Predicting if a COVID-19 Patient Requires Intensive Care Unit (ICU) Care

Nicholas Klapatch
Erick Garcia-Vargas
Thomas Olandt
Kyle Sacco

CSCI-4105: Knowledge Discovery and Data Mining, Fall 2021
November 27, 2021

Abstract:

In this study, we worked with a dataset regarding COVID-19 related health measures collected by a hospital. With this data, our goal was to identify whether a patient would require special care in an intensive care unit (ICU) or not based on the health measures provided in the dataset. Along with the health measures, there was a target variable that indicated whether the patient actually required ICU care or not. The dataset was organized in a way suited for a binary classification problem. By accurately predicting this target variable, we could improve the logistics involved with patient care in hospitals. This can lead to better care for patients who need it the most and help anyone working in healthcare to make more informed decisions.

The primary motivation for this project was to discover new knowledge from this dataset and to allow for others to apply it to related datasets in the future. The COVID-19 pandemic is undoubtedly a unique event. With our current technology and data collection capabilities, we can extract a lot of information that would have otherwise been impossible to. We originally wanted to classify patients into three groups based on risk of requiring ICU care: low, medium, and high. While reviewing the dataset, however, it became clear that it was not conducive to such a project. The dataset was very suitable for classifying low-risk and high-risk patients. What would have been a three-class classification problem became a binary classification problem.

For this study, we aimed to create a binary classification model that could accurately predict if a patient would require ICU care based on a few important health measures. This dataset was organized in a way that naturally allows for binary classification. Each health measure is a feature, and the ICU status is the target variable. Creating a model that accurately links the selected features with the target variable was the goal.

Introduction:

COVID-19 has been very intense. This intensity has led to many hospitals running out of ICU space, even well into 2021 [2]. Other than the main ways of combating the pandemic, (like social distancing, wearing a mask, getting vaccinated, etc.) there must be more ways to help prevent ICUs from reaching their capacities, especially if a patient has already been diagnosed with COVID-19 and admitted into a hospital. Medical professionals could use this information to save lives. With the data provided, how could the risk of a patient needing ICU care be determined? Could it be predicted if a patient will need ICU care or not? Some data features must tell more than others, and they can be used in a predictive binary classification model.

Each member of the team contributed to each phase of the project. As a group of four, we worked on the program code together, and we worked individually between meetings. We all contributed to the literature review, and report and presentation, both through writing and editing.

Methodology:

The methodology involved in this project began with preprocessing, where the most relevant data for the processing step was collected from the dataset. The next step was processing, where new knowledge was discovered from the data using a classification algorithm. Finally, the data was post processed. This step was where the performance of the classifier was analyzed, as well as conclusions made about its results.

Preprocessing:

To begin working with any dataset, it is necessary to understand how the data is organized. In this dataset, there was data recorded for 385 unique patients, each with 230 features excluding the classification of the patient's ICU status: presence in the ICU was classified by a 1, and absence by a 0. The data was

recorded over five different time slots for each patient, so there were 1925 rows, excluding the header. The data was delivered both normalized and cleaned, so fewer steps were required in preprocessing. Upon viewing the data, some time slots were missing more data than others. Missing values were denoted by a blank space. This was important for deciding what data would be used for the classification model, and it also had strong implications on what the model ultimately classified.

As mentioned before, there were five records for each patient recorded in this dataset. This meant that by working with all five, there could be duplicate, close to duplicate, or even incomplete data that could negatively influence the classifier. It is a good practice to remove these [1, p. 48]. Some of the time slots were missing more data than others, which influenced which data should be used to train the classifier. It is a common practice to drop all records that are missing data [1, p. 46]. For this reason, we decided it would be in our best interest to drop records that did not contain complete information. This way, the amount of data used for the classifier was appropriately maximized. Using too much data, however, can be troublesome for classification. It is necessary to collect a quality feature subset, especially to rule out any irrelevant features [1, p. 58]. It is also important to search for outliers, as they too can negatively influence the classifier [1, p. 46].

To decide which time slot to work with, we calculated how many rows in each window did not have any NAN values. This way, the number of missing values could be estimated for each feature. This has great implications on choosing the fifty best features because the number of training examples could remain sufficiently high. In Figure 1, it is shown that Window 5 (Above 12 Hours from Admission) contained the most data points, by far. It is important to note that this time slot was larger than the other four, and unbounded, which could impact the classifier. In this window, the patients were in hospital care for the longest period recorded. This could lead to more accurate results in the classifier.

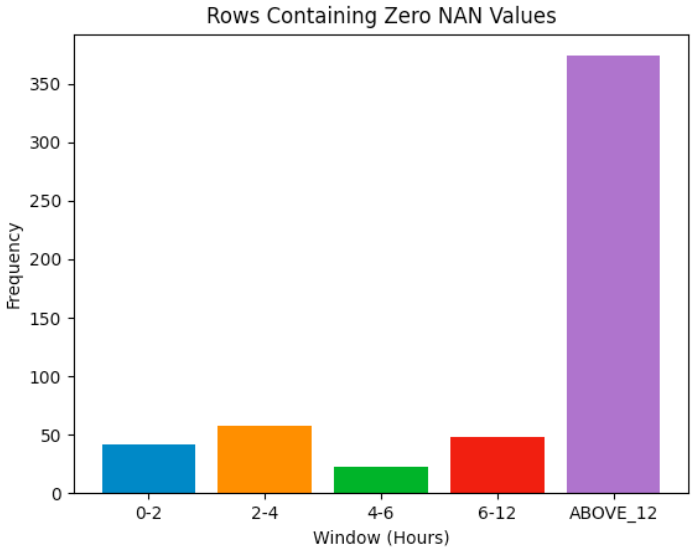


Figure 1

In our study, we wanted to find some of the data features that correlate most with the ICU target value. This is for finding a quality feature subset. After working specifically with the fifth window, we found the features that correlate most with the classification value. Including some of these features should enhance the performance of the classifier. In Figure 2, some of the unique features most correlated with the target value are shown. Of these, we chose: RESPIRATORY_RATE_DIFF_REL, BLOODPRESSURE_SISTOLIC_DIFF, HEART_RATE_DIFF, TEMPERATURE_DIFF_REL, and OXYGEN_SATURATION_MAX to use with the classifier. After choosing this Window and these features, it was determined that only 1 row had missing value(s). This row was simply dropped. For the extra credit option, a similar approach was followed. Except this time, the top fifty most correlated ICU

values were selected. These values are correlated by absolute value, so both highly positively and negatively correlated values were used. For the same reasons we expected the top five most highly positively correlated features with the target variable to enhance the performance of the classifier, we expected the top fifty to have similar effects.

```
25 most correlated features with ICU target variable:
Feature Correlation with ICU
RESPIRATORY_RATE_DIFF_REL          0.806281
RESPIRATORY_RATE_DIFF              0.802676
RESPIRATORY_RATE_MAX               0.770367
RESPIRATORY_RATE_MIN              -0.667276
BLOODPRESSURE_SISTOLIC_DIFF        0.616881
RESPIRATORY_RATE_MEAN              0.607315
BLOODPRESSURE_SISTOLIC_DIFF_REL    0.605053
BLOODPRESSURE_DIASTOLIC_MIN       -0.602732
BLOODPRESSURE_SISTOLIC_MAX         0.556393
LACTATE_MIN                       -0.549559
LACTATE_MAX                       -0.549559
LACTATE_MEAN                      -0.549559
LACTATE_MEDIAN                    -0.549559
BLOODPRESSURE_DIASTOLIC_DIFF_REL   0.547719
BLOODPRESSURE_DIASTOLIC_DIFF       0.526571
RESPIRATORY_RATE_MEDIAN            0.525372
HEART_RATE_DIFF                    0.506314
HEART_RATE_DIFF_REL                0.499959
TEMPERATURE_DIFF_REL               0.470005
TEMPERATURE_DIFF                   0.467971
OXYGEN_SATURATION_MAX              0.450272
TEMPERATURE_MIN                   -0.400927
HEART_RATE_MAX                     0.400081
BLOODPRESSURE_SISTOLIC_MIN        -0.393016
HEMOGLOBIN_MAX                    -0.381881
Name: ICU, dtype: float64
```

Figure 2

In the remaining data, we searched for outliers using a box plot. There are very few in the data, as shown in Figure 3.

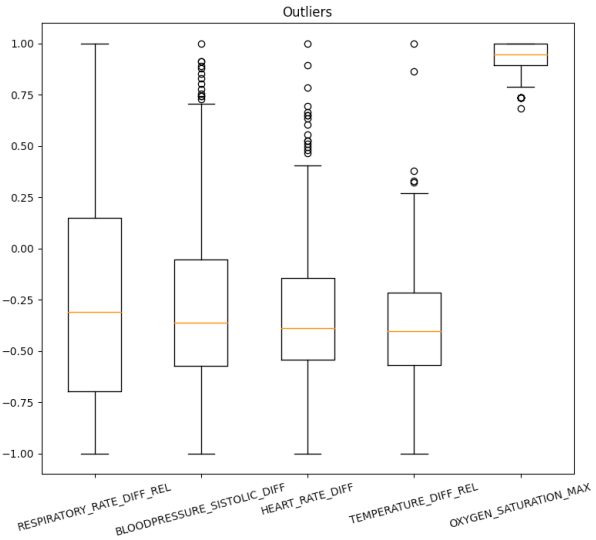
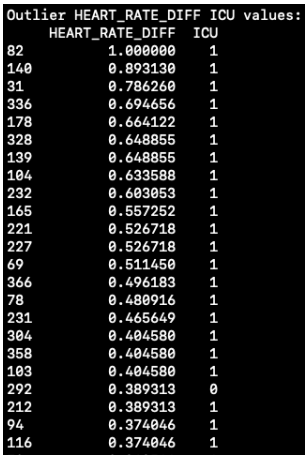


Figure 3

Upon reviewing the outliers, we decided that it was not worthwhile to remove them. First, this was because the data size was already limited to 384 samples. Second, when observing the outliers in HEART_RATE_DIFF, virtually all the outliers had the same target ICU value, as shown in Figure 4. This indicated that these values were meaningful, and it would be counterproductive to remove them. At this point, the data was preprocessed, stored in a Pandas Dataframe, and ready to be processed with a classification algorithm.



```
Outlier HEART_RATE_DIFF ICU values:
     HEART_RATE_DIFF  ICU
82          1.000000    1
140         0.893130    1
31          0.786260    1
336         0.694656    1
178         0.664122    1
328         0.648855    1
139         0.648855    1
104         0.633588    1
232         0.603053    1
165         0.557252    1
221         0.526718    1
227         0.526718    1
69          0.511450    1
366         0.496183    1
78          0.480916    1
231         0.465649    1
304         0.404580    1
358         0.404580    1
103         0.404580    1
292         0.389313    0
212         0.389313    1
94          0.374046    1
116         0.374046    1
```

Figure 4

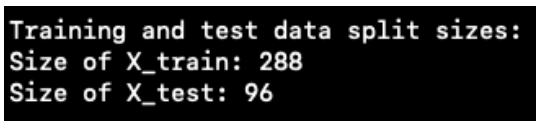
For preprocessing, two external Python libraries were used:

- Pandas: For preprocessing, calculating correlation, and Dataframes.
    - Openpyxl: Used for reading COVID_ICU_Prediction.xlsx into the Pandas Dataframe.
- MatPlotLib.pyplot: For displaying the bar chart and box plot.

Processing:

We originally intended on creating a predictive model that would classify three groups of patients based on risk of being admitted into an ICU. We later decided that the data is better suited for binary classification. As such, we studied different binary classification predictive models. Logistic regression was the technique that we settled on. The name implies that the model is used for regression, but it is not—it is used for binary classification [3, p. 205]. Logistic regression uses the sigmoid function (along with the features and their predicted coefficient values) which always returns a value between 0 and 1 [3, p. 197]. This was very convenient for our intentions, as the target variable in the data is already formatted this way.

The features that we used for the classifier were the ones that we decided on in preprocessing, and the target variable was the one provided in the data set. Using a training set of known features and target variables is an important step of learning a classifier and applying its results to a set of unknown values, or a test set, is important to evaluate its performance [1, p. 117]. The set of data points was divided into a training set, and a test set using a function from the Scikit-learn (sklearn) library. Sklearn is a Python machine learning library that interfaces well with the other libraries that we used in preprocessing, particularly Pandas. In Figure 5, the sizes of the training and test sets that the "train_test_split" function from the sklearn library generated are shown. From the training set, a logistic regression model would be created that would be used on the test set to determine the accuracy in the post processing phase.



```
Training and test data split sizes:
Size of X_train: 288
Size of X_test: 96
```

Figure 5

The implementation for the logistic regression model we used was also from the sklearn library (https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html). This allowed for an accurate implementation using five features from the dataset. The results could easily be assessed in the post processing phase. At this point, the data was processed, and the performance of the classifier was ready to be analyzed.

For processing, an additional external Python library was used:

- Scikit-learn (sklearn): For splitting the data into training and test sets and for a logistic regression implementation.

Results:

Obtaining the results was the post-processing phase of the study. In this phase, the accuracy, precision, and recall of the classifier were measured. Additionally, a confusion matrix was generated to determine exactly where the classifier exceled, and where it was limited. To calculate each, a function from the sklearn library was used. Each measurement is shown in Figure 6.

The accuracy, precision, and recall of the classifier were all very high. Additionally, the confusion matrix shows that the classifier produced very few false positives and false negatives. It is important to note that the function used to generate the training and test sets is randomized, so the results will be slightly different each time the program is executed, but not by much. For the extra credit option, the analysis of the classifier is shown in Figure 7. The classifier uses a different random training set than the previous classifier. Since different features were used, the sizes of the training and test datasets are also different because different rows have missing values in both.

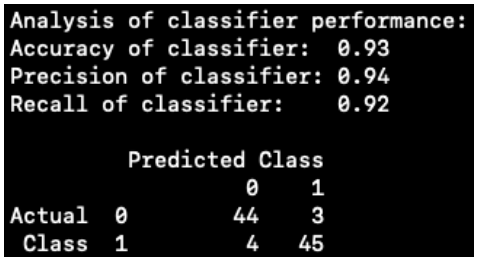


Figure 6

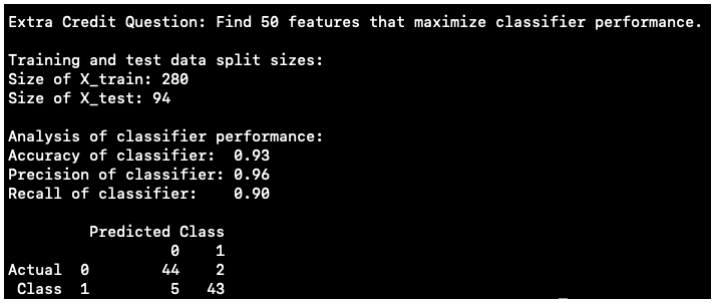


Figure 7

Conclusion:

In this project, we found that it was possible to accurately predict if a COVID-19 patient admitted into a hospital would require ICU care. The measurements provided in this dataset seemed to be easy to measure, and when placed into this classifier, a healthcare professional could make an informed decision. Not all of the features in the data were necessary to make this prediction. We used only five unrelated features, all from just one of the five time windows that were measured. Logistic regression was very suitable for this binary classification problem, and provided very strong results.

Reference List:

[1] Pang-Ning Tan, Michael Stenibach, Anuj Karpatne, Vipin Kumar. *Introduction to Data Mining*, 2nd ed. New York, NY, USA: Pearson Education Inc., 2019.

[2] Charlie Smart. "Covid Hospitalizations Hit Crisis Levels in Southern I.C.U.s." The New York Times. https://www.nytimes.com/interactive/2021/09/14/us/covid-hospital-icu-south.html (accessed November 25, 2021).

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Science + Business Media, LLC., 2006.

Appendix A. Grading Rubric for Project Paper

	A+ (100)	A (95)	B (85)	C (75)	D (65)	F (50)
Overall	Clear & consistent mastery; may have a few minor errors.	Reasonably consistent mastery; occasional errors or lapses in quality.	Adequate mastery; lapses in quality.	Developing mastery; shows one or more of the qualities below.	Little mastery; shows one or more of the qualities below.	No mastery; shows one or more of the qualities below.
Critical Thinking	Effective and insightful development of a point of view on the issue. Outstanding critical thinking; use of clear and appropriate examples and other evidence to support the position.	Effective development of a point of view on the issue. Strong critical thinking; use of appropriate examples and other evidence to support the position.	Adequate development of a point of view on the issue. Competent critical thinking; use of adequate examples and other evidence to support the position.	Some development of a point of view on the issue. Some critical thinking; inconsistent use of examples and other evidence to support the position.	Vague or limited development of a point of view on the issue. Weak critical thinking; insufficient or inappropriate use of examples and other evidence to support the position.	No development of a point of view on the issue. No use of examples and other evidence to support the position.
Organization	Well organized and clearly focused. Shows clear coherence and smooth progression of ideas.	Well organized and focused. Shows coherence and progression of ideas.	Generally organized and focused. Shows some coherence and progression of ideas.	Limited organization and focus. Shows lapses in coherence and progression of ideas.	Poor organization and focus. Shows serious problems with coherence or progression of ideas.	Disorganized and unfocused. No coherence or progression of ideas.
Language Use	Skillful use of language; varied, accurate, and apt vocabulary.	Competent use of language; appropriate vocabulary.	Adequate use of language; generally appropriate vocabulary.	Some adequate use of language; vocabulary may be weak or inappropriate.	Inadequate use of language; limited vocabulary or incorrect word choice.	Fundamental errors in vocabulary.
Prose Style	Meaningful variety in sentence structure.	Variety in sentence structure.	Some variety in sentence structure.	Lacks variety or demonstrates problems in sentence structure.	Frequent problems in sentence structure.	Severe flaws in sentence structure.
Mechanics	Free of most errors in grammar, usage, and mechanics.	A few errors in grammar, usage, and mechanics.	Some errors in grammar, usage, and mechanics.	Many errors in grammar, usage, and mechanics.	Errors in grammar, usage, and mechanics tend to obscure meaning.	Severe errors in grammar, usage, and mechanics obscure meaning.

Appendix B. Grading Rubric for Project Presentation

		1	2	3	4
Non-Verbal Skills	**Eye Contact while reading**	Student reads with no eye contact.	Student occasionally uses eye contact.	Generally looks at the audience, but generally to the teacher.	Student is able to present the project looking at the audience and making them feel included.
	Posture	Slumps or leans during presentation.	Sways or fidgets during much of presentation.	Occasionally sways or fidgets, but stands up straight with both feet on the ground most of the time.	Stands up straight and still with both feet on the ground, and moves the hands for emphasis.
Oral Skills	**Elocution**	Student mumbles, very low voice and do not use any tonal differences.	Student's voice is low-medium, but part of the audience still has some difficulty hearing presentation. Tonality barely changes.	Student's voice is clear, and most of the audience members can easily hear the presentation. The tone used changes.	Student uses a clear voice, rhythm and tone, so that all audience members can hear presentation.
	Pronunciation	Student does not do any effort regarding pronunciation.	Student pronounces incorrectly some terms, mostly vocabulary of the unit.	Pronunciation is good, but some constructions and terms are incorrect.	Student pronounces mostly everything clearly and correctly.
Contents	**Organization**	Audience cannot understand the presentation because there is no sequence of information.	Audience has some difficulty following presentation because student jumps around.	Students presents information in a logical sequence which audience can follow.	Student has a good hook and presents information in logical, interesting sequence which audience can easily follow.
	Subject Knowledge	Student does not appear to have a grasp of information; cannot answer questions about subject.	Student is comfortable with information, but is only able to answer simple questions.	Student is at ease with information and answers questions satisfactorily, but fails to elaborate.	Student demonstrates full knowledge and can answer and elaborate on most/all questions asked
Presentation	**Visual**	The presentation had small fonts and blurry pictures. It has been difficult to follow.	The images used changed from blurry to high-resolution. Text varied depending on parts.	The audience could read the slides and the images were generally good.	Visual aid showing effort and creativity is used thus improving overall presentation.
Teamwork	**Coordination**	The team did not know when to speak, or what role were having. Only one person leads the group.	One or two members of the group have focused most of the presentation. The rest of the group did not have clear instructions about their role.	The team was mostly coordinated, but there were some moments of doubt and/or unbalance. A minority of the members of the group did not know what to do.	The team run perfectly coordinated, with clear guidelines about each member's role. Each member has participated..