International Journal of Applied Engineering Research ISSN 0973-4562 Volume 16, Number 8 (2021) pp. 640-644

COVID-19 New and Dealth Case Forecasting Using Machine Learning Model

Shilpa Jackson
Department of Computer Applications
The Bhopal School of Social Sciences
Bhopal, India

Abstract

Forecasting mechanism specially using Machine Learning has been proved its importance for accurate and efficient outcomes. The ML models have for some time been utilized in numerous application areas which required the distinguishing proof and prioritization of the factors which are not in control. There exist several methods in ML which can be used for prediction of uncertainties of forecasting models. In this paper we present a model based on ML to forecast the number of patient will be get affected by COVID-19. This is the major threat as of now. The forecasting of the masses that will get affected will be more like saving some of them. In this paper we have utilized Hadoop framework for the storage and processing the huge amount of data collected for various countries. Specifically, three forecasting models like Support Vector Machine (SVM), Linear Regression (LR) and Least Absolute Shrinkage and Selection Operator (LASSO) are utilized in this paper to figure out the increasing cases of COVID-19. The two most common and life treat question are answered by the forecasting models. The first one is the, the number of new cases detected and another is the number of deaths occurs daily. The forecasting models produce promising results for the current situation of the COVID-19 pandemic. The Linear Regression and Least Absolute Shrinkage and Selection Operator models performs best in forecasting of new daily cases and death rates.

Keywords: Hadoop, MapReduce, COVID-19, SVM, LR, LASSO, Forecasting, Pandemic.

I. INTRODUCTION

Machine learning has substantiated itself as a noticeable field of study throughout the most recent decade by tackling numerous exceptionally complex and involved problem sets. Machine learning has dominated its roles in almost all the areas by its high usage on the application such as:

- Natural Language Processing: This subfield concerns with the effective and efficient interaction between computer and the human [1].
- Healthcare: Today health care is more important than anything else in life [2]. Central Government of India has announced Rs. 6,400 crore for health care alone for financial year 2020-21. This shows how important health care industry is. The use of machine learning helps the doctors in early detection and diagnosis of certain life threatening diseases. The application areas in healthcare includes, discovery of drugs and manufacturing, medical imaging diagnosis, personalized medicines for patients based on its health, for record keeping of public health records and analyzing them when required, clinical trials, outbreak prediction and research.
- Business application such as for prediction of stock markets [3].
- Intelligent robots: The smart and intelligent robots can solve various problems without having humans involved. Recently in India robots are utilized in restaurants and hotels for serving drinks and taking orders. These intelligent robots can fight the war, can drive a car or flew the plane. It can do almost whatever human can do.
- Gaming: ML is used by the gaming industry for realistic simulation of characters [4].
- Climate Modeling: ML are used by the Climatologist for prediction of weather accurately.
- Voice: The machine learning can able to distinguish the two different types of noises. The perfect example is the Cocktail party algorithm. Here the machine can distinguish the two person voice from the noise of the party.
- Image and Video Processing: The ML is most popular among image and video processing applications areas. In healthcare industry the analysis of Medical X-ray [5], detection of cancers [6], ulcers, diagnosis of kidneys and eyes are performed efficiently. In defense monitoring of CCTV cameras [7] and even the tracking of enemies through a satellite are done through ML.

The application areas which are discussed above are just a snapshot. Now a day's there are much of machine learning thing are utilized in each and every area. From health care, agriculture, automobile industry to defense. The list is endless.

Just opposite to the conventional method where the prediction is based on history, in machine learning it is obtained from the trial and error method. The old method just takes history of data and do some if-else like decision making [8]. While the ML relies on history, present data, decision making in the terms of various factors based on the history data.

This paper presents the forecasting of COVID-19 patients and death rates. This perdition will helps government in decision making or helps in applying tactics in their future actions. In machine learning accurate or almost accurate forecasting [9] is a biggest challenge. This paper deals with it with the help of

SVM, LR and LASSO algorithm along with Hadoop and Map Reduce framework.

II. HADOOP AND MAP REDUCE

Hadoop is a framework for storing and processing big data. Whereas Map Reduce is the programming model for analysing data by the user programmatically in language likes Java, Python.

A. Hadoop

Hadoop is the framework for storing and processing huge amount of data. Hadoop works on the distributed environment. The computer distributed for storing data are called cluster. These clusters are independent and are isolated from each other. The only known node is the master node. The master node knows all the clusters and has a pointer to them. It is responsible for transferring of data to all other nodes.

Hadoop is used for processing of big data. The term big data simply means unstructured, unarranged, non DBMS data. The areas where big data are generated such as:

- Social media data
- Stock market data
- Power grid data
- Transport data
- Search engine data

The Hadoop architecture has two major layers. Layers include computational layer known as Map Reduce and storage layer called HDFS. Figure 1 shows the Hadoop architecture.

B. Map Reduce

Map Reduce is the programming model for processing data. The user can program logic using Map Reduce with any one of the languages from Java and Python. The major advantages of using Map Reduce as the model is that it can process the multiple data from various sources parallel without needing large computational power. It has two parts. The Map part takes the data as key value pair and Reduce part produces more meaningful analysis of the data. Figure 2 shows the working of Map Reduce model.

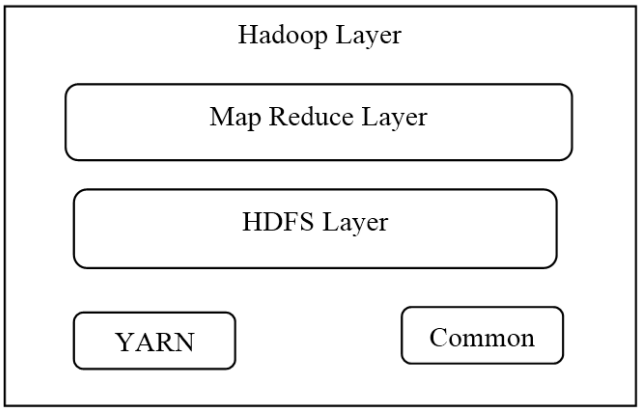

Figure 1: Hadoop architecture

The HDFS is the Hadoop Distrubuted File System which stores the data which are required to be analysed. It stores in the form of part file. Suppose we have 10 GB of social network data and we have 10 processing nodes. The Hadoop takes 10 GB of data and distribute to the rest of the nodes partially. Here in this scenario each node will get 1GB of data for processing. Hence the power of each node is utilized efficiently and the combined results are collected at one source by the reducers.

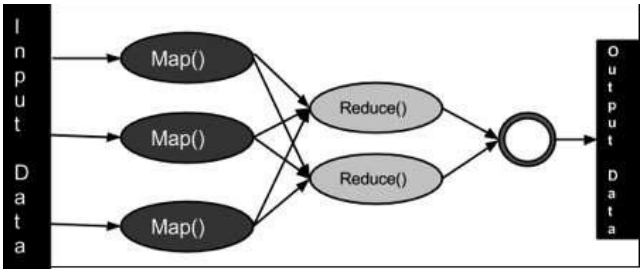

Figure 2: Map Reduce working

III. MACHINE LEARNING MODELS

In supervised learning model the dataset which are used for analysis are labeled. The labeled data means the COVID-19 data are tagged with the correct answer. These data are provided by the hospitals. It is like leaning from the labeled data by a teacher or supervisor. It helps to forecast the unforeseen data.

Three models are used in this paper for COVID-19 forecasting combined with Hadoop framework:

- Linear Regression
- LASSO
- SVM

A. Linear Regression

In this model, a class which are targeted are predicted based on the independent features. Independed and dependent variables are the findings of this algorithm. It highly depends on these two variables. The below equations shows the realtion between x and y variables.

$$y = \phi_0 + \phi_1 x + \epsilon \quad (1)$$

Or

$$E(y) = \phi_0 + \phi_1 x \quad (2)$$

Here, ϵ is the error term, x and y are variables and ϕ_0 and ϕ_1 are x and y axis intercepts respectively.

B. LASSO

LASSO belongs to the linear regression model family. The LASSO principal is based on the shrinkage of the data. The data who are extreme far from the central point are shrink. These shrinkages make a LASSO better predicting algorithm. I not only produce high accuracy output but also reduce the error

rate. It utilized regularization method for penalizing the extra features.

The LASSO based on the minimization of the below given objective function.

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\phi_j) + \lambda \sum_{j=1}^{p} |\phi_j| \quad (3)$$

Where, $x_{ij} and y_i$ are the coefficients and λ is the penalty.

C. SVM

SVM is ML technique which can be used for both classification and regression. It depends upon the mathematical function which predicts the value. It solves the problem using linear function. The function are classed kernel. The kernel transform the data into the desired form. The linear function is dipicted as below equation.

$$f(x) = x'\phi + b \quad (4)$$

Where, x is the input vector, ϕ is the features and b is the slope.

IV. METHODOLOGY

We have utilized Hadoop framework for the storage and processing the huge amount of data collected for various countries. Specifically, three forecasting models like Support Vector Machine (SVM), Linear Regression (LR) and Least Absolute Shrinkage and Selection Operator (LASSO) are utilized in this paper to figure out the increasing cases of COVID-19. The two most common and life threatening issues are answered by the forecasting models. The first one is the, the number of new cases detected and another is the number of deaths occurs daily. Figure 3 shows the methodology workflow.

The Hadoop first distributes the data to different Nodes. Each node executes Map and Reduce Function. Each node is responsible for forecasting of data and reducer combines forecasting data and does the final forecasting.

At each node, the framework first pre-processes the dataset which are acquired from the repository. The pre-processing step normalized the data in order to find out the number of deaths and number of new cases. After data pre-processing the training and test data are separated out. The training set contains 60 days of samples and test data contains 10 days of samples. The models LASSO, LR and SVM are applied to these samples. These models are discussed briefly in Machine Learning Model section.

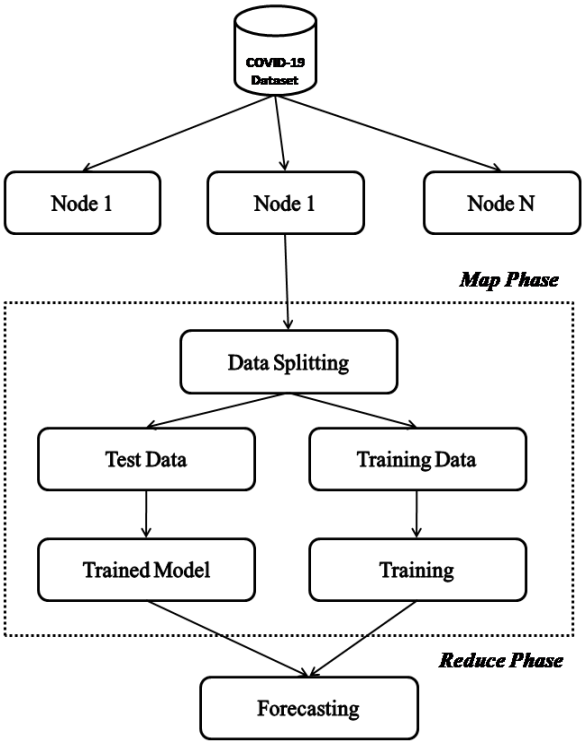

Figure 3: Shows the system workflow

V. DATASET

The point of this investigation is the future prediction of COVID-19 spread in on the quantity of new sure cases and the number of deaths. The dataset are collected from the GitHub repository which was uploaded by the J.Hopkins [10]. The GitHub repo contains day by day time arrangement outline tables, including the quantity of affirmed cases and deaths. All information is from the everyday case report and the update recurrence of information is one day. Information tests from the records are appeared in below tables.

Table I: COVID-19 Data Structure [10]

Column	Type	Description
ID	Char	Unique ID
Date	DateTime	Record date
Cases	Int	Cumulative cases
Cases New	Int	Number of new cases
Age	Char	Age of the reported case
Sex	Char	Gender of reported case

VI. RESULTS

The main objective of this system is to forecast the number of new cases and death occurs using machine learning models.

The dataset is quite accurate which is collected from the GitHub repository. It contains various information like date of the record, new cumulative cases, new daily cases, age and gender of patients etc. This paper predicts the new cases arise and death occurs in different parts of the world. This research does not focuses on any one country rather focuses on whole world.

Figure 4 to 6 shows the performance of algorithm for newly detected cases while figure 8 to 10 shows the performance for death rate prediction of different models.

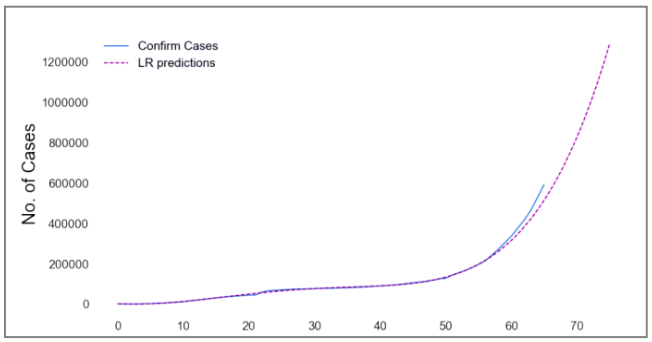

Figure 4: Confirm cases with LR model prediction for next 10 days.

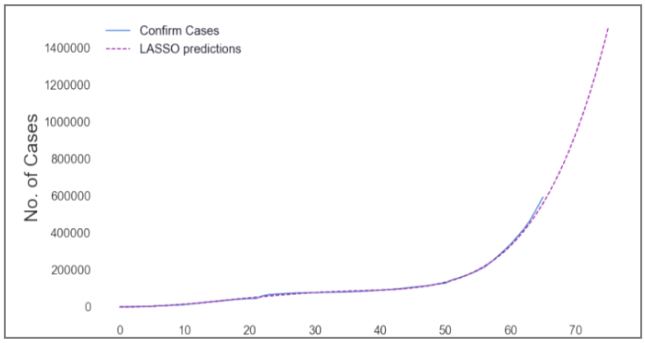

Figure 5: Confirm cases with LASSO model prediction for next 10 days.

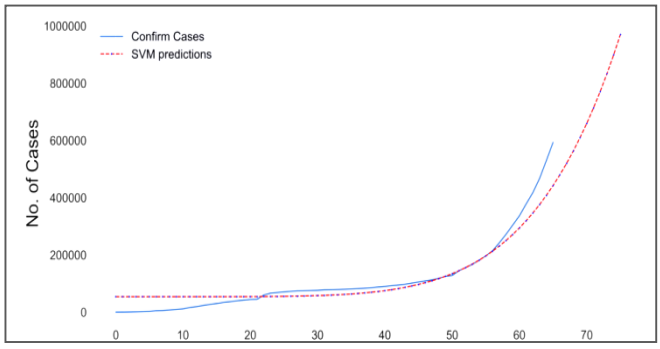

Figure 6: Confirm cases with SVM model prediction for next 10 days.

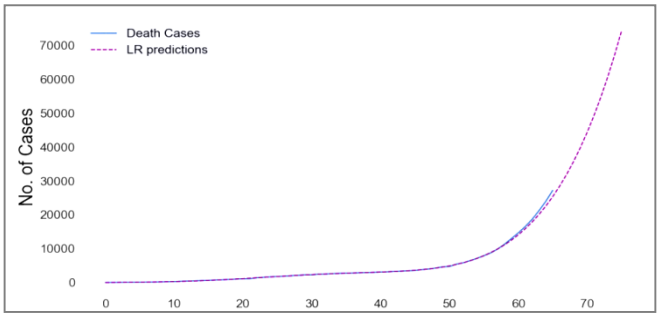

Figure 7: Death cases with LR model prediction for next 10 days.

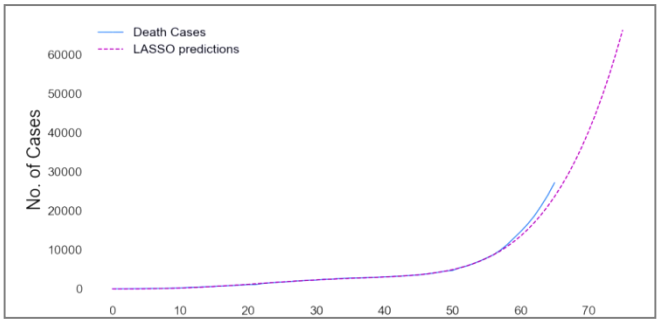

Figure 8: Death cases with LASSO model prediction for next 10 days.

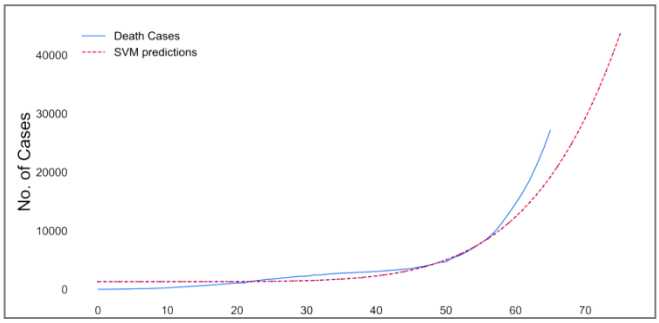

Figure 9: Death cases with SVM model prediction for next 10 days.

As shown in above figures, the forecasting made by the LR model for new confirmed cases is more accurate than SVM. However the forecasting of LASSO model for new confirmed cases is much accurate than SVM and LR.

For forecasting of deaths, LR and LASSO model outperforms the SVM model. Both LR and LASSO forecasts with higher accuracies.

VII. CONCLUSION

The point of this investigation is the future prediction of COVID-19 spread in on the quantity of new sure cases and the number of deaths. The COVID-19 dataset are given as input to the framework. The method LR and LASSO outperforms SVM in terms of accuracy. The LR and LASSO method

accuracy is much higher than that of SVM. Surely the vanilla SVM cannot be used for forecasting of COVID-19 new and death cases. With higher accuracy rate the LASSO and LR can be used for forecasting of COVID-19 cases.

In future we are planning to forecast the live real time dataset. Also we planned to work for India COVID-19 dataset, where recently peaks in the cases are observed.

REFERENCES

[1] A. Gelbukh, "Natural language processing," Fifth International Conference on Hybrid Intelligent Systems (HIS'05), Rio de Janeiro, Brazil, 2005, pp. 1 pp.-, doi: 10.1109/ICHIS.2005.79.

[2] K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.

[3] M. Billah, S. Waheed and A. Hanifa, "Stock market prediction using an improved training algorithm of neural network," 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), Rajshahi, Bangladesh, 2016, pp. 1-4, doi: 10.1109/ICECTE.2016.7879611.

[4] Y. Liu, D. Li and Y. Hu, "Machine Learning in Adversarial Game Using Flight Chess," 2011 Third International Conference on Multimedia Information Networking and Security, Shanghai, China, 2011, pp. 65-68, doi: 10.1109/MINES.2011.124.

[5] K. Ignatyev, P. Munro, R. Speller and A. Olivo, "First X-ray phase contrast images obtained with conventional X-ray source under exposure conditions compatible with real-world applications," IEEE Nuclear Science Symposuim & Medical Imaging Conference, Knoxville, TN, USA, 2010, pp. 889-891, doi: 10.1109/NSSMIC.2010.5873888.

[6] M. Vas and A. Dessai, "Lung cancer detection system using lung CT image processing," 2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, India, 2017, pp. 1-5, doi: 10.1109/ICCUBEA.2017.8463851.

[7] B. Eamthanakul, M. Ketcham and N. Chumuang, "The Traffic Congestion Investigating System by Image Processing from CCTV Camera," 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), Chiang Mai, Thailand, 2017, pp. 240-245, doi: 10.1109/ICDAMT.2017.7904969.

[8] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, ‘‘Statistical and machine learning forecasting methods: Concerns and ways forward,’’ PLoS ONE, vol. 13, no. 3, Mar. 2018, Art. no. e0194889

[9] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, ‘‘Machine learning strategies for time series forecasting,’’ in Proc. Eur. Bus. Intell. Summer School. Berlin, Germany: Springer, 2012, pp. 62–77.

[10] Johns Hopkins University Data Repository. Cssegisanddata. Accessed: Mar. 19, 2021. [Online]. Available: https://github.com/CSSEGISandData