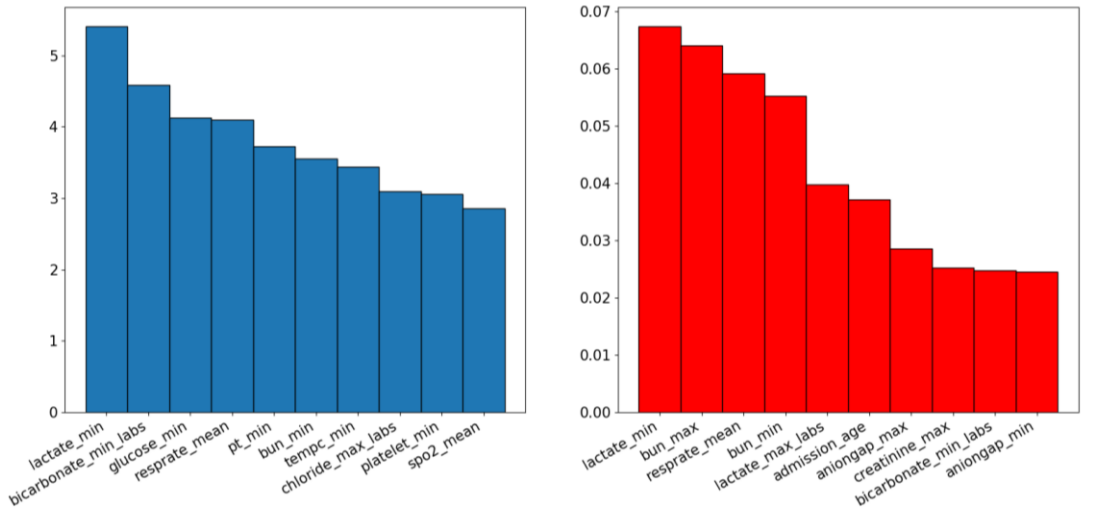
Algorithm	*AUC-ROC*	*Accuracy*	*Precision*	*Recall*	*False Negative*	*Parameters*
LR	0.838	0.75	0.73	0.77	0.26	C=10 Penalty=l1
SVM-linear	0.837	0.75	0.76	0.73	0.26	C=1 gamma=1e-05
SVM-rfb	**0.852**	**0.76**	**0.77**	**0.75**	0.25	C=1, gamma=0.1
Decision tree	0.762	0.70	0.69	0.73	0.28	max_depth=6
Random Forest	0.834	0.76	0.77	0.75	0.26	n_estimators=10max _depth=6
Neural Net	**0.858**	**0.77**	**0.76**	**0.81**	0.23	activation=tanh hidden_layers=(9,7)

Table 2. Model performance metrics



Although performing lower than the neural net the logistic regression model and the random forest we were able to provide useful insight to the importance of various features. As we can see there is a large degree of overlap between the significant features in both models.

Figure 2. Receiver-operating characteristics curves.



(a) (b)

Fig 3, shows top 10 features in terms of significance on learned model (a) absolute value of log-odds of LR (b) feature importance in RF

Discussion

In this report, we presented the challenging problem of predicting early mortality in ICU using a machine learning approach. We used six classifiers in this prediction problem and compared them with standard machine learning techniques. We demonstrated that all six, mainly Neural Net could offer good accuracy of prediction (AUC-ROC> 83%). Also, in the most relevant performance metric, high recall rate, Neural Nets outperformed all the other methods considered.

As the goal of this study is to derive important takeaways for clinicians to use in the clinical setting, not only the accuracy of the model but also the interpretability of the model was equally important. A study(Sadeghi, Banerjee, and Romine 2018) that explored early hospital mortality prediction using vital signs showed that Decision Trees provided the best interpretability and best accuracy out of the eight different learning strategies they tested. On the other hand, Ramon et al. 2007 reported that naïve Bayesian networks and Naive Bayesian networks performed better than Decision Tree. Similarly, Pirracchio et al. 2015 reported that a Bayesian Additive Regression Tree (BART) is the best candidate, while Random Forests (RF) outperformed all other candidates when using transformed variables. Our results confirm some of these findings, the random forest was the second-best classifier in terms of accuracy. More so, the logistic regression and random forest highlighted a few variables that had more impact on the model. A theory to why the minimum lactate across the 6-hour ICU stay had the greatest affect on mortality is based on the findings of Meakins (2018). They found that when oxygenated blood is not flowing adequately through the body, the amount of lactate in the blood increases. Both the lack of oxygenated blood flowing through the body and the accumulation of lactate are leading causes of death in hospitals (Meakins, 2018). Although this theory makes sense, we can not eliminate the possibilities of a confounding variable, without further analysis.

Our work highlights some important features that may need to be examined further to evaluate their effects on patient ICU mortality. Future studies can further analyse these important variables for confounds and rationalize their importance in the model. The time available for this project limited us from conducting further analysis and further identifying other critical predictive features

obtained from each algorithm considered. We also think that it would be interesting to explore the use of cost-sensitivity as a model performance measure for this study.

References

Awad, Aya, Mohamed Bader-El-Den, James McNicholas, and Jim Briggs. 2017. "Early Hospital Mortality Prediction of Intensive Care Unit Patients Using an Ensemble Learning Approach." *International Journal of Medical Informatics* 108: 185–95. https://doi.org/10.1016/j.ijmedinf.2017.10.002.

Information (CIHI), Canadian Institute for Health. 2016. "Care in Canadian ICUs." Text. August 11, 2016. https://secure.cihi.ca/estore/productFamily.htm?locale=en&pf=PFC3248.

Johnson, Alistair E. W., Tom J. Pollard, Lu Shen, Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. "MIMIC-III, a Freely Accessible Critical Care Database." *Scientific Data* 3 (May): 160035. https://doi.org/10.1038/sdata.2016.35.

Le Gall, J. R., S. Lemeshow, and F. Saulnier. 1993. "A New Simplified Acute Physiology Score (SAPS II) Based on a European/North American Multicenter Study." *JAMA* 270 (24): 2957–63.

Le Gall, Jean Roger, Anke Neumann, François Hemery, Jean Pierre Bleriot, Jean Pierre Fulgencio, Bernard Garrigues, Christian Gouzes, Eric Lepage, Pierre Moine, and Daniel Villers. 2005. "Mortality Prediction Using SAPS II: An Update for French Intensive Care Units." *Critical Care (London, England)* 9 (6): R645-652. https://doi.org/10.1186/cc3821.

Lemaitre, Guillaume, Fernando Nogueira, and Christos K. Aridas. 2016. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *ArXiv:1609.06570 [Cs]*, September. http://arxiv.org/abs/1609.06570.

Lemeshow, S., D. Teres, J. Klar, J. S. Avrunin, S. H. Gehlbach, and J. Rapoport. 1993. "Mortality Probability Models (MPM II) Based on an International Cohort of Intensive Care Unit Patients." *JAMA* 270 (20): 2478–86.

Meakins J, Long CNH. Oxygen consumption, oxygen debt and lactic acid in circulatory failure. J Clin Invest 1927; 4: 273.

Pirracchio, Romain, Maya L. Petersen, Marco Carone, Matthieu Resche Rigon, Sylvie Chevret, and Mark J. van der Laan. 2015. "Mortality Prediction in Intensive Care Units with the Super ICU Learner Algorithm (SICULA): A Population-Based Study." *The Lancet. Respiratory Medicine* 3 (1): 42–52. https://doi.org/10.1016/S2213-2600(14)70239-5.

Ramon, Jan, Daan Fierens, Fabian Güiza Grandas, Geert Meyfroidt, Hendrik Blockeel, Maurice Bruynooghe, and Greta Van den Berghe. 2007. "Mining Data from Intensive Care Patients." *Advanced Engineering Informatics* 21: 243–56. https://doi.org/10.1016/j.aei.2006.12.002.

Sadeghi, Reza, Tanvi Banerjee, and William Romine. 2018. "Early Hospital Mortality Prediction Using Vital Signals," March. https://arxiv.org/abs/1803.06589.

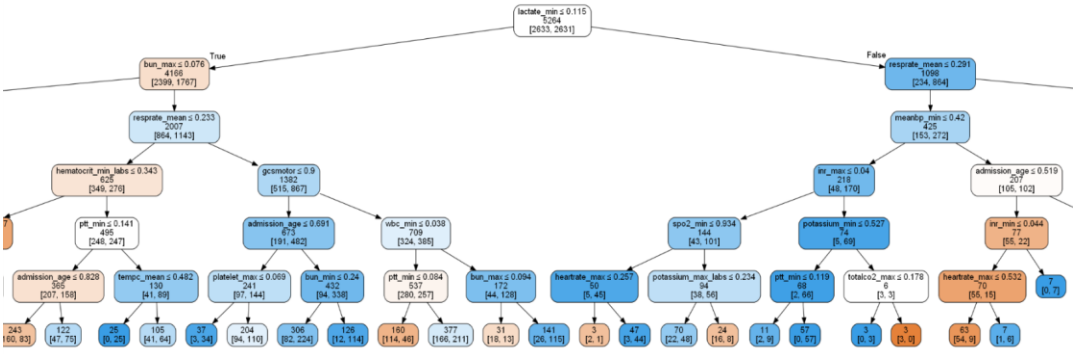
Silva, Ikaro, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. 2012. "Predicting In-Hospital Mortality of ICU Patients: The PhysioNet/Computing in Cardiology Challenge 2012." *Computing in Cardiology* 39: 245–48.

Appendix A: List of features included in the model development

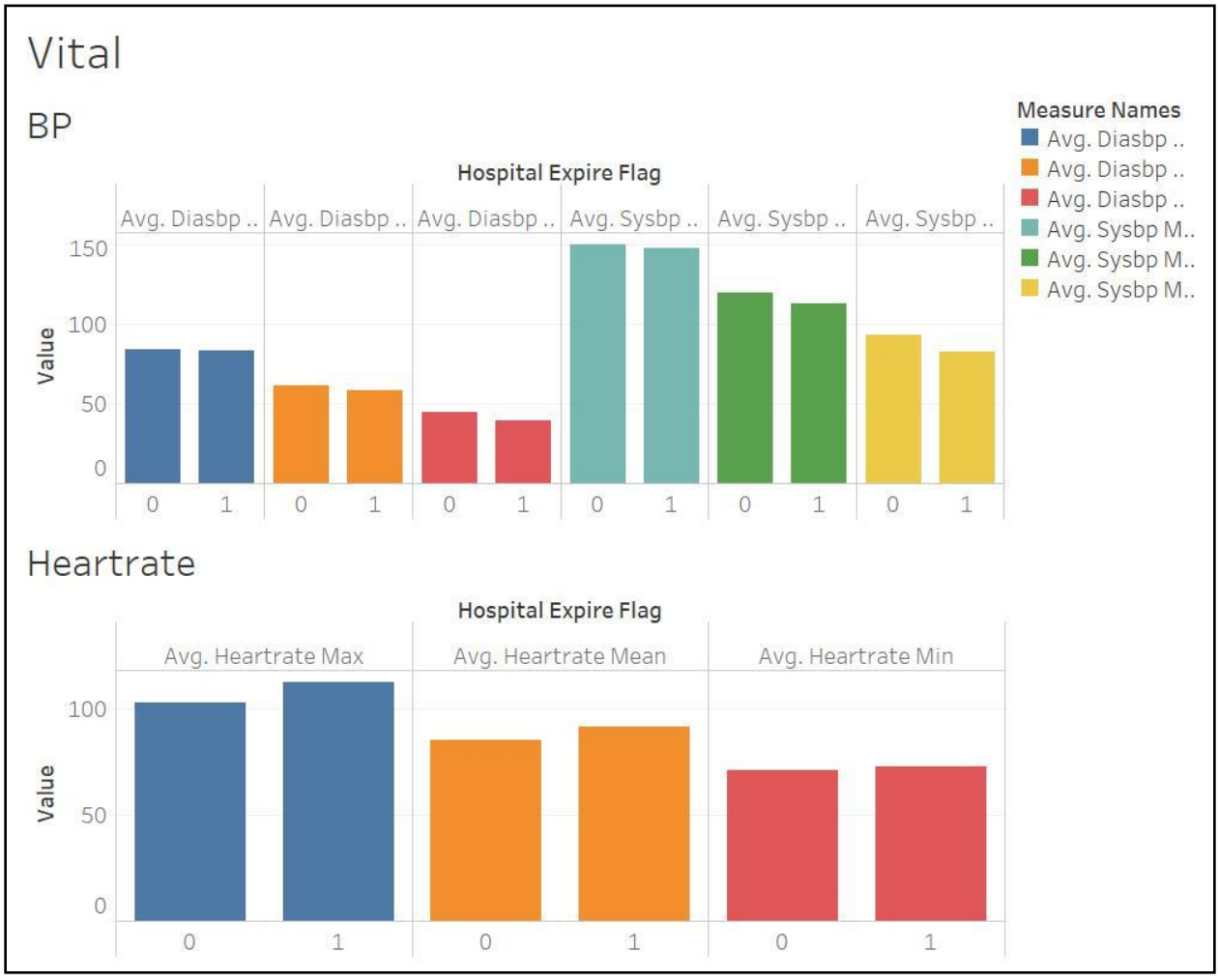
	Categories	Features		Categories	Features
1	Demographic	admission_age	39	Laboratory tests	hemoglobin_min_labs
2		height_first	40		hemoglobin_max_labs
3		weight_first	41		lactate_min_labs
4		height_meter	42		lactate_max_labs
5		BMI	43		platelet_min
6		gender	44		platelet_max
7	Clinical	gcseyes	45		potassium_min_labs
8		gcsverbal	46		potassium_max_labs
9		gcsmotor	47		ptt_min ptt_max
10		intubated	48		inr_min inr_max
11	Laboratory test	spo2_min	49		pt_min pt_max
12		spo2_max	50		sodium_min_labs
13		spo2_mean	51		sodium_max_labs
14		glucose_min	52		bun_min
15		glucose_max	53		bun_max
16		glucose_mean	54		wbc_min
17		totalco2_max	55		wbc_max
18		lactate_max	56		mingcs
19		pco2_max	57	Vital	heartrate_min
20		ph_max	58		heartrate_max
21		po2_max	59		heartrate_mean
22		potassium_max	60		sysbp_min
23		totalco2_min	61		sysbp_max
24		lactate_min	62		sysbp_mean
25		pco2_min	63		diasbp_min

26		ph_min	64		diasbp_max
27		po2_min	65		diasbp_mean
28		potassium_min	66		meanbp_min
29		aniongap_min	67		meanbp_max
30		aniongap_max	68		meanbp_mean
31		bicarbonate_min_labs	69		resprate_min
32		bicarbonate_max_labs	70		resprate_max
33		creatinine_min	71		resprate_mean
34		creatinine_max	72		tempc_min
35		chloride_min_labs	73		tempc_max
36		chloride_max_labs	74		tempc_mean
37		hematocrit_min_labs	75		respxvent
38		hematocrit_max_labs	76	Response	hospital_expire_flag]

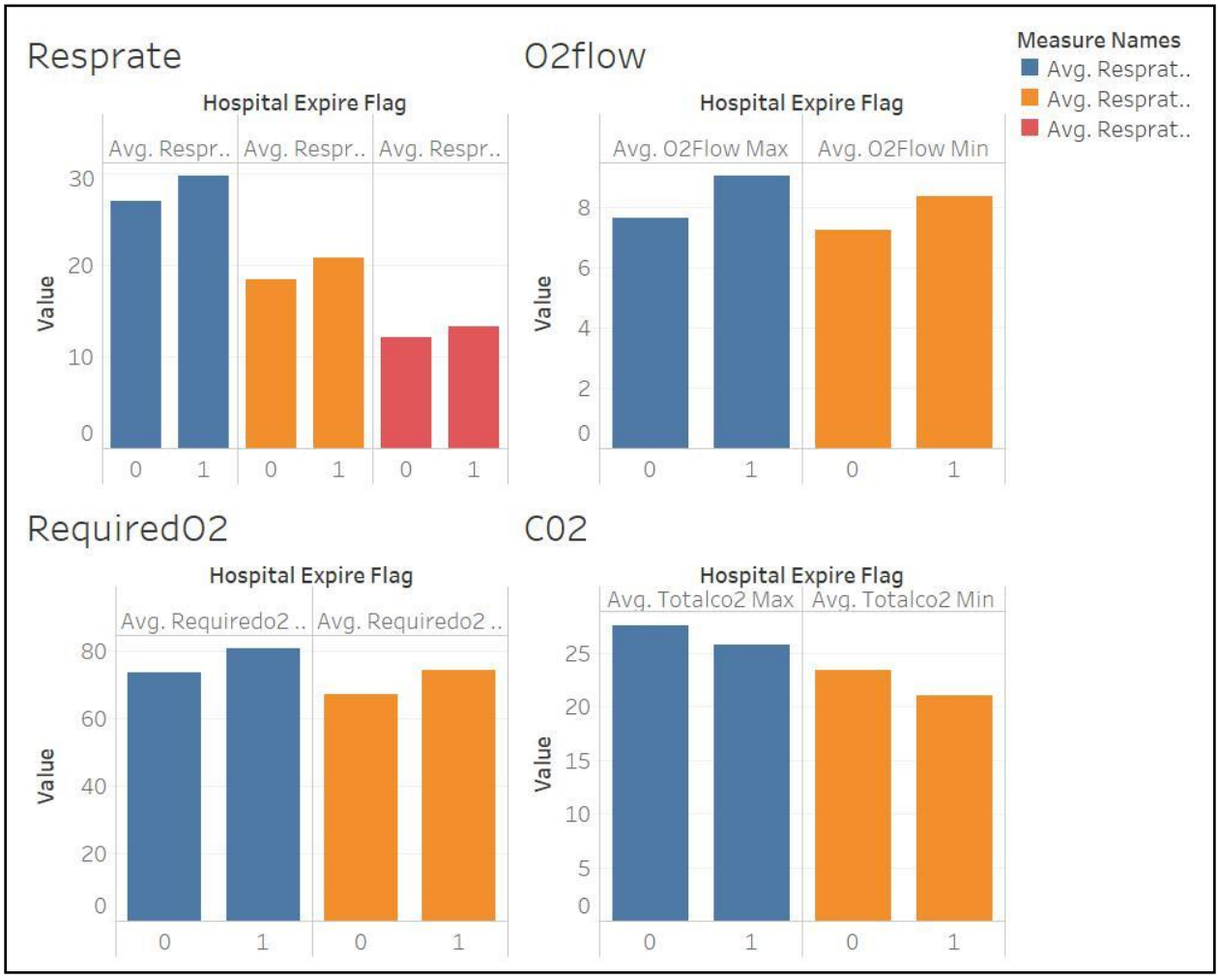
Appendix B: Sample of Visualization of decision tree



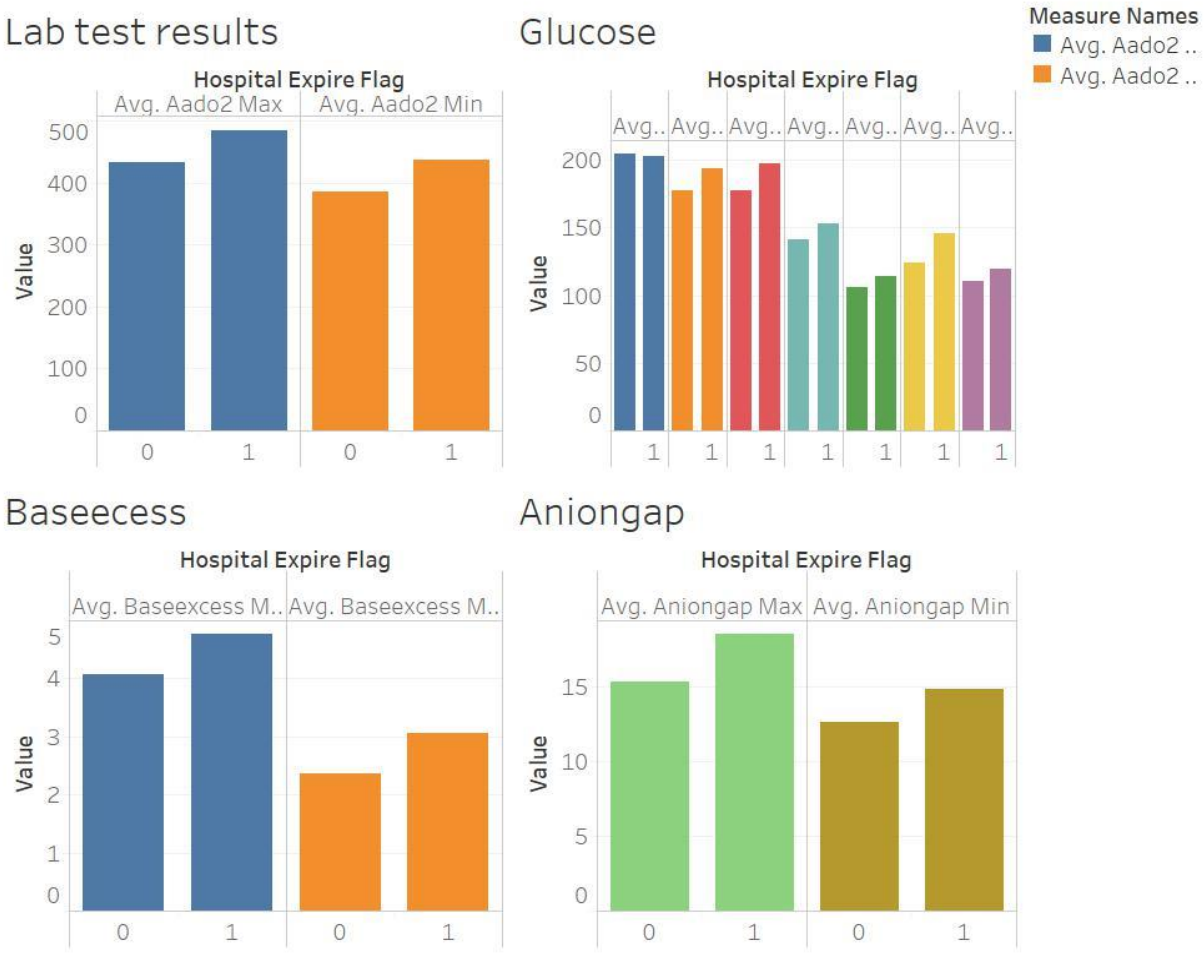
Appendix C: Average vital measures by response class



Appendix C: Average respiratory function test measures by response class



Appendix D: Average lab tests by response class


Lab test results
Hospital Expire Flag
Avg. Aado2 Max
Avg. Aado2 Min
Value
500
400
300
200
100
0
0
1
0
1
Glucose
Hospital Expire Flag
Avg..
Value
200
150
100
50
0
1
Measure Names
Avg. Aado2 ..
Avg. Aado2 ..
Baseecess
Hospital Expire Flag
Avg. Baseexcess M..
Value
5
4
3
2
1
0
Aniongap
Hospital Expire Flag
Avg. Aniongap Max
Avg. Aniongap Min
Value
15
10
5
0