Covid - 19 Data Analysis, Visualization and Machine Learning Implementation

Nishant Ranjan Verma, Letterkenny Institute of Technology, L00157133

***Abstract*— Big Data Analytics is mix bag of all the technologies come together when working with problems and analysis associated with Big Data. In here I am trying to attempt to do analysis, visualization and machine learning prediction on the Covid 19 dataset. I would use pyspark as the language to operate and Databricks as the platform to operate on achieving the requisite task.**

***Index Terms*— Attacks,Machine Learning,Adversarial Attacks,Artificial Intelligence**

I. Introduction

IN December 2019, the novel coronavirus (COVID-19) pandemic broke out in Wuhan, China, and has since spread across the world. COVID-19 pandemic disease was caused by a virus known as coronavirus 2, also known as extreme acute respiratory syndrome coronavirus 2, is a virus that causes severe acute respiratory syndrome. SARS-CoV-2 is a virus that causes SARS. Coronaviruses (CoV) are a broad group of viruses. Cold-related illnesses, such as the flu, are caused by viruses. Severe Middle East Respiratory Syndrome (MERS-CoV) and Middle East Respiratory Syndrome (MERS-CoV) Acute Respiratory Syndrome (ARS) is a condition that affects the lungs (SARS-CoV). COVID-19 is a new genus of Coronavirus that was discovered in 2019 and has never been found in humans before. In this project I aim at predicting the number of confirmed cases and fatalities in different regions of the world. I will apply the concepts of Big Data and make attempt to solve the following scenario using PySpark. I will make use of Linear Regression, Decision Tree, Random Forest in way to find the results.

II. Data

The Data for has been provided by John Hopkins CSSE. It is readily available on Kaggle[1] in train and test files to be used for any further research and analysis purposes. The training dataset file consists of corresponding Ids, Province State, Country Region, Date, Confirmed Cases and Fatalities as the required columns. The test dataset file consists of ID, Province State, Country Region and Date as the aforementioned columns. The format of the dataset will be changed accordingly in order to make better use of it. Also the data columns which consists of the NULL values will be handled in order better up the analysis of the whole set.

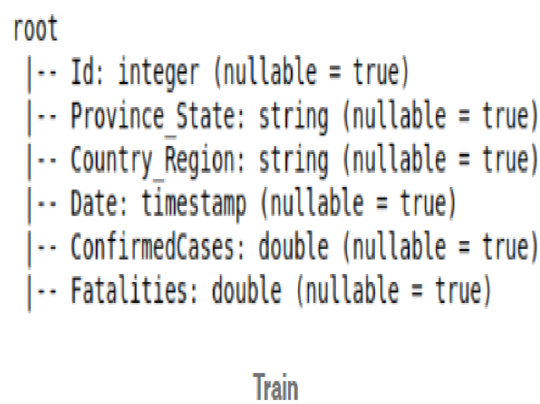


Fig 1: Train Data Schema

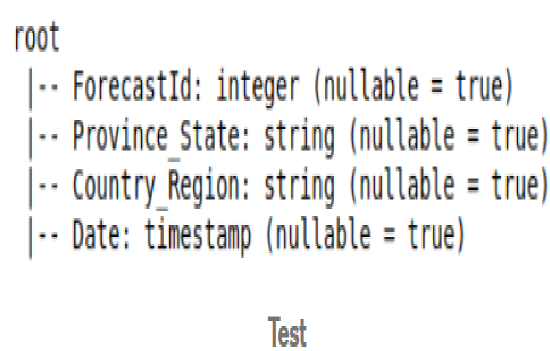


Fig 2: Test Data Schema

Id	Province_State	Country_Region	Date	ConfirmedCases	Fatalities
1	null	Afghanistan	2020-01-22 00:00:00	0.0	0.0
2	null	Afghanistan	2020-01-23 00:00:00	0.0	0.0
3	null	Afghanistan	2020-01-24 00:00:00	0.0	0.0
4	null	Afghanistan	2020-01-25 00:00:00	0.0	0.0
5	null	Afghanistan	2020-01-26 00:00:00	0.0	0.0
6	null	Afghanistan	2020-01-27 00:00:00	0.0	0.0
7	null	Afghanistan	2020-01-28 00:00:00	0.0	0.0
8	null	Afghanistan	2020-01-29 00:00:00	0.0	0.0
9	null	Afghanistan	2020-01-30 00:00:00	0.0	0.0
10	null	Afghanistan	2020-01-31 00:00:00	0.0	0.0
11	null	Afghanistan	2020-02-01 00:00:00	0.0	0.0
12	null	Afghanistan	2020-02-02 00:00:00	0.0	0.0
13	null	Afghanistan	2020-02-03 00:00:00	0.0	0.0
14	null	Afghanistan	2020-02-04 00:00:00	0.0	0.0
15	null	Afghanistan	2020-02-05 00:00:00	0.0	0.0
16	null	Afghanistan	2020-02-06 00:00:00	0.0	0.0
17	null	Afghanistan	2020-02-07 00:00:00	0.0	0.0
18	null	Afghanistan	2020-02-08 00:00:00	0.0	0.0
19	null	Afghanistan	2020-02-09 00:00:00	0.0	0.0
20	null	Afghanistan	2020-02-10 00:00:00	0.0	0.0

Some Training Records

Fig 3: Train Data

ForecastId	Province_State	Country_Region	Date
1	null	Afghanistan	2020-04-02 00:00:00
2	null	Afghanistan	2020-04-03 00:00:00
3	null	Afghanistan	2020-04-04 00:00:00
4	null	Afghanistan	2020-04-05 00:00:00
5	null	Afghanistan	2020-04-06 00:00:00
6	null	Afghanistan	2020-04-07 00:00:00
7	null	Afghanistan	2020-04-08 00:00:00
8	null	Afghanistan	2020-04-09 00:00:00
9	null	Afghanistan	2020-04-10 00:00:00
10	null	Afghanistan	2020-04-11 00:00:00
11	null	Afghanistan	2020-04-12 00:00:00
12	null	Afghanistan	2020-04-13 00:00:00
13	null	Afghanistan	2020-04-14 00:00:00
14	null	Afghanistan	2020-04-15 00:00:00
15	null	Afghanistan	2020-04-16 00:00:00
16	null	Afghanistan	2020-04-17 00:00:00
17	null	Afghanistan	2020-04-18 00:00:00
18	null	Afghanistan	2020-04-19 00:00:00
19	null	Afghanistan	2020-04-20 00:00:00
20	null	Afghanistan	2020-04-21 00:00:00

Some Test Records

Fig 4: Test Data

```
train_sdf.createOrReplaceTempView('train_sdf')
sDF = spark.sql('SELECT Country_Region FROM train_sdf GROUP BY Country_Region')
print('Total Countries: ',len(sDF.toPandas()['Country_Region']))
sDF = spark.sql("SELECT Date FROM train_sdf GROUP BY Date")
print('Total Days: ',len(sDF.toPandas()['Date']))
```

```
▸ (4) Spark Jobs
▸ sDF: pyspark.sql.dataframe.DataFrame = [Date: string]
Total Countries:  180
Total Days:  77
Command took 0.92 seconds -- by L00157133@student.lyit.ie at 3/6/2021, 6:35:52 PM on myfirstcluster
```

Cmd 12

```
test_sdf.createOrReplaceTempView('test_sdf')
sDF = spark.sql('SELECT Country_Region FROM test_sdf GROUP BY Country_Region')
print('Total Countries: ',len(sDF.toPandas()['Country_Region']))
sDF = spark.sql('SELECT Date FROM test_sdf GROUP BY Date')
print('Total Days: ',len(sDF.toPandas()['Date']))
```

```
▸ (4) Spark Jobs
▸ sDF: pyspark.sql.dataframe.DataFrame = [Date: string]
Total Countries:  180
Total Days:  43
Command took 1.54 seconds -- by L00157133@student.lyit.ie at 3/6/2021, 6:35:53 PM on myfirstcluster
```

Fig 5: Total countries

Since it is not possible to train a regression model with multiple outputs in PySpark, separate training and testing for Confirmed Cases and Fatalities were needed, followed by the merging of their individual outputs into a single file for submission.

III. Methodology

The problem involved forecasting confirmed cases and fatalities between April 1 and April 30 by region, the primary goal wasn't only to produce accurate forecasts. It was also to identify factors that appeared to impact the transmission rate of COVID-19. I later implied the following as a procedure to do the need full Visualization, Pre-Processing, Linear Regression, Decision Tree, Random Forest.

A. Visualisation

There were 32,707 training records and 13,158 evaluation records in the data collection, with the following schemas:

1. Cases with Confirmed Virus but No Fatalities:
Fig show different visualizations for cases with confirmed virus but no fatalities. China, the United States, Australia, Canada, and France are among the top five countries in Fig. The spike occurred after 30 days, and it began to go down after about 55 days, as shown in Fig.

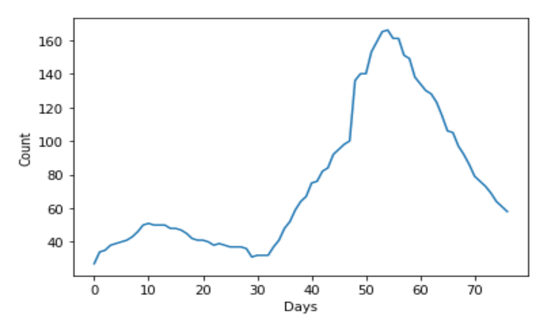
▸ (2) Spark Jobs

Country_Region	count(Fatalities)
China	1662
US	1003
Canada	122
Australia	117
France	110

Fig 5: Top 5 Confirmed

2) Sorted Confirmed Cases Per Day:
Fig. show various visualizations for cases where the virus has been reported as having caused fatalities or not. The increase in Fig. occurred after 30 days, but the data given shows that it never goes down.

▸ (4) Spark Jobs
Out[104]:



[<matplotlib.lines.Line2D at 0x7f3fc3533fd0>]

Command took 1.78 seconds -- by L00157133@student.lyit.ie at 3/6/2

Fig 6: Confirmed Graph

3) Sorted Per-Day Fatalities: Figures show various visualizations of per-day fatalities. The increase in Fig.16 occurred after 45 days, but the data given shows that it never

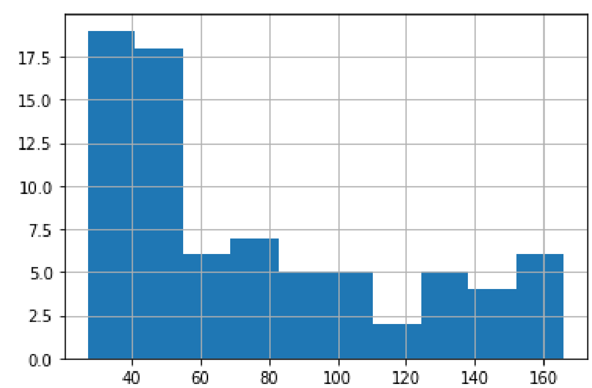
goes down.

```
▸ (4) Spark Jobs
Out[106]: count     77.000000
mean       75.584416
std        42.192962
min        27.000000
25%        41.000000
50%        59.000000
75%       100.000000
max       166.000000
Name: count(ConfirmedCases), dtype: float64

Command took 0.58 seconds -- by L00157133@student.lyi
```

Fig 7: Confirmed Describe

4) Country-wise Confirmed Cases: Figures show various visualizations for countrywide confirmed cases. China, the United States, Australia, Canada, and France are among the top five countries in Fig.20.

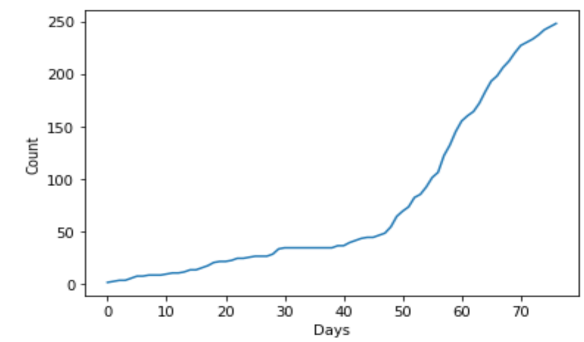



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3fc3
```

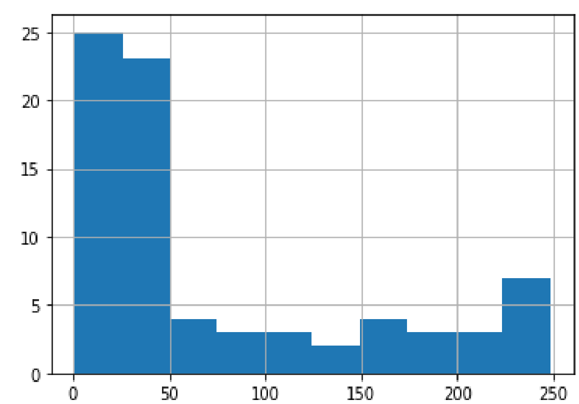
```
▸ (2) Spark Jobs
+------------------+-----------------+
|    Country_Region|count(Fatalities)|
+------------------+-----------------+
|Antigua and Barbuda|               1|
|            Belize|                2|
|             Benin|                2|
|          Barbados|                3|
|          Ethiopia|                3|
+------------------+-----------------+
```

5) Countrywide Fatalities: Figures depict various visualizations of countrywide reported cases. China, the United States, Canada, Australia, and France are the top five nations, according to Fig.23. I tried removing records with NULL values and removing Province State as a function to include all the records, but it didn't work. By eliminating Province State as a whole, I was able to achieve better results.

```
▸ (4) Spark Jobs
Out[108]:
```




```
[<matplotlib.lines.Line2D at 0x7f3fc31a7fd0>]
```




```
▸ (2) Spark Jobs
+--------------+-----------------+
|Country_Region|count(Fatalities)|
+--------------+-----------------+
|         China|             1662|
|            US|             1003|
|        Canada|              122|
|     Australia|              117|
|        France|              110|
+--------------+-----------------+
```

(2) Spark Jobs

Country_Region	count(Fatalities)
Antigua and Barbuda	1
Belize	2
Benin	2
Barbados	3
Ethiopia	3

Date	count(Fatalities)
2020-01-22	1
2020-01-23	2
2020-01-24	3
2020-01-25	3
2020-01-26	5
2020-01-27	7
2020-01-28	7
2020-01-29	8
2020-01-30	8
2020-01-31	8
2020-02-01	9
2020-02-02	10
2020-02-03	10
2020-02-04	11
2020-02-05	13
2020-02-06	13
2020-02-07	15
2020-02-08	17

Command took 0.69 seconds -- by L00157

(2) Spark Jobs

Country_Region	count(Fatalities)
Antigua and Barbuda	1
Belize	2
Benin	2
Barbados	3
Ethiopia	3

B. Pre-processing

Pre-processing tasks performed were as follows: Remove NULL values, Convert Timestamp to UnixTimestamp and Convert Categorical Attributes to Nominal.

C. Linear Regression

The most basic and widely used method of predictive analysis is linear regression. The aim of regression is to look at two things: Is it possible to predict an outcome (dependent) variable using a collection of predictor variables? The variables in particular are important predictors of the outcome variable, and how do they influence the outcome variable (as shown by the magnitude and sign of the beta estimates). These regression estimates are used to describe how one dependent variable and one or more independent variables are related. The simplest form of the regression equation with one dependent and one independent variable is y = c + b*x, where y represents the approximate dependent variable value, c represents the constant, b represents the regression coefficient, and x represents the score on the independent variable.

```
Coefficients: [0.015212651240759895,-2.615164941136036,1.396163125661749Ze-08]
Intercept: 0.0
numIterations: 67
objectiveHistory: [1703.8128481827157, 1008.4381795310941, 881.8521168284835, 873.4033753972781, 870.562866497902,
869.6423477066529, 869.0961177924488, 868.7580023550192, 868.5518679649857, 868.4258128385408, 868.362307099405З,
868.3622757117944, 868.3622713649924, 868.3622698170828, 868.3622536751863, 868.3622229769265, 868.362130176026, 8
68.3619070132879, 868.3612974706997, 868.3597352061977, 868.3558000188793, 868.3453563336791, 868.3177021950808, 8
68.2466438635433, 867.9980636431771, 867.5763666327385, 867.2856408247459, 867.2779354509449, 867.27512044611Z, 86
7.2750563020857, 867.274585656329l, 867.2736101070794, 867.2707283304234, 867.2630829422236, 867.2482899369921, 86
7.2081133412781, 867.1200218092373, 866.9130449030228, 866.861695733313Z, 866.1899452329901, 866.061985417251Z, 86
6.0202342595295, 866.0158221046959, 866.0143732559152, 866.0141153528201, 866.0140949017157, 866.0140917914835, 86
6.014090917095, 866.0140884913459, 866.0140776797706, 866.0140559252704, 866.0139367362285, 866.0136823880525, 86
6.01297728553З, 866.01124228920l, 866.006876791807, 865.9970988976153, 865.9786970449492, 865.9571861358104, 865.9
43414202851B, 865.9423457505122, 865.9374942422481, 865.937211531088, 865.9372007543752, 865.9371849941357, 865.93
71800280773, 865.9371798217732]
RMSE: 5337.372178
MAE: 645.756616
MSE: 28487541.761554
```

Fig 8: Train data

ForecastId	Province_State	Country_Region	Date	Province_StateIndex	Country_RegionIndex	features	prediction
345	Australian Capita...	Australia	1585162800	110.0	5.0	[110.0,5.0,1.5851...	10.729025426110711
346	Australian Capita...	Australia	1585249200	110.0	5.0	[110.0,5.0,1.5852...	10.73023171105128
347	Australian Capita...	Australia	1585335600	110.0	5.0	[110.0,5.0,1.5853...	10.731437995991852
348	Australian Capita...	Australia	1585422000	110.0	5.0	[110.0,5.0,1.5854...	10.732644280932425
349	Australian Capita...	Australia	1585508400	110.0	5.0	[110.0,5.0,1.5855...	10.733850565872997
350	Australian Capita...	Australia	1585594800	110.0	5.0	[110.0,5.0,1.5855...	10.735056850813569
351	Australian Capita...	Australia	1585681200	110.0	5.0	[110.0,5.0,1.5856...	10.736263135754141
352	Australian Capita...	Australia	1585767600	110.0	5.0	[110.0,5.0,1.5857...	10.73746942069471
353	Australian Capita...	Australia	1585854000	110.0	5.0	[110.0,5.0,1.5858...	10.738675705635282
354	Australian Capita...	Australia	1585940400	110.0	5.0	[110.0,5.0,1.5859...	10.739881990575855
355	Australian Capita...	Australia	1586026800	110.0	5.0	[110.0,5.0,1.5860...	10.741088275516427
356	Australian Capita...	Australia	1586113200	110.0	5.0	[110.0,5.0,1.5861...	10.742294560457
357	Australian Capita...	Australia	1586199600	110.0	5.0	[110.0,5.0,1.5861...	10.743500845397572
358	Australian Capita...	Australia	1586286000	110.0	5.0	[110.0,5.0,1.5862...	10.744707130338144
359	Australian Capita...	Australia	1586372400	110.0	5.0	[110.0,5.0,1.5863...	10.745913415278713
360	Australian Capita...	Australia	1586458800	110.0	5.0	[110.0,5.0,1.5864...	10.747119700219285
361	Australian Capita...	Australia	1586545200	110.0	5.0	[110.0,5.0,1.5865...	10.748325985159857
362	Australian Capita...	Australia	1586631600	110.0	5.0	[110.0,5.0,1.5866...	10.74953227010043
363	Australian Capita...	Australia	1586718000	110.0	5.0	[110.0,5.0,1.5867...	10.750738555041002
364	Australian Capita...	Australia	1586804400	110.0	5.0	[110.0,5.0,1.5868...	10.751944839981574

Fig 9: Test data

```
Coefficients: [5.449557612239998e-10,-5.6189620828913035e-09,1.2602663552243827e-17]
Intercept: 0.0
numIterations: 74
objectiveHistory: [1194.7435104344768, 28.272935594723386, 27.31932307279021, 27.317282278743725, 27.3171162631992
74, 27.316562298683372, 27.310225686189177, 27.29673505244733, 27.261498927121803, 27.183427561156883, 26.99008171
05179, 26.64803708319912, 26.574517458316404, 26.536126922807227, 26.53604616297333, 26.536039393579294, 26.536036
32907706, 26.53600181404246, 26.535913682008847, 26.535661461140602, 26.535023126341347, 26.533336802232814, 26.52
9029388336067, 26.51804042280626, 26.50782299088743, 26.47777348688777, 26.418244431802986, 26.39829872492616, 26.
398277981702897, 26.398259914663186, 26.398259477876433, 26.398257777653477, 26.398240889041244, 26.39820824913963
5, 26.39811165007662, 26.397871701551708, 26.397237196347522, 26.395620631542926, 26.391513005350216, 26.381198555
783502, 26.363672166086282, 26.32755292011903, 26.313841689475197, 26.313027258700078, 26.31249892173413, 26.31217
9968262424, 26.309538255751857, 26.306262883611645, 26.304462382768126, 26.304392881063695, 26.302989982448377, 2
6.302013971782795, 26.300975115127383, 26.300019518863945, 26.297707043657514, 26.29758662101335, 26.2974976197816
8, 26.29746402674083, 26.29745055516667, 26.297395725970247, 26.29731911043207, 26.29731698729302, 26.297316546053
217, 26.29731169370514, 26.297310468571023, 26.29730879744083, 26.297308331413884, 26.297308045345893, 26.29730800
7884435, 26.297307950165376, 26.297307860705033, 26.29730779168513, 26.297307769747047, 26.297307731067512]
RMSE: 211.088569
MAE: 19.479487
MSE: 44558.383973
```

Fig 10: Fatalities Train

ForecastId	Province_State	Country_Region	Date	Province_StateIndex	Country_RegionIndex	features	prediction
345	Australian Capita...	Australia	1585162800	110.0	5.0	[110.0,5.0,1.5851...	5.182759676411623E-8
346	Australian Capita...	Australia	1585249200	110.0	5.0	[110.0,5.0,1.5852...	5.182868563424714E-8
347	Australian Capita...	Australia	1585335600	110.0	5.0	[110.0,5.0,1.5853...	5.182977450437806E-8
348	Australian Capita...	Australia	1585422000	110.0	5.0	[110.0,5.0,1.5854...	5.183086337450897...
349	Australian Capita...	Australia	1585508400	110.0	5.0	[110.0,5.0,1.5855...	5.183195224463988...
350	Australian Capita...	Australia	1585594800	110.0	5.0	[110.0,5.0,1.5855...	5.18330411147708E-8
351	Australian Capita...	Australia	1585681200	110.0	5.0	[110.0,5.0,1.5856...	5.183412998490171E-8
352	Australian Capita...	Australia	1585767600	110.0	5.0	[110.0,5.0,1.5857...	5.183521885503263E-8
353	Australian Capita...	Australia	1585854000	110.0	5.0	[110.0,5.0,1.5858...	5.183630772516354...
354	Australian Capita...	Australia	1585940400	110.0	5.0	[110.0,5.0,1.5859...	5.183739659529446E-8
355	Australian Capita...	Australia	1586026800	110.0	5.0	[110.0,5.0,1.5860...	5.183848546542537E-8
356	Australian Capita...	Australia	1586113200	110.0	5.0	[110.0,5.0,1.5861...	5.183957433555628E-8
357	Australian Capita...	Australia	1586199600	110.0	5.0	[110.0,5.0,1.5861...	5.18406632056872E-8
358	Australian Capita...	Australia	1586286000	110.0	5.0	[110.0,5.0,1.5862...	5.184175207581811E-8
359	Australian Capita...	Australia	1586372400	110.0	5.0	[110.0,5.0,1.5863...	5.184284094594903E-8
360	Australian Capita...	Australia	1586458800	110.0	5.0	[110.0,5.0,1.5864...	5.184392981607994E-8
361	Australian Capita...	Australia	1586545200	110.0	5.0	[110.0,5.0,1.5865...	5.184501868621085E-8
362	Australian Capita...	Australia	1586631600	110.0	5.0	[110.0,5.0,1.5866...	5.184610755634176...
363	Australian Capita...	Australia	1586718000	110.0	5.0	[110.0,5.0,1.5867...	5.184719642647268E-8
364	Australian Capita...	Australia	1586804400	110.0	5.0	[110.0,5.0,1.5868...	5.18482852966036E-8

Fig 11: Fatalities Test

D. Decision Tree

By splitting the source set into subsets based on an attribute value test, a tree can be "learned." Recursive partitioning is the process of repeating this process on each derived subset. When all of the subsets at a node have the same value of the target variable, or when splitting no longer adds value to the predictions, the recursion is complete. Since the construction of a decision tree classifier does not necessitate domain awareness or parameter setting, it is suitable for exploratory knowledge exploration. High-dimensional data can be handled by decision trees. The accuracy of the decision tree classifier is generally fine. Decision tree induction is a popular inductive method for learning classification information.

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	1270.5553977272727
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	1270.5553977272727
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	1270.5553977272727
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	1270.5553977272727
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	1270.5553977272727
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	1270.5553977272727
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	1270.5553977272727
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	1270.5553977272727
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	1270.5553977272727
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	1270.5553977272727
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	1270.5553977272727
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	1270.5553977272727
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	1270.5553977272727
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	1270.5553977272727
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	1270.5553977272727
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	1270.5553977272727
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	1270.5553977272727
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	1270.5553977272727
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	1270.5553977272727
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	1270.5553977272727

Fig 12: DT Confirmed

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	4.038628472222222
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	4.038628472222222
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	4.038628472222222
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	4.038628472222222
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	4.038628472222222
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	4.038628472222222
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	4.038628472222222
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	4.038628472222222
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	4.038628472222222
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	4.038628472222222
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	4.038628472222222
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	4.038628472222222
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	4.038628472222222
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	4.038628472222222
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	4.038628472222222
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	4.038628472222222
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	4.038628472222222
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	4.038628472222222
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	4.038628472222222
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	4.038628472222222

Fig 13: DT Fatalities

Decision trees identify objects by sorting them down the tree from the root to a leaf node, which determines the object's classification. Starting at the root node of the tree, evaluate the attribute defined by this node, then move down the tree branch corresponding to the value of the attribute, an instance is categorized. The sub-tree rooted at the new node is then processed in the same way.

E. Random Forest

As the name suggests, a random forest is made up of a large number of individual decision trees that work together as an ensemble. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of our model.

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	858.2640841233557
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	1197.2721431851119
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	1261.8282497948262
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	1296.2584551241996
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	1296.2584551241996
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	1408.6840448779562
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	1519.9895258385063
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	6697.227883989891
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	6697.227883989891
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	6697.227883989891
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	6697.227883989891
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	6697.227883989891
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	6697.227883989891
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	6697.227883989891
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	6697.227883989891
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	6697.227883989891
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	6697.227883989891
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	6697.227883989891
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	6697.227883989891
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	6697.227883989891

Fig 14: RF Confirmed

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	33.358643996182145
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	45.18073419328118
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	47.589548340332385
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	50.61109559387538
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	50.61109559387538
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	62.25109495532659
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	65.12304361083373
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	80.06641098215732
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	80.06641098215732
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	80.06641098215732
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	80.06641098215732
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	80.06641098215732
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	80.06641098215732
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	80.06641098215732
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	80.06641098215732
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	80.06641098215732
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	80.06641098215732
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	80.06641098215732
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	80.06641098215732
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	80.06641098215732

Fig 15: RF Fatalities

F. Gradient Boosted

A gradient boosted model is a set of regression or classification tree models that have been combined. Both are forward-learning ensemble approaches that boost estimations over time to produce predictive results. Boosting is a versatile nonlinear regression technique that aids in improving the accuracy of nonlinear regression models a forest Applying poor classification algorithms in a sequential manner A series of decision trees are applied to the incrementally modified results. are made, resulting in a collection of weak predictions prototypes Although boosting trees improves their accuracy, it comes at a cost It also slows down the process and makes it more difficult for humans to understand. The incline To reduce these, the boosting method generalizes tree boosting problems.

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	691.1942022789012
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	691.1942022789012
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	691.1942022789012
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	691.1942022789012
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	589.0975839360901
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	463.2658822234095
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	316.7483464595758
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	338.8041600947142
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	338.8041600947142
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	338.8041600947142
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	338.8041600947142
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	338.8041600947142
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	338.8041600947142
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	338.8041600947142
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	338.8041600947142
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	338.8041600947142
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	338.8041600947142
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	338.8041600947142
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	338.8041600947142
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	338.8041600947142

Fig 16: Gradient Boost Confirmed

ForecastId	Country_Region	Date	Country_RegionIndex	features	prediction
1	Afghanistan	1585162800	38.0	[38.0,1.5851628E9]	10.329055445820146
2	Afghanistan	1585249200	38.0	[38.0,1.5852492E9]	20.117558048480596
3	Afghanistan	1585335600	38.0	[38.0,1.5853356E9]	20.117558048480596
4	Afghanistan	1585422000	38.0	[38.0,1.585422E9]	21.456640807870762
5	Afghanistan	1585508400	38.0	[38.0,1.5855084E9]	28.604925293360502
6	Afghanistan	1585594800	38.0	[38.0,1.5855948E9]	30.64581189807014
7	Afghanistan	1585681200	38.0	[38.0,1.5856812E9]	19.541957286661482
8	Afghanistan	1585767600	38.0	[38.0,1.5857676E9]	16.329802334099423
9	Afghanistan	1585854000	38.0	[38.0,1.585854E9]	16.329802334099423
10	Afghanistan	1585940400	38.0	[38.0,1.5859404E9]	16.329802334099423
11	Afghanistan	1586026800	38.0	[38.0,1.5860268E9]	16.329802334099423
12	Afghanistan	1586113200	38.0	[38.0,1.5861132E9]	16.329802334099423
13	Afghanistan	1586199600	38.0	[38.0,1.5861996E9]	16.329802334099423
14	Afghanistan	1586286000	38.0	[38.0,1.586286E9]	16.329802334099423
15	Afghanistan	1586372400	38.0	[38.0,1.5863724E9]	16.329802334099423
16	Afghanistan	1586458800	38.0	[38.0,1.5864588E9]	16.329802334099423
17	Afghanistan	1586545200	38.0	[38.0,1.5865452E9]	16.329802334099423
18	Afghanistan	1586631600	38.0	[38.0,1.5866316E9]	16.329802334099423
19	Afghanistan	1586718000	38.0	[38.0,1.586718E9]	16.329802334099423
20	Afghanistan	1586804400	38.0	[38.0,1.5868044E9]	16.329802334099423

Fig 17: Gradient Boost Fatalities

IV. Evaluation

I have used root mean square to perform the required evaluation on models.

V. Results

Below are the tabular results for all the methods of machine learning being employed on the dataset.

Linear Regression	Including Province State	Without Including Province State
Score	3.65539	3.35323

TABLE I
RESULTS USING LINEAR REGRESSION

Decision Tree	Depth = 3	Depth = 5
Score	2.49157	2.38298

m

TABLE II
RESULTS USING DECISION TREE

Random Forest	Trees = 2	Trees = 20	Trees = 100
Score	3.13457	3.15120	3.20820

TABLE III
RESULTS USING RANDOM FOREST

Depth	Score
3	2.54647
5	2.05806
7	2.01467
9	1.98333
30	1.98171

TABLE IV
RESULTS USING GRADIENT BOOSTED TREE

VI. CONCLUSION

There is multiple more analysis and implementation of the advance machine learning algorithms which can be achieved in here.

VII. REFERENCES

[1] https://www.kaggle.com/c/covid19-global-forecasting-week-3/overview