Global Journal of Computer Science and Technology: B
Cloud and Distributed
Volume 22 Issue 1 Version 1.0 Year 2022
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 | Print ISSN: 0975-4350 |

The State-of-the-Art Machine Learning in Prediction Covid-19 Fatality Cases

By Maleerat Maliyaem & Nguyen Minh Tuan
King Mongkut's University

Abstract- Day by day, the number of confirmed Covid-19 cases significantly increases all over the world. In India, the second wave of coronavirus has come back and created a disastrous impact. On April 3rd, India continuously recorded the highest number of daily cases globally, according to Financial Times, there was a scarcity of crematoriums and burial grounds due to the high number of corpses. The outbreak of death cases was an unprecedented circumstance, hence, there was a shortage of medical necessities. Prediction of death cases could help the government to manage the medical facilities such as beds and oxygen supply for the hospital. Machine learning could be used to analyze and predict fatality cases. PySpark library is used to process raw data and update new data each day, as the library allows the processing of a large amount of raw data efficiently. By using the Naïve Bayes algorithm available in PySpark, the prediction accuracy has increased to 81.3%.

Keywords: *pyspark, machine learning, Naïve bayes, fatality, covid-19.*

GJCST- B Classification: *H.1.2*

Strictly as per the compliance and regulations of:

The State-of-the-Art Machine Learning in Prediction Covid-19 Fatality Cases

Maleerat Maliyaem [α] & Nguyen Minh Tuan [σ]

Abstract- Day by day, the number of confirmed Covid-19 cases significantly increases all over the world. In India, the second wave of coronavirus has come back and created a disastrous impact. On April 3rd, India continuously recorded the highest number of daily cases globally, according to Financial Times, there was a scarcity of crematoriums and burial grounds due to the high number of corpses. The outbreak of death cases was an unprecedented circumstance, hence, there was a shortage of medical necessities. Prediction of death cases could help the government to manage the medical facilities such as beds and oxygen supply for the hospital. Machine learning could be used to analyze and predict fatality cases. PySpark library is used to process raw data and update new data each day, as the library allows the processing of a large amount of raw data efficiently. By using the Naïve Bayes algorithm available in PySpark, the prediction accuracy has increased to 81.3%.

Keywords: pyspark, machine learning, Naïve bayes, fatality, covid-19.

I. Introduction

The world has spent to the heart-rending day when fatalities passed 4 million people while the crisis becoming the race between vaccinating and new dangerous variants. Prediction is another way to control the Covid-19 situation and propose a new method to face the new stage of devastation coronavirus [18{21]. In paper [1], they used linear regression and polynomial regression to predict the results of fatalities. These two algorithms were applied to find the best fit line to estimate the average values of the two variables. These algorithms are dependent on the variation and dispersion of the data. The best fit line will divide the data into two parts with the same distance between the values of data from the best fit line. They also used root mean square error to estimate the accuracy of prediction. Root mean square error is a kind of metric to calculate the error when analyzing the data using regression algorithms. Root mean the square error will be calculated as the mean of the values and ensure the distances are the same as the points. The root means square error measures the variation and the concentration of the values around the mean. Many kinds of data could be expressed in Fig 1, the exactness belongs to the distribution of data.

In paper [2], they predicted the outbreak of Covid-19 in Ethiopia by comparing the Support Vector Machine (SVM) model and the Polynomial Regression (PR) model in the ScikitLearn library. The paper showed that SVM gets better performance than PR banked on evaluating graph performance and metric Mean Square Error (MSE), Mean Absolute Error (MAE) [3{7, 9]. With the same evaluation in paper [1], the results were also depending on the distribution of the data and this evaluation is just counted on the mean of the values that if the data is dense on the prediction, the mean of the values will be closed to the mean of prediction. This calculation usually makes the approximate values instead of exact values.

Table 1: Random Correlation Example for Data

	Total_ cases	new_ cases	new_cases _smoothed	total_ deaths	new_ deaths	new_deaths _smoothed
Total_cases	1.000000	0.862925	0.879309	0.984957	0.845896	0.871108
new_cases	0.862925	1.000000	0.989568	0.864804	0.928043	0.922668
_cases_smoothed	0.879309	0.989568	1.000000	0.879013	0.921921	0.937985
total_deaths	0.984957	0.864804	0.879013	1.000000	0.872722	0.898039
new_deaths	0.845896	0.928043	0.921921	0.872722	1.000000	0.976566
new_deaths_smoothed	0.871108	0.922668	0.937985	0.898039	0.976566	1.000000

In this paper, we considered the unformed data with the information in Fig 1. We calculated the correlation between the attributes of data and applied an accuracy metric to evaluate the exact values. We proposed the algorithms could solve with discrete and unformed data by calculating the correlation shown in Table 1.We defined very strong positive correlation when values are greater than or equal to 0.8, strong positive correlation when values are greater than or equal to 0.6 and smaller than 0.8, weak positive correlation when values are greater than or equal to 0.4, and smaller than 0.6. We omitted no correlation (values are in the interval of -0.4 to 0.4), weak negative correlation (Values are

Authors α σ: King Mongkut's University of Technology North Bangkok, Thailand. e-mails: maleerat.m@itd.kmutnb.ac.th, minh.tuan@itd.kmutnb.ac.th, http://kmutnb.ac.th/

smaller than or equal to -0.4 and greater than -0.6), strong negative correlation (Values are greater than -0.8 to values smaller than or equal to -0.6) and very strong negative correlation (Values are smaller or equal to -0.8). We tried models and chose the metric accuracy to calculate the true prediction and the percentage of the prediction. With this metric, we could evaluate exactly the number of predictions and depicted the records related to prediction. PySpark is one of the branches of Hadoop structure becoming strongly and easily in analyzing the data. With the powerful libraries, PySpark supplies the structure for direct and indirect processing, graph environment with ease of use, short time analyzing the big data. PySpark sponsors many sections with many kinds of functions such as Spark SQL, DataFrame, Streaming, MLlib, and Spark Core. PySpark could solve with big data and costs less time to analyze the classification problems. Table 2 shows details of the sections and functions in the PySpark library. The steps for analyzing data could not follow the sections but could form the data before applying the sections and functions (Fig 1). The data will be extracted feature and applied to the model to transform to right form data by choosing basic statistics. After that, we could confirm and make the kinds of problems such as classification, regression, or clustering problems. Finally, we applied evaluation metrics to estimate the models (Equations 1-4).

$$Accuracy = \frac{\sum_{i=1}^{n} T_{iV}}{\sum_{i=1}^{n} T_{iV} + \sum_{j=1}^{m} F_{jV}} \tag{1}$$

Where n; m are numbers of classes, TiV is a true value of prediction at class i; FjV is a false value of label at class j.

$$Precision = \frac{\sum_{i=1}^{n} T_{iP}}{\sum_{i=1}^{n} T_{iP} + \sum_{j=1}^{m} F_{jP}} \tag{2}$$

Where T_{iP} is the true positive at class i, and F_{iP} is false positive at class i. F_{jN} is false negative at class j.

$$Recall = \frac{\sum_{i=1}^{n} T_{iP}}{\sum_{i=1}^{n} T_{iP} + \sum_{j=1}^{m} F_{jN}} \tag{3}$$

$$F_1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

II. Literature Review

Nowadays, machine learning is becoming an essential part of computer science. PySpark is a strong application for analyzing the data with open-source libraries where we can run R, Python, Java, and Scala. PySpark is free for users and easy to use. PySpark supports two strong libraries with Spark MLlib and Spark ML packages where they can solve big data and analyze it in a very short time [17]. However, the processing for analyzing data could follow as Fig 2 shown. We summarized the algorithms used in the PySpark library shown the detail in Table 2.

III. Experiments

In this paper, we got data downloaded on June 10th, 2022 from the website https://ourworldindata.org/covid-deaths and updated every day (Table 2). The data totally consists of 59 attributes and we also chose the attribute with the

Table 2: Library support in PySpark

MLlib	Sections	Features
Data types	Local vector	The vector is formed by an integer or double or zero-based type. The data can be distributed densely or sparsely.
	Labeled point	A kind of local vector using supervised machine learning algorithms with data is labeled. Labels sometimes are 0 and 1 or start from 0, 1,2,... The data can be established in dense or sparse distribution.
	Distributed matrix	A kind of local matrix with long rows and columns. It is so difficult that we can't turn the matrix into another format matrix.
Basic Statistics	Summary Statistic	Giving the information of the instance consists of mean, max, min, variance, and nonzeros.
	Correlations	Depend on the input data, the output could be formed double or matrix.
Classification and Regression	Linear Models	Replied to the input data, the output could be formed double or matrix.MLlib supports the classification and regression algorithms. The classification consists of linear support vector and logistic regression. Regression consists of Lasso, and Ridge regression.
	Decision Trees	Solving any kind of classification problems, this ease of use and handle complicated tasks.
	Random Forest	Using the same model with Decision Trees, Random Forest uses the average of the values to improve the exact predictions but sometimes cost more time than Decision Trees
	Naïve Bayes Isotonic regression	Very strongly applied in labeled data and solving classification problems efficiently. It can be usually applied in sparse vectors conveniently.
Feature	StringIndexer	To turn string attributes into label attributes. If the column is a string, we could change it to a string column and label the string column.
	OneHotEncoder	Change a label column to binary column vectors
	VectorAssembler	Combining all kinds of categorical columns to build a vector column for model prediction.
Evaluation metrics	Classification model evaluation	Applied for binary and multiclass classification, the output could show the confusion matrix, accuracy, precision, recall, and F1-measure.
	Regression model evaluation	Applied for predicting continuous values with Mean Square Error, Root Mean Square Error, Mean Absolute Error, and Coefficient of Determinant.

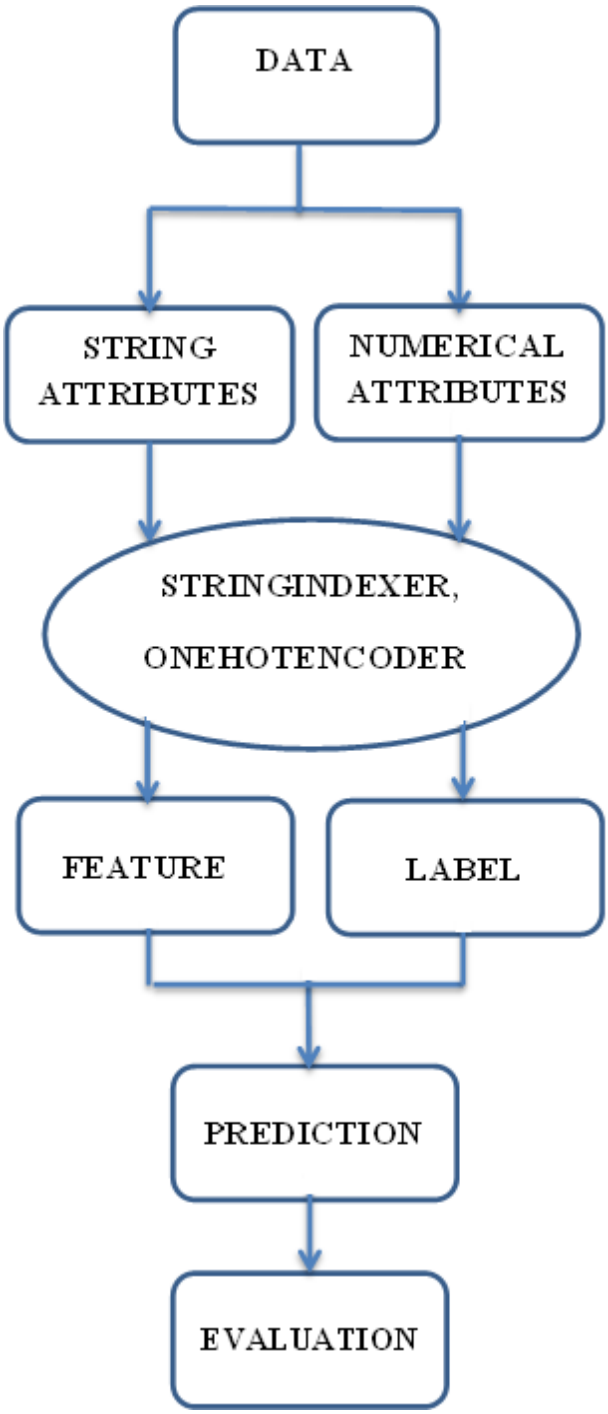

Fig. 1: Steps to Process Data

greatest correlation values in the set of very strong correlation values for building features combining location and total deaths is chosen as labels. The raw data chosen comprises about 208,111 instances and is cleaned by keeping specific character contributes.

As Fig 1 shown, we need to process the data in the right format by using PySpark libraries. The columns selected will be divided into two parts: One part for features and another for labels. We applied StringIndexer to change to the column labeled and applied OneHotEncoder to established binary vector and after that, we applied VectorAssembler to combine with total cases column to make column features for prediction. We also applied StringIndexer to turn total deaths into a label column for target prediction (see Table 2). Besides metrics accuracy to evaluate the ratio of right targets and total targets, we considered evaluating by Precision, Recall, and F1-score occupied great important units in the medical aspect. Precision is confirmed the rightly positive cases while Recall is to confirm rightly negative cases to decide the right method for curing. F1-score, calculated as the average

of Recall and Precision, is applied to confirm how much Recall is more important than Precision. In the medical branch, it is used to decide prior Recall or Precision to choose an appropriate patients' situation.

Compared to deep machine learning, we also analyze the data when trying with deep learning [8{13] such as LSTM, and GRU but get the worse results prediction shown such as the time costs too much time (5,435s/step), accuracy for the first step is 0.138 and the second step is 0.1384. The parameters for solving this data are a total of 202,878,594 parameters and the batch size is 1,318 parameters. PySpark has shown better performance with the best accuracy and least time to evaluate.

IV. Results

In this paper, we tried the models in PySpark and choose the models that could analyze the data. After trying the models in Spark.MLlib and Spark.ML, we got the results in Table 3. The results showed that Naïve Bayes has the best performance in predicting fatalities with an accuracy of 0.813. Following that was the Decision Tree model with an accuracy is 0.621. Table 4 shows some example prediction results with the models. Fig 2-3 showed the screen of prediction using Naïve Bayes and Random Forest.

Table 3: Prediction Results for All Models

Models	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.813	0.571	0.381	0.457
Random Forest	0.139	0.632	0.003	0.005
Decision Tree	0.621	0.824	0.013	0.026

Table 4: Example Results for Naïve Bayes Prediction

Label	Naïve Bayes	Random Forest	Decision Trees
8.0	5	4	3
20.0	32	3	16
23.0	16	1	8
38.0	28	31	12
43.0	46	3	41

V. Conclusion

To aim with getting the best prediction for Covid-19 fatalities, we applied Spark.MLlib and Spark.ml to get the best result in Naïve Bayes model [14{16]. We also try with deep machine learning like LSTM, GRU but get no better answers. With this paper, we hope to have a good prediction in India and other countries getting the Covid-19 cases increasing every day. With this result, we hope we and other researchers could get more information to continue facing the Covid-19 tornado. This result is also the basic report supplied for the government in repairing the utilities for Covid-19 cases. This also proposes a way tosolve big data in the new technology era with unformed data.

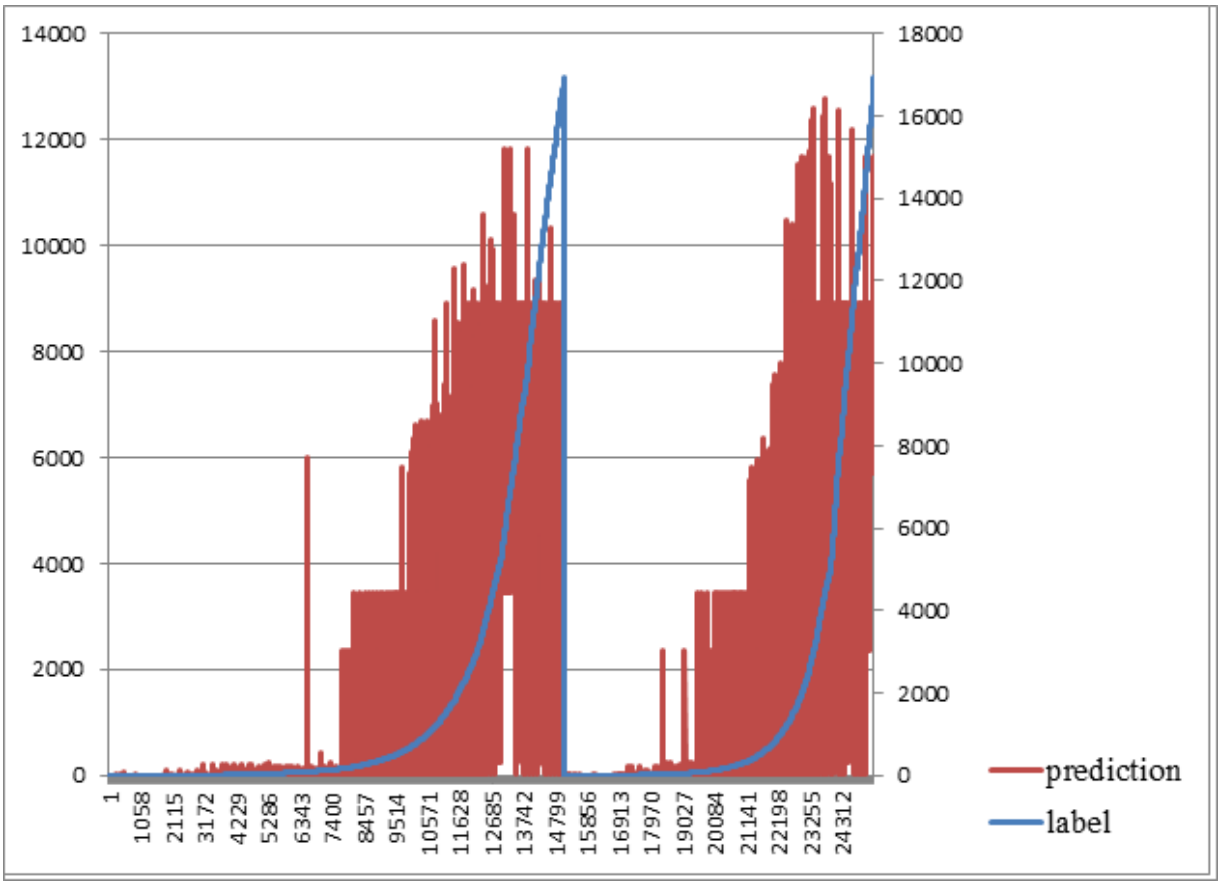

Fig. 2: A Prediction using Naïve Bayes

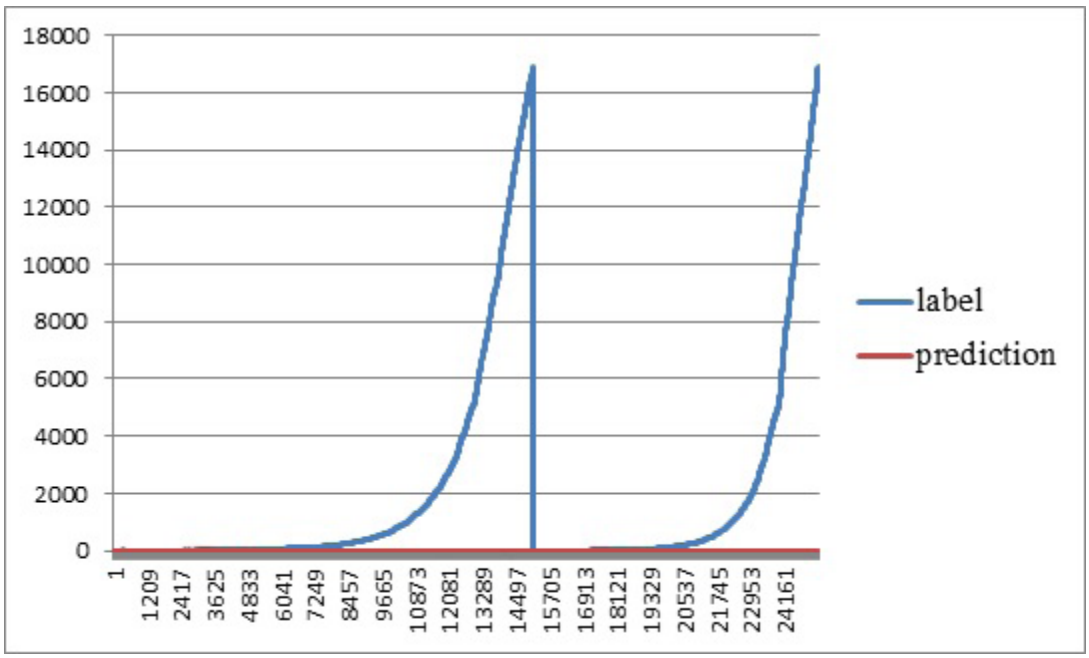

Fig. 3: A Prediction using Random Forest

References Références Referencias

1. Manpinder Singh, Saiba Dalmia, Prediction of number of fatalities due to Covid-19 using Machine Learning, IEEE 17th India Council International Conference (INDICON), 2020.
2. Sirage Zeynu Ahmed, Analysis and forcasting the outbreak of Covid-19 in Ethiopia using machine learning, .
3. Tamer Sh. Mazen, A novel machine learning based model for COVID-19 prediction, International Journal

of Avanced Computer Science and Applications, 2020.
4. Roseline Oluwaseun OGUNDOKUN, Joseph Bamidele AWOTUNDE, Machine learning prediction for Covid 19 pandemic in India, 2020.
5. Sudhir Bhandari, Ajit Singh Shaktawat, Amit Tak, Bhoopendra Patel, Jyotsna Shukla, Sanjay Singhal, Kapil Gupta, Jitendra Gupta, Shivankan Kakkar, Amitabh Dube, Logistic Regression Analysis to Predict Mortality Risk in COVID 19 Patients from Routine Hematologic Parameters, Ibnosina Journal of Medicine and Biomedical Sciences, 2020.
6. Arjun S. Yadaw, Yan-chak Li, Sonali Bose, Ravi Iyengar, Supinda Bunyavanich, and Gaurav Pandey, Clinical predictors of COVID-19 mortality: development and validation of a clinical prediction model" in Lancet Digit Health, 2020.
7. Shreshth Tuli, Shikhar Tuli, Rakesh Tuli, Sukhpal Singh Gill, predicting the growth and trend of Covid-19 pandemic using machine learning and cloud computing, Elsevier public health emergency collection, 2020.
8. Mohammad Behdad Jamshidi, Ali Lalbakhsh, Jakub Talla, Zden_ek Peroutka, Farimah Hadjilooei, Pedram Lalbakhsh, Morteza Jamshidi, Luigi La Spada, Mirhamed Mirmozafari, Mojgan Dehghani, Asal Sabet, Saeed Roshani, Sobhan Roshani, Nima Bayat-Makou, Bahare Mohamadzade, Zahra Malek, Alireza Jamshidi, Sarah Kiani, Hamed Hashemi-Dezaki, and Wahab Mohyuddin, Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis and Treatment, IEEE public health emergency collection, 2020.
9. Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, and Peter M. Atkinson, COVID-19 Outbreak Prediction with Machine Learning, MDPI, 2020.
10. Mohammadreza Nemati, Jamal Ansary, and Nazafarin Nemat, Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data, CellPress, 2020.
11. Batista AFM, Miraglia JL, Donato THR, Chiavegatto Filho ADP, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach, CSH, 2020.
12. Hongwei Zhao, Naveed N Merchant, Alyssa McNulty, Tifiany A Radclifi, Murray J Cote, Rebecca Fischer, Huiyan Sang, Marcia G Ory, COVID-19: Short term prediction model using daily incidence data, Plos One Collection, 2021.
13. Rajan Gupta, Gaurav Pandey, Poonam Chaudhary, Saibal K. Pal, Machine Learning Models for Government to Predict COVID-19 Outbreak, ACM Journal, 2020.
14. Kelly, Anthony; Johnson, Marc Anthony, Investigating the Statistical Assumptions of Naïve Bayes Classifiers, 55th Annual Conference on Information Sciences and Systems (CISS), 2021.
15. Sharmila, B S; Nagapadma, Rohini, Intrusion Detection System using Naive Bayes algorithm, IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), 2019.
16. B S Sharmila, Rohini Nagapadma, Intrusion Detection System using Naive Bayes algorithm, IEEE International WIE Conference on Electrical and Computer Engineering, 2019.
17. Tuan Nguyen M, Meesad P, Nguyen Ha H.C, English-Vietnamese Machine Translation Using Deep Learning, Recent Advances in Information and Communication Technology, 2021, https://doi.org/10.1007/978-3-030-79757-7 10.
18. Maliyaem M, Tuan Nguyen M, Lockhart D, Muenthong S, A Study of Using Machine Learning in Predicting COVID-19 Cases. Cloud Computing and Data Science, 2022.
19. Tuan Nguyen M, Phayung Meesad, A Study of Predicting the Sincerity of a Question Asked Using Machine Learning, 5th International Conference on Natural Language Processing and Information Retrieval (NLPIR), 2021.
20. Tuan Nguyen M, Machine Learning Performance on Predicting Banking Term Deposit, International Conference on Enterprise Information Systems (ICEIS), 2022.
21. Siti NurhidayahSharin, Mohamad KhairilRadzali, Muhamad Shirwan Abdullah-Sani, A network analysis and support vector regression approaches for visualizing and predicting the COVID-19 outbreak in Malaysia, ScienceDirect, 2022.