International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

Influence of Climatic Conditions on Spread of Covid'19-An Analysis using Machine Learning Methods

Sreedevi N

CMR Institute of Technology, Bengaluru

Abstract: Covid'19 is one of the deadliest infectious disease has been causing so many deaths around the globe for more than past one year. There are so many medical, habitual and other parameters available for the spread of this viral disease. This study has attempted to explore the relationship between number of new Covid'19 confirmed cases and some climatic factors and all these are done with machine learning framework. Seven locations are considered for the study as their pattern of spread differ and three main machine learning algorithms namely Linear model, Decision Tree and Random forest are exercised in the dataset. In order to improve the performance feature extraction and feature selection methods are used along with machine learning methods. Out of all three, Random Forest performed better by predicting the fact with an average of 96% of accuracy with respect to the dataset.

Keywords: Covid'19, Machine Learning Techniques, Linear Model, Decision Tree, Random Forest.

Introduction:

Viruses, a sort of microorganisms, are shaking the world today through diseases and it's really a challenging environment for human being to survive. There are several viral diseases spread these days, some of them in Figure 1, we come across in recent years are Swine flu, Ebola, Hanta, SARS, MERS recorded in the year 2013-2015 (Isra Al-Turaiki et al, 2016) and a variant of SARS and MERS (De Wit et al. 2016; Gautam and Hens 2020a) named SARS-CoV-2 which is named as COVID'19 and has labelled as PHEIC – Public Health Emergency of International Concern (WHO 2020b and Gautam and Hens 2020a, b; WHO 2020b). Ongoing research projects with updated databases on Covid'19 derives several results such as it transmits through air (Wang et al. 2020a, b),this airborne disease can survive in humans even without any symptoms and can cause death (WHO 2020a),maintaining personal hygiene like washing hands frequently and social distancing can be the best solution to stay away from this infectious disease(WHO 2020a). Initiatives of some countries with the cooperation of their citizens, reduced number of infected cases and deaths (Amarpreet Singh Arora et al 2020).

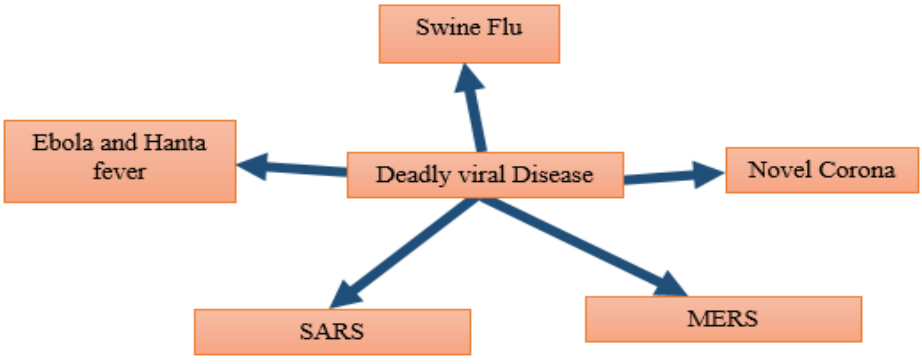


Figure 1. Different classification of deadly viral diseases

The research is an attempt to correlate some of the climatic factors with intense of infectious disease spread. Time series dataset with various stages of disease intensity of seven locations around the world are considered for comparison the study. In order to analyze the relationship between climatic factors and Covid'19, machine learning approaches are used, as its power of rapid processing and accuracy (Dasari and Prabaharan 2020, Dharun et al,2020). Techniques such as linear model, Decision Tree and Random forest are exercised on prediction and a variant of Random Forest performs better resulting 93% to 99% of accuracy in all of the considered datasets.

Contribution of the paper:

The research work includes feature extraction and feature selection into Random Forest, can be called as a variant of Random Forest. With respect to the case study, (i. e) some observation on application of machine learning on Covid'19 dataset are, according to this dataset and to the considered parameters relative humidity and snowfall plays superior role in increasing number of new cases than temperature and rainfall do. As a part of feature extraction, historical data (i. e) previous week's new admitted cases are fed as one of the parameters, accuracy significantly increased as per the data. If we have more data to examine, the prediction accuracy will also be improved.

Materials and Methods:

There are lots of research focuses on prediction using different perception of the disease (i. e) with different set of parameters. Few research works analyses the climatic variables for the outbreak (Ramon et al, 2020) exercising different machine learning approaches and their results show climatic contribution towards airborne disease is around 50% in the prediction. Hence climatic factors should also be considered to reach maximum accuracy in prediction.

About the dataset:

Figure 2 shows the region chosen for study namely Andorra (AD), Australia (AU), Czechia (CZ), India(IN), United States-Delaware(DE), Florida(FL), New York(NY) as per the availability in the chosen dataset from 1st Mar'20 to 6th Feb'21 and their respective parameters are a) Number of new confirmed cases, Temperature (Max, Min, Average), Rainfall, Dew point, Relative Humidity of daily data and historical data (i. e) previous week's new admitted cases are fed as one of the parameters (Ramon et al, 2020).

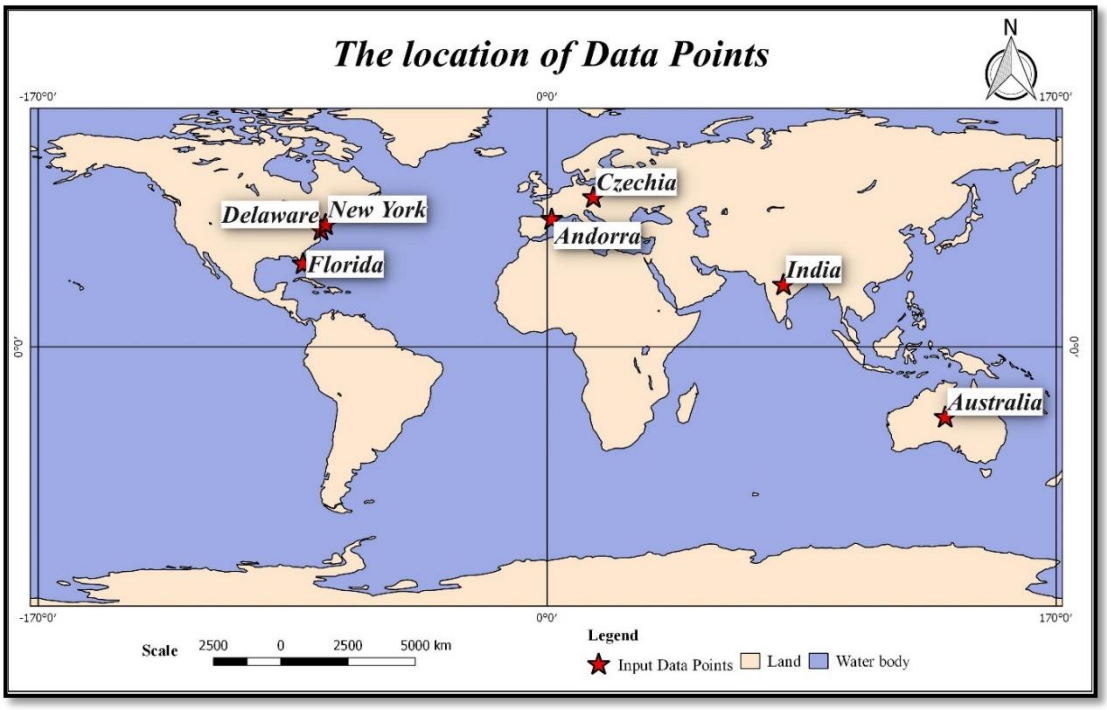


Figure 2. The location of data points

Methodology:

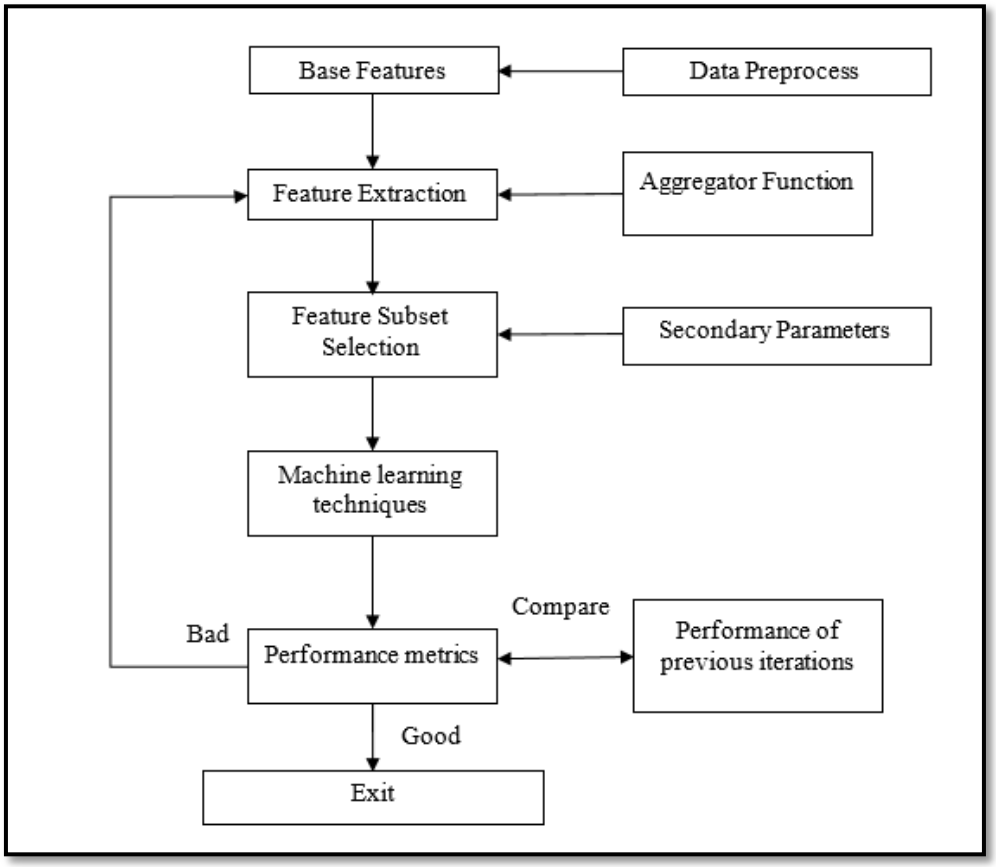
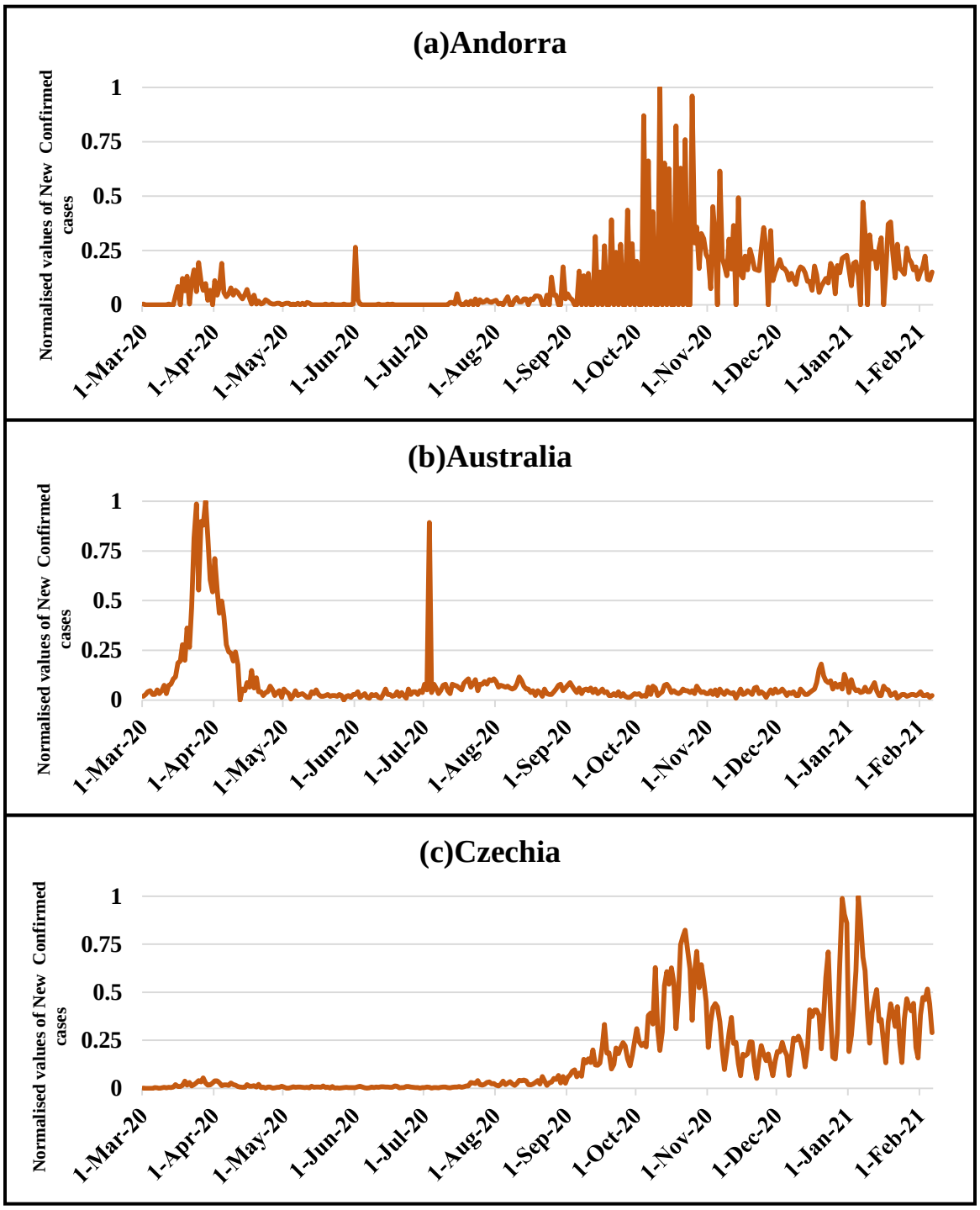
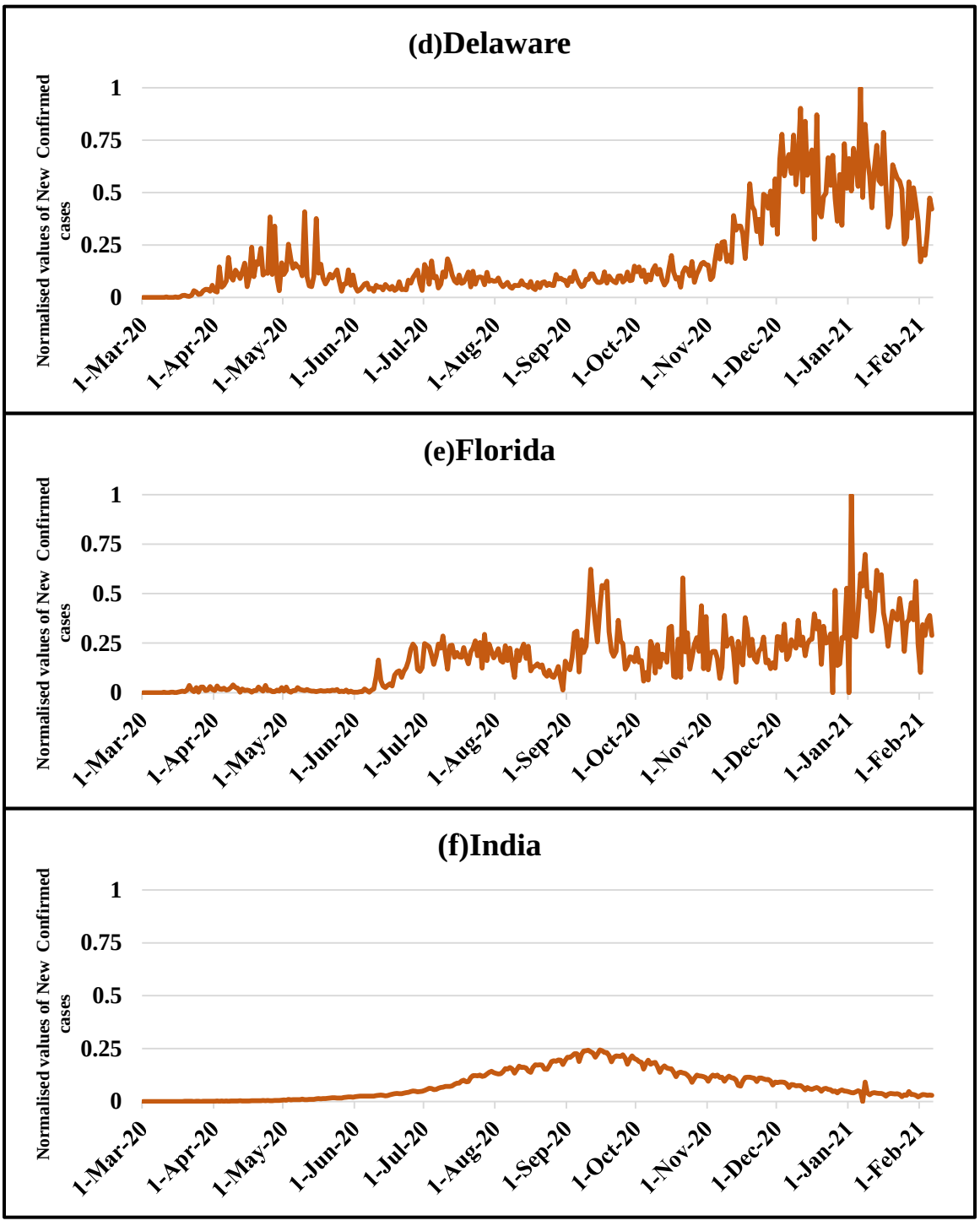


Figure 3. Workflow of the system

Data preprocessing is one of the inevitable steps in reaching exact prediction. Data Preprocessing includes outlier removal, filling up of missing data and applying normalization procedure. As handling missing data and outlier are not applicable in considered case study, data is normalized to the range 0 to 1. These normalized time series data are plotted in figure 3 and it is required to define target variable for prediction which is "number of new cases confirmed" , rest of the other variables stay independent or input variables. In order to make the prediction more understandable dependent variable is divided into five class namely, Class 1(Very low number of cases confirmed), Class 2(Low), Class 3(Moderate), Class 4(High), Class 5(Very High). Time series plots with normalized data of 7 datasets are found in the figure 4.


(a)Andorra
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21
(b)Australia
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21
(c)Czechia
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21


(d)Delaware
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21
(e)Florida
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21
(f)India
Normalised values of New Confirmed cases
1
0.75
0.5
0.25
0
1-Mar-20
1-Apr-20
1-May-20
1-Jun-20
1-Jul-20
1-Aug-20
1-Sep-20
1-Oct-20
1-Nov-20
1-Dec-20
1-Jan-21
1-Feb-21

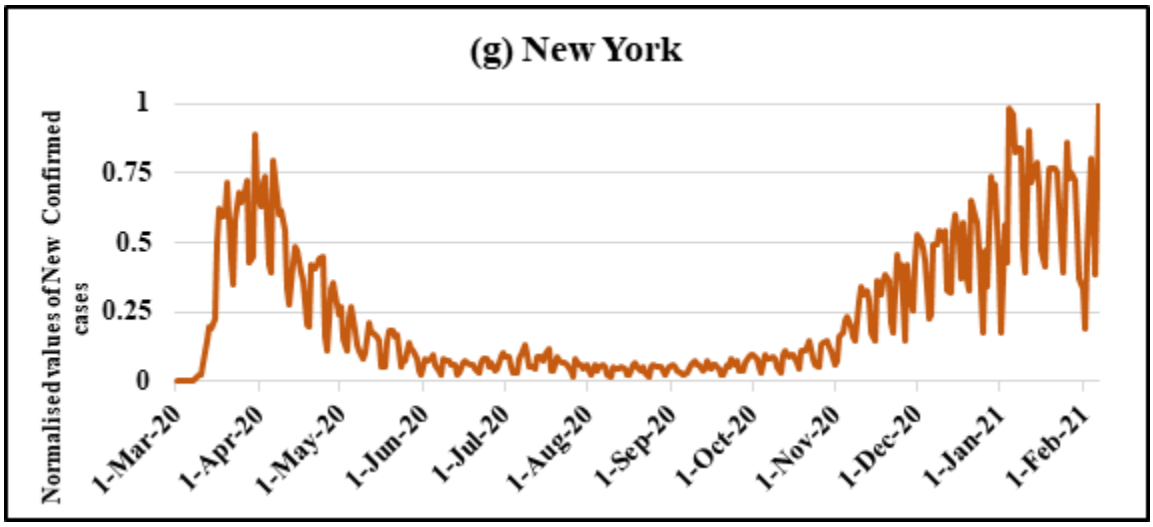


Figure 4.Time series plots of Covid'19 count of new cases.

There are several features that alter the spread of infectious disease. Though various countries have different reasons for rapid and slow rate of spread, this study focuses on contribution of basic climatic factors like temperature, rainfall, dew point, snowfall, etc.

Objective of the study is to make prediction using different algorithms of machine learning suggested by several scenarios of different research projects (Muhammad et al, 2020a,b,Milind and murukessan,2020, Victor Flores et al, 2021) namely Linear Regression, Decision tree, Random Forest. At initial iteration, Linear Regression, Support Vector Machine (SVM), Neural Network (David et al, 2020) and Decision Tree (DT) are involved but in later stages of analysis, performance of SVM and NN were not suitable to considered dataset.

Linear model:

Linear model, one of the basic techniques of machine learning, tries find a solution in a linear fashion (Rajani et al, 2021). To exercise Linear model, at least one independent variable should be specified and there can be more than one independent variable as shown in figure 5.

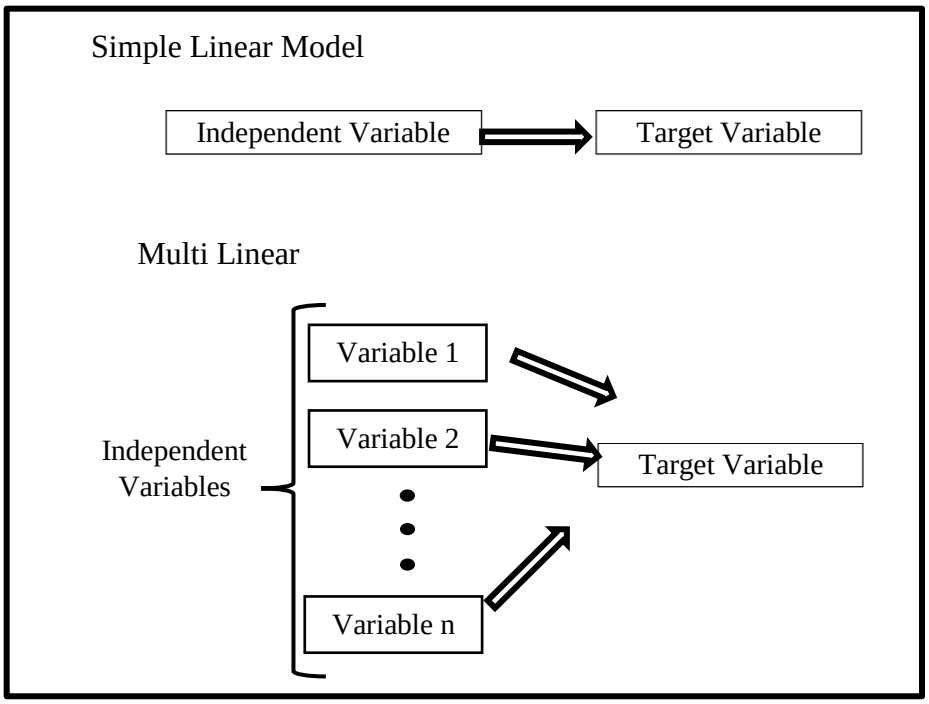


Figure 5. Linear Model

Decision Tree:

Decision Tree (DT), a supervised learning technique tries to find solution on the basis of set of rules or constraint learnt from training dataset as shown in figure 6. DT has a capability to handle uncertain scenarios by split mechanism. Research (Dasari and Prabaharan, 2020; Hu et al, 2012) proves that Decision tree is one of the best classifier techniques that can handle medical uncertainty.

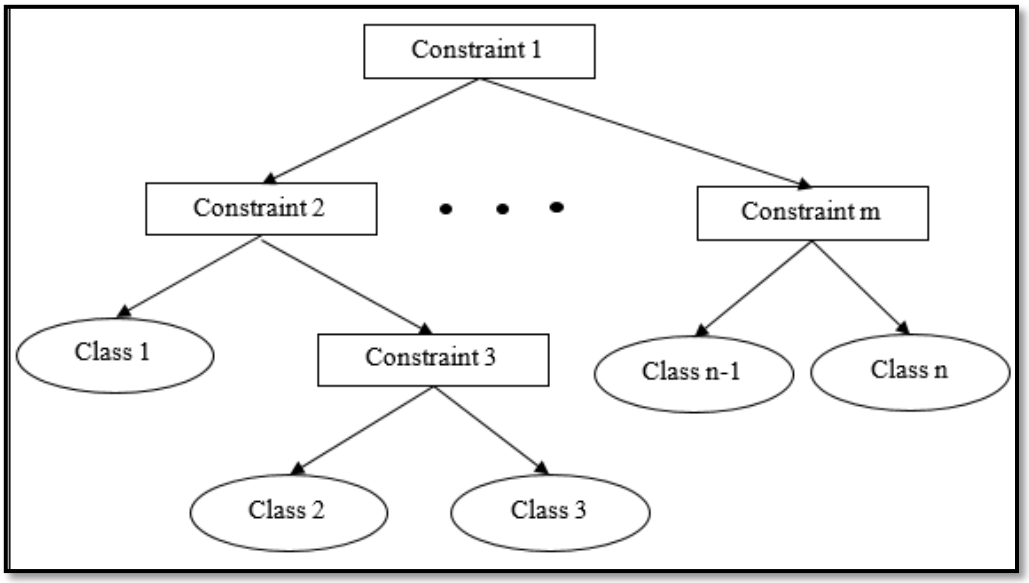


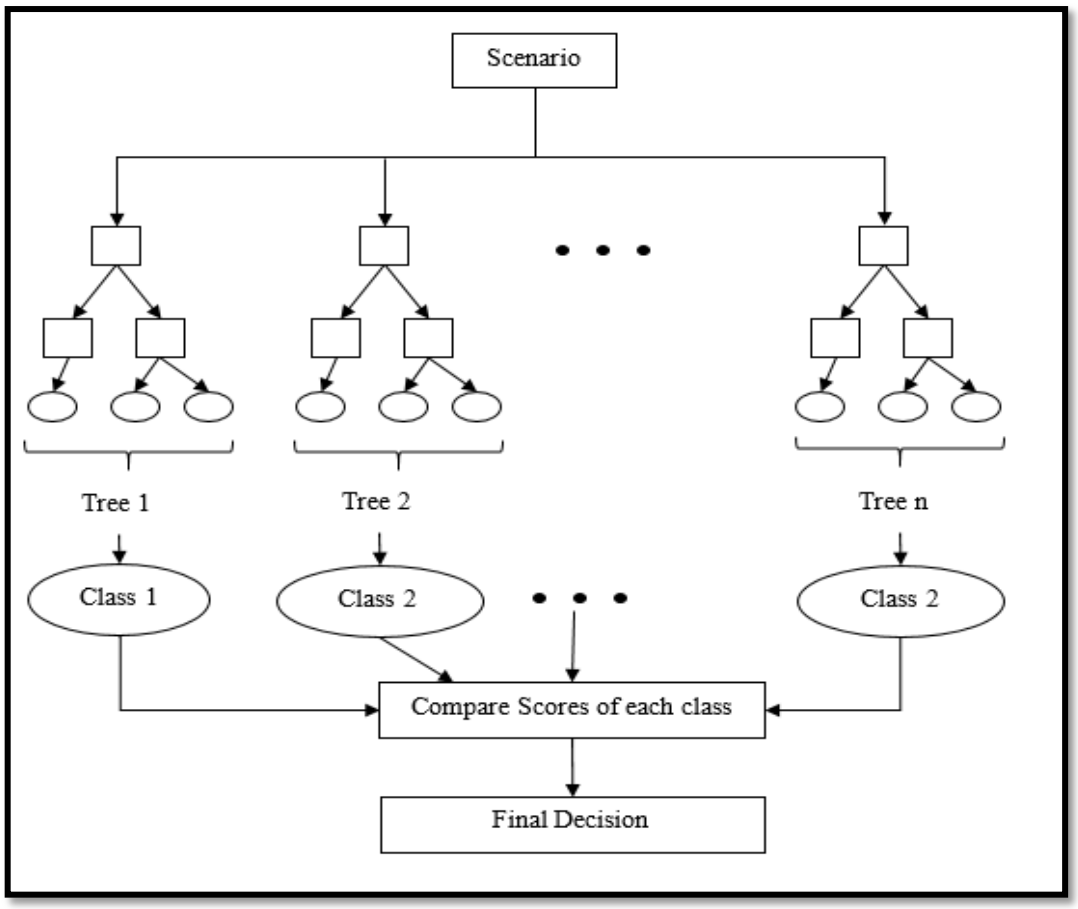
Figure 6. Decision Tree Model

Random Forest:

Random Forest is a developed version of Decision tree (i. e) it has several Decision trees in it as shown in the figure 7.

- One feature out of the 'p' features is picked up randomly, a tree is formed by placing target variable as the leaf node on assigning certain conditions to the chosen feature. Similarly, as many trees as it can be created with 'p' number of parameters, by assigning various condition and let the number of trees be 't'.
- Outcome of these 't' number of trees are predicted. An important point about this random forest is, the more number of trees in the forest will give more accuracy in prediction of target variable and in this way RF technique avoids overfitting problems unlike other algorithms.
  Figure 7: Random Forest Model

Random forest is proved (Iwendi et al , 2020) to be performing good in predicting Covid'19 using patients health monitoring attributes. A variant of Random Forest is used, which utilizes feature extraction and feature selection methods (i. e) using



several fuzzy aggregators are used to create more number of input parameters from existing and out of these derived parameters only selected parameters are utilized for the study based on their contribution towards resultant variable. These steps are followed before exercising machine learning methods to gain more accuracy on class prediction.

Feature Extraction and Feature Selection:

Feature extraction is a process of deriving new parameters from already existing and more contributing variable to the result. Aggregators are basic functions like Maximum, Minimum, Mean, and Weighted Average which are used to compute fuzzy membership functions. When fuzzy aggregators are applied on more contributing variables, new variables are derived and it is noted that not all the derived variables are capable of improving accuracy. So a methodology called feature selection is

exercised as a next step. Feature selection is a process of creating subsets of parameters from primary set of parameters and these newly created subsets will be utilized as the input of algorithms and performance are checked, subset which yields high accuracy will be considered.

Results & Discussion:

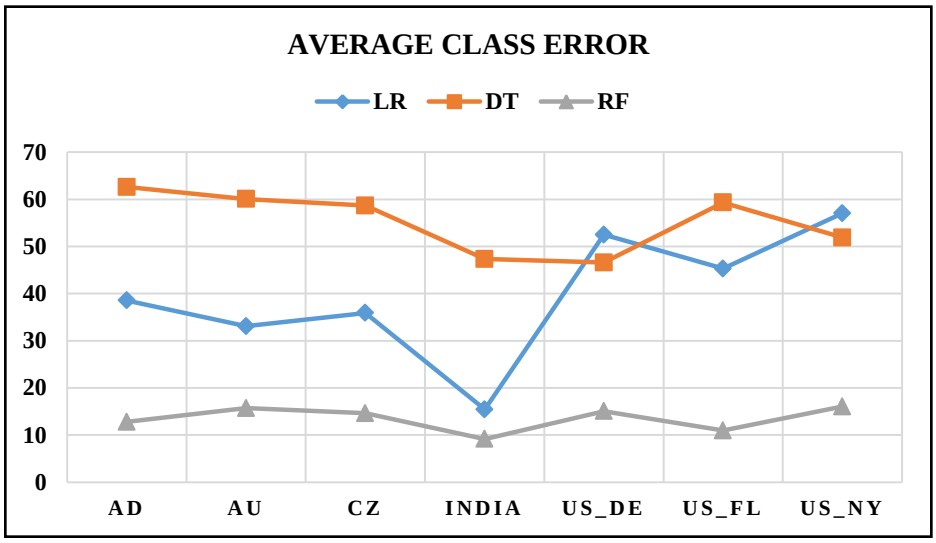


Figure 8. Average class error of Linear Model, Decision Tree and Random Forest

Machine Learning algorithms have different performance metrics to check stability of generated results. Confusion matrix is very helpful to know the number of correct and incorrect class predictions. The figure 8 shows average class error of applied algorithms for seven datasets.

Conclusion:

This study has attempted to explore the relationship between number of new Covid'19 confirmed cases and some climatic factors and all these are done with machine learning framework. Seven locations are considered for the study as their pattern of spread differ and three main machine learning algorithms namely Linear model, Decision Tree and Random forest are exercised in the dataset. In order to improve the performance feature extraction and feature selection methods are used along with machine learning methods. Out of all three, Random Forest performed better by predicting the fact with an average of 96% of accuracy with respect to the dataset.

References:

1. Amarpreet Singh Arora,Himadri Rajput,Rahil Changotra, "Current perspective of COVID-19 spread across South Korea: exploratory data analysis and containment of the pandemic", Environment, Development and Sustainability https://doi.org/10.1007/s10668-020-00883-y
2. David Oniani, Guoqian Jiang, Hongfang Liu, and Feichen Shen, Constructing co-occurrence network embeddings to assist association extraction for COVID-19 and other coronavirus infectious diseases, Journal of the American Medical Informatics Association, 27(8), 2020, 1259–1267 doi: 10.1093/jamia/ocaa117
3. Dasari Naga Vinod and S.R.S. Prabaharan, (2020).Data science and the role of artificial Intelligence in achieving the fast diagnosis of Covid-19, *Chaos, Solitons and Fractals*, 140 (2020) 110182, https://doi.org/10.1016/j.chaos.2020.110182.
4. De Wit, E., Van Doremalen, N., Falzarano, D., & Munster, V. J. (2020). SARS and MERS: Recent insights into emerging coronaviruses. Nature Reviews Microbiology, 14, 523–34.
5. Dharun Kasilingam, Sakthivel Puvaneswaran Sathiya Prabhakaran, Dinesh Kumar Rajendran, Varthini Rajagopal, Thangaraj Santhosh Kumar, Ajitha Soundararaj (2020),Exploring the growth of COVID-19 cases using exponential

modelling across 42 countries and predicting signs of early containment using machine learning, Transboundary and Emerging Diseases, https://doi.org/10.1111/tbed.13764.

6. Gautam, S. (2020). COVID-19: Air pollution remains low as people stay at home. Air Quality Atmosphere and Health, 13, 853–857
7. Gautam, S., & Hens, L. (2020a). COVID-19: Impact by and on the environment, health and economy. *Environment, Development and Sustainability, 22*, 4953–4954.
8. Gautam, S., & Hens, L. (2020b). SARS-CoV-2 pandemic in India: What might we expect? *Environment, Development and Sustainability, 22*, 3867–3869.
9. Isra Al-Turaiki, Mona Alshahrani, Tahani Almutairi, "Building predictive models for MERS-CoV infections using data mining techniques" Journal of Infection and Public Health (2016) 9, 744—748 http://dx.doi.org/10.1016/j.jiph.2016.09.007
10. Iwendi C, Bashir AK, Peshkar A, Sujatha R, Chatterjee JM, Pasupuleti S, Mishra R, Pillai S and Jo O (2020) COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. Front. Public Health 8:357. doi: 10.3389/fpubh.2020.00357
11. L. J. Muhammad, Ebrahem A. Algehyne, Sani Sharif Usman Abdulkadir Ahmad Chinmay Chakraborty,I. A. Mohammed, Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemiology Dataset, SN Computer Science (2021) 2:11 https://doi.org/10.1007/s42979-020-00394-7
12. Milind Yadav , Murukessan Perumal , Dr. M Srinivas, Analysis on novel coronavirus (COVID-19) using machine learning methods, Chaos, Solitons and Fractals 139 (2020) pp 1-12.
13. Muhammad, L.J., Algehyne, E.A., Usman, S.S., 2020. Predictive Supervised Machine Learning Models for Diabetes Mellitus, Predictive Supervised Machine Learning Models for Diabetes Mellitus, SN Computer Sci. 1, 1–10. https://doi.org/10.1007/s42979-020-00250-8
14. Rajani Kumari, Sandeep Kumar, Ramesh Chandra Paonia, Vijander Singh, Linesh Raja, Vaibhav Bhatnagar, and Pankaj Agarwal , BIG DATA MINING AND ANALYTICS ISSN 2096-0654 01107 pp65-75 Volume 4, Number 2, June 2021 dol:10.26599/BDMA.2020.9020013
15. Ramon Gomes da Silva, Matheus Henrique Dal Molin Ribeiro,Viviana Cocco Mariani, Leandro dos Santos Coelho, Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables*, Chaos, Solitons and Fractals* 139 (2020) 110027.
16. Victor Flores and Claudio Leiva, A Comparative Study on Supervised Machine Learning Algorithms for Copper Recovery Quality Prediction in a Leaching Process, Sensors 2021, 21, 2119. https://doi.org/10.3390/s21062119
17. Yuh-Jyh Hu, Tien-Hsiung Ku, Rong-Hong Jan, Kuochen Wang, Yu-Chee Tseng, Shu-Fen Yang,Decision tree-based learning to predict patient controlled analgesia consumption and readjustment, BMC Medical Informatics and Decision Making 2012, 12:131
18. Wang, C., Horby, P. W., Hayden, F. G., & Gao, G. F. (2020a). A novel coronavirus outbreak of global health concern. *The Lancet, 395*(10223), 470–473.
19. Wang, P., Chen, K., Zhu, S., Wang, P., & Zhang, H. (2020b). Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak. *Resources, Conservation and Recycling, 158*, 104814–104822
20. WHO. (2020a). Naming the coronavirus disease (COVID-19) and the virus that causes it. World Health Organization. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/. Accessed July 7, 2020
21. WHO. (2020b). Q&A on coronaviruses (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses. Accessed July 7, 2020.