# Fighting COVID-19 with Machine Learning

Anjali Agrawal
Center for Data Science, NYU
aa7513@nyu.edu

Armanda Lewis
Center for Data Science, NYU
aal861@nyu.edu

Dhara Atul Mungra*
Center for Data Science, NYU
dam797@nyu.edu

Praxal Suresh Patel
Center for Data Science, NYU
psp334@nyu.edu

Claire Saint-Donat
Center for Data Science, NYU
csd261@nyu.edu

## 1 INTRODUCTION

Since the worldwide outbreak of the Novel 2019 Coronavirus (COVID-19) in early 2020, over 4 million have contracted the disease, and over 300,000 succumbed. The disease has impacted economies on scales not seen since World War II, and altered daily life for billions of people. In response, data scientists and others have raced to discover patterns and build predictive models that will contribute to mitigating the detrimental health, financial, and social effects of this pandemic disease. Within only a few months, an impressive array of datasets, visualization dashboards, and scholarly articles have emerged to guide what we know about the disease and how we may fight it together [1–3, 5, 7, 9, 10].

We were motivated by data science's essential role in combating COVID-19 and aim to develop improved models for predicting the presence of COVID-19 in key countries, including Brazil, China, France, Germany, India, Italy, Russia, Spain, the United Kingdom, and the U.S.. We have two main goals: we will use an existing dataset and existing models that incorporate various features to develop an improved model that predicts COVID-19 cases. We will also look more closely at data on Non-pharmaceutical Interventions (NPIs), or "actions, apart from getting vaccinated and taking medicine, that people and communities can take to help slow the spread of illnesses like pandemic influenza (flu). NPIs are also known as community mitigation strategies"[6]. We will determine if features that rely on NPIs yield better models.

## 2 PROBLEM DEFINITION AND ALGORITHM

### 2.1 Task

We analyze the cumulative data of confirmed, deaths, and recovered, cases across the globe over time to analyze the spread trend of the coronavirus all over the world. We use the data for the previous days to predict the active cases for the next day. To understand the impact of the non-pharmaceutical interventions on the goal of flattening the curve, we also use various policies and interventions employed by the federal and the state government to predict the effect of these interventions on preventing the adversaries caused by the coronavirus.

### 2.2 Algorithm

For the experiments using CSSE data we have used following classifiers with the mentioned hyperparameter choice:

- Linear Regression(LR): Linear regression is used with normalized data and fit_intercept is set to False as we do not want to add any constant since our data is already centered.
- Support Vector Regression (SVR): Hyperparameter search using RandomizedCV is performed for SVR to search over hyperparamter space mentioned in Table 1.

| Hyperparameter | Value |
|---|---|
| iter | 10,20,30,50 |
| kernel | linear, rbf, polynomial |
| c | 0.01, 0.1, 1 |
| gamma | 0.01, 0.1, 1 |
| epsilon | 0.01, 0.1, 1 |
| shrinking | True, False |

**Table 1: SVR hyperparameters**

- Random Forest Regression (RFR):Hyperparameter search using RandomizedCV is performed for RFR to search over hyperparameter space mentioned in Table 2.

| Hyperparameter | Value |
|---|---|
| n_estimator | 50, 100, 150 |
| criterion | mse,mae |
| min_samples_split | 2, 3, 5 |
| max_depth | 3, 5, 7 |
| max_samples | 4, 6, 8 |

**Table 2: RFR hyperparameters**

- Bayesian Ridge Regression (BRR):Hyperparameter search using RandomizedCV is performed for BRR to search over hyperparameter space mentioned in Table 3.
- AutoRegressive Integrated Moving Average Model (ARIMA): Transforms a time series into stationary one(series without trend or seasonality) using differencing technique.

For the NPIs + CSSE combined data that looks at the impact of NPI policy, we utilized:

- Linear Regression: we predicted confirmed cases using linear regression regression line to identify the minimal sum of the squares of the differences between our NPI categorical data, rate of increase of cases, and other relevant data such as the size and nature of the county.

*Member Responsible for sumbission

| Hyperparameter | Value |
|---|---|
| tolerance | 1e-6, 1e-5, 1e-4, 1e-3, 1e-2 |
| alpha_1 | 1e-7, 1e-6, 1e-5, 1e-4, 1e-3 |
| alpha_2 | 1e-7, 1e-6, 1e-5, 1e-4, 1e-3 |
| lambda_1 | 1e-7, 1e-6, 1e-5, 1e-4, 1e-3 |
| lambda_2 | 1e-7, 1e-6, 1e-5, 1e-4, 1e-3 |

**Table 3: BRR hyperparameters**

- Bagging: To increase the robustness and accuracy of our regression model, we utilized bagging techniques. Parameters included:
  - n_estimators: enacted with a gridsearch to be with 1-75
  - max_samples: 20
  - max_features: default
  - n_jobs: defaults
  - random_state: set to true in order to compare different models

# 3 EXPERIMENTAL EVALUATION

## 3.1 Data

We have used two principle datasets. The first is the COVID-19 data repository of The Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, which aggregated various datastreams from U.S. city, state and federal governments, the World Health Organization, the China CDC, and other global data [10]. The other is from Keystone Strategy and contains NPI data at the county level for a selection of counties within the United States [8].

CSSE dataset is a time-series data from January 22, 2020, for all the countries in the world, that reported the COVID-19 cases, and is updated every day to include the most recently available data. The dataset has information about the total number of affected patients, recovered patients, and deaths in a country for each day. Using this data, we are generating four features namely-'Confirmed cases', 'Deaths', 'Recovered', and 'Day'. 'Day' here represents the number of days, having values in the range 0-118, that has passed from the date 22 January 2020. 'Confirmed cases', 'Deaths', and 'Recovered' are obtained by summing the values of these features for all the countries for a single day. 'Confirmed cases', 'Deaths', and 'Recovered' are further used to get the total number of active cases in a day using the given below equation:

$$active\_cases = confirmed\_cases - deaths - recovered$$

'Active cases' are then used as a label for the dataset and its value is predicted using machine learning classifiers. Since the scale of the number of cases is huge, we have transformed the features using centralization and scaling. Here, we only use the train test to standardise the data and then use the mean and variance of the train data to transform the test data to avoid any bias in the test data.

The Keystone Strategy NPIs data from April 16, 2020 consists of U.S. county data (4,226 rows) that contains the institution of policies to curb the spread of the disease. Though NPIs may include person interventions, such as frequent washing of hands and even informal group-based conventions such as general social distancing, the NPI

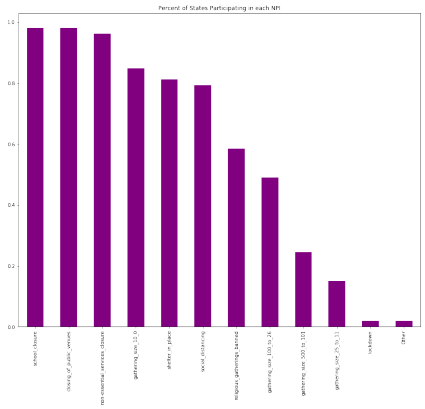


**Figure 1: NPI**

dataset highlights official government-sponsored NPI policies instituted publicly and at large scale. There are twelve main categories of NPIs utilized in the dataset: including limitation of gatherings of various sizes, closing of non-essential businesses, school closures, shelter-in-place orders, and even lockdown status. These data include the start dates when various NPIs were instantiated, and they cluster around mid-March 2020 (see Figure). There is a column for NPI policy end dates, though the latter has yet to be relevant. Of note is that the NPI data are meant to be used in conjunction with case-specific datasets such as the CSSE dataset mentioned above. In this report, we link CSSE data to the NPI data.

## 3.2 Exploratory Analysis

To understand the global impact of the coronavirus, we use various visualization tools to identify the regions that were affected by the pandemic. Figure 3 shows the countries that were affected by the virus. We analyse the data to find the countries that were highest hit by the pandemic using the number of confirmed cases and the number of fatalities in each country which is outlined in Figure 2 and 9 in Appendix A. Then, we analyse the countries with the highest number of recovered cases depicted in Figure 10 in Appendix A. We also explore the data to identify other factors as hospitalizations, number of hospital beds, testing rate etc. to understand the various aspects that could affect the predictions of the number of cases. This helps us in understanding the general trend of the increase and the effect of several interventions in preventing the spread of the virus. We also use the NPI data for each county in US to see how many counties implemented a specific policy and also compare these counties in terms of the number of cases within those dates.

## 3.3 Methodology

For the experiments using CSSE data we are splitting the data into train and test set into 80-20 ratio without shuffling. After splitting into train test set data is normalized. Train data is further split into 3 cross validation folds without shuffling to search over hyper parameter space and find the ones that gives the best results. We have used LR, RFR, SVR with linear, rbf and polynomial kernel, and BRR with degree 2 and 3. Model performance is evaluated

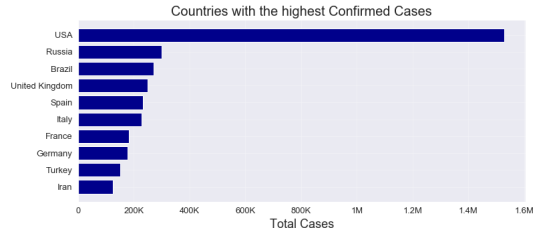


**Figure 2: Confirmed Cases**

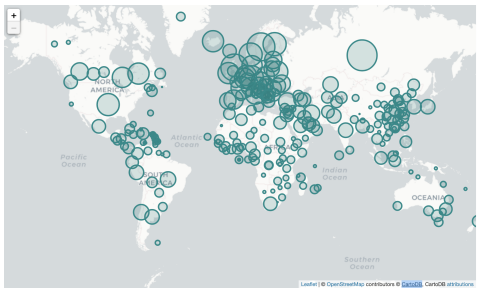

**Figure 3: Global Spread of the Coronavirus**

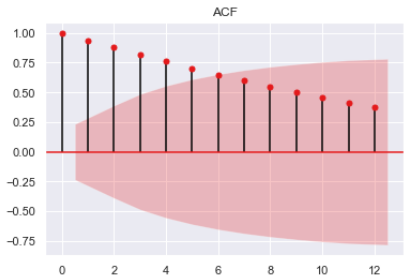


**Figure 4: ACF for global time series data**

using mean average error (MAE), mean square error (MSE), and $R^2$ metrics. Choice of these evaluation metrics is due to the following reasons:

- MAE - these method for evaluation is more robust to large change in values and in the experiments conducted for this project we observe exponential increase in the number of confirmed cases.
- MSE - we use this metric to aid MAE as we have exponentially increasing data and want model to capture this trend. Hence, if MSE value is too high than either our model is predicting very large or small value which not ideal.
- $R^2$ - This metric is used to measure the confidence level of the values predicted by the classifier.

data processing and modelling, linear models give best performance as after an initial exponential growth in the cases, number of cases increases linearly with some multiplicative factor with each passing day. Through this experiment we aim to identify how well our model learns from the previous data and predicts value for the next day based on these data.

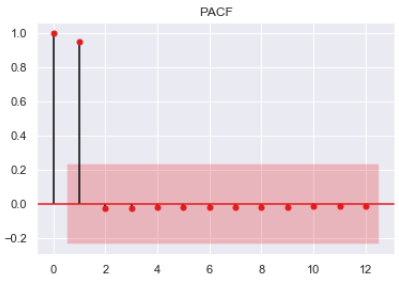


**Figure 5: PACF for global time series data**

We also tried to forecast the number of confirmed cases and deaths for a given period of time in the future. Given a time series data, we planned on implementing ARIMA which Transforms a time series into stationary one(series without trend or seasonality) using differencing technique. We also tried to explore the similarities between different countries based on the similarity in the trend of new confirmed cases. For analysis of how ARIMA model performs, we performed tests on global level as well as country level for certain countries where major laws had been framed and non pharmaceutical interventions had been carried for combating the spread of the pandemic. The time series we used was the number of confirmed cases for every day (strictly increasing as the previous cases were added to the current tally). Firstly, we decomposed our time series data for finding its properties. The seasonal decomposition using moving averages was performed using seasonal decompose function from Python library 'statsmodels'. Further, we employed the Dickey Fuller Test to check the stationarity of our data. We also plotted the ACF (Auto Correlation Function) and PACF (Partial Auto Correlation Function) as shown in Figure 4 and 5 while analyzing our time series to ensure that there was no random noise but presence of some pattern in the data. For feeding our model, we created features like Day of the week, week of the year, month and year from the date column.

## 3.4 Analysis of Results and Discussion

Prediction results from different classifiers for CSSE data is shown in the Figure 6. Best results are obtained for SVR with linear kernel followed by LR and BRR with degree 2. SVR with hyperparameter settings of 'shrinking':False, 'gamma': 1, 'epsilon': 0.01, and 'C': 1 gives predictions with 97 percent confidence having MAE 0.05 and MSE 0.008. MAE, MSE and R2 results for all the classifiers are mentioned in the Table 4.

As depicted in the Figure 6 and Table 4 results obtained in after modelling aligns with the hypotheis we proposed as linear models outperforms other models. It might be because of this reason that SVR gives best results with linear kernels and BRR gived best results with polynomial degree 2. High MAE and MSE scores with low $R^2$ value for RFR might be because of overfitting as we have small number of datasamples with few features. We trained our ARIMA model on the global data and calculated model results. We obtained the best Akaike information criterion (=312.548) results for ARIMA(5,2,5). We tried 216 different combinations for different values in the range of (0,6) for p (number of autoregressive terms),

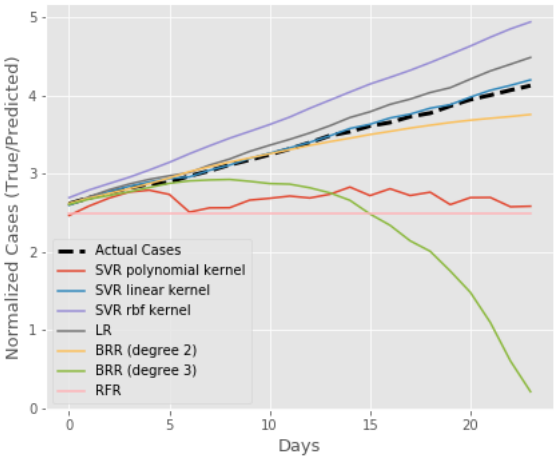


**Figure 6: Prediction Plot**

| Classifier | MAE | MSE | $R^2$ |
|---|---|---|---|
| LR | 0.15 | 0.03 | 0.84 |
| SVR | 0.05 | 0.008 | 0.97 |
| RFR | 0.87 | 0.97 | -3.5 |
| BRR | 0.1 | 0.02 | 0.89 |

**Table 4: BRR hyperparameters**

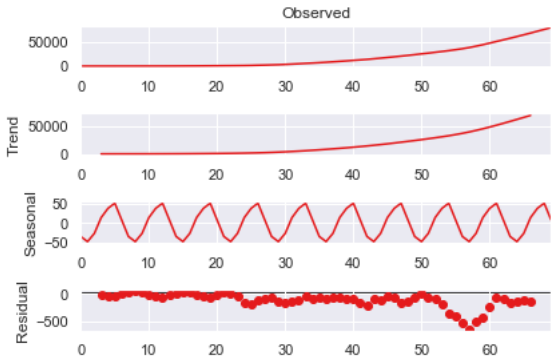


**Figure 7: Decomposition of confirmed cases for India**

d (number of nonseasonal differences needed for stationarity) and q(number of lagged forecast errors in the prediction equation).

Further, for identifying how Non Pharmaceutical Interventions had affected the number of confirmed cases, we trained our ARIMA model country wise and followed the same steps that we had performed for the global data. We analyzed for different countries like USA, China, France, Spain, Italy and India. The model gave reasonable results for most countries but we noticed something peculiar for India. While decomposing the time series for India, the residuals were really low for the initial days but the residuals peaked negatively for the later days as shown in the Figure 7. We inferred that because of the stringent rules and lockdown enforced

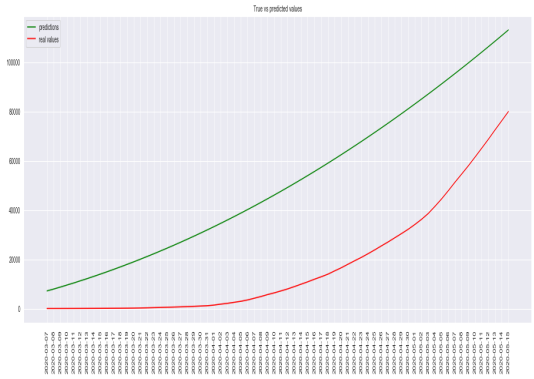

**Figure 8: Predictions vs Actual Cases when trained for data until 25th March 2020**

in India after March 25 (59 days), our model predicted higher values than the actual confirmed cases which decreased than the usual trend because of the intervention. We checked the particular time series of confirmed cases of India for stationarity and obtained p value of 0.983515 for Dickey Fuller Test, thus accepting the hypothesis about stationarity of data. We further trained the ARIMA model for data until 25th March 2020 (ie. not incorporating the data after nationwide lockdown). The predictions of the best model (ARIMA(1,2,5)) for the confirmed cases obtained were much higher than the actual values as shown in the Figure 8 , further validating our inference that lockdown had led to significant decrease in the number of cases. We obtained an AIC of 545.325 for the ARIMA(4, 2, 3) model. Subsequently, we included information till 11th April 2020, for observing if our model can learn better when we incorporate information of confirmed cases during the lockdown period. We obtained a much lower AIC value (=352.767) for much lesser test set (inversely proportional to the data size). From the experiments we suggest that Non Pharmaceutical Interventions worked in the favor of reducing the number of confirmed cases. The results for NPI+CSSE combined data using Linear Regression and Bagging are available on the mentioned Github repository.

## 4 CONCLUSION

For this work we are working on two datasets CSSE and NPI. We have done data processing, exploratory analysis, and modelling for three cases- CSSE, NPI, and CSSE and NPI both. Hypothesis proposed from the exploratory analysis and results obtained after modelling for CSSE data aligns with each till date. For CSSE data linear prediction models gives the best performance given the size and the nature of the data. These results, analysis and hypothesis might change given the precautionary measures taken as indicated in the results using ARIMA.

In terms of shortcomings, our data are being generated in real time, and thus we are largely analyzing a moving target. Many of the protocols in place to measure COVID-19 cases and to prevent cases (both pharmaceutical and non-pharmaceutical interventions) are actively being created and updates on a daily basis. Another shortcoming is that we would have to look more closely at regional

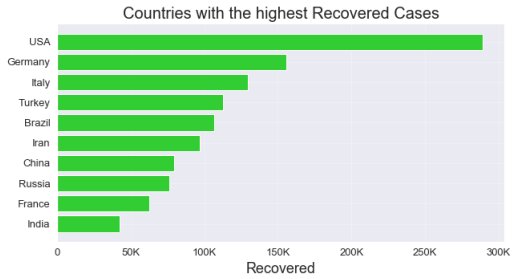


**Figure 10: Recovered Cases**

similarities that may make a more nuances exploration and analysis useful. The analysis of COVID-19 data has much potential for future studies, including integrating multidiscipinary data into our analysis, exploring additional ensemble methods, and delving more deeply into the reliability of certain datasets. More recently, reports of the benefits of citizen science and community-driven data have been highlighted [4]. We relied on numerical and related categorical data in our model, but there remains a host of features that would likely generate more robust explanatory and predictive models.

## REFERENCES

[1] [n.d.]. The COVID Tracking Project. https://covidtracking.com/about-data
[2] Sheryl L. Chang, Nathan Harding, Cameron Zachreson, Oliver M. Cliff, and Mikhail Prokopenko. 2020. Modelling transmission and control of the COVID-19 pandemic in Australia. https://arxiv.org/abs/2003.10218
[3] Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. 2020. The COVID-19 Social Media Infodemic. https://arxiv.org/abs/2003.05004
[4] Luca Ferretti, Chris Wymant, Michelle Kendall, Lele Zhao, Anel Nurtay, Lucie Abeler-Dorner, Michael Parker, David Bonsall, and Christophe Fraser. 2020. Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science* 368, 6491 (May 2020). https://doi.org/10.1126/science.abb6936
[5] NYU Center for Data Science. [n.d.]. NYU Mobilizes Against COVID. https://cds.nyu.edu/against-covid/
[6] Centers for Disease Control and Prevention. [n.d.]. Nonpharmaceutical Interventions (NPIs). https://www.cdc.gov/nonpharmaceutical-interventions/index.html
[7] University of Virginia Biocomplexity Institute. [n.d.]. COVID-19 Surveillance Dashboard. https://nssac.bii.virginia.edu/covid-19/dashboard/
[8] Keystone Strategy. [n.d.]. covid19-intervention-data repository. https://github.com/Keystone-Strategy/covid19-intervention-data
[9] The New York Times. [n.d.]. covid-19-data Repository. https://github.com/nytimes/covid-19-data/blob/master/PROBABLE-CASES-NOTE.md
[10] Johns Hopkins University. [n.d.]. Novel Coronavirus (COVID-19) Cases, provided by JHU CSSE. https://github.com/CSSEGISandData/COVID-19

# 5 APPENDIX

## A FIGURES FOR RECOVERED CASES AND DEATH CASES

Figure 9 and 10 show the highest death and and recovery cases.

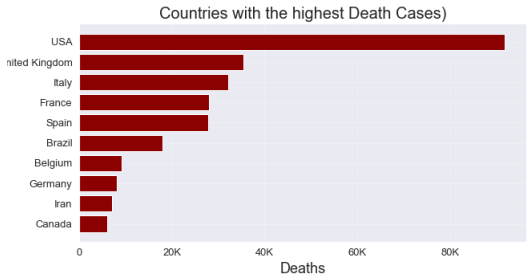


**Figure 9: Death cases**

## B REPRODUCIBLE RESULTS

The code for every detail is available on Github: https://github.com/aa7513/ML-Project-COVID-19