

Machine Learning Based Prediction for the Spread of Covid-19 in India

Ms.V.S Esther Pushpam^{#1}, A.Akash^{#2}, S.Divakar^{#3},
E.Karthick Rohan^{#4}

^{#1}Assistant professor, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

^{#2}UG Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, TamilNadu, India

^{#2}UG Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, TamilNadu, India

^{#4}UG Student, Department of CSE, Sri Ramakrishna Institute of Technology, Coimbatore, TamilNadu, India

Submitted: 10-04-2022

Revised: 26-04-2022

Accepted: 28-04-2022

ABSTRACT— The epidemic of coronavirus disease-2019 (COVID-19) establishes a medical emergency of worldwide concern with an exceptionally high danger of spread and affect the entire worldwide. The rapid spread of COVID-19 virus from China to other countries and outbreaks of disease require an epidemiological analysis of the disease in the shortest time and an increased awareness of effective interventions. In this project, the COVID19 progression in India will be analyzed and also, the three most affected states Maharashtra, Tamil Nadu and Andhra Pradesh. In this project, different models will be developed to capture the trend of a number of cases and also predict the cases in the upcoming days. The prediction model is to forecast the behavior of COVID-19 spread in the future months. The time series data will be used to make predictions. Linear Regression and Ridge Regression algorithms will be used for the model to predict the peak infective's and peak infective date for India. The analysis will be made for various states on India and the detailed results will be provided for top three affected states, then final performance and comparative analysis will be made between the predictive model and the actual result.

Keywords-Linear Regression, Ridge Regression, COVID-19.

I. INTRODUCTION

In 2020, the largest pandemic in recent history spread through the world Corona Virus Disease COVID-19. As of May 1st, 2021, there have already been 152 million cases and 3 million

deaths around the world. For the purpose of the ongoing pandemic and predictions for future pandemics, our project seeks to predict the spread of COVID-19 for future analysis. Given the potential benefits of saving lives, improving mental health, and keeping the world going, these advanced predictions could play a part in bringing this pandemic to a control and preventing further spreading in the future and hence effectively improving our collective well-being.

At present, due to the rapid spread of Coronavirus, the World Health Organization(WHO) had declared COVID-19 as a global pandemic. Due to this contagious deadly virus, many people lost their lives and there is massive damage in the global economy and even unemployment has increased. To prevent or to keep this situation in control, the prediction of COVID-19 cases in advance will be helpful for government and the society.

II. PROPOSED SYSTEM

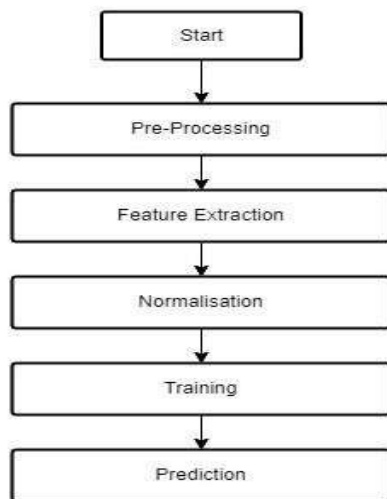
Our proposed model is a prediction before the COVID-19 disease gets more affected in the country. The proposed method predicts the COVID-19 affected cases by using Linear Regression and Ridge Regression algorithm and compares both the predicted model result and the actual result.

The project uses Scikit-learn for training and modelling using Deep Learning. By using these dataset, we are going to predict a model. The predicted model is then used to calculate the spread of COVID-19 in advance for a period of one month.

III. METHODOLOGY

Our proposed system is to predict the spread of COVID-19 disease before the country gets more affected. The proposed method uses the two different algorithms to make prediction and the result of the algorithms will be compared with the original COVID-19 affected cases for their performance on COVID -19 data. Firstly, the Linear Regression algorithm was chosen because of its strong independent assumptions Secondly, the Ridge Regression was selected because it can handle when we have a large dataset and their performance will be evaluated with its accuracy. Both has been selected based on its uniqueness to predict COVID -19 in India. India is a second populated country so the spread need to be avoided with caution. The Most affected states in India are Maharashtra, Tamil Nadu and Andhra Pradesh their trend also will be analyzed.

Working Module



A. Normalization

Normalization refers to a rescaling real-valued numeric attributes into a 0 to 1 range. Data normalization is used in machine learning to make a model training less sensitive to scale of features. This allows our model to converge to better weights and turn in leads to more accurate model.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

B.Linear Regression

Linear Regression is a machine learning algorithm based on the supervised learning. It performs a regression task and a target prediction value based on independent variables. It is mostly

used for finding out the relationship between variables and forecasting.

Linear regression is one of the mostly used predictive modelling techniques. It is represented by a equation $Y = a + bX + e$. This equation can be used to predict the value of a target variable based on the given predictor variables.

C.Ridge Regression

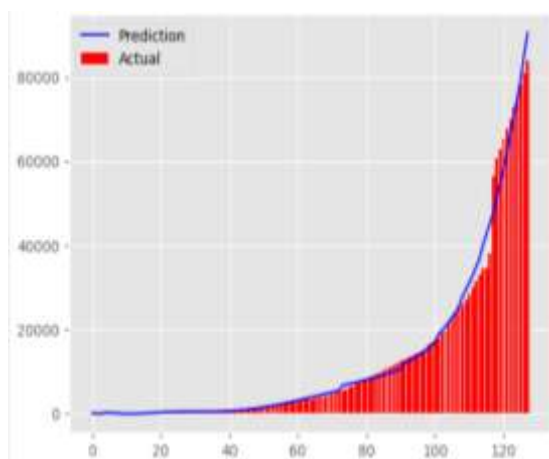
Ridge regression is suitable for analyzing when multiple regression data that suffers from multicollinearity. When multicollinearity occurs, the least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to regression estimates, ridge regression reduces the standard errors. It is hoped that the net effect will be to give estimates that are more reliable.

IV. BLOCK DIAGRAM



The model is fed with the collected datasets having parameters (number of COVID-19 affected cases, number of COVID-19 death cases, number of COVID-19 recovered cases etc..). Once the training is completed, the model is tested and validated.

V. OUTPUT



VI. CONCLUSION

The proposed model on prediction for the spread of COVID-19 in India is developed with Linear Regression and Ridge Regression algorithm. The datasets are trained using the Scikit-learn which gives a good result with perfect accuracy. The data is trained and tested for the model to gain good accuracy. The other researchers have many problems in output, using the dataset, only some were able to get better accuracy. Wrong data are removed successfully from this model since the dataset used is collected from various other sources. The parameters which is used in the dataset is pre-processed well to get better accuracy of the result. The Comparison is made between the Linear Regression and the actual COVID-19 confirmed cases, then the Graph will be displayed with the predictive accuracy, then the accuracy will be chosen and concluded.

REFERENCE

- [1] G Khan, M Sheek-Hussein, AR Al Suwaidi, K Idris and FM Abu-Zidan "Novel coronavirus pandemic: A global health threat", Turk J Emerg Med, vol. 20, pp. 55-62, 2020
- [2] Holshue, C DeBolt, S Lindquist et al., "First Case of 2019 Novel Coronavirus in the United States", N Engl J Med, Jan 2020
- [3] Coronavirus Disease (COVID-19) Situation Reports", Who.int, [online] Available: <https://www.who.int/emergencies/diseases/novel-coronavirus2019/situation-reports>.
- [4] R. M. Anderson, H. Heesterbeek, D. Klinkenberg and T. D. Hollingsworth, "How will country-based mitigation measures influence the course of the COVID-19 epidemic?", Lancet, vol. 395, no. 10228, pp. 931-934, 2020
- [5] Liu, A. A. Gayle, A. Wilder-Smith and J. Rocklöv, "The reproductive number of COVID-19 is higher compared to SARS coronavirus", J. Travel Med., vol. 27, no. 2, 2020.
- [6] P. Wang, X. Zheng, J. Li and B. Zhu, "Prediction of epidemic trends in with logistic model and machine learning technics", Chaos Solitons Fr, vol. 139, no. 110058, pp. 110058, 2020.
- [7] SS. Arun and G. NeelakantaIyer, "On the Analysis of COVID19 - Novel Corona Viral Disease Pandemic Spread Data Using Machine Learning Techniques", 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1222-1227, 2020.
- [8] M. Mousavi, R. Salgotra, D. Holloway and A. H. Gandomi, "COVID-19 Time Series Forecast Using Transmission Rate and Meteorological Parameters as Features", IEEE Computational Intelligence Magazine, vol. 15, no. 4, pp. 34-50, Nov. 2020.
- [9] P. Wang, X. Zheng, J. Li and B. Zhu, "Prediction of epidemic trends in with logistic model and machine learning technics", Chaos Solitons Fr, vol. 139, no. 110058, pp. 110058,
- [10] Zhao et al., "Epidemic growth and reproduction number for the novel Coronavirus disease (COVID-19) outbreak on the Diamond princess cruise ship from January 20 to February 19 2020: A preliminary data-driven analysis", SSRN Electron. J., 2020.