COVID-19 total case prediction: Case study for Prediction using Machine Learning Using Python.

COVID-19 has been a pandemic of high concern due to its effects on human health and the large amount of deaths caused by the SARS-CoV-2 virus widespread across the world in today's situation.

It is Very important for us to know the total number of cases that could have been on a particular day according to the daily updates of the new cases in that location.

Hence, the aim of this Study was to build a machine learning model with 2 of the extremely popular Regression Algorithms such as Random Forest Regressor as well as the Linear Regression Algorithm.

Here we not only predict the Total Cases but also do the Analysis of which model could be of best fit for the type of data we are supplying.

Here We are fetching the data from Owid Covid Data Website where we are using the latest data of 18th July,2020.

Due to Changes in the latest document, the results may vary.

The Description of data was as followed below:

Column	Mean	Mode	Median
Total_cases	136393.683417085	5734.0	0.0
New_cases	5044.381909547738	693.0	0.0
Total_deaths	3938.934673366834	166.0	0.0
New_deaths	128.6532663316583	27.0	0.0
total_cases_per_million	98.83563316582915	4.155	0.0
new_cases_per_million	3.6553417085427142	0.502	0.0
total_deaths_per_million	2.85429648241206	0.12	0.0
new_deaths_per_million	0.09321608040201008	0.02	0.0
total_tests	3757080.3421052634	2564154.0	there are 114 mode values)
new_tests	116709.27777777778	103523.0	there are 108 mode values)
total_tests_per_thousand	2.7225438596491225	1.8584999999999998	0.01
new_tests_per_thousand	0.08456481481481484	0.075	0.001
new_tests_smoothed	99516.63865546219	89872.0	1125.0

Here it can be observed that the data till date of 18th July ,2020 is as above.

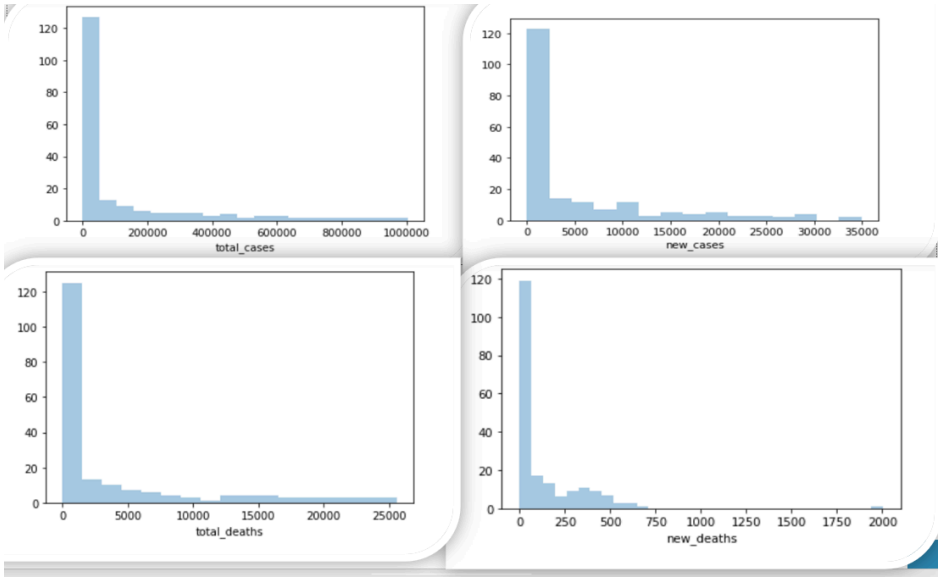
The actual columns which are used as the Independent or the Feature Variables are being showed above. Various Basic calculations such as Mean, Median, Mode are being calculated here.

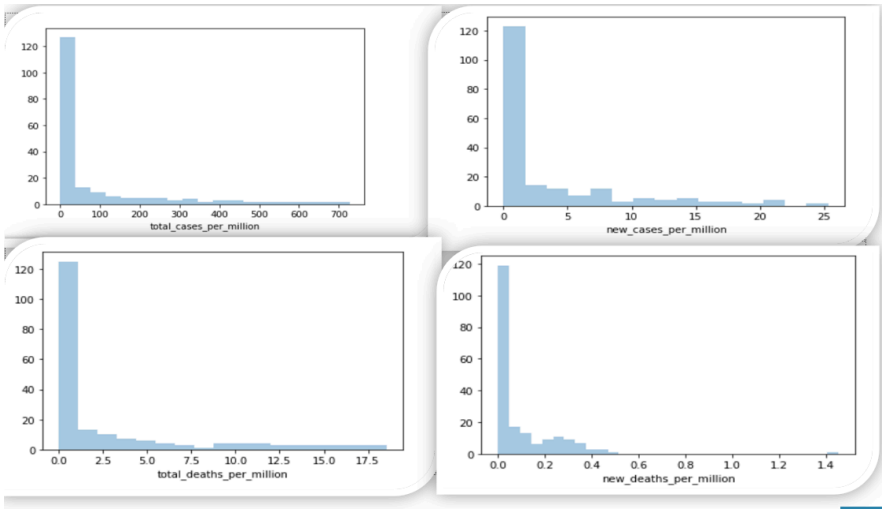
Various Analysis was done on the data:

1.Univariate analysis:

Histogram is one of the common univariate analysis methods where the distribution of the data can be observed. The Various feature variables used in our analysis are being plotted to find their distribution of data across the dataset.

The Analysis is as follows:

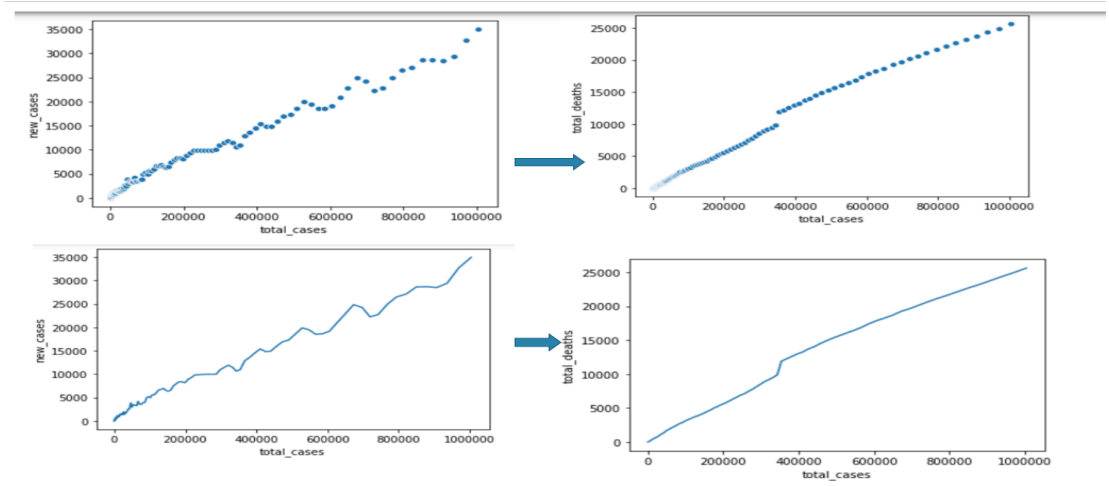


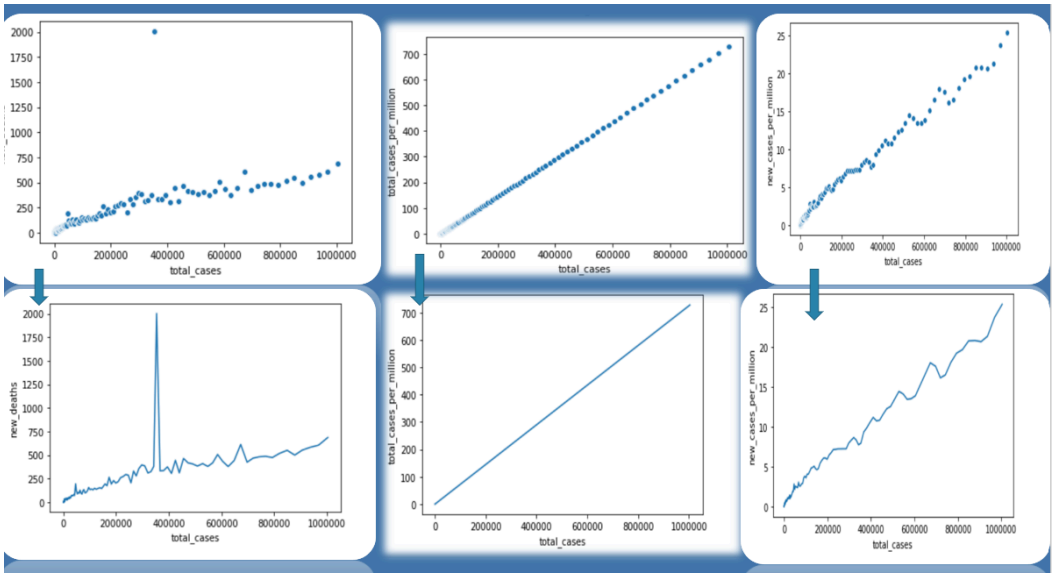


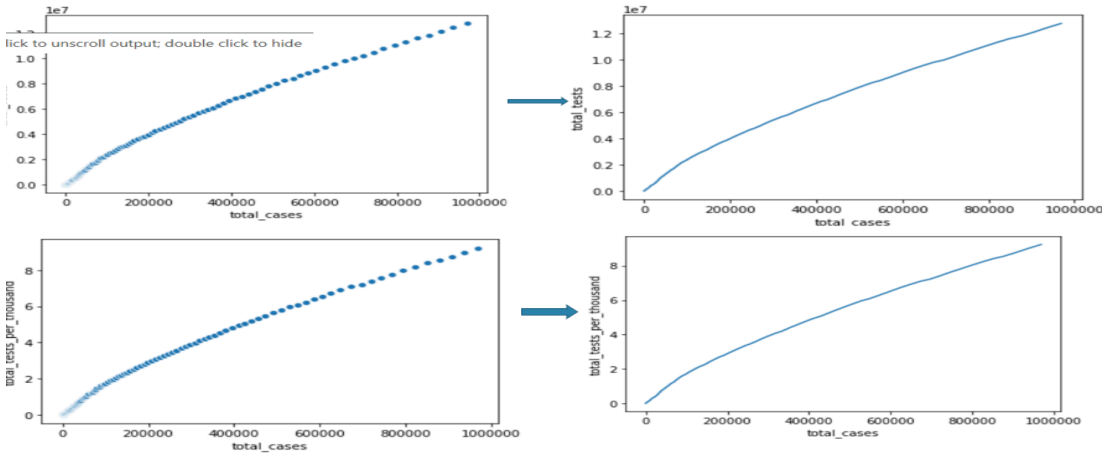
The Histograms tells us about the various values in data, the nature of data whether it's discrete or continuous, it also observes the nature of data distribution and so on.

Bivariate analysis:

- The Bivariate Analysis was done between data of any of the two Numerical columns from the dataset
- Bivariate analysis shows that many of the columns have linear relationships as well exponential relationships.
- Here the plots of Total Cases v/s other important features will be used







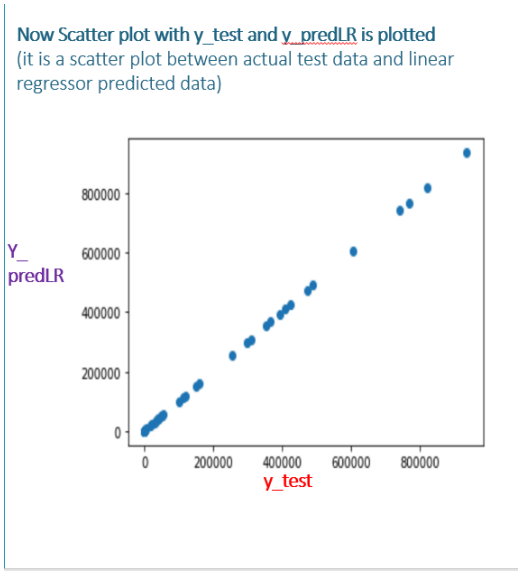
It can be observed from above that the Dependent or the target variable total_cases is majorly linear or exponential in nature.

The Linear Regression Model:

Linear Regression is one of the popular regression model built for continuous data prediction. The linear regression model fits a line

$Y=\beta_o+\beta_1 X_i$

- The Following Dataset after clearing the Null Values and Unnecessary Data columns the Model is fit to Linear Regression Model.
- Train test split was done and the training data is fit into the model.
- Then the testing is done accordingly.



Analysis of fit of Linear Model on the data:

The Various Tests or Validation was been done and their scores are as follows:

1. **Accuracy was found to be :**0.9999999999968615
2. RMSE score:

   0.4202825401794529

3.Five fold Cross validation score:

   0.9865092359544476

Conclusions:

The Conclusion is that the Model is quiet in good fit with data and can be expected to give good prediction results near to the originals.

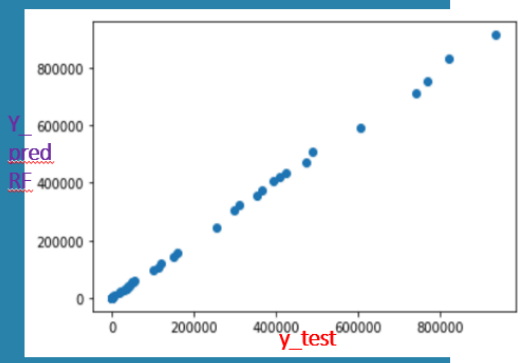
Random Forest Regressor:

Random Forest Regression is on of the Ensemble Techniques called as Bagging.

The Random Forest is an ensemble of Decision Trees where the Vote Casting is done to find the final output.

- The Following Dataset which was cleared for Linear regression from Null Values and Unnecessary Data columns is used .
- Model is fit to Linear Regression Model.
- Train test split was done and the training data is fit into the model.
- Then the testing is done accordingly.
- The testing prediction was scatter plotted against y_test in the data and deviations were found

Now Scatter plot with y_test and y_predRF is plotted
(it is a scatter plot between actual test data and linear regressor predicted data)



1. Accuracy was found to be :

   0.998994974199048

   2. RMSE score:

   0.4202825401794529

   3.Five fold Cross Val score:

   -3.59914

   Conclusions:

   The Model has good Accuracy and RMSE score but the Five fold Cross Val score is negative. Hence the prediction might be of less accuracy

   Prediction of data done on 17th July,2020:

   Here we are predicting the total_cases on 17th of July

   The Data is as follows:

Data prediction of Total Cases on 17th July,2020

The following data was given as input to the to the Linear Regression model and Random Forest Regression Model.

Date	737623(Ordinal value of 2020-07-17)
New_cases	34956
Total_deaths	25602
New_deaths	687
Total_cases_per_million	727.412
New_cases_per_million	25.33
Total_deaths_per_million	18.552
New_deaths_per_million	0.498
Total_tests	3.757080e+06
New_tests	116709.277778
total_tests_per_thousand	2.722544
new_tests_per_thousand	0.084565

Results:

1.Linear Regression Prediction on Total cases = 1003831.90603254

2.Random Forest Regression prediction on Total cases=813695.315

3.Total Number of Covid cases on 17th July in Actual= 1003832

(as per data)

So it can be concluded that Linear Regressor fits really well than that of Random Forest Regressor.