International Journal for Modern Trends in Science and Technology, 6(8S): 58-62, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTSTCIET12
Available online at: http://www.ijmtst.com/ncracse2020.html

Using Machine Learning Approach to Predict Covid-19 Progress

G Monica[1] | Dr.M.Bharathi Devi [2]

[1]Computer Science and Engineering, Blekinge Institute of Technology , Karlskona, Sweden, Europe
[2]Computer Science and Engineering, Rise Krishna Sai Gandhi Groups of Institutions , Ongole, Andhra Pradesh, India

To Cite this Article
G Monica, Dr.M.Bharathi Devi, "Using Machine Learning Approach to Predict Covid-19 Progress", *International Journal for Modern Trends in Science and Technology*, 6(8S): 58-62, 2020.

Article Info
Received on 16-July-2020, Revised on 15-August-2020, Accepted on 25-August-2020, Published on 28-August-2020.

ABSTRACT

The Corona Virus (COVID 19) has causes in excess of approximately 6 million cases on the planet up until now and with that number continuing to create. More research is proceeding to foresee and a remarkable number of machine learning models are being discussed to predict COVID-19 pandemic used by experts around the world to pick showed choices and support gigantic control measures .Among the standard models for COVID-19 overall pandemic desire, fundamental epidemiological and truthful models have gotten more thought by authorities, and they are notable in the media. Due to a huge degree of weakness and nonappearance of principal data, standard models have shown low exactness for long haul conjecture. This paper presents an examination of Machine learning models to anticipate the crown infection pandemic. This paper gives a basic benchmarking to show the capacity of AI for future research.

KEYWORDS: : *Corona Virus, Polynomial Regression, Decision Tree Regressor, Random Forest Regressor, RMSE, R2 Score*

I. INTRODUCTION

Coronavirus (COVID-19) gets the world to experience its misery stage. The first case was accounted for as unidentified pneumonia in China in the month of December, 2019[1], later identified as Coronavirus. It has been declared a pandemic by the world Health Organization (WHO) on 11 March 2020 due to the high contiguousness and spread rapidly across the countries. At that point as time streams, the quantity of coronavirus victims spread exponentially.

Initially COVID causes dry coughing, fever and shortness of breath. Earlier reports suggest COVID as an air borne and droplet infection [2]. This contagious infection spread while the people are nearer to one another. It also spread by means of latent interconnection with the beads and this may be causing respiratory sickness, throat contamination and influence the lungs which leads to death. Due to the nature of COVID transmission, it is spreading quickly over world. 213 Countries and Territories around the globe are under danger of this disease. There is a sum of 3,008,297 active cases and 367,227 are expired, along with 2,672,712 recovered are accounted as on 30-05-2020.

Some of the countries implementing lockdown, it worked out that the restriction of the peoples migration was viable in controlling pandemic diseases like COVID. Many components may impact the COVID pandemic, including social

distancing, topographical elements, climate variables and so forth.

Ajit Kumar Pasayat et al.[3] discussed the Prediction of COVID-19 positive cases in India with concern to Lockdown by using the exponential growth model and linear regression model through Machine Learning.Sina F. Ardabiliet al. [4] analysed a machine learning and soft computing models to predict the COVID-19 outbreak as an alternative to SIR and SEIR models.

In this paper, COVID 19 confirmed cases are predicted by Polynomial Regression, Decision Tree Regressorand Random Forest Regressor. Further comparative analysis between these three machine learning Regressor models are discussed.

II.METHODS AND MATERIALS:

A.Data Resource:

Data was extracted from JOHN HOPKINS GITHUB REPOSITORY https://github.com/CSSEGISandData/COVID-19 [5]. In this paperthe dataset consists of 7 fields and 27166 records is used which contains Confirmed cases, Deaths and Recovered cases of COVID-19 as on 20,May 2020.

B. Data Analysis:

Weekly wise data of confirmed cases, death cases and recovered cases of COVID-19 as on 20 May 2020 was analysed by python libraries such as numpy , pandas, matplotlib and seaborn

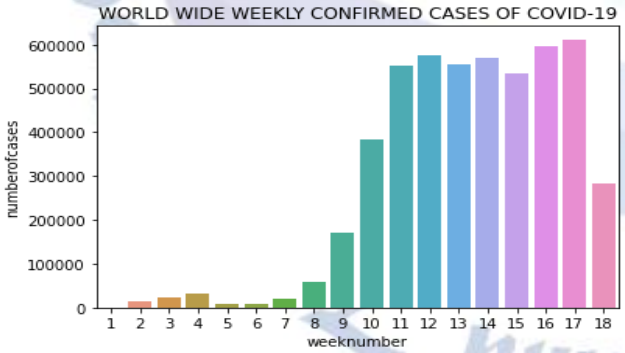

Figure 1: Weekly wise confirmed cases of COVID-19

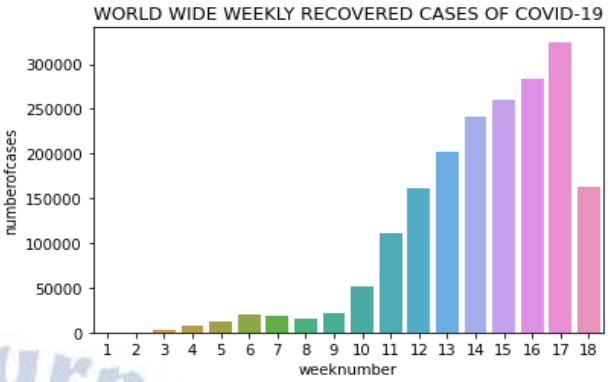

Figure 2: Weekly wise recovered cases of COVID-19

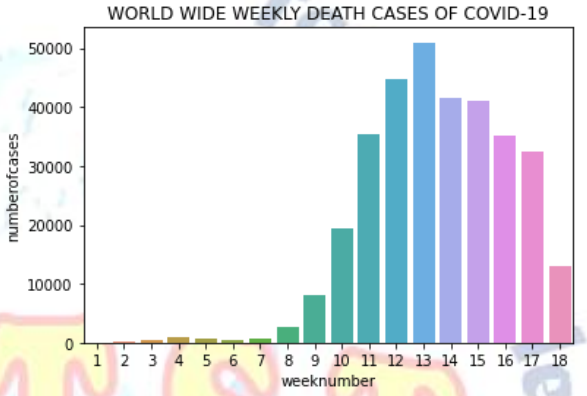

Figure 3: Weekly wise death cases of COVID-19

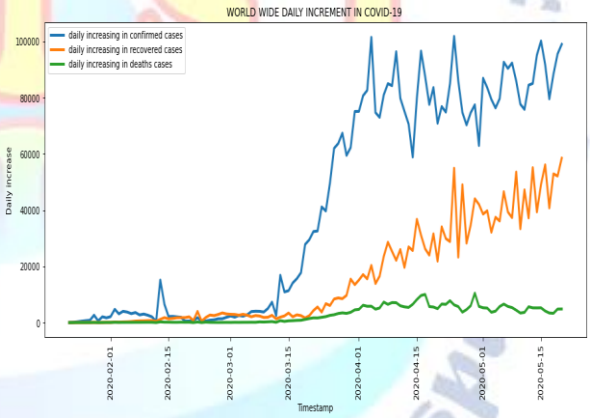

Figure 4: Distribution plot for week wise confirmed cases and death cases

Figure 4 indicates the bar plot of COVID-19 week wise confirmed cases from 22-01-2020 to 20-05-2020, in which x-axis and y-axis indicates the weeks and confirmed cases respectively. In this figure, we observed that up to 10th week COVID-19 confirmed cases are reached to 4 million approximately and then with in a week time it reaches nearly 5 millions.Figure 2 and 3 represents the bar plots of worldwide weekly progress of COVID 19 Recovered and Death cases respectively. In figure 4, daily increment of confirmed cases indicated by blue line, recovered cases by orange line and death cases by green line. From this figure, growth rate of confirmed cases is more

compare to the growth rate of recovered and death cases of COVID-19.

C.Models:

In this paper three different Machine Learning Models ,Polynomial Regression, Decision Tree Regressor and Random Forest Regressor are discussed for the prediction of the COVID-19 confirmed cases over the world wide.

In statistics, the Mean Squared Error (MSE) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors. The MSE is a measure of the quality of an estimator—it is always non-negative, and values closer to zero are better. The Mean Squared Error is given by the formula $\frac{1}{n}\sum_{i=1}^{n}(y_i -$ yi2 and Root Mean Squared Error (RMSE) is calculated by $\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y_i})^2}{n}}$.

D.Decision Tree Regression:

In Python, Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.Decision trees where the target variable can take continuous values (typically real numbers) are called regression tree. Similar in decision tree classification, however uses mean squared error or similar metrics instead of cross entropy or Gini impurity to determine split.

E.Random Forest Regression:

Every decision tree has high variance, but when combine all together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier but in regression problem, the final output is the mean of all the outputs. A Random Forest is a technique used to perform both regression and classification models with the support of multiple decision trees and is called Bootstrap and Aggregation, traditionally call as **bagging**. The basic idea of this technique is to combine multiple
decision trees in the process of getting final output rather than relying on individual decision trees .Random Forest has multiple decision trees as base learning models.
In Decision Tree and Random Forest models MSE is used to estimate the error.

Polynomial Regression:

Linear Regression is a basic model which used to show the linear relationship between a target variable y and one or more feature variables x. If we have an n number of feature variables, we use polynomial regression model and it assumes an equation $y = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots \ldots + a_nx^n$. In this model RMSE is used to estimate the error.

III.RESULTS AND DISCUSSIONS:

The Decision Tree of COVID-19 data has root node as an observation date and with a maximum depth 100 and is follows:

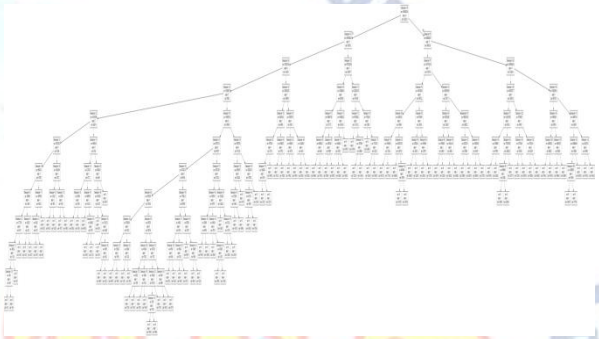

Figure 5 : Decision Tree for COVID-19 data

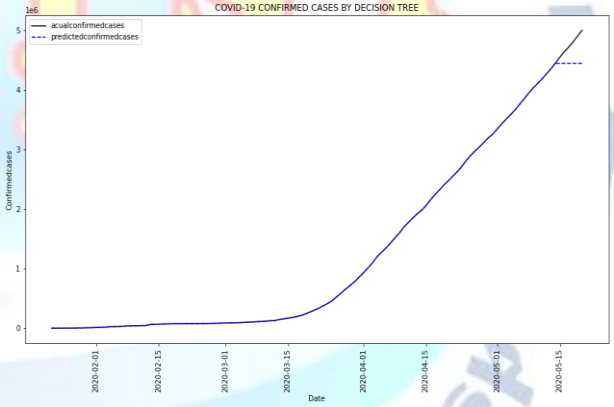

Figure 6 :Confirmed cases of COVID-19 by Decision Tree

In this figure, Y axis represented for Number of Confirmed cases and X axis is for the Date. The Number of Confirmed cases according to this model is represented with blue dotted line and the Actual Number of Confirmed cases represented with black line. In this model, the RMSE value is 356972 and the observation to be noted here is after 09 may,2020 the prediction became linear.

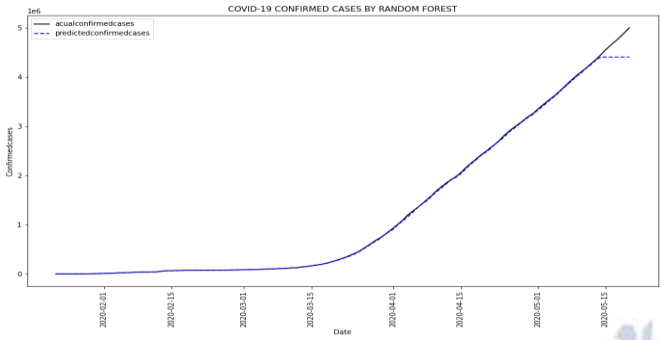

Figure 7 :Confirmed cases of COVID-19 by Random Forest

In figure 7, X axis represented the Date. And Y axis is for Number of Confirmed cases of COVID-19. The Number of Confirmed cases according to this model is represented with blue dotted line and the Actual Number of Confirmed cases represented with black line. The RMSE value by Random Forest is 406649 and after 09 may,2020 the prediction became linear is observed in this model also.

In Linear Regression, the relationship is in between one dependent variable and one or more independent variables. Here we are using an independent variable as the number of days and the dependent variable is the confirmed cases of COVID-19. By applying sklearn.preprocessing. Polynomial Features we convert a single independent feature to a 8-degree polynomial feature to the regression curve with the hypothesis. Then we divide our data into train and test sets with the ratio 95:5. The model is trained with training data and validated against test data by using the following equation generated by the regression model.

The Polynomial regression model with degree 8 assumes the equation $y = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5 + a_6x^6 + a_7x^7 + a_8x^8$and the intercept and corresponding coefficients are -32362.518686311785,3.00430246e+04,5.58065056e+03,4.31742664e+02,-1.59721503e+01,3.09535163e-01,-3.20326831e-03,1.68367007e-05, -3.54097020e-08.

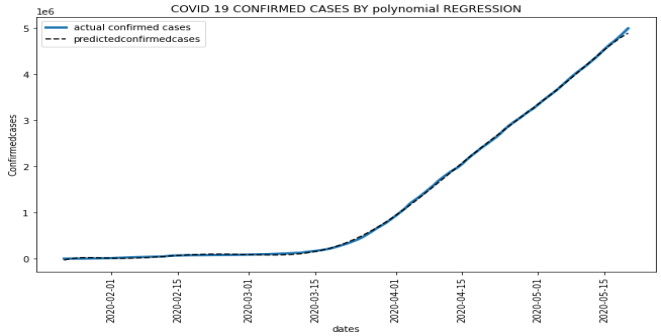

Figure 8 : COVID-19 confirmed cases through Polynomial Regression

In this figure, Y axis is represented for Number of Confirmed cases and X axis represented the Dates. Predicted Confirmed cases according to this model are represented with blue dotted line and the Actual Confirmed cases are represented with black line. Hence COVID-19 confirmed cases through Polynomial regression model is best fit with an accuracy of 90% by calculating r2 score and 53262.68 by RMSE value.

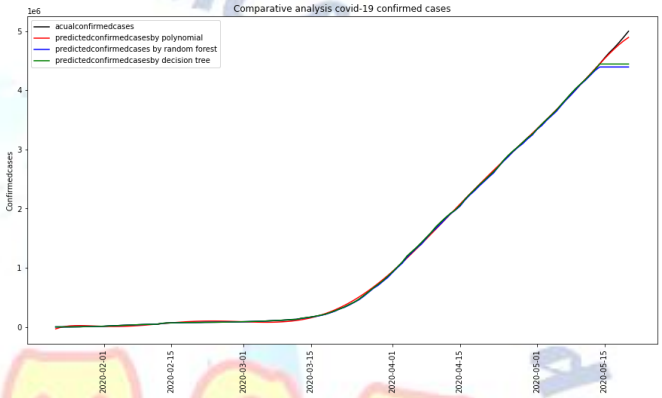

Figure 9: Comparative analysis of COVID-19 confirmed cases

Figure 9 explains comparative analysis of three machine learning models. In this figure, Y axis represents number of Confirmed cases and X axis represents the Date. The Number of Confirmed cases by decision tree is represented with green line, Number of Confirmed cases by random forest is represented with blue line ,Number of Confirmed cases by polynomial regression is represented with orange line and the Actual Number of Confirmed cases represented with black line. Here polynomial regression is best model to predict the COVID-19 confirmed cases compare to other two models. Polynomial regression is exactly coincides with actual confirmed cases when compared to the predicted confirmed cases by decision tree and random forest is observed.

IV. Conclusion

In this paper prediction of COVID-19 confirmed cases by Machine Learning models, Decision Tree, Random Forest and Polynomial Regression are studied. Further we concentrate on more ML models to predict the COVID-19 progress with more accuracy.

.

REFERENCES

[1] D. Fanelli and F. Piazza, "Analysis and forecast of covid-19 spreading in china, italy and france,"Chaos, Solitons& Fractals, vol. 134, p. 109761, 2020.

[2] Novel coronavirus diagnosis protocol. https://en.wikipedia.org/wiki/2019%E2%80%9320_coronavirus_pandemic

[3] Ajit Kumar Pasayat et al. - Predicting the COVID-19 positive cases in India with concern to Lockdown by using Mathematical and Machine Learning based Models,

[4] doi: https://doi.org/10.1101/2020.05.16.20104133

[5] Sina F. Ardabili et al. - COVID-19 Outbreak Prediction with Machine Learning,

[6] https://www.researchgate.net/publication/340782507

[7] JOHN HOPKINS GITHUB REPOSITORY https://github.com/CSSEGISandData/COVID-19