A DAILY COVID-19 CASES PREDICTION SYSTEM USING DATA MINING AND MACHINE LEARNING ALGORITHM

Yiqi Jack Gao[1] and Yu Sun[2]

[1]Sage High School, Newport Coast, CA 92657
[2]California State Polytechnic University, Pomona, CA, 91768

ABSTRACT

The start of 2020 marked the beginning of the deadly COVID-19 pandemic caused by the novel SARS-COV-2 from Wuhan, China. As of the time of writing, the virus had infected over 150 million people worldwide and resulted in more than 3.5 million global deaths. Accurate future predictions made through machine learning algorithms can be very useful as a guide for hospitals and policy makers to make adequate preparations and enact effective policies to combat the pandemic. This paper carries out a two pronged approach to analyzing COVID-19. First, the model utilizes the feature significance of random forest regressor to select eight of the most significant predictors (date, new tests, weekly hospital admissions, population density, total tests, total deaths, location, and total cases) for predicting daily increases of Covid-19 cases, highlighting potential target areas in order to achieve efficient pandemic responses. Then it utilizes machine learning algorithms such as linear regression, polynomial regression, and random forest regression to make accurate predictions of daily COVID-19 cases using a combination of this diverse range of predictors and proved to be competent at generating predictions with reasonable accuracy.

KEYWORDS

Covid-19 Case Prediction, Data Mining, Machine Learning Algorithm.

1. INTRODUCTION

On January 21st, 2020, Wuhan entered into lockdown after outbreaks of the novel coronavirus, SAR-COV-2 [1, 2] appeared, and not long after, on March 11st, 2020, the novel coronavirus pandemic was declared a global pandemic by the WHO [3]. Known for its severe pneumonia-like symptoms, the coronavirus rapidly spread throughout the world despite various preventative measures such as travel restrictions, social distancing mandates, and lockdowns [4, 5]. At the time of writing, the virus has infected over 150 million people worldwide and led to more than 3.5 million coronavirus related deaths [6]. The pandemic has placed a heavy burden on the world's medical systems, especially those already battling regional instabilities and lacking in adequate sanitation and medical supplies. However, many regions with advanced medical systems, such as Europe and the United States, were also hard hit because policy makers were not presented with enough information on the pandemic to adjust to the rapidly evolving situation [7, 8, 9]. Forecasts of the pandemic's development, though potentially inaccurate, help hospitals and policy makers make preparations to combat the spread of SAR-COV-2. In addition, while the lethality of the novel coronavirus cannot be understated, the virus also devastated the global economy due to repeated nationwide shutdowns and travel restrictions. The issue of shutdowns has since emerged as a heated point of political discourse regarding the efficacies of differing

David C. Wyld et al. (Eds): NLP, MLTEC, CLBD, SEAPP, NeTIoT, VLSIE, ITCS, ARIA, DMS - 2021
pp. 257-268, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112320

legislative approaches towards the pandemic [10, 11]. Therefore, it is necessary to weigh the importance of different factors, natural and legislative, so policy makers can make informed decisions to efficiently combat the current pandemic and future pandemics to come.

Many studies have utilized machine learning algorithms such as regularized ridge regression and deep learning models such as ARIMA and SARIMAX to accurately forecast the spread of COVID-19 [12, 13]. Others have utilized a combination of random forest and Bayesian models to forecast daily COVID deaths and cases using both recorded and forecasted data [14]. While such predictions can be accurate even over longer periods, most only utilize total cases, total days, and previously recorded daily new cases as predictors in their forecasting models. Including more diverse predictors such as population density and daily available COVID tests would theoretically increase the accuracy of such predictions. Using Random Forest Regression to analyze the importance of individual predictors may yield additional insight into how legislative policies and demographic information influences the spread of COVID-19.

In this paper, we follow the same line of research and utilize three regressive models from the sklearn machine learning library—linear regression, polynomial regression, and random forest regression—to predict daily changes in COVID-19 cases based on previously recorded data from the COVID-19 data set maintained by Our World in Data. The prediction models are coded in Python and the data file is read using pandas and modified using a label encoder, along with other manually written codes, to remove nan values and select specific parts of the data file based on its date. In comparison to other works, we fitted our model with a wider range of potential predictors such as the daily amount of COVID tests, ICU capacity, and population density. These additions should increase the accuracy of the predictions if they also have a correlation to the overall trend of recorded COVID-19 cases. The use of random forest regression to analyze the importance of various predictors may yield insight into the effects of aregion's demographic and societal factors on the recorded trends of COVID-19.

The results of the experiment were proven primarily using cross validation where a cross validation score was assigned based on the model, input data, and output data. Additionally, we qualitatively examined the accuracy of the model by making a prediction using the input data from a select day and later proved that the predictions made by the model were reasonably close to the actual recorded values for that day.

The rest of the paper is organized as follows: Section 2 focuses on the difficulties in organizing the code and building the model; Section 3 gives more details on how some of the difficulties were addressed and how the model functions; Section 4 further elaborates on our findings regarding the predictions we made with our model; Section 5 discusses related works.Finally, Section 6 provides concluding remarks and suggested future work for the project.

2. CHALLENGES

In order to use the feature significance of random forest regression to compare the influence of the individual predictors on the general trend of COVID-19, a few challenges have been identified as follows.

2.1. Challenge 1: Missing Data Points and "nan" Values

The data file we selected for the analysis was ambitious in that it included a plethora of relevant data points such as the density of handwashing facilities, the proportion of the population in certain age groups, the proportion of smokers within the population, and other factors. Though a

variety of data points is certainly beneficial for the analysis process, it also presented the greatest challenges in the form of "nan" values, which are added whenever a datapoint is not present for a certain country. Before continuing, it is worth mentioning that the data file referred to is a two-dimensional array with rows representing a day in a certain country and columns representing the data points on that day from each category listed. Since the data file is ambitious in its categories of data, every row of data contains numerous "nan" values where the data points cannot be found. First, we contemplated replacing the "nan" values with the mean value of the values above and below it. However, we discarded the idea since in most situations, the data points did not exist across all the rows representing days in the country. Replacing the datapoints with the mean value of the data points on the left and right would also cause inaccuracies since they would be from unrelated categories. Secondly, we considered replacing all the "nan" values with zeros, but this would also lead to inaccuracies since most data points are "nan" values not because they are zero, but because there are no data points from that category in that region to be found. An example of this is the number of handwashing facilities which exists as a "nan" value in roughly half of the rows. Handwashing facilities are clearly present in virtually every country, and thus replacing "nan" values with zeros in this case clearly does not make sense because it will negatively influence the correlation between the number of handwashing facilities and the number of daily COVID cases. In the end, we decided on a model where all rows containing "nan" values for its columns or categories of interest are removed. Although this is a crude way of organizing data, it is sufficient to generate accurate statistical correlations between categories of interest to which we can apply various models to generate predictions regarding daily COVID cases.

2.2. Challenge 2: The Introduction of Viral Mutations

The second challenge was concerning the approach to data analysis and allowing for the emergence of more infectious and potentially more deadly COVID-19 variants. The core purpose of this research was to compare the influences of various factors on daily COVID-19 cases and generate predictions based on the existing data points. However, the emergence and spread of new variants in late 2020 introduced an additional layer of unpredictability and possible error into the analysis. For example, the spike in COVID-19 cases between September of 2020 to December of 2020 may be attributed to a change in the virulence and contagiousness of the virus, though the spread of mutations operates on similar principles. An analysis of the data points without considering changes in the virus itself may yield a trend that is nonexistent or exaggerated. To account for this, the data points analyzed were divided into different time frames, and specifically focused on earlier time frames between the start of the pandemic until August, when relevant data points such as total cases and daily testing numbers had the clearest correlation with daily increases in COVID-19 cases.

2.3. Challenge 3: Selecting the Preferred Predictive Model

Finally, another challenge was selecting the preferred predictive model to use for analysis. Linear Regression, Polynomial Regression, and Random Forest Regression are easy to work with and potentially effective methods of data analysis. In the end, though it takes more computational power and some data structures are less compatible with certain models of analysis, we decided to use all three models and compare the accuracy of the various predictions generated. We believe having data from three predictive models is valuable to allow for comparison among their analytical methods and accuracy.

3. SOLUTION

In this paper, we focus on predicting the daily increases in COVID-19 cases through various predictors we provide to the model (see Figure 1). The overall experimental process can be separated into four distinct steps: data preparation, the creation of input and output data, prediction and validation, and the recording of results. Pandas is first used to convert the CSV file into a data frame where it is later turned into two dimensional lists through the values.tolist() command. The output depended on the predictors selected, which ranged from date and country name to new hospital admissions and total cases.

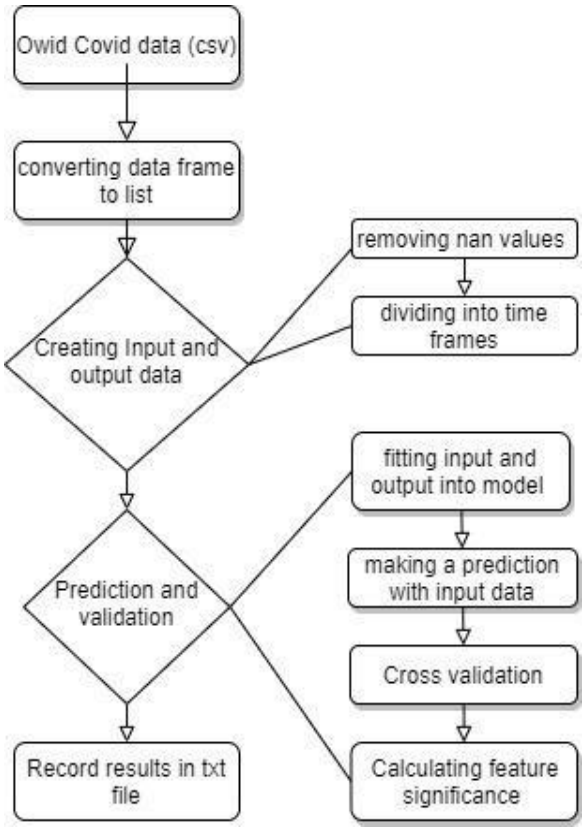

Figure 1. Schematic of the predictive model

Demographic data such as median age, density of handwashing facilities, and population density, along with other factors were also included. The date must be the first variable inputted since this is essential to assigning time frames to other data points. Any non-integer values such as date and country name are converted using the fit_transform function of the label encoder.

```
i = 0
while i < len(output_general):
    if math.isnan(output_general[i]) == True:
        output_general.pop(i)
        #output_general[i] = -1
    else:
        i += 1

print("Done cleaning up NaN values for output")

# Cleaning up NaN values for input
i = 0
while i < len(input_general):
    j = 0
    while j < len(input_general[i]):
        data = input_general[i][j]
        if not isinstance(data, str) and math.isnan(input_general[i][j]) == True:
            input_general.pop(i)
            #input_general[i][j] = -1
        else:
            j += 1
    i += 1
```

Figure 2. Nan values adjusted for output

Cleaning up nan values in each input and output data is done with while loops (see Figure 2). Nan values are found using the isnan() function of the math class, and if the row contains any nan values, the entire row will be removed as seen above in Figure 2. An alternative way of removing nan values is shown in Figure 3, where all nan values are replaced with -1. This is not preferable because it may potentially establish erroneous trends since the machine learning algorithm will use -1 as an input value when the actual value is not -1 and is simply unrecorded. The alternative method of removing nan values is only used when the primary way results in all rows being removed. This occurs with less documented predictors like the number of hand washing facilities or smokers in a population. Dividing the data into time frames requires another algorithm as shown in Figure 4. The input and output data are divided into three separate lists depending on experiment number: the function will output data from the start of pandemic to August of 2020 with an experiment number of 1; August of 2020 to December of 2020 with an experiment number of 2; and December of 2020 to latest time available, which by the time of testing was early March of 2021, when the experiment number is 3. The function is called as a part of the function of each individual model. This is done to account for the unquantifiable effects of the differing virulence of the different strains of COVID-19 such as the UK and South African variants, which were more deadly than the original COVID-19 variant. However, this difference is not seen within the separate tests possibly because the virulence of the different COVID-19 strains has little impact on the overall correlation between some predictors and daily COVID cases.

```
def organize_dates(experiment, input_data, output_data):
    new_input_data1a = []
    new_output_data1a = []

    new_input_data1b = []
    new_output_data1b = []

    new_input_data1c = []
    new_output_data1c = []
    for i in range(len(input_data)):
        date_ = input_data[i][0]
        year = int(date_[0:4])
        month = int(date_[5:6])
        if (year < 2021):
            if (month < 8):
                new_input_data1a.append(input_data[i])
                new_output_data1a.append(output_data[i])
            else:
                new_input_data1b.append(input_data[i])
                new_output_data1b.append(output_data[i])

        else:
            new_input_data1c.append(input_data[i])
            new_output_data1c.append(output_data[i])
```

Figure 3. Nan values replaced with -1

```
if experiment == 1:
    return new_input_data1a, new_output_data1a
elif experiment == 2:
    return new_input_data1b, new_output_data1b
elif experiment == 3:
    return new_input_data1c, new_output_data1c
else:
    return -1
```

Figure 4. Dividing data into time frames

The functions for each of the three machine learning algorithms are almost identical to the one shown in Figure 5, which shows the function for linear regression. First the model is defined to be linear regression and the organize_dates function is called, outputting two lists for input and output data depending on the time frame selected as indicated by the experiment number. Then the new_input_data and new_output_datais used to train the model. The different models are set up appropriately based on their specific design. The polynomial regression is implemented initially with a polynomial feature of 2, although that is changed if the trend can be more accurately predicted with a different value for the polynomial feature. The random forest regressor is implemented with max depth of 5 and random state of 0 (mostly due to limited computational power of the laptop used). The model is then used to make a prediction through the predict() function with a list of predictors. As this article is focused mainly as a proof of concept, the data given is a mix of real-world data selected from a date outside of the timeframe and realistic generated data meant to test the model's accuracy. To test the accuracy of the test, we decided to use K-fold cross validation at 5 folds instead of the default accuracy_score since the acuracy_score is more beneficial for analyzing classifiers rather than the regressive models used in the experiment. All predictions and cross validation scores are stored in a txt file to be accessed later for data analysis.

```
def linear_regression(test, experiment_number):
    print("Running Test 1A Linear Regression...")
    model = linear_model.LinearRegression()
    new_input_data, new_output_data = organize_dates(experiment_number, input_data, output_data)
    model.fit(new_input_data, new_output_data)
    # y_pred = model.predict(input_data)
    # accuracy_score(output_data, y_pred)
    print("Linear Prediction", model.predict([test]))
    score = cross_val_score(model, input_data, output_data, cv=5)
    score = score.mean()*100
    f = open("test_1A_results.txt", "a")
    print("Test 1A:\nLinear Regression\n Cross validation score =" + str(score))
    f.write("Test 1A:\nLinear Regression\n Cross validation score =" + str(score))
    f.close()
```

Figure 5. Function for linear regression

Finally, Figure 6 shows the main function that ties all the functions together and makes the experiment easier to debug and implement. The experimental process involves calling the main_experiment function and inputting the model type (model selection) and the experiment number (time frame selection) and results will be printed to the screen and recorded in a txt file. An example of this is shown in Figure 7.

```
def main_experiment(model_type, experiment_number):
    model_type = model_type.lower()
    test_data = le.fit_transform(prediction_test)
    print(test_data)
    if model_type == "linear regression":
        linear_regression_1a(test_data, experiment_number)
    elif model_type == "polynomial regression":
        polynomial_regression_1a(test_data, experiment_number)
    elif model_type == "random forest regression":
        random_forest_regression_1a(test_data, experiment_number)
    else:
        print("Not a valid model")
```

Figure 6. Main function that ties all other functions together

```
main_experiment("linear regression", 1)
main_experiment("polynomial regression", 1)
main_experiment("random forest regression", 1)
```

Figure 7. Results (txt file)

4. EXPERIMENT

The first part of the experiment aimed to test the feature significance of various predictors using feature significance. The model is used to perform predictions of daily COVID-19 cases based on real and artificially made data. Then, the feature significance of each element within the list of values used to make the prediction is calculated by a feature within random forest regression and given a value of 0 through 1. A value of 0 indicates that the predictor's correlation with the trend is statistically insignificant compared to the other predictors and a value of 1 indicates that the predictor can be used as the sole input value for the predictions. Out of all of the 66 potential predictors, eight were chosen as final contenders due to their feature significance, availability in the datafile, and similarity to output. Predictors such as number of handwashing facilities or vaccinations were not chosen despite moderate correlation due to their limited number of data points. Predictors such as new cases per million were also not used to predict new cases because they were too similar and thus would not reflect the model's ultimate goal. The eight predictors for predicting daily new cases were: date, new tests, weekly hospital admissions, population density, total tests, total deaths, location, and total cases. All these predictors, with few exceptions, are dynamic predictors, which is reasonable considering the value of the output is constantly changing.

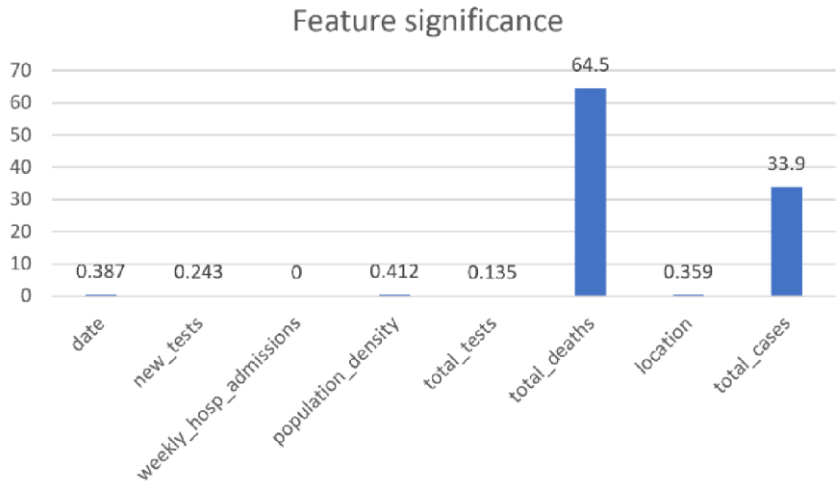

Figure 8. Mean value of feature significance after 10 tests with various realistic artificial data

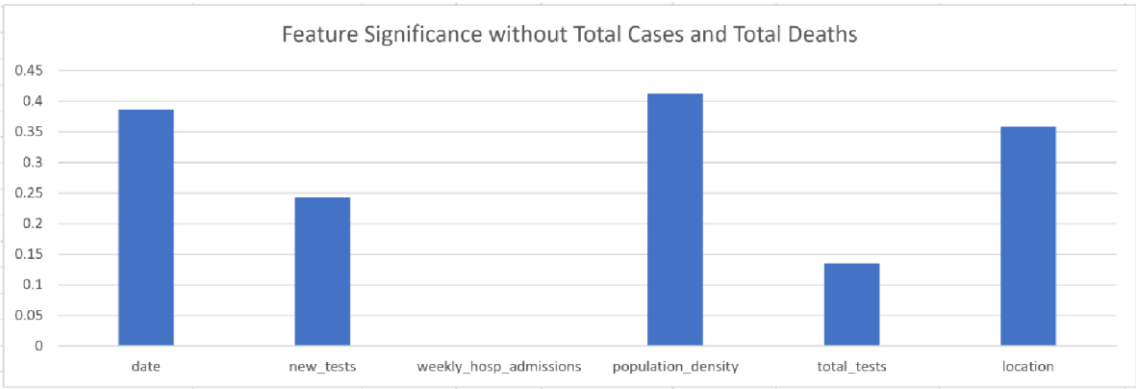

Figure 9. Mean value of feature significance without total deaths and total cases

Figure 8 shows the mean value of feature significance after 10 tests with various realistic artificial data. Figure 9 shows the mean value of feature significance without total deaths and total cases.

As seen Figures 8 and 9, the total cases and deaths combined is overwhelmingly more significant to the prediction than other factors. This is likely because predictors such as location and population density are stagnant. These predictors likely have a strong influence on the rate of increase of new cases rather than how many new cases there are. Predictors such as new tests, weekly hospitalizations, and total tests suffered due to the lack of data. Despite selecting the predictors with availability in mind, the preferred way of removing nan values removed every row within the list. Replacing every nan value with -1 dramatically harmed the accuracy. The date had a low feature importance since it increases linearly while daily new cases increases nonlinearly. The feature significance of total deaths was almost that of total cases. This was a surprise since there is often a considerable lag between cases and deaths. Our initial hypothesis was that total deaths would be a better predictor since deaths are better documented than cases and thus total deaths might be considered a better predictor than total cases for predicting adiseases' spread in a region. Further analysis is needed to provide a conclusive answer.

Daily Case Predictions using Real and Realistic Data

date	new tests	weekly hosp admissions	population density	total test	location	total deaths	total cases	linear prediction	polynomial prediction	random forest perdiction	actual value
300	155	122	50	400000	10	3000	150000	1564	2227	2407	N/A
300	1500	1220	50	1500000	10	20000	1500000	11733	15699	14981	N/A
100	120	65	50	150000	10	2000	30000	838	722	271	N/A
10	20	1	50	15000	10	0	12	533	-84	90	N/A
210	938376	N/A	35.608	1.16E+08		204531	7081895	76028	74200	46780	44673

Figure 10. Predictions

The second part of the experiment focuses on predicting daily Covid increases with the predictors shown in Figure 10. Displayed are five predictions made by each of the three models based on five sets of input, four artificially created and one taken from the USA on September 26th. It can be seen that the predictions made are mostly reasonable with random forest regression being the most accurate both logically and in terms of direct comparisons as seen with the predictions made with the fourth and fifth set of values.

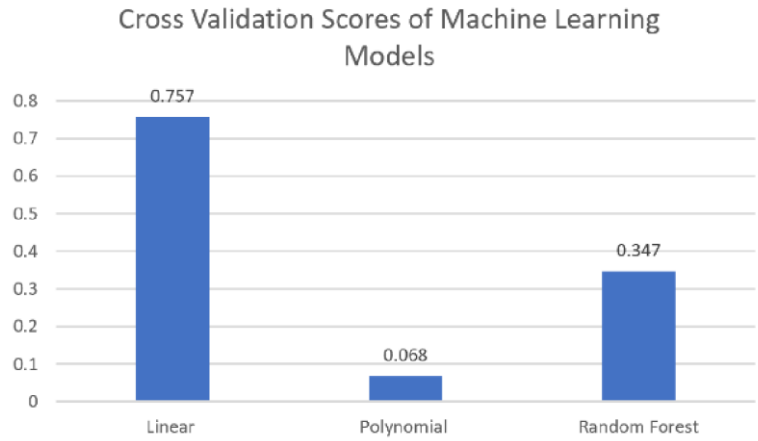

Figure 11. Results

Interestingly, K-fold cross validation values varied dramatically within the three regressive models despite little noticeable difference, especially between linear and polynomial regression, despite the two models having little appreciable difference in the predictions made. However, polynomial regression struggles to make predictions with numerically very small or very large inputs so that likely accounted for this significant difference. Additionally, the random forest regressor, despite being consistently more accurate when assessed qualitatively, received a much lower cross validation score than linear regression. This is likely due to overfitting where linear regression excels at making predictions based strictly on the training data, but is not as effective as random forest at analyzing new data outside the range of training data given to the model (See Figure 11).

The results of experiment one identified the eight most important predictors of daily increases in COVID-19 cases and the hierarchy of influences for each predictor within the selected group. Despite potential sources of error, this analysis contributes insight into the heated debate regarding the public's perception of the pandemic since it reveals that new tests and total tests do not have as significant an impact on new cases as some have suggested. This makes sense logically since the positivity rate remained relatively low on a national level for all intermediate periods of the pandemic and thus additional tests should not account for a significant influx of COVID cases. The results additionally highlight areas for future research in predicting the trends of total deaths and total cases. As they are the major predictors of daily increases in COVID-19 cases, factors that can significantly influence total deaths and total cases will likely have a significant impact on the spread of the pandemic as well. The second experiment shows the accuracy of the regressive models to make predictions on new cases of COVID when provided with both real and artificial data. Ultimately, despite certain inaccuracies and shortcomings, these experiments achieved their purpose of demonstrating the ability of machine learning models, especially the random forest regressor, to analyze and predict the trends of COVID-19, which affirms their potential in the field of epidemiology.

5. RELATED WORK

Solanki and Singh examined different machine learning models such as regularized ridge regression and deep learning (ARIMA and SARIMAX) to make predictions for COVID-19's spread in the form of cases and deaths over a set time period, which they did with commendable accuracy [15]. In particular, the study utilized a ridge regression model with given predictors such as "days since the first case," as well as "growth factors and growth ratios," which they estimated using a ridge polynomial regression to create a projection of active COVID cases in India over a seven-month period. In comparison, our experiment used regressive models to analyze a wider range of predictors such as daily available COVID-tests and hospitalizations to make next-day predictions rather than longer time periods.

Painuli, et al. developed a predictive model for individual COVID-19 cases in the major states of India using the ARIMA model and past COVID-19 infections with random forest and the extra tree classifier as well as a forecasting model of COVID-19 trends [16]. Those models were able to accomplish their respective tasks with accuracy. While our experiment is similar in our attempt to create a predictive model for COVID-19 cases, we utilized different machine learning models—polynomial regression, linear regression, and random forest regression—as opposed to univariate regression analysis models such as ARIMA. In addition, our experiment aimed to explore the differing significance of correlations between various predictors and the predicted values through random forest regression.

Watson, et al. utilized a robust combination of a Bayesian model for a location specific trajectory of COVID-19 and the random forest model for death predictions all within a compartmental

model to accurately forecastCovid cases, deaths, and recoveries based on observed and projected data. This parallels our attempt to predict daily Covid cases, though our methodology differs in that we focused on purely regressive models. In addition, we utilized a greater diversity of predictors in an attempt to not only produce accurate predictions, but also to compare the feature significance of the predictors.

6. CONCLUSION AND FUTURE WORK

In this paper, we explored a method of predicting daily COVID-19 cases with three regressive machine learning models—linear regression, polynomial regression, and random forest regression. These models were trained using a diverse variety of predictors ranging from total cases of COVID-19 and date to demographic data like median age and population density. The feature significance was calculated through random forest regression to compare the influence of the various predictors on the prediction being made. In the experiments, random forest regression and polynomial regression performed much better than linear regression, as confirmed by both comparisons with real world data and cross validations scores. This is reasonable because the trends are not linear in nature and an overall decreasing trend or increasing trend would lead the linear regression model to predict a number that is either far too small or far too big depending on the selected time frame. Ultimately, the purpose of the models is not to pinpoint accurate forecasts of a specific region, but rather to identify overarching, pandemic-related trends on a national level, estimate the daily increases in COVID-19 cases to a reasonable degree, and recognize the significant influencers of these trends. Though lacking specificity and precision, holistic views of the pandemic are arguably as important as precise community-specific forecasts, since they provide general insight into the national-level factors influencing the disease's spread through a diverse area. The handling of a pandemic requires effective local- and national-level responses, but knowledge of holistic trends may benefit decision making to help manage future, national-level pandemics.

REFERENCES

[1] Bernard Stoecklin, Sibylle, et al. "First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020." Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin vol. 25,6 (2020): 2000094.

[2] "China Coronavirus: Lockdown Measures Rise across Hubei Province." *BBC News*, BBC, 23 Jan. 2020, www.bbc.com/news/world-asia-china-51217455.

[3] Mahase, Elisabeth. "Covid-19: WHO Declares Pandemic Because of 'Alarming Levels' of Spread, Severity, and Inaction." *BMJ*, 2020, p. m1036.

[4] Jia, L., Li, K., Jiang, Y., Guo, X.: Prediction and analysis of coronavirus disease 2019. arXiv preprint arXiv:2003.05447 (2020)

[5] "Transmission of SARS-CoV-2: Implications for Infection Prevention Precautions." *World Health Organization*, World Health Organization, 9 July 2020, www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infect ion-prevention-precautions.

[6] "Coronavirus Cases:" *Worldometer*, www.worldometers.info/coronavirus/.

[7] Kulick, Debbie. "COVID Has Had an Impact on Emergency Medical Service Providers: Something to Think About." *Pocono Record*, Pocono Record, 28 Apr. 2021, www.poconorecord.com/story/lifestyle/columns/2021/04/28/debbie-kulick-covid-has-had-impact -emergency-medical-services/4860410001/.

[8] Rubin, Rita. "COVID-19's Crushing Effects on Medical Practices, Some of Which Might Not Survive." *JAMA*, vol. 324, no. 4, 2020, p. 321.

[9] "COVID-19 and the Least Developed Countries | Department of Economic and Social Affairs." *United Nations*, United Nations, 1 May 2020,

www.un.org/development/desa/dpad/publication/un-desa-policy-brief-66-covid-19-and-the-leastdeveloped-countries/.

[10] Munywoki, Gilbert. "Economic Effects of Novel Coronavirus (COVID – 19) on the Global Economy." *SSRN Electronic Journal*, 29 Oct. 2020.

[11] Boettke, Peter J., and Benjamin Powell. "The Political Economy of the COVID-19 Pandemic." *SSRN Electronic Journal*, 12 Feb. 2021

[12] Solanki, Arun, and Tarana Singh. "COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms." Emerging Technologies for Battling Covid-19: Applications and Innovations 16 Feb. 2021, vol. 324 57–78.

[13] Painuli, Deepak et al. "Forecast and prediction of COVID-19 using machine learning." Data Science for COVID-19 (2021): 381–397.

[14] Watson, Gregory L., et al. "Pandemic Velocity: Forecasting COVID-19 in the US with a Machine Learning & Bayesian Time Series Compartmental Model." PLOS Computational Biology, vol. 17, no. 3, 2021.

[15] Solanki, Arun, and Tarana Singh. "COVID-19 Epidemic Analysis and Prediction Using Machine Learning Algorithms." *Emerging Technologies for Battling Covid-19: Applications and Innovations* vol. 324 57–78. 16 Feb. 2021.

[16] Painuli, Deepak et al. "Forecast and prediction of COVID-19 using machine learning." *Data Science for COVID-19* (2021): 381–397.

[17] Watson, Gregory L., et al. "Pandemic Velocity: Forecasting COVID-19 in the US with a Machine Learning & Bayesian Time Series Compartmental Model." *PLOS Computational Biology*, vol. 17, no. 3, 2021