Prediction of growth in COVID-19 Cases in India based on Machine Learning Techniques

Aindrila Saha
Computer Science and Engineering
NIT Rourkela
Rourkela, India
aindrila.saha.kgec@gmail.com

Vartika Mishra
Computer Science and Engineering
NIT Rourkela
Rourkela, India
vartikamishra151@gmail.com

Santanu Kumar Rath
Computer Science and Engineering
NIT Rourkela
Rourkela, India
skrath@nitrkl.ac.in

***Abstract*—One of the biggest health challenges that the world has faced in recent times is the pandemic due to coronavirus disease known as SARS-CoV-2, or Covid-19 as officially named by the World Health Organization (WHO). To plan medical facilities in a certain location in order to combat the disease in near future, public health policy makers expect reliable prediction of the number of Covid-19 positive cases in that location. The requirement of reliable prediction gives rise to the need for studying growth in the number of Covid-19 positive cases in the past and predicting the growth in the number in near future. In this study, the growth in the number of Covid-19 positive cases have been modelled using several machine learning based regression techniques viz., Multiple Linear Regression, Decision Tree Regression and Support Vector Regression. Further, different feature selection techniques based on Filter and Wrapper methods have been applied to select the suitable features based on which prediction is to be done. This study proposes the best observed method for modelling the pattern of growth in number of Covid-19 cases in the near future for a locality and also the best selection method that can be employed for obtaining the optimal feature set. It has been observed that unregularized Multiple Linear regression model yields promising results on the test data set, compared to the other regression models, for predicting the future number of Covid-19 cases and Backward Elimination feature selection method performs better than other feature selection methods.**

***Index Terms*—Covid-19, coronavirus, regression analysis, feature selection**

I. Introduction

Coronavirus is a Ribo Nucleic Acid (RNA) virus that reportedly started in Wuhan, China during December, 2019 and spread all over the world by human-to-human transmission in an extremely short period of time [1]. By February 2020, it had taken the form of a pandemic. Since the start of the pandemic, various studies have been carried out and approaches have been proposed in order to model the growth in the number of Covid-19 positive cases. Epidemiological models such as the Susceptible-Exposed-Infectious-Recovered-model (SEIR model) are the mathematical ones to model the spread of infectious diseases [2]. Machine learning techniques are also being used and proposed to study epidemiological data. Studies have been done on global data, country level data and regional data. This study is focused on epidemiological data of India; India is one of the countries that have been worst hit by coronavirus [3]. The first case in India was reported on 30^{th} January, 2020 in Kerala. Thereafter there have been deadly waves of the coronavirus, taking the total number of people infected till 30^{th} October 2021 to 3,42,73,300 and the death toll standing at 4,58,186 [4]. Currently the situation is seemingly under control as 1,06,14,40,335 total doses of the vaccines Covaxin and Covishield, which includes first doses and second doses, have been administered till 30^{th} October, 2021 [4]. The rate of spread of coronavirus and the amount of impact it causes depends on several factors and therefore varies from region to region, country to country [5]. So, it is important to study the epidemiological data for a country, considering the factors specific to that country/location in order to understand the growth of the coronavirus in that country/location. In Fig. 1, the number of Covid-19 positive cases over four different time periods in between 1^{st} February, 2020 and 1^{st} November, 2021 in India have been plotted. Data of India for the period (1^{st} February, 2020 - 1^{st} November, 2021) have been analysed on the basis of features which have been found to have most impacted the growth in the number of Covid-19 cases in India.

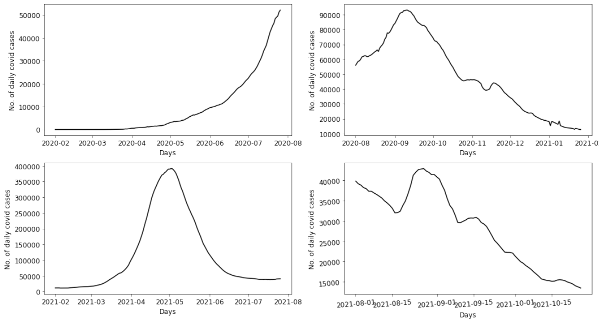

Fig. 1: Covid cases in India from February'20 - October'21

In this study, for the purpose of selecting optimal feature set from all the features present in the available dataset, two Filter methods namely Mutual Information (MI) score and Correlation Coefficient (CC) and two Wrapper methods namely Forward Feature Selection (FS) and Backward Feature Elimination (BE) been considered. Machine learning models based on Multiple Linear Regression, Decision Tree Regression and Support Vector Regression have been trained with the data to learn the growth of coronavirus. Regularization techniques such as Ridge, Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net have been applied on the Multiple Linear Regression model. Ensembling methods based on decision trees viz., Random Forest Regressor, Adaptive Boosting (AdaBoost) and Gradient Boosting have been implemented. For Support Vector Regression, multiple kernel functions viz., linear, poly, sigmoid and RBF (Radial Basis Function) have been used. A comparative analysis has been conducted between the performance of the various regression models on different feature sets from the dataset using different evaluation metrics.

This paper is organized as follows: Section 1 and Section 2 describe the introduction and literature review of few of the existing works on growth in number of Covid-19 positive cases respectively. Section 3 contains the description of the dataset. Section 4 addresses on the aspects of feature selection techniques, machine learning based regression techniques, time series analysis techniques applied on the dataset and the proposed methodology. Section 5 illustrates the results that have been obtained by performing the several machine learning techniques related to feature selection and regression analysis. Lastly, Section 6 summarizes the conclusion of the study.

II. Literature Review

An attempt has been made to make a comprehensive study on prediction of growth pattern in Covid-19 cases made by authors as available in literature. Serkan Balli [6] has analysed data of Covid-19 between 20^{th} January 2020 and 18^{th} September 2020 for Germany, USA and the global. He has proposed Time Series technique to model the pandemic curve and has predicted the same using machine learning methods. Results showed that Support Vector Machine (SVM) outperformed the other techniques applied such as Linear regression, Random Forest and Multi-Layer Perceptron. For the purpose of Covid-19 wave prediction, Amir Ahmad et al. [7] have categorized machine learning methods into four groups namely traditional machine learning regression, deep learning regression, network analysis and social media and search queries data based methods. Milind Yadav et al. [8] have applied linear regression model, polynomial regression model and SVM on Covid data of countries US, Mainland China, Italy, India and South Korea. SVM has been found to yield better results compared with the other regression methods. Andreou Andreas et al.[9] have compared the data predicted by linear, polynomial and exponential models using R-squared metric, to determine which one best models the data of cumulative number of infected and deaths due to Covid-19. Furqan Rustam et al.[10] have used linear regression, LASSO, SVR and Exponential Smoothing to predict the threatening factors of COVID-19. Exponential Smoothing performed the best while SVR performed poorly. Vartika Bhadana et al. [11] have performed a comparative study between Linear regression, LASSO, Decision tree, Random Forest and SVM on the data of India for the period 30^{th} January 2020 to 8^{th} October 2020. Results illustrated that poly linear regression and poly LASSO gave the best performance for predicting the number of Covid-19 cases.

From the above literature survey, some amount of research gap has been identified. Firstly, all aforementioned works are mainly based on Covid data in the year 2020. However, there have been lots of changes in the pattern of growth in the number of cases of Covid-19 thereafter and therefore, it is extremely important to study the growth patterns of the Covid-19 cases from time to time. Next, the aforementioned studies have mostly considered the number of confirmed positive cases and/or the count of days for their model training. There are several other factors such as mortality rate, recovery rate, vaccination numbers, number of tests done, etc. that should be considered for modelling epidemiological data of a country. Therefore, this study intends to overcome the research gap firstly by considering epidemiological data of year 2021 besides that of year 2020. Next, not just the number of cases of Covid-19 positive and/or the count of days, but also several other features, as aforementioned, have been used in this study to analyze the growth pattern of the Covid-19 virus.

III. Dataset considered

In order to analyse the Covid-19 data and predict the occurrences of Covid-19 cases in the nearby locality, related data has been collected from the GitHub repository [12]. This data is being provided by the Centre for Systems Science and Engineering, Johns Hopkins University. The dataset contains daily statistics about Covid-19 of various countries, since January 2020 and it gets updated every day with the latest data. The dataset contains 65 columns (features) and 125760 rows (records) with both numeric and non-numeric values. For this study, “new cases” is the target feature (whose value, model is expected to predict) and all other features are the predictor features.

Data for India have been filtered out for the study. The dataset has been sorted based on date, with the oldest data at the starting of the dataset. The ‘days’ feature has been added to the dataset to consider the number of days as a predictor for the growth in number of Covid-19 cases. For numeric data, missing values have been replaced with the median value of the feature. Mean value can also be used for this purpose, however mean is more sensitive to outliers, hence median has been chosen. Features with more than 90% null values and features with two or less unique values have been dropped from the dataset. Data for predictor features have been standardised as per min-max standardisation.

IV. Models and methods

This study intends to apply different feature selection techniques and regression analysis techniques for the prediction of Covid-19 cases of India. The following techniques have been applied in this study.

A. Feature Selection Techniques

Feature selection plays a crucial role in modelling the data, since not all the available features in a dataset contribute significantly in training the regression models. Filter and wrapper methods are broad categorizations of the feature selection methods. Filter methods [13] use the inherent properties of the features and measure them using statistical means while Wrapper methods [13] compare model predictive performance by executing the model on different subsets of the features to find out the optimal set of features.

1) Mutual Information Score: It is a filter method that measures the dependency between two features. Mutual Information score quantifies the amount of information one feature conveys about another feature. Thus, in this study, mutual information needs to be obtained in between each predictor feature and the target feature. It equals zero if and only if two features are independent, and higher the value of mutual information score, higher the dependency between two features [14].

2) Pearson Correlation Coefficient: It is a filter method that measures how strong a relationship is in between two features. If value is closer to 1 or -1, then the two features are strongly correlated. Therefore, predictor features that are more correlated with the target feature are more likely to get selected. If two features have strong correlation with the target feature then they should be uncorrelated among themselves [15].

3) Forward Feature Selection Technique: This is a wrapper method where at first every feature is trained one by one on the regression model; predictive performance of the model is evaluated and the best feature is retained. Next, the pair of features giving best predictive performance is obtained followed by the best triplet; the process is repeated with one more feature getting selected in every iteration. This process continues iteratively until a predefined maximum number of features are selected [16].

4) Backward Feature Elimination Technique: This is a wrapper method that works in a manner opposite to Forward Feature Selection Technique. In this technique, at first model is trained using all features and performance is evaluated. Next every feature is dropped one by one and model is trained and evaluated every time. The feature which on dropping from the process, model gives the least drop in predictive accuracy is eliminated from the process. With the remaining features, similar process is executed. This entire process is performed iteratively until a predefined minimum number of features are selected [16].

B. Models for regression analysis

For the purpose of modelling the growth of coronavirus, the following machine learning based regression techniques have been used.

1) Multiple Linear Regression: Multiple Linear Regression technique works on multiple predictor features and tries to capture the linear relationship between each of these features with the target feature. Regression coefficients have to be determined that minimise the difference between actual value and predicted value [17].

$$y^{(i)}_{pred} = \sum x_i \beta^{(i)}_j \tag{1}$$

and

$$Loss = \sum (y^{(i)}_{pred} - y^{(i)})^2 \tag{2}$$

where $y^{(i)}_{pred}$ gives the predicted value of the i^{th} observation, x_j stands for the j^{th} predictor feature, β_j is the regression coefficient of the j^{th} term and $y^{(i)}$ is the actual value of the i^{th} observation.

Linear regression model may become sensitive to inputs and possibly unstable if the estimated regression coefficients become large. One possible approach to address this problem is to change the loss function by adding penalty for a model with large regression coefficients. Multiple Linear regression model has therefore been regularized using LASSO, Ridge and Elastic Net.

2) LASSO Regression: LASSO overcomes the drawback of linear regression by lowering the variance of the model. It uses L1 penalty that that penalizes the model based on the sum of the absolute regression coefficient values [18].

$$LASSOLoss = \sum (y^{(i)}_{pred} - y^{(i)})^2 + \lambda \sum |\beta_j| \tag{3}$$

where $y^{(i)}_{pred}$ gives the predicted value of the i^{th} observation, $y^{(i)}$ denotes the actual value of the i^{th} observation, β_j is the regression coefficient of the j^{th} term and λ is the amount of L1 penalty.

3) Ridge Regression: LASSO eliminates few regression coefficients from the model in the process of minimizing them. Ridge regression prevents any regression coefficient from being eliminated from model. It uses L2 penalty that penalizes a model based on the sum of the squared regression coefficient values [19].

$$RidgeLoss = \sum (y^{(i)}_{pred} - y^{(i)})^2 + \lambda \sum \beta_j^2 \tag{4}$$

where $y^{(i)}_{pred}$ gives the predicted value of the i^{th} observation, $y^{(i)}$ denotes the actual value of the i^{th} observation, β_j is the regression coefficient of the j^{th} term and λ is the amount of L2 penalty.

4) Elastic Net Regression: Elastic Net combines characteristics of both LASSO and Ridge. It adds both L1 penalty and L2 penalty to the original multiple linear regression model [20].

$$ElasticNetLoss = \sum (y^{(i)}_{pred} - y^{(i)})^2 + \lambda \sum |\beta_j| + \lambda \sum \beta_j^2 \tag{5}$$

where $y^{(i)}_{pred}$ gives the predicted value of the i^{th} observation, $y^{(i)}$ denotes the actual value of the i^{th} observation, β_j is the regression coefficient of the j^{th} term and λ is the amount of penalty.

5) Decision Tree Regression: Decision trees build regression models in the form of a tree structure. The average of the values of the target feature in the resultant leaf node is considered as the predicted value in a decision tree regression [21].

However, a single decision tree may not be always be a good choice for predicting values. A simple decision tree might give large bias whereas a complex decision tree might give large variance. Therefore ensemble learning techniques have been used for achieving better predictive performance by combining several decision trees.

6) Random Forest Regression: Bagging is an ensemble learning technique that follows a parallel approach. Here results from multiple models are combined to obtain a generalized result. Random Forest is a bagging algorithm where multiple decision trees learn from different subsets of the data and the mean prediction of the individual trees yields the result [22].

7) AdaBoost: Another kind of ensembling technique is boosting which follows a sequential process. Here a model tries to correct the regression errors of the previous model. AdaBoost is a boosting algorithm where multiple short decision trees are added sequentially to the ensemble. Each subsequent tree added attempts to correct the predictions made by the previous tree in the ensemble, thereby boosting the predictive performance of the ensemble of models [23].

8) Gradient Boosting: Gradient Boosting is another boosting algorithm. Gradient Boosting has three main components: additive model, loss function and a weak learner [24].

9) Support Vector Regression: Support Vector Regression method aims at obtaining the best fit line within the decision boundary. The best fit line is the hyperplane that has a maximum number of points [25].

C. Hyperparameter Optimization

While regression models have to learn the model parameters from the dataset during the training procedure, hyperparameters are those which must be specified to the models before the training procedure.

1) Hyperparameters for Linear Regression based models: LASSO, Ridge and Elastic Net regression have regularization parameters as hyperparameters. These regularization parameters control the amount of penalty that is to be applied on the Multiple linear regression model. Elastic Net has another hyperparameter called mixing factor. It denotes the amount of L1 penalty and L2 penalty that have to be applied into the Elastic Net model.

2) Hyperparameters for Tree based regression models: Decision Tree regressor requires several hyperparameters viz., function to be chosen to select the best split, the maximum allowed depth of a tree, the maximum number of features allowed for the split in a tree, the maximum number of leaf nodes a tree is allowed to have, the minimum number of features a node must have before it is split, etc. Random Forest regression requires all these hyperparameters along with additional ones viz., estimators (which defines the number of trees the ensemble can have), etc. AdaBoost requires hyperparameters such as the type of base learner to be used, the number of base estimators, possible values of the learning rate (which controls the contribution of every estimator in the ensemble), etc.

3) Hyperparameters for Support Vector Regression: The kernel function, epsilon (the distance between the hyperplane and the decision boundary) and C (amount of error allowed) are the hyperparameters that have been considered in this study.

4) Optimal hyperparameters selection methods: Random Search and Grid Search are the two hyperparameter optimization algorithms that have been employed in this study to find out the optimal values of various hyperparameters of the regression models[26]. While Random Search algorithm uses random combinations of the hyperparameters to find the optimal hyperparameters for the regression model, Grid Search algorithm uses every possible combination of the hyperparameters to find out the optimal solution.

D. Regression metrics

To evaluate how well a regression model has learnt during the training procedure, evaluation metrics Mean Absolute Percentage Error (MAPE) and R-squared have been used in this study.

1) Mean Absolute Percentage Error (MAPE): Mean absolute error is the average of the errors (between the model predicted value and actual data) on test data [27]. MAPE is the percentage of this error.

$$MAPE = \frac{\sum ABS\left(Predicted\ value - Mean\ value\right)}{\sum ABS\left(Actual\ value\right)} \times 100 \tag{6}$$

2) R-squared metric: R-squared value represents the fraction of variance of the target feature that is being captured by the regression model [27].

$$R^2 = \frac{\sum(Predicted\ value - Mean\ value)^2}{\sum(Actual\ value - Mean\ value)^2} \tag{7}$$

E. Proposed Methodology

The data in the available dataset need to be preprocessed as mentioned in section 3. The five aforementioned feature selection methods are then applied on the preprocessed data to obtain the different optimal sets of features. Next, data is split into two sets: training set and testing set in 3:1 ratio. Regression models are being trained using training set data and validation done using testing set data. Before starting the model training process, the hyperparameters have been optimised for the different regression models. Once hyperparameter optimisation is done, the training process is being executed using training set data. Once the models learn, it needs to be evaluated how well the models have learnt. Therefore, the predictor features in the testing set are applied as input to the trained regression models, and the values predicted by the models are compared with the actual target values present in the testing set. The regression model yielding best results on the testing set is used for predicting the number of Covid-19 cases in future. Fig. 2 illustrates the methodology that has been followed in this study.

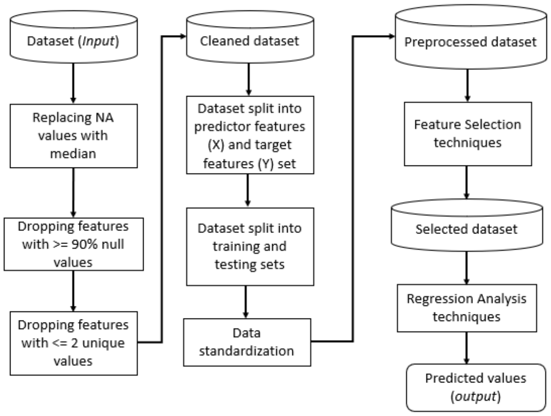
Dataset (*Input*)
Replacing NA values with median
Dropping features with >= 90% null values
Dropping features with <= 2 unique values
Cleaned dataset
Dataset split into predictor features (X) and target features (Y) set
Dataset split into training and testing sets
Data standardization
Preprocessed dataset
Feature Selection techniques
Selected dataset
Regression Analysis techniques
Predicted values (*output*)

Fig. 2: Methodology followed for modelling the growth in number of Covid-19 cases in India

V. Results and discussion

This study has been conducted on four different sets of data obtained by four different feature selection methods (Mutual Information, Pearson Correlation Coefficient, Forward Feature Selection and Backward Feature Elimination). Threshold for Mutual Information is chosen 1.2 by observing all the values of Mutual Information score obtained. For Pearson Correlation Coefficient score, it is being observed that many of the features are strongly correlated. Therefore 0.9 has been chosen as the threshold value for Pearson Correlation Coefficient in this study. Any two features with correlation coefficient score more than 0.9 have been considered, and the feature with higher correlation coefficient score with the target feature has been retained and the other one has been dropped. For Forward Feature Selection and Backward Feature Elimination methods, twelve best features have been selected, by repeating the methods on different number of features. For every regression technique, the Forward Feature Selection and Backward Feature Elimination methods have been performed separately and then the regression task has been executed on the best twelve features selected. Table 1 and Table 2 illustrate the MAPE values and R^2 that have been obtained by applying different regression techniques on different sets of features selected by different feature selection techniques respectively.

Among the linear regression based methods, unregularized multiple linear regression method has the minimum MAPE score, and therefore can be proposed as the best model for prediction. MAPE score $< 5\%$ is the most acceptable, and therefore all the linear regression based methods yield acceptable predicted outputs. R-squared value is also 0.999 for all the models, which too suggest the same. Among the decision tree based methods, Gradient boosting regression has scores the minimum MAPE value, and therefore can proposed as the best model for prediction. MAPE score $< 25\%$ is low but acceptable, and therefore all the decision tree based regression methods yield acceptable predicted values. R-squared value is in the range 0.8-0.9 for all the models, which too suggest the same. SVR – linear kernel has the minimum MAPE score, and therefore can be proposed as the best one for prediction, among all kernels. MAPE score $< 5\%$ is only for linear kernel for other kernels it is $> 25\%$ therefore not acceptable. R-squared value is in the range 0.8-1 for only linear kernel, which too suggest the same.

TABLE I: MAPE scores for different regression techniques

Model	***CC***	***MI***	***FS***	***BE***
Multiple Linear	35.730	0.003	0.003	0.003
LASSO	25.929	1.024	1.023	0.437
Ridge	35.710	0.0908	0.009	0.016
Elastic Net	25.958	1.157	1.156	1.241
Decision Tree	39.369	17.817	17.817	17.817
Random Forest	36.366	24.215	22.506	23.421
AdaBoost	47.472	23.293	22.301	26.828
Gradient Boosting	36.620	14.507	14.540	14.689
SVR-Linear	36.717	5.840	2.000	0.968
SVR-RBF	58.028	52.88	37.992	61.962
SVR-Sigmoid	38.287	>100	>100	57.386
SVR-Poly	72.254	>100	>80	>80

TABLE II: R^2 metrics scores for different regression techniques

Model	***CC***	***MI***	***FS***	***BE***
Multiple Linear	0.828	0.999	0.999	0.999
LASSO	0.889	0.999	0.999	0.999
Ridge	0.828	0.999	0.999	0.999
Elastic Net	0.889	0.999	0.999	0.998
Decision Tree	0.804	0.939	0.939	0.939
Random Forest	0.824	0.913	0.920	0.868
AdaBoost	0.559	0.898	0.914	0.856
Gradient Boosting	0.830	0.946	0.945	0.954
SVR-Linear	0.821	0.995	0.999	0.999
SVR-RBF	0.585	0.633	0.764	0.527
SVR-Sigmoid	0.747	<0	<0	0.506
SVR-Poly	0.006	<0	<0	<0

When comparing all the regression techniques, it is being observed that unregularized Multiple Linear Regression yields best MAPE and R-squared scores on test data set and therefore can be proposed as the most suitable model for modelling Covid-19 data of India. LASSO and Ridge regularized Multiple Linear Regression and linear-kernel based Support Vector Regression model are next best suited for modelling the Covid-19 data. Backward Feature Elimination method yields best results for most of the regression algorithms and hence can be considered as the most suitable method for feature selection from Covid-19 dataset of India. Fig. 3 displays the plot of actual and predicted number of Covid-19 cases over time, based on the test data set.

VI. Conclusion

In this study, it is being observed that varying the feature set varies the outcome of the regression model training. By training the same regression model with features selected

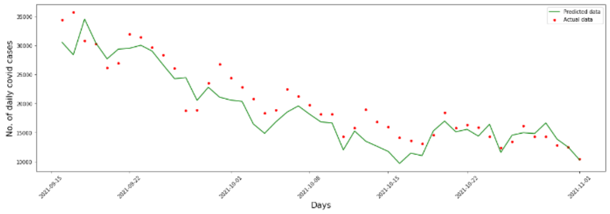

Fig. 3: Actual vs Predicted data

by different feature selection methods, different prediction results are being obtained. Therefore, feature selection is an important task in the process of model training. Next, different regression techniques have to be applied on the training data set and the ones that give minimum MAPE (value must be $<$5) and maximum R^2 score (value should be > 0 and as close as possible to 1) can be proposed as the most acceptable one for modelling the epidemiological data and predicting future number of Covid-19 cases. Also both linear and non linear regression techniques need to be applied on the dataset. Comparing the regression analysis results mentioned in the literature survey, it can be concluded that the epidemiological data varies from time to time and from region to region and therefore region-wise study and studying the data at different periods of time are of utmost importance, to model the growth of the number of Covid-19 cases with maximum accuracy.

VII. Future Scope

It is important to mention that for predicting Covid-19 cases in the future, the values of different predictor features in the future have to be known. However the values of the different predictor features in the future are unknown and they have to be computed using time series analysis methods. Once the values of the predictor features in nearby future become available, the number of Covid-19 cases in nearby future can be predicted using the best suitable regression model.Therefore as part of future scope of work, time series analysis models such as Auto Regressive (AR) model, Moving Average (MA) model, Auto Regressive Integrated Moving Average (ARIMA) model, etc. can be used to predict values of the predictor variables and then regression analysis applied on it.

References

[1] World Health Organization. ”WHO-convened global study of origins of SARS-CoV-2: China Part”, 2021.

[2] Mohamadou, Youssoufa, Aminou Halidou, and Pascalin Tiam Kapen. ”A review of mathematical modeling, artificial intelligence and datasets used in the study, prediction and management of COVID-19.” Applied Intelligence vol. 50, no. 11, pp. 3913-3925, 2020.

[3] Samarasekera, Udani. ”India grapples with second wave of COVID-19.” The Lancet Microbe, vol. 2, no. 6, pages e238, 2021.

[4] News item regarding spread of Covid-19 pandemic: https://www.livemint.com/news/india/india-covid-19-updates-14-313-new-cases-549-deaths-reported-in-24-hrs-11635566257293.html.

[5] Şahin, Utkucan, and Tezcan Şahin. ”Forecasting the cumulative number of confirmed cases of COVID-19 in Italy, UK and USA using fractional nonlinear grey Bernoulli model.” Chaos, Solitons Fractals, vol. 138, page 109948, 2020.

[6] Balli, Serkan. ”Data analysis of Covid-19 pandemic and short-term cumulative case predicting using machine learning time series methods.” Chaos, Solitons Fractals, vol. 142, page 110512, 2021.

[7] Ahmad, Amir, Sunita Garhwal, Santosh Kumar Ray, Gagan Kumar, Sharaf Jameel Malebary, and Omar Mohammed Barukab. ”The number of confirmed cases of Covid-19 by using machine learning: Methods and challenges.” Archives of Computational Methods in Engineering, vol. 28, no. 4, pp. 2645-2653, 2021.

[8] Yadav, Milind, Murukessan Perumal, and M. Srinivas. ”Analysis on novel coronavirus (COVID-19) using machine learning methods.” Chaos, Solitons Fractals, vol. 139, page 110050, 2020.

[9] Andreas, Andreou, Constandinos X. Mavromoustakis, George Mastorakis, Shahid Mumtaz, Jordi Mongay Batalla, and Evangelos Pallis. ”Modified machine learning Techique for curve fitting on regression models for COVID-19 projections.” IEEE, In 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), pp. 1-6, 2020.

[10] Rustam, Furqan, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won On, Waqar Aslam, and Gyu Sang Choi. IEEE ”COVID-19 future forecasting using supervised machine learning models”, vol. 8, pp. 101489-101499, 2020.

[11] Bhadana, Vartika, Anand Singh Jalal, and Pooja Pathak. ”A Comparative Study of Machine Learning Models for COVID-19 prediction in India.” IEEE In 4th Conference on Information Communication Technology (CICT), pp. 1-7, 2020.

[12] Johns Hopkins University Data Repository. [Online].Available: https://github.com/CSSEGISandData

[13] S. Das, “Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection,” Proc. 18th Int’l Conf. Machine Learning, pp. 74- 81, 2001.

[14] J. R. Vergara and P. A. Estévez, “A review of feature selection methods based on mutual information,” Neural Comput. Appl., vol. 24, no. 1, pp. 175–186, 2014.

[15] Liu, Yaqing, Yong Mu, Keyu Chen, Yiming Li, and Jinghuan Guo. ”Daily activity feature selection in smart homes based on pearson correlation coefficient.” Neural Processing Letters, pp. 1-17, 2020.

[16] Aha, David W., and Richard L. Bankert. ”A comparative evaluation of sequential feature selection algorithms.” Learning from data, New York, NY, pp. 199-206, 1996.

[17] Rath, Smita, Alakananda Tripathy, and Alok Ranjan Tripathy. ”Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model.” Diabetes Metabolic Syndrome: Clinical Research Reviews, vol. 14, no. 5, pp. 1467-1474, 2020.

[18] Tibshirani, Robert. ”Regression shrinkage and selection via the lasso.” Journal of the Royal Statistical Society: Series B (Methodological) vol. 58, no. 1, pp. 267-288, 1996.

[19] Hoerl, Arthur E., and Robert W. Kennard. ”Ridge regression: applications to nonorthogonal problems.” Technometrics, vol. 12, no. 1, pp. 69-82, 1970.

[20] Zou, Hui, and Trevor Hastie. ”Regularization and variable selection via the elastic net.” Journal of the royal statistical society: series B (statistical methodology), vol. 67, no. 2, pp. 301-320, 2005.

[21] Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. ”Cart.” Classification and Regression Trees; Wadsworth and Brooks/Cole: Monterey, CA, USA, 1984.

[22] Zhou, Xudong, Xinkai Zhu, Zhaodi Dong, and Wenshan Guo. ”Estimation of biomass in wheat using random forest regression algorithm and remote sensing data.” The Crop Journal, vol. 4, no. 3, pp. 212-219, 2016.

[23] D. P. Solomatine and D. L. Shrestha, “AdaBoost.RT: A boosting algorithm for regression problems,” in Proc. IEEE Int. Joint Conf. Neural Netw., pp. 1163–1168, 2004.

[24] Bentéjac, Candice, Anna Csörgő, and Gonzalo Martínez-Muñoz. ”A comparative analysis of gradient boosting algorithms.” Artificial Intelligence Review, vol. 54, no. 3, pp. 1937-1967, 2021.

[25] Kavitha, S., S. Varuna, and R. Ramya. ”A comparative analysis on linear regression and support vector regression.” In 2016 online international conference on green engineering and technologies (IC-GET), pp. 1-5. IEEE, 2016.

[26] P. Liashchynskyi and P. Liashchynskyi, “Grid search, random search, genetic algorithm: A big comparison for NAS,” ,arXiv:1912.06059.[Online].Available: http://arxiv.org/abs/1912.06059.

[27] Chicco, Davide, Matthijs J. Warrens, and Giuseppe Jurman. ”The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation.” PeerJ Computer Science, vol. 7, page e623, 2021.