COVID-19 FUTURE FORECASTING USING SUPERVISED MACHINE LEARNING MODELS

R Sai Santosh, saisantoshrayana@gmail.com
M Ram Charan, charanram1998@gmail.com
A Sarayu, sarayu.achampeta@gmail.com

ABSTRACT: Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.

Keywords: ***COVID-19, exponential smoothing method, future forecasting, supervised machine learning.***

1. INTRODUCTION

Machine learning (ML) has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle (AV), business applications, natural language processing (NLP), intelligent robots, gaming, climate modeling, voice, and image processing. ML algorithms' learning is typically based on trial and error method quite opposite of conventional algorithms, which follows the programming instructions based on decision statements like if-else [1]. One of the most significant areas of ML is forecasting [2], numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis. Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease [3]. There are lots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease [4], cardiovascular disease prediction [5], and breast cancer prediction [6]. In particular, the study [7] is focused on live forecasting of COVID-19 confirmed cases and study [8] is also focused on the forecast of COVID- 19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively.

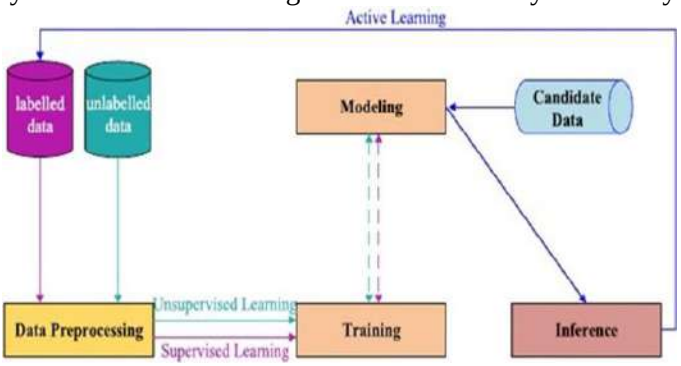

Fig.1: Flowchart for training process machine learning tasks

This study aims to provide an early forecast model for the spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO) [9]. COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia [10]. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration [11]. Hundreds of thousands of people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close person to person physical contacts, by respiratory droplets, or by touching the contaminated surfaces.

2. EXISTING SYSTEM

Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity.

3. PROPOSED SYSTEM

To contribute to the current human crisis our attempt in this study is to develop a forecasting system for COVID-19. The forecasting is done for the three important variables of the disease for the coming 10 days:

1) The number 0f New confirmed cases.
2) The number of death cases
3) The number of recoveries.

This problem of forecasting has been considered as a regression problem in this study, so the study is based on some state-of-art supervised ML regression models such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES). The learning models have been trained using the COVID-19 patient stats dataset provided by Johns Hopkins.

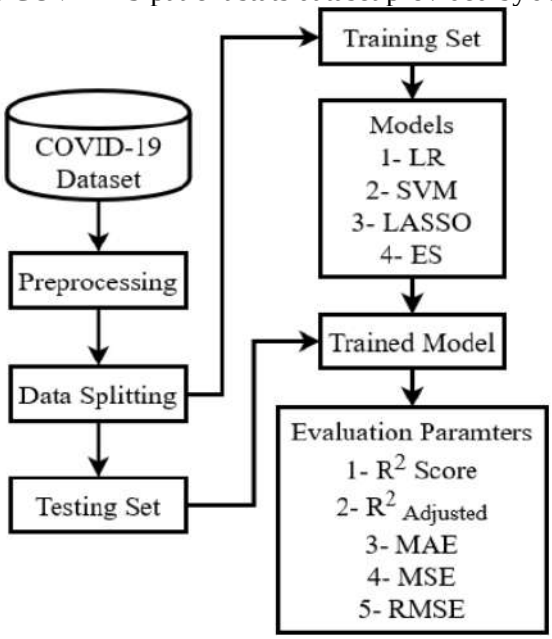

Fig.2: Workflow diagram

The study is about novel coronavirus also known as COVID- 19 predictions. The COVID-19 has proved a present potential threat to human life. It causes tens of thousands of deaths and the death rate is increasing day by day throughout the globe. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 10 days. The forecasting has been done by using four ML approaches that are appropriate to this context. The dataset used in the study contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries in the past number of days from which the pandemic started. Initially, the dataset has been preprocessed for this study to find the global statistics of the daily number of deaths, confirmed cases, and recoveries.

4. RELATED WORK

4.1 Using machine learning algorithms for breast cancer risk prediction and diagnosis.

Breast cancer represents one of the diseases that make a high number of deaths every year. It is the most common type of all cancers and the main cause of women's deaths worldwide. Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions. In this paper, a performance comparison between different machine learning algorithms: Support Vector Machine (SVM), Decision Tree (C4.5), Naive Bayes (NB) and k Nearest Neighbors (k-NN) on the Wisconsin Breast Cancer (original) datasets is conducted. The main objective is to assess the correctness in classifying data with respect to efficiency and effectiveness of each algorithm in terms of accuracy, precision, sensitivity and specificity. Experimental results show that SVM gives the highest accuracy (97.13%) with lowest error rate. All experiments are executed within a simulation environment and conducted in WEKA data mining tool.

4.2 Forecasting the novel coronavirus covid-19:

What will be the global impact of the novel coronavirus (COVID-19)? Answering this question requires accurate forecasting the spread of confirmed cases as well as analysis of the number of deaths and recoveries. Forecasting, however, requires ample historical data. At the same time, no prediction is certain as the future rarely repeats itself in the same way as the past. Moreover, forecasts are influenced by the reliability of the data, vested interests, and what variables are being predicted. Also, psychological factors play a significant role in how people perceive and react to the danger from the disease and the fear that it may affect them personally. This paper introduces an objective approach to predicting the continuation of the COVID-19 using a simple, but powerful method to do so. Assuming that the data used is reliable and that the future will continue to follow the past pattern of the disease, our forecasts suggest a continuing increase in the confirmed COVID-19 cases with sizable associated uncertainty. The risks are far from symmetric as underestimating its spread like a pandemic and not doing enough to contain it is much more severe than overspending and being over careful when it will not be needed. This paper describes the timeline of a live forecasting exercise with massive potential implications for planning and decision making and provides objective forecasts for the confirmed cases of COVID-19.

4.3 Identification of a new human coronavirus:

Three human coronaviruses are known to exist: human coronavirus 229E (HCoV-229E), HCoV-OC43 and severe acute respiratory syndrome (SARS)-associated coronavirus (SARS-CoV). Here we report the identification of a fourth human coronavirus, HCoV-NL63, using a new method of virus discovery. The virus was isolated from a 7-month-old child suffering from bronchiolitis and conjunctivitis. The complete genome sequence indicates that this virus is not a recombinant, but rather a new group 1 coronavirus. The in vitro host cell range of HCoV-NL63 is notable because it replicates on tertiary monkey kidney cells and the monkey kidney LLC-MK2 cell line. The viral genome contains distinctive features, including a unique N-terminal fragment within the spike protein. Screening of clinical specimens from individuals suffering from respiratory illness identified seven additional HCoV-NL63-infected individuals, indicating that the virus was widely spread within the human population.

5. IMPLEMENTATION

In this paper author using various machine learning algorithms such as SVM, Linear Regression, Lasso and ES (Exponential Smoothing) to forecast COVID-19 disease. Among all algorithms ES is giving better forecast result

compare to other algorithms and to implement this dataset author has COVID-19 dataset from John Hopkins GitHub Page and this dataset contains lots of attributes but we need total cases and DATE columns to forest confirmed cases, death and recovered cases.

SVM algorithm: SVM is a supervised machine learning algorithm which can be used for classification or regression problems. It uses a technique called the kernel trick to transform your data and then based on these transformations it finds an optimal boundary between the possible outputs.

Linear Regression Algorithm: Linear Regression Algorithm is a machine learning algorithm based on supervised learning. ... Regression analysis is used for three types of applications: Finding out the effect of Input variables on Target variable. Finding out the change in Target variable with respect to one or more input variable.

LASSO: Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean.

Exponential smoothing: is a time series forecasting method for univariate data that can be extended to support data with a systematic trend or seasonal component. It is a powerful forecasting method that may be used as an alternative to the popular Box-Jenkins ARIMA family of methods.

6. EXPERIMENTAL RESULTS

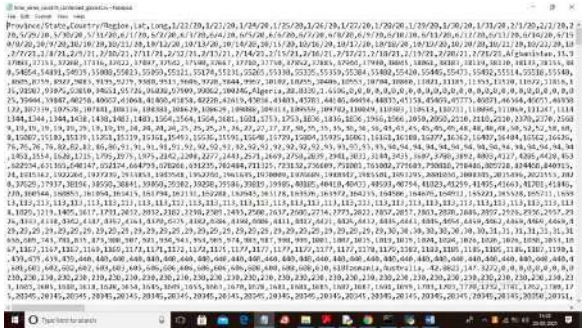

Fig.3: Dataset

using above dataset to trained machine learning algorithms and then using trained machine learning model to forecast next 30 days records and then we are comparing forecast and actual data to find out error rate.

Fig.4: Home screen

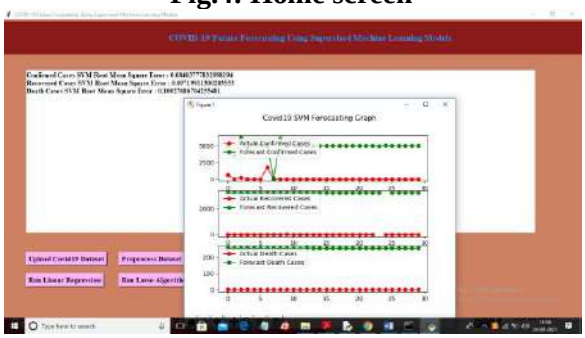

Fig.5: SVM algorithm

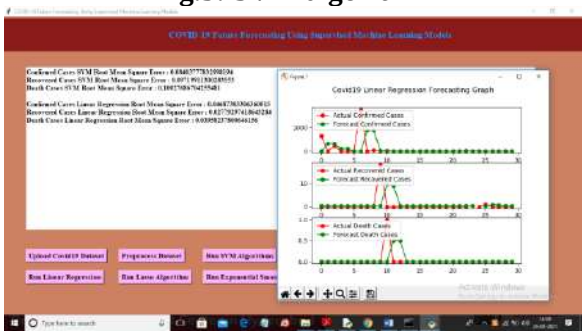

Fig.6: Linear regression algorithm

Fig.7: LASSO algorithm

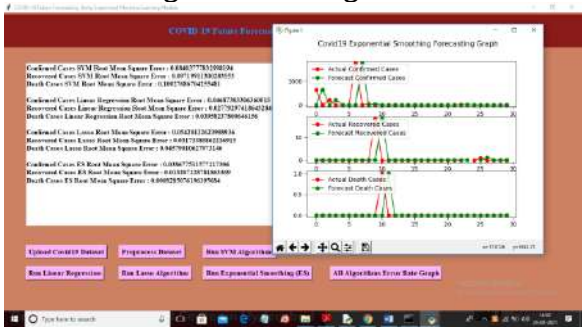

Fig.8: Exponential smoothing algorithm

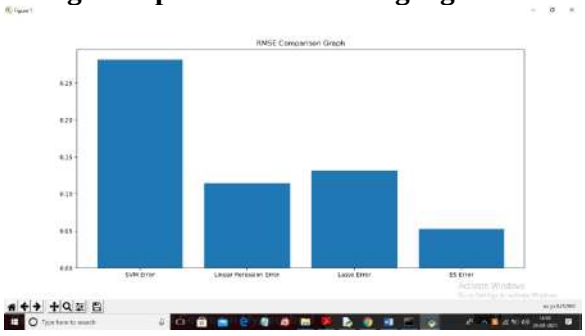

Fig.9: All algorithms error rate graph

7. CONCLUSION

The results of the study prove that ES performs best in the current forecasting domain given the nature and size of the dataset. LR and LASSO also perform well for forecasting to some extent to predict death rate and confirm cases. According to the results of these two models, the death rates will increase in upcoming days, and recoveries rate will be slowed down. SVM produces poor results in all scenarios because of the ups and downs in the dataset

values. It was very difficult to put an accurate hyperplane between the given values of the dataset. Overall we conclude that model predictions according to the current scenario are correct which may be helpful to understand the upcoming situation. The study forecasts thus can also be of great help for the authorities to take timely actions and make decisions to contain the COVID-19 crisis.

8. FUTURE SCOPE

This study will be enhanced continuously in the future course, next we plan to explore the prediction methodology using the updated dataset and use the most accurate and appropriate ML methods for forecasting. Real-time live forecasting will be one of the primary focuses in our future work.

REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," PloS one, vol. 13, no. 3, 2018.
[2] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in European business intelligence summer school. Springer, 2012, pp. 62–77.
[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, "Regression models for prognostic prediction: advantages, problems, and suggested solutions." Cancer treatment reports, vol. 69, no. 10, pp. 1071–1077, 1985.
[4] P. Lapuerta, S. P. Azen, and L. LaBree, "Use of neural networks in predicting the risk of coronary artery disease," Computers and Biomedical Research, vol. 28, no. 1, pp. 38–52, 1995.
[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," American heart journal, vol. 121, no. 1, pp. 293–298, 1991.
[6] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," Procedia Computer Science, vol. 83, pp. 1064–1069, 2016.
[7] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," Plos one, vol. 15, no. 3, p. e0231236, 2020.
[8] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response," Jama, 2020.
[9] WHO. Naming the coronavirus disease (covid-19) and the virus that causes it. [Online]. Available: https://www.who.int/emergencies/diseases/novelcoronavirus-2019/technical-guidance/naming-the-coronavirus-disease- (covid-2019)-and-the-virus-that-causes-it
[10] C. P. E. R. E. Novel et al., "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (covid-19) in china," Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi, vol. 41, no. 2, p. 145, 2020.
[11] L. van der Hoek, K. Pyrc, M. F. Jebbink, W. Vermeulen-Oost, R. J. Berkhout, K. C. Wolthers, P. M. Wertheim-van Dillen, J. Kaandorp, J. Spaargaren, and B. Berkhout, "Identification of a new human coronavirus," Nature medicine, vol. 10, no. 4, pp. 368–373, 2004.
[12] J. H. U. data repository. Cssegisanddata. [Online]. Available: https://github.com/CSSEGISandData
[13] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Chapter 1 - analytics defined," in Information Security Analytics, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston: Syngress, 2015, pp. 1 – 12. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128002070000010
[14] H.-L. Hwa, W.-H. Kuo, L.-Y. Chang, M.-Y. Wang, T.-H. Tung, K.-J. Chang, and F.-J. Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models," Journal of evaluation in clinical practice, vol. 14, no. 2, pp. 275–280, 2008.
[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996