ASSIGNMENT REPORT:

COVID-19 ANALYSIS AND CASES PREDICTION

NIK FAIZ AFIQ BIN NIK AB RAHMAN

WQD7006, MACHINE LEARNING COURSE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2020/2021 SEMESTER 1

TABLE OF CONTENTS

Table of Contents 2

INTRODUCTION 3

CHAPTER 1: ANALYSIS AND DESIGN 4

1.1 Project Design 4

1.2 Data Preprocessing /Preparation 5

1.3 Exploratory Data Analysis (for secondary objective) 6

1.4 Modeling and Evaluation 7

CHAPTER 2: EXPERIMENT RESULT 8

2.1 Original Input data 8

2.2 Output for objective Secondary objective from EDA 9

2.3 Output for Main Objective 12

CHAPTER 3: DISCUSSION (IMPROVEMENT AND LIMITATIONS) 16

3.1 Discussion on the Exploratory Data Analysis section 16

3.2 Discussion the Modeling section 17

CONCLUSION 18

References 19

INTRODUCTION

The outbreak of coronavirus disease (COVID-19) has been declared a Public Health Emergency of International Concern (PHEIC) and the virus has now spread to many countries and territories. Based on recent data, it has affected more than 90 million people around the world and causing close to 2 million death. (Bender, 2020, pp. 1–3)

COVID-19 is a disease caused by a new strain of coronavirus. The COVID-19 virus is a new virus linked to the same family of viruses as Severe Acute Respiratory Syndrome (SARS) and some types of common cold. Symptoms can include fever, cough and shortness of breath. In more severe cases, infection can cause pneumonia or breathing difficulties. More rarely, the disease can be fatal. The virus is transmitted through direct contact with respiratory droplets of an infected person (generated through coughing and sneezing) (Bender, 2020, pp. 1–3).

The main objective of this project is to predict the number of new COVID-19 Cases based on previous history of cases. This will enable us to see the current trend of this cases and take an appropriate action towards it. The secondary objective of this project is to classify the coronavirus trend for each of the countries based on the countries' new cases in last 7 days. From this, we can see which of the countries are handling covid-19 effectively and which country is doing really bad. After all, the first step to any improvement is realizing the current situation of COVID-19 in every country.

For this, we are using the data provided by John Hopkins University that is publicly available to the people. People can access it through this given link: https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

CHAPTER 1: ANALYSIS AND DESIGN

1.1 Project Design

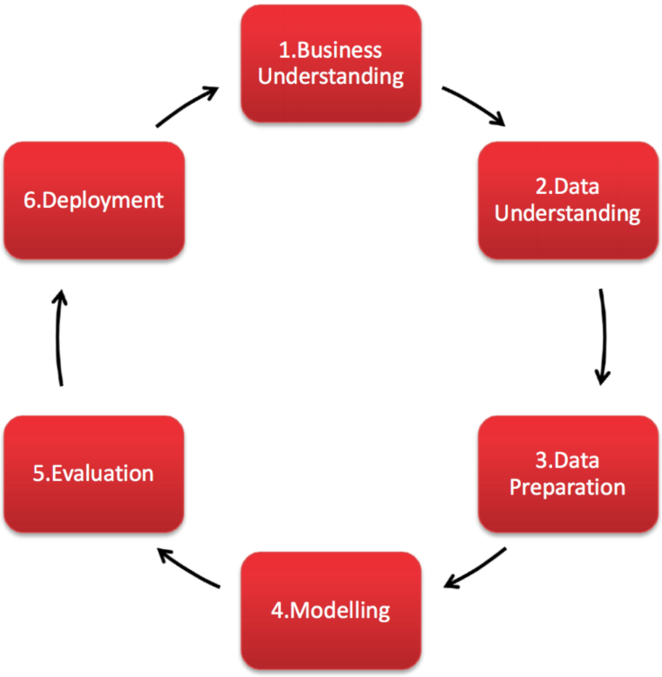


Figure 1-1: CRISP-DM Methodology

This assignment report is designed to follow the CRISP-DM methodology that provides a structured approach to planning a data mining project (Smart Vision Europe, 2020). In business understanding, we try to understand the issues and provide an objective based on that issue. Then from there, we search for a suitable dataset to tackle this issue and achieve the objective (data understanding). The data is then prepared and undergoing pre-processing process (data preparation). Then we model the data using machine learning algorithm (modeling). We will then evaluate the models accordingly (evaluation). However, this model is yet to be deployed to be used in real life yet. All the codes that is used in this assignment report is available in this GitHub link: https://github.com/n-Faiz/WQD7006_ML_Covid19.

1.2 Data Preprocessing /Preparation

The data input consists of 3 csv files, each contains COVID-19 number of confirmed cases, number of death and number of people recovered respectively. Each of the data have list of countries as the rows with the dates as the column. Since we need the data to be in a nice tabular format with column as the attributes, we will need to do some pre-processing on the data. First, we will need to unpivot the date columns so all the numbers for that attribute is under the same column. Unpivoting here will be done on all there of the dataset. Here, instead of the columns originally have different dates as column, we will now convert it so that the dates now occupy only 1 of the columns and the other columns fills the value corresponding to the dates and the countries.

After that, all 3 tables now need to be merged together for easier analysis. The data is then checked for any null values and an appropriate action are taken for these null values. Since the dataset only consist of “total new confirmed cases”, “total death” and “total people recovered”, a new column: “the total active cases” are then calculated by using the formula: total active = total confirmed – total death – total recovered. It is also useful to include daily new confirmed cases, daily new death and daily new recovered, so we include that in by taking the current total cases minus the cases on the previous day for that country.

In summary, for preprocessing, these steps are taken for data pre-processing:

1. Data Merging
2. Needs to be unpivoted (merge the dates to 1 column only)
3. Check for missing value
4. Calculate for new columns: Total Active
5. Calculate for new columns: New Confirmed, New Death, and New Recovered

1.3 Exploratory Data Analysis (for secondary objective)

Our secondary objective of this project is to classify the coronavirus trend for each of the countries based on the countries' new cases in last 7 days. For that, we need to use the numbers of new cases only and include the data only for the last seven days. Then we standardize it by dividing it with the maximum number of cases in the last 7 days for each country. This is done so that the range of y axis is the same that is from 0 to 1 for every country and the gradient produced on the next step will be based on standardized unit. Next the gradient is calculated based on the 7 data point for each country. Then, categorized the trend of the cases in each country based on the gradient. Here we defined that the trend is "increasing" if the gradient value is more than 0.2, "around the same" if the gradient value is between -0.2 and 0.2, "decreasing" if the value is below -0.2 and "No Cases" if there are no new cases observed for the last 7 days. Then the number of countries with each status is visualized using bar chart. After that, we will plot the performance of the top 10 of "decreasing" trend to see the graph of the top 10 countries that performs best to combat covid-19 within the last 7 days and the same for "increasing" trend.

Summary of Exploratory Data Analysis steps taken:

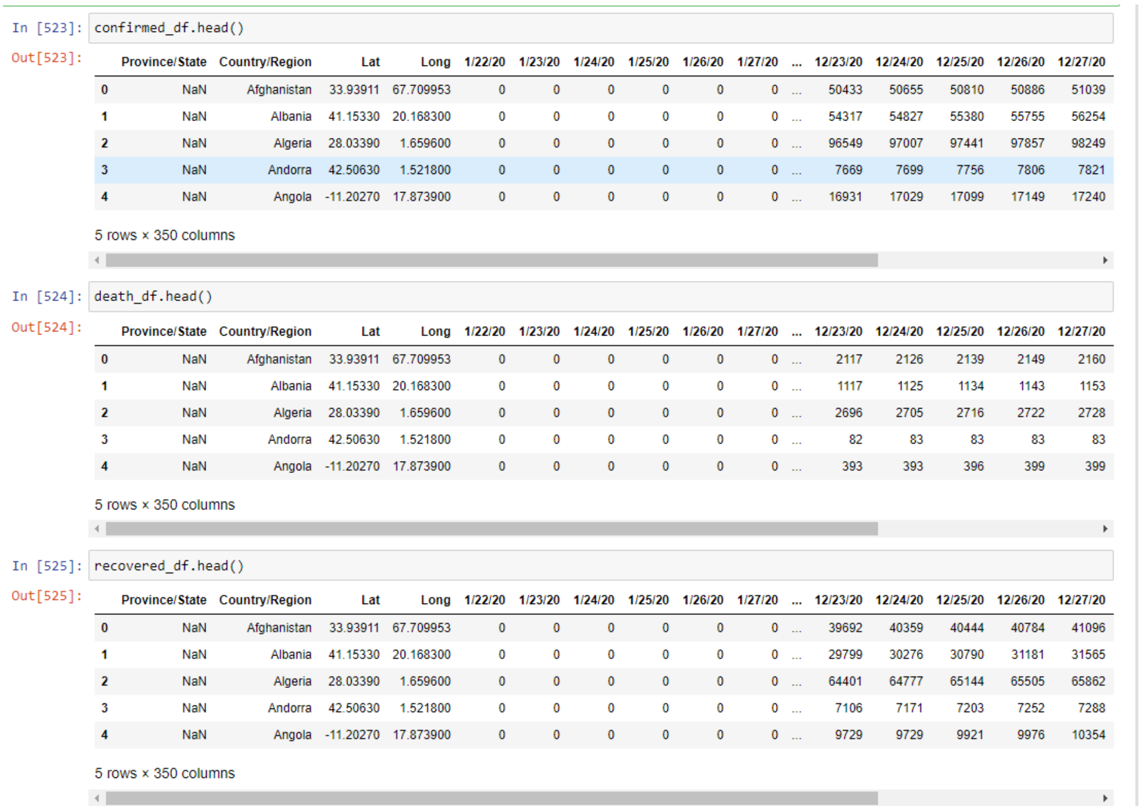
1. Filter the data to only include last 7 days
2. Standardize it based on the country before calculating the gradient (the new confirmed cases are standardized so it falls within 0 and 1 in each country)
3. Calculate the gradient for each of the country based on their last 7 new cases
4. Categorize the status based on the gradient
5. Visualize total country that have increasing, decreasing or same gradient
6. Plot top 10 performers and bottom 10 performers

1.4 Modeling and Evaluation

The models that we used here for prediction and forecasting are: Support Vector Machine (SVM), Simple Linear Regression, Polynomial Linear Regression and Decision Tree regression. But before the modeling is done, we need to split the data into train and test split. Here we are using 70% of the data as training while the remaining 30% of data as testing. We also provide 20 extra points that is used to forecasting 20 days into the future. Instead of using the date as the X value, integer values are used to make it easy for calculation. The model is then evaluated using 3 score evaluation: Mean Absolute Value, Root Mean Square Error, and R^2 score. The result is also projected into a graph so we can see the points clearly.

CHAPTER 2: EXPERIMENT RESULT

2.1 Original Input data



In [523]:
```
confirmed_df.head()
```

Out[523]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/23/20	12/24/20	12/25/20	12/26/20	12/27/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	50433	50655	50810	50886	51039
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	54317	54827	55380	55755	56254
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	96549	97007	97441	97857	98249
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	7669	7699	7756	7806	7821
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	16931	17029	17099	17149	17240

5 rows × 350 columns

In [524]:
```
death_df.head()
```

Out[524]:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/23/20	12/24/20	12/25/20	12/26/20	12/27/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	2117	2126	2139	2149	2160
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	1117	1125	1134	1143	1153
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	2696	2705	2716	2722	2728
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	82	83	83	83	83
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	393	393	396	399	399

5 rows × 350 columns

In [525]:
```
recovered_df.head()
```

Out[525]:

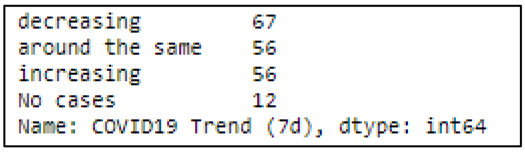
	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	12/23/20	12/24/20	12/25/20	12/26/20	12/27/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	39692	40359	40444	40784	41096
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	29799	30276	30790	31181	31565
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	64401	64777	65144	65505	65862
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	7106	7171	7203	7252	7288
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	9729	9729	9921	9976	10354

5 rows × 350 columns

Figure 2-1: Original Data

Our source files are originally the 3 comma separated value format file. The first one contains the confirmed cases of covid-19, the second file contains the number of deaths due to COVID-19 in each country and the last one contains the number of recovered people from COVID-19. Each of the files have the countries and regions as the rows and the dates as the column.

2.2 Output for objective Secondary objective from EDA



```
decreasing          67
around the same     56
increasing          56
No cases            12
Name: COVID19 Trend (7d), dtype: int64
```

Figure 2-2: COVID-19 trend in each country

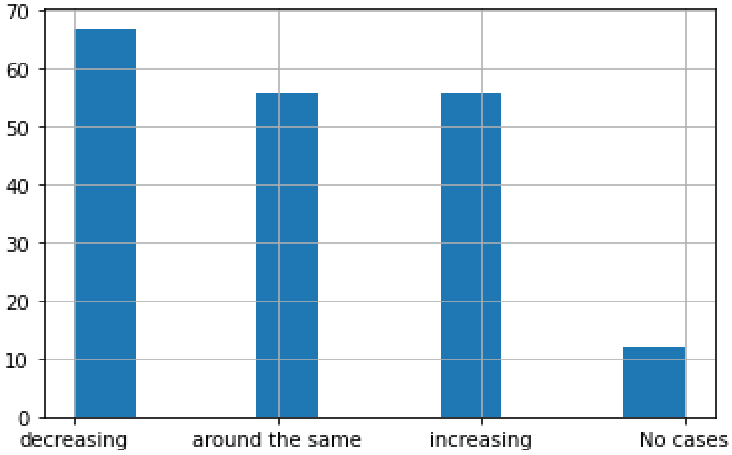


Figure 2-3: COVID-19 Trend in each country

As of 18th January 2021, we can see that most of the countries have decreasing trend of new COVID-19 cases. There are 67 countries that have decreasing trend, 56 countries have flat trend (around the same), 67 countries that have increasing COVID-19 new cases trend, and 12 countries that have no new cases at all.

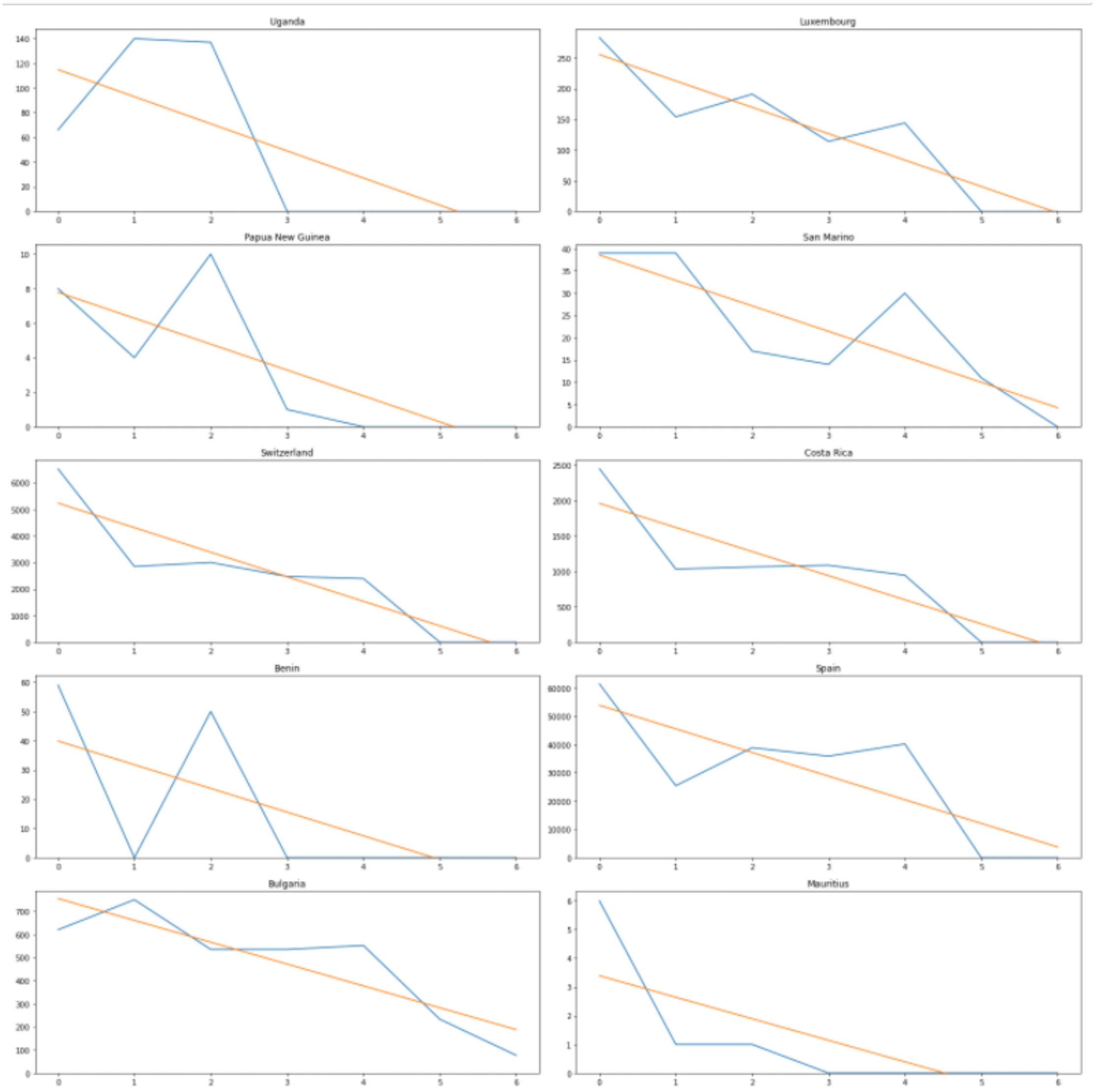


Figure 2-4: Top 10 countries with decreasing trend of New COVID-19 cases

List of top 10 countries with decreasing trend of new COVID-19 cases: Uganda, Luxembourg, Papua New Guinea, San Marino, Switzerland, Costa Rica, Benin, Spain, Bulgaria and Mauritius. Some of them even have COVID-19 cases dropped to 0 new cases for the last couple of days, a sign that these countries are combating COVID-19 effectively.

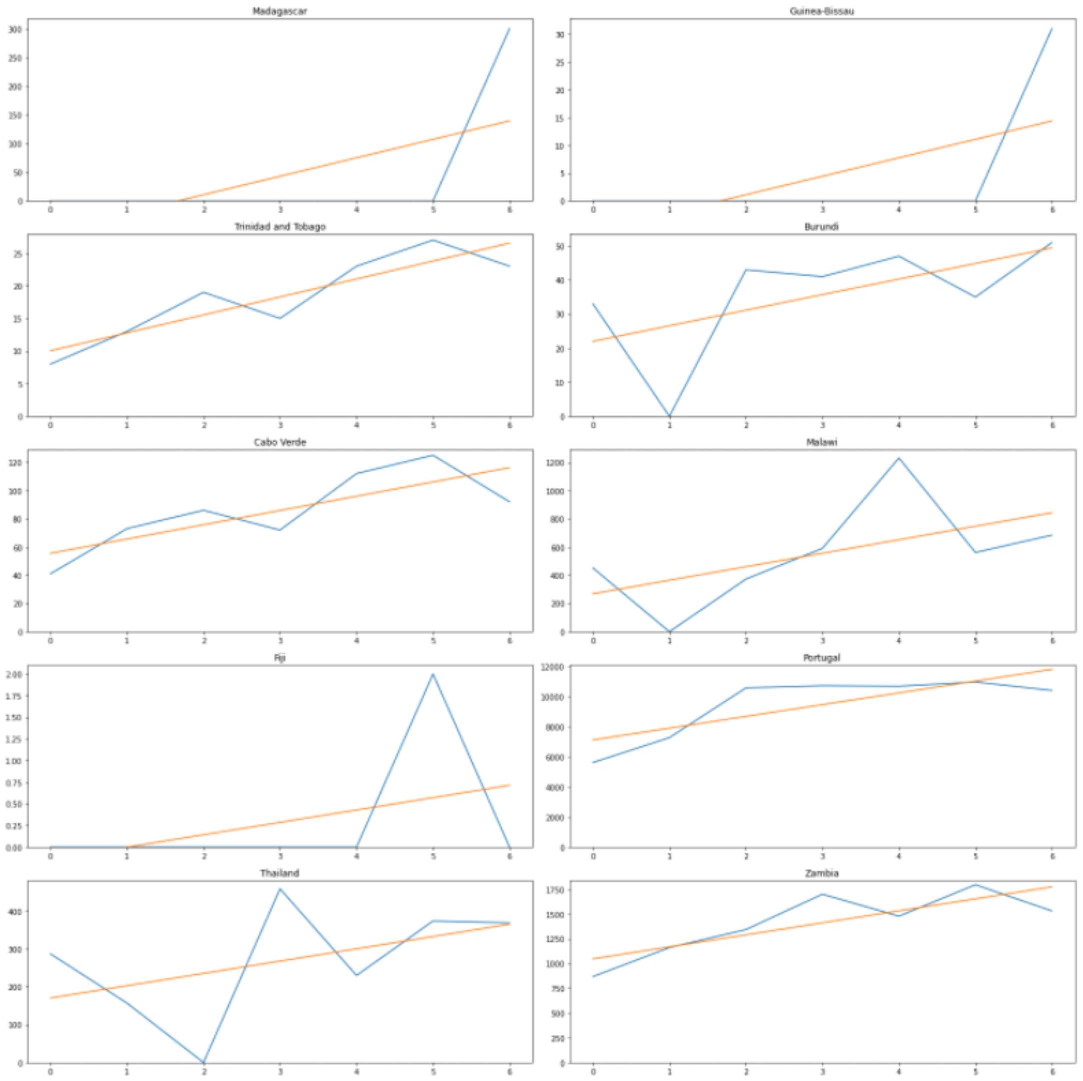


Figure 2-5: Bottom 10 Countries with increasing trend of COVID-19 New Cases

These countries that have increasing trend for COVID-19 new cases in the last 7 days. These countries might need to take extra precautions to prevent further increase in COVID-19 cases.

2.3 Output for Main Objective

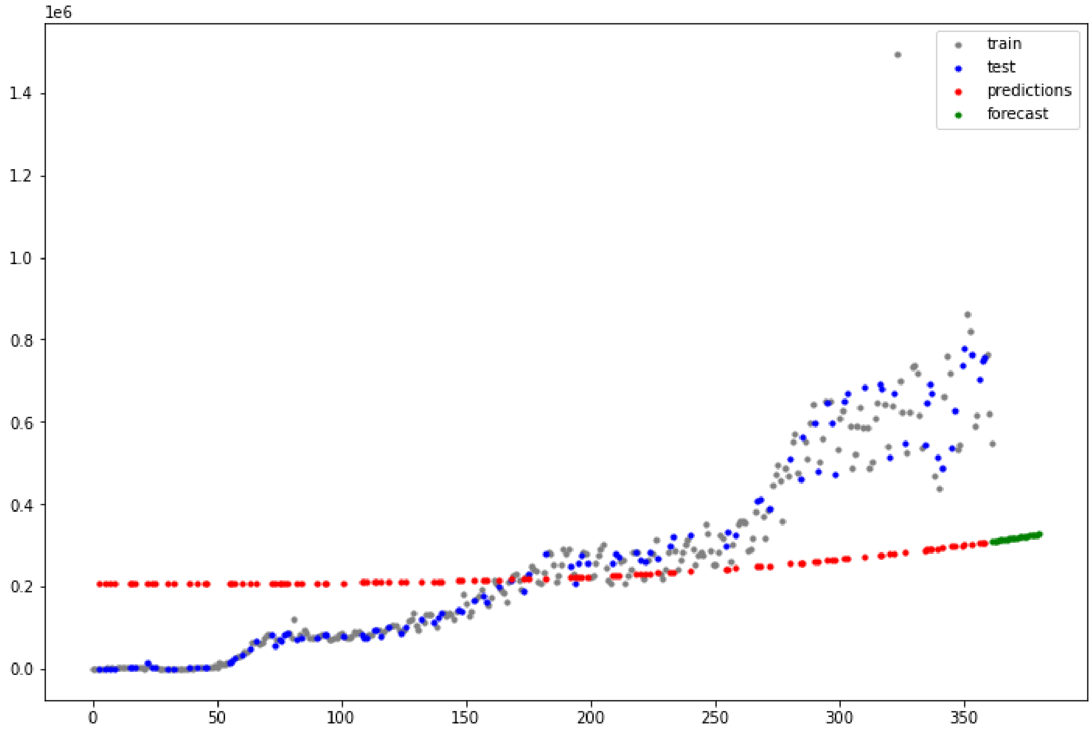


Figure 2-6: SVR modeling and forecasting

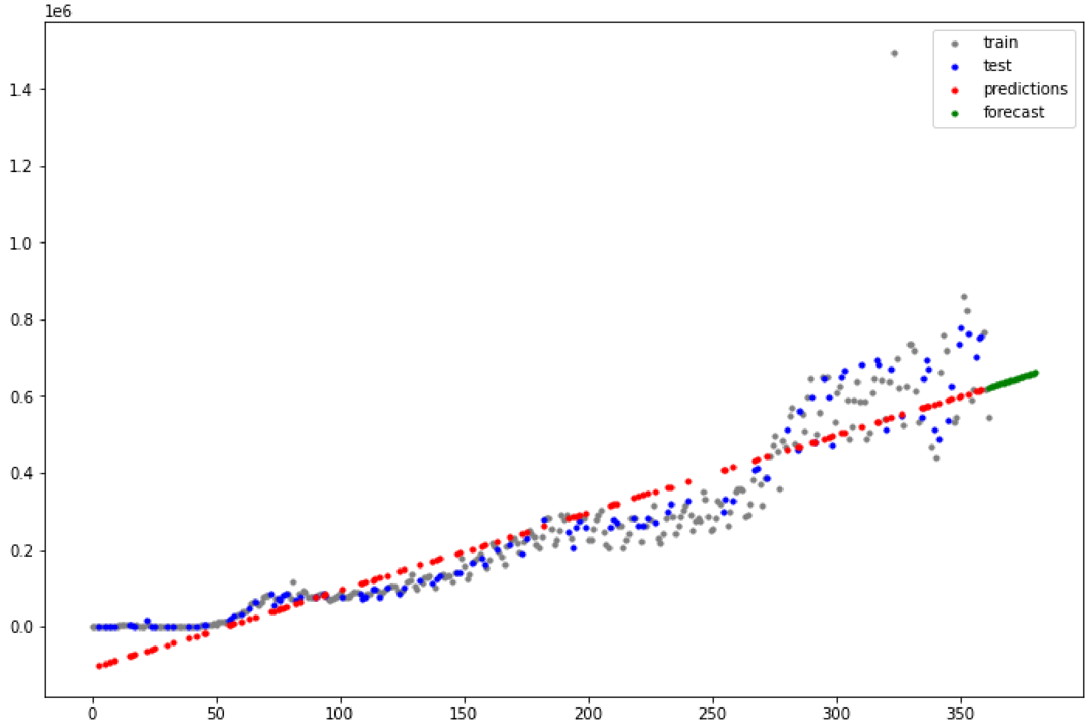


Figure 2-7: Linear regression modeling and forecasting

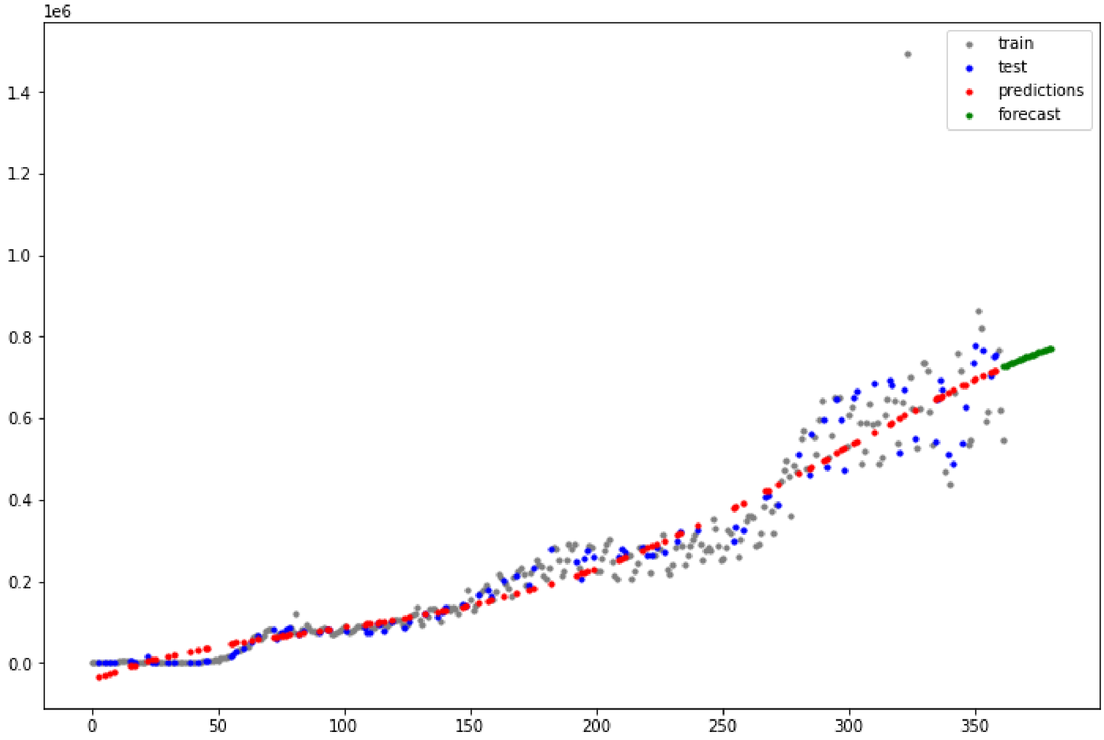


Figure 2-8: Polynomial Regression modeling and forecasting

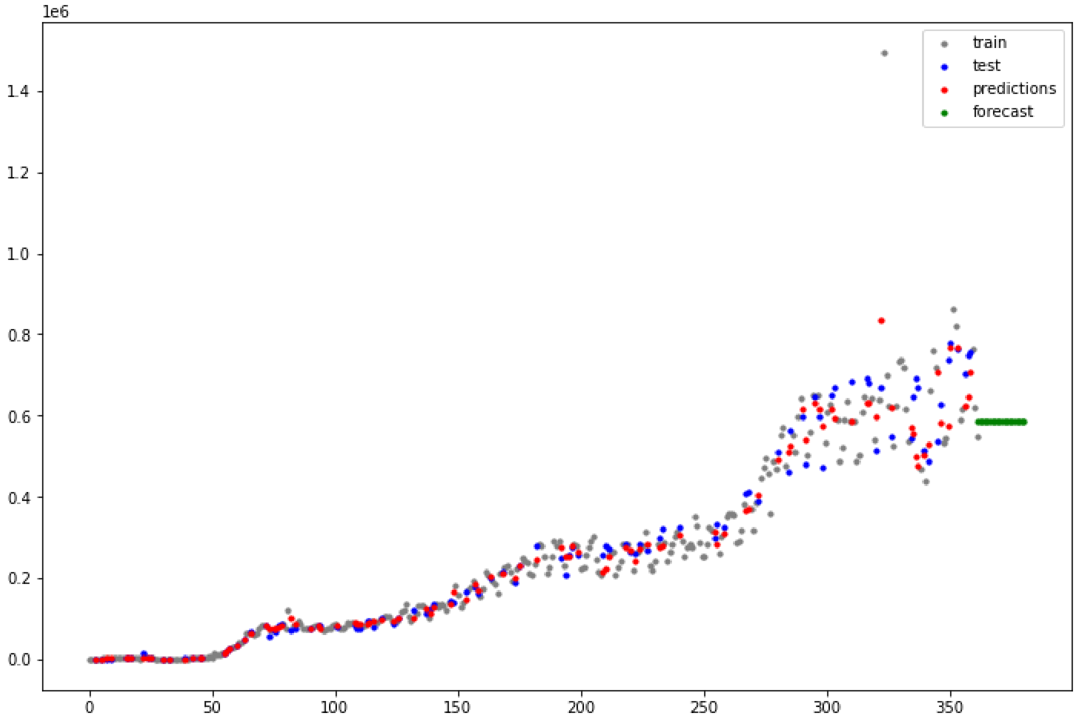


Figure 2-9: Decision Tree Regressor Modeling and forecasting

Figure 2-7 to 2-10 shows the graph of the models that include the train value, the test value, the prediction of the test value and the forecasting value, colored in grey, blue, red and green respectively.

Regression Tree and Polynomial Linear Regression seems to be doing good for the prediction the test value. SVR seems to be fail for this prediction and forecasting. Further research needs to be done to find out why this SVR is not reliable to predict this data. For the Simple Linear Regression, the result is okay but can be improved more and polynomial Linear Regression is a better version of Simple Linear Regression that does a good job at predicting the test value and also for the forecasting. For the decision tree model, it does a pretty great job at predicting the values in between, but for the forecasting future value, the value predicted seems to be a constant value, although we can see from the figure that the real trend of COVID-19 new cases seems to be an increasing trend.

Table 2-1: Evaluation Score across different models

Model	Mean Absolute Error (MAE)	Root Mean Square Error (RMSE)	R^2 Score
Support Vector Machine	177688.41	215668.40	-44.62
Simple Linear Regression	58131.40	72334.89	0.90
Polynomial Linear Regression	34931.38	51009.30	0.95
Random Forest	27827.08	49588.08	0.96

Based on the evaluation score, Random Forest has the best score for all 3 of the evaluation score : lowest Mean Absolute Error, lowest Root Mean Square Error and highest R2 score. Here we can conclude that Random Forest is the best model for predicting the test case number. However, if we also include the forecasting result from the figure before this, Polynomial Linear Regression is the best model for predictions and forecasting.

CHAPTER 3:

DISCUSSION (IMPROVEMENT AND LIMITATIONS)

There are many areas that this studies can be improves on and this section mainly focuses on this.

3.1 Discussion on the Exploratory Data Analysis section

For this part, the analysis mainly studies the new cases on the last 7 days. However, this can be expanded further to include more days, even months of data that is used for the gradient. 7 days of data might not be conclusive enough to summarize the trend for the countries.

Additionally, same calculations also can be done on the other features, namely the number of deaths, the number of recovered cases, and the number of active cases. More analysis can mean more meaningful insights that can be deduced from the data.

3.2 Discussion the Modeling section

From the result we can conclude that the polynomial linear regression model does the best job at predicting and forecasting the cases. However, here we only include four regression models. Some of them doing a very bad job at modeling the test data. Further studies need to be done to study the behavior of the models and the reason it does not performing well.

Additionally, for this assignment report, we only use the default hyperparameter for each of the model. However, we can actually use different hyperparameter for these models and further test needs to be done to find the best parameter that performs best for this model.

Next, since the data is a time-series data, the splitting of the training and testing data might not be suitable to be used. In real case scenario, we know all the past historic data and all the future data are unknown. There are more suitable ways to test time-series related data. For example: Start with fitting with 10 data point, predict next 10 data point. Then fit 20 data point and predict next 10 data point and so on.

Lastly, this modeling can also be used for the other feature, namely: total cases, total death, total recovered, total active, new death, new recovered and new active. This model can also be used to predict COVID-19 cases for each individual country using the same method.

CONCLUSION

From our model, we can conclude that Decision Tree performs the best among all of the models tested if we are only predicting the test data. However, polynomial linear regression model preforms better if we include the forecasting.

On the other hand, from the total new cases that are observed worldwide, we can see that there is an increasing trend and this is quite worrying.

From the secondary objective, we know that there are still 56 countries that have increasing trend of COVID-19 new cases. However, we can see many of the countries also starts to have decreasing trend and this shows that countries around the world are combating COVID-19 effectively.

We hoped that the vaccine that are already introduced and are used world-wide, can negate the increasing trend and thus, reduce the cases world-wide. However, even if the vaccine works perfectly, billions of people, 60-70% of the global population must be immune to stop the virus spreading easily - a concept known as herd immunity (Gallagher, 2021)

REFERENCES

Bender, L. (2020). *Key Messages and Actions for COVID-19 Prevention and Control in Schools*. UNICEF New York. https://www.who.int/docs/default-source/coronaviruse/key-messages-and-actions-for-covid-19-prevention-and-control-in-schools-march-2020.pdf?sfvrsn=baf81d52_4

Gallagher, B. J. (2021, January 11). *Covid vaccine update: When will others be ready?* BBC News. https://www.bbc.com/news/health-51665497

John Hopkins University. (n.d.). *JHU CSSE COVID-19 Dataset*. GitHub. Retrieved January 18, 2021, from https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data

Smart Vision Europe. (2020, June 17). *Crisp DM methodology*. https://www.sv-europe.com/crisp-dm-methodology/

Taheri, Z. (2020, July). *Spread Visualization and Prediction of the Novel Coronavirus Disease COVID-19 Using Machine Learning*. https://www.researchgate.net/profile/Zahra-Taheri/publication/343685138_Project_report/links/5f3beaf392851cd302019189/Project-report.pdf