SVKM’s NMIMS
Mukesh Patel School of Technology Management &
Engineering

A REPORT ON PREDICTION OF COVID-19

By

ABHI GOYAL

A REPORT ON

PREDICTION OF COVID-19

BY

ABHI GOYAL

A Report submitted in partial fulfillment of the requirements of 5 years Integrated MBA (Tech) Program of Mukesh Patel School

of Technology Management & Engineering, NMIMS

Completion Certificate

This is to certify that Mr. Abhi Goyal Roll No.- N-224
Has completed training & project as a part of Technical Internship in our company as mentioned below and the Report is also submitted.

(i) Project Title: Prediction of Covid-19
(ii) Date of Joining:
(iii) Date of Completion:
In partial fulfillment of XII Semester Technical Internship for MBA (Tech) program of Mukesh Patel School of Technology Management & Engineering, Narsee Monjee Institute of Management Studies (NMIMS)(Deemed-to-be University), Mumbai.

....
Industry Mentor / Faculty Mentor
Date:
Place:
Company /Institution Seal:

ACKNOWLEDGEMENT

I would like to thank our college SVKM's NMIMS Mukesh Patel School of Technology Management and Engineering, Dr. **R. S. Gaud**, Director (SVKM's NMIMS, Shirpur Campus), **Dr. Nikhilesh Kumar Sharma**, Director (Engineering Program, Shirpur Campus),**Dr. Nitin Choubey**, Associate Dean (MPSTME, Shirpur Campus), **Dr. Chandak**, Dean(MPSTME, Shirpur Campus) and **Dr. Rambola**,HOD(MPSTME, ShirpurCampus,Computer Science) for introducing me to this Technical Internship Programme, which exposed me to the corporate world besides giving me a platform to have hands-on experience to learn new technologies and apply them.

My faculty mentor Professor **BHUSAN INJE** has given me constant support and guidance besides being a source of inspiration through her words of encouragement. Her continuous supervision and much needed guidance has really helped me in conceptualizing my project and executing the same.

Last but not the least, I place a deep sense of gratitude towards my family members and my friends who have been a constant source of inspiration during the internship.

I perceive this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives.

TABLE OF CONTENT

List of Figures

Abstract

Covid-19 is one of the most widespread illnesses declared as a pandemic which has stopped the world and has put billions of people under lockdown. The corona virus or SARS-CoV-2 causes this illness. There is much uncertainty regarding the spread of this virus and the period for which people will have to be under lockdown.

India first reported a COVID-19 case in a student who returned from Wuhan, China on January 30, 2020. For future references, 2020 is the default year for all the dates, unless mentioned otherwise. Since then, there has been a gradual rise in the number of infections. However, countries like India are at a greater risk because of a very large population density.

Formal, quantitative approaches are now widely used to make predictions about the likelihood of an infectious disease outbreak, how the disease will spread, and how to control it. Several well-established methodologies are available, including risk factor analysis, risk modeling and dynamic modeling. Even so, predictive modeling is very much the ‘art of the possible’, which tends to drive research effort towards some areas and away from others which may be at least as important.

Whatever methodology is used, a key challenge in making any kind of prediction is to establish the extent to which the past is likely to be an accurate guide to the future.Through this project, we will predict the possible number of cases in the near future in India.

Introduction

India first reported a COVID-19 case in a student who returned from Wuhan, China on January 30, 2020. For future references, 2020 is the default year for all the dates, unless mentioned otherwise. Since then, there has been a gradual rise in the number of infections with 1,251 cases on March 30, among which there are 1,117 active cases, 102 recovered cases and 32 deaths. In response, India has implemented international travel bans and a strict lockdown. However, countries like India are at a greater risk because of a very large population density, limited infrastructure, and healthcare systems to cater to very large demands.

A large proportion of the young population, and possible immunity due to BCG vaccinations, may favor India. Most of these studies are preliminary and correlation-based.

I hereby present project to predict the spread of COVID-19 in India. Most of pandemics follow an exponential curve during the initial spread and eventually flatten out. The current models are thus based on an exponential fit and logistic regression for short term and long-term predictions, respectively. Further, Susceptible-Infectious-Recovered (SIR) compartment model is used to include considerations for susceptible, Infectious, and recovered or deceased individuals.

OBJECTIVES

In the past decades, there is an increasing interest in predicting the numbers of corona patients and recovered cases in order to do the valid research for the beneficial of the society. The objective of the proposed work is to study and implement the supervised learning’s to predict the corona patients in India.

TECHNICAL OBJECTIVE

The technical objectives will be implemented in Python. The system must be able to access a list of dates with the details of corona patient numbers. It must calculate the estimated number of patients infected and recovered from 28 January, 2020. It must also provide an instantaneous visualization of corona in India.

- To add to the academic understanding of corona prediction, this project will focus exclusively on predicting the daily numbers of patients. The project will also analyze the accuracies of these predictions.

Understanding Project

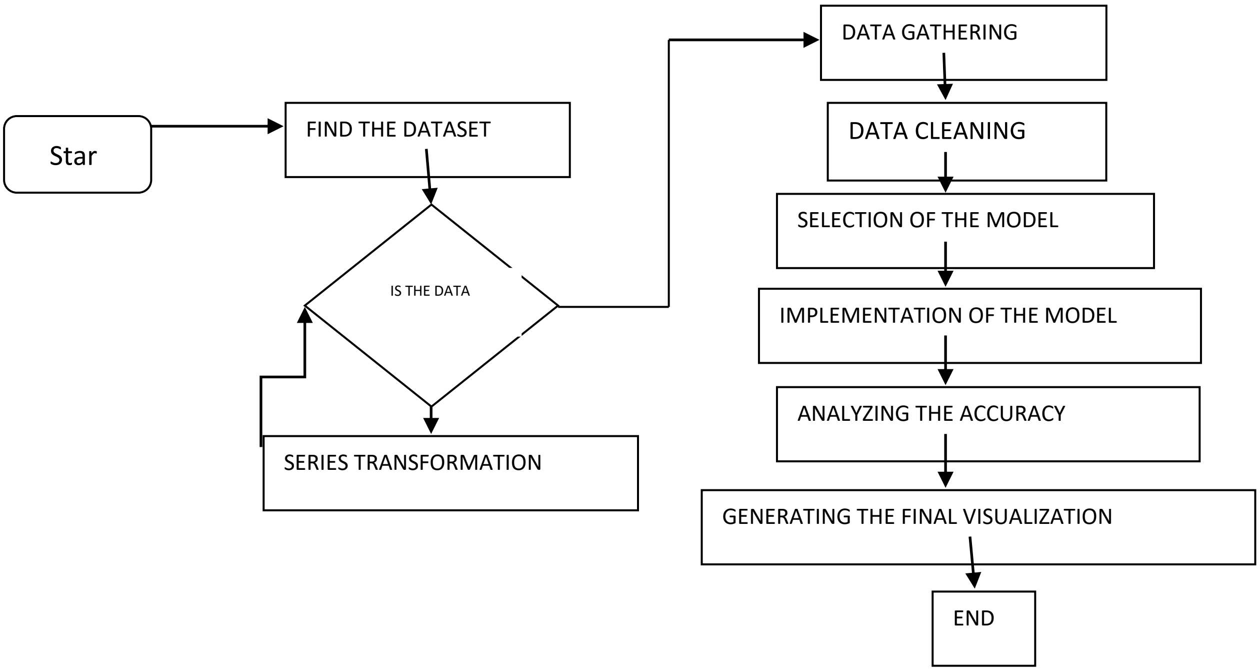
All the countries across the world are trying to halt the virus. The accepted way to attenuate the growth is to practice social distancing. Major steps are taken by various governments across the globe, and the most critical steps are imposing lockdowns on countries. With varied implementation of lockdown, there have been some outbreaks, including in Italy and Spain and this work consists of the analysis and prediction of such measures.

Similar measures have been taken in India to minimize social contacts, including the shut-down of various organizations, lockdown of schools and colleges, Junta Curfew, and a 21-day lockdown. Since these measures have immense pressure on economy and is important for containing the Corona virus, quantitative estimates are imperative to learn the impact of spread which will help in planning policies. Given the paucity of such quantitative estimates, the predictions given in this paper becomes critical and to know when the changes are required.

This paper consists of evaluating these metrics and thereby come up with quantitative estimates using customized Susceptible-Infected -Recovered (SIR). Previous work for prediction of the spread of COVID 19 has been done using many different machines learning algorithms, including neural networks for deep learning, polynomial fitting, exponential smoothing and ARIMA. Neural Networks end up over fitting the data, and polynomial fitting using a third-degree polynomial is also found to be over fitting with a very high bias. This is because the trend in the epidemic spread assumes different nature in different phase's overtime.

A polynomial fit will give an appropriate result for an instance, but as soon as the society relaxes lockdown or changes social-contact pattern in any way, this modelcauses a lot of deviations. A more appropriate method is found out to be an analytical solution to the differential system of equations using a three-compartment system SIR model described below.

I will be using a 6-step approach to build our model then we will predict stock prices with the help of the model.



PYTHON

It has been chosen as the language of choice for this project. This was an easy decision for the multiple reasons.

1. Python as a language has an enormous community behind it. Any problems that might be encountered can be easily solved with a trip to Stack Overflow. Python is among the most popular languages on the site which makes it very likely there will be a direct answer to any query.
2. Python has an abundance of powerful tools ready for scientific computing. Packages such as NumPy, Pandas, and Scikit are freely available and well documented.Packagessuchasthesecandramaticallyreduce,andsimplifythe code needed to write a given program. This makes iteration quick.

Python as a language is forgiving and allows for programs that look like pseudo code. This is useful when pseudocode given in academic papers needs to be implemented and tested. Using Python, this step is usually reasonably trivial.

However, Python is not without its flaws. The language is dynamically typed and packages are notorious for Duck Typing. This can be frustrating when a package method returns something that, for example, looks like an array rather than being an actual array. Coupled with the fact that standard Python documentation does not explicitly state the return type of a method, this can lead to a lot of trials and error testing that would not otherwise happen in a strongly typed language. This is an issue that makes learning to use a new Python package or library more difficult than it otherwise could be.

PYTHON LIBRARIES

- ***NUMPY***

It is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data.

- ***PANDAS***

It is an open-source library that is built on top of NumPy library. It is a Python package that offers various data structures and operations for manipulating numerical data and time series. It is mainly popular for importing and analyzing data much easier. Pandas is fast and it has high-performance & productivity for users.

- ***MATPLOTLIB***

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002.

One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

- ***SCIKIT-LEARN***

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machine, random forest, gradient boosting, k-means etc. It is mainly designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

Scikit-learn is largely written in Python, with some core algorithms written in Cython to achieve performance. Support vector machines are implemented by a Python wrapper around LIBSVM.i.e., logistic regression and linear support vector machines by a similar wrapper around LIBLINEAR.

- ***SCIPY***

SciPy refers to several related but distinct entities:

- The *SciPy ecosystem*, a collection of open source software for scientific computing in Python.
- The *community* of people who use and develop this stack.
- Several *conferences* dedicated to scientific computing in Python - SciPy, EuroSciPy, andSciPy.in.
- The SciPy library, one component of the SciPy stack, providing many numerical routines.

Procedure

1. *DATA GATHERING*

- Collected Covid data on daily basis: I searched on a number of platforms for Indiandata; we have taken the data from 28 January, 2020.

- Proper study of the datasets: After selection of the proper data we have assigned in them in proper schemas and used it for the prediction.

2. *DATA CLEANING*

- Removed duplicate Data: During data cleaning we deleted the duplicate data in order to get best possible prediction with least error count.

- Removed missing Data: On a number of days there were no cases of corona so we have not taken them for the prediction purposes as it may manipulate the prediction.

- Assigning Proper data types to every column

3. *SELECTION OF THE MODEL*

- Study of various models: Previous work for prediction of the escalation of COVID 19 has been done using many different machines learning algorithms, including neural networks for deep learning, polynomial fitting, exponential smoothing and ARIMA. Neural Networks end up over fitting the data, and polynomial fitting using a third-degree polynomial is also found to be over fitting with a very high bias. This is because the trend in the epidemic spread assumes different nature in different phase’s overtime. A polynomial fit will give an appropriate result for an instance, but as soon as the society relaxes lockdown or changes social-contact pattern in any way, these models cause a lot of deviations. A more appropriate method is found out to be an analytical solution to the differential system of equations using a three-compartment system SIR model described below.

- Selection of model: I selected the SIR model, because The SIR model is one of the simplest compartmental models, and many models are derivatives of this basic form. The model consists of three compartments:
  - S: The number of susceptible individuals. When a susceptible and an infectious individual come into "infectious contact", the susceptible individual contracts the disease and transitions to the infectious compartment.
  - I: The number of infectious individuals. These are individuals who have been infected and are capable of infecting susceptible individuals.
  - R for the number of removed (and immune) or deceased individuals. These are individuals who have been infected and have either recovered from the disease and entered the removed compartment, or died. It is assumed that the number of deaths is negligible with respect to the total population. This compartment may also be called "recovered" or "resistant".

  This model is reasonably predictive for infectious diseases that are transmitted from human to human, and where recovery confers lasting resistance, such as measles, mumps and rubella.

  These variables (S, I, and R) represent the number of people in each compartment at a particular time. To represent that the number of susceptible, infectious and removed individuals may vary over time (even if the total population size remains constant), we make the precise numbers a function of *t* (time): S(*t*), I(*t*) and R(*t*). For a specific disease in a specific population, these functions may be worked out in order to predict possible outbreaks and bring them under control.

4. *IMPLEMENTATION OF THE MODEL*

- Fitting the dataset in the model: The cleaned data was used to proceed with the prediction and to use SIR on it to get the predictions.

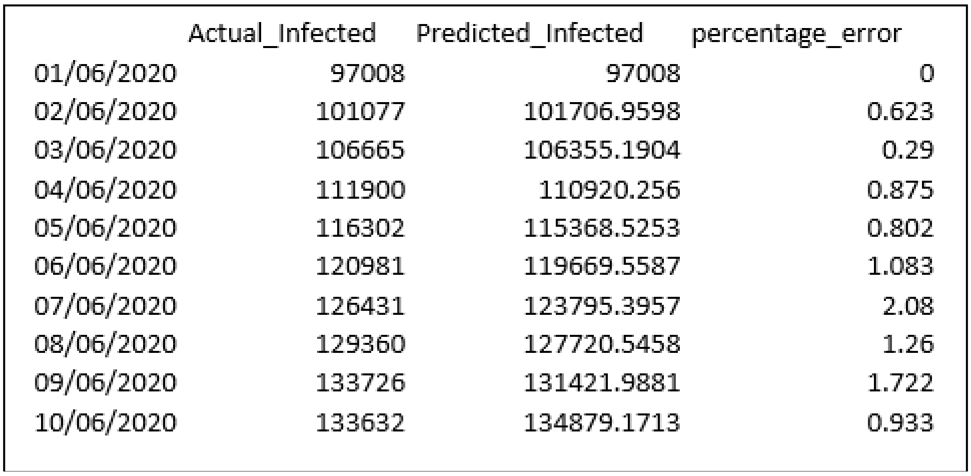
- Training of the model to get the values of the parameters alpha and beta: As alpha beta taken were assumed values so in order to get better results a number of tests were done.

- Predicting the upcoming Covid cases.

5. *ANALYZING THE ACCURACY*

- In this step we analyzed the data of upcoming days and compared it with our predicted data, and calculated the accuracy for the next 7-8 days and changed the parameters value to make it more accurate.

- At last we generated the error percentage to get the knowledge of error for each day, at first, we were having high amount of error in prediction but as soon as more data has been trained the predicted values started to get low errors.



	Actual_Infected	Predicted_Infected	percentage_error
01/06/2020	97008	97008	0
02/06/2020	101077	101706.9598	0.623
03/06/2020	106665	106355.1904	0.29
04/06/2020	111900	110920.256	0.875
05/06/2020	116302	115368.5253	0.802
06/06/2020	120981	119669.5587	1.083
07/06/2020	126431	123795.3957	2.08
08/06/2020	129360	127720.5458	1.26
09/06/2020	133726	131421.9881	1.722
10/06/2020	133632	134879.1713	0.933

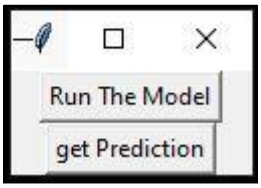
Fig.0

6. GENERATING VISUALIZATIONS

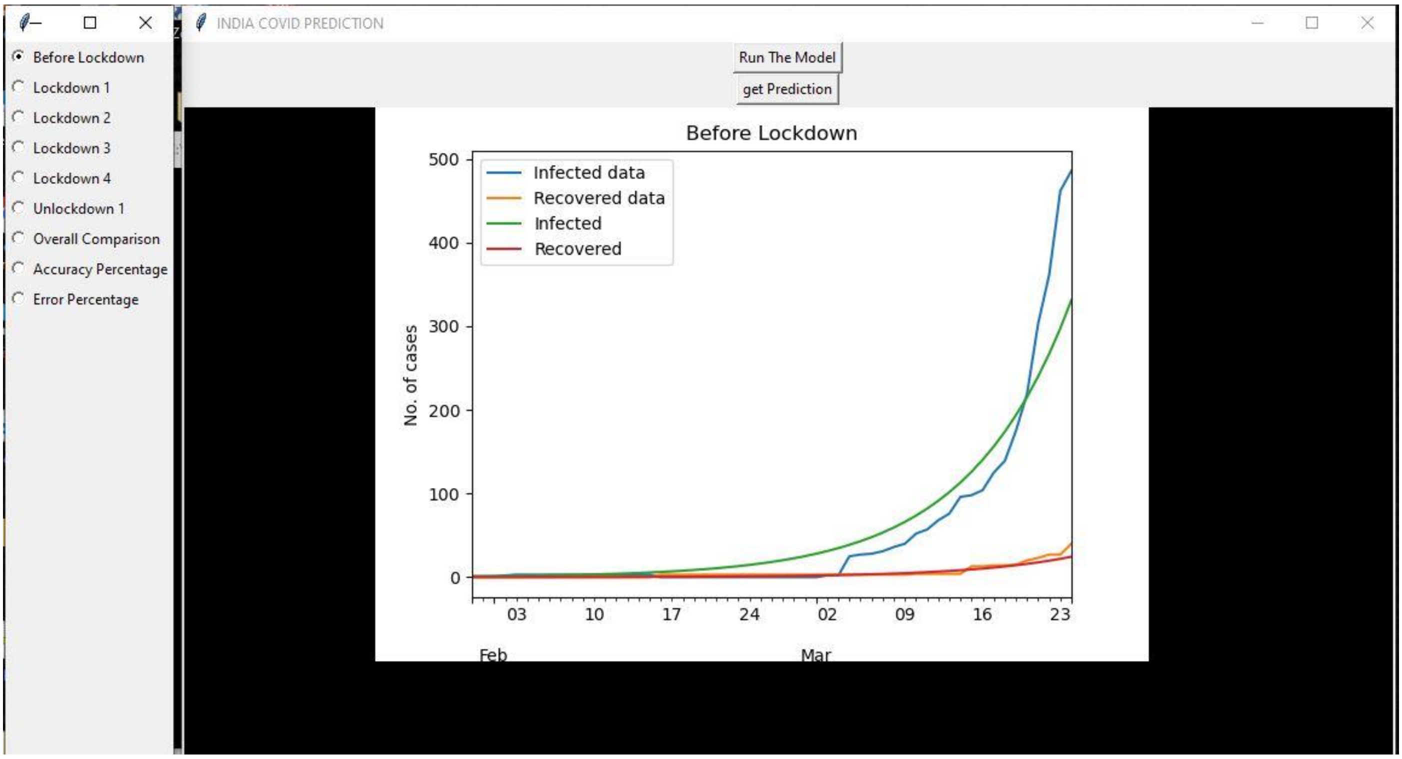
- In this last process after getting less error rate we generated the graphs of our prediction to that of the actual value to get the proper visuals how our model is performing.

Visualizations

- Fig.1 Initial GUI

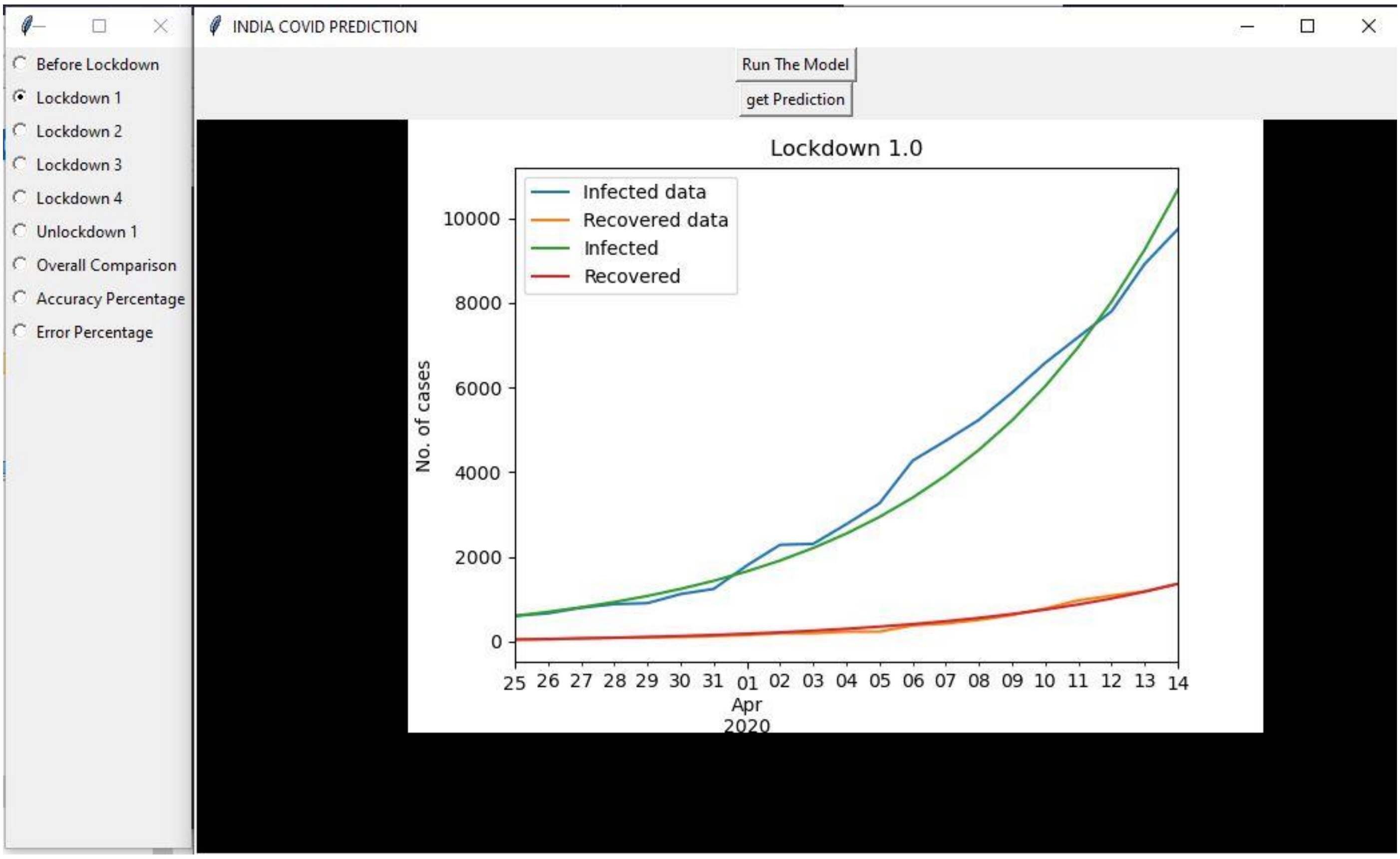


- Fig.2Visualization Before Lockdown



This is a prediction graph before the first lockdown. As such the cases were very low the model had a very minor data on which it was trained shows such deviation in the infected data curve.

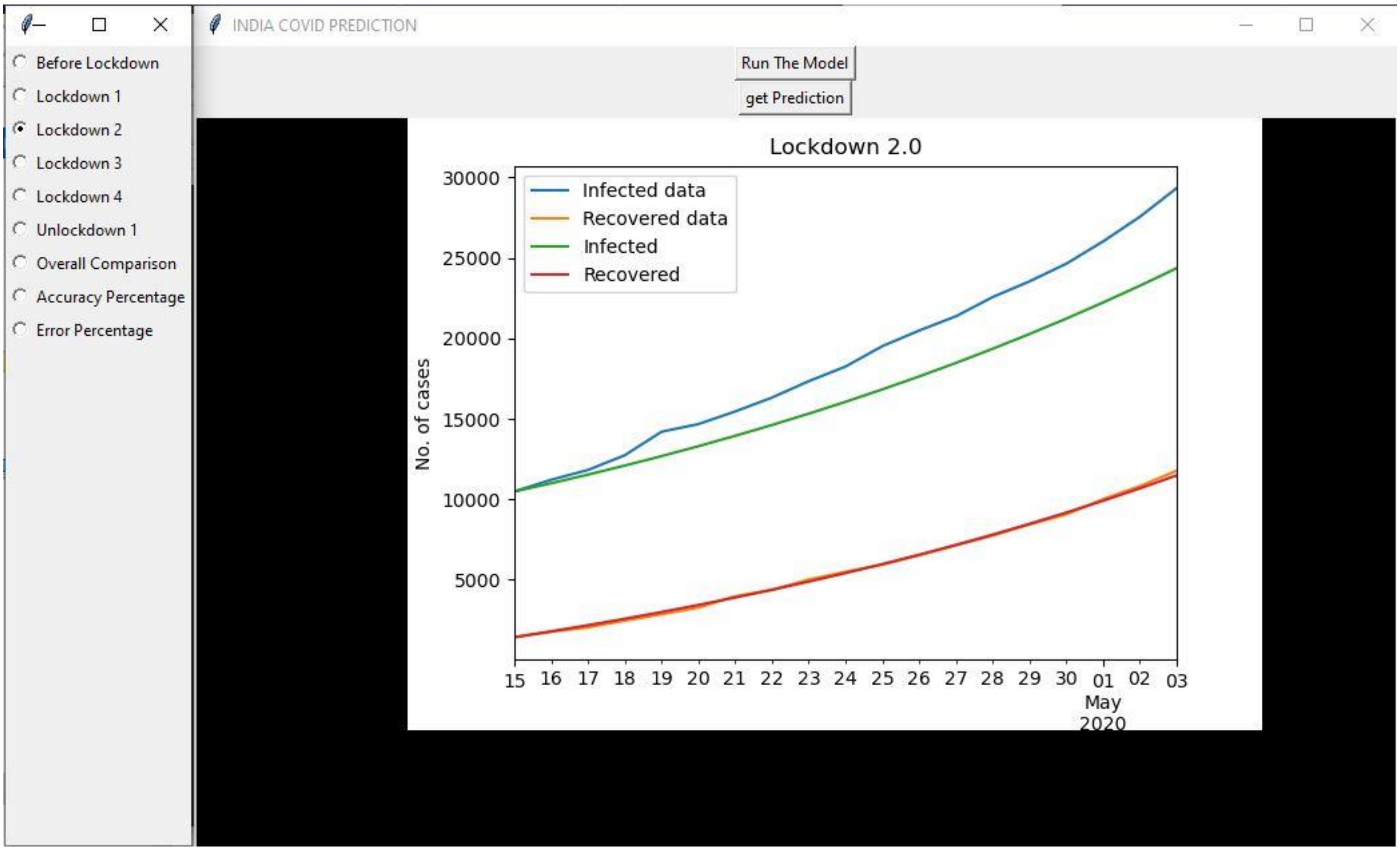
- Fig.3Visualization of Lockdown 1.0



○ Before Lockdown
● Lockdown 1
○ Lockdown 2
○ Lockdown 3
○ Lockdown 4
○ Unlockdown 1
○ Overall Comparison
○ Accuracy Percentage
○ Error Percentage

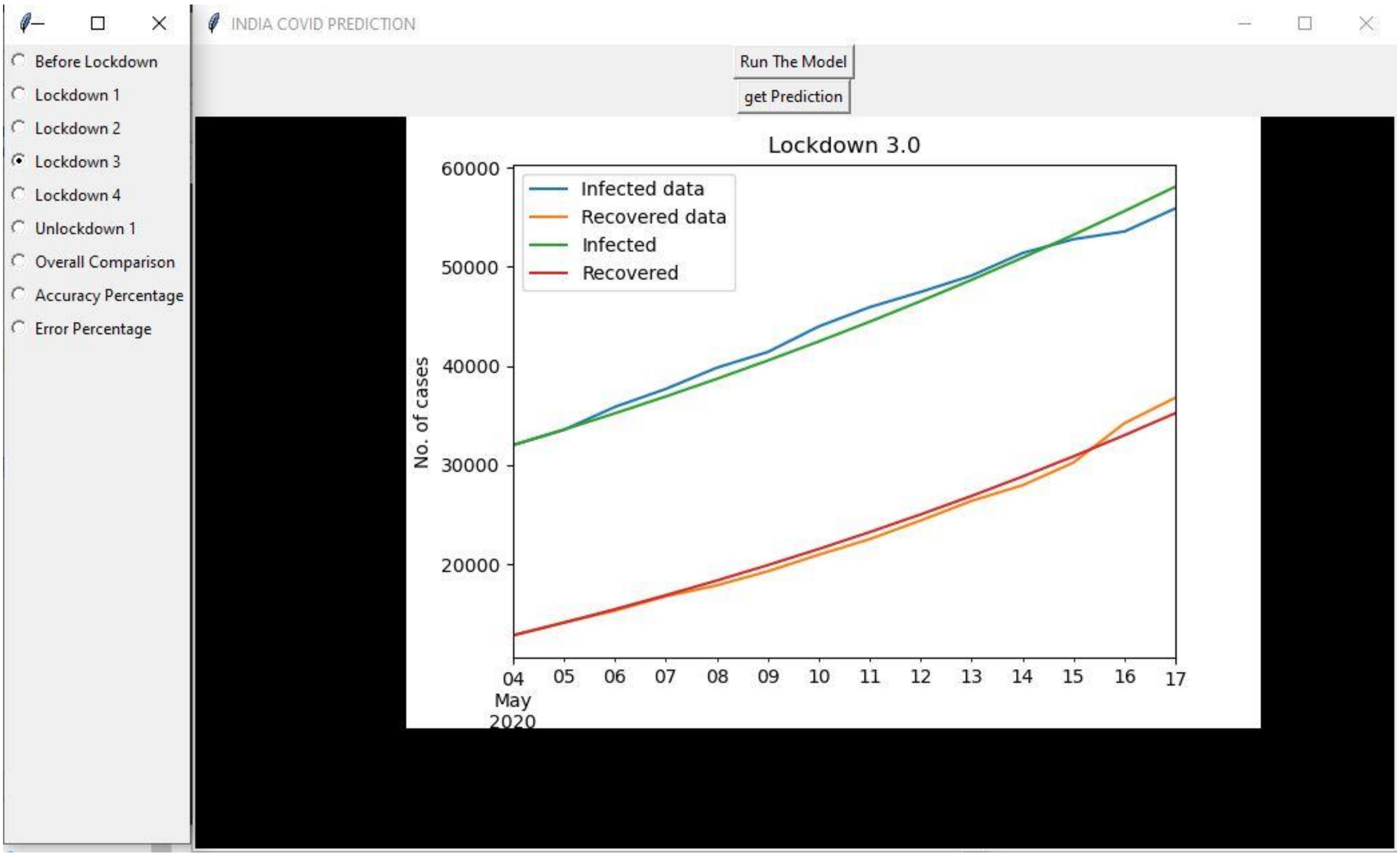
As the first lockdown initiated and so the greater number of cases started appearing and we were able to get appropriate data in larger number. Since now our model had enough data to train, we were able to get more accurate result.

- Fig.4Visualization of Lockdown 2.0



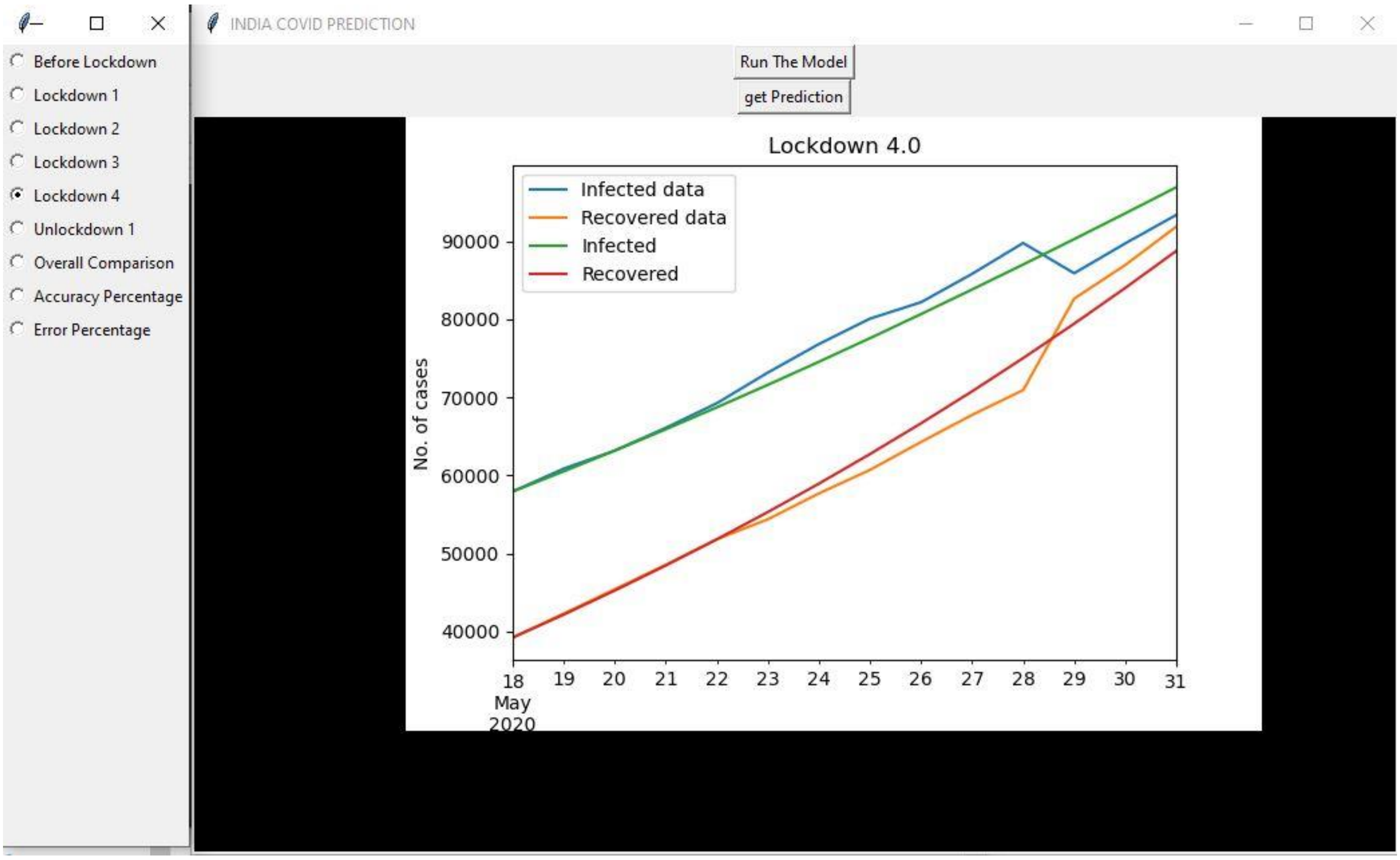
As in lockdown 2.0, greater data brought in greater error and we had re-train the model on the regular basis.

- Fig.5Visualization of Lockdown 3.0



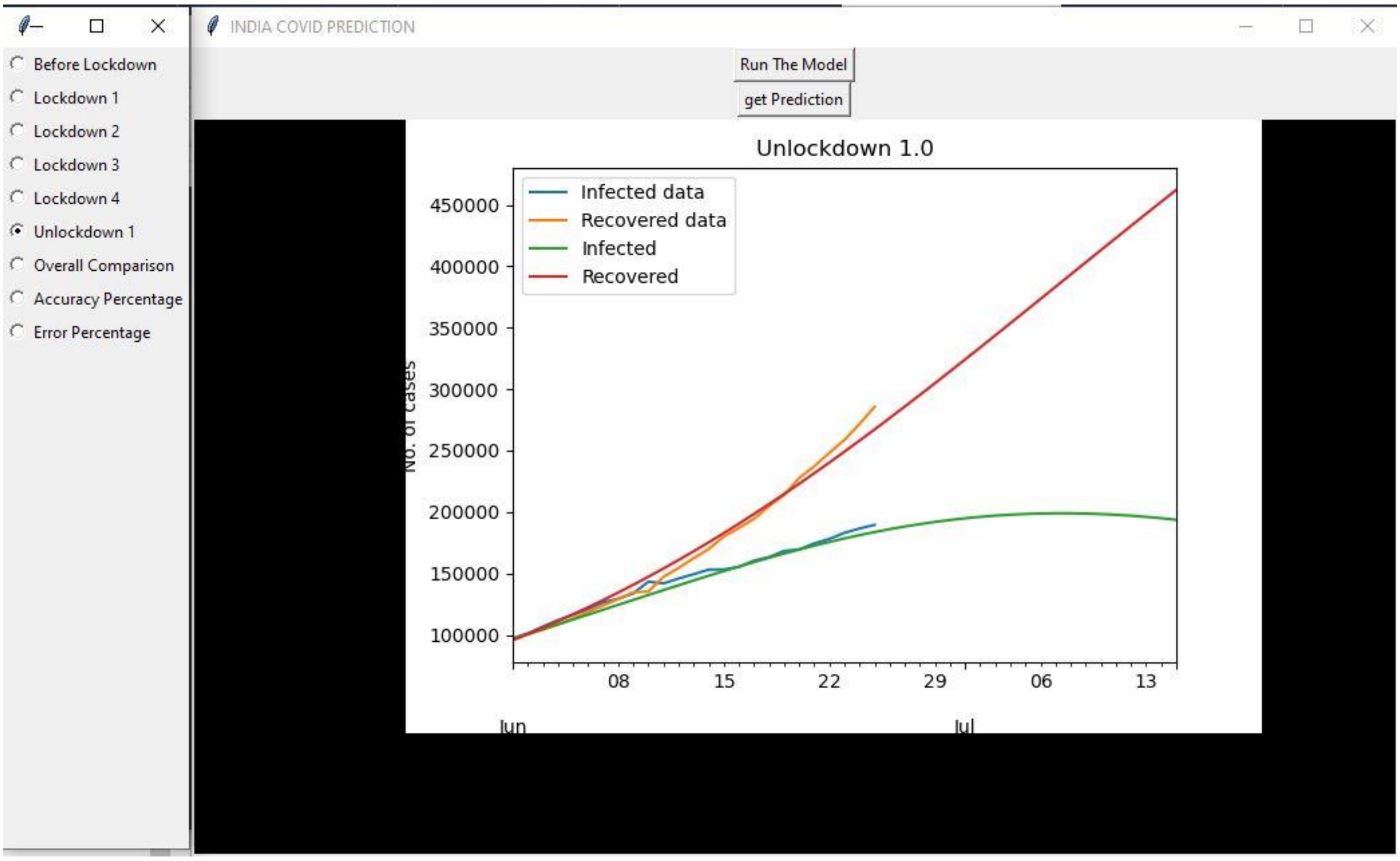
As we can see in the prediction of third lockdown, we were able to achieve a greater accuracy. As the model was getting trained by more no. of data it achieved greater accuracy.

- Fig.6Visualization of Lockdown 4.0



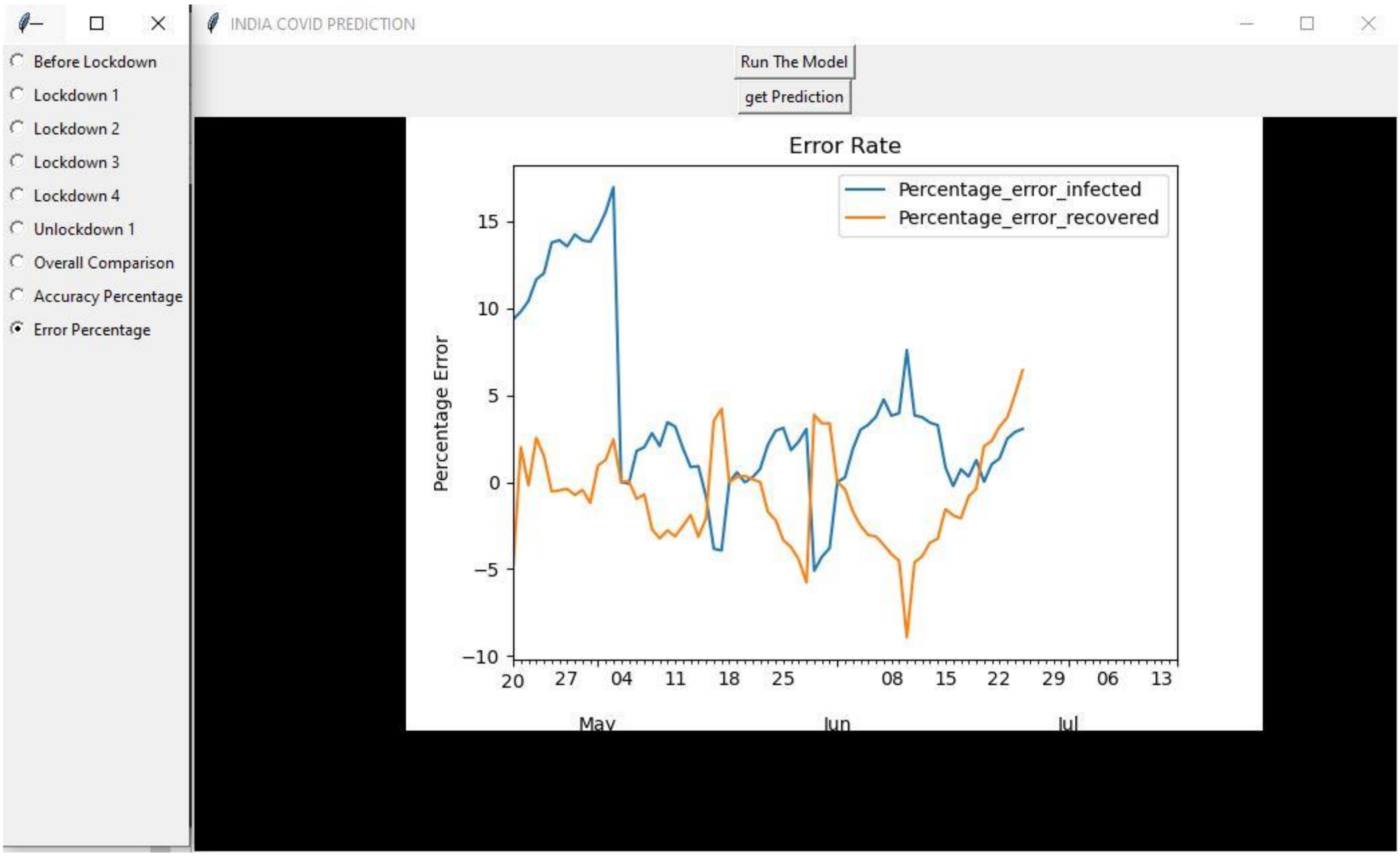
In lockdown 4.0, as people got more liberty and started getting back to their jobs, we saw a fluctuation in the graph due to the abrupt eruption of cases across the nation.

- Fig.7Visualization of Unlock 1.0 and Predicted value of next 20 days (as on 25-06-2020)

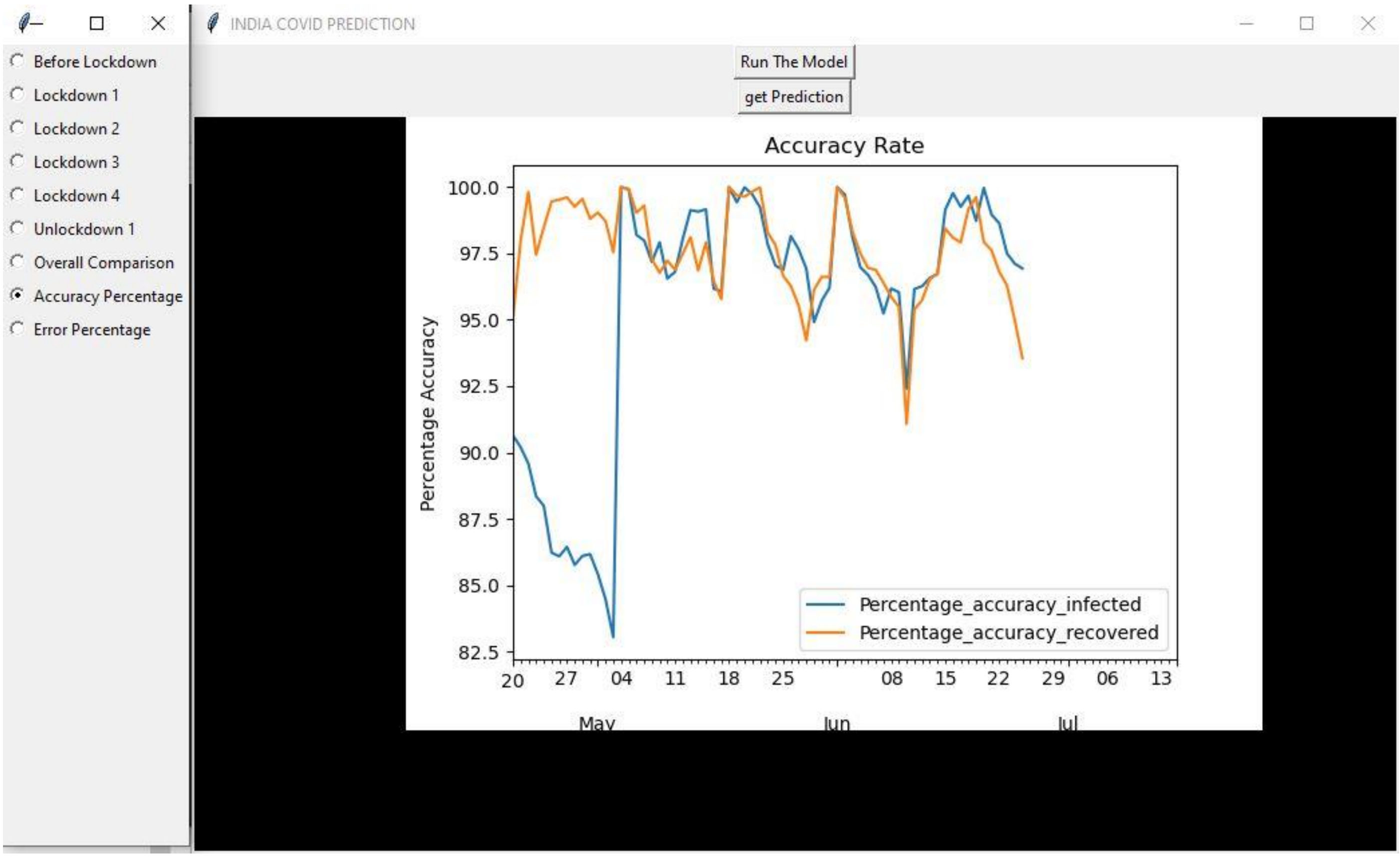


As model got trained, we were able to achieve a greater accuracy and predicted the no. of cases for next 20 days, we saw that the reading we achieved was very close to the actual ones and were able to cut down the error to the minimum.

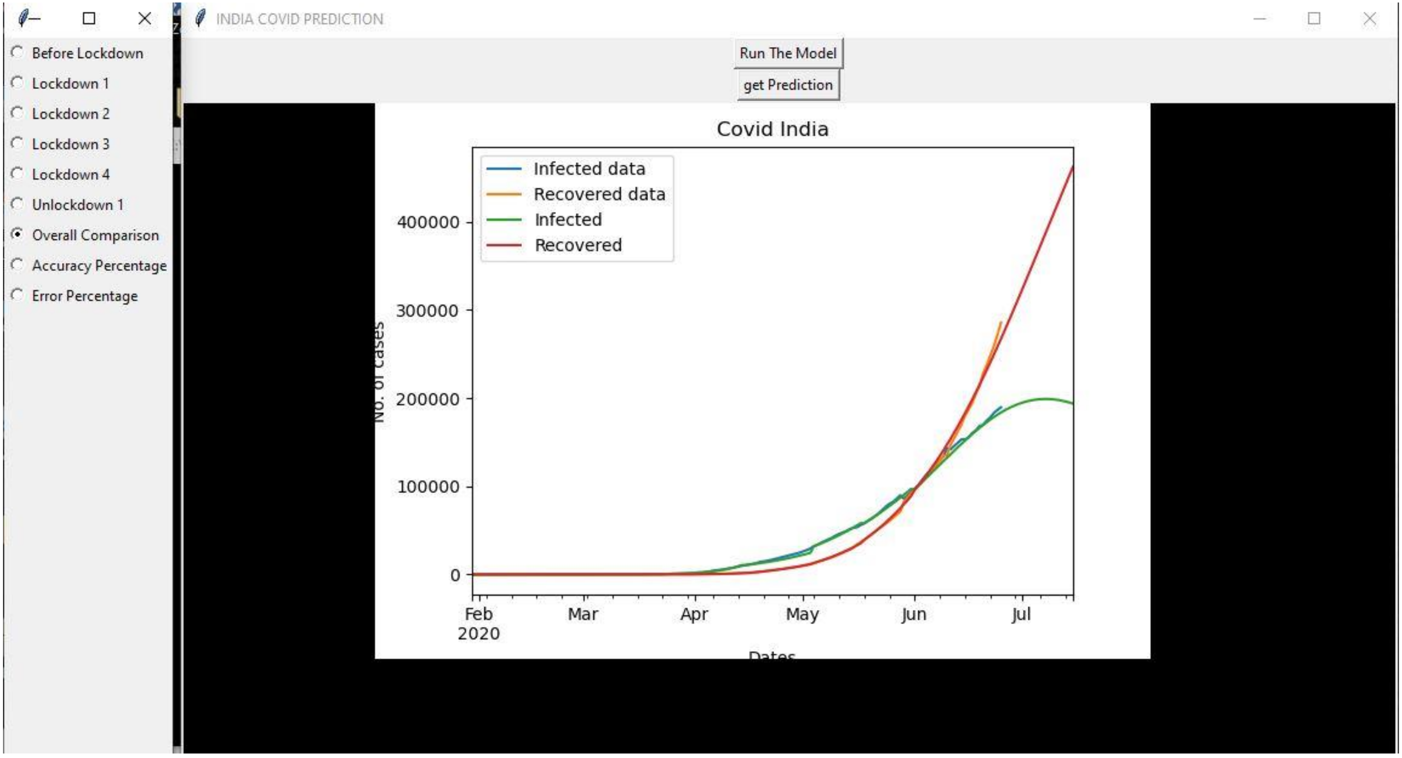
- Fig.8Visualization of error rate



- Fig.9Visualization of rate of accuracy



- Fig.10 Overall Comparison



○ Before Lockdown
○ Lockdown 1
○ Lockdown 2
○ Lockdown 3
○ Lockdown 4
○ Unlockdown 1
● Overall Comparison
○ Accuracy Percentage
○ Error Percentage

Conclusion

In conclusion we would like to say that rate of recovery will surpass the rate of infection and the peak will likely be achieved in the late August or September. The prediction can be made more accurate and specific if the data is made available. I can predict things like in what age spectrum it is more likely to spread, in what age spectrum does death more likely, what preconditions makes people more vulnerable (i.e. what health issues makes people more vulnerable like diabetes, cholesterol etc.), we can get prediction of specific areas. This can be achieved with a larger amount of data and longer period of training the model. As such for the 2-month period predicting this much detail would be very unlikely, hence a basic model predicting the escalation and downfall of the number of cases and recoveries was feasible and developed.

References

1. “Coronavirus Cases:” Worldometer, www.worldometers.info/coronavirus/.

2. “COVID-19 dataset on githubhttps://github.com/CSSEGISandData/COVID-19

3. “The Sir epidemic model” https://scipython.com/book/chapter-8-scipy/additional-examples/the-sir-epidemic-model/

4. Srk. “Novel Corona Virus, 2019 Dataset.” Kaggle, 14 May, 2020, www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset.

5. Tanu Singhal. A review of coronavirus disease-2019 (COVID-19). The Indian Journal of Pediatrics, pages 1–6, 2020.

6. Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the chinese center for disease control and prevention. Jama, 2020.

7. Yueling Ma, Yadong Zhao, Jiangtao Liu, Xiaotao He, Bo Wang, Shihua Fu, Jun Yan, Jingping Niu, and Bin Luo.

8. Effects of temperature variation and humidity on the mortality of covid-19 in Wuhan. medRxiv, 2020.

9. Miguel B. Araujo and Babak Naimi. Spread of SARS-CoV-2 Coronavirus likely to be constrained by climate. medRxiv, 2020.

10. Samar Salman and Mohammed Labib Salem. The mystery behind childhood sparing by COVID-19. International Journal of Cancer and Biomedical Research, 2020.