CMPE 255 - Data Mining

COVID-19 Spread and Symptoms Analysis

Project Group 13

Submitted to

Prof. Jorjeta Jetcheva

on

15/05/2021

by

Sarthak Khanna (015252211)
Ishan Kanungo (015260726)
Abhijeet Padwal (015219958)

Github Link: https://github.com/sarthakkhanna05/CMPE-255-Term-Project-Team-13

TABLE OF CONTENTS

Section 1: INTRODUCTION

1.1. Motivation:

Corona**Vi**rus **D**isease 20**19** has infected over 160M individuals worldwide and killed about 3.3M. COVID-19 is caused by a coronavirus called SARS-CoV-2. Older adults and people who have severe underlying medical conditions like heart or lung disease or diabetes seem to be at higher risk for developing more serious complications from COVID-19 illness. COVID-19 affects different people in different ways. Infected people have had a wide range of symptoms reported – from mild symptoms to severe illness. Our motivation here was to flatten the curve by using Data Mining techniques.

1.2. Objective

With the data collected from different countries around the globe, we intend to analyze the general trend infection spread follows. Also to understand the significance of specific symptoms and derive the relation between symptom development and probability of those symptoms cause is actually COVID-19. Based on the numbers reported in the past few days, we also intend to predict the most probable conditions that will follow. This is to understand and anticipate the severity of the situation that the countries/regions will be facing in the coming days and be prepared to respond accordingly and efficiently, and help flatten the curve and provide the necessary preventive measures to be taken.

1.3. Approach

Firstly we focused on the nature of symptoms and tried to analyze the symptoms and covid relation data, we focused on general symptom features, special attributes like travel/proximity that are relevant to covid infections. We drew symptom significance and figured out the most important and relevant symptoms. Then we analyzed the world infections counts like daily new cases, daily recovered cases and daily new deaths, then analyzed the patterns that surface based on country/region location, population. Then based on the exponentially increasing infection rate in India, we started analysing the infection data for india. Then we used various prediction models to determine the statewise covid cases. We achieved good accuracy with predictions.

1.4 Literature/Market Review

1. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7310678/
2. https://www.nature.com/articles/s41598-020-77292-8
3. https://www.frontiersin.org/articles/10.3389/fdata.2020.00004/full

Section 2: SYSTEM DESIGN & IMPLEMENTATION

2.1 Algorithms Selected

Symptom analysis and Covid prediction

1. Support Vector Machine
2. Decision Tree Classifier
3. KNN Classifier
4. Random Forest Classifier
5. Logistic Regression
6. Guassian Naive Bayes
7. Neural Networks

World Data Country Wise Analysis and Trend Prediction

1. Support Vector Machine
2. Polynomial Regression
3. Bayesian Ridge Regression

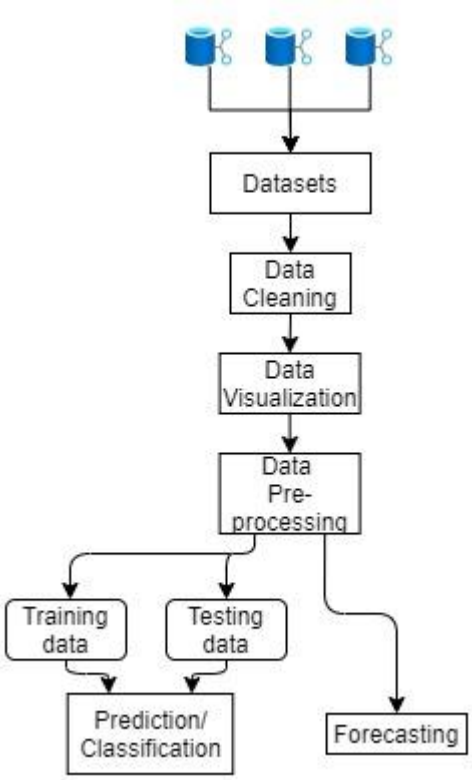
India Analysis and Trend Prediction

1. Auto-Arima
2. FBProphet
3. Holt-Winters

2.2 Technologies, Tools Used

1. **Editor**: Jupyter Notebooks, Google Colab
2. **Model libraries**: Scikit-learn, Tensorflow, Statsmodel, Fbprophet, Keras
3. **Plotting libraries**: Matplotlib, Plotly, Seaborn, Altair
4. **Evaluation**: Mean squared error, Mean absolute error, accuracy score
5. **Other**: pandas, numpy, time, datetime, dataprep.eda

2.3 System Design and Architecture



Chapter 3: EXPERIMENTS / POC EVALUATION

3.1 Datasets

We have considered 3 datasets for our project:

<u>Kaggle WHO possible symptoms and covid presence Dataset:</u>

1. The dataset contains symptoms developed by individuals and infection test results. Symptom Columns: Breathing Problem, Fever, Dry Cough, Sore throat, Running Nose, etc.
2. Proximity/exposure Columns: Abroad travel, Contact with COVID Patient, Attended Large Gathering, Visited Public Exposed Places, Family working in Public Exposed Places
3. Infection Column(Target Column) : COVID-19

<u>COVID-19 Data Repository by CSSE at Johns Hopkins University:</u>

1. About: Country wise time series dataset with confirmed, deaths and recovered cases
2. Duration: 22nd Jan, 2020 - 8th May, 2021
3. Features: Country/Region, Geographical coordinates, Date, Confirmed, Deaths, Recovered Date-wise count

COVID19INDIA.ORG Statewise data:

1. The dataset contains COVID-19 across different states/union territories in India.
2. Duration: 30th January, 2020 - 8th May, 2021
3. Features: Date, Time, State/Union Territory, confirmedIndianNational, ConfirmedForeignnational, Cured, Deaths, Confirmed

3.2 Data Preprocessing

3.2.1 Data Preprocessing for Symptom Analysis

1. The data was collected from Kaggle WHO symptom significance dataset. The dataset encompassed columns which extensively explained about the candidate's condition. For example: Heart diseases, Diabetes etc.
2. Data shape: The dataset contained 5434 rows and 21 columns.
3. Data cleaning: The dataset contained some columns which had little to no correlation with the target column. We removed those columns since we did not need them for modelling. We also deleted columns which had missing values. The final result of our data cleaning process was a clean dataset with highly correlated attributes.
4. Label Encoding: Since the columns were in Yes and No strings, we used LabelEncoder from Sklearn for encoding to 1 and 0.

3.2.2 Data Preprocessing for Global Cases Analysis

1. **Raw Data Gathering**: The data was loaded into a data frame using the pandas library, the dataset was referred from a remote github repository from JHU.
2. **Data shape**: The data was flattened out because of the date columns. Also the active, confirmed and death cases data was in different csv files
3. **Data cleaning**: There were some anomalies in the data i.e. the ship data was present in country wise data, also the conventions used by some of the countries were different from most of other countries i.e. Canada did not report recovered cases state wise.
4. **Aggregation**: after combining different csvs into dataframes, we aggregated the data country wise and then we used that information to calculate new columns like Country wise daily new cases.

3.2.3 Data Preprocessing for India State-Wise Analysis

1. The data was loaded into a data frame using the pandas library, the dataset was referred from a remote github repository from the Ministry Department of Health India. Data shape is 1490 rows and 9 columns.
2. **Top 6 countries:** Top 6 countries with most COVID cases were considered for evaluation
3. **Daily New cases from total cases:** Daily reported cases were calculated from total aggregated cases till date.
4. **Moving Average:** We have calculated the moving average for top 6 states to get a uniform data for future prediction.
5. **Normalize:** We have also normalized the data for top 6 states to get a better idea of mean square error and mean absolute error.

3.3 Methodology

3.3.1 Symptoms Prediction

The attributes of the dataset were based on two main factors, first being the symptoms and the second being the external factors that influence presence of Covid-19. Through visualization we analysed the behaviour of our data and proceeded to implement various models on it. From the visualizations we inferred that most of the highly correlating factors to the target column, had the most number of positive tested people. For prediction we used regression models and neural networks.

1. Train and test data - The dataset was split into training and testing data, where testing data was 20% of the total data and the models were trained on 80% of the data.
2. Classification modelling- Since the data was encoded in a binary format, the accuracies of all the models were highly comparable. The models used were classification models, such as SVM classifier, Logistic Regression, Decision Tree classifier, K- neighbors classifier, Random Forest classifier, to

name a few. The worst performing model turned out to be the Gaussian Naive Bayes model and the best performing was SVM classifier, until we implemented the keras neural networks.

3. Neural Networks- The keras sequential model contained 3 dense layers with activation functions including relu and sigmoid. The input shape to the sequential model was (10, None) since the data set after cleaning contained 10 attributes. The problem of overfitting was prevalent which is why two dropout layers were introduced between the dense layers with each of the layers having a dropout of 0.45. To further combat the problem of overfitting the learning rate was reduced to 0.0001. Reducing the batch size for training to 2 provided better results. The neural networks were validated at each epoch using a validation . x_test and y_test were used as validation. The model was run for 200 epochs.

3.3.2 Global Cases Prediction

After the data cleanup, preparation and data transformation the data was ready to be staged for visualization. The country wise grouping of the data and application of aggregation to a certain level, viable visualizations were made possible. The visualizations allowed us to filter out the insignificantly affected countries and help bring prominently impacted countries into the light. Looking at all the visual trends in the past data, applying regression models was the only logical choice to predict/forecast for coming days. Considering the data size which was not sufficient for trend analysis, we anticipated the Bayesian models to work the best which eventually turned out to be correct after modelling. For the purpose of trend analysis we applied three different machine learning models including Support Vector Machine, Polynomial Regression and Bayesian Ridge Regression. For the training and testing purpose we used 97/3 percent split considering size and purpose of trend analysis. We experimented with various parameters, kernels to get the better predictions and verification of results spanned over days. Finally we made predictions for 25 coming days, where Bayesian Ridge Regression results looked promising as they were appropriately weighing the time based significance of the cases and were good at assigning correct weight values for recent cases. Randomized search cv with 4 folds, helped finding best estimator.

3.3.3 INDIA Daily new Cases Prediction

1. **Auto-ARIMA**: Auto Regressive Integrated Moving Average is a model that forecasts future values based on some past values provided to the algorithm. This is a univariate algorithm which is widely used in the industry to forecast future values. ARIMA has three important parameters- p,d and q. After trying different values of p,d,q in auto-arima, the best values that the model got was (5,1,0) i.e. 5th order auto regressive model, 1 difference between response variable data, we took 0 moving average as we have already done that. Arima model works for seasonal as well as non seasonal data. We can specify the seasonality and order in a parameter. The model was used for prediction on 0.1 % test data as forecasting models need a lot of data to train.
2. **FBprophet Model:** Fbprophet is developed by facebook and works on data that has a pattern or seasonality. FBprophet doesn't predict well on data that doesn't have any trend. That's why in our case, we trained it on only 90 days of data with 10% testing data and 90% training data.
3. **Holt-Winters:** Holt winters is a time series exponential model. It can take three parameters into account, average, trend and seasonality. Averaging is necessary to smoothen the data. Since there isn't any trend in our data, we have used a smoothing factor of 0.8 and smoothing trend of 0.2.

3.4 Analysis of results: (Values highlighted in yellow are the best)

1. **COVID-19 prediction:**

Metric/Model	Logistic Regression	SVM	Gaussian NB	Decision Tree Classifier	KNN Classifier	Random Forest Classifier	Gradient Boost Classifier	Neural Networks
Test Accuracy	96.04	96.96	77.27	96.964	96.41	96.87	96.78	97.14

2. **Global cases prediction**

Metrics\Model	SVM	Polynomial Regression	Bayesian Ridge Reg.
Mean Absolute Error	26.25 e+06	9.6 e+06	6.1 e+06
Mean Squared Error	6.9 e+14	9.6 e+13	3.8 e+13

3. **India state-wise prediction:**

Metrics\Model	ARIMA	Fbprophet	Holt-Winters
Mean Absolute Error	0.0180	0.1229	0.0640
Mean Squared Error	0.0007	0.0354	0.0103
Time (Mins)	0.28	1.5900	0.0042

3.5 Graphical Comparison:

1. **COVID-19 prediction:**

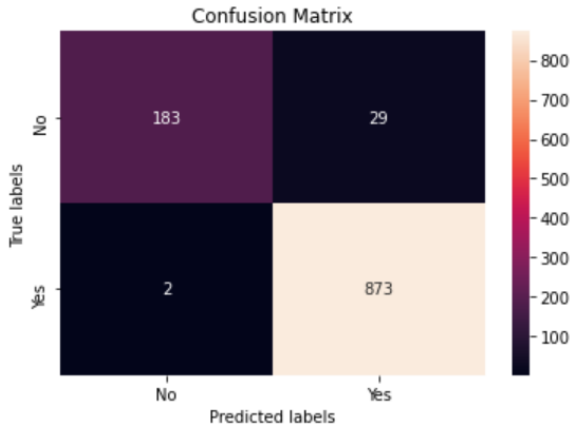
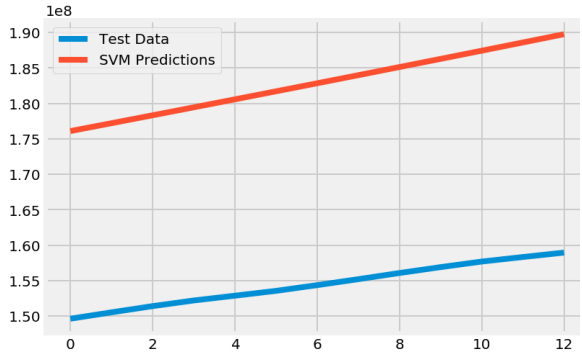
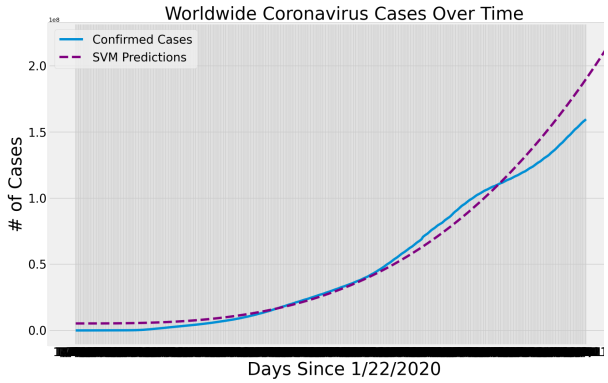


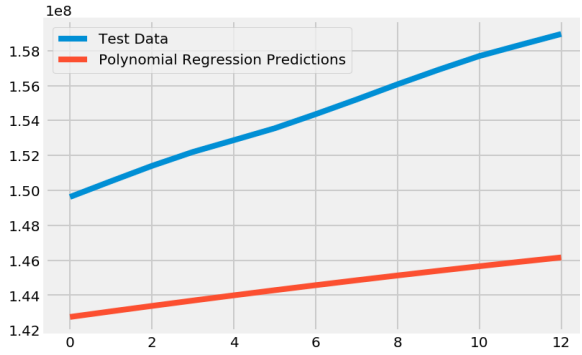
Figure 3.5.1: Confusion Matrix

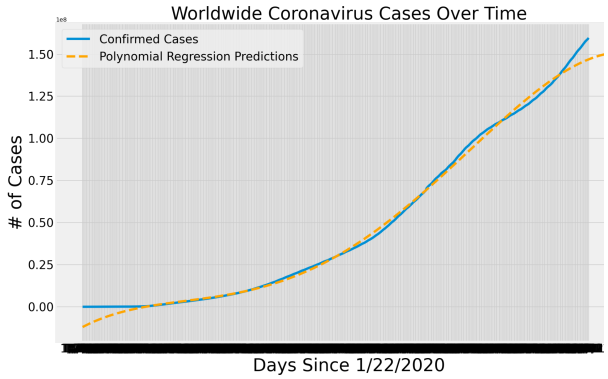
- The covid symptom analysis neural network model wrongly predicted 29 Covid negative individuals, and wrongly predicted 2 Covid positive individuals.
- The model wrongly correctly predicted 873 Covid positive individuals and correctly predicted 183 Covid negative individuals.

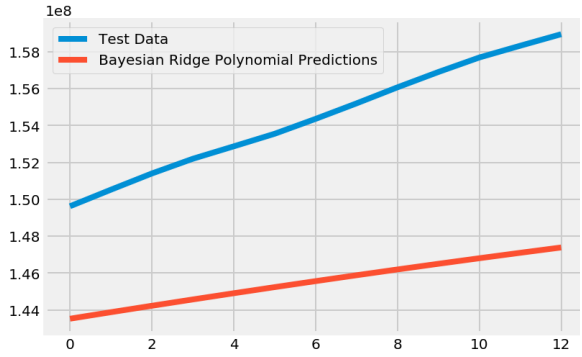
2. Global cases prediction

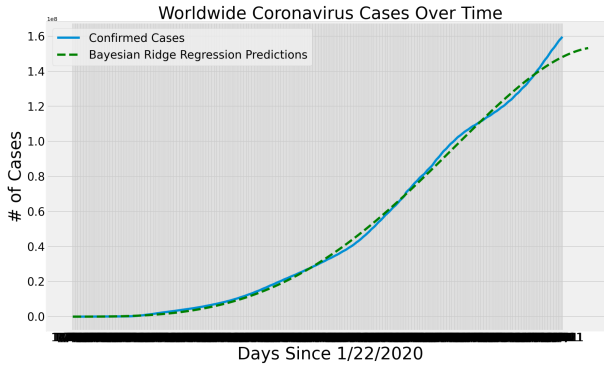












- In the graphs on the left side, the blue line indicates the test data values and orange lines indicate the model test results.
- In the graphs on the right, the solid line indicates the worldwide corona cases over time and the dotted line indicates the forecasted trend that will follow for next 25 days
- After modelling with 3 different machine learning models, Bayesian Ridge Regression produced the best results which were anticipated based on the training data size.

3. **India state-wise prediction and forecasting:**

Blue: Actual values
Green: Predicted values

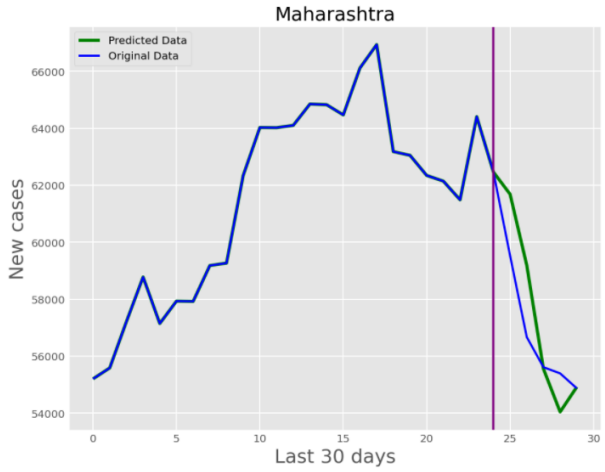


Figure 3.4.3.1: ARIMA

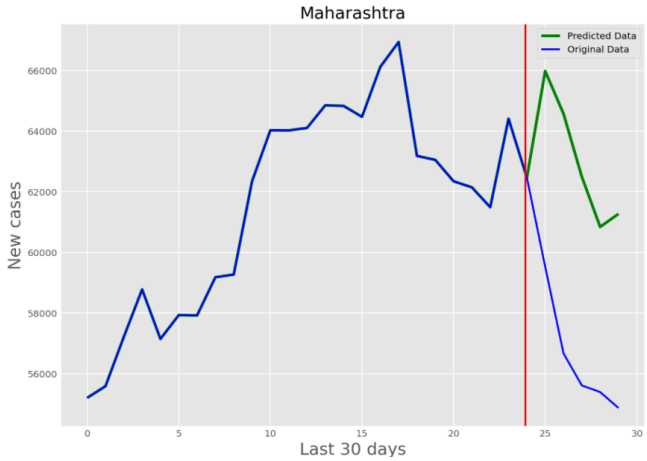


Figure 3.4.3.2: Fbprophet

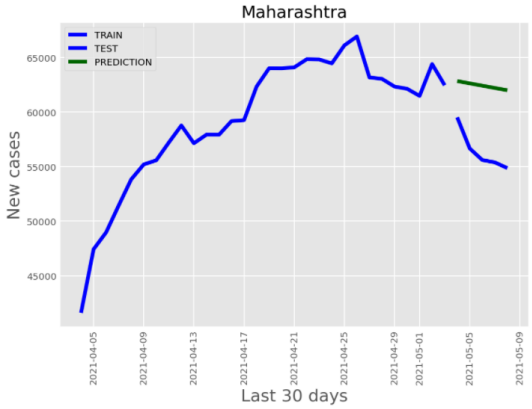


Figure 3.4.3.3: Holt-Weathers

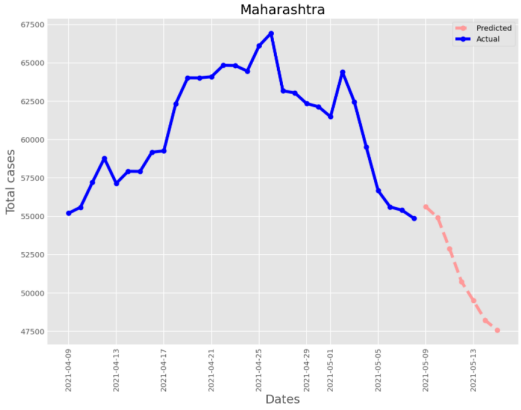


Figure 3.4.3.3: Forecasting using ARIMA

- Blue line shows the actual data and the green line represents predicted values.
- Figure 3.4.3.3 shows forecasting of new COVID cases for the next 7 days in the state of Maharashtra
- The x axis shows the data for the last 30 days and the Y axis represents daily new cases in the last 30 days.
- Note that the data is smoothed using moving averages.
- The data that we have does not have any trend/ seasonality and that's why fbprophet and holt weathers are not able to perform well.
- You can see that the best results are given by ARIMA as it can work on data without seasonality.

Chapter 4: Discussions and Conclusions

4.1. Decisions Made

1. The team spent a good amount of time researching the appropriate datasets for COVID-19 analysis that were good fit for our problem statement.
2. After confirming the viability of the datasets, the team decided to structure the flow and architecture of the modelling analysis.
3. The team constantly brainstormed and reconsidered the data engineering steps to be able to generate more accurate predictions and forecast models.

4. Finally, the team also decided to take the project one step further by moving from individual level to a large population level. We did strenuous analysis on India, as the sudden rise in the covid cases in India was a more important topic, which was added to the scope for forecasting and predictions analysis.

4.2. Difficulties Faced

1. The dataset repositories were maintained by different authorities, and while working on the project there were changes introduced in the schema which the team had to rework on and make compatible with existing schema.
2. For the forecasting part, since new cases per day is not a uniform thing, we had to apply different moving average techniques to smoothen the data.
3. Some algorithms were giving a problem because the date was not in datetime format, we had to change that as well.
4. The team faced the difficulties narrowing down the machine learning models to focus on, in order to achieve the best forecasting and prediction results. Finding out the most relevant features was also a very difficult task.

4.3. Things that worked well

1. The data gathering part was important for the project, but finding the helpful and viable sources of data was given great deal of attention at the very beginning which resulted in minimal efforts to be put in the later stages of project development.
2. Working on model tuning and parameter search was done very diligently by each member of the team, so, very good modelling results were achieved.
3. Forecasting worked well as it gave highly accurate results, even when compared to the real-life covid cases of the next day in a state. This proved our project's real life usefulness and formulated an appropriate use-case..

4.4 Things that didn't work well

1. Some datasets that were chosen early in the stage of data collection were turned out to be irrelevant, so significant time was lost in the initial stage.
2. The datasets were highly convoluted. During the project when we were exploring datasets we usually found them to have multiple irregularities. For example, when we found India's state wise Covid cases dataset, the numbers also included people stranded on ships in India.
3. The finetuning for each and every model proved to be strenuous. The models had to be tuned based on the parameters individually since most of them have different requirements and some even different inputs.

4.5 Conclusion

1. While developing this project, we learned a number of data preprocessing, modelling and visualization techniques from ML libraries. Specifically the impact of various data modelling parameters on the predictions generated. As we worked on the project we came across very detailed analysis of models and experts suggestions for fine tuning models.
2. Covid 19 Symptom Analysis- The covid analysis project depicts that the presence of covid is highly dependent on the short term symptoms such as Sore Throat, Dry Cough, Fever, and not much dependent on the long term effects such as, Asthma, Diabetes, Gastrointestinal. The external factors play a large role in identifying a covid positive person. Therefore it can be concluded that monitoring of these factors are equally as important, if not more. The accuracy of our model is high which also proves the conjectures made above.
3. We tried to analyse covid situations in the world and found out that India is currently experiencing a high number of cases. Data-pre processing was required to get the states with most cases and to generate daily new cases from total confirmed cases. While using ARIMA, Fbprophet and Holt-winters modelling techniques, ARIMA turned out to be the most accurate. We then performed forecasting for the next 7 days using ARIMA.

Chapter 5: Project Plan/Task Distribution

5.1 Who was assigned to what task

Task	Responsibility
Dataset Selection	All
Data Exploration and Cleaning	All
Data Preprocessing	All
Research on Algorithms	All
Symptoms Significance Analysis	Ishan
Global Cases trend Analysis	Abhijeet
India state wise case analysis	Sarthak
Models Tuning and parameter experiments	All
Documentation and Report	All
PPT	All

5.2 Who ended up doing what task

The team divided the workflow in the start as mentioned above. Each of the individual parts were fulfilled as assigned to each other and each member showcased utmost team working skills.

References:

1. https://www.kaggle.com/datasnaek/youtube-new
2. https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
3. https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html#exercise-2-sentiment-analysis-on-movie-reviews
4. https://regexr.com/
5. https://www.kaggle.com/yanpapadakis/trending-youtube-video-metadata-analysis