IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

COVID-19 Future Predictions Using 4 Supervised Machine Learning Models

Aditi Vadhavkar[1], Pratiksha Thombare[2], Priyanka Bhalerao[3], Utkarsha Auti[4]
Students, Department of Computer Engineering[1,2,3,4]
D. Y. Patil Institute of Engineering & Technology, Pune, Maharashtra, India

Abstract: *Forecasting Mechanisms like Machine Learning (ML) models having been proving their significance to anticipate perioperative outcomes in the domain of decision making on the future course of actions. Many application domains have witnessed the use of ML models for identification and prioritization of adverse factors for a threat. The spread of COVID-19 has proven to be a great threat to a mankind announcing it a worldwide pandemic throughout. Many assets throughout the world has faced enormous infectivity and contagiousness of this illness. To look at the figure of undermining components of COVID-19 we've specifically used four Machine Learning Models Linear Regression (LR), Least shrinkage and determination administrator (LASSO), Support vector machine (SVM) and Exponential smoothing (ES). The results depict that the ES performs best among the four models employed in this study, followed by LR and LASSO which performs well in forecasting the newly confirmed cases, death rates yet recovery rates, but SVM performs poorly all told the prediction scenarios given the available dataset.*

Keywords: Forecasting, Machine Learning, COVID-19, Supervised Machine Learning

I. Introduction

Machine Learning continuously have been proving itself as a prominent field of study over the decade by solving many very complex and sophisticated real-world problems. The real-world domains including healthcare, autonomous vehicle (AV), business applications, natural language processing (NLP), intelligent robots, gaming, climate, modelling, voice and image processing are the application areas of the ML. Unlike conventional algorithms', which follows the programming instructions, ML algorithms are quite opposite working on trial and error method. Forecasting being the most significant areas of ML, numerous standard algorithms have been used in this area to guide the future course of actions in the areas including weather forecasting, disease prognosis and various other forecasting domains. This study in particular focuses on the live forecasting of COVID-19 confirmed cases and outbreak with early response. These predictions can be very helpful in decision making to handle the present scenarios & guide early interventions to manage the spread of this pandemic effectively.

II. Motivation And Problem Definition

2.1 Motivation

The increasing spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization, has been affecting the lives of millions of people all over the world. It has impacted lifestyles, work-culture, education and many more. Studying its outbreak and early response by the number of new positive cases, the number of deaths, and the number of recoveries will helps us figure out better solutions to tackle the global pandemic.

2.2 Problem Definition

To study the future forecasting of COVID-19 outbreak focusing on the quantity of newly infected cases, deaths, and recoveries.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

III. MATERIALS AND METHODS

3.1 Dataset

The focus of this study is on the quantity of the positive cases, the number of deaths, and also the number of recoveries of COVID-19. The dataset employed in this study is available online on covidtracking.com which is the leading stats of COVID-19 within the States.

3.2 Supervised Machine Learning Models

Supervised Machine Learning is a subcategory of machine learning and computing accustomed labelled datasets to train algorithms that to classify data or predict outcomes accurately. These models will be employed in the context of classification, once we want to map input to output labels, or regression, once we want to map input to a continual output. The regression models used here are:

- Linear regression
- LASSO regression
- Support Vector Machine
- Exponential Smoothing

A. Linear regression

Linear Regression is a ML algorithm based on supervised learning which performs a regression task and target prediction value supported independent variables. It's most likely probably used for locating out the connection between variables and forecasting.

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$E(y) = \beta_0 + \beta_1 x$$

Where, *ε points to the error term of linear regression followed by two factors (x, y)representing input training dataset and sophistication labels within the input dataset respectively.*

B. LASSO

LASSO or least absolute shrinkage and selection operator could be a methodology that performs each variable choice and regularization therefore on boost prediction accuracy and interpretability of the ensuing applied math model. Shrinkage here means that the action of data values contracted towards a central purpose as a result of the mean ensuing to less errors and additional stable model.

C. Support Vector Machine

SVM could be a supervise machine learning algorithm which might be used for both classification or regression challenges within the domain. It's a model that is able to generalize between two different classes if the set of labelled data is provided within the training set to the algorithm, though, the most function maps to the pliability to inform apart between two classes. It is the ability to resolve linear and non-linear problems and work well for several practical problems.

D. Exponential Smoothing

Exponential smoothing in machine learning might be a family of prediction models that uses weighted averages of past observations to forecast new values giving the conception to administer a lot of importance to recent values inside the series. As a result, once observations get mature (in time), the importance of those values get exponentially smaller. ES specially supports data point knowledge with seasonal parts, or say, systematic trends wherever it is used once the observations to create anticipation.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

3.3 Evaluation Parameters

To evaluate the performance of each of the learning models, we will be using some evaluation parameters like R-squared score, Adjusted R-square, MAE, MSE and root mean square error. These metrics guide us to how accurate our predictions are and, what is the amount of deviation from the actual values in our data.

A. R-Squared Score

In straightforward terms, the R-squared score maps to the proportion of the variance within the dependent variables that's certain from the freelance variables. thus if it's 100%, the 2 variables area unit utterly related with no variance the least bit.

B. Adjusted R-square

Being a modified version of R-square, this has been adjusted for the number of predictors in the model. It increases when the new term is witnessing the model more than would be expected by chance and reduces (decreases) when a predictor improves the model by less than expected.

C. Mean Absolute Error

In statistics, mean absolute error could also be alive of errors between paired observations expressing the same development and put together a model analysis metric used with regression models. With relevance a take a glance at set the mean of completely the values of the individual prediction errors on over all instances at intervals the take a glance at set. It measures the everyday magnitude of the errors throughout a group of forecasts, whereas not considering their direction.

D. Root Mean Square Error

RMSE is that the basis of the mean of the square of all of the error giving a good live of accuracy, but alone to examine prediction errors of varied models. In general, a lower RMSD value us on top of ensuing one.

IV. METHODOLOGY

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 20780 entries, 0 to 20779
Data columns (total 55 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   date                        20780 non-null  datetime64[ns]
 1   state                       20780 non-null  object
 2   positive                    20592 non-null  float64
 3   probableCases               9271 non-null   float64
 4   negative                    13290 non-null  float64
 5   pending                     2138 non-null   float64
 6   totalTestResultsSource      20780 non-null  object
 7   totalTestResults            20614 non-null  float64
 8   hospitalizedCurrently       17339 non-null  float64
 9   hospitalizedCumulative      12382 non-null  float64
 10  inIcuCurrently              11636 non-null  float64
 11  inIcuCumulative             3789 non-null   float64
 12  onVentilatorCurrently       9126 non-null   float64
 13  onVentilatorCumulative      1290 non-null   float64
 14  recovered                   12003 non-null  float64
 15  lastUpdateEt                20164 non-null  object
 16  dateModified                20164 non-null  object
 17  checkTimeEt                 20164 non-null  object
 18  death                       19930 non-null  float64
 19  hospitalized                12382 non-null  float64
 20  hospitalizedDischarged      3070 non-null   float64
 21  totalTestsViral             14516 non-null  float64
 22  positiveTestsViral          8958 non-null   float64
 23  negativeTestsViral          5024 non-null   float64
 24  positiveCasesViral          14246 non-null  float64
 25  deathConfirmed              9422 non-null   float64
 26  deathProbable               7593 non-null   float64
 27  totalTestEncountersViral    5231 non-null   float64
 28  totalTestsPeopleViral       9197 non-null   float64
```

Figure 4.1: Sample of the dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

index	date	state	positive	probableCases	negative	pending	totalTestResultsSource	totalTestResults	hospitalizedCurrently	hospitalizedCumulative	inIcuCurrently	inIcuCumulativ
0	20210307	AK	56886.0	NaN	NaN	NaN	totalTestsViral	1731628.0	33.0	1293.0	NaN	Nal
1	20210307	AL	499819.0	107742.0	1931711.0	NaN	totalTestsPeopleViral	2323788.0	494.0	45976.0	NaN	2676.
2	20210307	AR	324818.0	69092.0	2480716.0	NaN	totalTestsViral	2736442.0	335.0	14926.0	141.0	Nal
3	20210307	AS	0.0	NaN	2140.0	NaN	totalTestsViral	2140.0	NaN	NaN	NaN	Nal
4	20210307	AZ	826454.0	56519.0	3073010.0	NaN	totalTestsViral	7908105.0	963.0	57907.0	273.0	Nal

Figure 4.2: Organized dataset

This study renders the aim of COVID-19 predictions also known as novel coronavirus. Causing tens of thousands of deaths and death rates increasing drastically each day throughout the globe has proved itself as a potential threat to humankind. To contribute to this pandemic, this study attempts to perform future forecasting of the newly infected cases, death rates, daily number of cases and the recovery each day.

The workflow starts by the pre-processing of dataset followed by data Splitting and Testing Set. The training set derived from Data Splitting uses the models LR, SVM, LASSO and ES. Whereas, the trained model created using different supervised machine learning models will now use the evaluation parameters R-square score, Adjusted R-square, MAE, MSE and RMSE for accurate predictions.

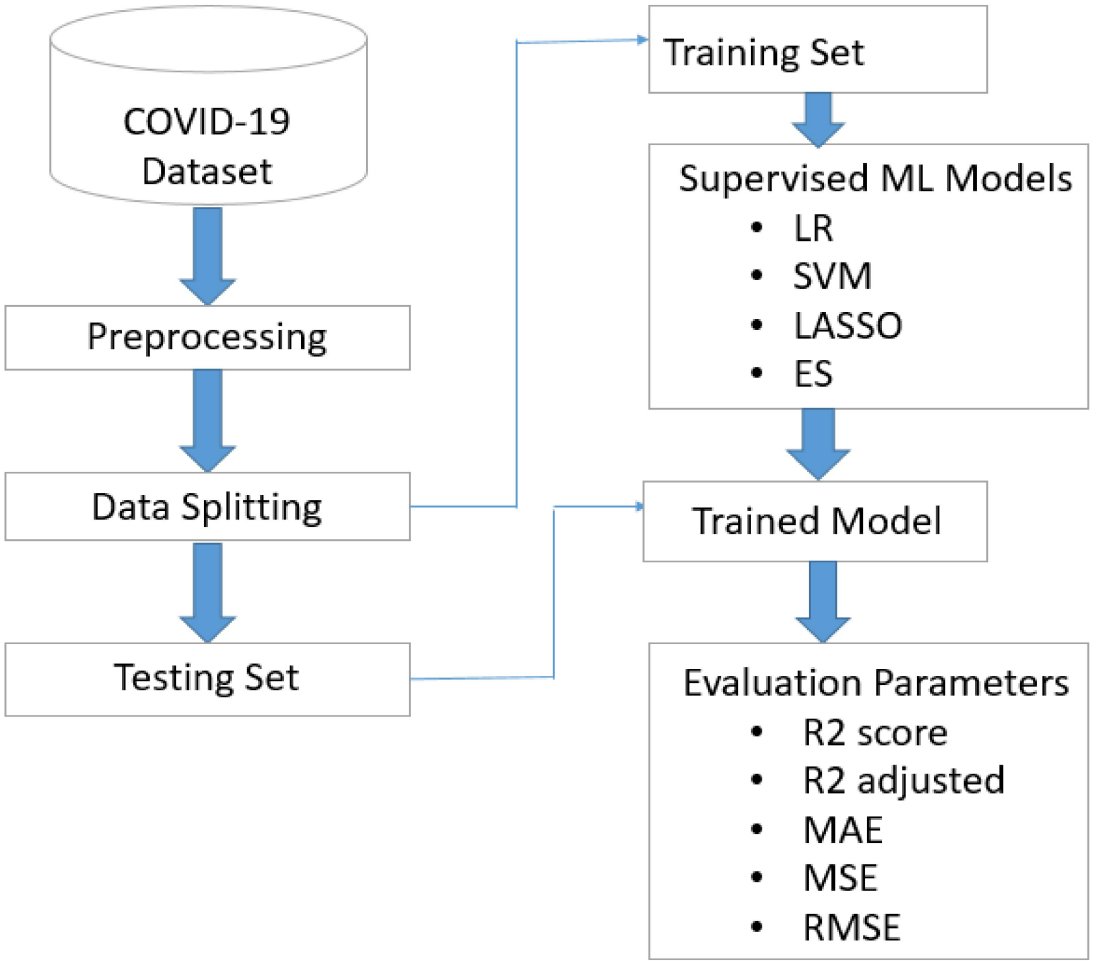

Figure 4.3: Workflow diagram

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

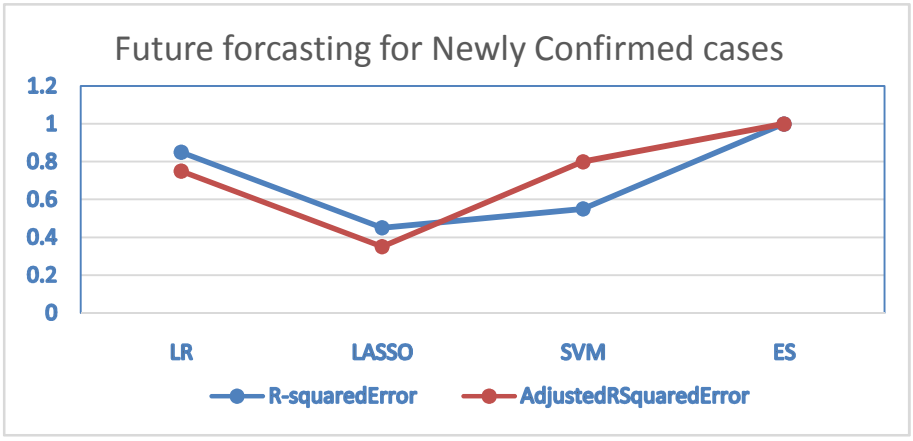

Figure 4.4: Forecasting on the newly confirmed cases

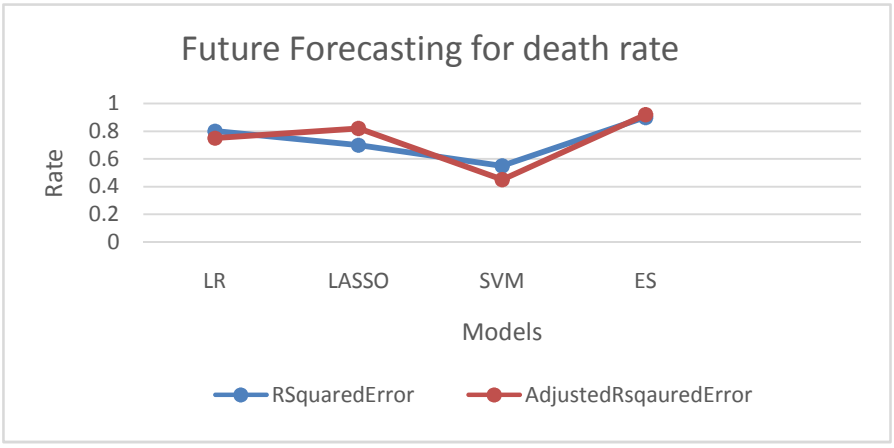

Figure 4.5: Forecasting for the death rate

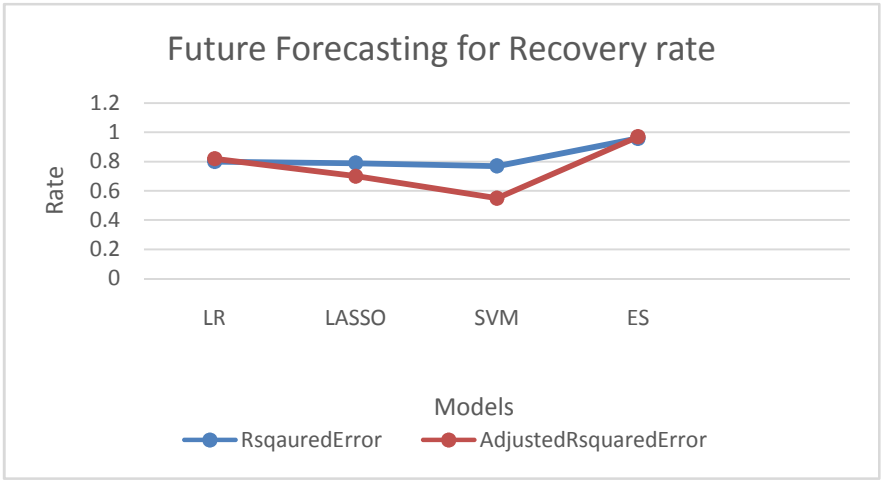

Figure 4.6: Forecasting for the recovery rate

V. CONCLUSION

The randomness of the COVID-19 pandemic can cause a massive global crisis. Throughout the world some researchers have found that the pandemic can affect large proportions of the mankind. In this study, a prediction system working on ML algorithms has been proposed for the historical as well as future predictions of COVID-19 pandemic.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 1, January 2022
Impact Factor: 5.731

The system analyses dataset containing the past data and makes predictions for upcoming situation using machine learning algorithms. Above results prove that ES performs best in the current forecasting domain followed by LR and LASSO in prediction of death rates and confirm cases respectively. Overall we assume that the model prediction accurate according to the dataset values which may prove to be a guiding light towards better decision making throughout.

ACKNOWLEDGMENT

We would like to express our deep and sincere gratitude to our coordinator, guide, faculty, HOD Principal of Department of Computer Engineering SPPU for their unflagging support and continuous encouragement throughout the project work. Without their guidance and persistent help this would not have been possible.

REFERENCES

[1]. G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, "Machine learning strategies for time series forecasting," in Proc. Eur. Bus. Intell. Summer School. Berlin, Germany: Springer, 2012, pp. 62–77.

[2]. 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 Treat. Rep., vol. 69, no. 10, pp. 1071–1077, 1985.

[3]. P. Lapuerta, S. P. Azen, and L. Labree, "Use of neural networks in predicting the risk of coronary artery disease," Comput. Biomed. Res., vol. 28, no. 1, pp. 38–52, Feb. 1995.

[4]. K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, "Cardiovascular disease risk profiles," Amer. heart J., vol. 121, no. 1, pp. 293–298, 1991.

[5]. F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus COVID 19," PLoS ONE, vol. 15, no. 3, Mar. 2020, Art. no. e0231236.

[6]. 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, vol. 323, no. 16, p. 1545, Apr. 2020.

[7]. WHO. Naming the Coronavirus Disease (Covid-19) and the Virus That Causes it. Accessed: Apr. 1, 2020. [Online]. Available: https:// www.who.int/emergencies/diseases/novel coronavirus-2019/technic% alguidance/ naming-the-coronavirus-disease-(covid-2019)- and-the-virusthat-cau%ses-it

[8]. C. P. E. R. E. Novel, "The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (Covid-19) in China," Zhonghua Liu Xing Bing XueZaZhi= ZhonghuaLiuxingbingxueZazhi, vol. 41, no. 2, p. 145, 2020.

[9]. M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Analytics defined," in Information Security Analytics, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston, MA, USA: Syngress, 2015, pp. 1–12. [Online]. Available: http://www. sciencedirect.com/science/article/pii/ B9780128002070000010.

[10]. X. F. Du, S. C. H. Leung, J. L. Zhang, and K. K. Lai, "Demand forecasting of perishable farm products using support vector machine," Int. J. Syst. Sci., vol. 44, no. 3, pp. 556–567, Mar. 2013.

[11]. J. Lupón, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galán, E. Zamora, J. Vila, J. Peñafiel, A. Urrutia, E. Ferrer, N. Vallejo, J. L. Januzzi, and A. BayesGenis, "Biomarker-assist score for reverse remodeling prediction in heart failure: The ST2-R2 score," Int. J. Cardiol., vol. 184, pp. 337–343, Apr. 2015.