INTERNATIONAL RESEARCH JOURNAL ON ADVANCED SCIENCE HUB
e-ISSN : 2582 - 4376
Open Access
RSP SCIENCE HUB
(The Hub of Research Ideas)
Available online at www.rspsciencehub.com

Special Issue of Second International Conference on Advances in Science Hub (ICASH 2021)

Analyzing and Predicting Covid-19 Dataset in India using Data Mining with Regression Analysis

Dr. P. Rajesh [1], Vetrivel Govindarasu [2]

[1]Assistant Professor, PG Department of Computer Science, Government Arts College, C.Mutlur, Chidambaram, Tamil Nadu, India. (Deputed from Department of Computer and Information Science Annamalai University, Chidambaram, India)

[2] Senior Software Architect, Cognizant Technology Solutions, Thoraipakkam, Chennai, Tamil Nadu, India.

rajeshdatamining@gmail.com

Abstract

COVID-19 is a disease caused by coronavirus. 'CO' stands for corona, 'VI' for virus, and 'D' for disease. Formerly, this disease was referred to as '2019 novel coronavirus. The data mining is the best tools for analyzing and predicting the hidden information with the help of pre-existing dataset. The covid analysis and prediction for consider different related parameters namely name of the states, total cases, today cases, active cases, discharged cases, today discharged cases, overall death and today deaths. In this paper, taking consideration into analyzing and predicting covid dataset using statistical techniques namely regression model. Numerical illustrations also provide to prove the results and discussions.

Keywords: Covid-19, Data Mining, Regression model and Forecasting

1. Introduction

Data mining is the process of analyzing hidden patterns for using pre-existing data. Data mining is also known as data discovery and knowledge discovery for handing advanced data analysis. The major steps involved in a data mining process namely locate the data, data collection, data cleaning, integration, data selection, data transformation and discovering the knowledge. In data mining techniques, normalization is one of the most important concepts for prepare a well suitable dataset with unique format. Data mining is the process of analyzing hidden patterns for using pre-existing data. Data mining is also known as data discovery and knowledge discovery for handing advanced data analysis [1]. The major steps involved in a data mining process namely locate the data, data collection, data cleaning, integration, data selection, data transformation and discovering the knowledge [2]. The area of weather forecasting is used to collecting hugs amount of data as possible to find the current weather state of the atmosphere metrics namely temperature, humidity, and wind conditions [3]. Data mining techniques is easy to understand the atmospheric condition and to determine how to find the future atmosphere conditions using regression analysis [4]. In data mining techniques, normalization is one of the most important concepts for prepare a well suitable dataset with unique format. After using the normalization techniques various scales of information converted into similar scale of information. Various normalization techniques are also used to handling the data analysis, one of the most popular normalization techniques called maxima and minima normalization [5 - 8].

2. Experimental Methods or Methodology

Regression analysis is a statistical tool to launch a relationship between two or more variables. Likewise, one of these variables named as predictor variable which means value is collected via experiments. Another variable is named as response variable which means derived from the predictor. The general mathematical equation for a linear regression is,

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

Where y is the response variable, x is the predictor variable and a, b are constants which are called the coefficients.

Table.1. Covid-19 overall dataset in India with different cases

States	Total Cases	25.6.2021 Cases	Active Cases	Discharged	25.6.2021 Discharged	Deaths	25.6.2021 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
Puducherry	115925	298	3077	111114	276	1734	3
Manipur	66171	549	9174	55912	655	1085	11
Tripura	63868	369	3828	59378	400	662	2
Chandigarh	61542	22	247	60488	42	807	0
Meghalaya	46878	420	4424	41647	298	807	10
Arunachal_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
Lakshadweep	9601	42	322	9232	60	47	0
Andaman_Nicobar	7440	2	99	7214	4	127	0

Table.2: Statistical observations in Covid-19 dataset in India

Statistics	Total Cases	25.6.2021 Cases	Active Cases	Discharged	25.6.2021 Discharged	Deaths	25.6.2021 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. Regression model accuracy for training (80%) and testing (20%)

x	y	Accuracy
Total cases	Today cases	0.7309
Total cases	Discharged	0.9999
25.06.2021 Cases	25.06.2021 Discharged	0.9342

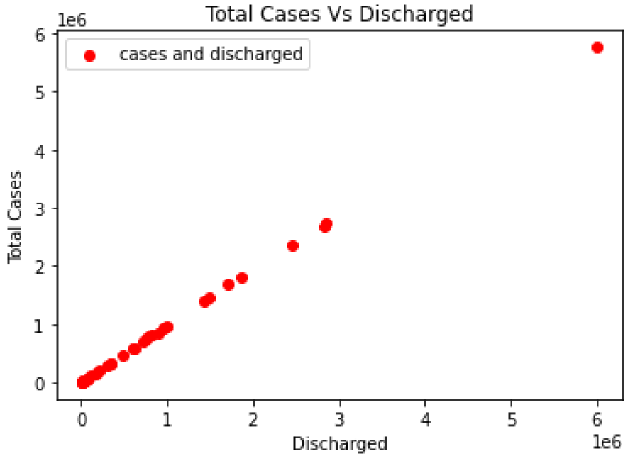

Chart.2. Regression model accuracy for total cases and discharged cases

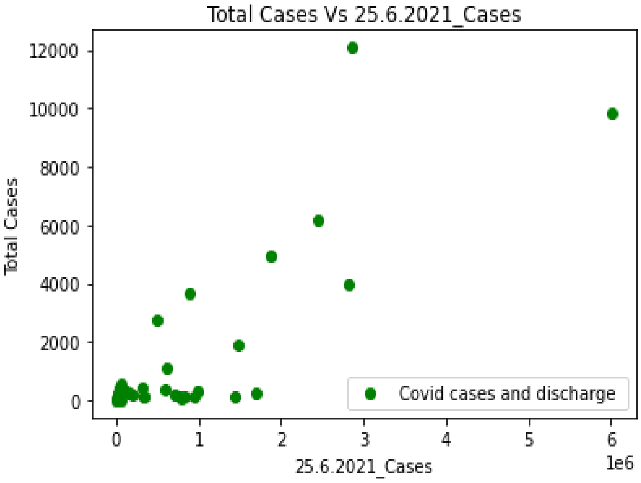

Chart.1. Regression model accuracy for total cases and today cases

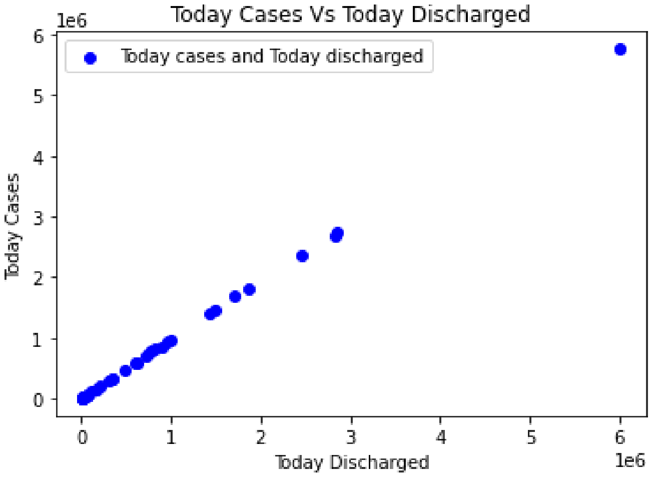

Chart.3. Regression model accuracy for today cases and today discharged cases

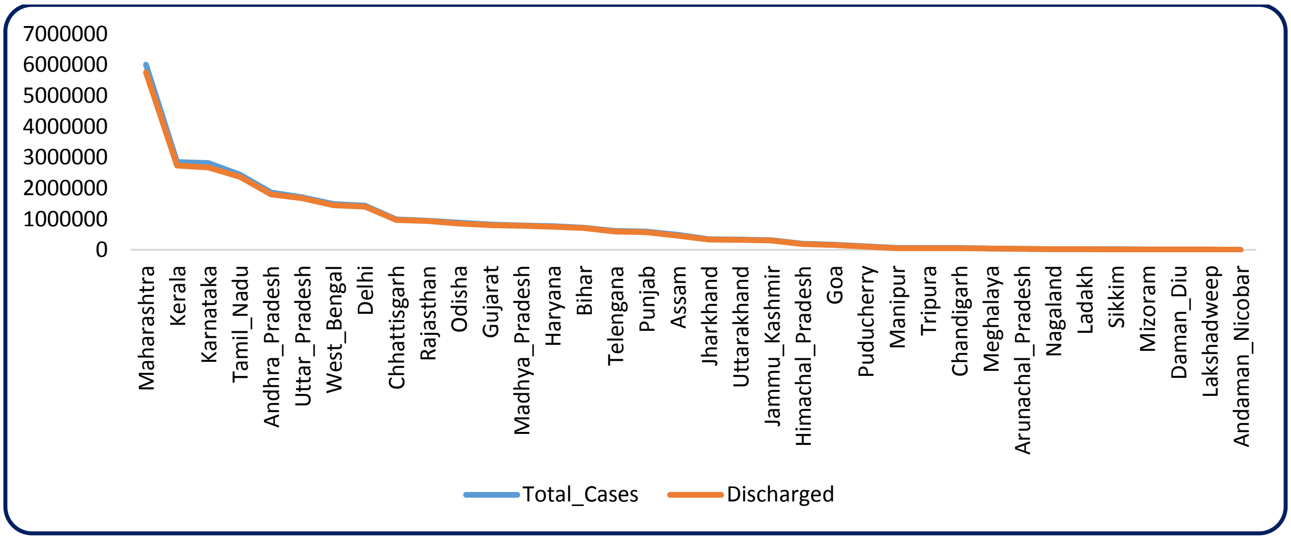

Chart.4. Linear machining in total cases and discharged

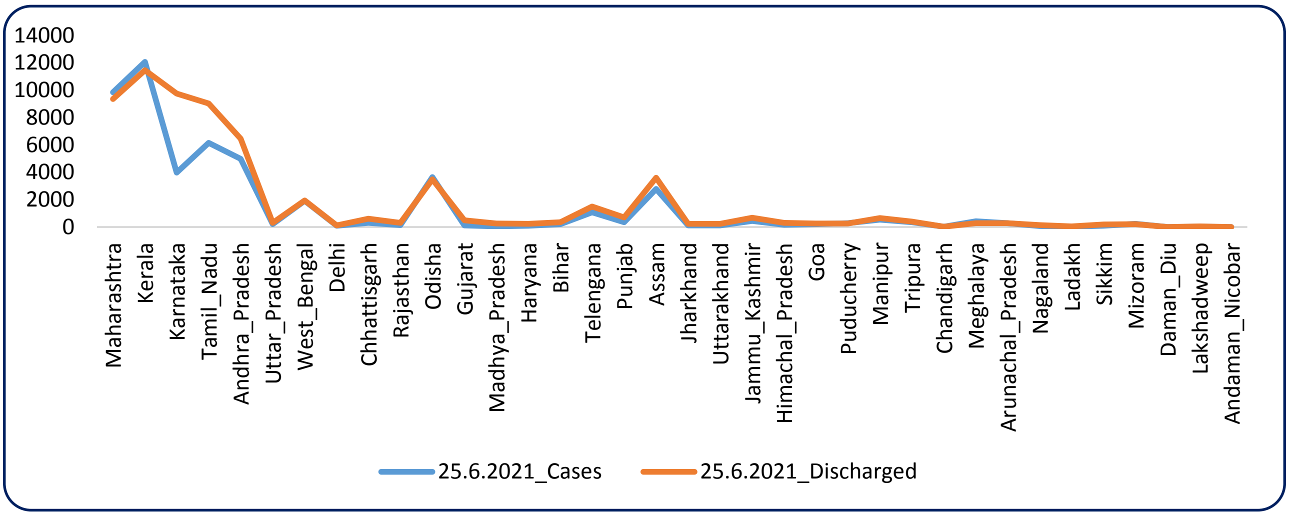

Chart.5. Linear machining of covid cases on 25.06.2021 (positive cases and discharged cases)

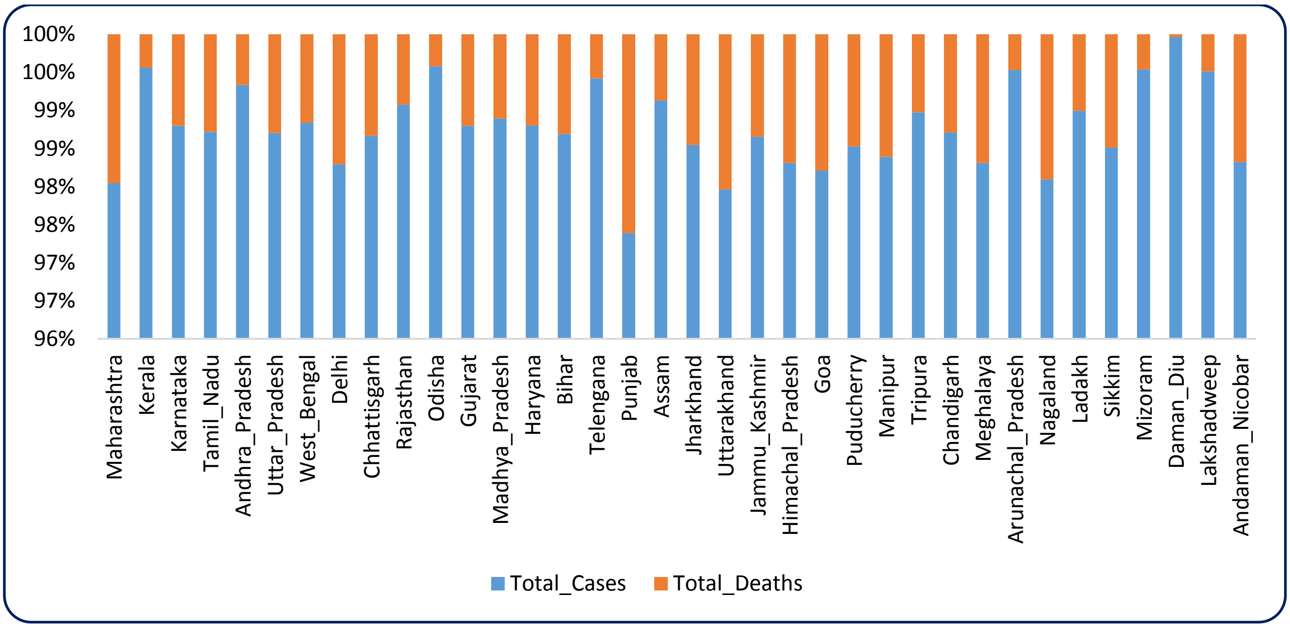

Chart.6. Linearity missing and data mining approaches in total cases and total deaths

3. Results and Discussion

The secondary Covid’19 related data collected from official website of Government of India (https://www.covid19india.org/). The website having various information related regarding Covid’19.

Conclusions

In this paper, consider different parameters namely states, total cases, today cases, active cases, discharged cases, today discharged cases, overall death and today deaths. Related dataset shows in table1. In table 2, indicate descriptive statistics, which is used to finding the average cases in India and how to deviate the in different parameters.

Regression model explain how to find the similarity or linearity with different parameters using table 3 and charts 1, 2 and 3. Numerical illustrations in table 3 and chart 3, the regression approaches only 73% in total cases Vs today cases. Based on numerical illustrations shows in chart 4 and 5, how to influence in the parameters which is satisfied the linear regression model and how many percentages occur the linearity. In chart 6, indicate some hidden information regarding in Kerala, positive cases are maximum at the rate of death cases compare to other states is also minimum. In Punjab, the positive cases are low compared to others. But the death ratio also high compared to other states. The graphical representation highlighted in chart 6. The regression model approach in today cases Vs today discharged (25.06.2021). In this case, the model accuracy having 93%. The regression model highly consider total cases Vs discharged cases. In this case, the model accuracy having 99%. In this research conclude that total cases Vs discharged cases having for better performance in future predictions.

References

[1]. Jiawei Han, Micheline Kamber and Jian Pei. (2011). “Data Mining: Concepts and Techniques”, The Morgan Kaufmann Series in Data Management Systems, Morgan Kaufmann Publishers, (2011).

[2]. Bocca, F. F., Henrique, L. and Rodrigues, A. (2016). The Effect of Tuning, Feature Engineering and Feature Selection in Data Mining Applied to Rainfed Sugarcane Yield Modelling. Computer and Electronic in Agriculture, 128, 67–76.

[3]. Rajesh, P. and Karthikeyan, M. (2019). Data Mining Approaches to Predict the Factors that Affect the Agriculture Growth using Stochastic Model. International Journal of Computer Sciences and Engineering, 7, 18-23.

[4]. Rajesh, P. and Karthikeyan, M. (2019). Data Mining Approaches to Predict the Factors that Affect the Groundwater Level using Stochastic Model. AIP Conference Proceedings, 2177, 20079-1–020079-11.

[5]. Rathod, S., et al. (2018). Modeling and Forecasting of Oilseed Production of India through Artificial Intelligence Techniques. Indian J. Agric. Sci., 88(1), 22–27.

[6]. Teixeira de Lima, G. R. and Stephany, S. (2013). A New Classification Approach for Detecting Severe Weather Patterns. Comput. Geosci., 57, 158–165.

[7]. Rajesh, P. and Karthikeyan, M. (2019). Data Assimilation of Gross Domestic Product (GDP) in India using Stochastic Data Mining Approach. Journal of Computational and Theoretical Nanoscience, 16(4), 1478–1484.

[8]. Rajesh, P., Karthikeyan, M., and Arulpavai, R. (2018). Predication of Labour Demand in Agriculture based on Comparative Study of Different Data using Data Mining and Stochastic Approach. International Journal of Engineering Science Invention, 2, 86-97.