A Comparative Study of Machine Learning Models for COVID-19 prediction in India

Vartika Bhadana
Dept. of Computer Engineering and Applications
GLA University
Mathura - INDIA
vartika.bhadana_mtcs19@gla.ac.in

Anand Singh Jalal
Dept. of Computer Engineering and Applications
GLA University
Mathura- INDIA
asjalal@gla.ac.in

Pooja Pathak
Dept. of Mathematics
GLA University
Mathura- INDIA
pooja.pathak@gla.ac.in

***Abstract*—Machine learning is commonly being used in every field. Forecasting systems based on machine learning (ML) have shown their importance in interpreting perioperative effects to accelerate decision-making on the potential course of action. In several technology domains, ML models have been used long to define and prioritize adverse threat variables. To manage forecasting challenges, many prediction approaches are widely used. The paper shows the ability of ML models to estimate the amount of forthcoming COVID-19-affected patients that is now considered a serious threat to civilization. In paper this, we have performed a comparative study of five machine learning standard models like Linear regression (LR), decision tree, least absolute shrinkage and selection operator (LASSO), random forest and support vector machine (SVM) to forecast the threatening variables of COVID-19. Each of the models makes three forms of forecasts, i.e. the total active cases, the total deaths, and the total recoveries in the next five days. The findings provided by the paper suggest that the use of these techniques for the current COVID-19 the pandemic scenario is a promising strategy. For better accuracy, we have used a six-degree polynomial. Experiment results illustrate that poly LR and poly LASSO gives the best results followed by LR, LASSO, random forest, and decision tree. SVM shows the poor result in the prediction of COVID-19.**

ey ords— Ma hine learning models, COVID-1 India, future fore asting, polynomial features.

I. INTRODUCTION

Coronavirus is an RNA virus that was caused by birds and mammals [d]. The novel coronavirus was originated from SARS-COV-d (severe acute respiratory syndrome). The virus was first originated in Wuhan, China in December d0d9. The virus spread all over the world in a very short period of time [d]. The latest virus is really infectious and has rapid progress of spread all over the world. The WHO announce this as epidemic a public health emergency of international concern (PHEIC) as it extended to nearly d9 nations before d0 January d0d0 [d]. The emergence of this new virus in India, person-to-person spread include the families and medical workers was observed [4]. As the whole world is suffering from this pandemic situation India is one of the third countries which is affected by COVID-d9 [5]. Every day, many new people are recorded as positive in India. The virus spreads mainly by people close to each other through physical touches, by sneeze, through holding the infected areas. The very difficult feature of the propagation is that for several days a human will carry the virus without displaying any symptoms. Health experts around the world are actively trying to identify a safe vaccination and treatments for the disease.

In India, the first case came on 30th January 2020 in Kerala. After that till March 2020 cases were only 200, to overcome the pandemic a lockdown was planned for 21 days. As unlock 1.0 came on 1st June 2020 the cases started increasing rapidly. In recent years, machine learning (ML) has shown to be a leading technology of research by solving so many complicated and complex actual problems. The fields of research include nearly all actual realms such as medical, independent vehicles (AV), business solutions, Natural Language processing (NLP), smart robotics, games, environment simulation, speech and image recognition. ML algorithm's is usually based on hit and trial approaches very unlike standard algorithms, its totally based on programming like if-else rules [6].

The objective of this study is to predict the cases for the next 5 days that will be helpful for the doctors and government for preparing their plans. Ramjeet et al. [7] utilized the six regression techniques that are quadratic, three-degree, four-degree, five degree, exponential and sixth degree polynomial. The sixth polynomial regression model shows the less values of best fit model as compare to another model.

Ahmad et al. [8] present the classification that groups them into four categories. The four themes are as follows: deep learning regression, regression of conventional machine learning, system design, based on social media and queries-data.

The Car et al [9]. used a public dataset and this dataset is time-series data that is been converted to the regression dataset to use the Multilayer perceptron (MLP) which is feed-forward of Artificial Neural networks (ANN). The R2 shows different values for confirmed (0.98599), recovered (0.97941), and deceased (0.99429). Future work is to apply other models then MLP.

Of all the precautions, "being informed" and among all the facts of COVID-19 is deemed to be important. Contributing to the global human tragedy is our effort in this analysis to build a COVID-19 forecasting framework. The forecasting is done on three bases: 1) confirmed cases 2) recovered cases 3) death cases for the upcoming 5 days. The problem is basically

978-0-7381-2447-6/20/$31.00 © 2020 IEEE

basing on supervised machine learning techniques. The study states the regressions model such as Least Absolute Shrinkage and Selection Operator (LASSO), Decision tree, Linear Regression, Random Forest, and polynomial regression. The prediction of the patient status is done from the dataset taken from Covidindia.org. the data is been pre-processed and divided into training and testing sets. Significant measures like the R-squared score (score), Mean absolute error (MSE), Mean Square Error (MAE), and root means square error (RMSE) was used to assess the result.

The rest paper contains 5 sections: Section 1 is about Introduction; section 2 contains material and methods. Section 3 contains methodology of paper; the outcome is outlined in section 4 and lastly section 5 summarizes the paper and conclusion.

II. MATERIAL AND METHOD

A. Supervised machine learning model

When an unknown instance of feedback is given, a supervised model of learning is built to make a decision [10]. In supervised learning, we practice the programme using a dataset that has been classified, indicating that this knowledge is identified with the right answer. Through labelled training outcomes, a supervised learning algorithm learns and helps one to forecast unexpected data effects. [11]. For predictive model growth, this paper uses regression models.

We have used six regression models to forecast the study of COVID-19:

- Least Absolute Shrinkage and Selection Operation (LASSO)
- Random forest
- Decision tree regressor
- Linear regression
- Support vector machine
- Polynomial regression

1. LASSO

Lasso is a multivariate statistical regression analysis tool that works on attribute selection and validation in addition to enhancing prediction accuracy and evaluate the ability of a mathematical model generated by it [12]. It was initially designed for linear regression analysis and this basic case shows a considerable amount about the estimator's behaviour, includes its relationship to ridge regression and best collection of subsets, interactions between lasso coefficients and so-called soft threshold.

The 'lasso' reduces the number of squares, equivalent to the number of fixed value of the coefficients which is low. [13]. It tends to generate certain coefficients that are exactly zero due to the existence of this restriction and thus gives interpretable models. It is the most stable regression and the shrinkage make LASSO better and reduces the error. That implies that LASSO regression works in order to minimize the following;

$$\sum_{k=1}^{n}(ak-\sum_{m}bkm\beta m)^2+\lambda\sum_{m=1}^{q}|\beta m| \quad (1)$$

The coefficient is set where λ is a concept of penalty that is minimum squares of residuals sum), $(ak-\sum_m bkm\beta m)^2$ is the residual sum of squares, and $|\beta m|$ is the sum of absolute value.

2. Support Vector Machine (SVM)

It is a very particular category of algorithms that are characterized by decision function energy storage, use of kernel functions, and solution sparsity [14]. System of machines are SVMs, which means that a regression problem is solved by the function. The input vector, x, when dealing with non-linear regression, is converted into a high-dimensional feature space by a nonlinear transformation, and then linear regression is done in that space [15]. Putting the definition as a series of observed responses in the sense of ML with a multivariate training data with N number of observations [16]:

$$g(x)=y'\beta+c \quad (2)$$

y' is the how far the value is, β is slope or gradient function, c is value of x when y' is zero.

3. Decision Tree Regressor

Strategies for the creation of discriminant analysis, also named classification and regression problems trees, have been established over the past twenty years [17]. Researchers in the machine-learning have been developing methods for immediately stimulating decision trees from data sets [18]. A decision tree, which is effective method to solve classification and regression problems. Building a regression tree is also focused on binary recursive partitioning that is an ongoing process that divides into partitions. [19].

4. Linear Regression

The most functional statistical methodology is a form of regression modelling, for the analysis of machine learning. [20]. It establishes causal link between the contingent and independent causes. In regression modelling, the independent characteristics are predicated on a target class [21]. To estimate a value for a dependent variable (y), linear regression performs a task based on a given independent variable (x). Thus, this method, a linear association is observed between x (input) and y (output). Equation indicates, then, how y is connected to x:

$$y=\theta_1+\theta_{2.x}+\varepsilon \quad (3)$$

Here, θ_1 represents the intercept where as θ_2 represents the coefficient of x, ε is define as error term. The machine learning algorithm aims to find the right values to get the best-fit regression line for (intercept) and (coefficient). The discrepancy between the true values and expected values should be small to get the best match, so this issue of minimization can be expressed as:

$$\text{minimise}\frac{1}{n}\sum_{i=1}^{n}(pred_i-y_i)^2 \quad (4)$$

$$h = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \qquad (5)$$

here, h is the cost function of linear regression in the root mean square (RMSE) between predicted value ($pred_i$)and real value (y_i), the total number of data points is denoted by n.

5. Random forest

It's a evaluation which compares multiple decision trees for numerous datasets sub- samples and average to boost prediction results is used and unnecessary power. The algorithm for ensemble learning that integrates a broad variety of regression trees is the RF regression algorithm. A regression tree describes a collection of hierarchical requirements and constraints that are extended from root to tree leaf. [22].

6. Polynomial regression

It is the component of artificial intelligence technology, in different meaning, it's a set of algorithms for machine learning is regression analysis. The research includes a set of computer vision which require a constant outcome variable (Y) dependent on the value of one or more predictor variables to be predicted (X) [23]. We have used the six-degree polynomial which is been used to predict the better accuracy of the model.

B. Methodology

This study is about the COVID-19 forecast also known as a novel coronavirus. In human life, COVID-19 was a threat. The rate of death is raising every day in India, contributing to thousands of deaths. In order to resolve this pandemic power, the confirmed, and active cases still rise day by day, this study aims to estimate death rates, the number of active cases and, cases of confirmed and the total that are recovered in the coming days. Prevision is produced by the use of approaches to machine learning that are beneficial in this respect. The dataset reports the day the pandemic began, confirmed cases, recovered cases, deaths, and active situations. To find the details of known cases, rescued cases, deaths cases, the data is initially pre-processed.

After the pre-processing step is done, the data is parted in two sets: a training set for training model, and for testing model test set is used. Learning models used are LASSO, SVM, decision tree, random forest regression, Linear regression. The techniques are trained randomly. Models are then evaluated with the help of the following parameters like R^2 score, mean absolute Error, Root Mean Square Error, mean square error, as shown in the results. Fig.1 shows the proposed work.

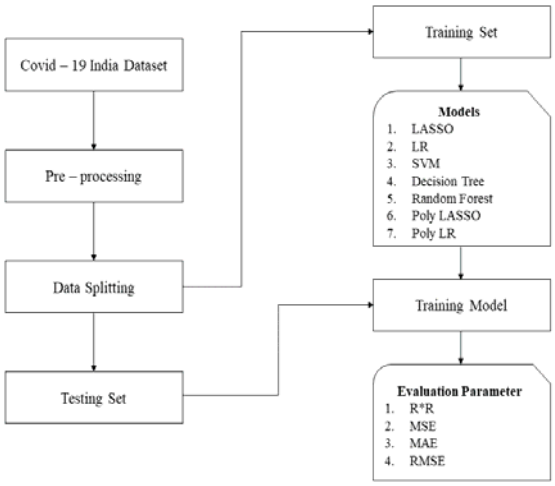

Fig.1. Proposed flow diagram

III. RESULT AND DISCUSSIONS

A. Dataset

Objective is to know about the cases, the total deaths, and total reported cases. Api.covid19india.org which is the official website of India from which data is taken [24]. The attributes of dataset contain in form of announced date, patient number, state code, entry id, age, gender, number of cases, current status, detected city, detected state. The data is taken from the 30th of January 2020 to 8th October 2020. The labels which is used for training the models are num cases, announced date, age, detected city, detected district, detected state, gender and current status. The Fig.2. shows number of detected cases in particular state.

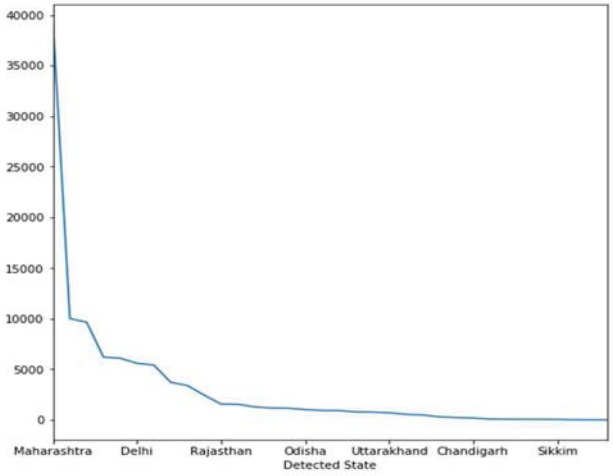

Fig.2. Graph of states in which cases are detected

B. Evalution Parameter

For this study, we evaluate each learning model's performance in terms R^2 score, RMSE, MAE, and MSE.

1. R squared score

The score is a statistics test used in the regression models' performance estimates. The numbers indicate the percentage of variation of the variable that determines the independent variable jointly [25], [26]. It tests on an accurate scale from 0 to 100 percent the strength of the relationship of the dependent's component to the regression model.

0 percent score means that the response variable does not have any uncertainty in the mean and 100 percent means that there is all the confusion about the reaction element. Its equation is as follows:

$$R^2 = \frac{\text{variance of model}}{\text{total variance}} \qquad (6)$$

2. *Mean Square Error (MSE)*

In mathematics, it estimates to determines the mean response time of the error and average square difference of the predicted values and real values [27]. Mean Square Error is a risk function that corresponds to the expected value of the loss of square error. The reason that Mean Square Error is almost always purely positive is due to randomness or that the estimator does not take into account details which might yield a more precise calculation. The bigger the MSE means the closest you are to finding the right match side.

$$\text{MSE} = \frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{y_j})^2 \qquad (7)$$

m observed points, actual value is y_j, and predicted value is $\hat{y_j}$.

3. *Mean Absolute Error (MAE)*

It is commonly used in model assessments [27] [28]. It is the statistics is a calculation of errors that reflect a certain performance. Y versus X provide measurements of expected vs. actual, concurrent time versus original time, reliability method vs. an alternate [26].

$$\text{MAE} = \frac{1}{m}\sum_{i}^{m}\left|y_i - \hat{y_i}\right| \qquad (8)$$

the number of data points is m, $|y_i - \hat{y_i}|$ is the absolute value of residual.

4. *Root Mean Square Error (RMSE)*

It's the standard residual deviation (prediction errors). RMSE is an example of how these variables are removed. Root mean square error is usually shown in climate science, prediction, and linear regressions to verify the research results. It is an error rate define as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p-a)^2} \qquad (9)$$

n is the observed points, estimated value is p, and actual value is a.

C. Result

This paper aims to build a framework using machine learning techniques for potential predictions of cases caused from coronavirus. The analysis provides details on regular estimates of total of active person, the total recovered cases, and death occurred from coronavirus in India. The way total death and active cases increase day by day which is a worrying condition for India.

This research is an attempt to estimate the number of human beings who could be affected by new infection person, death and the total predicted recovery for upcoming 5 days. For estimating total new active, total of death and total of recoveries, five machine learning models LR, Least Absolute shrinkage and selection Operation, SVM, random forest and a decision tree is used.

1) Active Cases Forecasting:

As in India the active COVID-19 cases are increasing every day, the result of cases is shown in table 1 below. In this decision tree shows the best R^2 value, followed by random forest, and polynomial LR gives a better result than poly LASSO, LR and LASSO. SVM gives the poor result of it. Table 1 shows the performance of the model.

Table: 1 EFFICIENCY OF MODELS FOR FUTURE ACTIVE PREDICTIONS

Models	R^2	MSE	MAE	RMSE
Poly LASSO	98.15	20089397.86	3365.81	4482.12
Poly LR	98.64	14754167.34	2191.13	3841.12
Decision tree	100	0.00	0.00	0.00
Random forest	99.9	1309454.26	610.9	1144.31
SVM	8.366	1179969500.2	26343.17	34350.6
LR	86.90	142215249.58	10225.9	11925.2
LASSO	87.20	142215249.58	10225.96	11925.4

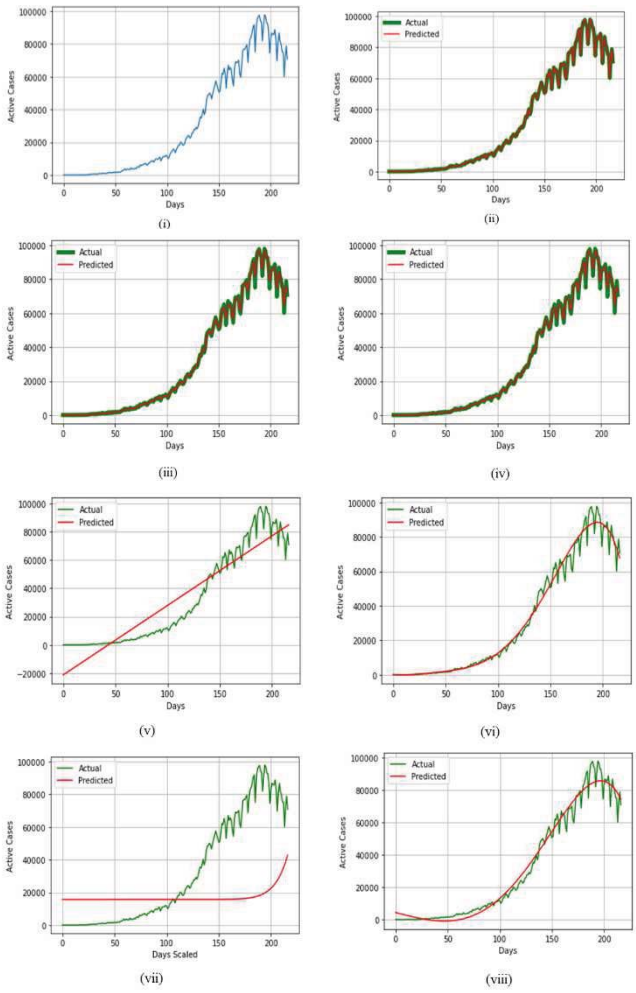

Fig.3. Shows the graph of (i) Total Active Cases, (ii) active case prediction by Decision Tree, (iii) active cases prediction by Random Forest, (iv) active cases prediction by LR (v) active cases prediction by LASSO, (vi) active cases prediction by poly LR, (vii) active cases prediction by SVM, (viii) active cases prediction by poly LASSO.

2) *Death rate Forecasting*

Prediction of total death occurred and from the prediction the best result is given by decision tree and random forest performs equal results achieve the approx. same R^2 value but they are been overfitted. Poly LASSO and poly LR shoes the next best result. SVM performs poor results. Table 2 shows the result:

Table: 2 EFFICIENCY OF MODELS FOR FUTURE DEATH PREDICTIONS

Models	R^2	MSE	MAE	RMSE
Poly LASSO	89.49	18839.5	54.3	137.22
Poly LR	89.59	18657.7	55.39	136.59
Decision tree	100	0.0	0.0	0.0
Random forest	98.719	3236.82	20.16	62.87
SVM	10.29	160056.1	341.17	400.0
LR	86.61	24002.19	86.85	154.93
LASSO	86.42	24002.19	86.85	154.62

The Fig.4. (i, ii, iii, iv, v, vi, vii, viii) shows the results of death rate with the ML approaches. The graphs show the increase in death rate day by day which is a very serious situation in India.

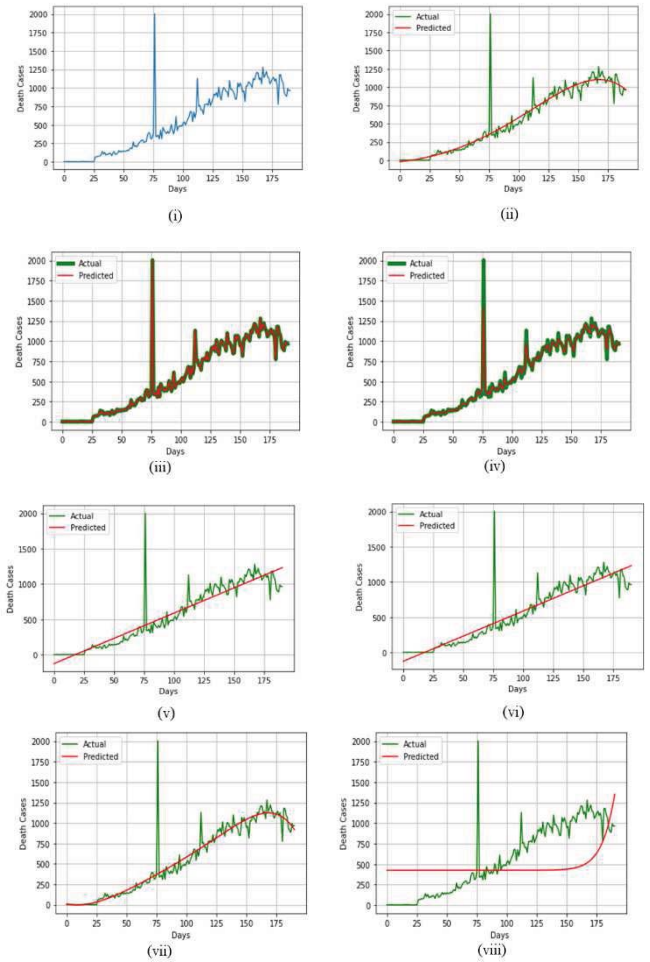

Fig.4. Shows the graph of (i) Total Death Cases, (ii) death case prediction by poly LASSO, (iii) death cases prediction by Decision Tree, (iv) death cases prediction by Random Forest (v) death cases prediction by LR, (vi) death cases prediction by LASSO, (vii) death cases prediction by poly LR, (viii) death cases prediction by SVM.

3) *Recovery rate forecasting*

The prediction of recovery rate result is shown in the table, the best one is decision tree and random forest but they both are been over fitted, then the best result is poly LR which is followed by poly LASSO, Linear regression, LASSO, and SVM. Fig.5. (i, ii, iii, iv, v, vi, vii, viii) shows the prediction graphs. Table 3 shows the models forecasting results:

Table 3: EFFICIENCY OF MODELS FOR FUTURE RECOVERY PREDICTION

Models	R^2	MSE	MAE	MSE
LASSO	87.12	123656340.7	9385,07	11120.09
LR	87.54	123656340.1	9385.07	11120.09
Decision tree	100	0.0	0.0	0.0
Poly LASSO	98.38	16045032.92	2882.65	4005.63
SVM	-7.30	1066758218.3	24852.9	32661.2
Poly LR	98.62	13621960.52	2278.28	3690.79
Random forest	99.75	1323532.34	717.23	1150.45

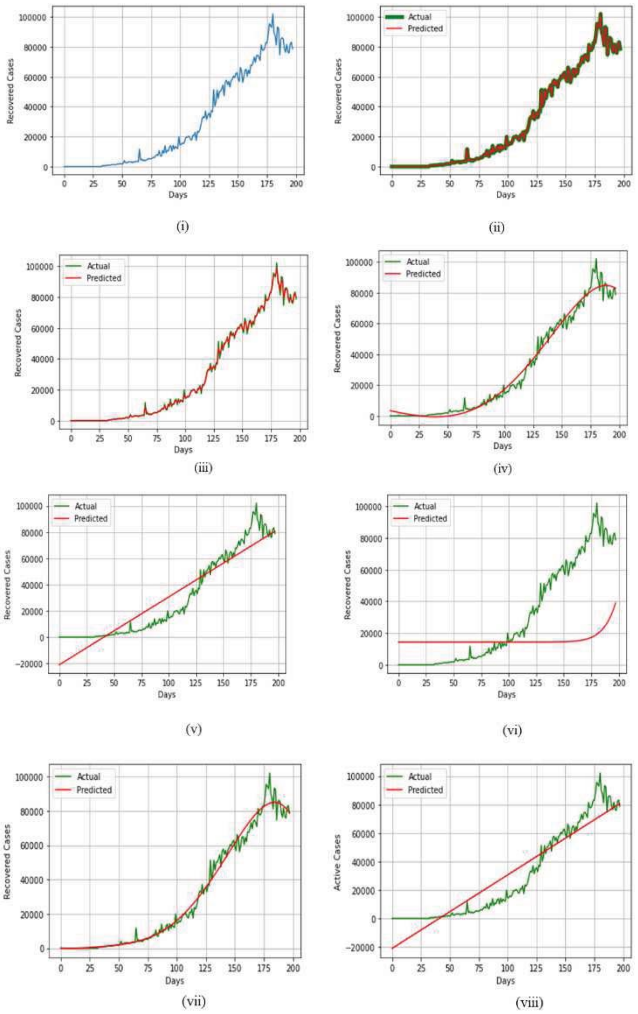

Fig.5. Shows the graph of (i) Total Recovered Cases, (ii) recovery case prediction by Decision Tree, (iii) recovery cases prediction by Random Forest, (iv) recovery cases prediction by poly LASSO(v) recovery cases prediction by LR, (vi) recovery cases prediction by SVM, (vii) recovery cases prediction by poly LR, (viii) recovery cases prediction by LASSO.

IV. CONCLUSION

A major environmental crisis could be sparked by the hazardous of the COVID-19 pandemic. Researchers and government organizations around the world are concerned that a significant percentage of the world's population will be impacted by the pandemic. In this study, we have used ML approaches for the prediction of the COVID-19 outbreak globally. The dataset contains the past data, and the data is in the day, month, and year format and make the prediction for upcoming days. The best result in the active, death and recovered cases is given by polynomial LASSO and polynomial LR. The other models also show good results in the forecasting of the data like LR, LASSO, Decision tree, Random forest. Death rate, and recovery rate will increase in the upcoming days, according to the effect of these two models Polynomial LASSO and polynomial LR. SVM shows the poor result among along because of them up, and down in dataset values, it difficult to divide the values by a hyperplane. Overall, we assume that, according to current model forecasts, Scenarios that could be useful to consider the future situation are right. Therefore, the research projections can be a great benefit for the people to take appropriate steps, and to take rights to manage the crisis of COVID-19. In this work, the two models Decision tree and Random forest do not show the promising result as they are been overfit.

REFERENCES

[1] Wikipedia contributors. "Coronavirus." *Wikipedia, The Free Encyclopedia*. Wikipedia, [online] The Free Encyclopedia, 24 Sep. 2020. Web. 29 Sep. 2020.

[2] Z. Cakir, and H. B. Savas, "A Mathematical Modelling Approach in the Spread of the Novel 2019 Coronavirus SARS-CoV-2 (COVID-19) Pandemic", Electron J Gen Med, vol 17, pp. 1-3, 2020.

[3] N.S Punn, S. K Sonbhadra, and S. Agarwal, "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms," *medRxiv preprint,* 2020.

[4] S. B. Stoecklin, P. Rolland, Y. Silue, A. Mailles, C. Campese, A. Simondon, "First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures", *Eurosurveillance* 25, no. 6, 2020.

[5] The Hindu, 'coronavirus news', 2020. [online]. Available:https://www.thehindu.com/news/national/coronavirus-indias-rise-in-cases-third-fastest globally/article31829312.ece. [Accessed: 26-Sep-2020].

[6] S. Makridakis, E. Spiliotis, and V. Assimakopoulos," Statistical and machine learning forecasting methods:Concerns and ways forward.", PloS one13.3, vol.18, pp. 1-26, 2018

[7] R. S Yadav, "Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India.", *International Journal of Information Technology,* 2020, pp. 1-10.

[8] A. Ahmad, S Garhwal, S K Ray, G Kumar, S J Malebary, and O M Barukab, "The number of confirmed cases of covid-19 by using machine learning: Methods and challenges.", *Archives of Computational Methods in Engineering*, 2020, pp. 1-9.

[9] Z Car, S Baressi Šegota, N Anđelić, I Lorencin, and V Mrzljak, "Modeling the Spread of COVID-19 Infection Using a Multilayer Perceptron.", *Computational and Mathematical Methods in Medicine,* vol 20, pp. 1-10, 2020.

[10] S .A. Diwani, and A. Sam, "Diabetes forecasting using supervised learning technique.", *Advances in Computer Science: An International Journal* 3.5, pp. 10-18, 2014.

[11] H. A Guvenir, B. Acar, G. Demiroz, and A. Cekin, "A supervised machine learning algorithm for arrhythmia analysis.", In *Computers in Cardiology*, pp. 433-436, 1997.

[12] R. Tibshirani, "Regression shrinkage and selection via lasso.", Journal of the royal statistical society: Series B 58.1, pp. 267-288, 1996.

[13] A.E. Hoerl, and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems.", techno metrics 12, no. 1, pp. 55-67, 1970.

[14] W. Huang, Y. Nakamori, and SY Wang, "Forecasting stock market movement direction with support vector machine.", *Computers & operations research* ,32.10, 2005, pp. 2513-2522

[15] X. F. Du, SCH Leung, J.L. Zhang, and K.K. Lai," Demand forecasting of perishable farm products using SVM.", International Journal of system science 44, no. 3, 2013, pp.556-567.

[16] F. Rustam, A.A Reshi, A. Mehmood, S. Ullah, B. On, and G. S. Choi," COVID-19 future forecasting using supervised machine learning models.", IEEE access, vol.8, pp. 101489-101499,2020.

[17] L. Breiman, "Classification and Regression Trees, Wadsworth International Group, Belmont, California", *Google Scholar,* 1984.

[18] W.J Long, J L. Griffith, H. P. Selker, and R.B. D'agostino, "A comparison of logistic regression to decision-tree induction in a medical domain.", *Computers and Biomedical Research*, vol. 26, no. 1, 1993, pp. 74-97.

[19] M. Xu, P. Watanachaturaporn, P. K. Varshney, and M. K. Arora, "Decision tree regression for soft classification of remote sensing data.", *Remote Sensing of Environment* 97, no. 3, 2005, pp. 322-336.

[20] D. Sztahó, G. Kiss, and K. Vicsi, "Estimating the severity of Parkinson's disease from speech using linear regression and database partitioning.", *Sixteenth Annual Conference of the International Speech Communication Association*, pp. 498-502, 2015.

[21] H. L. Hwa, WH Kuo, LY Chang, MY Wang, TH Tung, KJ Chang, and FJ Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumor markers using logistic regression models.", *Journal of evaluation in clinical practice* 14, no. 2, 2008, pp. 275-280.

[22] X Zhou, X. Zhu, Z Dong, and W Guo, "Estimation of biomass in wheat using random forest regression algorithm and remote sensing data.", *The Crop Journal* 4, no. 3, 2016, pp. 212-219.

[23] T Chai, and RR. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE) Arguments against avoiding RMSE in the literature.", *Geoscientific model development* 7.3, 2014, pp. 1247-1250.

[24] @misc covid19indiaorg2020tracker, author COVID-19 India Org Data Operation Group {Accessedon10.09.2020} Available: https://api.covid19india.org, 2020.

[25] J. Lupon, H. k. Gaggin, M De Antonio, M. Domingo, A. Galan, E. Zamora," Biomarker- assist score for reversers

modeling prediction in heart failure: the STR-R2 score.", International journal of cardiology 184, 2015, pp. 337-343.
[26] J. H. Han, and SY Chi, "Consideration of manufacturing data to apply machine learning methods for predictive manufacturing.", In *2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN) IEEE*, pp. 109-113, 2016.
[27] C. J Willmott., and K Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.", *Climate research* 30.1, 2005, pp. 79-82.
[28] R Kaundal, A S. Kapoor, and GPS Raghava, "Machine learning techniques in disease forecasting: a case study on rice blast prediction.", *BMC bioinformatics* 7, no. 1, 2006 pp. 485.