IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 10, Issue.5, May 2022 E-ISSN: 2347-2693

Data Based Model for Predicting COVID-19 Incidence Using Data Mining

Rachana Yadav[1*], Amit Kumar Manjhwar[2]

[1,2]Computer Science and Engineering, M.I.T.S. Gwalior, India

**Corresponding Author: Ry199928@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v10i5.6573 | Available online at: www.ijcseonline.org

Received: 22/Apr/2022, Accepted: 08/May/2022, Published: 31/May/2022

Abstract— Since, covid19 is affecting many countries in the world therefore, to take the necessary steps in order to control the outbreak or incidence of covid19, which is possible if we know the outbreak or incidence of covid19 which is possible with machine learning. Therefore, in this study we are analyzing the covid19 data of India and performing EDA (exploratory data analysis) and proposing various machine learning algorithm in order to predict the outbreak of covid19. We are using various machine learning algorithms like Linear regression, Gaussian naïve bayes, Decision tree and ensemble learning like random forest, gradient boosting and then finding the best algorithm by comparing their accuracy score. With the help of best algorithm, the outbreak of covid19 to manage the health crisis in each country can be controlled by taking the essential steps.

Keywords— Covid19,Machine Learning, EDA (Exploratory data analysis), Linear Regression, Gaussian Naïve Bayes, Decision Tree, Ensemble learning, Random forest, Gradient boosting

I. INTRODUCTION

Coronaviruses (CoV) are a family of viruses that affect several members of the animal kingdom. Middle East Respiratory Syndrome CoV (MERS-CoV), Severe Acute Respiratory Syndrome CoV (SARS-CoV), and SARS-CoV2 are members of the CoV family that seldom cause infections in humans. However, the SARS-CoV2 epidemic in humans in January 2020 prompted WHO to proclaim the disease COVID-19 as a public health emergency of international concern. As of February 2022, more than 409 million cases of COVID-19 infection have been reported all over the world, with a death toll of over 5.8 million globally. 16 million new cases and 75000 deaths were reported in the second week of February 2022. The number of new COVID cases reported each week is slowly decreasing, but the number of deaths remains constant. COVID-19 Weekly Epidemiological Update, 15 February 2022, n.d.. As COVID-19 is spreading rapidly prediction models can help in health resource management and planning for prevention purposes. The prevalence of the disease requires diligent prediction of both global prevalence as well as incidence data. This will help health professionals of public bodies and government services to make informed decisions to manage this global pandemic effectively and efficiently. Novel coronavirus patients are treated with antibiotics, antivirals, chloroquine and vitamin C supplements. If the virus is not managed and controlled the global health care infrastructure will be severely over-burdened. Therefore, if we know the future value of covid incidence, then we can find effective solutions for covid 19 patients through the non-clinical approaches. Data-mining augmented intelligence and various types of AIs can help to provide the best possible means to tackle COVID-19. Predictive models can be used to improve resource management, prevent outbreaks and loss of life due to the virus. Using the data from COVID-19 patients to develop practical tools to predict the course of the disease progression in patients can improve treatment success and prevent death. The lack of specialized human resources in the health sector and sensitivity due to the value of maintaining human life using these models can reduce the cost effects of long-term treatment on hospitals and insurance.

The major objective of this research is to use data mining and several machine learning techniques to predict covid occurrence and to determine the optimal algorithm for doing so.

II. RELATED WORK

Ayyoubzadeh et al. (2020), the global spread of the covid-19 virus has been affecting almost every country in a rapid manner which has been making a prediction of the disease very important. Models are used for the prediction of the covid-19 incidence which assists in the management of healthcare resources so that effective planning can be done with the prime objective of prevention. However, predictions cannot be extremely precise but they have the capability to establish ideas which can assist in constructing accurate frameworks as well as models.

More importantly, **Ayyoubzadeh et al. (2020),** stress that lagging in predictions has been increasing the worries of people as well as demand to have solutions for prevention is rising because of public sensitivity towards health care.

This demand is rising because people are hearing their nation's epidemic news. Hence, models of data mining can help the policymakers as well as the health managers in effectively planning the resources of health care along they can even control the prevention related to the epidemic.

Ahouz & Golabpour (2021), found that prevention planning along with the allocation of essential resources is possible when information about covid-19 is stored accurately with the aim of incidence prediction. In addition to that, palliative caring, as well as treatment, are possible with the prediction of covid 19 incidence. More importantly, prediction of covid 19 prevalence together with the incidence around the globe is important so that the advantages can be availed, namely assistance to the professionals in healthcare for making important decisions as well as to ensure effective management of the disease
In addition to that, **Ahouz & Golabpour (2021),** further emphasizes that predicting situations in the present pandemic is the key towards containment of the threat. This is because the prediction of covid-19 incidence has the capability to assist decision-makers in making timely decisions about the effectiveness as well as efficient measures that can help the public at large. In addition to that, these measures include facilitating medical facilities and managing the allocation of resources as well as even sending healthcare personnel into areas that are termed as high-risk. Additionally, prediction of covid 19 incidence even helps in decision-making about the borders if they need to be closed or if they require restarting the traffic. Along with those decisions resuming or suspending services in the community can be made effectively through the prediction of covid 19 incidence.

Muhammad et al. (2021), stress that prediction of covid 19 incidence through data mining as well as machine learning assist in deeper learning via automation which helps the workers of healthcare in making an efficient decision in their clinical practices with high-performance precision.

Safdari (2021), determines that prediction of covid 19 incidence helps in physicians as well as the societies along with healthcare policymakers in gaining the knowledge about the unknown characters of new and next potential pandemics.

Moulaei et al. (2021), covid-19 incidence prediction is important because if it is unpredicted, unmanaged, or uncontrolled timely the outcome can be critical for a country's healthcare system as well as the public can even encounter serious issues. This makes it important to have predictive models for predicting the covid-19 incidence so that efficient management of healthcare resources can be done along with that prevention of death and outbreak by the virus can be even made.

Shakeel et al. (2021), stress that forecasting the numbers of covid-19 cases requires covid 19 incidence prediction which can help healthcare professionals in making essential arrangements so that physical distancing can be imposed along with lockdowns. Moreover, covid 19 incidence prediction helps in facilitating the management of healthcare through the arrangement of facilities in the hospitals as well as for equipment arrangement so that care for patients can be improved.

Zhao (2021), prediction of covid 19 incidence is important because it can help in monitoring the present caseload as well as it even projects the nature at which the covid-19 is spreading so that awareness can be created about the health of the public. In addition to that, response as well as the preparedness of an adequate number of supplies including PPE (Personal Protective Equipment) can be made possible with the prediction of covid 19 incidence so that global societies would not have to face further challenges. This preparedness can even assist the healthcare workers with satisfactoriness during the pandemic.

III. DATA MINING AND MACHINE LEARNING

Data mining is extracting useful information (databases, web pages etc), patterns and trends from large volumes of data. Data mining provides automated tools that search for associations within huge databases consisting of many different data types.

The four basic steps of knowledge discovery in databases are:

1. Select relevant data to be included in the warehouse (data cleaning).
2. Integration of results obtained by applying a set of data analysis techniques (data integration).
3. Database exploration techniques to discover patterns within the integrated database(data modeling).
4. Application of validation criteria - e.g., statistical tests - for assessing discovered patterns (outlier detection, estimation accuracy etc.)

A large amount of data needs to be processed before any interesting patterns can be discovered and understood by people; this process is called data pre-processing.

Data preprocessing includes:

1. Data cleaning - getting rid of noise (outliers), errors (typos), inconsistencies
2. Data integration - combining information that has been obtained from independent sources
3. Data reduction - reducing number of dimensions/features or compressing the data
4. Data transformation - converting scale
5. Feature selection - selecting relevant features for future use
6. Feature extraction - creating new features based on combinations of existing ones

Feature engineering is applying domain knowledge to make data more amenable for use with algorithms that learn from data.

The process of using a machine-learning algorithm to discover patterns in a set of training examples aims to generalize the patterns to new data points. That is, given a sufficiently large and representative dataset, an appropriate machine learning model can be trained from it. Each modeling or prediction method comes from a family of methods that share common characteristics; they are inductive because they build models from examples, instead of abstract axioms as deductive methods do; and they attempt to find regularities among the examples which will generalize to new ones - making them capable of extrapolation.
Machine learning techniques include:

1) **Supervised learning** - each example is labeled, and the goal is to discover a function that maps inputs to labels.
2) **Unsupervised learning** - There are no labels in this algorithm therefore, hidden structure of the unlabeled data is being identified in this algorithm.
3) **Reinforcement learning -** Each action has a consequence. The goal is to learn actions that maximize the consequences (used mostly in robotics for training an agent).

Machine learning can be supervised, unsupervised or reinforcement (depending on problem definition). Supervised machine learning algorithms build a model from labeled training examples; unsupervised machine learning algorithms build models without labeled training examples; Reinforcement machine learning agents direct their algorithmic search for good policies through trial-and-error.

Examples of machine learning techniques include: gaussian naïve bayes, decision tree, linear regression, k-nearest neighbor algorithm, support vector machine, cluster analysis and neural networks etc.

Supervised learning is a branch of machine learning concerned with prediction. In supervised learning problems given training data containing the correct answer to the problem being solved for each example in the training set, a model can be learned from this information that efficiently predicts the answer to new unseen examples.

A. ALGORITHM

Algorithms taken into consideration are -
Algorithms that are taken into consideration for this study are:

Linear Regression

Linear regression is a supervised machine learning technique which predicts the value based on some input features. Training is done at the beginning with input and output features. Linear regression is of two types.

I. Simple linear regression

In this case, only one input feature is taken into consideration with one output variable.
Y (Output Variable) = m(Parameter)X (Input Variable) + C (Bias)

II. Multiple linear regression

Multiple input variables are there in this technique with one outvariable.
Y = aX1 + bX2 + cX3 + … + C(Bias)

Gaussian Naive Bayes

This falls under the category of Naïve Bayes that supports continuous data. Therefore, having an idea of Naïve Bayes is necessary before going in depth of Gaussian Naïve Bayes technique.

Naïve Bayes are a group of classification algorithms, which is simple, yet consists of high functionality The inputs with high dimensionality find use of this classification technique. Even complex classifications can be done using this method.

Now, talking about Gaussian Naïve Bayes technique of classification, it comes handy for continuous data. When a set of continuous data is provided, it is usually assumed that each class has got values that are distributed according to Gaussian distribution. Hence, Gaussian Naïve Bayes is possible for continuous valued distributions (John, 2013) and it is named as the Gaussian or normal distribution.

For the creation of a simple model, one has to assume that the data is distributed as per the gaussian systempand contains no independent dimensions between dimensions, that is, it should have zero covariance. This can be achieved by getting the mean and standard deviation of the points of each target, which is enough to name such a type of distribution.

Hence, it is found that a slightly different approach is sought by Gaussian Naïve Bayes method of classification and it can be used efficiently for various purposes.

Decision Tree Regressor

Decision tree is a popular machine learning algorithm which is generally used for classification problems but here we are dealing with continuous data so, python allows us to use decision trees as regressor as well where it created a tree by breaking down the dataset into smaller subsets. Decision tree is generally created with several if-else conditions based on the impurity of the data. Impurity measures the homogeneity of the features and then creates conditions. Impurity is being measured using several methods:

1. ENtropy
2. Gini Index/ Gini Impurity
3. Standard deviation

Entropy

It basically describes whether all elements in the features are similar and these elements are divided to split the data based on certain conditions.

$$Entropy = -\sum_{i=1}^{n} p_i \times log(p_i)$$

Gini Index/ Gini Impurity

This measure basically helps to check whether based on various splits, which split is best to build a pure decision tree.

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2$$

Random Forest Regressor

Random forest comes under the type of supervised learning process which is generally used for classification and regression. This algorithm is also called as a type of ensemble learning because it is made up of various decision tree algorithms therefore, it builds multiple decision trees based upon the samples in classification. In regression, the trees are built as per the average of the samples.This with these algorithms we can predict results from continuous (regression) as well as discrete (classification) data but having said that, this algorithm gives better results for classification problems as compared to regression (Torgo, 1996).

To be able to understand the working of random forest methods, a concept called the ensemble technique should be known. Ensemble is basically combining two or more models. Hence, predictions are made based upon a set of models, instead of individual models. Ensemble utilizes two types of methods:

1. Bagging ensemble learning: As mentioned above, various algorithms are taken into account in ensemble learning but in bagging algorithm, first dataset is being splitted into a number of subsets and then the same multiple algorithms are used on each subset and majority voting determines the final output data. Random forest uses this method for classification.
2. Boosting ensemble learning: In this method of ensemble, the final model created gives the maximum accuracy. This is done by generating a sequential model upon combining weak learners with strong learners. Examples of this include ADA BOOST, XGBOOST.

Having understood the ensemble methods, let’s look at the random forest technique, which uses the first method, that is, bagging. The following steps are taken into consideration in random first algorithm and these are as follows::

Step 1: The dataset is split into various subsets where each element is taken randomly.

Step 2: Each subset is trained with one single algorithm, in case of random forest decision tree is used for training the subsets.

Step 3: An output is generated from all individual decision trees.

Step4: The final output is being formulated by taking the output of each predictor and then finding the output which comes larger and then taking the average for the regression which decides the final output.

Gradient Boosting Regressor

Gradient boosting is a very powerful algorithm of machine learning. In gradient boosting, many weak algorithms are combined to form a strong algorithm. Gradient boosting can be optimized by changing its learning rate. In this algorithm, we will check our model accuracy with multiple numbers of different learning rates and among those learning rates, we will find the best model for our problem.

IV. METHODOLOGY

In this section, we are proposing the flow of our methodology where the steps take into consideration for this study is as follow:

Step 1 - Start

Step 2 - Importing Data

Data Preprocessing Steps

Step 3 - Converting date into datetime64 data type

Step 4 - Converting numerical features into int64 data type.

Step 5 - Converting categorical features into numerical features.

Step 6 – Data Visualization

Modelling

Step 7 – Data Cleaning

Step 8 - Split data into training and testing

Step 9 – Calculating accuracy of model using metric like mean squared error value.

Step 10 – Stop

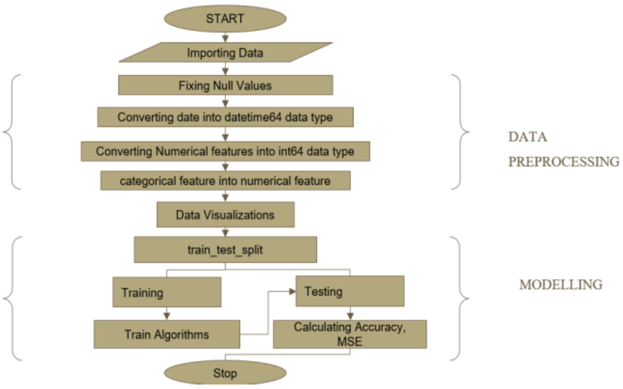

START
Importing Data
Fixing Null Values
Converting date into datetime64 data type
Converting Numerical features into int64 data type
categorical feature into numerical feature
Data Visualizations
train_test_split
Training
Testing
Train Algorithms
Calculating Accuracy, MSE
Stop
DATA PREPROCESSING
MODELLING

Fig1. Proposed process flow chart of this study

1. Data collection

For this study we have collected our data from kaggle. Kaggle allows users to use published data for exploring the insights of data and modeling to practice users and even various competitions are held on kaggle where many users specifically who are more interested in data science and machine learning take part which makes kaggle a good source for providing the data. Covid cases around the world are increasing day by day and various state and union territories data is being studied in this study. Data is collected from 2020-01-30 to 2021-08-21. Included columns are -

Date - Data by which data is collected.

Time - Time by which data is collected.

State/UnionTerrireties - Name of the state/Unionterritory.

ConfirmedIndianNationals - The observed number of confirmed Indian nationals.
ConfirmedForeignNationals - The observed number of confirmed foreign nationals.
Cured - Collected observation of number of cured peoples.
Deaths - observed number of death cases.
Confirmed - observed number of confirmed cases.

2. Data preprocessing

Data preprocessing is an important step in data mining. Since we are collecting raw data from the source, we have to convert this raw unstructured data into structured data and this process is called data preprocessing. This process is so important before using the data for the final model as machine learning models require clean data for prediction. Various processes are involved in order to convert a structured data into structured data. For example - Feature transformation which includes steps like (*missing value imputation, handling categorical features, outlier detection, feature scaling, etc)*, Feature construction or feature splitting, Feature selection, Feature extraction *(Principal component analysis)*.

2.1 Feature Transformation

Checking null values

Presence of null values can reduce the performance of machine learning. To reduce the biases of algorithms, it is good practice to remove the null columns or impute the values using central tendency. After exploring our data we find that there are no null values present in our data.

Changing data type

Collected data contain features with numerical values but the data type of these columns are object type. So first we will convert the data type of these columns by int64. Data type of change also changed from object to datetime64.

Categorical feature

Machine learning algorithms work best on numerical features, but a dataset can have both numerical as well as categorical features. Therefore, the machine learning model is not going to perform well if we don't do something about categorical features. It is best practice to encode every categorical feature into numerical features. In python we have an awesome library known as "LabelEncoder" which we can encode categorical features into numerical features. We will encode the State/Uniotterritories column as this feature is important for our machine learning model

3. Data visualizations

Data visualization is creating graphical pictures to represent the data. Data visualization helps to tell stories to the users by organizing the data into a form easier to understand. We will create various graphical representations for understanding the covid impact on India.

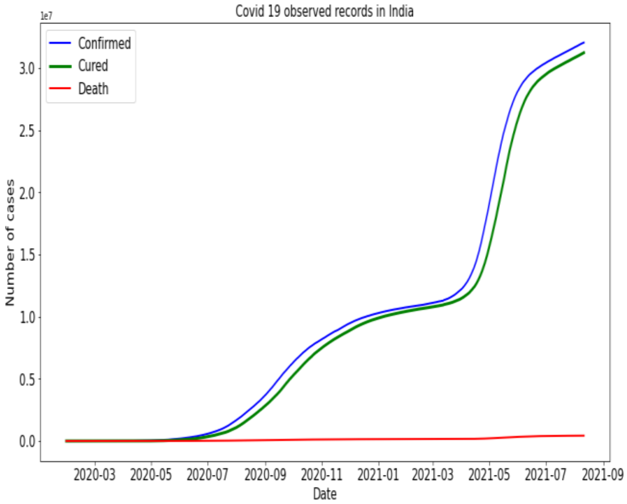


Fig. 2. Observed records in India

Above figure shows the visual effect of confirmed, cured and death cases in India from 2020-01-30 to 2021-08-21. There is an increasing trend in confirmed cases and cured cases and the death rate is considerably low as compared to confirmed and cured cases. Number of cases started to explode from June, lockdown happened in India at may 27 and because of lockdown, these were the results and India could have suffered more if lockdown had not happened at that time. **Fig. 2.** Shows the covid cases in top 15 countries. Between the interval 2020-01-30 to 2021-08-21, Maharashtra was on top in confirmed cases, while South India states like Karnataka , Kerala, Tamilnadu showed huge response of covid cases as well. What was satisfactory in covid was the cured cases were good enough as well for each country.

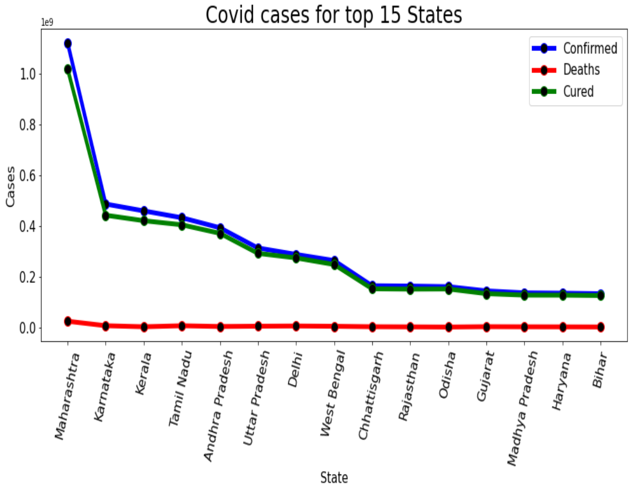


Fig. 3. Covid cases for top 15 states in India

Below **Fig. 3**. Shows the Cumulative number of confirmed Indian nationals for top 15 countries and being compared with Cumulative number confirmed foreign nationals. For Indian Nationals, Maharashtra was on top (see **Fig. 3**) while for foreign nationals, Haryana was on the top (see **Fig. 4**).

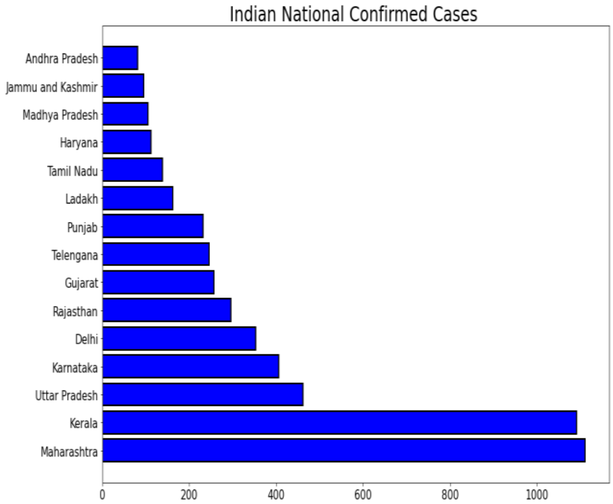


Fig. 4. Indian National confirmed cases

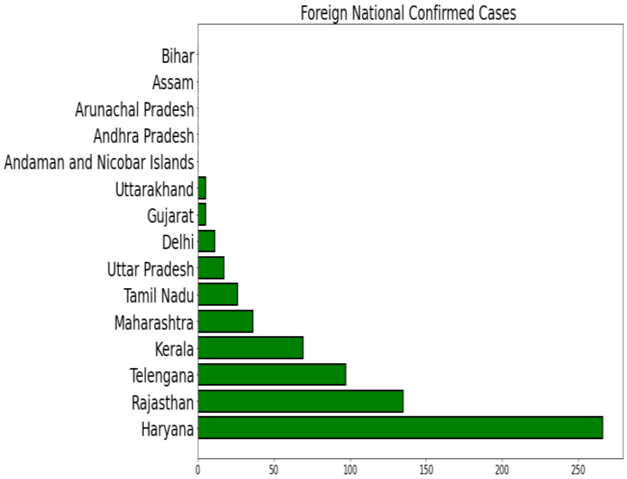


Fig. 5. Foreign National

confirmed cases

Comparison between confirmed cases, cured cases and death cases of covid are shown below in **Fig. 5**, **Fig. 6**, and **Fig. 7**. Maharashtra with sound Indian states like Karnantaka, Kerala, Tamil Nadu are showing huge results in all three cases. That's why many precautions have been taken at that time in order to reduce the impact of covid 19 in these states

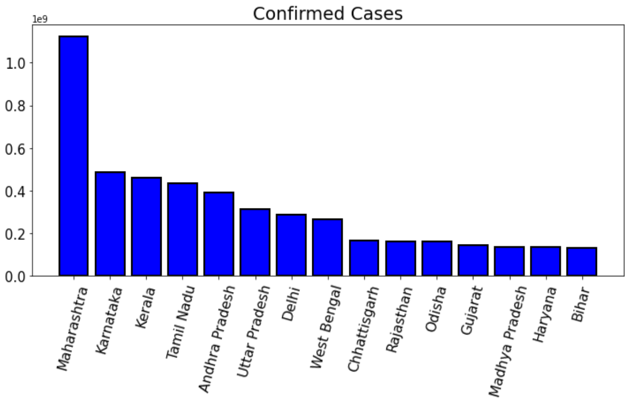


Fig. 6. Top 15 states of Confirmed cases

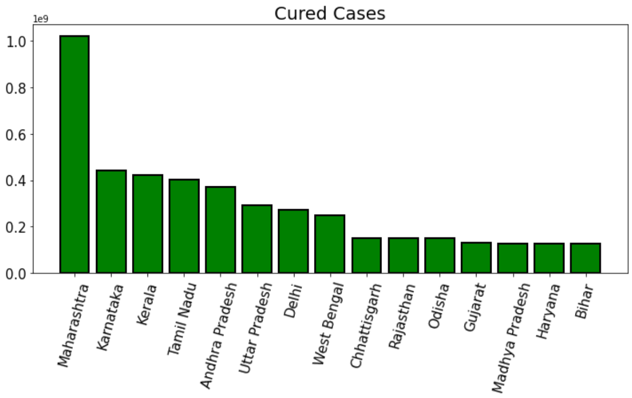


Fig. 7. Top 15 states of cured cases

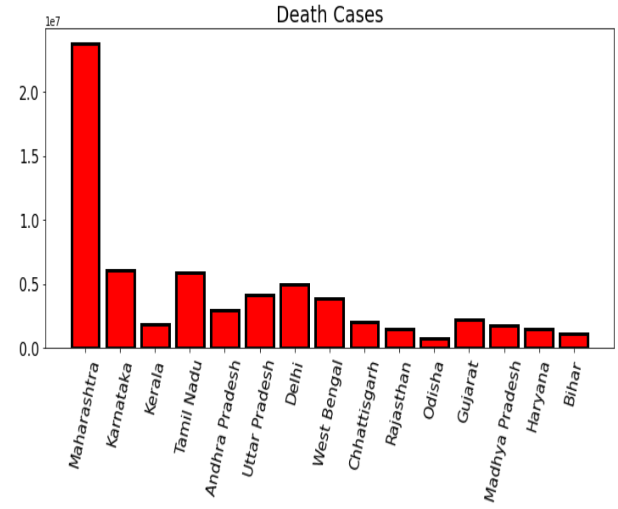


Fig 8. Top 15 states of death cases

4. Model preparation

The concept behind learning in groups is to build the foundation for a prediction model employing features from the base models of a group which are less complex, called weak learners [24]. Each step is when the group adjusts each individual learner to the difference between the observed response and the collective prediction of the learners constructed prior to. The most commonly used loss function can be described as the lesser-squares (LS) error that is [25].

The model was based on an individual set of Least-squares-increasing (LSBoost) learners who wanted to decrease MSE (or the (MSE).

4.1. Identifying dependent and independent variables.

We are using supervised machine learning algorithms, we have to train our model using dependent and independent variables.

4.2. Independent Variables

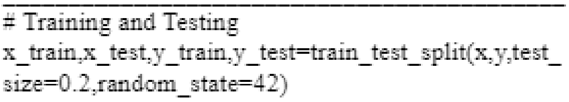
Independent variables are the input variables which do not correlate with each other. Chosen features for inputs are - (Inputs) *X* = *State/Unionterritories, dae, month, Cured, Deaths, ConfirmedIndianNational, ConfirmedForeignNational*

4.3 Dependent Variable

Dependent variables are those variables which depend upon independent or input variables. For this case, Confirmed cases will be our dependent feature.
(Output) *y = Confirmed*

4.4. Training and testing

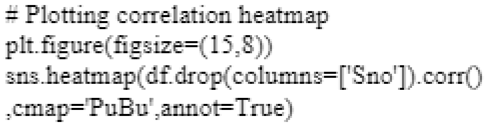
machine learning. Machine learning learns with the training data without being explicitly programmed. After training the data, we test each machine learning algorithm to check the accuracy. Accuracy is being calculated by comparing the predicted values and actual values. For training we are selecting 80% of data randomly and 20% of the remaining data is chosen for testing.



```
# Training and Testing
x_train,x_test,y_train,y_test=train_test_split(x,y,test_
size=0.2,random_state=42)
```

4.5. Correlation heatmap

Let's see the relation of every column with each other, for that we will use correlation heatmap.
Correlation basically tells how one feature is changing with another feature. Its value ranges from -1 to 1.If it is +1 then it means one feature is increasing with another feature and if the value is -1 then one feature is decreasing while the other one is increasing.



```
# Plotting correlation heatmap
plt.figure(figsize=(15,8))
sns.heatmap(df.drop(columns=['Sno']).corr()
,cmap='PuBu',annot=True)
```

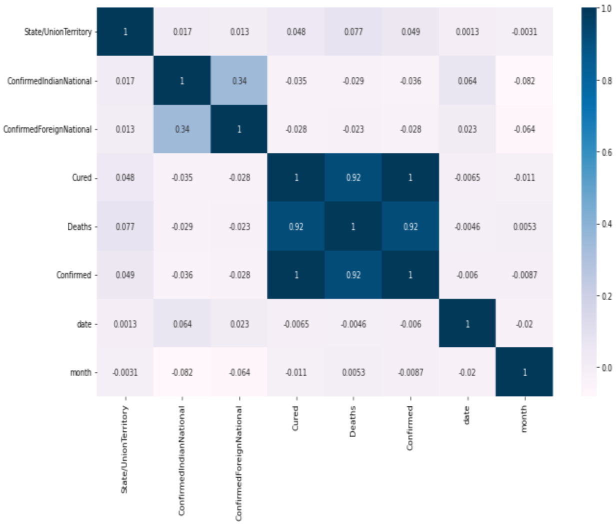


Fig 9. Correlation heatmap

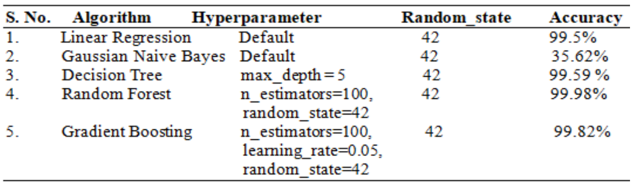
Data mining is capable of creating an accurate model, as well as providing new knowledge from the old data. The method in which data is processed, as well as the variables that are used had an effect on the acquisition of new knowledge. There are numerous methods for data mining that could be utilized to detect an outbreak. In terms of health, COVID-19 had already developed into one of the world's most severe issues. This study was designed to examine its spread in India and to predict the incidence of covid using various machine learning algorithms like Linear regression, Gaussian Naive bayes, Decision tree, Random forest and Gradient descent that predicts the spread of COVID-19 based from the data collected retrospectively. Obtained results are shown in the next sub section of this section.

V. RESULT AND DISCUSSION

Data mining is capable of creating an accurate model, as well as providing new knowledge from the old data. The method in which data is processed, as well as the variables that are used had an effect on the acquisition of new knowledge. There are numerous methods for data mining that could be utilized to detect an outbreak. In terms of health, COVID-19 had already developed into one of the world's most severe issues. This study was designed to examine its spread in India and to predict the incidence of covid using various machine learning algorithms like Linear regression, Gaussian Naive bayes, Decision tree, Random forest and Gradient descent that predicts the spread of COVID-19 based from the data collected retrospectively. Obtained results are shown in the next sub section of this section

1. Accuracy

Accuracy score is used to check the performance of classifiers. Accuracy score is calculated by actual predicted values over the total number of values. The more the accuracy score will be, the better will be the performance. Below table shows at which randomstate we obtained the accuracy. Hyperparameter for linear regression and gaussian naive bayes is set as default, for decision tree, max_depth is chosen as 5 while for the random forest, a total 100 number of estimators are taken into consideration and for gradient descent, 100 estimators with 0.05 learning rate is chosen with the random state of 42



S. No.	Algorithm	Hyperparameter	Random_state	Accuracy
1.	Linear Regression	Default	42	99.5%
2.	Gaussian Naive Bayes	Default	42	35.62%
3.	Decision Tree	max_depth = 5	42	99.59 %

4.	Random Forest	n_estimators=100,
		random_state=42
			42	99.98%

5.	Gradient Boosting	n_estimators=100,
		learning_rate=0.05,
		random_state=42
			42	99.82%

Fig10. Accuracy score for each algorithm

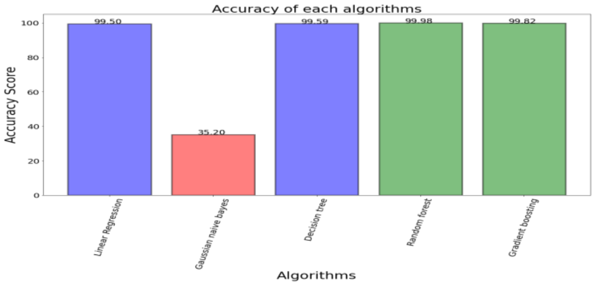


Fig 11. Accuracy score for each algorithm

VI. CONCLUSION AND FUTURE SCOPE

The above discussion brings us to the conclusion that machine learning and data mining are crucially effective in the domain of covid incidence prediction. The above ML model uses all techniques (Linear regression, Decision tree, Gaussian naive bayes, Random forest, and Gradient boosting). Out of these, the Random forest algorithm turned out to be most accurate with an accuracy of around 99.98%. The model predicted many aspects regarding the incidence of covid among the population. The 21st century is the 'era of technology' where efficiency has a special link to computerization. AIs and MLS form the future of computers taking over complex processes and forming new ways and methods to predict what might happen and prepare in advance. The degree and accuracy of predictions vary from time to time due to government restrictions, public practices like masks, sanitizers, etc. ML techniques are being evolved continuously by data scientists to account for these changing parameters and create a system of adapted computers to account for them and predict accurately.

On the other hand, data mining is also a supplementary field to ML which if used in combination, like in the above model, can bring out efficient outcomes like Random forest techniques with 99% accuracy. Data mining cleans and classifies the data, analyses the patterns whereas machine learning teaches the systems how to interpret and predict them. Therefore, data mining and machine learning are the best we have to predict covid incidence. Continuous developments and techniques are in development and the coming era might see the infusion of ML and AI to predict life and better handling of contingencies.

REFERENCES

[1] Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. **91:157–160, 2020.**

[2] Zhao Y, Shou M, Wang Z. Prediction of the Number of Patients Infected with COVID-19 Based on Rolling Grey Verhulst Models. Int J Environ Res Public Health. **17(12):4582, 2020.**

[3] Leeuwenberg AM, Schuit E. Prediction models for COVID-19 clinical decision making. The Lancet Digital Health. **2:e496–e497, 2020.**

[4] M. I. Jordan1 and T. M. Mitchell, Machine learning: Trends, perspectives, and prospects (sciencemag.org) **2019**.

[5] Godio A, Pace F, Vergnano A. SEIR Modeling of the Italian Epidemic of SARS-CoV-2 Using Computational Swarm Intelligence. Int J Environ Res Public Health. **17(10):3535, 2020.**

[6] Sharma, V.K., Nigam, U.: Modelling and forecasting of COVID-19 growth curve in India. Trans. Indian Nat. Acad. Eng. **5, 697–710, 2020.**

[7] Saeed, S., Humayun, M.: Quantitative analysis of COVID-19 patients: a preliminary statistical result of deep learning artificial intelligence framework. In: Book Series: ICT Solutions for Improving Smart Communities in Asia, IGI Gobal, **pp. 218–242, 2021.**

[8] Ghosal, S., Sengupta, S., Majumder, M., Sinha, B.: Linear regression analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020). Diabetes Metab. Syndr. Clin. Res. Rev. 14(January), **311–315, 2020.**

[9] Rushing J, Ramachandran R, Nair U, Graves S, Welch R, Lin H. ADaM: a data mining toolkit for scientists and engineers. Computers & Geosciences **31(5):607-618, Jun. 2005.**

[10] Dangare CS, Apte SS. Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. IJCA 30;**47(10):44-48, Jun 2012.**

[11] R.S. Yadav, Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. Int. J. Inf. Technol., **pp. 1-10, 2020.**

[12] S.F. Ardabili, A. Mosavi, P. Ghamisi, F. Ferdinand, A.R. Varkonyi-Koczy, U. Reuter, et al., COVID-19 outbreak prediction with machine learning, medRxiv (2020), 2020.04.17.20070094

[13] P. Bedi, S. Dhiman, P. Gole, N. Gupta, V. Jindal, Prediction of COVID-19 trend in India and its four worst-affected states using modified SEIRD and LSTM models, SN Comput. Sci., **2(3), p. 224, 2021.**

[14] T. Chakraborty, A. Bhattacharyya, M. Pattnaik, Theta autoregressive neural network model for COVID-19 outbreak predictions, medRxiv (2020), 2020.10.01.20205021

[15] A. Chatterjee, M.W. Gerdes, S.G. Martinez, Statistical explorations and univariate timeseries analysis on COVID-19 datasets to understand the trend of disease spreading and death, Sensors, **20(11), 2020.**

[16] N.P. Dharani, P. Bojja, P. Raja Kumari, Evaluation of performance of an LR and SVR models to predict COVID-19 pandemic, Mater. Today Proc. **2021**.

[17] S.W. Doe, T.R. Seekins, D. Fitzpatrick, D. Blanchard, S.Y. Sekeh, Adaptive County Level COVID-19 Forecast Models: Analysis and Improvement (**2020**), arXiv preprint arXiv:200612617

[18] S. Ganiny, O. Nisar, Mathematical modeling and a month ahead forecast of the coronavirus disease 2019 (COVID-19) pandemic: an Indian scenario, medRxiv (2020), 2020.09.10.20192195

[19] K.K.A. Ghany, H.M. Zawbaa, H.M. Sabri, COVID-19 prediction using LSTM algorithm: GCC case study, Inform. Med. Unlocked, **23, p. 100566, 2021.**

[20] N.M. Ghazaly, M.A. Abdel-Fattah, A.A. Abd El-Aziz, Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network, Int. J. Adv. Sci. Technol., **29**, **5** Special Issue, **pp. 1831-1849, 2020.**

[21] Q. Guo, Z. He, Prediction of the confirmed cases and deaths of global COVID-19 using artificial intelligence, Environ. Sci. Pollut. Res. Int., **28 (9), pp. 11672-11682, 2021.**

[22] K.T. Hasan, M.M. Rahman, M.M. Ahmmed, A.A. Chowdhury, M.K. Islam, 4P model for dynamic prediction of COVID-19: a statistical and machine learning approach, Cognit. Comput., **pp. 1-14, 2021.**

[23] B.B. Hazarika, D. Gupta Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks, Appl. Soft Comput., **96, p. 106626, 2020.**

[24] M. Hawas, Generated time-series prediction data of COVID-19's daily infections in Brazil by using recurrent neural networks, Data Brief, **p.106175, 2020.**

[25] B. Heni, COVID-19, Bacille Calmette-Guerin (BCG) and tuberculosis: cases and recovery previsions with deep learning sequence prediction, Ingénierie Des. Systèmes Inf., **25 (2), pp. 165-172, 2020.**

[26] A.-E.E. Hridoy, M. Naim, N.U. Emon, I.H. Tipo, S. Alam, A. Al Mamun, et al., Forecasting COVID-19 dynamics and endpoint in Bangladesh: a data-driven approach, medRxiv (2020), 2020.06.26.20140905

AUTHORS PROFILE

Miss Rachana Yadav completed B.E in computer science engineering from MITS Gwalior india in 2020. & pursuing m.tech CSE from MITS Gwalior.

Mr. Amit Kumar Manjhvar completed B.E. in Computer Engineering from SGSITS Indore, India in 2007 & M.Tech in Software System from SATI Vidisha in 2012. He is currently working at the Department of Information Technology in Madhav Institute of Technology & Science Gwalior, & his research experience of 5 years. He has published 21 research papers in different refereed journals & 03 papers presented in different IEEE conferences. He is member of IETE & IAENG societies.