17BCE1290
Akshay Kumar

WEB MINING PROJECT

Introduction

The Novel Corona Virus also known as Covid-19 is the infectious disease caused by the most recently discovered corona virus. Covid-19 has caused the ongoing pandemic that has shocked the world with its ability to easily spread from one human to another through physical contact or through droplets produced while sneezing or coughing. What makes the Covid-19 virus even more dangerous is that its incubation period can last up to 15 days. The challenge in combating the covid-19 is that is ability to easily spread from one human to another even if the disease carrier is asymptomatic.

Till date, there is no existing vaccine nor effective treatment for combating the covid-19 virus. And only the biggest remedy to control the spread the virus is to wear PPEs such as masks, maintain personal hygiene such as wash hands with soap for 20s and sanitizing hands using alcohol-based hand sanitizers with 70% alcohol and to practice social distancing.

This report gives us the insights on the novel corona virus that has affect the state of Tamil Nadu in India. The report also gives a predictive analysis of the positive tested covis-19 cases and the deaths occurred due to it. The reports also forecast the covid-19 cases based on the prediction model. The report also gives a predictive analysis on the number of tests conducted. The report also briefly contains informative graphs on the trend in covis-19 cases and also contains the district wise covid-19 cases.

Dataset

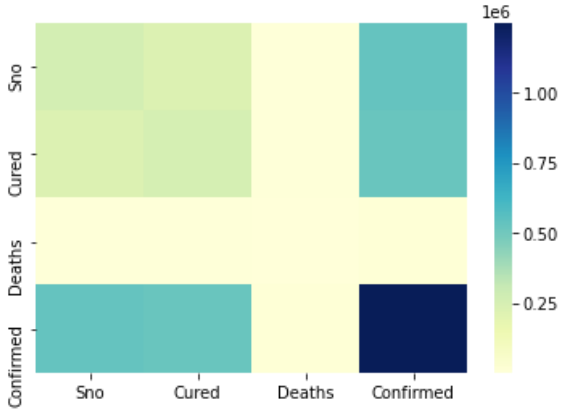
The dataset used for this project was taken from Kaggle, where State level data comes from Ministry of Health & Family Welfare & the Individual level data comes from covid19india. The dataset is a time series dataset where individual cases are recoded with details such as the date & time, state, district, the number of positive cases, deceased and the number of patients cured from the disease, the amount tests conducted throughout India for covis-19.

Preliminary Data Analysis

The dataset contains no duplicate values and does not contain any null values. Hence the data was already a cleaned dataset.

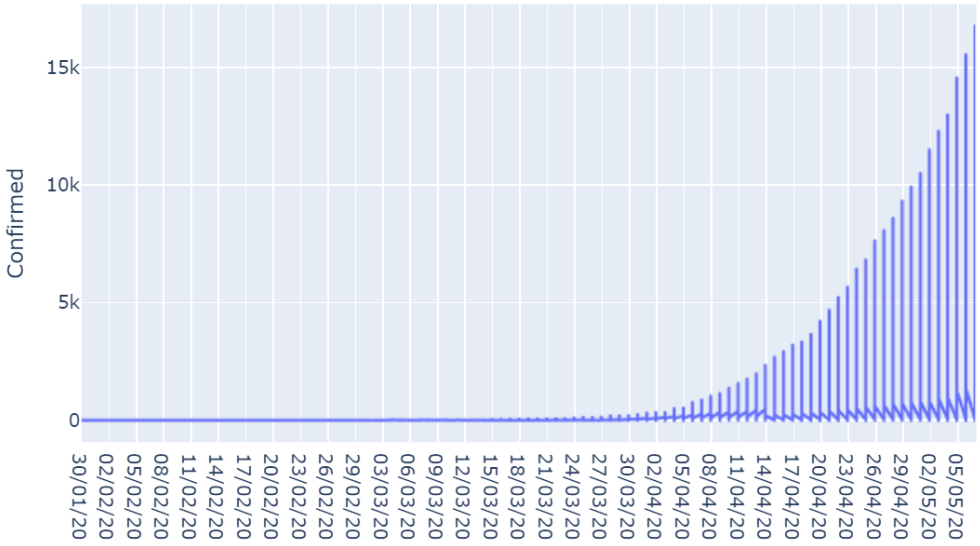
To check the relation between the attributes in the dataset, a correlation analysis was performed and plotted with the help of the correlation matrix that tells us the relation between the attributes in the dataset.

Visualization of correlation matrix

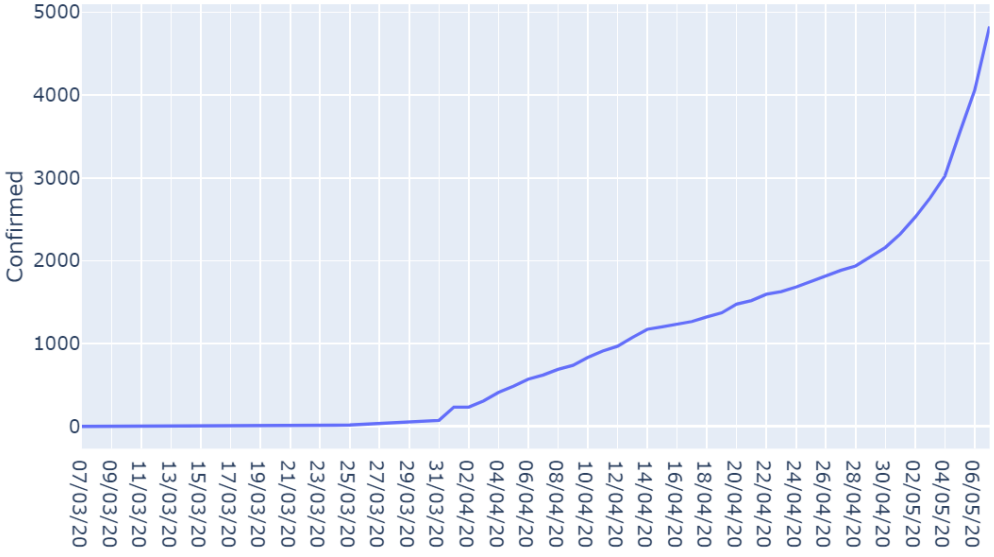


There were not much meaningful insights from the correlation analysis as the attributes are not much related.

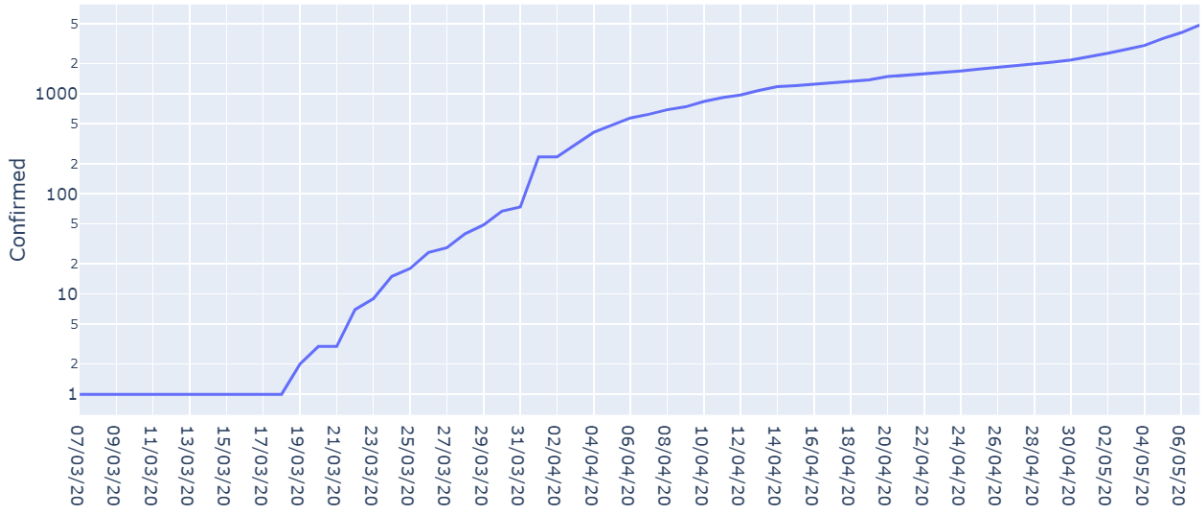
Graph plotted on the overall covid-19 cases throughout India.



Graph plotted on the overall covid-19 cases throughout the state of Tamil Nadu, India.

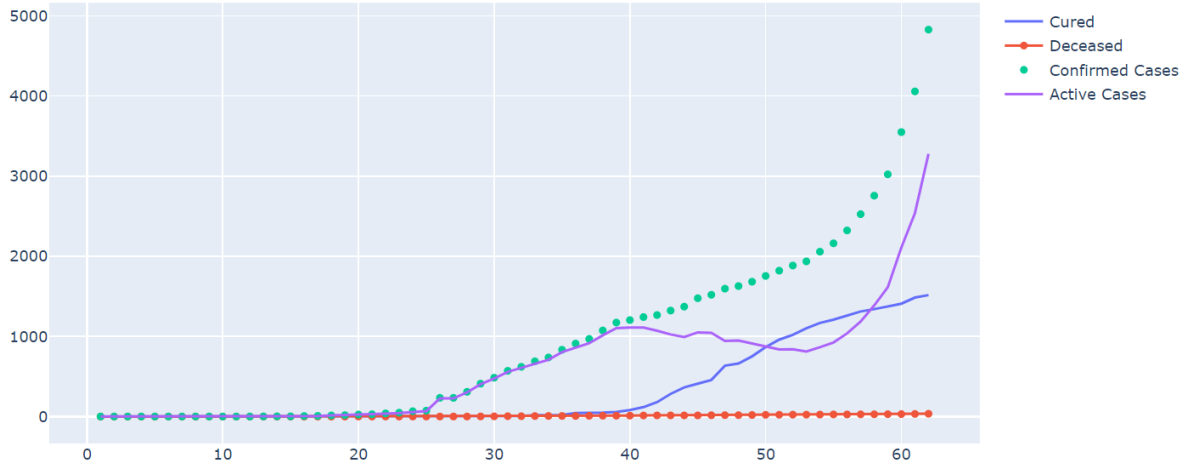


The same graph was plotted by using the logarithmic scale.



The reason why the logarithmic scale was used to plot the covid-19 cases is that it considers the multiplication factor in the covid-19 cases trend in Tamil Nadu. The graph above is not strictly linear indicating that the cases are increasing but not multiplying at an alarming rate. But from 4th to 6th May there is an increase on the graph, indicating a sudden spike in Covid-19 +ve cases.

Plotting the Cured, Deceased, Confirmed Covid-19 cases and Active Covid-19 cases in Tamil Nadu.

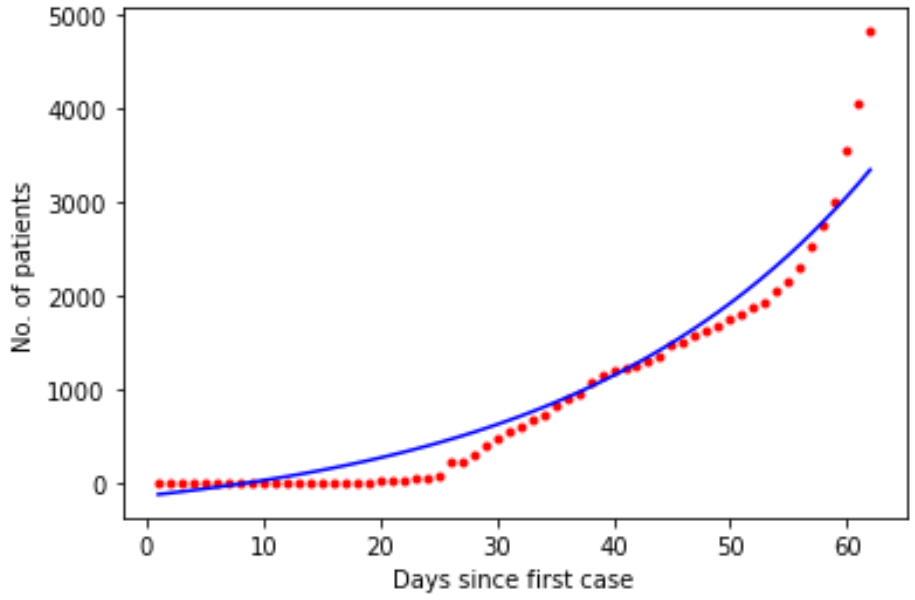


This gives an overall picture of the active covid-19 cases, the patients who are cured from covid-19 , the number of cases reported daily and the deaths due to the virus. This graph shows that it takes about 20 to 30 days for patients to get cured from the disease while the death rates are extremely low in Tamil Nadu.

Predictive Analysis

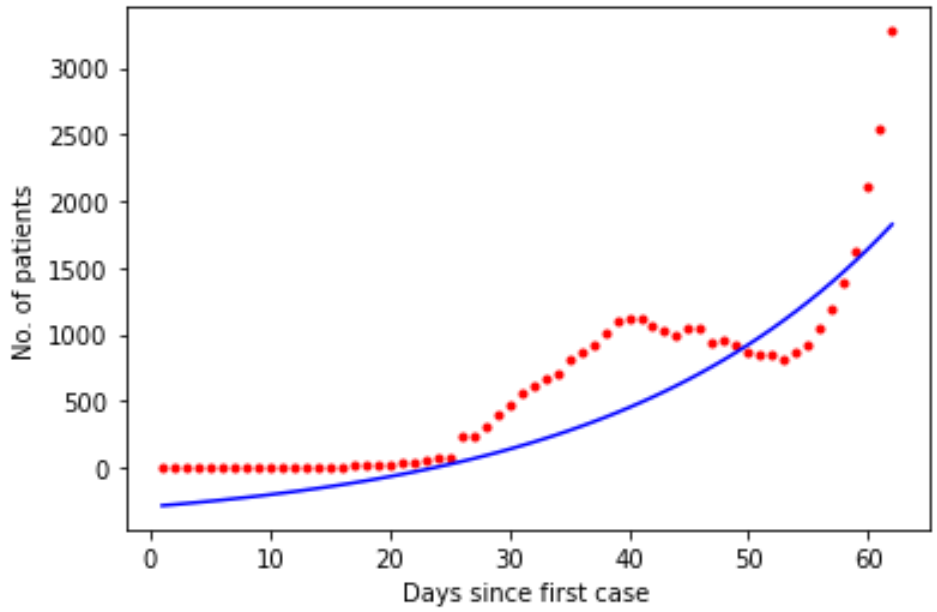
Using machine learning algorithm logistic regression by using gradient descent for each of the predictive analysis performed. While the curve-fit () function is used to fit the predicted points to form a line passing through the points.

Prediction of Confirmed Covid-19 Cases in Tamil Nadu



Red dotted line denotes the actual data
Blue Line denotes the predicted value

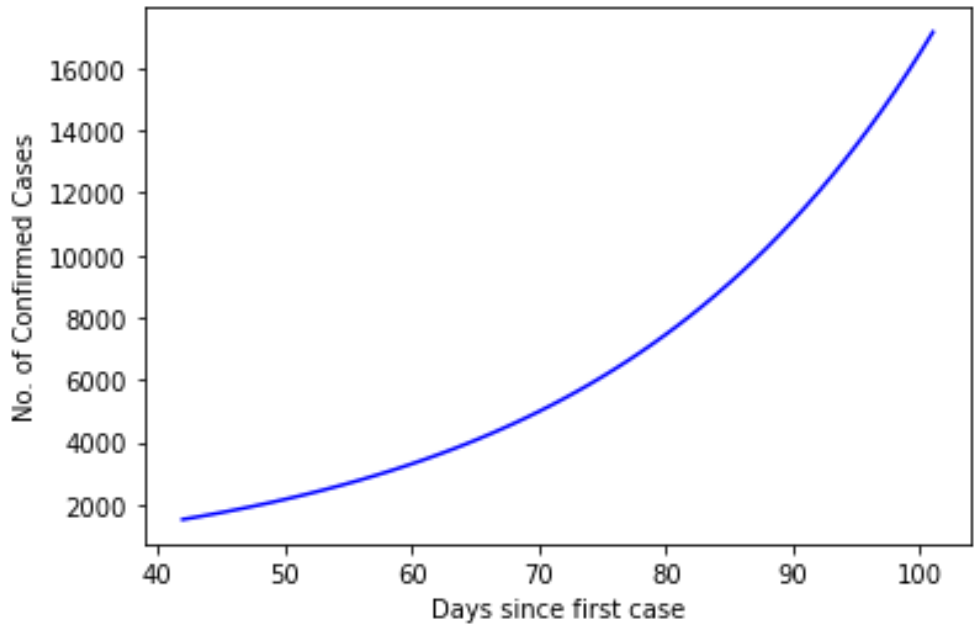
Prediction of the Deceased in Tamil Nadu due to Covis-19



Red dotted line denotes the actual data
Blue Line denotes the predicted value

The predicted of deaths is low due to the data being ununiform (no pattern) and high variation in the death data. Therefore, forecasting will give incorrect result.

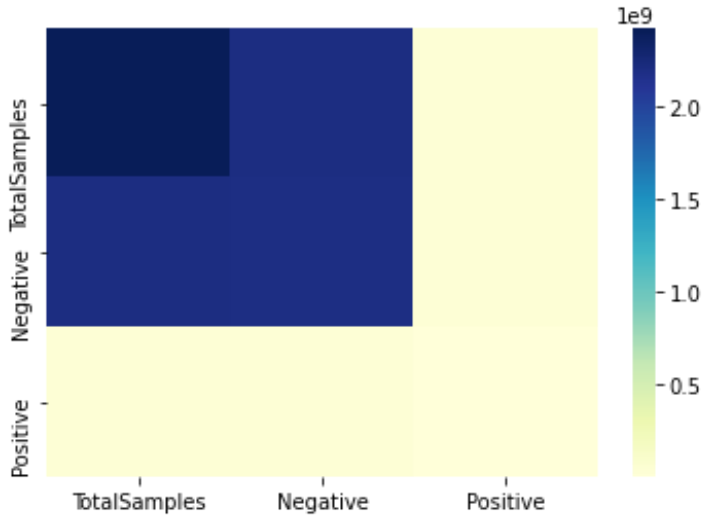
Forecasting Covid-19 Cases in Tamil Nadu



The forecasting shows that the number of confirmed cases in Tamil Nadu will increase to about 7,750 Covid-19 cases by June 2020.

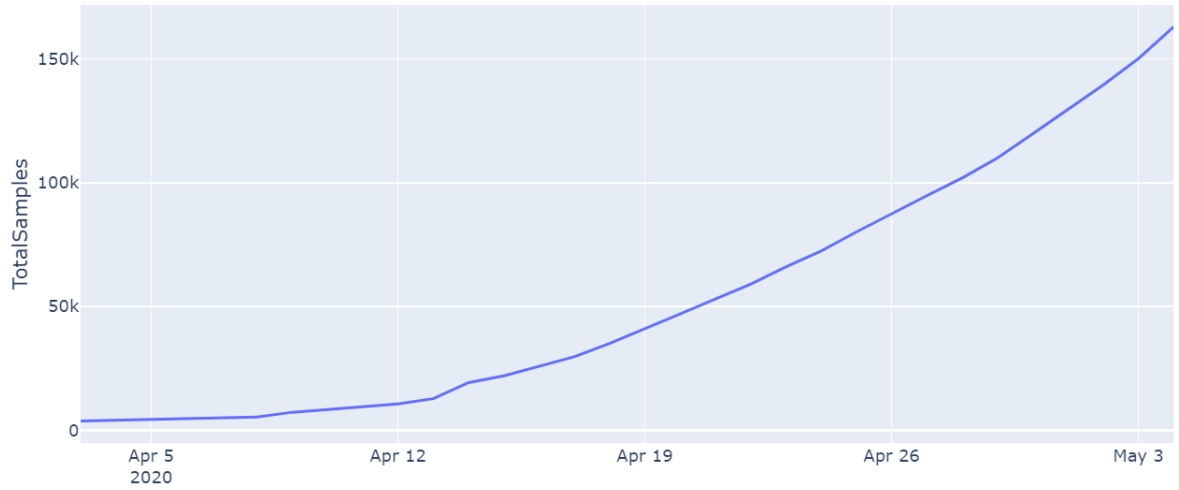
<u>Tamil Nadu Testing analysis</u>

Finding relations between the attributes.

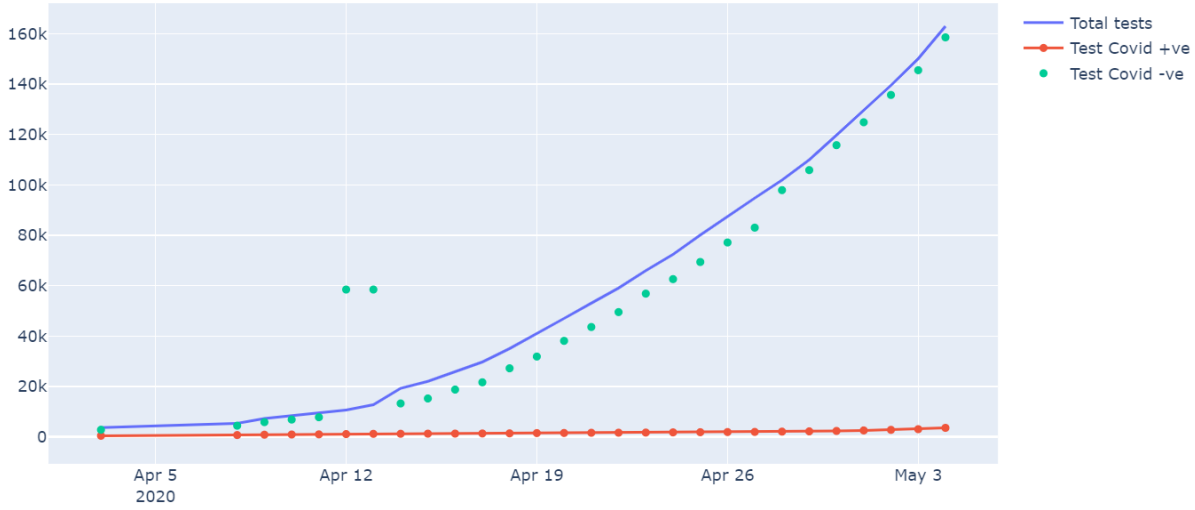


The correlation matrix indicates that most of tests performed in Tamil Nadu turns out as negative for Covid-19.

Plotting the cumulative Covis-19 tests performed in Tamil Nadu till 4th May 2020.



Plotting the graph for covid-19 on the total number of tests performed, tests that turned out to be positive & negative



The graph shows us that a greater number of samples, more the chances of the test being negative.

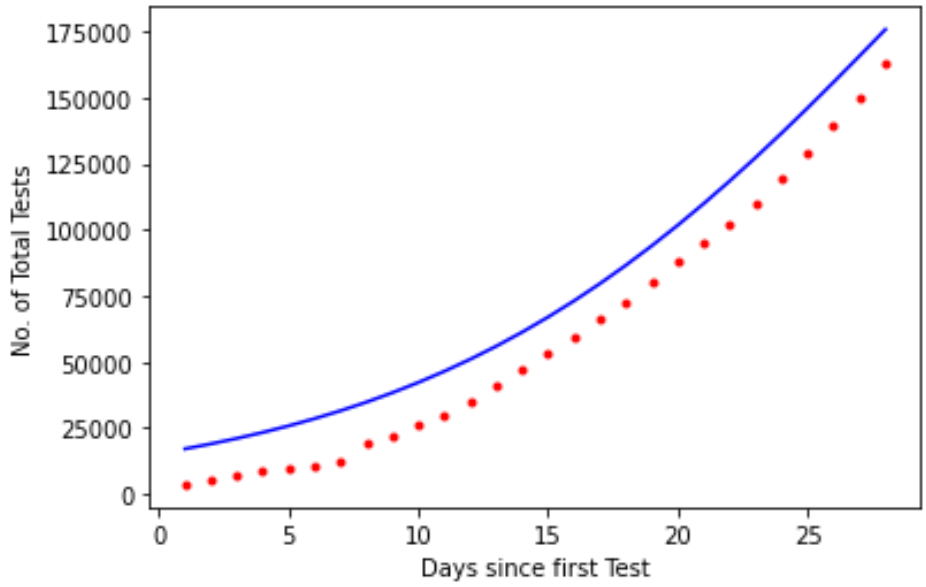
Testing Rate

Population of Tamil Nadu (2011 census)	72147030
Total testing by May 4th	162970

The Test Rate per Million in Tamil Nadu was found to be 2258.86

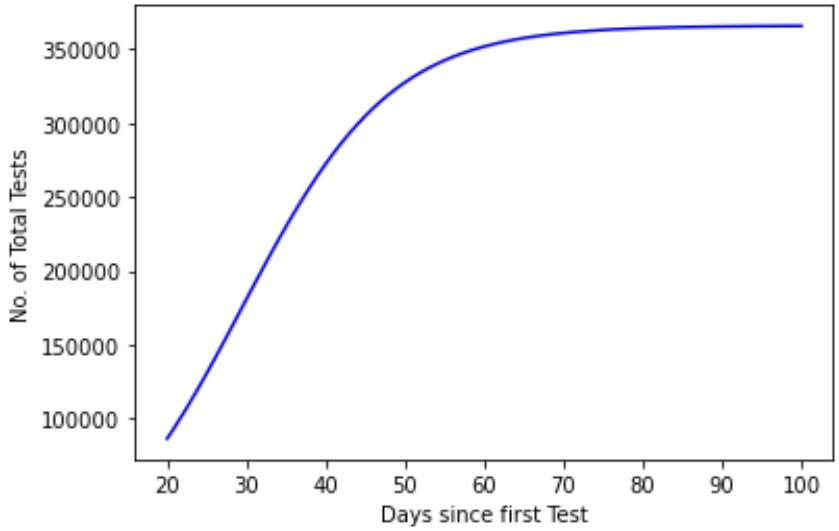
By comparing the number of positively number of cases to total number of tests performed.
It is found that only 2.18 % of the tests are turned out to be positive.

Tamil Nadu Testing Prediction using ML algorithm



Red dotted line denotes the actual data
Blue Line denotes the predicted value

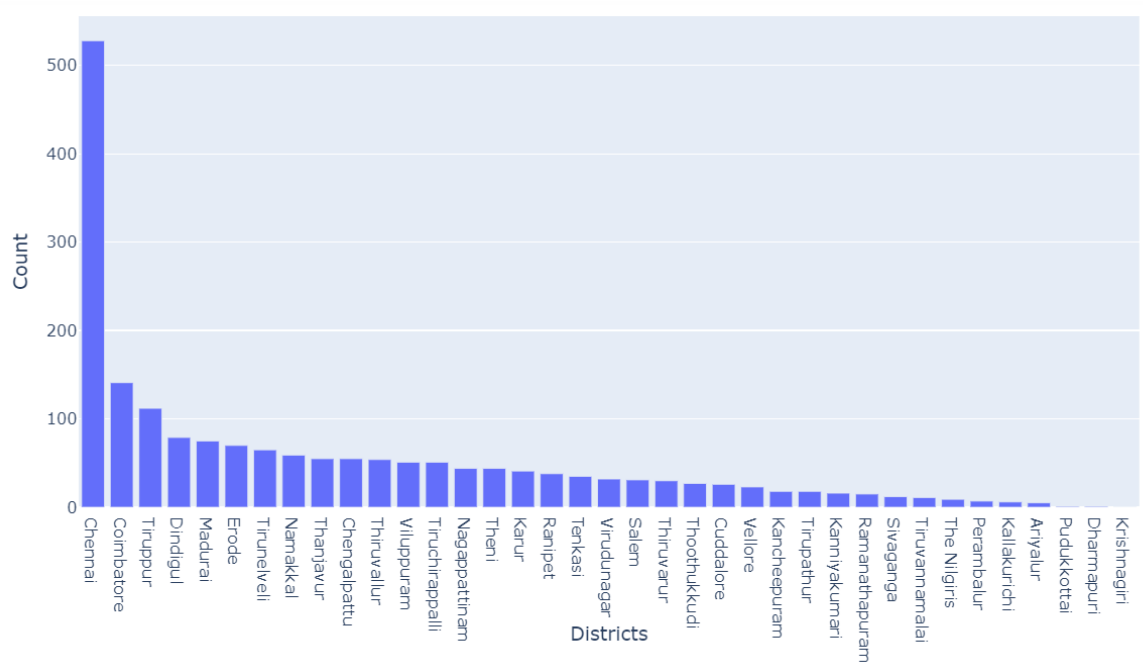
Tamil Nadu Testing Forecasting



From the above forecasting, a total number of 3 lakh tests are expected to be performed by the end of May-2020.

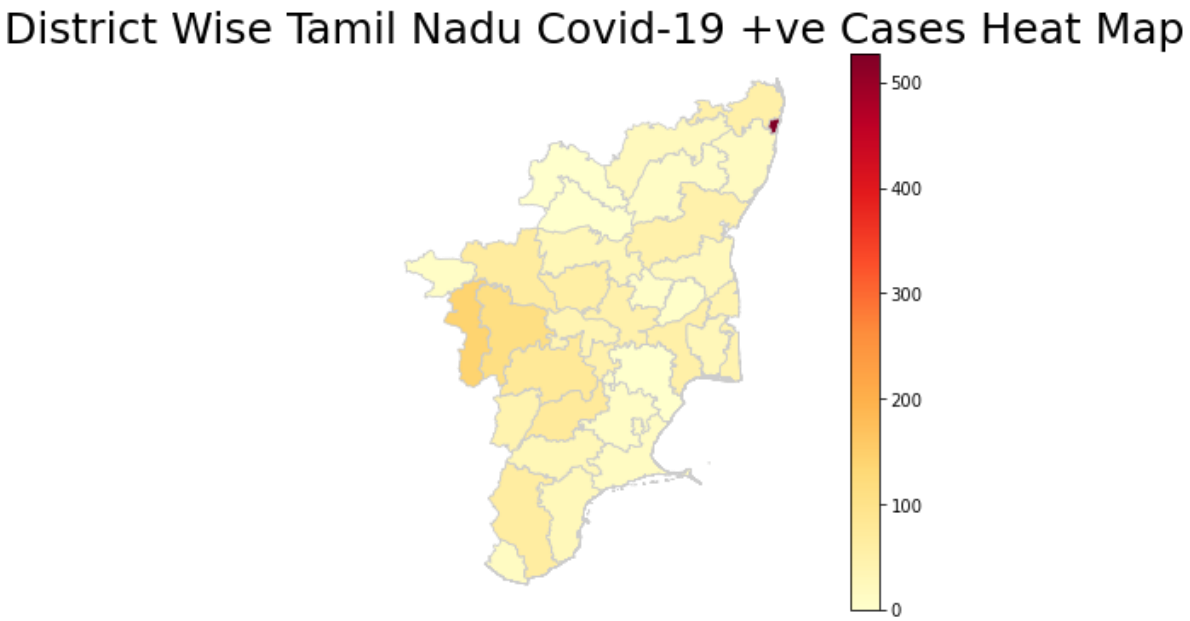
The number of testing is set to flatten after reaching 3.5 lakhs tests.

District wise Covid-19 cases



This shows that Chennai having the number of Covid-19 cases followed Coimbatore and Tirupur districts.

Plotting the district wise Covid-19 cases in map.



Algorithm

Simple Gradient Descent Algorithm used for the prediction and forecasting of the data.

```
theta=0.4
n=len(f2)
s2=np.empty(n)
def yp2(f,tau,theta,gam,b):
    z=tau/(1+np.exp(-gam*(f-theta)))+b
    return(z)
p0=[max(y21), np.median(f21),1,min(y21)]
print(yp2(f21,max(y21), np.median(f21),1,min(y21)))
```

Conclusion

The report tells us that the total number of Covid-19 cases are still on the rise in Tamil Nadu and is predicted to reach 16,000 by July 2020 based on the current trend predict by the ML Model.
Due to the wide scale and increased amount of testing done in Tamil Nadu, the number of cases is set to increase for the coming days even though the positive test rate in below 2.18 %. And by the demographics, Chennai and Coimbatore are set to retain the highest number of confirmed Covid-19 cases in Tamil Nadu.