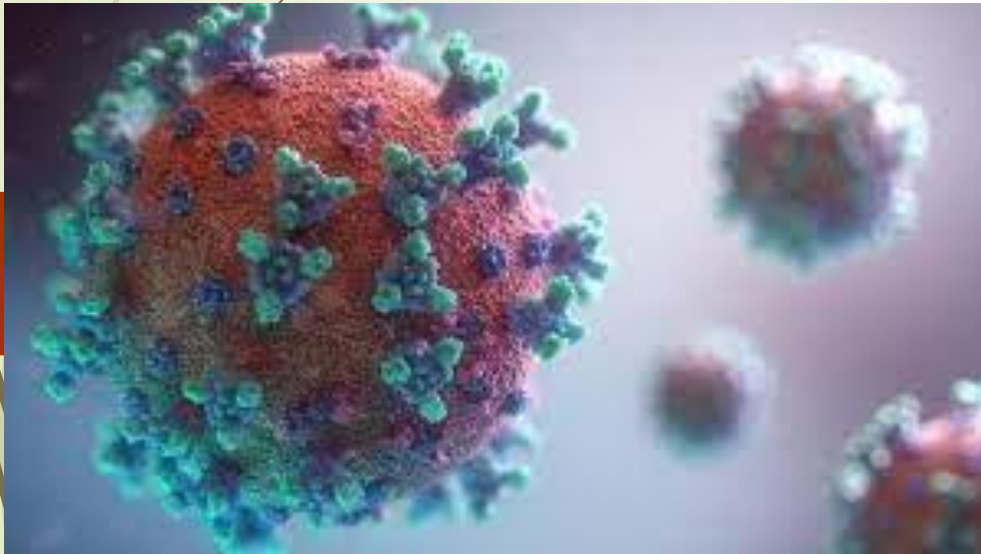


AMRUTA INSTITUTE OF ENGINEERING AND MANAGEMENT SCIENCES

Forecasting and Severity Analysis of COVID-19 using Machine Learning Approach




Under The Guidance of :
Asst. Prof. Ravi

TEAM MEMBERS:

NAGAJYOTHI P(1AR18CS024)
MAMATHA U (1AR18CS018)
RASHMITHA R(1AR18CS035)
ASHWINI (1AR18CS004)

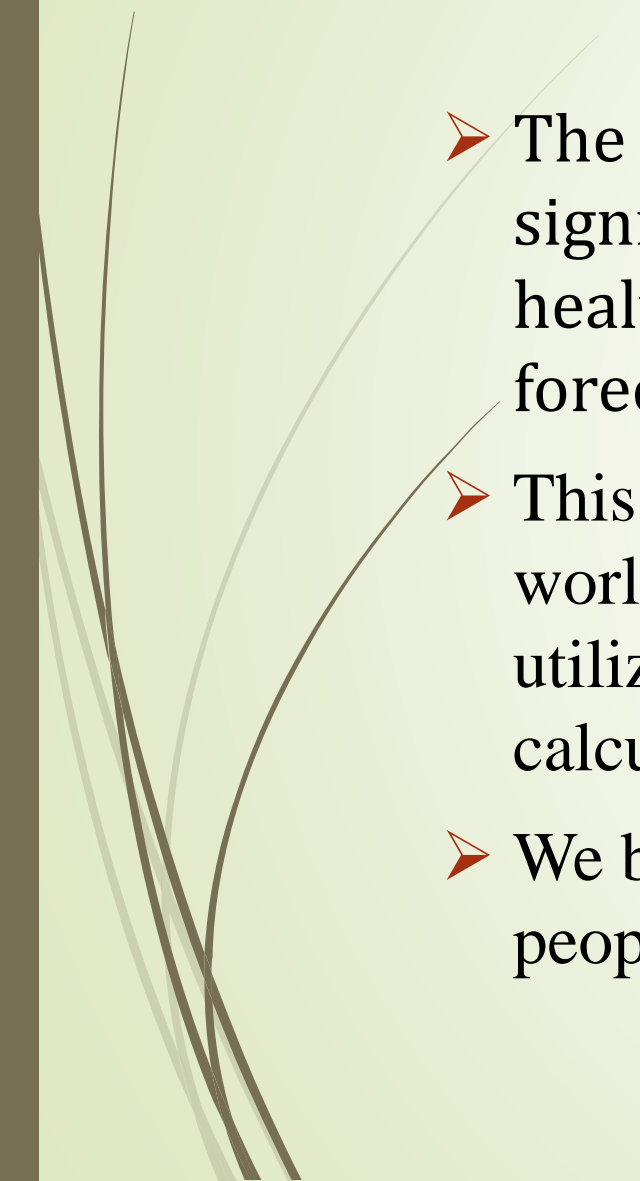


AGENDA

- 
- 1. Introduction**
 - 2. Literature survey / Existing system and drawbacks**
 - 3. Problem identification**
 - 4. Objectives**
 - 5. Requirements (hardware and software)**
 - 6. Methodology**
 - 7. System Representation (Input Dataset, Data Analysis. Linear Algorithm, Clustering Algorithm, ARIMA Model, Forecasting Model.**
 - 8. Modules**
 - 9. Implementation**
 - 10. Applications**
 - 11. Conclusion and Future enhancement**
 - 12. Result analysis**



INTRODUCTION

- The global spread of the COVID-19 pandemic has caused significant losses. The most critical issues, medical and healthcare departments are facing is the fact that the COVID-19 forecasting is important.
 - This project focused on the coronavirus pandemic situation in the world region and its related effects and future status. We have utilized different information representation and machine learning calculations to recreate the affirmed, recuperated, and passing cases.
 - We believe the research will help scientists, researchers, and ordinary people predict and analyze this pandemic's impact.
- 

LITERATURE SURVEY

Early Detection of COVID-19 Hotspots Using Spatio-Temporal Data	Predicting COVID-19 in China Using Hybrid AI Model	A greedy-based oversampling approach to improve prediction of mortality in MERS patients	A Weakly-Supervised Framework for COVID-19 Classification and Lesion Localization From Chest CT
<p>The Centers for Disease Control and Prevention (CDC) with other federal agencies have identify counties with a significant increase in COVID-19 incidence (hotspots), which offers a unique opportunity to investigate the spatio-temporal dynamics between the identified hotspots.</p>	<p>The corona virus disease 2019 (COVID-19) breaking out in late December 2019 is gradually being controlled in China, but it is still spreading rapidly in many other countries and regions worldwide. It is urgent to conduct prediction research on the development and spread of the epidemic. In this article, a hybrid artificial-intelligence (AI) model is proposed for COVID-19 prediction.</p>	<p>Predicting mortality of Middle East respiratory syndrome (MERS) patients with identified outcomes is a core goal for hospitals in deciding whether a new patient should be hospitalized or not in the presence of limited resources of the hospitals. We present an oversampling approach that we call Greedy-Based Oversampling Approach (GBOA).</p>	<p>Accurate and rapid diagnosis of COVID-19 suspected cases plays a crucial role in timely quarantine and medical treatment. Developing a deep learning-based model for automatic COVID-19 diagnosis on chest CT is helpful to counter the outbreak of SARS-CoV-2. A weakly-supervised deep learning framework was developed using 3D CT volumes for COVID-19 classification and lesion localization.</p>

PROBLEM IDENTIFICATION

- Multiple companies have launched the vaccines in different countries. But, to be fully vaccinated in the world is a time taking process and, roughly it takes approximately five years for the full-fledged immunization to the world's population. So, there are two possible solutions, either the vaccine is available for everyone or the people follow the SOPs to avoid the spread of the coronavirus. Furthermore, on the other side, the possible solution is to follow the SOPs set by the
- World Health Organization (WHO) for the prevention of this disease. Forecasting helps in prevention of the disease.



OBJECTIVE

- This project is a push to dissect the aggregate information of affirmed passings and recouped cases over the long run, which is examined in the information investigation area. The primary center is to investigate the spread pattern of this infection around the country.
- This project proposing two algorithms, i.e., Logistic Regression algorithm & Random Forest Regression forecasting algorithms to measure the daily increase in confirmed, recovered, death cases and growth factor.

HARDWARE REQUIREMENT

- System Processor : i7 / i5 / i3 processor
- Hard Disk : 500 GB.
- Ram : 8 GB / 12 GB.
- *Any desktop / Laptop system with above configuration or higher level.*



SOFTWARE REQUIREMENT

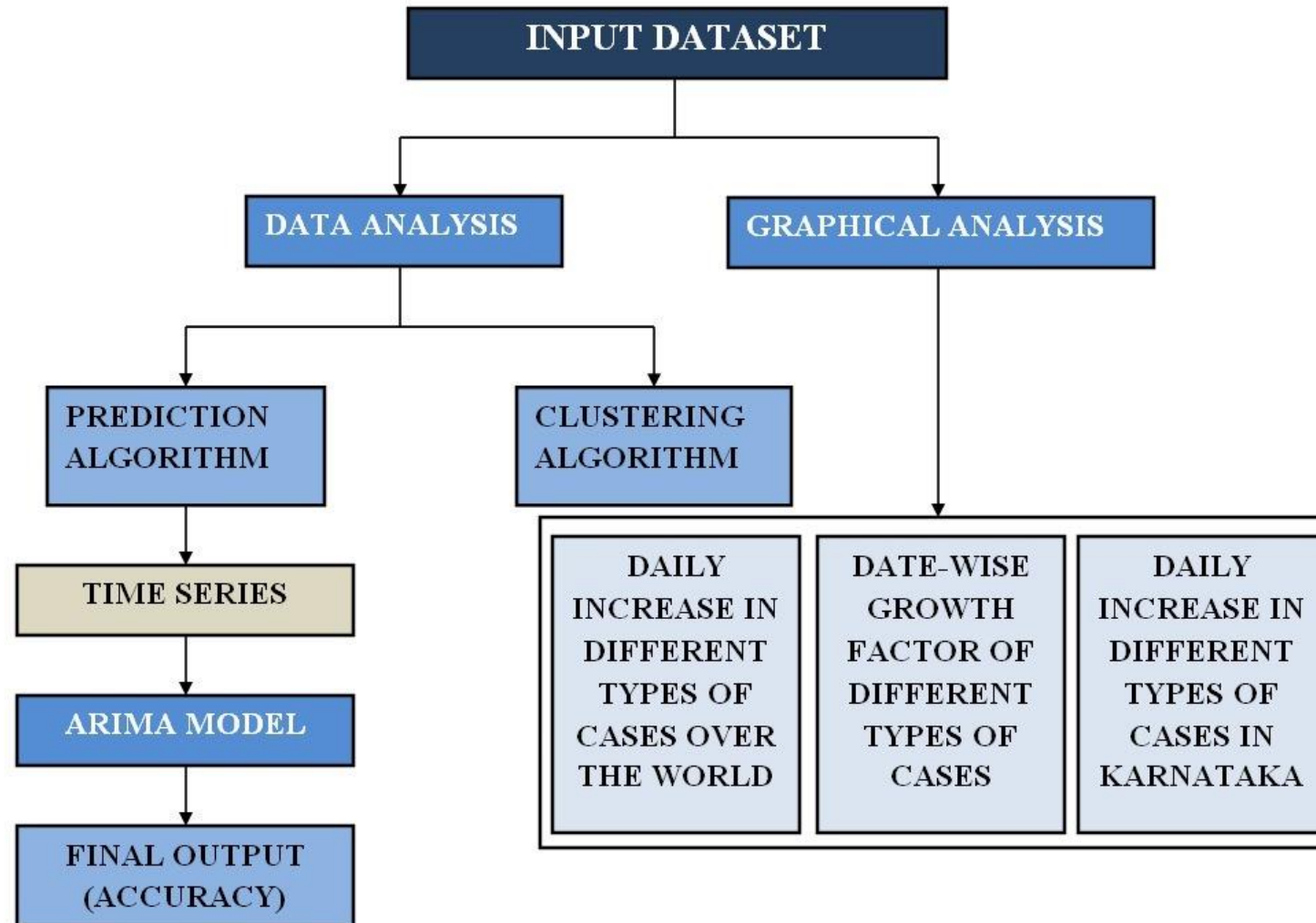
- **Operating system** : Windows 8 / 10 (64 bits OS)
- **Programming Language** : Python 3
- **Framework** : Anaconda
- **Libraries** : SKLEARN, MATPLOTLIB, NUMPY, PANDAS
- ➡ **IDE** : Jupyter Notebook


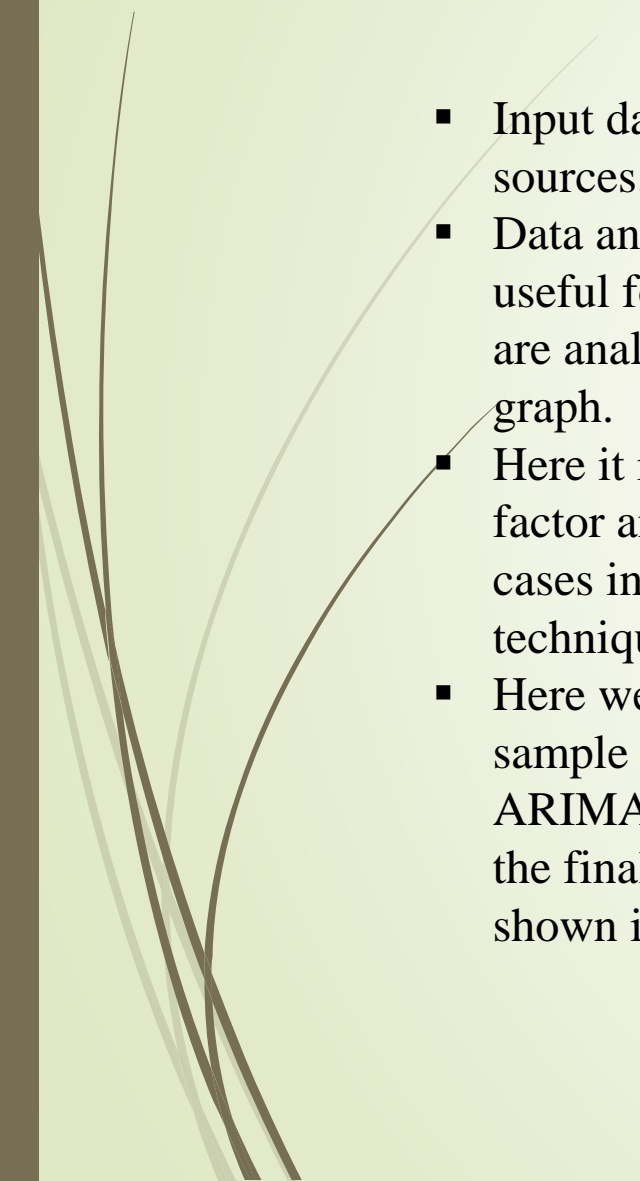


METHODOLOGY

- **Data Collection**
- **Data Preprocessing**
- **Data Visualization using Matplotlib**
- **Data Splitting**
- **Data Training using Logistic Regression / Random Forest Regression**
- **Result Analysis**

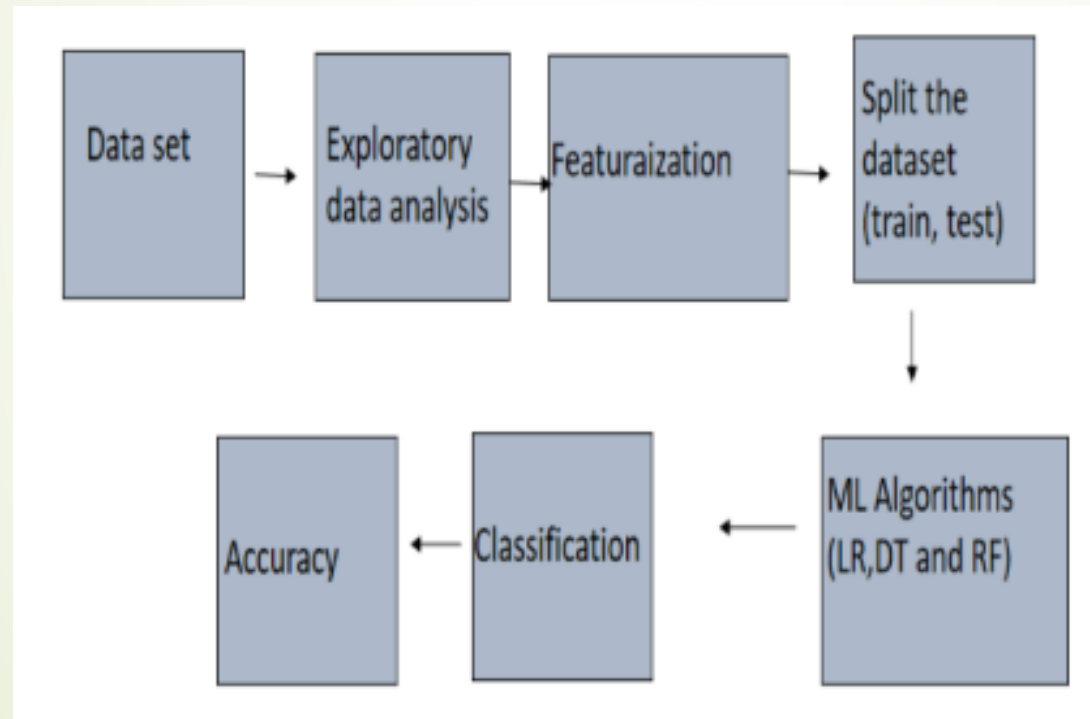
METHODOLOGY



- 
- 
- Input dataset refer to collection of dataset and the dataset can be downloaded from various sources.
 - Data analysis, is a process for obtaining raw data, and subsequently converting it into information useful for decision-making by users. Data, is collected and it has been analyzed. Graph Analysis are analytic tools used to determine strength and direction of relationships between objects in a graph.
 - Here it is broadly classified into three categories i.e., daily increase in case, date wise growth factor and daily increase in case in Karnataka. Growth factor is calculated by dividing the new cases in the present day to the new cases in the previous day. Prediction algorithm is a statistical technique using machine learning and data mining to predict and forecast
 - Here were using linear regression algorithm (tell about it). A time series is a data set that tracks a sample over time. Time series analysis can be useful to see how a given asset changes over time. ARIMA model is a class of statistical models for analyzing and forecasting time series data. So, the final output will be in accuracy and we got the accuracy of (%) and even the final output is shown in statistical graph also.

INPUT DATASET

Data is the foundation for any machine learning project. Each of these phases can be split into several steps.



DATA ANALYSIS

- We have utilized the perception time, region/state, nation/locale, last perception date, affirmed cases, recuperated cases and, passing cases on date-wise. We have collected the daily increase in confirmed, recovered, and death cases. We have used the date(observation) in the x-axis and the number of cases (confirmed, recovered, deaths) on the y-axis.
- At that point we are going to estimate the development factor, which is an amount increases itself after some time. The formula we have applied: $G=E/P$
- Here, G = Growth Factor, E = Consistently's new (affirmed, recuperated, passing) cases, and P = New (affirmed, recuperated, demise) cases on the earlier day. A development factor consistent at one assigns there is no variety in any cases. We have used the date(observation) in the x-axis and the growth factor in the y-axis. Similarly, we estimate the day by day increment in various sorts of cases.

CLUSTERING ALGORITHM (K-Means)

- K-means clustering is a method of quantization vectors, which is common in data harvesting of cluster analysis. Unpredictable algorithms usually deduct data sets using only input vectors without referring to displayed or specified consequences. The approach presents a hierarchical and fast way to evaluate a given collection of data across a clustering algorithm (implying k clusters) concluded by the centering of data.
- Finally, the aim of using k means clustering for decreasing an outside function acknowledged as the squared error function, which is given:

$$M = \sum_{j=1}^k \sum_{i=1}^n \|x^{(j)}_i - y_j\|^2$$

CLUSTERING ALGORITHM (K-Means)

1.Distance Measure:

- To figure out the similarities between the alternatives, this distance between the points takes as a standard metric. The basic separation computation utilizes the Euclidean measurement, that characterizes the span among a^i and b^i , where:

$$d=[\Sigma(ai-bi)^2]^{1/2}$$

2. Algorithm Steps:

The following approaches to this algorithm:

- Here K points are objecting state effective internal control, and the initial position is labeled.
- Each group object assigns to the nearest centroid.
- Recalculate the position of the K centroids after assigning all the objects.
- Phases 2 and 3 likely proceed. And why the category should differ from the one that minimizes the metric.

LINEAR REGRESSION ALGORITHM

- There are many machine learning algorithms, and here, the equation is formed by Linear construction on the data with a direct curve correlation. This relationship is created between the objective variable and the free factors. The method is to acquire a regression model from a set of input values (calculated values).
- This process is a very straightforward machine learning approach, where the dependent variable is modeled as a linear order of predictors. More high-level regression methods and models include regression and multivariable regression, which is an acknowledged linear regression algorithm.

LINEAR REGRESSION ALGORITHM

- At a simple linear regression method, one feature or single independent variable. So, the formula for linear regression is:

$$y = \theta_0 + \theta_1 X_1$$

- Suppose, there is only one variable, the model represented relationship formula is:

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \theta_3 X^3 + \dots + \theta_m X^n$$

- Here, n is the linear equation degree. As n (degree) increases, the feature complexity increases. The stated linear regression assessment used to forecast each number of reported case scenarios. The θ_m is the polynomial achieved by applying a sci-kit-learn Machine Learning algorithm in Python.

ARIMA MODEL

- ARIMA, a model set up by Box and Jenkins, is furthermore called the B-J model, and it's a mix of the AR (Auto-Regressive) model and the MA (Moving Average) model.

Econometricians normally use ARIMA models applied to figure time arrangement, and the ARIMA model is known as the most unpredictable and progressed time arrangement determining technique in the global. The generalized equation of the ARIMA model is:

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \dots + \alpha_p X_{t-p} + \mu_t - \beta_1 \mu_{t-1} - \beta_2 \mu_{t-2} - \dots - \beta_q \mu_{t-q} \parallel X_t = \nabla^d y_t$$

- Here, X_t = new time series after steps difference of d , $\alpha_1, \alpha_2, \dots, \alpha_p$ = auto regressive coefficient, $\beta_1, \beta_2, \dots, \beta_q$ = moving average coefficient, ∇ = backward difference operators, μ_t = random disturbance and y_t = time series.

FORECASTING MODEL

- It utilized for anticipating time with three fundamental model parts: pattern, irregularity, and occasions. They consolidated in the accompanying condition:

$$Z(a)=P(a)+Q(a)+R(a)+\xi(a)$$

- Here, $P(a)$ = piecewise straight or strategic development bend shaping non-intermittent time arrangement shifts, $Q(a)$ = periodical variations(e.g. week after week/yearly irregularity), $R(a)$ = effects of occasions (client gave) with unpredictable timetables, $\xi(a)$ = blunder term considering any uncommon varieties not upheld by the model.

FORECASTING MODEL

1. Piecewise straight or strategic development:

- The underlying one is called Nonlinear, Saturating Growth. It expressed in the form of the logistic growth model:

$$P(a)=C/(1+e^{(-k(a-m))})$$

- Here, C= carrying capacity which is the maximum value of the curve, k= growth rate that reflects curve steepness, and m= a variable offset. This calculated condition empowers non-straight displaying development with immersion when the development pace of significant worth decreases with its development.

FORECASTING MODEL

2. Seasonality:

- Seasonal effects $q(a)$ are approximated by the following function:

$$\boxed{Q} = \sum_{n=1}^N \left(a_n \cos \left(\frac{2\pi nt}{P} \right) + b_n \sin \left(\frac{2\pi nt}{P} \right) \right)$$

- P is the period (365.25 for yearly data and 7 for weekly data),
- Parameters $[a_1, b_1, \dots, a_N, b_N]$ need to be estimated for a given N to model seasonality.

3. Occasion:

- Occasion incur predictable shocks to a time series.

MODULES

1. Data Acquisition and Preprocessing:

- Machine learning needs two things to work, data (lots of it) and models. When acquiring the data, be sure to have enough features (aspect of data that can help for a prediction, like the surface of the house to predict its price) populated to train correctly your learning model. In general, the more data you have the better so make to come with enough rows.
- The primary data collected from the online sources remains in the raw form of statements, digits and qualitative terms. The raw data contains error, omissions and inconsistencies. It requires corrections after careful scrutinizing the completed questionnaires. The following steps are involved in the processing of primary data. A huge volume of raw data collected through field survey needs to be grouped for similar details of individual responses.
- Data Preprocessing is a technique that is used to convert the raw data into a clean data set.

MODULES

2. Feature Selection and Data Preparation:

- Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process.
- Feature engineering is the most important art in machine learning which creates the huge difference between a good model and a bad model. Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.
- The process of organizing data into groups and classes on the basis of certain characteristics is known as the classification of data. Classification helps in making comparisons among the categories of observations. It can be either according to numerical characteristics or according to attributes. So here we need to visualize the prepared data to find whether the training data contains the correct label, which is known as a target or target attribute.

MODULES

3. Model Construction and Model Training:

- The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process. The training data must contain the correct answer, which is known as a target or target attribute. The learning algorithm finds patterns in the training data that map the input data attributes to the target (the answer that you want to predict), and it outputs an ML model that captures these patterns.

MODULES

4. Model Validation and Result Analysis:

- In testing phase the model is applied to new set of data. The training and test data are two different datasets. The goal in building a machine learning model is to have the model perform well. On the training set, as well as generalize well on new data in the test set. Once the build model is tested then we will pass real time data for the prediction. Once prediction is done then we will analyzes the output to find out the crucial information.
- Never train on test data. If you are seeing surprisingly good results on your evaluation metrics, it might be a sign that you are accidentally training on the test set. For example, high accuracy might indicate that test data has leaked into the training set.
- The average delays in departure and arrival has been predicted using three basic statistical parameters: Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Coefficient of Determination (CD). The Mean Absolute Error helps to determine how close the predicted outcomes are to the consequent outcomes. It is a more natural measure of average error.

IMPLEMENTATION

Implementing:

- In this work, a business intelligent model has been developed, to classify Dataset based on a specific business structure deal with Drug recommendation using a suitable machine learning technique. The model was evaluated by a scientific approach to measure accuracy, build our model.

Import Libraries:

- **Numpy:** NumPy is a Python library used for working with arrays. It is an open source project.
- **Pandas:** Pandas is a Python library used for working with data sets.
- **Matplotlib:** Matplotlib is a low level graph plotting library in python that serves as a visualization utility. It is open source and we can use it freely.

IMPLEMENTATION

To select one state:

- Here we can choose the state to predict the active cases. Here dataset contains all state data which is related to active cases, newly active cases etc. And here we can predict the active cases of all the status Which we select.

```
In [8]: df['Name of State / UT'].unique()
Out[8]: array(['Kerala', 'Delhi', 'Telangana', 'Haryana', 'Rajasthan',
               'Uttar Pradesh', 'Tamil Nadu', 'Ladakh', 'Karnataka',
               'Maharashtra', 'Punjab', 'Jammu and Kashmir', 'Andhra Pradesh',
               'Uttarakhand', 'Odisha', 'Puducherry', 'West Bengal',
               'Chhattisgarh', 'Chandigarh', 'Gujarat', 'Himachal Pradesh',
               'Madhya Pradesh', 'Bihar', 'Manipur', 'Mizoram',
               'Andaman and Nicobar Islands', 'Goa', 'Assam', 'Jharkhand',
               'Arunachal Pradesh', 'Tripura', 'Meghalaya',
               'Dadara & Nagar Haveli', 'Sikkim', 'Nagaland'], dtype=object)
```

Name of state consider in the dataset



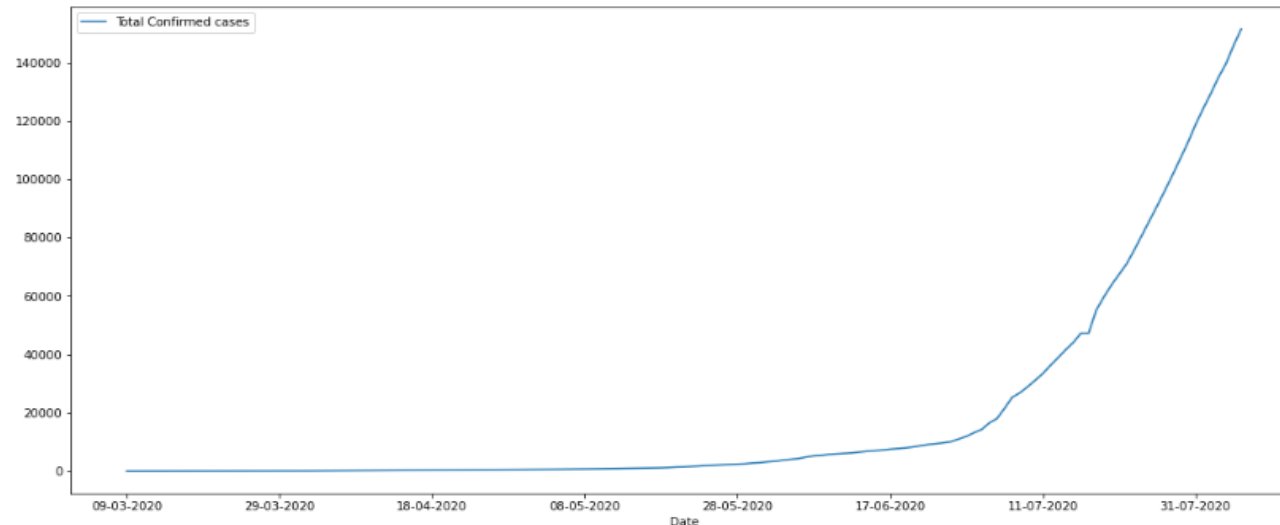
APPLICATIONS

- In Covid-19 Control system.
- In Health Care Domain for other diseases prediction also it may useful.

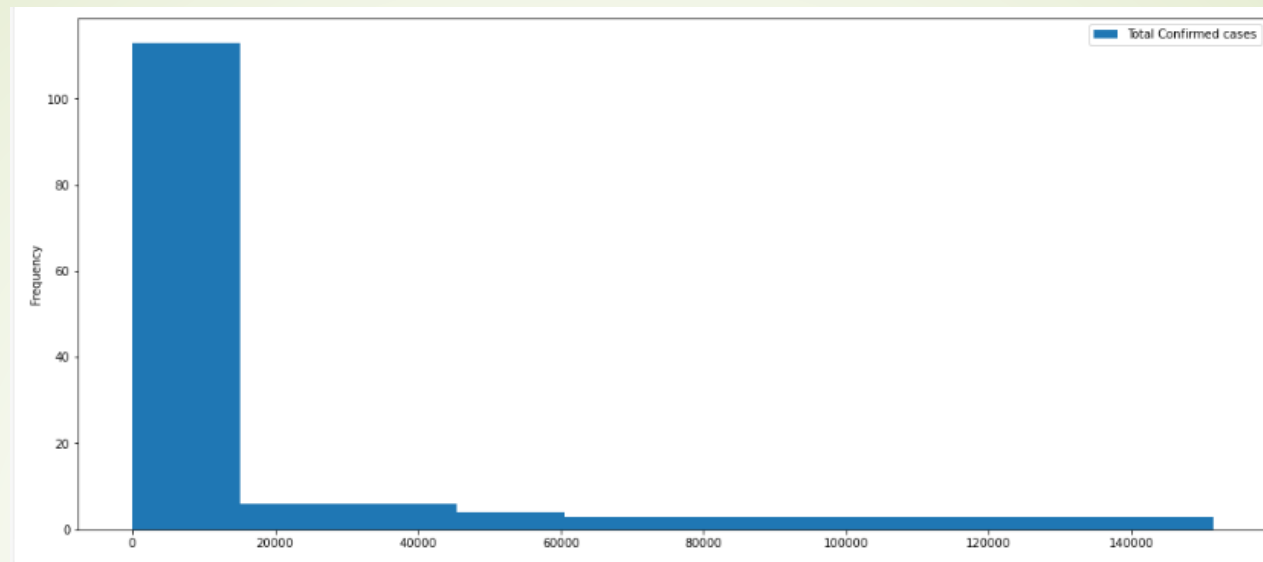
RESULT ANALYSIS

	Date	Name of State / UT	Latitude	Longitude	Total Confirmed cases	Death	Cured/Discharged/Migrated	New cases	New deaths	New recovered
74	09-03-2020	Karnataka	15.3173	75.7139	1	0	0	0	0	0
86	10-03-2020	Karnataka	15.3173	75.7139	4	0	0	3	0	0
98	11-03-2020	Karnataka	15.3173	75.7139	4	0	0	0	0	0
111	12-03-2020	Karnataka	15.3173	75.7139	4	1	0	0	0	0
124	13-03-2020	Karnataka	15.3173	75.7139	5	1	0	1	0	0

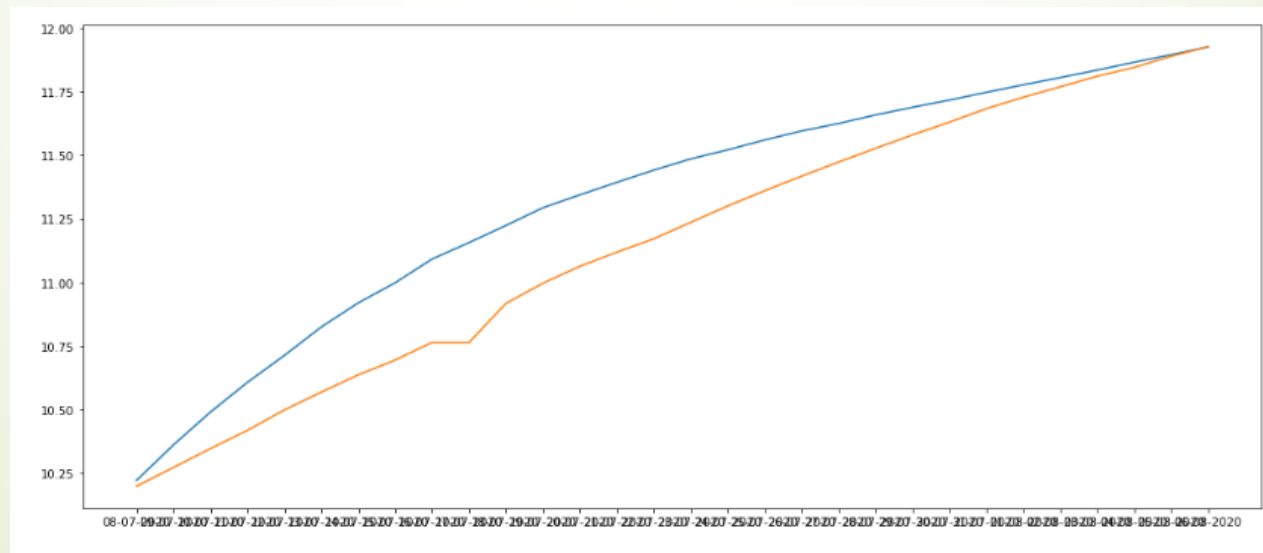
Data set



Total confirmed cases



Histogram



Final output

CONCLUSION AND FUTURE ENHANCEMENT

- In this research, we evaluate the state of COVID-19 as a result of its rapid growth in the affected cases and forecast new numbers for the following week with different models. The various forecasting techniques clearly show the increasing rate of the newly confirmed cases. With all the researched methods, we have found Facebook's Prophet model to be the perfect one to show the accurate forecasting with the least RMSE value.
- This is just the start of Project. Further approaches could use bigrams (sequences of two words), using neural networks like LSTMs (Long Short-Term Memory) to extend the relationships among growth factor.

THANK YOU

