e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

PREDICTING THE GROWTH OF COVID PANDEMICUSING MACHINE LEARNING

Radhika Tikone*1, Dipali Salunke*2, Bhakti Devre*3,
Rajnandini Somvamshi*4, Mrs. Shital. A. Karande*5
*1,2,3,4,5Department Of Computer Engineering, Bharati Vidyapeeth's College Of Engineering For Women, Savitribai Phule Pune University, India.

ABSTRACT

Nowadays Machine learning has proved its significance in solving various problems of prediction. This could help in reducing the stress on health care systems and administrations by helping them plan better. In this paper the datasets used are obtained from the John Hopkins University's [10] publicly available datasets to develop a forecasting model of COVID-19 outbreak. We have incorporated data-driven estimations and time series analysis to predict the trends in coming days such as the number of cases confirmed positive, number of deaths caused by the virus and number of people recovered from the novel coronavirus. The ML models have been used long in many application domains which needed identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. In particular, four standard forecasting models, such as linear regression (LR) [1], polynomial regression (PR) [2] and least absolute shrinkage and selection operator (LASSO [3] and support vector machine (SVM) [4] have been used in this study to forecast the threatening factors of COVID-19. Three types of predictions are made by each of the models, suchas the total number of newly infected cases, total number of deaths, and total number of recoveries in the next 10 days.

Keywords: COVID-19, Data Analysis, Machine Learning, Supervised Learning, Regression,Prediction, Accuracy.

I. INTRODUCTION

The World has been affected by a highly contagious virus called the Corona virus. This virus firstly found in Wuhan, Hubei province of China during December 2019. This virus quickly spread to almost every country within a span of 3 months causing over 400,000 deaths with more than 9 million people affected globally. This virus has caused very distressing times across all the countries and significant disruptions in global economies.

Several measure actions were taken by government such as quarantining people to stop the spread of the virus. Coronavirus being a contagious and infectious disease like the flu with certain growthpatterns, such patterns are noted to be non-linear and dynamic in nature. Data is Dynamic in natureas the cases might differ based on the seasons, populations etc.

Machine learning (ML) has proved itself as a prominent field of study over the last decade by solving many very complex and sophisticated real-world problems. The application areas included almost all the real-world domains such as healthcare, autonomous vehicle (AV), business applications, natural language processing (NLP), intelligent robots, gaming, climate modeling, voice, and image processing. One of the most significant areas of ML Forecasting, numerous standard ML algorithms have been used in this area to guide the future course of actions needed in many application areas including weather forecasting, disease forecasting, stock market forecasting as well as disease prognosis. Various regression and neural network models have wide applicability in predicting the conditions of patients in the future with a specific disease. There arelots of studies performed for the prediction of different diseases using machine learning techniques such as coronary artery disease, cardiovascular disease prediction, and breast cancer prediction. In our project we are using ML technology in which algorithms like polynomial regression and linear regression had been used for predicting total, recovered and death cases in coming 10 days.

The main transmitting methods of COVID-19 are breathing and close contact. Moreover, many governments have issued rigid regulations based on social distance and wearing masks in closed environment. In general, these regulations help the governments sector to control the exponential growth of infected people. However, up to date, COVID-19 still present and a huge number of researchers investigate the best approaches to deal with this virus either preparing new vaccines orissuing new regulations. Detecting the existence of COVID-19 is a challenging task. Machine learning provides excellent methods to detect it. Moreover, Data Mining (DM) can

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

be used to extract meaningful information from big data and perform complex tasks to discover hidden knowledge, especially for medical datasets. In general, there are many methods that can be used to determine the existence of COVID-19, such as Linear Regression (LR) [1], PolynomialRegression [2] and Lasso Regression [3].

II. LITERATURE SURVEY

To diagnosis the disease researchers have developed many approaches like machine learning for diagnosing the disease. Machine learning, in addition to clinical approaches, assists in disease diagnosis by using textual data and images [4]. In addition to diagnosis, it is essential to keep trackof the growth of COVID-19. The number of infected cases that can be estimated in the coming days can be predicted using a genetic algorithm, a logistic growth regression model, and a sigmoid model. In addition to that, the existing approaches is applicable for a particular region and gives less accuracy in forecasting the future cases globally and analyses only small amount of data. In order to improve accuracy in prediction and to analyze large dataset, supervised machine learningalgorithms especially time series forecasting algorithm called Holt's Winter algorithm is preferredover other algorithms.

Linear Regression (LR) [1], Polynomial Regression [2] Lasso Regression [3] and Support Vector (SVR) [4] Supervised Machine Learning algorithms can be used in statistical analysis to forecast predicted numerical values. Time series Forecasting algorithm, an example of classifier supervisedmachine learning algorithm aids in forecasting values of a time series at multiple time points in the future.

To train the model for predicting the total number of positive cases in the world in the coming days, the proposed work employs both regression and time series algorithms. The dataset, which includes the total number of confirmed, recovered, and death cases around the world, is taken, preprocessed to minimize analysis time, and given to the model as training and testing information.

As a result, the current research focuses on using the clinical text COVID - 19 dataset to forecast an improvement in the number of global COVID - 19 positive cases in the future (Fig. 1).

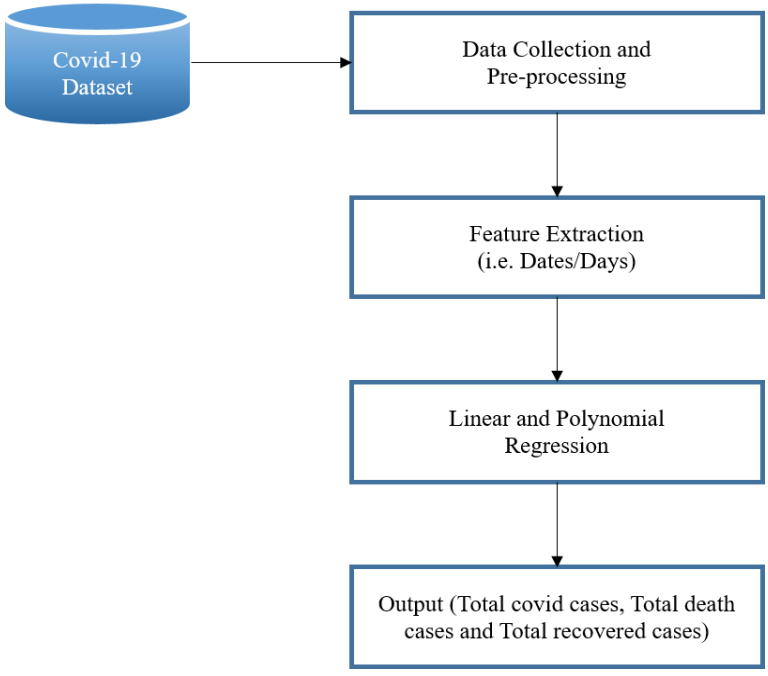

Fig 1: Flow Diagram

The above diagram represents the flow diagram of Predicting the number of future infected casesusing machine learning algorithms. There are different phases namely pre-processing, feature extraction and prediction for detecting forest smokes. Initially, the dataset which is the form of rows and columns that are gathered from different sources with different data, are converted to a CSV (Comma-Separated Values) file. It is then passed to pre-processing where the dataset with different attributes is being converted to standardized dataset containing the date, state/province, country, number of confirmed, recovered and death cases. Then, feature

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

extraction is done by LR

[1] and SVR [4]. Feature extraction includes extracting necessary columns or data to be given as input to the model. Then, prediction is made by applying respective algorithms to the model and future confirmed cases are being obtained.

2.1 Comparative Analysis of Literature

Author	Methodology	Features	Challenges/ Futurescope
Sina F. Ardabili,Amir Mosavi	Machine Learning[5]	No. of infection andnew cases	Fluctuations in theflow of cases
Quoc-Viet Pham,Dinh C. Nguyen	AI, Big Data, DeepLearning, Data Analytics [6]	No of global covid 19 cases, confirmed ,dead and active cases	1. Regulation 2. Lack of standard dataset 3. Privacy andsecurity
Furquan, Rustam,Aijaz Reshi	Supervised Machine Learning [7]	Positive, Death and recovered covid cases	High speed network required
L. J. Muhammad,Ebrahem A. Alegehyne	Supervised Machine Learning [8]	Patient count	NA
Sivaramakrishnan Rajendar	Decision Tree, Machin learning [9]	Diabetic patientcount	Exceeded Symptoms showing major no of Diseases to be predict .

2.2 Performance Evaluation of Various Parameters

Sr. No.	Research Paper's	Performance Measures					
		Type of Input Data set	Preprocessing Technique	Feature Extraction	Classifier	Accuracy	Date
1	Covid 19 outbreak prediction with ML [5]	No of infection	ML	GA , PSO , GWO	Logistic Regression model , microbial growth model	91.7%	April 22, 2020
2.	AI and Big Data for coronavirus pandemic a survey on the state of the arts [6]	No of global covid 19 cases	AI , Big data (Consists of multiple analysis phases through big data)	NA	Uses CNN classifier	89.03%	-
3	Covid 19future forecasting using Supervised Machine Learning Models [7]	Covid cases	Supervised ML	NA	LR, LASSO, SVM	88.76%	4 May, 2020

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

4	Supervised Machine Learning Models for Prediction of COVID-19 [8]	Covid-19 patients dataset	ML	NA	SVM, Naïve Bayes, LR,Decision Tree, ANN	94.99%	Oct 26,2020
5	Comparative analysis of classifier models for the early prediction oftype2 diabetes [9]	Datasetof diabetic patients (968 female record)	ML	NA	SVM, Logistic Regression , RF	86.70%	May2020

III. METHODS AND MODEL

3.1 Supervised machine learning models

The regressor is used for the regression model. For the development of predictive models, Regression techniques and classification algorithms study method used here. In this COVID 19 prediction analysis, three regression models are used:

- Linear Regression
- LASSO Regression
- Support Vector Machine

A. Linear Regression

The aim class concentrates on individual regression simulation characteristics. It may also be usedto define and model the relationship between independent variables and dependent. The most useful computer method for mathematical analysis of the machine learn is linear regression type regression simulation. A linear regression observation relies on two values, one on the dependence and one on the isolation. Linear Regression defines a linear relation between these variables' dependency and independence. Two variables (x, y) are necessary for the linear regression search.This equation indicates how y is associated with ×, which is called regression.

$$y = \beta 0 + \beta 1x + \varepsilon \quad (1)$$

$$E(y) = \beta 0 + \beta 1x \quad (2)$$

This is the linear term for error regression. This error term takes into account the variability between × and y, β0 is the y-intercept and ß1 is the pitch.

A class mark is specified in the input data set for the purpose of the model × training of the linear regression in the context of machine study. The aim is to find the optimum values for β0 (intercept) and β1 (coefficient) to get the best regression line. The difference between the actual values and the values predicted should be minimum to make sure that this minimising problem is presented:

$$\text{Minimize} \frac{1}{2n}\sum i=(\text{pred}_i - y_i)^2 \quad (3)$$

here, g, which is the mean root square of the expected value for y (pred_i) and y (y_i), n is the cumulative number of data points. g is called the cost function.

B. Lasso Regression

LASSO is a regression model that is part of a linear regression method that uses shrinkage. Shrinking means reducing the extreme data sample values to the key values in this case. This strengthens and stabilises LASSO and reduces the error by the shrinking process. For multi-linear situations, LASSO is considered a more fitting model. LASSO thus makes the regression smoother in terms of the amount of functions it uses. It uses a form of regularisation to penalise additional tasks automatically.

However, the LASSO regression tries one at a time, because it does not add importance of zero if the new function would not boost the penalty term's fit with that function. The power ofregularisation is therefore to

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

automatically pick for us by adding the penalty for extra functions. Therefore, in this case of regularisation the models become sparse with few coefficients as the method removes values are zero. This regression LASSO acts to reduce the coefficient, which can be known by the square residual β slope), where, β slope is a concept of penalty.

C. Support Vector Machine

The SVM is a type of ML managed algorithm for reverse and regression classification. The SVM regression [4] depends on a variety of statistical functions as a non-parametric technician. The set of the kernel function converts data input into the form you like. In order to overcome regression problems using a linear function, SVM maps the vector(x) input(s) in the n-dimensional space called the function space(z) when dealing with non-linear regression problems. After liner's regression is implemented in the space, non-linear mapping techniques are used for this mapping.Put the concept into an ML context using a number of observations with y from a multivariate training dataset (x) to N. The goal is therefore that the value of f (x) with (βrβ) as the minimum standard values is found as flat as possible. The dilemma then blends in with the minimization function. If the value of all residues is not greater than p, as in the following equation:

Predictions are rendered based on data from previous times of exponential family smoothing techniques. As previous data findings get older, their effect declines exponentially. The weights are therefore geometrically reduced to the different lag values. Particularly for univariate is a timeseries provision. Ft 1 is the preview value of the previous prediction for the present period (Ft) inthe ES. The prediction is as follows. Ft 1 is the real value in the preceding time frame, in which Ft1 is the expected value in the prediction.

3.2 System Architecture

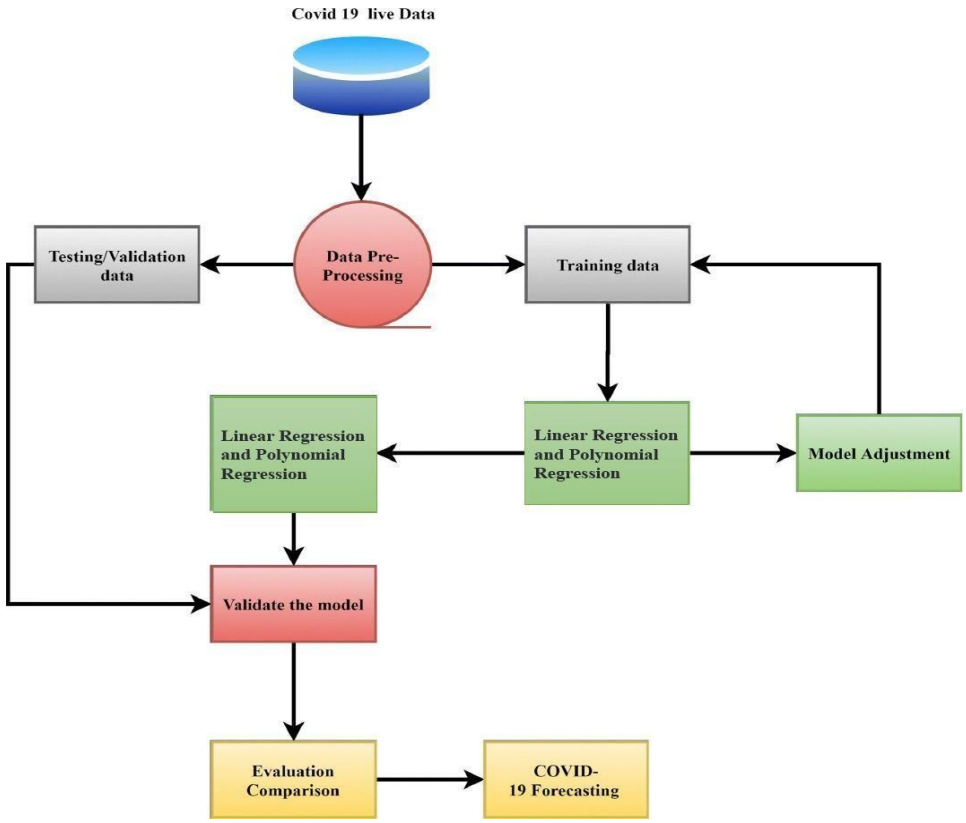

Fig 2: System Architecture

a. Data collection

We have used covid-19 dataset which was taken from John Hopkin's university repository [10]. Serial Number, Observation Date, State/Province, Country/Region, Last Update, Confirmed, Deaths, and Recovered are the eight attributes. The information was gathered between January and December 2020. The dataset includes 172,479 documents from the 6th of December 2020 to the 4th of June 2022. The number of confirmed cases expected by the model for the next few days is compared to the real-world covid-19 confirmed cases data to determine the model's accuracy.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

b. Data Pre-processing

The unstructured text must be refined by which machine learning can be made. Different kinds of steps are followed in this phase. Unnecessary texts are removed so that the text is being cleaned. Lemmatization and punctuation are used to further refine the results. Stop terms, ties, and icons are omitted in order to increase classification and distinction accuracy.

c. Feature extraction

Various features are extracted from the pre-processed clinical reports and are converted into probabilistic values as per the semantics. We use NumPy and Pandas library for extracting relevant features. We identified relevant features like confirmed cases, data wise confirmed case, death, recovered cases, etc., by which the classification can be achieved. Corresponding weight to the feature was given and the machine learning algorithms are supplied with same inputs.

d. Prediction

To forecast the number of confirmed cases worldwide in the future, we employ supervised Machine learning algorithms like LR [1], PR [2] and Lasso [3]. Time series forecasting linear model to train the model by providing the necessary feature as training input.

e. Predicting the global confirmed cases, death cases and recovered cases

With the help of Polynomial Regression model, we were able to predict the future number of globally confirmed cases, death cases and recovered cases very accurately. This method was betterwhen we compare to both Linear Regression and Polynomial Regression.

f. Dataset used

The dataset of COVID-19 is taken from John Hopkin's University repository [10] which contains the data observation, name of the country, name of the state or province, update time, for the particular day, the number of confirmed cases, recovered cases, and death cases (Table 1).

Table 1

Province /State	Country /Region	Lat	Long	1/22/20	1/23/20	. . .	3/27/20
NaN	Afghan	33.00	65.00	0	0	. . .	74
Victoria	Australia	-37. 81	144. 96	0	0	. . .	411
NaN	Algeria	28.03	1.65	0	0	. . .	264

IV. RESULT

4.1 Login and Registration:

Registration page will allow user to create his Username and Password so that he can use those same credentials for the future login, After creating account data of user will be stored in Sqlite3 databse.

Fig 3: Login Page

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

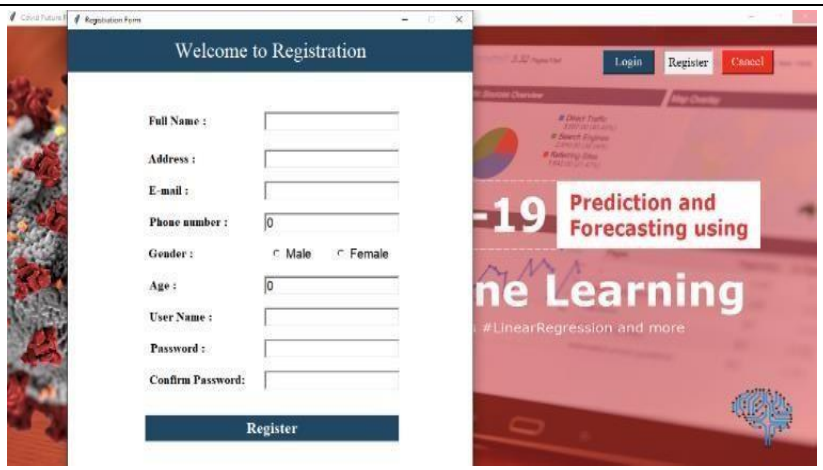

Fig 4: Registration Page

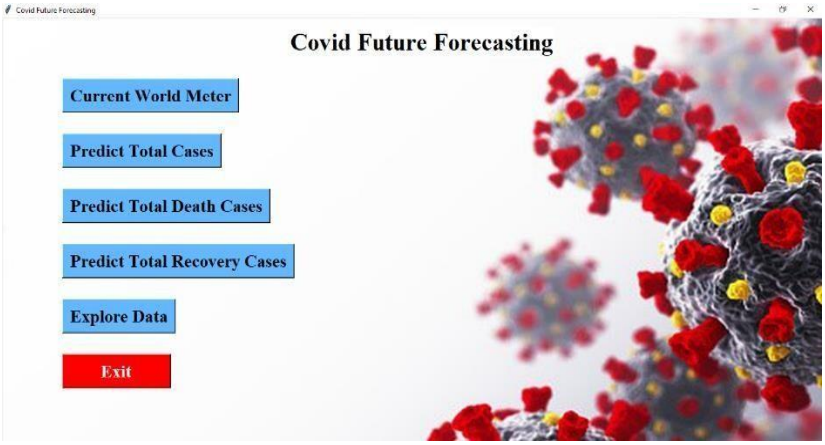

Fig 5: Main Page

As Fig. 5 will be shown after successful Login and Registration. Where the elements will becontaining

1. Current World Meter
2. Predicted total covid cases for next 10 days.
3. Predicted total death cases for next 10 days.
4. Predicted total recovery cases for next 10 days.
5. Explore data: This will visualize data through graphical representation.

4.2 Current World Meter:

In this current world meter it will be fetching the current total cases, death cases and recoveredcases from www.worldometer.com [11] and shows in the form of table as shown in Fig. 6.

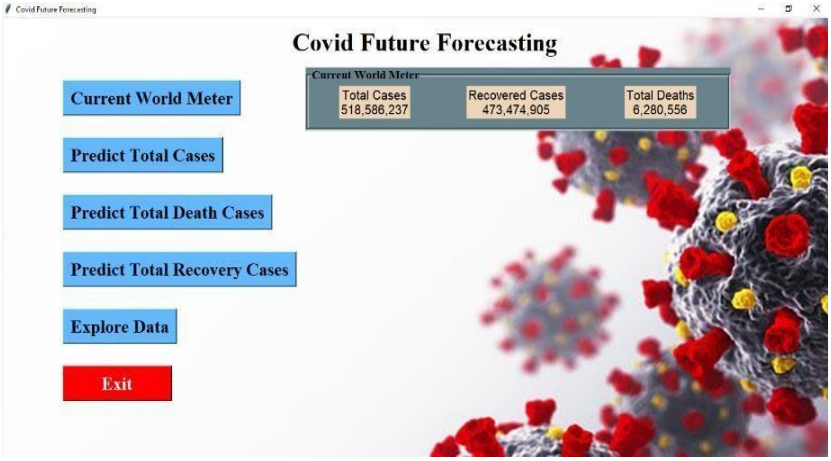

Fig 6: Current world meter

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

4.3 Predicted Total Future Cases:

All the available data is utilized in training the model. The pre-processing and training steps are the same as explained above. The fully trained model is used to predict confirmed positive virus cases for next 10 coming days. In order to create charts containing historical and predicted cases the date index of the data frame is extended. Fig. 7 shows the predicted total cases.

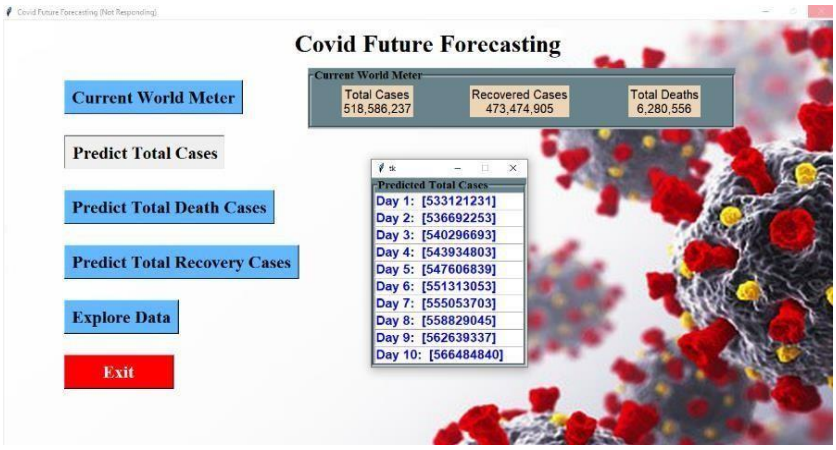

Fig 7: Predict total cases

4.4 Predicted Total Death Cases:

The total death cases in upcoming 10 days based on present dataset of COVID cases.

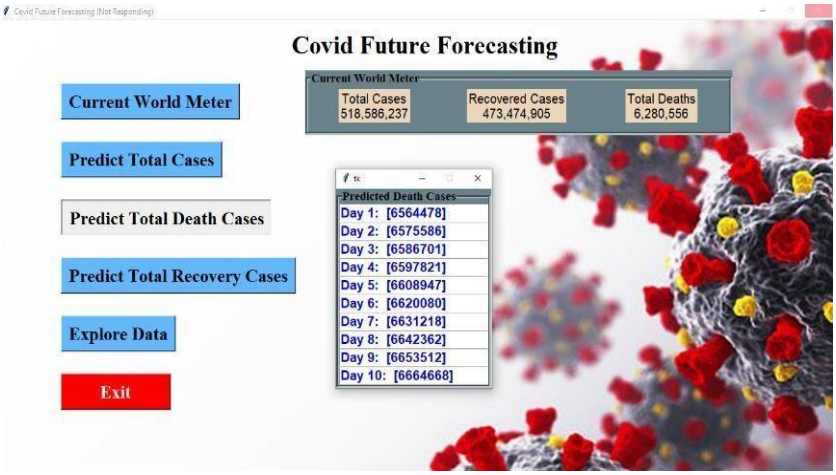

Fig 8: Predict death cases

4.5 Predicted Total Recovered Cases:

Total recovered cases in upcoming 10 days based on present dataset of COVID cases.

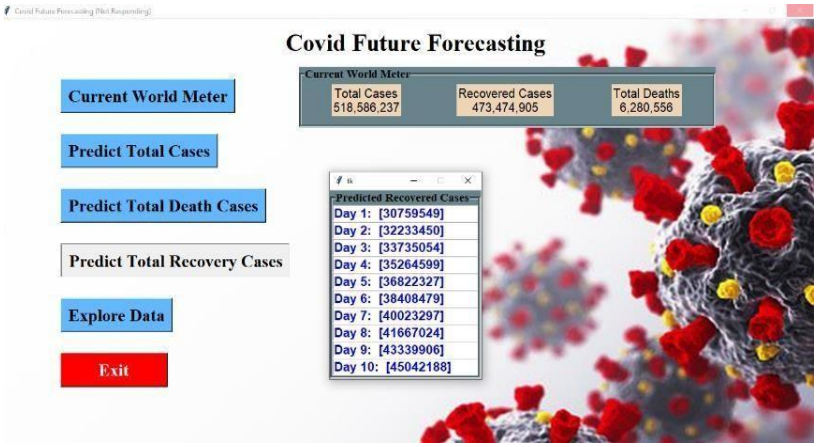

Fig 9: Predict recovered cases

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

4.6 Explore Data:

4.6.1 Growth of cases:

Growth of cases, particularly Total cases, Recovered and death cases in next 10 days across thedifferent areas in the world is shown here in a graphical form.

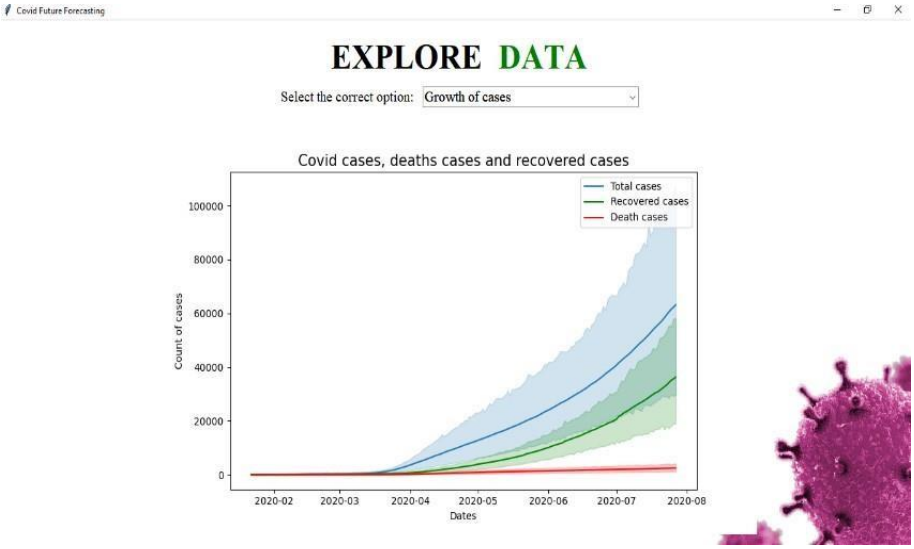

Fig 10: Growth of total covid, death and recovered cases

4.6.2 Country wise covid cases:

This will show the country wise covid cases across the world.

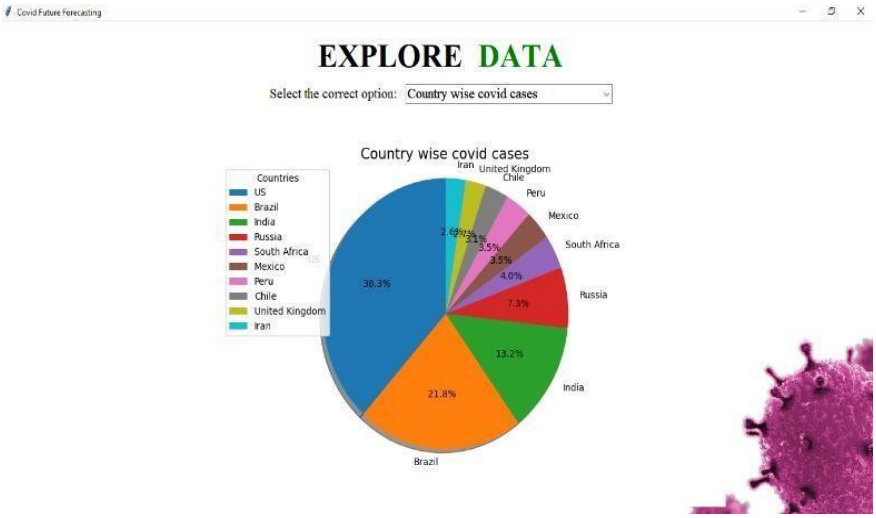

Fig 11: Country wise covid cases

In this we can see the countries like US, India Russia, Brazil, etc. The highest colored part indicates the highest growth of covid cases in that country.

4.6.3 Most vaccinated state in India:

This will show the Vaccination status across the states in India, so with the help of that we can get the Vaccination status of covid.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:11/November-2022 Impact Factor- 6.752 www.irjmets.com

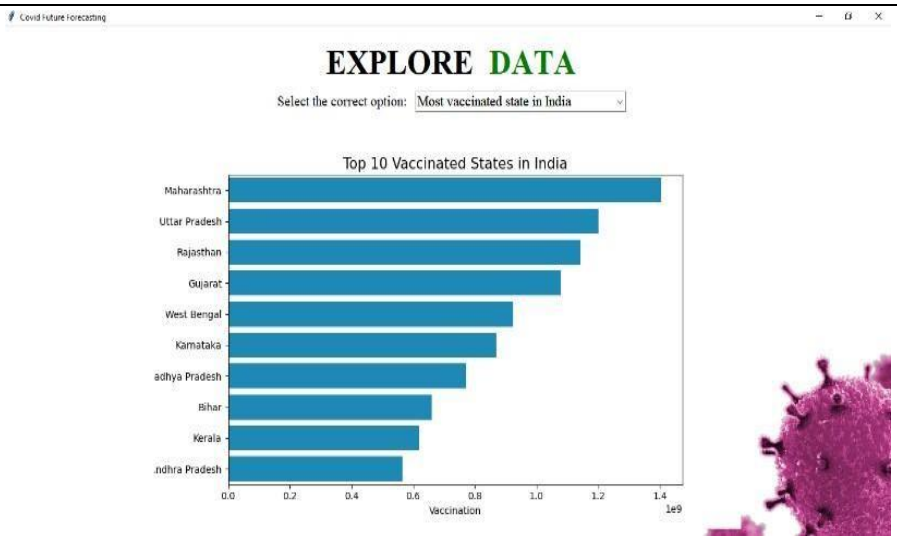

Fig 12: Most Vaccinated state in India

4.6.4 Male vs Female Vaccination status:

Vaccination status of male and females in india has been shown in this data .

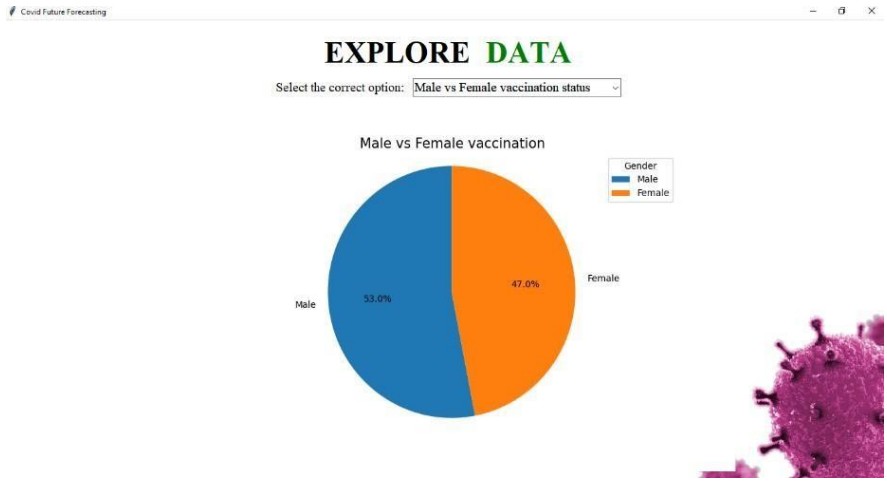

Fig 13: Male vs Female vaccination status

V. CONCLUSION

Predicting COVID-19 cases has immense significance in the present dire scenario. In this work the growth patterns of the disease have been analyzed, data-driven estimations have been incorporated. Deep learning model based on Linear regression, Polynomial regression have been used to predict the trends in coming 10 days such as the number of confirmed positive viral cases, number of deaths caused by the virus and number of people recovered from the novel corona virus number of predicted total cases with some explore data points like Growthof cases, Country wise COVID cases, Vaccination status, etc. Encouraging experimental results have been obtained on the dataset used.

VI. FUTURE SCOPE

The problem of predicting Covid-19 related data such as future cases, recovered cases and deaths is difficult, since we are amidst an outbreak. The future trends and patterns may vary widely based on myriad external conditions like quarantine measures, new behavior of the virus strain, population of a country etc., as the dataset becomes larger and we have more data to train our model, we can improve the accuracy. The same model can be used to predict any future pandemicsthat are similar in nature to SARS COVID-19. This model can be integrated with an application that streams live data from government sites to view real time graphs of COVID-19 related data. Hope that everything will recover and get back to normal soon.

VII. REFERENCES

[1] Smita Rath, Alakananda Tripathy and Alok Ranjan Tripathy "Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model",2020.

[2] Pinkin Sagar, Prinima Gupta & Indu Kashyap "A forecasting method with efficient selection of variables in multivariate data sets", 2021.

[3] Jammbe Z Musoro, Aeilko H Zwinderman, Milo A Puhan, Gerben ter Riet & Ronald B Geskus "Validation of prediction models based on lasso regression with multiply imputed data", 2014.

[4] A.M.U.D. Khanday, Q.R. Khan, S.T. Rabani, SVMBPI: support vector machine based propaganda identification. SN Appl. Sci.(accepted).

[5] Sina F. Ardabili 1, Amir Mosavi "COVID-19 Outbreak Prediction with Machine Learning" 2020.

[6] Artificial Intelligence (AI) and Big Data for Coronavirus (COVID19) Pandemic: A Survey onthe State-of-the-Arts, 2020.

[7] Furqan Rustam, Aijaz Ahmad Reshi, (Member IEEE), Arif Mehmood, Saleem Ullah, Byung- Won ON, Waqar Aslam, (Member, IEEE), And Gyu Sang Choi "COVID-10 Future Forecasting Using Supervised Machine Learning Models", 2020.

[8] L. J. Muhammad, Ebrahem A. Alegehyne, Sani Sharif Usman, Abdulkadir Ahmad, Chinmay chakraborty, L.A. Mohammed "Supervised Machine Learning Models for Prediction of COVID-19 Infection using Epidemilogy Dataset", 2020.

[9] Sivaramakrishnan Rajendar, Rajasekaran Thangaraj, Jayasheelan Palanisamy, Vishnu Kumar Kaliappan "Comparative Analysis of Classifier Models for the Early Prediction of Type2 Diabetes", 2020.

[10] John Hopkins University repository, https://coronavirus.jhu.edu/data.

[11] Wordometer website: www.worldometer.com

[12] L¨angkvist, Martin, Lars Karlsson, and Amy Loutfi. "A review of unsupervised feature learning and deep learning for time-series modeling". Pattern Recognition Letters 42 (2014): 11-24.

[13] Taieb, Souhaib Ben, et al. "A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition". Expert systems with applications 39.8 (2012): 7067-7083.

[14] Yang, Zifeng, et al. "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions." Journal of Thoracic Disease 12.3 (2020): 165.

[15] Peng, Liangrong, et al. "Epidemic analysis of COVID-19 in China by dynamical modeling." arXiv preprint arXiv:2002.06563 (2020).

[16] Jay S Sevak, Aerika D. Kapadia, Jaiminkumar B. Chavda, Arpita Shah, Mrugendrasinh Rahevar. "Survey on semantic image segmentation techniques", 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017.

[17] Hui He, Ran Hu, Ying Zhang, Runhai Jiao, Honglu Zhu. "Chapter 18 Hourly Day-Ahead Power Forecasting for PV Plant Based on Bidirectional LSTM", Springer Science and Business Media LLC, 2019.