Model-based machine learning to explore the nexus between COVID-19 and environmental factors in the United States

[a]Munir, T., [b]Hudson, I. L., [c]Cheema, S. A. [a]Muhammad, R., [a]Shafqat, M., and [a]Kifayat, T.

[a] Department of Statistics, Quaid-i-Azam University Islamabad, Pakistan. [b] Mathematical Sciences, College of STEM, School of Science, Royal Melbourne Institute of Technology, Melbourne, Australia. [c] Department of Applied Sciences, National Textile University Faisalabad, Pakistan.

Email: *irene.hudson@rmit.edu.au*

Abstract: The aim of this study is to demonstrate the applicability of machine learning methods to understand the transmission of the viral flow of COVID-19 with respect to various environmental factors. Daily update data of **new** COVID-19 related reported cases from six states of the United State (US), dated from 1st March 2020 to 30th November 2020, across 6 US states - *New York*, *New Jersey*, *Illinois*, *Massachusetts*, *Georgia* and *Michigan* are examined. The daily COVID-19 update data are assembled from the US health department and Weather Underground Company (WUC) official websites. A diverse set of environmental factors, including temperature, humidity, dew point, wind speed, atmospheric pressure and precipitation are used to express possible environmental determinants. Asymmetric distributions of daily reported new cases of COVID-19 with respect to all states is evident. The average numbers of new reported cases of COVID-19 patients remains highest in *Illinois*. Whereas maximum numbers of affected cases in a single day were reported in *Georgia*. The lowest of the average new cases is found in *Massachusetts* state.

We test six most used model-based machine learning methods, namely, linear discriminant analysis (LDA), classification and regression trees (CART), k-nearest neighbours (KNN), support vector machines (SVM), random forest (RF) and the naïve bayes (NB) method. The comparative performance of these ML schemes is expressed using statistics, such as kappa, balanced accuracy, detection rate, information preservation rate, accuracy, sensitivity, and specificity. Moreover, predictive orderings of the environmental factors, for each state with respect to the most promising ML method, are also reported to highlight the hierarchical significance of climatic determinants. The performance orderings of the ML approaches vary across states with the RF model the most promising in exploring the underlying nexus of between the environment covariates and case numbers across all states, the ML hierarchies are: ***New York:*** $P_{RF} > P_{KNN} = P_{CART} = P_{SVM} > P_{LDA} > P_{NB}$, ***New Jersey*** : $P_{RF} > P_{LDA} = P_{SVM} > P_{NB} = P_{CART} > P_{KNN}$, ***Illinois:*** $P_{RF} > P_{KNN} = P_{SVM} > P_{NB} = P_{CART} > P_{LDA}$, ***Massachusetts:*** $P_{RF} > P_{SVM} > P_{CART} > P_{KNN} > P_{NB} > P_{LDA}$, ***Georgia:*** $P_{RF} > P_{SVM} > P_{CART} > P_{KNN} = P_{NB} > P_{LDA}$ and ***Michigan:*** $P_{RF} > P_{KNN} > P_{SVM} > P_{CART} = P_{NB} > P_{LDA}$. Noting that procedures such as CART, NB and LDA show questionable performance where *Michigan* state is concerned.

Across the states, average temperature emerges as the most important candidate in explaining the underlying nexus between environment and COVID-19 numbers, consistent with Shahzad et al. (2020). However, we have found that other climate variables such as dewpoint, is a close second in *Georgia* and *Michigan* states, and humidity and wind speed play a similarly important role to dewpoint in *Illinois* and *Michigan*. Note *Georgia* and *Michigan* states have highest average temperature and dew point, and both states record low average wind speed. *Michigan* has a reported high black community, as does Georgia. For *Illinois*, temperature is dominant, but followed by dew point, and then closely by **both** humidity and wind speed, with *Illinois* having lowest average wind speed and low temperature. There is less evidence for an association between air pressure and precipitation and COVID-19 cases in all states. Finally, based on the outcomes of this research, we believe that a more rigorous study targeting other variables, such as population density, mobility, air quality, nature of travel bans, race, and the degree of health interventions, is required. Furthermore, understanding the potential for seasonality and the association with weather is particularly relevant for further work given the longer time series of COVID-19 information now available in 2021, as is modelling new cases, transmission, along with deaths, reproduction number and severity levels of COVID-19. Given the skewed nature of the distribution of number of reported cases in each state, future work could likewise employ the quintile regression approach.

Keywords: *COVID-19, environmental covariates, machine learning model*

1. INTRODUCTION

1.1 Motivation

The emergence of the novel coronavirus has led to enormous research efforts to understand how several environmental and non-environmental factors affect transmission. The research community from every corner of the globe is responding to the COVID-19 crisis in diverse analytic methods and using both local and global data inputs, and analysis is often based on a variety of outcomes, new cases, number of deaths, transmission rates, reproduction number and inputs. For example, Huang et al. (2020) and Guan et al. (2020) initiated the systematic mapping of the clinical nature of the virus at the start of the first wave. Further, Wild-Smith and Freedman (2020) and Cheema et al. (2020) focused on the quantification of the degree of effectiveness of various health interventions such as, quarantine, social distancing, and travel bans. Moreover, Kang et al. (2020) and Kawohl and Nordt (2020) investigated the long run socio-economic cost of the pandemic. Furthermore, Shahzad et al. (2020) studied environmental aspects of the pandemic in China as a case study. There is need for more inclusive and interactive efforts to assemble the so-called collective wisdom, towards equity and possible consensus, see Maas et al. (2020). Despite the launch of rigorous preventive measures, the viral transmission is escalating, confounded by new variants emerging. As the United States has led the path in terms of case incidence, various studies have focussed in the first instance on how weather variables may impact the spread of the disease (Bashir et al. (2020)), as environmental factors are known to affect the epidemiological transmission of many infectious diseases (Majumder & Ray, 2021). Machine learning is a branch of Artificial Intelligence (AI) shown to provide powerful predictive capabilities and superiority over conventional statistical modelling (Beam & Kohane, 2018, Hudson 2021, Boutaba et al. (2018). Given the high predictive power of these algorithms ML is becoming more widely used in public health data analysis and promises to aid in overall new strategies and interventions in public health.

The aim of this study is to demonstrate the applicability of machine learning methods to understand the transmission of the viral flow with respect to various environmental factors. Firstly, we study the underlying nexus between in the flow of viral spread and environmental factors for six major US cities. The US remains an attractive candidate in this regard as the number of COVID-19 active patients in US remained higher than China – the early sufferer of the pandemic, see WHO (2020) report. To this end we study the daily update data of new COVID-19 reported cases from six states of the United State (US), dated from 1st March 2020 to 30th November 2020, across 6 US states - *New York*, *New Jersey*, *Illinois*, *Massachusetts*, *Georgia* and *Michigan*. Furthermore, a diverse set of environmental factors, including temperature, humidity, dew point, wind speed, atmospheric pressure and precipitation are used as the environmental determinants. The daily update data is assembled from the US health department and Weather Underground Company (WUC) official websites. Secondly, we explore the applicability of the model-based machine learning (ML) methods to understand the nature of the health emergency. Comparative performance of the ML schemes is expressed using numerous relevant statistics, such as kappa, balanced accuracy, detection rate, information preservation rate, sensitivity, and specificity. Moreover, predictive orderings of the environmental factors, for each state with respect to the most efficient ML method are reported to highlight the hierarchical significance of the climatic determinants. Results and discussion are followed by areas for future research in Sections 2 and 3.

1.2 Machine Learning Tools

In this sub-section, we briefly summarise the ML methods investigated, namely, linear discrimination analysis (LDA), classification and regression trees (CART), k-nearest neighbours (KNN), support vector machines (SVM), random forest (RF) and naïve bayes (NB) methods).

Linear discrimination analysis (LDA): The LDA approaches the problem by assuming that $P(x|y=0)$ and $P(x|y=1)$, that is the conditional density functions are normally distributed with (μ_o, Σ) and (μ_1, Σ), where Σ is the covariance matrix which is Hermitian. The deletion criterion is launched on the threshold $w.x > c$, where $w = \Sigma^{-1}(\mu_1 - \mu_o)$. LDA has been used in various multidisciplinary areas, such as health surveillance, pattern recognition and marketing, for more details see Miller and Busby-Earle (2017) and Hudson (2021).

Naïve Bayes classifier (NB): Based on the rule of picking the most probable hypothesis and assuming conditional independence amongst the features, the NBC assigns class labels $\hat{y} = C_k$ as follows,

$$\hat{y} = \underset{k \in \{1,2,\dots,k\}}{argmax} p(C_k) \prod_{i=1}^{n} p(x_i|C_k),$$

where, $\mathbf{x}$ represents the vector of features and k are the possible outcomes. The utility of above conditional probability model is well documented in applied research literature, see for example Hudson (2021).

k-nearest neighbour (KNN): A non-parametric approach aims at partitioning set of d-dimensional vector of n observations, such as $(x_1, x_2, \dots, x_n)$ into $k \leq n$ set of partitions, that is, $S = \{S_1, S_2, \dots, S_k\}$ by giving each neighbor a contribution weight. The objective is then to find

$$\underset{S}{argmin} \sum_{i=1}^{k} \sum_{x \in s_i} \|x - \mu_i\|^2,$$

where, μ_i is mean of points in s_i. For more details see Nigsch et al. (2006) and Rajeswari et al. (2017).

Random Forest (RF): An off-the-shelf data mining procedure which aims to reduce the variance by employing bootstrap aggregation to tree learning. Given a training set, $\mathbf{x} = (x_1, x_2, \dots, x_n)$ with responses $\mathbf{y} = (y_1, y_2, \dots, y_n)$, random samples are selected from training data through bootstrapping to make predictions for unseen samples. The predictions from the training set are averaged such as $\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(x^{/})$, where B denotes bagging. The efficacy of the random forest procedure is discussed in Breiman (2001) and Fawagreh et al. (2014).

Support Vector Machine (SVM): A member of the supervised learning model family, SVM focusses on the pattern identification by creating hyper planes or set of hyper planes in high dimensional space. In its simplest form, given some training data, D, of n points of the form; $D = [(x_i, y_i) | x_i \in \mathcal{R}^p{}_n, y \in \{-1,1\}]$, where y_i represents the class to which point x_i belongs. The SVM aims to attain maximum-margin hyper plane dividing the points with $y = 1$ from those $y = -1$, orthogonally. For details see Simon et al. (2015).

Classification and Regression Tree (CART): A commonly used procedure in data mining whose prime goal is to establish a model with best predictive influx. Variance reduction in CART is usually conceptualised as the reduction of variance of the target variable due to a split, such as,

$$I_v(N) = \frac{1}{|s|^2}\sum_{i \in s}\sum_{j \in s}\frac{1}{2}\left(x_i - x_j\right)^2 - \left\{\frac{1}{|s_i|^2}\sum_{i \in s_i}\sum_{j \in s_i}\frac{1}{2}\left(x_i - x_j\right)^2 + \frac{1}{|s_f|^2}\sum_{i \in s_f}\sum_{j \in s_f}\frac{1}{2}\left(x_i - x_j\right)^2\right\},$$

where, N is the splitting node. Also, s, s_i and s_f represent set of presplit sample indices, indices for which the split is true and sample indices for which the split is false, respectively. A detailed account is given by Sharma et al. (2021).

2. MATERIALS

2.1 The Data

The data for this study are assembled from daily updates available on US health department official website reporting numbers of new COVID-19 cases from 1st March 2020 to 30th November 2020. Daily average values of environmental covariates are compiled from the official website of Weather Underground Company (WUC), these include, temperature (°F), humidity (%), wind speed (Mph.), atmospheric pressure (Hg.), precipitation (in.) and dew point. Moreover, the data is based on six US states, *New York*, *New Jersey*, *Illinois*, *Massachusetts*, *Georgia*, and *Michigan*. The state-wise summaries of all variables studied are presented in Table 1. From Table 1, from 1st March to 30th November, the average numbers of new reported cases of COVID-19 patients remains highest in Illinois. Whereas maximum numbers of affected cases in a single day were reported in Georgia. Further, the highest average temperature value is associated with Georgia and Michigan, with minimal averages observed in Massachusetts. Similarly, Massachusetts overall is the most humid state, New York exhibiting the minimal average humidity. At the same time, the highest values of average wind speed occur in New York state with Illinois state showing the lowest average wind speed. New Jersey exhibits the highest atmospheric pressure, whilst Georgia the minimum average air pressure. In contrast the state of Georgia exhibits the highest average precipitation in contrast to Massachusetts at the lower end. Similar trends are evident with respect to dew point with maximum average dew point in Georgia and Michigan, and minimum average dewpoint in New York.

Table 1. Summary statistics

Variable	*Minimum*	*Q1*	*Mean*	*Median*	*Q3*	*Maximum*
			New York			
New Cases	0.00	687.50	2357.84	1048.00	3169.50	11434.00
Temperature (°F)	35.5.00	52.55	64.27	64.40	75.80	90.30
Humidity (%)	26.10	47.15	59.27	58.10	70.70	91.80
Wind Speed (Mph)	2.30	7.85	10.29	9.40	12.10	21.60
Pressure (Hg)	29.30	29.90	29.99	30.00	30.10	30.60
Precipitation (in)	0.00	0.00	0.11	0.00	0.07	2.47
Dew.	13.70	36.35	48.37	49.90	62.25	72.70

New Jersey						
New Cases	0.00	326.00	1228.22	559.00	1929.00	4669.00
Temperature (°F)	33.90	52.10	63.25	63.20	75.25	87.90
Humidity (%)	30.40	52.15	64.26	64.80	76.35	96.90
Wind Speed (Mph)	1.50	6.75	9.25	8.70	11.30	22.30
Pressure (Hg)	29.00	29.90	30.00	30.00	30.10	30.60
Precipitation (in)	0.00	0.00	0.14	0.00	0.07	2.78
Dew.	13.40	36.95	49.41	50.80	63.80	73.30
Illinois						
New Cases	0.00	977.00	2643.01	1617.00	2482.00	15415.00
Temperature (°F)	34.00	52.25	62.73	64.50	74.05	83.80
Humidity (%)	30.50	66.00	74.77	76.90	84.50	95.40
Wind Speed (Mph)	0.50	3.65	6.45	6.10	8.70	23.20
Pressure (Hg)	14.10	29.50	29.53	29.60	29.70	30.10
Precipitation (in)	0.00	0.00	0.12	0.00	0.09	2.20
Dew.	22.90	41.95	53.54	56.20	66.40	73.80
Massachusetts						
New Cases	0.00	247.00	824.27	440.00	1192.00	4658.00
Temperature (°F)	27.10	47.60	57.73	58.90	67.75	81.10
Humidity (%)	38.60	68.35	78.49	82.10	90.30	98.80
Wind Speed (Mph)	1.90	6.85	9.84	9.50	12.15	23.90
Pressure (Hg)	29.30	29.80	29.95	29.90	30.10	30.60
Precipitation (in)	0.00	0.00	0.08	0.00	0.035	2.80
Dew.	9.10	37.55	50.29	51.20	63.65	74.80
Georgia						
New Cases	0.00	671.50	1711.54	1275.00	2407.50	31605.00
Temperature (°F)	43.70	61.60	69.45	70.60	77.55	85.20
Humidity (%)	31.70	61.05	68.46	69.40	78.25	91.60
Wind Speed (Mph)	1.80	6.00	7.92	7.30	9.50	19.20
Pressure (Hg)	28.60	28.90	28.97	28.90	29.10	29.40
Precipitation(in)	0.00	0.00	0.17	0.00	0.07	4.04
Dew.	21.00	50.50	57.60	60.90	67.90	72.90
Michigan						
New Cases	0.00	360.50	1412.38	717.00	1221.00	17368.00
Temperature (°F)	25.80	42.82	56.59	56.90	70.85	81.80
Humidity (%)	33.90	60.80	68.43	67.85	77.30	96.80
Wind Speed (Mph)	1.40	5.50	8.22	7.90	10.47	22.10
Pressure (Hg)	28.60	29.00	29.09	29.10	29.20	29.60
Precipitation (in)	0.00	0.00	0.11	0.00	0.04	2.12
Dew.	14.60	32.20	45.05	46.10	58.35	71.60

2.2 Results and Discussions

This section details data aspects along with major findings. The summary statistics are supported by appropriate tabular and graphical displays, as is the contribution hierarchy of the environmental factors in predicting new cases. The performance hierarchy of contemporary models is assessed using various relevant statistics, such as accuracy, true-false split of test data, degree of retaining the information and marginal homogeneity test for training and testing data. Table 2 presents the performance ordering of all considered machine learning models for each state. Next, we elaborate state wise findings of our study. Moreover, Table 3 displays the contribution ladder of the covariates in explaining the number of new COVID-19 cases for each state, using the most predictive ML model in each case.

Firstly, ***New York*** state reveals the performance ordering such as, $P_{RF} > P_{KNN} = P_{CART} = P_{SVM} > P_{LDA} > P_{NB}$. The RF model is the most promising candidate among the assembly of rival ML procedures in exploring the underlying nexus of environment covariates and the numbers of COVID-19 patients. We observe noticeably tight 95% confidence interval demonstrating the accuracy of the split of the test data into the fundamental categories. The accuracy remained bounded in the interval $(0.93, 1.00)$. The extent of retrieving the information is quantified through the null hypothesis that $H_o: accuracy \geq no\ information\ rate$. The associated p-value for RF is highly significant $p < 0.0001$. The next procedures in line are KNN, CART and SVM with equal potential, and accuracy bounded in the interval $(0.77, 0.96)$ along with p-value < 0.0001. LDA takes third position with accuracy interval $(0.73, 0.93)$ with significant p-value < 0.0001 (Table 2). Lastly, for the NB model accuracy remained bounded in $(0.66, 0.89)$ with significant p-value of information test. To ascertain explanatory and predictive contribution of the environmental covariates, we launched RF as the best choice ML method. Temperature emerges as the most important contributing covariate followed by dew point then to a lesser extent humidity, see gini index based on RF (Table 3).

The performance ordering of models in the state of ***New Jersey*** is $P_{RF} > P_{LDA} = P_{SVM} > P_{NB} = P_{CART} > P_{KNN}$. The RF maintains its position of prominence with accuracy interval of $(0.93, 1.00)$. It is to be noted that the 95% interval takes the same value as in the case of New York state. The reason for this is that in both cases the value of test statistics evaluating the accuracy attains maximum value of 1.0. This observation stays consistent throughout the study. Further, the significant p-value associated with the information test highlights the capability of the RF in retaining informative data splitting. Further, LDA and SVM have equal performance with confidence interval $(0.80, 0.97)$ and significant p-value of information retention. Next place belongs to NB and CART where accuracy is bounded in $(0.77, 0.96)$ with significant p-value of the test of information. Similarly, for KNN the accuracy is in the interval of $(0.73, 0.94)$, with significant p-value (Table 2). The climatic variable contribution patterns also are consistent, with temperature the leading covariate and dew point following, as in ***New York*** state, see gini index based on RF (Table 3).

Table 2. Confidence intervals for state-wise accuracy of models (grey shaded show insignificance of model)

New York	RF	KNN	CART	SVM	LDA	NB
	(0.934,1.00)	(0.773,0.958)	(0.773,0.958)	(0.773,0.958)	(0.728,0.933)	(0.664,0.893)
New Jersey	RF	LDA	SVM	NB	CART	KNN
	(0.934,1.00)	(0.797,0.969)	(0.797,0.969)	(0.773,0.958)	(0.773,0.958)	(0.728,0.938)
Illinois	RF	KNN	SVM	NB	CART	LDA
	(0.934,1.00)	(0.623,0.865)	(0.623,0.865)	(0.525,0.789)	(0.525,0.789)	(0.468,0.740)
Massachusetts	RF	SVM	CART	KNN	NB	LDA
	(0.934,1.00)	(0.821,0.979)	(0.773,0.958)	(0.751,0.946)	(0.623,0.865)	(0.583,0.853)
Georgia	RF	SVM	CART	KNN	NB	LDA
	(0.934,1.00)	(0.821,0.979)	(0.775,0.985)	(0.751,0.946)	(0.751,0.946)	(0.664,0.893)
Michigan	RF	KNN	SVM	CART	NB	LDA
	(0.934,1.00)	(0.623,0.865)	(0.487,0.757)	(0.360,0.639)	(0.360,0.639)	(0.343,0.621)

Next, for ***Illinois*** state the performance ordering is observed as, $P_{RF} > P_{KNN} = P_{SVM} > P_{NB} = P_{CART} > P_{LDA}$. Again, the RF model outperforms other techniques while obtaining maximum accuracy with 95% confidence interval of $(0.93, 1.00)$ and significant p-value of information test. RF is followed equally by KNN and SVM with interval of $(0.63, 0.87)$. At equal third position, NB and CART with associated confidence interval $(0.53, 0.79)$ with significant p-value of information test. The LDA remains the poorest performer with interval of $(0.47, 0.74)$, LDA shows insignificant ability in retaining the desired information from the training data (p-value of 0.07). We also observe varying patterns in contribution ordering of the climatic covariates (Table 2). For ***Illinois***, temperature is dominant but followed by dew point, and then closely followed by **both** humidity and wind speed, differing to the patterns above (Table 3). Noting ***Illinois*** state has the lowest average wind speed and low temperature.

For the states of ***Massachusetts*** and ***Georgia*** states, we found the ordering such as $P_{RF} > P_{SVM} > P_{CART} > P_{KNN} > P_{NB} > P_{LDA}$ and $P_{RF} > P_{SVM} > P_{CART} > P_{KNN} = P_{NB} > P_{LDA}$, respectively. In both instances, RF approach again is dominant. Again, the RF model outperforms other techniques while obtaining maximum accuracy with 95% confidence interval of $(0.93, 1.00)$ and significant p-value of information test. For both ***Massachusetts*** and ***Georgia*** states SVM, CART and KNN accuracies are significant and equal across both states, in order being (0.821,0.979), (0.775,0.985), and (0.751,0.946), with NB and LDA significant but lower in Massachusetts (Table 2). The climatic variable contribution patterns for ***Massachusetts*** have temperature the leading covariate and dew point following, see RF based gini index (Table 3). Noting ***Illinois*** state has the lowest average wind speed and low temperature. The climatic variable contribution patterns for ***Georgia*** temperature are followed very by closely by dew point and then to a slightly lesser degree by both humidity and wind speed, which also differs to the patterns discussed above (Table 3). However, ***Michigan*** state demonstrates a distinctive profile regarding **both** ML procedure hierarchy and regarding important climate predictors. The performance ordering is $P_{RF} > P_{KNN} > P_{SVM} > P_{CART} = P_{NB} > P_{LDA}$, with RF again the most promising, but followed here by KNN and then SVM, with associated intervals of $(0.63, 0.87)$ and $(0.49, 0.76)$, respectively (Table 2). The remaining three models, CART, NB and LDA have questionable performances with respect to the test of the null hypothesis of $H_o: accuracy \geq no\ information\ rate$. Their related p-values are 0.55 and 0.66, respectively, indicating a lower degree of information sustainability. This is further supported by Mc Nemar's test for testing the marginal homogeneity among training split and testing data, with a p-value of 0.47. For ***Michigan***, temperature is followed very by closely by dew point and humidity and then to a slightly lesser degree by wind speed, which again differs to the patterns discussed above (Table 3). Note ***Georgia*** and ***Michigan*** states have highest average temperature and dew point, and both states record low average wind speed. ***Michigan*** state has maximal value of average temperature among all considered states and reported highest black community, as is ***Georgia*** state.

Table 3. Gini Index to highlight the importance of climatic variables per state: RF based.

State	Temp.	Dew.	Hum.	Wind.	Pressure.	Preci.
New York	68.476	25.559	18.083	11.856	9.525	3.509
New Jersey	48.621	34.061	20.365	16.598	11.101	6.076
Illinois	37.057	29.447	25.592	23.572	11.349	9.963
Massachusetts	49.928	34.845	17.939	17.217	10.936	5.945
Georgia	38.601	36.551	23.155	20.212	11.175	7.019
Michigan	29.88	29.35	28.48	25.2	12.18	11.38

Figure 1 shows the error rate, out-of-class sample bagging, and cross entropy against the number of trees in RF to be used to obtain optimal classification of the data. All curves show 50 as the optimal number of trees to attain the most suitable model, noting that RF assumes that each tree is identically distributed.

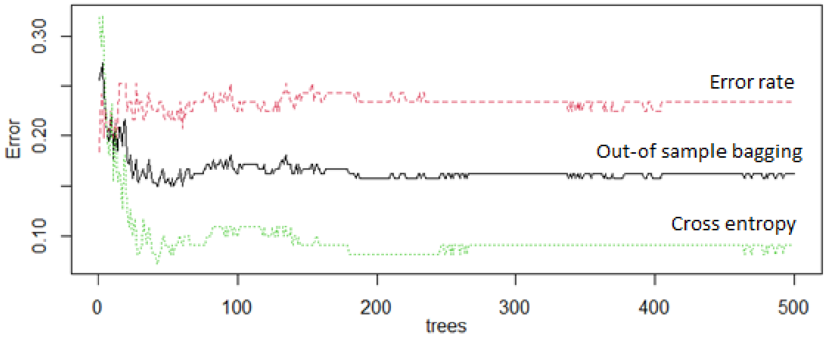

Figure 1. Error rate, out-of-class sample bagging and cross entropy *versus* tree number using RF.

3. SUMMARY

This article primarily focuses on the study of interlinks between the environmental factors and the flow of COVID-19 outbreak by using machine learning tools. The objectives are achieved by conducting in-depth exploration of six US states' data of daily updates of new reported cases of COVID-19 patients and daily average values of environmental covariates, dated form 1st March 2020 to 30th November 2020. In all six states, the RF model was found to be the optimal approach to capture the probabilistic features of the data, among the assembly of the 6 ML techniques. This performance ordering is established by employing various relevant statistics, such as accuracy, degree of classification, marginal homogeneity, and maintenance of information rate. Further, in five instances out of six, RF is followed by the SVM procedure; Michigan state appears to be the only exception in this regard, with KNN following RF. The contribution hierarchy of the environmental factors in predicting new COVID-19 case numbers is reported using only RF. Temperature emerges as the most important candidate in explaining the underlying nexus of environment and COVID-19 related numbers, consistent with Shahzad et al. (2020) who used a generalized linear model approach. However, we have found that other climate variables such as dewpoint, and humidity play a significant role. Also, by considering the skewed nature of the distribution of number of reported cases in each state, future work could employ the quintile regression approach (see Sim & Zhou, 2015). It is anticipated that, instead of studying the probabilistic behaviours of the distribution based on averages, analysis at various quintiles may be more informative. Temperature and humidity are significant factors in virus transmission and seasonality for several reasons, as they determine virus survivability and persistence in the air and on surfaces (Aboubakr et al. 2020). Across the states, average temperature emerges as the most important candidate, as in Shahzad et al. (2020). But variables such as dewpoint, is a close second in Georgia and Michigan, and humidity and wind speed play a similarly important role to dewpoint in Illinois and Michigan. There is less evidence for an association between air pressure and precipitation and COVID-19 cases. A fuller treatment of previous day/weeks cases and interactions of the (sub) populations, which leverage time-series approaches that use inputs from previous time periods, as in Hudson (2018) and also in Hudson & Keatley (2013) who modelled flowering and budding pheno-phases with respect to lagged climate/pheno-phases is underway. Factors such as population density, mobility, race, air quality, wind direction, interventions, will add further insights, as will modelling the interaction of populations, specifically with lockdowns.

REFERENCES

Aboubakr, H. A., Sharafeldin, T. A. and Goyal, S.M. (2020). Stability of SARS-CoV-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: A review. *Transbound Emerg. Dis.* doi: 10.1111/tbed.13707

Bashir, F. M., Benjiang, M. A, Komal, B., Bashir, A. M., Tan, D. and Bashir, M. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA, *Science of The Total Environment,* Vol. 728, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138835.

Beam, A. L. and Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319, 1317–1318.

Boutaba, R., Salahuddin, M. A., Limam, N., Ayoubi, S., et al. (2018). A comprehensive summary on machine learning for networking: evolution, applications, research opportunities. *J of Internet Services*, 4(16), 1257-1270.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45 (1): 5–32.

Cheema, S. A., Kifayat, T., Abdu R Rahman, Khan, U., Zaib. A., Khan. I, and Kottakkaran, S. N., (2020), Is social distancing, and quarantine effective in restricting COVID-19 outbreak? Statistical evidence from Wuhan, China. *Computers, Materials & Continua*. 66(2), 1977-1985.

Cioffi, R., Travanglioni, M., Piscitelli, G., Petrillo, A. and Felice, F. D. (2020), Artificial intelligence and machine learning applications in smart production: Progress, trends and directions. *Sustainability*, 12, 492-516.

Fawagreh, K., Gaber, M. M. and Elyan, E. (2014). Random forest: from early developments to recent advancements. *System Sciences & Control Engineering*, 2(1), 602-609.

Guan, J. W., Ni, Y. Z., Hu, Y., Liang, W., Ou, C. et al. (2020). Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine.* 382 (1708-1720).

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y. et al. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet*. 395 (497-506).

Hudson, I. (2021). Data Integration Using Advances in Machine Learning in Drug Discovery and Molecular Biology. In: Artificial Neural Networks, *Methods in Molecular Biology,* vol 2190 (7), 167-184.

Hudson I.L. (2018). Stochastic trend detection by a broad class of state-space models: application to phenology. Phenology 2018 Conference, International Society Biometeorology Phenology Commission (ISB-PC) 23-27/9/2018, Melbourne.

Hudson IL and Keatley MR. (2013). Scoping the budding and climate impacts on Eucalypt flowering: nonlinear time series In: Piantadosi, J., Anderssen, R.S. & Boland J. (eds) 20th Intlternational MODSIM, Australia & NZ, 1582-1588.

Kang, L., Ma, S., and Chen, M. (2020). Impact on mental health and perceptions of psychological care among medical and nursing staff in Wuhan during the 2019 novel coronavirus disease outbreak: A cross-sectional study. *Brain Behaviour. Immunology.* 87, 11–17.

Kawohl, W., and Nordt, C. (2020). COVID-19, unemployment, and suicide. *Lancet Psych*. 7, 389–390.

Liang, W., Guan, W., Chen, R., Wang, W., KeXu, J. et al. (2020). Cancer patients in SARS-CoV-2 infection: a nationwide analysis in China. *The Lancet*, 21(3), 335-337.

Maas, B., Grogan, K. E., Chirango, Y., Harris, N., Lievano-Latorre, L. F. et al. (2020). Academic leaders must support inclusive communities during COVID-19. *Nature Ecology & Evolution*, 4, 997-998.

Majumder, P., Ray, P.P. (2021). A systematic review and meta-analysis on correlation of weather with COVID-19. *Sci Rep* **11,** 10746 https://doi.org/10.1038/s41598-021-90300-9

McClymont H, Hu W. Weather Variability and COVID-19 Transmission: A Review of Recent Research. *Int J Environ Res Public Health*. 2021;18(2):396. Published 2021 Jan 6. doi:10.3390/ijerph18020396

Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E. and Mitchell, J. B. (2006). Melting point prediction employing KNN algorithms and genetic parameter optimization. *Journal of Chemical Information &Modeling*, 46(6), 2412–2422.

Rajeswari, S. Suthendran, K. and Rajakumar, K. (2017). A smart agricultural model by integrating IoT, mobile and cloud-based big data analytics. *International Journal of Pure and Applied Mathematics*, 118(8), 365-370.

Shahzad, F., Shahzad, U., Fareed, Z., Iqbal, N., Hashmi, H. S., and Ahmad, F. (2020). Asymmetric nexus between Temperature and COVID-19 in top ten affected provinces of China: A current application of quantile-on-quantile approach. *Science of the Total Environment*, 736, 139-146.

Sharma, N, Sharma, N. and Jindal, N. (2021). Machine learning and deep learning application-A vision. *Global Transitions Proceedings*, 2(1), 24-28.

Sim, N., Zhou, H., (2015). Oil prices, US stock return, dependence between their quantiles. *J. Bank. Finance*,55, 1–8.

Simon, A., Deo, M. S., Venkatesen, S. and Babu, D. R. M. (2015). An overview of machine learning and its applications. *International Journal of Electrical Sciences & Engineering*, 1(1), 22-24.

Sousa, A. D., Mohandas, E. and Javed, A. (2020). Psychological interventions during COVID-19: challenges for low and middle-income countries. *Asian Journal of Psychiatry*, 51, 102-106.

Wilder-Smith, A. and Freedman, D.O. (2020). Isolation, quarantine, social distancing, and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak. *Journal Travel Medicine*. 2, 13-27.

World Health Organization (WHO). Coronavirus disease (COVID-2020) situation reports. Official website of World Health Organizationhttps://www.who.int/emergencies/dis-eases/novel-coronavirus-2020/situation-reports