PREDICTION OF DISEASE SEVERITY AND OUTCOME IN COVID-19 PATIENTS USING A SUPERVISED MACHINE LEARNING-BASED MODEL

ILANI DAYANA BINTI NOOR AZMAN

S2003292

https://github.com/Ilani-wq/Prediction-of-Covid-19-Using-Supervised-Machine-Learning-/tree/WQD-7006---Machine-Learning

FACULTY OF COMPUTER SCIENCE AND TECHNOLOGY
MASTER OF DATA SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2022

TABLE OF CONTENTS

1.0 INTRODUCTION

In December 2019, the first cases of the coronavirus Covid-19 (Coronavirus Disease 2019), later dubbed SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2), were reported in Wuhan, Hubei Province, China. Covid-19 is a dangerous respiratory condition that causes cough, fever, nausea, low pulse oxygen level, and trouble breathing, among other symptoms. Patients with Covid-19 who do not have any symptoms are more likely to be treated late, resulting in a larger proportion of deaths. This ever-increasing demand places a greater strain on the healthcare system. The new coronavirus outbreak was declared a Public Health Emergency of International Concern (PHEIC) by the World Health Organization (WHO) in January 2020. (Bogoch et al., 2020; World Health Organization). The official name of this disease has been chosen by the World Health Organization (WHO).

On November 26, 2021, the World Health Organization (WHO) announced the discovery of a novel COVID-19 variant called Omicron B.1.1.529, based on several mutations. Although the knowledge of the Omicron variant is limited, WHO has stated that it is easily spreadable and that it is believed to have spread widely in many countries, despite the fact that it has yet to be detected. According to early findings, the Omicron variant is spreading faster than other variants, but it appears to be less severe than the Delta variant.

The Omicron variant is now the most significant concern until the pandemic's end. According to scientists, the Omicron variant does have more than 30 mutations that might spike proteins that cover the exterior of the virus, vaccination targets, and monoclonal antibody therapies. Because every time it duplicates itself, there's a high possibility it might change the structure, the variant is always shifting and developing. These changes are referred to as "mutations," and a virus that has one or more mutations is defined as a "variant" of the original virus.

Despite the fact that Covid-19 is a significant disease in the United States, worry has developed for the higher-risk patients who are also Covid-19 patients (Johns Hopkins University and Medicine Coronovarus Resource Center., 2020). Until date, no one in the world has been able to find a specific vaccination or treatment that directly deals with this Covid-19, despite the fact that various drug firms have developed several vaccines that have shown strong reactions and have been approved by the WHO for use worldwide (Mahmood et al. 2020). According to the most

recent figures, the United States has the highest number of Covid-19 deaths in the country, at 823,390. However, did these people die solely as a result of Covid-19, or did they die as a result of a combination of conditions that they had when Covid-19 was positive?

The goal of this research project is to use a supervised machine learning model to predict the early detection of Covid-19 patients' outcomes based on their disease features. A well-timed prediction of Covid-19 at-risk patients, combined with preventative measures, is likely to improve patient survival while lowering the death rate. By implementing this machine learning method, healthcare providers will be able to better understand which patients should be admitted first. Also, by using Tree Viewer, which will be displayed in the results, this project will be able to determine whether a person with a positive Covid-19 has become worse or not.

The supervised machine learning techniques such as Nave Bayes, Logistic Regression, Decision Tree, and Random Forest are used to extract features, handle missing values, eliminate worthless features, and choose target variables in data preprocessing and prediction. The dataset has been analyzed and shared using Orange technologies, which will benefit the research community in the future.

2.0 OBJECTIVES

The objectives of this project are:

1. To explore the severity of diseases among Covid-19 patients.
2. To predict the severity of diseases in Covid-19 patients discharged and deceased.
3. To develop a prediction model based on the severity of the disease in Covid-19 patients.

3.0 RESEARCH BACKGROUND

3.1 Introduction

Artificial Intelligence (AI) is now widely used in a variety of fields, including the detection of Covid-19 instances, which can be aided by a machine learning algorithm that can improve accuracy, save time, and speed. AI may be used to deal with a wide range of problems, and it has proven to be the most effective solution (Tayarani-N 2020). The use of machine learning and statistical models to do various activities without explicit commands demonstrates that they are systematic reviews utilising computers (Bishop 2006).

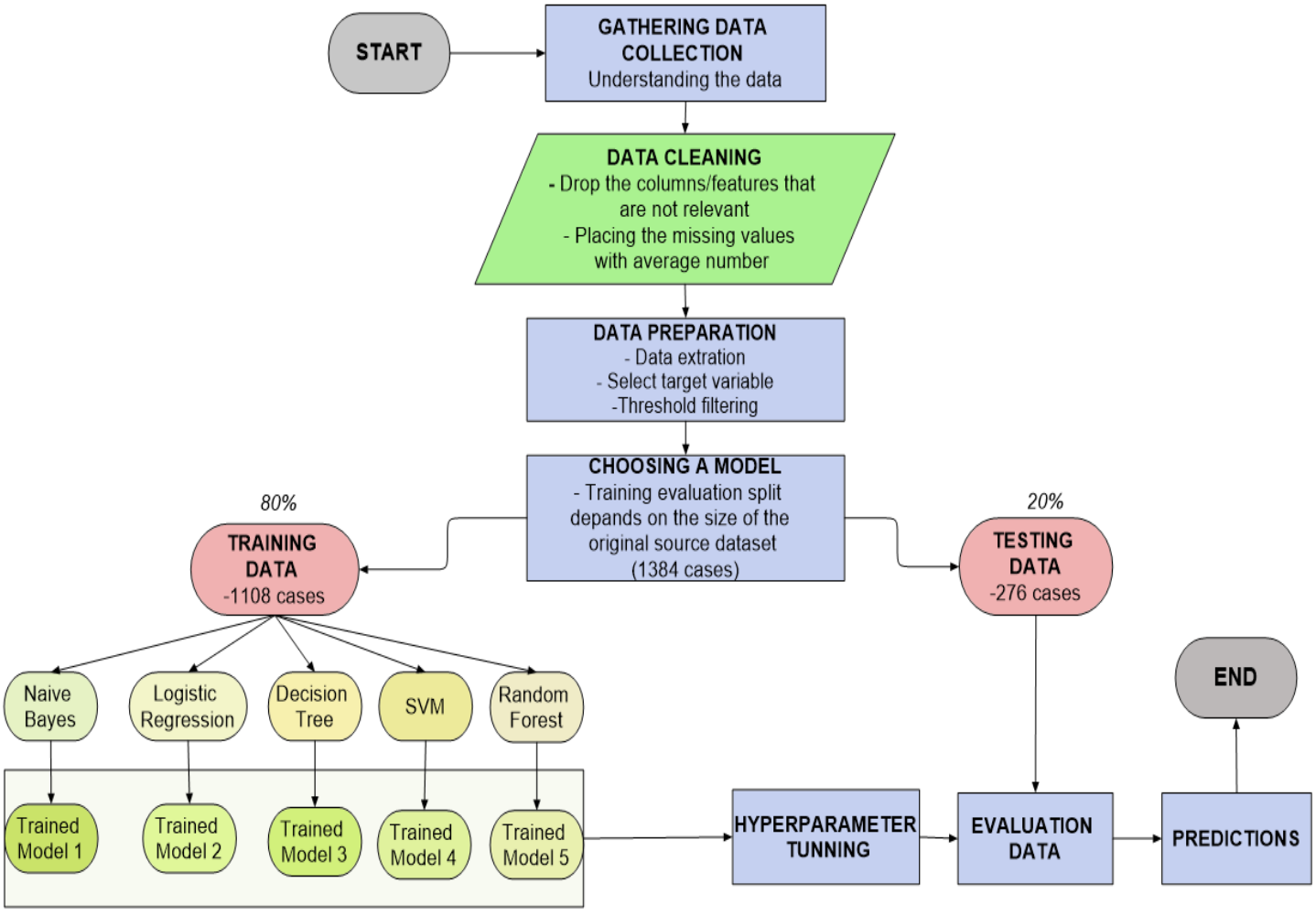
Both the number of discharged and deceased patients with positive Covid-19 who had serious Covid-19 symptoms and needed to be alerted, and the number of discharged and deceased patients with positive Covid-19 who had serious Covid-19 symptoms and needed to be alerted, must be considered in this study. Patients who test positive for Covid-19 in any of these categories have a variety of dangerous symptoms that can harm their bodies and pose a substantial health risk. Understanding those with a high-risk level for this severe Covid-19 symptom has a lot of ramifications. Covid-19 is a highly dangerous disease that has caused havoc on healthcare systems around the globe.

This Covid-19 dataset was collected from patients who tested positive for Covid-19 at Stony Brook University. The Renaissance School of Medicine at Stony Brook University enabled the data gathering, and the Office of the Dean spearheaded the data quality endeavor, which was supported by the Department of Biomedical Informatics. This project case applies the Orange tool, and significant outcomes will be shown in Chapters 5.

4.0 ANALYSIS AND DESIGN

4.1 Flowchart

Figure 1 depicts the basic structure of the supervised machine learning model. At the end of the day, data cleaning is required. The average number has been used to replace the missing value. To keep a strategic distance from any inclination in training and testing, the dataset is segregated into training and testing datasets at this point. The data was used for training in 80% of the cases, and the testing data was used in 20% of the cases for the proposed action's performance.



4.2 Supervised Machine Leaning Models

When working with labelled data, a supervised learning model is used to plan, observe, and direct the execution of a task, project, or activity by 'training' the model. The model is trained with labelled data to predict the outcome of out-of-sample data. Machine learning is one of the most promising methods for classification (Hossain 2019). In order to investigate Covid-19 predictions, four categorization models were used: Naïve Bayes, Logistic Regression, Decision Tree, and Random Forest.

4.2.1 Naïve Bayes

The Nave Bayes technique is a basic and effective classification technique that aids in the construction of a fast machine learning model capable of making a quick prediction. Also known as Bayes' Rule or Bayes' Law, this formula is used to determine the likelihood of a hypothesis with previous knowledge, which is based on the conditional probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

where,

$P(A)$= Prior probability (probability of hypothesis before observing the evidence).

$P(B)$= Marginal probability (probability of evidence).

$P(A|B)$ = posterior probability (probability of hypothesis A on the observed event B).

$P(B|A)$= Likelihood probability (probability of the evidence given that the probability of a hypothesis is true).

4.2.2 Logistic Regression

The classification algorithm logistic regression is used to assign observations to a discrete set of classes. It is utilised for classification problems, predictive analytic algorithms, and is based on the concept of probability in supervised machine learning

algorithms. Logistic regression is a classification procedure that is used to allocate observations to a particular group or category of individuals based on one or more predicted variables (Li et al. 2020). Unlike linear regression, which generates a continuous number, logistic regression transforms the result into a probability value that may be mapped to two or more discrete groups using the logistic sigmoid function.

In order to map expected probabilities, the Sigmoid function or Activation function is used to translate the outcome into a categorical value. Having a feature that may be taken as a real number and mapped to a range of 0-1 shaped like an "S" in a mathematical function. In equation (2), the sigmoid function, also known as a logistic function, is used to assess the connection between the independent and dependent variables or attributes in a dataset (Naw Safrin et al. 2019).

$$s(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

where,

$s(x)$= sigmoid function which is donate an output between 0 and 1 (probability estimate).

x= denoted the input to the function such as $mx + b$.

e= Eucler's number (real number constant).

4.2.3 Decision Tree

In statistics, data mining, and machine learning, the decision tree classification algorithm is one of the predictive modelling approaches. It is a well-known machine learning tool that employs a tree-like model of decision making and probable outcomes.

Internal node for feature or attribute testing, leaf node for decision node or class label, root node for base splitter, branches for feature conjunctions that lead to those class labels, and paths from root to leaf are classification rules. Pruning strategies are one of the ways utilised in this project research to remove noise from the dataset and avoid overfitting in the decision tree algorithm.

4.2.4 Random Forest

Random forest is a type of supervised machine learning that is commonly utilised in classification and regression problems. It creates decision trees based on different samples and uses the majority of the trees for classification and the average for regression. Importantly, random forest can handle datasets with both continuous and categorical variables (in the case of regression) (vase of classification). A random forest classifier separates a training dataset into subsets, which are then provided to each decision tree in the random forest system. The outputs of each decision tree are unique to it. The ultimate output of a random forest is based on an average or majority rating and is built from subsets of data. As a result, the issue of overfitting is resolved.

It is, however, slower than decision trees and is only successful if the trees are diversified and acceptable. Before using the random forest in this research endeavor, the ensemble technique was investigated, which uses numerous models to generate a prediction rather than a single model. Bagging techniques or bootstrap aggregation were used to construct various subsets from sample training data with replacement and final output based on majority vote.

4.3 Confusion Matrix

The confusion matrix is a strong tool for predictive analysis in machine learning that may be used to evaluate the effectiveness of a model-based categorization. Furthermore, the confusion matrix is a table that summarizes the number of correct and wrong predictions produced by the classification model for binary classification tasks. In the shape of a square matrix, the column displays the actual values while the row depicts the model's predicted value, and vice versa. As shown in the confusion matrix, there are four types of terminology and their derivations, as shown below.

- True Positive (TP): The actual value was positive, and the model predicted a positive value.
- True Negative (TN): The actual value was negative, and the model predicted a negative value.

- False Positive (FP): Predict as positive and it was false (Type I error).
- False Negative (FN): Predict as negative, and the result also false (Type II error).

i. Accuracy

- Measure how many predictions is correct in model made for the complete test dataset. it's a nice basic metric to measure the performance of the model. However, in unbalanced datasets, accuracy becomes a poor metric.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

ii. Precision

- Tells how many of the correctly predicted case turned out to be positive which will determine the model is reliable or not. Its useful metric in case the false positive is higher concern than false negative.

$$Precision = \frac{TP}{TP+TN} \quad (4)$$

iii. Recall

- Tells how many actual positive cases is where able to predict correctly with the model that useful metric in case false negative trumps false negative.
- A higher recall score means the most positive cases (TP+FN) will be labeled as positive (TP) which is leas to a higher number of false positive measurement and a lower overall accuracy.
- A lower recall score means have a higher number of false negatives which should have positive but labeled as negative. Also, means that have more certainty if found a positive case, this is like to be true positive.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

iv. F1 Score

- Is a harmonic mean of Precision and Recall. When to try increase the precision of model, the recall grows down and vice-versa. It is maximum when precision is equal to recall.
-

$$F1 - Score = \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{6}$$

5.0 EXPERIMENT RESULT

5.1 Data Description

This project investigates a strategy for predicting and classifying people who are Covid-19 patients with a variety of disorders. All positive patients collected by Stony Brook University throughout the required data collection period are included in the dataset. There are 1384 Covid-19 patients in the dataset, with 127 different features, and 26.9% of the data is missing. Table 1 shows the 11 characteristics of Covid-19 patients' disease including one target variable that were chosen for this research endeavor.

Table 1: Description of the Dataset with Missing Value

#	Features Name	Label	Role	Missing value
1.	age_splits	Years	Feature	0%
2.	gender_concept_name	Female/male	Feature	2%
3.	cough_v	Yes/no	Feature	23%
4.	nausea_v	Yes/no	Feature	26%
5.	vomiting_v	Yes/no	Feature	25%
6.	diarrhea_v	Yes/no	Feature	25%
7.	fever_v	Yes/no	Feature	22%
8.	temperature.over38	True/false	Feature	2%
9.	pulseOx.under90	True/false	Feature	0%
10.	Respiration.over24	True/false	Feature	0%
11.	HeartRate.over100	True/false	Feature	0%
12.	Last.status	Discharged/deceased	Target variable	0%

5.2 Data Analysis and Pre-Processing

The results of categorical and numeric sort columns are included in this dataset. Because the machine learning model requires numerical input in the dataset, label-encoding was done. Since there are few missing values in the dataset, the Impute widgets have started replacing them with the average or most frequent value.

The data has been cleaned up and the missing value has been filled in using the Impute widget. Using the Feature Statistics widget can quickly explore a dataset pattern that includes variables such as variance, mean, bias, percentiles, missing values, and more.

In the collection, there are 1384 Covid-19 patient records with a binary class designation, such as "discharged" or "deceased." There are 1201 patients who have been discharged, with 183 of them deceased. Figure 1 depicts the distribution of cases by class label.

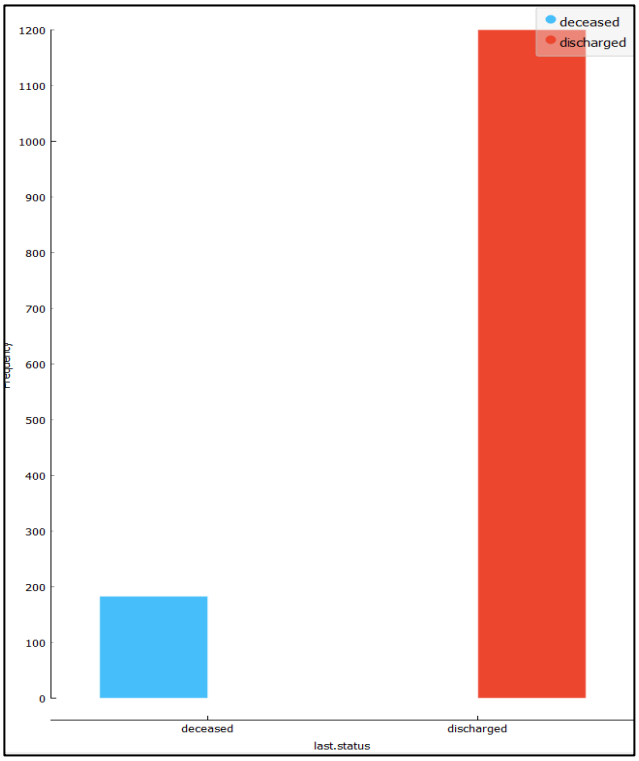


Figure 1: Distributions of Instances Per Class

Figure 2: Data Pattern using Feature Statistics Widget

The FreeViz widget, a sophisticated and dynamic visualisation for showing numerous variables at the same time, was used to gain a better understanding of the feature selected. The step of extracting and selecting features in the FreeViz widget to achieve the best results from the model that can help to demonstrate the information's underlying structure is referred to as "feature selection." Rather than utilising scatter plots, the FreeViz widget is an interactive visualisation for charting several attributes on a 2-D plane in order to fully comprehend the characteristics of data. As shown in Figure 3, patients with Covid-19 exhibit 11 disease. The colours blue and red denote the deceased and discharged, respectively.

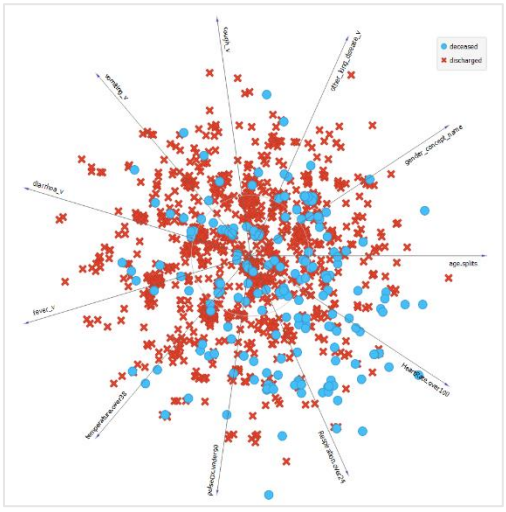


Figure 3: Plotting Multiple Features using FreeViz widget

To provide a fuller description of the characteristics chosen, Table 2 shows the features, label, and value of "Discharged" and "Deceased" for each feature. Table 2 also shows how to acquire a more exact selection of features by converting some categorical data into numerical values using visualisation. The Continuize widget is used to normalise various features to the interval ['Yes': 1, 'No': 0] in order to encode categorical data into numeric. Following feature selection, the Distribution widget is used to run each feature contained in the dataset.

Table 2: Description of the Dataset with Discharged and Deceased

#	Features Name	Label	Discharged	Deceased
1.	age_splits	18-59	735	30
		60-74	284	66
		75-90	182	87
2.	gender_concept_name	Female	537	55
		Male	662	98
3.	cough_v	1:Yes	1040	153
		0:No	161	30
4.	nausea_v	1:Yes	255	13
		0:No	946	170
5.	vomiting_v	1:Yes	138	11
		0:No	1063	172
6.	diarrhea_v	1:Yes	314	34
		0:No	887	149
7.	fever_v	1:Yes	1017	148
		0:No	184	35
8.	temperature.over38	True	305	29
		false	896	154
9.	pulseOx.under90	True	134	72
		False	1067	111
10.	Respiration.over24	True	197	80
		False	1004	103
11.	HeartRate.over100	True	526	73
		False	675	110

5.3 Splitting Data Into Training Data And Testing Data

The data is divided into an 80:20 ratio, with 80% of the data being used to train the model and 20% being used to test it. A machine learning method was used to perform the classifications.

To evaluate the model's supposition from preparing information to hiding information and reduce the risk of overfitting, the dataset has been subdivided into multiple training and test subsets.

At this point, the standardised information is arranged by a proposed method, which is 10-cross validation techniques are utilised for the training set to assess the best classifier. Also, to help the precision of the model, the information tested is not utilised during the training process. The Data Sampler widget was used and works in Orange's tool to implement several means of sampling data from an output channel. Using the Data Sampler widget, the input is the data set to be sampled, and the outputs are a set of sampled data instances (Data Sample) and all other data instances from the input data set that are not included in the sample (Remaining Data).

For the fixed proportion of data, which is set at 80%, it uses for returns a selected percentage of the entire data set. In the Data Sampler widget, replicable sampling was used to maintain sampling patterns that could be carried across users. Now, the data has been split into training data, which has 1108 instances, and testing data, which has 276 instances. The Save Data widget is used for saving the data table that will be used in the next step.

5.4 Evaluation Results

After splitting the data, the Test & Score widget was used by connecting two lines for train data and the remaining data. The training model of machine learning classifications has been utilised and consists of; Naïve Bayes, Logistic Regression, Decision Tree classifier, and Random Forest. Since the dataset could not be an imbalanced dataset, the essential measurement for the examination is F1 Score, Accuracy, Precision, and Recall.

The set size for training was set to 70%, the repeat train/test was set to 10, and the number of folds was set to 10. To avoid overfitting, the Test & Score widget was used, and then tested on a separate test dataset. This procedure was repeated multiple times and then reported on the average accuracy as Figure 4 shows below.

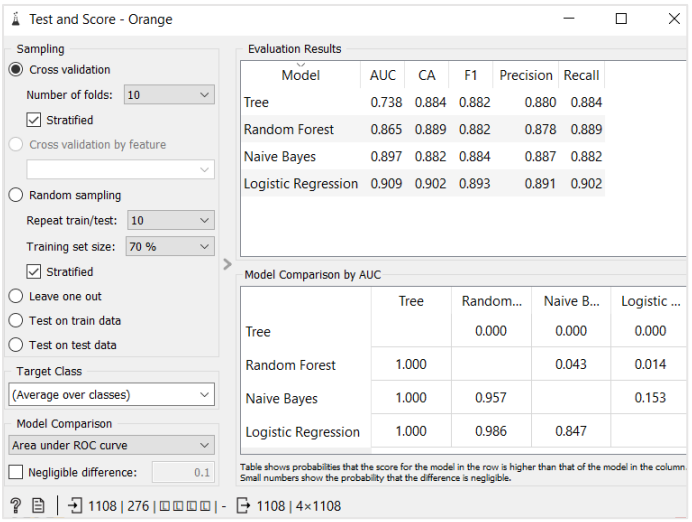


Figure 4: Evaluation Result for Each Model

In this project study, cross-validation spits the data into 10 subsets, using nine for training the model and the remaining subsets for testing it. This procedure was repeated nine more times, each time using a different subset for testing. The classification accuracy (CA) in the second column report on the proportion of correctly data instances shows that Logistic Regression looks like it has a higher classification accuracy, which is 0.902 (90.2%) than other classifiers. The same is true for the area under the ROC (AUC). Logistic Regression has an excellent score that is nearly 1.00 (100%), which is 0.909 (90.9%). The F1 score that is a weighted average of Precision and Recall is 0.893 (89.3%), 0.891 (89.1%) for Precision, which is a pretty good score, and 0.902 (90.2%) for Recall score.

5.5 Prediction

The confusion matrix is a summary of prediction results for a classification task that is crucial for understanding other classification metrics such as Precision and Recall. By displaying the number of correct and incorrect predictions, which were broken down by class and summarised with count values.

The confusion matrix for each classifier is a simple aggregate classifier result evaluated over the folds and is one of the powerful tools for machine learning classification performance

testing. In this project study, the confusion matrix was used to provide a rapid summary of important predictive findings such as accuracy, recall, specificity, and precision.

It's also useful because it compares values such as True Positive, False Positive, True Negative, and False Negative directly. Based on Table 3, which shows the result rate for discharged and deceased of every classifier that gives the proportion of instances between the predicted and actual class.

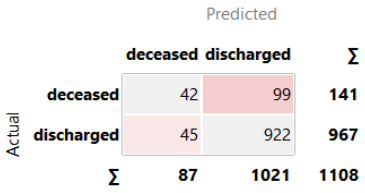


Figure 5: Naïve Bayes

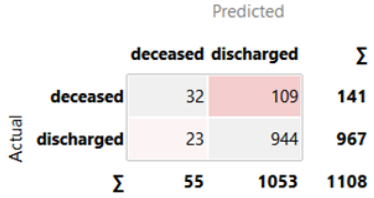


Figure 6: Logistic Regression

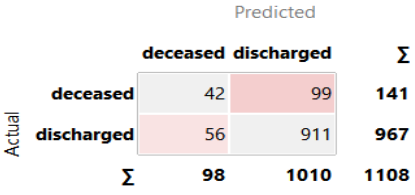


Figure 7: Decision Tree

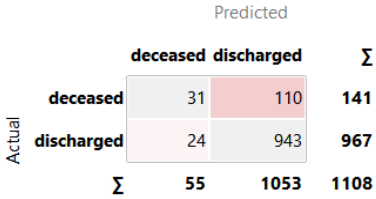


Figure 8: Random Forest

"Discharged" and "Deceased" are two probable predictions. The terms "Discharged" indicate that the Covid-19 patients will be discharged based on the 11 diseases chosen, while "Deceased" indicates that the Covid-19 patients will die as a result of that disease. A total of 1108 predictions were made by the classifier, which were then tested for the presence of the disease.

Table 3: Descriptions of Classifier Predictions Based on Confusion Matrix Result

Model	TP	TN	FP	FN	Descriptions of Classifier Predictions
Naïve Bayes	42	922	45	99	Predicted 87 patients deceased and 1021 discharged
Logistic Regression	32	944	23	109	Predicted 55 patients deceased and 1053 discharged
Decision Tree	42	911	56	99	Predicted 98 patients deceased and 1010 discharged
Random Forest	31	943	24	110	Predicted 55 patients deceased and 1053 discharged

To find a best view based on model classifier, Figure 9 shows the comparisons of predictions as below.

	Naive Bayes	Tree	Logistic Regression	Random Forest	last.status	age.splits	nder_concept_nar	cough_v	nausea_v	vomiting_v	diarrhea_v	fever_v
1	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
2	0.03 : 0.97 → dischar...	0.20 : 0.80 → dischar...	0.03 : 0.97 → dischar...	0.17 : 0.83 → dischar...	discharged	2	1	1	1	0	1	1
3	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
4	0.00 : 1.00 → dischar...	0.04 : 0.96 → dischar...	0.01 : 0.99 → dischar...	0.11 : 0.89 → dischar...	discharged	1	0	1	1	0	1	1
5	0.00 : 1.00 → dischar...	0.04 : 0.96 → dischar...	0.01 : 0.99 → dischar...	0.00 : 1.00 → dischar...	discharged	1	0	0	1	1	1	1
6	0.11 : 0.89 → dischar...	0.00 : 1.00 → dischar...	0.06 : 0.94 → dischar...	0.02 : 0.98 → dischar...	discharged	2	1	0	0	0	0	1
7	0.89 : 0.11 → deceased	0.83 : 0.17 → deceased	0.84 : 0.16 → deceased	1.00 : 0.00 → deceased	deceased	3	0	1	0	0	0	1
8	0.04 : 0.96 → dischar...	0.04 : 0.96 → dischar...	0.05 : 0.95 → dischar...	0.06 : 0.94 → dischar...	discharged	1	0	1	0	0	1	1
9	0.00 : 1.00 → dischar...	0.04 : 0.96 → dischar...	0.01 : 0.99 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	0	1	1	0	1
10	0.16 : 0.84 → dischar...	0.18 : 0.82 → dischar...	0.12 : 0.88 → dischar...	0.19 : 0.81 → dischar...	deceased	2	0	1	0	0	0	1
11	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
12	0.10 : 0.90 → dischar...	0.10 : 0.90 → dischar...	0.08 : 0.92 → dischar...	0.00 : 1.00 → dischar...	discharged	2	1	1	0	0	0	1
13	0.20 : 0.80 → dischar...	0.15 : 0.85 → dischar...	0.20 : 0.80 → dischar...	0.21 : 0.79 → dischar...	discharged	3	1	1	0	0	0	1
14	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
15	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
16	0.07 : 0.93 → dischar...	0.00 : 1.00 → dischar...	0.06 : 0.94 → dischar...	0.00 : 1.00 → dischar...	discharged	2	0	1	1	0	0	1
17	0.04 : 0.96 → dischar...	0.04 : 0.96 → dischar...	0.04 : 0.96 → dischar...	0.00 : 1.00 → dischar...	discharged	1	0	1	0	0	0	0
18	0.06 : 0.94 → dischar...	0.00 : 1.00 → dischar...	0.09 : 0.91 → dischar...	0.05 : 0.95 → dischar...	discharged	3	1	0	1	0	1	1
19	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	1	0	0	0	1
20	0.03 : 0.97 → dischar...	0.33 : 0.67 → dischar...	0.08 : 0.92 → dischar...	0.28 : 0.72 → dischar...	discharged	2	0	1	1	1	1	1
21	0.03 : 0.97 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	1	0	0	0	0	0
22	0.34 : 0.66 → dischar...	0.22 : 0.78 → dischar...	0.29 : 0.71 → dischar...	0.18 : 0.82 → dischar...	discharged	3	0	0	0	0	0	0
23	0.09 : 0.91 → dischar...	0.25 : 0.75 → dischar...	0.06 : 0.94 → dischar...	0.07 : 0.93 → dischar...	discharged	2	0	1	0	0	0	1
24	0.01 : 0.99 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.00 : 1.00 → dischar...	discharged	1	0	0	0	0	1	1
25	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	0.02 : 0.98 → dischar...	0.04 : 0.96 → dischar...	discharged	1	0	1	0	0	0	1
26	0.15 : 0.85 → dischar...	0.23 : 0.77 → dischar...	0.13 : 0.87 → dischar...	0.18 : 0.82 → dischar...	discharged	2	0	1	0	0	0	1

. **Figure 9: Predictions Table Based on Model**

Figure 10 shows that Random Forest has the greatest AUC score of 0.928 (92.8%), which indicates how effectively the predictions are ranked. It also has the highest CA score of 0.904 (90.4%), Precision of 0.895 (889.5%), and Recall of 0.904 (90.4%). (90.4%). The highest F1 score is 0.891 for Decision Tree or Tree classification (89.1%).

Model	AUC	CA	F1	Precision	Recall
Naive Bayes	0.837	0.875	0.860	0.854	0.875
Tree	0.861	0.900	0.891	0.888	0.900
Logistic Regression	0.849	0.881	0.860	0.857	0.881
Random Forest	0.928	0.904	0.888	0.895	0.904

Figure 10: Comparisons of Supervised Learning Model

The Tree Viewer widget, as shown in Figure 11, provides an interactive visualization of instances where the relationship between features and outcome or whether features interact with each other for a better understanding of the predictive result. To prevent the Tree from falling to the problem of overfitting, which causes it to grow out of control, the Tree's edge width and depth were set to control, and the Tree was pruned after it learned in order to remove some of the details it had taken up. The Tree's depth is set to six levels, and the edge width is proportional to the root.

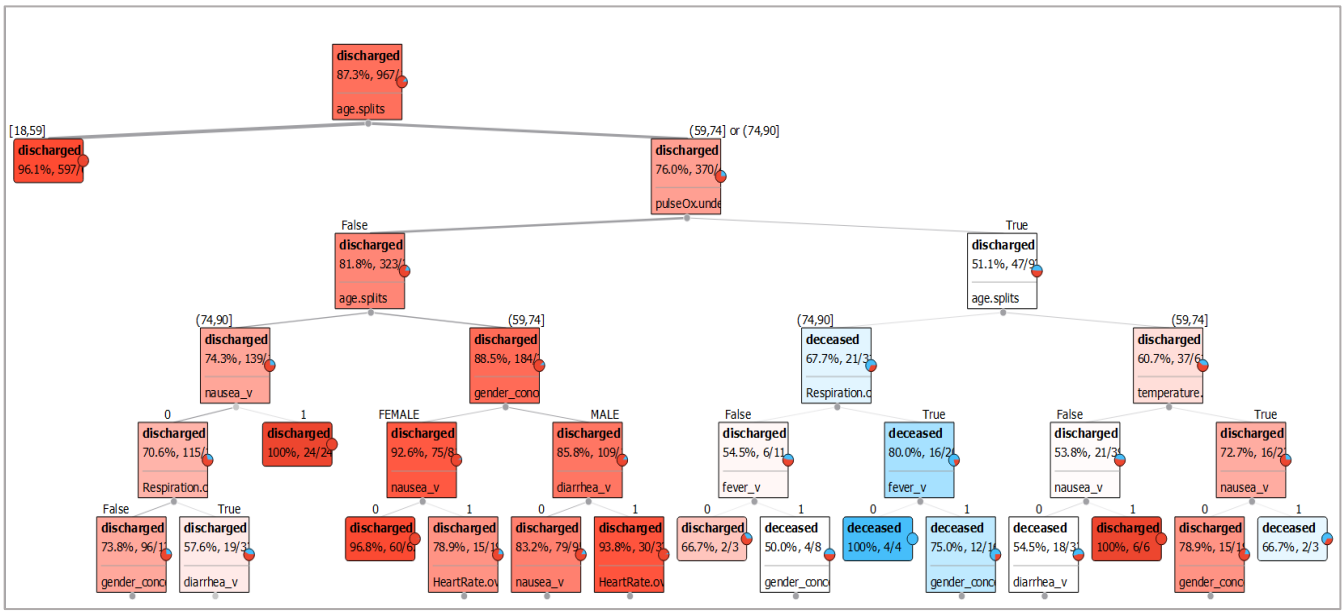


Figure 11: Tree Viewer widget

According to the Tree Viewer results, the split point was divided depending on age, with 96.1% of patients aged 18 to 59 being discharged and also 76.0% of patients aged 60 to 74 and 75 to 90 being discharged, according to the Tree Viewer results. The internal node for pulse oxygen under 90 in the second split node shows discharged patients who are 51.1 percent "True" for pulse oxygen under 90 in white and 81.8 percent "False" for pulse oxygen under 90 in red. Finally, for discharged and deceased patients with disease selected in Tree Viewer, there are three categories and predictions, which are:

a) Prediction of Discharge Based on Disease Severity

   i) Patients with an average age of 74–90 whose pulse oxygen is not under 90 and have nausea predict 100% will be discharged.

   ii) Patients with an age range of 59–74, whose pulse oxygen is not under 90, are female, have nausea of 96.8%, and males without nausea but with diarrhea of 93.8% predicted to be discharged.

   iii) Covid-19 patients aged 59–74 who have a pulse oxygen level of less than 90, a temperature of less than 38, and nausea have a 100% chance of being discharged.

b) Prediction of Deceased Based on Disease Severity

   i) Patients with an age between 74-90, pulse oxygen under 90, respiration over 24, and who do not have a fever have a predicted 100% chance to deceased.

   ii) Patients aged 59–74 with pulse oxygen under 90 and a temperature over 38 have a nausea scoring risk of 66.7%.

   iii) Patients with an age between 74-90, respiration over 24, and a fever have a risk to deceased is 75.0%.

6.0 DISSCUSION

The strength of this project study is that it is the first-time data from Stony Brook University has been used to do an analysis using a supervised machine learning algorithm that can be used as a preference and direction in the future. As it contains a big sample size and some critical features that can be very valuable for future analysis, this dataset is ideal for prediction. The prediction result based on this project study may be used as a well-timed prediction of at-risk Covid-19 patients with preventive actions using machine learning methodologies, which is projected to raise the survival rate of patients and lower the deceased rate in the future.

The limitation of this project study is that it does not dive deep into the details of the attributes, which are usually only understood by those with medical experience. When it comes to gender, cough, nausea, vomiting, diarrhea, fever, and temperature, the data has certain missing values, which can be used to count the severity of each patient's disease and predict discharge and diseases more accurately in cases of Covid-19.

All the models are suitable for use for prediction since the Training and Testing results are almost the same in accuracy. Using a Decision Tree is very powerful as a predictive model for people without a data science background since it is very simple to understand, can perform well with large datasets, is highly possible to validate a model using a statistical test, and is able to handle both numerical and categorical data. A Random Forest is the optimum match to the modification of bagging where Trees are designed to reduce correlation, as Decision Trees have a high risk of growing very deep and learning highly irregular patterns that can be overfit the training sets that have low bias but high variation. A Random Forest is a mechanism for averaging numerous depths in Decision Trees and is trained on different parts of the same training set with the purpose of reducing variation. Since it predicts the categorical dependent variables "Discharged" and "Deceased," Logistic Regression has the highest accuracy for prediction in this project study.

7.0 CONCLUSION

In conclusion, the project study discovered that age, pulse oxygen levels, respiration, temperature, nausea, fever, and gender were the most relevant predictors of Covid-19 discharged and deceased. The most accurate machine learning model for predicting the outcome with good accuracy compared to other models for the dataset chosen was Logistic Regression, while the highest score for F1 was Decision Tree. In this project study, all of the objectives were achieved, including exploring illness severity, predicting disease severity, and developing a prediction model based on disease severity in Covid-19 patients gathered by Stony Brook University.

It has been proved that even though those between the ages of 59-74 and 74-90 have a significant risk of diseases, those under the age of 59-74 and 74-90 should not be negligent when having severe diseases. If there is a severe infectious disease, society's awareness should be raised so that everyone feels responsible for battling the pandemic together. Evaluating the population level of discharged and deceased owing to severe diseases with positive Covid-19 could provide crucial insights not only in the United States, but throughout the world, in order to prevent the rising number of people dying due to a lack of preventive measures.

The severity of the disease was classified by gender and age, so researchers can determine whether males or females are more afflicted by Covid-19, as well as which age groups are most affected. Increasing the amount of data and characteristics in a model will improve its performance, resulting in a greater outcome. Collaboration between individuals with data analysis and medical backgrounds is important, and a deep learning technique can be used in the future to increase performance. Predictive modelling techniques such as SVM and Neural Networks may be used in the future to help predict future outcomes in order to gain a more in-depth understanding and make better decisions.

8.0 REFERENCES

Bishop, C.M. (2006). "Pattern Recognition and Machine Learning Springer-Verlag New York." *Inc. Secaucus, NJ, USA*, 2006.

Hossain B., Morooka, T., Okuno, M., Nii, M., Yoshiya, S., & Kobashi, S. (2019). Surgical Outcome Prediction in Total Knee Arthroplasty Using Machine Learning. *Intelligent Automation and Soft Computing,* 25(1),105–115.

Johns Hopkins University and Medicine Coronovarus Resource Center. (2020). COVID-19 United States Cases by County.

Mahmood, A., Eqan, M., Pervez, S., Alghamdi, H. A., Tabinda, A. B., Yasar, A., ... & Pugazhendhi, A. (2020). COVID-19 and Frequent Use of Hand Sanitizers; Human Health and Environmental Hazards by Exposure Pathways. *Science of the Total Environment*, 742, 140561.

McConghy, T., Pon, B., & Anderson, E. (2020). *"When Does Hospital Capacity Get Overwhelmed in USA? Germany? A Model of Beds Needed and Available for Coronavirus Patients" trent.st*.

Tayarani-N, M.-H (2020) Applications of Artificial Intelligence in Battling Against Covid-19: A Literature Review. *Chaos Solitons Fractals* 142:110338.

World Health Organization, 2. (2020). Statement on the Second Meeting of the International Health Regulation Worldwide-By-Country/s (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-nCoV).