IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117101

Covid-19 Analysis using Machine Learning for Indian states

Bhavya Sharma[1], Garima Choudhary[2], Neeta Verma[3]

[1, 2] B. Tech student, Inderprastha Engineering College (IPEC), Ghaziabad, UP

[3] Professor, Inderprastha Engineering College (IPEC), Ghaziabad, UP

Abstract: During this global pandemic urgency, scientists, healthcare specialists and researchers continue to look for the high quality feasible treatment of COVID-19 disease. The existence of Artificial Intelligence (AI) and Machine Learning (ML) majorly contributed in finding solutions for ongoing novel Coronavirus pandemic outbreak.
Various machine learning algorithms like K-means Algorithm, Support Vector Machine, Decision Tree, have proved their importance in forecasting, predicting, analyzing, and visualizing spread and cautious effects of coronavirus in India.

PROPOSED MODEL

For our project on Covid-19 analysis using machine learning in Indian states. We will be using different machine learning algorithms such as Artificial
Intelligence (AI) and Supervised Machine Learning (ML) Algorithms like Decision
Tree, KNN (K-nearest neighbors), Naive
Bayes theorem, Support Vector Machine (SVM) algorithm .
Further we will be calculating the accuracy of the model of COVID-19 disease outbreak using various Artificial Intelligence (AI) and Supervised Machine Learning (ML).
Algorithms with highest accuracy will be selected and used for the prediction.
The accuracy evaluation metric shows the percentage of the dataset instances correctly predicted by the model developed by the machine learning algorithm.

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fn + fp}.$$

tp = True Positive - It's an outcome where the model predicts correctly the positive class.
tn = True Negative - It's an outcome where the model correctly predicts the negative class.
fn = False Negative - It's an outcome where the model incorrectly predicts the negative class.
fp = False Positive - It's an outcome where the model incorrectly predicts the positive class.

Keywords: COVID-19, Machine Learning (ML), Artificial Intelligence (AI), Deep Learning (DL), SARS-CoV-2, Pandemic, World Health Organization (WHO), Supervised Learning Algorithms, Regression, RAT, RT/PCR, Classification, Prediction, Decision Tree, Support Vector Machine (SVM), K- nearest neighbors, Naive Bayes.

DATASET USED

Dataset : A collection of data pieces that can be treated by a computer as a single unit for analytic and prediction.
Here in our project we have selected Indian states data which are majorly affected by Covid-19 cases . On which we have applied several ML algorithms such as KNN, Naive Bayes, Decision Tree.
Here we have separated our Dataset on
basis of state wise positivity rates such as :-
Case 1:- Positivity Rate < 5%
Case 2:-Positivity Rate >=5% to <= 10%
Case 3:-Positivity Rate >= 10%

Attributes we have used here are :
1. % Contribution of testing by RAT
2. % Contribution of testing byRT/PCR
3. Positivity

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022
DOI: 10.17148/IJARCCE.2022.117101

Here after evaluating each state wise positivity rate, we have separately analyzed % Contribution of RAT, RT/PCR on positivity of different states, after that we have applied ML algorithms in order to obtain accuracy in Covid-19 positive cases.

INTRODUCTION

As per the current situation caused due to COVID- 19 outbreak, it became necessary to find a solution for prediction of disease by observing the trend of its spread. Machine learning (ML) has proved to be an important field of study to address complicated problems in the real world. Several standard algorithms of ML were used in different application domains including stock market prediction, disease analysis and prediction, weather forecasting. Various models of regression, classification technique and neural network are broadly applicable to forecast future prediction of disease spread, rate of recovery and death cause in patients. Prediction focuses on the decision making process in order to direct earlier interventions in the efficient management of the disease.

The use of Machine Learning models have increased in recent years due to the pandemic. Machine learning is widely used in today's world because of increasing computational power and the availability of large datasets on open-source tools.

The main aim of the research is to increase prediction accuracy for the spread of new coronaviruses known as the SARS-CoV-2. At present the threat to human life is COVID-19 Worldwide It has a spectrum from severe acute respiratory conditions and organ collapse leading to death.

Around a million people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people tested positive every coming day from countries across the world. The virus spreads primarily through close physical contacts or by touching the contaminated surfaces. The most challenging aspect of its spread is that a person can possess the virus for many days without even showing any symptoms of the disease. The causes of its spread and considering its danger, almost all the countries have declared strict lockdowns in affected areas and regions

To contribute to the current human crisis our attempt in this study is to develop a system with more accurate prediction or increasing the accuracy of prediction of COVID-19 spread in Indian states.

LITERATURE REVIEW

From our literary survey of various journals highlighting the Covid-19 outbreak causes in different regions of the world we have found the following key points :

The model predicted suspicious casualties would be minimized around 1000 days and infected death people around 300 days. [1] "A Novel Parametric Model for the Prediction and Analysis of the COVID-19 Casualties" has proposed the SIR model to analyze and predict Covid-19 casualties. According to LR and LASSO, the death rate will increase and recovery rate will slow down. LR performs better than LASSO and SVM in predicting death rate and casualties as proposed in [2] "COVID-19 Future Forecasting Using Supervised Machine Learning Model".

The paper [3] "Predicting and Analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR model" was proposed with an objective to fit and predict epidemic situations in China where it was found that none of the three models outperformed.

To employ data-driven mathematically proven model for prediction SEIR/SIR model was used where it was found that prediction model that fall under the scope of SEIR/SIR, agent based and curve fitting approaches barely include aforementioned fact as proposed in [4] "COVID-19 Prediction Models and Unexploited Data". As proposed in [5] "COVID-19 Prediction and Detection using Deep Learning" CNN and LSTM was used to detect the Covid-19 recoveries and deaths using AI.

SVM, Deep CNN and Random Forest Algorithm was used in screening and forecasting and drug development in Sars-Cov-2 pandemic where it was found that the accuracy rate of SVM was 77.5%, CNN was 87.02% and Random Forest Algorithm was 95.95% as proposed by [6] "Applications of ML and AI for Covid-19 (Sars-Cov-2) Pandemic".

Linear Regression is more accurate than SVM so for predicting confirmed Covid-19 cases Linear Regression was used as proposed in [7] "Covid-19 future predictions using ML algorithms".

To find the accuracy rate of prediction and analysis of Covid-19 SVM with an accuracy of 96%, KNN with an accuracy of 98.34%, Decision tree with an accuracy of 91% and Random Forest Classifier with and accuracy of 81% was used as proposed in [8] "A Comparative Approach to Predict Corona Virus using ML".

Long short-term(LSTM)was used to detect Covid-19 using X-rays images and CNN is used for deep feature extraction and detection is performed using LSTM as proposed by [9]"A systematic review on AI/ML approaches against COVID-19 outbreak".

A systematic search of PubMed, Web of Science and CINAHL databases was performed according to the PRISMA(Preferred Reporting Items for Systematic Reviews and Meta-Analysis) as proposed by [10]"Role of Machine Learning Techniques to tackle the COVID-19 Crisis: Systematic Review". A forecasting model was developed through

DOI: 10.17148/IJARCCE.2022.117101

the application of SVM model which is then combined with RBF Kernel so as to forecast the overall readmission rate as proposed by [11]"Classification of COVID-19 by using supervised optimized machine learning technique".
The real novelty in outbreak prediction can be realized through integrating machine learning and SEIR models as proposed by [12]"COVID-19 Outbreak Prediction with Machine Learning".
An integrated multi criteria decision making(MCDM) method is employed to evaluate and benchmarking different diagnostic models for Covid-19 with respect to evaluation criteria as proposed by [13]"Benchmarking Methodology for Selection of Optimal COVID- 19 diagnostic Model Based on Entropy and TOPSIS Methods".
Most of the Covid-19 patients were recognized by using algorithms like convolutional neural network(CNN) model and support vector machine(SVM) as proposed by [14]"A Systematic Review on the Use of AI and ML for Fighting the COVID-19 Pandemic". Natural language processing was used for understanding and classification of clinical reports and analyze and extract information from unstructured data and machine learning and deep learning was used for diagnosing and detecting disease as proposed by [15] "A Comprehensive Study of Artificial Intelligence and Machine Learning Approaches in Confronting the Coronavirus (COVID-19) Pandemic".
Linear Regression was used for the purpose of focusing on confirmed cases, death cases and recovered cases daily as proposed in [16] "Machine Learning Approaches for Tackling Novel Coronavirus (COVID-19) Pandemic".
SVM, K-Means clustering, Naive Bayes and many other models were used for to learn various data trends as proposed by [17] "Battling COVID-19 using machine learning: A review".
Comparative analysis of machine learning and soft computing models used to predict Covid-19 outbreak used SIR and SEIR models as proposed by [18]"COVID-19 Outbreak Prediction with Machine Learning".

METHODOLOGY USED :

Decision Tree

A decision tree (source - Sharma and Kumar, 2016) is a tree whose internal nodes can be as tests (on operant data patterns) and whose terminal nodes can take as categories (of such pattern). It contains tree-like structure and are constructed by using an algorithmic approach to identify the ways to split a dataset depending upon the given conditions. It maps all possible outcomes. It starts with a single node and further branches into outcomes (leaf nodes) and the branches keep dividing into further possible outcomes nodes. In such a way it seems to be a tree-like structure deciding outcomes for solving problems.

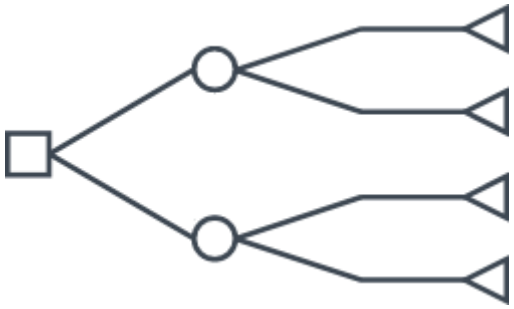

Fig1: Decision tree representation

(source : lucidchart.com)

Support Vector Machine (SVM)

Support Vector Machine (source geeksforgeeks) is a supervised machine learning algorithm used for both classification and regression. This method is best suited for classification problems. The SVM finds a hyperplane in an Ndimensional space that clearly classifies the data points.
The number of features decides the dimension of the hyperplane. If the number of input features is two, then the hyperplane is just a line and if the input features are three, the hyperplane is a 2-D plane.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117101

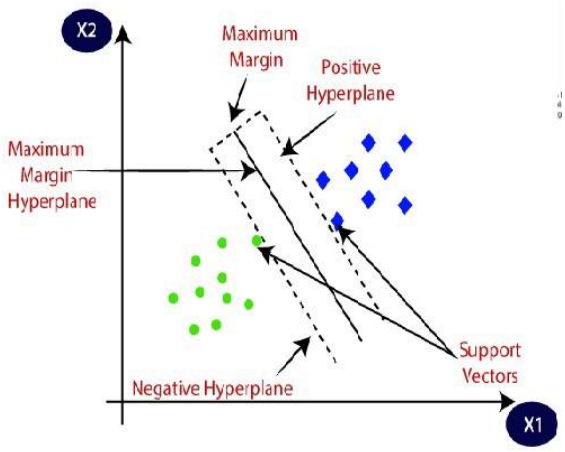

Fig2: SVM representation (source: javapoints.com)

K-Nearest Neighbour (KNN)

K-NN (Cai et al., 2010) is a sort of learning that depends upon the examples, where the capacity is just approximated locally and all calculations are acknowledged until grouping. This calculation can be very convenient to appropriate the loads to the organization with the end goal that the closer of the neighbors have more association to the normal than the ones that are further away. The neighbors are chosen from the lot for which the class is recognisable.

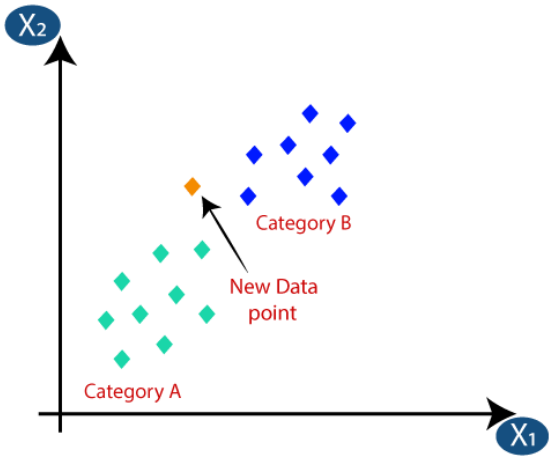

Fig3 : KNN representation (source: javapoint.com)

Naive Bayes

It is a technique used for constructing classifiers (source-wikipedia). This algorithm is suitable for binary and multiclass classification. For getting best results using naive bayes, categorical input variables proves to be a good choice as compared to numerical variables.

Formula: P(A|B) = P(B|A) P(A)P(B)

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117101

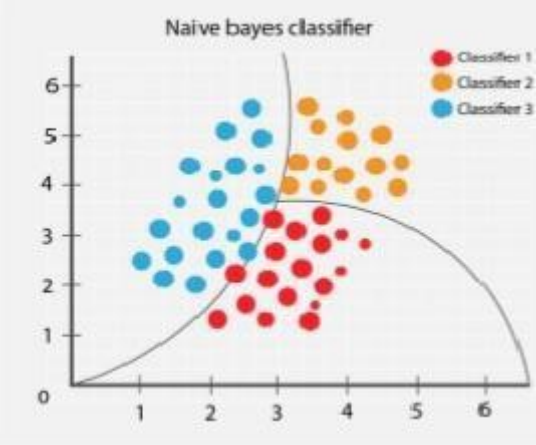

Fig4:Naive bayes theorem representation
(Source: https://thatware.co/naive-bayes/)

RESULTS

From our experiment on data set of Indian states (Maharashtra, Haryana, Bihar and Kerala) and using supervised learning techniques (Decision Tree, Naive Bayes, KNN) we found the following accuracy pattern:

1) On Maharashtra data, KNN and Naive Bayes predicted positivity rate well than decision tree with the same prediction accuracy.
2) On Haryana and Kerala data, all three algorithms gave the same accuracy of prediction.
3) On Bihar data, Naives Bayes predicted the accuracy well in comparison to other two.

Bar graph to show the pattern and percentage of accuracy:

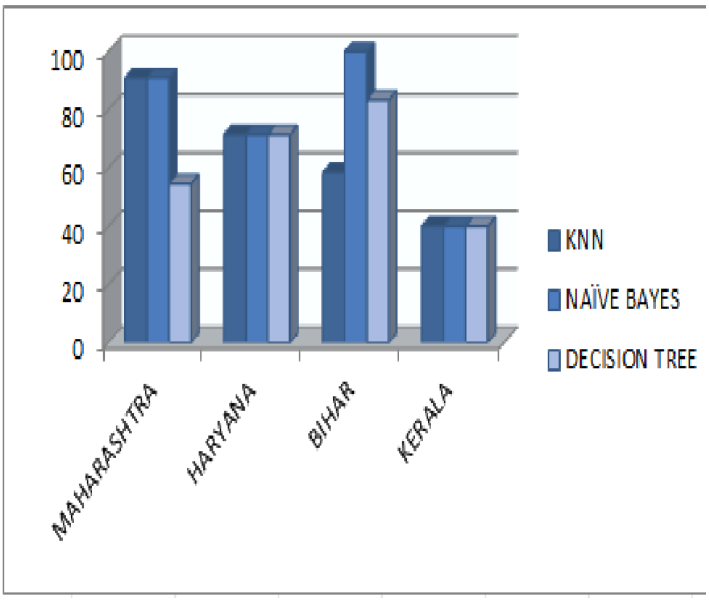

Fig 5 : Bar Graph of prediction on Indian States

CONCLUSION

From our findings and implementation, it is clear that the Naive Bayes algorithm has the highest prediction accuracy with an average accuracy prediction rate of 75.58 %. So, on our dataset of various Indian states, Naive Bayes Algorithm worked and performed well in comparison to the Decision Tree (accuracy = 62.32%) and KNN (accuracy = 65.16%) method.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.117101

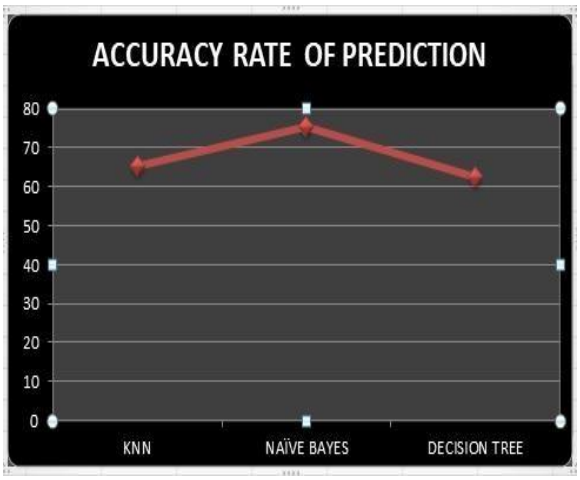

Fig 6 : Line Graph to show accuracy rate

REFERENCES

1. Ionder Tutsoy, Şule Çolak, Adem Polat, Kemal Balicki, "A Novel Parametric Model for the Prediction and Analyiss of the COVI-19 Casualties", IEEE ACCESS, October 2020, https://ieeexplore.ieee.org/document /9235320.
2. Furqan Rustam, Aijaz Ahmad Reshi, Arif Mehmood, Saleem Ullah, Byung-Won, Waqar Aslam, Gyu Sang Choi, "COVID-19 Future Forecasting Using Supervised MachineLearning Model", IEEE ACCESS, May2020, https://ieeexplore.ieee.org/stamp/sta mp.jsp?arnumber=9099302.
3. Fenglin Liu, Jie Wang, Jiawen Liu, Yue Li, Dagong Liu, Junliang Tong, Zhuoqun Li, Dan Yu, Yifan Fan, Xiaohui Bi, Xueting Zhang, Steven Mo, "Predicting and analyzing the COVID-19 epidemic in China: Based on SEIRD, LSTM and GWR model", PLOS ONE, August 2020, https://pubmed.ncbi.nlm.nih.gov/32853285 /
4. K.C.Santosh, "COVID-19 Prediction Models and Unexploited Data",Journal of medical systems , August 2020,https://link.springer.com/article/10.1007/s10916-020-01645-z
5. Moutaz Alazab, Albara Awajan, Abdelwadood Mesleh, Ajith Abraham, Vansh Jatana, SalahAlhyari, "COVID-19 Prediction and Detection Using Deep Learning", JCISIM, May 2020, https://www.researchgate.net/publica tion/341980921_COVID-19_Predicti on_and_Detection_Using_Deep_Le arning
6. S. Lalmuanawma, J. Hussain and L. Chhakchhuak, "Applications of ML and AI for COVID-19 (SARS-CoV-2) pandemic : A review", Elsevier, June 2020.
7. Jubin James, Tushar Kumar, Fauzan Mahzaib, Prashant Johri, "Covid-19 future predictions using ML algorithms", Turkish Journal ofComputer and Mathematics Education, May 2021, https://9lib.net/document/7qvnvx0z-v iew-covid-future-predictions-using-m achine-learning-algorithms.html
8. Rohini.M, Naveena K.R, Jothipriya. G, Kameshwaran.S, Jagadeeswari M, "A Comparative approach to predict corona virus using ML", IEEE XPLORE, April 2021.
9. Onur Dogan, Sanju Tiwari, M.A. Jabbar, Shankru Guggari, "A systematic review on AI/ML approaches against COVID-19 outbreak", Springer, July 2021,https://www.ncbi.nlm.nih.gov/pmc/art icles/PMC8256231/ 10. Hafsa Bareen Syeda, Mahanazuddin Syed,Kevin Wayne Sexton, Shorabuddin Syed, Salma Begum, Farhanuddin Syed, Fred Prior, Feliciano Yu Jr, "Role of Machine Learning Techniques to tackle the COVID-19 Crisis: Systematic Review", Medinform, August2020, https://pubmed.ncbi.nlm.nih.gov/33326405/
11. Dilip Kumar Sharma, Muthukumar Subramanian, Pacha.Malyadri, Bojja Suryanarayana Reddy, Mukta Sharma, Madiha Tahreem, "Classification of COVID-19 by using supervised optimized machine learning technique", Elsevier, November 2021, https://pubmed.ncbi.nlm.nih.gov/34868886/
12. Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R.Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk, Peter M. Atkinson, "COVID-19 Outbreak Prediction with Machine Learning", Research Gate, March 2020,https://www.researchgate.net/publication/344441018_COVID-19_Outbrea k_Prediction_with_Machine_Learnin g
13. Mazin Abed Mohammed, Karrar Hemeed Abdul Kareem, Alaa S. AL-Waisy, Salama A.Mostafa, Shumoos AL-FahdaWI, Ahmed Musa Dinar, Wajdi Alhakami, Abdullah Baz, Mohammed Nasser AL-Mhiqani, Hosam Alhakami, Nureize Arbaiy, Mashael S. Maashi, Ammar Awad Mutlag, Begona Garcia-Zapirain, Isabel De La Torre Diez, "Benchmarking Methodology for Selection of Optimal COVID- 19 diagnostic Model Based on Entropy and TOPSIS Methods",May 2020, https://ieeexplore.ieee.org/stamp/sta mp.jsp?arnumber=9096375

DOI: 10.17148/IJARCCE.2022.117101

14. Muhammad Nazrul Islam, Toki Tahmid Inan, Suzzana Rafi, Syeda Sabrina Akter, Iqbal H. Sarker, A. K. M. Najmul Islam, "A Systematic Review on the Use of AI and ML for Fighting the COVID-19 Pandemic", December 2020, https://ieeexplore.ieee.org/document /9366414
15. Md Mijanur Rahman, Fatema Khatun, Ashik Uzzaman, SadiaIslam Sami, Md Al-Amin Bhuiyan, Tiong Sieh Kiong, "A Comprehensive Study of Artificial Intelligence and Machine Learning Approaches in Confronting the Coronavirus (COVID-19) Pandemic", May 2021, https://pubmed.ncbi.nlm.nih.gov/339 99732/
16. Md Marufur Rahman, Md Milon Islam, Md Motaleb Hossen Manik, Md Rabiul Islam, Mabrook S. Al-Rakhami, "Machine Learning Approaches for Tackling Novel Coronavirus (COVID-19) Pandemic", July 2021, https://www.ncbi.nlm.nih.gov/pmc/art icles/PMC8287848/
17. Krishnaraj Chadaga ORCID, Srikanth Prabhu ORCID, Bhat K Vivekananda ORCID, S. Niranjana ORCID & Shashikiran Umakanth D T Pham, "Battling COVID-19 using machine learning: A review", August 2021, https://www.tandfonline.com/doi/full/ 10.1080/23311916.2021.1958666
18. Sina F. Ardabili, Amir Mosavi, Pedram Ghamisi, Filip Ferdinand, Annamaria R. Varkonyi-Koczy, Uwe Reuter, Timon Rabczuk and Peter M. Atkinson, "COVID - 19 Outbreak Prediction with Machine Learning", October 2021, https://www.mdpi.com/1999-4893/13/10/249