Global Journals LaTeX JournalKaleidoscope™

Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

CrossRef DOI of original article:

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

Maleerat Maliyaem[1] and Nguyen Minh Tuan[2]

[1] King Mongkut?s University of Technology North Bangkok

Received: 1 January 1970 Accepted: 1 January 1970 Published: 1 January 1970

Abstract

Day by day, the number of confirmed Covid-19 cases signif- icantly increases all over the world. In India, the second wave of coro- navirus 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 govern- ment to manage the medical facilities such as beds and oxygen supply for the hospital.

Index terms— pyspark, machine learning, Naïve bayes, fatality, covid-19.

1 I. Introduction

In paper [1], they used linear regression and polynomial regression to predict the results of fatalities. These two algorithms were applied to nd the best t line to estimate the average values of the two variables. These algorithms are dependent on the variation and dispersion of the data. The best t line will divide the data into two parts with the same distance between the values of data from the best t 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.

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

2 Authors

: 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/ unformed data by calculating the correlation shown in Table **??**.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. 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 **??** 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).

(2)

3 Where

is the true positive at class , and is false positive at class . is false negative at class j.

(

(4)Accuracy = n i=1 T iV n i=1 T iV + m j=1 F jV P recision = n i=1 T iP n i=1 T iP + m j=1 F jP T iP i i F iP F jN Recall = n i=1 T iP n i=1 T iP + m j=1 F jN F 1 ? Score = 2 × P recision × Recall P recision + Recall 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 **??**.

4 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 **??**). The data totally consists of 59 attributes and we also chose the attribute with the 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.

5 MLlib Sections Features

Data types

6 Local vector

The vector is formed by an integer or double or zero-based type. The data can be distributed densely or sparsely.

7 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. 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 **??**). 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.

8 Distributed

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.

9 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. [1]

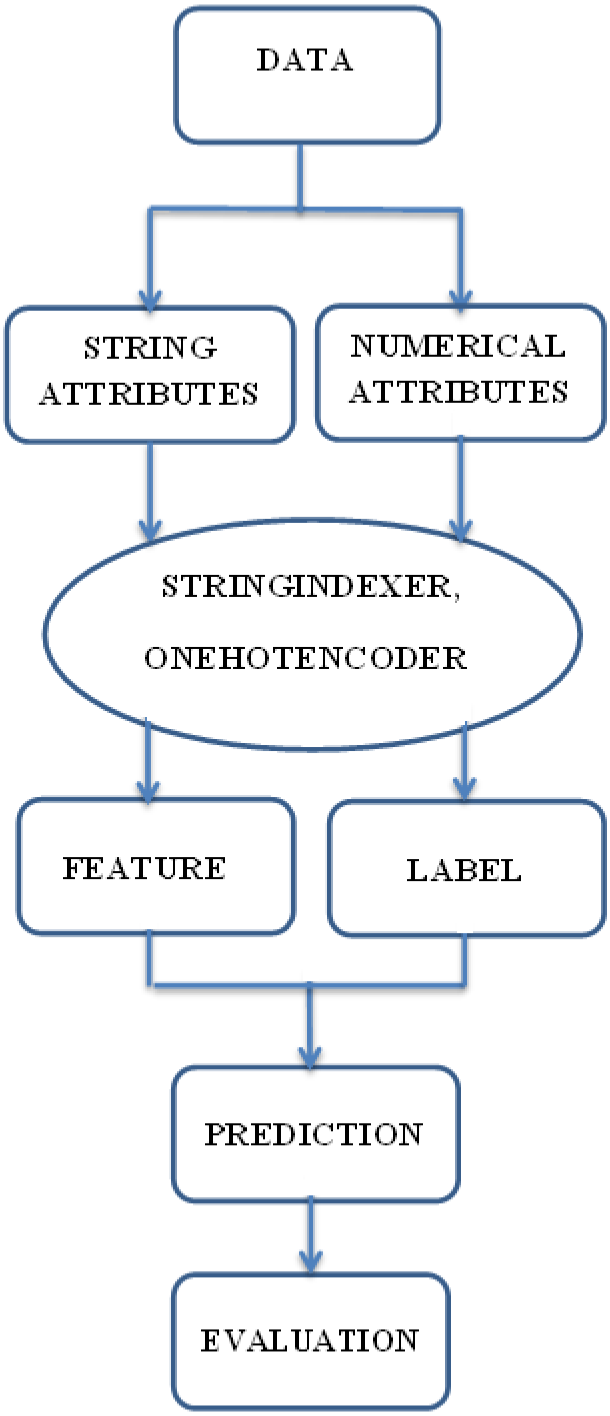

DATA
STRING ATTRIBUTES
NUMERICAL ATTRIBUTES
STRINGINDEXER,
ONEHOTENCODER
FEATURE
LABEL
PREDICTION
EVALUATION

Figure 1:

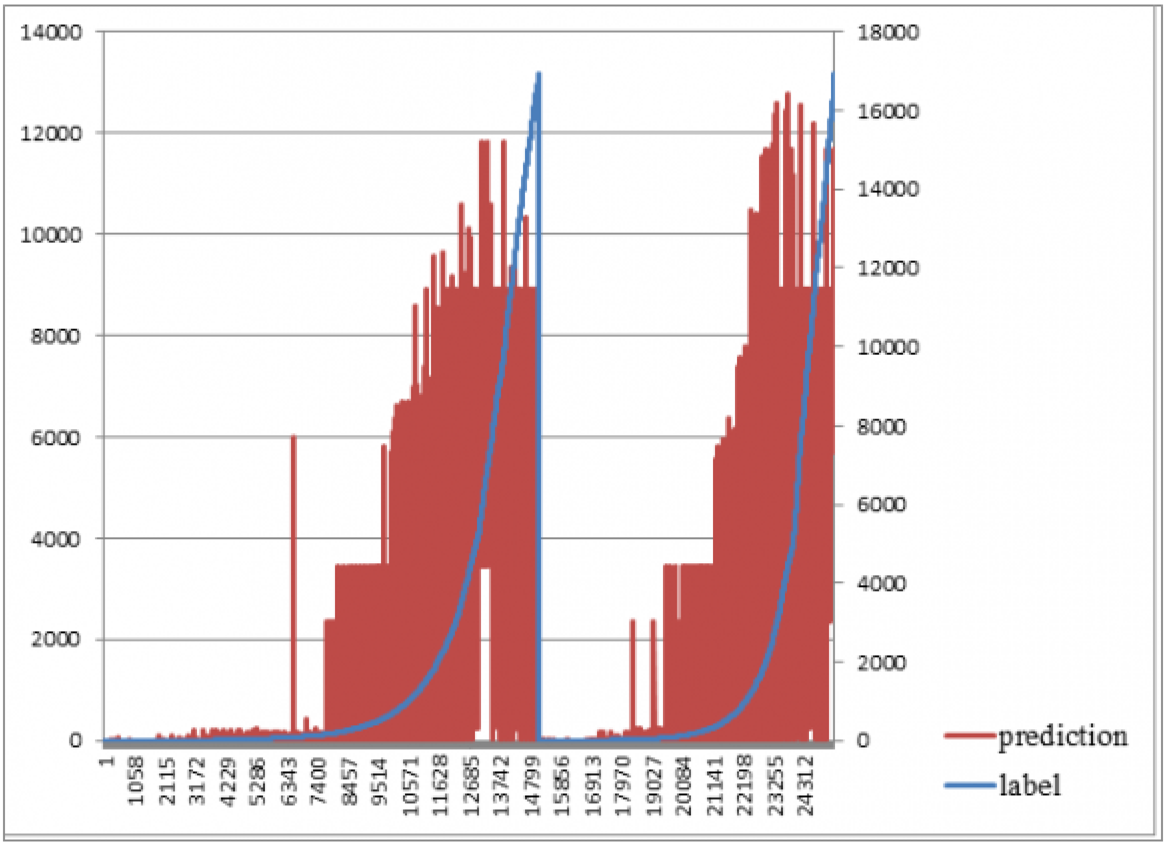
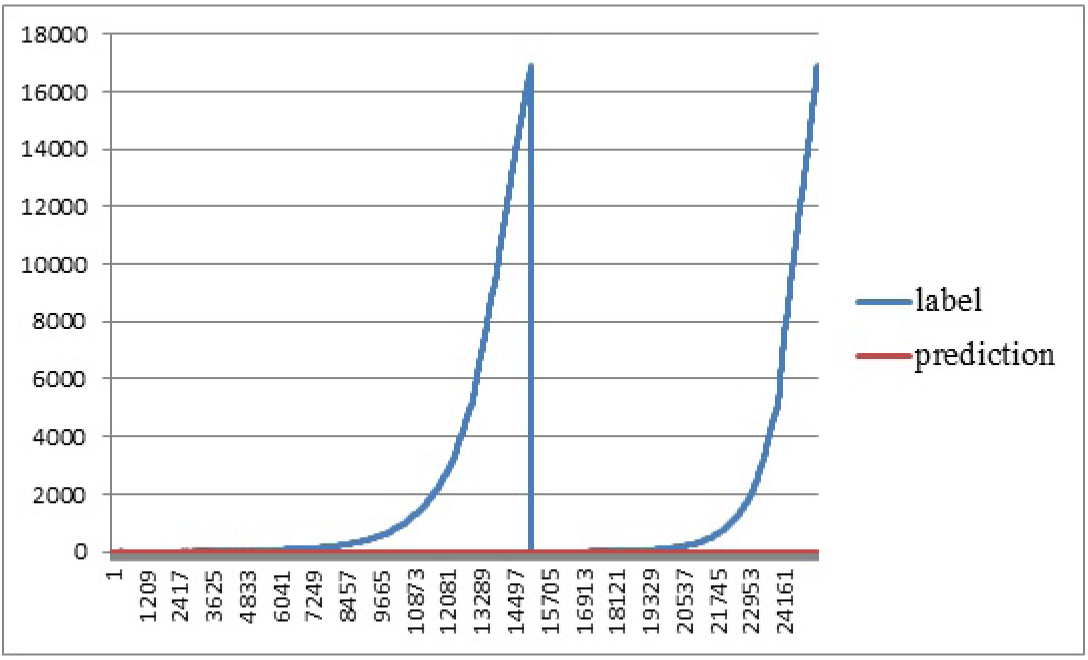


Figure 2:



1

Figure 3: Fig. 1 :

3

Figure 4: Table 3 :

4

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

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

Figure 5: Table 4 :

[Kelly] , Anthony Kelly .

[Sh and Mazen ()] 'A novel machine learning based model for COVID-19 prediction'. Tamer Sh , Mazen . *International Journal of Avanced Computer Science and Applications* 2020.

[Nguyen and Meesad] 'A Study of Predicting the Sincerity of a Question Asked Using Machine Learning'. Tuan Nguyen , M , Phayung Meesad . *5th International Conference on Natural Language Processing and Information Retrieval (NLPIR)*, p. 2021.

[Maliyaem et al. ()] *A Study of Using Machine Learning in Predicting COVID-19 Cases. Cloud Computing and Data Science*, M Maliyaem , Tuan Nguyen , M Lockhart , D Muenthong , S . 2022.

[Sirage Zeynu] *Analysis and forcasting the outbreak of Covid-19 in Ethiopia using machine learning*, Ahmed Sirage Zeynu .

[Ardabili et al. ()] Sina F Ardabili , Amir Mosavi , Pedram Ghamisi , Filip Ferdinand , Annamaria R Varkonyi-Koczy , Uwe Reuter , Timon Rabczuk , Peter M Atkinson . *COVID-19 Outbreak Prediction with Machine Learning*, 2020. MDPI.

[Batista et al. ()] *Chiavegatto Filho ADP, COVID-19 diagnosis prediction in emergency care patients: a machine learning approach*, Afm Batista , J L Miraglia , Thr Donato . 2020. CSH.

[Arjun et al. ()] *Clinical predictors of COVID-19 mortality: development and validation of a clinical prediction model" in Lancet Digit Health*, S Arjun , Yan-Chak Yadaw , Sonali Li , Ravi Bose , Supinda Iyengar , Gaurav Bunyavanich , Pandey . 2020.

[Zhao et al. ()] 'COVID-19: Short term prediction model using daily incidence data'. Hongwei Zhao , N Naveed , Alyssa Merchant , Mcnulty , A Tifiany , Radclifi , J Murray , Rebecca Cote , Huiyan Fischer , Sang , G Marcia , Ory . *Plos One Collection* 2021.

[Nguyen et al. ()] *English-Vietnamese Machine Translation Using Deep Learning, Recent Advances in Information and Communication Technology*, Tuan Nguyen , M Meesad , P , Nguyen Ha , H . 10.1007/978-3-030-79757-710. https://doi.org/10.1007/978-3-030-79757-710 2021.

[Sharmila and Nagapadma ()] 'Intrusion Detection System using Naive Bayes algorithm'. B S; Sharmila , Rohini Nagapadma . *IEEE International WIE Conference on Electrical and Computer Engineering*, 2019. WIECON-ECE

[B S Sharmila and Nagapadma ()] 'Intrusion Detection System using Naive Bayes algorithm'. Rohini B S Sharmila , Nagapadma . *IEEE International WIE Conference on Electrical and Computer Engineering*, 2019.

[Johnson and Anthony] 'Investigating the Statistical Assumptions of Naïve Bayes Classifiers'. Marc Johnson , Anthony . *55th Annual Conference on Information Sciences and Systems (CISS)*, p. 2021.

[Jamshidi et al.] Mohammad Behdad Jamshidi , Ali Lalbakhsh , Jakub Talla , Zden_Ek Peroutka , Farimah Hadjilooei , Pedram Lalbakhsh , Morteza Jamshidi , Luigi La Spada , Mirhamed Mirmozafari , Mojgan Dehghani . *Deep Learning Approaches for Diagnosis and Treatment*, Asal Sabet, Saeed Roshani, Sobhan Roshani, Nima Bayat-Makou, Bahare Mohamadzade, Zahra Malek, Alireza Jamshidi, Sarah Kiani, Hamed Hashemi-Dezaki, Wahab Mohyuddin (ed.) 19 p. 2020.

[Bhandari et al. ()] 'Logistic Regression Analysis to Predict Mortality Risk in COVID 19 Patients from Routine Hematologic Parameters'. Sudhir Bhandari , Ajit Singh Shaktawat , Amit Tak , Bhoopendra Patel , Jyotsna Shukla , Sanjay Singhal , Kapil Gupta , Jitendra Gupta , Shivankan Kakkar , Amitabh Dube . *Ibnosina Journal of Medicine and Biomedical Sciences* 2020.

[Gupta et al. ()] 'Machine Learning Models for Government to Predict COVID-19 Outbreak'. Rajan Gupta , Gaurav Pandey , Poonam Chaudhary , K Saibal , Pal . *ACM Journal* 2020.

[Oluwaseun and Bamidele ()] *Machine learning prediction for Covid 19 pandemic in India*, Roseline Oluwaseun , Ogundokun , Joseph Bamidele , Awotunde . 2020.

[Nemati et al. ()] *Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data*, Mohammadreza Nemati , Jamal Ansary , Nazafarin Nemat . 2020. CellPress.

[Siti Nurhidayahsharin and Khairilradzali] *Muhamad Shirwan Abdullah-Sani, A network analysis and support vector regression approaches for visualizing and predicting the COVID-19 outbreak in Malaysia*, Mohamad Siti Nurhidayahsharin , Khairilradzali . ScienceDirect. p. 2022.

[Nguyen] Tuan Nguyen , M . *Machine Learning Performance on Predicting Banking Term Deposit, International Conference on Enterprise Information Systems (ICEIS)*, p. 2022.

[Singh and Dalmia] 'Prediction of number of fatalities due to Covid-19 using Machine Learning'. Manpinder Singh , Saiba Dalmia . *IEEE 17th India Council International Conference (INDICON)*, p. 2020.

[Tuli et al. ()] *Sukhpal Singh Gill, predicting the growth and trend of Covid-19 pandemic using machine learning and cloud computing*, Shreshth Tuli , Shikhar Tuli , Rakesh Tuli . 2020. Elsevier public health emergency collection.