ZKG INTERNATIONAL

ISSN: 2366-1313

SUPERVISED MACHINE LEARNING MODELS FOR COVID-19 FUTURE FORECASTING

M. LAKSHMI BAI[1], A. UMA KARTHIK[2], K. RAJESH [3], B. VINEELA[4], K. HARSHITH CHOWDARY[5].

[1] Associate Professor, CSE,Chalapathi Institute of Technology,Guntur, India

[2]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[3]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[4]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

[5]UG Student,CSE,Chalapathi Institute of Technology,Guntur, India

ABSTRACT: Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in preoperative 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 Random Forest(RF), Extreme Gradient Boosting(XGBoost), Light Gradient Boosting (LGBMR)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 prove it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the performs best among all the used models followed by RF and LGBMR which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while XGBoost performs poorly in all the prediction scenarios given the available dataset.

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

ISSN: 2366-1313

[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 COVID19 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. This study aims to provide an early forecast model for the spread of novel corona virus, 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. The most challenging aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared either partial or strict lockdowns throughout the affected regions and cities. 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. 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. The dataset has been preprocessed and divided into two subsets: training set (85% records) and testing set (15% records). The performance evaluation has been done in terms of important measures including R-squared score (R2 score), Adjusted R-squared Score (R2 adjusted), mean square error (MSE), mean absolute error (MAE), and root mean square error (RMSE). This study has some key findings which are listed below: • ES performs best when the time-series dataset has very limited entries. • Different ML algorithms seem to perform better in different class predictions.

• Most of the ML algorithms require an ample amount of data to predict the future, as the size of the dataset increases the model performances improve.

ISSN: 2366-1313

• ML model based forecasting can be very useful for decision-makers to contain pandemics like COVID-19.

2. LITERATURE SURVEY

The COVID-19 outbreaks have become a disaster for several nations. However, the recovery rate of COVID-19 in India is more than 88%. In this study, we have proposed the piecewise linear regression based machine learning approach for the prediction of actual positive cases and recovery cases of five different states in India. The main novelty of the proposed scheme is that we have applied piecewise linear regression method instead of simple linear regression. As a result, the proposed scheme produces an accurately predicted result for both cases. Henceforth, it may be concluded that our model could be applicable for other parameters of COVID-19 also in any state or country as well.In the future, we will focus on developing various ML- and DL-based model to enhance the performance to combat COVID-19 as well as other pandemic may be. The main challenge to implement the piecewise linear regression is to find the point of partition of data .In this paper, the partition has been done by observing the slope of the point heuristically and we have considered the past 7 days data to predict the next day. In future, our aim is to solve the problem to find an optimal partitioning point such that the error becomes minimum.

3.EXISTING SYSTEM

Comparative analysis of machine learning models to predict the outbreak of COVID-19 in various countries. Their study and analysis demonstrate the potential of machine learning models for the prediction of COVID19. In this article they predicted the disease in an individual by using the clinical information and CT scan to predict the outbreak. Our work was based entirely on the outbreak of positive cases and deaths in future days.

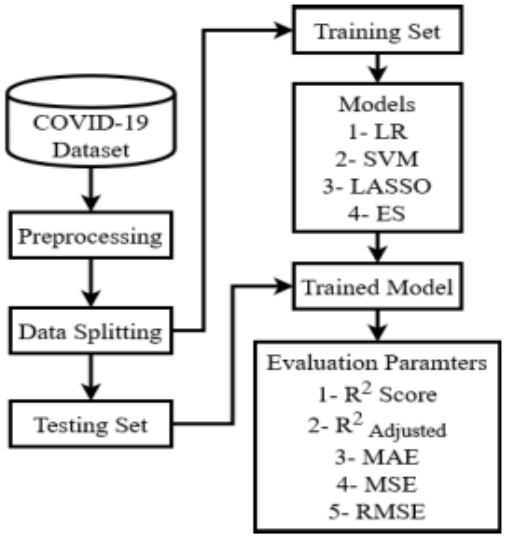
4. PROPOSED SYSTEM

We will analyze the datasets using Machine learning techniques. Then to predict the total number of confirmed and recovered cases machine learning strategies such as polynomial regression are performed using various python libraries.

It is also to train the data of COVID-19 by collecting the raw data from the various sources.

1. This system is capable to train the dataset of both persons wearing masks and without wearing masks.
2. After training the model the system can predict the number of positive cases in the upcoming days.
3. It also can predict the number of deaths in future days.

5. ARCHITECTUREDIAGRAM



6. IMPLEMENTATION

Data Set:

This is the step where we collect the raw data of COVID-19 from the website.

ISSN: 2366-1313

CSV (Comma Seperated Values): A CSV file is a text file that has a specific format which allows data to be saved in a table structured format.

Pre Processing:

It is a technique which is used to transfer the raw data into a useful or efficient format such that to avoid duplication and to reduce the size of the rows.

Training the Data: The main purpose is to search the some important information in the raw data .We have used neural network technologies for training the data. Training is nothing but feature extraction.

Predicting the Outbreak: Here, We have used a regression technique to predict the output. We made use of polynomial regression to find the number of cases in future days.

Preprocessing

In any Machine Learning process, Data Preprocessing is that step in which the data gets transformed, or Encoded, to bring it to such a state that now the machine can easily parse it. In other words, the features of the data can now be easily interpreted by the algorithm.

Supervised learning

Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs. The data is known as training data, and consists of a set of training examples. Each training example has one or more inputs and the desired output, also known as a supervisory signal. In the mathematical model, each training example is represented by an array or vector, sometimes called a feature vector, and the training data is represented by a matrix. Through iterative optimization of an objective function, supervised learning algorithms learn a function that can be used to predict the output associated with new inputs. An optimal function will allow the algorithm to correctly determine the output for inputs that were not a part of the training data. An algorithm that improves the accuracy of its outputs or predictions over time is said to have learned to perform that task.

Supervised learning algorithms include classification and regression. Classif ication algorithms are used when the outputs are restricted to a limited set of values, and regression algorithms are used when the outputs may have any numerical value within a range. Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is to learn from examples using a similarity function that measures how similar or related two objects are. It has applications in ranking, recommendation systems, visual identity tracking, face verification, and speaker verification.

Supervised learning can be grouped further in two categories of algorithms:

1.Classification

2.Regression

Unsupervised learning

Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. The algorithms, therefore, learn from test data that has not been labeled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. A central application of unsupervised learning is in the field of density estimation in statistics, though unsupervised learning encompasses other domains involving summarizing and explaining data features.

ALGORITHM USED

SUPERVISED MACHINE LEARNING MODELS

A supervised learning model is built to make a prediction when it is provided with an

ISSN: 2366-1313

unknown input instance. Thus in this learning technique, the learning algorithm takes a dataset with input instances along with their corresponding regressor to train the regression model. The trained model then generates a prediction for the given unforeseen input data or test dataset [13]. This learning method may use regression techniques and classification algorithms for predictive models' development Four regression models have been used in this study of COVID-19 future forecasting: • Linear Regression • LASSO Regression • Support Vector Machine • Exponential Smoothing 1) Linear Regression In regression modeling, a target class is predicated on the independent features [14]. This method can be thus used to find out the relationship between independent and dependent variables and also for forecasting. Linear regression a type of regression modeling is the most usable statistical technique for predictive analysis in machine learning. Each observation in linear regression depends on two values, one is the dependent variable and the second is the independent variable. Linear regression determines a linear relationship between these dependent and independent variables.

To put the concept of linear regression in the machine learning context, in order to train the model x is represented as input training dataset, y represents the class labels present in the input dataset. The goal of the machine learning algorithm then is to find the best values for β0 (intercept) and β1(coefficient) to get the best-fit regression line. To get the best fit implies the difference between the actual values and predicted values should be minimum.

7. METHODOLOGY

The study is about novel corona virus also known as COVID19 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.

8.SCRREN SHOTS

9.CONCLUSION FUTURE SCOPE

The precariousness of the COVID-19 pandemic can ignite a massive global crisis. Some researchers and government agencies throughout the world have apprehensions that the pandemic can affect a large proportion of the world population [26], [27]. In this study, an ML-based prediction system has been proposed for predicting the risk of COVID19 outbreak globally. The system analyses dataset containing the day-wise actual past data and makes predictions for upcoming days using machine learning algorithms. The results of the study prove

ISSN: 2366-1313

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 hyper plane 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. This study will be enhanced continuously.

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. 4

[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,” Zhonghualiuxingbingxue za zhi= Zhonghualiuxingbingxuezazhi, 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.

ISSN: 2366-1313

[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.