CSE303: Statistics for Data Science
[Fall 2021]

Project Report

Group No. – 09 (Section 1)

Submitted by:

Student ID	Student Name	Contribution	Signature
2019-1-60-167	Plabon Banik	28%	
2019-1-60-168	Amit Mojumder	27%	
2019-1-60-170	Tamzid Inam Tasin	25%	
2018-2-60-015	Shajnin Alam Sarnali	20%	

- ## Introduction

We have worked on different types of model such as Linear Regression, Polynomial regression**.**

- ## Data Pre-processing

There is Null values. No dropping duplicates. There are no missing values. We use Label Encoder for encoding technique

- ## Dataset Characteristics and Exploratory Data Analysis

For the Covid dataset:

Day: It shows the dates of the coronavirus confirmed cases were reported. The virus was confirmed to have spread to Bangladesh in **April 2020.** Since then, the pandemic has spread day by day over the whole nation and the number of affected people has been increasing. During this period, 2 new death was reported across the country which takes the number of deaths from COVID-19.

Lab Test: The number of tests that detects the presence of COVID19. In the beginning approximately about 111 tests were conducted in Bangladesh. In April 2020, there were a total of 434 individuals tested. The number of lab test started increasing day by day.

Confirmed cases: The number of cases that came positive after having a confirmatory viral test of COVID 19. In April 2020, there were a total of 434 individuals tested among whom there were 9 individuals detected as positive.

Death Cases: The number of death cases on the according dates. During this time the number of death cases also started increasing. On 4th April 2020, 2 death cases has been reported. The first two months the number of death cases was less than 10. But on May 2020, the number of death cases started growing rapidly.

For the Covid_first_dose:

Number of Vaccination (1st dose): Bangladesh began the administration of COVID-19 vaccines on 27 January 2021. Initially it focused on a pilot program of 500 health workers. Then the mass vaccination started on 7 February 2021. It was planned that 6 million doses would be administered in the first month, and a further 5 million the following month.

For the Covid_second_dose:

Number of Vaccination (2nd dose): The **second shot as close to the recommended 3-week or 4-week interval as possible.** On April 2020, those who completed their first dose vaccine around this recommended time period they started to take their second shot of vaccination.

- ## Machine Learning Models

Linear Regression:

Linear regression is a statistical technique. By fitting a linear equation to observed data, linear regression seeks to model the connection between two variables. One variable is regarded as an explanatory variable, while the other is regarded as a dependent variable. For example, could

wish to use a linear regression model to match people's weights to their heights. It is also known as dependent and independent variables. When there is only one explanatory variable, a basic linear regression is employed; when there are more than one, a multiple linear regression is utilized. This is not the case with multivariate linear regression, which predicts numerous linked components rather than a single scalar variable.

Polynomial regression:

Polynomial regression is a type of statistical regression analysis in which an nth degree polynomial is used to describe the relationship between the independent variable X and the dependent variable Y.

Lasso Regression:

Lasso regression is a type of regularization. For a more accurate forecast, it is preferred over regression approaches. Shrinkage is used in this model. Data values are shrunk towards a central point known as the mean in shrinkage. Simple, sparse models are encouraged by the lasso approach. This type of regression is ideal for models with a lot of multicollinearities or when you wish to automate elements of the model selection process, such as variable selection and parameter removal.

Ridge Regression:

In situations when linearly independent variables are heavily correlated, ridge regression is a method of calculating the coefficients of multiple-regression models. When there is a subset of true coefficients that are small or even zero, ridge regression performs well.

Elasticnet Regression:

The penalties from both the lasso and ridge approaches are used to regularize regression models in elastic net linear regression. The strategy combines the lasso and ridge regression methods by learning from their flaws to better statistical model regularization.

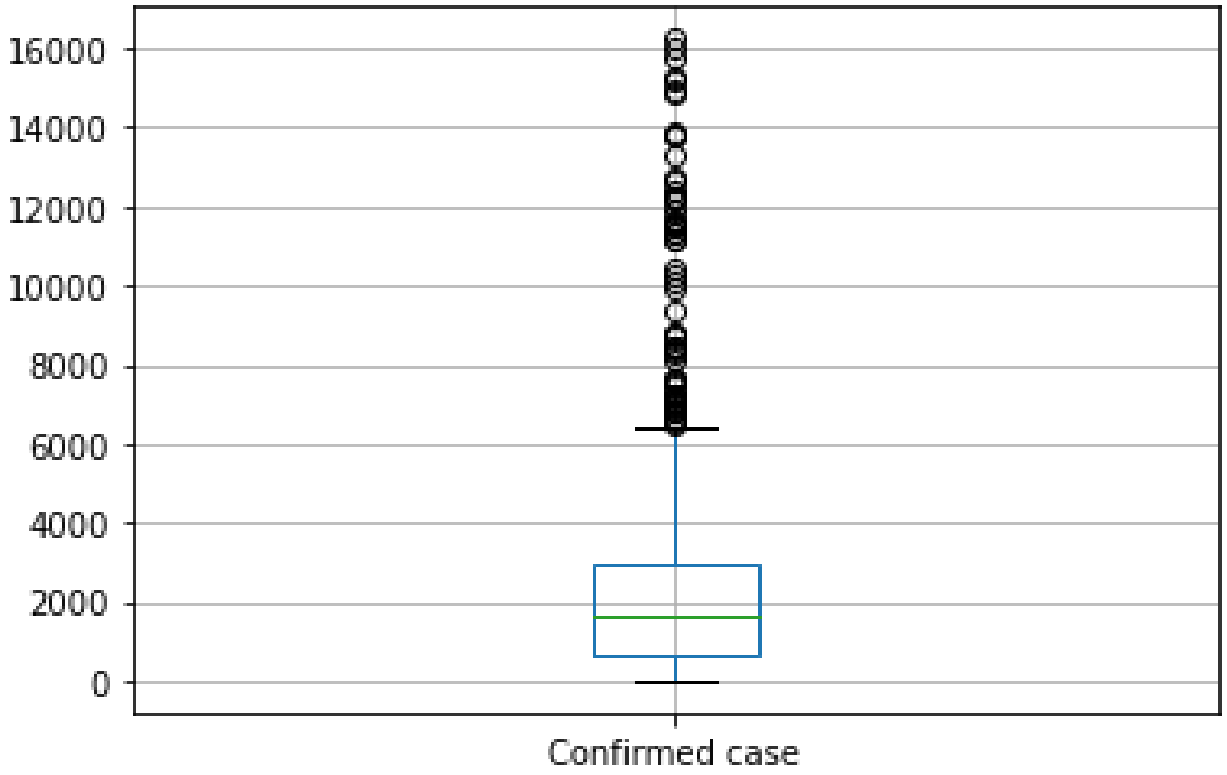
Logistic Regression:

Logistic regression is a statistical model that uses a logistic function to represent a binary dependent variable in its most basic form, though there are many more advanced variants. Logistic regression is a type of regression analysis that involves estimating the parameters of a logistic model.

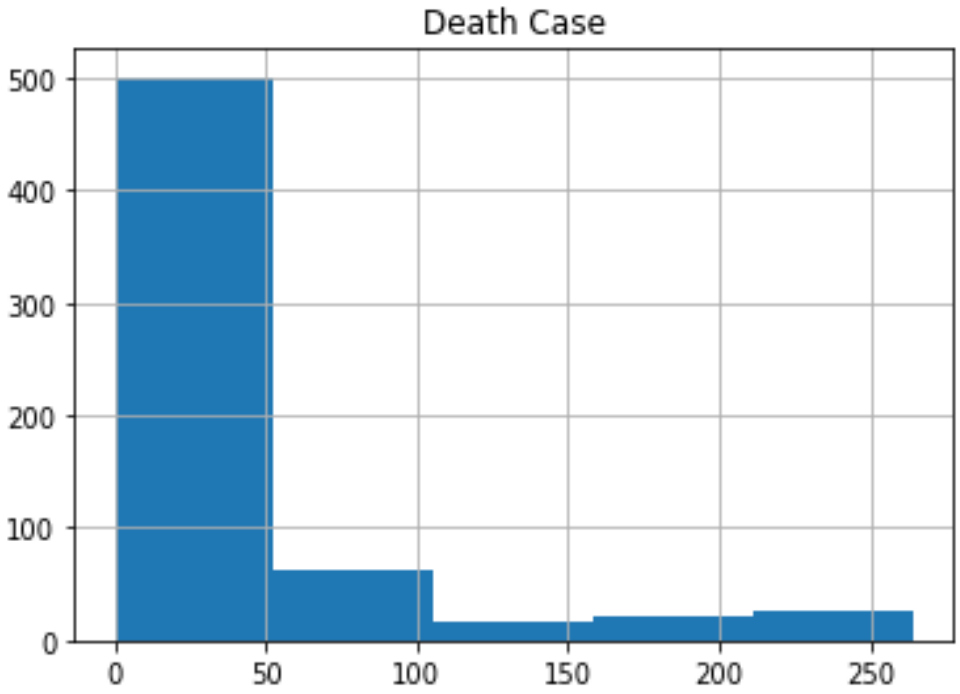
Linear SVC (Support Vector Classifier) :

A Linear SVC (Support Vector Classifier) is designed to fit to the data you provide and provide a "best fit" hyperplane that divides or categorizes your data. Following that, you may input some features to your classifier to check what the "predicted" class is after you've obtained the hyperplane.

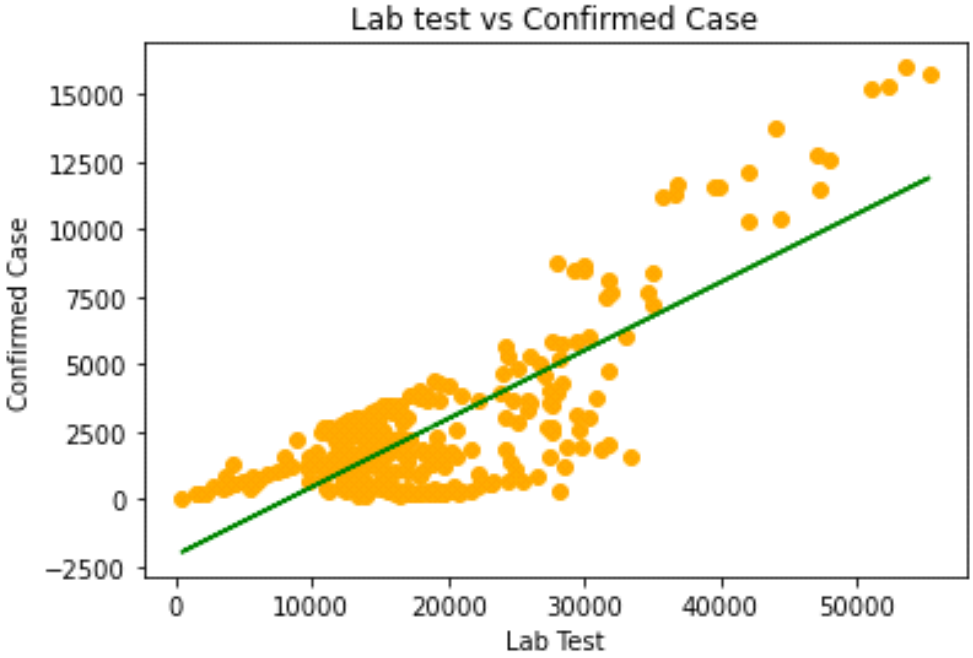
Box Plot of Confirmed Case:



Histogram of Death Case:

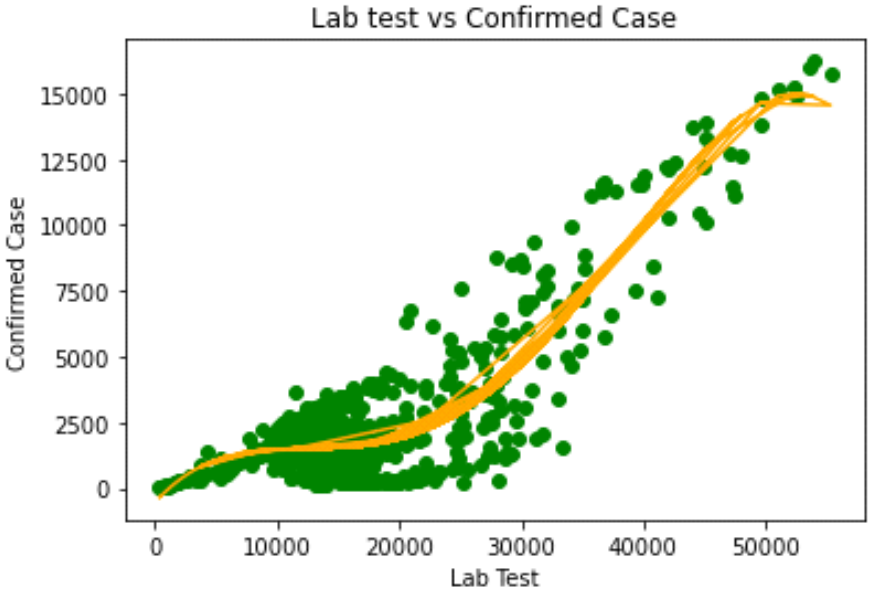


Linear Regression:



Mean Absolute Error: 2724.9002152932117
Mean Squared Error: 15684691.291847384
R-square: -0.8230095621703548

Polynomial Regression:



Mean Absolute Error: 1023.4790747473065
Mean Squared Error: 1801805.5947169072
R-square: 0.7907508397292874

Lasso Regression:

```
In [40]: runcell('Lasso regression', 'C:/Users/tamzi/OneDrive/Desktop/project/project.py')
Accuracy of test dataset 64.8163902925962 %
Accuracy of train dataset 64.13266046070633 %
```

Ridge Regression:

```
In [41]: runcell('Ridge regression', 'C:/Users/tamzi/OneDrive/Desktop/project/project.py')
Accuracy of test dataset 64.81639019282485 %
Accuracy of train dataset 64.13266046070648 %
```

Elasticnet Regression:

```
In [42]: runcell('lasticNet regression', 'C:/Users/tamzi/OneDrive/Desktop/project/project.py')
Accuracy of test dataset 64.8163902553213 %
Accuracy of train dataset 64.13266046070643 %
```

Logistic Regression:

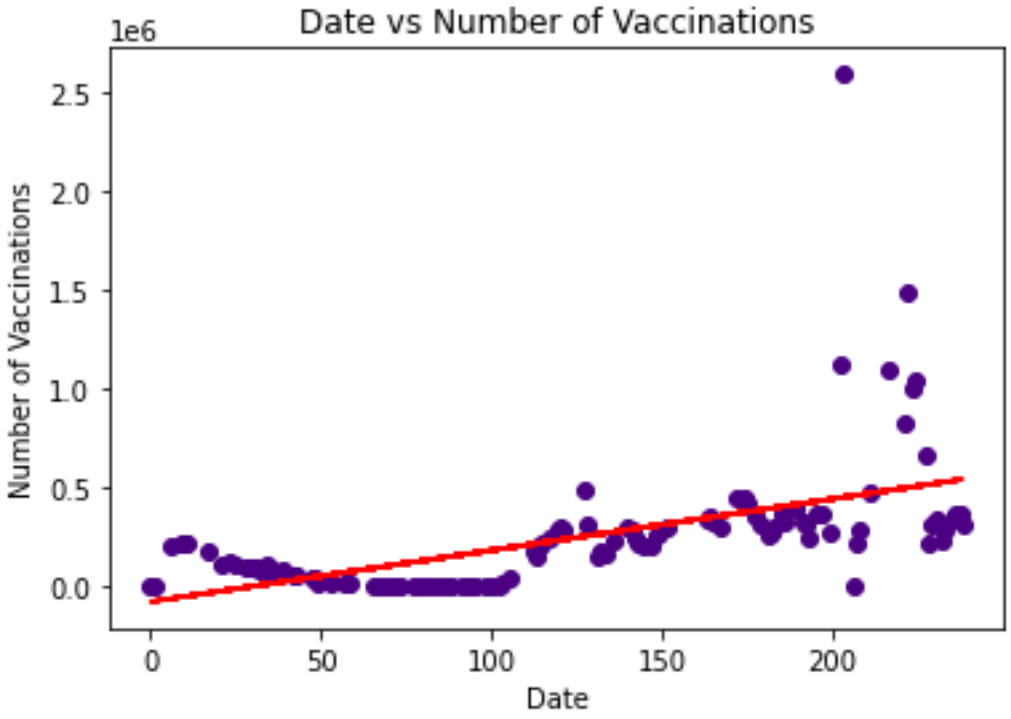
```
In [45]: runcell(19, 'C:/Users/tamzi/OneDrive/Desktop/project/project.py')
Accuracy rate with Logistic regression  3.827751196172249 %
C:\Users\tamzi\Anaconda3\lib\site-packages\sklearn\linear_model\_logistic.py:763: ConvergenceWarning:
lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
  n_iter_i = _check_optimize_result(
```

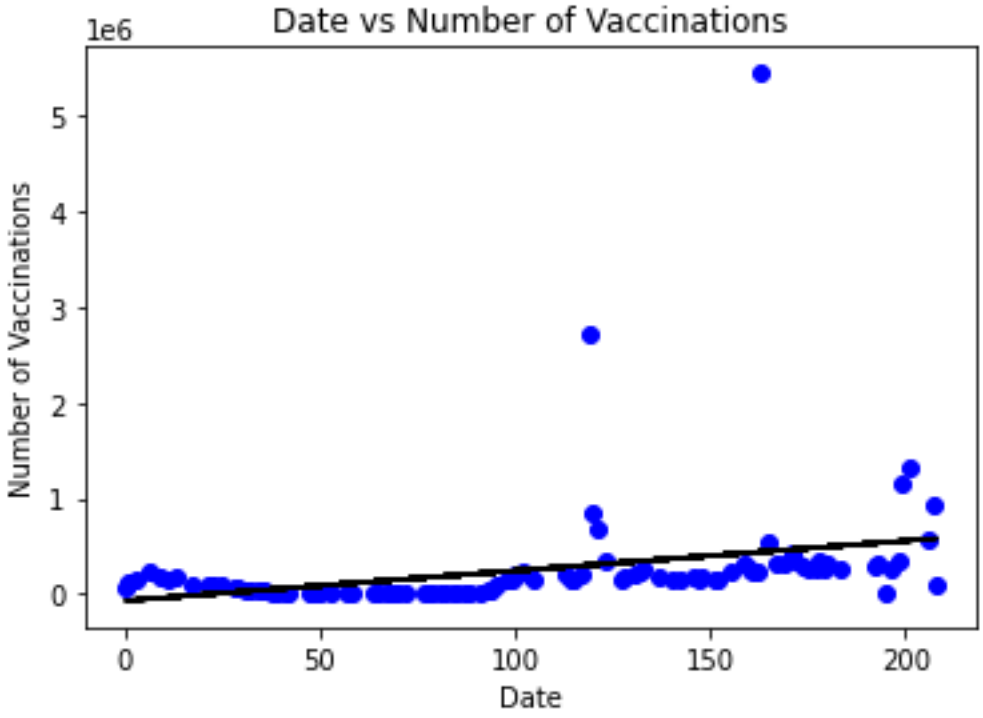
Linear SVC (Support Vector Classifier):

```
In [46]: runcell(20, 'C:/Users/tamzi/OneDrive/Desktop/project/project.py')
Accuracy rate with Support Vector Classifer  3.349282296650718 %
```

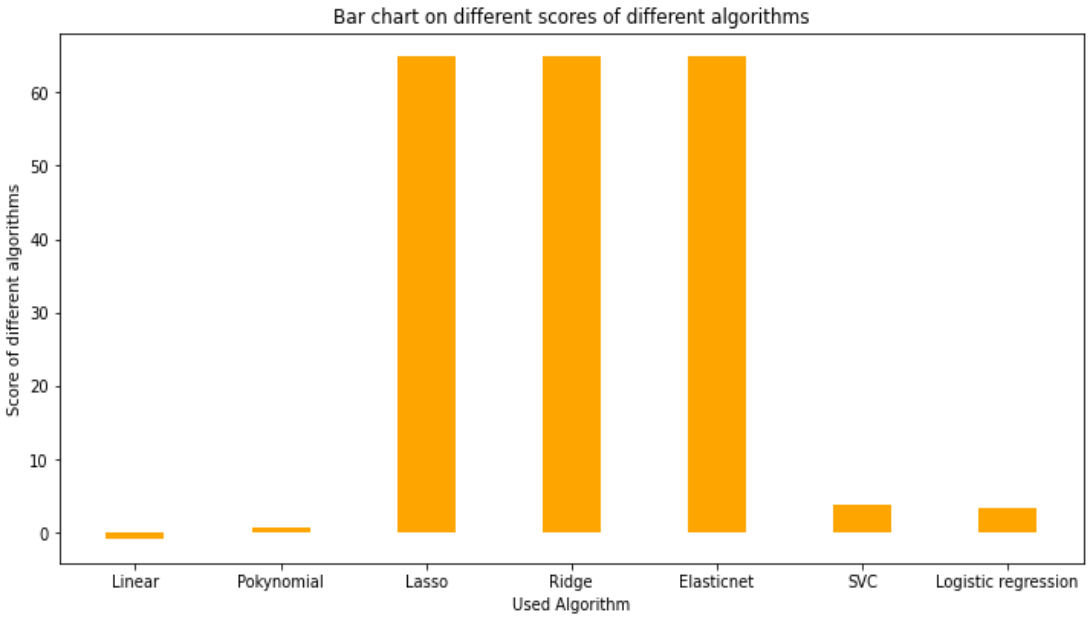
Scatter plot of Linear Regression model for "covid_first_dose" dataset:



Scatter plot of Linear Regression model for "covid_second_dose" dataset:



Performance Evaluation:



Here we notice that Lasso, Ridge and Elasticnet algorithms score are same. But SVC, Logistic, Polynomial and Linear algorithms scores are highly decreasing this graph we see that, Lasso, Ridge and Elasticnet have the highest score and polynomial have the lowest score.

Discussion:

From all of the models Lasso Regression, Ridge Regression and Elasticnet Regression performed better than other models. Because in Linear and Polynomial Regression the train accuracy was good enough, but test accuracy was poor. So, it overfitting the dataset. But in terms of Lasso Regression, Ridge Regression and Elasticnet Regression the test and train accuracy were not underfit or overfit. Test and train accuracy were 64.8163902925962% 64.13266046070635% respectively. SVC, Logistic Regression performed better than Linear and Polynomial but not much than Lasso, Ridge and Elasticnet. So, Lasso, Ridge and Elasticnet performed better than others.

Conclusion:

Firstly, we want to express our gratitude to you, sir, for providing us with this fantastic chance to work on this fantastic machine learning project. We were able to deploy these models and determine their correctness with your aid and some YouTube instructions. This investigation has broadened our understanding of building various types of models and predicting the accuracy level in relation to various types of data information's. We had a lot of fun assessing the accuracy level of this entire dataset. It

provides us with a It's a fantastic concept to forecast any form of pleasure or rating. We've learnt when and how to employ are several sorts of models We are really grateful to you for providing us with this chance. improve our understanding of various sorts of datasets