Journal of
Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7765918

FUTURE PREDICTION IN COVID-19 BASED ON VARIABLE SELECTION WITH THE SUPPORT OF DATA MINING AND MACHINE LEARNING

N. Sankar [1] and S. Manikandan [2]

[1] Research Scholar, Department of Computer and Information Science, Faculty of Science, Annamalai University, Annamalainagar – 608 002, Tamil Nadu, India,
Email: nsankarraj@gmail.com

[2]Assistant Professor, PG Department of Computer Science, Government Arts College, Chidambaram - 608 102, Tamil Nadu, India,
Email: us.mani.s.mca@gmail.com

Abstract

Novel Coronavirus is an infection caused by SARS CoV 2 that began in China in 2019. The data mining is the important and dedicated tools for forecasting the hidden knowledge with the help of pre-existing dataset. The covid analysis and vaticination for consider different affiliated parameters like name of the sates (countries), total affected cases, cases in particular date, active cases, discharged cases, discharged cases in particular date, overall death, and deaths in particular date. In this paper was aimed at tracing the suitable variable selection for future prediction and subsequently based on suitable variable to predict the future trends related to total cases admitted and total cases discharged on expected date using overall total cases. Similarly predict discharged cases using total cases admitted in India. Numerical illustrations also give to prove the results and conclusions using regression model and its accuracy parameters.

Keywords: Covid-19, Data Mining, Regression model, coefficient of determination and prediction

1. Introduction and Related Work

Data mining is one of the most beneficial procedures that help entrepreneurs, researchers, and individuals to extract crucial knowledge from huge categories of data. The approach requires the knowledge of the data mining is also called Knowledge Discovery in Database (KDD). The knowledge discovery procedure includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation. Data mining encompasses all fields of Data mining such as applications, Data mining vs Machine learning, Data mining tools, Social Media Data mining, Data mining techniques, Clustering in data mining, Challenges in Data mining, etc. The technique of extracting data to identify patterns, trends, and advantageous data that would enable the business to take the data-driven decision from huge categories of data is called Data Mining [1].

Data mining in healthcare has excellent capability to improve the health system. It uses data and analytics for better understandings and to identify best practices that will boost health care services and alleviate costs. Analysts use data mining techniques such as Machine learning, multi-dimensional database, Data visualization, soft computing, and statistics. Data Mining can be employed to forecast patients in each category. The procedures guarantee that the patients get intensive care at the right place and at the right time. Data mining also facilitates healthcare insurers to recognize fraud and abuse [2]. Analysis and accuracy of data mining algorithms for various decision tree approaches using WEKA tool to stumble on important parameters of the tree structure. Seven classification algorithms such as J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy [3].

AI applications based on data mining and machine learning (ML) algorithms for detecting and diagnosing COVID-19. Overview of this critical virus, address the limitations of utilising data mining and ML algorithms, and provide the health sector with the benefits of this technique. We used five databases, namely, IEEE Xplore, Web of Science, PubMed, ScienceDirect and Scopus and performed three sequences of search queries between 2010 and 2020 [4]. to run tests on real-world data, and four output classification algorithms (Decision Tree, K-nearest neighbors, Random Tree, and Naive Bayes) are used to analyze and draw conclusions. The comparison is based on accuracy and performance period, and it was discovered that the Decision Tree outperforms other algorithms in terms of time and accuracy [5].

ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) have been used for forecast the factors of COVID-19 [6]. extracting risk factors from clinical data of early COVID-19 infected patients and utilizing four types of traditional machine learning approaches including logistic regression (LR), support vector machine (SVM), decision tree(DT), random forest(RF) and a deep learning-based method for diagnosis of early COVID-19. The higher AUC of our LR-base predictive model makes it a more conducive method for assisting COVID-19 diagnosis. The optimal model has been encapsulated as a mobile application (APP) and implemented in some hospitals in Zhejiang Province [7].

2. Methodology

In this section explain different description of method is provided. In this paper, consider six tasks were performed and concepts working based on linear regression model and its model accuracy.

2.1 Linear Regression Model

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables: (1) One variable, denoted *x*, is regarded as the predictor, explanatory, or independent variable. (2) The other variable, denoted *y*, is regarded as the response, outcome, or dependent variable. Because the other terms are used less frequently today, we'll use the "predictor" and "response" terms to refer to the variables encountered in this course [8].

$$y=a_x+b \qquad \dots (1)$$

2.2 Coefficient of Determination

The coefficient of determination is a dimension used to explain how important variability of one factor can be caused by its relationship to another affiliated factor. This correlation called goodness of fit is represented by values between 0.0 and 1.0. A value of 1.0 indicates a perfect fit and is therefore a largely dependable model for future prediction, while a value of 0.0 would indicate that the computation fails to accurately model the data at all. But a value of 0.20, for illustration, suggests that 20% of the dependent variable is predicted using independent variable, while a value of 0.50 suggests that 50% of the dependent variables is predicted using independent variable [9].

$$\boldsymbol{r} = \frac{\boldsymbol{n}(\sum \boldsymbol{xy}) - (\sum \boldsymbol{x})\,(\sum \boldsymbol{y})}{\sqrt{[\boldsymbol{n}\sum \boldsymbol{x}^2 - (\sum \boldsymbol{x})^2] - [\boldsymbol{n}\sum \boldsymbol{y}^2 - (\sum \boldsymbol{y})^2]}} \quad \ldots (2)$$

2.3 Mean Absolute Error (MAE)

Mean Absolute Error calculates the average disparity between the quantified values and actual values. It is also known as scale-dependent precision as it quantifies error in observations taken on the identical scale. It is adopted as assessment metrics for regression patterns in machine learning. It calculates errors between actual values and values foreseen by the model. It is utilized to predict the correctness of the machine learning model.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad \ldots (3)$$

where Σ: Summation, y_i: Actual value for the i^{th} observation, x_i: Calculated value for the i^{th} observation and n: Total number of observations [10].

2.4 Mean Square Error (MSE)

The mean squared error quantifies how close a regression line is to a set of information points. It is a risk function that corresponds to the expected significance of the squared error loss. The mean squared error is evaluated by taking the average, specifically the mean, of the squared errors from data relevant to a function. A larger MSE shows that the data points are commonly scattered around its central moment (mean), while a smaller MSE suggests the opposite. A smaller MSE is preferred as it demonstrates that your data points are closely spaced around its central moment (mean). It reflects the centralized distribution of your data values, the fact that it is unbiased, and most importantly, has less error [11].

$$MSE = \left(\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}\right)^2 \quad \ldots (4)$$

where Σ: for summation, y_i: actual value for the i^{th} observation, x_i: calculated value for the i_{th} observation and n: total number of observations

2.5 Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is the standard deviation of the prediction error. Residuals are a quantify of how far data points are from the regression line; RMSE is a quantify of how distributed these residuals are. It tells you how concentrated the data is around the line of best fit. The mean square error is frequently used in climatology, predicting and regression model evaluation to verify empirical results [12].

$$RMSE = \sqrt{\left(\frac{\sum_{i=1}^{n} |y_i - x_i|}{n}\right)^2} \quad \ldots (5)$$

where Σ: for summation, y_i: actual value for the i^{th} observation, x_i: calculated value for the i_{th} observation and n: total number of observations

2.6 Dataset

Initially dataset was obtained from https://www.covid19india.org/ and Ministry of Health and Family Welfare, Government of India. This research work utilized government of India publicly available datasets [13]. The datasets have entries from different state of India up to June 2022. A brief description about them is mentioned in Table 1.

Table 1. Covid-19 dataset in India upto June 2021

State	Cumulative Cases	Today Cases	Active Cases	Cumulative Discharged	Today Discharged	Cumulative Deaths	Today Deaths
Maharashtra	6007431	9844	124911	5762661	9371	119859	556
Kerala	2854325	12078	100308	2741436	11469	12581	136
Karnataka	2823444	3979	110546	2678473	9768	34425	138
Tamil Nadu	2449577	6162	49845	2367831	9046	31901	155
Andhra Pradesh	1867017	4981	49683	1804844	6464	12490	38
Uttar Pradesh	1705014	224	3552	1679096	308	22366	30
West Bengal	1489286	1923	22308	1449462	1952	17516	41
Delhi	1433475	109	1767	1406760	131	24948	8
Chhattisgarh	992391	317	7314	971662	605	13415	8
Rajasthan	951695	147	2019	940771	306	8905	0
Odisha	890596	3650	30337	856498	3486	3761	44
Gujarat	822887	129	4427	808418	507	10042	2
Madhya Pradesh	789561	62	1280	779432	255	8849	22
Haryana	768002	102	1990	756679	253	9333	19
Bihar	720717	212	2558	708586	355	9573	4
Telengana	617776	1088	16030	598139	1511	3607	9
Punjab	593941	369	5274	572723	715	15944	21
Assam	493688	2781	31014	458330	3604	4344	34
Jharkhand	345028	114	1224	338698	252	5106	2
Uttarakhand	339245	118	2739	329432	250	7074	6
Jammu Kashmir	313476	448	6537	302655	682	4284	11

Himachal Pradesh	201210	161	2123	195624	323	3463	2
Goa	165426	229	2727	159677	258	3022	9
Puducher ry	115925	298	3077	111114	276	1734	3
Manipur	66171	549	9174	55912	655	1085	11
Tripura	63868	369	3828	59378	400	662	2
Chandiga rh	61542	22	247	60488	42	807	0
Meghala ya	46878	420	4424	41647	298	807	10
Arunacha l Pradesh	34214	298	2565	31487	298	162	2
Nagaland	24629	88	1509	22641	155	479	2
Ladakh	19903	22	314	19387	46	202	0
Sikkim	19681	92	2282	17101	198	298	2
Mizoram	18859	235	4455	14316	220	88	2
Daman Diu	10526	3	59	10463	4	4	0
Lakshad weep	9601	42	322	9232	60	47	0
Andaman Nicobar	7440	2	99	7214	4	127	0

3. Numerical Illustrations

The result of determination a trend in the data and find the accuracy based on combination of different parameters mentioned in table 2 with descriptive statistics like mean, median and mode. Means used to find the average for different parameters, median measure the middle values for all and finally, standard deviation (SD) is used to find the deviation trends between corresponding columns [14 - 15]. Different The descriptive statistics is used to writ the basic interpretation using table 2. Python scripts were used to find all necessary accuracy parameters mentioned in equation (1) to (3). Python script was used to predict the future based on different combination data from various state upto June 2021with numerical illustrations mentioned in table 2 and table 3 and subsequently the related figure mentioned in figure 1 to figure 5.

Table 2: Descriptive statistics in Covid-19 dataset

Statistics	Cumulative Cases	Today Cases	Active Cases	Cumulative Discharged	Today Discharged	Cumulative Deaths	Today Deaths
Mean	837067.91	1435.19	17024.11	809118.58	1792.417	10925	36.91
Median	419358	232	3314.5	398514	307	4314	8
SD	1197052.50	2801.78	31802.81	1150068.83	3194.97	20740	97.21

Table 3: Linear regression model and its accuracy parameters with training (80%) and testing (20%)

x	y	Mean	Median	SD	Coefficients	Variance	R2 Score
Cumulative Cases	**Today Cases**	140.3304	135.8628	919.0967	0.0020	844738.6729	0.6270
Cumulative Cases	**Active Cases**	10396.1984	4096.1262	16908.4934	0.0214	285897148.7579	0.6749
Cumulative Cases	**Cumulative Discharged**	-10209.6270	-5706.1173	12819.8146	0.9620	164347646.4819	0.9996
Cumulative Cases	**Today Discharged**	779.9161	174.4825	1401.4507	0.0020	1964064.1243	0.7300
Cumulative Cases	**Cumulative Death**	-186.5714	2356.4442	5887.9504	0.0165	34667959.7839	0.7099
Cumulative Cases	**Today Death**	5.7169	10.0410	25.0886	0.0001	629.4359	0.6436

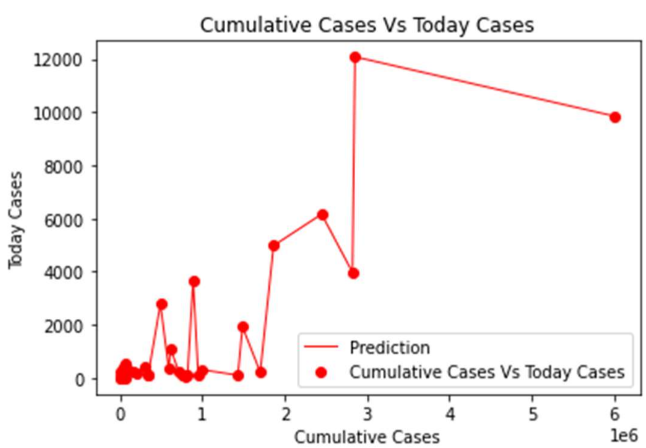

Fig. 1. Linear regression between cumulative cases and today cases

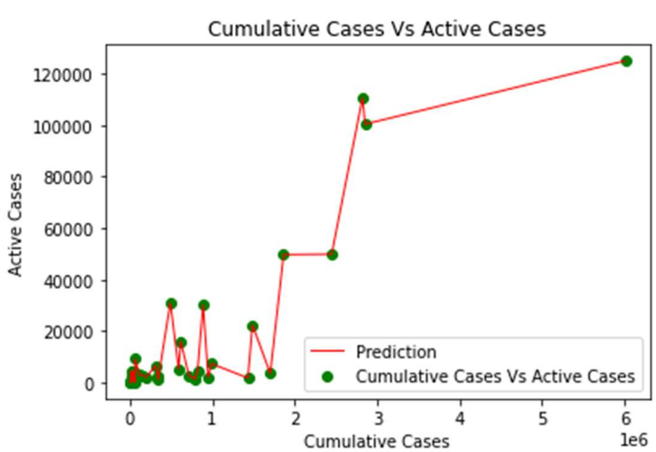

Fig. 2. Linear regression between cumulative cases and active cases

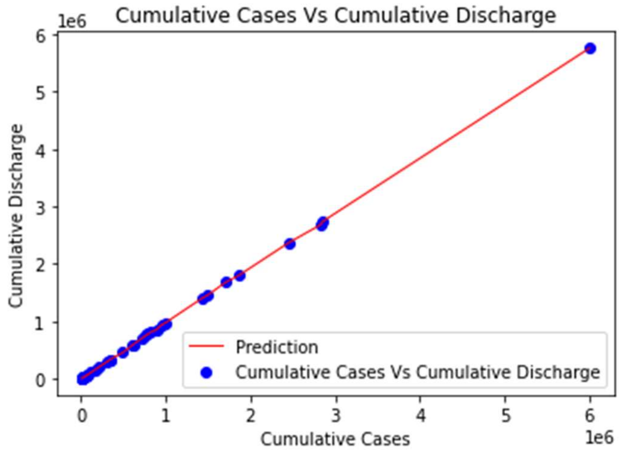

Fig. 3. Linear regression between cumulative cases and cumulative discharge cases

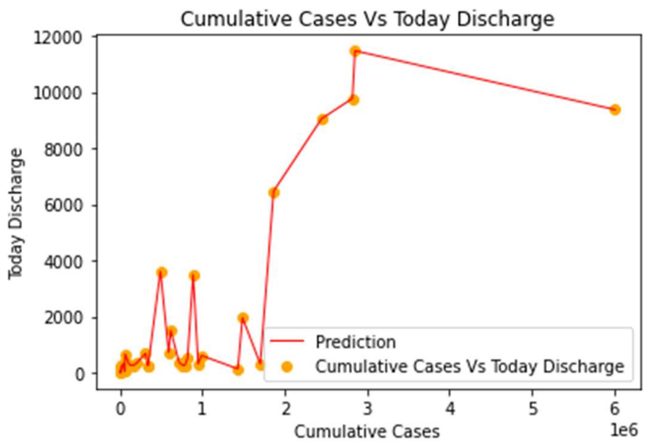

Fig. 4: Linear regression between cumulative cases and today discharge cases

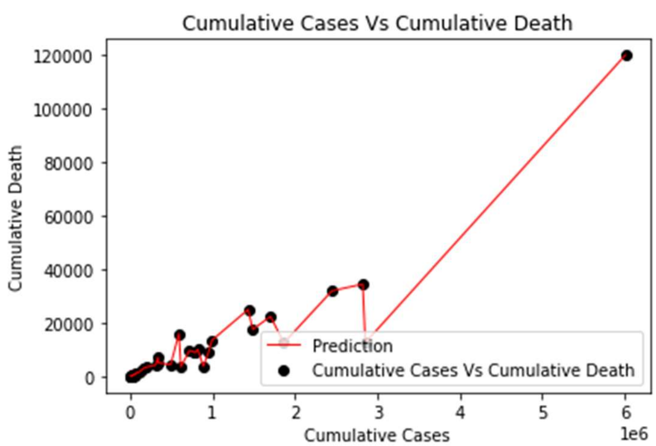

Fig. 5: Linear regression between cumulative cases and cumulative death cases

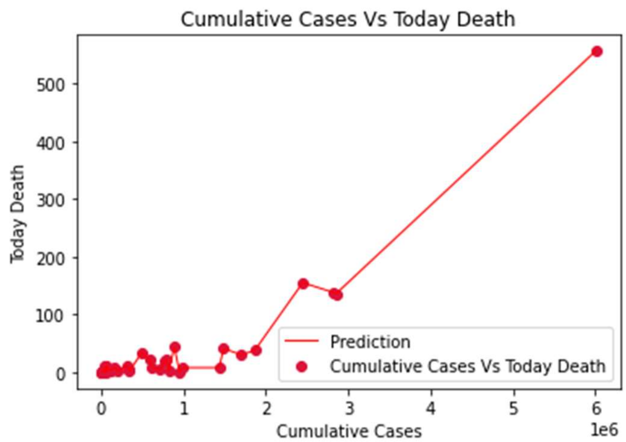

Fig. 6: Linear regression between cumulative cases and today death cases

Table 4: Prediction parameters and its accuracy using R2 score

Prediction Parameters	R2 Score	MAE	MSE	RMSE
Cumulative Cases Vs Today Cases	0.6270	617.6060	864431.2840	929.7480
Cumulative Cases Vs Active Cases	0.6749	11453.7746	393978090.1128	19848.8813
Cumulative Cases Vs Cumulative Discharged	**0.9996**	10209.6270	268584129.6927	16388.5365
Cumulative Cases Vs Today Discharged	0.7300	16388.5365	2572333.2731	1603.8495
Cumulative Cases Vs Cumulative Death	0.7099	5113.1159	34702768.6794	5890.9056
Cumulative Cases Vs Today Death	0.6436	20.9751	662.1190	25.7317

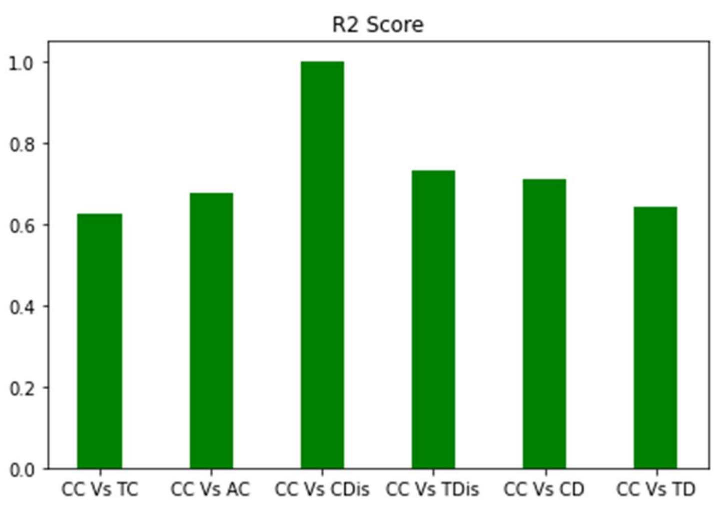

Fig. 7: Combination of different prediction accuracy using R2 score

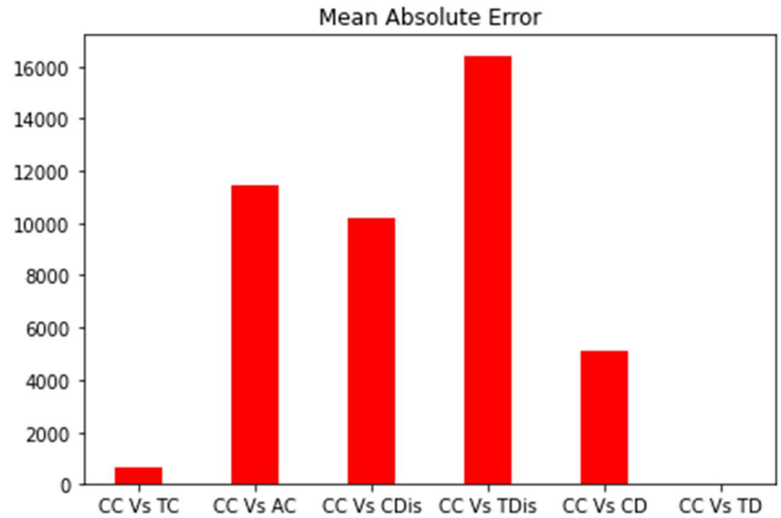

Fig. 8: Combination of different prediction accuracy using MAE

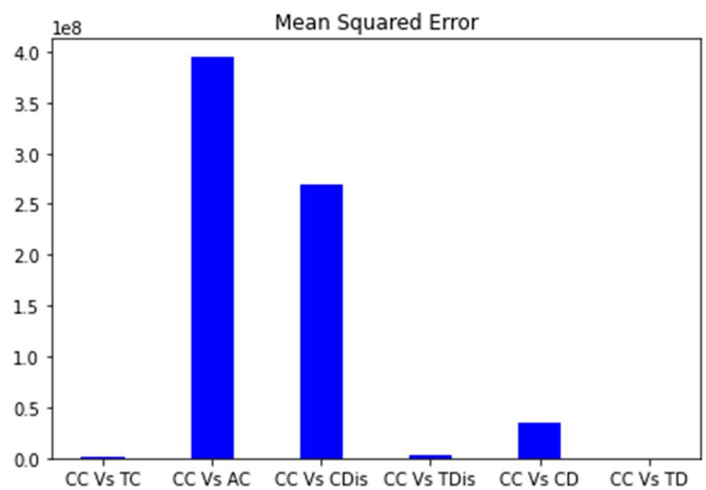

Fig. 9: Combination of different prediction accuracy using MSE

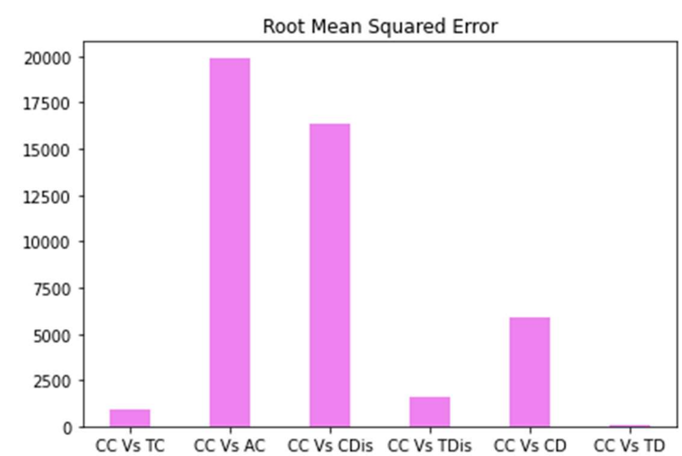

Fig. 10: Combination of different prediction accuracy using RMSE

Table 5: Prediction accuracy for actual cumulative discharge and predicted cumulative discharge using R2 score with 0.9996

Cumulative Cases	Actual Cumulative Discharge	Predicted Cumulative Discharge
6007431	5762661	5785506.7269
2854325	2741436	2752078.4259
2823444	2678473	2722369.5296
2449577	2367831	2362692.8366
1867017	1804844	1802244.1898
1705014	1679096	1646390.0938
1489286	1449462	1438850.1569
1433475	1406760	1385157.4910
992391	971662	960815.0525
951695	940771	921663.6894
890596	856498	862883.7322
822887	808418	797744.6608
789561	779432	765683.5656
768002	756679	744942.8481
720717	708586	699452.5737
617776	598139	600418.7503
593941	572723	577488.4195
493688	458330	481040.5718
345028	338698	338023.0357
339245	329432	332459.5323
313476	302655	307668.6074
201210	195624	199663.7190
165426	159677	165237.9185
115925	111114	117615.7534
66171	55912	69750.1911
63868	59378	67534.6026
61542	60488	65296.8871
46878	41647	51189.4665
34214	31487	39006.1349
24629	22641	29784.9383
19903	19387	25238.3160
19681	17101	25024.7421
18859	14316	24233.9415
10526	10463	16217.2246
9601	9232	15327.3335
7440	7214	13248.3553

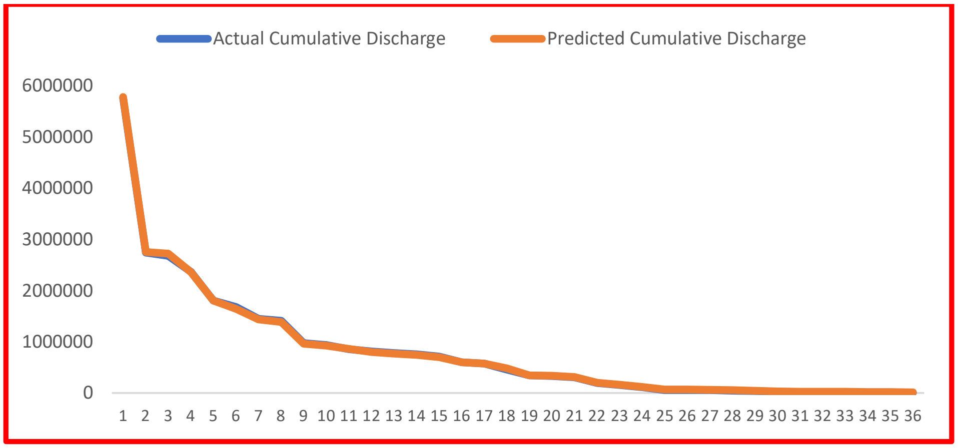

Fig. 11: Relationship between actual cumulative discharge and predicted cumulative discharge with R2 score 0.9996

4. Result and Discussion

In our study, the result of determining a trend in the Covid'19 dataset and its accuracy parameters shows in table 1 and table 2. The data were analysed using table 2, the descriptive statistics namely mean, median and SD (standard deviation) which is used to find the average covid'19 cases, average active cases, average discharge cases and average death cases. The main factors in table 3, which show the linear relationship model and its accuracy with 80% of training and 20% of testing. In this study clearly explain through table 5, the mean, median, SD, coefficients, variance, and accuracy are indicated.

According to the interpretation of table 3, figure 1 and figure 2 can say that the variable selection for cumulative cases and today cases having 0.6270 and cumulative cases and active cases relationship return 0.6749 with R2 score. In this combination the accuracy parameters return nearly 62% to 67% is not a suitable variable selection for predict the future.

According to the result of table 3 shows, linear regression model and its accuracy parameters. In this table explain different R2 score like 0.6270, 0.6236, 0.6749, 0.7099, 0.7300 and 0.9996. In this case, the R2 scores 0.9996 clearly indicate the strong relationship between cumulative cases and cumulative discharge. The aim is to establish a linear relationship between predictor variable and the response variable combination return R2 score nearly 1. Based on accuracy parameters the cumulative cases and cumulative discharge is the best variable for predict the future. The related numerical illustrations show in figure 3.

Combination of different prediction parameters and its accuracy mentioned in table 4 and figure 11. In this case, explain different accuracy parameter namely R2 score, mean absolute error, mean squared error, and root mean squared error clearly say that different types of errors. Based on this interpretation the relationship between cumulative cases versus today death having moderate error. Which means the relationship between predictor variable and the response variable difference is very low compared to others, but in this case the R2 score is 0.6436.

Based on figure 4, figure 5 and figure 6, the relationship between cumulative cases with the combinations parameters namely today discharged, cumulative death and today death. In this case, the R2 score return 0.7300, 0.7099 and 0.6436. The relationship returns the accuracy level only 64% to 73%. In these observations not suitable variable combination for predict the future. In according to the table 5 and figure 7, the relationship between actual cumulative discharge and predicted discharge with R2 score 0.9996. Based on numerical illustrations as mentioned in table 4, figure 8, figure 9 and figure 10, the accuracy parameters Cumulative Cases and Today Death having minimum errors and R2 score is 0.6436 as shown in figure 7.

5. Conclusion and Further Research

The main conclusions of this work are drawn together and presented in this section; the linear regression model makes it possible to acquire the variable selection using different accuracy parameters. Features of the data because it predicts Covid-19 cases using different parameters, the combination of cumulative cases and cumulative discharge is the best variable for predict the future. The model had 99% of the overall data for training, testing, and validation, respectively. Therefore, the prediction performance was seen to be high, as was the validation accuracy. The future work should incorporate to find the accuracy using different combination of the parameters. This should be considered in future experiments.

References

1. Jiawei Han, Micheline Kamber and Jian Pei. “Data Mining: Concepts and Techniques.” The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, 2011.
2. Mohan, S., Abugabah, A., Kumar Singh, S., Kashif Bashir, A. and Sanzogni, L., 2022. An approach to forecast impact of Covid-19 using supervised machine learning model. Software: Practice and Experience, 52(4), pp.824-840.
3. Rajesh, P., and Karthikeyan, M., 2017. A comparative study of data mining algorithms for decision tree approaches using WEKA tool. Advances in Natural and Applied Sciences, 11(9), pp. 230-243.
4. Albahri, A.S., Hamid, R.A., Al-qays, Z.T., Zaidan, A.A., Zaidan, B.B., Albahri, A.O., AlAmoodi, A.H., Khlaf, J.M., Almahdi, E.M., Thabet, E. and Hadi, S.M., 2020. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. Journal of medical systems, 44(7), pp.1-11.
5. Abdulkareem, N.M., Abdulazeez, A.M., Zeebaree, D.Q. and Hasan, D.A., 2021. COVID-19 world vaccination progress using machine learning classification algorithms. Qubahan Academic Journal, 1(2), pp.100-105.
6. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. and Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. IEEE access, 8, pp.101489-101499.
7. Sun, N.N., Yang, Y., Tang, L.L., Dai, Y.N., Gao, H.N., Pan, H.Y. and Ju, B., 2020. A prediction model based on machine learning for diagnosing the early COVID-19 patients. MedRxiv.
8. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. and Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. IEEE access, 8, pp.101489-101499.
9. Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C. and Mohammed, I.A., 2021. Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset. SN computer science, 2(1), pp.1-13.

10. Ballı, S., 2021. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos, Solitons & Fractals, 142, p.110512.
11. Ayyoub Zadeh, S.M., Ayyoubzadeh, S.M., Zahedi, H., Ahmadi, M. and Kalhori, S.R.N., 2020. Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study. JMIR public health and surveillance, 6(2), p.e18828.
12. Rustam, F., Reshi, A.A., Mehmood, A., Ullah, S., On, B.W., Aslam, W. and Choi, G.S., 2020. COVID-19 future forecasting using supervised machine learning models. IEEE access, 8, pp.101489-101499.
13. Covid 19 dataset was obtained from Government of India https://www.covid19india.org/
14. Ouf, S. and Hamza, N., 2021. The role of machine learning to fight COVID-19, International Journal of Intelligent Engineering and Systems, 14(2), pp. 121-135.
15. Sujath, R.A.A., Chatterjee, J.M. and Hassanien, A.E., 2020. A machine learning forecasting model for COVID-19 pandemic in India. Stochastic Environmental Research and Risk Assessment, 34(7), pp.959-972.