International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 05 | May 2022 www.irjet.net p-ISSN: 2395-0072

COVID-19 FUTURE FORECASTING USING SUPERVISED MACHINE LEARNING MODELS

S. LAXMAN PRASAD[1], V. TANUJA[2], T.SAI REVATHI[3], V. PRADEEP SIVARAM[4]

[1,2,3,4] Final Year B.Tech, Department of Computer Science Engineering, SVPE COLLEGE, Andhra Pradesh, India
Guided by G.Sandhya, Associate Professor, SVPEC, Visakhapatnam, A.P, India.

ABSTRACT: *Machine learning (ML) based forecasting has proved compelling to prepare for outcomes to enhance the decision making on the future course of actions. previously, many applications have been developed using Machine Learning in various fields .Certain prediction methods are being popularly used to handle forecasting issues. This study Tells the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is nowadays considered a danger to mankind. Especially, four standard forecasting models, such as support vector machine, linear regression, least absolute shrinkage, selection operator, and exponential smoothing have been used in this study to forecast the cautionary factors of covid-19. Few types of predictions are made by the models, such as the number of newly infected cases, number of deaths, and number of recoveries in the next 10 days. The outcome produced by the study proves it a promising tool to use these methods for the current scenario of the COVID-19 pandemic. The outcomes confirm that the ES performs superior among all models followed by LR and LASSO which performs excellently in forecasting the new confirmed cases, death rate as well as healing rate, while SVM performs inadequately in all the prediction methods given in the general dataset*.

Keywords: ***COVID-19, Machine Learning, Prediction, Pandemic, future forecasting, Supervised Learning***

1. INTRODUCTION

Machine Learning may be a rapidly evolving and continuously developing field. Recently, many applications are developed using Machine Learning in various fields like healthcare, banking, military equipment, space, autonomous vehicle (AV), business applications, tongue processing (NLP), intelligent robots, gaming, and climate modeling.ML algorithms' learning is usually supported an effort and error method quite opposing conventional algorithms, which follow the programming instructions supported decision statements like if-else. particularly, the study is concentrated on live forecasting of COVID-19 confirmed cases, and also the study is additionally focused on the forecast of COVID-19 outbreak and early response. These prediction systems will be helpful in decision-making to handle the current scenario very virtually. This study desires to supply an earlier forecast model for the spread of novel coronavirus, also referred to as SARS-CoV-2, officially named COVID-19 by the planet Health Organization (WHO). COVID-19 is presently a really extreme danger to human life everywhere the planet. At the tip of 2019, the virus was first identified in a very city of China called Wuhan, when an vast number of individuals developed symptoms like pneumonia.. Coronavirus spread from one city to an entire country in precisely 30 days. On Feb 11, it had been named COVID-19 by World Health Organisation (WHO).. many thousands of individuals are stricken by this pandemic throughout the planet with thousands of deaths every coming day. Thousands of recent people are reported to be positive on a daily basis from Countries across the globe.The virus spreads primarily through close person-to-person physical contact, by respiratory droplets, or by touching the contaminated surfaces. The most difficult aspect of its spread is that a person can possess the virus for many days without showing symptoms. The causes of its spread and considering its danger, almost all the countries have declared strict lockdowns throughout the affected regions and cities. COVID-19 has spread across the globe with around 213 countries and territories affected. The rise in the number of cases of infected coronavirus quickly outnumbered the available medical resources in hospitals, resulting in a significant burden on the health care systems. Due to the limited availability of resources at hospitals and the time delay for the results of the medical tests, it is a distinct situation for health workers to give proper medical treatment to the patients. As the number of cases to test for coronavirus is increasing fast day by day, it is not possible to test due to the time and cost factors. we use machine learning techniques to predict the infection of the coronavirus in patient.

2. METHODOLOGY

The study is about novel coronavirus also referred to as COVID-19 predictions. It causes tens of thousands of deaths and also the death rate is increasing day by day throughout the world. To contribute to the present pandemic situation control

Firstly, we performed a scientific literature review where we carefully analyzed the literature and from the results, we conducted an experiment for research question 1 through which we identified appropriate machine learning techniques for prediction. For research question 2, we conducted an experiment, where we determined what features would influence the results of the prediction of COVID-19 further as this study attempts to perform future forecasting on death. rate, the amount of daily confirmed infected cases, and therefore the number of recovery cases within the upcoming 10 days. The forecasting has been done by using four ML approaches.

3. EXISTING SYSTEMS

Great research has already been done using various computer science for diagnosing and predicting COVID-19 infection and recovery. within the work of information mining predictive model for COVID-19 patient's recovery was developed with four data processing algorithms but however, but among them, a model product of the choice tree has the very best precision of 99.85%. Medical researchers throughout the world are currently involved to get appropriate vaccines and medications for the disease. Since there's no approved medication now for killing the virus therefore the governments of all countries are that specialize in the precautions to prevent the spread. To contribute to the current aspect of knowledge, multiple researchers are studying the varied dimensions of the pandemic and producing the outcomes to assist humanity. within the work, the machine learning-based approach was designed for a real-time forecast of the 2019-nCoV outbreak using news alerts reported by Media Cloud, an officer heath report from the Chinese Center Disease for Control and Prevention, and internet search activity from Baidu.

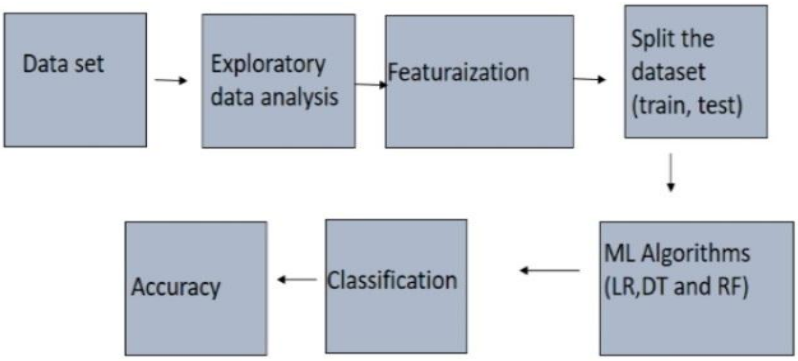

Fig.1: Backend Module Diagram

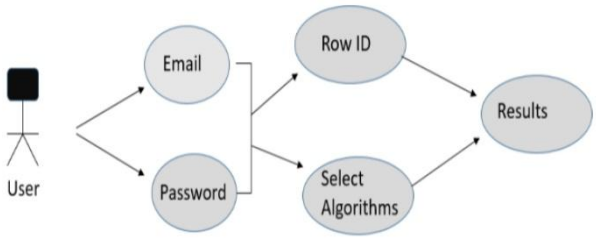

Fig.2:Frontend Module Diagram

4. PROPOSED SYSTEMS

To contribute to the present humanitarian crisis.The forecasting is completed for the three important variables of the disease for the approaching 10 days1) The amount of New confirmed cases2) The amount of death cases3) The quantity of recoveries This problem of forecasting has been considered a regression problem during this study, so the study relies on supervised ML regression models like linear regression, least absolute shrinkage and selection operator, support vector machine, and exponential smoothing. the educational models are trained using the COVID-19 patient stats dataset provided by Johns Hopkins. The study is about novel coronavirus also called COVID- 19 predictions. COVID-19 has proved a gift potential danger to human life. It causes tens of thousands of deaths and also the death rate is increasing day by day

throughout the world. The forecasting has been done by using four ML approaches that are appropriate to the present context. The dataset utilized in the study contains daily statistic summary tables, including the amount of confirmed cases, deaths, and recoveries within the past number of days from which the pandemic started. Initially, the dataset has been preprocessed for this study to seek out the world statistics on the daily number of deaths, confirmed cases, and recoveries. After the initial data preprocessing step, the dataset has been divided into two subsets: a training set (56 days) to coach the models and a testing set (10 days). the training models have then been evaluated supported important metrics like R2-score, R2 adjusted score MSE, RMSE, and MAE and reported within the results. The proposed approach utilized in the study has been shown as a diagram.

5. WORKING

5.1Dataset preparation and Data preprocessing

The data set that was wont to train the model to predict COVID-19 was gathered from open-source data shared by Yanyan Xu. the information set contained information about hospitalized patients with COVID-19. It included demographic data, signs and symptoms, previous medical records,and laboratory values that were extracted from electronic records. The data-set could be a combined multi-dimensional data. a number of the information gives information on whether the patient is diagnosed with a specific disease within the past like Renal Diseases, Digestive Diseases, and other data containing precise clinical values obtained previously. Textual data was encoded with integer values for the experimental setup

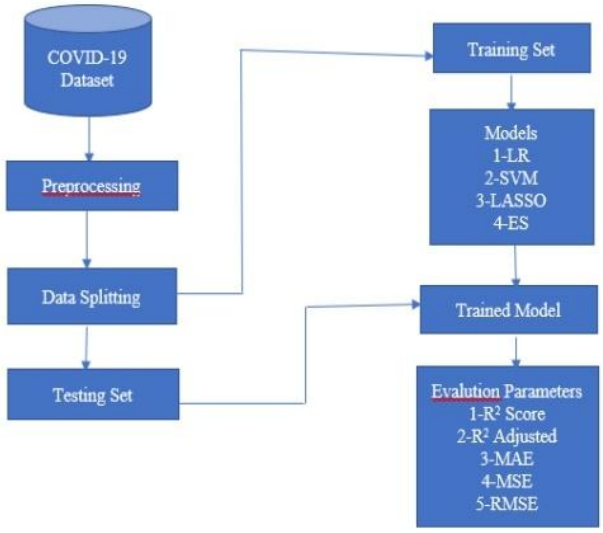

Fig.3: Workflow diagram

Data preprocessing is a very important process within the development of a machine learning model. the information collected is usually loosely controlled with out-of-range values, missing values, etc. Such data can mislead the results of the experiment. The hint of missing values - In our data, missing values are handled by employing a simple imputer from the sklearn python package. The missing values are replaced by using the mean strategy. Encoding Categorical Data - We used the package of OneHotEncoder in python, this package handles categorical data by one-hot or dummy encoding scheme.

5.2 Data Collection

Data collection was a necessary and lengthy process. Anyhow within the field of research, the accuracy of the information collection is critical to keep up cohesion. because the clinical information of patients wasn't publicly obtainable, it absolutely was a tricky and tedious process to gather the information. Various Hospitals and Health Institutes were approached to induce the foremost accurate data but thanks to the current situation at hospitals with a significant inflow of patients with COVID-19, we couldn't get access to direct information. An intense search was conducted on various databases to assemble open-source clinical details of patients diagnosed with COVID-19. It's time for a knowledge analyst to select up the baton and guide the thanks to machine learning implementation. The job of an information analyst is to

search out ways and sources of collecting relevant and total data, interpreting it, and analyzing outcomes with the assistance of statistical methods. the sort of information relies on what you would like to predict. there's no exact answer to the question "How much data is needed?" because each machine learning problem is exclusive. In turn, the quantity of attributes data scientists will use when building a predictive model depends on the attributes' predictive value. 'The better, the more useful approach is affordable for this phase. It's difficult to estimate which a part of the info will provide the foremost accurate results until the model training begins. That's why it's significant to gather and store all data internal and open, structured and unstructured. The tools for collecting internal data depend upon the industry and business infrastructure. A weblog file, additionally, is an honest source of internal data. It stores data about users and their online behavior time and length of visit, viewed pages or objects, and placement.

Data formating:-

The significance of knowledge formatting rises when data is achieved from various sources by different people. the earlier task for an information scientist is to standardize record formats. A specialist studies whether variables defining each attribute are recorded within the same way. Titles of products and services, prices, dates formats, and addresses are samples of variables. The principle of knowledge consistency also applies to attributes defined by numeric ranges.

Data cleaning:-

This set of operations permits removing noise and fixing inconsistencies in data. a knowledge scientist can fill in skipping data using imputation techniques, e.g. replacing missing values with mean attributes. A specialist also notices outliers — observations that fluctuate significantly from the remainder of the allocation. If an outlier indicates incorrect data, a knowledge scientist deletes or updates them if feasible. This stage also contains removing incomplete and useless data objects.

Data anonymization:-

Sometimes a knowledge scientist must anonymize or ban attributes describing sensitive information (i.e. when working with healthcare).

Data sampling:-

Big datasets require more additional time and computational power for analysis. If a dataset is too enormous, applying data sampling is the way to go. A data scientist uses this technique to select a shorter but representative data sample to build and run models much quicker, and at the same time produce exact outputs.

Featuraization:-

Featuraization may be a method to change some type of data into a numerical vector. Featuraization is distinct from feature engineering. Feature engineering is simply transforming the numerical features. In feature engineering, features are already in numerical form. Whereas in Featuraization data doesn't have to be within the style of a numerical vector. The machine learning model cannot operate with row text data directly. In the end, machine learning models work with numerical features. So it's important to vary some styles of data into a numerical vector in order that we are able to leverage the full power of algebra and statistics tools with other varieties of data also.

Data splitting:-

A dataset operated for machine learning should be partitioned into three subsets — training, test,and validation sets. Training set: - an information scientist utilizes a training set to coach a model and define its optimal parameters — parameters it's to know from data.

Test Set:-

A test set is required for an evaluation of the trained model and its ability for generalization. The latter means a model's ability to spot patterns in new unseen data after having been trained over training data. It's crucial to use other subsets for training and testing to avoid model overfitting, which is that the incapability for generalization we mentioned above.

Modeling:-

During this stage, a knowledge scientist drills numerous models to define which one amongst them provides the foremost valid predictions

Model training:-

After an information scientist has preprocessed the gathered data and split it into three subsets, he or she will be able to proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and result from a model that may find a target value (attribute) in new data — an answer you wish to induce with predictive analysis. the target of model training is to develop a model. Two model training styles are commonest — supervised and unsupervised learning. the choice of every style relies on whether you need to predict specific attributes or group data objects by resemblances.

Supervised learning: Supervised learning allows for processing data with target attributes or labeled data. These attributes are mapped in recorded data before the training begins. With supervised learning, an information scientist can decode classification and regression problems.

Unsupervised learning: During this training style, an algorithm analyzes unlabeled data. The destination of model training is to search out hidden interconnections between data objects and structure objects by likenesses or dissimilarities. Unsupervised learning aims at solving such issues as clustering, association rule learning, and dimensionality reduction.

Model Testing:- the aim of this stage is to style the foremost straightforward model capable to develop a mark value quickly and nicely enough. a knowledge scientist can accomplish this plan through model tuning. That's the optimization of model parameters to achieve an algorithm's best version. one in every of the higher efficient methods for model evaluation and tuning is cross-validation.

Cross-validation:- Cross-validation is that the considerable commonly used tuning method. It entails dividing a training dataset into ten equal parts (folds). A given model is trained on only nine folds then tested on the tenth one (the one previously left out). Training resumes until every fold is left aside and used for testing. As an outcome of the model implementation measure, a specialist estimates a cross-validated score for every set of hyperparameters. a knowledge scientist trains models with additional sets of hyperparameters to define which model has the best prediction exactness. The cross-validated score points average model routine across ten hold-out folds.. There are different error metrics for machine learning tasks.

SOFTWARE REQUIREMENTS

IDE: Anaconda Jupyter

Programming Language: Python

HARDWARE NECESSITY

PROCESSOR: Dual Core 2 Duos.

HARD DISK: 250 GB

RAM: 4 GB DD RAM

ALGORITHM / TECHNIQUE USED

Four regression models are employed in this study of COVID-19 future forecasting:

- Linear regression
- LASSO Regression
- Support Vector Machine
- Exponential Smoothing

1) Linear Regression: In regression modeling, a mark class relies on the separated features. This process are often thus accustomed discover the association between separated and dependent variables and also for forecasting. rectilinear regression a kind of regression modeling is that the most usable statistical technique for predictive analysis in machine learning.

There are two factors (x, and y) that are involved in rectilinear regression analysis. The equation below shows how y is expounded to x called regression.

y = β0 + β1x + ε (1) or equivalently E(y) = β0 + β1x (2) Here, ε is that the error term of regression toward the mean.

The error term here uses to account for the variability between both x and y, β0 represents y-intercept, and β1 represents slope.

2) LASSO: LASSO could be a regression model that belongs to the regression procedure which uses shrinkage. Shrinkage during this context guides to the shrinking of excessive values of an information sample towards central values.

The shrinkage procedure thus makes LASSO better and more stable and also reduces the error. Since the model performs L1 regularization and therefore the penalty added during this case is adequate to the magnitude

Thus the models are made sparse with few coefficients during this case of regularization since the method eliminates the coefficients when their values are adequate to zero. works on an objective to reduce the following: Xn i=1 (yi − Xj xijβj) 2 + λ Xp j=1 |βj | It sets the coefficient, which may be interpreted as min(sum of square residuals + λ |slope|), where, λ |slope| is penalty term.

3) Support Vector Machine: Support Vector Machines perform sorting by constructing an N-dimensional hyperplane that separates the info into two categories.

In SVM, the variable is named an attribute and also the transformed attribute is termed a feature. selecting the various suitable sample data is termed feature selection. a collection of parts describing one case is named a vector putting the concept in ML context with a multivariate training dataset (xn) with N number of observations with yn as a group of observed responses. The linear function will be depicted as f(x)=x0β+b.

6. CONCLUSION

The outcomes of the study prove that ES performs most useful within the current forecasting domain given the character and size of the dataset. LR and LASSO also execute nicely for forecasting to some extent to expect death rate and make sure cases. per the results of those two models, the death rates will rise within the upcoming days, and therefore the recoveries rate are stalled. SVM has poor leads to all scenarios thanks to the ups and downs within the dataset values... during this study, an ML-based forecast system has been proposed for predicting the chance of COVID- 19 outbreak globally.ML methods for forecasting. Real-time live forecasting are going to be one amongst the first focuses in our forthcoming work.

7. Future Scope

In this study, a scientific literature review has been accomplished to see the appropriate algorithm for the prediction of COVID-19 in patients. There was no pure proof found to summarize one algorithm because the suitable technique for prediction. there's plenty of range for Machine Learning in Healthcare.ML methods for forecasting. Real-time live forecasting are going to be one in all the preliminary focuses of our future work.

REFERENCES

[1] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," PloS one, vol. 13, no. 3, 2018.

[2] C. P. E. R. E. Novel et al., "The epidemiological characteristics of an out-break of 2019 novel coronavirus diseases (covid-19) in china," Zhonghualiu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi, vol. 41, no. 2,p. 145, 2020.

[3] Y. Grushka-Cockayne and V. R. R. Jose, "Combining prediction intervals within the m4 competition," International Journal of Forecasting, vol. 36, no. 1,pp. 178–185, 2020

[4] J.-H. Han and S.-Y. Chi, "Consideration of producing data to use machine learning methods for predictive manufacturing," in 2016 Eighth International Conference on Ubiquitous and Future Networks (ICUFN). IEEE, 2016, pp. 109–113

[5] J. Lupón, H. K. Gaggin, M. de Antonio, M. Domingo, A. Galán, E. Zamora, J. Vila, J. Peñafiel, A. Urrutia, E. Ferrer et al., "Biomarker- assist score for reverse remodeling prediction in heart failure: the st2-r2 score," International journal of cardiology, vol. 184, pp. 337–343, 2015.

[6] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996

[7] C. P. E. R. E. Novel et al., "The epidemiological characteristics of a deadly disease of 2019 novel coronavirus diseases (covid-19) in china," Zhonghua liu xing bing xue za zhi=Zhonghua liuxingbingxue zazhi, vol. 41, no. 2, p. 145, 2020.

[8] G. Grasselli, A. Pesenti, and M. Cecconi, "Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response," Jama, 2020.

[9] WHO. Naming the coronavirus disease (covid-19) and also the virus that causes it. [Online]. Available: https://www.who.int/emergencies/diseases/novel coronavirus-2019/technical-guidance/namingthe-coronavirus-disease- (covid-2019)-and-thevirus-that-causes-it

[10] F. Petropoulos and S. Makridakis, "Forecasting the novel coronavirus covid-19," Plos one, vol. 15, no. 3, p. e0231236, 2020.

[11] J. H. U. data repository. Cssegisanddata. [Online].**[12]** K. M. Anderson, P. M. Odell, P. W. Wilson, and W.B.Kannel,"Cardiovasculardiseaserisprofiles,Americanheartjournal,vol.121,no.1,pp.293–298,1991.

BIOGRAPHIES

G.SANDHYA
Currently working as associate professor from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

V. TANUJA
Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College

S. LAXMAN PRASAD
Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College.

T. SAI REVATHI
Pursuing B-tech from Department of computer science and Engineering at Sanketika Vidya Parishad Engineering College