Author: Angelica Butler

Instructor: James Meredith

Institution: UTA QuickStart Bootcamp March 2024 Cohort

Active Project Dates: September 22, 2024

Title:

Predictive Analysis for Identifying COVID-19 Patients needing Intensive Care Unit (ICU) Admission.

Abstract

During this research, I was able to develop some machine learning models that will predict which COVID-19 patients are most likely to need to be admitted to the ICU. Using the data gathered from the patients, I explored which components are more critical when building an effective and efficient machine-learning model for predictive analysis. This conclusion was discovered using exploratory data analysis (EDA), feature selections, model development, and hyperparameter tuning techniques. The results from the machine learning model will show that I have obtained an excellent accuracy score and feel confident about the overall predictability. The implementation of this model will enhance the clinical decision-making process.

Introduction/Business Problem

The COVID-19 pandemic has had a profound impact on healthcare systems worldwide. Healthcare workers face unprecedented pressure and challenges as they navigate the chaos of the crisis. Key factors contributing to the strain include:

- ICU Bed Availability
- Access to Personal Protective Equipment (PPE)
- Staffing and Personnel Shortages
- Overall Healthcare Resource Constraints

This research is mainly about the number one strain on the list: ICU Bed Availability. It is not only about availability but also about assessing which patients need this type of care. It is critical to allocate the space available appropriately. Building a model that can predict the patient's need for ICU admission in a timely manner can assist with improving the rates. This paper will explore how I build a machine-learning algorithm that can support and predict ICU admissions using patient-required data obtained during the patient's initial admission into the hospital.

Data

The dataset used in this research was the Sírio-Libanês data for AI and Analytics by Data Intelligence Team from [https://www.kaggle.com/datasets/S%C3%ADrio-Libanes/covid19]. This dataset contains COVID-19 clinical data to assess the initiate diagnosis. The feature in the dataset includes demographic information, patient previously grouped diseases, blood results, vital signs, and blood gases. Before any changes to the data, there is a total of 385 patients.

Data Cleaning

Once the data was loaded into a data frame and reviewed, I begin the data cleaning and preprocessing.

Label Encoder:

I first recognized the different data types that make up this data. This data primarily consists of integers and floats. There were only two columns that were object types, "AGE_Precentil" and "WINDOW." I used the label encoder processor from Sci-kit Learn to turn the non-numerical data into numerical data (https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html).

Missing/Null Values:

There was a significant amount of missing data. I use the "fill" function to fill the missing values with values derived from the previous value or the next value surrounding the data (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.fillna.html). A few rows pertaining to one patient had missing data in columns, and others were believed to be necessary. It was decided to delete that patient information, decreasing the total number of patients from 384 to 384.

Exploratory Data Analysis (EDA)

The EDA process was conducted to analyze and under the various features and distribution of the data. The data was initially cleaned and scaled according to the Min-Max Scaler to fit between -1 and 1. I use most histograms, violin plots, and box plots to explore and visualize the relationship between the features.

One of the most important distributions I could see was in *exhibit* 1, the distribution of the amount of treatment time between non-ICU admits and ICU admits, which can show a significant difference. In *exhibit* 2, the violin plots show the age distribution of both the non-ICU admits and ICU admits.

Exhibit 1.

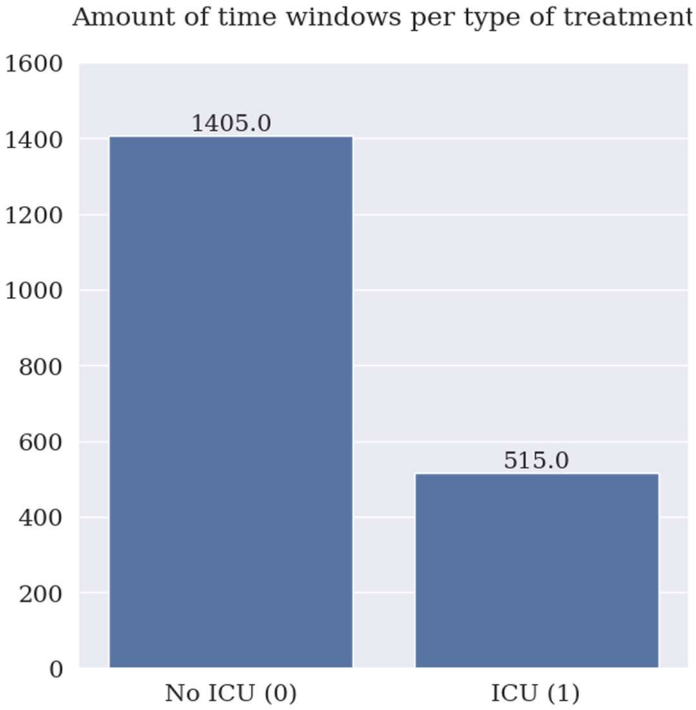
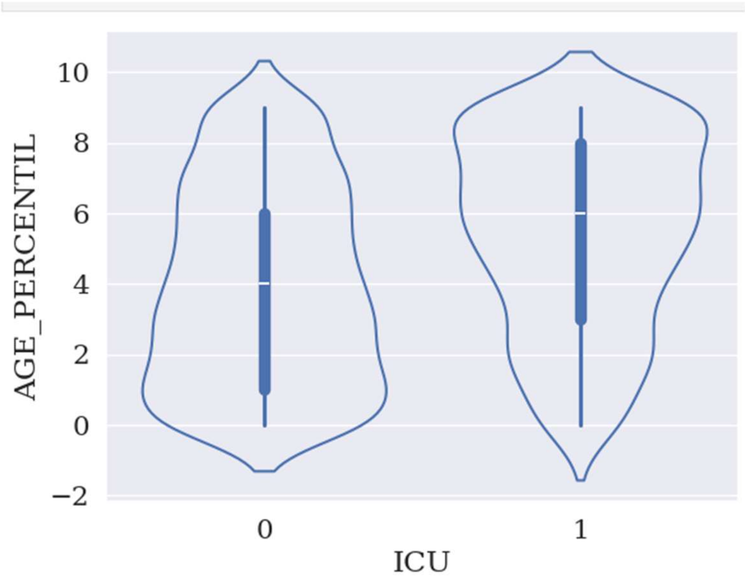


Exhibit 2.



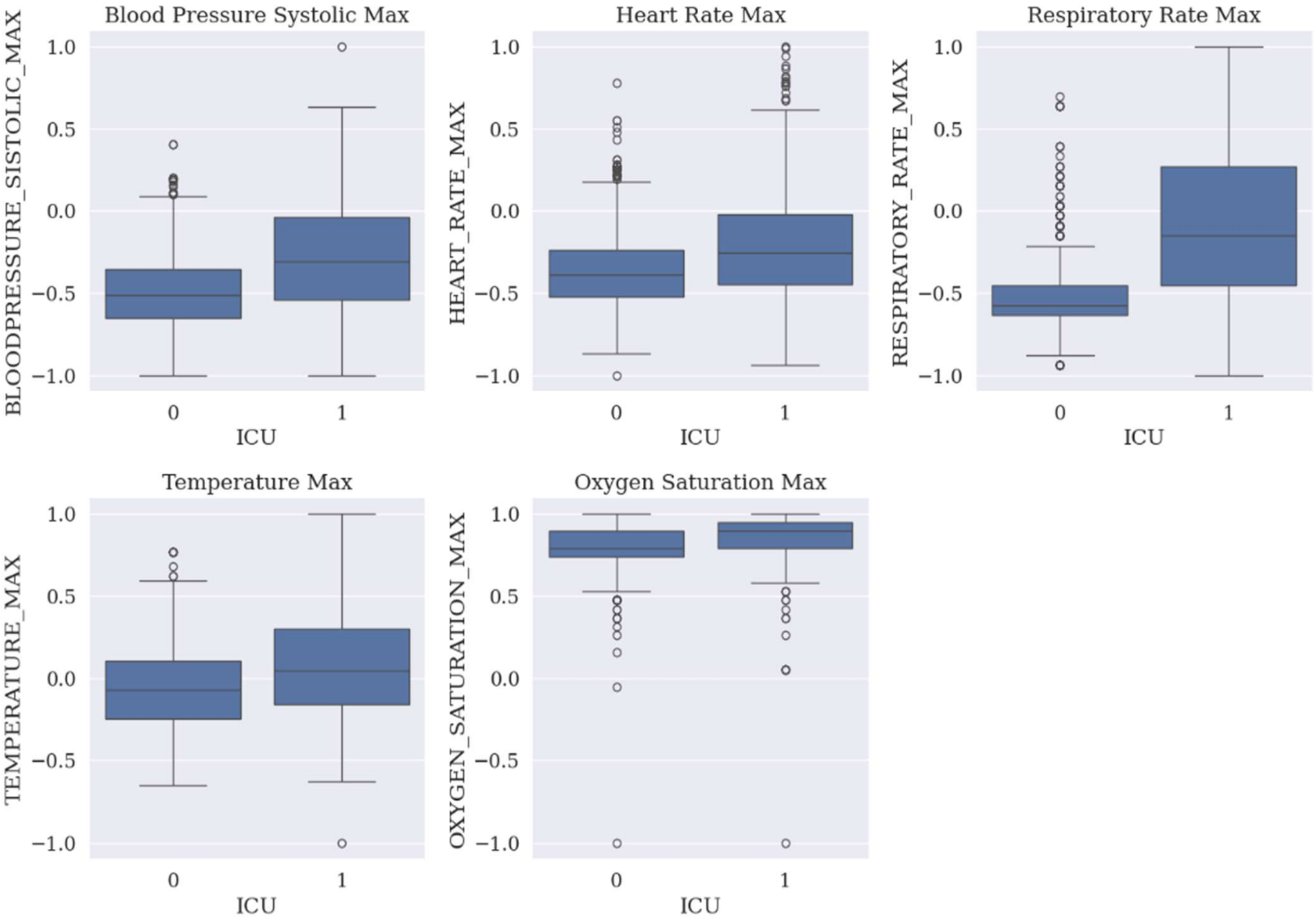
Because there were over 230 columns with data, except for the “AGE_Precentil,” “Gender,” and “Window” columns, subplots were used to analyze the rest of the columns simultaneously. I was able to use those histograms to identify the ones with the most variation. I took the five feature columns, which I believe are the most important when determining whether a patient should be in the ICU.

The box plots in *exhibit* 3 concluded the following:

Higher variability in ICU patients is evident for heart rate, respiratory rate, and systolic blood pressure max, which suggests that patients in the ICU tend to have more extreme and varied physiological responses.

Oxygen saturation and temperature don't differ as much between ICU and non-ICU patients, suggesting these variables may not be as critical in distinguishing between the two groups in this dataset.

Exhibit 3.



Model

Several machine learning models were used in testing to predict ICU admission. Each model was trained on 70%, with the remaining 20% reserved for testing. The model performance was evaluated using the metrics included in the classification report: accuracy, precision, recall, and f1-score.

The results for each model concluded as follows:

Logistic Regression:

Accuracy 86%

Precision (0) 87% and (1) 84%

Recall (0) 95% and (1) 64%

F1 Score (0) 91% and (1) 73%

Decision Trees:

Accuracy 82%

Precision (0) 87% and (1) 69%

Recall (0) 88% and (1) 69%

F1 Score (0) 87% and (1) 69%

Random Forest:

Accuracy 82%

Precision (0) 80% and (1) 94%

Recall (0) 99% and (1) 41%

F1 Score (0) 89% and (1) 57%

Overall, each model performs very well, with an accuracy rate above 80%. There are some imbalances between classification rates, which is expected when the data contains one variable versus another.

Hyperparameter Tuning

The decision tree model and the random forest model both produced an accuracy score of 82%. Because there are so many features/variables being fed into the model at once, it was safe to assume there was a chance to get a higher accuracy rate if we were able to narrow it down to only the features that really drove this model's performance. Random Search CV was used to optimize the hyperparameters of the random forest. After multiple runs of the model using different hyperparameters and cross-validation, the accuracy score increased to 86%.

The hyperparameters of the best-fitting model are the following:

- n_estimators: 300
- min_samples_split: 2
- min_samples_leaf: 1

- max_features: 40
- max_depth: 10

Conclusion

In conclusion, this research successfully built a machine-learning model that can reasonably predict ICU admission using clinical data. After the proper hyperparameters are included and tuned accordingly, the random forest model yields the highest overall accuracy rate. Hospitals can adopt and integrate this model into their systems to achieve a faster and more accurate ICU diagnosis when treating their patients.

Future Improvements

Some limitations of the study include the dataset's inherent biases (e.g., missing values or measurement errors) and the features' static nature, which does not capture real-time data changes. Future research should explore integrating real-time monitoring data to improve predictive accuracy.

A very important detail was the amount of data that was actually collected compared to what needed to be used to increase the model's accuracy. In the future, we can study other hyperparameters in the estimators that are used. That will make the process much more efficient because not as much data will need to be feed into the system on the front end.