Enhanced Knowledge in Sciences and Technology Vol. 2 No. 2 (2022) 021-031
© Universiti Tun Hussein Onn Malaysia Publisher's Office

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN : 2773-6385

Predictive Analysis On Recovery of Covid-19 Patients in Singapore by Using Data Mining Techniques

Koa Kar Mun[1], **Sabariah Saharan**[2]*

[1 2] Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology,
Universiti Tun Hussein Onn Malaysia, Pagoh Edu Hub, 84600, Johor, MALAYSIA.

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2022.02.02.003
Received 1 January 2022; Accepted 30 September 2022; Available online 23 November 2022

Abstract: Whilst the researchers are actively searching for infections and recovery data across countries actively, the information of patients' recovery on COVID-19 disease is poorly recognized. There is a lot of uncertainty of the mild or asymptomatic COVID-19 cases in the clinical presentation that may never present to healthcare services. The purpose of this study is to determine the recovery and stability of the newly infected persons with pandemic of COVID-19 based on the Age, Gender, Causes of Infection Cases and Number of Days for Hospitalized by using a descriptive analysis. In addition, it is also aims to develop predictive model for prediction of COVID-19 infected patients by using Decision Tree, K-Nearest Neighbors and Naïve Bayes. Lastly, to evaluate the most accurate predictive model in estimating the recovery of the infected patients from COVID-19 between these predictive models by using accuracy evaluation. Data Mining techniques such as Decision Tree, K-Nearest Neighbors and Naïve Bayes help to predict the causes of infection cases that are more likely to recover from COVID-19; the maximum and minimum number of days for the patients to recover, and the recovery rate of the different age groups. The result of this research shows the Decision Tree model has been proven as the most efficient data mining technique in predicting the recovery of COVID-19 patients in Singapore with the greatest percentage of accuracy of 78.95% among other predictive models. In future research, it is suggested to use Malaysian data with more relevant attributes and input samples so that it can reflect a better performance to the government about the trend of the COVID-19 patient recovery.

Keywords: Decision Tree, K-Nearest Neighbors, Naïve Bayes, COVID-19

**Corresponding author: sabaria@uthm.edu.my*

1. Introduction

Coronavirus disease 2019 is an infection which is caused by SARS-CoV-2 virus, and it can spread to a person through close contact, droplets of saliva, mucus from nose and mouth during coughing and sneezing [1]. According to [2] the novel coronavirus 2019 named as COVID-19 was emerged from Wuhan City, Hubei Province, China on 31 December 2019. It has been identified as a powerful outbreak in many cities of China, spreading globally to at least 25 countries as of February 2, 2020 including United State, United Kingdom, Italy, Spain, Japan, Korea, India, Singapore and also Malaysia. The World Health Organization [3] has officially named this disease as coronavirus disease 2019 (COVID-19) on 11 February 2020. This powerful outbreak of COVID-19 has infected more than 3 million patients in 187 countries with a 4.20% mortality rate which has become the biggest global threat [4]. According to the update of [5] on 5 April 2021, the number of confirmed deaths due to COVID-19 for our country is up to 1295 people (0.37%), while the total number of cases confirmed is 352029, which means the percentage of death will continually increase.

COVID-19 is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is similar to the zoonotic severe acute respiratory syndrome (SARS) and coronavirus (SARS-CoV) in 2002 [6]. Whilst the researchers are actively searching for infections and recovery data across countries actively, the information of patients' recovery on COVID-19 disease is poorly recognized. According to [7], there is a lot of uncertainty of the mild or asymptomatic COVID-19 cases in the clinical presentation that may never present to healthcare services. Based on the current data available, there is still no specific antiviral treatment that can hundred percent cure for COVID-19, and the infection control guidance for COVID-19 is more likely be based on COVID-19 vaccines while the antiviral pill for COVID-19 still undergoing investigation. The rapid spreading of COVID-19 is caused by the changes in population such as the rapid growth of tourism, changing geography of migration and also the long distance taken by people for family reunion [8]. The COVID-19 patients have symptoms which are similar to normal illnesses such as fever, cough, nausea and diarrhoea. The most significant symptoms for COVID-19 are anosmia, sore throat and cough [9], while some other symptoms like decreased smell have also been reported [10].

The purpose of this study is to determine the recovery and stability of the newly infected persons with pandemic of COVID-19 based on the Age, Gender, Causes of Infection Cases and Number of Days for Hospitalized by using a descriptive analysis. In addition, it is also aims to develop predictive model for prediction of COVID-19 infected patients by using Decision Tree, K-Nearest Neighbors and Naïve Bayes. Lastly, to evaluate the most accurate predictive model to predict the possibility of recovery of the infected patients from COVID-19 pandemic between the predictive models by using accuracy evaluation. The scope of this study is limited to the patients of COVID-19 from Singapore between the age from 0 to 96 and the duration of the study is from January to February in 2020.

2. Materials and Methods

The dataset was obtained from California School Employees Association [11] which was available on Kaggle Website. The dataset was prepared and cleaned where only relevant attributes were extracted from the original dataset. Since missing values caused biased estimations which lead to inaccurate conclusions, thus mice imputation techniques were used to handle the missing values in this dataset. The extracted dataset has 93 samples, 465 data instances with 5 attributes which includes gender, age, causes of infection cases, number of days for hospitalized and state of patients which shown as Table 1. The 465 data instances are enough for data mining since it can predict different possibility and sufficient to conduct significant statistics. In addition, the attribute of causes of infection cases shown the reason that the patients get infected while the number of days for hospitalization indicated the duration of the patients being hospitalized. Lastly, the state attribute shown the final outcome of the patients, whether get recovered or unrecovered which is death due to COVID-19 infection.

Table 1: Sample of the instances of the Singapore COVID-19 patients' recovery dataset.

Gender	Age	Causes of Infection Cases	Number of Day for Hospitalization	State
Male	66	From Wuhan	28	Recovered
Male	28	Grace Assembly of God	18	Recovered
Male	53	Church Cluster	25	Recovered
Female	38	Private Meeting	10	Recovered
Male	36	Close contact with patients	13	Recovered

2.1 Decision Tree

Decision Tree is one of the data mining techniques that is widely used for solving regression and classification problems due to its simplicity and transparency [12]. According to [13], a Decision Tree represents a tree-structured classifier that performs a split test in its internal node and predicts a target class of an example in its leaf node.

In this research, Gini Index is used as a measure for attribute selection and act as a binary split to calculate for each attribute. It is used to measure the degree of probability variable being wrongly classified when it is randomly chosen [14]. The Gini Index for each attribute was calculated by subtracting the sum of squared probabilities of each class from one, and the lowest Gini index value is picked as the root node. According to [15], the Gini Index is defined as:

$$Gini(K) = 1 - \sum_{i=1}^{n}(pi)^2 \qquad Eq. 1$$

$$Gini_A(K) = \frac{N_1}{N} Gini\ (K_1) + \frac{N_2}{N} Gini\ (K_2) \qquad Eq. 2$$

$$\Delta Gini(A) = Gini(K) - Gini_A(K) \qquad Eq. 3$$

According to the equations of Gini Index in Eq. 1, *pi* is the relative frequency and *n* is the number of attributes. If the dataset is split on attribute A into two subsets K_1 and K_2, Gini Index is calculated as Eq. 2, where N_1 and N_2 explained the sizes of subsets. Thus, the reduction in impurity is calculated as Eq. 3.

2.2 K-Nearest Neighbors

K-Nearest Neighbors model is the simplest techniques which can classifies a new data point based on the similarity and categorizes the new point to the most frequent classes. Before applying K-NN algorithms, normalization of data is required to change the column with numerical values to a common scale to create better performance of accuracy. The formula of normalization for numerical data is shown as below [16].

$$X_s = \frac{X - Min}{Max - Min} \qquad Eq. 4$$

According to the Eq. 4, *Min* explained the minimum distance while *Max* is the maximum distance in training set. Furthermore, X_s is the standardized distance and *X* is known as distance. In this study, the age of the patients acted as X_i, while the number of days for hospitalized acted as Y_i.

By using the techniques of Euclidean Distance, it can help to calculate the distances between the two data and determine their similarity. The group with shortest distance can be classified as a group and it can classify the patients' recovery states based on the age of patients and the number of days for hospitalized. The formula of Euclidean Distance as shown below [17].

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad Eq. 5$$

Based on Eq.5, *d* is the distance between two points, while *n* is the number of points. Generally, the label of *x* is denoted as the position of *x* coordinate while *y* is the position of *y* coordinate.

2.3 Naïve Bayes

Naïve Bayes is a probabilistic and statistical machine learning algorithm which can be used to differentiate the dataset instances based on attributes and features [17]. By using the Naïve Bayes theorem, the probability of *A* happening given that *B* has occurred. The formula of Naïve Bayes is defined as below [18].

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)} \qquad Eq. 6$$

$$B = (b_1, b_2, b_3 \dots, b_n) \qquad Eq. 7$$

$$P(A|b_1, \dots, b_n) = \frac{P(b1|A)P(b2|A)\dots P(b_n|A)P(A)}{P(b_1)P(b_2)\dots P(b_n)} \qquad Eq. 8$$

Based on the equation above, $P(A|B)$ is explained on how often *A* happens given that *B* happens while $P(B|A)$ is illustrated how often *B* happens given that *A* happens. In addition, *P(A)* is explained the probability of *A* occurs and the formula of $(b_1, b_2, b_3 \dots, b_n)$ are the features.

2.4 Performance evaluation

After building the Decision Tree, K-Nearest Neighbors and Naïve Bayes models, the confusion matrix is created to measure the performance of each model. The confusion matrix only shows correct outcome prediction while the incorrect prediction is located outside the diagonal of the confusion matrix. In order to determine the predictability, accuracy values of the data mining models are evaluated using evaluation techniques [19]. It is given as the percentage of total correct predictions divided by the total number of instances. The higher the accuracy of the model, the better the performance of the model. The formula of accuracy was shown as below [20].

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \qquad Eq. 9$$

Based on the equation above, *TP* is known as the number of cases belongs to a class and actually belong to it while *FP* is known as number of cases belong to a class but reality does not. Oppositely, *TN* is explained on the number of cases does not belong to a class and actually does not belong to it but *FN* explained in it does not belong to a class but reality it does.

3. Results and Discussion

3.1 Descriptive Analysis

In this research, the descriptive analysis was applied to see the overall visualization of the data.

3.1.1 Comparison of Age Attribute with Status of Patients' Recovery

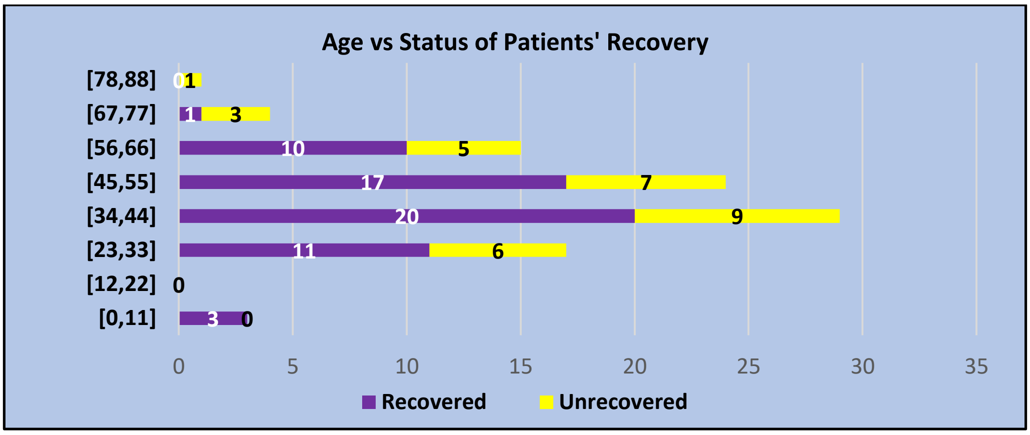

Figure 1: Frequency of Age vs Status of Patients' Recovery Attribute

Based on Figure 1, there were a higher number of patients recovered from COVID-19 in middle age between 23 years old to 66 years old. For example, the age group of patients between 34 to 44 years old had the highest recovery rate with 68.97%, since there were 20 patients who recovered out of 29 patients. However, the patients' age above 67 years old had a lower recovery rate, 20%, whereby there were only 1 patient who recovered from COVID-19 out of 5 patients.

3.1.2 Comparison of Number of Days for Hospitalized Attribute with Status of Patients' Recovery

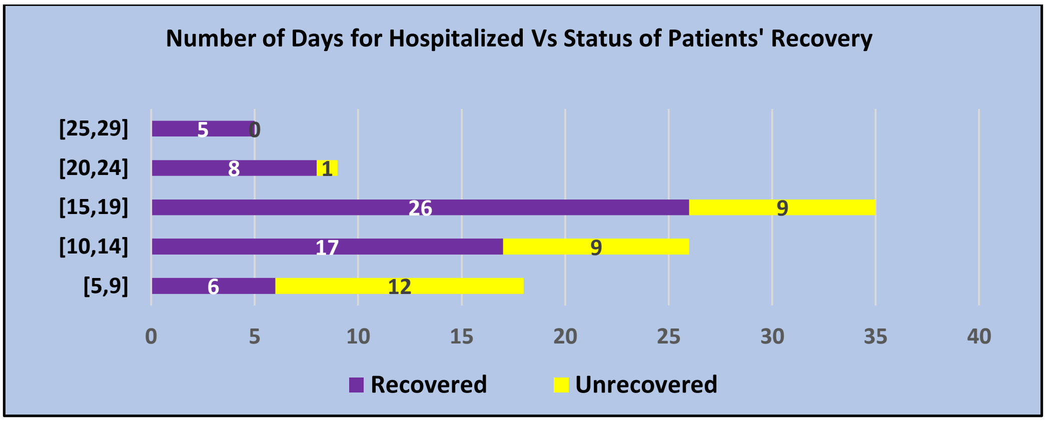

Figure 2: Number of days for hospitalized vs status of patients' recovery attribute

According to Figure 2, the patients with 15 to 19 days of hospitalization had the highest frequency of patients' recovery, 74.28% whereby there were 26 patients who had successfully recovered from COVID-19 out of 35 patients. Besides, there were 100% of patients' recovery rate for those who were staying 25 days to 29 days in hospital, that means, the patients with longer duration for hospitalization were more likely to be recovered. However, the patients with shorter hospitalized duration (5 to 9 hospitalized days) had high risk not to recover since the rate of recovery was only 33.3% and only 6 patients could recover out of 18 patients.

3.1.3 Comparison of the Causes of Infection Cases with Status of Patients' Recovery

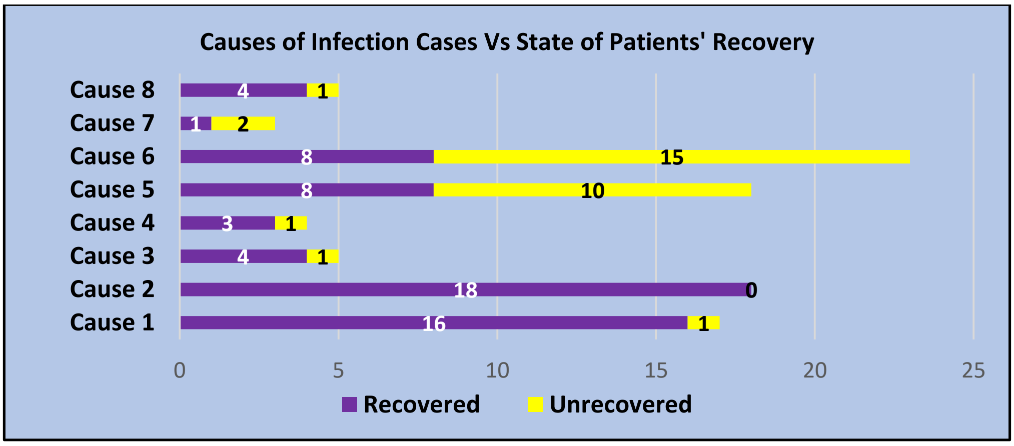

Figure 3: Frequency based on the causes of infection cases vs status of patients' recovery attributes

Based on Figure 3, the patients infected in Cause 1 (From Wuhan) and Cause 2 (Visited Wuhan) are more likely to recover because the recovery rate for the patients in both infection causes were 94.11% and 100% respectively. However, the patients from Cause 6 (Church Cluster) shows 15 patients are unrecovered out of 23 patients, 65.21%, which means there are more patients dying due to COVID-19 disease for the same type of causes. This is same for the patients in Cause 5 (Close contact with patients) since there are around 10 patients recorded as unrecovered, which is the cause of infection cases that contained the second highest frequency of patients' unrecovered.

3.2 Decision Tree

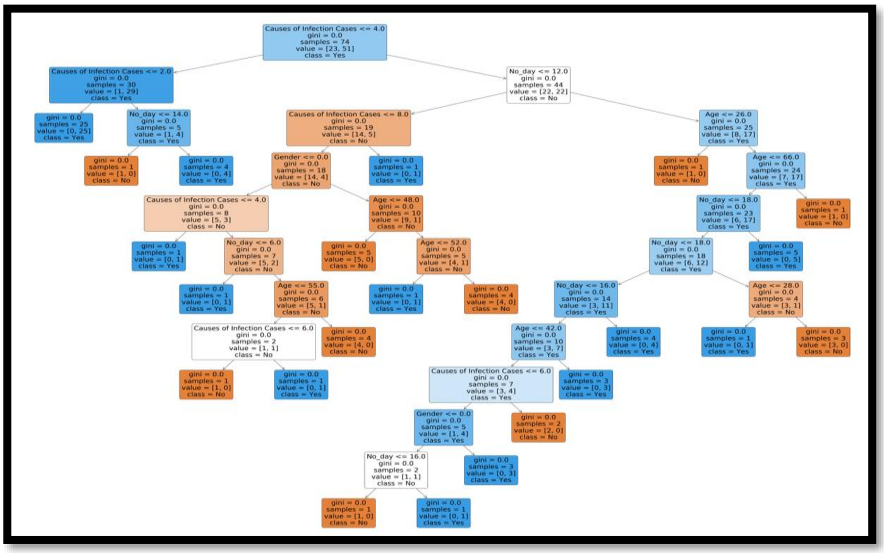
Causes of Infection Cases <= 4.0 gini = 0.0 samples = 74 value = [23, 51] class = Yes
Causes of Infection Cases <= 2.0 gini = 0.0 samples = 30 value = [1, 29] class = Yes
No_day <= 12.0 gini = 0.0 samples = 44 value = [22, 22] class = No
gini = 0.0 samples = 25 value = [0, 25] class = Yes
No_day <= 14.0 gini = 0.0 samples = 5 value = [1, 4] class = Yes
Causes of Infection Cases <= 8.0 gini = 0.0 samples = 19 value = [14, 5] class = No
Age <= 26.0 gini = 0.0 samples = 25 value = [8, 17] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 4 value = [0, 4] class = Yes
Gender <= 0.0 gini = 0.0 samples = 18 value = [14, 4] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
Age <= 66.0 gini = 0.0 samples = 24 value = [7, 17] class = Yes
Causes of Infection Cases <= 4.0 gini = 0.0 samples = 8 value = [5, 3] class = No
Age <= 48.0 gini = 0.0 samples = 10 value = [9, 1] class = No
No_day <= 18.0 gini = 0.0 samples = 23 value = [6, 17] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
No_day <= 6.0 gini = 0.0 samples = 7 value = [5, 2] class = No
gini = 0.0 samples = 5 value = [5, 0] class = No
Age <= 52.0 gini = 0.0 samples = 5 value = [4, 1] class = No
No_day <= 18.0 gini = 0.0 samples = 18 value = [6, 12] class = Yes
gini = 0.0 samples = 5 value = [0, 5] class = Yes
gini = 0.0 samples = 1 value = [0, 1] class = Yes
Age <= 55.0 gini = 0.0 samples = 6 value = [5, 1] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 4 value = [4, 0] class = No
No_day <= 16.0 gini = 0.0 samples = 14 value = [3, 11] class = Yes
Age <= 28.0 gini = 0.0 samples = 4 value = [3, 1] class = No
Causes of Infection Cases <= 6.0 gini = 0.0 samples = 2 value = [1, 1] class = No
gini = 0.0 samples = 4 value = [4, 0] class = No
Age <= 42.0 gini = 0.0 samples = 10 value = [3, 7] class = Yes
gini = 0.0 samples = 4 value = [0, 4] class = Yes
gini = 0.0 samples = 1 value = [0, 1] class = Yes
gini = 0.0 samples = 3 value = [3, 0] class = No
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes
Causes of Infection Cases <= 6.0 gini = 0.0 samples = 7 value = [3, 4] class = Yes
gini = 0.0 samples = 3 value = [0, 3] class = Yes
Gender <= 0.0 gini = 0.0 samples = 5 value = [1, 4] class = Yes
gini = 0.0 samples = 2 value = [2, 0] class = No
No_day <= 16.0 gini = 0.0 samples = 2 value = [1, 1] class = No
gini = 0.0 samples = 3 value = [0, 3] class = Yes
gini = 0.0 samples = 1 value = [1, 0] class = No
gini = 0.0 samples = 1 value = [0, 1] class = Yes

Figure 4: Decision Tree model for the recovery of COVID-19 patients.

According to the Decision Tree in Figure 4, the first splitting of the Decision Tree model is the Cause of Infection Cases which is the most contributed attribute to the recovery of COVID-19, followed by No_day (Number of Days for Hospitalized), Age group and Gender of the patients. The model estimated the patients infected due to Cause 1 to 4 (From Wuhan, Visited Wuhan, Cluster of Shop Tour Visited, and Private Meeting) are more likely to be recovered especially the patients infected due to Cause 1 (From Wuhan) and Cause 2 (Visited Wuhan) of infection cases. On the other hand, this model also predicted a minimum of 12 days as the number of days for the infected patients to be recovered from COVID-19. The patients in all type of Causes of Infected Cases with less than 12 days for hospitalized will mostly classified as unrecovered even the age is less than 48. Thus, the patients with longer duration for hospitalization are more likely to recover from COVID-19 if the number of days for hospitalization is above 13 days. However, in some cases, there are also some patients who are unable to recover from the pandemic virus even if they are below 28 years old. As a result, the Decision Tree model had an accuracy rate of 78.95% in predicting COVID-19 patients' recovery for the testing dataset.

3.3 K-Nearest Neighbors Model

First of all, the data had undergone normalization to change the values in the numerical column to a common scale because the different ranges of the dataset cause the distortion of data. To evaluate the model's performance, the patients' predicted recovery state, *Test_pred* is matched up with the patients' actual recovery state, *Test_actual* to check for the model's accuracy as shown in Table 2. According to Table 2, there were more patients predicted as recovered (Yes) while only 2 patients are classified as unrecovered (No). Furthermore, the accuracy of the model is checked by using a confusion matrix which the result is shown in Table 3.

Table 2: K-Nearest Neighbors model for the recovery of COVID-19 patients.

S/N	*Test_actual*	*Test_pred*
1	Yes	Yes
2	No	Yes
3	Yes	No
4	Yes	Yes
5	Yes	Yes
6	No	Yes
7	Yes	No
8	No	Yes
9	Yes	Yes
10	No	Yes
11	Yes	Yes
12	Yes	Yes
13	Yes	Yes
14	Yes	Yes
15	Yes	Yes
16	No	Yes
17	Yes	Yes
18	No	Yes
19	Yes	Yes

Table 3: Confusion matrix of K-Nearest Neighbors model.

Test_actual	*Test_pred*		Row Total
	No	Yes	
No	0	6	6
Yes	2	11	13
Column Total	2	17	19

Whilst from the confusion matrix table, there are a total 17 patients predicted as recovered while only 2 patients are classified as unrecovered. According to Table 3, there is none of the patients has been falsely identified as unrecovered (No) after the predicted recovery state (*Test_pred*) is matched with the actual recovery state (*Test_actual*). However, this model has false identification for some of the patients whose actual unrecovered but classified as recovered and vice versa. In addition, Table 3 shows 2 patients have been falsely identified as unrecovered and 6 patients have been wrongly classified as recovered which are misclassified from actual result. Thus, there is overall 42.1% of misclassification rate in predicting the COVID-19 patients' recovery for the testing dataset which means the total accuracy value for K-Nearest Neighbors is only 57.90%.

3.3 Naïve Bayes

Table 4: A-Priori probabilities of Naïve Bayes model.

Y_pred	**No**	**Yes**
	0.3784	0.6216

Table 5: Conditional probabilities of gender for Naïve Bayes model.

Y_pred	Gender	
	Female	Male
No	0.4286	0.5714
Yes	0.3478	0.6522

Table 6: Conditional probabilities of causes of infection cases for Naïve Bayes model.

Y_pred	Causes of Infection Cases							
	1	2	3	4	5	6	7	8
No	0.0357	0.0000	0.0357	0.0357	0.2857	0.5000	0.0714	0.0357
Yes	0.2609	0.3261	0.0435	0.0435	0.0870	0.1522	0.0217	0.0322

Based on Table 4, the A-priori probabilities show a higher probability for the patients to recover from COVID-19 which is 0.6216, while the probability of unrecovered is only 0.3784. Moreover, the conditional probabilities for each attribute are calculated to predict the likelihood of an event or outcome occurring given that event has occurred. For the example in Table 5, the probability of males given recovered from COVID-19, *P(Male|Recovered)* is 0.6522 while the probability of female given recovered from COVID-19, *P(Female|Recovered)* is 0.3478 which indicated the male patients had a higher recovery rate compared to females. Similarly, the new patients can also be predicted based on their Age, Gender, Causes of Infection Cases and Number of Days for Hospitalized. Based on Table 6, the patients infected due to Cause 2 of Infection Case (Visited Wuhan) records as the highest probability to recover which is 0.3261 followed by the patients infected due to Cause 1 of Infection Case (From Wuhan) with the probability of 0.2609. However, the Naïve Bayes model has predicted that the patients who get infected due to Cause 6 (Cluster of Church) have the lowest recovery rate since the probability of recovery is 1.522 and the probability of unrecovered is 0.5000.

Table 7: Confusion matrix of Naïve Bayes classifier for each predictors.

Test_actual	*Y_pred*		Row Total
	No	Yes	
No	2	1	3
Yes	5	11	16
Column Total	7	12	19

According to the confusion matrix in Table 7, it shows there are 2 patients classified as not recovered out of 7 patients while there are 11 patients classified as recovered out of 12 patients. However, there were also some misclassifications occurred during the prediction of patients' recovery whereby there are a total 5 patients being misclassified as recovered and 1 patient was misclassified as unrecovered from COVID-19. Thus, there is an overall 31.57% misclassification rate which indicated the accuracy value in predicting the COVID-19 patients' recovery for the testing dataset is 68.42%.

3.4 Performance Evaluation

In this study, the three predictive data mining models are analyzed and compared based on their accuracy value. The Table 8 shown the performance evaluation of the three different predictive data mining models.

Table 8: Performance evaluation results of the models.

Predictive Model	Accuracy Values (%)
Decision Tree	78.95
Naïve Bayes	68.42
K-Nearest Neighbors	57.90

The accuracy of a classifier was identified as the percentage of total correct prediction divided by the total number of instances. According to Table 8, the Decision Tree shows the most efficient result with the highest percentage of accuracy of 78.95% followed by Naïve Bayes with 68.42%, and the K-Nearest Neighbors had the lowest accuracy values which was only 57.90%. Since the performance of Decision Tree was the most accurate among others predictive models, thus it was the most suitable data mining technique to be used for further data analysis such as classification in financial analysis and predictive diagnostic of medical research.

4. Conclusion

In this research, data mining models such as Decision Tree, K-Nearest Neighbors and Naïve Bayes were built for the prediction of COVID-19 infected patients' recovery. The model built with Decision Tree was recorded as the most efficient data mining technique in predicting the recovery of COVID-19 patients in Singapore with the greatest percentage of accuracy of 78.95% among other predictive models. With the aid of the Decision Tree, it determined 'Causes of Infection Cases' as the most contributed attribute to the patients' recovery of COVID-19. In addition, the COVID-19 patients with longer period of hospitalized days are classified to has higher recovery rate compared to the patients with shorter hospitalized duration such as below 12 days. Furthermore, the recovery and stability of the newly infected patients with the COVID-19 is highly related to the number of days for hospitalization, whereby there are more patients being recovered in the days between 15 to 19 hospitalized and the age of the patients between 34 to 44 years old have the highest recovery rate.

Data Mining has the potential to have greater impact on COVID-19-related investigations, thus, the created models would be extremely useful in healthcare to fight against COVID-19 disease. In future, it is suggested to use the dataset with more attributes by increasing the size of the tree to make the prediction more accurate. Involving more data and attributes in a training set always adds information and increases the accuracy of the model. Outlier detection is an important component to improve the classification accuracy of the model since the existence of outlier values might affect model's accuracy and cause bias. Thus, removing and mean imputation methods are suitable to deal with the outliers to decrease the variability in the dataset and increase the statistical power.

Acknowledgement

The authors would like to thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia for its support and provide basic facilities to carry out this present research work.

References

[1] R. Dhand and J. Li, "Coughs and Sneezes: Their Role in Transmission of Respiratory Viral Infections, Including SARS-CoV-2," American Journal of Respiratory and critical Care Medicine, vol. 202, no. 5, pp. 651-659, 2020, doi:org/10.1164/rccm.202004-1263PP.

[2] Y. C. Wu, C. S. Chen, and Y. J. Chan, "The outbreak of COVID-19: An overview," Wolters Kluwer Public Health Emergency Collection, vol. 83, no. 3, pp. 217-220, 2020, doi: 10.1097/JCMA.0000000000000270.

[3] World Health Organization, WHO, "Naming the coronavirus disease (COVID-19) and the virus that causes it," February 11, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it [Accessed April 10, 2021].

[4] World Health Organization, WHO, "Coronavirus disease (COVID-19) pandemic,". March 20, 2020. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 [Accessed April 25, 2021].

[5] L. L. Thomala, "COVID-19 confirmed and death case development in China 2020-2022," January 2, 2020. [Online], Available: https://www.statista.com/statistics/1092918/china-wuhan-coronavirus-2019ncov-confirmed-and-deceased-number/ [Accessed April 12, 2021].

[6] J. W. Chan, C.K. Ng, Y. H. Chan, T. Y. Mok, S. Lee, S. Y. Chu, W. L. Law, M. P. Lee and P.C. Li, "Short term outcome and risk factors for adverse clinical outcomes in adults with severe acute respiratory (SARS)," Thorax, vol. 58, no. 8, pp. 686–689, 2003, doi: 10.1136/thorax.58.8.686.

[7] J. W. Tang, P. A. Tambyah and D. S. C. Hui, "Emergence of a novel coronavirus causing respiratory illness from Wuhan, China," Journal of Infection, vol. 80, no. 3, pp. 350-371, 2020, doi: 10.1016/j.jinf.2020.01.014.

[8] Q. J. Shi, D. Dorling, G. Z. Cao, and T. Liu, "Changes in population movement make COVID-19 spread differently from SARS," Social Science & Medicine, vol. 255, 2020, doi:10.1016/j.socscimed.2020.113036.

[9] E. Elibol, "Otolaryngological symptoms in COVID-19," European Archives of Oto-Rhino-Laryngology, vol. 278, no. 4, pp. 1233–1236, 2021, doi:10.1007/s00405-020-06319-7.

[10] A. R. Sedaghat, I. Gengler and M. M. Speth, “Olfactory Dysfunction: A Highly Prevalent Symptom of COVID-19 With Public Health Significance,” Otolaryngol Head Neck Surg, vol. 163, no. 1, pp. 12-15, 2020, doi: 10.1177/0194599820926464.

[11] COVID-19, CSEA, “Day by Day Information on COVID-19 Affected Cases,” Jan 22, 2020. [Online]. Available: https://www.kaggle.com/pratik1235/covid19-csea [Accessed January. 18, 2021]

[12] L. M. Wang, X. L. Li, C. H. Cao and S. M. Yuan, “Combining decision tree and Naive Bayes for classification,” Knowledge-Based Systems, vol. 19, no. 7, pp. 511–515, 2016, doi.org/10.1016/j.knosys.2005.10.013.

[13] M. J. Aitkenhead, “A co-evolving decision tree classification method,” Expert Systems with Applications, vol. 34, no. 1, pp. 18–25, 2018, doi:10.1016/j.eswa.2006.08.008.

[14] E. Turban, R. Sharda, and D. Delen, “Business intelligence and analytics: systems for decision support. 10th ed,” New York: Pearson Higher, 2014.

[15] T. Daniya, M. Geetha and K. Suresh Kumar, “Classification and Regression Trees with Gini Index,” Advances in Mathematics: Scientific Journal, vol. 9, no. 10, pp. 8237-8247, 2020 doi:10.37418/amsj.9.10.53.

[16] T. Srivastava, “Introduction to k-Nearest Neighbours: A powerful Machine Learning Algorithm (with implementation in Python & R),” March 26, 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/ [Accessed May 28, 2021].

[17] L. J. Muhammad, M. M. Islam, S. S. Usman and S. I. Ayon, “Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients' Recovery,” SN computer science, vol. 1, no. 4, pp. 206, 2020, doi: 10.1007/s42979-020-00216-w.

[18] R. Gandhi, “Naive Bayes classifier, towards data science,” May 6, 2018. [Online]. Available: https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c [Accessed Mei. 28, 2021].

[19] S. Lakshmanan, “How, When, and Why Should You Normalize / Standardize / Rescale Your Data,” May 16, 2019. [Online]. Available: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff. [Accessed June. 6, 2021].

[20] Y. Gao, “Using Decision Tree to Analyze the Turnover of Employees,” Uppsala University: Master's Thesis, 2017.