**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 will need to be admitted to the ICU. Using the data gathered from the patients, I explored which components are more important when building an effective and efficient machine-learning model for predictive analysis. This conclusion was discussed using techniques such as exploratory da data analysis (EDA), feature selection, model development, and hyperparameter tuning. The results from the machine learning model will show I have obtained an excellent accuracy score and feel confident with 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 about the number one strain on the list, which is ICU Bed Availability. It is critical to allocate the space available appropriately. Building a model that can timely predict the patient's need for ICU admission 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 previous 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 it, I begin the data cleaning and preprocessing.

**Label Encoder:**

I first recognized the different data types that make up this data. Primarily this data consists of integers and floats. There were only 2 columns that were object types which are “AGE\_Precentil” and “WINDOW”. I use label encoder processor from scikit learn to turn the non-numerical data into numerical (<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html>).

**Missing/Null Values:**

The was a significant amount missing value. I use the fillna 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>). There was a few rows pertaining to one patient that had missing data in columns other that is believe to had importance. It was decided to delete that patient information in which decreasing the total number of patients from 384 to 384.

**Exploratory Data Analysis (EDA)**

The EDA process was conducted to analyses 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.

To important distribution I were able to see was in exhibit 1, the distribution of the amount of treatment time between non-ICU admits and ICU admits can show signification a difference. In exhibit 2, the violine plots shows the age distribution of both the non-ICU admits and ICU admits.

***Exhibit 1.***

A graph of a number of windows

Description automatically generated

***Exhibit 2.***

**A graph showing the number of objects

Description automatically generated with medium confidence**

Because there were over 230 columns with data, with the exception of the “AGE\_Precentil”, “Gender” and “Window column, subplots was used to analyze the reset of the columns simultaneously. I was able to use those histograms to identify the ones that had the most variation. I took the 5 I believe are the most important.

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 show as much difference 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.***

***A group of graphs showing different types of data

Description automatically generated with medium confidence***

**Model**

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

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 with an accuracy rate above 80%. There are some imbalances between classification rates, which is normal when the data contains many one variable versus another variable.

**Hyperparameter Tuning**

The decision tree model and the random forest model both produced an accuracy score of 82%. Because contained so many features/variables, I knew there was a chance to get a higher accuracy rate if we are able to narrow down the features that really drives this model performance. Random search was used to optimize the hyperparameters of the random forest. After multiple times of running the model with using different hyperparameters and cross-validation, the accuracy score was able to increase to 86%.

The hyperparameters of the best fitting model is the following:

* n\_estimators: 300
* min\_samples\_split: 2
* min\_samples\_leaf: 1
* max\_features: 40
* max\_depth: 10

**Conclusion**

In conclusion this research was successful in building a machine learning model that will be able to predict ICU admission using clinical data within a reasonable accuracy rate. The random forest model yields the highest overall accuracy rate after the proper hyperparameters was included and tune accordingly.

**Future Improvements**

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

The amount of data that was actually collected verses what needed to be use to increase the accuracy in the model was a very important detail. In the future we can study other hyperparameter in the estimators that was use.