# A COVID-19 PREDICTION BASED ON MACHINE LEARNING ALGORITHMS – A LITERATURE REVIEW

**Dr. AGUSTHIYAR R [1], S. SARANYA [2 ]**

[1] Prof & Head, Department of Computer Applications, SRM Institute of Science and Technology, Ramapuram, Chennai.

[2] Research Scholar - Part time(External),Department of Computer Applications, SRM Institute of Science and Technology, Ramapuram, Chennai.

**E-mail**: [1]hod.dca.rmp@srmist.edu.in, [2]saranyasivapragasam88@gmail.com .

## *ABSTRACT:*

*The SAR.CoV2 disease 2019(covid-19) pandemic affected many countries of the world. Actually, almost all the countries presented Covid-19 positive cases and governments are choosing different health policies to stop the infection and many programs are conducted with aware common people. Then number of positive cases increasing rapidly everyday around the world. This paper is going to propose a prediction on what basis common people getting affected and how to reduce the spreading of disease. Machine learning algorithms have been used in all the fields in predicting. Especially in medicine and enriches the applications of machine learning which are accurate and robust in selecting attributes. Here we investigate some of the machine learning models namely Decision Tree, Random Forest, Adaboost and Logistic Regression to predict accuracy of getting affected. In our experiment shows prediction result accuracies 70.1, 70.3, 67.9, 70.6 respectively.*

***Keywords:*** *SARS, Decision tree, Random Forest, Logistic Regression, AdaBoost, Renal Chronic, Intubed.*

## 1. INTRODUCTION:

A SAR. Cov 2 was identified 2019 December in Wuhan (China)[1]. Now a days the whole world faces Covid-19 pandemic. It has very globally spread that affects the people and leads to death and also all countries are economically down. In addition the world has faced many other diseases like (SARS) Severe Acute Respiratory Syndrome, Middle east Respiratory Syndrome related(MERS) likewise now Corono virus[2] . So that it really become important to understand the Characteristics and impact of this disease and to predict the further spread of this disease among the people. So far several methods are used for forecasting the disease impacts. During the year 2022 May 53crore people get affected by covid-19 and the death rate was 62.9 Lakhs.

For instance, some journals showed that men are more easily affected to the disease than women[3], whereas in others no significant difference in gender-related[3]. Zhang et al.[3] claimed that children aged less than 14 years were less affected to the disease compared to age between 15 to 64 years, and elderly people aged more than 65 years were more affected.

Although, Bi et al.[4] reported that infection rates in children were no lower than the population. Later, some articles reported an increased risk of testing positive cases for Covid-19 among Black, Hispanic, and Asian patients compared to white patients[5], yet Pan et al.[6] found no ethnicity related results from China. From a strategic and healthcare management perspective, a susceptibility classification model of individuals is of great importance to governments and decision makers, as it facilitates imposing personalized protective strategies, which can save lives and minimize social and economic consequences. Further, such models can be useful for authority when vaccination becomes available to decide who is more susceptible and hence should be vaccinated first.

Many models are used to predict the influenced of different disease like Diabetics, Cancer, Asthma, etc[3]. One of the best models is machine learning. Machine learning algorithm has been using much prediction application. Trained models are then used to predict the covid-19 patients are at the greater risk of death. Machine learning methods can also helps to distinguish covid-19 infection from the communication.

This work aims to predict the use of ML models to forecast the chances of getting affected by Covid-19. Mexican Datasets are used for prediction. 5,66,602 records with 23 attributes found in Mexican Dataset. We use well know ML models namely Decision Tree(DT), Random Forest(RF), Adaboost(AB), Logistic Regression(LR). These models achieved the moderate results.

## 2. RELATED WORK:

*Analysis and Prediction of Covid-19 trajectory: A Machine Learning Approach:* This paper builds predictive models that can predict the number of positive cases with higher accuracy. Regression-based, Decision tree based and Random forest based models have been built on the data from China and are Validated on Indian sample[7].

*Prediction and Analysis of Data mining Models for Student underlying issues during Novel Coronovirus(covid-19):* This paper proposed the student problem considered for research to find out the desire solution to reduce the depression of students. Data collected from 734 students from Bangalore through Questioner. Using KNN, Decision tree, SVM, Random forest, Logistic recursion he reached the maximum of 94.85% accuracy [8].

*Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models:* This work has tried to forecast the situation of India amid the outbreak of global pandemic COVID-19.The outbreak of COVID-19 in India was analyzed using approximate mathematical modeling and then by an enhanced version of the SIR epidemic model. Instead of assuming values of parameters for SIRD modelling, sophisticated methods were applied implementing curve fitting on existing data and linking mortality rate $\delta$ to the recovery rate $\gamma$, in order to obtain more realistic results. The initial value of susceptible patients was based on Zone bifurcation implemented by the Government of India to get optimal results[9].

*Analysis And Study Of K-Means Clustering Algorithm:* Study of this paper describes the behavior of Kmeans algorithm. This paper we have try to overcome the limitations of K-means algorithm by proposed algorithm. That's why the proposed clustering concept comes into picture to provide quick and efficient clustering technique on large data set. In this paper

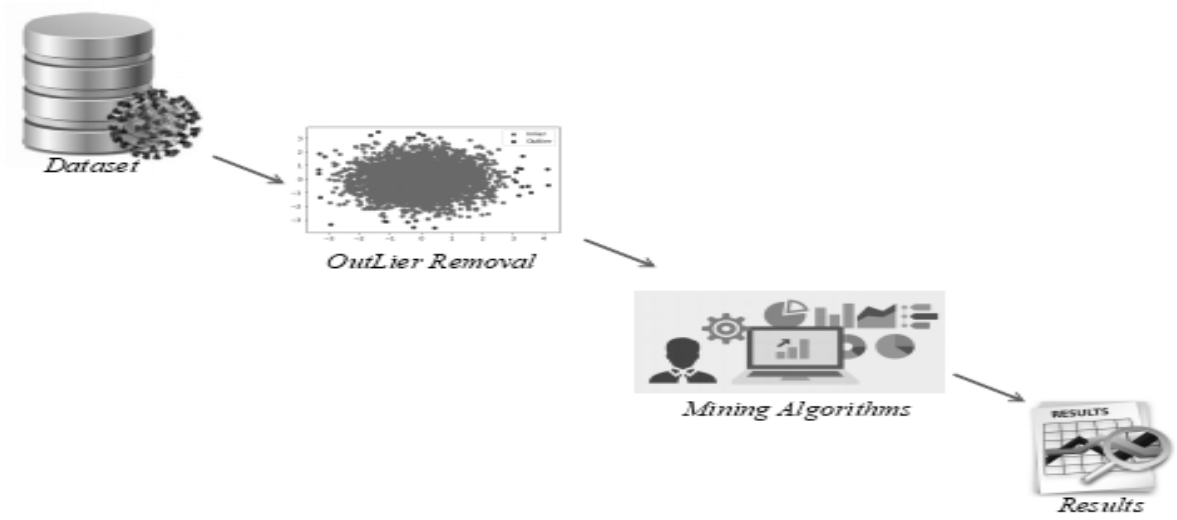
performance evaluation is done for proposed algorithm using Hospital Diabetic Patient Dataset[10]

*Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis And Treatment:* We discussed that they are desirable because of their potential for creating a workspace while AI Experts and physicians could work side by side. However, It should be noted while AI speeds up them conquer COVID-19, real experiments should happen because a full understanding of advantages and limitations of AI-based Methods for COVID-19 is yet to be achieved, and novel Approaches have to be in place for problems of this level of complexity. Succeeding in the combat against COVID-19. Toward its eventual demise is highly dependent on building an arsenal of platforms, methods, approaches, and tools that Converge to achieve the sought goals and realizes saving more lives[11]

*Cloud based framework for diagnosis of diabetes mellitus using K-means clustering:* In this work, we have compared both k-means and hierarchical clustering algorithm based on performance, runtime and quality. From this analysis, k-means clustering algorithm is good for handling large data set in cloud computing platform and it more efficient when comparing to hierarchical clustering algorithm. We mainly analyzed the diabetes dataset using hadoop framework by considering the attributes such as age, gender and family history. The results found that, the age group under 45–64 are more diagnosed with diabetes[12].

# 3 METHODOLOGY:

**Figure 1: Data Processing**

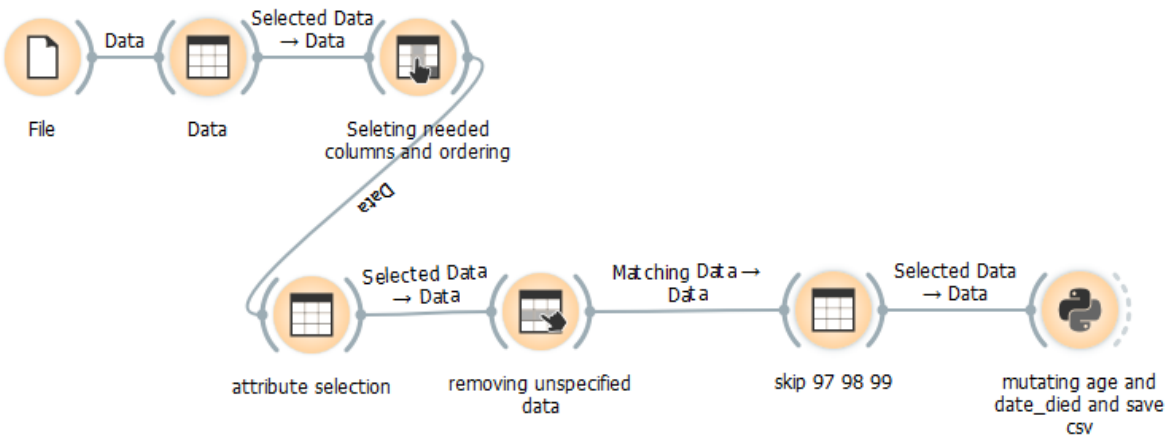



## 3.1 DATASET DESCRIPTION

We use a real dataset of Mexican people which was received from Kaggle.com. The dataset has 5,66,602 records and it consist of test records of all people who were admitted in the hospital. The data includes patient id, sex, patient type, entry date, date of symptoms, intubed, pneumonia, age, pregnancy, diabetics, asthma, date of died, Chronic Obstructive pulmonary disease(COPD), immunity suppression, hyper tension, other diseases, cardio vascular, obesity, renal chronic, tobacco, contact with other covid, result, icu.

### 3.1.1 DATA PREPROCESSING:

**Figure 2: Data PreProcessing**




Further many data cleaning and pre processing procedures were followed. Irrelevant confirmatory results (result-3) are removed. We have selected only 16 out of 23 attributes. The removed attributes are patient id, patient type, entry date, date of symptoms, other diseases, contact with other covid, renal chronic. The reason for removing those attributes are hospital information (patient id, patient type, entry date, date of symptoms) not related to covid (other diseases, renal chronic ) and spread(contact with other covid) and also removal of unspecified data(97,98,99) from all the above attributes except pregnancy. The data in the date of died, age are mutated.

### 3.1.2 DATASET DESCRIPTION- AFTER PREPROCESSING:

After pre processing the Mexican dataset we got 1,06,310 data out of 5,66,602 . The detailed information about data based on attributes are stated below:-

**Table 1: Attributes List**

| S. NO | ATTRIBUTES | YES | NO | S. NO | ATTRIBUTES | | YES | NO |
|---|---|---|---|---|---|---|---|---|
| 1 | Pneumonia | 65252 | 41058 | 9 | Tobacco | | 9562 | 96748 |
| 2 | Diabetics | 31164 | 75146 | 10 | Date of died | | 30300 | 76010 |
| 3 | COPD | 4966 | 101344 | 11 | ICU | | 8812 | 97498 |
| 4 | Asthma | 2809 | 103501 | 12 | Pregnancy | | 963 | 105347 |
| 5 | Immune suppression | 4239 | 102071 | 13 | Sex | | 42540 | 63770 |
| 6 | Hyper tension | 35325 | 70985 | 14 | Intubed | | 8844 | 97466 |
| 7 | Cardio vascular | 5605 | 100705 | 15 | Result | | 67488 | 38822 |
| 8 | Obesity | 22280 | 84030 | 16 | Age | 5489(<18) | 27294 | 72897(>45) |

## 4 CLASSIFICATION MODELS:

We have applied 4 well popular ML models namely, DT, RF, AB, LR to predict getting infected by Covid-19 data.

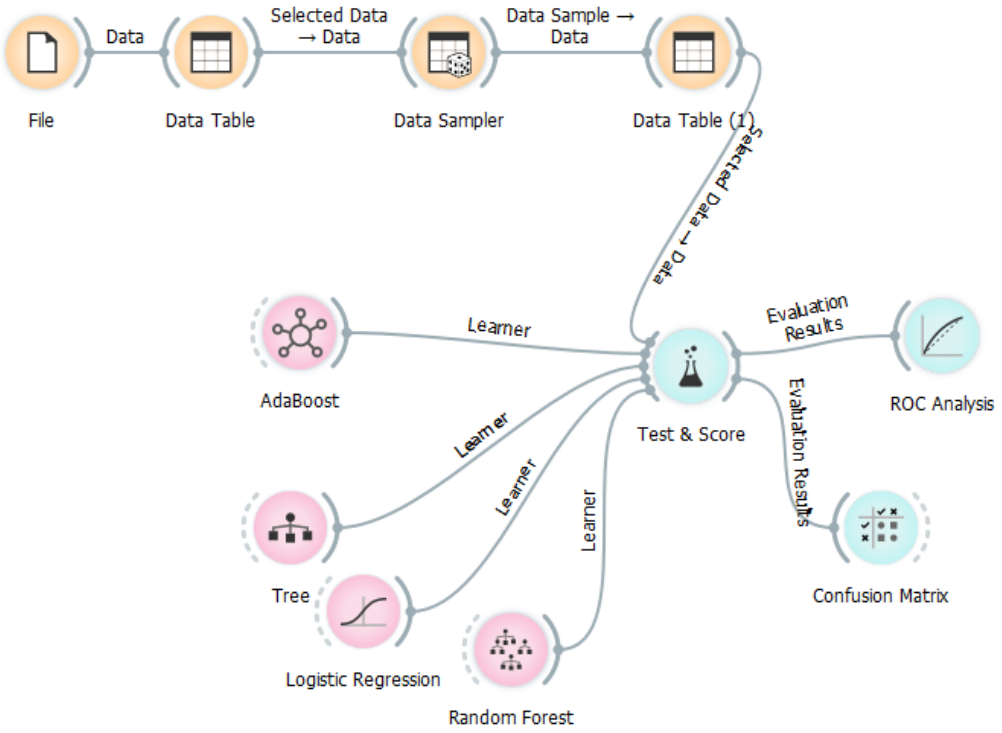
DT algorithms are a non-parametric supervised learning model used for regression and classification. The aim of the model is to predict the value of target variable by learning simple decision rules inferred from the data features [13].

RF is an sample learning method for classification, regression and other task that operates by constructing a multitude of decision trees at training time [13].

AB is one of most promising algorithm and it is statistical classification algorithm. It used in conjunction with other types of learning algorithm to improvise the result [13].

LR is also a statistical model used to find the probability of an event taking place by having the log-odds for the linear combination of more than one independent variable [13].

**Figure 3: Classification Model Tools**

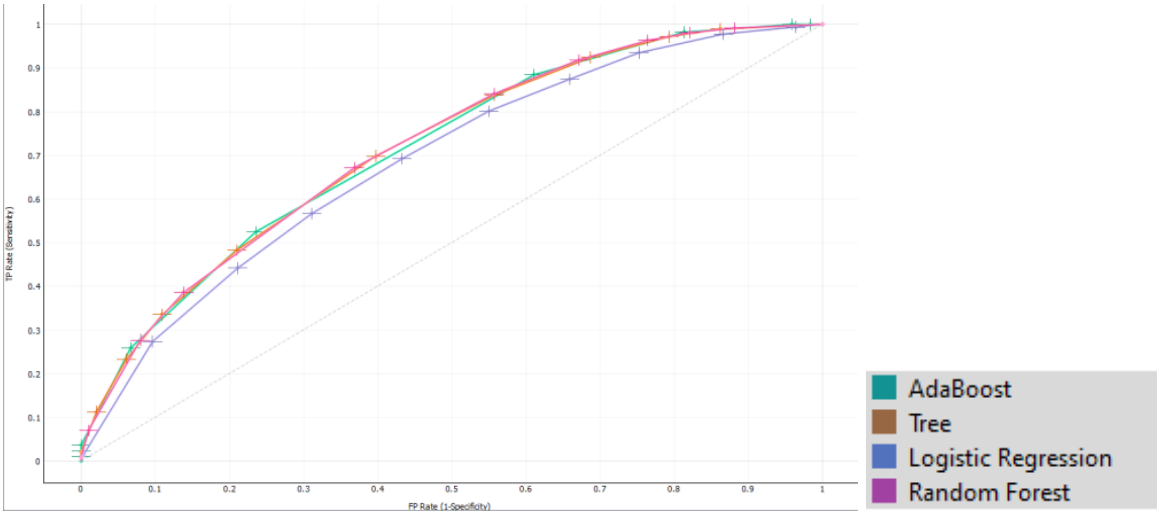



## 5 RESULTS:

Based on the above-mentioned algorithm the result of our experiments. The table shows the performances of ML models with respect to CA, Precision, Recall and F1 score. The best performance score is bolded.

### 5.1 ML MODEL PERFORMANCE SCORE:

| PERFORMANCE MATRIC | DT | RF | AB | LR |
|---|---|---|---|---|
| CA | 70.1 | 70.3 | **70.6** | 67.9 |
| PRECISION | 70 | 70.1 | **70.4** | 66.5 |
| RECALL | 70.1 | 70.3 | **70.6** | 67.9 |
| F1 | 66.6 | 67.1 | **67.4** | 65.4 |




Overall, the ML models shows the good performance, where AB, RF seems to best in performance among others with our dataset.

## CONCLUSION AND FUTURE WORK:

The work aimed to predict an individual getting infected to covid-19 based on demographic data including age, gender and disease infected using ML models. The favourable results were obtained. The results of a present work enhance our understanding of covid -19. This work will be helpful to take necessary action against this pandemic. In future we will develop new algorithm to achieve best findings.

## REFEENCE:

1. Amir Ahmad , Ourooj Safi , Sharaf Malebary , Sami Alesawi , and Entisar Alkayal ," Decision Tree Ensembles to Predict Coronavirus Disease 2019 Infection: A Comparative Study", Hindawi Complexity Volume 2021.
2. Othman Istaiteh , Tala Owais , Nailah Al-Madi, Saleh Abu-Soud, "Machine Learning Approaches for COVID-19 Forecasting", 2020 International Conference on Intelligent Data Science Technologies and Applications.
3. Alhanoof Althniana, Afnan Abou Elwafa, Nourah Alobou, Hend Alrasheed , Heba Kurdi, "Prediction of COVID-19 Individual Susceptibility using Demographic Data: A Case Study on Saudi Arabia", ScienceDirect ELSEVIER , November 2-5, 2020.
4. Bi, Q., Wu, Y., Mei, S., Ye, C., Zou, X., Zhang, Z., Liu, X., Wei, L., Truelove, S.A., Zhang, T. and Gao, W. (2020). "Epidemiology and Transmission of COVID-19 in Shenzhen China: Analysis of 391 cases and 1,286 of their close contacts". MedRxiv.
5. de Lusignan, S., Dorward, J., Correa, A., Jones, N., Akinyemi, O., Amirthalingam, G., Andrews, N., Byford, R., Dabrera, G., Elliot, A. and Ellis, J. (2020). "Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study". The Lancet Infectious Diseases.

6. Pan, D., Sze, S., Minhas, J.S., Bangash, M.N., Pareek, N., Divall, P., Williams, C.M., Oggioni, M.R., Squire, I.B., Nellums, L.B. and Hanif, W. (2020). "The impact of ethnicity on clinical outcomes in COVID-19: A systematic review". EClinicalMedicine 23.
7. Ritanjali Majhi,Rahul ,Thangeda,Renu Prasad Sugasi,Niraj Kumar, "Analysis and Prediction of Covid-19 trajectory: A Machine Learning Approach", Wiley October 2020 .
8. Krishna, Praveen Kumar V, "Prediction and Analysis of Data mining Models for Student underlying issues during Novel Coronovirus(covid-19)", 2020 International Journal of Engineering Research & Technology (IJERT) .
9. Siddharth Singh,Pivush Raj, Raman Kumar,Rishu Chaujar, "Prediction and forecast for COVID-19 Outbreak in India based on Enhanced Epidemiological Models ",IEEE: Proceedings of the Second International Conference on Inventive Research in Computing Applications (ICIRCA-2020)
10. Sudhir Singh and Nasib Singh Gill, "Analysis And Study Of K-Means Clustering Algorithm ", International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 7, July - 2013
11. Alilalbakhsh, Mohammad(Behdad)Jamshid ,"Artificial Intelligence and COVID-19: Deep Learning Approaches for Diagnosis And Treatment ",IEEE/ May 2020
12. P. Mohamed shakeel, S. Baskar , V. R. Sarma dhulipala and Mustafa Musa Jaber ," Cloud based framework for diagnosis of diabetes mellitus using K-means clustering ",IEEE/2018.
13. Wikipedia "https://en.wikipedia.org/wiki/"