**Data Analysis of Covid-19 using MLP for India**

**A PROJECT REPORT**

*for*

**Technical Answers for Real World Problems (ITE3999)**

*in*

**B.Tech (Information Technology) Programme**

*by*

**GAYATHRI B (17BIT0049)**

**HARNITHA (17BIT0085)**

**Fall Semester 2020 - 2021**

*Under the Guidance of*

**Prof. B. VALARMATHI**

Associate Professor (Senior), SITE



**School of Information Technology and Engineering**

October, 2020

**DECLARATION BY THE CANDIDATE**

We here by declare that the project report entitled **“Data Analysis of Covid-19 using MLP for India”** submitted by us to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Technical Answers for Real World Problems (ITE3999)** is a record of bonafide project work carried out by us under the guidance of **Prof. B.Valarmathi.** We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore Signature

Date :



**School of Information Technology & Engineering [SITE]**

**CERTIFICATE**

This is to certify that the project report entitled **“Data Analysis of Covid-19 using MLP for India”** submitted by **Student Name 1 (Reg. No), Student Name 2 (Reg.No)** to Vellore Institute of Technology University, Vellore in partial fulfillment of the requirement for the award of the course **Technical Answers for Real World Problems (ITE3999)** is a record of bonafide work carried out by them under my guidance.

**Prof. B.Valarmathi**

**GUIDE**

**Associate Professor (Senior), SITE**

**Data Analysis of Covid-19 using MLP for India**

**GAYATHRI B1, HARNITHA2**

**Abstract**

Coronavirus has made the world to a standstill by affecting billions of people worldwide. Doctors, nurses and other frontline workers are struggling hard to handle the situation. This project aims to implement the future scope of the research paper “Yadav, R. S. (2020). Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. International Journal of Information Technology, 1-10.” The future scope of the paper is to create a regression based neural network model for prediction of total COVID-19 cases in India the next day. The project aims to develop a regression based artificial neural network model that automatically estimate the number of total cases for the next day in India. The uniqueness in this project is that we combine real time analytics and prediction. It is mentioned real-time because this project uses API to get data that gets updated every day. Thus, the dashboard and the prediction model updates and provides updated results automatically, without manually updating each day. The prediction model uses time series data and converts that to supervised learning format, thus converting it to a regression-based problem. Validation is performed both on training and testing model. Validation on testing data follows an approach known as walk forward validation. Thus, this model is a combination of both supervised based training and time series-based validation on time series data. This project also compares the proposed prediction model with different hyper-parameters, activation functions and optimizers so as to get a view of the best performing model from all. A comparison is also done with the proposed model and the model proposed in the base paper. A dashboard is also created and deployed in a cloud-based platform Heroku so as to be available to everyone at any time. The dashboard illustrates various graphs that help doctors to better analyse and understand the present situation in an easier way.

**Keywords –** *neural network, regression, dashboard, walk forward validation, series to supervised, covid-19 prediction*

1. **INTRODUCTION**

Coronavirus virus is a severe threat now that impacts the daily lives of people worldwide. The economy of several countries are pushed into a tailspin, and the governments are struggling to control the pandemic. The capacity of the virus to reach very quickly and rapidly gives the decision-makers significantly less time to plan and implement preventive measures. Without proper analytics, it is not easy to understand the behaviour of the spread. Technology has been a boon, in this case, helping in unmanned tasks and planning tasks. Since at present, no vaccine is available, a supreme component in tracking is to visualise the trends past and present and come up with a reasonable prediction to plan effectively. Machine learning and data analytics had helped technologist, doctors and others to create models and dashboard to trace the spread of the virus efficiently.

There are different algorithms, methods being implemented to create an efficacious model to predict future cases. This project uses MLP(Multilayer Perceptron), and the performance of our model is evaluated using Root Mean Square Error(RMSE) for the prediction.

Concerning this, the project aims to visualize past and present trends in the form of a dashboard, furthermore create a prediction model to predict the cases for the next day in real-time.

1. **BACKGROUND**

Using deep learning techniques like multi-layer perceptron to predict the possible cases using regression based neural network model in the future by utilizing the real-time information from the covid19 API . The project also aims to provide a dashboard for doctors and frontline workers that contains graphs that visualizes the past, current and predicted analysis . We also show the EDA with various chart like pie chart,bar chart , line chart and the also using the world map . In the EDA we displayed

* India total and daily cases-
  + line chart for total deceased , total recovered , total confirmed
  + Pie chart for total deceased , total recovered , total confirmed
  + Pie chart for daily deceased , daily recovered , daily confirmed
* Statewise Comparison
  + Line chart for Daily cases for Tamil Nadu- confirmed , death and recovered
  + Total Confirmed cases of all the states
  + Barchart of confirmed ,recovered,deceased and active cases of all the states
* Statewise testing
* World data
  + Barchart for the total confirmed cases for all the country and top 25 countries
  + Barchart of the top 25 deceased country
  + Barchart for recovered
  + Choropleth(world map)

1. **Literature Survey**

The following table suggest the diverse applications in Covid 19 analyses available offered by various techniques.

1. A python based support vector regression model for prediction of COVID19 cases in India

The paper uses SVR to predict the number of covid cases in India. It uses 60:40 train test split and StandardScaler to normalize data.it used SVR with rbf kernel and degree 3 and used MSE,RMSE,R2 and accuracy to measure performance. The model received an accuracy score of 97% on testing data and 87% on predicting daily cases. The proposed methodolgy predicts the total number of COVID19 infected cases, total number of daily new cases, total number of deaths and total number of daily new deaths

1. Analyzing the epidemiological outbreak of COVID‐19: A visual exploratory data analysis approach

In this paper, an exploratory data analysis with visualizations has been made to understand the number of different cases reported (confirmed, death, and recovered) in different provinces of China and outside of China. The visualization are made by using plotly API. Visualization model like map view and tree map view provides an interactive interface and visualize each and every raw fact in a comprehensive manner. The paper suggests to implement machine learning algorithms to perform prediction.

1. An Exploratory Data Analysis of COVID-19 in India

The paper performs EDA on john Hopkin university dataset for india. The graphs shows spread in india vs other countries ,age wise spread of covid in india, state-wise analysis of covid and a barchart for symptoms of covid 19.the paper used ARIMA model to web scrap time series dataset.As part of future work , it aims to use machine learning algorithms to predict daily cases.

1. Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India

The paper proposed six regression analysis based machine learning models for prediction of the COVID-2019 outbreak datasets of India. The models basically regression analysis based exponential, quadratic, third degree, fourth degree, fifth degree and sixth degree Polynomials. These models also predict the outbreak of the COVID-2019 in India for the next 7 days. The value of for proposed models namely sixth degreepolynomial is very close to the confirmed case or actual results regarding training dataset of the COVID-2019

1. Dynamic model of COVID-19 disease with exploratory data analysis

The paper propose and analyze a non- linear mathematical model to investigate the effect of healthy sanitation and awareness on the transmission dynamics of Coronavirus disease (COVID-19) prevalence. This study reveals that the probability of getting infected p 1 , maximum transmission rate βmax of COVID-19 and restriction rate γ are most sensi- tive. In addition, it was also discovered that the propagation of information on good hygiene over time induce behavioral change in individuals leading to significant reduction in the number of quarantined and infected individuals.

1. Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 Patients

This paper analyses covid-19 screening data. The screening data consists of nucleic acid and hematology data from 2510 patients for retrospective analysis. From the analysis the paper concludes that The positive rate of influenza A and B infection was higher than that of COVID-19. When pharyngeal swab collection may cause infection, fecal samples can be examined. Evaluation of pharyngeal swab and fecalsamples can improve the positive rate of nucleic acid detection. The COVID-19 can cause some hematological indices changes. results show that during the same period, the positive rate of influenza A/B infection, which lead to influenza outbreaks, was higher than that of COVID-19 infection. Therefore, COVID-19 must be differentiated from influenza A/B to ensure appropriate patient management.No patients diagnosed with COVID-19 and influenza A/B infection had coexisting conditions.

1. Prediction for the spread of COVID-19 in India and effectiveness of preventive measures

In this paper, data-driven estimation methods like long short-term memory (LSTM) and curve fitting for prediction of the number of COVID-19 cases in India 30 days ahead and effect of preventive measures like social isolation and lockdown on the spread of COVID-19. The prediction of various parameters (number of positive cases, number of recovered cases, etc.) obtained by the proposed method is accurate within a certain range and will be a beneficial tool for administrators and health officials.

1. Real-time estimation and prediction of mortality caused by COVID-19 with patient information-based algorithm

This paper uses The Patient Information Based Algorithm (PIBA) to estimate the death rate of a disease in real-time. The PIBA model accurately predicted a case fatality of 1.6% for symptomatic patients in China at a very early stage in the Covid-19 pandemic. The model can be generalized to predict case fatality for any infection (including asymptomatic), to predict the rate of severe disease, and to predict the death rate for patients who develop severe disease. These early, accurate predictions inform the public, society, and governments to estimate the extent of the disease's harm and to develop suitable strategies. PIBA is capable of accurately estimating the disease mortality and the number of future deaths

1. Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis

The paper focus on 2 things 1. generating short term (real-time) forecasts of the future COVID-19 cases for multiple countries 2. risk assessment (in terms of case fatality rate) of the novel COVID-19 for some profoundly affected countries by finding various important demographic characteristics of the countries along with some disease characteristics .The paper uses hybrid approach based on autoregressive integrated moving average model and Wavelet-based forecasting model that can generate short-term (ten days ahead) forecasts of the number of daily con- firmed cases for Canada, France, India, South Korea, and the UK.

1. Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis

It is of crucial importance to estimate the reproductive number (R0) of the novel virus in the early stage of outbreak and make a prediction of daily new cases on the ship. The paper fitted the reported serial interval (mean and standard deviation) with a gamma distribution and applied “earlyR” package in R to estimate the R0 in the early stage of COVID-19 outbreak and applied “projections” package in R to simulate the plausible cumulative epidemic trajectories and future daily incidence by fitting the data of existing daily incidence, a serial interval distribution, and the estimated R0 into a model based on the assumption that daily incidence obeys approximately Poisson distribution determined by daily infectiousness.

1. Forecast and evaluation of COVID-19 spreading in USA with reduce d-space Gaussian process regression

The paper analyzes historical and forecast infections for COVID-19 death based on Reduced-Space Gaussian Process Regression associated to chaotic Dynamical Systems with information obtained in 82 days with continuous learning, day by day, from January 21 th , 2020 to April 12 th .The paper proposes a Reduced- Space Gaussian Process Regression model predicts that the epidemic will reach saturation in USA on July 2020. The findings suggest, new quarantine actions with more restrictions for containment strategies implemented in USA could be successfully, but in a late period, it could generate critical rate infections and death for the next 2 month.

1. A model based study on the dynamics of COVID-19: Prediction and control

In this paper a mathematical model introducing a quarantine class and governmental intervention measures to mitigate disease transmission.

The paper performs sensitivity analysis of the essential reproduction number and found that reducing the contact of exposed and susceptible humans is the most critical factor in achieving disease control. forecast a short-term trend of COVID-19 for the three highly affected states, Maharashtra, Delhi, and Tamil Nadu, in India, and it suggests that the first two states need further monitoring of control measures to reduce the contact of exposed and susceptible humans.

1. Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries

The paper employed a multivariate prediction model, based on up-to-date transmission and clinical parameters, to simulate outbreak trajectories in 16 countries, from diverse regions and economic categories. In each country, they modelled the impacts on intensive care unit (ICU) admissions and deaths over an 18-month period for following scenarios: (1) no intervention, (2) consecutive cycles of mitigation measures followed by a relaxation period, and (3) consecutive cycles of suppression measures followed by a relaxation period. This multi-country analysis demonstrates that intermittent reductions of R below 1 through a potential combination of suppression interventions and relaxation can be an effective strategy for COVID-19 pandemic control.

1. Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study

Data were obtained from the Google Trends website. Linear regression and long short-term memory (LSTM) models were used to estimate the number of positive COVID-19 cases. All models were evaluated using 10-fold cross-validation, and root mean square error (RMSE) was used as the performance metric.

1. Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan

This paper using Logistic Regression , SVM and multi-layer perceptron to predict novel covid 19 cases in Jordan. For performance measures, accuracy, sensitivity, specificity, g\_mean and Precision are used to compare the models and find the best performing models.in that Multi layer perception had highest accuracy and specificity. SVM had the highest precision .

1. Comparison of Regression Models on Covid-19 Cases

The basis of this study is to identify and create an optimal Death Prediction System (DPS) which predicts the death rate by processing and applying an effective, optimized and convenient regression methods and measures that allows the government and the health workers for planning the eradication of the threatening disease. The paper uses Polynomial Regression, Support Vector Regression, Random Forest Regression, Decision Tree Regression to predict and Adjusted-R, Root Mean Square Error for performance measures. In which polynomial regression had the least RMSE.

1. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach

This paper uses Hybrid Multi-Layered Perceptron-Imperialist Competitive Algorithm (MLP-ICA) and ANFIS to predict covid cases in hungary. The models predict that by late May, the outbreak and the total morality will drop substantially. MLP-ICA outperformed ANFIS with delivering accurate results on validation samples.

1. Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms

This paper does a comparison analysis Decision Tree Classifier, Gaussian Naïve Bayesian Classifier, KNN+NCA, Logistic Regression, Machine Learning,, Multilinear Regression, SVM, XGBoost Classifier to predict covid cases.in which SVM,DT classifier and random forest has the highest accuracy. The experiments reveal the persons of age groups 20-30, 30- 40 and 40-50 are suffered with COVID-19.

1. COVID-19 Outbreak Prediction with Machine Learning

This paper presents a comparative analysis of machine learning and soft computing models to predict the COVID-19 outbreak as an alternative to SIR and SEIR models. Among a wide range of machine learning models investigated, two models showed promising results (i.e., multi-layered perceptron, MLP, and adaptive network-based fuzzy inference system, ANFIS).

1. COVID-19 Detection using Artificial Intelligence

This paper uses xray images to predict whether a person has a covid 19 or not. For the proposed model development, validation, and testing 260 images available from the repository of GitHub and Kaggle were used. The Images consists of 130 COVID-19 (ignoring SARS, MERS and ARDS) and 130 Normal X-ray images. The proposed model achieved: sensitivity of 100%, specificity of 100%, accuracy of 100%, PPV of 100%, and NPV of 100% in the dataset.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Author** | **Objectives** | **Model** | **Data Source** | **Evaluation Method** | **Data**  **set** | **Accuracy** | **Precision** | **Recall** | **F-**  **measure** |
| 1  Debanjan Parbat , Monisha Chakraborty | To predict the total number of deaths,recovered cases,cumulative number of confirmed cases and number of daily cases. | Uses support vector regression model | Novel Coronavirus dataset  http://https//www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset | The mean square error (MSE), root mean sqaure error (RMSE), R2 score and percentage accuracy are calculated | 35043 cases | 97%-predicting deaths, recovered, cumulative no.of confirmed cases .  87% -accuracy in predicting daily new cases. | NA | NA | NA |
| 2  Samrat K. Dey, Md. Mahbubur Rahman,  Umme R. Siddiqi,  Arpita Howlader | to understand the number of different cases reported in and outside of China. | Using exploratory data analysis (EDA) methods and visualization model | 2019 Coronavirus dataset (January‐February 2020)  <https://www.kaggle.com/benhamner/jhucovid19> | NA | 77 262 case | NA | NA | NA | NA |
| 3  Sarvam Mittal | to study and analyze the reported COVID-19 cases in India | ARIMA MODEL  Exploratory Data Analysis (EDA) | Ministry of Health and family Welfare,COVID-19 India website”, “John Hopkins GitHub repository” [8], “Worldometer” | statistics from the cases of different  countries using “Bailey’s model” | 3245 CASES | NA | NA | NA | NA |
| 4  Ramjeet Singh Yadav | Finding the rate of spread of the disease i | machine learning-based SIX regression analysis model | India (COVID-19) Datasets (2020). Retrieved 13 Apr 2020.  https://www.kaggle.com/sudalairajkumar/covid19-in-india/data | root mean square | 16,116 cases | NA | NA | NA | NA |
| 5  Michael O. Adeniyi ,Matthew I. Ekum , Iluno C , Ogunsanya A. S , Akinyemi J. A , Segun I. Oke , Matadi M. B | To analyse the no. of infected people | nonlinear mathematical model- | https://www.worldometers.info/coronavirus/country/italy/ | variational iteration method (IVM), Modified Initial Guess Variational Iterative Method (MIGVIM) and Euler’s method | NA | NA | NA | NA | NA |
| 6  Hu Yunb , Zhuoran Suna , Jun Wua , Aiguo Tanga , Min Hua , Zhongyuan Xianga, | analyze and compare the laboratory data of patients with COVID-19. | RETROSPECTIVE ANALYSIS | Department of Laboratory Medicine, The Second Xiangya Hospital, Central South University,  https://doi.org/10.1016/j.cca.2020.04.018 | median and interquartile range | 2510 patients | 100% | less than 5% | NA | NA |
| 7  Anuradha Tomar,  Neeraj Gupta | Prediction of the number of COVID-19 cases in India 30 days ahead and effect of preventive measures . | Data driven model :  long short-term memory (LSTM) | https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public | NA | - | NA | NA | NA | NA |
| 8  Lishi Wang , Jing Li , Sumin Guo , Ning Xie d , Lan Yao , Yanhong Cao, Sara W. Day g , Scott C. Howard g , J. Carolyn Graff , Tianshu Gu h , Jiafu Ji , Weikuan Gu, Dianjun Sunf, | To estimate the death rate of a disease in real-time | PIBA model | <http://www.nhc.gov.cn/yjb/new_index.shtml>  http://wjw.hubei.gov.cn/fbjd/dtyw/ | NA | 33 CASES | NA | NA | NA | NA |
| 9  [TanujitChakraborty](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302502?via%3Dihub#!),  [ndrajitGhosh](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302502?via%3Dihub#!) | To generate short term (real-time) forecasts of the future COVID-19 cases for multiple countries and risk assessment | (autoregressive integrated moving average  )ARIMA and wavelet-based forecasting techniques. | Global Change Data Lab  https://github.com/indrajitg-r/COVID | regression tree (RT)  mean square error | 59 769 | high accuracy | NA | NA | NA |
| 10  [MengYuanDia](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub#!)  [WenboYu](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub#!)  [LeiPei](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub#!)  [ZhaofenLin](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub#!)  [DechangChen](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub#!) | to estimate the reproductive number (R0) of the novel virus in the early stage of outbreak and make a prediction of daily new cases on the ship | serial interval (mean and standard deviation) with a gamma distribution  and applied “earlyR” package in R to estimate the R0 in the early stage of COVID-19 outbreak. | [Ministry of Health, Labour and Welfare of Japan, 2020](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S1201971220300916?via%3Dihub" \l "bib0030) | mean and standard deviation) with a gamma distribution | 3711 people | 95% | NA | NA | NA |
| 11  [Ricardo Manuel Arias Velásquez](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920303234?via%3Dihub#!),  [Jennifer VanessaMejía Lara](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920303234?via%3Dihub#!) | analyze historical and forecast infections for COVID-19 death | Reduced-Space Gaussian Process Regression model | <https://www.worldometers.info/coronavirus/> | Reduced-Space Gaussian Process Regression ,  Bayesian method |  | 99.19% | NA | NA | NA |
| 12  [ManotoshMandal](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!)  [SoovoojeetJan](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!)  [Swapan KumarNandi](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!)  [AnupamKhatu](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!)  [SayaniAdak](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!)  [T.K.Kar](http://www.sciencedirect.com.egateway.vit.ac.in/science/article/pii/S0960077920302897?via%3Dihub#!) | to mitigate disease transmission. | mathematical model | [https://stopcorona.tn.gov.in](https://stopcorona.tn.gov.in/),  <https://www.worldometers.info/coronavirus/>, | fourth order Runge-kutta method |  | NA | NA | NA | NA |
| 13  Rajiv C, Kevin H, Md Shajedur Rahman S, Daisy O, Carolina Ochoa-Rosales, Valentina Gonzalez-J, Abbas B, Daniel R , Shamini P, Sara S, Christian L. Althaus, Nathalia Gonzalez-J, Oscar H. Franco | (1) prevent critical care overload and deaths, (2) gain time to develop preventive and clinical measures, and (3) reduce economic hardship globally. | multivariate prediction model  -susceptible-exposed-infected-recovered (SEIR) compartmental model | <https://www.worldometers.info/coronavirus/>  <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports>. |  |  | NA | NA | NA | NA |
| 14  Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N | This study aimed to predict the incidence of COVID-19 in Iran. | Linear regression and long short-term memory (LSTM) models | <http://worldometers.info/> | 10-fold cross-validation, and root mean square error (RMSE) |  | NA | NA | NA | NA |
| 15  Fayyoumi, E., Idwan, S., & AboShindi, H. | to predict potential patients of COVID-19 based on their signs and symptoms. | models based on statistical model (Logistic Regression, LR) and machine learning model (Support Vector Machine, SVM, and Multi-Layer Perceptron, MLP | Institutional Review Board at the Hashemite University | statistical model we used Logistic Regression (LR), while in machine learning model we invoked Support Vector Machine (SVM), and Mulit-Layer Perceptron (MLP). | 105 | MLP has shown the best accuracy (91.62%)  LR model was as high as 85.00% | SVM has shown the best precision 91.67%.  LR-66.67% | NA | NA |
| 16  Joseph George,  Ranjeesh R Chandran | s to identify and create an optimal Death Prediction System (DPS) which predicts the death rate | Decision tree regression models ID3  random forest model | <https://covid19.who.int/?gclid=CjwKCAjw34n5BRA9EiwA2u9k3xIzPepuDFYn2LoJYNliP9YKgzw5cNSDJ9Nu7W0OAx7wFiMSwCKmzxoCHQAQAvD_BwE> | k-fold validations, evaluated using the r-Squared value. |  | 93% | NA | NA | NA |
| 17  Gergo Pinter ,  Imre Felde ,  Amir Mosavi , Pedram Ghamisi  and Richard Gloaguen | a hybrid machine learning approach to predict the COVID-19 | Neural Networks (NN)  MLP | https://www.worldometers.info/coronavirus/country/hungary/ | ML and AI methods  the standard values for determination coefficient, mean absolute percentage error root mean square error. | 2500 |  | NA | NA | NA |
| 18  Kolla Bhanu Prakash  S. Sagar Imambi,  Mohammed Ismail,  T Pavan Kumar,  YVR Naga Pawan | To diagnose and predict COVID-19 | Random Forest Regressor and Random Forest Classifier | https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset?select=covid\_19\_data.csv | molecular analysis (PCR) of respiratory tract samples |  | 96% | NA | NA | NA |
| 19  Sina F.Ardabili, Amir Mosavi , Pedram Ghamisi ,  Filip Ferdinand, Annamaria R,  Varkonyi-Koczy, Uwe Reuter ,  Timon Rabczuk , Peter M. Atkinson | prediction can be realized through integrating machine learning and SEIR models. | multi-layered perceptron, MLP, and adaptive network-based fuzzy inference system, ANFIS | https://www.worldometers.info/coronavirus/country | root mean square error (RMSE) and correlation coefficient |  | 86 | NA | NA | NA |
| 20  Fatima M. Salman, Samy S. Abu-Naser, Eman Alajrami, Bassem S. Abu-Nasser, Belal A. M. Ashqar | Detecting COVID 19 using X-rays images | deep learning - CNN model | COVID-19 X-ray images are collected from the GitHub repository shared by Dr. Joseph Cohen  <https://github.com/> | data augmentation-accuracy, sensitivity, specificity, PPV and NPV | 260 images  (contains 130 of COVID-19 and 130 Normal X-ray images) | 100% | 100% | 100% | 100% |

1. **Dataset description & Sample data**

For line chart and piecharts for total india cases and for prediction: <https://api.covid19india.org/csv/latest/case_time_series.csv>

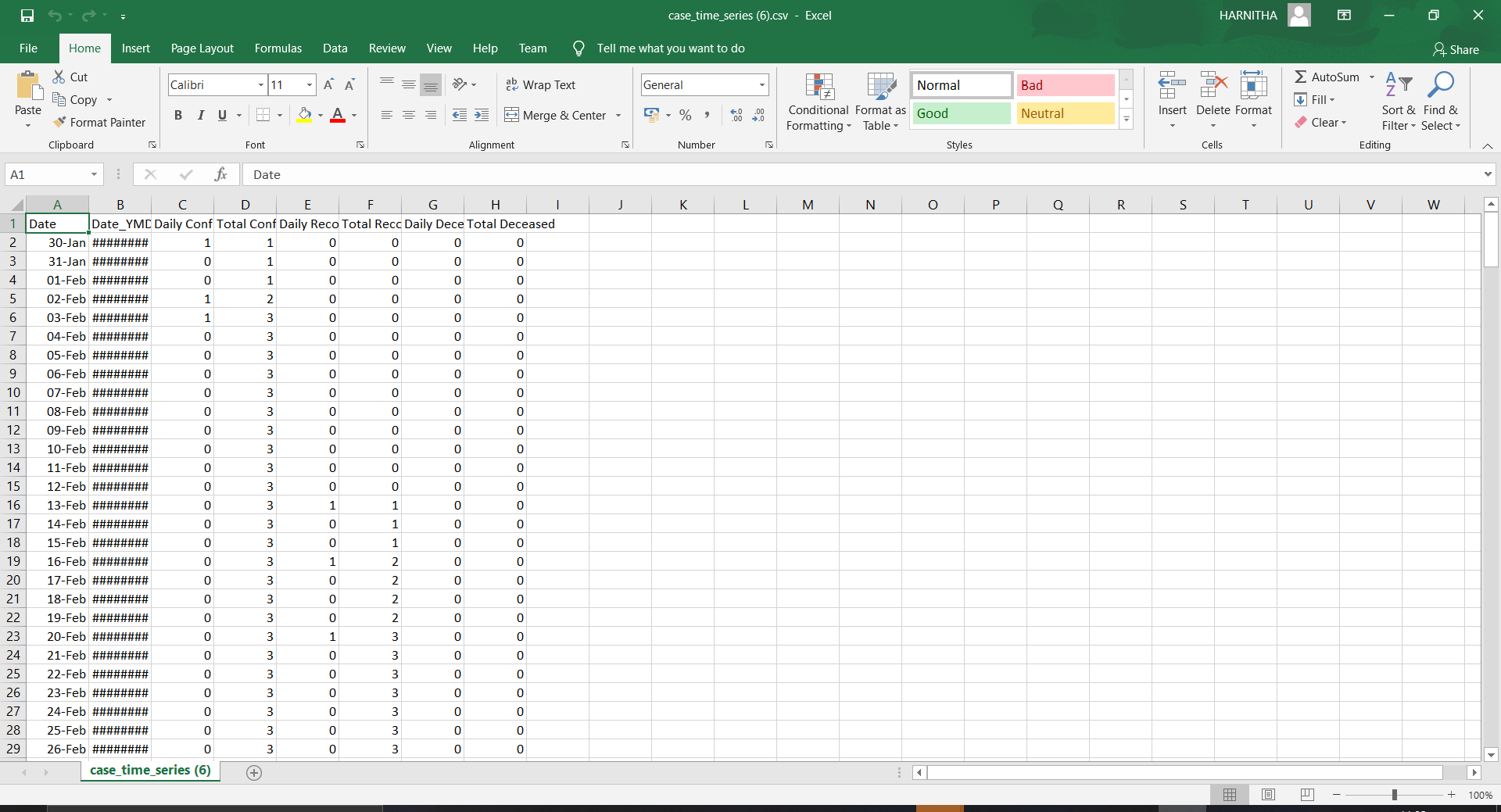


Figure 1: case\_time\_series.csv

The dataset consists of date, date in YMD format, Daily Confirmed, Total Confirmed, Daily Recovered, Total Recovered, Daily Deceased and Total Deceased Cases in time series for INDIA.

For visualization of state-wide records for India: <https://api.covid19india.org/csv/latest/state_wise.csv>

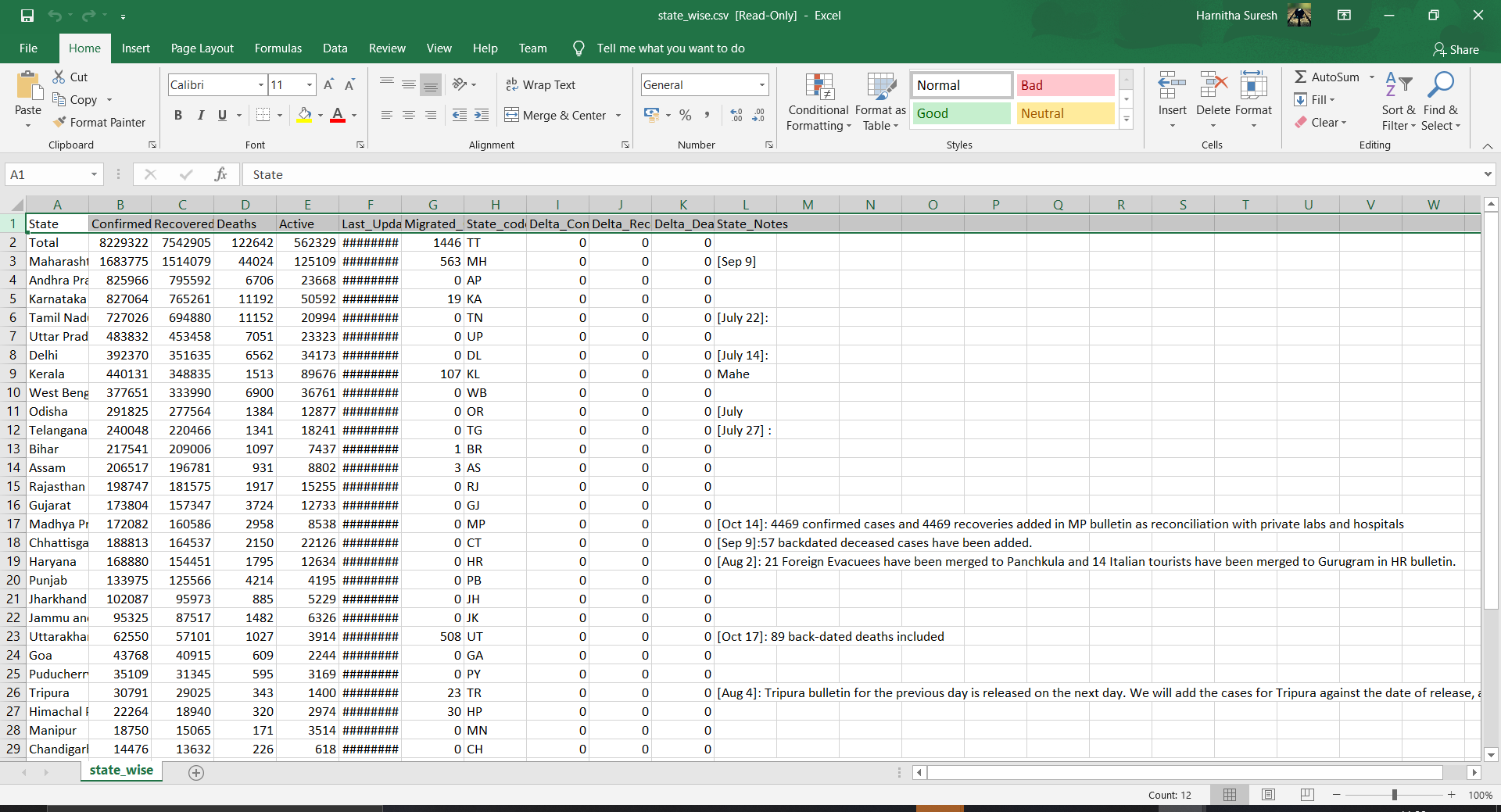


Figure 2: statewise.csv

The dataset consists of State of india , Confirmed cases, Recovered cases, Deaths cases, Active cases, Last\_Updated\_Time, Migrated Cases, State\_code, Difference in Confirmed Cases compared to previous day, Recovered Cases compared to previous day, Difference in Dealth Cases compared to previous day and State\_Notes.

For visualization of state-wide testing records for India: <https://api.covid19india.org/csv/latest/statewise_tested_numbers_data.csv>

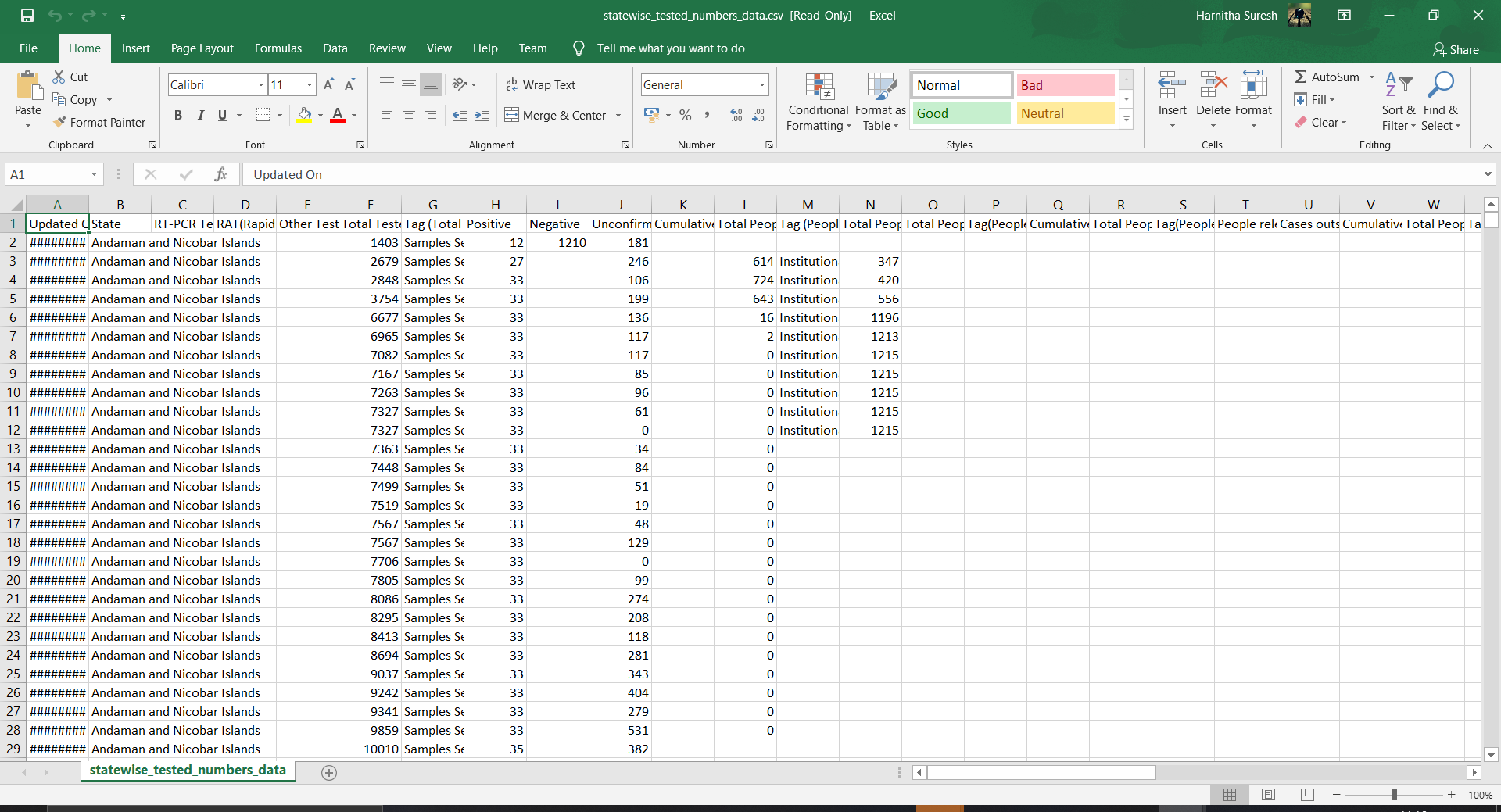


Figure 3: statewise\_tested\_numbers\_data.csv

The dataset consists of Date, State,RT-PCR TestNumbers (Includes TrueNat,CBNAAT,CRISPR), RAT(Rapid Antigen Test)numbers, Other Tests Numbers, Total Tested, Tag (Total Tested), Positive Cases, Negative Cases, Unconfirmed Cases,Cumulative People In Quarantine, Total People Currently in Quarantine, Tag (People in Quarantine), Total People Released From Quarantine, Total People Currently in Quarantine, Tag(People in Quarantine), Cumulative People in Home Isolation, Total People Currently in Isolation, Tag(People in Isolation), People released from Home Isolation, Cases outside, Home Isolation i.e Institutional Isolation,Hospital Etc., Cumulative People in Surveillance ,Total People Currently in Surveillance, Tag(People in Surveillance), People released from Surveillance ,Cumilative People in Observation, People released from Observation, People on ICU Beds, People on Ventilator, People on O2 Beds, Beds Occupied(Normal/Isolation), Total Num ICU Beds, Total Num Ventilators, Total Num of O2 Beds, Total Num Beds (Normal/Isolation), Total PPE Total N95 Masks, COVID Enquiry Calls, Number of Containment Zones, Cumuilative Number of Symptomatic Cases, Cumuilative Number of Asymptomatic Cases, Unknown(Nature of Cases), Tag(Nature of Cases), Male(Confirmed Cases), Female(Confirmed Cases), Transgender(Confirmed Cases), Unknown(Confirmed Cases), Male(Deceased Cases), Female(Deceased Cases), Transgender(Deceased Cases), Unknown(Deceased Cases), Source1, Source2, Source3.

World API :

world\_confirmed :

<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv>

world\_death:

<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv>

world\_recovered:

<https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv>

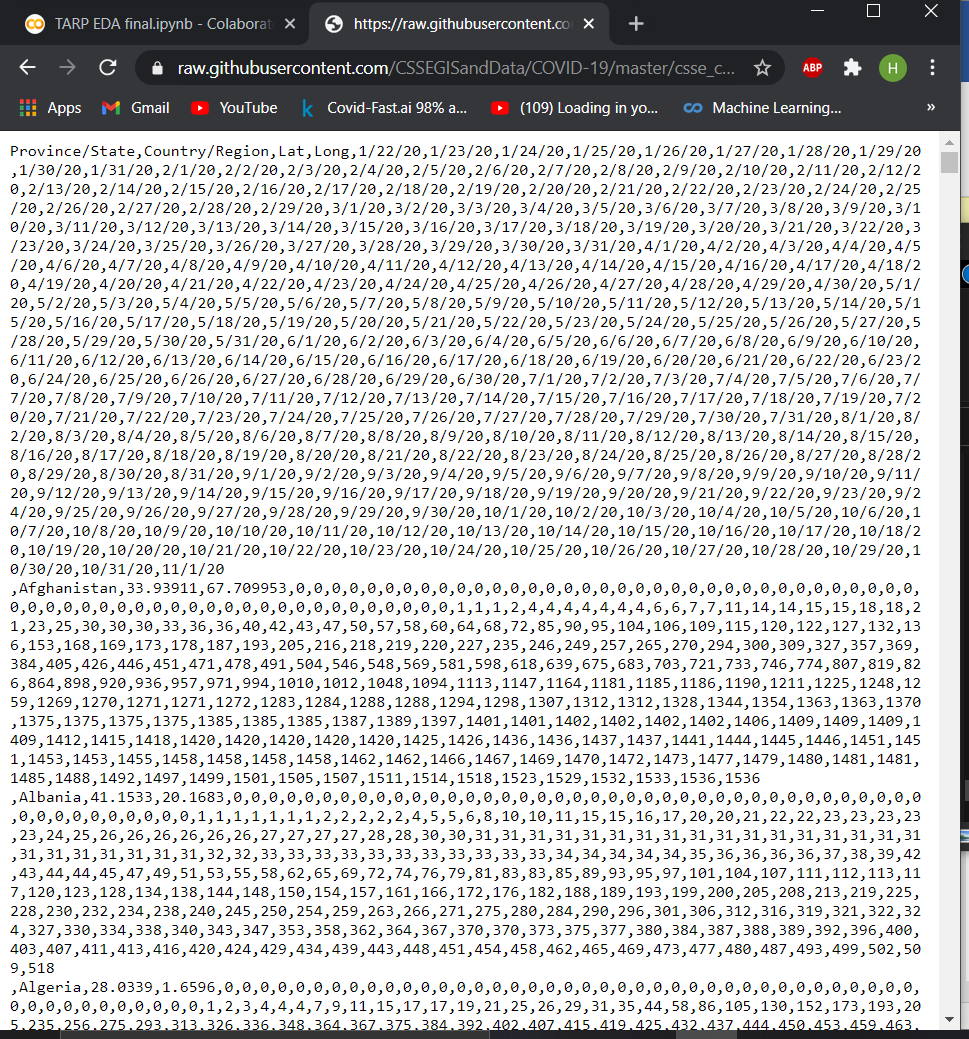
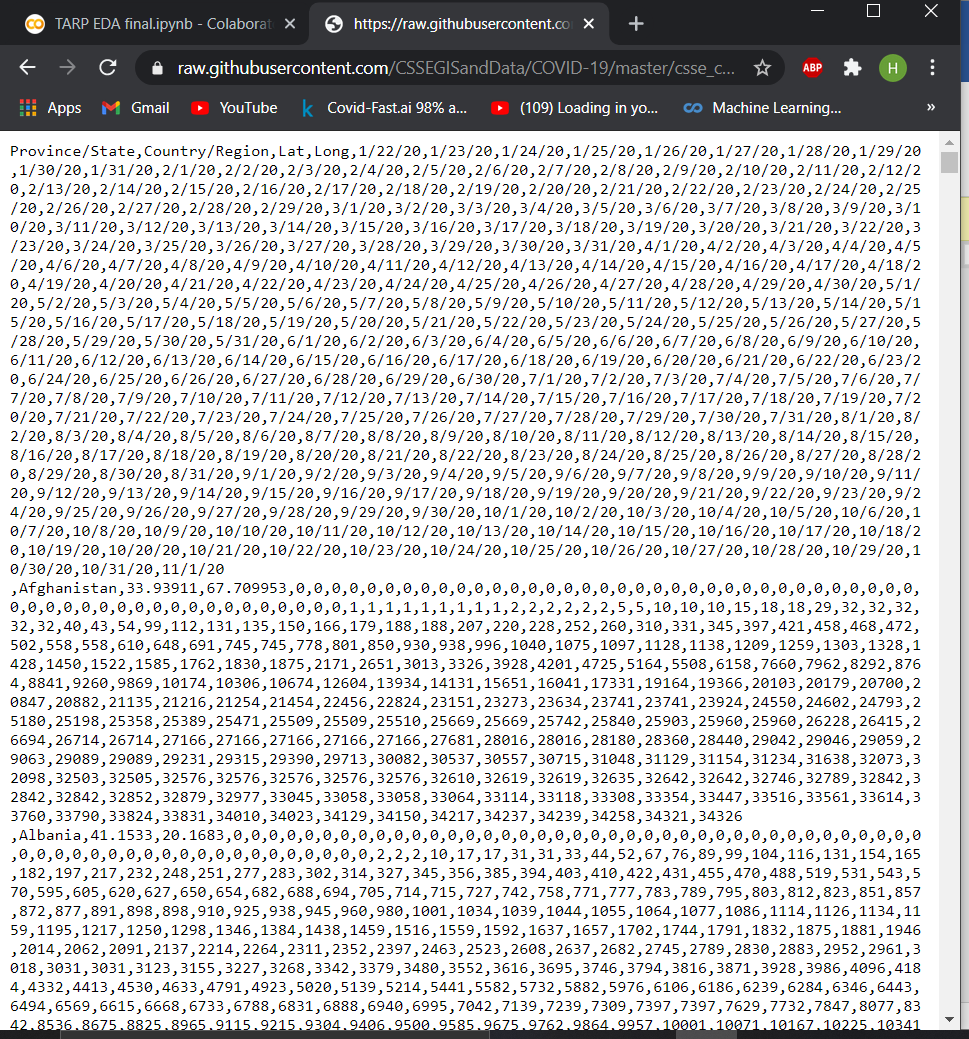
 

Figure 4: world-confirmed Figure 5: world-deaths

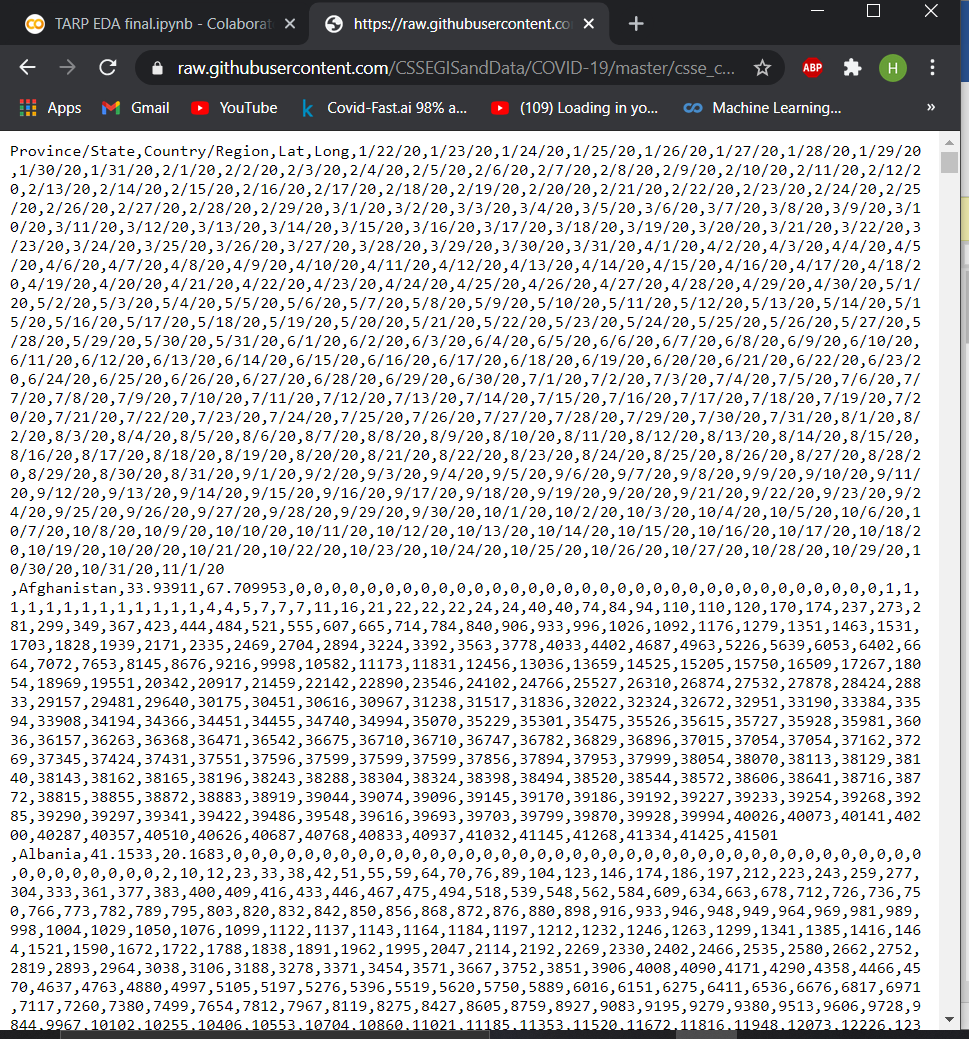


Figure 6: world-recovered

The dataset consists of Province/State name, Country name and the dates with the cases

For world Map:

<https://raw.githubusercontent.com/imdevskp/covid_19_jhu_data_web_scrap_and_cleaning/master/covid_19_clean_complete.csv>

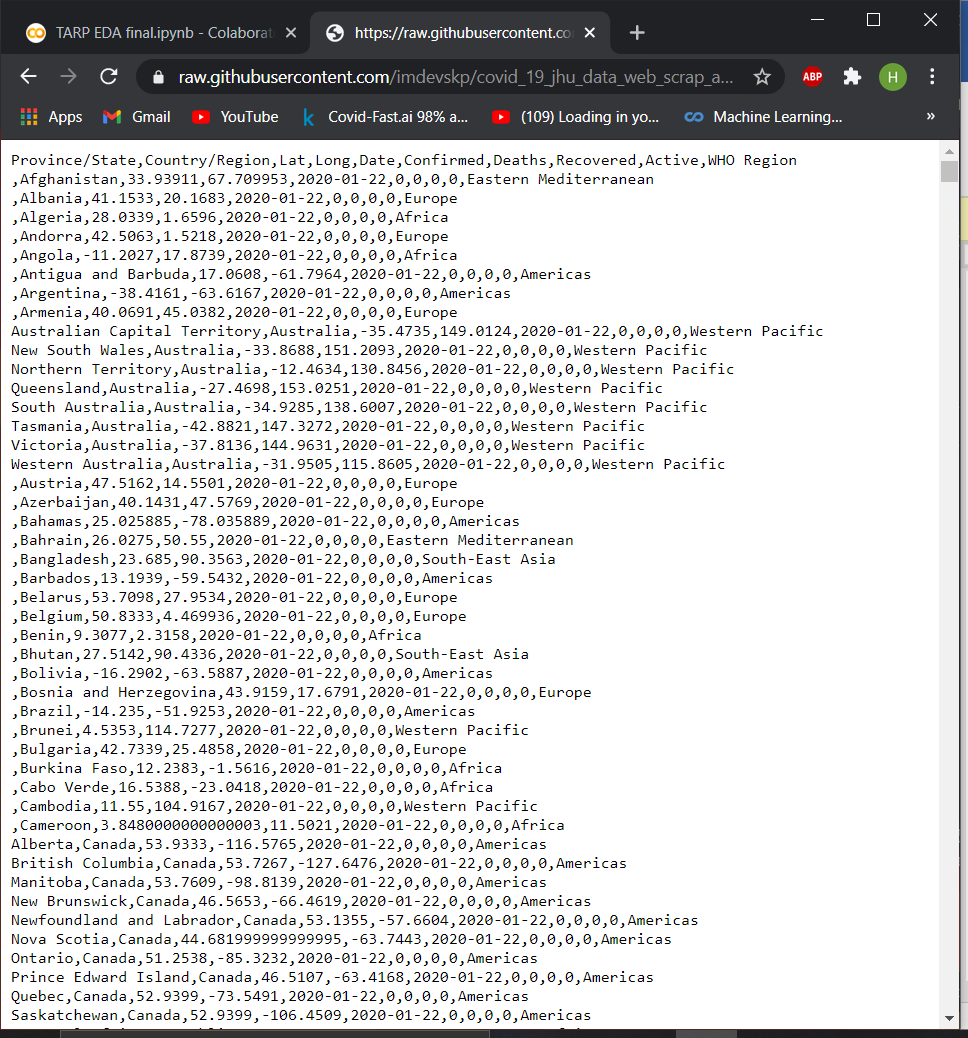


Figure 7: world confirmed cases with location coordinates

The dataset consists of Province/State, Country/Region, Lat, Long, Date, Confirmed, Deaths, Recovered, Active, WHO Region

1. **PROPOSED ALGORITHM WITH FLOWCHART**

**FLOWCHART:**

**FOR DASHBOARD:**

India Totals:

Statewise(India):

State wise Testing(India):

World data:

**PREDICTION:**

The dataset used for this purpose is obtained from covid19india website. The aim is to create a regression based neural network model to predict the total number of confirmed cases in India the next day. The dataset is a time series dataset. In order to convert it into a regression problem. The training dataset is converted from series format to supervised learning format. A multi linear perceptron is trained having 1 hidden layer and 24 input nodes and 1 hidden layer with 500 nodes. The performance measure metric used in this model is root mean square error (RMSE).Validation is done both for training as well as for testing data. Validation on testing data follows an approach known as walk forward validation. Walk forward validation is an approach in which at each iteration where the testing data is fed into the model the result is updated to the training dataset so as to get a better prediction when next data is fed to the model. In this way the model trains even when tested.

**Import libraries**

import the necessary libraries needed to run the code.

Pandas, sklearn.metrics, keras, numpy

**Read dataset**

Read dataset and store it in a dataframe.

**Dataframe to series**

Extract only the Total Confirmed Cases and store it in array format.

**Train and Test dataset split**

Split the dataset into training and Testing dataset.

**Convert Train data from Series to Supervised format**

In order to convert into a regression problem, the series format of training data is converted into supervised format with independent variables and dependent variables.

**Create ML model**

Created a neural network model with 1 hidden layer. Activation functions, Optimizers and loss Functions are defined. Validation testing is defined with 33% split.

**Train model**

The created model is trained using the training data.

**Walk forward validation**

Walk forward validation is an approach in which at each iteration where the testing data is fed into the model the result is updated to the training dataset so as to get a better prediction when next data is fed to the model. In this way the model trains even when tested.

**Calculate RMSE**

Calculate the RMSE score to find the performance of the model.

**Display Results**

Display the loss vs epoch graph to find whether the model overfits or underfits. Display the validation loss and training loss. Display the Prediction Total Cases for the next day in India.

1. **EXPERIMENTS RESULTS**

In this section, experiments are executed to evaluate the performance of proposed technique using COLAB(PYTHON). In this paper, Heroku was used for the deployment of the COVID-19 analysis of India and the World using various graphs.

**DASHBOARD LINK: https://india-world-covid.herokuapp.com/**

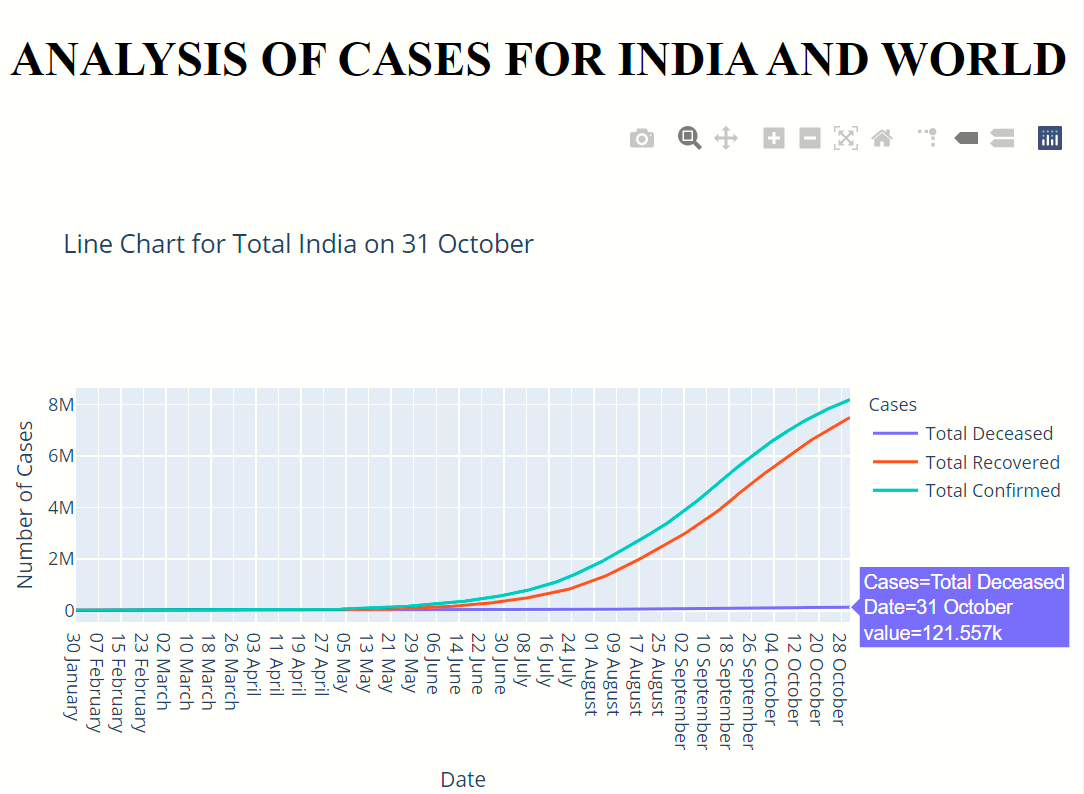
ANALYSIS GRAPHS

Figure 8: Line chart for India on 31 October

The above graph explains that the trend is exponential initially from 13 May , but the slope of the curve is very slowly increasing, confirmed and recovered are approximately in the same rate. and death is very low compared to confirmed and recovered. Thus Indians are more immune and are recovering well.

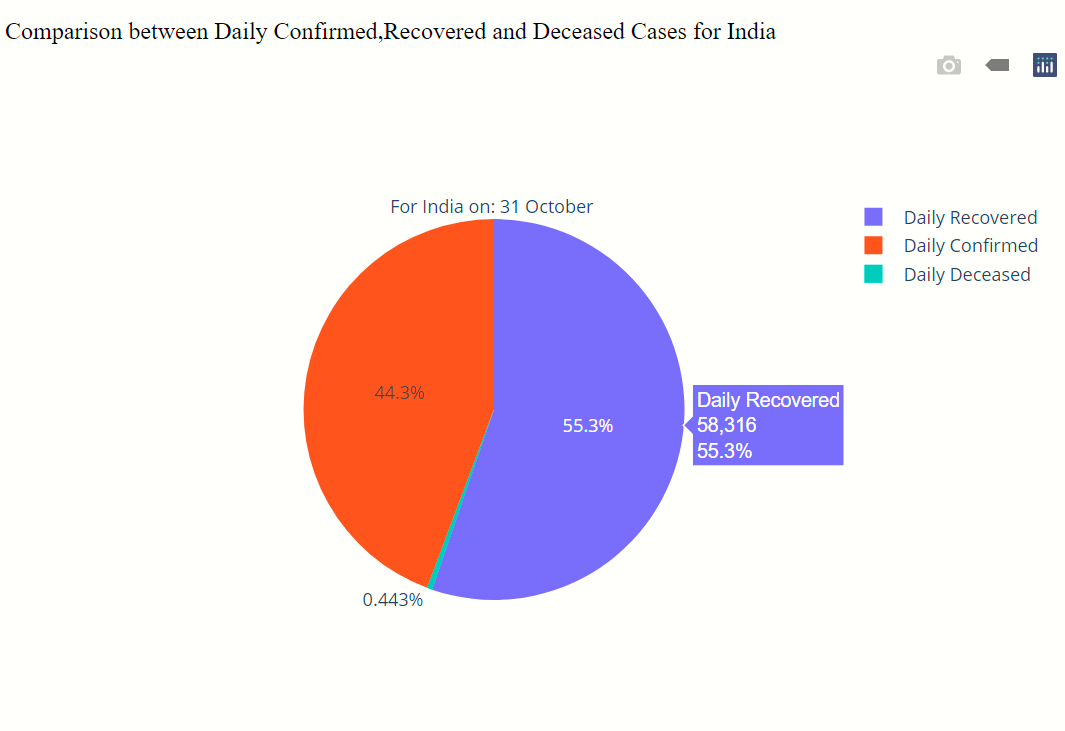


Figure 9: comparison between daily confirmed, recovered and deceased cases for india.

The above graph compares the share of Daily Recovered, Daily Confirmed and Daily Deceased Cases. The graph illustrates that the share of people recovering daily in india is more than the confirmed cases recorded. Thus, this is a positive sign that India is recovering.

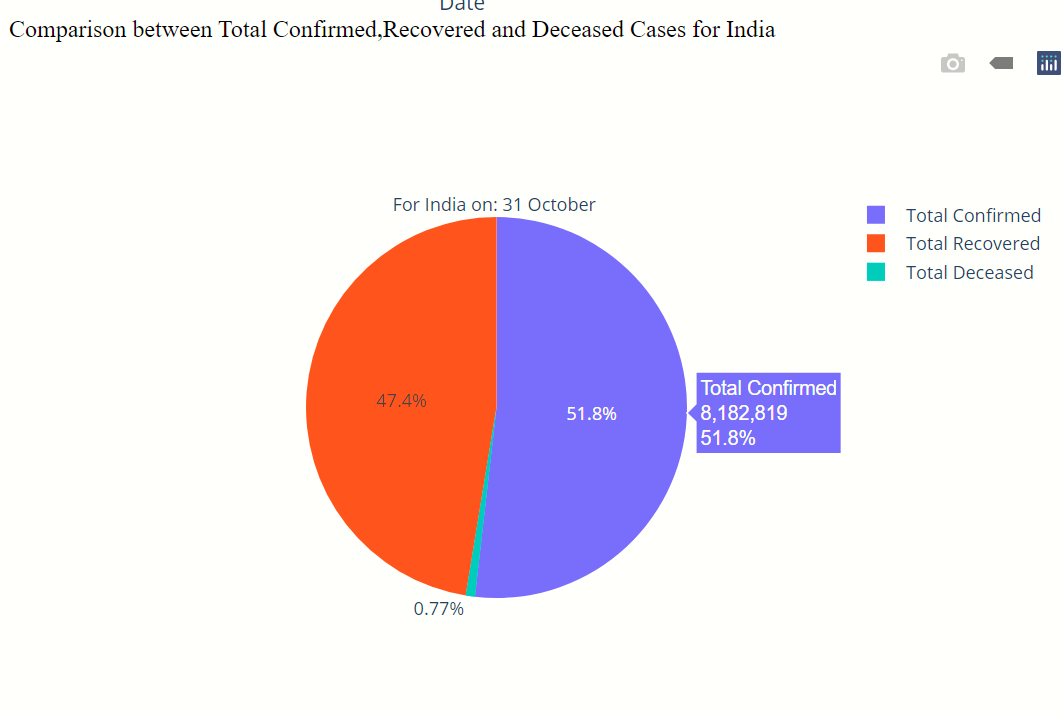


Figure 10: comparison between Total confirmed, recovered and deceased cases for india.

The above graph compares the share of Total Recovered, Total Confirmed and Total Deceased Cases. The graph illustrates that the share of people recovering daily in india is more than the confirmed cases recorded. Thus, this is a positive sign that India is recovering.

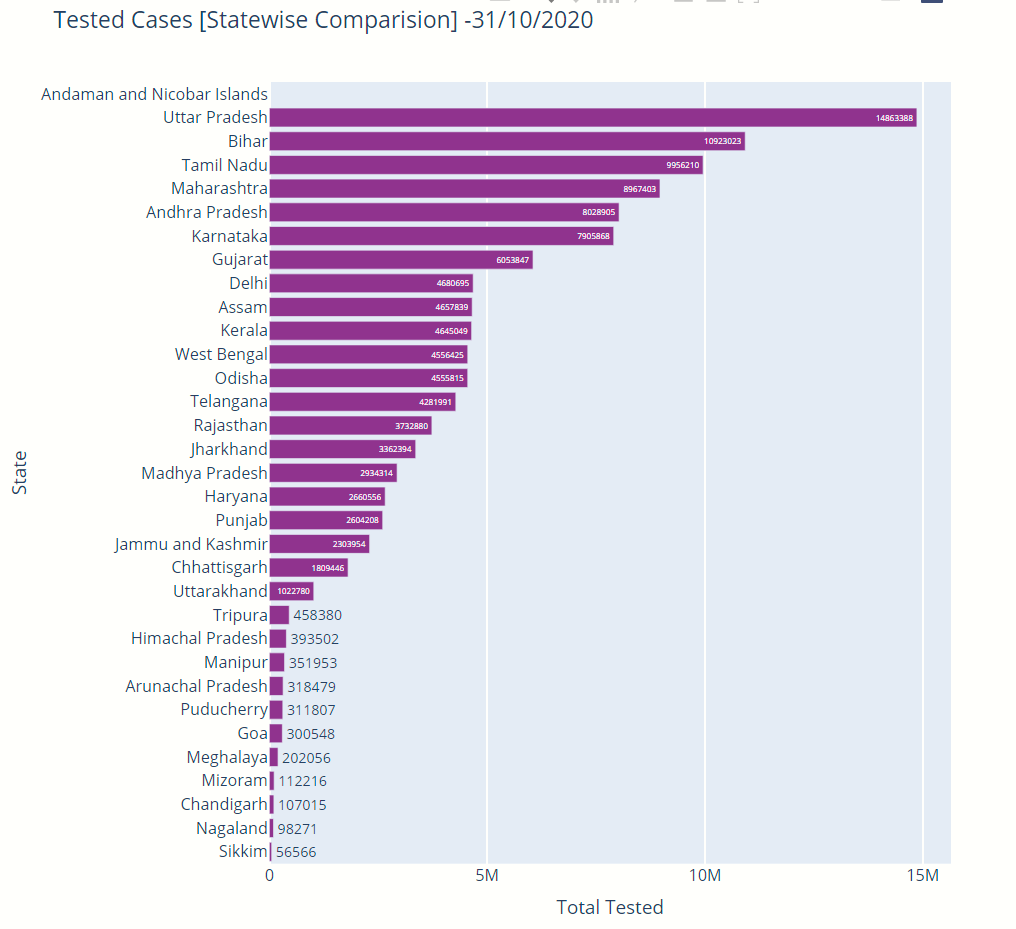


Figure 11: Tested Cases [Statewise comparision]

The above graph illustrates the Total Tested cases in India state wise in decreasing order of the number of cases. From the graph we can see that UP has leading in testing followed by Bihar and TN. Thus UP is performing better in Testing when compared to others and Sikkim is the poorest in Testing.

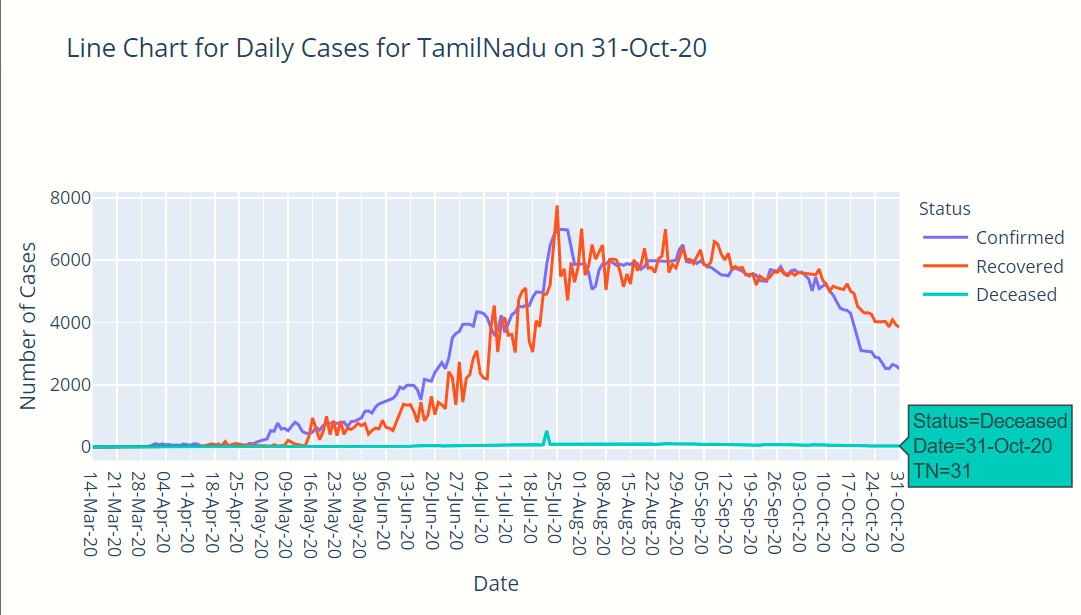


Figure 12: Line chart for Daily Cases for Tamilnadu

The above graph explains that the trend is exponential initially from 2nd May , but the slope of the curve is steep initially and the curves of confirmed and recovered tend to flattern. After 3 oct , the confirmed cases becomes drastically low and the recovered cases is higher than confirmed after 10th oct. Thus, TN is recovering very rapidly.

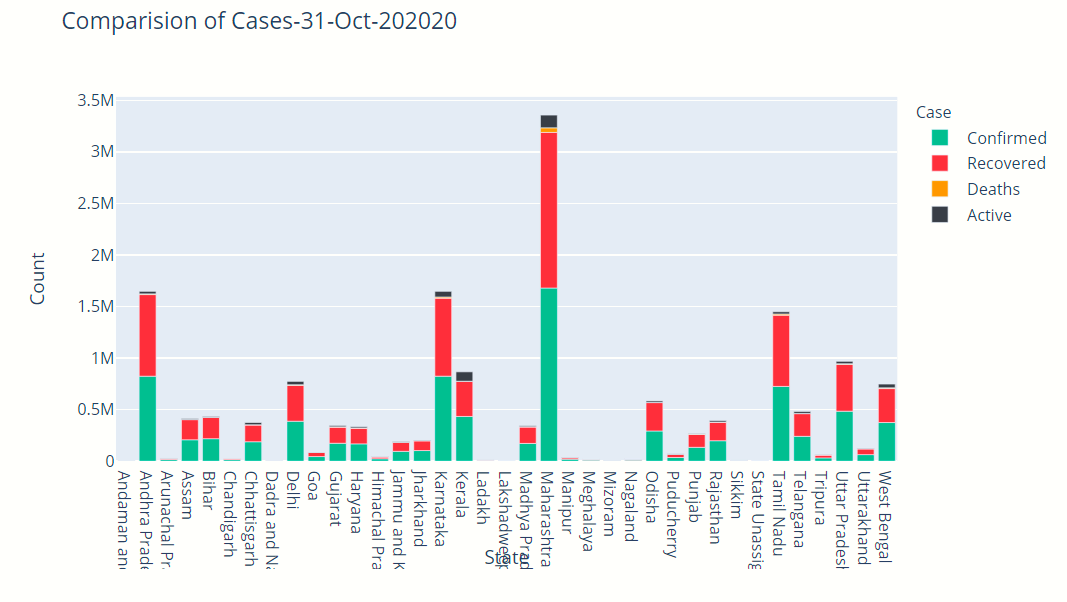


Figure 13: comparision of cases statewise

The graph shows the stacked variation of bar chart in depicts the confirmed, Recovered, Death and Active cases for each State in India. Maharashtra has the highest share of confirmed and recovered than other states and Sikkam, Lakshadweep and Mizoram has the lowest share of the same.

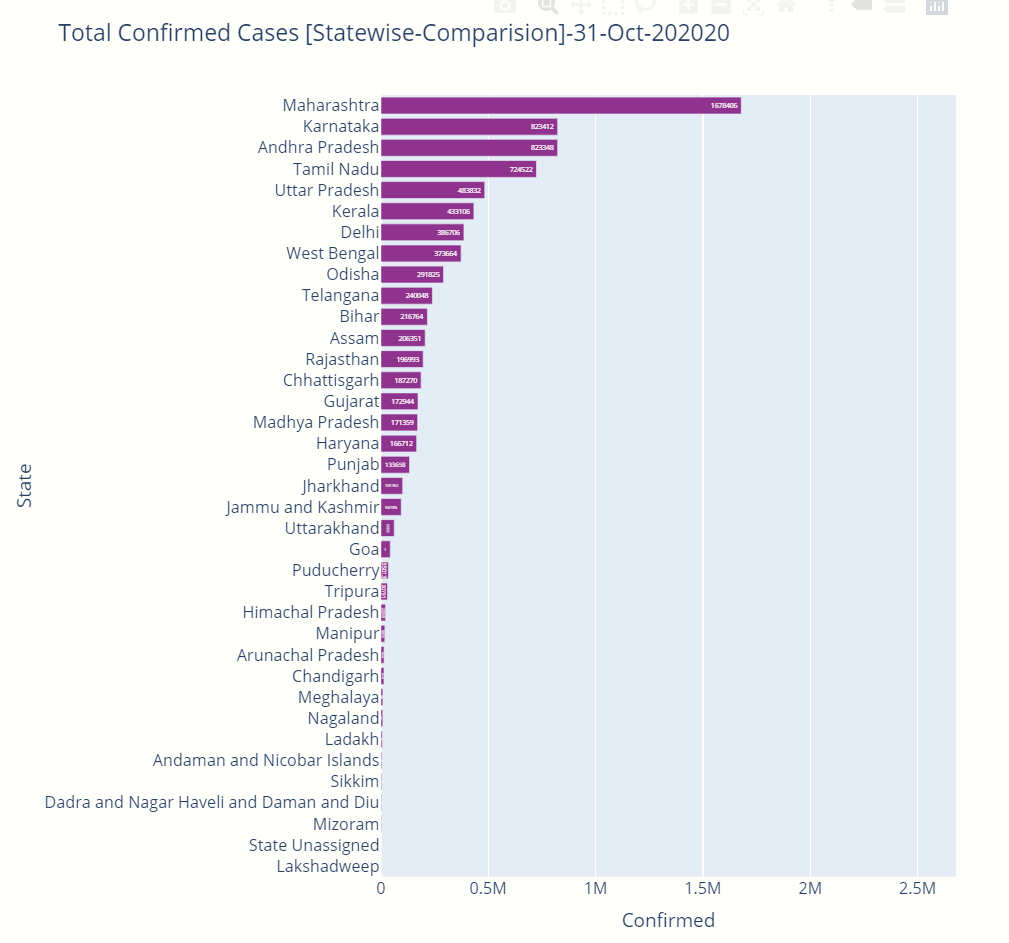


Figure 14: Total Confirmed Cases [State wise Comparison ]

The above graph illustrates the Total Confirmed cases in India state wise in decreasing order of the number of cases. From the graph we can see that Maharashtra has leading in Confirmed Cases followed by Karnataka and Andhra Pradesh. Thus Maharashtra should take more measures to control the spread of the virus

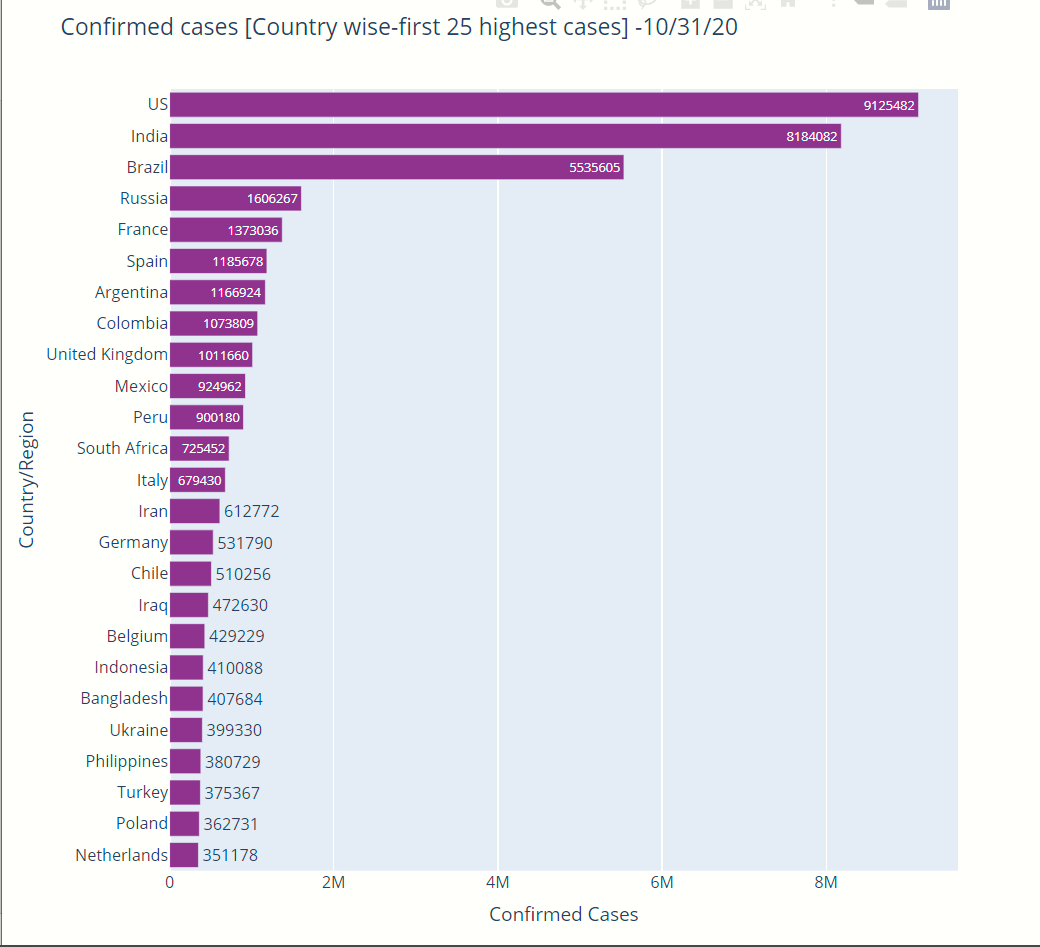


Figure 15: Total Confirmed Cases worldwide

The above graph illustrates the Total Confirmed cases in World in decreasing order of the number of cases. From the graph we can see that US is leading in Confirmed Cases followed by India and Brazil. Thus US should take more measures to control the spread of the virus.

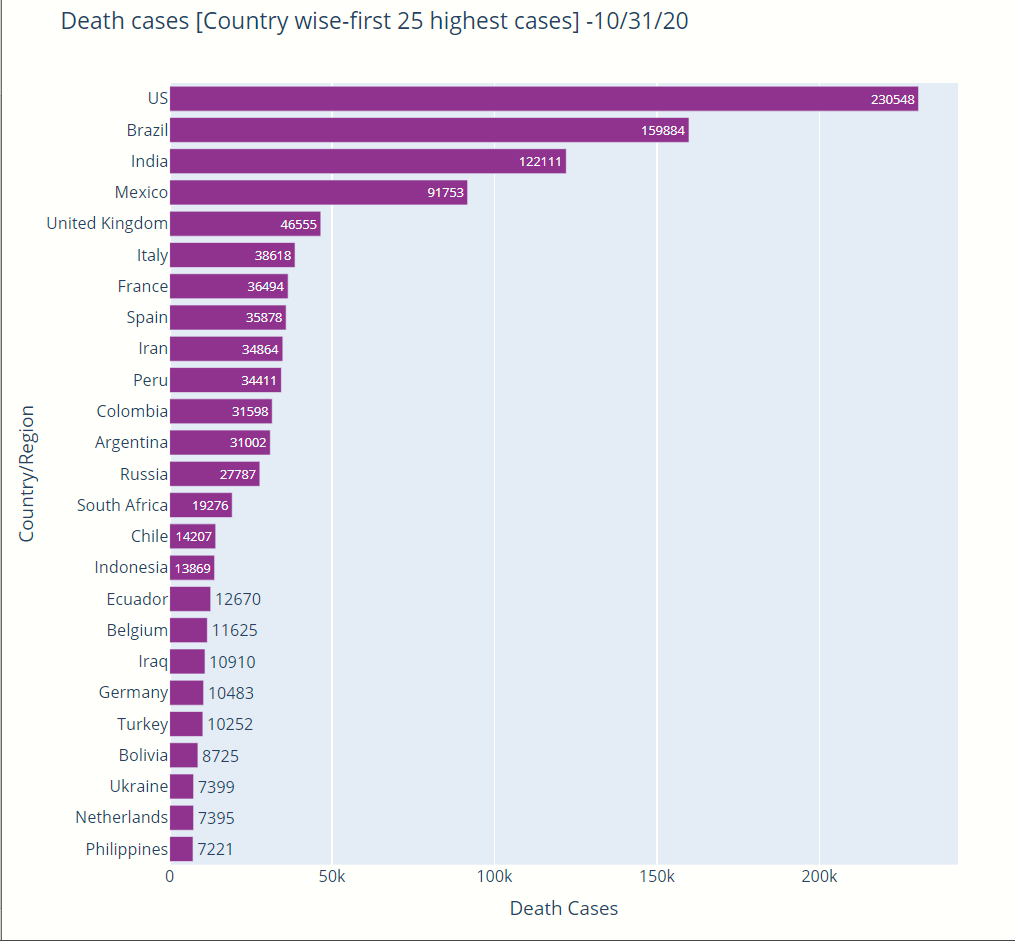


Figure 16: Total Death Cases worldwide

The above graph illustrates the Total Death cases in World in decreasing order of the number of cases. From the graph we can see that US is leading in Death Cases followed by Brazil and India. Thus US should take more measures to control the death cases.

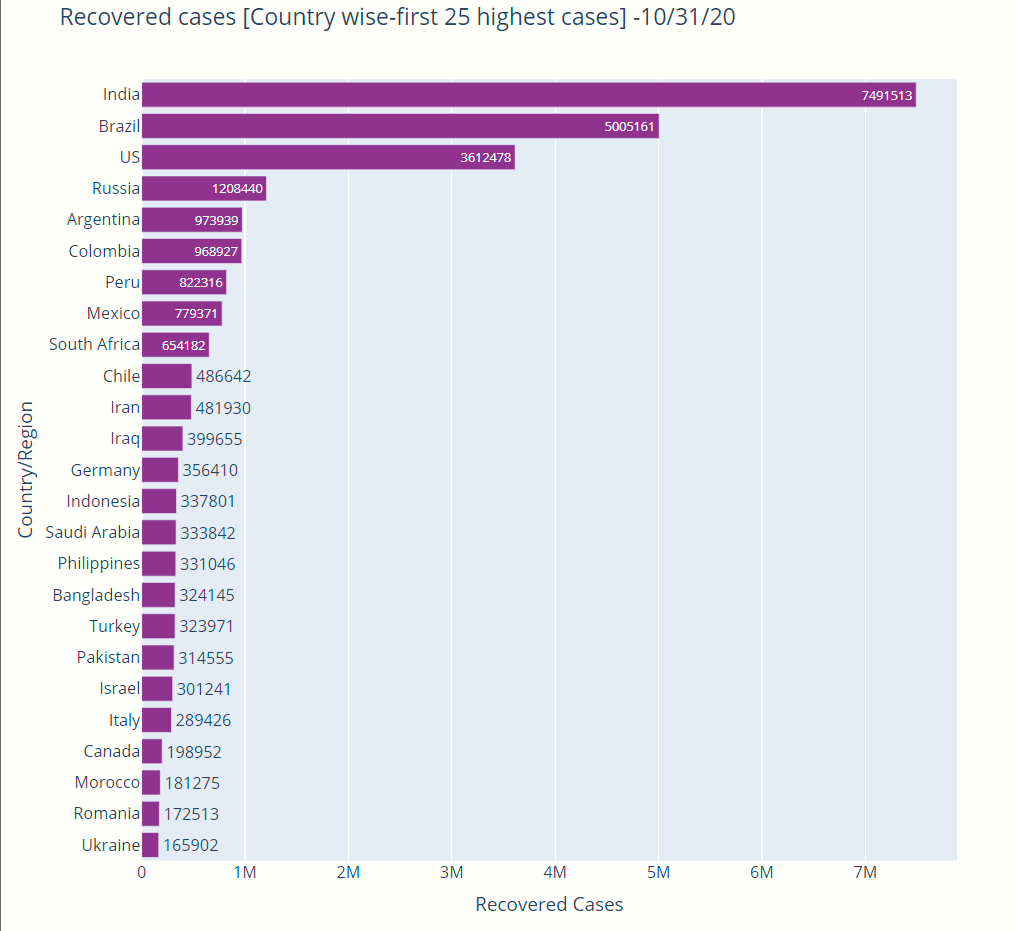
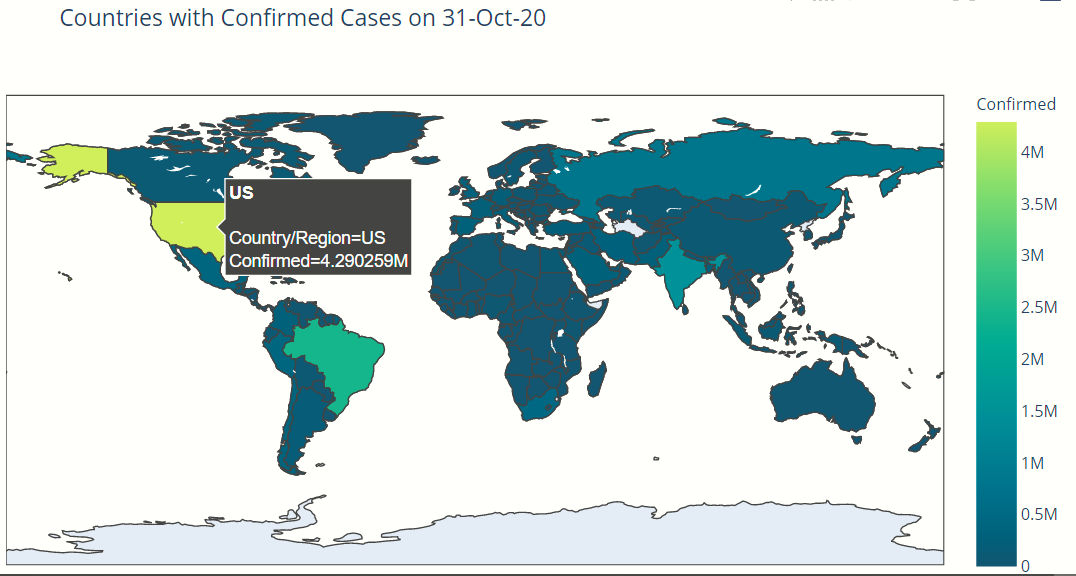


Figure 17: Total Recovered Cases worldwide

The above graph illustrates the Total Recovered cases in World in decreasing order of the number of cases. From the graph we can see that India is leading in Recovered Cases followed by Brazil and US. Thus India is more immune to the virus.

Figure 18: Total Confirmed Cases worldwide -MapView

The graph illiterates the Total Confirmed Cases in a Map View. Scroll and view in one graph.

**PREDICTION MODEL:**





The above table illustrates the comparison of the neural network model trained using different optimizers, hyperparameters and activation functions. The table concludes that the model trained using selu activation function , adam optimizer and hyperparameters with input nodes as 25, number of nodes in hidden layer as 500, the epochs to be 100 and batch size to be 110 performed better than other models with rmse value to be 20,091.

1. **COMPARATIVE STUDY / RESULTS AND DISCUSSION**

**Comparison between our model ( MLP ) and the proposed model in base paper**

config = [24, 500, 200, 110]

input= 24 nodes, hidden layer nodes=500, epochs=200, batch=110

|  |  |  |
| --- | --- | --- |
| Actual Cases | Multi Layer Perceptron | Base Paper Result |
| 8447 | 10229 | 8301 |
| 9352 | 11654 | 9254 |
| 10815 | 13166 | 9960 |
| 11933 | 14967 | 11600 |
| 12759 | 17000 | 12953 |
| 13835 | 19170 | 13975 |
| 17792 | 21382 | 17490 |
|  |  |  |

config = [8, 500, 300, 150]

input= 8 nodes, hidden layer nodes=500, epochs=300, batch=150

|  |  |  |
| --- | --- | --- |
| Actual Cases | Multi Layer Perceptron | Base Paper Result |
| 8447 | 9253 | 8301 |
| 9352 | 10297 | 9254 |
| 10815 | 11689 | 9960 |
| 11933 | 13331 | 11600 |
| 12759 | 14939 | 12953 |
| 13835 | 16061 | 13975 |
| 17792 | 17444 | 17490 |

The above tables illustrate the comparison between the prediction cases from the proposed model from this project and the model depicted in the base research paper (polynomial regression). Thus the table concludes that our model performs much similar to the base paper model when the number of input nodes= 8 nodes, hidden layer nodes=500, epochs=300, batch=150 with activation function as selu and optimizer as adam.

Neural networks produce better results when more data is fed to train. Since the above comparison uses only limited dataset from 1st March 2020 to 11th April 2020 it is expected to perform lower. But the results say that the proposed model does predict near equal values to the base paper results with small dataset. Thus, the proposed model in this paper will surely produce better results than base paper model when fed with more data.

1. **CONCLUSION AND FUTURE WORK**

The project aimed to visualize the current scenario of covid world-wise, country-wise(for India) and statewise (Tamilnadu) as part of Exploratory Data Analysis in real-time and also provide Prediction of cases in the next day for India with analysis of different models trained with different optimizers and activation functions. RMSE was used to analyse and compare different models. Validation Testing was performed to check whether the model overfits or underfits. The study revealed that model trained using SELU activation function and ADAM optimisers with input nodes, hidden nodes, epochs and batch to be 24,500,100,110 respectively performs better than others.

As part of the future scope, the project can be further tried out with more complex neural networks with more hidden layers and other components.

1. **REFERENCES**
2. Parbat, Debanjan. "A Python based Support Vector Regression Model for prediction of Covid19 cases in India." *Chaos, Solitons & Fractals* (2020): 109942.
3. Dey, Samrat K., et al. "Analyzing the epidemiological outbreak of COVID‐19: A visual exploratory data analysis approach." *Journal of medical virology* 92.6 (2020): 632-638.
4. Mittal, S. An Exploratory Data Analysis of COVID-19 in India. *International Journal of Engineering Research & Technology (IJERT), ISSN*, 2278-0181.
5. Yadav, R. S. (2020). Data analysis of COVID-2019 epidemic using machine learning methods: a case study of India. *International Journal of Information Technology*, 1-10.
6. Adeniyi, M. O., Ekum, M. I., Iluno, C., & Oke, S. I. (2020). Dynamic Model of COVID-19 Disease with Exploratory Data Analysis. *Scientific African*, e00477.
7. Yun, H., Sun, Z., Wu, J., Tang, A., Hu, M., & Xiang, Z. (2020). Laboratory data analysis of novel coronavirus (COVID-19) screening in 2510 patients. *Clinica Chimica Acta*.
8. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, 138762.
9. Wang, L., Li, J., Guo, S., Xie, N., Yao, L., Cao, Y., ... & Ji, J. (2020). Real-time estimation and prediction of mortality caused by COVID-19 with patient information based algorithm. *Science of the Total Environment*, 138394.
10. Chakraborty, T., & Ghosh, I. (2020). Real-time forecasts and risk assessment of novel coronavirus (COVID-19) cases: A data-driven analysis. *Chaos, Solitons & Fractals*, 109850.
11. Zhang, S., Diao, M., Yu, W., Pei, L., Lin, Z., & Chen, D. (2020). Estimation of the reproductive number of novel coronavirus (COVID-19) and the probable outbreak size on the Diamond Princess cruise ship: A data-driven analysis. *International journal of infectious diseases*, *93*, 201-204.
12. Velásquez, R. M. A., & Lara, J. V. M. (2020). Forecast and evaluation of COVID-19 spreading in USA with Reduced-space Gaussian process regression. *Chaos, Solitons & Fractals*, 109924.
13. Mandal, M., Jana, S., Nandi, S. K., Khatua, A., Adak, S., & Kar, T. K. (2020). A model based study on the dynamics of COVID-19: Prediction and control. *Chaos, Solitons & Fractals*, 109889.
14. Chowdhury, R., Heng, K., Shawon, M. S. R., Goh, G., Okonofua, D., Ochoa-Rosales, C., ... & Shahzad, S. (2020). Dynamic interventions to control COVID-19 pandemic: a multivariate prediction modelling study comparing 16 worldwide countries. *European journal of epidemiology*, *35*(5), 389-399.
15. Ayyoubzadeh, S. M., Ayyoubzadeh, S. M., Zahedi, H., Ahmadi, M., & Kalhori, S. R. N. (2020). Predicting COVID-19 incidence through analysis of google trends data in iran: data mining and deep learning pilot study. JMIR Public Health and Surveillance, 6(2), e18828.
16. Fayyoumi, E., Idwan, S., & AboShindi, H. (2020). Machine Learning and Statistical Modelling for Prediction of Novel COVID-19 Patients Case Study: Jordan. Machine Learning, 11(5).
17. X Joseph ,G., Ranjeesh, RC., (2020) .Comparison of Regression Models on Covid-19 Cases . International Research Journal of Engineering and Technology (IRJET)
18. Pinter, G., Felde, I., Mosavi, A., Ghamisi, P., & Gloaguen, R. (2020). COVID-19 Pandemic Prediction for Hungary; a Hybrid Machine Learning Approach. Mathematics, 8(6), 890.
19. Prakash, K. B., Imambi, S. S., Ismail, M., Kumar, T. P., & Pawan, Y. N. (2020). Analysis, Prediction and Evaluation of COVID-19 Datasets using Machine Learning Algorithms. International Journal, 8(5).
20. Ardabili, S. F., Mosavi, A., Ghamisi, P., Ferdinand, F., Varkonyi-Koczy, A. R., Reuter, U., ... & Atkinson, P. M. (2020). Covid-19 outbreak prediction with machine learning. Available at SSRN 3580188.
21. Salman, F. M., Abu-Naser, S. S., Alajrami, E., Abu-Nasser, B. S., & Alashqar, B. A. (2020). Covid-19 detection using artificial intelligence.