UniversePG

Publisher homepage: www.universepg.com, ISSN: 2663-7804 (Online) & 2663-7790 (Print)

https://doi.org/10.34104/ajeit.022.01300139

Australian Journal of Engineering and Innovative Technology

Journal homepage: www.universepg.com/journal/ajeit

Study and Prediction of Covid-19 Cases and Vaccinations Using Machine Learning in Bangladesh

Md. Anwar Hossain[1]*, Ebrahim Hossen[1], and Md Asraful[2]

[1,2]Department of Information and Communication Engineering, Pabna University of Science and Technology, Pabna-6600, Bangladesh.

*Correspondence: manwar.ice@pust.ac.bd (Md. Anwar Hossain, Associate Professor, Department of Information and Communication Engineering, Pabna-6600, Bangladesh).

ABSTRACT

Coronavirus disease 2019 (COVID-19) is a contagious disease caused by a virus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This first known case was identified in Wuhan, China in December 2019. The first case of COVID-19 was detected on 8th March 2020 in Bangladesh. Subsequently, COVID-19 cases have been increasing at an alarming rate in Bangladesh because of the high-density population with limited resources. However, to protect against this type of disaster there is an urgent need for high-quality data and forecasting for national preparedness and action plans, which is currently unavailable. Therefore, we developed a machine learning-based project that provides us with trend analysis and predicts the likely number of cases and deaths in upcoming days, Distributed Bangladesh in different zone according to the number of cases to increase people's awareness, finding the correlation between coronavirus and any other disease. For this investigation, we applied linear regression, polynomial regression, Extra Tree Classifier, Decision Tree Regression, and chi-square test machine learning algorithm with high and best accuracy.

Keywords: Data preprocessing, Chi-square test, Linear regression, and COVID-19 vaccinations.

INTRODUCTION:

COVID-19 is the most global epidemic recorded in recent times. It first broke out in Wuhan, China in December 2019 and has since spread throughout the world. Bangladesh first detected COVID-19 on 8th March 2020 (Wikipedia, 2021). Since its, invention the virus is increasing rapidly in Bangladesh because of the high population density and people's unawareness. Here below is a map of **Fig. 1**, Dhaka and Narayangonj are the most affected by the virus that's defined as deep red color and Rangamati is less affected by the virus that's defined as light red colored. Deep colored also represented the dangerous area for COVID-19 because of the increase in new cases and new death, on the other hand, Light colored are also represented the safe area for COVID-19 because of fewer new cases and new death (Susanto *et al*., 2022).

In Bangladesh death are increasing according to the increase of new cases and we have also lost 29,359 people in Bangladesh. Not only death but also virus destroys our economic policy. In this study using the machine learning algorithm, we predict new cases and new death to increase people's awareness and also find the correlation between new death and other's attribute and last we discuss COVID-19 vaccinations.

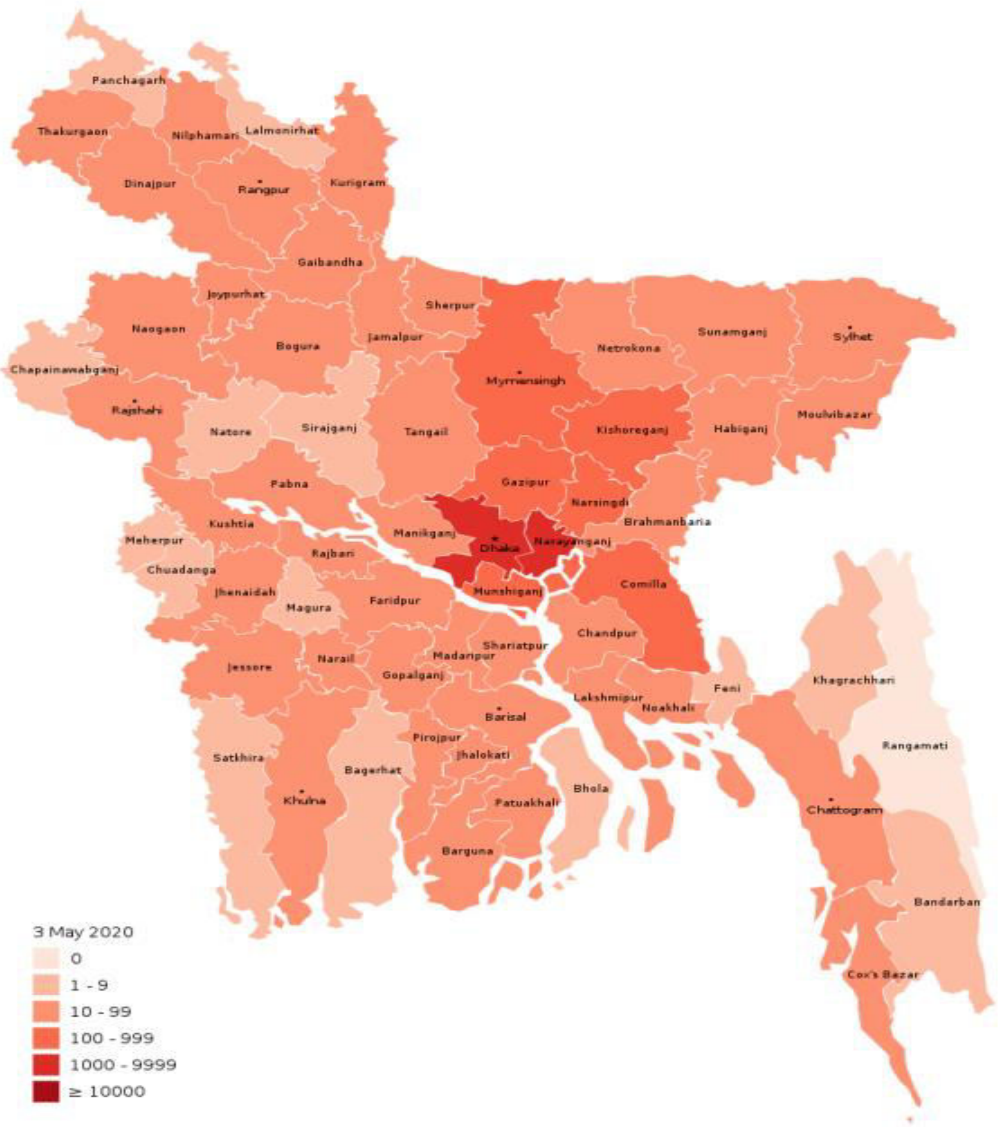

Fig. 1: District wise COVID-19 cases in Bangladesh (Symum *et al*., 2021; Mohiuddin, A. K, 2020).

METHODOLOGY:

Architecture

During the COVID-19 Outbreak, the number of new cases is increasing according to the increase in the test cases. To know how rapidly increase the virus and predicted the new cases if we increase the test cases. For this, we collect data set from online and applied machine learning-based model. In this section we broadly describe data collection, data preprocessing, machine learning algorithms, data predicting, and data visualization.

Data collection

In machine learning, machines learn from previous data so datasets are the most important things in machine learning algorithms. To find more informative and suitable feature data we collected this our-world-in data-covid19-dataset dataset (Our World in Data - COVID-19). From this dataset, we separate Bangladesh data which contains 67 columns and 941 rows of data.

Data Preprocessing

Firstly, check the null value and fill the null value with the mean value of this column. Using Extra Tree Classifier algorithm for feature selection. Our machine-learning algorithm separates the suitable column into dependent (new cases) and independent (new test cases) variables. After finding the feature spilt the data into train and test data using the test size of 30%.

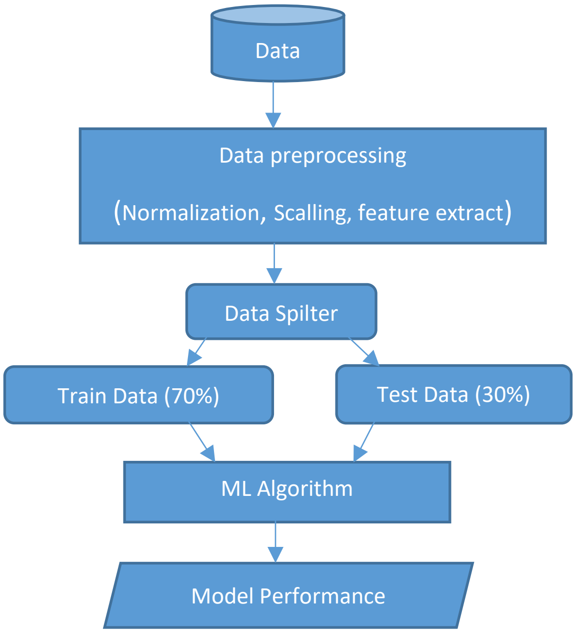

Fig. 2: Structure of preprocessing and fit of a machine learning algorithm.

Polynomial Regression (PR) algorithms

When the range of independent data fluctuates a lot and data patterns cannot be predicted in Linear Regression then a Polynomial algorithm is used. In this model independent variable is an nth-degree polynomial of the dependent variable. Polynomial algorithms are used to know how to spread the disease. Here below is the equation of polynomial regression

$$y = b_0 + b_1x_1 + b_2x_1^2 + b_3x_1^3 + \cdots + b_nx_1^n \dots\dots\dots\dots..(i)$$

Here,

x_i^j represent the input terms

y represents the predicted value

b_0 represents the intercept value

b_1-b_n represent the coefficient value

Data Prediction

For prediction value using a polynomial equation first evaluate the intercept and coefficient and update them. Here below is an example of polynomial regression (Finding coefficients of a polynomial.

We are looking for a third-degree polynomial,

$$P(x) = a_1 + a_2 * x + a_3 * x^2 + a_4 * x^3 \dots\dots\dots\dots..(ii)$$

Where, a_1, a_2, a_3, and a_4 are unknown and y = P(x).

We have this table of five values of x and y,

x:	-2	-1	1	3	6
y:	-5.8	.9	1.1	-12.3	4

We have the following five equations in four unknown.

$$1 * a_1 + -2 * a_2 + (-2)^2 * a_3 + (-2)^3 * a_4 = -5.8 \dots\dots(iii)$$
$$1 * a_1 + (-1) * a_2 + (-1)^2 * a_3 + (-1)^3 * a_4 = 0.9 \dots\dots(iv)$$
$$1 * a_1 + 1 * a_2 + 1^2 * a_3 + 1^4 * a_4 = 1.1 \dots\dots\dots\dots..(v)$$
$$1 * a_1 + 3 * a_2 + 3^2 * a_3 + 3^3 * a_4 = -12.3 \dots\dots\dots(vi)$$
$$1 * a_1 + 6 * a_2 + 6^2 * a_3 + 6^3 * a_4 = 4 \dots\dots\dots\dots\dots.(vii)$$

In matrix notation, this looks as follows

$$[\mathrm{C}] * [\mathrm{X}] = [\mathrm{D}] \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.(\mathrm{viii})$$

Now multiply both sides (on the left) by $([C]^T [C])^{-1}$:

$$([C]^T[C])^{-1}([C]^T[C])[X] = ([C]^T[C])^{-1}[C]^T[D] \dots..(ix)$$

This yield,

$$[X] = ([C]^T[C])^{-1}[C]^T[D] \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.(x)$$

So the best-fitting polynomial of degree three is

$y = P(x) = 3.523884615 – 1.797352564x – 2.474461538x^2 + .4640833333x^3$

Data Visualization and Discussion

Graphical representation of data is data visualization. Here the below graph **Fig. 3** red line represents the predicted value and the blue point is the actual value. We see that when we increase the test cases value then the value of the new cases is also increasing. The increasing number of new cases means the disease is rapidly spreading out in Bangladesh.

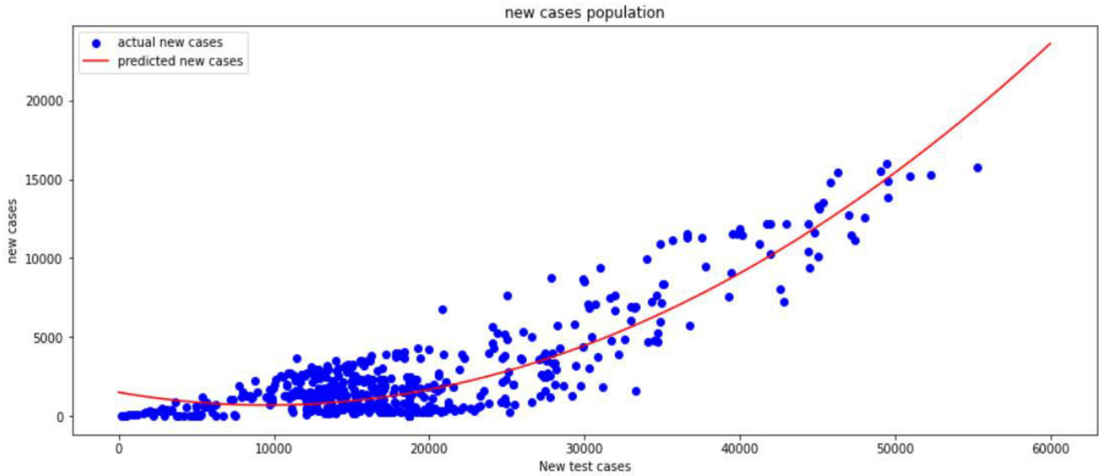

Fig. 3: Predicting new cases according to test cases using polynomial regression algorithm.

There are another graph **Fig. 4** number of new cases according to per day. In the below graph maximum number of cases are found in Bangladesh in August 2021 nearly 16000 cases per day.

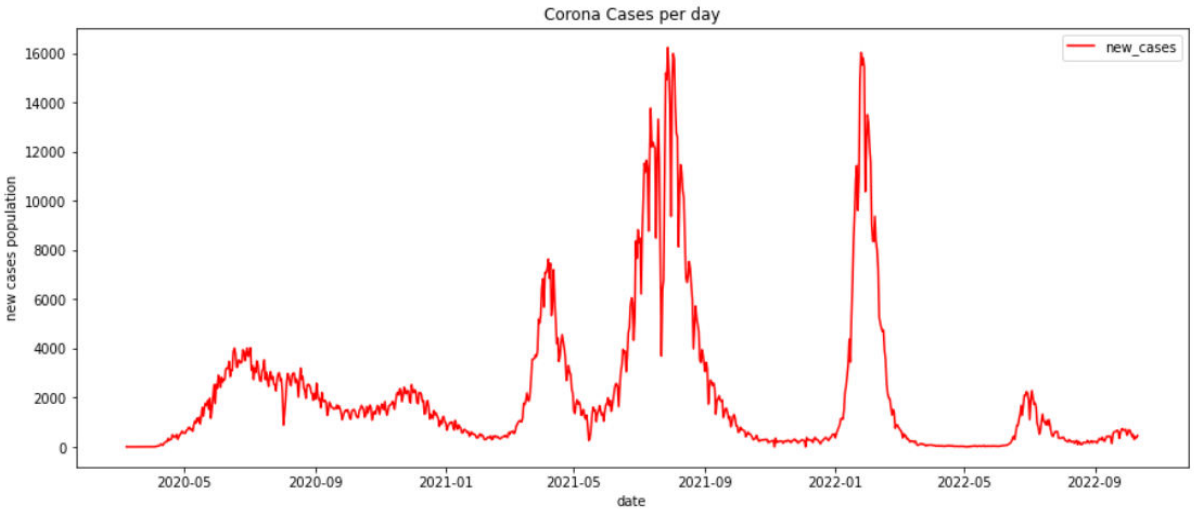

Fig. 4: Graph presenting the corona cases per day in Bangladesh from the dataset (World in Data - COVID-19).

Performance

Table 1: Performance of Polynomial Regression Algorithm.

Root mean square error	Accuracy
1319.78	0.8007

Finding Correlation of new death

After the invention of COVID-19 in Bangladesh, the number of death increases day by day. It's more important things to know why the death occurred. To find the reason for death and what's are the correlator for this we used a statistical test that's Chi-Square test. For this Test, we separate those column data (new cases, new death, medium age, aged 65 older, aged 70 older, cardiovascular death rates, and diabetes prevalence) from the dataset (Our World in Data - COVID-19. In this section, we describe the Chi-square test; calculate the p-value, and data visualization using a correlation matrix.

Chi-square test

The Chi-square test is a statistical test that is used to find out the difference between the observed and the expected data we can also use this test to find the correlation between categorical variables in our data.

The formula for chi-square is

$$(X^2) = \sum (O_i - E_i)^{2/E_i} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . . (xi)$$

Here,
Oi=Observed-value
Ei = Expected valueH0 (Null Hypothesis) = The 2 variables to be compared are independent.
H1 (Alternate Hypothesis) = The 2 variables are dependent.
Now, if the p-value obtained after conducting the test is less than 0.05 we reject the Null hypothesis and accept the Alternate hypothesis and if the p-value is greater than 0.05 we accept the Null hypothesis and reject the Alternate hypothesis (Shinde, 2021).

P value

The P-value is known as the probability value. The P-value is used as an alternative to the rejection point to provide the least significance at which the null hypothesis would be rejected. If the P-value is small, then there is stronger evidence in favor of the alternative hypothesis.

The formula for the calculation of the P-value is

$$z = \frac{\hat{p} - p0}{\sqrt{\frac{po(1-p0)}{n}}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (xii)$$

Where,
p^=Sample Proportion
P0 = assumed population proportion in the null hypothesis
n= sample size

Data visualization and Discussion

In the below heat map **Fig. 5**, we observe that there is a correlation between new cases and new death. The color box represents their relationship and the white place represents there is no relation between x label attributes and y label attributes. The value of 1 represents 100% related and 0.78 means 78% related that's means new death is 78% related to new cases according to the dataset (Our World in Data - COVID-19).

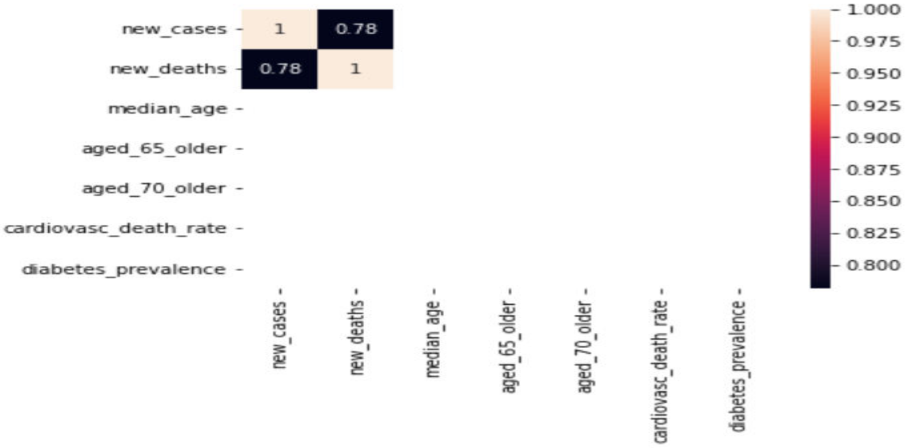

Fig. 5: Finding the relation between attributes using the chi-square test from the dataset (World in Data - COVID-19).

Predicting the new Death

In this COVID-19 situation, Bangladesh has lost a total of 29,388 people that are heartrending for our nation. We have also lost many reputable persons whose achievements are latch on in our country. To prevent this type of humdrum death we need more awareness for this we applied a machine learning-based algorithm that predicted new death according to total cases, new cases, and positive rates. In this section we describe data collection, data preprocessing, and machine learning algorithms, data prediction, and data visualization.

Data collection

We collect data from the our-world-in-data-covid19-dataset dataset (Our World in Data - COVID-19). From this dataset, separate Bangladesh data contain 67 columns and 941 rows of data. For our machine learning algorithms, we separate the suitable column into dependent (new death) and independent (total cases, new cases, positive rate) variables.

Data Preprocessing

Check the null value in this dataset and fill the null value with the mean value of this column. Then applied a feature selection algorithm (Extra Tree Classifier) for finding suitable columns for our model. We normalized the variable and spilt it into train and test data using a test size of 30%.

Decision Tree Regression

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes (Decision Tree Regression).

Data Prediction

How many people are dead in this situation to evaluate the applied Decision Tree Regression algorithm using predict method.

Data visualization and Discussion

Here below graph **Fig. 6**, the yellow point represents the actual death and the green point represents the predicted death. From the graph, we can see that the actual death and predicted death are almost the same and we also see that new death increases when new cases increase. Using this machine learning model we predict new death in a particular number of cases helps us to increase people's awareness and immediately take action plans.

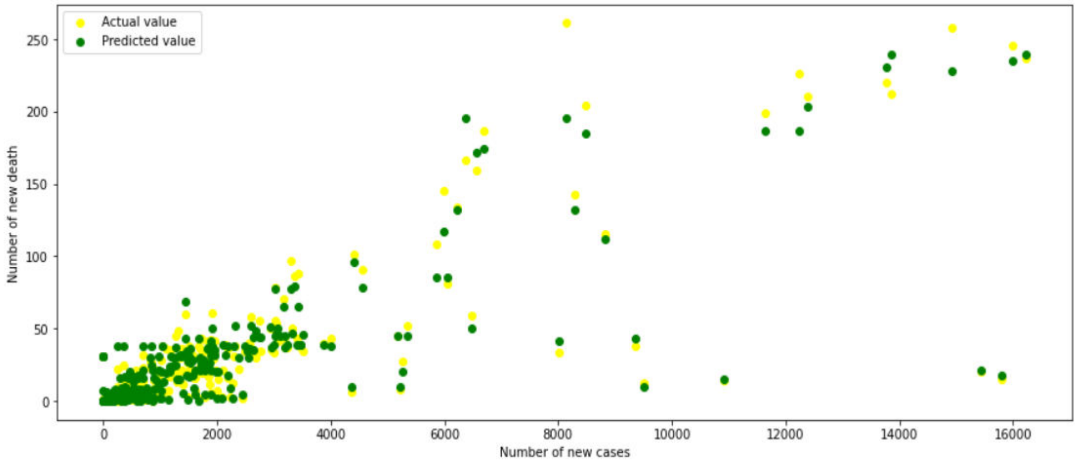

Fig. 6: Predicting the new death (green) according to the new cases from the dataset (World in Data- COVID-19).

Here another graph **Fig. 7** that's represents new death per day. From the graph, we can see that the maximum number of death that occurred in Bangladesh in August 2021 was nearly 250 deaths per day.

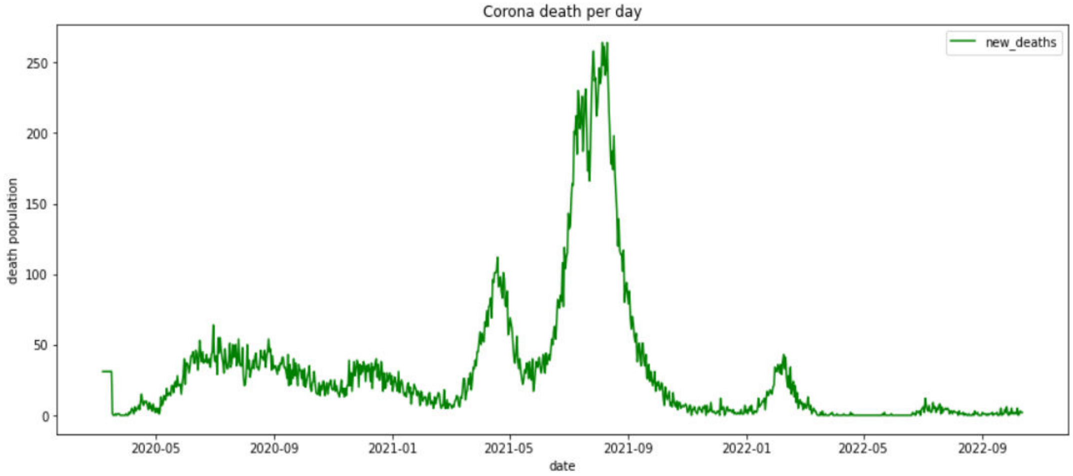

Fig. 7: Graphs are represented the number of death per day in Bangladesh from the dataset (World in Data - COVID-19).

This graph **Fig. 8** represents the corona-positive rate per day. From this graph, we can see that the maximum number of positive rates occurred in Bangladesh in August 2021 nearly 0.30% per day.

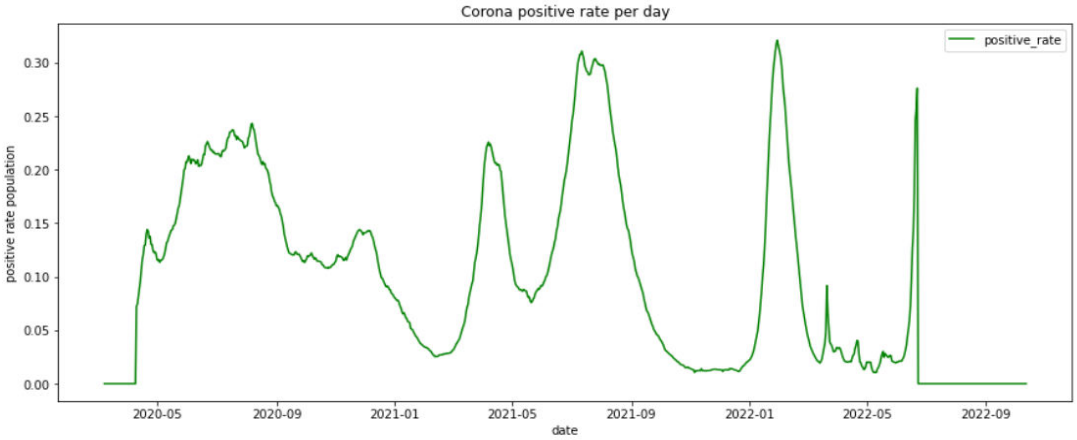

Fig. 8: Graph is represented rate of corona positive per day in Bangladesh from the dataset (World in Data - COVID-19).

Performance

Table 2: Performance of Decision Tree Regression.

Root Mean Square Error	Accuracy
9.11	0.9628

Covid-19 Vaccination

During the outbreak of COVID-19 there badly needed safe and effective vaccines are available that provide strong protection against serious illness, hospitalize-tion, and death from COVID-19. Here is a world map **Fig. 9** that provides how to spread out the vaccination in the world. The black color represents the bad position and the yellow color represents the good position.

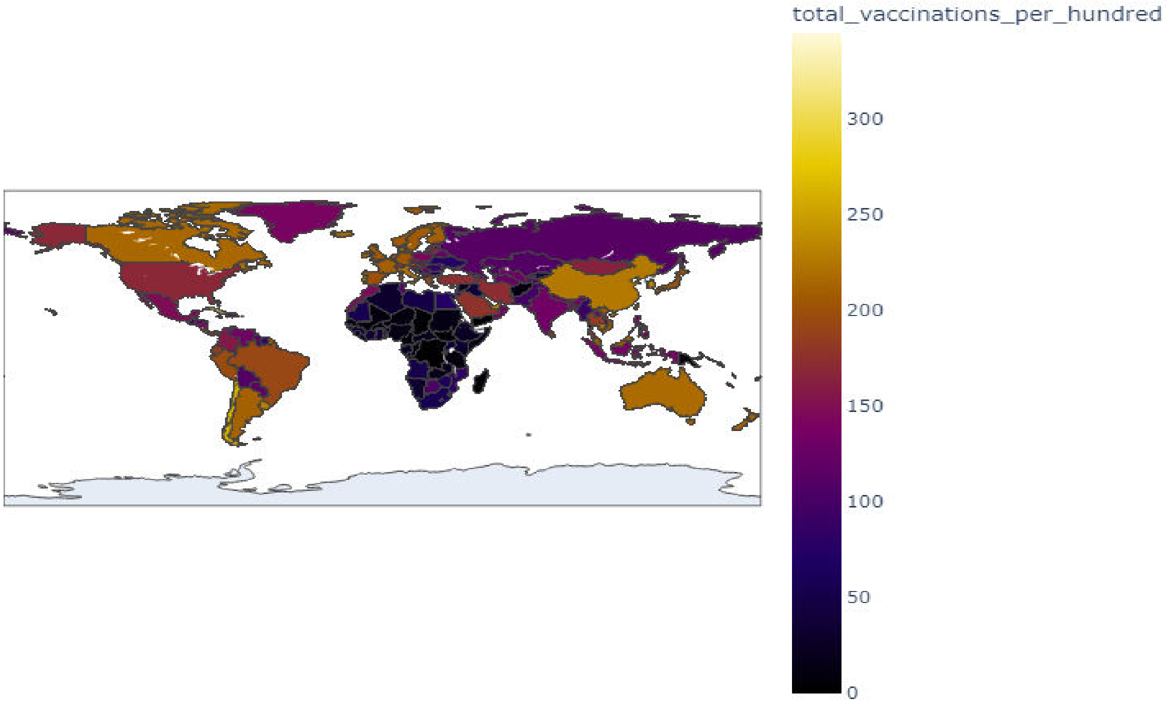

Fig. 9: Representing the vaccinations condition overall in the world from the dataset (COVID-19 World Vaccination Progress).

Bangladesh began the administration of COVID-19 vaccines on 27 January 2021 while vaccination started on 7 February.

Predicting people fully vaccinated in Bangladesh

Data Collection

For our machine learning algorithm, we collect the Covid-world-vaccination-progress dataset (COVID-19 World Vaccination Progress (n.d.)). In this dataset, we separate the Bangladesh data which contains 15 columns with 428-row data. Divided the columns into dependent (people fully vaccinated) and independent (people vaccinated, daily vaccinations, people fully vaccinated per hundred) variables for our model.

Data Preprocessing

Check the null value and fill it with the mean value. We applied a feature selection algorithm (Extra Tree Classifier) for finding suitable columns for our model and normalized the value and spilt it into train and test data using a test size of 30%.

Linear Regression algorithm

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. In this section, we used a linear regression algorithm to predict how many need independent (people vaccinated, daily vaccinations, people fully vaccinated per hundred) variables to cover fully vaccinated in Bangladesh.

Hypothesis function for Linear Regression

$$y = \theta_1 + \theta_2 * x \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (xiii)$$

Cost Function (J)

$$(minimize)\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (xiv)$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (xv)$$

Data Visualization and Discussion

From the linear regression graph **Fig. 10** we see that the curve linearly increases that means if we increase the independent (people vaccinated, daily vaccinations, people fully vaccinated per hundred) variable the predicted value (people fully vaccinated) also increases linearly that represent a red line graph. Using this machine learning model we predict how many daily vaccinations and vaccinated per hundred to complete fully vaccinated people in Bangladesh.

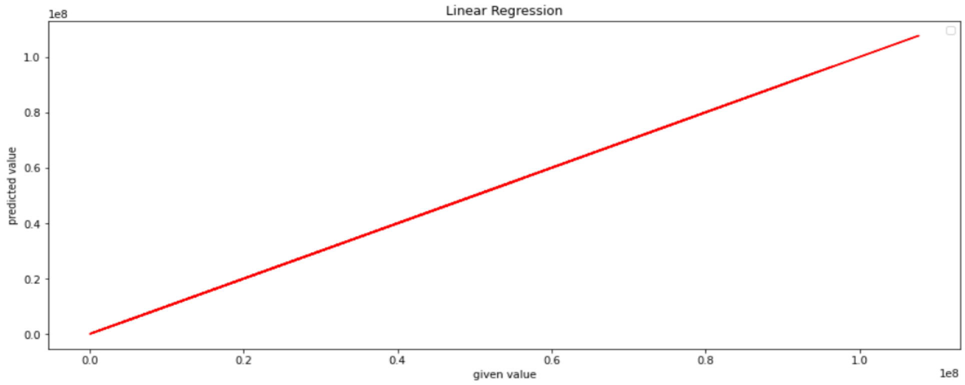

Fig. 10: Predicted people fully vaccinated use linear regression algorithms from the dataset (COVID-19 World Vaccination Progress (n.d.)).

Performance

Table 3: Performance of Linear Regression algorithm.

Root Mean Square Error	Accuracy
21843.67	0.9999

Vaccinations Condition in Bangladesh

Here is a graph **Fig. 11** of daily vaccinations in Bangladesh. We can see that number of vaccination is increasing day by day after 2022. The Red line represents the total number of vaccinations per hundred, the Green line represents the people vaccinated per hundred and the Blue line represents the people fully vaccinated per hundred.

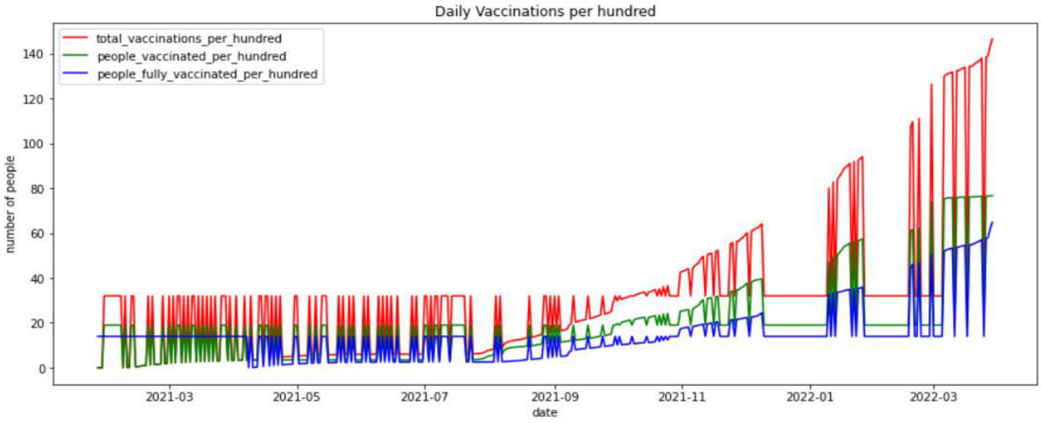

Fig. 11: Graph shows that the vaccinations condition in Bangladesh from the dataset (COVID-19 World Vaccination Progress (n.d.)).

We have found the top 10 countries in **Fig. 12** that used the maximum number of vaccines. In this place, Bangladesh has positioned at the 7th number. We used around 0.25X1009 vaccines.

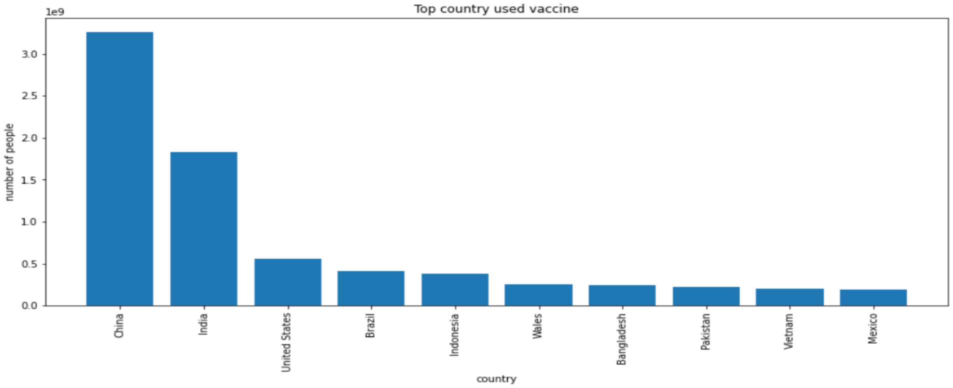

Fig. 12: Ranking the top 10 country according to the total vaccination in the world from the dataset (COVID-19 World Vaccination Progress (n.d.)).

Bangladesh is a high-density populated country there in around 164.7 million people here. To fully people vaccinated we need more vaccinations here below is a pie chart **Fig. 13** that provides how many people are vaccinated according to the dataset. Today we have covered 48% of vaccinated.

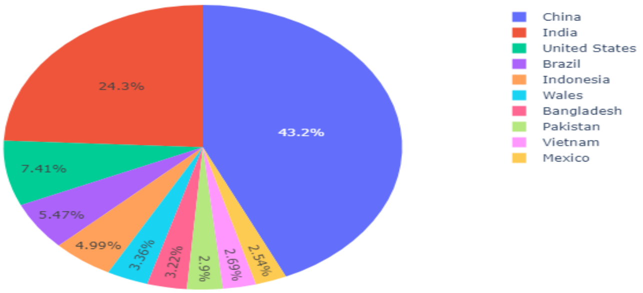

Fig. 13: Pie plots are indicated how many people are vaccinated from the dataset (COVID-19 World Vaccination Progress (n.d.)).

CONCLUSION:

COVID-19 has already taken over 6.6 million lives. The global pandemic caused by this virus has influenced almost everyone's lives in the year 2020 and the primary target everyone is combating is to dispose of. In this current situation, Bangladesh is one of the most important phases in the battle against COVID-19 because of the increasing number of new cases and new death. In this study machine learning provide us with very accurate information on how to pandemic situation will affect us. Using machine learning we can predict new cases that inform us how many viruses are spread out and we can predict new death that increases our people's awareness about the COVID-19 gruesomeness and find the correlation of this virus. We also predict how many vaccinations are needed and what the vaccination condition in Bangladesh is. This valuable information helps us our government or authority will know what decision to make a month in advance and what kinds of action plans are needed. This information helps us to protect against immature death and economical losses.

ACKNOWLEGDEMENT:

We are grateful to Pabna University of Science and Technology (PUST) for the support of the research

CONFLICTS OF INTEREST:

The authors state that there is no potential conflict of interest in publishing this research article.

REFERENCES:

1) COVID-19 World Vaccination Progress. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/gpreda/covid-world-vaccination-progress
2) DecisionTreeRegression.(n.d.). https://www.saedsayad.com/decision_tree_reg.htm#:~:text=Decision%20tree%20builds%20regression%20or
3) Finding coefficients of a polynomial. (n.d.). Sofia.nmsu.edu.
4) Mohiuddin, A. K. (2020). Covid-19 Situation in Bangladesh. https://doi.org/10.20944/preprints202005.0094.v1
5) OurWorldinData-COVID-19.(n.d.). https://www.kaggle.com/datasets/caesarmario/our-world-in-data-covid19-dataset
6) Satu, M. S., Howlader, K. C., & Islam, S. M. S. (2020). Machine Learning-Based Approaches for Forecasting COVID-19 Cases in Bangladesh. https://ssrn.com/abstract=3614675
7) Symum H, Hiya HK, & Ali KM. (2021). Impact of COVID-19 pandemic on population-level interest in skincare: evidence from a Google trends. *Eur. J. Med. Health Sci.*, 3(6), 147-160. https://doi.org/10.34104/ejmhs.021.01470160
8) Shinde, Y. (2021). Chi-Square Test-Use, Implementation and Visualization. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/decoding-the-chi-square-test%E2%80%8A-%E-2%-80%8Ause-along-with-implementation-and-visualization
9) Susanto F, Arefin MS, and Badiruzzaman M. (2022). Reflective critical thinking on education and teaching during the COVID-19 pandemic. *Int. J. Agric. Vet. Sci.*, **4**(2), 26-38. https://doi.org/10.34104/ijavs.022.026038
10) Wikipedia. (2021). COVID-19. Wikipedia. https://en.wikipedia.org/wiki/COVID-19

Citation: Hossain MA, Hossen E, and Asraful M. (2022). Study and prediction of covid-19 cases and vaccinations using machine learning in Bangladesh. *Aust. J. Eng. Innov. Technol.*, **4**(6), 130-139. https://doi.org/10.34104/ajeit.022.01300139