Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

An Efficient Framework for Classifying Novel Corona Cases based on Machine Learning and Data Mining Techniques

Baida Abdulredha Hamdan
Department of Computer Science, College of Education for Pure Sciences, University of Thi-Qar, Iraq.
baida.alkinza66@utq.edu.iq

DOI: 10.5281/zenodo.7004517

ABSTRACT

Hospitals with the disease of coronavirus (COVID) are always at risk of dying. COVID hospitalized patients may benefit from the application of machine learning and data mining techniques to predict their death. Therefore, the purpose of this paper was to evaluate many machine learning and data mining algorithms to COVID mortality prediction utilizing patient data at the time of first admission and select the algorithm that performs best as a decision-making tool. Its signs and symptoms resemble those of the regular flu consisting fever, cough, shortness of breath, fatigue, and muscle pain. This paper proposed running the most important algorithms from data mining and machine learning such as naïve bayes, models of decision tree (ID3 and C4.5), support vector machine, and logistic regression to classify corona cases. To test the proposed framework, the confusion matrix has been used. From the confusion matrix the important performance measures have been computed such as accuracy, recall, precision, balance accuracy, and AUC. The experimental findings of this paper supported the notion that the supported vector machine algorithm had good performance and high accuracy in classifying corona disease.

Keywords: COVID, data mining, machine learning, confusion matrix.

Cite as: Baida Abdulredha Hamdan (2022). An Efficient Framework for Classifying Novel Corona Cases based on Machine Learning and Data Mining Techniques. *LC International Journal of STEM (ISSN: 2708-7123), 3*(2), 213-216. DOI: 10.5281/zenodo.7004517

INTRODUCTION

On December 8, 2019, the Chinese government revealed that 41 patients in Wuhan were hospitalized with an unexplained etiology, including one patient who passed away [1]. The new coronavirus epidemic respiratory disease was started by this cluster. Although the wet market was associated with the earliest cases, human-to-human transmission has caused a widespread outbreak of the virus across the country. The World Health Organization (WHO) classified corona as a public health emergency on January 30, 2020. (PHEIC) [2].

Considering the disease's widespread reach and severity, the World Health Organization's Director-General proclaimed the outbreak of corona pandemic on March 11, 2020. With the pandemic's rapid expansion outside of China, it entered a new phase [3].

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Healthcare systems around the world are trying to apply classifiers of data mining in response to the above-mentioned issues for obtaining appropriate decision-making by removing doctors' subjective assessments. Machine learning is a subfield of artificial intelligence (AI) that enables the extraction of useful big datasets for high-quality prediction models. It is an important instrument that is used ever more in medical research to enhance predictive modelling and identify fresh causes of a particular target outcome. By providing evidence-based medicine, machine learning algorithms can eliminate ambiguity and uncertainty for screening, risk analysis, care plans, and prediction; it encourages trustworthy clinical judgment and aspires to raise patient outcomes and level of treatment [4].

Classifier models for the prediction of corona mortality were created. The relevant papers in this field were studied in order to choose the best machine learning algorithms during the modelling stage [5, 6, 7, 8, 9, 10, and 11] in addition to taking into account the chosen dataset's kind and quality. Figure 1 offer a block diagram of AI to predict the corona.

The rest of this study as follows: section two shows the literature survey of the proposed system or framework that contained theoretical background of data mining, machine learning, SVM, NB, ID3, C4.5, and LR. Section tree contain the brief discussion of the proposed simple and reliable block diagram for the framework. Section four contains the results and analysis also the confusion matrix of the techniques that have been used. Finally section five contain the conclusions.

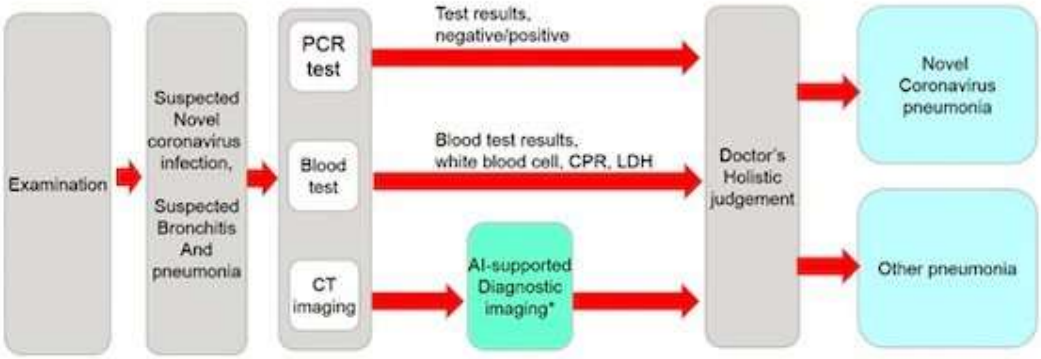


Figure 1: AI help in forecasting of Corona

LITERATURE REVIEW

Machine Learning and Data Mining

Data mining is the process of sorting through large data sets to uncover patterns and relationships that can be applied to data analysis to help in find solution of challenges for a business. Businesses can make better business decisions by employing techniques and tools of data mining [12].

Data mining is a fundamental data science subject and critical component of data analytics, it searches through data collections for relevant information using modern analytics techniques. To put it another way, the data mining process is a step in the Knowledge Discovery in Database (KDD) process, a strategy for collecting, processing, and interpreting data using data science. It's customary to use the terms data mining and KDD interchangeably, yet they're usually seen as distinct concepts [13].

Machine learning is an area of research focused on developing and comprehending techniques that 'learn' that is, approaches that make use of data to enhance assessment in a certain set of circumstances.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Considered to be a component of artificial intelligence. Sample data, referred to as training data, and is used by machine learning algorithms to develop a model, without being specifically programmed to do so, in order to make decisions or predictions. A wide range of applications make use of machine learning methods, such as in email filtering, medicine, computer vision, and speech recognition in cases where standard algorithms are insufficient or infeasible [14, 15].
Computational statistics and a subset of machine learning are closely linked concepts, which is concerned with utilizing computers to make predictions, however not all machine learning can be considered statistical. The discipline of machine learning benefits from the methodologies, theories, and application fields that mathematical optimization research provides. Exploratory data analysis and unsupervised learning are two linked fields of research. Some systems of machine learning replicate the functioning of a brain's biological by using data and neural networks. When used to address business problems, machine learning is often referred to as predictive analytics [16, 17].

Support Vector Machine

A supervised machine learning approach called Support Vector Machine (SVM) is applied for both regression and classification. It's best for classification, even if we call regression problems by this name. Finding a hyper-plane in a space of N-dimensional. The SVM approach aims to classify the data points in a transparent manner. According to the number of features, the hyper-dimensions planes can be calculated. When there are only two input characteristics, the hyper-plane is just a line. The hyper-plane transforms into a two-dimensional plane when three input features are used. It becomes difficult to imagine something having more than three features [18].
Let's consider one dependent variable (either a blue or red circle) and 2 independent variables x2, x1. Figure 2 makes it quite evident that there are several lines (For the sake of simplicity, we're merely taking into account the two input properties listed above (x1, x2). Our data points into red and blue circles, or performs a categorization [19].

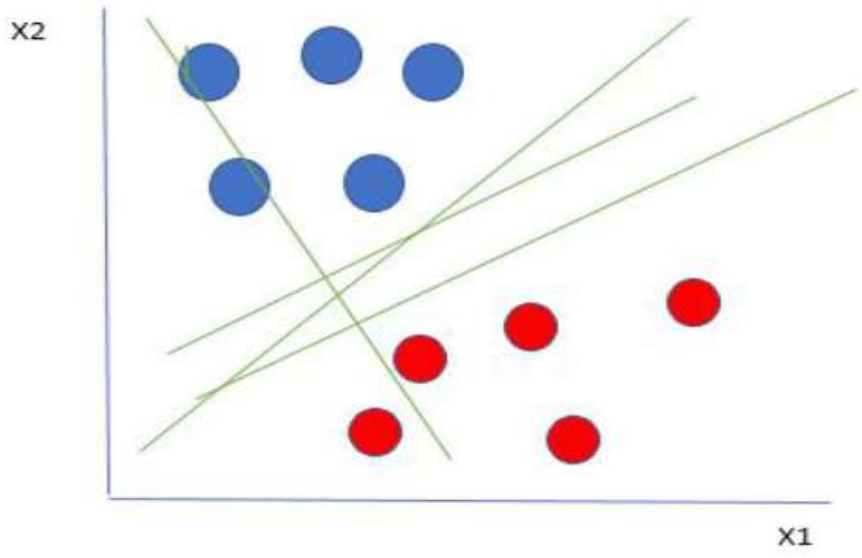


Figure 2. Linearly Separable Data points

Naïve Bayes

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Depend on the Bayes Theorem, Nave Bayes is a probabilistic machine learning method, used to classify a wide range of things. We shall comprehend the Naive Bayes algorithm and the crucial ideas in Figure 3 so that there are no ambiguities [20]. Figure 3 offers the flow chart of NB.

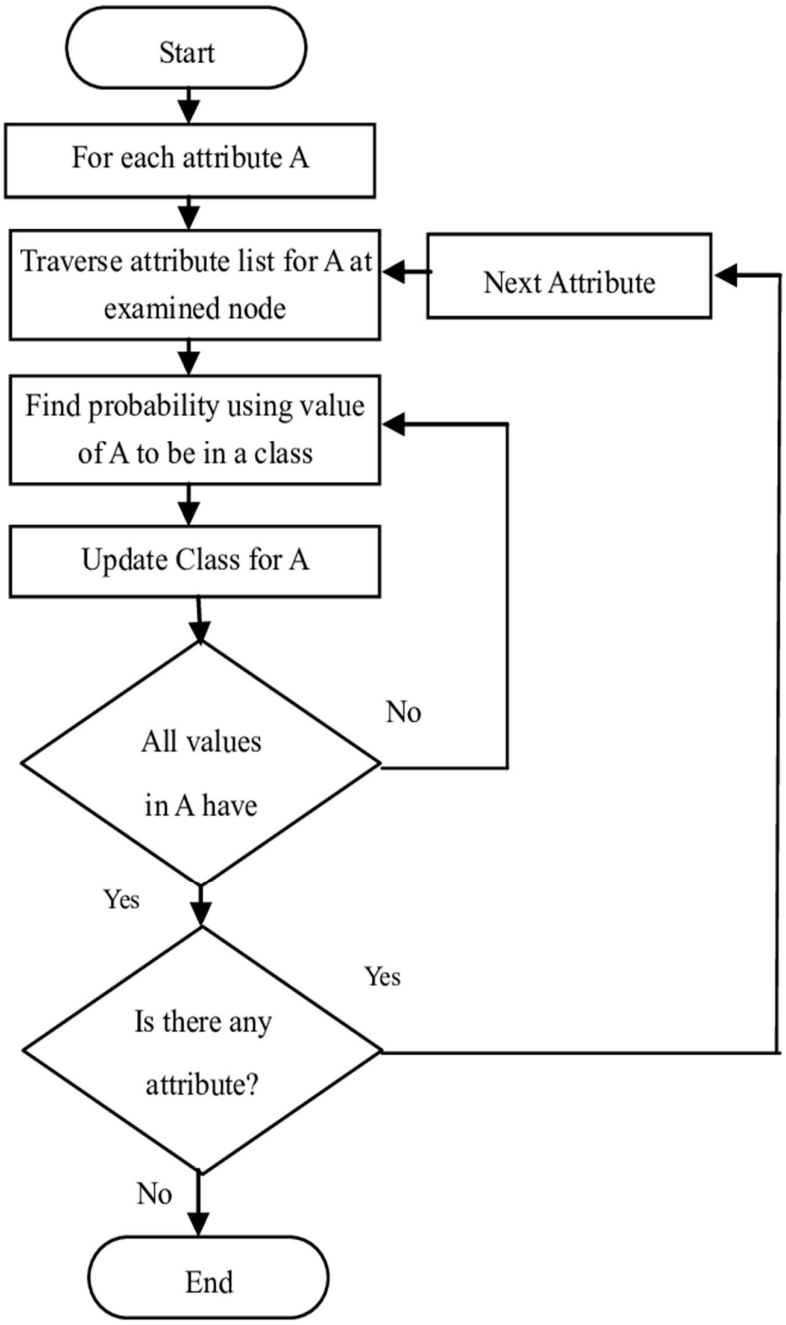


Figure 3. Flowchart of Naïve Bayes

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

ID3 and C4.5

Iterative Dichotomiser 3 (ID3) is the acronym for the ID3 algorithm, is a classification algorithm that builds a decision tree by picking the best attribute in a greedy manner that generates minimum Entropy (H) or maximum Information Gain (IG) [21].

Quinlan Ross added C4.5 as an enhancement to ID3. Additionally, it utilizes Hunt's algorithm. To construct a decision tree, C4.5 utilizes both categorical and continuous features. C4.5 divides the attribute values into two parts to handle continuous attributes the values that are higher than the threshold are treated as a single leaf, while the values below the threshold are treated as separate leafs. It addresses missing attribute values as well. In order to generate a decision tree, C4.5 uses the Gain Ratio as an attribute selection metric. When there are numerous result values for a single attribute, it eliminates the biasness of information gain. To begin, figure out the gain ratio for each of each attribute's properties. The characteristic with the highest gain ratio will be the root node. Pessimistic pruning is used by C4.5 to remove nonessential decision tree branches in order to enhance classification accuracy. Figure 4 offers the model of processing for Decision Tree (DT) [22].

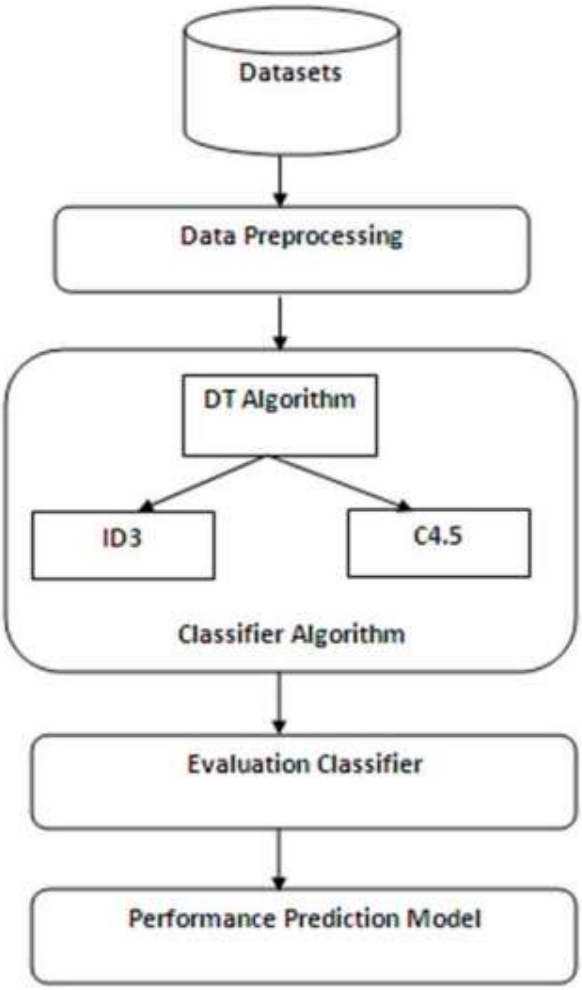


Figure 4. Model of Processing

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Linear Regression

Modeling the relationship between two continuous variables using simple linear regression. An output variable (or response value)'s can often be predicted using the value of an input variable.

We frequently want to understand how different variables are related to one another. There are several ways to look at possible correlations between two variables applying measures of scatterplot. Correlation measures the linear connection between two variables, however, correlation does not indicate more complicated interactions [23]. Figure 5 offers three scatterplots.

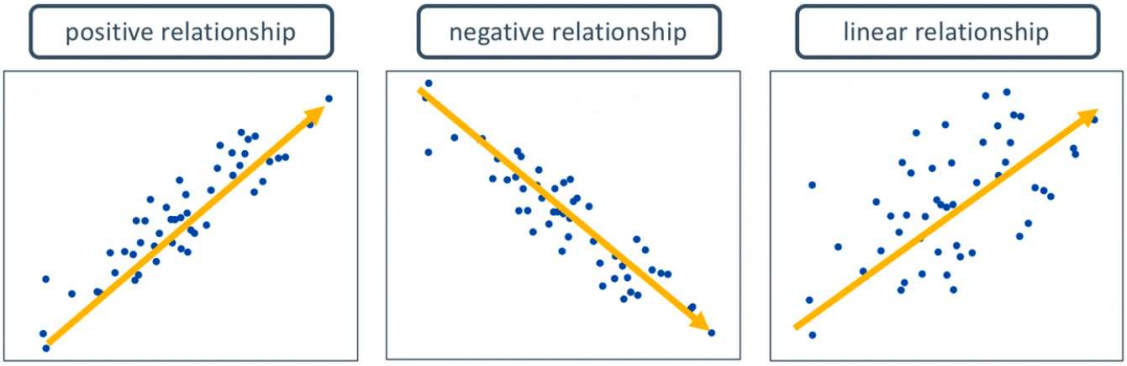


Figure 5. Three Scatterplots

METHODOLOGY

Based on machine learning and data mining, we hope to improve the predictive model by using SVM, NB, ID3, C4.5, and LR to forecast whether a case will be negative or positive. Using pre-processing, we were able to improve classification accuracy while simultaneously reducing processing time. Finding a framework technique to use as a guide is crucial for achieving this goal, understanding and collecting new corona data are two essential steps in this process, new corona data preparation and preprocessing, modeling & experiments, testing & evaluating. Figure 6 depicts the method a step-by-step: Prior to that, we gather data from reference [24, 25] and clean the data. As a result of using preprocessing techniques, we are able to clean up the data, as well as continuing to train those algorithms for each patient in a data collection, the model used ML and DM algorithms to forecast what the state is; Positive status indicates the patient has been infected with a new strain of coronavirus, If the patient does not have new corona infection, the patient's status will be negative. Finally, we looked at the model that was proposed and we are tested it by using measures of performance such as accuracy, Recall, balance accuracy, precision, and AUC.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

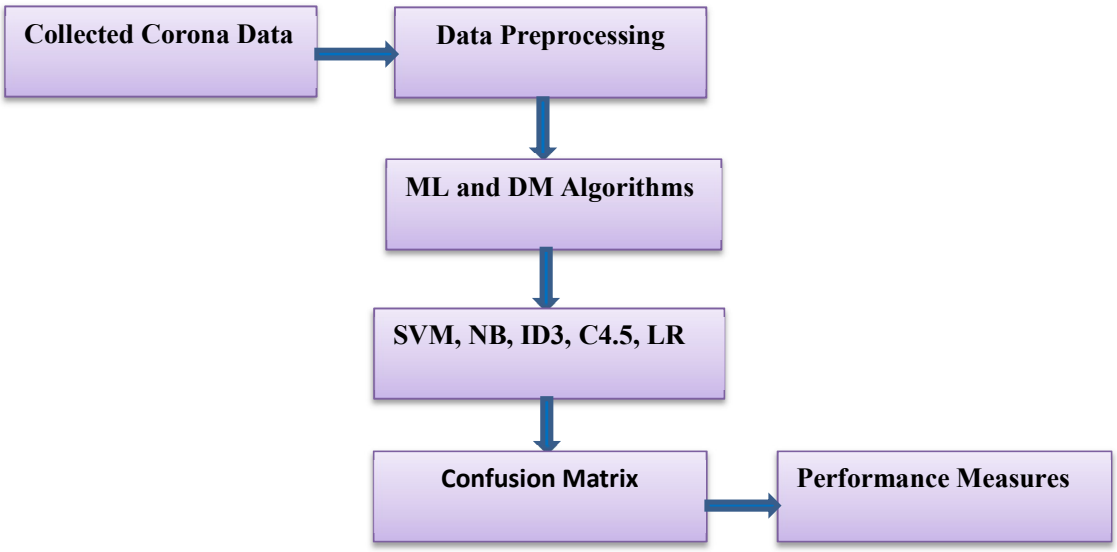


Figure 6. Block Diagram of the Proposed Framework

DATA ANALYSIS AND RESULTS

Python v3.8.0, RAM: 8GB, does a performance evaluation of machine learning and data mining algorithms in this part, and implementing on OS Build 19041.488 (Windows 10 Home). The optimal method for interpreting these novel corona data is also determined by comparing different algorithms, make sure they can accurately identify the patient's condition, and identify the technique whose accuracy is inappropriate for this data's analysis. The performance of machine learning and data mining techniques is estimated using the confusion matrix, which has been offered in Figure 7. From the confusion matrix, we can calculated classical metrics such as accuracy, balance accuracy, recall, and precision respectively. Table 1 and Figure 7 show the results of SVM, NB, ID3, C4.5, and LR in comparison to each other.

Confusion matric contains true negative, true positive, false positive, and false negative. AUC is a measurement that takes false-positive rate (FPR) and true-positive rate (TPR) into account (TPR). It assesses a classification model's accuracy throughout the entire range of categorization thresholds. A higher AUC indicates that the classifier is better able to distinguish across classes, this may be used to discern between different forms of data and assess whether or not a corona is present or absent in a given scenario (positive or negative).

Tables 2,3,4,5 and 6 and Figures 8, 9, 10, 11 and 12 offers the confusion matrix of all models that are used in predicting the case of the patient. From the results we are conclude that SVM is outperform all the models in analyzing data of novel corona, which is achieved accuracy = 87% while NB achieved 72.3%, ID3 achieved 74.1%, C4.5 achieved 79.4% and LR achieved 80.1%.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

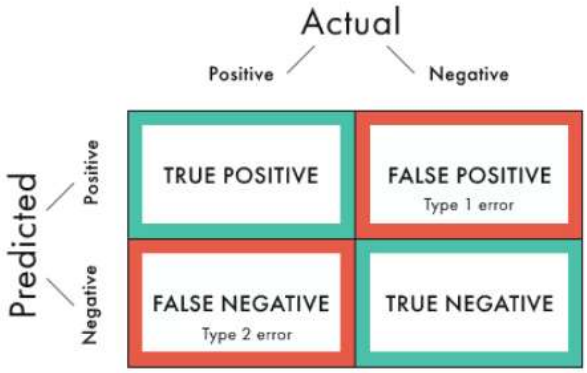


Figure 7. Confusion Matrix

Table 1. Performance Measures

Technique	Accuracy%	Balance Accuracy%	Recall%	Precision%	AUC%
SVM	87	85	86.9	86	97.1
NB	72.3	70	71.2	73	92.2
ID3	74.1	73.4	73.9	75.9	93
C4.5	79.4	81	80.2	80.9	95.4
LR	80.1	82.4	82	84	96.7

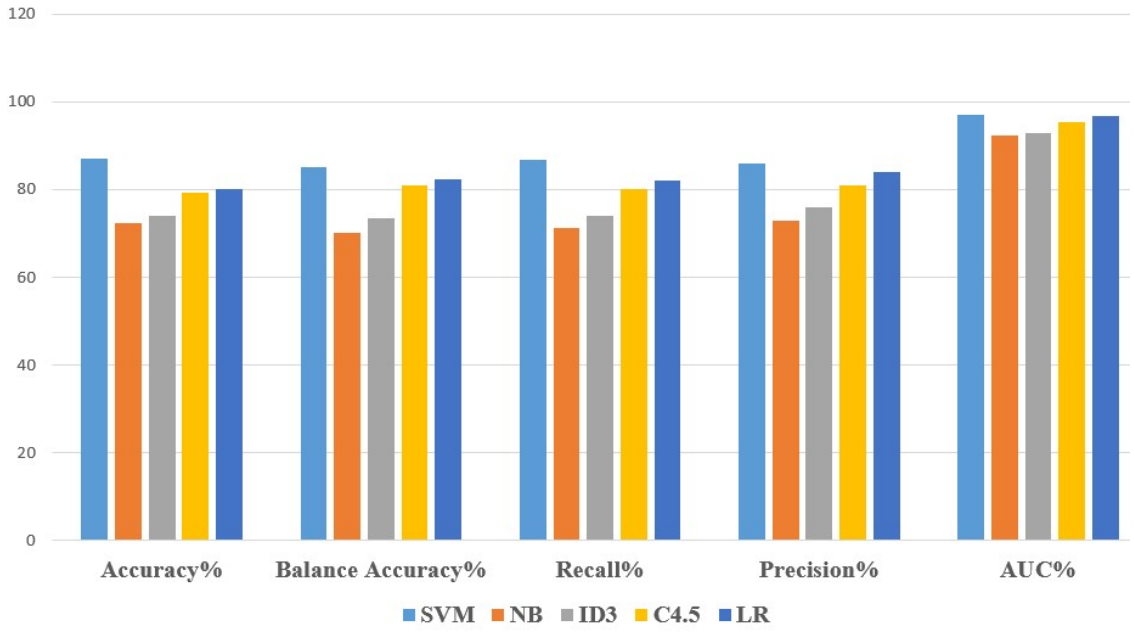


Figure 8. Performance Measures

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Table 2. Confusion Matrix of SVM

	Predicted Class		
Actual Class		Positive	Negative
	Positive	0.59	0.29
	Negative	0.24	0.33

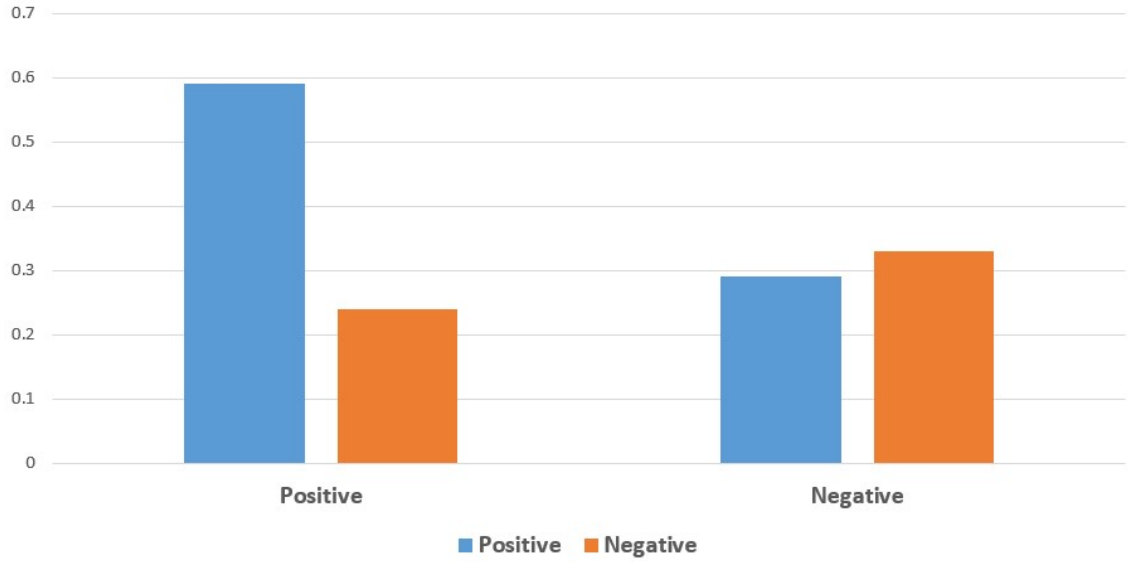


Figure 9. Confusion Matrix of SVM

Table 3. Confusion Matrix of NB

	Predicted Class		
Actual Class		Positive	Negative
	Positive	0.44	0.34
	Negative	0.29	0.32

Table 4. Confusion Matrix of ID3

	Predicted Class		
Actual Class		Positive	Negative
	Positive	0.49	0.22
	Negative	0.32	0.34

Table 5.Condusion Matrix of C4.5

	Predicted Class		
Actual Class		Positive	Negative
	Positive	0.52	0.27
	Negative	0.23	0.30

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Table 6. Confusion Matrix of LR

	Predicted Class		
Actual Class		Positive	Negative
	Positive	0.44	0.25
	Negative	0.24	0.27

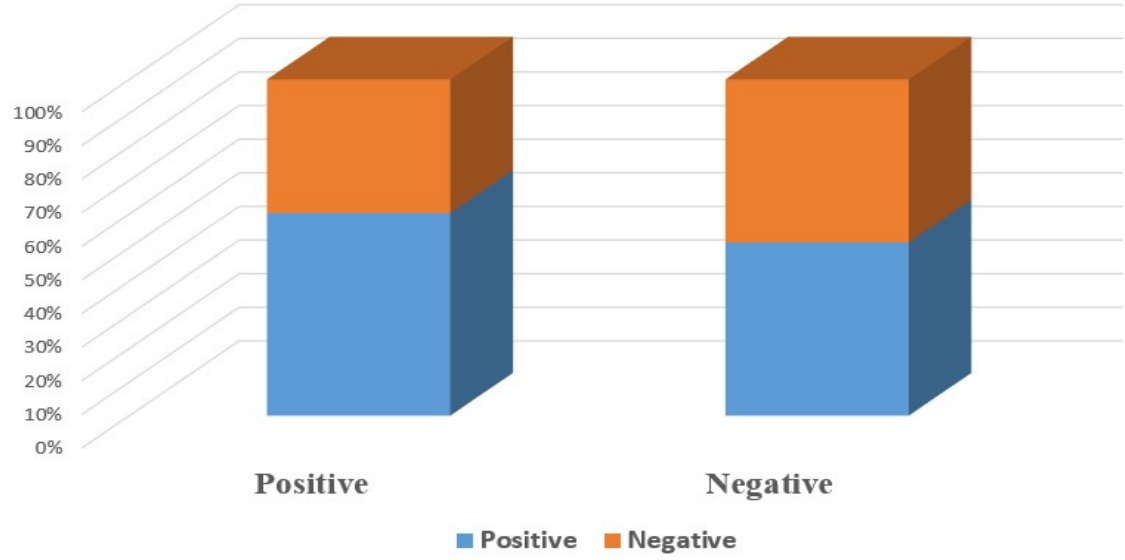
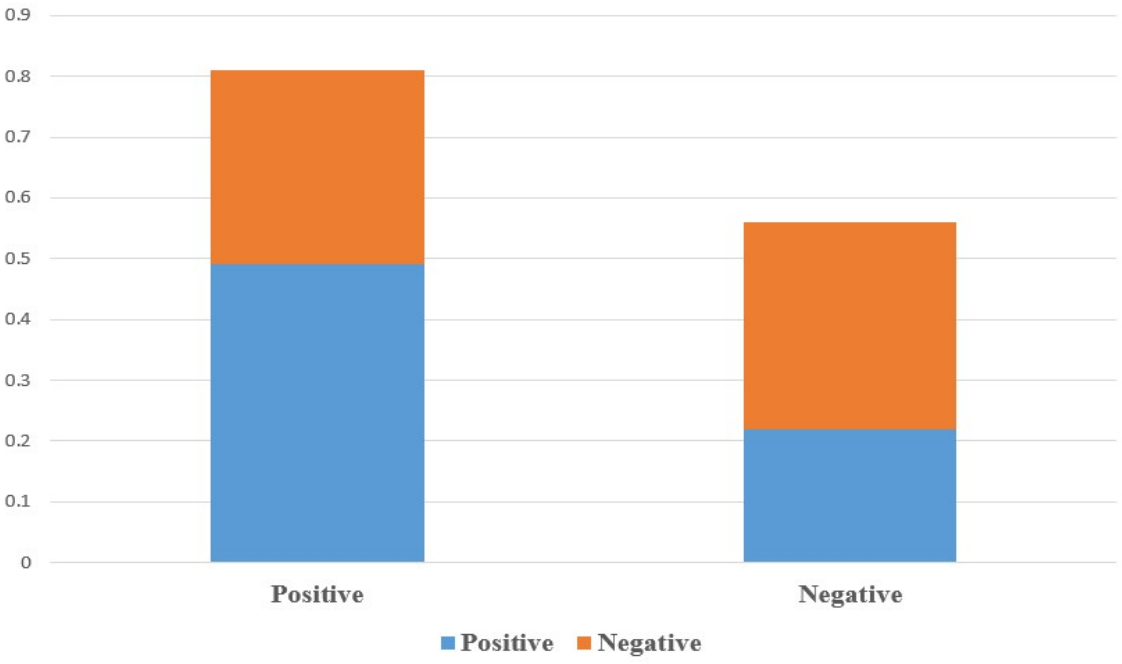


Figure 10. Confusion Matrix of NB



Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

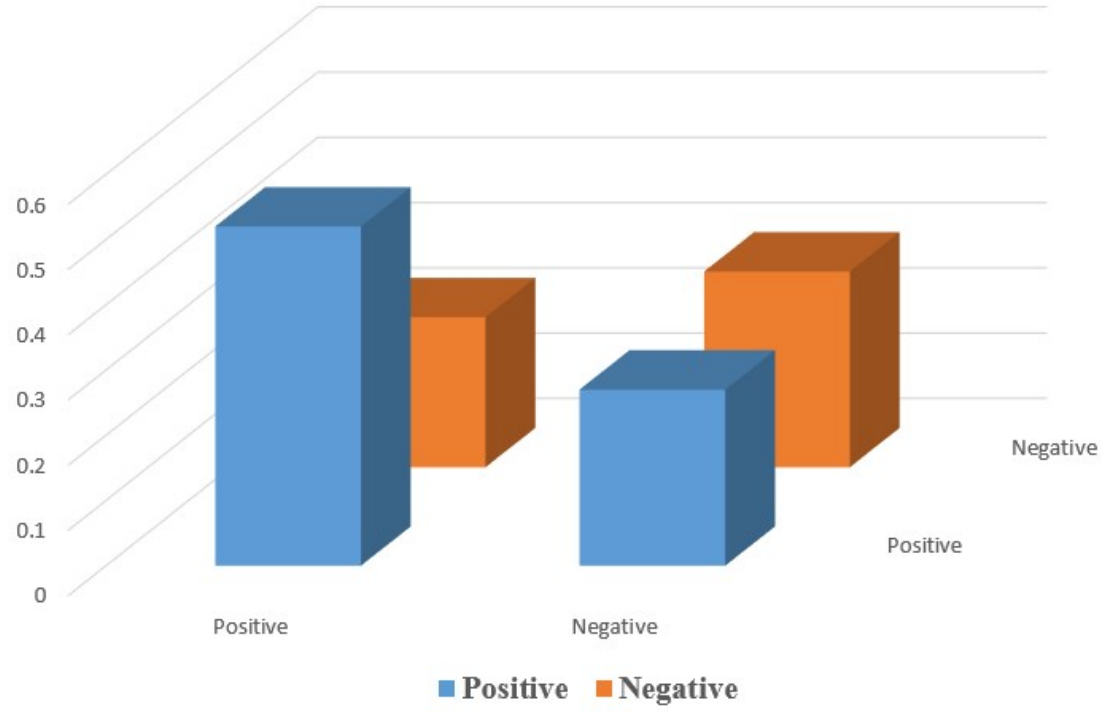


Figure 12. Confusion Matrix of C4.5

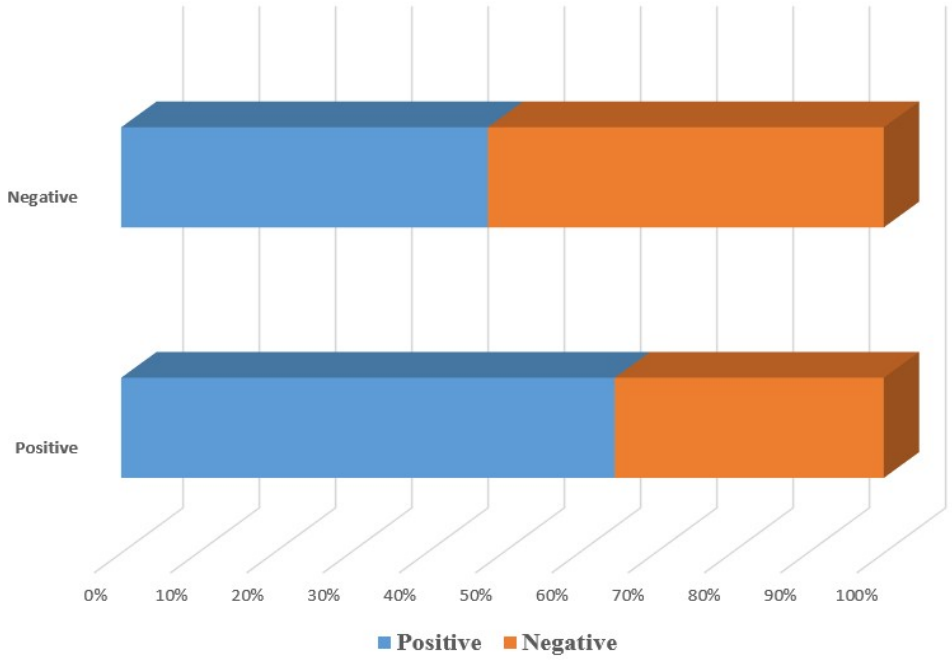


Figure 13. Confusion Matrix of LR

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

CONCLUSION

In this study, a set of corona data from people is classified using machine learning approaches. with positive or negative. The SVM approach has been proven to be the most advantageous in terms of accuracy for corona detection after a number of experiments. The NB algorithm performs fairly well. The worst performance is that of the NB algorithm despite it gives satisfactory and convincing findings in this paper. Because accuracy and AUC values are crucial for assessing the degree of implementation of machine learning classifiers, they are the primary focus of this work among the two targets of corona cases (positive and negative).

REFERENCES

[1] Mosa, A. M., Hamed, E. A., Hussein, Z., & Jaleel, R. A. (2022, April). Improved Smart Forecasting Model to Combat Coronavirus using Machine Learning. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 1953-1957). IEEE.
[2] Fadhil, Z. M., & Jaleel, R. A. (2022, April). Multiple Efficient Data Mining Algorithms with Genetic Selection for Prediction of SARS-CoV2. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2016-2020). IEEE.
[3] Adday, B. N., Shaban, F. A. J., Jawad, M. R., Jaleel, R. A., & Zahra, M. M. A. (2021, October). Enhanced Vaccine Recommender System to prevent COVID-19 based on Clustering and Classification. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)* (pp. 1-6). IEEE.
[4] Jawad, M. R., Qasim, M. A., Cazzato, G., Zahra, M. M. A., Kapula, P. R., Gherabi, N., & Jaleel, R. A. (2021). Advancement of artificial intelligence techniques based lexicon emotion analysis for vaccine of COVID-19. *Periodicals of Engineering and Natural Sciences (PEN)*, *9*(4), 580-588.
[5]Zhao, Z., Chen, A., Hou, W., Graham, J. M., Li, H., Richman, P. S., ... & Duong, T. Q. (2020). Prediction model and risk scores of ICU admission and mortality in COVID-19. *PloS one*, *15*(7), e0236618.
[6] Hu, H., Yao, N., & Qiu, Y. (2020). Comparing rapid scoring systems in mortality prediction of critically ill patients with novel coronavirus disease. *Academic Emergency Medicine*, *27*(6), 461-468.
[7] Ryan, L., Lam, C., Mataraso, S., Allen, A., Green-Saxena, A., Pellegrini, E., ... & Das, R. (2020). Mortality prediction model for the triage of COVID-19, pneumonia, and mechanically ventilated ICU patients: A retrospective study. *Annals of Medicine and Surgery*, *59*, 207-216.
[8] Yan, L., Zhang, H. T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., ... & Yuan, Y. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nature machine intelligence*, *2*(5), 283-288.
[9] Gao, Y., Cai, G. Y., Fang, W., Li, H. Y., Wang, S. Y., Chen, L., ... & Gao, Q. L. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature communications*, *11*(1), 1-10.
[10] Booth, A. L., Abels, E., & McCaffrey, P. (2021). Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Modern Pathology*, *34*(3), 522-531.
[11] Moulaei, K., Ghasemian, F., Bahaadinbeigy, K., Sarbi, R. E., & Taghiabad, Z. M. (2021). Predicting mortality of COVID-19 patients based on data mining techniques. *Journal of Biomedical Physics & Engineering*, *11*(5), 653.
[12] Abed, A. S., Hassan, H. F., Aldulaimi, M. H., Zahra, M. M. A., & Jaleel, R. A. (2022, April). An Effective Framework for Enhancing Performance of Internet of Things using Ant Colony Meta-

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

Heuristic and Machine Learning Algorithms. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2498-2502). IEEE.
[13] Albahri, A. S., Hamid, R. A., Al-qays, Z. T., Zaidan, A. A., Zaidan, B. B., Albahri, A. O., ... & Madhloom, H. T. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *Journal of medical systems*, *44*(7), 1-11.
[14] De Felice, F., & Polimeni, A. (2020). Coronavirus disease (COVID-19): a machine learning bibliometric analysis. *in vivo*, *34*(3 suppl), 1613-1617.
[15] Iqbal, S. Z., & Saghar, K. (2020). Improving Software Cost Estimation With Function Points Analysis Using Fuzzy Logic Method. *LC International Journal of STEM (ISSN: 2708-7123)*, *1*(1), 10-19.
[16] Asif, S., Ambreen, M., Muhammad, Z., ur Rahman, H., & Iqbal, S. Z. (2022). Cloud Computing in Healthcare-Investigation of Threats, Vulnerabilities, Future Challenges and Counter Measure. *LC International Journal of STEM (ISSN: 2708-7123)*, *3*(1), 63-74.
[17] De Felice, F., & Polimeni, A. (2020). Coronavirus disease (COVID-19): a machine learning bibliometric analysis. *in vivo*, *34*(3 suppl), 1613-1617.
[18] Le, D. N., Parvathy, V. S., Gupta, D., Khanna, A., Rodrigues, J. J., & Shankar, K. (2021). IoT enabled depthwise separable convolution neural network with deep support vector machine for COVID-19 diagnosis and classification. *International journal of machine learning and cybernetics*, *12*(11), 3235-3248.
[19] Guhathakurata, S., Kundu, S., Chakraborty, A., & Banerjee, J. S. (2021). A novel approach to predict COVID-19 using support vector machine. In *Data Science for COVID-19* (pp. 351-364). Academic Press.
[20] Mansour, N. A., Saleh, A. I., Badawy, M., & Ali, H. A. (2022). Accurate detection of Covid-19 patients based on Feature Correlated Naïve Bayes (FCNB) classification strategy. *Journal of ambient intelligence and humanized computing*, *13*(1), 41-73.
[21] Fadli, A., Nugraha, A. W. W., Aliim, M. S., Taryana, A., Kurniawan, Y. I., & Purnomo, W. H. (2020, December). Simple correlation between weather and COVID-19 pandemic using data mining algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 982, No. 1, p. 012015). IOP Publishing.
[22] Ramadhan, A. (2022). Sistem Pendukung Keputusan Evaluasi Problematika Pendampingan Pembelajaran Daring dengan Algoritma C4. 5. *Jurnal Sistim Informasi dan Teknologi*, 58-63.
[23] Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *14*(5), 1467-1474.
[24] Ali, N. G., Abed, S. D., Shaban, F. A. J., Tongkachok, K., Ray, S., & Jaleel, R. A. (2021). Hybrid of K-Means and partitioning around medoids for predicting COVID-19 cases: Iraq case study. *Periodicals of Engineering and Natural Sciences (PEN)*, *9*(4), 569-579.
[25] Jaleel, R. A., Burhan, I. M., & Jalookh, A. M. (2021, June). A Proposed Model for Prediction of COVID-19 Depend on K-Nearest Neighbors Classifier: Iraq Case Study. In *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (pp. 1-6). IEEE.

Logical Creations Education Research Institute
LC INTERNATIONAL JOURNAL OF STEM
E-ISSN: 2708-7123
Web: www.lcjstem.com | Email: editor@lcjstem.com
Volume-03 | Issue-02 | June-2022

BIOGRAPHY

Baida Abdulredha Hamdan received the B.S. and M.S. degrees from Thi-Qar University of Iraq and Ferdowsi University of Mashhad in 2005 and 2018 respectively. Currently, she is a lecturer in Thi-Qar University. She is current research interests include deep metric learning, and deep neural networks.