International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

COVID-19 FUTURE FORECASTING USING SUPERVISED MACHINE LEARNING MODELS

Bouna das Y[1], Nis shammini Nimsha R[2], Dayana R[3]

[1,2],Dept of Computer Science Engineering, Jeppiaar Institute of Technology,
[3] Assist Professor. Dayana, Dept. of computer science Engineering, Jeppiaar Institute of Technology ,Tamilnadu, India

---

Abstract - *Machine learning (ML) based forecasting mechanisms have proved their significance to anticipate in perioperative outcomes to improve the decision making on the future course of actions. The ML models have long been used in many application domains which needed the identification and prioritization of adverse factors for a threat. Several prediction methods are being popularly used to handle forecasting problems. This study demonstrates the capability of ML models to forecast the number of upcoming patients affected by COVID-19 which is presently considered as a potential threat to mankind. In particular, four standard forecasting models, such as linear regression (LR), least absolute shrinkage and selection operator (LASSO), support vector machine (SVM), and exponential smoothing (ES) 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, such as the number of newly infected cases, the number of deaths, and the number of recoveries in the next 10 days. The results produced by the study proves it a promising mechanism to use these methods for the current scenario of the COVID-19 pandemic. The results prove that the ES performs best among all the used models followed by LR and LASSO which performs well in forecasting the new confirmed cases, death rate as well as recovery rate, while SVM performs poorly in all the prediction scenarios given the available dataset.*

Key Words*: *COVID-19, exponential smoothing method, future forecasting, Adjusted R2 score, supervised machine learning

1.INTRODUCTION

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. ML algorithms' learning is typically based on trial and error method quite opposite of conventional algorithms, which follows the programming instructions based on decision statements like if-else. One of the most significant areas of ML is 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 are lots 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 particular, the study is focused on live forecasting of COVID-19 confirmed cases and study is also focused on the forecast of COVID- 19 outbreak and early response. These prediction systems can be very helpful in decision making to handle the present scenario to guide early interventions to manage these diseases very effectively.

This study aims to provide an early forecast model for the spread of novel coronavirus, also known as SARS-CoV-2, officially named as COVID-19 by the World Health Organization (WHO). COVID-19 is presently a very serious threat to human life all over the world. At the end of 2019, the virus was first identified in a city of China called Wuhan, when a large number of people developed symptoms like pneumonia. It has a diverse effect on the human body, including severe acute respiratory syndrome and multi-organ failure which can ultimately lead to death in a very short duration. Hundreds of thousands of people are affected by this pandemic throughout the world with thousands of deaths every coming day. Thousands of new people are reported to be positive every day from countries across the world. The virus spreads primarily through close person to person physical contacts, by respiratory droplets, or by touching the contaminated surfaces. The most challenging 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 either partial or strict lockdowns throughout

the affected regions and cities. Medical researchers throughout the globe are currently involved to discover an appropriate vaccine and medications for the disease. Since there is no approved medication till now for killing the virus so the governments of all countries are focusing on the precautions which can stop the spread. Out of all precautions, "be informed" about all the aspects of COVID-19 is considered extremely important. To contribute to this aspect of information, numerous researchers are studying the different dimensions of the pandemic and produce the results to help humanity

1.1METHODOLOGY

The study is about novel coronavirus also known as COVID-19 predictions. The COVID-19 has proved a present potential threat to human life. It causes tens of thousands of deaths and the death rate is increasing day by day throughout the globe. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death. rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 10 days. The forecasting has been done by using four ML approaches.

1.2 EXISTING SYSTEM

Much research has already been done using various artificial intelligence for diagnosing and predicting COVID-19 infection and recovery. In the work of data mining predictive model for COVID-19 patient's recovery were developed with four data mining algorithms but however among them, model made of the decision tree has the highest accuracy of 99.85%. In the work of convolutional neural networks that predict novel coronavirus with x-ray images were developed. The deep learning technique, which is one of the sub-branches of ML, inspired by the structure of the human brain is used for the automatic prediction of 2019-nCoV patients. Dataset with chest x-ray images were used, and pre-trained models including InceptionV3, ResNet50 and Inception ResNetV2 were trained and tested on the dataset. The performance result of the study showed that the RestNet pre-trained model gave the highest accuracy among the three models: 98%. Therefore, this shows that the model can help health workers to make decisions in clinical practice with high-performance accuracy, which can also detect 2019-nCoV in the early stages of infection. In the work of a modified susceptible-exposed-infectious- removed (SEIR) Model and ML Model for prediction of the trend of the 2019-nCoV pandemic in China were developed under public health interventions. The models were effective in predicting the pandemic peaks and size. Population migration data before and after 23rd January 2020 and updated 2019-nCoV epidemiological data were integrate into the SEIR Model to derive the pandemic curve. The ML approach was trained on 2003 SARS data to predict the pandemic. In the work of data mining and a deep learning pilot study were carried out to predict 2019-nCoV incidence by levering Google trend data in Iran. Long Short-Term Memory and Linear Regression Models were used to estimate the number of 2019-nCoV positive cases. The models were evaluated with root mean square error (RMSE) metric and 10 folds cross-validation techniques, respectively.The RMSE of long short-term memory and linear regression Models were 27.187 and 7.562, respectively. Moreover, the study predicted the trend of the 2019-nCoV outbreak. Such predictions can support healthcare managers and policy makers with planning, allocating and deploying healthcare resources effectively. Reference identified an intrinsic 2019-nCoV genomic signature using an ML-based alignment-free approach. This approach incorporated ML-controlled digital signal analysis of genome analysis, augmented decision-making process and Spearman's rank correlation coefficient analysis for validation result. The result of the study corroborates the research hypothesis of a bat as the origin 2019-nCoV pandemic and the study further classifies the pandemic as Sarbecovirus within beta coronavirus. More than 5000 unique genomic sequences from the dataset, totaling 61:8 million bp were analyzed with more than 90% accuracy. In the work of the machine learning-based approach was developed for a real-time forecast of 2019-nCoV outbreak using news alerts reported by Media Cloud and official heath report from Chinese Center Disease for Control and Prevention, internet search activity from Baidu and daily forecast from GLEAM (an agent-based mechanistic model)

1.3 PROPOSED SYSTEM

The study is about novel coronavirus also known as COVID-19 predictions. The COVID-19 has proved a present potential threat to human life. It causes tens of thousands of deaths and the death rate is increasing day by day throughout the globe. To contribute to this pandemic situation control, this study attempts to perform future forecasting on the death rate, the number of daily confirmed infected cases and the number of recovery cases in the upcoming 10 days. The forecasting has been done by using four ML approaches that are appropriate to this context. The dataset used in the study contains daily time series summary tables, including the number of confirmed cases, deaths, and recoveries in the past number of days from which the pandemic started. Initially, the dataset has been preprocessed for this study to find the global statistics of the daily number of deaths,

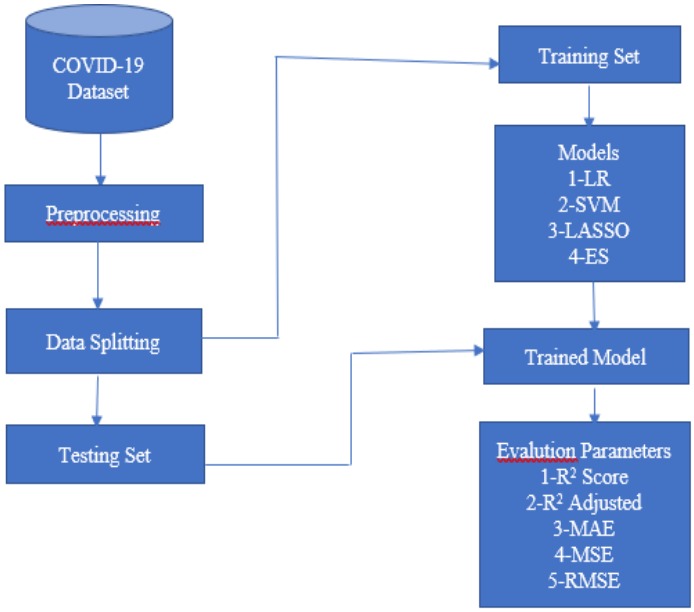
confirmed cases, and recoveries. The resulted time-series has been extracted from the reported data as shown in Table 4, the samples of the resulted dataset are shown in Tables 5, 6, 7 respectively. After the initial data preprocessing step, the dataset has been divided into two subsets: a training set (56 days) to train the models and testing set (10 days). The learning models such as SVM, LR, LASSO, and ES have been used in this study. These models have been trained on the days and newly confirmed cases, recovery, and death patterns. The learning models have then been evaluated based on important metrics such as R2-score, R2 adjusted score MSE, RMSE, and MAE and reported in the results. The proposed approach used in the study has been shown as a block diagram.

This study attempts to develop a system for the future forecasting of the number of cases affected by COVID-19 using machine learning methods. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries, and the number of deaths due to COVID-19 worldwide. As the death rate and confirmed cases are increasing day by day which is an alarming situation for the world. The number of people who can be affected by the COVID-19 pandemic in different countries of the world is not well known. This study is an attempt to forecast the number of people that can be affected in terms of new infected cases and deaths including the number of expected recoveries for the upcoming 10 days. Four machine learning models LR, LASSO, SVM, and ES have been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries.

2. WORKING

Dataset preparation and preprocessing:

Data is the foundation for any machine learning project. The second stage of project implementation is complex and involves data collection, selection, preprocessing, and transformation. Each of these phases can be split into several steps.



Data collection:

It's time for a data analyst to pick up the baton and lead the way to machine learning implementation. The job of a data analyst is to find ways and sources of collecting relevant and comprehensive data, interpreting it, and analyzing results with the help of statistical techniques.

The type of data depends on what you want to predict.

There is no exact answer to the question "How much data is needed?" because each machine learning problem is unique. In turn, the number of attributes data scientists will use when building a predictive model depends on the attributes' predictive value.

'The more, the better' approach is reasonable for this phase. Some data scientists suggest considering that less than one-third of collected data may be useful. It's difficult to estimate which part of the data will provide the most accurate results until the model training begins. That's why it's important to collect and store all data — internal and open, structured and unstructured.

The tools for collecting internal data depend on the industry and business infrastructure. For example, those who run an online-only business and want to launch a personalization campaign can try out such web analytic tools as Mixpanel, Hotjar, CrazyEgg, well-known Google analytics, etc. A web log file, in addition, can be a good source of internal data. It stores data about users and their online behavior: time and length of visit, viewed pages or objects, and location.

Companies can also complement their own data with publicly available datasets. For instance, Kaggle, Github contributors, AWS provide free datasets for analysis.

Data preprocessing:

The purpose of preprocessing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling.

Data formating: - The importance of data formatting grows when data is acquired from various sources by different people. The first task for a data scientist is to standardize record formats. A specialist checks whether variables representing each attribute are recorded in the same way. Titles of products and services, prices, date formats, and addresses are examples of variables. The principle of data consistency also applies to attributes represented by numeric ranges.

Data cleaning: - This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible. This stage also includes removing incomplete and useless data objects.

Data anonymization: - Sometimes a data scientist must anonymize or exclude attributes representing sensitive information (i.e. when working with healthcare and banking data).

Data sampling: - Big datasets require more time and computational power for analysis. If a dataset is too large, applying data sampling is the way to go. A data scientist uses this technique to select a smaller but representative data sample to build and run models much faster, and at the same time to produce accurate outcomes.

Featuraization:-

Featuraization is a way to change some form of data (text data, graph data, time-series data…) into a numerical vector.

Featuraization is different from feature engineering. Feature engineering is just transforming the numerical features somehow so that the machine learning models work well. In feature engineering, features are already in the numerical form. Whereas in Featuraization data not need to be in the form of numerical vector.

The machine learning model cannot work with row text data directly. In the end, machine learning models work with numerical (categorical, real…) features. So it is import to change some type of data into numerical vector so that we can leverage the whole power of linear algebra (making the decision boundary between data points) and statistics tools with other types of data also.

Data splitting:-

A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

Training set: - A data scientist uses a training set to train a model and define its optimal parameters — parameters it has to learn from data.

Test set: - A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model's ability to identify patterns in new unseen data after having been trained over a training data. It's crucial to use different subsets for training and testing to avoid model over fitting, which is the incapacity for generalization we mentioned above.

Modeling:-

During this stage, a data scientist trains numerous models to define which one of them provides the most accurate predictions.

Model training:-

After a data scientist has preprocessed the collected data and split it into three subsets, he or she can proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that is able to find a target value (attribute) in new data — an answer you want to get with predictive analysis. The purpose of model training is to develop a model.

Two model training styles are most common — supervised and unsupervised learning. The choice of each style depends on whether you must forecast specific attributes or group data objects by similarities.

Supervised learning: - Supervised learning allows for processing data with target attributes or labeled data. These attributes are mapped in historical data before the training

begins. With supervised learning, a data scientist can solve classification and regression problems.

Unsupervised learning: - During this training style, an algorithm analyzes unlabeled data. The goal of model training is to find hidden interconnections between data objects and structure objects by similarities or differences. Unsupervised learning aims at solving such problems as clustering, association rule learning, and dimensionality reduction. For instance, it can be applied at the data preprocessing stage to reduce data complexity.

Hyper parameter Tuning:-

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters.

However, there is another kind of parameters, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Some examples of model hyperparameters include:

- The penalty in Logistic Regression Classifier i.e. L1 or L2 regularization
- The learning rate for training a neural network.
- The C and sigma hyperparameters for support vector machines.
- The k in k-nearest neighbors.
- The aim of this module is to explore various strategies to tune hyperparameter for Machine learning model.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Two best strategies for Hyperparameter tuning are:

- GridSearchCV
- RandomizedSearchCV

Model Testing:-

The goal of this step is to develop the simplest model able to formulate a target value fast and well enough. A data scientist can achieve this goal through model tuning. That's the optimization of model parameters to achieve an algorithm's best performance.

One of the more efficient methods for model evaluation and tuning is cross-validation

Cross-validation: - Cross-validation is the most commonly used tuning method. It entails splitting a training dataset into ten equal parts (folds). A given model is trained on only nine folds and then tested on the tenth one (the one previously left out). Training continues until every fold is left aside and used for testing. As a result of model performance measure, a specialist calculates a cross-validated score for each set of hyperparameters. A data scientist trains models with different sets of hyperparameters to define which model has the highest prediction accuracy. The cross-validated score indicates average model performance across ten hold-out folds. Then a data science specialist tests models with a set of hyperparameter values that received the best cross-validated score. There are various error metrics for machine learning tasks.

SOFTWARE USED

IDE : Anaconda Jupyter

Programming Language : Python

HARDWARE REQUIREMENTS

PROCESSOR : Dual Core 2 Duos.

RAM : 4 GB DD RAM

HARD DISK : 250 GB

ALGORITHM / TECHNIQUE USED

Four regression models have been used in this study of COVID-19 future forecasting:

• Linear Regression

• LASSO Regression

• Support Vector Machine

• Exponential Smoothing

1) Linear Regression

In regression modeling, a target class is predicated on the independent features [14]. This method can be thus used to find out the relationship between independent and dependent variables and also for forecasting. Linear regression a type of regression modeling is the most usable statistical technique for predictive analysis in machine learning. Each observation in linear regression depends on

two values, one is the de- pendent variable and the second is the independent variable. Linear regression determines a linear relationship between these dependent and independent variables. There are two factors (x, y) that are involved in linear regression analysis.

The equation below shows how y is related to x known as regression.

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

or equivalently

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

Here, ε is the error term of linear regression. The error term here uses to account the variability between both x and y, β0 represents y-intercept, β1 represents slope. To put the concept of linear regression in the machine learning context, in order to train the model x is represented as input training dataset, y represents the class labels present in the input dataset. The goal of the machine learning algo- rithm then is to find the best values for β0 (intercept) and β1(coefficient) to get the best-fit regression line. To get the best fit implies the difference between the actual values and predicted values should be minimum, so this minimization problem can be represented as:

$$\text{minimize} 1\ n\ Xn\ i=1\ (predi - yi)\ 2 \quad (3)$$

$$g = 1\ \ n\ Xn\ i=1\ (predi - yi)\ 2 \quad (4)$$

Here, g is called a cost function, which is the root mean square of the predicted value of y (predi) and actual y (yi), n is the total number of data points.

2) LASSO

LASSO is a regression model belongs to the linear regres- sion technique which uses shrinkage [15]. Shrinkage in this context refers to the shrinking of extreme values of a data sample towards central values. The shrinkage process thus makes LASSO better and more stable and also reduces the error [16]. LASSO is considered as a more suitable model for multicollinearity scenarios. Since the model performs L1 regularization and the penalty added in this case is equal to the magnitude of. So LASSO makes the regres- sion simpler in terms of the number of features it is using. It uses a regularization method for automatically penalizing the extra features. That is, the features that cannot help the regression results enough can be set to a very small value potentially zero. An ordinary multivariate regression uses all the features

available to it and will assign each one a coefficient of regression. In contrast, the LASSO regression attempts to add them one at a time and if the new feature does not improve the fit enough to out-way the penalty term by including that feature then it could not be added meaning as zero. Thus the power of regularization by applying the penalty term for the extra features is that it can automatically do the selection for

us. Thus the models are made sparse with few coefficients in this case of regularization since the process eliminates the coefficients when their values are equal to zero. That means LASSO regression works on an objective to minimize the following:

$$Xn\ i=1\ (yi - Xj\ xij\beta j\ )\ 2 + \lambda\ Xp\ j=1\ |\beta j\ | \quad (5)$$

It sets the coefficient, which can be interpreted as min(sum of square residuals + λ |slope|), where, λ |slope| is penalty term.

3) Support Vector Machine

A support vector machine (SVM) is a type of supervised ML algorithm used for both regression and classification [17], [18]. SVM regression being a non-parametric technique de- pends on a set of mathematical functions. The set of functions called kernel transforms the data inputs into the desired form. SVM solves the regression problems using a linear function, so while dealing with problems of non-linear regression, it maps the input vector(x) to n-dimensional space called a feature space (z). This mapping is done by non-linear mapping techniques after that linear regression is applied to space. Putting the concept in ML context with a multivariate training dataset (xn) with N number of observations with yn as a set of observed responses. The linear function can be depicted as:

$$f(x)=x0\beta+b \quad (6)$$

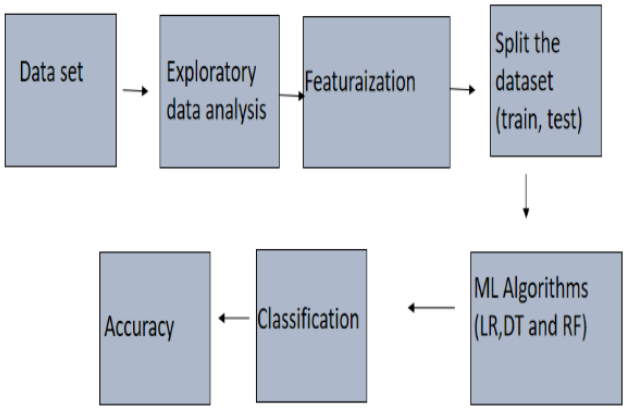


Fig-1: Back end module diagram

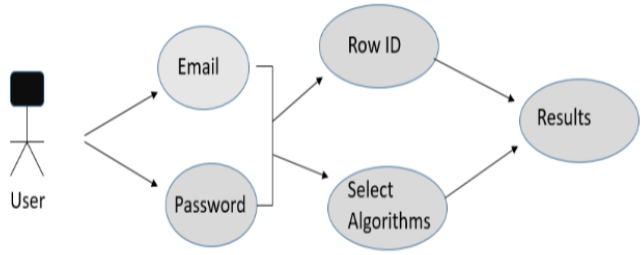


Fig -2: Front end module diagram

3. CONCLUSIONS

The precariousness of the COVID-19 pandemic can ignitea massive global crisis. Some researchers and government agencies throughout the world have apprehensions that the pandemic can affect a large proportion of the world pop- ulation [26], [27]. In this study, an ML-based prediction system has been proposed for predicting the risk of COVID- 19 outbreak globally. The system analyses dataset contain- ing the day-wise actual past data and makes predictions for upcoming days using machine learning algorithms. The results of the study prove that ES performs best in the current forecasting domain given the nature and size of the dataset. LR and LASSO also perform well for forecasting to some extent to predict death rate and confirm cases. According to the results of these two models, the death rates will increase in upcoming days, and recoveries rate will be slowed down. SVM produces poor results in all scenarios because of the ups and downs in the dataset values. It was very difficult to put an accurate hyperplane between the given values of the dataset. Overall we conclude that model predictions according to the current scenario are correct which may be helpful to under- stand the upcoming situation. The study forecasts thus can also be of great help for the authorities to take timely actions and make decisions to contain the COVID-19 crisis. This study will be enhanced continuously in the future course, next we plan to explore the prediction methodology using the updated dataset and use the most accurate and appropriate ML methods for forecasting. Real-time live forecasting will be one of the primary focuses in our future work

RESULT AND OUTPUT

This study attempts to develop a system for the future fore-casting of the number of cases affected by COVID-19 using machine learning methods. The dataset used for the study contains information about the daily reports of the number of newly infected cases, the number of recoveries, and the number of deaths due to COVID-19 worldwide. As the death rate and confirmed cases are increasing day by day which is an alarming situation for the world. The number of people who can be affected by the COVID-19 pandemic in different countries of the world is not well known. This study is an attempt to forecast the number of people that can be affected in terms of new infected cases and deaths including the number of expected recoveries for the upcoming 10 days. Four machine learning models LR, LASSO, SVM, and ES have been used to predict the number of newly infected cases, the number of deaths, and the number of recoveries

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

[4] 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.

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

[6] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," Climate research, vol. 30, no. 1, pp. 79–82, 2005

[7] R. Kaundal, A. S. Kapoor, and G. P. Raghava, "Machine learning tech-niques in disease forecasting: a case study on rice blast prediction," BMC bioinformatics, vol. 7, no. 1, p. 485, 2006.

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

BIOGRAPHIES

Mrs.Dayana R is currently an Assistant Professor in the Department of Computer Science and Engineering at Jeppiaar Institute of Technology, Chennai.

Ms.Bouna Das Y is currently pursuing her bachelor's degree in the field of Computer Science and Engineering at Jeppiaar Institute of Technology, Kanchipuram, Tamil Nadu, India. She did her schooling in Kanyakumari. she is particularly interested in ML and Data Base Management.

Ms.Nis shammini nimsha R is currently pursuing her bachelor's degree in the field of Computer Science and Engineeringat Jeppiaar Institute of Technology, Kanchipuram, Tamil Nadu, India. She did her schooling in Kanyakumari. She is particularly interested in ML.