Predicting Covid-19 Trends Using Machine Learning based on Linear and Multiple Linear Regression Model

Ashwani Kumar[1] Kondam Deeksha[2] Gudidevuni Sai Pooja[3]
Thippireddy Tarun Reddy[4] Tera Akhil Reddy[5]
[1]Department of CSE, Sreyas Institute of Engineering and Technology
ashwani.kumarcse@gmail.com[1]

Abstract

The COVID-19 pandemic has adversely affected the health and economy of almost all the countries in the world including India. Almost thousands of people are getting affected by this daily. In this paper, analysis of the daily statistics of people who got affected and this proposed work is going to predict the future trend of the active cases in Odisha and India. Machine Learning based forecasting algorithms have proved their significance in generating predictive outcomes which are used to make decisions on actions that are going to happen in the future. ML algorithms have been using for a long time to do this kind of task. This proposed work is going to do analysis and prediction on the dataset which was created by COVID India organization. Linear and Multiple Linear Regression models are used to predict the future trend of active cases and also the number of active cases in fore coming days and to visualize the trend of future active cases. Here, the performance of Linear and Multiple Linear regression models are compared by using the R^2 score. Linear and Multiple Linear regression got 0.99 and 1.0 as R^2 scores respectively which shows that these are the strongest prediction models that are used to predict the future active cases of COVID - 19. Both these models acquired remarkable accuracy in COVID - 19 prediction. A strong correlation factor shows that there is a very strong relationship between a dependent variable (Active cases) and independent variables (positive, deceases, recovered cases).

Keywords: *Coronavirus, India, Odisha, Linear regression, Multiple linear regression, Correlation coefficient, R^2Score.*

1.Introduction

Today the world has got badly affected by COVID-19 which is claimed to be originated from Wuhan city, China. Almost every sector right from agriculture to the software industry is affected by this pandemic. As this disease is contagious, it is being spread by having close contact with the infected people [1]. Specifically, the transmission occurs through droplets that come during sneezing, cough, or through the saliva droplets while talking [2]. This contagious disease is deadly unless we follow social distancing and measures put forth by government bodies [3]. There is also evidences that this air-borne disease is harmful to old age people and even it may affect the person twice i.e., even after the first attack. Although the government has taken necessary measures, India is one of the most densely populated countries, the conditions have gone worse here. So far nearly 9.6 million people got affected by COVID-19 in India and whereas it is 65 million in the world. But the only good sign in India is the recovery rate which is over 93% [4]. The vaccines are yet to come to public use. This paper can be used to forecast active cases earlier. So that proper arrangements can be done to reduce the fatality rate. This paper can be used to forecast active cases earlier [5]. So that proper arrangements can be done to reduce the fatality rate. The ML models are widely used in many areas where identification of major factors for a threat and prioritizing them accordingly is a major task [6].

The main aim of this paper is to develop a system that predicts the future active cases of COVID and the trend of active cases of COVID with better accuracy using machine learning models.

- Using linear and multiple regression models.
- Using the Indian and Odisha state data sets.
- Validate the results using R^2 score
- Depict the results and comparisons.

2. Related Work

A similar approach to this paper is predicting the same COVID-19 trends but using different methods. This uses the multi-layer perceptron and exponential smoothing along with regression models [7]. Where A feedforward artificial neural network with multiple layers is known as a multilayer perceptron (ANN). MLP is distinguished from linear perceptron by its numerous layers and non-linear activation [8]. It can tell the difference between data that isn't linearly separable and data that is. Using the exponential window function, exponential smoothing is a rule-of-thumb technique for smoothing time series data [9-12]. The accuracy that was achieved with these models is about 80-90 percent. These methods are complex to implement when compared to the proposed system. These methods of working performs well in the training phase but when it comes to the end phase it suffers from several difficulties. Stochastic Environmental Research and Risk Assessment by Springler, May 2020 "A machine learning forecasting model for COVID-19 pandemic in India," by R. Sujath et al [13]. In this case, multilayer perceptron and vector auto-regression are applied. Kumar et proposed object detection techniques using machine learning algorithms [14-17]. It's tricky to use and only gives mediocre accuracy. COVID-19 Future Forecasting Using Supervised Machine Learning Models," IEEE Access, IEEE, 2020. 'Furqan Rustam et al.' is an acronym for Furqan Rustam and his collaborators. They recommended using a linear regression model with exponential smoothing for prediction. It is 83 percent accurate [18]. It's simple to use and has an 83 percent accuracy rating.

3. Proposed Method

The proposed system consists of two phases namely linear regression and multiple linear regression. Both the methods are performed and are validated using the appropriate validation mechanism. The accuracy that can be obtained using this approach is above 95%. It works well both in the training and testing phase. The approach is also much simpler than the existing system.

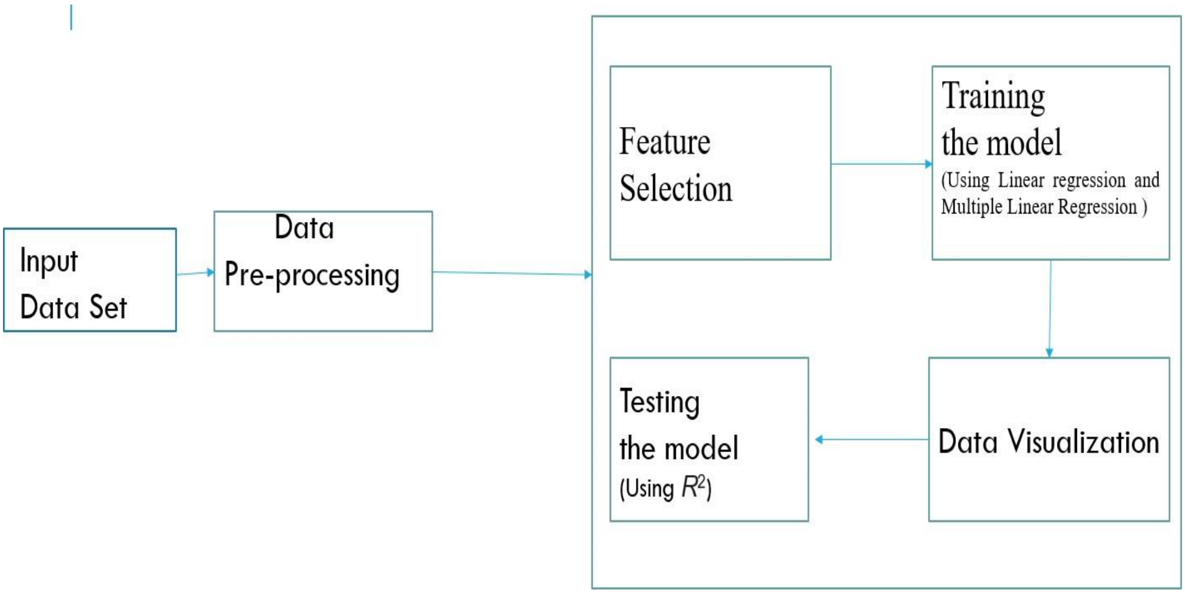

Figure 1. Architecture diagram of proposed approach

3.1 Data Pre-processing

Data Pre-processing is a process of setting up the crude information and making it appropriate for an AI and machine learning model. It is the first and essential advance step while making an AI and machine learning model. While making an AI and machine learning paper, it isn't generally a case that we confess all and designed information. And keeping in mind that doing any activity with information, it is required to clean it and put in an organized manner. So, for this, the data pre-processing task is used. A genuine information for the most part contains noises, missing values, and perhaps in an unusable configuration which can't be straightforwardly utilized for AI and machine learning models. Data pre-processing is a required task for cleaning the information and making it reasonable for an AI model which additionally builds the exactness and proficiency of an AI model. To make an AI and machine learning model, the primary thing required is a dataset as an AI and machine learning model totally chips away at information. The gathered information for a specific issue in a legitimate arrangement is known as the data set.

Here, in this case, the dataset is taken from the Covid-19 india.org website. As this is a delicate task, the data is obtained without any errors. The zero values in the dataset indicate that no case was recorded on that particular day. In this work, a python code is used for pre-processing, by which the cumulative data will be obtained as output. This facilitates the training of models. Rather than this everything was fine with the data that has been taken from the web.

3.2 Feature Selection

The team had worked with two models here namely Linear and Multiple linear regression. Each Model requires its own features. The linear regression model needs only one feature. So only the Number of Confirmed Cases is taken into the consideration while multiple linear regression needs multiple features to get better accuracy. Hence, the Number of Confirmed cases, Number of Recovered cases and Number of Deaths are taken into consideration.

3.3 Linear Regression

Linear Regression is one of the famous and most underrated Regression algorithms. It is a factual strategy that is utilized for prescient examination. Linear Regression makes predictions for continuous variables or numeric variables, for example, sales, pay, age, item cost, and so on. Linear regression algorithm predicts the best-fitted line that shows a linear relationship between dependant and independent variables. Therefore, it is called Linear regression. As the dependent variable depends on another variable which was taken into measure when there is a change in the independent variable Linear regression algorithm predicts the value of the dependent variable accordingly. This model gives a slanted straight line speaking to the connection between the factors.

Mathematically, Linear regression is represented as:

$$Y = a+b\ (X)+c$$

Y = Dependent variable

X = Independent variable

a = Intercept of a regressor line (also states degree of freedom)

b = regressor coefficient

c = Random error

The values for x and y variables are training datasets for Linear Regression model Regression.

3.4 Multiple Linear Regression

There are many cases where a reliant variable depends on more than one independent variable, for these cases Multiple Linear regression is used. Multiple Linear Regression is a regression algorithm which forms a linear relationship between a single dependent continuous variable and more than one independent variable. As it depends on more than one independent variable we can say it is an extension for Linear Regression.

For MLR,

2.3.1 dependent variable - Continuous/real

2.3.2 independent variables - Continuous or categorical

2.3.3 There should be a direct relationship between dependant and independent variables.

$$\mu y=\beta o+\beta 1x1+\ldots\ldots+\ \beta_p x_p$$

μy is a dependent variable.

X1, X2, …., Xn are independent variables.

β0, β1,…, βn are the regression coefficient.

3.5 R^2 Coefficient

R-squared measure states about how well the regression model get fitted. It is also known as the coefficient of determination. The outcome of the measure varies from 0 - 1. The value closer to 1 states the model is best fitted. It is measured as the ratio between the sum of squares of residuals (SSres) and the total sum of squares (SStot).

Coefficient of Determination → $R^2 = (SSR \div SST) = 1-(SSE \div SST)$
Sum of Squares Total → $SST = \sum(y-\bar{y})2$
Sum of squares Regression → $SSR = \sum(y_1-\bar{y}_1)2$
Sum of Squares Error → $SSE = \sum(y-\bar{y}_1)2$

3.6 Results Analysis

The Hardware Requirements used in this paper are PC/Laptop, RAM (minimum of 4GB). The Software Requirements used in this paper are Python Language, Jupyter Notebook (IDE) and the libraries used are Pandas, numpy, sklearn, matplotlib.

4. BAR GRAPHS

The below Bar graph is for Odisha state obtained by applying linear regression Where Blue indicates the Actual values and orange indicates the predicted values of active COVID cases.

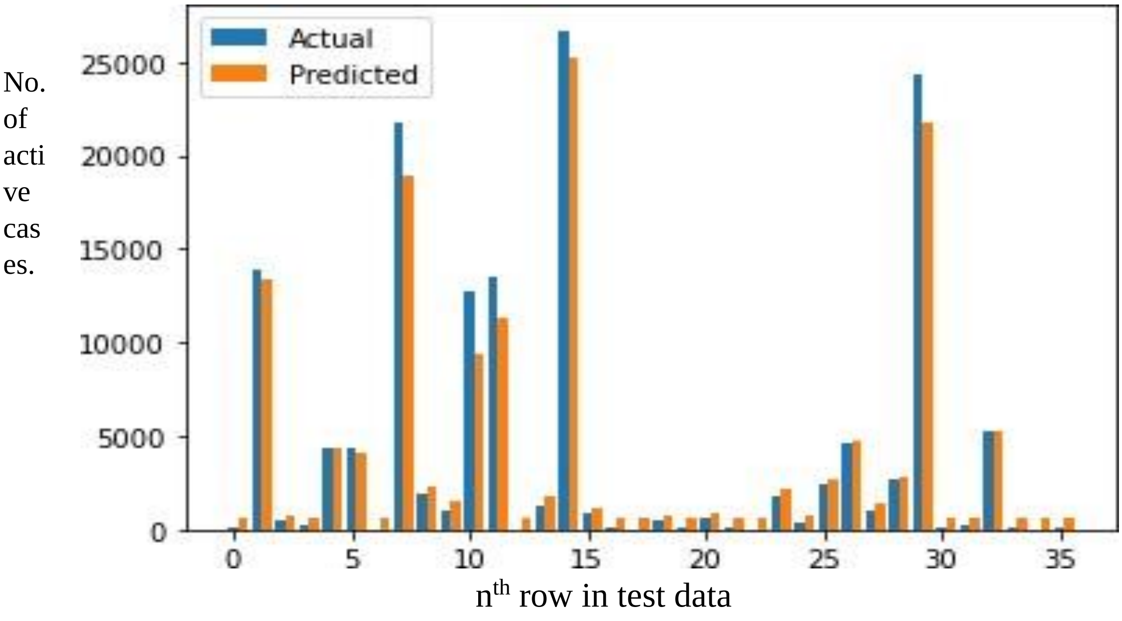

Figure.2.Odisha Linear Bar graph

The below Bar graph is for Odisha state obtained by applying Multiple linear regression Where Blue indicates the Actual values and orange indicates the predicted values of active COVID cases.

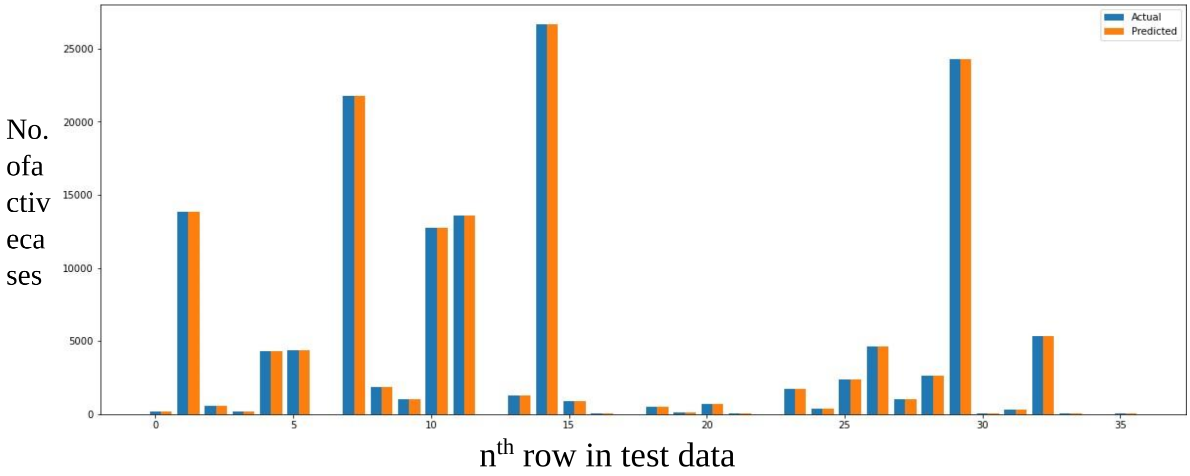

Figure.3. Odisha Multiple Linear Bar graph

The below Bar graph is for India obtained by applying linear regression Where Blue indicates the Actual values and orange indicates the predicted values of active COVID cases.

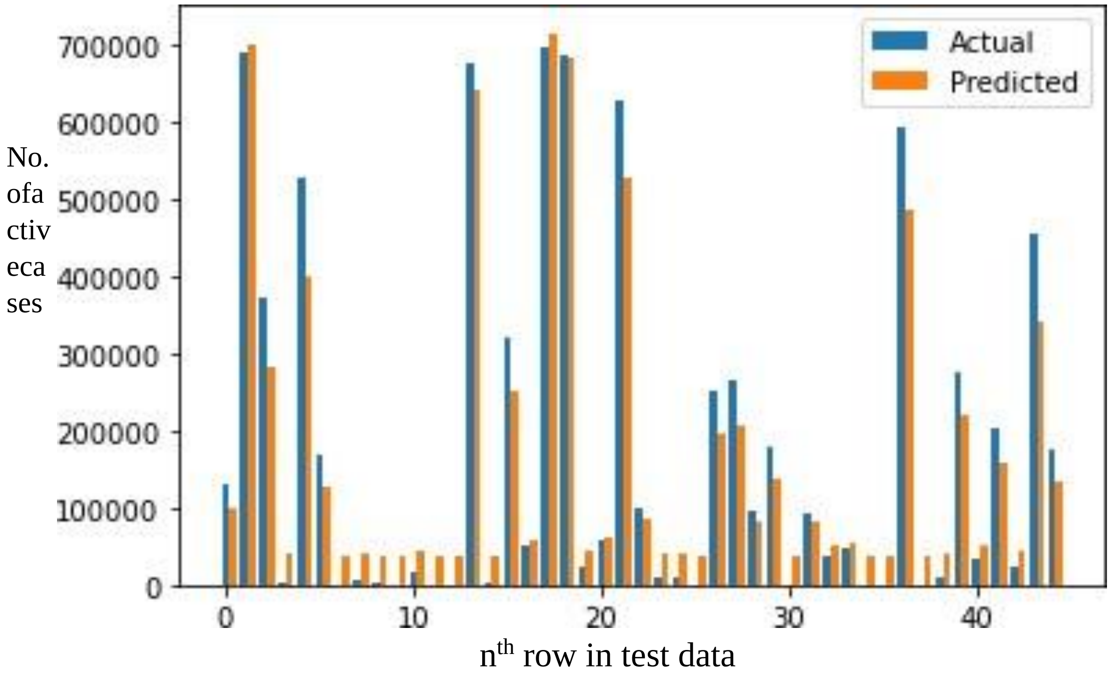

Figure.4.India Linear Bar graph

The below Bar graph is for India obtained by applying Multiple linear regression Where Blue indicates the Actual values and orange indicates the predicted values of active COVID cases.

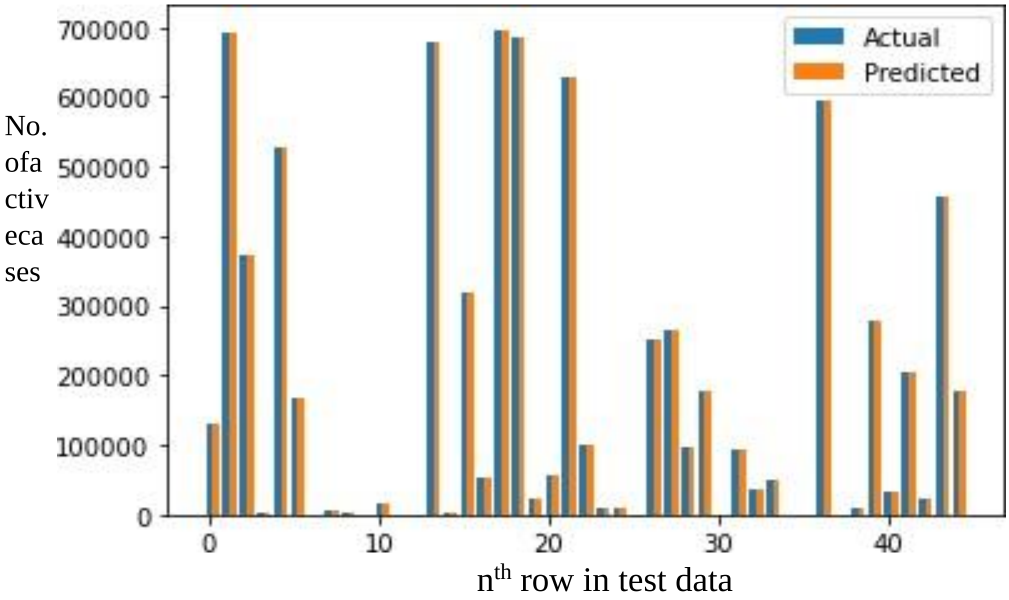

Figure.5.India Multiple Linear Bar graph

After training with a linear regression prediction model, the values obtained are:

Dataset	Intercept	Coefficient	Score(R2)	Mean Absolute Error(MAE)	Mean Squared Error(MSE)	Root Mean Squared Error(RMSE)
Orissa	582.1591456175147	0.2517285	0.978749223901475	683.0527208901069	1074966.0167184807	1036.8056793432802
India	38298.9116493169	0.22788256	0.95484835860697	39539.38665011771	2382881574.1232305	48814.767992106965

Table(1) Linear Regression Values

After training with a multiple linear regression prediction model, the values obtained are:

Dataset	Intercept	Coefficient	Score (R2)	Mean Absolute Error(MAE)	Mean Squared Error(MSE)	Root Mean Squared Error(RMSE)
Orissa	3.637978807091713e-^{12}	[1.,-1.,-1.]	1.0	5.561674577671106e^{-12}	8.343621147381854e^{-23}	9.1343424215330 65e^{-12}
India	4.94765117764473e^{-10}	[1.,-1.,-1.]	1.0	4.496757336861871e^{-10}	2.607399398593064e^{-19}	5.106270065902375e^{-10}

Table(2) Multiple Linear Regression Values

Case 1:

As previously explained, Linear regression and Multiple linear regression fits the regressor line effectively and the give the best suited values for intercept and coefficient. So, from the results obtained after performing Linear regression it can be said that for every one – unit rise in positive case there is an increase of 25% of active cases in Odisha and 22% of active cases in India. A base error of 582.159 & 38298.911 in case of Odisha and India
respectively. The value of R_2 scores are 0.97 and 0.95 which says this a strong predictor model. But the MAE, MSE, RMSE values are a bit higher than expected.

Case 2:

From the results obtained after performing Multiple linear regression it can be said that for every one-unit change in independent variables there is a change in one – unit of active cases in both Odisha and India. The values of R2 scores higher than linear regression which shows that a better predictor model than the previous one and also the values of MAE, MSE, RMSE errors are also very low. Therefore, MLR is the best model as compared to Linear regression.

5. Conclusions

Both linear and multiple linear regression are successfully implemented. The future active cases were predicted successfully with very less error and also the trend of active cases of COVID – 19 accurately.The COVID-19 Pandemic has adversely affected the health and economy of almost all the countries.so in this proposed work we used Machine Learning Forecasting Algorithms. we used Regression models in which the object predicts the output in a continuous values.In this study we predict the future active cases of covid-19 using Linear and Multiple Linear Regression models.First we collected two Data Sets from WHO and India Covid organization of odisha state and India.Then Data Processing is done where raw data is transformed to machine understandable format. Then Training of the data is done using Linear and Multiple Linear Regression Models. Testing is done using R-square Score. By comparing the R-square scores of Linear and Multiple Linear regression models,the values of R-square score of Multiple Linear Regresion model is higher than Linear Regression model which shows that a better predictor model than the previous one.While, linear regression is taken, which in turn considered only one feature so the desired outputs are not obtained and some of the error values are high. Hence, multiple linear regression was considered which takes more features as input, in-turn desired outputs are obtained.So, considering multiple factors had helped to achieve more efficient and accurate results. An interface needs to be developed for this model and also, this work can be extended to the remaining states as well. Also, some more features shall be added to this model based on the change in requirements.

REFERENCES

[1] *Wang W, Tang J, and Wei F. Updated understanding of the outbreak of 2019 novelcoronavirus (2019-nCoV) in Wuhan, China. Journal of medical virology 2020.*

[2] *Coronavirus disease 2019 (COVID-19): situation report. World Health Organization 2020.*

[3] *Modes of transmission of virus causing COVID-19: implications for IPC precaution recommendations: scientific brief. World Health Organization. [Online] Available at : 27 March 2020 (No. WHO/2019nCoV/Sci_Brief/Transmission_modes/2020.1).*

[4] *INDIA COVID-19 TRACKER. 2020 [Online] Available at: https://www.covid19india.org/. Accessed on: 11th July 2020.*

[5] *Centers for Disease Control and Prevention. Symptoms of coronavirus 2020[Online] Available at: https://www. cdc. gov/coronavirus/2019ncov/symptoms- testing/symptoms. Accessed on: April 21, 2020.*

[6] *Barkur G, and Vibha GBK. Sentiment analysis of nationwide lockdown due to COVID 19 outbreak: Evidence from India. Asian journal of psychiatry 2020.*

[7] *Syazali M, Putra F, Rinaldi A, Utami L, Widayanti W, Umam R, &Jermsittiparsert K. (2019). Partial correlation analysis using multiple linear regression: Impact on business environment of digital marketing interest in the era of industrial revolution 4.0.*

[8] *Salleh FHM, Zainudin S, &Arif SM. Multiple linear regression for reconstruction of gene regulatory networks in solving cascade error problems. Advances in Bioinformatics 2017.*

[9] *Uyanık GK, Güler N. A study on multiple linear regression analysis. Procedia-Social and Behavioral Sciences 2013.*

[10] *Khademi F, Jamal SM, Deshpande N, Londhe S. Predicting strength of recycled aggregate concrete using artificial neural network, adaptive neurofuzzy inference system and multiple linear regression. International Journal of Sustainable Built Environment 2016.*

[11] *Hosseinzadeh A, Baziar M, Alidadi H, Zhou JL, Altaee A, Najafpoor AA, Jafarpour S. Application of artificial neural network and multiple linear regression in modeling nutrient recovery in vermicompost under different conditions. Bioresource Technology 2020.*

[12] *Luu C, von Meding J, Mojtahedi M. Analyzing Vietnam's national disaster loss database for flood risk assessment using multiple linear regressionTOPSIS. International Journal of Disaster Risk Reduction 2019.*

[13] *Duu Z, Hu Y, Buttar NA. Analysis of mechanical properties for tea stem using grey relational analysis coupled with multiple linear regression. Scientia Horticulturae.*

[14] *Kumar, Ashwani. "A cloud-based buyer-seller watermarking protocol (CB-BSWP) using semi-trusted third party for copy deterrence and privacy preserving." Multimedia Tools and Applications (2022): 1-32.*

[15] *Kumar, A., Design of secure image fusion technique using cloud for privacy-preserving and copyright protection. International Journal of Cloud Applications and Computing (IJCAC), 2019. 9(3): p. 22-36.*

[16] *Ashwani Kumar, "A Review on Implementation of Digital Image Watermarking Techniques Using LSB and DWT" in the Third International Conference on Information and Communication Technology for Sustainable Development (ICT4SD 2018), held during August 30-31,2018 at Hotel Vivanta by Taj, GOA, INDIA.*

[17] *Kumar, A., Z.J. Zhang, and H. Lyu, Object detection in real time based on improved single shot multi-box detector algorithm. EURASIP Journal on Wireless Communications and Networking, 2020. 2020(1): p. 1-18.*

[18] *Kadam AK, Wagh VM, Muley AA, Umrikar BN, Sankhua, RN. Prediction of water quality index using artificial neural network and multiple linear regression modelling approach in Shivganga River basin, India. Modelling Earth Systems and Environment 2019.*