Student ID: 21485094
Beatriz Manso

ICU ADMISSIONS - EXECUTIVE SUMMARY REPORT

Introduction

Among the most promising areas of research is healthcare data analysis. Clinicians collect and store electronic health records, which contain patient records collected throughout the treatment process.

It is very difficult to manage raw data manually and, as a result, machine learning is becoming an increasingly important tool for analysing this data. With the help of machine learning, healthcare data can be predicted more accurately using various statistical techniques and advanced algorithms, such as supervised, unsupervised and reinforcement learning.

In this report, we will analyse the clinical data collected from positive covid-19 patients from the Brazil's Hospital Sírio-Libanês and to assess if a patient will need to be admitted to the ICU.

Exploratory Data Analysis

In the Exploratory Data Analysis, we take a quick glance of our data to understand what information it contains and how we will handle it to get better predictions.

From a quick look at our data, we observe that there are 1925 rows and 231 columns that have important information about 385 patients from the Brazilian hospital.

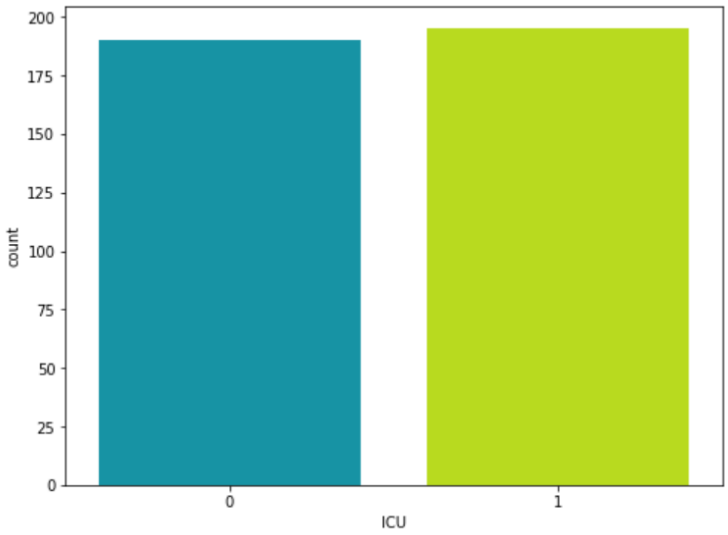

Figure 1. Number of ICU admissions

From the 385 patients in our dataset, we can see that 195 patients were admitted to the ICU and 190 weren't. We can consider this distribution of data close enough and our target variable is well balanced.

Student ID: 21485094
Beatriz Manso

Figure 2. Sample of some features with high percentage of missing values

The graph above shows just a sample of 10 features with the highest missing values percentage. These 10 features show that there are almost 60% of missing data.

From our analysis, in all our data set there are 225 variables with missing values and in total, 223863 NaN values.

The missingness of values in our dataset is a problem that we will have to deal because a machine learning model can't learn from values that aren't there.

Missing values in medical data is a common issue. However, according to the authors of the dataset, we can assume the patients that don't have a measurement recorded are clinically stable and there is a possibility that they have vital signs and blood labs similar to neighbouring windows. Therefore, in the data preparation stage, we dealt with missing values by filling them using the next or previous entry.

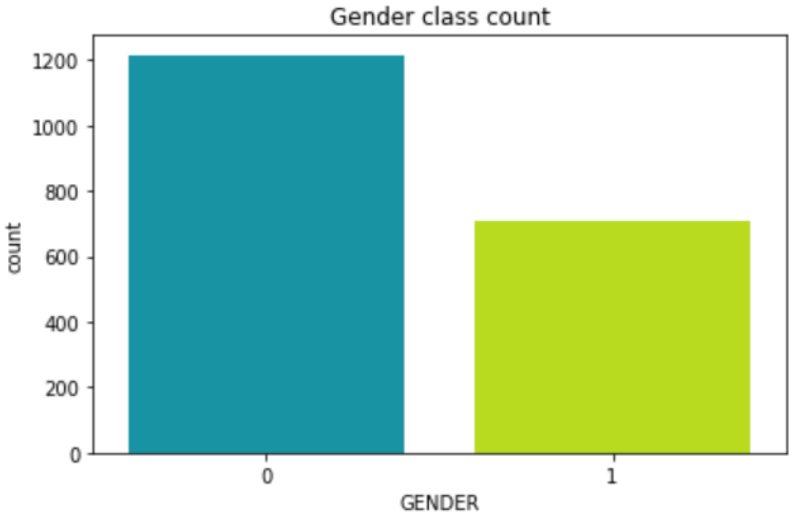

Figure 3. Gender distribution

In our dataset there are 243 males and 142 females. We conclude that our dataset is not very well balanced in the gender feature, as there are almost the double number of males comparing to the number of females. Ideally this feature should be balanced.

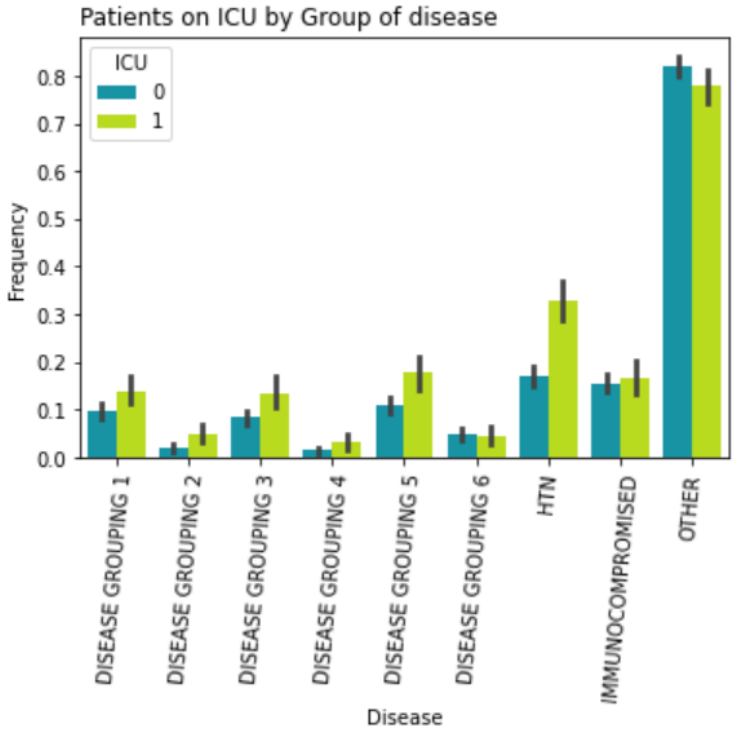

Figure 4. Frequency of Patients admitted to the ICU by disease

The diseases are anonymised, because of the data protection policy, but we can infer that, except for the disease group 6, patients with some sort of disease have a higher ICU admissions frequency.

Student ID: 21485094
Beatriz Manso

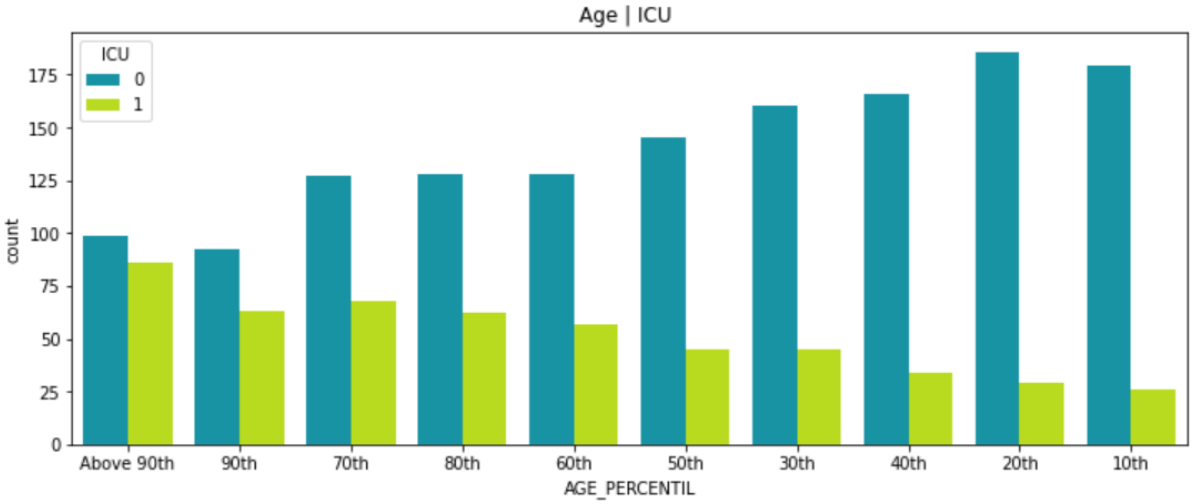

Figure 5. ICU admissions based on the Age percentile

We can observe that the number of ICU admissions is higher as the percentile of age goes up as well. Patients with age above 90 have almost a 50 per cent chance of having to be admitted to the ICU.

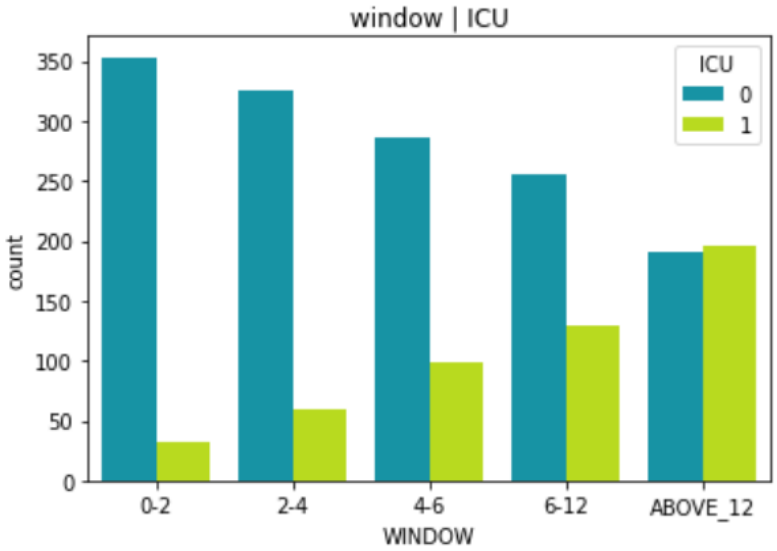

Figure 6. Number of ICU admissions by window of time

From the graph above we can observe that the number of people that were admitted in the ICU, the window of above 12 hours has the highest percentage of people being admitted to the ICU.

The percentage of ICU admissions goes up as the window of hours since admission into the hospital increase, which is expected.

Student ID: 21485094
Beatriz Manso

Data Preparation

During the data preparation step, the data is cleaned to be able to fit into our machine learning model.

In our case, for this specific dataset, we created a column called ICU_SUM to identify if a patient ever went to the ICU, for our model to be trained with this column. So, the ICU_SUM it our target column.

Then, we dropped the rows that had ICU equal to 1 because their data might have been collected after the admission so it's not useful for our model to learn as stipulated by the author.

The authors also recommended to use the window of time between 0-2 hour of admission to the hospital because it is better to be able to predict if a patient will ever need ICU intervention as soon as possible, so we selected the data from the window 0-2 to fit into our model.

The patient ID column, window column as well as ICU column were then dropped because they don't hold value for our model.

Finally, we encoded the AGE_PERCENTIL column, which transforms the categorical values of this feature into numerical, because a machine learning model can only learn from numerical values.

At the end of our data preparation step, 1572 rows and 2 columns have been dropped, and we end up with 353 rows and 229 columns. We don't have a lot of data to get a very accurate ML model so ideally, we would get more data from other sources.

Machine Learning Model Comparison

Now that our data is ready to fit a ML model, we will use some classification models to see which performs the best with our cleaned dataset.

Below are the results for each model used:

- Logistic Regression - AUC test/ train: 0.69 - 0.84
- Decision Tree Classifier - AUC test/ train: 0.62 - 1.00
- Random Forest - AUC test/ train: 0.67 - 1.00
- Support Vector Machine - AUC test/ train: 0.72 - 0.78

AUC is a measure of a classifier's ability to discriminate between ICU admission being positive or not.

As we can see from the results obtained, the results values are not too different between each model, but the worst performer was Decision Tree and SVM was the model that had the best AUC value.

We can now further evaluate our model, using the confusion matrix and some evaluation metrics such as accuracy, precision, recall and f1-score.

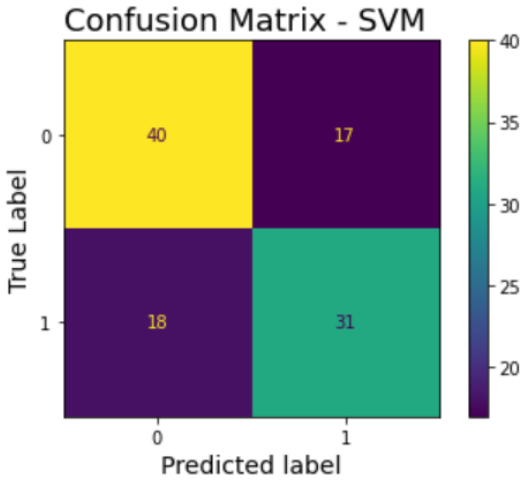

Figure 7. Confusion Matrix for SVM model

An overview of the classification model is provided by the confusion matrix. In order to achieve better performance, TP(True positives), TN(True negatives) should be high and FN (false negatives), FP(false positives) should be as low as possible.

TP – 31%
TN – 18%
FN – 40%
FP – 17%
We can see that our model is predicting a lot of false values that do not correspond to reality.

	precision	recall	f1-score	support
0	0.69	0.70	0.70	57
1	0.65	0.63	0.64	49
accuracy			0.67	106
macro avg	0.67	0.67	0.67	106
weighted avg	0.67	0.67	0.67	106

Figure 8. Classification report

Precision – True positives are compared with total predicted positives to find the precision ratio.

Recall – Also called sensitivity, identifies how many true positives are being classified correctly.

F1-score – Calculates the harmonic mean between the precision and recall.

$$\text{F1 Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Figure 9. F1 score calculation

Accuracy – Calculates the percentage of correct predictions compared with the total number of input samples

Student ID: 21485094
Beatriz Manso

Conclusion

We were able to create a model that predicts the covid19 patients' admission to the ICU with a accuracy of 67%. However, this could be improved by enhancing our model.

To make our model a better predictor we should find more data for training and do hyperparameter tuning to understand which features contribute the most to a good prediction.