COVID-19 Spreading Speed and mortality rate Prediction using leave one out cross validation

Ahmad Arrabi
Computer Engineering department
Princess Sumaya University for Technology
Amman, Jordan
AhmadZahiArrabi@gmail.com

Amjed Al-Mousa
Computer Engineering department
Princess Sumaya University for Technology
Amman, Jordan
a.almousa@psut.edu.jo

***Abstract*—Ever since COVID-19 was declared as a global pandemic, the world has been enduring many hardships from different sectors. These hardships produced the need for solutions to help aid the emergency. We have developed a method that assesses the rate at which the virus spreads in a country. Our method was to predict how long did it take for a country's cases to reach 1%, 2%, and 5% of its population. The prediction was conducted by Regularized Linear Regression models and Support Vector Machine (SVM) Regression using the kernel trick. Leave one out cross validation was implemented to train the models. The highest median and mean prediction accuracy achieved was using an Elastic Net model, they were 95.1% and 93.65%, respectively. Another prediction was for the mortality rate of COVID-19 cases. A ridge regression model was trained and predicted with a median and mean accuracies of 82.7% and 73.33%.**

***Index Terms*—COVID-19, Machine Learning, SVM, Kernel trick, Regularized Linear Regression, Leave one out cross validation**

I. Introduction

On 11 March 2020, the World Health Organization (WHO) announced COVID-19 to be a global pandemic [1]. Since then, the world has been facing one of toughest challenges in modern times. Financial losses grew considerably, health systems collapsed, and social repercussions arose [2]. Consequently, the need for solutions to fight the virus also grew. Machine learning (ML) can be a powerful tool that helps governments and health sectors to predict when the next outbreak is going to happen [3].

This works aims to predict when COVID-19 cases reach a certain percentage of a country's population as well as the mortality rate of the virus in a country. The study was exploratory in nature, our approach was to collect data from different sources to create one inclusive dataset. The collected data were applied to train different regression ML models to predict the target variables. Each dataset that was used in the training process, and its source, can be found in Table I in section III. Data were last updated in late March 2021.

A method of measuring the speed of the virus's spread was to predict when will cases reach 1%, 2%, and 5% of a country's population. We believe that this prediction may help countries that still have not reached these percentages yet. The predicted time will allow governments to create some timetable to prepare for the virus' spread. This is also a great opportunity to discover unnoticed correlations between different features and the spreading speed of the virus.

The other prediction was concerning the mortality rate of the virus. This prediction can aid health sectors in the fight against the pandemic. As the features used were relating to a country's nature and characteristics, in addition to some data regarding the virus's status in the country.

The paper is organized as follows: Section III describes the experimental setup, what pre-processing techniques were used, how the train test split was implemented, and a thorough investigation of the dataset. Section IV provides a description of the algorithms used to develop the ML training models. Section V emphasizes on the results of our work. A comparison between the accuracy of different models was made. In addition to a comparison between the results of predicting different targets.

II. Literature Review

An increasing number of studies have explored the impact of machine learning and artificial intelligence (AI) in the fight against COVID-19. In [4] researchers concluded that these approaches encouraged healthcare systems and governments to take the suitable actions towards the pandemic.

In [5] the authors presented an improved mathematical model to predict the growth of the pandemic. The improved model was a ML-based one. They used a cloud computing platform for more accurate predictions and real-time analysis.

Singh et al. [6] investigated time series data to develop a support vector machine (SVM) prediction model. The model predicted the confirmed, deceased, and recovered cases. They used two common features that we used, longitude and latitude of the predicted country.

The work in [7] performed clinical predictive models that predicted if a patient was prone to getting infected with COVID-19. They developed their models using deep learning and laboratory data. The models were tested on 18 laboratory findings from 600 patients and validated with 10-fold cross-validation. The results gave 91.89% F1-score, 86.75% precision, and 99.42% recall.

Malki et al. [8] concluded that the temperature, a feature we used, and humidity are important features for predicting COVID-19 mortality rate. To reach this conclusion, they used several regressor ML models to extract the relationship between different factors and the spreading rate of the virus.

In [9] it has been found that lockdowns help in reducing the virus' spreading speed. The study offers initial evidence that the pandemic can be suppressed by a lockdown. We will be using the lockdown date and type as features in our work.

Studies in [10] explored the challenges that may affect the accuracy of machine learning methods when predicting the number of confirmed cases. It was shown that many countries are not doing enough testing, leading to inaccurate data regarding the number of confirmed cases. We saw this inaccuracy in our dataset in some developing or politically unstable countries, so they were neglected.

Most research has tended to focus on predicting COVID-19 cases. Time series data was the most dominant in these types of research. We focused on predicting how fast the virus spreads rather than the number of cases. This approach has the potential to help countries to prepare for the next outbreak of the virus.

III. Experimental Setup

A. Dataset Overview

The dataset used was derived from 7 different datasets (Sources can be found in Table I). The features used in the dataset mainly highlighted general characteristics of the countries e.g., climate, poverty, population, ...etc. In addition to features that were strictly related to the virus e.g., reproduction rate and stringency index. This variation led to analysing unnoticed correlations between different features and the spreading speed of the virus as well as the mortality rate.

For the prediction of the spreading speed of the virus, 3 subsets of the dataset were constructed. The first subset represented countries that reached a total of cases equal to 1% of its population. This dataset included 91 countries. The second and third subsets represented countries that reached a total of cases equal to 2% and 5% of its population, respectively. Countries that reached the 2% point were 66 in total, while countries that reached the 5% point were only 23.

A key distinction between the first two datasets and the third (Countries that reached the 5% threshold), was that the latter consisted of mainly European countries. Fig. 1 illustrates this distribution.

Correspondingly, a fourth subset was created for the prediction of the mortality rate of the virus. Only certain features were chosen to be a part of this dataset. The features were only those with high correlation with the target, as will be represented in the following sections. The number of countries in this dataset was 27. This was due to the data being noisy and inconsistent, as the cases reporting in unstable or developing countries may be inaccurate [10]. This inaccuracy affected the performance of the models. Thus, the dataset was filtered and only countries with relatively high human development and freedom (above 0.8 and 6.5 respectively) were chosen.

B. Preprocessing and Preparing Data for ML Model

The targets (predictions) were the speed of the virus' spread and the mortality rate of it. Thus, two metrics that represent the targets were devised.

The first metric indicated the time it takes for a country's total COVID-19 cases to reach a certain threshold of its population. To calculate this metric. First, a new feature, which is the date of reaching the threshold, was constructed. Then, using eq. (1), the difference between the calculated date and the date of the first confirmed case was computed. The computed time was represented in days.

$$T_{spread} = Date_{threshold} - Date_{first\ case} \tag{1}$$

To measure the mortality rate, the total death count was divided by the total number of cases, see eq. 2. Fig. 5 represents the distribution of the mortality rate.

$$Mortality\ Rate = \frac{Total\ deaths}{Total\ cases} \tag{2}$$

Another feature created was the time for lockdown. It indicates how long did a country wait to declare a lockdown, if any, since announcing the first confirmed case. Refer to Eq. (3) to calculate the lockdown time.

$$T_{lockdown} = Date_{lockdown} - Date_{first\ case} \tag{3}$$

To gain a deeper insight of the datasets, Figs.2-5 illustrate the distribution of the targets in each dataset.

1) Dealing with NULL values: In any data science related work, it is inevitable to have NULL values in your dataset. To solve this problem, an approach of calculating the mean of each continuous feature grouped by the continent was implemented. The NULL values were filled with the calculated mean values. E.g., all NULL values of European countries were filled with the same data.

Moreover, categorical NULL values were dealt with in a similar manner. Filling them with the most frequent value grouped by continent.

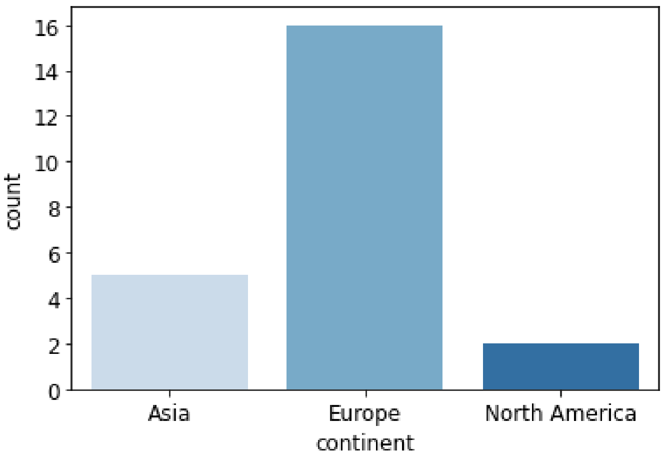

Fig. 1. Continent distribution of the 5% dataset

TABLE I
FEATURES USED

Feature	Data type	Description	Source
Country	String	Country name	US government, CIA the World Factbook [11].
Climate	Integer	Number between 1-4 to represent the climate	
Total cases	Float	Accumulative sum of cases	the World Health Organization, Coronavirus source data [12].
Date	datetime	Date of each day	
Total tests	Float	Accumulative sum of tests	
Tests per case	Float	Number of tests per total cases	
Extreme poverty	Float	Share of the population living in extreme poverty, most recent year available since 2010	
Human development index	Float	A number between 0-1 representing human development	
Hospital beds per thousand	Float	Hospital beds per a thousand population	
Total deaths	Float	Accumulative sum of deaths	
Positive rate	Float	The share of COVID-19 tests that are positive, given as a rolling 7-day average	
Continent	String	Name of continent	
Date of cases to reach 1%	Datetime	Calculated feature, see eq. (1)	
Reproduction rate	Float	Real-time estimate of the effective reproduction rate (R) of COVID-19 [13]	
stringency index	Float	Government Response Stringency Index: composite measure based on 9 response indicators including school closures, workplace closures, and travel bans, re-scaled to a value from 0 to 100 (100 = strictest response)	
Aged 70 and older	Float	Share of the population that is 70 years and older in 2015	
median age	Float	Median age of the population, UN projection for 2020	
Diabetes Prevalence	Float	Diabetes prevalence (% of population aged 20 to 79) in 2017	
Mortality rate	Float	Total deaths/cases	
HF (human freedom) score	Float	Number between 0-10 to represent the degree of freedom in a country. Economic freedom, freedom of speech, religion, discrimination, freedom of movement are all factors affecting the final score.	Cato Institute and the Fraser Institute [14].
Tourism	Float	International tourism, number of arrivals 2018	Multiple governmental websites. Gathered from kaggle [15].
Latitude	Float	Latitude of the country	
Longitude	Float	Longitude of the country	
First confirmed case date	Datetime	Date of first confirmed case	
Lockdown date	Datetime	Date when the lockdown started	
Lockdown type	String	Full or partial	
Population	Integer	Population of a country	United Nations Population Division [16].
Population density	Integer	Population/Area	
GDP	Float	Gross domestic product of a country (2018)	The World Bank [17].
Literacy rate	Float	Percentage of literate population	Wikipedia [18].
Time for spread	Float	Target column	Calculated from eq.(1)
Time for lockdown	Float	Time till declaring a lockdown since announcing the first confirmed case	Calculated from eq.(3)

2) Categorical Data: Only numerical data could be applied to the used ML models. One hot encoder was implemented on the categorical data to transform them into numerical ones. The following three features were transformed: climate, lockdown type, and continent.

3) Feature Scaling: The data was originally not scaled. Hence, min-max transformation, also called normalization, was implemented to scale the data. This transformation rescales each feature so that its values end up ranging from 0 to 1 [19]. The transformation was done according to eq. (4). Min-Max scaling preserves the shape of the original distribution, leading to the conservation of outliers.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{4}$$

4) Train Test Split: Due to the small size of the dataset, leave one out cross validation (LOOCV) was used to assess the accuracy of the predictions. Data would be split into folds equal to the number of samples in the dataset. This way the training would be done on all samples except one, which is then used as a testing sample to evaluate the model. The significance of this method can be represented as follows: It ensured that there would be no bias in the train test split, each sample got evaluated equally, and all data were used in training and testing.

Other splits were evaluated but led to biased results, e.g., 80% training and 20% testing. This bias led to inconsistent readings, i.e., any slight change in the training sample would lead to considerable changes in the testing results. Hence, LOOCV was selected to split the dataset in favor of other methods.

C. Analyzing Correlations

To investigate correlations between different features and the target variables, scikit-learn's correlation function was used [20]. The method of measurement was Pearson's standard correlation coefficient. This method quantifies the linear correlation between two features [21].

First, an analysis was done regarding the time for spread targets. Tables II-IV represent the correlation between all features and the targets. Due to the 2% and 5% datasets being

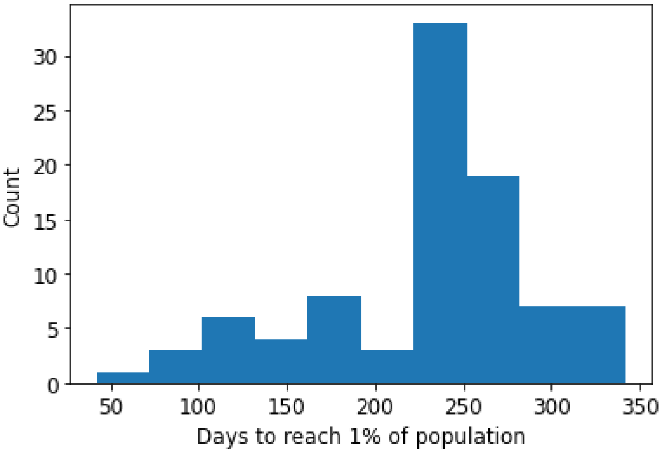

Fig. 2. Distribution of the time it took to reach the 1% point

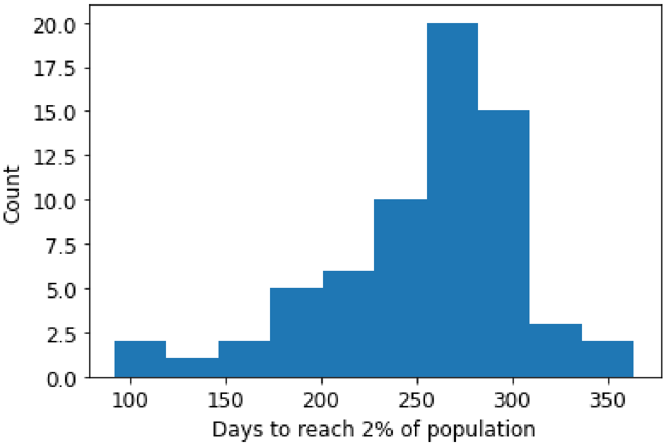

Fig. 3. Distribution of the time it took to reach the 2% point

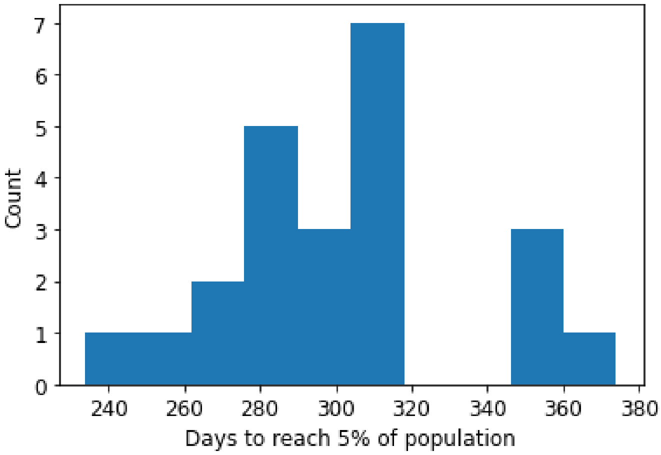

Fig. 4. Distribution of the time it took to reach the 5% point

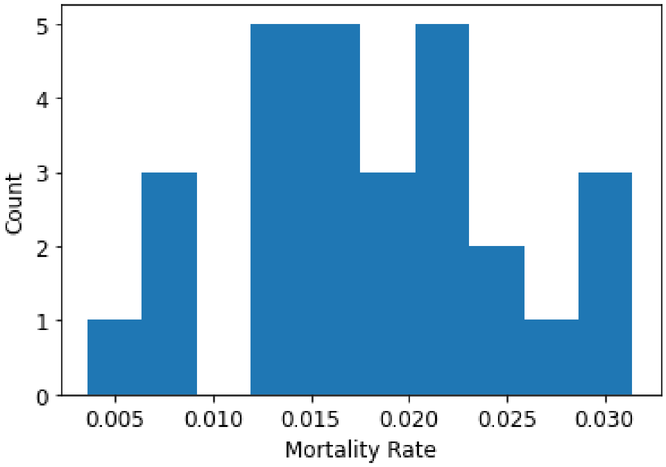

Fig. 5. Distribution of the mortality rate

subsets of the 1% dataset, multiple features had the same trend but with different impact on the target throughout the datasets. E.g., the time for lockdown feature had a 0.34, 0.40, and 0.75 correlation in the 1%, 2%, and 5% datasets, respectively. These samples are all positive but with different effect on the target. Notice that not all features acted this way, e.g., population had a positive correlation in the 5% dataset, while having a negative one in the remaining ones. This was due to the dataset being more conclusive with less noise, as the countries in the 5% dataset were mostly the ones with the higher human development and freedom index, which may indicate that the data would be more precise.

Moreover, the time for a country to declare a lockdown since its first case has the highest positive correlation with the targets in the 2% and 5% datasets. I.e., The more a country waits to declare a lockdown the more time the virus would take to reach 5% of its population, thus, slower spread of the virus.

Inspired by [8], climate was correlated with the virus' spreading speed. Notice that dry areas had a negative correlation with the time for spread. Hence, countries with dry climate tend to allow a faster spread of the virus than other countries.

The second analysis was performed on the mortality rate target. Table V indicates the correlation between all features and the mortality rate. The features with the highest positive correlation were time for lockdown, stringency index, number of people aged 70 or older, tourism, median age, reproduction rate, and weather a country is European. While the features with the highest negative correlation were diabetes prevalence, density, the dry climate of a country, and weather it is an Asian country.

When all features were fed to train the models, unreliable predictions were observed. The models predicted the same value for all samples. This inaccurate prediction led to filtering out the features and using only the above-mentioned ones (the ones with highest correlation with the target). After this filtration, the results were more accurate and logical.

IV. Algorithms

Multiple ML regression algorithms were applied to train our models. The highest number of datapoints was only 91,

TABLE II
CORRELATIONS WITH TIME FOR SPREAD IN THE 1% DATASET

Feature	Correlation
Reproduction rate	0.347167
Africa	0.223720
Tests per case	0.202058
Time for lockdown	0.166314
Hospital beds per thousand	0.111057
Longitude	0.108781
Latitude	0.095912
Tourism	0.083677
Partial lockdown	0.081741
Total tests	0.080791
Europe	0.078070
Wet tropical	0.076152
Positive rate	0.059972
Density	0.055610
HF score 2017	0.035075
Wet winters	0.012655
Humid subtropical	0.005718
Extreme poverty	-0.000264
North America	-0.036053
Human development index	-0.052033
Total deaths	-0.061432
Full lockdown	-0.081741
Dry	-0.088302
Population	-0.098405
Total cases	-0.099041
South America	-0.105785
Asia	-0.136463
Literacy rate	-0.166647
GDP 2018	-0.171983

TABLE III
CORRELATIONS WITH TIME FOR SPREAD IN THE 2% DATASET

Feature	Correlation
time for lockdown	0.407243
Latitude	0.380231
hospital beds per thousand	0.294899
Europe	0.264645
reproduction rate	0.244301
HF score 2017	0.189730
positive rate	0.163862
Tourism	0.154895
tests per case	0.146898
total tests	0.137600
Africa	0.126552
wet winters	0.105188
Longitude	0.093732
Full lockdown	0.088607
humid subtropical	0.078423
Density	0.071502
human development index	0.065885
North America	0.051150
literacy rate	0.028427
extreme poverty	0.024488
Wet tropical	-0.000424
total cases	-0.024564
Population	-0.025677
total deaths	-0.084898
Partial lockdown	-0.088607
Dry	-0.190031
GDP 2018	-0.193473
Asia	-0.198200
South America	-0.303463

a relatively small number in the ML world. It is necessary to take this impediment into consideration when choosing the algorithms.

Regularization had an impact on our analysis. Hence, constrains were applied to the models to overcome overfitting. Linear Regression was implemented as well as three regularized versions of it, Lasso Regression, Ridge Regression, and Elastic Net.

SVM Regression was also implemented using the kernel trick, where different kernel transformations were applied to the model [19]. We chose the following kernels: Polynomial, Radial Basis Function (RBF), and Linear.

A. Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) Regression is a penalized regression method. It identifies the important features associated with the target. This identification is done by eliminating the weights of unimportant features (shrinking the weights). Eq. (5) represents the cost function of Lasso Regression, notice the added term to the error.

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \tag{5}$$

B. Ridge Regression

Ridge Regression is a regularized linear regression model. Similar to Lasso, Ridge Regression adds a regularization term to the cost function. This additional term minimizes the weights thus constraining the model. Eq. (6) represents the cost function of Ridge Regression.

$$J(\theta) = MSE(\theta) + \frac{1}{2}\alpha \sum_{i=1}^{n} \theta_i^2 \tag{6}$$

Mean Squared Error (MSE) was used in the cost functions. However, it is still acceptable to use different error metrics as performance measures in the training and test sets.

C. Elastic Net

Elastic Net has characteristics from both Lasso and Ridge Regression. The regularization term is a combination of the terms of Lasso and Ridge.

$$J(\theta) = MSE(\theta) + r\alpha \sum_{i=1}^{n} |\theta_i| + \frac{\alpha(1-r)}{2} \sum_{i=1}^{n} \theta_i^2 \tag{7}$$

Eq. (7) represents the cost function of the Elastic Net model. To investigate the relationship between Elastic Net and the other regularized models. Let $r = 0$, the cost function is equivalent to the one in Ridge Regression. If $r = 1$, then it will be equivalent to Lasso Regression.

D. Support Vector Machine (SVM) Regression

SVM is a supervised model that can be used for various purposes, it was used for regression in this work. SVM

TABLE IV
CORRELATIONS WITH TIME FOR SPREAD IN THE 5% DATASET

Feature	Correlation
time for lockdown	0.756118
Tourism	0.548128
human development index	0.369697
reproduction rate	0.332031
Latitude	0.276913
positive rate	0.239234
wet winters	0.203171
Europe	0.199672
HF score 2017	0.184683
total deaths	0.145731
Full lockdown	0.096376
literacy rate	0.087256
total cases	0.080640
Population	0.079642
Longitude	0.024074
hospital beds per thousand	0.017393
total tests	0.000079
GDP 2018	-0.018195
tests per case	-0.030201
humid subtropical	-0.039023
Partial lockdown	-0.096376
Wet tropical	-0.099224
Asia	-0.101801
Dry	-0.110210
North America	-0.177041
extreme poverty	-0.189994
Density	-0.193281

TABLE V
CORRELATIONS WITH THE MORTALITY RATE

Feature	Correlation
time for lockdown	0.521332
stringency index	0.407612
aged 70 older	0.404424
Tourism	0.383587
median age	0.367227
Europe	0.243851
reproduction rate	0.235265
South America	0.224678
humid subtropical	0.222886
total deaths	0.214290
literacy rate	0.213887
hospital beds per thousand	0.198040
Population	0.195492
HF score 2017	0.179583
life expectancy	0.163713
total cases	0.128338
human development index	0.103788
Wet tropical	-0.026962
Latitude	-0.099651
GDP 2018	-0.119436
extreme poverty	-0.162475
cardiovasc death rate	-0.217634
Longitude	-0.320243
diabetes prevalence	-0.398307
Dry	-0.413128
Density	-0.426061
Asia	-0.511501

Regression tries to fit as many data points as possible on a hyperplane in an N-dimensional space (N is the number of features).

The kernel trick is a method that obtains similar results of adding similarity features. This addition does not lead to any heavy computational load on the model. The following three kernel functions were applied on the SVM Regression model to get better results and to compare them.

1. Polynomial Kernel
2. Radial Basis Function (RBF) Kernel
3. Linear Kernel

V. Results and Analysis

The aforementioned algorithms were used to train our models. LOOCV was implemented to evaluate the predictions of the models. This method led to the evaluation of only one test sample for each fold. The number of folds was equal to the number of data points.

A. Performance Measures

The Mean Absolute Error (MAE) was used as a metric to quantify the performance of the predictors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{8}$$

In eq. (8), x is the target (independent variable) while $\hat{x}$ is the predicted value (dependent variable). In our case, $n = 1$ as the test sample is only one.

To further the analysis, an accuracy metric was defined using the MAE. The MAE was divided by the test target sample, this value represented the percentage error. Then, subtracted the percentage error from 100 get the accuracy, see eq. (9).

$$Accuracy = 100 - \frac{MAE}{target\ test\ sample} * 100 \tag{9}$$

The mean and the median of the accuracies of all prediction samples were finally computed in order to generalize the accuracy evaluation.

$$Mean = \frac{\sum_{i=1}^{n} x_i}{n} \tag{10}$$

B. Time for spread prediction

1) Regularized Linear Regression Models: Table VI summarizes the mean and median accuracies of the predictions. It is apparent that there was a general rise in the accuracy as the dataset compressed. The median accuracy was not as affected with the dataset change as the mean accuracy. This concludes that there exists some outliers in the results.

The regularized linear regression models had an advantage over linear regression in the last dataset. Linear regression gave a mean accuracy of 84.43% while the other models were above that by at least 6%. On the other hand, the median accuracy was not as different between the models.

TABLE VI
MEAN ACCURACY OF THE PREDICTION RESULTS

Mean Accuracy (%)				
percentage of population	Linear Regression	Lasso Regression	Ridge Regression	Elastic Net
1 %	73.05	74.80	75.72	75.70
2 %	85.32	85.70	84.50	84.80
5 %	84.43	93.55	91.20	93.65
Median Accuracy (%)				
percentage of population	Linear Regression	Lasso Regression	Ridge Regression	Elastic Net
1 %	85.23	87.23	87.51	84.75
2 %	93.01	92.50	91.60	91.62
5 %	94.20	95.01	92.07	95.10

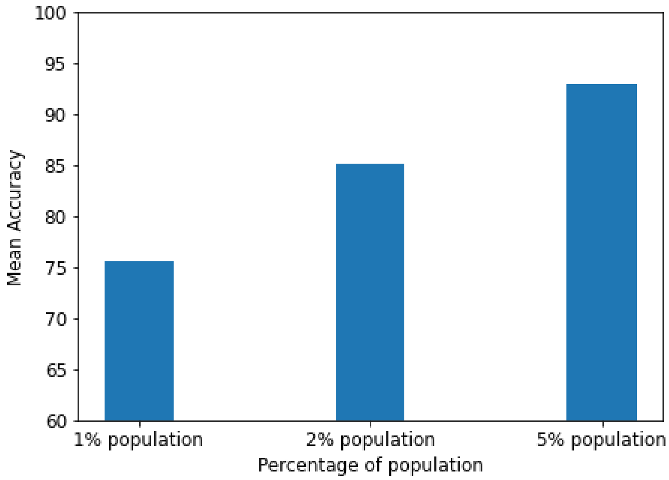

Fig. 6. Mean accuracy of the ensemble method's predictions

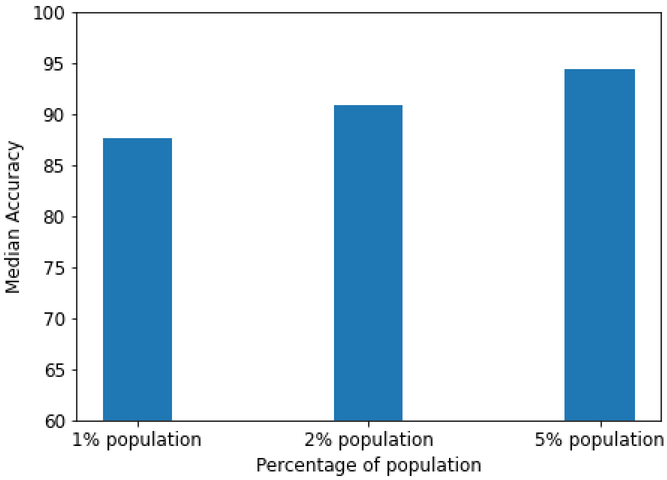

Fig. 7. Median accuracy of the ensemble method's predictions

When predicting the time for the virus to reach 1% of a population ridge regression produced the best results. On the other hand, if it was 2% or 5% of a population then Lasso regression and Elastic Net were more accurate, respectively.

As a pursuit to generate better results, an ensemble method was applied. The ensemble method would be to compute the average prediction between the three regularized linear regression models. However, this only led to similar results and non with higher accuracy from what was previously predicted. Figs. 6-7 illustrate the mean and median predictions accuracy of the ensemble method.

2) Support Vector Machine (SVM) Regression: Moving on now to consider SVM Regression. Utilizing the kernel trick to apply different similarity features leads to better model performance. To investigate this statement a comparison of the results of three SVM kernels was executed.

Figs. 8-9 illustrate a visual comparison between the three kernels used. The median accuracy of the results was close, while the mean accuracy had an increase as the dataset decreased, as in the previous section.

In the 1% population dataset, considerable disparity existed between the mean and median accuracy. This was due to the outliers in the predictions. Meaning that in some samples, the models performed poorly with low accuracy.

The RBF kernel was the most accurate model in terms of

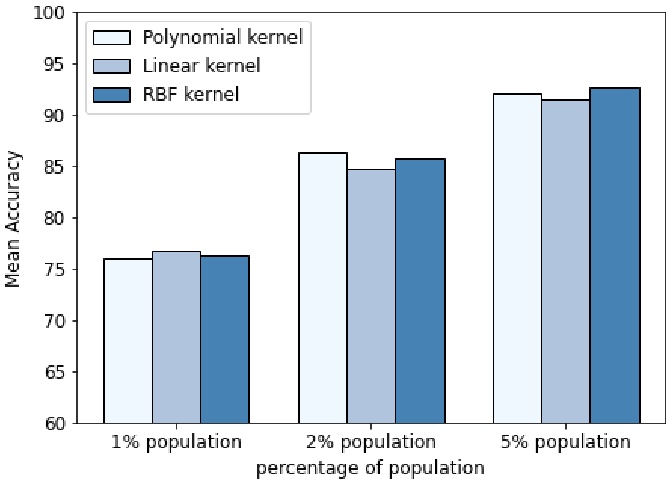

Fig. 8. Mean accuracy of different SVM kernels

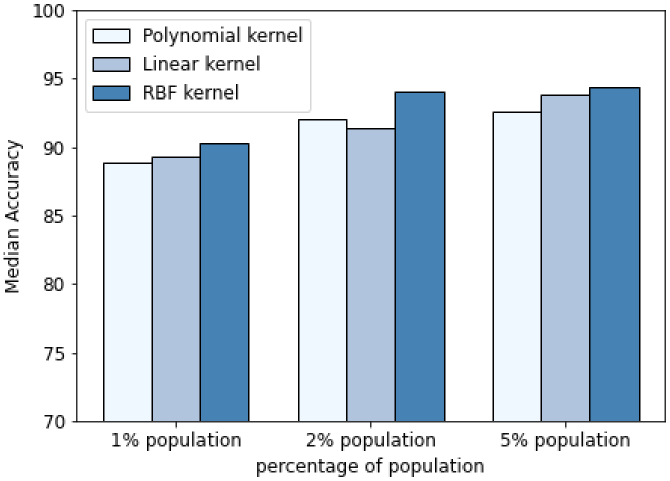

Fig. 9. Median accuracy of different SVM kernels

median accuracy in all datasets. In the 5% dataset, the median prediction accuracy of it was 94.36%, the highest of them all. In the 2% dataset it was 94.01%, while in the 1% dataset it was 90.3%. Regarding the mean prediction accuracy, the RBF kernel's mean prediction accuracy was the highest in the 5% dataset, where it was 92.64%. While the polynomial kernel performed with the highest accuracy, 86.3%, in the 2% dataset.

C. Mortality Rate Prediction

By only using the mentioned features in section III.C, the highest prediction accuracy achieved was by applying ridge regression. The mean accuracy reached was 73.33% and the median accuracy was 82.7%. Other models were not able to predict accurately and produced poor results.

VI. Conclusion

We devised a novel method in assessing the rate at which COVID-19 spreads in a country. Our work has led us to predict, with a median prediction accuracy of 90.3%, 94.01%, and 95.1%, In addition to a mean prediction accuracy of 76.75%, 86.3%, and 93.65%, when the virus' total cases reach 1%, 2%, and 5% of a country's population, respectively. The findings of this study indicate that SVM Regression with an RBF kernel gives the most accurate results.

A mortality rate of COVID-19 cases prediction model was developed using ridge regression, the achieved median accuracy of the model was 82.7% and the mean accuracy was 73.33%.

References

[1] the World Health Organization. WHO announces COVID-19 outbreak a pandemic. Accessed: Oct. 28, 2020. [Online]. Available: https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/news/news/2020/3/who-announces-covid-19-outbreak-a-pandemic

[2] N. Donthu and A. Gustafsson, "Effects of covid-19 on business and research," vol. 117, pp. 1–15, 2020.

[3] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for covid-19 (sars-cov-2) pandemic: A review," *Chaos, Solitons & Fractals*, vol. 139, p. 110059, 2020.

[4] van der Schaar Mihaela, A. M. Alaa, A. Floto, A. Gimson, S. Scholtes, A. Wood, E. McKinney, D. Jarrett, P. Lio, and A. Ercole, "How artificial intelligence and machine learning can help healthcare systems respond to covid-19," *Machine Learning*, 2020.

[5] S. Tuli, S. Tuli, R. Tuli, and S. S. Gill, "Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing," *Internet of Things*, vol. 11, p. 100222, 2020.

[6] V. Singh, R. C. Poonia, S. Kumar, P. Dass, P. Agarwal, V. Bhatnagar, and L. Raja, "Prediction of covid-19 corona virus pandemic based on time series data using support vector machine," pp. 1–15, 2020.

[7] T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict covid-19 infection," *Chaos, Solitons and Fractals*, vol. 140, p. 110120, 2020.

[8] Z. Malki, E.-S. Atlam, A. E. Hassanien, G. Dagnew, M. A. Elhosseini, and I. Gad, "Association between weather data and covid-19 pandemic predicting mortality rate: Machine learning approaches," *Chaos, Solitons and Fractals*, vol. 138, p. 110137, 2020.

[9] A. Atalan, "Is the lockdown important to prevent the covid-19 pandemic? effects on psychology, environment and economy-perspective," *Annals of Medicine and Surgery*, vol. 56, pp. 38–42, 2020.

[10] A. Ahmad, S. Garhwal, S. K. Ray, G. Kumar, S. J. Malebary, and O. M. Barukab, "The number of confirmed cases of covid-19 by using machine learning: Methods and challenges," *Archives of Computational Methods in Engineering*, 2020.

[11] The World Factbook. Accessed: Nov. 12, 2020. [Online]. Available: https://www.cia.gov/the-world-factbook/

[12] the World Health Organization (WHO). Coronavirus Source Data. Accessed: April. 10, 2021. [Online]. Available: https://ourworldindata.org/coronavirus-source-data

[13] F. A. Marioli, F. Bullano, S. Kucinskas, and C. Rondón-Moreno, "Tracking r of covid-19: A new real-time estimation using the kalman filter," *Available at SSRN: https://ssrn.com/abstract=3581633 or http://dx.doi.org/10.2139/ssrn.3581633*, 2020.

[14] I. Vásquez and F. McMahon, *The Human Freedom Index 2020: A Global Measurement of Personal, Civil, and Economic Freedom*. Washington: Cato Institute and the Fraser Institute, 2020.

[15] COVID-19 useful features by country. Accessed: Nov. 12, 2020. [Online]. Available: https://www.kaggle.com/ishivinal/covid19-useful-features-by-country

[16] Countries in the world by population. Accessed: Nov. 12, 2020. [Online]. Available: https://www.worldometers.info/world-population/population-by-country/

[17] the World Bank. GDP per capita. Accessed: Nov. 12, 2020. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD

[18] Youth, Unemployment, GDP and Literacy percentage. Accessed: Nov. 12, 2020. [Online]. Available: https://www.kaggle.com/niyamatalmass/youth-unemployment-gdp-and-literacy-percentage/metadata

[19] A. Geron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media, 2017.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] W. Kirch, Ed., *Pearson's Correlation Coefficient*. Dordrecht: Springer Netherlands, 2008, pp. 1090–1091. [Online]. Available: https://doi.org/10.1007/978-1-4020-5614-7_2569