Chapter 10

COVID-19 Analysis, Prediction, and Misconceptions:

A Computational Machine Learning Model as a New Paradigm in Scientific Research

Balachandran Krishnan
https://orcid.org/0000-0002-9051-8801
CHRIST University (Deemed), India

Sujatha Arun Kokatnoor
CHRIST University (Deemed), India

Vandana Reddy
CHRIST University (Deemed), India

Boppuru Rudra Prathap
https://orcid.org/0000-0002-5161-4972
CHRIST University (Deemed), India

ABSTRACT

COVID-19 is an infectious disease of the newly discovered coronavirus (CoV). The importance and value of open access (OA) resources are critical in the context of the COVID-19 epidemic. OA aided in the development of a vaccine and informed public health actions necessary to stop the virus from spreading. Many publishers implicitly acknowledged that OA was vital to promote science in the fight against the disease. Accordingly, publishers have committed to OA publication and scholarly communication of disease-related scientific research. This chapter covers three issues based on the modeling of the CoV dataset. First, an exploratory data analysis is done to detect the hidden facts and the relevant information patterns about the affected, recovered, death cases caused by the CoV and the vaccination details. Second, a predictive model is developed using machine learning techniques to effectively predict the number of COVID-19 positive cases in India. In the last step, a hybrid computational model is developed to identify the misconceptions that are spread through social media networks.

DOI: 10.4018/978-1-7998-9805-4.ch010

INTRODUCTION

Scholarly journals have been turned into online publications/journals with the advent of the internet, and have developed numerous beneficial capabilities such as online submission, searching, indexing and referring to many items beyond merely citation referencing for improved scholarly communication. The importance and value of Open Access are critical in the face of the COVID-19 epidemic. Open access to scientific information and open data aids in the development of a vaccine and informs public health actions necessary to stop the virus from spreading. Open access resources keep citizens informed and educated about the virus, ensuring that they follow public health recommendations and allowing for distance study.

The novel coronavirus (COVID-19) was widely replicated in China at the end of 2019, infecting a substantial proportion of people. The coronavirus is a family of viruses capable of causing a variety of diseases that are life threatening to humans, including common and more severe forms of cold. The signs and symptoms of the disease may occur within two to 14 days after exposure. This time referred to as the incubation period is the time after exposure and before symptoms. The general signs and symptoms include fever, cough, tiredness, breathing difficulty, sore throat, running nose, headache and chest pain (Sear, R. F. et al., 2020). Other less common signs also include rash, nausea, vomiting and diarrhea. Some people may only have a few symptoms and some may not have any symptoms at all. These cases are referred to as cases, symptomatic and asymptomatic respectively.

As per the World Health Organization (WHO), data have shown that the virus spreads from person to person (about 6 feet or 2 meters) among the people in close contact. The virus spreads through respiratory droplets when someone is coughing, sneezing or talking. Such droplets may be inhaled or landed in a nearby person's mouth or nose. It can also spread when a person touches a surface and touches his or her mouth, nose or eyes, but this is not a major way of spreading the virus as per WHO reports (Saba, T. et al., 2021). In the case of symptoms (symptomatic), a person with the virus is the most infectious – and this is the time that they are most likely to transmit the virus – according to the Center for Disease Control and Prevention (CDC) trusted Source. But even before they start showing symptoms (asymptomatic) of the disease itself, someone can spread it.

India had the world's second highest (after the US), with 29.3 million cases of COVID-19 infections documented, and the third largest number of COVID-19 deaths (after the US and Brazil) with 367.081 deaths as of 12th June 2021. A second wave started in March 2021 with shortage of vaccines, hospital beds, oxygen cylinders and others in the various sections of the country being significantly larger than that of the first one. India led the globe in new and active cases by the end of April. In a 24-hour period on 30th April 2021, the country was the first to record more than 400,000 new cases. Health experts feel that India has underreported its data owing to a number of circumstances.

This chapter aims to study over time cumulative data on confirmed cases, deaths and recovered cases, and to analyze the transmission of this virus across India in the first step. It is feasible to acquire insight into how each state performed in COVID-19 using this data. During what time period was the particular condition successful, so that other Indian states might learn from their processes during that time period. In the second step, a predictive model is developed using the machine learning techniques to effectively predict the number of COVID_19 positive cases in India. AutoRegressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average with eXogenous factors (SARIMAX), FBProphet, Logistic Regression, Linear Regression, Ridge Regression, Decision Trees, Random Forest and Neural Networks are used for the predictive analysis in this chapter. In the last step, a hybrid computational model is developed to identify the misconceptions that are spread through social media networks through public's tweets. The efficacy of the Misconception Detection System is tested on Corona Pandemic Dataset extracted from Twitter posts. For categorizing the dataset into two classes, FST and a weighted TF-IDF Model are utilized, followed by a supervised classifier: one with COVID-19 virus misconceptions, and the other with genuine and authorized information.

BACKGROUND

The literature review is carried out based on the sub-topics addressed in this chapter.

Exploratory Data Analysis:

Exploratory Data Analysis (EDA) is a data analysis approach that enables the discovery of hidden information within a data collection. This technique is frequently used to derive inferences from data. COVID-19 data are publicly accessible via the standard dataset repository. These widely available datasets are used to derive conclusions (DSouza, J., & Velan, S. S. 2020). Data visualization helps to understand the impacts of the pandemic on the variables/labels in the dataset.

According to experts, the number of confirmed, recovered, and deadly COVID-19 cases in India is anticipated to increase (Mahdavi, M. et al., 2021). Predictions are made using correlation coefficients and Multiple Linear Regression, with autocorrelation and autoregression employed to enhance prediction accuracy. The investigation is based on occurrences in several Indian states and is presented in chronological sequence (Varshney, D., & Vishwakarma, D. K., 2021). After data preprocessing, prediction analysis is done using Random Forest, Linear Regression Model, Support Vector Machine (SVM), Decision Tree, Neural Networks, Random Forests and so on. The Susceptible-Infected-Removed (SIR) model (Tutsoy, O. et al., 2020) is commonly used to estimate COVID-19 casualties.

The predicting techniques are generally categorized into two types: mathematical theory and stochastic theory (data science / machine learning techniques). The study generally includes statistical, analytical, mathematical and medical parameters (symptomatic and asymptomatic). The parameters cover various reasons behind the cause of coronavirus disease amongst people. Asymptomatic parameters include people's details who didn't show any signs of the disease yet they had it and the symptomatic parameters include people's details with fever, cough, tiredness and difficulty in breathing (Shinde, G. R. et al., 2020).

For anticipation of the disease epidemiological trend and rate of COVID-19 in India, Linear Regression, Multilayer Perceptron and the Vector Auto - Regression models are used. The prediction model is based on the cases which are in primitive stages and the Spearman's correlation is used to find the similarity between the features present in the dataset (Sujath, R. et al., 2020). As the dataset considered is non-linear, and dependent on each other, Spearman's Rank coefficient has led to inaccurate forecasting of the spread of the disease.

Several technologies including Blockchain technology, Internet of Things, Artificial Intelligence, Machine Learning, 5G and Unmanned Aerial Vehicles are used to reduce the impact of corona virus disease outbreak by analyzing the datasets available (Chamola, V. et al., 2020).

COVID-19 Prediction Analysis

(Bharadwaj, S. et al., 2013) highlighted how recent developments using Machine Learning (ML) and Artificial Intelligence (AI) are used for COVID-19 analysis and prediction. (Shorten, C. et al., 2021) suggested a deep learning model for the analysis of COVID-19 outbreak. (Mahdavi, M. et al., 2021) discussed the COVID-19 crisis using IoT and ML algorithms. According to (Alsunaidi, S. J. et al., 2021), important multisource urban variables (including temperature, relative humidity, air quality, and influx rate) affect daily new confirmed cases during early pandemic transmission stages. Another recent research (ArunKumar, K. E. et al., 2021) shows how machine learning algorithms can estimate the amount of incoming COVID-19 cases. Researchers used four forecasting models to anticipate COVID-19's risk variables: Linear Regression, Least Absolute Shrinkage and Selection Operator, Support Vector Machine (SVM), and Exponential Smoothing (ES). Each model anticipates additional infections, deaths, and hospitalizations.

In (Mahdavi, M. et al., 2021) article, three SVM models are created and evaluated on three separate groups of people: invasive, non-invasive, and both. Non-invasive factors provide mortality estimates equivalent to intrusive features and the combined model. Also, the model outperformed the invasive model with fewer features based on SVM-RFE (Recursive Feature Elimination) and sparsity analysis, revealing predictive information content in terms of SPO2 (Oxygen Saturation) and cardiovascular diseases. Time-series analysis and machine learning algorithms are used (Li, L. et al., 2020) for analyzing infected cases and fatalities caused due to COVID-19.

To analyze multivariate time series evolution, a cluster-based method named Hierarchical clustering is used for the COVID-19 pandemic. Countries are divided into clusters on a daily basis, according to their cases and death numbers. Algorithmically, the total number of clusters and the membership of individual countries is determined. This analysis gives new insights into COVID-19 ‘s spread across countries and through time (Rustam, F. et al., 2020). Hierarchic clustering seldom provides the best solution, as it involves a lot of arbitrary choices, does not work with missing data, works poorly with mixed data types, is doesn't work well on huge data sets, and is commonly misinterpreted with its main output, the dendrogram.

In the data set from different regions of China, obtained from the WHO, the K-means clustering based machine learning method is used. Within the original WHO data set the temperature area is included to demonstrate the effect of temperature on each region within three separate COVID-19 perspectives – suspected, verified, and death (Abd-Alrazaq, A. et al., 2020). It is observed that temperature is not the only factor for the spread of the corona disease. There are several other factors for the spreading if included as attributes for the data analysis, a better model of avoidance can be emerged.

Identification of CoV Misconceptions in Social Media Networks

People utilize social networking sites such as Twitter® to express themselves, report events, and provide a worldwide perspective. During the COVID-19 outbreak, users used Twitter® to share data visualizations from news outlets and government agencies, as well as their own. During the COVID-19 epidemic, few people were also bombarded with incorrect and misleading information. To study a framework that can automate methods of combating the COVID-19 epidemic in smart cities, Mohammed N. (Alenezi, M. N. & Alqenaei, Z. M., 2021) proposed viable models for detecting misinformation. The suggested models include Long Short-Term Memory (LSTM) networks, a subclass of Recurrent Neural Networks (RNN);

Multichannel Convolutional Neural Networks (MC-CNNs) and K-nearest Neighbours Networks (KNN). (Kowsari, K. et al., 2019) examined Machine Learning and Deep Learning approaches for the identification of misleading information. D. (Kokatnoor, S. A. & Krishnan, B., 2020) suggested a technique which utilizes Context Knowledge, Distance Metric, and Word Resemblance to select crucial evidence based on news item titles and content found on the top 10 Google search results relevant to the COVID-19 information spread. This study created a COVID fake news dataset for future research and assessment.

Automated extraction of social media and Natural Language Processing (NLP) discussion of COVID-19 is done based on topic modeling for detecting topics relevant to COVID-19 from public views. In addition, LSTM's Recurrent Neural Network is explored for characterizing COVID-19 emotions. Results illustrate the importance of using public views and relevant computational methods to consider and educate the decision-making process in connection with COVID-19 issues (James, N. & Menzies, M., 2020)

Machine learning is used to quantify COVID-19 contents of establishment of health guidance, especially vaccines amongst online opponents. User's posts on Facebook are analyzed for both anti vaccination and pro-vaccination communities. Snowball's approach is used for scraping user's posts which discuss either vaccines or policies about vaccination or an argument on pro and anti-vaccination for the COVID-19 disease. Latent Dirichlet Allocation algorithm is used for analyzing the appearance and involvement of topics on COVID-19 (Schaar, M. V. D. et al., 2021).

Situational information from social media data on COVID-19 is identified, analyzed and classified using Natural Language Processing techniques into seven types of situational information. They are cautions and advice, measures taken, donations, emotional support, seeking help, criticizing and rumor spreading. The dataset is manually labeled, and later SVM, Naïve Bayes and Random Forest algorithms are used for the classification (Hossain, T. et al., 2020). The limitations are that the social media data doesn't come with a label and manual labeling is very time consuming and is limited to one's domain expertise.

Latent Dirichlet Allocation (LDA) which is a topic modeling algorithm is used in the grouping of similar tweets which occur in the same user to user communication channel. Cosine Similarity is used for extracting the topmost ten tweets (Alvarez-Melis, D. & Saveski, M., 2016). The grouping done by considering hashtags caused duplication of the tweets and thus took a lengthier training time thereby reducing the performance of the model.

The important topics posted by the public in twitter are identified using the online LDA topic model. A total of twelve different topics are identified which are consolidated into four main categories: Virus Origin, Virus resources, Virus Impact Factor on the Public, Countries and the Economy and the last category is the identification of ways of mitigating the risk of infection (Siddiqui, M. K. et al., 2020). The regular online LDA uncorrelated topics could not be captured due to the topic's distribution in the tweets collected. The number of topics in the dataset are specified by the authors which is subjective and doesn't always highlight the true distribution of topics.

MAIN FOCUS OF THE CHAPTER

Statistics and data, such as health and geolocation, can be used to combat COVID-19 in a variety of ways, including mapping the outbreak's location, tracking COVID-19 deaths and recovered cases, tracking vaccination details and disease spread, evaluating the impact of governments' virus-containment efforts, and providing targeted information in high-risk areas. There is a need for transparent and open communication among researchers. In procedures that develop answers to the grave health threat that is likely to cause substantial hardships to humanity, it is critical to utilize scientific advances and support ideals of openness and participation (Waltman, L. et al., 2021).

Therefore, scholarly scientific information, COVID-19 data, experimentation and analysis are essential components in the development of new theoretical information. It is critical to recognize that the creation of new theoretical information to address emergency risk management requires an open and an unconditional access to and sharing of scientific contents, technologies, and processes by the entire scientific community from both developed and developing countries. To find a cure for the ongoing crisis, access to verified and peer-reviewed data and journal papers is essential. Verified knowledge and sophisticated scientific study can also keep the public informed about the situation and assuage worries sparked by ignorance or misinformation.

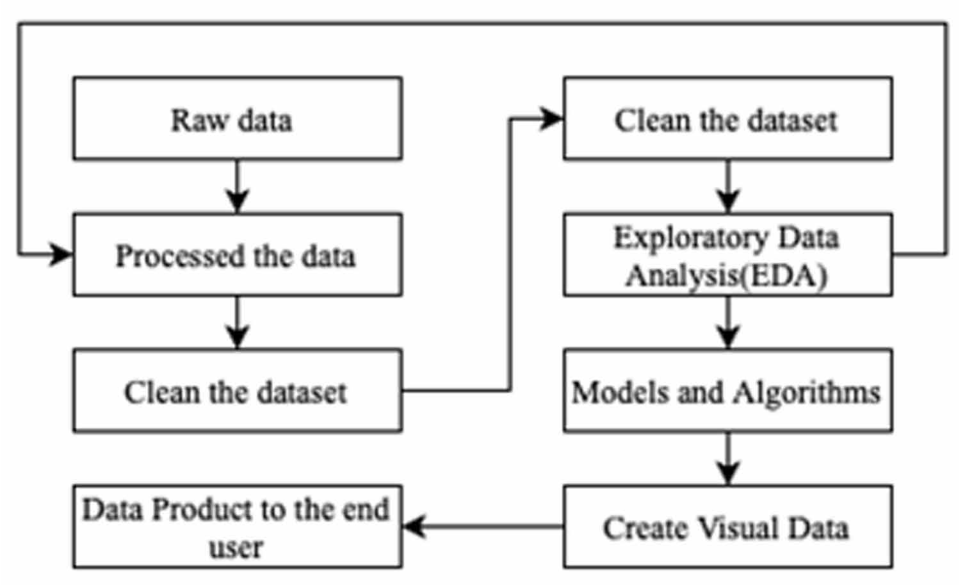
The chapter mainly focuses on three things:

1. Exploratory Data Analysis of CoV dataset
2. COVID-19 Prediction analysis
3. Identification of CoV misconceptions in social media networks.

Exploratory Data Analysis (EDA)

EDA is termed as numerical/graphical work which is required to be one in the initial stage of data processing and analysis. Converting the statistical data into graphical data makes the record readable and EDA is useful in this perspective. EDA is the first step that lays an accurate foundation to start the data analysis. Figure 1 gives the raw data processing in the data science field. To understand the number of distinct cases reported (confirmed, death, and recovered) in different Indian provinces, an EDA with visualizations is conducted.

Figure 1. Process of raw data in data science

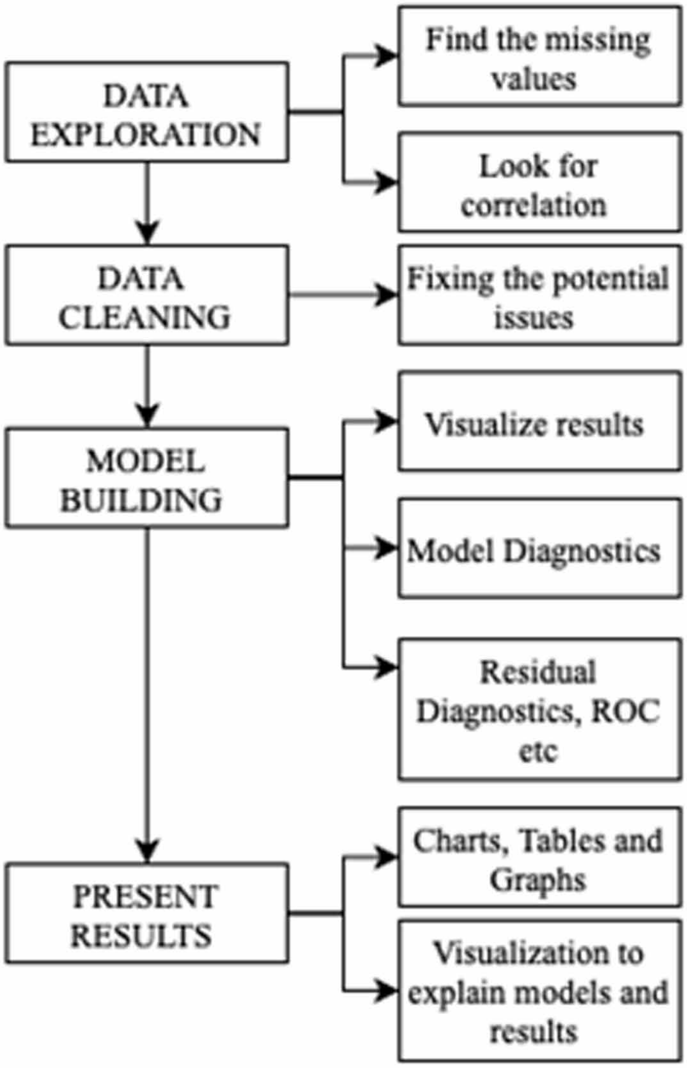


The process of EDA is shown in Figure 2, which briefs the steps and procedures present in EDA. The process of EDA begins with data exploration, which happens by finding out the missing values and looking for the correlation between the available values. The datasets having complete values helps in correlation of the data and this is achieved by cleaning and preprocessing it in the second step. The third step in the process is to visualize the dataset and run the model diagnostics and present the analyzed values. Finally with the help of models, visualized results analysis would be completed (DSouza, J., & Velan, S. S., 2020).

There are many stages for EDA and they are as follows:

- Description of the Information: Various types of information and different insights of the information are to be known before proceeding with the analysis. The description begins with the depict () function in Python. In Pandas, depict () is applied on a DataFrame which helps in creating error-free insights that sum up the focal inclination, scattering, and state of a dataset's circulation, excluding Not a Number (NaN) esteems. The lower percentile is considered a value of 25 and the upper percentile is considered as 75. The 50 percentile is equivalent to the middle.
- Dealing with Missing Information: Information in reality is infrequently spotless and homogeneous. Information can be absent during information extraction. Missing values should be taken care of cautiously on the grounds that they do not diminish the nature of the analysis being carried out. It can likewise prompt wrong expectation or order and can likewise cause a high inclination for some random model being utilized. There are a few alternatives for dealing with missing qualities. The following are few strategies adopted to address missing values:
    - Drop NULL or Missing Values: This is the quickest and simplest method to deal with missing values. This technique decreases the nature of the considered model as it lessens test size since it works by erasing any remaining perceptions where any of the values are absent.
    - Fill Missing Values: This is the most widely recognized technique for taking care of missing values. The missing values are replaced with a statistical measure like mean, median or mode of the specific attribute.
    - Predict Missing values with a ML Algorithm: This is done using outstanding and most proficient techniques for taking care of missing information. Contingent upon the class of information that is missing, one can either utilize a relapse or order model to predict missing information.

Figure 2. Processes in EDA



- Dealing with Anomalies: An anomaly is something which is independent or not the same as the group of data. Anomalies can be an aftereffect of an error during information assortment or it very well may be only a sign of fluctuation in the information. A portion of the strategies for recognizing and dealing with anomalies are as follows:
  - BoxPlot: It is a standardized method of depicting data distribution using a five-number summary: minimum value, first quartile (Q1), median, third quartile (Q3), and maximum value. It can reveal the values of outliers. It can also determine whether or not the data is symmetrical, how tightly the data is clustered, and whether or not the data is skewed.
  - Scatterplot: Dots are used to represent values for two different numeric variables in a scatter plot. The values for each data point are indicated by the position of each dot on the horizontal

and vertical axes. Scatter plots are used to see how variables relate to one another. The data points that are a long way from the populace are identified as outliers.

   - Z-score: The Z-score is the marked number of standard deviations by which the worth of a perception or information point is over the mean worth of what is being noticed or estimated. While ascertaining the Z-score the focus of the information is rescaled and searched for information that are excessively far from nothing. These information focuses which are excessively far from zero are treated as the anomalies. In the vast majority of the cases an edge of 3 or - 3 is utilized. If the Z-score value is more noteworthy than or under 3 or - 3 individually, that information point will be distinguished as anomalies.
   - IQR: The Inter Quartile Range (IQR) is a proportion of measurable scattering, being equivalent to the distinction somewhere in the range of 75th and 25th percentiles, or among upper and lower quartiles.
- Data Visualization: The approaches used to express data or information by encoding it as image representation using points, lines, or bars, are referred to as data visualization. Histograms, Bar graphs, Line charts, Pie charts, HeatMaps and so on are used for data visualization.

Relationship between Data Science and EDA

The principal motivation behind EDA is to help check information prior to making any presumptions. It can assist with recognizing clear errors, comprehend designs inside the information, identify anomalies and discover intriguing relations among the factors (Shorten, C. et al., 2021). Data scientists can utilize exploratory examination to guarantee the outcomes they produce for legitimacy and materialize it to any ideal business results and objectives. When EDA is completed and experiences are drawn, its provisions would then be able to be utilized for more refined information investigation or displaying, including AI.

Tools Required for EDA

- Python: Python is a high-level programming language with dynamic semantics that is interpreted and object-oriented. Its high-level built-in data structures, together with dynamic typing and dynamic binding, make it ideal for Rapid Application Development as well as use as a scripting or glue language to link together existing components. Python's basic, easy-to-learn syntax prioritizes readability, lowering software maintenance costs. Python facilitates program flexibility and code reuse by supporting modules and packages.
- R: An open-source programming language and free programming environment for measurable registering and designs upheld by the R Foundation for Statistical Computing. The R language is generally utilized among analysts in information science in creating factual perceptions and information investigation.

Types of EDA

There are basically four types of EDA that are summarized in Table 1.

Table 1. Types of EDA

Sl. No.	Type	Definition	Available Graph Form
1	Univariate non-graphical	This is the easiest type of information examination, where the information being examined comprises only one variable. Since it's a single variable, it doesn't understand the relationships with other variables in the dataset. The primary reason for univariate analysis is to portray the information and discover designs that exist inside it.	NA
2	Univariate graphical	It gives summary statistics for each field in the raw data set (or a single variable summary).	Stem-and-Leaf plots, which show all information points and the state of the appropriation. Histograms, a bar plot in which each bar addresses the recurrence (count) or extent of cases for a scope of values. Box plots, which graphically portray the five-number analysis: outline of least value, first quartile, middle, third quartile, and most extreme value.
3	Multivariate nongraphical	Multivariate data emerges from more than one variable. Multivariate non-graphical EDA strategies show the relationship between at least two attributes of the dataset through cross-classification or insights.	NA
4	Multivariate graphical	Multivariate information utilizes illustrations to show relationships between at least two attributes of information.	Scatter plot, which is utilized to plot information, focuses on an upward pivot to show the amount one variable is influenced by another. Multivariate chart, which is a graphical portrayal of the connections among factors and a reaction. Run chart, which is a line diagram of information plotted. Bubble chart, which is an information representation that shows numerous circles (rises) in a two-dimensional plot. Heat Map, which is a graphical portrayal of information where data points are portrayed by shading.

EDA Graphical Representation of COVID-19 in INDIA

For this study, three datasets are extracted from Kaggle.com. Table 2, Table 3 and Table 4 shows the retrieved dataset and its associated data files, along with their attribute descriptions.

Table 2. Description of attributes in CoV dataset

Sl. No.	Attribute	Description
1.	Date	Date of the observation in DD-MM-YYYY format
2.	Time	Time of the observation in HH:MM format
3.	State / Union Territory	India State or Union Territory
4.	Confirmed	Number of confirmed cases
5.	Cured	Number of cured cases
6.	Deaths	Number of death cases
7.	Latitude	Latitude value
8.	Longitude	Longitude value

Table 3. Description of attributes in statewise testing details dataset

Sl. No.	Attribute	Description
1.	Date	Date of the observation in DD-MM-YYYY format
2.	State / Union Territory	India State or Union Territory
3.	Total Samples	Total number of CoV samples collected
4.	Negative	Total number of negative samples
5.	Positive	Total number of positive samples

Table 4. Description of statewise vaccination details

Sl. No.	Attribute	Description
1.	Date	Date of the observation in DD-MM-YYYY format
2.	State / Union Territory	India State or Union Territory
3.	Total Doses Administered	Total number of vaccination doses administered
4.	Total Sessions Conducted	Total number of sessions conducted
5.	Total Sites	Total number of sites
6.	First Dose Administered	Total number of first dose administered
7.	Second Dose Administered	Total number of second dose administered
8.	Male (Individuals Vaccinated)	Total number of male vaccinated details
9.	Female (Individuals Vaccinated)	Total number of female vaccinated details
10.	Transgender (Individuals Vaccinated)	Total number of transgender vaccinated details
11.	Total Covaxin Administered	Total number of Covaxin Administered
12.	Total CoviShield Administered	Total number of CoviShield Administered
13.	Total Sputnik V Administered	Total number of Sputnik V Administered

Sl. No.	Attribute	Description
14.	Adverse event following immunization (AEFI)	Total number of AEFI Administered
15.	18-45 years (Age)	Vaccination details of people between 18-45 years of age
16.	45-60 years (Age)	Vaccination details of people between 45-60 years of age
17.	60+ years (Age)	Vaccination details of people above 60 years of age
18.	Total Individuals Vaccinated	Total number of individuals vaccinated

The Table 5 gives the state-wise statistical data of Indian people affected with CoV. The data is captured between the period of January 2020 to July 2021. The Recovery Rate is calculated as shown in equation (1) and the Mortality Rate is calculated as shown in equation (2).

$$Recovery\_Rate = \left(\frac{Cured}{Confirmed} * 100 \right). \tag{1}$$

$$Mortality\_Rate = \left(\frac{Deaths}{Confirmed} * 100 \right). \tag{2}$$

During the EDA analysis, it is observed that Maharashtra reported the highest number of confirmed cases (6113335) and death cases (123531) and the second highest being Kerala. Telangana and Mizoram have the slowest recovery rate of 81.6% and 82.97% respectively and Punjab state has the highest mortality rate of 2.7% followed by Uttarakhand state of 2.15%. The top five states with active number of coronavirus cases are Maharashtra, Karnataka, Kerala, Tamilnadu and Uttar Pradesh and the lowest cases are reported by Daman & Diu, Dadra & Nagar Haveli, Andaman & Nicobar Islands, Lakshadweep and Arunachal Pradesh. Maharashtra again recorded the highest number of deaths followed by Karnataka, Tamilnadu, Delhi and Uttar Pradesh.

Table 5. Statistical details of CoV in India

State/ Union Territory	Confirmed	Cured	Deaths	Recovery Rate (%)	Mortality Rate (%)
Maharashtra	6113335	5872268	123531	96.056702	2.020681
Kerala	2996094	2877557	13960	96.043615	0.46594
Karnataka	2859595	2784030	35526	97.357493	1.242344
Tamil Nadu	2503481	2435872	33132	97.2994	1.323437
Andhra Pradesh	1908065	1861937	12898	97.582472	0.675973
Uttar Pradesh	1706818	1682130	22656	98.553566	1.327382
West Bengal	1507241	1472132	17834	97.670645	1.183222
Delhi	1434687	1408853	25001	98.199328	1.74261
Chhattisgarh	996359	977893	13462	98.146652	1.351119
Rajasthan	952836	942882	8942	98.955329	0.938462
Odisha	927186	897362	4299	96.783385	0.463661
Gujarat	823964	811699	10072	98.511464	1.222384
Madhya Pradesh	790042	780578	9017	98.802089	1.141332
Haryana	769030	758442	9506	98.623201	1.236103
Bihar	722746	711913	9612	98.501133	1.329928
Bihar	715730	701234	9452	97.974655	1.32061
Telangana	628282	613124	3703	97.587389	0.589385
Punjab	596736	578590	16131	96.959124	2.703205
Assam	522267	493306	4717	94.454752	0.903178
Jharkhand	346038	340365	5118	98.360585	1.479028
Uttarakhand	340882	332006	7338	97.396166	2.152651
Jammu and Kashmir	317481	309554	4345	97.503158	1.368586
Himachal Pradesh	202945	198134	3485	97.629407	1.717214
Goa	167823	162787	3079	96.999219	1.834671
Puducherry	118227	114673	1763	96.993918	1.491199
Manipur	73581	66132	1218	89.876463	1.655319
Tripura	68612	63964	701	93.225675	1.021687
Chandigarh	61752	60837	809	98.518267	1.310079
Meghalaya	52358	47173	880	90.097024	1.680736
Arunachal Pradesh	37879	34525	181	91.14549	0.477837
Nagaland	25619	23982	503	93.610211	1.963387
Mizoram	22155	18383	98	82.974498	0.442338
Sikkim	21403	19200	309	89.70705	1.443723
Ladakh	20137	19733	204	97.993743	1.013061
Dadra and Nagar Haveli and Daman and Diu	10575	10532	4	99.593381	0.037825
Lakshadweep	9947	9643	49	96.943802	0.492611
Andaman and Nicobar Islands	7487	7343	128	98.076666	1.70963

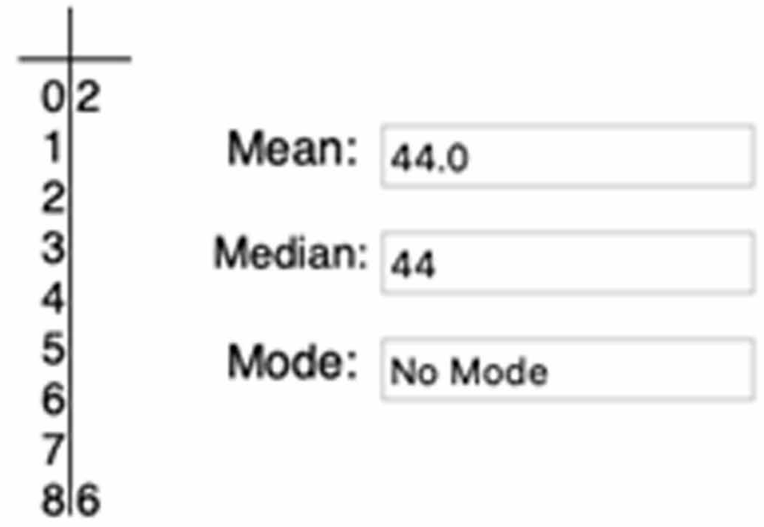
The basic CoV statistics observed while EDA is as follows:

- Total number of States with Disease Spread: 42
- Total number of Confirmed Cases: 30663665
- Total number of Recovered Cases: 29799534
- Total number of Deaths Cases: 404211
- Total number of Active Cases: 459920
- Total number of Closed Cases: 30203745
- Approximate number of Confirmed Cases per Day: 58407.0
- Approximate number of Recovered Cases per Day: 56761.0
- Approximate number of Death Cases per Day: 770.0
- Approximate number of Confirmed Cases per hour: 2434.0
- Approximate number of Recovered Cases per hour: 2365.0
- Approximate number of Death Cases per hour: 32.0
- Number of Confirmed Cases in last 24 hours: 43733
- Number of Recovered Cases in last 24 hours: 47240
- Number of Death Cases in last 24 hours: 930

The univariate graphical representation is done using Stem and Leaf, BoxPlot and Bar Graph methods. Figure 3 is the Stem and Leaf representation of COVID-19 data of India as per Table 2 data. The Stem and Leaf representation are done only for the confirmed cases in India. The representation is shown in Figure 3.

The Figure 3 also has Mean, Median and Mode calculated for the considered dataset. The Figure 4 represents the BoxPlot representation of COVID-19 data as per Table 2. The BoxPlot representation is done only for the confirmed and cured cases in India. The Median, Quartile and Outlier calculation is as shown in Table 6.

Figure 3. Stem and leaf representation of CoV dataset.



0	2
1	
2	
3	
4	
5	
6	
7	
8	6

Mean: 44.0

Median: 44

Mode: No Mode

The Statewise Testing details are depicted in Figure 5. Uttar Pradesh has recorded the highest total samples of 59.33166 million, the second being Maharashtra with 42.90829M and the last being Diu and Daman with total samples of 7241000.

Figure 4. BoxPlot representation of confirmed and cured cases

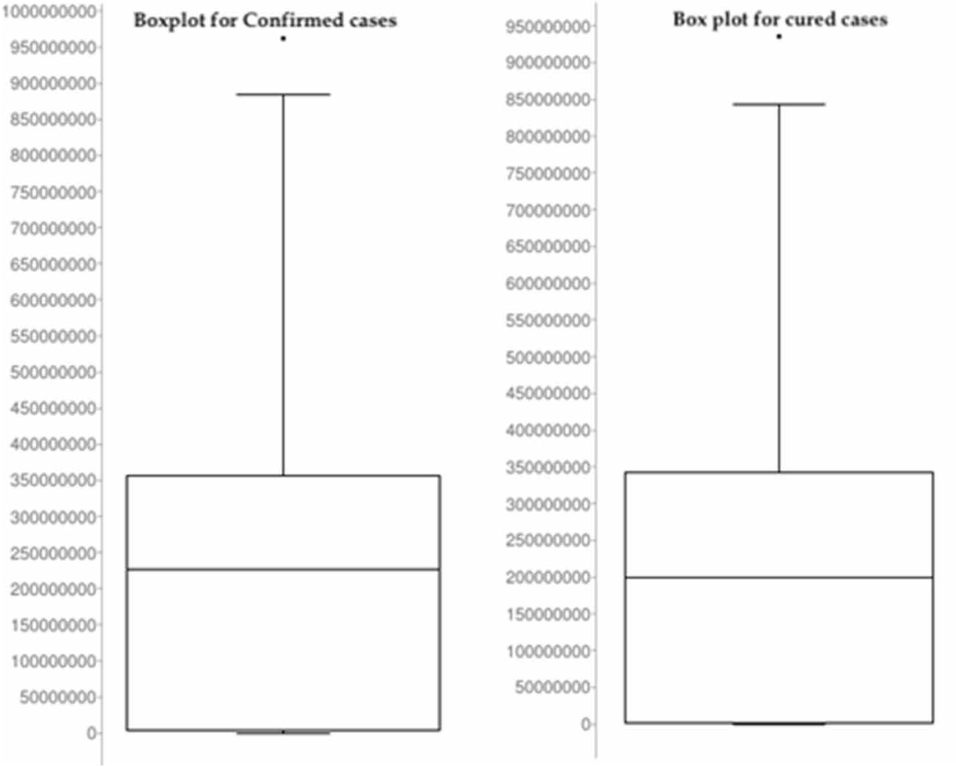


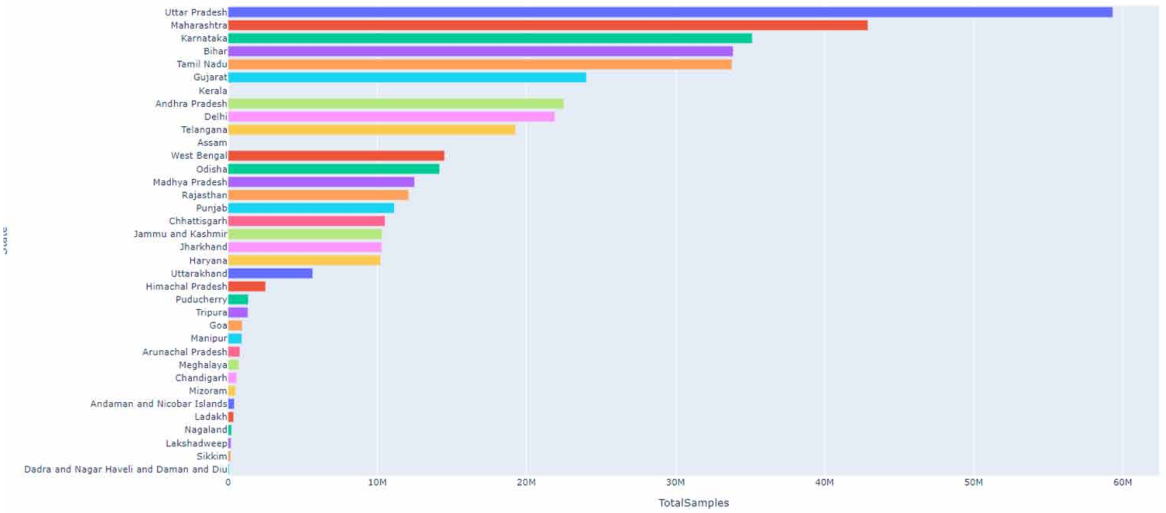
Table 6. Median and quartile calculations for COVID -19 Indian data

Sl. No.	Confirmed Cases	Cured Cases
1.	Median: 226770312	Median: 198824412
2.	Minimum: 2	Minimum: 0
3.	Maximum: 961636364	Maximum: 935289657
4.	First quartile: 2938234	First quartile: 1133341
5.	Third quartile: 356305616	Third quartile: 342616397
6.	Interquartile Range: 353367382	Interquartile Range: 341483056
7.	Outlier: 961636364	Outlier: 935289657

Prime Minister Narendra Modi announced a national lockdown on 24 March 2020 and the details are as follows:

- No lockdown= 2020-01-30 to 2020-03-24
- lockdown 1= 2020-03-24 to 2020-07-15
- Lockdown_2= 2020-07-15 to 2020-11-04
- Lockdown_3= 2020-11-04 to 2021-02-19
- Lockdown_4= 2021-02-19 to 2021-05-31
- Unlock_1= 2020-06-01 to 2020-06-30
- Unlock_2= 2020-07-01 to present

Figure 5. Statewise testing details



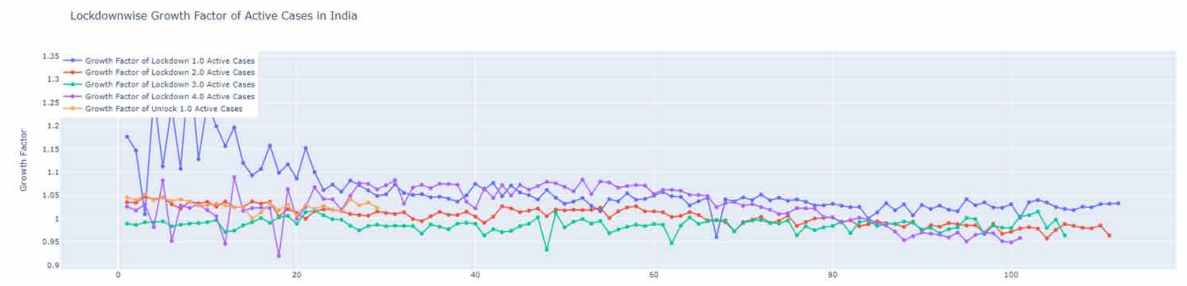
The growth rate during lockdown and unlock period is as follows:

- Average Active Cases growth rate in Lockdown 1.0: 1.06
- Median Active Cases growth rate in Lockdown 1.0: 1.04
- Average Active Cases growth rate in Lockdown 2.0: 1.00
- Median Active Cases growth rate in Lockdown 2.0: 1.01
- Average Active Cases growth rate in Lockdown 3.0: 0.99
- Median Active Cases growth rate in Lockdown 3.0: 0.99
- Average Active Cases growth rate in Lockdown 4.0: 1.03
- Median Active Cases growth rate in Lockdown 4.0: 1.03
- Average Active Cases growth rate in Unlock 1.0: 1.03
- Median Active Cases growth rate in Unlock 1.0: 1.03

Figure 6 gives the graphical representation of lockdown wise growth factor of active cases in India. Active cases are calculated as shown in equation (3).

$$Active\_Cases = Confirmed - Cured - Deaths. \qquad (3)$$

Figure 6. Lockdown wise growth factor of active cases in India



The vaccination details too are analyzed. The male to female vaccinations are in a ratio of 53.2% and 46.8% respectively. The top five vaccinated states are Maharashtra, Uttar Pradesh, Rajasthan, Gujarat and West Bengal. The results are shown in Figure 7. Figure 8 gives the statistics of Covaxin and Cov-iShield Vaccinated Details and Figure 9 gives the top five vaccinated states of India. Figure 10 depicts the variations of COVID-19 positive test results (percentage) from April to August 2020.

Since the COVID-19 epidemic began in early 2020, governments all over the world have taken various methods to deal with it. Countries established a variety of laws and restrictions to stop the virus from spreading, decrease the outbreak's effects, and provide effective control measures. Despite the fact that the pandemic has been ongoing for over a year, few researchers have looked into its long-term effects.

Figure 7. Male and female vaccinated ratio

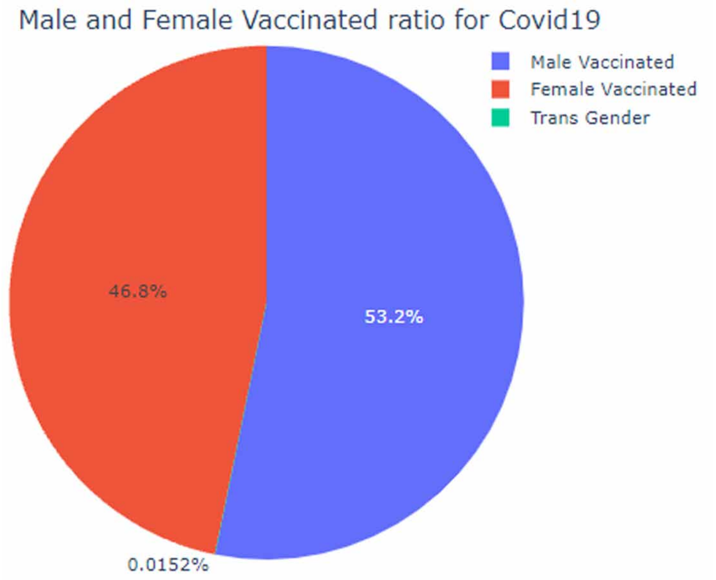


Figure 8. Covaxin and CoviShield vaccinated details

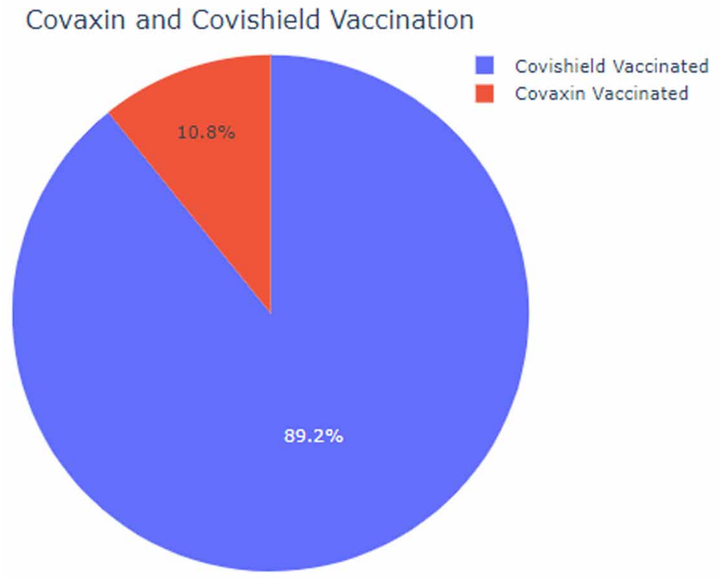


Figure 9. Topmost five vaccinated states of India

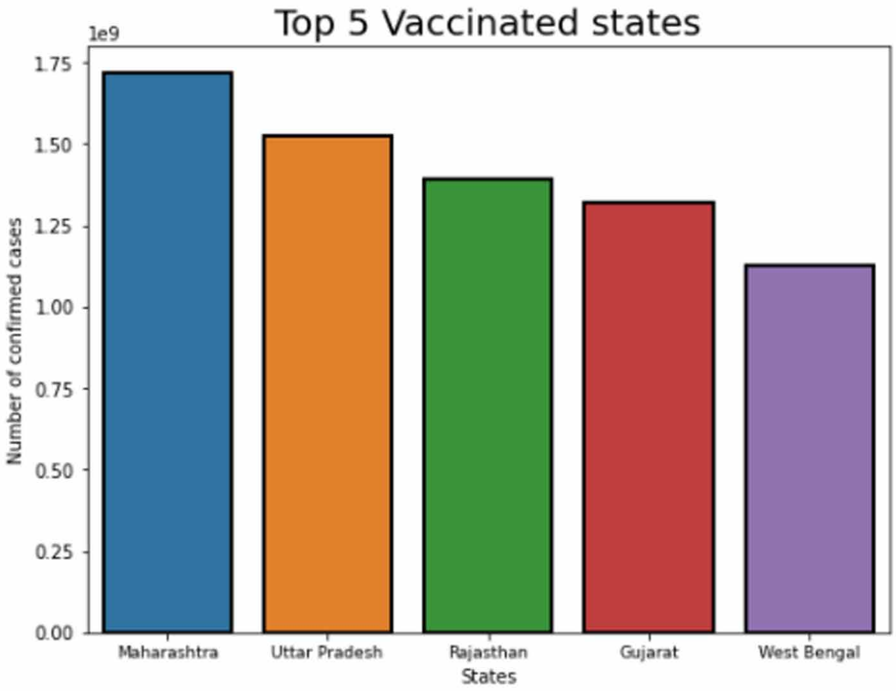
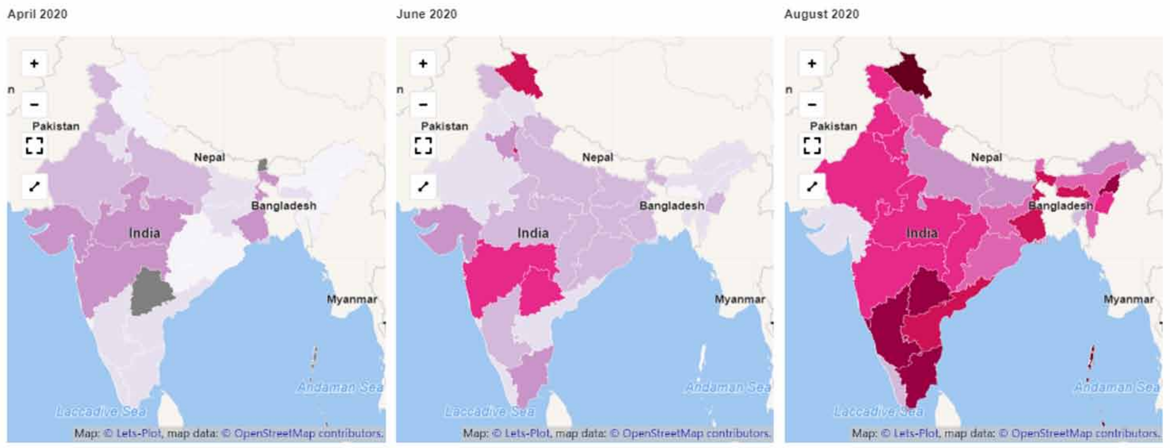


Figure 10. COVID-19 Positive Test Results (Percentage) from April to August 2020



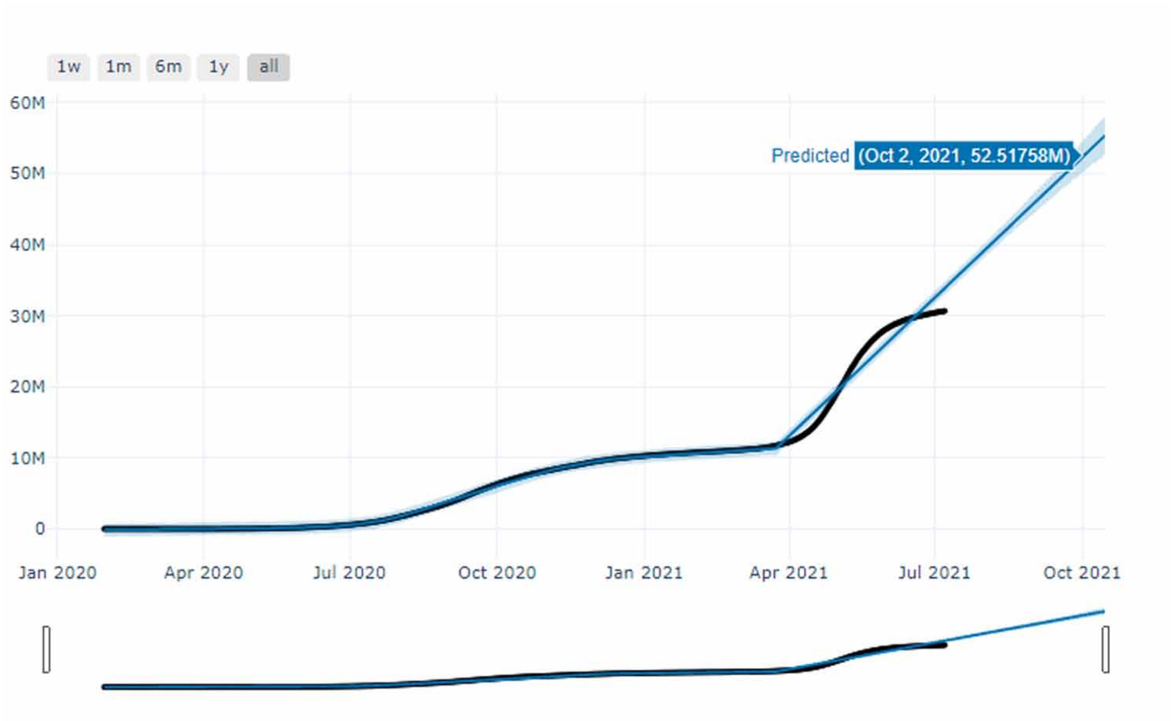
The current COVID-19 outbreak has prompted an EDA using Python on the datasets obtained, with the goal of studying the spread and trend of COVID-19 in different Indian states. The EDA dataset goes through normalization, filtering to select critical columns, deriving new columns, and presenting the data in a graphical way. To process and extract information from the given dataset, this study used the Python data processing tool and Pandas packages. For better visualization, appropriate graphs are constructed, and the Python tools Matplotlib and Seaborn are used for the same.

COVID-19 Prediction Analysis

After the effective visualization of the statistical data through EDA, the second step is to present the predictions of the data that has been read. The predictions on the data are done with the help of ML techniques (Shinde, G. R. et al., 2020). Some of the important prediction algorithms that are used in this study are AutoRegressive Integrated Moving Average (ARIMA), Seasonal Auto Regressive Integrated Moving Average (SARIMA) (ArunKumar, K. E. et al., 2021), FBProphet, Polynomial Regression, Linear Regression, Support Vector Regression, AutoRegression, Moving Average, Holt's Linear and Holt's Winter models. These mentioned techniques are discussed in detail in the following subsections with the example of the COVID-19 India data.

1. *FBProphet:* The FBProphet library, which is created by Facebook and is primarily used for time series forecasting, is used in the prediction analysis in this study (Darapaneni, N. et al., 2020). FBProphet is a time series data forecasting process based on an additive model in which non-linear trends are fit with yearly, weekly, and daily seasonality, as well as holiday impacts. It works best with time series with substantial seasonal influences and historical data from multiple seasons. Using FBProphet, the prediction is done for 60 days (till 10th October 2021). As per 2nd October 2021, the actual confirmed cases observed is 6312584. And the prediction accuracy by FBProphet is 83.19%. The prediction analysis is shown in Figure 11.

Figure 11. COVID-19 prediction analysis using FBProphet model



2. *Linear Regression:* On the basis of independent variables, regression models are statistical sets of processes that are used to estimate or forecast the target or dependent variable. There are many different types of regression models, including Linear Regression, Ridge Regression, Stepwise Regression, and Polynomial Regression. A simple model for determining the relationship between a dependent and an independent variable is Linear Regression (Darapaneni, N. et al., 2020). The association between a dependent (COVID-19 Confirmed Cases) and independent variables is shown in equation 4. In the Linear Regression model, each univariate analysis is utilized to indicate how much each independent variable will be predicted by the dependent variable.

$$C = \beta 0 + \beta 1\mathrm{x}_{1\ +}\ldots + \beta \mathrm{nxn} + \epsilon \tag{4}$$

where C is the total number of Confirmed COVID-19 cases in India, x1,$_{\mathrm{x}}$2,$_{\ldots}$,xn $_{a}$re n independent variables, β *i*s the intercept and coefficients and ϵ is the error of the Linear Regression Model. Figure 12 gives the prediction analysis of the CoV dataset using Linear Regression Model.

The Linear Regression Model appears to be collapsing in Figure 12. As can be seen, the trend of Confirmed Cases is far from linear. In order to overcome this drawback, Polynomial Regression is used in this study. The prediction results are depicted in Figure 13.

Figure 12. COVID-19 prediction analysis using linear regression model

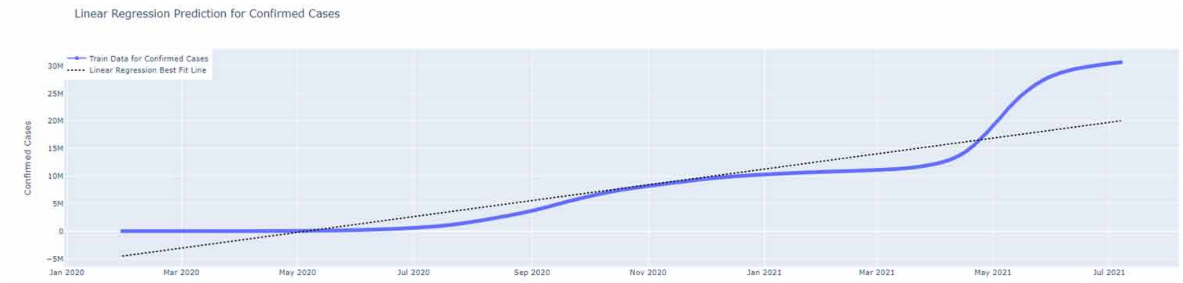
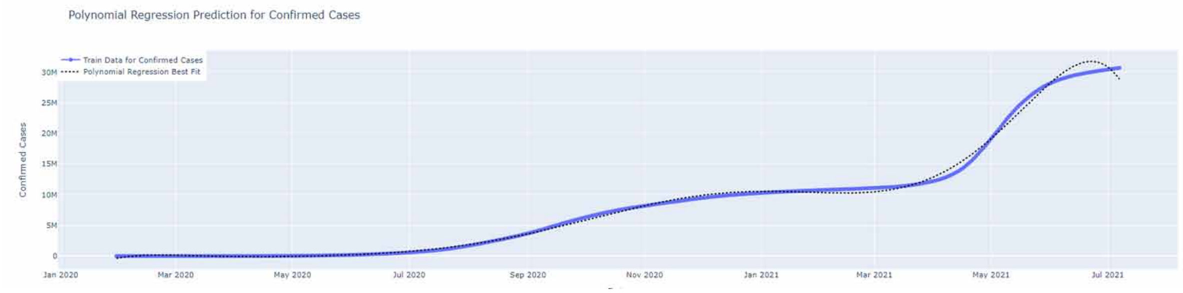


Figure 13. COVID-19 prediction analysis using polynomial regression model



3. *Support Vector Regression:* For both linear and nonlinear regression types, Support Vector Regression (SVR) is a common choice for prediction and curve fitting. SVR is built on Support Vector Machine (SVM) elements, where support vectors are essentially closer points towards the created hyperplane in an n-dimensional feature space that distinguishes the data points regarding the hyperplane (Saba, T. et al., 2021). Because the cost function for developing the model is unconcerned with training points outside the margin, the model built by classification of support vectors is solely dependent on a subset of training data. Similarly, because the cost function ignores samples whose prediction is close to their objective, the model built by Support Vector Regression is based entirely on a subset of training data. The equation for the hyperplane is given in equation (5). CoV prediction analysis using SVR is shown in Figure 14.

$$y = \beta \mathrm{x} + \varepsilon \tag{5}$$

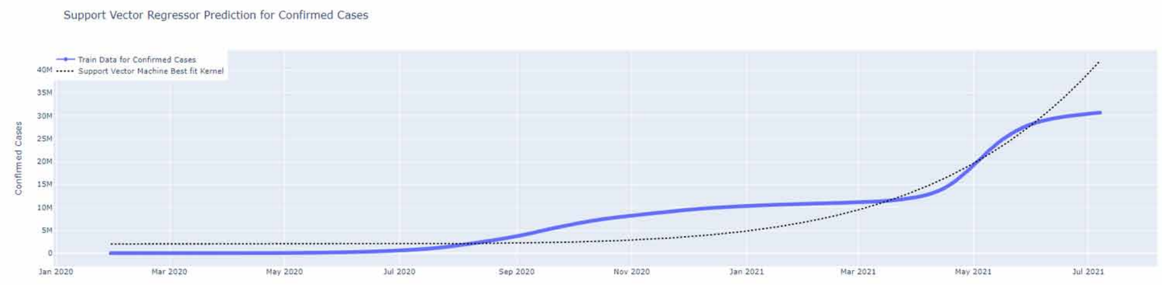
where y is the total number of Confirmed Cases, x is the independent variable, β is the intercept and ε is the error term. The decision boundary equations are shown in the equation (6):

$$\beta \mathrm{x} + \varepsilon = +a \textit{ and } \beta \mathrm{x} + \varepsilon = \text{-a}. \tag{6}$$

As a result, equation (7) shows the hyperplane equation that should satisfy SVR.

$$-a < y - \beta x - \varepsilon < +a. \tag{7}$$

Figure 14. COVID-19 prediction analysis using support vector regression model



It is observed from Figure 14 that the SVR model isn't producing good results, since the forecasts are either overshooting or falling short of expectations.

4. Ho*lt's Linear Model: S*tatistical and structural models are the two types of forecasting models. The functional link between future and actual values of the time series, as well as external influences, is set analytically in statistical models. The following are the different types of statistical models – Regression, AutoRegressive and Exponential Smoothing models. Holt's Linear Model is an exponential smoothing method for smoothing time series in which the computational operation comprises the processing of all prior observations while taking into account the aging of data as it approaches the forecast period (Maurya, S. & Singh, S., 2020). The exponential smoothing method allows for the estimation of trend parameters that characterize the trend that has formed since the last observation, rather than the average level of the process. Three equations make Holt's model. The first is the equation for data smoothing. The trend smoothing equation is the second, and the forecast equation for the period t = k is the last. Equations (8), (9), and (10) are the formulae, respectively.

$$\mathrm{a}t = \alpha \mathrm{yt}\ _{+}\ (1 - \alpha)\left(a_{t-1} + b_{t-1}\right) \tag{8}$$

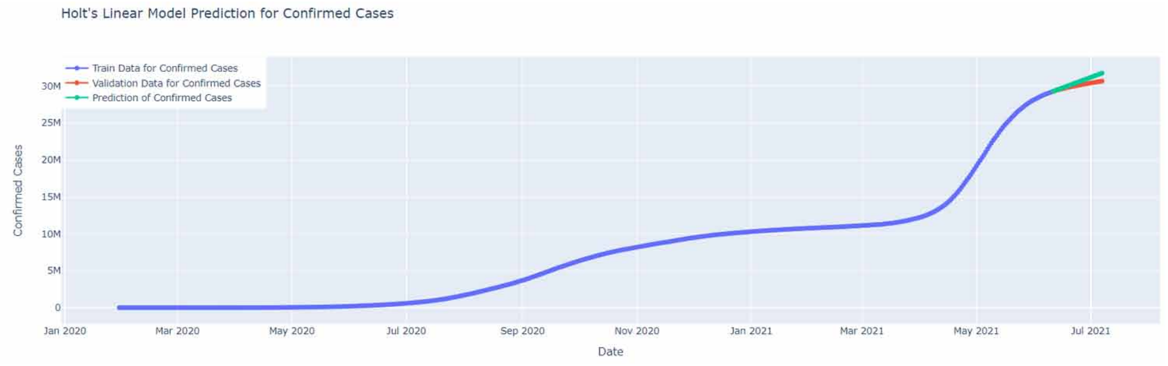
$$b_t = \beta\left(a_t - a_{t-1}\right) + \left(1 - \beta\right) b_{t-1} \tag{9}$$

$$y_{t+k} = a_t + b_t k \tag{10}$$

where a_t is the smoothed value of the anticipated indicator for period *t, b_t* is the growth trend estimate, α is the smoothing parameter ($0 \le \alpha \le 1$), β is the parameter used for smoothing ($0 \le \beta \le 0$), and k is the number of time periods for which the forecast is produced.

The smoothing parameters α and β are chosen subjectively by the forecaster based on previous forecasting experience or by minimizing forecast error. When the smoothing parameters are large, which tend to zero, the model's response to changes in the data is stronger, and the structure of the smoothed values is less even. When the smoothing parameters are small, which tend to zero, the model's response to changes in the data is weaker, and the structure of the smoothed values is less even. Figure 15 demonstrates CoV prediction analysis using Holt's Linear Model.

Figure 15. COVID-19 prediction analysis using Holt's Linear Model

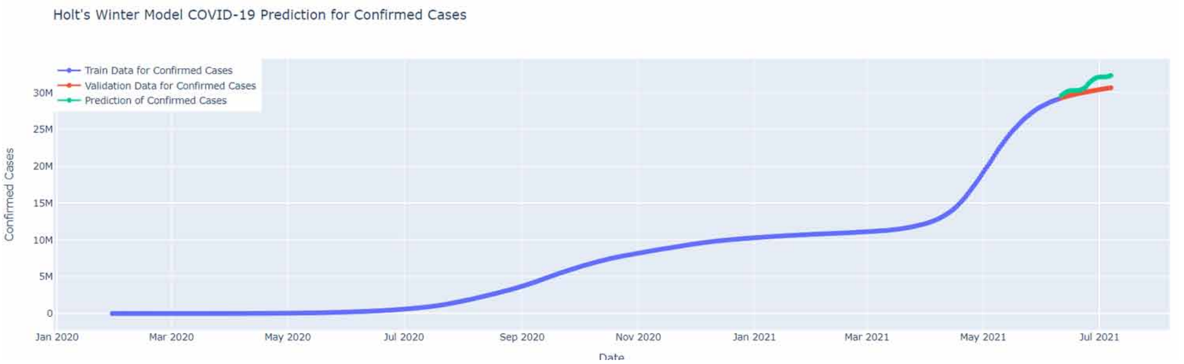


5. *Holt's Winter Model:* The Holt-Winters technique is a popular time series forecasting approach that can account for both trend and seasonality. The Holt-Winters approach is made up of three other smoothing methods (Maurya, S. & Singh, S., 2020).

   ◦ Simple Exponential Smoothing (SES) presupposes that the level of the time series remains constant. As a result, it can't be utilized with series that have both trend and seasonality.
   ◦ Holt's Exponential Smoothing (HES): HES is a step up from simple exponential smoothing because it includes a trend component in the time series data.
   ◦ Winter's Exponential Smoothing (WES): WES is a Holt's exponential smoothing extension that finally incorporates seasonality. The Holt-Winters method is the name given to Winter's exponential smoothing.

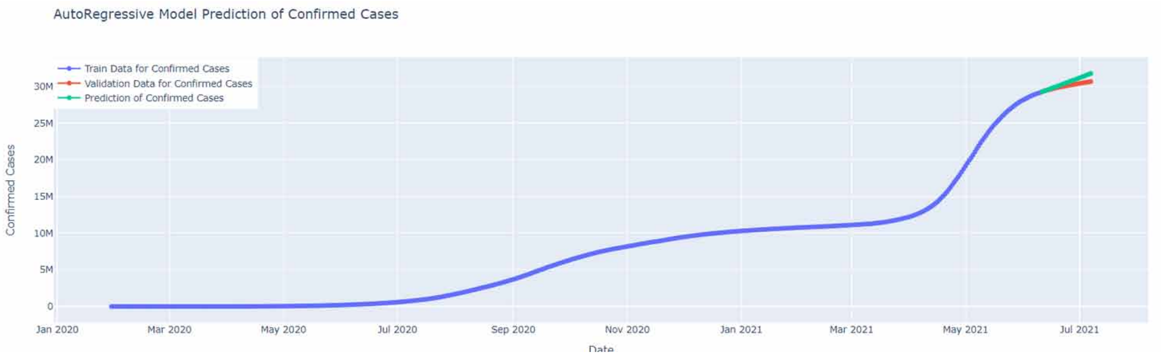
As a result, the Holt-Winters approach is frequently referred to as triple exponential smoothing, because it is essentially a combination of three smoothing methods stacked on top of one another. Figure 16 demonstrates CoV prediction analysis using Holt's Winter Model.

Figure 16. COVID-19 prediction analysis using Holt's Winter Model



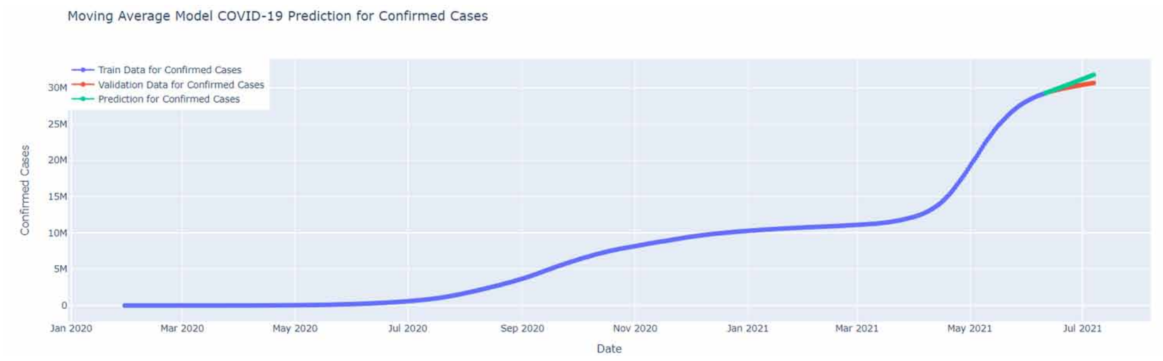
6. *Auto Regressive Model:* An autoregression is a time series model that predicts the value at the next time step by using observations from past time steps as input to a regression equation. It's a simple concept that can produce reliable forecasts for a variety of time series issues (ArunKumar, K. E. et al., 2021). The prediction analysis using this model is shown in Figure 17.

Figure 17. COVID-19 prediction analysis using AutoRegressive Model



7. *Moving Average Model:* A moving average is a method of calculating and analyzing data in statistics and economics by providing a series of averages of various subsets of the dataset. A Simple Moving Average (SMA) is defined as the unweighted mean of preceding data or an equal number of data points on either side of a center value (in science or engineering). SMA too is used for prediction analysis in this chapter. The prediction results are shown in Figure 18.

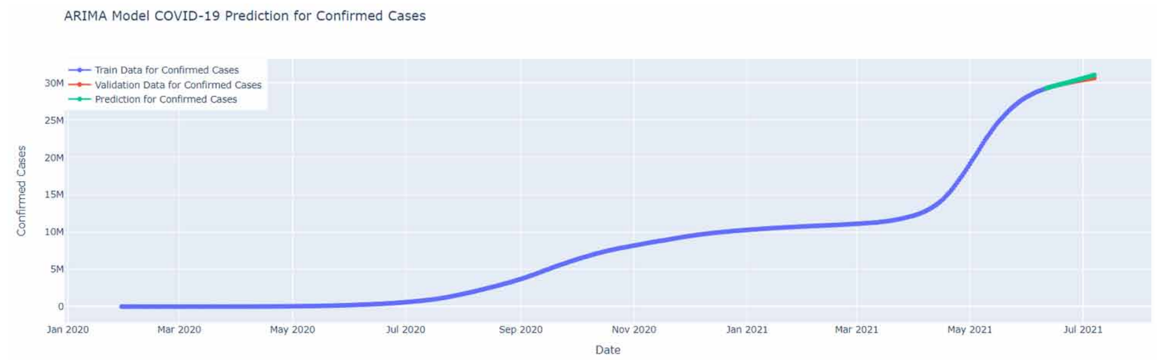
Figure 18. COVID-19 prediction analysis using Moving Average Model



8. *AutoRegressive Integrated Moving Average (ARIMA Model):* Data on a given occurrence is collected via time series forecasting, and a model is developed to depict the underlying link between the variables. The model is then used to extrapolate time series data to forecast future event values. This approach is useful for forecasting future behaviour when there isn't a meaningful link between the two factors. The most often used time series model is the Autoregressive Integrated Moving Average (ARIMA) model (ArunKumar, K. E. et al., 2021).

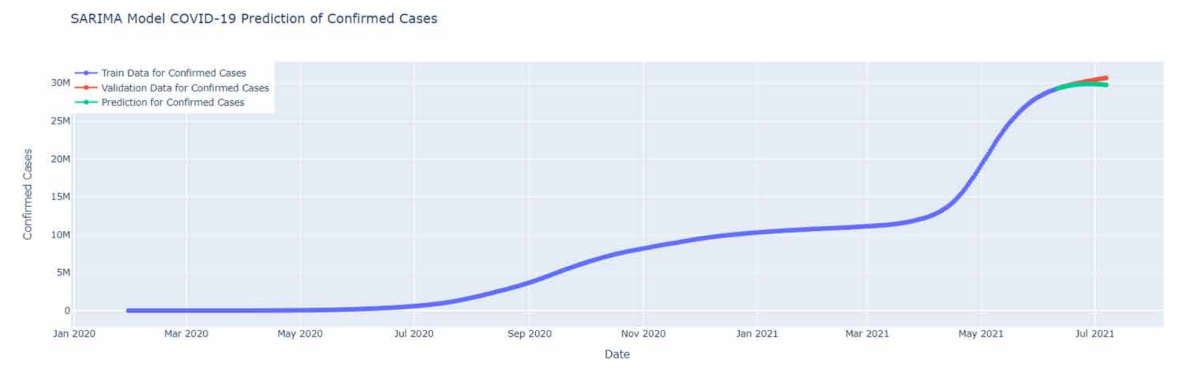
ARIMA captures extremely complicated relationships as it includes error factors and delayed data. These models are created by regressing a variable against its previous values (ArunKumar, K. E. et al., 2021). The ARIMA model is based on the notion that previous time points in a series can impact present and future time points. ARIMA model's CoV prediction analysis is shown in Figure 19.

Figure 19. COVID-19 prediction analysis using ARIMA Model



9. *Seasonal AutoRegressive Integrated Moving Average (SARIMA Model):* This model is distinct from an ARIMA model in that it is based on seasonal patterns rather than time. Seasonal impacts are widespread and may be extremely large in many time series data sets. Figure 20 demonstrates SARIMA model's prediction analysis for COVID-19 dataset (ArunKumar, K. E. et al., 2021).

Figure 20. COVID-19 prediction analysis using SARIMA Model



Comparative Analysis of Prediction Models

Performance of all 10 models is compared using a statistical measure namely Root Mean Squared Error. Table 7 shows the comparative analysis. It is observed from Table 7 that AR performs better when compared to other models. The performance predictions of MA and SARIMA are close to each other and Linear Regression has predicted less when compared to other models.

The lack of data on COVID-19 makes modeling and prediction difficult. The data from cumulative confirmed cases in India is modeled using 10 distinct methodologies in this study. According to the findings, the Auto Regressive Model has a substantially higher success rate than ARIMA, Moving Average, and the other models included in the study.

Table 7. Comparative analysis of RMSE values of COVID-19 prediction models

Model Name	Root Mean Squared Error (RMSE)
Auto Regressive Model (AR)	176891.262332
Moving Average Model (MA)	427826.5262134
SARIMA Model	437422.329164
Holt’s Linear Model	574341.509304
Holt’s Winter Model	592496.781198
ARIMA Model	719831.008192
Facebook’s Prophet Model	719831.008192
Polynomial Regression Model	1411334.739950
Support Vector Machine Regressor Model	6990167.647466
Linear Regression Model	10624435.202866

Identification of CoV Misconceptions in Social Media Networks

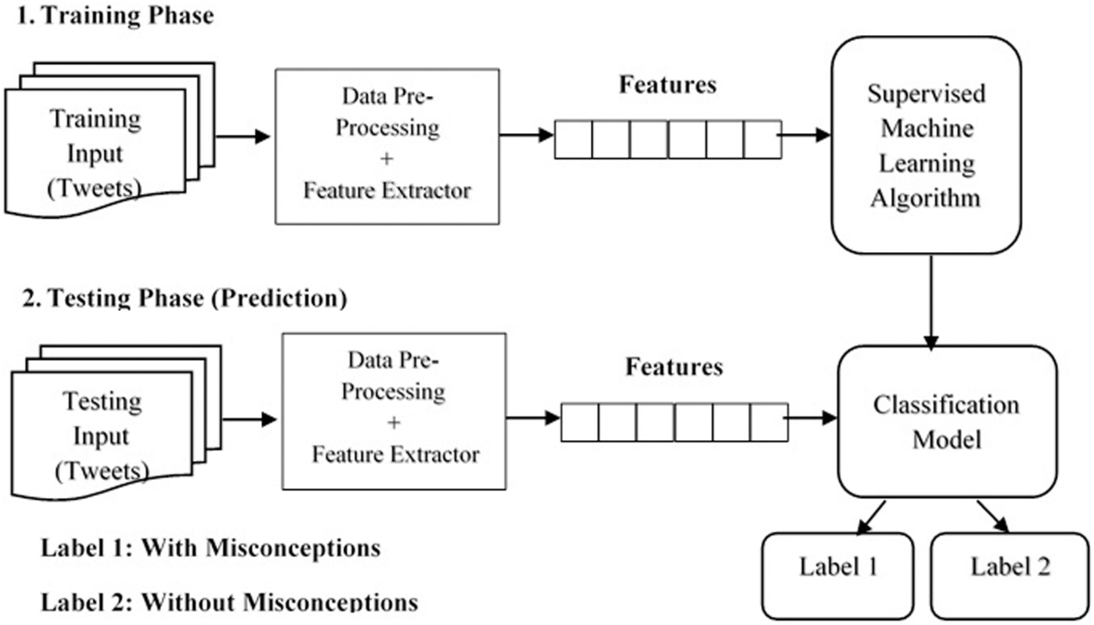
This sub-section aims primarily to recognize misconceptions on COVID-19 that are shared in Twitter®. COVID-19 is a disease attributed to a newly identified virus namely coronavirus. People have expressed their thoughts and views regarding the onset of coronavirus. Some of them are right, while others are myths. Most negative knowledge is found in myths. Relevant and reliable knowledge for this study is sought from the World Health Organization (WHO).

Introduction to Coronavirus Misconceptions: In this case study, the effectiveness of the identification systems of misconceptions has been checked on corona pandemic dataset from Twitter® messages. In the classification of the dataset in two groups, a combination of Forward Scan Trigrams and a weighted TF-IDF model is subjected to a supervised classification: one with misconceptions about the COVID-19 virus and the other containing the correct and authenticated details (Kokatnoor, S. A. & Krishnan, B., 2020).

Identification of Misconceptions Architecture: The classification of texts including unstructured text data must take place through various phases, including preprocessing, input text transformation into a vector of features, identification of meaningful patterns and final analysis of the model. The proposed architecture is as shown in Figure 21. The text corpus is divided into two sets, training and testing datasets in the ratio of 80:20 respectively. The training dataset after preprocessing and Feature Engineering process (Kokatnoor, S. A. & Krishnan, B., 2020) is converted into Vector Space Model (VSM). This VSM trains and builds a model using the standard supervised machine learning classification algorithm. The model built is tested on the testing dataset to accurately classify into binary classes, one with misconceptions and the other being labeled as correct information.

Collection of Tweets on Coronavirus: To identify the misconceptions about coronavirus outbreak, Twint Python Library is used. Using this library, 1568 tweets are collected between 1st March 2021 and 15th September 2021. In which 31 tweets are manually deleted from the file which is posted in other languages using English. Later on, the tweets are manually annotated in two labels 0 and 1 using the information from legitimate and authenticated sources like the World Health Organization (WHO), British Broadcasting Corporation (BBC) and mygov.in. A ’0’ demonstrates tweets that include misconceptions of coronavirus and ’1’ shows the authenticated details.

Figure 21. Architectural diagram for the identification CoV misconceptions


1. Training Phase
Training Input (Tweets)
Data Pre-Processing + Feature Extractor
Features
Supervised Machine Learning Algorithm
2. Testing Phase (Prediction)
Testing Input (Tweets)
Data Pre-Processing + Feature Extractor
Features
Classification Model
Label 1
Label 2
Label 1: With Misconceptions
Label 2: Without Misconceptions

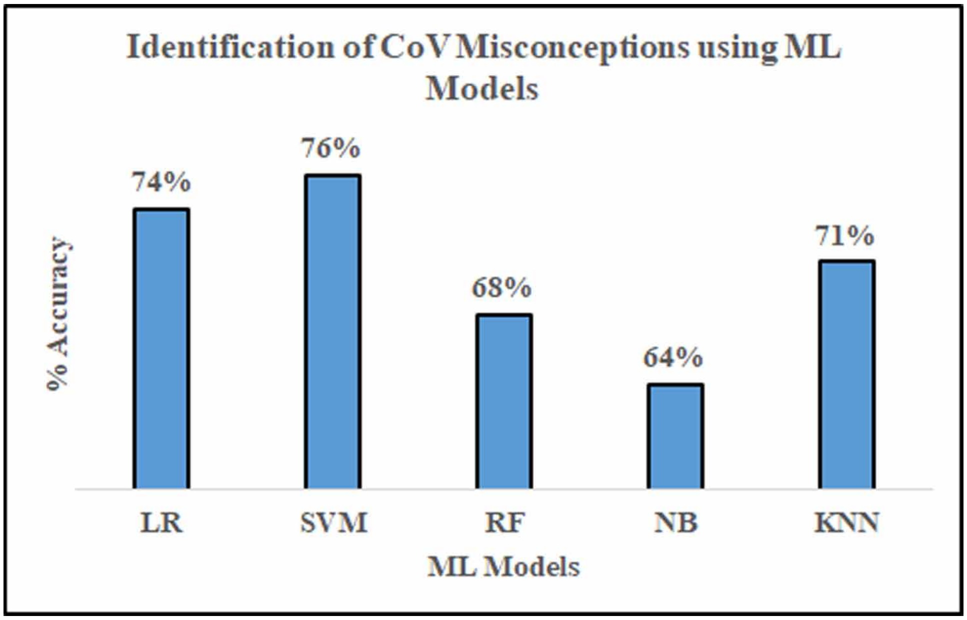
Text-Preprocessing and Feature Engineering: Extracted text corpus from Twitter® is preprocessed using NLTK 3.1 tool (Saad, S. E. & Yang, J., 2019). For creating a standard text dataset, the following procedures are used. The text is transformed to a lower case to reduce the text dataset volume. The special characters and whitespaces are removed. The stop words of the dataset with no insight into the semantic content of the document are deleted. The words that have similar semantic characteristics, but have different forms, are reduced to a generic root word. For the preprocessed text corpus, a combination of Forward Scan Trigrams and weighted TF-IDF method is applied for text vectorization (Kokatnoor, S. A. & Krishnan, B., 2020). With this an efficient VSM is created which is input to five supervised machine learning algorithms for comparative studies, namely Naïve Bayes (NB), K-Nearest Neighbour (KNN), Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF) (Varshney, D., & Vishwakarma, D. K., 2021).

Experimental Results*:* The following parameters are used for the chosen classifiers:

- The regularization parameter C is chosen as the default value 1.0 in the Support Vector Machine (SVM) standard classifier. Where C specifies how many samples contribute to the total error within the margin. With a low C, samples are penalized less within the margins than with a greater value of C. Radial Basis Function (RBF), a kernel function is chosen with SVM during the experimentation process.
- The Logistic Regression (LR) model uses L1 regularization (Lasso) or L2 (Ridge). In this study, an L2 penalty is chosen to avoid overfitting problems. Limited memory Broyden Fletcher Goldfarb Shanno (LBFGS) algorithm is used for optimizing the results. With LBFGS, the second derivative matrix updates are approximated with gradient evaluations. It saves memory only with the last few updates.
- For the K-Nearest Neighbor (KNN) classifier, K=5 is chosen. Where K is the number of neighbors. In order to find the proximity measure, p=2 is chosen where p is the power parameter for the Minkowski similarity distance calculation.
- The number of trees chosen in the Random Forest (RF) classifier is 100. Gini measure is used to find the node impurity. It measures the cumulative decrease of the node impurity over all trees of the ensemble (weighted by the likelihood of reaching this node (approximated by the proportion of samples reaching this node). The Gini impurity criterion for the two descending nodes is less than the parent node each time the split of one node is performed on a variable. Adding up the Gini decreases for each individual variable over all trees in the forest. During the experimental process the minimum number of samples required to split an internal node is 2.

SVM has yielded good results when compared to other models during the experimentation process. The output results are displayed using a statistical metric, namely Accuracy, where Accuracy is defined as the ratio between the correctly predicted model values and the total predictions number. The results are shown in Figure 22.

Figure 22. Comparison of ML models' accuracy scores



To split the dataset into two classes, SVM uses the RBF kernel function and nonlinear hyperplanes, thereby accurately classifying the dataset into anomalous and non-anomalous. This along with improvised Feature Engineering has increased the performance of SVM classifiers in terms of its accuracy. Logistic Regression divides the input by a linear boundary into two classes (anomalous and non-anomalous), one for each class. The data considered must therefore be linearly separated. But the text corpus based on COVID-19 misconceptions taken from Twitter® is imbalanced and is not linearly separable. Logistic Regression between independent variables requires moderate or no multicollinearity. This means that only one of them can be used if two independent variables have a high correlation. Repeated information in the input VSM has caused the weights parameter wrongly trained while minimizing the cost function.

Besides, the presence of data values in the text corpus that vary from the expected range has led to wrong results and hence low accuracy value, as LR is sensitive to anomalies. Based on the provided training dataset, KNN created a highly complex resolution border. Due to the initial metric vector chosen and the lack of precise discrete classes, KNN is less successful. Under the conditional independence theory of Naive-Bayes, as all the probabilities are combined, a negative value is obtained as its outcome. So, Naive Bayes too performed less in terms of its accuracy. Since RF is an ensemble model, when compared to an individual Decision Tree, it is essentially uninterpretable. This ensemble model is trained with a wide variety of decision trees, which uses more memory and increased time complexity due to which it resulted in less performance.

CONCLUSION

Throughout history, humans have been subjected to epidemics and pandemics. Often, such infectious outbreaks have resulted in entire civilizations facing extinction. Despite recent clinical advances and technology innovations, challenges such as disregarded sustainability and poor public hygiene practices, among others, have given a setting for the COVID-19 pandemic to emerge. In this context, scientific and scholarly communication utilizing a variety of open access platforms could play an important role in efforts toward preparedness and control, as well as the implementation of prompt corrective actions in the battle against epidemics and pandemics. These technologies aid in increasing understanding of scientific options for reducing infectious disease outbreaks, hence enhancing social immunity.

COVID-19 research is advancing open access and research forward at a faster rate than ever before. Preprints, for example, allow academics to disclose their study results quickly, sometimes along with their datasets. Furthermore, cross-disciplinary collaboration occurs on a regular basis, such as when a physicist is analyzing the COVID-19 dataset and suggesting a model. Data science plays a crucial role in bringing trans-disciplinary discussion between researchers and the public to address this challenge.

According to COVID-19, India's population and poor hygiene standards among the bulk of the country's people are the most concerning problems. Another issue that may come back to haunt India is a lack of medical equipment, outdated medical technology, and medical facility negligence, all of which could play a key role in this pandemic. The lack of testing and the unavailability of medical hospitals may only add to those concerns.

According to the observations made with the help of EDA, the number still seems good right now when considering the population and India. There is a silver lining, however: India implemented a Nationwide Lockdown at the appropriate time. Another good to take away is "unity in diversity," where people are working to aid others who are poor, and people are donating money to the government to combat this epidemic, which might play a huge role in this pandemic.

The people of this country will determine the course of this pandemic; forecasts may appear reasonable in contrast to other countries, but that picture might change in a matter of days. It all relies on how closely people adhere to the rules and restrictions enforced by the Indian government. The vaccination details based on the EDA looks promising. Social distancing in public settings, self-isolation if any COVID-19 symptoms are observed, quarantine of CoV positive patients, lockdown, and other measures are the only possible and effective COVID-19 precautions.

As observed through EDA, COVID-19 has a low mortality rate, which is the most favourable takeaway. Furthermore, a robust Recovery Rate indicates that the condition is treatable. The only cause for concern is the infection's exponential growth rate. Since the last few days, the number of confirmed and fatality cases has appeared to have slowed. This is a very good indicator. There should be no new country emerging as the new COVID-19 epicentre, much as the United States did for a brief while. If a new country emerges as an epicentre, the number of confirmed cases will increase dramatically.

The struggle with COVID begins and ends with the people. This pandemic can be overcome by following the steps given below:

- When everyone goes out of the house, everyone should wear a mask. Face masks are Covid-19's first line of defence.
- It is vital to maintain a gap between persons of more than 6 feet. Citizens who buy food and other important products should keep their physical distance secure.
- Unless obligatory, children under 10 years of age and older adults above 60 years of age should avoid going out.

- For flu/influenza symptoms like fever, cough, sore throat, fluid nose, difficulty breathing, headache and bodily pain, contact the government health center in the nearest city and get treatment in advance.
- Soap and hand washing facilities/sanatorium shall be supplied for the work spaces. There should be a sufficient physical distance between staff.
- The public must avoid unnecessary travel. In the absence of an unavoidable precaution, the use of face masks, frequent hand washing, sanitizer, safe physical distance etc. shall assure every measure of safety.
- Comorbid conditions such as high blood pressure, diabetes, cardiac disease, chronic kidney disease, chronic obstructive pulmonary conditions, cancer and/or all other chronic illnesses, are requested to remain indoor and avoid traveling in a way which avoids exposure to COVID-19 except for the medical treatment.

FUTURE RESEARCH DIRECTIONS

The future scope of this study can include finding the reasons behind the fresh cases of COVID-19 from the public's perception for data specific to India. The analysis can be focused on finding the reasons attributed to spread and impact of the disease, by using machine learning and deep learning approaches and validating the inferences with medical professionals. Improvised K-means clustering algorithm can be used for clustering similar data, based on the public posts from Twitter®. Then the LDA topic model can be applied for discovering the trigram topics relevant to the reasons behind the increase of fresh COVID-19 cases. The future scope may also include automated identification of root causes not only from the textual posts, but also from the emoticons, images and videos posted by the public in OSM.

REFERENCES

Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hamdi, M., & Shah, Z. (2020). Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study. *Journal of Medical Internet Research*, *22*(4), 1–9. doi:10.2196/19016 PMID:32287039

Alenezi, M. N., & Alqenaei, Z. M. (2021). Machine Learning in Detecting COVID-19 Misinformation on Twitter. *Future Internet*, *13*(10), 1–20. doi:10.3390/fi13100244

Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., Khan, I. U., Aslam, N., & Alshahrani, M. S. (2021). Applications of Big Data Analytics to Control COVID-19 Pandemic. *Sensors (Basel)*, *21*(7), 2282. doi:10.339021072282 PMID:33805218

Alvarez-Melis, D., & Saveski, M. (2016). *Topic Modeling in Twitter: Aggregating Tweets by Conversations. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*. The AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/14817

ArunKumar, K. E., Kalaga, D. V., Kumar, M. S., Chilkoor, G., Kawaji, M. & Brenza, T. M. (2021). Forecasting the dynamics of cumulative COVID-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: ARIMA and SARIMA. *Applied Soft Computing*, *103*, 107161. doi:10.1016/j.asoc.2021.107161 PMID:33584158

Bharadwaj, S., Srivastava, S., & Gupta, J. R. P. (2013). Pattern-Similarity-Based Model for Time Series Prediction. *Computational Intelligence*, *31*(1), 1–9. doi:10.1111/coin.12015

Chamola, V., Hassija, V., Gupta, V., & Guizani, M. (2020). A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 90225–90265. doi:10.1109/ACCESS.2020.2992341

Darapaneni, N., Jain, P., Khattar, R., Chawla, M., Vaish, R., & Paduri, A. R. (2020). Analysis and Prediction of COVID-19 Pandemic in India. In *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 291-296). IEEE.

Dsouza, J., & Velan, S. S. (2020). Using Exploratory Data Analysis for Generating Inferences on the Correlation of COVID-19 cases. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1-6). Kharagpur, India: IEEE. doi:10.1109/ICCCNT49239.2020.9225621

Hossain, T., Logan, R. L., IV, Ugarte, A., Matsubara, Y., Young, S., & Singh, S. (2020). COVIDLies: Detecting COVID-19 Misinformation on Social Media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020* (pp. 1-11). London, UK: Association for Computational Linguistics. 10.18653/v1/2020.nlpcovid19-2.11

James, N., & Menzies, M. (2020). Cluster-based dual evolution for multivariate time series: Analyzing COVID-19. *An Interdisciplinary Journal of Nonlinear Science*, *30*(6), 1–11. doi:10.1063/5.0013156 PMID:32611104

Kokatnoor, S. A., & Krishnan, B. (2020). Identification of Misconceptions about Corona Outbreak Using Trigrams and Weighted TF-IDF Model. *Journal of Advanced Research in Dynamical and Control Systems*, *12*(05), 524–533. doi:10.5373/JARDCS/V12SP5/20201788

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. *Information (Basel)*, *10*(4), 1–68. doi:10.3390/info10040150

Li, L., Zhang, Q., Wang, X., Zhang, J., Wang, T., Gao, T., Duan, W., Tsoi, K. K., & Wang, F. (2020). Characterizing the Propagation of Situational Information in Social Media During COVID-19 Epidemic: A Case Study on Weibo. *IEEE Transactions on Computational Social Systems*, *7*(2), 556–562. doi:10.1109/TCSS.2020.2980007

Mahdavi, M., Choubdar, H., Zabeh, E., Rieder, M., Safavi-Naeini, S., Jobbagy, Z., Ghorbani, A., Abedini, A., Kiani, A., Khanlarzadeh, V., Lashgari, Z., & Kamrani, E. (2021). A machine learning based exploration of COVID-19 mortality risk. *PLoS One*, *16*(7), 1–20. doi:10.1371/journal.pone.0252384 PMID:34214101

Maurya, S., & Singh, S. (2020). Time Series Analysis of the Covid-19 Datasets. In *2020 IEEE International Conference for Innovation in Technology (INOCON)* (pp. 1-6). Bengaluru, India: IEEE. doi:10.1109/INOCON50539.2020.9298390

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B., Aslam, W., & Choi, G. S. (2020). COVID-19 Future Forecasting Using Supervised Machine Learning Models. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 101489–101499. doi:10.1109/ACCESS.2020.2997311

Saad, S. E., & Yang, J. (2019). Twitter Sentiment Analysis Based on Ordinal Regression. *IEEE Access: Practical Innovations, Open Solutions*, *7*(1), 163677–163685. doi:10.1109/ACCESS.2019.2952127

Saba, T., Abunadi, I., Shahzad, M. N., & Khan, A. R. (2021). Machine learning techniques to detect and forecast the daily total COVID-19 infected and deaths cases under different lockdown types. *Microscopy Research and Technique*, *84*(7), 1462–1474. doi:10.1002/jemt.23702 PMID:33522669

Schaar, M. V. D., Alaa, A. M., Floto, A., Gimson, A., Scholtes, S., Wood, A., McKinney, E., Jarrett, D., Lio, P., & Ercole, A. (2021). How artificial intelligence and machine learning can help healthcare systems respond to COVID-19. *Machine Learning*, *110*(1), 1–14. doi:10.100710994-020-05928-x PMID:33318723

Sear, R. F., Velasquez, N., Leahy, R., Restrepo, N. J., Oud, S. E., Gabriel, N., Lupu, Y., & Johnson, N. F. (2020). Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 91886–91893. doi:10.1109/ACCESS.2020.2993967 PMID:34192099

Shinde, G. R., Kalamkar, A. B., Mahalle, P. N., Dey, N., Chaki, J., & Hassanein, A. E. (2020). Forecasting Models for Coronavirus Disease (COVID-19): A Survey of the State-of-the-Art. *Springer Nature Computer Science*, *197*(4), 1–15. doi:10.100742979-020-00209-9 PMID:33063048

Shorten, C., Khoshgoftaar, T. M., & Furht, B. (2021). Deep Learning applications for COVID-19. *Journal of Big Data*, *8*(1), 18. doi:10.118640537-020-00392-9 PMID:33457181

Siddiqui, M. K., Morales-Menendez, R., Gupta, P. K., Iqbal, H. M. N., Hussain, F., Khatoon, K., & Ahmad, S. (2020). Correlation Between Temperature and COVID-19 (Suspected, Confirmed and Death) Cases based on Machine Learning Analysis. *Journal of Pure & Applied Microbiology*, *14*(suppl 1), 1017–1024. doi:10.22207/JPAM.14.SPL1.40

Sujath, R., Chatterjee, J. M., & Hassanien, A. E. (2020). A machine learning forecasting model for COVID-19 pandemic in India. *Stochastic Environmental Research and Risk Assessment*, *34*(1), 959–972. doi:10.100700477-020-01827-8 PMID:32837309

Tutsoy, O., Colak, S., Polat, A., & Balikci, K. (2020). A Novel Parametric Model for the Prediction and Analysis of the COVID-19 Casualties. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 193898–193906. doi:10.1109/ACCESS.2020.3033146 PMID:34976560

Varshney, D., & Vishwakarma, D. K. (2021). Analysing and Identifying Crucial Evidences for the prediction of False Information proliferated during COVID-19 Outbreak: A Case Study. In *2021 8th International Conference on Smart Computing and Communications (ICSCC)* (pp. 47-51). 10.1109/ICSCC51209.2021.9528205

WaltmanL.PinfieldS.RzayevaN.OliveiraH. S.FangZ.BrumbergJ. (2021): Scholarly communication in times of crisis: The response of the scholarly communication system to the COVID-19 pandemic. *Research on Research Institute Report.* doi:10.6084/m9.figshare.17125394.v1

ADDITIONAL READING

Boyle, P. (2021). 'Covid-19 underlines the need for full open access'. Times Higher Education. https://www.timeshighereducation.com/blog/covid-19-underlines-need-full-open-access

Gagan, M. (2020). 'Science communication as a preventative tool in the COVID19 pandemic'. *Humanities and Social Sciences Communications*, *7*(159), 1–14. doi:10.105741599-020-00645-1

Hongwei, Z., Naveed, N. M., Alyssa, M., Tiffany, A. R., Murray, J. C., Rebecca, S. B. F., Huiyan, S., & Marcia, G. O. (2021). 'COVID-19: Short term prediction model using daily incidence data'. *PLoS One*, *16*(4), e0250110. doi:10.1371/journal.pone.0250110 PMID:33852642

Kazuhiro, H. (2021). ' How Could COVID-19 Change Scholarly Communication to a New Normal in the Open Science Paradigm?'. *Patterns*, *2*(1), 100191. Advance online publication. doi:10.1016/j.patter.2020.100191

Kiley, R. (2020). 'Open access: how COVID-19 will change the way research findings are shared'. Wellcome. https://wellcome.org/news/open-access-how-covid-19-will-change-way-research-findings-are-shared

Matthew, Z. D., Roger, B., Janesse, B., & Daniel, A. S. (2021). 'Walking the Tightrope: Re-evaluating science communication in the era of COVID-19 vaccines'. *Vaccine*, *39*(39), 5453–5455. doi:10.1016/j.vaccine.2021.08.037 PMID:34446317

Misra, D. P., & Agarwal, V. (2016). Open Access Publishing in India: Coverage, Relevance, and Future Perspectives. *Journal of Korean Medical Science*, *34*(27), e180. doi:10.3346/jkms.2019.34.e180 PMID:31293108

Robert, C. M., & Jillian, C. T. (2021). 'Scholarly Publishing in the Wake of COVID-19'. *International Journal of Radiation Oncology, Biology, Physics*, *108*(2), 491–495. doi:10.1016/j.ijrobp.2020.06.048 PMID:34044094

Robinson-Garcia, N., van Leeuwen, T. N., & Torres-Salinas, D. (2020). Measuring Open Access Uptake: Data Sources, Expectations, and Misconceptions. *Scholarly Assessment Reports*, *2*(1), 15. doi:10.29024ar.23

Tavernier, W. (2020). 'COVID-19 demonstrates the value of open access: What happens next?' Association of College and Research Libraries. https://crln.acrl.org/index.php/crlnews/article/view/24414/32251

KEY TERMS AND DEFINITIONS

Artificial Intelligence (AI): AI refers to a computer or a robot controlled by a computer's capacity to do jobs that are normally performed by people because they require human intellect and judgement.

AutoRegressive Integrated Moving Average (ARIMA): ARIMA is a statistical analysis model that uses time-series data to understand a data set better or predict future patterns in the data set. Autoregressive statistical models predict future values based on the previous values.

Centers for Disease Control and Prevention (CDC): The Centers for Disease Control and Prevention (CDC) is the nation's health protection agency, operating around the clock to keep America safe from foreign and domestic health and safety risks. The CDC improves our country's health security.

Data Science (DS): Data science is an interdisciplinary subject that combines scientific techniques, procedures, algorithms, and systems to extract knowledge and insights from noisy, structured, and unstructured data.

Decision Trees (DTs): DTs are used for classification and regression in non-parametric supervised learning. The objective is to learn basic decision rules using data attributes to forecast the value of a target variable. A tree is a constant piecewise approximation.

Exponential Smoothing (ES): ES is a univariate time series forecasting approach that may be expanded to accommodate data with a systematic trend or seasonal component. It is a strong forecasting approach that may be used in place of the popular Box-Jenkins ARIMA family of algorithms.

Exploratory Data Analysis (EDA): EDA is a data analysis approach that enables the discovery of hidden information within a data collection. This technique is frequently used to derive inferences from data.

FBProphet (FP): The FBProphet library, which is created by Facebook and is primarily used for time series forecasting, is used in the prediction analysis.

Holt's Linear Model (HLM): A prominent smoothing technique for predicting data with trend is Holt's two-parameter model, sometimes known as linear exponential smoothing. Holt's model consists of three different equations that interact to provide a final forecast.

Holt's Winter Model (HWL): HWM is a time series behavior model. Forecasting usually necessitates the use of a model, and Holt-Winters is a method for modelling three components of a time series: a typical value (average), a slope (trend) across time, and a cyclical repeating pattern (seasonality).

Latent Dirichlet Allocation (LDA): The LDA is a generative statistical model that allows unobserved groups to explain why some parts of the data are similar.

Least Absolute Shrinkage and Selection Operator (LASSO): LASSO is a regression analysis approach that uses attribute selection and regularization to improve the predictability and interpretability of the final statistical model.

Linear Regression: A linear approach to modeling the relationship between a scalar response and one or more explanatory factors is known as linear regression (also known as dependent and independent variables).

Logistic Regression: Logistic regression is a statistical model that uses a logistic function to represent a binary dependent variable in its most basic form; however, many more advanced extensions exist. Logistic regression (or logit regression) in regression analysis is used to estimate the parameters of a logistic model (a form of binary regression).

LSTM-Regression: The LSTM model is a Gated Recurrent Neural Network, and bidirectional LSTM is simply an extension of that model. The crucial aspect is that these networks may save information for future cell processing.

Machine Learning (ML): ML is a sort of artificial intelligence (AI) that allows software programs to improve their prediction accuracy without being expressly designed to do so. In order to forecast new output values, machine learning algorithms use past data as input.

Neural Network (NN): A neural network is a set of algorithms that attempts to detect underlying relationships in a batch of data using a technique similar to how the human brain works. In this context, neural networks are systems of neurons that might be biological or artificial in origin.

Precision and Recall: Precision (also known as positive predictive value) is the proportion of relevant examples discovered among the recovered instances, whereas recall (also known as sensitivity) is the proportion of non-relevant instances found among the retrieved instances. Precision and recall are two different concepts. As a result, relevance determines the precision and recall of an experiment.

Predictive Analytics (PA): PA is a subset of advanced analytics that predicts future events by combining historical data with statistical modelling, data mining tools, and machine learning. Companies use predictive analytics to discover hazards and opportunities by looking for trends in data.

Random Forest (RF): Random Forest is a Supervised Machine Learning Algorithm frequently utilised in Classification and Regression applications. It constructs decision trees from several samples and uses their majority vote for classification and average for regression.

Ridge Regression: Ridge regression is a technique for estimating the coefficients of multiple-regression models when the variables are linearly independent but highly linked. It has been applied in various domains such as econometrics, chemistry, engineering, etc.

Seasonal Auto-Regressive Integrated Moving Average With eXogenous Factors (SARIMAX): SARIMAX used to forecast daily Covid-19 cases in this chapter.

World Health Organization (WHO): The WHO, dedicated to the well-being of all people and informed by science, leads and champions worldwide efforts to provide everyone, everywhere, an equal chance to live a healthy life.