Predictive Data Mining Models for Novel Coronavirus (COVID-19) Infected Patients Recovery

L. J. Muhammad (lawan.jibril@fukashere.edu.ng)
Federal University of Kashere, P.M.B. 0182, Gombe, Nigeria https://orcid.org/0000-0003-4175-424X

Md. Milon Islam
Khulna University of Engineering & Technology, Khulna-9203, Bangladesh https://orcid.org/0000-0002-4535-5978

Usman Sani Sharif
Federal University of Kashere, P.M.B. 0182, Gombe, Nigeria https://orcid.org/0000-0002-7965-8678

Safial Islam Ayon
Khulna University of Engineering & Technology, Khulna-9203, Bangladesh

Research Article

Keywords: Coronavirus, pandemic, decision tree, data mining, patients recovery

Posted Date: June 5th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-33247/v1

Version of Record: A version of this preprint was published at SN Computer Science on June 21st, 2020.
See the published version at https://doi.org/10.1007/s42979-020-00216-w.

Abstract

Novel coronavirus (COVID-19 or 2019-nCoV) pandemic has neither clinically proven vaccine nor drugs; however, its patients are recovering with the aid of antibiotics medications, anti-viral drugs, and chloroquine as well as vitamin C supplementation. It is now evident that the world needs a speedy and quicker solution to contain and tackle the further spread of COVID-19 across the world with the aid of non-clinical approaches such as data mining approaches, augmented intelligence and other artificial intelligence techniques so as to mitigate the huge burden on the healthcare system while providing the best possible means for patients' diagnosis and prognosis of the 2019-nCoV pandemic effectively. In this study, data mining models were developed for the prediction of COVID-19 infected patients' recovery using epidemiological dataset of COVID-19 patients of South Korea. The decision tree, support vector machine, naive Bayes, logistic regression, random forest, and K-nearest neighbor algorithms were applied directly on the dataset using python programming language to develop the models. The model predicted a minimum and maximum number of days for COVID-19 patients to recover from the virus, the age group of patients who are of high risk not to recover from the COVID-19 pandemic, those who are likely to recover and those who might be likely to recover quickly from COVID-19 pandemic. The results of the present study have shown that the model developed with decision tree data mining algorithm is more efficient to predict the possibility of recovery of the infected patients from COVID-19 pandemic with the overall accuracy of 99.85 % which stands to be the best model developed among the models developed with other algorithms including support vector machine, naive Bayes, logistic regression, random forest, and K-nearest neighbor.

Introduction

Severe Acute Respiratory Syndrome Coronavirus two (SARS-CoV-2), the causative agent of novel coronavirus (COVID-19 or 2019-nCoV), has emerged in late 2019 which is believed to be originated from Hubei Province, China called Wuhan [16], [25]. 2019-nCoV or COVID-19 is rapidly spreading in humans which is evidently believed that it was first derived from bats, transmitted to humans through intermediate hosts probably the raccoon dog (*Nyctereutes procyonoides*) and palms civet (*Paguma larvata*) [8], [18], [21]. The major symptoms of SARS-CoV-2 include fever, cough, and shortness of breath which in many instances appeared to be similar to that flu [16]. COVID-19 had since reached a decisive point and pandemic potential which claimed the lives of many people across the world and human-to-human transmission of COVID-19 from infected individuals with mild symptoms have been reported [20], [16]. According to Worldometers, COVID-19 pandemic affects 210 countries and territories around the world and two (02) international conveyances with 6,033,875 confirmed cases, 2,661,213 recovered cases, and 366,894 deaths as of May 30th, 2020, 05:37 GMT [27]. However, there is no drug or vaccine clinically proven to treat COVID-19 pandemic, therefore other non- clinical or non-medical therapeutic techniques are urgently needed to contain and prevent further outbreak of COVID-19 pandemic such as

data mining techniques, machine learning and expert system among other artificial intelligence techniques.

Data Mining (DM) is an advanced Artificial Intelligence (AI) technique that is used for discovering novel, useful, and valid hidden patterns or knowledge from dataset [6], [14]. The technique reveals relationships and knowledge or patterns among the dataset in several or single datasets [15], [16]. It has also widely used for the prognosis and diagnosis of many diseases including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and Middle East Respiratory Syndrome Coronavirus (MERS-CoV) that were so far discovered in 2003 and 2012 respectively [16]. As huge dataset generated around the world related to 2019-nCoV pandemic everyday is a treasured resource to be mined and analyzed for useful, valid, and novel knowledge or patterns extraction for better decision making to contain the outbreak of COVID-19 pandemic. In healthcare sector, data mining has been widely applied in many different applications such as predicting patient outcomes, modeling health outcomes, hospital ranking, and evaluation of treatment effectiveness and infection control, stability, and recovery [1], [23], [29].

In this study, we developed several data mining models for the prediction of 2019-nCoV infected patients' recovery. The models predict when COVID-19 infected patients would be recovered and released from isolation centers as well as patients that may likely not be recovered and lost their lives to COVID-19 pandemic. The models help the health workers to determine the recovery and stability of the newly infected persons with pandemic COVID-19. The models are developed with the dataset obtained from Korea Centers for Disease and Prevention (KCDC) and dataset instances of the death and recovery records of the infected 2019-nCoV pandemic were considered. Data mining algorithm which includes decision tree, support vector machine, naive Bayes, logistic regression random forest, and K-nearest neighbor were applied directly on the dataset using python programming language to develop the models.

The rest of the paper is organized as follows. Section 2 describes the overall methodology of the proposed system including data collection and preparation with data mining techniques. The experimental results with detailed discussions are described in Section 3. Lastly, Section 4 concludes the paper.

Methodology

Dataset collection and description

The dataset was obtained from KCDC which was made available on Kagge Website [3]. We used the epidemiological dataset of COVID-19 patients of South Korea. The dataset has three thousand two hundred and fifty-four (3,254) instances with eight (8) attributes which include patient ID, global number (the number given by KCDC), sex, birth year, age, country, province, city, disease (TRUE: underlying disease / FALSE: no disease) infection case, infection order (the order of infection), infected by (the ID of who infected the patient), contact number (the number of contacts with people) symptom onset date, the date of symptom onset, confirmed date, the date of being confirmed, released date (the date of being released)

deceased date (the date of being deceased) and state (state of the patients isolated / released / deceased).

Dataset preparation

The dataset was prepared, and cleaned where only relevant attributes were extracted from the original dataset. The extracted dataset has only one thousand five hundred and five (1,505) data instances with five (5) attributes which include gender, age, infection_case and no_days (the number of days between the date the disease was confirmed and date the patients was released or die) and state of the patients (released or deceased). We considered only two states of the patient which include released and deceased while isolation state was excluded. Table 1 revealed the data type of the attributes and Table 2 showed the sample of some instances of the dataset. The missing value in the dataset reduces the prediction power and produces biased estimates leading to an invalid conclusion [28]. Therefore, we used last observation carried forward imputation technique to handle the missing values in the dataset. Figure 1-5 showed the frequency of each attribute of the dataset.

Table 1: Data type of each attribute

S/N	Attribute	Data type
1	Gender	Object
2	Age	Object
3	Infection_case	Object
4	no_day	Int64
5	State	Object

Table 2: Sample of the instances of the dataset

S/N	Sex	Age	Infection_case	No_day	State
1	male	50s	overseas inflow	13	released
2	male	30s	overseas inflow	32	released
3	male	50s	contact with patient	20	released
4	male	20s	overseas inflow	16	released
5	female	20s	contact with patient	24	released
6	female	50s	contact with patient	19	released
7	male	20s	contact with patient	10	released
8	male	20s	overseas inflow	22	released
9	male	30s	overseas inflow	16	released
10	female	60s	contact with patient	24	released
11	female	50s	overseas inflow	23	released
12	male	20s	overseas inflow	20	released
13	male	80s	contact with patient	11	released
14	female	60s	contact with patient	25	released
15	male	70s	contact with patient	21	released

Data Mining Techniques/Algorithms

Logistic Regression (LR)

Logistic Regression (LR) is used to determine the association between categorical dependent variables against the independent variables [9]. LR is used when the dependent variable has two values such as 0 and 1, yes and no or true and false and thus it is called Binary Logistic Regression [22]. However, when the dependent variable has more than two values, multinomial logistic regression is used. A mathematical model of a set of explanatory variables for LR is used to predict a transformation of the dependent variables. LR transformation is written mathematically as:-

$$i = \text{LogisticRegression}(p) = 1n\left(\frac{p}{1-p}\right) \qquad (1)$$

Let, assume the dependent values are numerical of 1 and 0 where 0 represents negative value and 1 positive value as a binary variable. Therefore, the mean of the binary variable will the proportion of

positive values. If p is the proportion of observations with an outcome of 1, then 1-p is the probability of an outcome of 0. The ratio p/ (1-p) is called the odds and the LG is the logarithm of the odds or just log odds.

Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the supervised learning algorithms used for classification and regression [24]. For classification task in SVM involves testing and training data which contain some instances of the data [10]. Each instance in the training dataset contains one or more target values; therefore the main goal of SVM is to produce a model that will predict target value or values [24]. For regression, SVM applied by the introduction of alternative loss function which can be linear or nonlinear [10].

Decision Tree (DT)

Decision Tree (DT) is used for classification tasks in data mining and successful technique due to its ability to handle both categorical and continuous data, simplicity and comprehensibility, DT builds tree into phases which include growth and pruning phases respectively [14-15], [26]. In the first phase, a tree is built by partitioning data into a smaller set until each partition is pure, however, the split type of the data is solely dependent on the data type [19]. The splits for a numerical attribute C form the value of (C) $\leq$ y, where y is value in C domain. For splitting a categorical D, form the values of (D), B$\in$G, where G is a subset of domain (D) [7]. To remove the noise in the dataset, the pruning technique is used to get the final tree built when it is fully grown [11]. The growth phase is however computationally more expensive than the pruning phase of the decision tree [6].

Naive Bayes (NB)

Naive Bayes is one kind of data mining classification algorithm and used to discriminate dataset instances based on specified features or attributes [13]. NB is a probabilistic classifier and uses Bayes theorem for classification tasks [5]. Below is the Bayes theorem:

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} \qquad (2)$$

Random Forest (RF)

Random Forest (RF) algorithm is an ensemble learning technique for data mining classification and regression tasks. The algorithm constructs a multitude of decision trees at training time and outputting [12]. RF data mining algorithm is the best to be used for any decision tree with overfitting to its training dataset [13].

K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) is a non-parametric and supervised data mining classifier used for regression and classification tasks [2]. In both tasks, the input variables consist of the K closes training dataset in the feature space. K-NN relies on labeled input data to learn a function so that to produce appropriate output when inputted unlabeled data [17]. In K-NN classification, the output is a class membership in which data instances is classified by a plurality vote of its neighbors, with the data instance being assigned to the class most common among its K-nearest neighbors while in K-NN regression, the output is the property value of data instance and this value is the average of the value of K-nearest neighbors [4].

Experimental Results

Python programming language was used for data mining predictive tasks. Python is a well-known general purpose and dynamic programming language that is being used for different fields such as data mining [30], machine learning [31], [32], and internet of things [33], [34]. Data mining algorithms are being implemented using python with the help of the special purpose libraries. The models were developed using 5 fold cross-validation.

In decision tree, the model revealed in Figure 6, the No. of days attribute appeared to be the first splitting attribute which indicates the most important attribute. The model predicted a minimum of 5 days and a maximum of 35 days as the number days for COVID-19 patients to recover from the pandemic virus. The model has shown that another important attribute for predicting recovery of the COVID-19 patients is age attribute. The patients of age between 65 – 85 years are of high risk not to recover from the COVID-19 pandemic, patients of age between 26 – 64 years are likely to recover while patients of age between 1 – 24 years are recovered quickly from COVID-19 pandemic. From the model, we concluded old patients are at high risk of developing COVID-19 complications which may result in death.

Performance Evaluation of Model

Data mining models are evaluated using evaluation techniques in order to determine their accuracy [14]. The techniques determine the quality and efficiency of the model using the data mining algorithm or machine learning algorithms. These main performance evaluation techniques for the data mining model include specificity, sensitivity, and accuracy. However, in this study, only accuracy is considered to evaluate the developed models.

Accuracy determines the percentage of the dataset instances correctly classified for the model developed by the data mining algorithm. Expressed as:

$$Accuracy = \frac{tp + tn}{tp + tn + fn + fp} \qquad (3)$$

Where tp = = True Positive, tn == true negative, fp = false positive while fn = false negative

The model developed with DT happened to be the most efficient with the highest percentage of accuracy of 99.85 %, followed by RF with 99.60 %, then SVM with 98.85 % accuracy, then K-NN with 98.06 % accuracy, then NB with 97.52 % accuracy and LR with 97.49 % accuracy. Table 3 has shown the results of the performance evaluation of the developed models.

Table 3: Performance Evaluation of Predictive Data Mining Models

S/N	Predictive Data Mining Models	Accuracy (%)
1	Decision Tree	99.85
2	Support Vector Machine	98.12
3	Naive Bayes	97.52
4	Logistic Regression	97.49
5	Random Forest	99.60
6	K-Nearest Neighbor	98.06

Findings and Discussions

The data mining algorithm which includes DT, SVM, NB, LR, RF, and K-NN were applied directly on the dataset using python programming language. However, the model developed with DT algorithm was found to be the most accurate with 99.85 % accuracy which appears to be the highest among others as shown in Figure 7 below:

The model predicted a minimum and maximum number of days for COVID-19 patients to recover from the virus. The model also predicted the age group of patients who are at high risk not to recover from the COVID-19 pandemic, those who are likely to recover and those who might be likely to recover quickly from COVID-19 pandemic. From the performance evaluation result of the models, the model developed with DT data mining algorithm is efficiently capable of predicting the possibility of recovery of infected patients from COVID-19 pandemic with the overall accuracy of 99.85 % when compared with RF, SVM, K-NN, NB and LR with the overall accuracy of 99.60 %, 98.85 %, 98.06 %, 97.52 %, and 97.49 % respectively.

Conclusion

In the present study, data mining models were developed for the prediction of COVID-19 infected patients' recovery using epidemiological dataset of COVID-19 patients of South Korea. DT, SVM, NB, LR, RF, and K-NN algorithms were applied directly on the dataset using python programming language. The model developed with DT was found to be the most efficient with the highest percentage of accuracy of 99.85%, followed by RF with 99.60% accuracy, then SVM with 98.85% accuracy, then K-NN with 98.06% accuracy, then NB with 97.52% accuracy and LR with 97.49% accuracy. The developed models would be very helpful in healthcare fo mthe the combat against COVID-19.

Declarations

Funding

No funding sources

Conflict of interest

Authors have declared that no conflict of interest exists.

References

[1] Al-Turaiki I, Alshahrani M, Almutairi T. Building predictive models for MERS-CoV infections using data mining techniques. Journal of Infection and Public Health 2016; 9: 744-748.

[2] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression (PDF). The American Statistician. 1992; 46:3,175–185.

[3] Coronavirus dataset **of** Korea Centers for Disease Control & Prevention (KCDC) retrieved from https://www.kaggle.com/kimjihoo/coronavirusdataset/data

[4] Everitt BS et al. Miscellaneous Clustering Methods in Cluster Analysis. 5th Edition, John Wiley & Sons, Ltd., Chichester, UK. 2011.

[5] Gandhi R. Naive Bayes Classifier, Towards Data Science. 2018. https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c accessed 25th April, 2020

[6] Hussain S, et al. Performance Evaluation of Various Data Mining Algorithms on Road Traffic Accident Dataset. In: Satapathy S, Joshi A. (eds) Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies. 2019; 106.

[7] Kohavi R, Quinlan R. Decision Tree Discovery, 1999.

[8] Li Y et al. A machine learning-based model for survival prediction in patients with severe COVID-19 infection medRxiv 2020.02.27.20028027; doi: https://doi.org/10.1101/2020.02.27.20028027 15th May, 2020

[9] Logistic Regression, NCSS Statistical Software, Chapter 231

[10] Mavroforakis ME, Theodoridis S. A geometric approach to Support Vector Machine (SVM) classification. IEEE Transactions on Neural Networks. 2006;17-3,671-682.

[11] Mehta M, Agrawal R, Risanen J. SLIQ: A Fast Scalable Classifier for Data Mining. In: Apers, P, Bouzeghoub M, Ardarin G. Proceedings of the 5th international conference on Extending Database Technology. Berlin: Pringer-Verlag, 1996; 18-32.

[12] Haque MR, Islam MM, Iqbal H, Reza MS, Hasan MK (2018) Performance Evaluation of Random Forests and Artificial Neural Networks for the Classification of Liver Disorder. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). IEEE, pp 1–5.

[13] Muhammad L.J et al. Performance Evaluation of Classification Data Mining Algorithms on Coronary Artery Disease Dataset: IEEE 9th International Conference on Computer and Knowledge Engineering (ICCKE 2019), Ferdowsi University of Mashhad. . 2019.

[14] Muhammad L.J et al. Performance Evaluation of Classification Data Mining Algorithms on Coronary Artery Disease Dataset: IEEE 9th International Conference on Computer and Knowledge Engineering (ICCKE 2019), Ferdowsi University of Mashhad 978-1-7281-5075-8/19/$31.00 ©2019 IEEE

[15] Muhammad LJ et al. Using Decision Tree Data Mining Algorithm to Predict Causes of Road Traffic Accidents, its Prone Locations and Time along Kano –Wudil Highway. International Journal of Database Theory and Application. 2017; 10:11,197-208

[16] Muhammad LJ, Usman SS. Power of Artificial Intelligence to Diagnose and Prevent Further COVID-19 Outbreak: A Short Communication.2020. ArXiv: 2004.12463 [cs.CY] accessed 15th May, 2020

[17] Onel H. Machine Learning Basics with the K-Nearest Neighbors Algorithm, Towards Data Science, 2018 https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761 accessed 26th April, 2020

[18] Raphael Dolin MD, Stanley Perlman MD, Novel Coronavirus from Wuhan China, 2019-20, Chapter 155, Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, Ninth Edition, Elsevier. 2020.

[19] Rong C, Lizhen X. Improved C4.5 Algorithm for the Analysis of Sales. Proceeding of the 6th Web Information System and Application Conference, IEEE. 2009; 173-179

[20] Rothe C. et al. Transmission of 2019-nCoV infection from an asymptomatic contact in Germany. N. Engl. J. Med. 2020; 382:970–971.

[21] Shang J et al. Structural basis of receptor recognition by SARS-CoV-2. Nature (2020).

[22] Ayon SI, Islam MM, Hossain MR (2020) Coronary Artery Heart Disease Prediction: A Comparative Study of Computational Intelligence Techniques. IETE J Res 1–20.

[23] Rahaman A, Islam M, Islam M, Sadi M, Nooruddin S (2019) Developing IoT Based Smart Health Monitoring Systems: A Review. Rev d'Intelligence Artif 33:435–440.

[24] Islam MM, Iqbal H, Haque MR, Hasan MK (2017) Prediction of breast cancer using support vector machine and K-Nearest neighbors. In: 2017 IEEE Region 10 Humanitarian Technology Conference (R10-

HTC). IEEE, pp 226–229.

[25] Wölfel R et al. Virological assessment of hospitalized patients with COVID-2019. Nature. 2020.

[26] Yahaya BZ et al. An Improved C4.5 Algorithm Using L' Hospital Rule for Large Dataset, Indian Journal of Science and Technology. 2018; 11:47.

[27] https://www.worldometers.info/coronavirus/ accessed 30th May, 2020

[28] Hyun K. The prevention and handling of the missing data. Korean J Anesthesiol. 2013; 64(5): 402–406.

[29] Islam, M.M., Rahaman, A. & Islam, M.R. Development of Smart Healthcare Monitoring System in IoT Environment. SN Comput. Sci. 1, 185 (2020).

[30] Hasan MK, Islam MM, Hashem MMA (2016) Mathematical model development to detect breast cancer using multigene genetic programming. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV). IEEE, pp 574–579

[31] Islam Ayon S, Milon Islam M (2019) Diabetes Prediction: A Deep Learning Approach. Int J Inf Eng Electron Bus 11:21–27.

[32] Hasan M, Islam MM, Zarif MII, Hashem MMA (2019) Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. Internet of Things 7:100059.

[33] Islam M, Neom N, Imtiaz M, Nooruddin S, Islam M, Islam M (2019) A Review on Fall Detection Systems Using Data from Smartphone Sensors. Ingénierie des systèmes d Inf 24:569–576.

[34] Nooruddin S, Islam MM, Sharna FA (2020) An IoT based device-type invariant fall detection system. Internet of Things 9:100130.

Figures

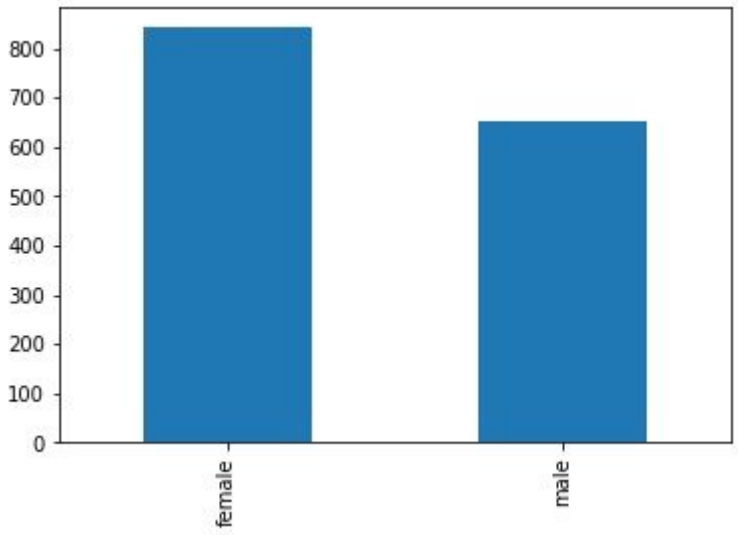

Figure 1

Frequency of sex attribute

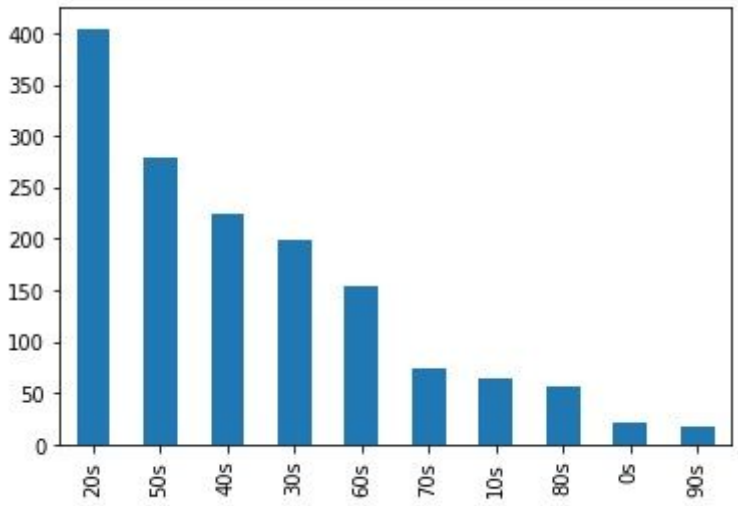

Figure 2

Frequency of age attribute

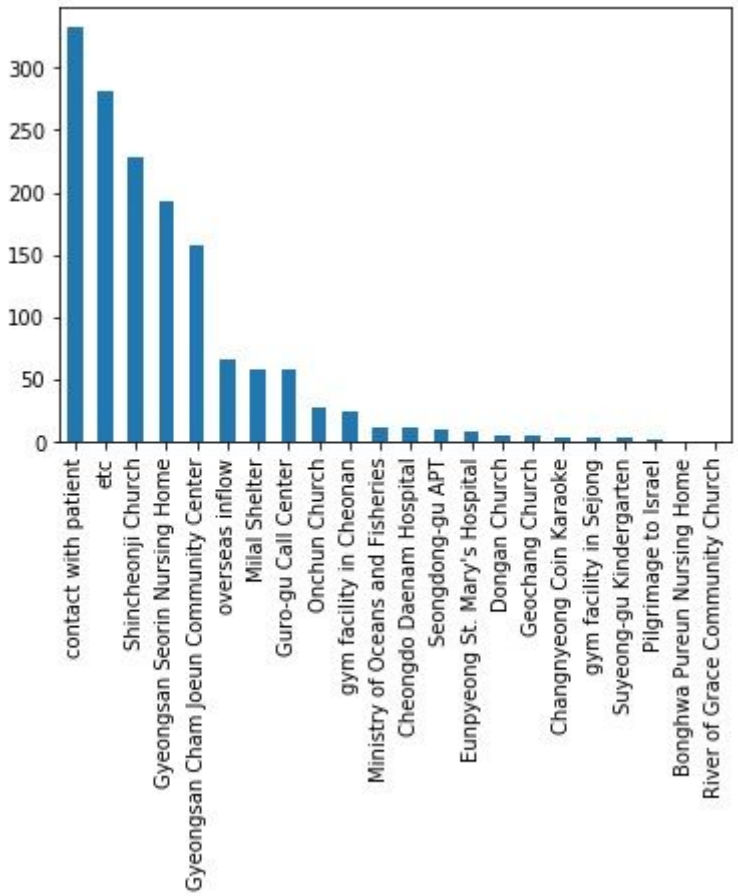

Figure 3

Frequency of infection_case attribute

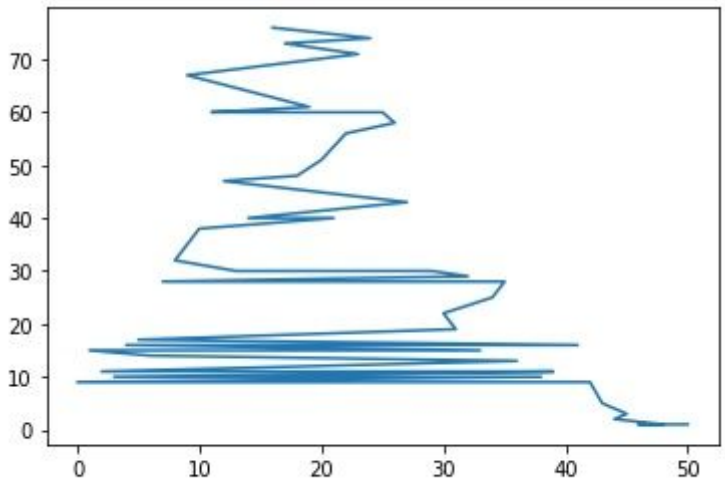

Figure 4

Frequency of no_days attribute

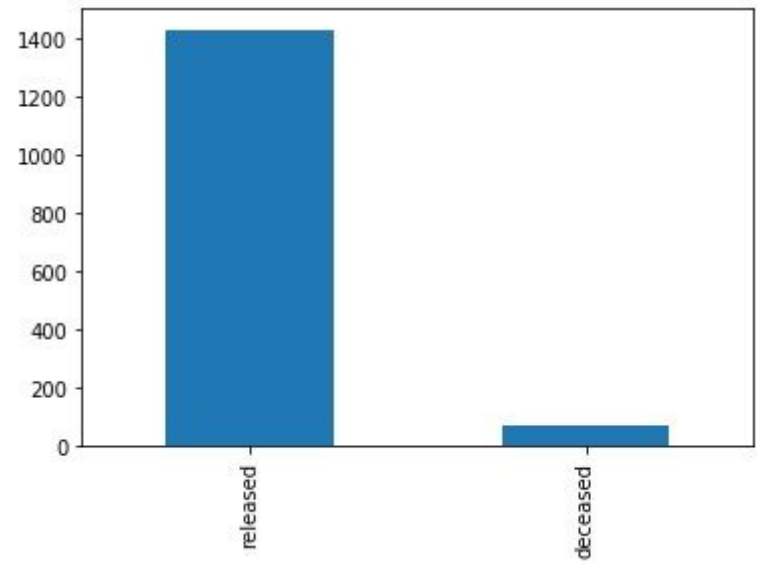

Figure 5

Frequency of state attribute

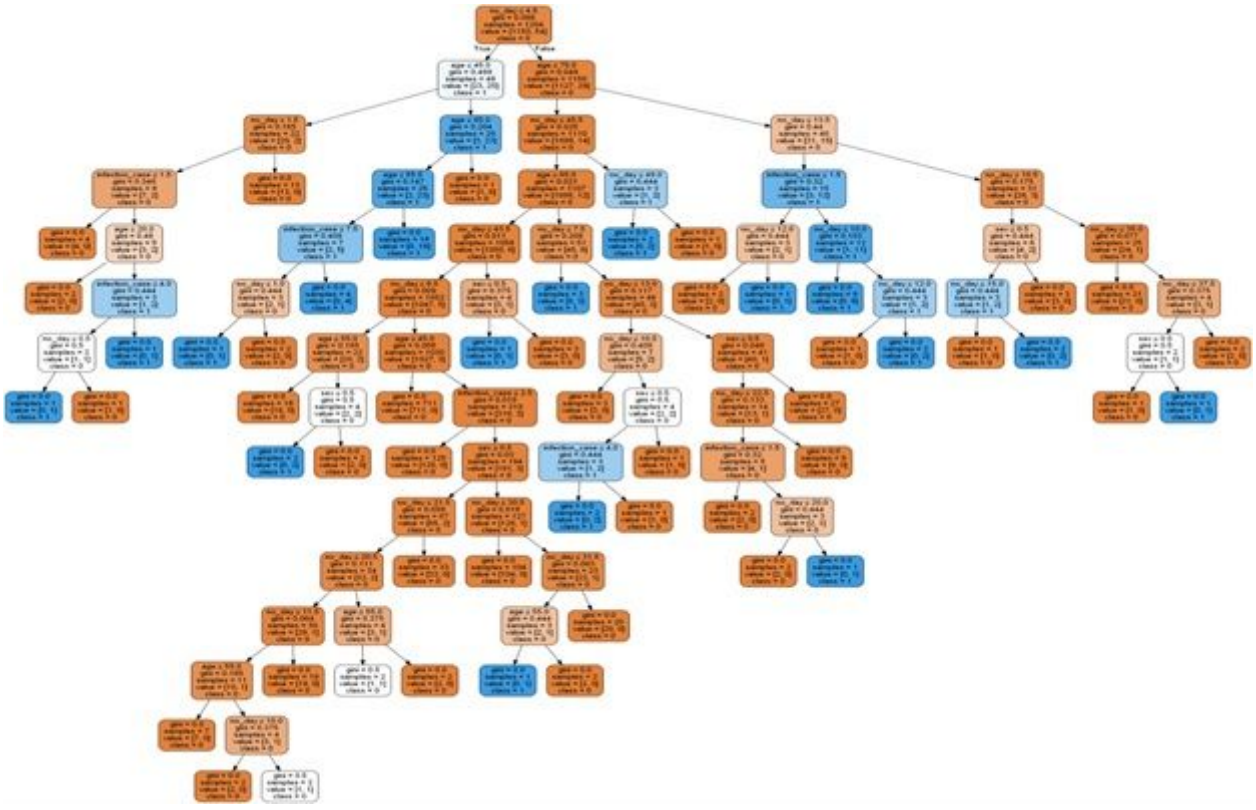

Figure 6

Decision Tree model for COVID-19 infected patients recovery

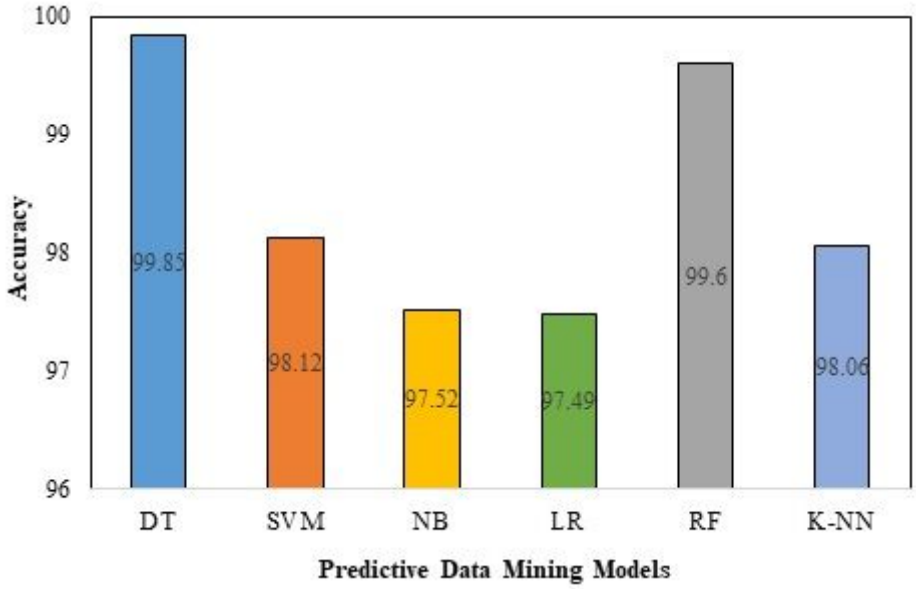

Figure 7

Performance evaluation results of the models