



(REVIEW ARTICLE)



## Prediction of COVID-19 based on machine learning using cartographic variables

Kavitha Soppari \*, K. Chinmayi, G. Srikanth, K. Bhavani Reddy and CH Ramanavasulu

*Department of Computer Science and Engineering, ACE Engineering College Hyderabad Telangana, India.*

International Journal of Science and Research Archive, 2023, 09(02), 163–170

Publication history: Received on 19 May 2023; revised on 07 July 2023; accepted on 10 July 2023

Article DOI: <https://doi.org/10.30574/ijrsra.2023.9.2.0511>

### Abstract

The Coronavirus has rapidly spread to all parts of the world. Research is continuing to find a cure for this disease while there is no exact reason for this outbreak. As the number of cases to test for Coronavirus is increasing rapidly day by day, it is impossible to test due to the time and cost factors. Over recent years, machine learning has turned very reliable in the medical field. Using machine learning to predict COVID-19 in patients will reduce the time delay for the results of the medical tests and modulate health workers to give proper medical treatment to them.

A Systematic Literature Review is performed to identify the most suitable algorithms for the prediction model. Then through the findings of the literature study, an experimental model is developed for prediction of COVID-19 and to identify the features that impact the model.

Based on cartographic variables, it is possible to make predictions about the spread and impact of COVID-19 in different regions. These variables can include population density, age distribution, healthcare infrastructure, and mobility patterns.

Overall, a combination of cartographic variables can be used to develop predictive models that can help public health officials and policymakers better understand the trajectory of the pandemic and make informed decisions about resource allocation and mitigation strategies.

**Keywords:** COVID-19; Machine Learning; Prediction; Classification Algorithms Techniques; Naive bayes classifier algorithm

### 1. Introduction

The COVID-19 pandemic has had a profound impact on global health, economies, and societies worldwide. As efforts continue to combat the spread of the virus and find effective treatments, researchers are increasingly turning to machine learning techniques to aid in the prediction and management of COVID-19 cases. By cartographic variables and advanced data analysis, these predictive models hold the potential to provide valuable insights into the spread and severity of the disease.

Recent advancements in machine learning have demonstrated their effectiveness in various medical domains, including disease prediction and diagnosis. Applying machine learning algorithms to COVID-19 prediction models offers several advantages, such as reducing the time delay in obtaining test results and enabling healthcare workers to provide prompt and appropriate medical interventions to patients. Furthermore, integrating cartographic variables can enhance the accuracy of these models by capturing the geographic and demographic factors that play a significant role in the spread of the virus.

\* Corresponding author: Soppari Kavitha

The utilization of cartographic variables offers a comprehensive perspective on the pandemic's dynamics, as it takes into account the spatial distribution and contextual factors that influence disease transmission and severity. By integrating these variables into the predictive models, public health officials and policymakers can gain valuable insights into the potential trajectory of the pandemic, aiding in resource allocation and the formulation of effective mitigation strategies tailored to specific regions.

The objective of this research is to develop a predictive model that enhances existing machine learning techniques and cartographic variables to enhance our understanding of the spread and severity of COVID-19. By addressing the gaps in the existing literature and utilizing robust methodologies, this study aims to contribute to the growing body of knowledge in this critical area, ultimately supporting evidence-based decision-making and enabling more effective management of the ongoing pandemic.

---

## 2. Methodology

### 2.1. Existing methodology

Machine Learning, a subset of Artificial Intelligence (AI), has emerged from the domain of pattern recognition and has found applications in various fields such as healthcare. As a rapidly evolving discipline, Machine Learning aims to optimize computer performance by programming them to learn from data and past experiences. By automatically extracting patterns and trends from given data, Machine Learning can generate accurate predictions and build mathematical models based on statistical analysis[1].

In the context of COVID-19 prediction, existing methodologies involve the use of Regression and Naive Bayes classifier algorithms[2] Regression algorithms [3] are employed to make predictions with high accuracy, leveraging cartographic variables such as population density, healthcare infrastructure, and other relevant geographic data. By determining statistically significant relationships between variables, Regression algorithms facilitate the prediction of one variable based on the values of others.

On the other hand, the Naive Bayes classifier algorithm separates data into different classes using Bayes' Theorem, assuming that all predictors are independent of each other. This algorithm is particularly useful in COVID-19 prediction as it can identify relationships between features within a class, aiding in the classification of data into different categories.

These methodologies can suffer from over fitting, leading to inaccurate predictions, while the assumption of independence in Naive Bayes classifier may not hold in complex scenarios. Additionally, these methodologies may lack flexibility to adapt to evolving situations, and the quality and representativeness of the available data can impact their reliability. These algorithms don't use cartographic variables as major concern. The cartographic variables in machine learning models for COVID-19 prediction can limit the accuracy and comprehensiveness of the predictions. Cartographic variables provide valuable contextual information about population density, geographical features, and healthcare infrastructure, which are crucial in understanding the spatial dynamics of the disease. Without these variables, the models may overlook important factors that influence transmission patterns and regional disparities, hindering their ability to accurately predict the spread and impact of COVID-19.

### 2.2. Proposed methodology

In proposed methodology the regression algorithm is enhanced by using cartographic variables in the regression algorithm is to capture the cartographic patterns and geographical context that can influence the spread and severity of COVID-19. Cartographic variables, such as population density, healthcare infrastructure, and transportation networks, provide valuable information about the local conditions and dynamics of the disease. By incorporating these variables into the regression algorithm, we can account for the cartographic heterogeneity and variations in COVID-19 transmission and severity across different regions. This enhances the predictive power of the algorithm by considering the cartographic context and potential cartographic dependencies in the data. Moreover, the inclusion of cartographic variables allows for a more comprehensive understanding of the factors contributing to the spread and severity of COVID-19, enabling better predictions and insights for public health interventions and resource allocation. The working principle of enhanced regression algorithm is described in algorithm 2.1.

### 2.2.1. Regression Algorithm

#### Algorithm 2.1 Enhanced Regression Algorithm using Cartographic variables

- **Gather Data:** Collect a dataset with COVID-19-related variables and cartographic variables. Let the dataset be denoted as  $D$ .
- **Preprocess Data:** Clean and preprocess the dataset  $D$  by handling missing values, normalizing numerical variables, and encoding categorical variables.
- **Feature Selection:** Determine the relevant cartographic variables for regression analysis. Let the selected cartographic variables be denoted as  $X_c$ .
- **Split the Dataset:** Divide the preprocessed dataset  $D$  into training and testing subsets:  $D_{train}$  and  $D_{test}$ .
- **Train the Regression Model:** Build a regression model, such as linear regression, using the training dataset  $D_{train}$ . The model can be represented as  $y = f(X, X_c, \theta)$ , where  $y$  represents the COVID-19 outcome variable,  $X$  denotes the other relevant variables,  $X_c$  represents the cartographic variables, and  $\theta$  represents the model parameters. The goal is to learn the optimal parameter values that minimize the prediction error.
- **Evaluate the Regression Model:** Use the testing dataset  $D_{test}$  to evaluate the performance of the trained regression model. Calculate evaluation metrics, such as mean squared error (MSE) or R-squared, to assess the model's accuracy in predicting COVID-19 outcomes.
- **Refine the Feature Set:** Assess the impact of the cartographic variables by examining their coefficients or importance scores. Refine the feature set by selecting the most influential cartographic variables and excluding any irrelevant or redundant ones.
- **Repeat and Optimize:** Iterate the steps 4-7, refining the regression model and feature set, until the desired performance is achieved. Consider employing techniques like cross-validation or regularization to enhance the model's generalization capabilities.
- **Prediction on New Data:** Apply the trained regression model, including the refined cartographic variables, to predict COVID-19 outcomes on new or unseen data. Utilize the equation  $y = f(X, X_c, \theta)$  to obtain the predicted values.
- **Monitor and Update:** Continuously monitor the model's performance and update it as new data becomes available. Incorporate additional cartographic variables or adjust the feature set if necessary to improve the model's accuracy and relevance.

### 2.2.2. Naïve Bayes Classifier Algorithm

The Naive Bayes classifier separates data into different classes according to the Bayes' Theorem, along with the assumption that all the predictors are independent of one another. It assumes that a particular feature in a class is not related to the presence of other features. The working principle of enhanced Naïve Bayes Classifier algorithm is described in algorithm 2.2

#### Algorithm 2.2 Enhanced Naïve Bayes Classifier Algorithm using Cartographic variables

- **Data Preparation:** Gather a labeled dataset consisting of COVID-19-related variables, including cartographic variables, and the corresponding class labels indicating the spread or severity of the disease.
- **Data Preprocessing:** Clean the dataset by handling missing values and encoding categorical variables if necessary.
- **Feature Extraction:** Extract relevant features from the dataset, including both COVID-19-related variables and cartographic variables, that will be used to train the classifier.
- **Training Phase:**
  - Calculate the prior probabilities for each class label, denoted as  $P(C)$ , by counting the occurrences of each class in the training dataset.

- For each feature, calculate the likelihood probabilities, denoted as  $P(F_i | C)$ , by estimating the probability distribution based on the occurrences of each feature value given each class label. Depending on the type of feature, different probability estimation techniques can be used. For example, for continuous features, the Gaussian distribution can be used, and for discrete features, the multinomial distribution can be applied. The likelihood probabilities can be calculated using the following equation:
- $P(F_i | C) = (\text{Count}(F_i, C) + 1) / (\text{Count}(C) + N)$
- where  $\text{Count}(F_i, C)$  is the number of occurrences of feature  $F_i$  with class label  $C$ ,  $\text{Count}(C)$  is the total number of occurrences of class  $C$ , and  $N$  is the total number of distinct feature values for  $F_i$ .
- Optionally, apply smoothing techniques such as Laplace smoothing to handle cases where a feature value is unseen in the training data. This can be done by adjusting the numerator and denominator of the likelihood probability calculation accordingly.

**1. Classification Phase:**

- Given a new data instance with COVID-19-related and cartographic feature values, calculate the posterior probability for each class label using Bayes' theorem, considering both types of features. The posterior probability, denoted as  $P(C | F_1, F_2, \dots, F_n)$ , can be calculated using the following equation:

$$P(C | F_1, F_2, \dots, F_n) = (P(C) * P(F_1 | C) * P(F_2 | C) * \dots * P(F_n | C)) / P(F_1, F_2, \dots, F_n)$$

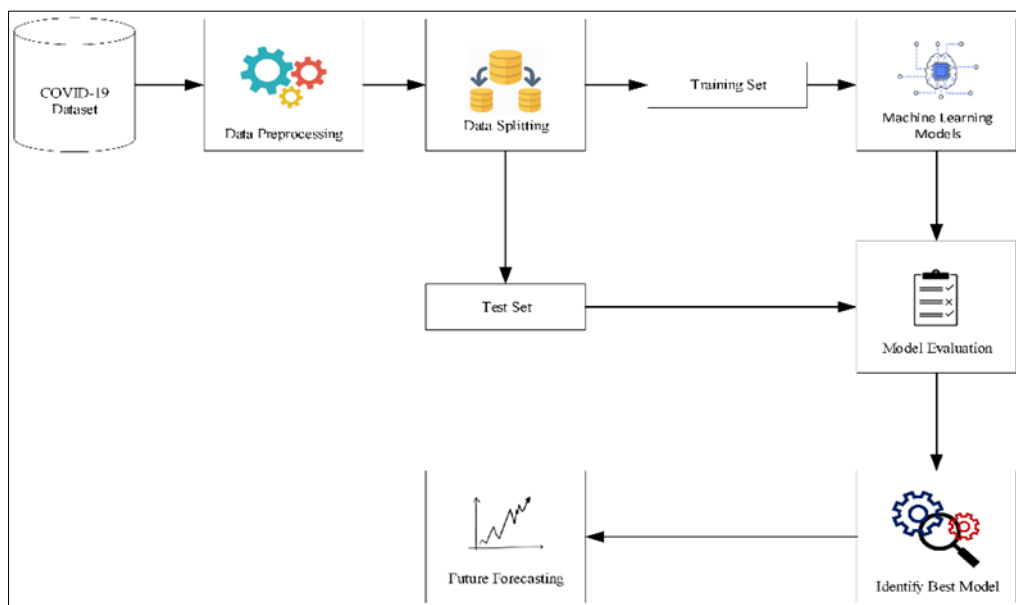
where  $P(F_1, F_2, \dots, F_n)$  is the probability of observing the given feature values, calculated as the sum of the joint probabilities for each class:

$$P(F_1, F_2, \dots, F_n) = \sum P(C) * P(F_1 | C) * P(F_2 | C) * \dots * P(F_n | C)$$

- Select the class label with the highest posterior probability as the predicted class for the new instance.

By incorporating cartographic variables into the Naïve Bayes Classifier algorithm, we enhance its ability to predict COVID-19 spread and severity by considering cartographic variables.

**2.3. Architectural diagram**



**Figure 1** Architectural diagram

**2.4. Data collection**

Collect data on Covid-19 cases, deaths, and recoveries from various sources, such as government health agencies, hospitals, and news reports. In addition, collect data on relevant cartographic variables that could impact the spread of the virus, such as population density, climate, air quality, and transportation infrastructure.

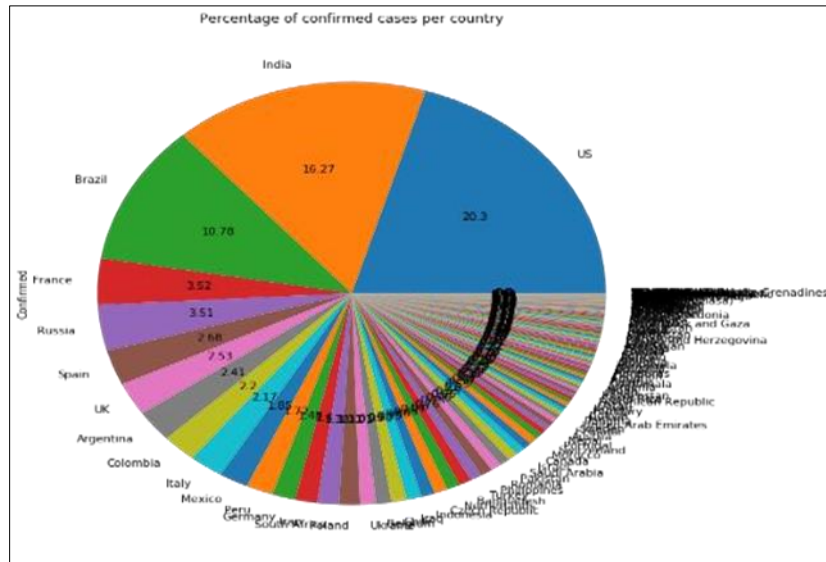


Figure 2 Worldwide cases of covid-19

### 3. Case studies in India

The first case of coronavirus in India was identified on 30 January 2020. By 3 February, the number of cases increased to 3. On 4 March, 22 new cases were identified, of which 14 were from a group of tourists who had arrived from Italy (The Week 2020). In March, India also reported its first coronavirus-related death. The number of confirmed cases in India crossed 1000 on 29 March, 30,000 on 28 April, and 180,000 on 30 May. The death toll crossed 50 on 1 April, 1000 on 28 April, and 5000 on 30 May. As of 25th November 2020, the numbers of infected cases and deaths are 9,227,557 and 134,804, respectively (World meter 2020b). On 24 March 2020, the Government of India under Prime Minister Narendra Modi ordered a nationwide lockdown for 21 days, limiting movement of the entire 1.3 billion population of India as a preventive measure against the COVID-19 pandemic in India. It was ordered after a 14-hour voluntary public curfew on 22 March, followed by enforcement of a series of regulations in the country's COVID-19 affected regions.

Observing the cases in India. Confirmed cases are increasing in India each day. There is a need to get a flatter curve for confirmed cases which currently is in upswing with a steep increase since past few days.

There are 3 datasets being used here

- Number of Confirmed cases
- Number of Deaths
- Number of Recovered cases

All this three data sets includes:

- Province/State
- Country/Region , Both of them includes the places around the world

#### 3.1. Source of data set

Collection of data and Importing The dataset used for this project was directly obtained from Kaggle, an open source repository. This are the following datasets are obtained:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
154	#####	6:00 PM	Maharash		32	0	0	0	32														
155	#####	6:00 PM	Punjab		1	0	0	0	1														
156	#####	6:00 PM	Rajasthan		2	2	3	0	4														
157	#####	6:00 PM	Tamil Nad		1	0	0	0	1														
158	#####	6:00 PM	Telangana		3	0	1	0	3														
159	#####	6:00 PM	Jammu an		2	0	0	0	2														
160	#####	6:00 PM	Ladakh		3	0	0	0	3														
161	#####	6:00 PM	Uttar Prad		12	1	4	0	13														
162	#####	6:00 PM	Uttarakha		1	0	0	0	1														
163	#####	6:00 PM	Andhra Pri		1	0	0	0	1														
164	#####	6:00 PM	Delhi		7	0	2	1	7														
165	#####	6:00 PM	Haryana		0	14	0	0	14														
166	#####	6:00 PM	Karnataka		6	0	0	1	6														
167	#####	6:00 PM	Kerala		23	0	3	0	23														
168	#####	6:00 PM	Maharash		32	0	0	0	32														
169	#####	6:00 PM	Odisha		1	0	0	0	1														
170	#####	6:00 PM	Punjab		1	0	0	0	1														
171	#####	6:00 PM	Rajasthan		2	2	3	0	4														
172	#####	6:00 PM	Tamil Nad		1	0	0	0	1														
173	#####	6:00 PM	Telangana		3	0	1	0	3														
174	#####	6:00 PM	Jammu an		3	0	0	0	3														
175	#####	6:00 PM	Ladakh		4	0	0	0	4														
176	#####	6:00 PM	Uttar Prad		12	1	4	0	13														
177	#####	6:00 PM	Uttarakha		1	0	0	0	1														
178	#####	6:00 PM	Andhra Pri		1	0	0	0	1														
179	#####	6:00 PM	Delhi		8	0	2	1	8														
180	#####	6:00 PM	Haryana		1	14	0	0	15														

Figure 3 Sample Data Set

### 3.2. Confirmed cases increased in India till 15<sup>th</sup> Nov 2020

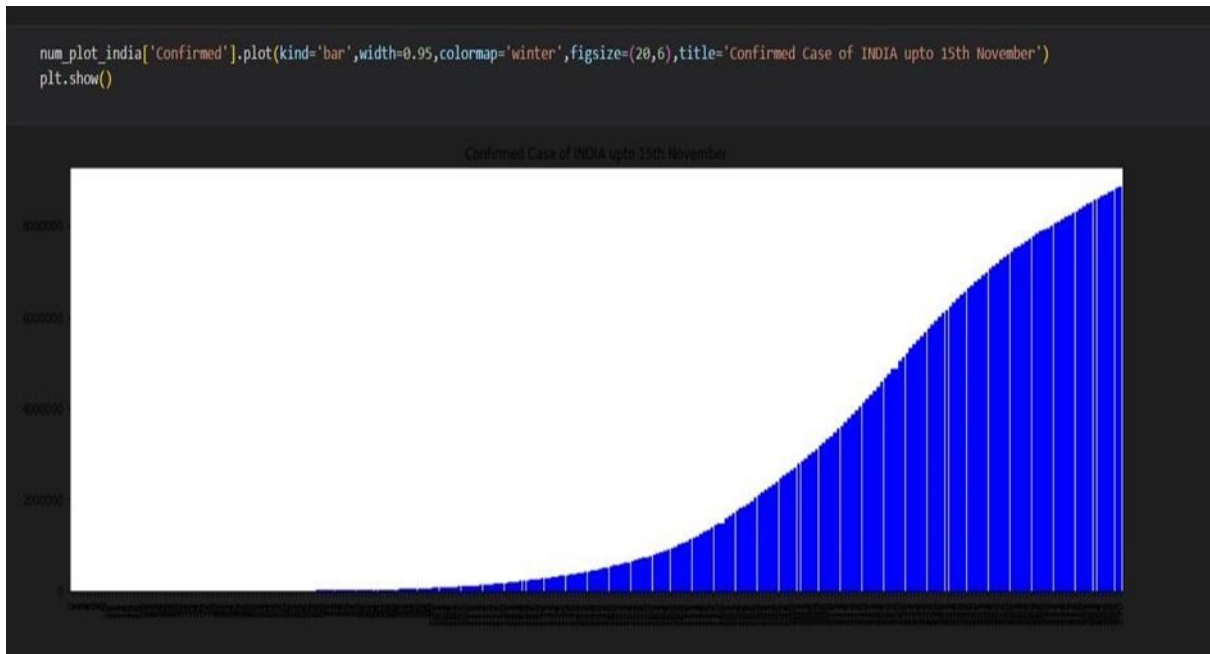
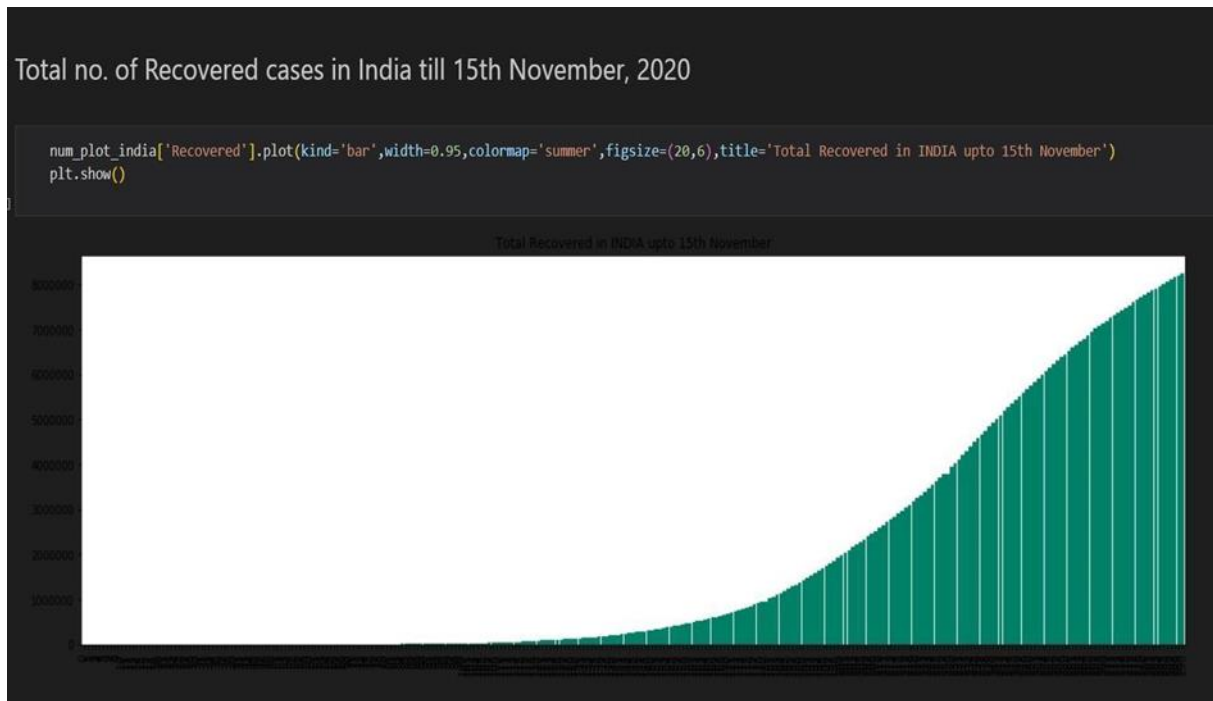


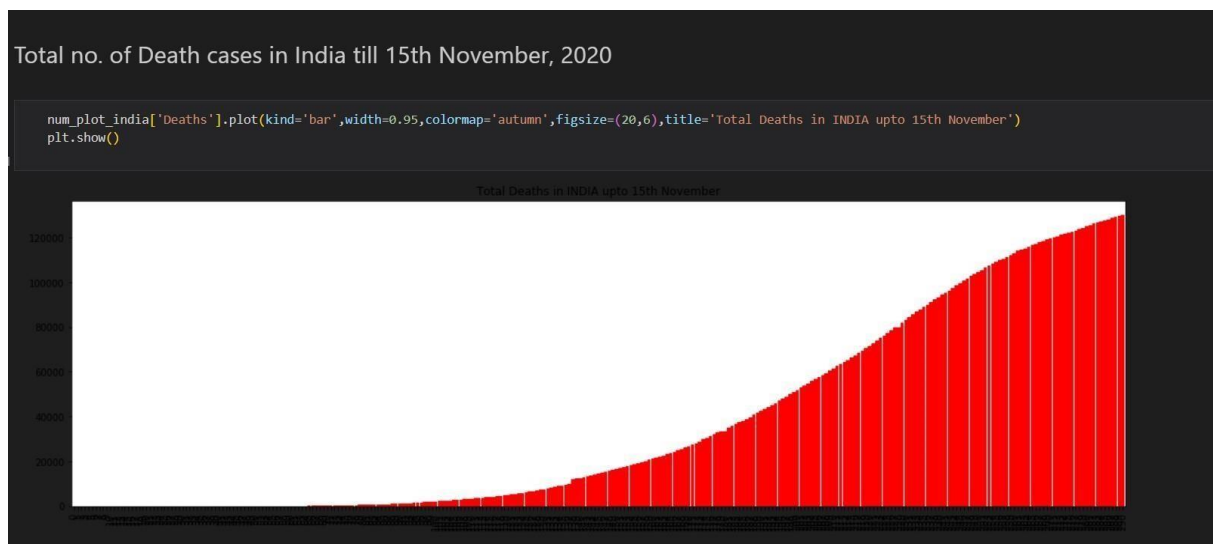
Figure 4 Confirmed Cases in India till 15<sup>th</sup> Nov 2020

### 3.3. Recovered cases increased in India till 15<sup>th</sup> Nov 2020



**Figure 5** Recovered Cases increases in India till 15<sup>th</sup> Nov 2020

### 3.4. Death cases increased in India till 15th Nov 2020



**Figure 6** Death Cases increased in India till 15<sup>th</sup> Nov 2020

### 3.5. Role of cartographic variables for prediction

A visual variable, in cartographic design, graphic design, and data visualization, is an aspect of a graphical object that can visually differentiate it from other objects, and can be controlled during the design process.

Cartographic variables play an essential role in prediction.

The properties of cartographic variables are applied to the geometric elements used to visualize geographical information: position, shape, size, value, texture and orientation.

#### 4. Conclusion

The objective of this paper is to predict the COVID-19 based on cartographic variables. COVID19 Prediction on different techniques are applied Regression, Decision trees and their accuracy and performance has been compared. It is obtained that Random COVID-19 prediction with accuracy.

#### Compliance with ethical standards



##### *Disclosure of conflict of interest*

No conflict of interest to be disclosed.

#### References

- [1] Malki, Zohair, et al. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. *Chaos, Solitons & Fractals* 138 (2020): 110137.
- [2] Kwekha-Rashid, Ameer Sardar, Heamn N. Abduljabbar, and Bilal Alhayani. "Coronavirus disease (COVID-19) cases analysis using machine-learning applications." *Applied Nanoscience* 13.3 (2023): 2013-2025.
- [3] Vanaja, M. R., Pravallika, M. S., Venkateswarlu, M. M., Yuvakiran, M. G., & Rakesh, M. D. EARLY DETECTION AND PREVENTION OF CARVICAL CANCER
- [4] Sharma, S.K. and Paliwal, M., 2023, February. Overview of data mining with Python modules. In *AIP Conference Proceedings* (Vol. 2427, No. 1). AIP Publishing.
- [5] Satpathy, Parmeshwar, Sanjeev Kumar, and Pankaj Prasad. "Suitability of Google Trends™ for digital surveillance during ongoing COVID-19 epidemic: a case study from India." *Disaster medicine and public health preparedness* 17 (2023): e28.
- [6] Satpathy P, Kumar S, Prasad P. Suitability of Google Trends™ for digital surveillance during ongoing COVID-19 epidemic: a case study from India. *Disaster medicine and public health preparedness*. 2023;17:e28
- [7] Rinner, C. (2023). *Pandemic Open Data: Blessing or Curse*.
- [8] Fuest, S., Shkedova, O., & Sester, M. (2023). Promoting favorable routes through visual communication: a design study for creating 'Social' route maps for the case of air pollution. *International Journal of Cartography*, 1-26.

#### Authors short biography

	<b>Dr. Kavitha Soppari</b> holds Ph.D in CSE from JNTUH. She has around 25 years of Teaching Experience in various Engineering Colleges. She is currently working as Associate Professor in Department of CSE, ACE engineering College. Her Area of Interests include Machine Learning , Artificial Intelligence, Network Security, Image Processing, etc.,
	<b>Ms.K.Chinmayi</b> holds B.Tech in CSE from JNTUH.She is the student of ACE Engineering College Who has profound intrest in area of Machine Learning and Artifical Intelligence.