iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** XII **Month of publication:** December 2021

DOI: https://doi.org/10.22214/ijraset.2021.39410

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

Linear Regression Algorithm in Machine Learning through MATLAB

Kalva Sindhu Priya
Department of EEE, Assistant professor, JNTU University, Hyderabad

Abstract: ***In the present scenario, it is quite aware that almost every field is moving into machine based automation right from fundamentals to master level systems. Among them, Machine Learning (ML) is one of the important tool which is most similar to Artificial Intelligence (AI) by allowing some well known data or past experience in order to improve automatically or estimate the behavior or status of the given data through various algorithms. Modeling a system or data through Machine Learning is important and advantageous as it helps in the development of later and newer versions. Today most of the information technology giants such as Facebook, Uber, Google maps made Machine learning as a critical part of their ongoing operations for the better view of users. In this paper, various available algorithms in ML is given briefly and out of all the existing different algorithms, Linear Regression algorithm is used to predict a new set of values by taking older data as reference. However, a detailed predicted model is discussed clearly by building a code with the help of Machine Learning and Deep Learning tool in MATLAB/ SIMULINK.***

Keywords: Machine Learning (ML), Linear Regression algorithm, Curve fitting, Root Mean Squared Error.

I. INTRODUCTION

Machine Learning [1] is said to be subset of Artificial Intelligence (AI) [2], [3] and also a branch of computational science which analysis, interpret the data (can be a form of anything like pattern reorganization, tracing, tracking of data) to improve a system and helps in making decisions with less or no human interference as shown in the Fig:1. If a particular algorithm is developed in Machine Learning, it suggests giving recommendations and decisions with the reference of input data and if any modifications are identified, the designed model should be capable to tune itself for better decision making until the algorithm output satisfies the known data. ML extends to numerous applications and advances like automation, improving client's experiences, manufacturing, health care in diagnosing of diseases through biological reference and life sciences, financial services in management, time series forecasting like Electrical load forecasting. Some of the other real time applications are cyber security- identification of threats to personal systems, personal information, self driving vehicle application in foreign countries, Digital advances and Artificial Intelligence applications like Replica, siri, Alexa, Google Assistant with the voice and text commands, Taxi applications like Uber, Ola which estimates the cost of trip based on timings, weather and location, Similarly, Email and fraud detection like automatically moving certain mails into spam folder, Image and Pattern recognization which is used in mobiles, advanced security cameras, Automatic medical test home devices like diabetic, pulse monitoring, Marketing applications like amazon, other shopping applications automatically sends recommended purchases as notification and many more.

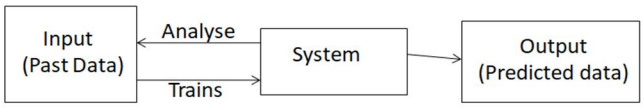


Fig: 1 Machine Learning Process

In order to design or train a model, there are many types of algorithms in Machine Learning and it is required to choose exact algorithm based on the application. On a whole, the main types of Machine Learning are supervised learning, unsupervised learning and Reinforced learning as shown in the Fig: 2 below.

Supervised learning [4-6] helps to solve real- world computational problem by supervising a system or data as a mentor. It is preferred to use when the data is labeled i.e., some predefined exact data with correct output is stored as training data in advance and produces correct output. For Example, In order to find whether the patient is covid positive or not, first it is required to load some correct training data of previous patients who have been infected such as age, temperature, level of throat infection, pulse rate, diabetic or non-diabetic, viral load etc.,

©IJRASET: All Rights are Reserved

All these data which has been loaded acts as training data and machine develops a own model. Now, a new patient biological values which is acting as testing data is to be incorporated to a newly developed model to predict whether the person is infected or not. Based on the biological and medical values, the model predicts the correct output. If not, once again train a model with more datasets and redesign it for better result with fewer errors. Similarly, this type of algorithm is applied to true or false categorization like fault taken place- True or False, switching states in electronic converters [7] - ON or OFF. There is further classification in supervised learning such as Linear Regression [8], Logistic Regression, Classification, Naïve bytes classification, k- Neural Network, Decision trees, Support Vector machine algorithm.

Unsupervised learning [9] is opposite to supervised learning in which there is no supervisor along with no labeled data, no classification, no data points. The concept of unsupervised learning is that it works on identification of patterns, classification and labeling within the provided datasets based on the differences and similarities which fed earlier. For Example, this type of learning algorithm is mostly used to recognize patterns and figures. Suppose if some group of person's features is loaded on security cameras or confidential biometric software system, the machine will learn features of each person. Now for an instance, if a person whose features are already been loaded on the system would appear after a long time with different outfit and appearance, still a model can recognize the person based on unsupervised learning. The two important concepts in unsupervised learning are clustering- where one would go for grouping persons based on skills and identification. And the other is association- when one need for division or segregation of rules. The other available algorithms in unsupervised learning are Exclusive clustering, Agglomerative clustering, overlapping, probabilistic clustering, Hierarchical clustering, k-means clustering, Principal component analysis, Singular value decomposition, Independent component analysis.

Reinforced learning [10] is different from both supervised and unsupervised learning. It makes to identify a suitable solution in a maximum way to avail the rewards or punishments in between the paths by an agent. The aim of reinforced learning is to find a best possible ways out of all available ways to reach a target. Some of the applications in reinforced learning are gaming, Robot automation, Navigation system, Actuation system [11], [12], Stock trading etc. The available algorithms in reinforcement learning are Q-learning, R learning, TD learning.

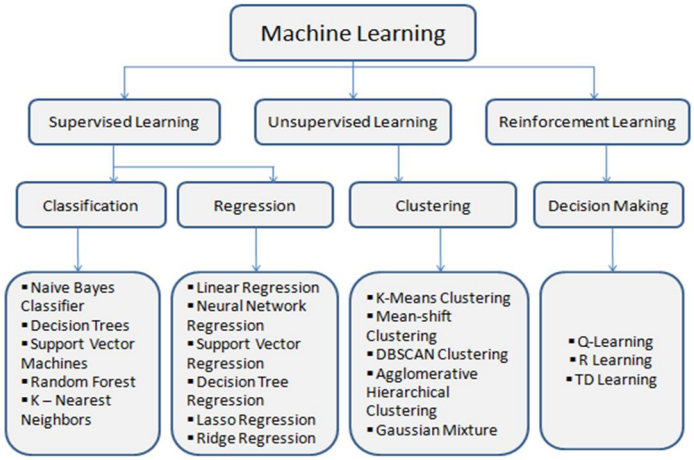


Fig: 2 Classification of Machine Learning algorithms

II. LINEAR REGRESSION ANALYSIS

Linear Regression Analysis [13] is a statistical model or statistical approach which establishes a relationship between two variables (One act as Dependent Variable and the other acts as Independent Variable) as shown in the Fig: 3. From the Fig: 3, it is noticed that the slope or characteristics of given regression line steeply increasing linearly and can be known as positive regression line and for suppose, if the regression line tend to have steeply decreasing characteristic then it is said to be negative regression line. It is a predictive modeling technique to determine the relationship between two variables like trend forecasting, forecasting an effect, determination of economic growth, price determination of a products, house sales, rain and weather prediction, score predictions, predictions of covid vaccinations etc., The variations in linear Regression model are Simple linear regression, multiple linear regressions, polynomial or non linear regression. Linear Regression (LR) mainly depends on two factors- which variables acts as predictors for an outcome and How much all the predictions made are accurate.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

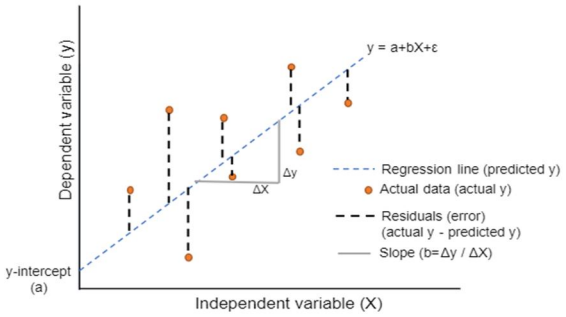


Fig: 3 Linear Regression plot

If the predictions are done with only one single variable then it is treated as simple linear regression whose expression is given below which is also called as Hypothesis equation and the same analysis is presented in this paper.

$$Y=a+bX+\varepsilon$$

Where, Y is the response or output or Dependent variable, a is the intercept, b is slope of linear regression line, X is independent variable and ε is the error or residual of model.

Similarly, if the same predictions are carried out for more than one variable, then it is referred as multiple linear regressions and the expression goes as follows:

$$Y=a+bX_1+cX_2+dX_3+\ldots\ldots\ldots\ldots+\varepsilon$$

Where, Y is the response, X_1, X_2, X_3 and b, c, d are independent variables and their slope respectively as it has multiple regression lines, a is the intercept and ε is sum of residual errors calculated for all regression lines. The most common factor in both simple and multiple linear regression lines is error 'ε'. The error should be minimum as such as it can, as it may result to better accurate model. Certain mathematical methods are adopted to reduce the error. Some of the techniques include Root Mean Squared Error (RMSE), Minimum Squared Error (MSE), Minimum Absolute Error (MAE), R squared, Ordinary least squares method, Sum of absolute errors, Gradient descent method. Out of all these methods, the most common and comfortable method is RMSE method which is the root of squares of difference between predicted and true values of a model and the same technique has been carried over in this paper. However, the equations for calculating errors in different methods and their formulae is as listed below.

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Ypred-Ytrue)\,2}$$

Where, N is the number of observations or iterations to calculate error, Y_{pred} is the predicted values of dependent values and Y is true or actual value. True values are the values which are fed as input to trained model and predicted values are the values obtained after performing LR analysis.

Mean Absolute Error (MAE) is expressed as the difference between predicted and true responses and Mean Squared Error (MSE) is defined as the squares of difference between predicted and true responses which are given below.

$$MAE=\frac{1}{N}\sum_{i=1}^{N}(Ypred-Ytrue)$$

$$MSE=\frac{1}{N}\sum_{i=1}^{N}(Ypred-Ytrue)^2$$

The coefficient of R squared gives the difference in variance with dependent variables. There is no scale in measuring these error and generally the value of this error lies within the unity, irrespective of size of data. Similarly, Adjusted R squared is newest version of R square and it is adjusted to dependent variables number in a model and generally less than R Squared error.

$$R^2=1-\frac{\text{Residual Sum of squares (RSS)}}{\text{Total Sum of squares (TSS)}}$$

$$\text{Adjusted } R^2=1-\{\frac{(1-R^2)(n-1)}{n-k-1}\}$$

Where, n is the number of observations, k is number of dependent variables in the dataset. The low the value of calculated residual value, the high the accuracy is. The accuracy of LR model not only depends on calculation of RMSE, but also the slope of regression line- how best the line is fitted with the available data points and it can be altered by curve fitting tool or using optimization techniques on MATLAB.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue XII Dec 2021- Available at www.ijraset.com

III. LR THROUGH MATLAB

All the Machine Learning algorithms can be executed or analyzed through different platforms. Machine learning using python [14], [15] for different applications and case studies were discussed in most of the research papers. Here, in this paper, LR algorithm is executed and implemented through MATLAB as it is offers compatibility and flexibility for non computer science persons too. However, the workflow of Linear Regression model in MATLAB is as shown in the below steps.

1) *Step: 1.* Import required dataset into workspace of MATLAB.
2) *Step: 2.* Explore data- According to LR algorithm divide the data into training set and testing set. Training set is used to achieve objective of function by building desired algorithm. The algorithm will analyze the training datasets, classifies and understands repeatedly on its own and try to develop a model which is enough sufficient to meet the desired outcome. After building a model, testing data is fed to model to make accurate predictions. Generally, it is recommended to have more than half of data (As example: 85 percent of data or 90 percent of data) for training set for accuracy and of course the remaining data is treated as testing data. These divisions in data set can be done by importing dataset to workspace and then by coding as shown in the Fig.4.
3) *Step: 3.* based on the training set, train a model on MATLAB by clicking on “Regression Learner” in ‘Machine Learning and Deep Learning’ under Apps which can be seen on tool bar of MATLAB.
4) *Step: 4.* Attain regression line characteristic by following given steps.

Click on ‘New Session’ and chose a variable for which one need to get regression characteristics. Next, Choose a cross validation (generally preferred) and click on’ Start Session’.

Validation data incorporates new data into the model which are not evaluated earlier, so that one can understand how well the model is predicted based on new data. Generally, Cross validation is used when there is a smaller data and Holdout validation is used when there is big data.

```
Independent=LinearRegressionSheet1.X
Dependent=LinearRegressionSheet1.Y
figure
plot(Independent,Dependent)
grid on
title('Actual plot')
xlabel('Number of observations')
ylabel('Actual value')

N=floor(0.9*numel(Dependent))
Xtrain=Dependent(1:N)
Ytrain=Dependent(2:N+1)

Xtest=Dependent(N+1:end-1)
Ytest=Dependent(N+2:end)
lindata=[Xtrain Ytrain]


trainedmodel=trainRegressionModel(lindata)
yp=trainedmodel.predictFcn(Xtest);
rmse=sqrt(mean((yp-Ytest).^2))

Ntest=numel(Xtest);
figure
plot(Dependent(1:N))
hold on
idn=N:(N+Ntest);
plot(idn, [Dependent(N); yp], '.-')
title('predicted graph')
xlabel('Number of observaations')
ylabel('Actual and predicted values')
legend(["Actual" "Predicted"])
```

Fig: 4. Coding on MATLAB

©IJRASET: All Rights are Reserved

5) *Step: 5.* The model can now able to give simple linear regression line by clicking on 'All quick to train', so that the given data runs through different techniques like Interaction linear, robust linear, stepwise linear, all linear, Gaussian process and regression models along with the display of calculated errors of each method at the left side of the screen. MATLAB automatically chooses a better technique by reading all the calculated Root Mean Squared errors of each and every technique as shown in Fig: 5 below.
6) *Step: 6.* The trained model can give different plots such as Response plot, Residual plot, Predicted vs actual plot, Minimum MSE plot. If most of the data points are lying on regression line, then the model is said to be perfect prediction.

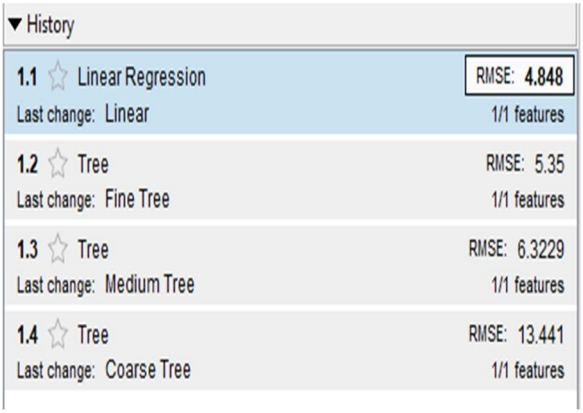


Fig: 5. RMSE calculation

7) *Step: 7.* Predict the test results by incorporating some testing data into trained model and observe the predictions and resultant errors. Visualize the strength of errors, measure accuracy and analyze the model whether the predicted data is satisfactory or not. If the results are not remarkable, one can train the model once again by incorporating lot of datasets, new datasets or train a model and work for better regression line by optimization techniques.
8) *Step: 8.* The above discussed workflow and the trained model can be exported by generating a live script for further results advances, tuning and optimizing the line of regression by curve fitting tool on MATLAB.

IV. RESULTS

In this paper, in order to learn Linear Regression algorithm, it is required to import dataset on MATLAB as discussed in the previous sections. In the proposed paper, required linear regression data set is downloaded from kabble.com [16] - which offers different and wide range of customizable datasets from a data analyst programmer. The supervised learning workflow which has been discussed in earlier sections is carried forward for results. The actual plot for number of observations taken as Independent variables and the true values are as shown in below Fig: 6. Train a model with by taking training data as reference and check the trained model with testing data. The response plot of both actual and predicted data is as shown in the Fig: 7, where, blue points are actual data and the yellow indicates predicted data and even the errors can also be shown which is the square of distance between actual and predicted data.

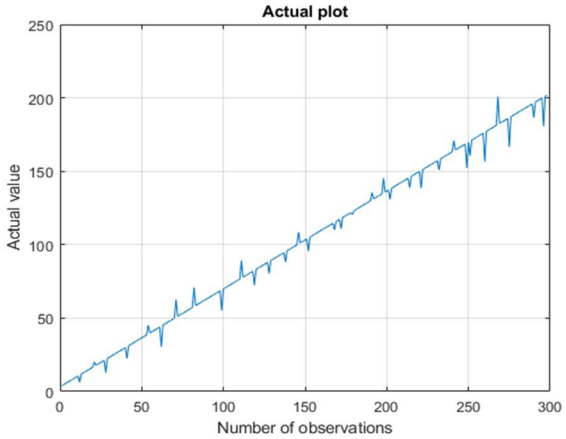


Fig: 6. Actual Plot

©IJRASET: All Rights are Reserved

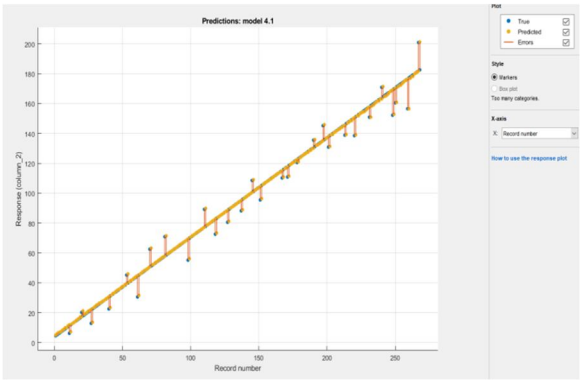


Fig: 7. Response plot with RMSE errors

After training a model, Regression learner also displays the calculated errors like RMSE error, R-squared, Minimum Squared Error (MSE), Mean Absolute Error (MAE) as shown in the following Fig: 8. Similarly, actual vs. predicted plot is shown in Fig:9 for the perfect prediction. Though R- squared error is less than the other errors, it is recommended to use mostly RMSE value for regression model as this type of error is fitted exactly and moreover, it is noted that R squared error value should be always less than unity irrespective of the data and model.

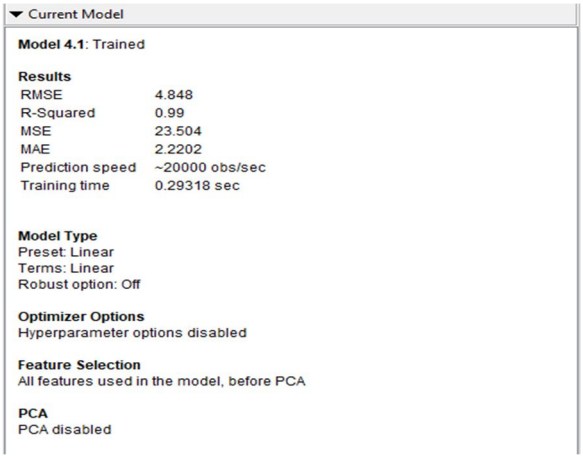


Fig: 8. Trained model calculated parameters

Linear Regression algorithm also predicts a future data which can be used in most of the applications like Electrical load forecasting, weather forecasting. Predicted values of dependent variable are also shown in the Fig: 10 below along with the true data.

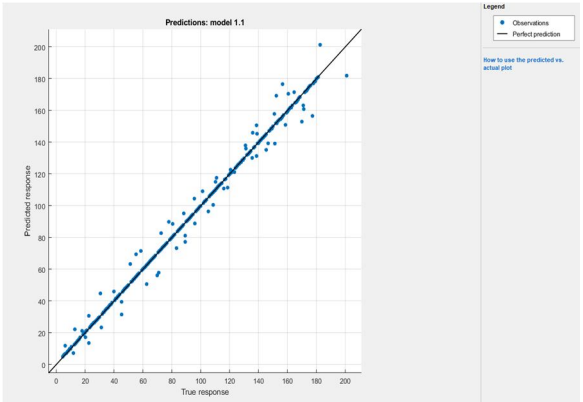


Fig: 9. Actual vs. Predicted plot

©IJRASET: All Rights are Reserved

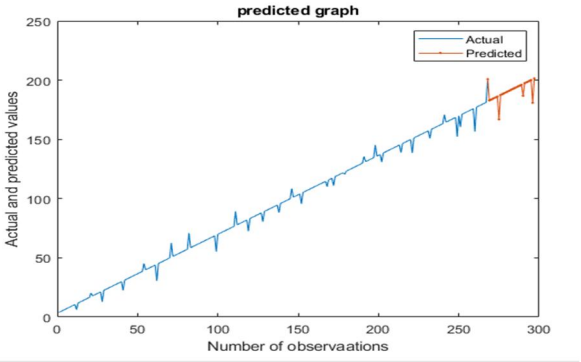


Fig: 10. Predicted/ forecasted values

V. CONCLUSION AND FUTURE SCOPE

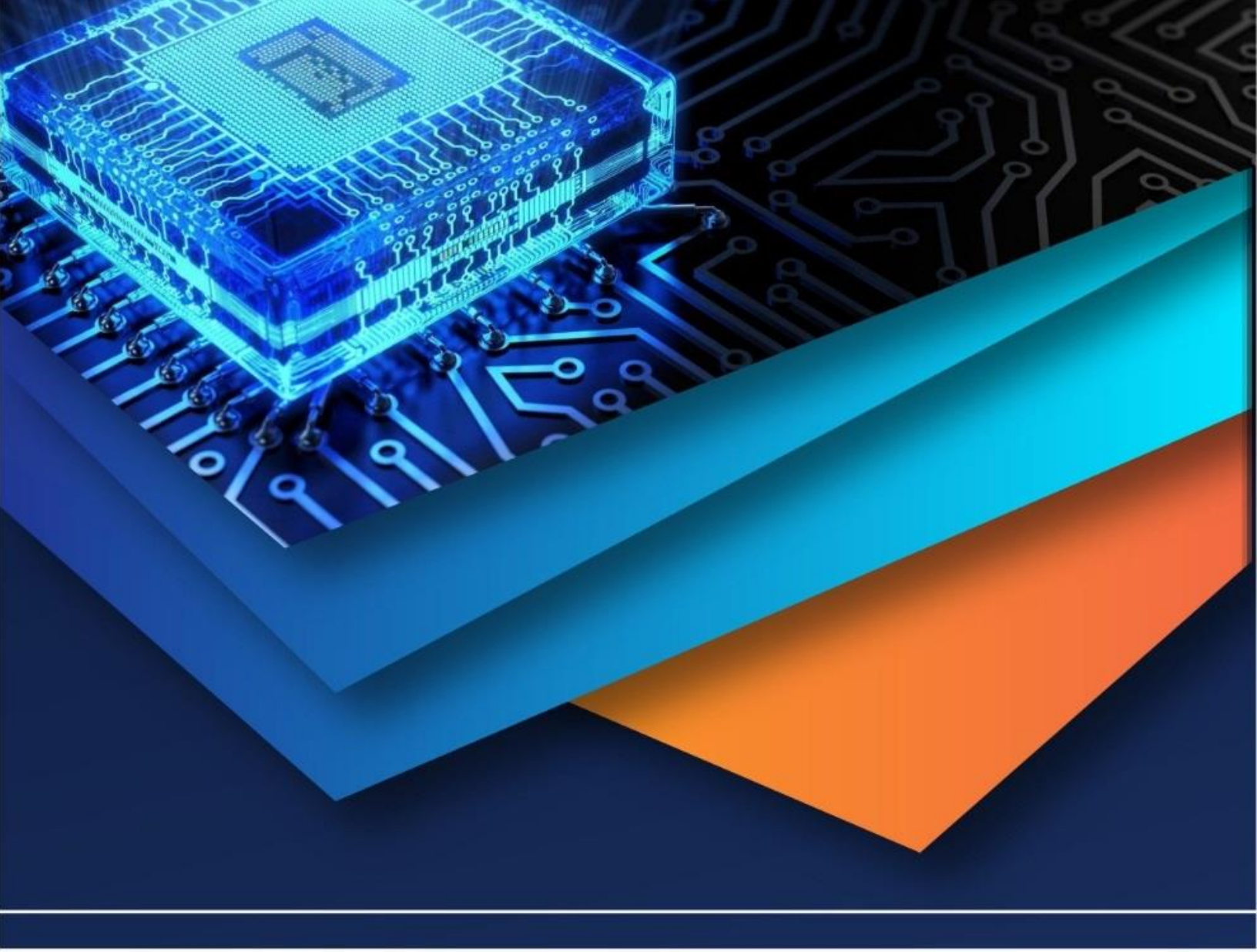
Machine Learning is an advancement or later version of Artificial Intelligence which is rapidly growing field in the streams of automation, designing, prediction, Image and sound processing. It also offers real time and practical information along with the security with wide range of data. The paper herein discussed about Linear Regression algorithm in Machine learning by importing a available dataset from kabble.com. This type of algorithm is preferred to use when the data is in linear structure so that it yields better accuracy and better prediction. The complete analysis which was presented in this paper is done in MATLAB and the predicted results were also given. By doing this, one can go for forecasting of dependent variables with respect to independent variables which gives a clear idea to optimize as well as tuning purposes.

The same model can be exported and correlate to numerous applications like weather and rainfall prediction, Electrical load forecasting, Trends forecasting, and even can be applied to national demonstrations like Total Economic income of various countries, Stock market prediction, market- sales predictions, banking and finance stream, Entertainment purposes in gaming, advancement in technology like robotics, quantum Computing, Big data and many more.

REFERENCES

[1] Zhang XD. (2020) Machine Learning. In: A Matrix Algebra Approach to Artificial Intelligence. Springer, Singapore. https://doi.org/10.1007/978-981-15-2770-8_6.
[2] Arel I,Rose D C,Karnowski T P., “Deep machine learning-A new frontier in artifical intelligence esearch[J]”, Computational Intelligence Magazine, IEEE,2010,5(4):13-18.
[3] Elemasetty Uday kiran, “ Usage of neural networks in communication links with structural inverted vee antenna ”, International journal of engineering Research and applications (IJERA), vol.8, no.9, 2018, pp. 65-69.
[4] Leonidas Akritidis, Panayiotis Bozanis. (2013) A supervised machine learning classification algorithm for research articles. In Proceedings of the 28th Annual ACM Symposium on Applied Computing, Coimbra, Portugal.
[5] Vladimir Nasteski, “An overview of the supervised machine learning methods”, Research gate, DOI: 10.20544/HORIZONS.B.04.1.17.P05 December 2017.
[6] Nagaraju Kolla, M. Giridhar Kumar, “Supervised Learning Algorithms of Machine Learning: Prediction of Brand Loyalty”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-11, September 2019.
[7] M. Vanisri , K. Sindhu Priya , G. Chandra Shekar, “Comparison of Level Shifting Modulation Techniques using Designed Seven Level Multilevel Inverter”, International Journal of Engineering Research & Technology (IJERT) Volume 10, Issue 03, March 2021.
[8] Shen Rong, Zhang Bao-wen, “The research of regression model in machine learning field”, MATEC web of conferences 176, 010113, 2018.
[9] Memoona Khanam, Tahira Mahboob, Warda Imtiaz, Humaraia Abdul Ghafoor, “A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance”, International Journal of Computer Applications, june 2015.
[10] Ahmad Hammoudeh, “A Concise Introduction to Reinforcement Learning”, Research gate February 2018.
[11] Suresh Kumar B., Ravi Kumar B.V., Sindhu Priya K. (2019) Modeling and Simulation of Dual Redundant Power Inverter Stage to BLDCM for MEA Application. In: Saini H., Singh R., Kumar G., Rather G., Santhi K. (eds) Innovations in Electronics and Communication Engineering. Lecture Notes in Networks and Systems, vol 65. Springer, Singapore. https://doi.org/10.1007/978-981-13-3765-9_18.
[12] Elemasetty Uday kiran, “Imaginary axis on logarithmic with singularity transformation hyperbolic function in Arithmetical equations”, Quest journal of research in apllied mathematics, vol.05, n0o. 02, 2018, pp, 29-33.
[13] Sankranti Srinivasa Rao, “Stock Prediction Analysis by using Linear Regression Machine Learning Algorithm”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-9 Issue-4, February 2020.
[14] Sebastian Raschka, Joshua patterson, corey nolet, ‘Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence, Information (Switzerland), April 2020.
[15] Pinky sodhi, Naman Awasthi, Vishal Sharma, “Introduction to Machine Learning and its basic application in phyton”, proceedings of 10th International conference on Digital strategies for Organizational Success, April 2019.
[16] https://www.kaggle.com/tanuprabhu/linear-regression-dataset.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)