Predicting the Severity of Health Status of an Infected COVID- 19 Individual

Vadiwoo Karuppiah
Department of Computer Science
Syracuse University
New York, USA
vkaruppi@syr.edu

ABSTRACT

Globally, Coronavirus is on everyone's mind and forcing almost all of us to stay at home to prevent the spread of the virus. Coronavirus also is known as COVID-19 appeared in Wuhan, China in December 2019, and since then it has immediately become a serious public health problem worldwide and has killed 430K people [1] as of June 14, 2020. A vaccine to prevent coronavirus disease 2019 (COVID-19) would be the best hope for ending this pandemic but there are no effective treatments to counteract the effects of the illness yet. While healthcare researchers are racing to create a vaccine, machine learning researchers are combining their efforts to collect data and develop solutions to minimize the spread and the impact of this pandemic. Machine learning has proven to be invaluable in predicting risks in many healthcare systems, and in this study, we are using machine learning techniques to predict the health status of uninfected Covid-19 individuals, to be Critical or Stable based on demographic and health history of that Individual.

Keywords

COVID-19; Decision Tree; Random Forest; Naive Bayes; Gradient Boosting Machine; Random Forest; Ensemble

1. INTRODUCTION

Machine learning is an important tool in fighting the current COVID-19 pandemic. If we take this opportunity to collect data, pool our knowledge, and combine our skills, we can save many lives — both now and in the future. In a normal situation, hospitals would take time to test the tool on hundreds of patients, refine the algorithm underlying it, and then adjust care practices to implement it in their clinics. However, with Covid-19, hospitals do not have the luxury of following these practices, they need to make much faster decisions to predict who is at highest risk so that the hospitals can be prepared with the necessary treatments. Once a person or group has become infected, we need to predict the risk of that person or group developing complications or requiring advanced medical care. Knowing ahead of time whether mechanical ventilation might be necessary is helpful because doctors can ensure that an ICU bed and a ventilator or other breathing assistance is available. Many people experience only mild symptoms, while others develop severe lung disease or acute respiratory distress syndrome (ARDS), which is potentially deadly. It's not possible to treat and closely monitor everyone with mild symptoms, but it's far better to start treatment early if more severe symptoms are likely to develop.

2. STUDY OBJECTIVES

The goal of this study is to predict the health status of Covid-19 individual, to be critical or stable based on demographic data and health history of that Individual. We will also analyze what are the most relevant features to predict if a patient infected with Covid-19, will be in stable or critical condition? Can we find the best combination of features to reach a high level of accuracy? We will also investigate which Machine Learning Algorithm will produce better accuracy?

3. DATA PREPROCESSING

3.1 Data Understanding

Johns Hopkins University has made an excellent dashboard[2] using the affected case data from COVID-19. The data was extracted from associated google sheets and made available on Kaggle to be used by the data science community. The main dataset has daily level information on the number of affected cases, deaths, and recovery from the 2019 novel coronavirus. The Kaggle link also added new files with individual-level information, and the COVID19_2020_open_line_list.csv [3] dataset is the one I am using for this study.

This dataset contains 929729 rows and 33 columns of data. These columns consist of 2 numerical features, 2 logical features, and 29 features are categorical. The description of each of these features was included in Appendix A.' The outcome' variable is used as the target variable.

3.2 Challenges in Data Preprocessing

Data preprocessing is an important feature in machine learning. Data that is used in any machine learning system is often taken from multiple sources which are normally not too reliable and come in a different format. There may be problems due to human error, limitations of measuring devices, or flaws in the data collection process. It is simply unrealistic to expect that the data will be perfect, however, data preprocessing can be done to transform this data into quality data that can be fed into our machine learning modeling systems.

3.2.1 Large NA Values

Commonly, datasets contain missing values due to a variety of reasons. The COVID19_2020_open_line_list contains a huge number of missing values. There are several strategies for dealing with missing data, each of which is appropriate in certain circumstances. These strategies are listed below: -
.

I. Excluding Missing Values from Analyses
II. Impute missing values with mean, median, or mode.
III. Replace missing values with 0.

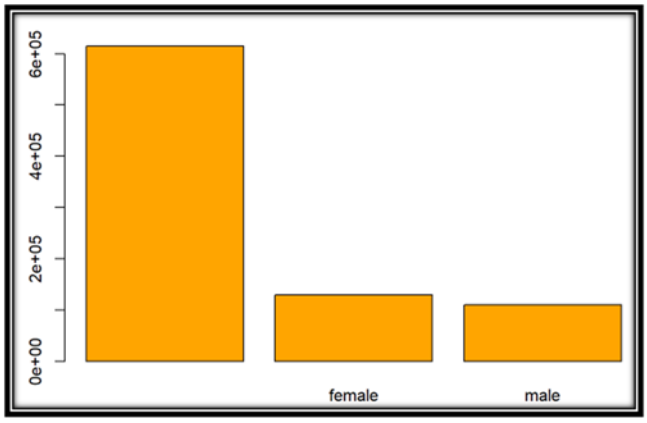

Figure 3.2.1: NA values in the feature 'sex'

Figure 3.2.1 shows the large NA values that exist in the 'sex' feature. This value can be imputed with the mode values. Excluding missing values is not a great idea for this feature, so I imputed this feature with mode values.

3.2.2 Outliers in data

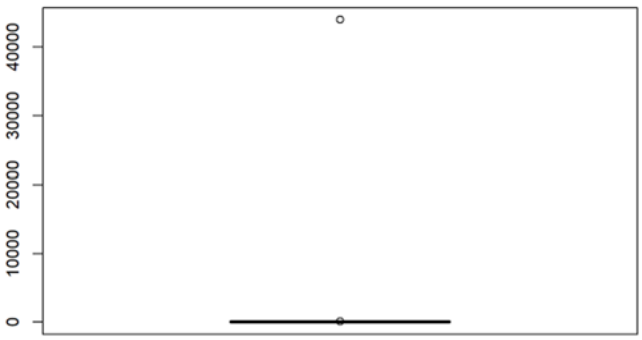

Figure 3.2.2: boxplot of 'age' feature before and after removing the outlier

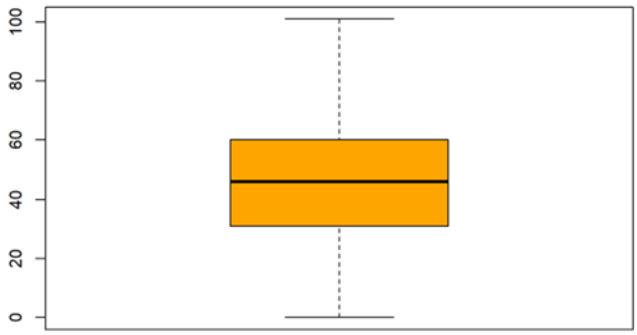

Figure 3.2.3: boxplot of 'age' feature before and after removing the outliers.

Outliers could yield misleading results and provide incorrect model predictions. The quickest and easiest way to identify outliers is by visualizing them using plots. As shown in Figure 3.2.2, the feature 'age' had outliers and after removing it, we can see a uniform range of values in Figure 3.2.3.

3.2.3 Data Aggregation

This is a crucial step since the accuracy of insights from data analysis depends heavily on the amount and quality of data used. It is important to gather high-quality accurate data and a large enough amount to create relevant results. Since I intended to build a binary classification model, I needed the 'outcome' variable from the dataset into 'Stable' or 'Critical'. The initial factor level of this variable was 13, and it was aggregated into binary classes. The data aggregation was also done for age, chronic disease, and symptoms.

3.3 Feature Selection

Feature selection is the process of choosing variables that are useful in predicting the target variable. The train.csv dataset contains 82 predictive variables, and it is considered a good practice to identify which features are important when building predictive models. There are various numbers of feature selection approaches that can be implemented in R and we used Boruta and VarImp() function from the caret package to identify the important target variables.

3.3.1 Feature Selection Algorithm: Boruta

Boruta is a feature ranking and selection algorithm based on a RandomForest algorithm. Figure 3.2.2.1 shows the selected features from Boruta. The values in green are the most important to the algorithm. The attributes closer to 0 are the least important for the accuracy of the dataset.

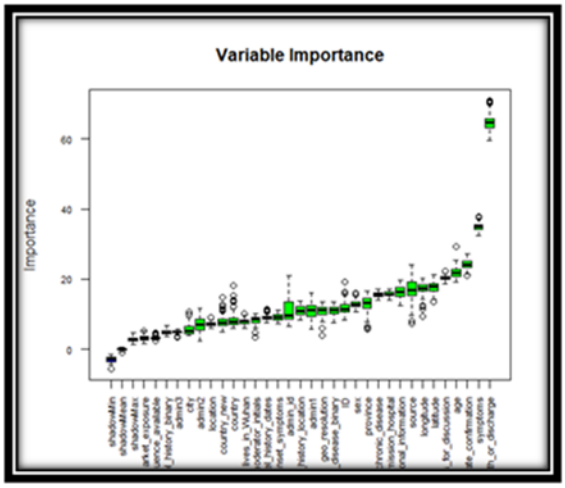

Figure 3.2.2.1: Feature selection from the Boruta method.

3.3.2 Feature selection using machine learning algorithms

Another way to look at feature selection is to consider variables most used by various machine learning algorithms the most to be important. We used the summary function from gbm package. At first, we have to train our model using gbm package and then use the summary to determine the feature importance.

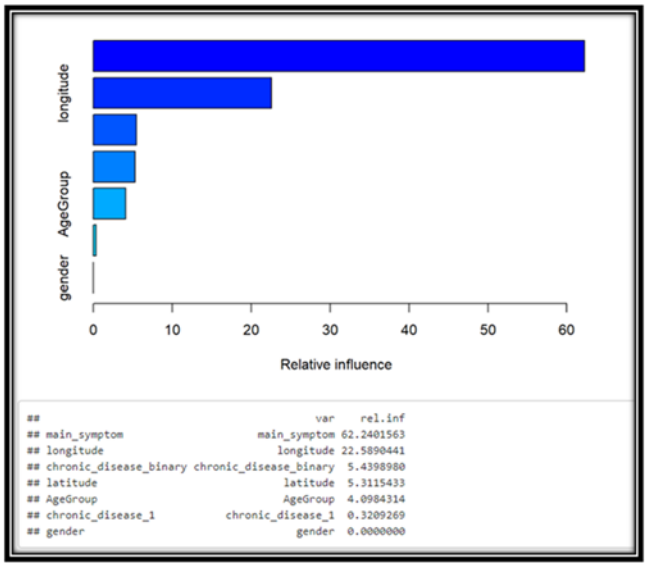

Figure 3.2.2.2: Feature selection from gbm package

3.3.3 CFS Algorithm

The CFS algorithm finds the usefulness of individual features for predicting the class label along with the level of intercorrelation among them. The selected feature subsets contain features highly correlated with predictive of the class yet uncorrelated with each other.

Table 3.3.3: Features Selected at Each Iteration Using CFS Algorithm

Iteration	Selected features
1	1) symptoms 2) date_death_or_discharge 3) notes for discussion
2	4) age
3	5) latitude 6) date_admission_hospital 7) travel_history_location 8) additional_information 9) chronic_disease_binary 10) chronic_disease 11) data_moderator_initials
4	12) sex 13) longitude 14) date_onset_symptoms 15) travel_history_dates 16) location
5	17) country 18) lives_in_Wuhan
6	19) country_new
7	20) geo_resolution 21) date_confirmation 22) reported_market_exposure 23) admin_id 24) travel_history_binary
8	25) province 26) source

3.4 Data Visualization

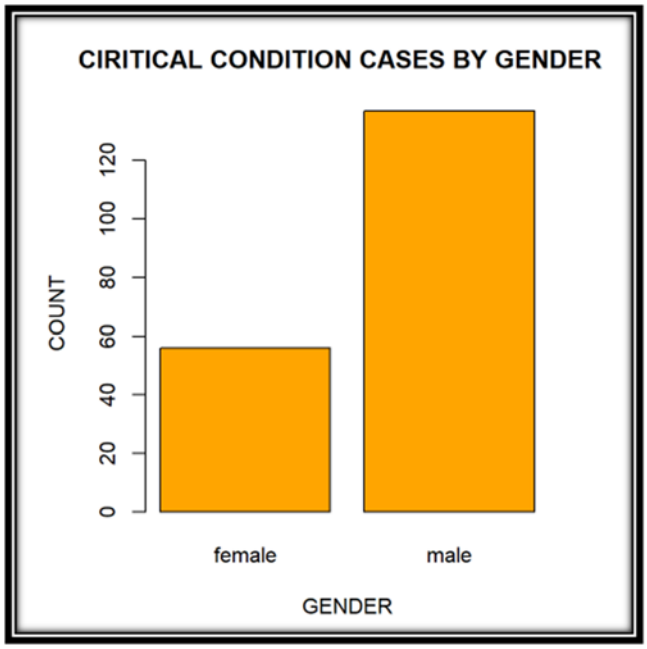

Figure 3.3.1: Critical Condition Cases by Gender

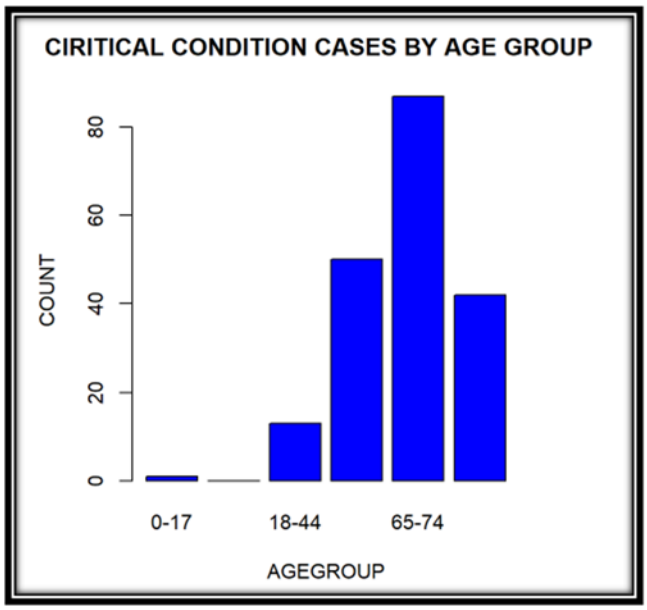

Figure 3.3.2: Critical Condition Cases by Age Group

Figure 3.3.1 and Figure 3.3.2 show some data visualization that we are conducting the aforementioned preprocessing steps. from selected attributes. From Figure 3.3.1 we could COVID infected individuals that are male and with age group between 65-74 are at higher risk of being at critical condition.

4. Modeling and Evaluation

In this section, we have discussed how predictive models have been built and evaluated. Selecting appropriate algorithms is an important decision in data mining and machine learning, and it requires knowledge of both the data set and the classification algorithm. We have selected Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting Machine modeling for this study. The data splitting was done to 80% to train data and 20% to test data.

4.1 Decision Tree

Decision Tree Classification model is one of the most common data mining models that's easy to understand. The classification uses Hunt's Algorithm where attributes are split using a recursive partitioning approach. It is implemented in RStudio using the rpart package. This algorithm is well suited while dealing with categorical attributes and binary targets.

Visualizing the results

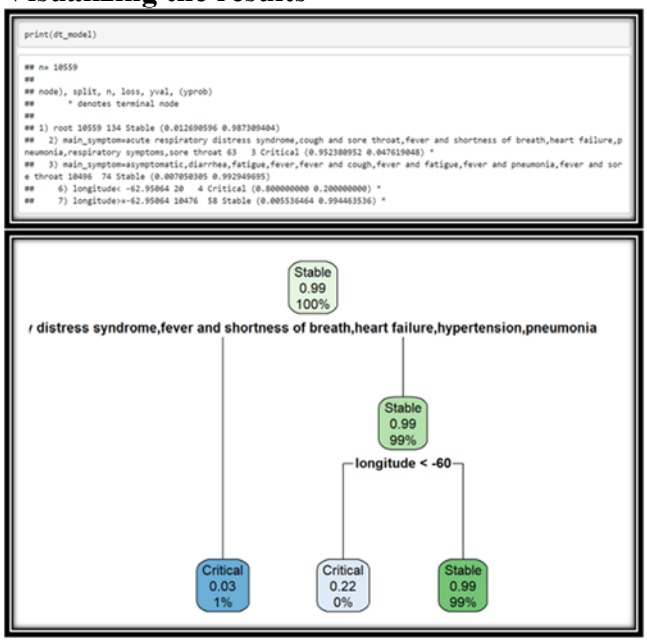

Figure 3.4.1.1: Decision Tree Model

4.2 Random Forest

Random Forest is a supervised algorithm that consists of multiple decision trees. The number of trees and the accuracy in the Random Forest algorithm have a direct relationship where the more trees there are in the model the more accurate it will be. A big difference between the Random Forest and Decision Tree algorithm is the process of finding root nodes and splitting of the attribute nodes is done randomly. Some major advantages of Random Forest are that it cannot be used for both classification and regression. The classifiers can handle missing values and can be modeled for categorical values. The Random Forest

algorithm from CRAN implements Breiman's random forest which can also be used in an unsupervised mode for assessing proximities among data points.
The attributes with over 53 levels were removed as the CRAN library for random forest cannot handle over 53 levels. The next step in preprocessing was to look for the percentage of missing values in the attributes.

Visualizing the results

```
##
## Call:
##  randomForest(formula = HealthStatus ~ ., data = traindata, ntree = 300,      mtry = 7, importance = TRUE, proximity = TR
UE)
##                Type of random forest: classification
##                      Number of trees: 300
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 0.81%
## Confusion matrix:
##          Critical Stable class.error
## Critical       79     61 0.455223881
## Stable         24  10401 0.002302158
```

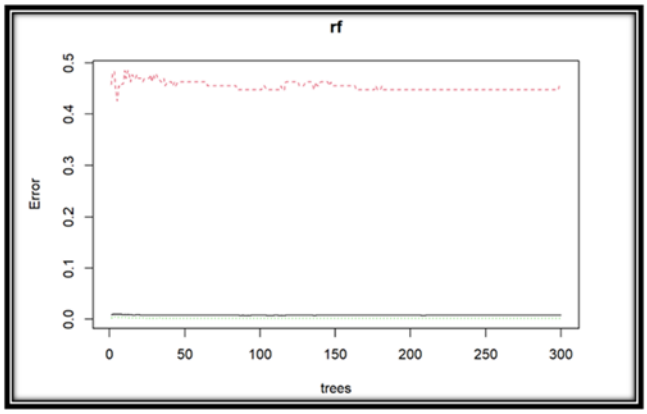

Figure 3.4.2.2: Random Forest plot.

4.3 Naive Bayes

Naive Bayes Algorithm is one of the popular classification machine learning algorithms that helps to classify the data based upon the conditional probability values computation. It implements the Bayes theorem for the computation and uses class levels represented as feature values or vectors of predictors for classification. Naive Bayes Algorithm is a fast algorithm for classification problems. This algorithm is a good fit for real-time prediction, multi-class prediction, recommendation system, text classification, and sentiment analysis use cases. Naive Bayes Algorithm can be built using Gaussian, Multinomial, and Bernoulli distribution. This algorithm is scalable and easy to implement for the large data set.

Visualizing the results

```
##
## ================================ Naive Bayes ================================
##
## - Call: naive_bayes.formula(formula = traindata$HealthStatus ~ ., data = traindata)
## - Laplace: 0
## - Classes: 2
## - Samples: 10559
## - Features: 7
## - Conditional distributions:
##     - Bernoulli: 2
##     - Categorical: 3
##     - Gaussian: 2
## - Prior probabilities:
##     - Critical: 0.0127
##     - Stable: 0.9873
##
## ------------------------------------------------------------------------------
```

4.4 Gradient Boosting Machine

Gradient boosted machines (GBMs) are an extremely popular machine learning algorithm that has proven successful across many domains and is one of the leading methods for winning Kaggle competitions. Whereas random forests build an ensemble of deep independent trees, GBMs build an ensemble of shallow and weak successive trees with each tree learning and improving on the previous. When combined, these many weak successive trees produce a powerful "committee" that is often hard to beat with other algorithms. This tutorial will cover the fundamentals of GBMs for regression problems.

Visualizing the results

```
## gbm(formula = HealthStatus ~ ., distribution = "multinomial",
##     data = traindata, n.trees = 200, n.minobsinnode = 10, shrinkage = 0.01,
##     cv.folds = 10)
## A gradient boosted model with multinomial loss function.
## 200 iterations were performed.
## The best cross-validation iteration was 200.
## There were 7 predictors of which 6 had non-zero influence.
```

5. Performance Measurement

5.1 Accuracy

Classification Model	Accuracy (%)
Decision Tree	99.7
Random Forest	99.38
Naive Bayes	97.43
Gradient Boosting Machine (GBM)	99.6

Figure 5.1 Accuracy of predictive models

As shown in Figure 5.1, we managed to obtain high accuracy on the test data for all the models. However, for

classification models not only overall accuracy can be used to measure the performance but the Receiver Operating Characteristic (ROC) curves can also be used to characterize the model performance.

5.2 ROC/AUC

A useful tool when predicting the probability of a binary outcome is the Receiver Operating Characteristic curve or ROC curve. The true positive rate is calculated as the number of true positives divided by the sum of the number of true positives and the number of false negatives. It describes how good the model is at predicting the positive class when the actual outcome is positive

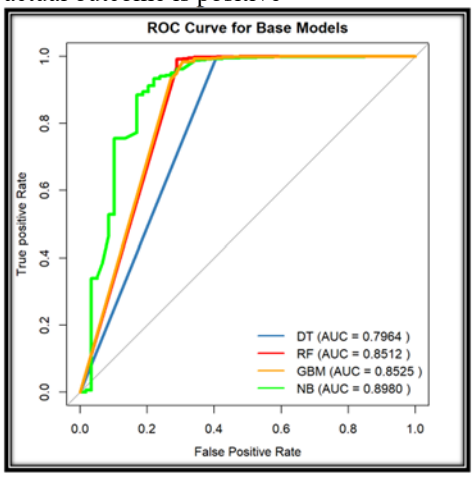

Figure 5.2: ROC plot for all base classifiers

5.3 Ensemble Modelling

Ensemble methods is a machine learning technique that combines several base models to produce one optimal predictive model. Ensembles can be more accurate when the prediction errors made by the individual predictors are (somewhat) uncorrelated. Almost always you gain accuracy improvements with ensembles, but in this study, if ensembles can improve the Area Under Curve. An ensemble can be broadly categorized into four types as data related, features related, output related, and setting related. In this paper, we have used data related and features related ensemble and the results are as shown in Figure 5.31 and 5.3.2. The data related ensemble was done using subsampling training examples and the features related ensemble was done using different features set.

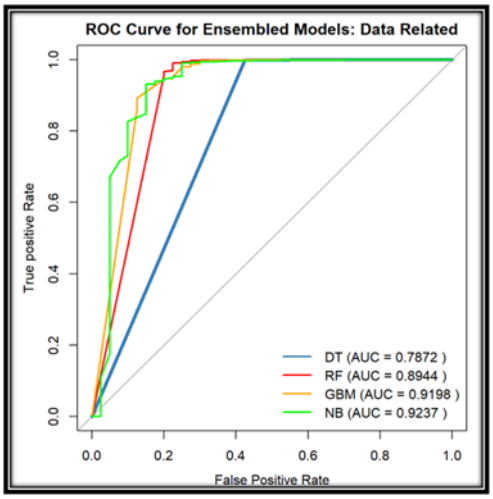

Figure 5.3.2: ROC plot for all Ensemble classifiers (Data Related)

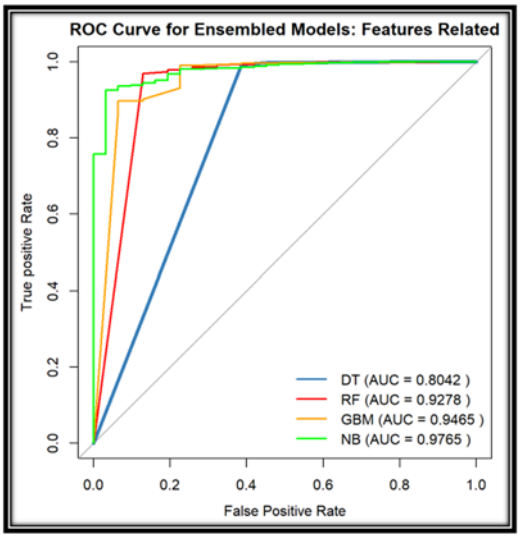

Figure 5.3.3: ROC plot for all Ensemble classifiers (Features Related)

We used the Majority Voting method to combine the ensemble results, and the accuracy maintained at the range of 98% to 99%. The figures also show that The Area Under Curve has improved. So, I have great accuracy, great ROC curves, and AUC values. Can we conclude from these results that I have the best predictive models? The answer is NO, and we will discuss it further in section 6.

A comparison of the time taken to train base models and ensemble models also included in this study. Figure 5.3.2 Ensemble Model

5.4 Model Training Time

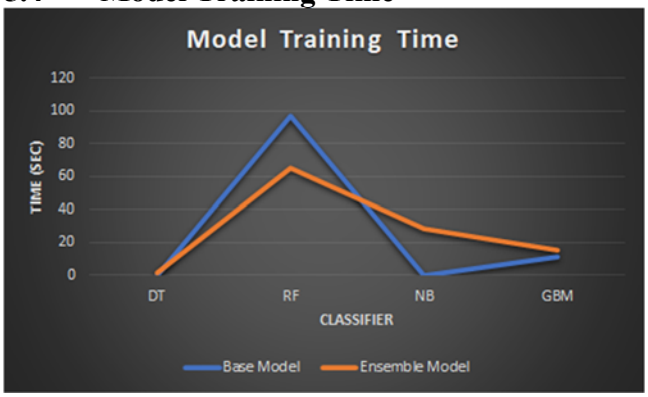

Figure 5.3.2: Model Training Time

6. What is the problem?

The problem that finally encountered was the Imbalanced Class problem. For binary classification, the accuracy measurement could be misleading when the classes are not balanced.[10]

```
##
##            Critical Stable
##   Critical       21     15
##   Stable         19   2960
```

Figure 6.1: Imbalance Classes in the dataset

According to an article about imbalance classes [16], because accuracy can be misleading in imbalance classes, we should be cautious about other metrics such as recall and precision and the kappa value.

6.1 Recall and Precision

This article [16] also described that, for a given class, the different combinations of recall and precision have a different meaning as shown in Table 6.1.1

Precision	Recall	Outcome
HIGH	HIGH	The class is perfectly handled by the model
LOW	HIGH	The model can't detect the class well, but it's highly trustable when it does
HIGH	LOW	The class is well detected but the model also includes points of other classes in it
LOW	LOW	The class is poorly handled by the model

Table 6.1.1: Precision and Recall outcome on Imbalanced Classes

	Recall	Precision
Stable	0.88889	0.9932
Critical	0.58333	0.47772

Table 6.1.1: Sensitivity and Precision Analysis on Naive Bayes Model

The analysis of recall and precision values for Naive Bayes And Gradient Boosting Machine Models that were used in this study are as shown in Table 6.1.1. and Table 6.1.2 respectively.

	Recall	Precision
Stable	0.9993	0.9927
Critical	0.4444	0.8889

Table 6.1.2: Sensitivity and Precision Analysis of Gradient Boosting Machine (GBM) Model

In conclusion, Gradient Boosting Machine perfectly handles the 'Stable' class, and the 'Critical' class is also well detected but included points from other classes. Further studies about this will be included in future work.

6.2 A Related Study for Imbalanced Classes

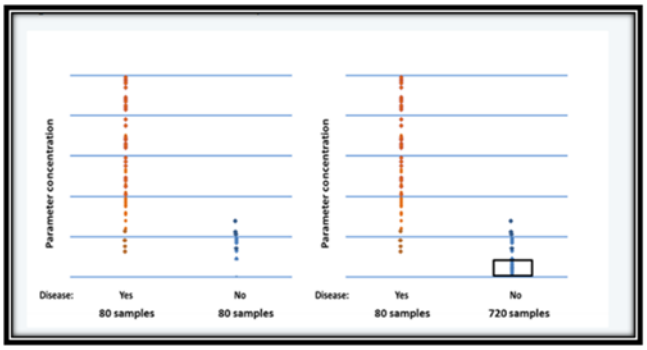

Figure 6.2.1 Imbalances Classes in a sample diagnosis test

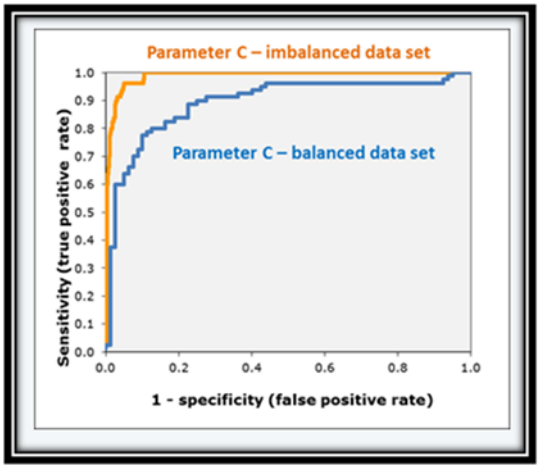

Figure 6.2.2 PRC plot for Imbalanced Classes

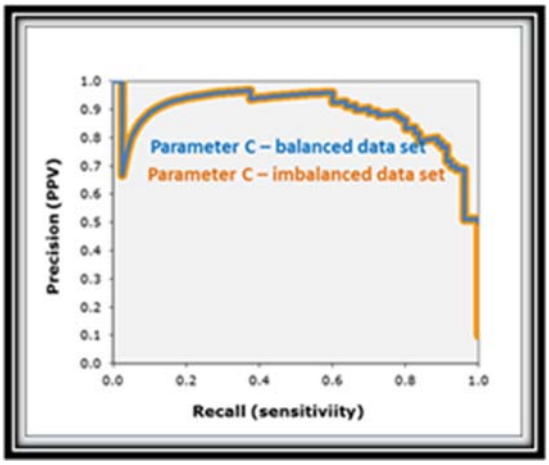

Figure 6.2.3 PRC plot for Imbalanced Classes

A related study on imbalance classes [11], was done with a sample imbalanced dataset, the distribution of the classes is as shown in Figure 6.2.1. According to this study, the imbalanced data set gives a much better ROC curve as shown in Figure 6.2.2 as compared to the balanced data set.

The study also shows that the general assumption has so far been that if you compared the area under ROC curves for two tests, you would see the real differences in model performances. However, if you look at the Precision-Recall Curves (PRC), as shown in Figure 6.2.3 the two curves are completely overlapping.

Based on their models for how to compare the performance the researchers of this study concluded that changing the main evaluation method from ROC to PRC may influence many studies. This study highly recommended using PRC as a supplement to the routinely used ROC curves to get the full picture when evaluating and comparing performances of models.

6.3 Kappa

Kappa is another metric that can be measured, especially for imbalanced classes. Kappa or Cohen's Kappa is similar in the same way as classification accuracy, except that it is normalized at the baseline of random chance on the dataset. Figure 6.3 shows that the kappa metrics for classifiers GBM and Random Forest have a higher value. Further studies about how to interpret the kappa values in terms of model performance will be added for future work.

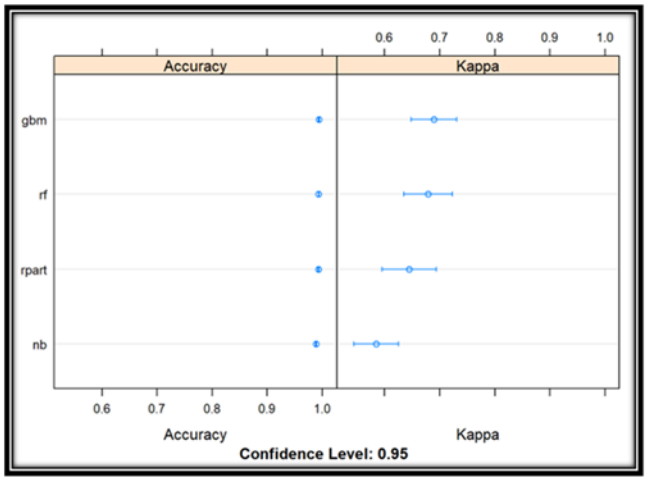

Figure 6.3: Accuracy and Kappa metrics for all classifiers

7. A Review of Related Studies

A similar study was conducted [11] to predict Coronavirus clinical severity in the experiment, the predictive power for developing Acute Respiratory Distress Syndrome (ARDS) which is also a deadly symptom that was observed from my analysis. The results of the performance of the classifier are as shown in Figure 7.1. My analysis produces high accuracy with the feature age, gender, symptoms, and history of chronic disease as the main features for predicting the severity. However, from this study, the researchers were surprised to find that patterns seen in lung images, fever, immune response or age, and gender were not contributing factors of the severity, instead, they found that levels of the liver enzyme alanine aminotransferase (ALT), reported myalgia, and hemoglobin levels were most accurately predictive of severity of the patient's condition. With these three features, the team reported being able to predict the risk of ARDS with up to 80 percent accuracy.

Table 5: Predictive algorithms accuracy

Predictive Algorithm	Accuracy
Logistic Regression	50%
KNN (k=5)	80%
Decision Tree (based on Gain Ratio)	70%
Decision Tree (based on Gini Index)	70%
Random Forests	70%
Support Vector Machine	80%

Figure 7.1 Predictive Results from Related Work

The team also concluded that the limitation of the dataset which consists of only 53 patients with some incomplete data was a clear limitation of their study, and a larger dataset would help increase the performance of their analysis.

Another study was conducted by Li Yan et al. [1] to make prediction of criticality in Covid-19 patients using clinical data in Wuhan. They used clinical features from 2,799 patients admitted in Tongji Hospital from January 10th to February 18th, 2020 to study a machine learning-based prognostic model. Among these patients, only 375 patient's data which includes 201 survivors, were used for training the model because only these patients have complete data materials. The researchers of this study built a prognostic prediction model based on XGBoost machine learning algorithm and then tested 29 patients (including 3 patients from other hospitals) who were cleared after February 19th. From this study, they found that the mean age of the 375 patients was 58.83 years old with 58.7% of males and fever was the most common initial symptom (49.9%), followed by cough (13.9%), fatigue (3.7%), and dyspnea (2.1%). The model also identified three key clinical features, as the main predictive features of Covid-19 patients, they were lactic dehydrogenase (LDH), lymphocyte and High-sensitivity C-reactive protein (hs-CRP), from a pool of more than 300 features. Similar findings were also detected from a study by Jiao Gong et.al [4], they constructed an effective model to identify cases at high risk of progression to severe COVID-19. They found that the features contributed to the severity of COVID-19 were older age, higher serum lactate dehydrogenase, C-reactive protein, coefficient of variation of red blood cell distribution width, blood urea nitrogen, and direct bilirubin and lower albumin. In this study, 372 hospitalized patients with nonsevere COVID-19 were followed for 15 days and then assigned to the severe and nonsevere groups, respectively. Based on baseline data of the 2 groups, they constructed a risk prediction nomogram for severe COVID-19 and evaluated its performance. The results show that among all cases, 72 (19.4%) patients developed severe COVID-19. The researchers generated the nomogram for identifying severe COVID-19 in the training cohort with area under the curve AUC, 0.912 sensitivity 85.7%, specificity 87.6%) and the validation cohort with area under the curve, AUC, 0.853, sensitivity 77.5% and specificity 78.4%.

However, Chi Zhang et. al. [8] moved a step further from these studies, they not only found the risk factors of severity but they built a scoring system to predict the severity of Covid-19 patients. The researchers conducted a study of 80 hospitalized patients with laboratory-confirmed COVID- 19 to establish a predictive model for disease severity. All these candidates were divided into "mild" and "severe" disease according to the clinical definitions from the National Health Commission of China. A total of 48 indicators were collected from the candidates at the initial stage of COVID-19 infection, including age, gender, pre-existing conditions (respiratory disease, cardiac disease, hypertension, hyperlipemia, diabetes, kidney disease, liver disease, post-operative, and more than two kinds of diseases), presenting symptoms (fever, cough, expectoration, vomit, and diarrhea) and laboratory detections at the initial stage of COVID-19 infection included pH, partial pressure of carbon dioxide (PCO2), partial pressure of oxygen (PO2), blood oxygen saturation (SaO2), white blood cell count (WBC), hemoglobin (HGB). The study concluded that age, pre-existing conditions (cardiac disease, hypertension, and more than two comorbidities), and 1st Laboratory detection (white blood cell count (WBC), absolute value of neutrophil (NEU), lymphocyte percentage (LYM%) , neutrophil percentage (NEU%), ratio of neutrophil to lymphocyte (NLR), fibrinogen content (FIB), c-reactive protein (CRP),total bilirubin (TBIL), albumin (ALB) , glomerular filtration rate (GRF) , creatine kinase isoenzyme-MB (CK-MB), Myoglobin, and Troponin) were identified as the predictors of the severity of disease by univariate analysis. Amongst them, age, WBC, NEU, GFR, and Myoglobin were selected by multivariate analysis as candidates of a scoring system for prediction of disease severity in COVID-19. Each variable selected by multivariate analysis was assigned diverse scores according to their hazard ratio (HR). Patients with age above 59 years old were assigned a score of 1; and the level of WBC above 6.09, the value of neutrophil above 2.89 were given a score of 2; GFR below 103.75 and myoglobin above 43 were assigned score 1. Finally, a scoring system was designed, which ranged from 0 to 7 by calculating each patient's score. Individuals with scores of 0–4 were defined to be at low risk of severity, and 5–7 at high risk.

The accuracy of prediction in severity was evaluated and found that the AUC was 0.958, sensitivity of prediction was 100%, and the specificity was 88.9%. They also claimed that the biomarkers used in the scoring system such as white blood cell count, absolute value of neutrophil, GFR and myoglobin are routine clinical detection in hospitals, which could be obtained on the first day of hospital admission. The availability of these biomarkers indicates this scoring system could be used in an outpatient setting to classify patients in high or low risk of severity and receiving different therapy strategies. These researchers claimed that they are the first to create a score-based severity prediction model, but interestingly, another score-based risk model not only has been developed but has been already adopted by the

Israeli ministry of health as its risk classification tool for COVID-19 lab tests prioritization and for targeting its instructions on risk management during the lockdown exit strategy. This model was studied by Noa Dagan et.al.[10] to create a score-based risk classification tool to predict the severity in Covid-19 patients. This tool combines 10 risk factors to predict the risk of patients for severe COVID-19 illness to be at basic risk, high risk and very high risk. This was a retrospective cohort study based on the data of Clalit Health Services (CHS) and the Israeli Ministry of Health (MOH). The variables chosen to be included in the model were cardiovascular disease or congestive heart failure, diabetes mellitus, overweight (BMI $\geq$ 30), active or recent malignancy, immunosuppression, chronic obstructive pulmonary disease or over 10 smoking pack-years, chronic hepatic, renal or neurological disease and hospital admissions in the last 3 years (with the exception of those for normal delivery). From the analysis of the data, age was found to be the main driver of risk of severe disease, and age groups were thus addressed as an interaction variable in this model, sub-classified by the total number of risk points accumulated. The model performed well by correctly identifying 92% of patients who will experience a severe COVID-19 infection or death as having an elevated risk, while only classifying 18% of the total population as such. The researchers found that this is in contrast to the CDC's list of risk criteria, which classify nearly two times as many patients (35%) at elevated risk, for a small 4% gain in sensitivity (96%).

Another study to predict the severity using clinical data was done by Weifeng Shang et. al. [6] by studying clinical data of 443 patients with COVID-19 admitted and divided them into nonsevere group and severe group according to their condition. Among the data they used were leukocyte, neutrophils, lymphocytes, neutrophil-to-lymphocyte ratio (NLR), hemoglobin, platelets, D-dimer, erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), procalcitonin, lactate dehydrogenase (LDH), uric acid, creatinine, albumin, chest computed tomographic (CT) scans. The analysis shows that NLR in the severe group was significantly higher than that in the nonsevere group where the area under the ROC curve of the NLR predicting the severity of COVID-19 was the largest. The optimal working point was 4.283, and the sensitivity and specificity to predict the severity of COVID-19 were 56.3% and 83.7%, respectively. The researchers also found that platelet count was lower in the severe group than in the nonsevere group. Binary logistic regression analysis showed platelet count was an independent protective factor for severe COVID-19. The study concluded that NLR, CRP, and platelets can effectively predict the severity of COVID-19, and among these predictors, NLR is the best predictor of severe COVID-19. This study has several limitations, the study was a single center retrospective study, which may affect the generalization of the results due to the limitation of enrolled patients. Second, some patients are still hospitalized among the 443 cases. Therefore, it is difficult to assess risk factors for poor outcomes. Third, the sample size was relatively small, which may have some impact on the statistical results.

A study done by Celestine Iwendi et.al. [2] proposes a fine-tuned Random Forest model boosted by the AdaBoost algorithm to predict COVID-19 patient's health. They used the dataset from Kaggle and they used COVID-19 patient's geographical, travel, health, and demographic data to predict the severity of the case to be either recovery or death. The data analysis of this study shows a positive correlation between patients' gender and deaths, and also indicates that the majority of patients are aged between 20 and 70 years. The study also found that fever, cough, cold, fatigue, body pain, and malaise were the most common symptoms of Covid-19 patients. The models included in this study include Decision Tree Classifier, Support Vector Classifier, Gaussian Naïve Bayes Classifier, and Boosted Random Forest Classifier. The researchers of this study used evaluation metrics such as Accuracy, Precision, Recall, F1 Score and stated that since the dataset of their study can be an imbalanced dataset, they will use F1 Score as the primary metric for comparison of the models. The study concluded that Boosted Random Forest performs better while predicting COVID-19 patient deaths. After Hyperparameter tuning he model has an accuracy of 94% and a F1 Score of 0.86. Even Though this study has highlighted that gender is an contributing feature of severity in patients, a study done by Char Leung [5], on analyzing the risk factors for predicting mortality in elderly patients with COVID-19, stated that gender did not appear to be a mortality risk factor to predict the severity. However, this study shows that age played an important role as a risk factor of severity. In this study a total of 154 individual cases in 26 provinces, including 89 deceased patients and 65 surviving patients, were identified from 86 sources originating from the official web pages of Chinese health authorities and the media under the supervision of the Chinese government such as Xinhua, Sina and the Paper. The data collected includes gender, age, travel history to Hubei, time from symptom onset to admission, time from admission to discharge/death, symptoms on admission and comorbidities. Statistical tests on the difference in measures on the data between the deceased and surviving patient groups were performed and a logistic regression model was estimated to identify risk factors for mortality with the stepwise regression procedure for independent variable selection. Age, fever and diarrhea were selected by the stepwise regression procedure in the logistic regression model and they found that age and fever were mortality risk factors. However, they found that for older patients, fever was less likely to occur in deceased patients. The most commonly observed comorbidities in deceased patients were hypertension (53.2 %), cardiovascular and cerebrovascular disease (42.0 %), and diabetes (37.8 %). The major limitation of the present work is that only baseline characteristics were considered as mortality risk factors. While this study was focused only on

elderly patients, Sally H. Adams et.al.[9] conducted a study to find the medical vulnerability of young adults to severe COVID-19 illness. The objective of this study was to determine the overall medical vulnerability, to determine medical vulnerability for nonsmokers and to determine individual vulnerability indicators. Logistic regressions were conducted to compare subgroup differences (sex, race/ethnicity, income, and insurance status). This study used a young adult subsample aged 18 - 25 years which consists a total of 8,405 participants. Some important findings from this study was that nearly one in three young adults are medically vulnerable to severe COVID-19 illness (32%). However, in the nonsmoking young adult group, only about one in six is medically vulnerable to severe COVID-19 illness (16%). This difference was caused by the large portion of young adults who reported that they engaged in smoking. The findings from the analysis by sex show that overall severe COVID-19 illness medical vulnerability was higher for men than women. Higher rates of engagement in smoking (cigarettes and/or cigars) and e-cigarette use in men than women. Among nonsmokers, by contrast, females were significantly more likely vulnerable to severe illness. Analysis of differences by race/ethnicity generally showed higher vulnerability to severe COVID-19 illness for the white subgroup compared with the black, Hispanic, and Asian subgroups. The researchers concluded that the most prevalent factor conferring medical vulnerability to severe COVID-19 illness among young adults is smoking. Notably, the risk of being medically vulnerable is halved when smokers, including e-cigarette users, are removed from the sample.

Patrick Schwab et. al [3] conducted a study about predictive models to identify three types of Covid-19 Victims; who will be tested positive, who will be hospitalized and who will be admitted to the Intensive Care Unit. They used clinical, demographic and blood analysis data for this study to build the model and also studied which features were most important for each clinical prediction task. The researchers of this study used the data from a cohort of 5644 patients seen at the Hospital Israelita Albert Einstein in São Paulo, Brazil in the early months of 2020. During the collection of data, the rate of SARS-CoV-2 positive patients at the hospital was around 10% of which around 6.5% and 2.5% required hospitalization and critical care.
The models of this study were built using Logistic Regression (LR), Neural Network (NN), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (XGB). The

researchers found that different models provide best results on the three different categories of the prediction. To predict who will be tested positive, the XGB model performed well with a sensitivity of 75% and a specificity of 49%. The researchers found that 71.7% of the importance for the best XGB model for predicting SARS-CoV-2 test results was assigned to the missing indicator corresponding to the Arterial Lactic Acid measurement which shows much of the marginal predictive performance gain of the XGB model was attributed to whether or not the Arterial Lactic Acid test had been ordered. Beyond Arterial Lactic Acid being missing, age, leukocyte count, platelet count, and creatinine were implied to be associated with a positive SARS-CoV-2 test result by the best encountered predictive model, which further substantiates recent independent reports of those factors being potentially associated SARS-CoV-2. The RandomForest model performed well on predicting hospitalization of COVID-19 patients with 0.92 AUC and SVM model outperformed other models on predicting which patients require critical care with 0.98 AUC. However, this study has two limitations; the data of this study collected from a single study site and also limitation to access to mortality data for the analyzed cohort, and we were therefore not able to correlate our predicted individual risk scores with patient mortality.

Todd J. Levy et. al. [7] developed and validated a clinical tool to predict 7-day survival in patients hospitalized with COVID-19. They used data from the electronic health record (EHR), forgoing symptom-related records and radiology reads for developing this tool. The researchers of this study claimed that, even some clinical prediction tools have been established to estimate survival in patients with pneumonia or hospitalized with severe illness, including the Sequential Organ Failure Assessment (SOFA) and CURB-65 Scores, but these tools have not been validated in patients with COVID-19. This study involved patients admitted to 11 of 12 acute care facilities in the Northwell Health system between March 1 and April 23, 2020. Among the collected data for this study were patient demographic information, comorbidities, laboratory values, and outcomes (i.e., death, length of stay, discharge). The model was developed by analyzing 42 potential predictors for the patients, and Least Absolute Shrinkage and Selection Operator (LASSO) regression was used to identify predictors that, when linearly combined, predict the survival of hospitalized patients with COVID-19. This study is the first to develop a model, the NOCOS Calculator, that predicts survival of patients hospitalized with COVID-19 in the United States. We created and validated the Northwell COVID-19 Survival (NOCOS) Calculator with data on almost 14000 patients, using only 6 clinical data points typically available to clinicians within the first 60 minutes of patient presentation. The 6 predictors were Serum blood urea nitrogen, age, absolute neutrophil count, red cell distribution width, oxygen saturation, and serum sodium. All these data points are available as discrete inputs in most commercial EHRs, supporting that this calculator could be readily incorporated into tools to support clinical decisions.

8. CONCLUSION

This study shows that the accuracy metric is not the only performance metric of a predictive model, instead, other important metrics such as ROC and PRC and kappa values should be taken into account, especially for imbalanced datasets. It was also observed that the accuracy of the model is depending on data-preprocessing, feature selection, classification algorithm, and the tuning parameters of the model.

The present study used demographic data, health history and symptoms of an infected individual to predict the severity and found that age, gender, symptoms, and history of chronic disease to be the best predictive variable. However, the review of related studies shows that the attention for predicting the severity of COVID-19 patients is more focused with the usage of clinical data. The issues of these studies have been always around limitation of the data.

9. ACKNOWLEDGMENTS

I would like to thank Professor. Alsmadi for giving the opportunity to work on this project. It has given us great exposure to machine learning topics and techniques.

10. REFERENCES

[1] Li Yan, M.D., Hai-Tao Zhang, Ph.D., Yang Xiao, Ph.D., Maolin Wang, Yuqi Guo, Chuan Sun, Xiuchuan Tang, Liang Jing, Shusheng Li, Mingyang Zhang, Ying Xiao, Haosen Cao, Yanyan Chen, Tongxin Ren, Junyang Jin, Ph.D., Fang Wang, Yanru Xiao, Sufang Huang, Xi Tan, Niannian Huang, Bo Jiao, Yong Zhang, Ph.D., Ailin Luo,
M.D., Zhiguo Cao, Ph.D., Hui Xu, M.D., and Ye Yuan, Ph.D.
Prediction of criticality in patients with severe Covid-19 infection using three
clinical features: a machine learning-based prognostic model with clinical data in
Wuhan
https://www.medrxiv.org/content/10.1101/2020.02.27.20028027v2.full.pdf

[2]Celestine Iwendi, Ali Kashif Bashir, Atharva Peshkar, R. Sujatha,
Jyotir Moy Chatterjee, Swetha Pasupuleti, Rishita Mishra, Sofia Pillai, and Oyun Jo
COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm
https://www.frontiersin.org/articles/10.3389/fpubh.2020.00357/full

[3]Patrick Schwab, August DuMont Schütte, Benedikt Dietz, and Stefan Bauer,predCOVID-19: A Systematic Study of Clinical Predictive Models for Coronavirus Disease 2019
https://arxiv.org/abs/2005.08302

[4] Jiao Gong, Jingyi Ou, Xueping Qiu, Yusheng Jie, Yaqiong Chen, Lianxiong Yuan, Jing Cao, Mingkai Tan, Wenxiong Xu, Fang Zheng, Yaling Shi, and Bo Hu, A Tool for Early Prediction of Severe Coronavirus Disease 2019 (COVID-19): A Multicenter Study Using the Risk Nomogram in Wuhan and Guangdong, China
https://academic.oup.com/cid/article/doi/10.1093/cid/ciaa443/5820684

[5] Char Leung, Risk factors for predicting mortality in elderly patients with COVID-19: A
review of clinical data in China
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7184979/

[6] Shang W, Dong J, Ren Y, et al. The value of clinical parameters in predicting the severity of COVID-19. J Med Virol. 2020;1–5. https://doi.org/10.1002/jmv.26031
https://onlinelibrary.wiley.com/doi/epdf/10.1002/jmv.26031

[7] Todd J. Levy, MS; Safiya Richardson, MD, MPH; Kevin Coppad, BS; Douglas P.
Barnaby, MD, MSc; Thomas McGinnbc, MD, MPH; Lance B. Beckerab, MD; Karina
W. Davidson, Ph.D., MASc; Stuart L. Cohenbc, MD; Jamie S. Hirsch, MD, MA,
MSB; Theodoros P. Zanos, PhD
Development and Validation of a Survival Calculator for Hospitalized Patients
with COVID-19
https://www.medrxiv.org/content/10.1101/2020.04.22.20075416v3

[8] Chi Zhang, Ling Qin, Kang Li, Qi Wang, Yan Zhao, Bin Xu, Lianchun Liang,
Yanchao Dai, Yingmei Feng, Jianping Sun, Xuemei Li, Zhongjie Hu, Haiping Xiang,
Tao Dong, Ronghua Jin1 and Yonghong Zhang, A Novel Scoring System for Prediction of Disease Severity in COVID-19
https://www.frontiersin.org/articles/10.3389/fcimb.2020.00318/full

[9] Sally H. Adams, Ph.D., M. Jane Park, M.P.H., Jason P. Schaub, M.P.H.,
Claire D. Brindis, Dr.P.H., and Charles E. Irwin Jr., M.D., Medical Vulnerability of Young Adults to Severe COVID-19 Illness- Data From the National Health Interview Survey
https://www.jahonline.org/action/showPdf?pii=S1054-139X%2820%2930338-4

[10] Noa Dagan, Noam Barda, Dan Riesel, Itamar Grotto, Siegal Sadetzki, Ran Balicer, A score-based risk model for predicting severe COVID-19 infection as a key component of lockdown exit strategy
https://www.medrxiv.org/content/10.1101/2020.05.20.20108571v1

[11] Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity,Xiangao Jiang , Megan Coffee , Anasse Bari , Junzhang Wang , Xinyue Jiang , Jianping Huang , Jichan Shi1 , Jianyi Dai1 , Jing Cai , Tianxiao Zhang, Zhengxing Wu1 , Guiqing He1 and Yitong Huang,https://www.techscience.com/cmc/v63n1/38464

[12] https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset

[13]https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

[14]https://docs.google.com/spreadsheets/d/1itaohdPiAeniCXNlntNztZ_oRvjh0HsGuJXUJWET008/edit#gid=0/

[15] https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30243-7/fulltext#seccestitle220

[16] https://www.worldometers.info/coronavirus/coronavirus-age-sex-demographics/

[17] https://blog.revolutionanalytics.com/2016/05/using-caret-to-compare-models.html

[18] https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

[19] https://www.blopig.com/blog/2017/04/a-very-basic-introduction-to-random-forests-using-r/

[20] https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/

[21]https://machinelearningmastery.com/?s=ensembling+in+r&post_type=post&submit=Search

[22] Precision-recall curves- what are they and how are they used? https://acutecaretesting.org/en/articles/precision-recall-curves-what-are-they-and-how-are-they-used

[23] http://www.edii.uclm.es/~useR-2013//Tutorials/kuhn/user_caret_2up.pdf

[24] https://machinelearningmastery.com/machine-learning-evaluation-metrics-in-r/#:~:text=Accuracy%20and%20Kappa,class%20classification%20datasets%20in%20caret.&text=Learn%20more%20about%20Accuracy%20here,random%20chance%20on%20your%20dataset.

[25] http://www.edii.uclm.es/~useR-2013//Tutorials/kuhn/user_caret_2up.pdf

[26] https://towardsdatascience.com/handling-imbalanced-datasets-in-machine-learning-7a0e84220f28

Appendix A:

ID - Unique identifier for reported case. Currently ID is run concurrently for cases reported from Hubei, China and cases reported outside of Hubei, China. ID order does not necessarily reflect epidemiological progression, or reporting date, and should not be used to order cases in temporal progression.

age - Age of the case reported in years. When not reported, N/A is used. Age ranges are recorded as start_ age-end_age e.g. 50–59.

sex - Sex of the case. When not reported, N/A is used.

city – Initial generic geographic metadata is reported here. Subsequently standardized via lookup with a geographic reference table.

province – Initial entry of name of the first administrative division in which the case is reported. Subsequently standardized via lookup with a geographic reference table.

country - Name of country in which the case is reported. Note that imported cases will be assigned to the country in which confirmation occurred - this is typically in the arrival country, rather than the site of infection.

Travel_ history_location will describe other locations of travel for such instances.

wuhan(0)_not_wuhan(1) - Binary fag to distinguish cases from Wuhan, Hubei, China, from all other cases. 0 denotes a case is reported in Wuhan, 1 denotes a case reported elsewhere in the world.

latitude - The latitude of the specific location (denoted as point in geo_resolution) where the case was reported, or the latitude of a representative location (denoted as admin in geo_resolution) within the administrative unit the case is reported.

longitude - The longitude of the specific location (denoted as point in geo_resolution) where the case was reported, or the longitude of a representative location (denoted as admin in geo_resolution) within the administrative unit the case is reported.

geo_resolution - An indicative field in which the spatial representativeness of latitude and longitude are described. point indicates that a specific location is being represented by these coordinates.

admin denotes that the coordinates are representative of the administrative unit in which coordinates lie.

Subsequent **admin3 , admin2 , admin** 1 and corresponding admin_id and shapefile will allow for a more specific representation to be had.

date_onset_symptoms - Date when the reported case was recorded to have become symptomatic. Specific dates are reported as DD.MM.YYYY. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start or finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

date_admission_hospital - Date when the reported case was recorded to have been hospitalized. Specific dates are reported as DD.MM.YYYY. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start or finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

date_confirmation - Date when the reported case was confirmed as having COVID-19 using rt-PCR. Confirmation accuracy is contingent on the data source used. Specific dates are reported as DD.MM.YYYY. Scientific Data | (2020) 7:106 | https://doi.org/10.1038/s41597-020-0448-0 4 www.nature.com/scientificdata/ www.nature.com/scientificdata Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM.YYYY - respectively.

symptoms - List of symptoms recorded in the description of the case.

lives_in_Wuhan - Recorded relationship of patient with city of Wuhan, Hubei, China. yes indicates that the case was a resident of Wuhan. no indicates that the case is not a resident of Wuhan (residential). No information indicates that no data was available.

travel_history_dates - Recorded travel dates to and from Wuhan. Specific dates are reported as DD.MM.YYYY and indicate the date when the individual lef Wuhan. Ranges are recorded as DD.MM.YYYY - DD.MM.YYYY when both are available. Ranges with uncertain start of finish dates are recorded as - DD.MM.YYYY and DD.MM. YYYY - respectively.

travel_history_location - An open field describing the recent recorded travel history of the case.

reported_market_exposure - An open field indicating yes if there was reported market exposure and no if there was not. N/A indicates that no information is provided.

additional_information - Any additional information that may be informative about the case, such as the occupation of the patient, the purpose of their travels, the hospital they were admitted to, etc.

chronic_disease_binary - 0 represents a case that was reported to have no chronic disease and 1

represents cases that reported a chronic disease

chronic_disease - Reported chronic condition(s) of the reported case. source - URL identifying the source of this information

sequence_available - If there was a genomic sequence available the accession number is inserted here.

outcome - Patients outcome, as either died or discharged from hospital.

date_death_or_discharge - Reported date of death or discharge in DD.MM.YYYY format.

location – Location of the reported case.

admin3 – Administrative unit level 3 (e.g., zip code) of where the case was reported.

admin2 – Administrative unit level 2 (e.g., county) of where the case was reported.

admin1 – Administrative unit level 1 (e.g., province) of where the case was reported.

country_new – Administrative unit level 0 (e.g., country) of where the case was reported.

admin_id – Administrative unit ID of the lowest level available for the case reported.