# Exploring COVID-19 Cases

**An analysis of the progression of COVID-19 viral infection around the world**

**Riley Kendall -** Introduction, Data Visualization, Model Selection, Future Plans
**Duc Le -** Topic Selection, Data Preprocessing, Random Forest Model Construction
**Matthew Nwermn -** Logistic Regression Model Construction and Code Compilation

**Data Mining**
**May 24, 2020**

## Introduction

Our group worked on a dataset called “NCoV2019” found on github.[1] This dataset contains information about the health statuses and demographics of Covid-19 patients around the world. Specific contributors to this dataset include “individuals and organizations” that the publisher describes as credible. We assume that doctors and health-based organizations specifically contributed, as some of the information provided in the dataset would only be known by medical personnel. Prior to cleaning and preprocessing, the dataset is fairly large, with 476,126 samples and 23 features. Our group narrowed down our interest to 5 features in particular: age, sex, country, chronic disease status, and patient outcome. We chose these features due to their relevance in the media and in safety guidelines for Covid-19. The independent variables include age, sex, country, and chronic disease status. Age is a continuous variable and sex is a binary variable where “0” indicates male and “1” indicates female. The country variables were assigned arbitrary values representing four particular countries of interest: China, Italy, the United States, and Vietnam. Chronic disease status is a binary variable where “0” indicates that the patient does not have a chronic disease and “1” indicates that the patient does. The dependent variable in this dataset is the outcome of the patient. Outcome is a binary variable where “0” indicates that the patient has not recovered and “1” indicates that a patient has recovered. “Not recovered” means that a patient had either passed away or was in critical condition by the time data was collected, while

“recovered” means that a patient was released from the hospital/quarantine or was in stable condition by the time data was collected.

## Problem

Given our dataset, we aim to predict the outcome of a patient diagnosed with COVID-19. We approached this problem by applying two different classification machine learning algorithms: Decision Trees (Random Forest Classifier) and Logistic Regression. With the given features, we want to predict whether the patient will or will not successfully recover. As stated previously, we consider a successfully recovered patient as one who is completely discharged from a hospital, released from quarantine or in stable condition. On the other hand, non-recovered patients are ones who die or are in critical/severe condition. Some assumptions were made for the classification model. An intended assumption was that the outcome associated with an age range regresses towards the mean of that age range. We assumed this so that we could work around some cases where patient age was provided as a range. An implicit assumption was that the data set is correctly and reliably recorded by credible sources. In addition, we assumed that critical condition suggests that a patient will not recover. With common knowledge and provided facts, we know that age, sex and underlying chronic health issues play a significant role in the outcome of patients infected with COVID-19. In addition, the data set was not completed for all the features thus we had to decide on a features vs. samples tradeoff. Had we included more features into our model, we would have had to sacrifice a large amount of samples due to the incompleteness of the data

set. Lastly, by using these features mentioned along with the countries of these cases, we seek to discover a pattern that accurately classifies the recovered from non-recovered patients.

## Data Preprocessing

The following were problems found within the original data set:

- Missing values.
- Qualitative, categorical variables.
- Features recorded in ranges. Ex. Age: 19-44.
- Extra/unnecessary features. Ex: Additional comments/notes on the patients.
- Unscaled data.
- High dimensionality.

Preprocessing of the dataset was completed by following the steps listed in Table 1:

**Table 1)** Data Preprocessing Functions and Steps

| Function | Method | Reason |
|---|---|---|
| Removal of NaN's | df.dropna(subset = [column_names]) | Removing missing values from the dataset prevents them from potentially skewing our results. Dropping missing values is more beneficial than filling them in with mean/median because assumed values could affect the results of Random Forest Classifier as the algorithm is dependent on selecting the best splitting value. |
| Conversion of binary qualitative variables | Replaced categorical values with numerical ones. (ex. True = 1. False = 0). | Logistic Regression does not accept qualitative variables. |

| | | |
|---|---|---|
| Conversion of age ranges | Replaced values with age range with the mean. Ex: age = 20-29-> mean[20,29] | Done under the assumption mentioned previously in order to keep as many data samples as possible. |
| Removal of extra features | Selected a subset of the original dataset: Train = data[desired_features]. | Columns containing notes and comments are not necessary for our model. |
| Appropriate scaling of the data | Normalized the dataset. | The data set was normalized for other attempted models. |

## Data Visualization Post Preprocessing

A visualization of the data in terms of sex, chronic disease, and age is shown in Figure 1 below. Country was not included because its assigned values are arbitrary. Blue points symbolize non-recovered patients and orange symbolize recovered patients.

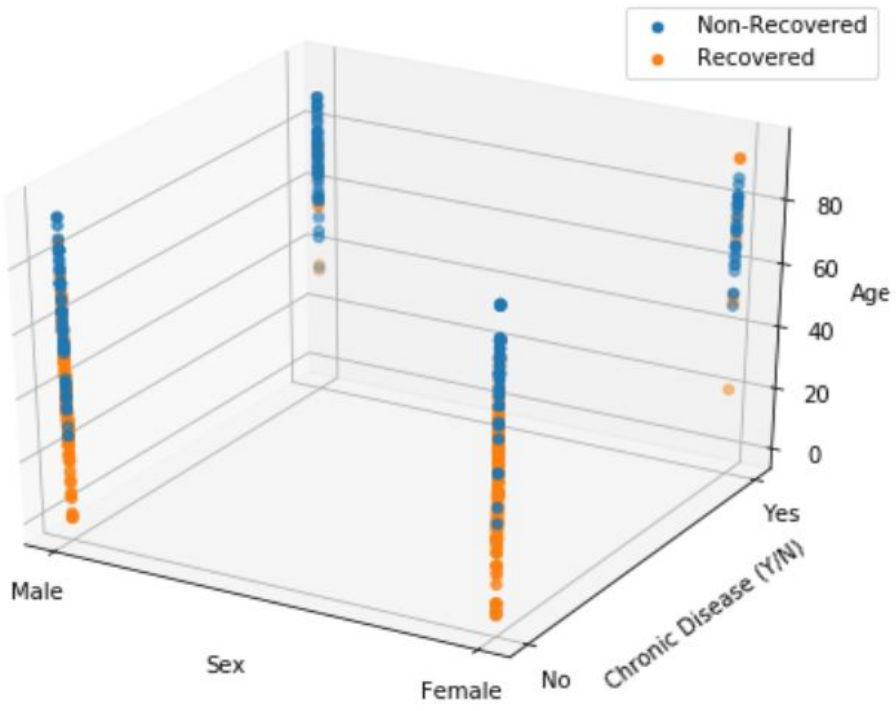


*Figure 1. Initial Visualization of Dataset*

Before designing solutions to predict a patient's outcome, it was important that we first visualized the features of the dataset in order to understand their unique relations to the outcome. We plotted each feature against outcome and checked for unexpected or surprising relations.

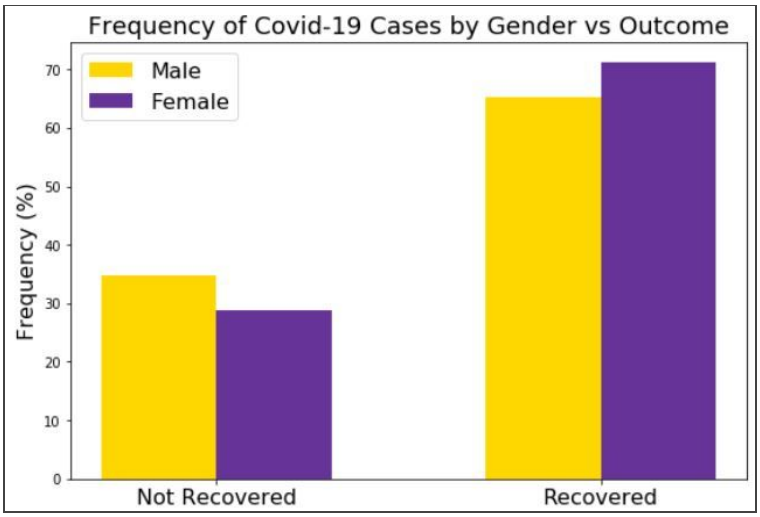


*Figure 2. Gender vs Outcome plot*

Figure 2 demonstrates the plot results for Gender vs Outcome. The results show that more females than males recovered, and more males than females did not recover. However, the difference in the results between the two genders is not significant: there is only ~5% frequency difference. What we can extract from this is that sex does not appear to have a significant effect on the recoverability of a patient.

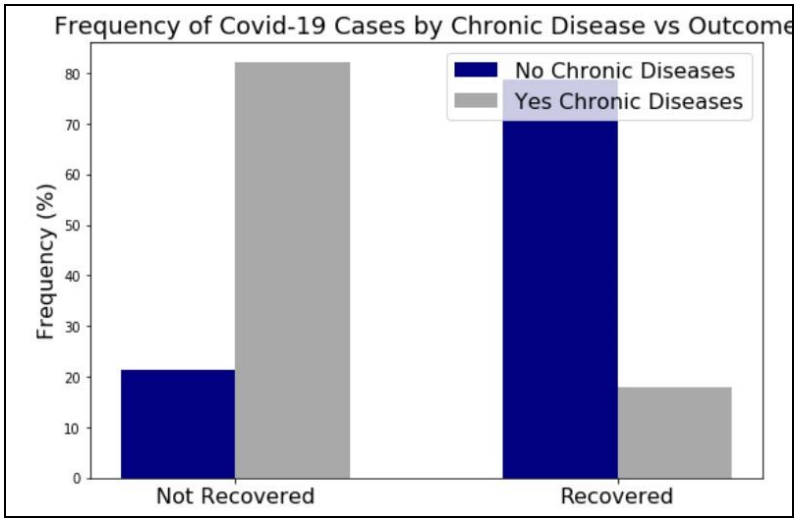


*Figure 3. Chronic Disease vs Outcome plot*

Figure 3 demonstrates the plot results for Chronic Disease vs Outcome. The results show that far more patients without chronic disease recovered than patients with chronic disease, and far more patients with chronic disease

did not recover than patients without. In particular, the differences in the frequency results for patients with vs without chronic disease that recovered is ~20%, and for patients that did not recover, ~80%. What we can extract from this is that chronic disease does appear to have a significant effect on the recoverability of a patient.

Figure 4 demonstrates the plot results for Country vs Outcome. Most strikingly, the results show that in Vietnam, all patients made full recoveries and no patients did not recover. Whether these results are due to the measures taken by Vietnam to reduce Covid-19 spread, high quality medical care in Vietnam, chance, or a lack of records for patients that did not recover in Vietnam is unknown. The results also show that China has a high ratio of recovered patients to not recovered patients, Italy has a low ratio of recovered patients to not recovered patients, and the United States has the lowest ratio of recovered patients to not recovered patients according to this dataset. What we can extract from this is that residency in Vietnam or the United States are likely key predictors in determining whether a patient will recover.

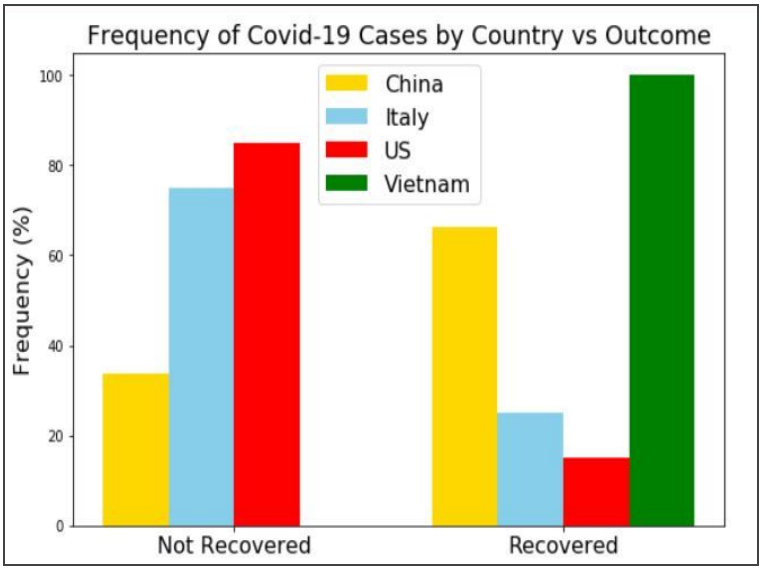


*Figure 4. Country vs Outcome plot*

Figure 5 demonstrates the plot results for Age vs Outcome. Most obviously, the results show that the mean age of recovered patients is ~40 years of age and the mean age of

patients who did not recover is ~70 years of age for the whole dataset. This suggests that younger ages recover more frequently than older ages. The wings associated with the box plots span a larger age range for recovered patients than not recovered patients. This tells us that, even if a patient is older, they are still able to recover. The recovered box plot demonstrates one outlier at ~100 years of age, suggesting that extremely old ages infrequently recover. What we can extract from all of this is that age does appear to have a significant effect on the recoverability of a patient.

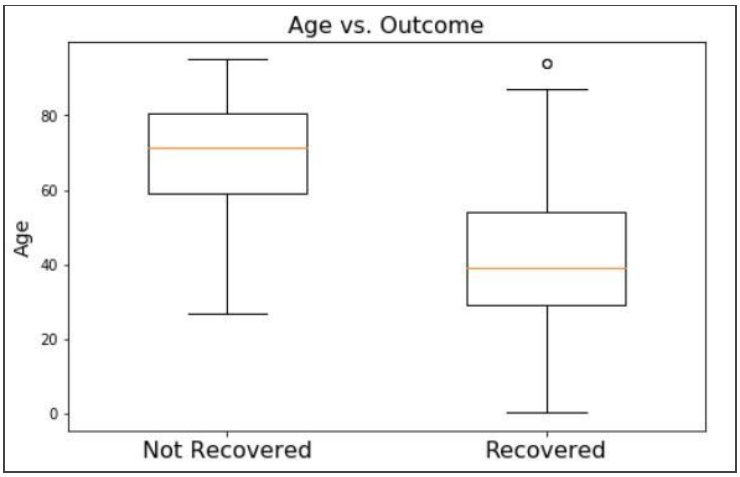


*Figure 5. Age vs Outcome plot*

We also visualized the features of the dataset against each other in order to see if any relations existed between the different features and if there were any unexpected or surprising relations. Four plots are demonstrated in Figures 6-8 representing what we found to be the most interesting feature vs feature relationships.

Figure 6 demonstrates the plot results for Gender vs Chronic Disease. The results show that the male patients more frequently had a chronic disease and that the female patients more frequently did not have chronic disease. This may explain the results that we found in Figure 2 when analyzing Gender vs Outcome. The slightly higher frequency of males that did not recover than females may be more of a reflection on

chronic disease's effect on outcome than gender's, as more males had a chronic disease and Figure 3 suggests that chronic disease plays a significant role in a patient's outcome. This theory further supports that sex does not significantly affect a patient's outcome.

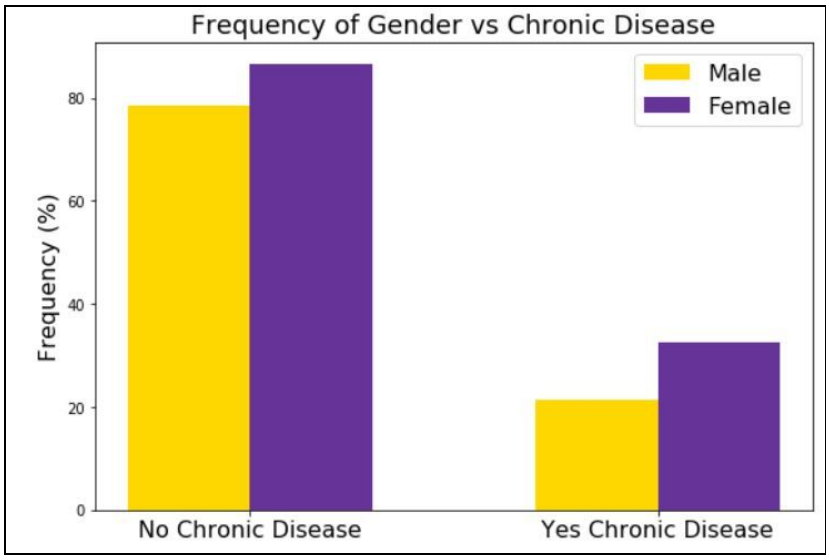


*Figure 6. Gender vs Chronic Disease plot*

Figure 7 demonstrates the plot results for Age vs Chronic Disease. The results show that the mean age of patients with chronic disease is ~70 years of age and the mean age of patients without chronic disease is ~45 years of age. This suggests that older ages more frequently have chronic disease than younger ages. There are some outliers detected for younger patients with chronic disease, further suggesting that younger patients rarely have chronic disease.

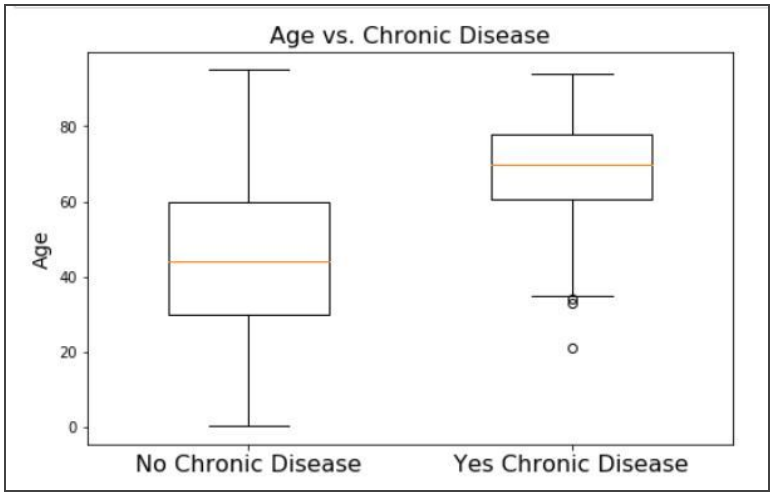


*Figure 7. Age vs Chronic Disease plot*

The wings associated with the box plots span a larger age range for patients without chronic disease than with chronic disease. This tells us that, if a patient is older, they are not necessarily expected to have chronic disease. What we can extract from all of this is that older patients tended to have chronic disease more frequently than younger patients, but a wide age range of patients did not have chronic disease.

Figure 8 demonstrates the plot results for Country vs Chronic Disease. The results show that patients in Vietnam and China were recorded as having chronic disease, while no patients in Italy or the United States were recorded as having chronic disease. Whether these results are due to chance or a lack of records for patients that had chronic diseases in Italy and the United States is unknown. However, these results are interesting when one considers the results of Figure 4 (in which Vietnam and China had patients more frequently recover than not) alongside the results of Figure 3 (which suggest that chronic disease may play a significant role in a patient's outcome). We would have expected countries with higher levels of patients with chronic disease to have lower frequencies of recovery, but this is not what these results demonstrate. It is possible that Italy and the

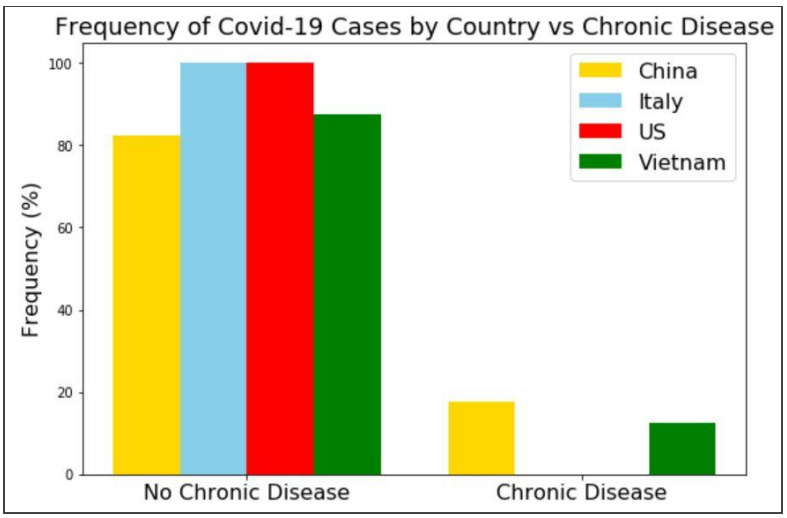


*Figure 8. Country vs Chronic Disease plot*

United States do have higher levels of patients with chronic disease than Vietnam and China but information on this feature was not properly recorded. For all of these reasons, it is especially important to keep these findings and theories in mind when predicting patient outcome based on country.

Figure 9 demonstrates the plot results for Age vs Country. The results show that the mean age of patients recorded in Vietnam is ~30 years of age, the mean age of patients recorded in the United States is ~70 years of age, the mean age of patients recorded in Italy is ~70 years of age as well (with a 30 year old outlier), and the mean age of patients recorded in China is ~50 years of age. Considering these results alongside the results of Figure 5 (which suggest that age may have a significant effect on recoverability) enables us to theorize why Vietnam and China were found to have more frequent patient recoveries than Italy and the United States in Figure 4. If the average age of patients recorded in the dataset is younger for one country than another, that country's patients may recover more frequently if age does in fact significantly affect a patient's recoverability.

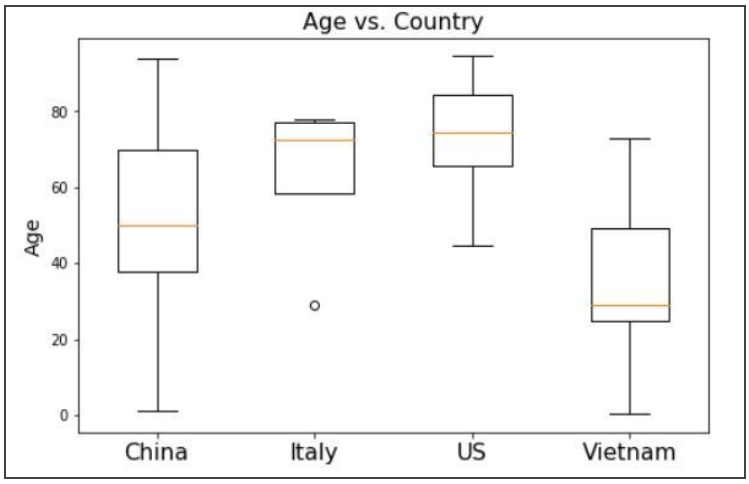


*Figure 9. Age vs Country plot*

## Model Construction

### Model Comparisons

Prior to constructing our classification model solutions, it was important that different model algorithms were compared to one another so that the most accurate models could be determined. To do this, we first compared four classification models against each other: AdaBoost, Naive Bayes, Random Forest, and Logistic Regression. We didn't investigate Neural Networks because our dataset size post cleaning and pre-preprocessing was small (673 samples x 4 features), and Neural Networks predict best for large datasets. We also didn't investigate KNN because all of the features we are working with besides age are categorical and qualitative, and the numbers assigned to them are arbitrary. Since the KNN algorithm classifies based on distances defined between continuous variables, it wouldn't make sense to apply KNN to this kind of dataset, with only one continuous variable. For these reasons, AdaBoost, Naive Bayes, Random Forest, and Logistic Regression were compared to determine which models would best predict the outcome. The models were plotted on an ROC Curve and their AUC and accuracy scores were determined, as seen in Figure 10 and Table 2.

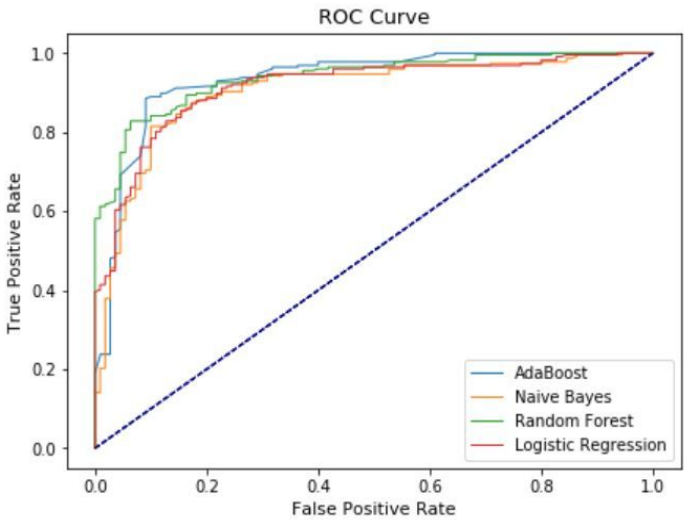


*Figure 10. ROC Curve for different methods*

Table 2. AUC and ACC Scores

| | AUC | Accuracy |
|---|---|---|
| **AdaBoost** | 0.9332 | 0.8724 |
| **Naive Bayes** | 0.9037 | 0.8516 |
| **Random Forest** | 0.9381 | 0.8724 |
| **Logistic Regression** | 0.9142 | 0.86 |

Methods highlighted in green, Random Forest and Logistic Regression, were selected as methods to base our models on. These were selected not only for their high AUC and accuracy scores, but for additional reasons pertaining to their unique handling of variables that will be explained in further detail later. Table 3 demonstrates the probabilities that a patient is recovered for each of the four different methods we are interested in investigating. It is apparent that Naive Bayes has very extreme probabilities that tend to swing almost entirely to one outcome or another. The rest of the methods have less extreme probability values associated with each case.

Table 3. Different Model Probabilities of Recovery

| | AdaBoost | Naive Bayes | Random Forest | Logistic Regression |
|---|---|---|---|---|
| 0 | 0.319929 | 0.263565 | 0.003528 | 0.154838 |
| 1 | 0.491697 | 0.000057 | 0.402116 | 0.109747 |
| 2 | 0.505661 | 0.988383 | 0.706578 | 0.861700 |
| 3 | 0.477913 | 0.000037 | 0.020690 | 0.073315 |
| 4 | 0.512293 | 0.960779 | 0.989493 | 0.912687 |
| ... | ... | ... | ... | ... |
| 332 | 0.477913 | 0.000041 | 0.025703 | 0.079574 |
| 333 | 0.543160 | 0.987780 | 1.000000 | 0.962505 |
| 334 | 0.539485 | 0.996522 | 0.995900 | 0.973414 |
| 335 | 0.813762 | 0.999683 | 0.879980 | 0.965943 |
| 336 | 0.313793 | 0.000048 | 0.001167 | 0.058233 |

## Random Forest Solution

The first model decided for this classification problem was Random Forest. We chose Random Forest, a supervised learning algorithm, because it handles categorical variables well and it requires no assumptions on the distribution of the data. Random

Forest also works efficiently if it encounters the issue of collinearity and it provides an understandable explanation of the predictions. Random Forest was picked over other CART methods such as Adaboost or regular Decision Trees because it is less prone to overfitting due to its predictions made by plurality. Lastly, with Adaboost, the errors of past predictions affect the following ones in contrast to Random Forest, where the decisions are made independently from each other. With this in mind, we preferred Random Forest over Adaboost as we do not want an outcome of a particular patient to affect the prediction of another. .

**Parameters**

We built a random forest of 500 decision trees (n_estimators = 500). As we increased the amount of trees in the forest from the default setting (100 trees), the accuracy consistently improved and converged to approximately 91-92% once we surpassed 500 trees. We bootstrapped and shuffled the data to increase the diversity of trees made to achieve better results. Figure 11 represents the AUC achieved by different parameters for Random Forest. The first set of parameters is the ideal and chosen parameter for our model. It achieved an accuracy of 89% and AUC of 94%. These parameters include 500 random trees, minimum samples leaf of 1, and our data to be bootstrapped and shuffled. The

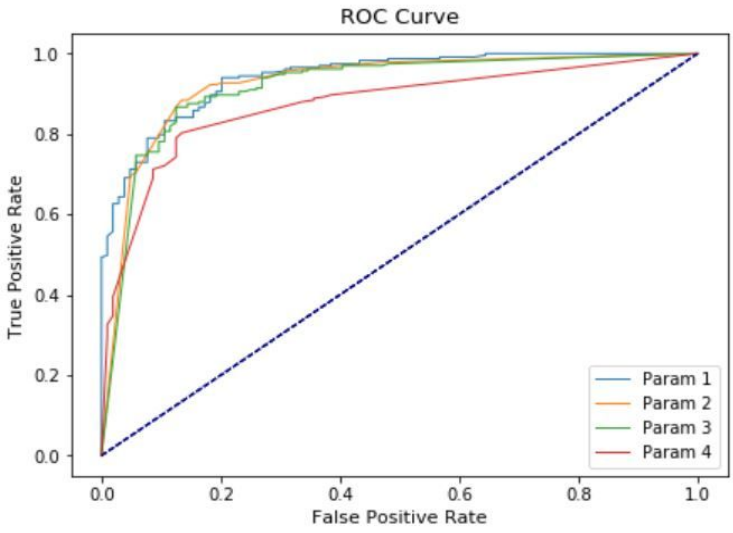


*Figure 11. RF ROC Curve for Differing Parameters*

second set of parameters reduced the number of trees to 100 and the change brought down the AUC to 92% and accuracy to 88%. For our third experiment, we kept the trees at 500 but turned off bootstrapping and saw that although the AUC did not suffer (91%), the accuracy of the predictions was cut to 86%. The reasoning behind this is simply that bootstrapping was absolutely necessary with our limited dataset. It allowed Random Forest to manufacture extra data to accommodate the amount of trees built. Lastly, we increased the minimum sample leafs to 100 and saw a significant decrease in the AUC (87%) and accuracy (81%).

**Results**

Random Forest concluded that age was the most important feature for classifying COVID-19 patients, followed by country of the case, then chronic disease history and sex. Random Forest was able to achieve an average accuracy of 89% and AUC of 94%. It consistently outperformed other models tested on this data set.

Confusion matrices of the chosen parameters vs. tested parameters are in Figure 12.

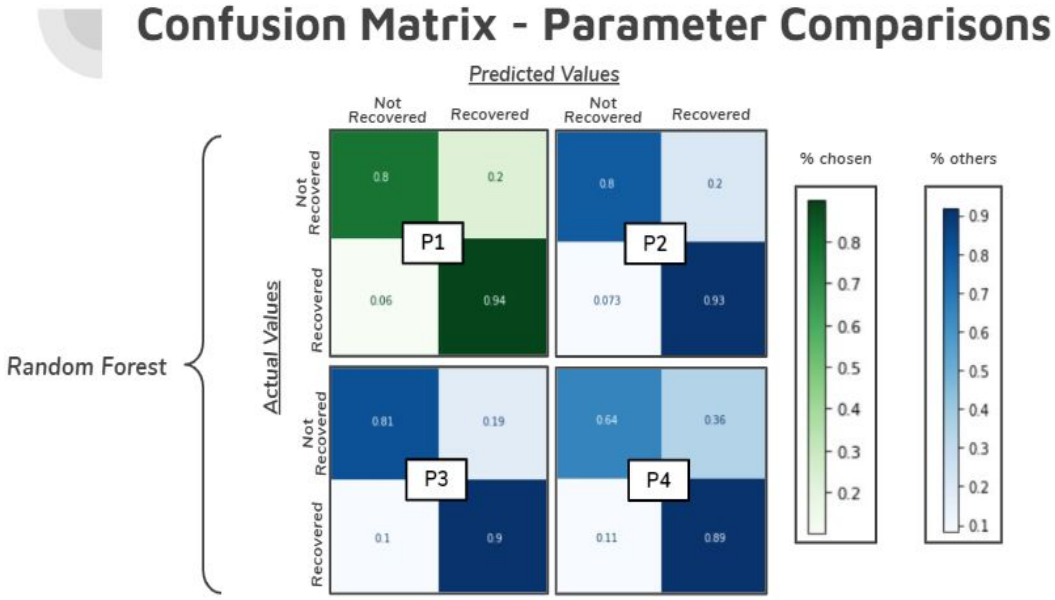


*Figure 12: Confusion Matrices for Random Forest Differing Parameters*

The chosen parameters (P1) produced the highest percentage of true positives and true negatives, and is therefore an ideal option for Random Forest parameters.

## Logistic Regression Solution

The second model that we decided on using was Logistic Regression. We chose Logistic Regression because it requires no assumptions regarding normally distributed data. In addition, its simplicity and ability to only give values between 0 and 1 was perfect for our data set, a classification prediction in which we looked into only two outcomes. Logistic regression parameters gave intuitive ways to explain the direction and intensity of significance of features over our outcomes. Logistic regression was chosen over other methods such as Decision Trees or Naive Bayes for a couple reasons. It was chosen over Decision Trees because it can derive the significance of features, and trees generally have a harder time coming up with calibrated probabilities. As for Naive Bayes, Logistic Regression performs better upon collinearity, due to Naive Bayes expectancy that all features are independent.

### Parameters

We built a logistic regression model with the following parameters (penalty = 'elasticnet', solver ='saga', l1_ratio= 0.5). When changing the parameters of our model, we found that substitution of any penalty (“l2, l1 or elastic net”) or solver (“liblinear, newton-cg, saga, or lbfgs”, we found that the performance did not change. Between all variations, accuracy scores ranged between 0.8546-0.8576.

Figure 13 represents the AUC achieved by two different parameter sets. Param1 is (penalty = 'elasticnet', solver ='saga', l1_ratio= 0.5), and Param2 is (penalty = 'l2' solver = 'lbfgs', l1_ratio = None). The first set of parameters produced an AUC of 91% and an accuracy of 85.4%, while the second set of parameters achieved an AUC of 91.1% and accuracy of 85.7%.

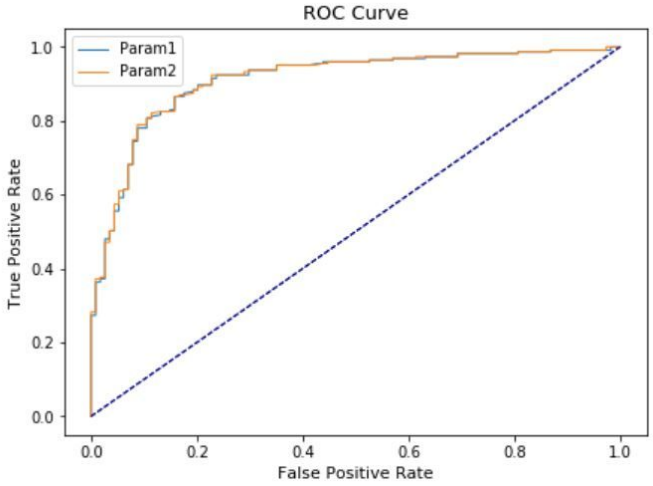

*Figure 13. Logistic Regression ROC Curve*

**Results**

Logstic Regression concluded that chronic disease was the most important feature for classifying COVID-19 patients, followed by sex, then age and country of the case. Although the outcomes of the two parameters are close, and in fact Param2 had minimally improved results, we chose Param1 to be labeled as our "best" logistic regression model. Due to the small differences, our explanation for this lies within the parameters themselves. For Param1, Elastic net penalty increases variance, and due to the bias-variance trade-off, this makes our predictions more generalizable for the test set and thus decreasing variance. In this case, a better generalized model should perform better with an unknown test set from our boss later on. Figure 13 demonstrates confusion matrices of the chosen parameter vs. tested parameter.

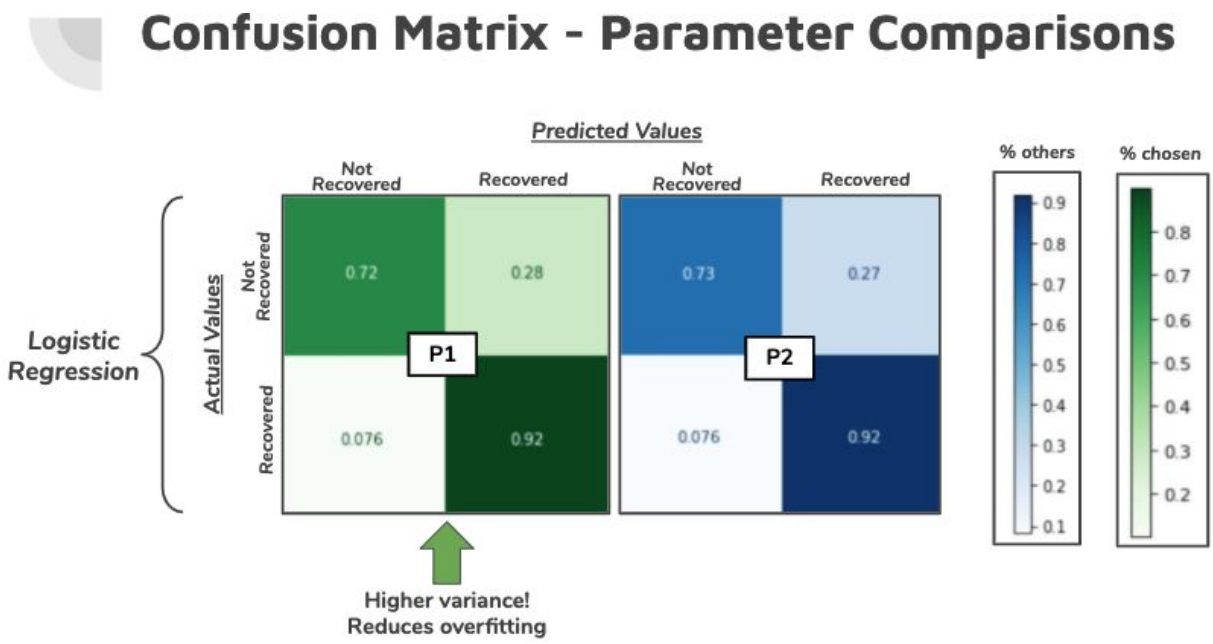


*Figure 14. Confusion Matrices for Logistic Regression Differing Parameters*

## Future Plans

Although Random Forest and Logistic Regression have provided great results despite working within a relatively limited data set, there are several additions that could enhance future predictions. First, it would be beneficial to have more supportive features that can improve the quality of the predictions and possibly provide more details on the outcome of the patients such as probability of re-infection, most influential features behind an infection, etc. Additional features that would help us do this include the status of a patient's critical condition, whether a patient was hospitalized or not, the time taken for a patient to make a full recovery, if a patient practiced social distancing or not, what a patient's travel history was like, and the severity of a patient's symptoms. Other preprocessing techniques that can be tried in future plans include removing the notable outliers that were found after data visualization post preprocessing. We could

also try grouping all ages into ranges as opposed to making age a continuous variable. If provided more data and samples, we plan to construct a neural network that is more capable and versatile to learn and make more accurate predictions.

## References

(1) beoutbreakprepared. *Beoutbreakprepared/NCoV2019*; 2020. https://github.com/beoutbreakprepared/nCoV2019/tree/master/latest_data