COVID-19 Analysis and Prediction of Death

Arushi Sharma, Spatika Krishnan, Vidhey Oza
Northeastern University
DS 5110 Project Report

April 20, 2020

1 Summary

COVID-19, known as the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), belongs to a family of coronaviruses. The outbreak was first identified in Wuhan, China in December 2019 and has been recognized as a pandemic by the World Health Organization on March 11, 2020. The virus is spread between people primarily via respiratory droplets produced during coughing.

1.1 Goals

We aim to perform statistical analysis and use exploratory and confirmatory data analysis to explore relationships between different underlying factors related to the pandemic. We also create visualizations to answer some questions like - the most affected countries, growth rate of confirmed cases, recovery and deaths for different countries and days under which first death happened country-wise among others. We analyze and predict the progress of COVID-19 in the USA and build a model to predict the death of COVID-19 patients. The aim of this prediction is to facilitate the healthcare establishments to prioritize patients for specialized medical attention, at an early stage.

1.2 About Data

We have taken the dataset [4] from Kaggle open dataset, which was then logically separated into two separate datasets. The first one consists of day-wise data points on the number of affected cases, deaths and recovery from COVID-19. This data is available from 22nd January, 2020 to 8th April, 2020. Whereas the second one has patient-wise data points on confirmed cases, and their current status as infected, recovered or dead. This data has information of about 260,000 patients. Both files are being updated daily. Few of the features actively used for analysis and modeling are country, states, confirmed/death/recovered cases, observation date, age, sex, symptoms, travel history, etc.

1.3 Project Overview

The dataset covers a holistic view of how COVID-19 has impacted different countries and people. The project covers multiple analysis that uses the given data to explore the trend of COVID-19 confirmed cases, recovery and deaths across countries. From exploratory data analysis, we see which countries took preventive measures and were able to contain the spread of the disease and flattened the curve by actively maintaining social distancing. We also observe which countries are struggling the most and where the growth of cases is exponential. Further, the project used SIR modeling, a popular epidemiology model that predicted growth and decline of the number of confirmed cases and deaths over the course of the epidemic for USA. Our second model predicts the fatality outcome of patients based on extracted features from the dataset like age, gender, COVID symptoms, etc. Random Forest was the best model to predict the death. Holistically, the project provides insights on where we stand right now as a community.

2 Methods

2.1 Data Cleaning and Preprocessing

The first step before doing exploratory and confirmatory data analysis is to clean and preprocess the data to bring it in a format from where it can be used for further analysis. In our COVID-19

country-wise dataset, we first checked for missing values and found that only *Province/State* has a missing value. We imputed it with a constant because this variable is necessary for visualizing data. We then converted *"Observation Date"* and *"Last Update"* object to DateTime format and *"Confirmed, Recovered and Death Cases"* to a numeric format. We observed an anomaly in the date format of the *"Observation Date"* column and to fix that, we selected only those rows and converted them to a proper date format. We also created new variables for values which were hidden in the existing data. For example, if Taiwan was the province then its country would also be Taiwan whereas in our data it was given to be China. In our COVID-19 patient's data, we formatted the *"age"* and *date* columns to their respective formats and either removed the missing values or added them in a separate column to avoid losing out on data while analyzing.

2.2 Feature Selection

It is essential to engineer new features that uncover hidden patterns buried inside the data before applying machine learning algorithms on it to find better data insights. To predict the progress of COVID-19 in the United States, we engineered few features from our country-wise dataset namely *Country/Region*, *Province/State*, *Confirmed*, *Recovered and Death cases*. We also calculated the population of each state using 2016 US census data [3] to find the cases per capita which was used in our modeling. Table 1 shows a sample of the features created using Cases and Deaths of different affected states in the United States.

States	Cases	Cases Per Cap	Deaths	Deaths Per Cap	Mortality
New York	139875	706.59	5489	27.73	3.924%
New Jersey	44416	495.82	1232	13.75	2.774%
Louisiana	16284	348.64	582	12.46	3.574%
Massachusetts	15202	223.74	356	5.24	2.342%
Connecticut	7781	216.69	277	7.71	3.560%
Michigan	18970	191.18	845	8.52	4.454%
District of Columbia	1211	180.15	22	3.27	1.817%
Washington	8692	121.22	400	5.58	4.602%
Rhode Island	1229	116.35	30	2.84	2.441%
Pennsylvania	14853	116.02	247	1.93	1.663%

Table 1: State wise Analysis of USA

For COVID-19 patient's dataset, we created few features namely *hasCOVIDSymptoms*, *daysFrom-FirstOccurrence*, *travelHistory* and a binary-valued response variable, *death*. We plotted all the selected features using correlation matrix. Correlation matrix is a table that shows correlation between variables and enables us to see the mutual relationship of each feature with the response variable. As the dataset consists of both categorical as well as continuous variables, we used hetcor package of R to plot it. Figure 7b in appendix shows the correlation plot for the filtered features of CoVid-19 patient's dataset. We observed that the features having the strongest correlation with death are *age* and *hasCOVIDSymptoms*. For selecting the features to be included in the model, we analyzed the relation between different features against death and finalized *age*, *sex*, *travel-history*, *hasCOVIDSymptoms* and *daysToFirstoccurence* variables to go as predictor variables for our initial model to predict death. Most of the variables were transformed into factors and the dataset was split between train and test data in the ratio 70:30, as shown in C.5.

2.3 Modeling

2.3.1 SIR

Our first model is SIR [8] which is an epidemiology model that predicts the development of the disease in the population of the USA. The SIR model is a simple linear system of differential equations that models the population compartmented into Susceptible(S), Infectious(I) and Recovered(R). The

equations are:

$$\begin{aligned} \frac{dS}{dt} &= -bSI/N \\ \frac{dI}{dt} &= bSI/N - cI \\ \frac{dR}{dt} &= cI \end{aligned} \tag{1}$$

Here, N represents the total population considered. One condition that SIR equations must satisfy for all t is:

$$N = S + I + R \tag{2}$$

As there was no data for the susceptible column, this data was derived using the above condition based on the number of infected and recovered people. So, the value of N is required to make an assumption of the total "at risk population in the USA". The model itself assumes an equal probability of infection that everyone within the tally of N would be equally exposed. b and c are coefficients to be determined. A final assumption is that the SIR state is one-way i.e. a susceptible would progress to become infected and then to recovered.

2.3.2 Death Prediction

Our second model predicts fatality outcome of a COVID-19 patient. This is a binary classification problem and we used various algorithms to model it. The majority class prediction was used as the baseline model, which always output majority class label. So, the baseline model will always predict no death for all the patients. We use Logistic Regression [6], Support Vector Machine (SVM) [2] and Random Forest [1] to model. We compared the performance of each model by evaluating on sensitivity, specificity and accuracy on the binary classification task. As the label distribution was heavily skewed, i.e highly imbalanced class distribution (refer C.5), we also treated this problem as novelty detection and modeled using SVM one-class classification [7].

Logistic Regression is similar to Linear Regression but for a categorical response. It is a type of generalized linear model.

$$Y_i = \beta_0 + \beta_1 X_i + ... + \beta_n X_n \tag{3}$$

Here, Y_i is the response variable. X_i to X_n are the features used for modeling. β_0 is the death intercept and β_1 to β_n are feature coefficients.

SVM is a linear model for classification and regression problems and it creates a hyperplane that separates the data into classes. A good separation is achieved by the hyperplane that has the largest distance to the nearest training-point of any class, since in general the larger the margin, the lower is the generalization error of the classifier. We used different kernels using the function *svm* in R [5] to see which kernel works best for this dataset while varying the hyper-parameters.

$$g(x) = w^T x + b \tag{4}$$

Maximize k(RHS) such that $-w^T x + b >= k$ for $d_i == 1$
Here, x is the vector of features and w represents hyperplane.

We used **Random Forest** for their high accuracy and the ability to handle features with small samples without overfitting. It creates decision trees on randomly selected data samples, gets a prediction from each tree and selects the best solution by means of voting.

$$\sum_{i=1}^{C} f_i(1 - f_i) \tag{5}$$

Here, f_i is the frequency of label i at a node and C is the number of unique labels.

SVM One-class classification was used for predicting death as it can be used for binary classification task with a severely skewed class distribution. This model can be fit on the input examples from the a class in the training dataset, then evaluated on a holdout test dataset and is quite effective for imbalanced classification dataset where there are none or very few examples of the other class.

3 Results

3.1 Exploratory Data Analysis

3.1.1 Most Affected Countries

Figure 1 represents the top five countries that are most affected by COVID-19 based on the count of confirmed cases. From the plot, we could note that US has the highest number of confirmed cases followed by Spain, Italy, France and Germany. The stats are represented for the data till 8th April. We also observed that China was in the plot of top five countries initially when COVID-19 started spreading but later it flattened the curve of the disease and saw a decline in the number of cases. This happened as they implemented a proper execution of testing & social distancing.

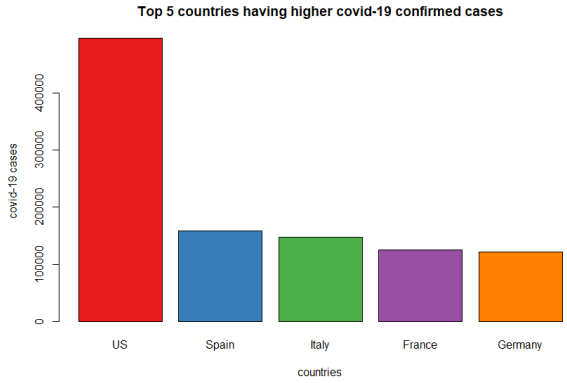
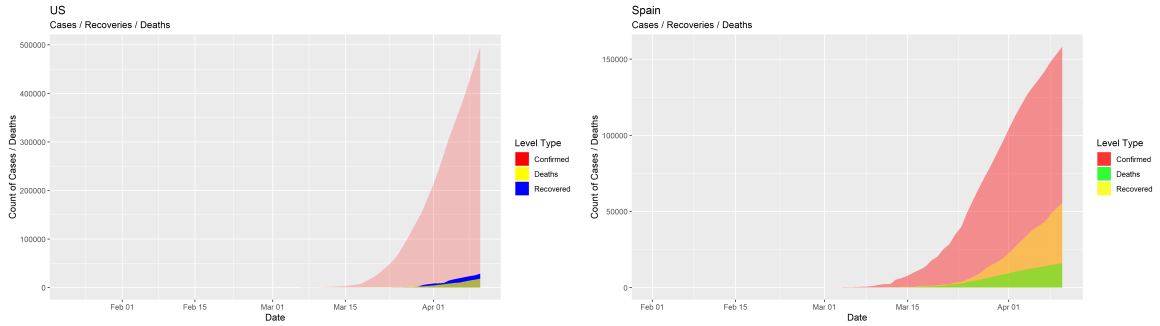


Figure 1: Most Affected Countries

3.1.2 Country-wise Analysis of Cases, Recoveries and Deaths

We analysed the confirmed, recovered and death cases for each of the countries. From figure 2a, we can infer that USA's response to COVID-19 has been slow as it has the least recovered to confirmed ratio whereas Spain (Figure 2b) and Germany (Figure 11a in the appendix section has a good recovered to confirmed ratio. Italy (Figure 10a saw the most number of deaths till 8th April.



(a) Cases, Recoveries and Deaths in USA

(b) Cases, Recoveries and Deaths in Spain

Figure 2: Case details in USA and Spain

3.1.3 Comparing containment of countries (Cases)

Figure 3 compares the containment of countries with respect to the number of daily confirmed cases among the ten most affected countries. We can note that USA, UK, Turkey are still on the path of exponential growth whereas Italy, Spain, Germany and Iran are attempting to contain the growth effectively. China has shown tremendous response within 3 months by reducing the number of daily new cases to zero. We can also infer that even though France saw a high drop in the number of new cases but it is still experiencing deaths (Refer Figure 10b in Appendix).

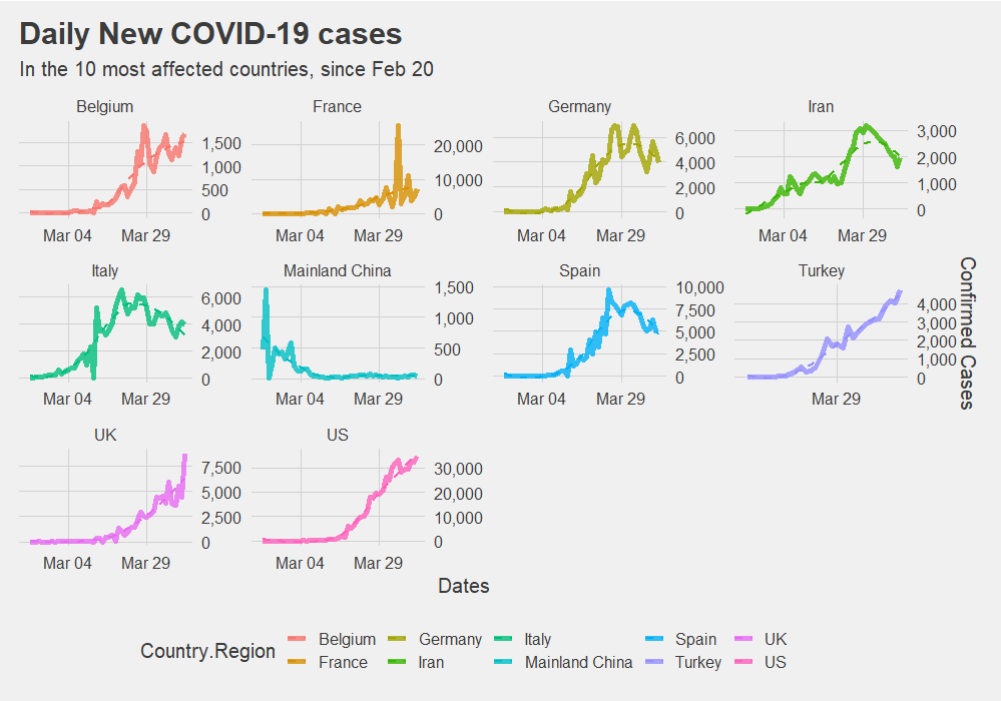


Figure 3: Daily New Cases

3.1.4 Removing time from the axes

Based on what we learned about the data and the growth rate of cases worldwide and country-wise, we realized that the number of cases added everyday (or new cases) is more correlated with the total number of cases on that given day than with progression of time. We used this simple idea to design a unique plot shown in Figure 4 that clearly shows the current position of five countries.

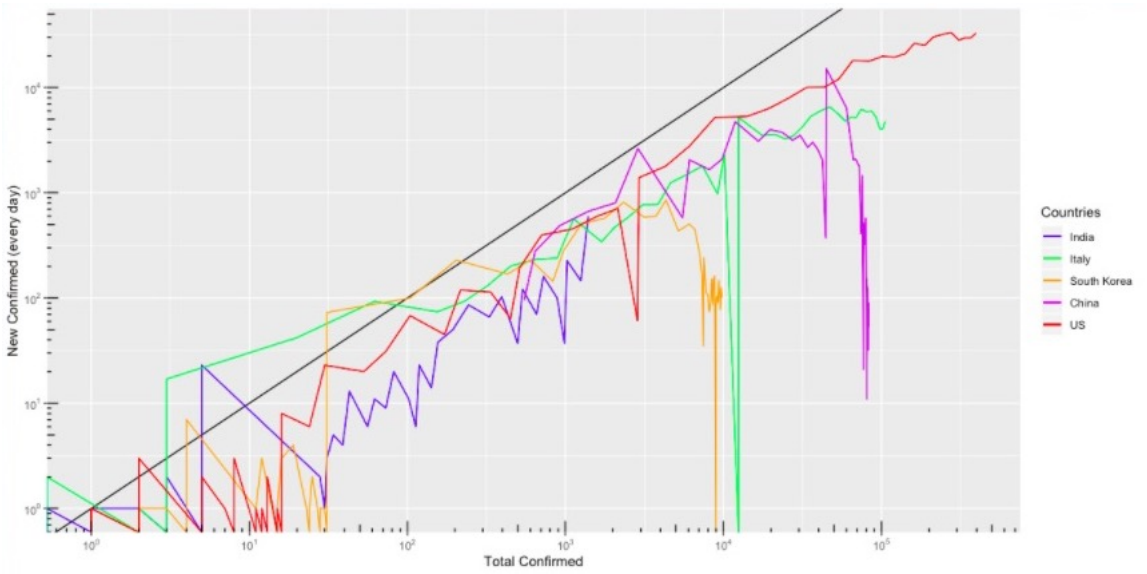


Figure 4: Total Confirmed Cases vs Daily New Cases

The first is the point zero of contact: China. On 22nd January, it had around 600 cases which quickly spreaded throughout the country. But they managed to test and isolate as fast as possible, and currently, they have fallen off the exponential line (marked in black) in Figure 4. The next stories are failures to contain the spread: US and Italy. We clearly see that South Korea is a success story. With their previous experience in handling epidemic spread during the MERS outbreak in

the last decade, their system was well designed to test and isolate any suspected cases, which they started doing before the number of cases reached one hundred.

3.2 Modeling

3.2.1 COVID-19 Modeling for the USA using SIR model

SIR model is an epidemiological model which assumes a closed population over time. This model was applied to compute the expected number of people infected with COVID-19 in the USA population, which was taken from 2016 US census data.

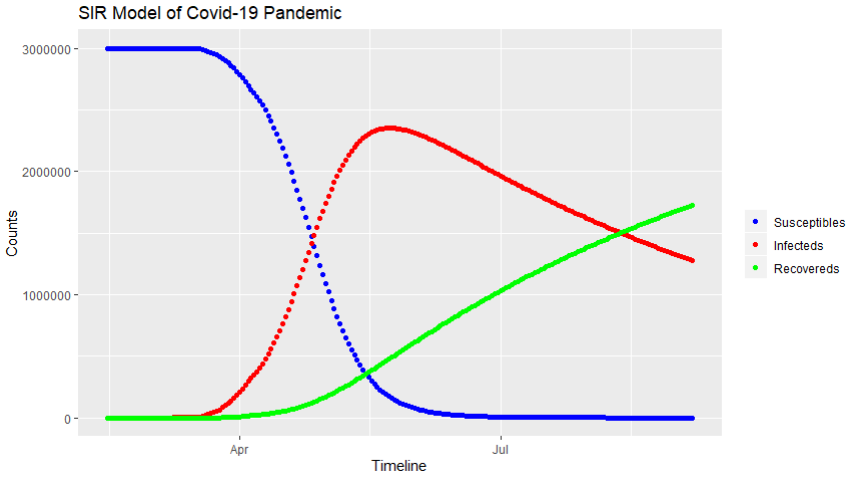


Figure 5: SIR PLOT of the USA

Fig. 5 shows the output of the SIR model. The peak of infection appears to be in the mid of May, with the total infected cases going approximately to 2.6 million. This wave depicts that we are likely to see a decline by the end of may and it would tail off towards August, assuming no further separate waves of infection occur. We will now look at the actual data shown in Table 2 to compare with our predictions.

Dates Predicted	Infected Predicted	Infected Actual	Recovered Predicted	Recovered Actual
2020-04-07	1115373	1501931	476034	474261
2020-04-08	1611038	1545036	504738	511019
2020-04-09	1682091	1610540	534771	542107

Table 2: Prediction to Actual cases

The Table 2 shows that the infection rates (including death) and recovery rates appears to be much closer to the actual numbers due to the increased information now available from the pandemic progression. This proves that our graph for SIR plot is an accurate prediction of what we are going to see in the future.

3.2.2 Death Prediction

The models used for predicting death are Logistic Regression, SVM, Random Forest and SVM one-class classification. For death prediction, sensitivity is the ability of a test to correctly predict no death (True Positive Rate), whereas specificity is the ability of the test to correctly predict death (True Negative Rate). In our case, for a model to perform well, it's specificity needs to be high which indicates we are predicting death correctly.

In figure 6, we see that Random Forest performed the best out of all applied models. We experimented with SVM one-class classification model and saw an increase in specificity but accuracy and sensitivity decreased.

Table 3 shows the performance of all models. In SVM, we used different kernels namely polynomial, sigmoid and radial with hyperparameters tuning. SVM's radial kernel performed the best.

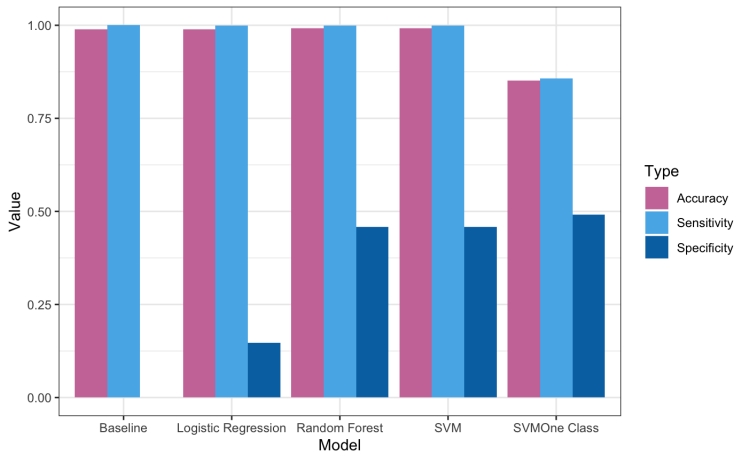


Figure 6: Models

Model	Accuracy	Specificity	Sensitivity
Baseline	0.9893	0.0000	1.0000
Logistic Regression	0.9884	0.1475	0.9997
SVM ($kernel : polynomial$)	0.9915	0.3934	0.9995
SVM ($kernel : sigmoid$)	0.9775	0.0163	0.9905
SVM ($kernel : radial$)	0.9921	0.4590	0.9993
Random Forest	**0.9923**	**0.4590**	**0.9995**
SVM ($one - class$)	0.8520	0.4918	0.8569

Table 3: Performance of different models on *test* split

4 Results & Discussion

The exploratory data analysis showed that all countries are on different stages of containing the virus but the dread of exponentiality has hit every country. SIR model predicts that USA is yet to see the peak likely in the first week of May but there is an expected decline in the infected cases by the end of the May. In this project, we used SIR modeling to predict only for the USA as it is the most affected country with the highest number of cases. There is a scope to calculate the total risk of population(N) of the world and this model can be tuned further to predict the progress of COVID-19 globally.

Random Forest performed best to predict death of patients among all models that we tested on but we can focus more on model optimization techniques and hyper parameter tuning next time to account for better performing models. Secondly, the classifier model can be designed in such a way that the class imbalance problem (Fig. 14 from Appendix) bears no influence on the overall performance of the model. In this project, we modeled using SVM one-class classification to handle data imbalance but the results weren't good. That can be improved if we handle the temporal information of the disease in a better way. Currently we are finding the disease duration for the patient by taking the difference of the first reported date with the patient's confirmed date of being COVID positive. There is also a scope of having an interactive dashboard that provides a user friendly interface to interact and analyze the data in a much efficient way.

5 Statement of Contributions

The team worked on the topic research and created a concrete plan from the dataset to achieve feasible goals. The project consists of two different analysis and modeling on each type of dataset i.e country-wise and patient-wise.

- **Spatika Krishnan** worked on data cleaning and pre-processing of the country-wise dataset. She also did exploratory data analysis and SIR modeling to predict the progress of COVID-19 in United States.

- **Vidhey Oza** worked on data pre-processing and focused on analyzing the country-wise trends of confirmed cases and daily new cases. He also worked on exploratory data analysis of new cases as a measure against total cases.
- **Arushi Sharma** was responsible for data cleaning and pre-processing of patient-level dataset. She analyzed and implemented different modeling techniques to see which model predicts the patient's death with more accuracy.

References

[1] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[3] Census Data. 2016 us census data. `https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2016/`.

[4] Kaggle. Novel corona virus 2019 dataset. `https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset`.

[5] Medium. Support vector machines(svm) — an overview. `https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989`, Jun 16, 2018.

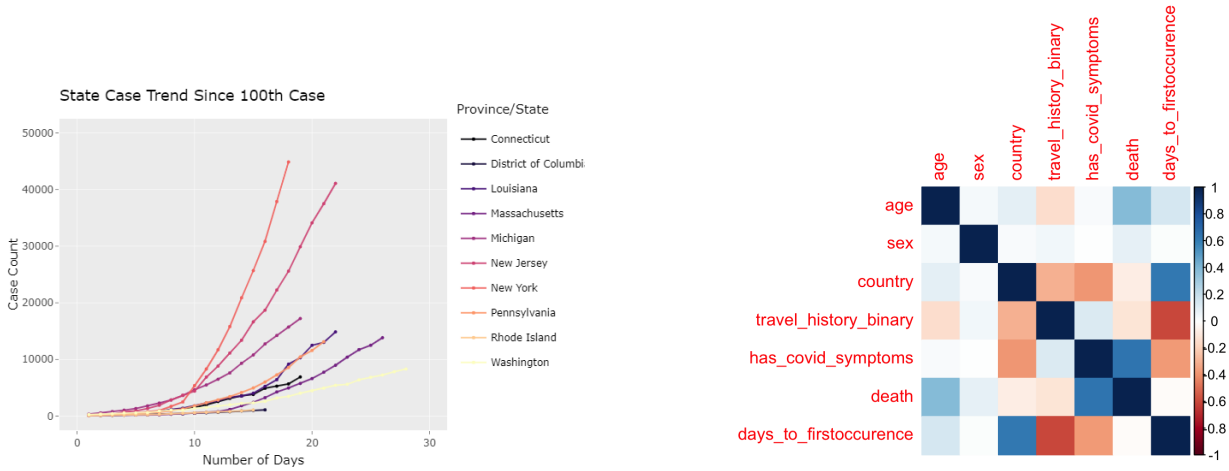
[6] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[7] Bernhard Schölkopf, Robert Williamson, Alex Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 582–588, Cambridge, MA, USA, 1999. MIT Press.

[8] David Smith and Lang Moore. The sir model for spread of disease: The differential equation model. *Loci.(originally Convergence.) https://www. maa. org/press/periodicals/loci/joma/the-sir-model-for-spread-of-disease-the-differential-equation-model*, 2004.

Appendix

A Feature Selection



(a) State wise trends of US

(b) Correlation Matrix

Figure 7: Feature Selection

B EDA

B.1 Confirmed Cases and Deaths Worldwide

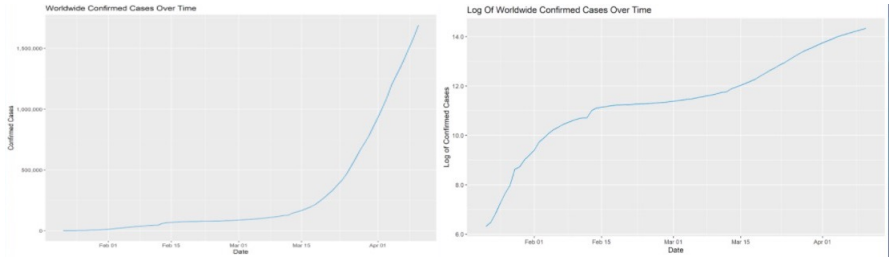


Figure 8: Total Confirmed Cases Worldwide

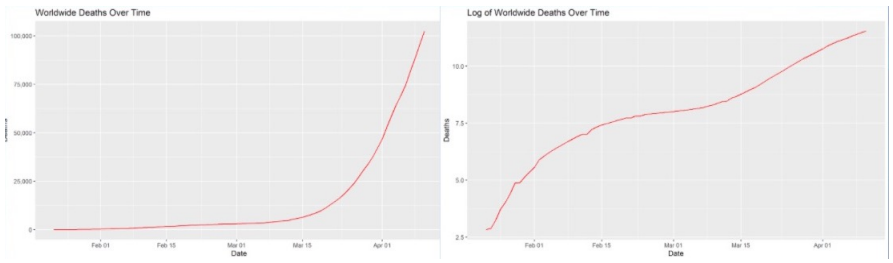


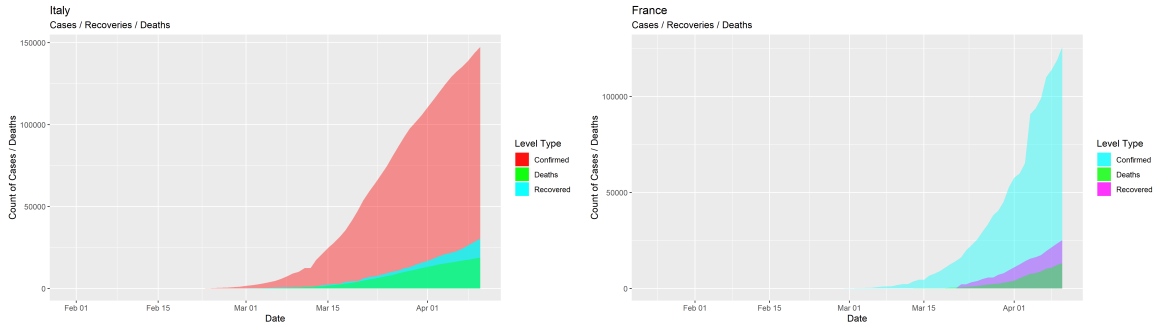
Figure 9: Total Deaths Worldwide

Fig. 8 contains two subplots. The left plot is the total confirmed cases vs time, which is increasing exponentially, as all pandemics begin. But with that exponentiality we also lose interpretability, since the earlier half looks negligible in front of the latter half so no proper trend analysis can be done, and due to the steep increase it is really difficult to know whether we've escaped the exponential part of growth yet. To solve that, the graph on the right includes the same data but plotted on a

logarithmic scale. A linear graph here shows exponential growth, so it is much easier to detect when we do in fact grow slower than exponentially.

Fig. 9 from Appendix is a similar 2-subplot of total worldwide deaths. An interesting thing to note here is the similarity of graphs in Fig. 8 and Fig. 9. This signals to a near constant ratio of confirmed vs deaths, which has been observed to be at around 5-10% in different countries.

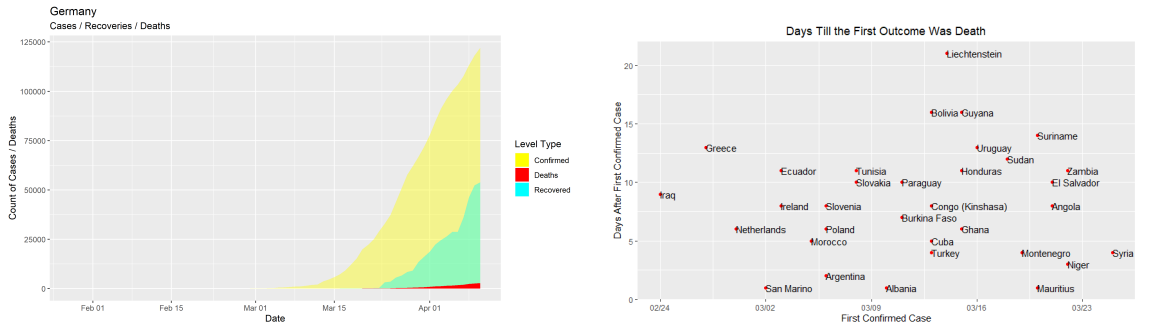
B.2 Cases, Recoveries and Death Cases in Different Countries



(a) Cases,Recoveries and Deaths of Italy

(b) Cases,Recoveries and Deaths of France

Figure 10: Cases in Italy and France



(a) Cases,Recoveries and Deaths of Germany

(b) First death across countries

Figure 11: Cases in Germany and First Death across Countries

B.3 First death across countries

The figure 11b represents the countries that faced their first death few days after their first confirmed case. The plot shows many third world countries where the situation worsened because of the inadequacy in testing and isolation scheme. There are also some European countries like Ireland, Netherlands who were not initially prepared to face this pandemic.

B.4 Comparing containment of countries (Recovered)

The figure 13 compares the containment of countries with respect to the number of daily recoveries occurring among the ten most affected countries. Among other countries there was an inclusion of South Korea in having the most number of recovered cases. This counts as a success story in this pandemic as South Korea has managed to recover all their patients by effective testing and isolation.

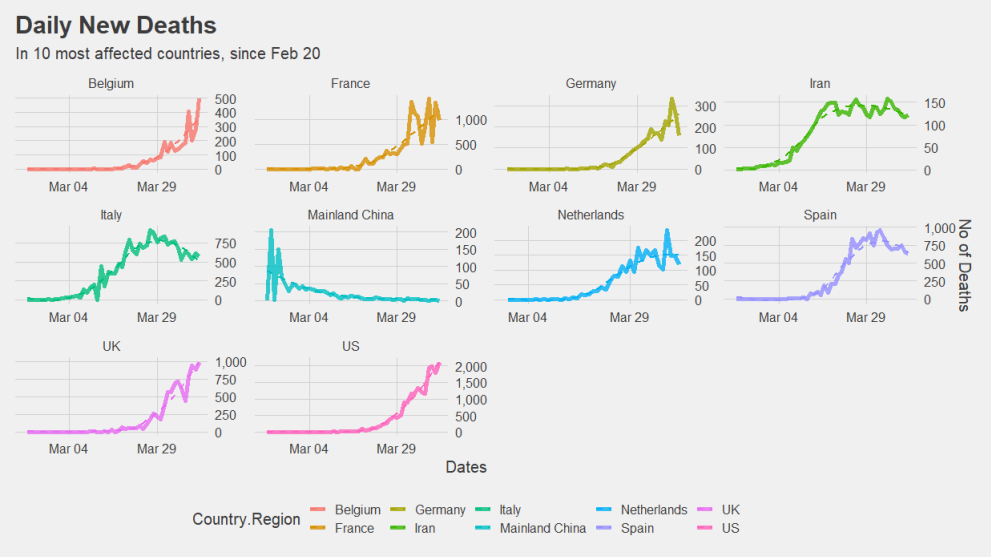


Figure 12: Deaths among countries

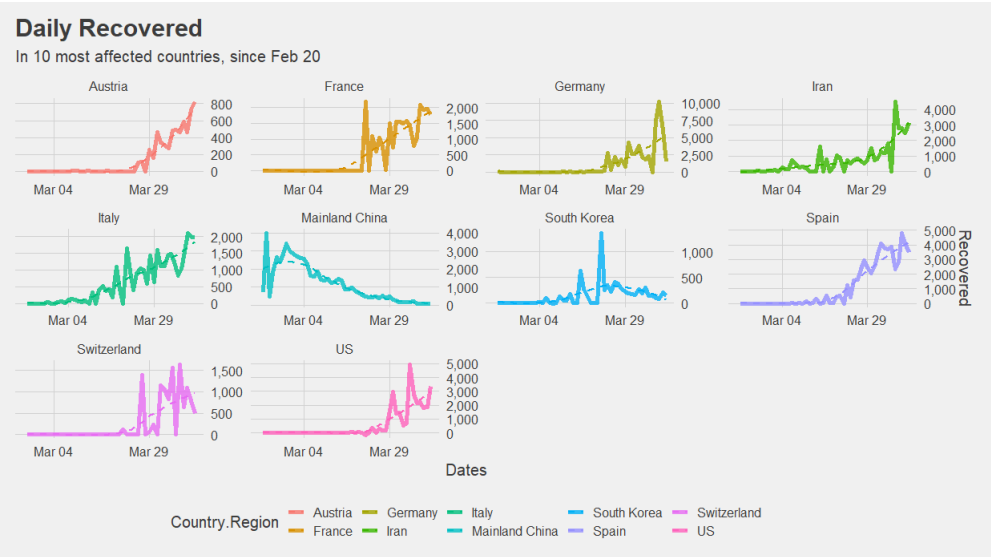


Figure 13: Recoveries among countries

C Modeling Statistics

C.1 Logistic Regression

```
Call:
glm(formula = death ~ age + sex + has_covid_symptoms + days_to_firstoccurence,
    family = binomial(link = "logit"), data = trainingData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6118  -0.1089  -0.0623  -0.0408   4.0049

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -10.951156   0.538204 -20.348   <2e-16 ***
age                       0.080994   0.006647  12.185   <2e-16 ***
sexmale                   0.502198   0.196342   2.558   0.0105 *
has_covid_symptoms1       3.736761   0.222900  16.764   <2e-16 ***
days_to_firstoccurence    0.011979   0.005307   2.257   0.0240 *
---
```

C.2 Random Forest

```
Call:
 randomForest(formula = death ~ age + sex + has_covid_symptoms +days_to_firstoccurence,
 data = predicted_trainData, importance = TRUE)
Type of random forest: classification
Number of trees: 500
No. of variables tried at each split: 2

OOB estimate of  error rate: 0.68%
Confusion matrix:
      0  1  class.error
0 10576  6 0.0005670006
1    67 73 0.4785714286
----
```

C.3 SVM (Radial Kernel)

```
Call:
svm(formula = death ~ age + sex + hasCovidSymptoms + daysToFirstoccurence, data = predictedTrainData,
    type = "C-classification", kernel = "radial", gamma = 0.25, nu = 0.5)
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
Number of Support Vectors:  258
 ( 106 152 )
Number of Classes:  2
Levels:
 0 1
---
```

C.4 One-Class SVM

```
Call:
svm(formula = death ~ age + sex + has_covid_symptoms + days_to_firstoccurence,
data = predicted_trainData[predicted_trainData$death == 1, ],
type = "one-classification", kernel = "radial", gamma = 0.25, nu = 0.5)
```

```
Parameters:
SVM-Type:  one-classification
SVM-Kernel:  radial
gamma:  0.25
nu:  0.5
Number of Support Vectors:  73
Number of Classes: 1
---
```

C.5 Data Characteristics

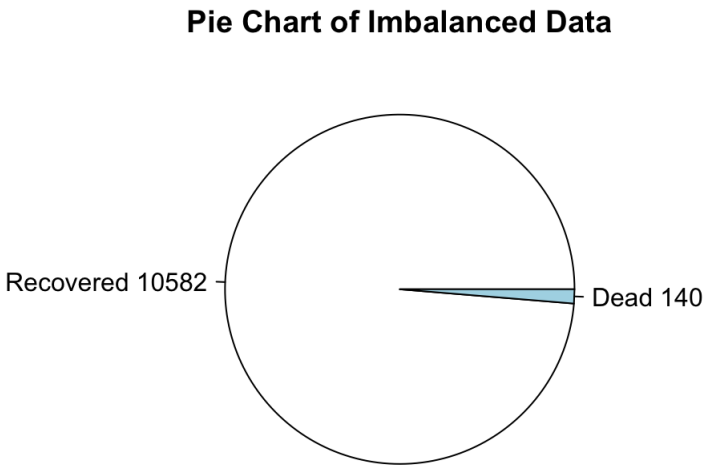


Figure 14: Data Imbalance

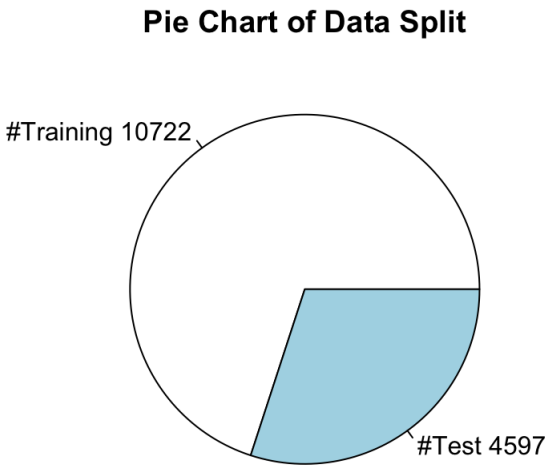


Figure 15: Data Split