International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Exploratory Data Analysis, Prediction of COVID-19 Pandemic

Bhargav Ambati[1], Boja Yashoda Krishna[2], Akash Makane[3], Jyotsna Pyarasani[4]

[1]UG Student, Dept. of Electronics & Communication, KG Reddy College of Engineering & Technology, Telangana, India
[2,3,4]UG Student, Dept. of Computer Science, KG Reddy College of Engineering & Technology, Telangana, India

---

Abstract - *COVID-19, Corona virus Disease-2019, belongs to the genus of Coronaviridae. A virus with no vaccine, creating unpredictable havoc in the human lives and financial and economic systems in every country throughout the world. COVID-19 global pandemic prediction, simple epidemiological and statistical models have received more attention from authorities, and these models are popular in the media. Due to a high level of uncertainty and lack of essential data, standard models have shown low accuracy for long-term prediction. Although the literature includes several attempts to address this issue, the essential generalization and robustness abilities of existing models need to be improved.*

This study presents the current situation of COVID-19 spread in India along with the major factors that lead to the rapid spread with the inclusion of major impacts of the Outbreak on various sectors and Domains. The current Data is used as a source for Data-centric comparison of COVID-19 with the previous pandemics. An Exploratory Data analysis((EDA) technique is implemented to study and analyze the reports. Different prediction models are built using machine learning algorithms and their performances are computed and evaluated. Linear Regression, Polynomial Linear Regression and Recurrent Neural Network (RNN).

Key Words: COVID-19, Linear Regression, Polynomial Linear Regression and Recurrent Neural Network (RNN).

1. INTRODUCTION

This COVID-19 epidemic occurred in December 2019 in Wuhan China. World Health Organization (WHO) declared on January 30, 2020 the outbreak as an emergency and Pandemic for public health. COVID-19's clinical symptoms are respiratory disorder, fatigue, dry cough, tiredness, etc. while 80 percent of patients heal without any care. On March 6, the total amount of cumulated infection cases over the world was 97,000 and 3,400 deaths [WHO]. On March 11, the virus outbreak was declared a pandemic by the World Health Organization, as the virus spread to 114 countries, totaling over 118,000 recorded cases and 4,300 deaths [WHO].

Furthermore, recent reports expose an astonishing case fatality rate of 61.5 percentage for critical cases, increasing sharply with age and for patients with underlying comorbidities3. Both the reach and severity of cases are putting great pressure on the medical services and readily lead to a shortage of intensive care resources. Access to accurate outbreak prediction models is essential to obtain insights into the likely spread and consequences of infectious diseases.

Governments and other legislative bodies rely on insights from prediction models to suggest new policies and to assess the effectiveness of the enforced policies. The novel corona virus disease (COVID-19) has been reported to have infected more than 2 million people, with more than 132,000 confirmed deaths worldwide. The recent global COVID-19 pandemic has exhibited a nonlinear and complex nature. During such an emergent situation, it is important for clinicians to effectively and efficiently triage patients. In recent months, studies have proposed several machine learning models that can accurately predict COVID-19 disease severity.

Many of these studies have been successful in generating a high-performing model. However, until now, these models have only been trained on one kind of machine learning algorithm, and many researchers have limited the evaluation of their models' performance to the area under the curve (AUC) analysis.

2. SYMPTOMS

The most common symptoms of COVID-19 are flu-like symptoms. The details are tabulated in Fig 1. Due to mild and unspecific symptoms, it is becoming difficult to identify and quarantine.

Table -1: Symptom of COVID-19

Most common	Moderate	Severe
Fever, dry cough & tiredness	Aches and pains, Sore throat, conjunctivitis, headache, loss of taste or smell, a rash on the skin, or discoloration of fingers or toes	Difficulty in breathing or shortness of breath, chest pain or pressure, loss of speech or movement

COVID-19 affects different people in different ways. Most infected people will develop mild to moderate illness and recover without hospitalization.

A. Most common symptoms:

- fever
- dry cough
- tiredness

B. Less common symptoms:

- aches and pains
- Sore throat
- diarrhea
- conjunctivitis
- headache
- loss of taste or smell
- a rash on the skin, or discoloration of fingers or toes

C. Serious symptoms:

- difficulty breathing or shortness of breath
- chest pain or pressure
- loss of speech or movement

Seek immediate medical attention if you have serious symptoms. Always call before visiting your doctor or health facility. People with mild symptoms who are otherwise healthy should manage their symptoms at home. On average, it takes 5–6 days from when someone is infected with the virus for symptoms to show, however it can take up to 14 days. [1]

3. IDENTIFICATION AND CONFIRMATION

Viral tests notify about infection with SARS-CoV-2, the virus which triggers COVID-19. If a test results as positive it indicates the person is infected. The diagnostic test is dependent on the affected person's geographic location [2].

Rapid Diagnostic Test (RDT) tests for the existence of proteins of the virus, called antigens, developed by the SARSCoV-2 virus in the respiratory tract of a person. Usually within 30 minutes, if the SARS-CoV-2 antigen exists in sufficient concentrations in the collected sample, it can bind to numerous antibodies attached to a paper strip in a plastic case. It generates a signal which is easily detectable. The RDT tests are used for diagnosing the acute or early infections of SARSCoV-2, as the developed antigens are released only when the virus replicates successfully. These tests are considered reliable for diagnosing COVID-19 [3].

Another specific form of RDT advertised for COVID-19; a test that measures the existence of antibodies in the blood of those suspected to have been COVID-19 infected. Antibodies grow within days to weeks after infection with the virus.

4. TREATMENT

The COVID-19 infected patients have no defined treatment. The medication is given based on symptoms. It may include pain relievers, cough syrup, rest and fluid intake. If the patients have mild symptoms, they may stay at home and take treatment in isolation. Otherwise, treatment in the hospital is evident [3].

A. Self-care

Asymptomatic cases, mild cases of COVID-19

- Isolate yourself in a well-ventilated room.
- Use a triple layer medical mask, discard the mask after 8 hours of use or earlier if they become wet or visibly soiled. In the event of a caregiver entering the room, both caregiver and patient may consider using N 95 masks.
- Mask should be discarded only after disinfecting it with 1 percentage Sodium Hypochlorite.
- Take rest and drink a lot of fluids to maintain adequate hydration.
- Follow respiratory etiquette at all times.
- Frequent hand washing with soap and water for at least 40 seconds, or clean with an alcohol-based sanitizer.
- Don't share personal items with other people in the household.
- Ensure cleaning of surfaces in the room that are touched often (tabletops, doorknobs, handles, etc.) with 1 percentage hypochlorite solution.
- Monitor temperature daily.
- Monitor oxygen saturation with a pulse oximeter daily.
- Connect with the treating physician promptly if any deterioration of symptoms is noticed.

B. Instructions for caregivers:

- Mask: The caregiver should wear a triple layer medical mask. N95 mask may be considered when in the same room with the ill person.
- Hand hygiene: Hand hygiene must be ensured following contact with an ill person or patient's immediate environment.
- Exposure to patient/patient's environment: Avoid direct contact with body fluids of the patient, particularly oral or respiratory secretions. Use disposable gloves while handling the patient. Perform hand hygiene before and after removing gloves.

5. MEDICAL TREATMENTS

Treatment for patients with mild/asymptomatic disease in home isolation

- Patients must be in communication with a treating physician and promptly report in case of any worsening.
- Continue the medications for other co-morbid illnesses after consulting the treating physician.
- Patients to follow symptomatic management for fever, running nose and cough, as warranted.
- Patients may perform warm water gargles or take steam inhalation twice a day.

When to seek immediate medical attention:

- Difficulty in breathing
- Dip in oxygen saturation (SpO2 ¡ 94 percentage on room air)
- Persistent pain/pressure in the chest
- Mental confusion or inability to arouse. [4]

6. ALGORITHM

A. Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Mathematically, we can represent a linear regression as:

$$y= a_0 + a_1x + \varepsilon$$

Here,

- Y= Dependent Variable (Target Variable)
- X= Independent Variable (predictor Variable)
- a_0= intercept of the line (Gives an additional degree of freedom)
- a_1 = Linear regression coefficient (scale factor to each input value).
- ε = random error

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

Fig. 1: Linear Regression flowchart

For Linear Regression, we use the Mean Squared Error (MSE) cost function, which is the average of squared error that occurred between the predicted values and actual values.

For the above linear equation, MSE can be calculated as:

$$MSE= 1\frac{1}{N}\sum_{i=1}^{n}(y_i - (a_1x_i + a_0))^2$$

Where,

- N=Total number of observation
- Y_i = Actual value
- $(a_1x_i + a_0)$= predicted value. [5]

B. Polynomial Regression

Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modeled as an nth degree polynomial.

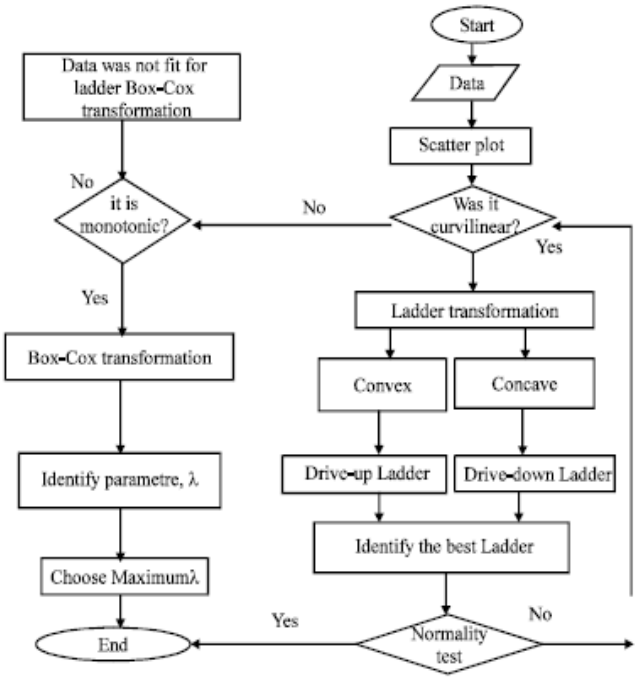

Fig. 2: Polynomial Regression flowchart

It is used in many experimental procedures to produce the outcome using this equation. It provides a great defined relationship between the independent and dependent variables. It is used to study the isotopes of the sediments. [6]

Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y —x)
Polynomial Regression equation:

$$y= b_0 + b_1x + b_2x^2 + b_3x^3 + \dots + b_nx^n$$

C. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a type of Neural Network where the output from previous step are fed as input to the current step. In traditional neural networks, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words.

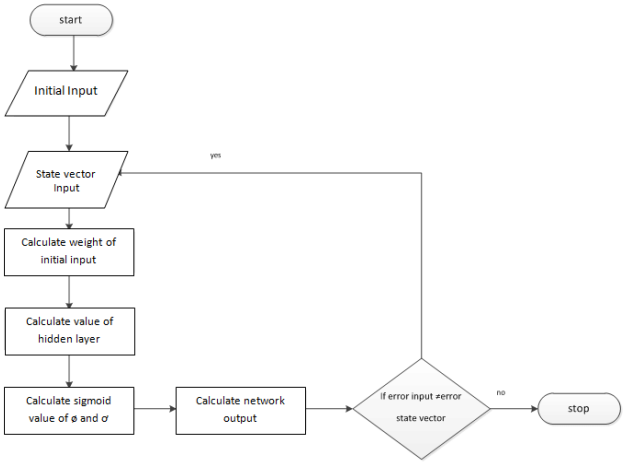

Fig. 3: Recurrent Neural Network flowchart

Thus, RNN came into existence, which solved this issue with the help of a Hidden Layer. The main and most important feature of RNN is hidden state, which remembers some information about a sequence.

$$h_t = \int(h_{t-1}, X-t)$$

- h_t = Current state
- h_{t-1} = Previous state
- X - t = Input state

RNN converts the independent activations into dependent activations by providing the same weights and biases to all the layers, thus reducing the complexity of increasing parameters and memorizing each previous outputs by giving each output as input to the next hidden layer. [7]

7. RESULT AND ANALYSIS

The Machine Learning Techniques Linear Regression, Polynomial Linear Regression and Recurrent Neural Network (RNN) are used to understand the COVID-19 affecting people, its confirmation and recovery predictions. The datasets Covid-19 India is used to analyze their features and to build ML models for performance assessment.

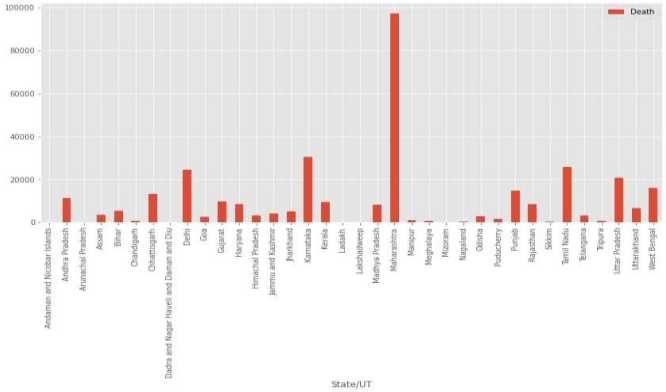

Fig. 4: Representation of the total deaths recorded in each state

The above bar graph represents the relationship between the total deaths recorded in each state. Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack.

	State/UT	Confirmed Cases	Active Cases	Cured/Discharged	Death	Total cases
0	Andaman and Nicobar Islands	7070	131	6820	119	14021
1	Andhra Pradesh	1728577	138912	1578452	11213	3445941
2	Arunachal Pradesh	28382	3843	24420	119	56645
3	Assam	424385	51881	368981	3523	845247
4	Bihar	710199	11431	693472	5296	1415102
5	Chandigarh	60399	1135	58502	762	120036
6	Chhattisgarh	976760	29378	934243	13139	1940381
7	Dadra and Nagar Haveli and Daman and Diu	10345	250	10091	4	20686
8	Delhi	1427926	8748	1394731	24447	2831405
9	Goa	157847	9700	145437	2710	312984
10	Gujarat	813270	24404	778976	9890	1616650
11	Haryana	760019	12688	738799	8532	1511506
12	Himachal Pradesh	193137	11057	178847	3233	383041
13	Jammu and Kashmir	295879	30657	261230	3992	587766
14	Jharkhand	339930	7537	327372	5021	674839
15	Karnataka	2653446	286819	2336096	30531	5276361
16	Kerala	2584853	184699	2390779	9375	5160331
17	Ladakh	18954	1505	17256	193	37715
18	Lakshadweep	8479	1355	7089	35	16923
19	Madhya Pradesh	782945	14186	760552	8207	1557683
20	Maharashtra	5791413	207813	5486206	97394	11485432
21	Manipur	52899	8863	43187	849	104949
22	Meghalaya	37149	6352	30172	625	73673
23	Mizoram	13064	3415	9602	47	26081
24	Nagaland	22240	4711	17125	404	44076
25	Odisha	790970	75042	713055	2873	1579067
26	Puducherry	107114	10015	95516	1583	212645
27	Punjab	574114	28673	530601	14840	1133388
28	Rajasthan	943494	27408	907527	8559	1878429

Fig. 5: Gradient depth according to cases

The above figure shows the rigid plot. A rigid graph is an embedding of a graph in a Euclidean space which is structurally rigid. That is, a graph is rigid if the structure formed by replacing the edges by rigid rods and the vertices by flexible hinges is rigid.

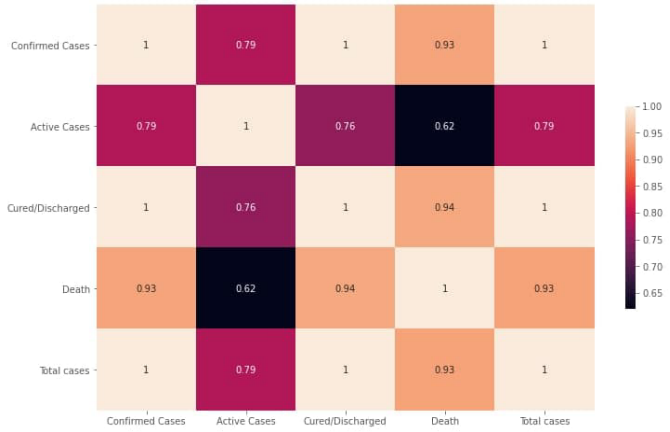

Fig. 6: Heatmap

A heatmap contains values representing various shades of the same color for each value to be plotted. Usually the darker shades of the chart represent higher values than the lighter shade.

Fig. 7: Indian map with COVID-19 situations using Folium

Folium is a Python library used for visualizing geospatial data. It is easy to use and yet a powerful library. Folium is a Python wrapper for Leaflet.

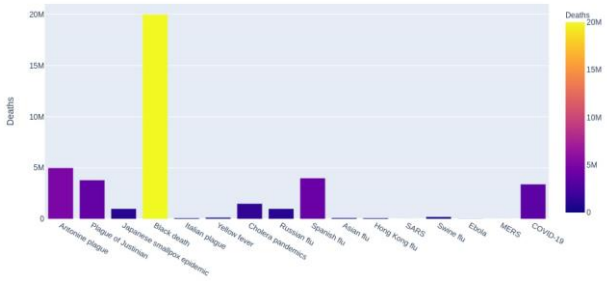

Fig. 8: comparison of COVID-19 with previous pandemics

The plague is a serious bacterial infection that can be deadly. Sometimes referred to as the “black plague,” the disease is caused by a bacterial strain called Yersinia pestis. This bacterium is found in animals throughout the world and is usually transmitted to humans through fleas.Bubonic plague infects your lymphatic system (a part of the immune system), causing inflammation in your lymph nodes. Untreated, it can move into the blood (causing septicemic plague) or to the lungs (causing pneumonic plague).Black plague has registered more number of deaths when compared with the COVID-19 and previous pandemics.

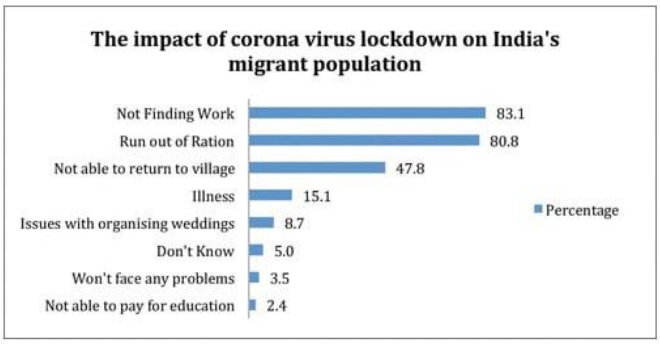

Fig. 6: impact on various sectors

On 16 March 2020, the union government ordered the closure of schools and colleges. Only a few educational institutions in India have been able to effectively adapt to e-learning and remote learning; the digital divide is further impacted by serious electricity issues and lack of internet connectivity.

CONCLUSIONS

Using the above methods from libraries and plots, we obtained clear analysis and visualizations of the data. Where in everyone can understand the range, Spread and Mortality rate of the cases by observing the bar, scatter plots, pie chart, Matrix table and the Map representation. The correlation matrices are built to understand the relationship between the features of the datasets. The feature importance is computed for the classifiers built. Along the classifiers and repressors are also built for prediction.

The results show that the Linear Regression, Polynomial Linear Regression and Recurrent Neural Network (RNN) Classifier has outperformed other models in terms of CoD and Accuracy. In the future, more ML classifiers and Regressors are evaluated on the evolving COVID-19 datasets.

REFERENCES

[1] WorldHealthOrganization https://www.who.int/emergencies/diseases/novelcoronavirus-2019/question-and-answers-hub/q-a-detail/coronavirusdisease-covid-19””::text=symptoms
[2] Corona virus disease 2019 (COVID-19), https://www.mayoclinic.org/diseasesconditions/coronavirus/diagnosis-treatment/drc20479976
[3] Advice on the use of point-of-care immunodiagnostic tests for COVID-19, https://www.who.int/newsroom/commentaries/detail/adviceon-the-use-of-pointof-care-immunodiagnostic-tests-for-covid-19
[4] Government of India Ministry of Health and Family Welfare https://www.mohfw.gov.in/pdf/RevisedguidelinesforHomeIsolationofmildasymptomaticCOVID19cases.pdf
[5] LinearRegression https://www.javatpoint.com/linear-regression-inmachine-learning.
[6] PolynomialRegression https://www.javatpoint.com/machine-learningpolynomial-regression
[7] Recurrent Neural Network (RNN) https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network

BIOGRAPHIES

Ambati Bhargav
Student
Dept of Electronic & Communications
KGRCET

Boja Yashoda Krishna
Student
Dept of Computer Science
KGRCET

Akash Makane
Student
Dept of Computer Science
KGRCET

Jyotsna Pyarasai
Student
Dept of Computer Science
KGRCET