Grenze International Journal of Engineering and Technology, Jan Issue

Comparative Analysis of Parameters using Regression Supervised Machine Learning on Dataset

Saurabh Shrivastava[1], Dr. Jitendra Singh Kushwah[2] and Dr. Aditya Vidyarthi[3]
[1-3]Institute of Technology and Management Gwalior-MP- India
Email: saurabhshrivastava.it@itmgoi.in, jitendra.singhkushwah@itmgoi.in, aditya.vidyarthi@itmgoi.in

***Abstract*—Machine learning Calculations can provide logic for how to accomplish important jobs by adding up examples. The goal of this study is to contrast several machine learning algorithms. Both experience and theory-based learning can be used in machine learning. The most common regression algorithms, such as linear regression, were utilized in this paper, and the effectiveness of the algorithms was tested using different sizes of training and testing datasets. Regression performs better than the other approaches by a significant margin, as shown by a relative study of the classifiers.**

***Index Terms*— Machine learning algorithms, Regression, Dataset, Training Testing.**

I. INTRODUCTION

Machine learning is a model that may be used to learn from the past (in this case, historical data) to enhance performance in the future. This field focuses on reflex learning techniques for individuals. Learning refers to the modification or improvement of an algorithm based on past "knowledge" automatically without any outer help from the human.[1]

According to Arthur Samuel, the scientific subject that enables computers to learn without explicit programming is known as machine learning. It was commonly known that Arthur Samuel had a checkers-playing program. Machine learning (ML) teaches computers how to handle data more skillfully. Sometimes, even after carefully examining the data, we are unable to discern any useful information. We then use machine learning in that situation. The availability of numerous datasets has increased the demand for machine learning. Machine learning is used by many firms to extract the right data. Machine learning is persistent in learning from the data. Numerous mathematicians and programmers use a variety of techniques to solve this challenge, which involves handling enormous amounts of data.[3] as shown in fig 1.

A subset of machine learning and artificial intelligence is supervised learning, commonly referred to as supervised machine learning. It is distinguished by the way it trains computers to accurately classify data or predict outcomes using labeled datasets. The model modifies its weights as input data is fed into it until the model has been properly fitted, which takes place as part of the cross-validation process. Such as classifying spam in a different folder from your email, supervised learning assists enterprises in finding scalable solutions to several real-world issues. A subset of machine learning is supervised learning. It makes use of tagged data and a trained algorithm to classify the data and forecast the results. Adjust the weight of the data as it is put into this model so that it fits properly. Assisted learning assists in resolving issues in the actual world.[4]

Supervised learning is a sort of machine learning in which the output is predicted by the machines using well-

Grenze ID: 01.GIJET.9.1.49_2

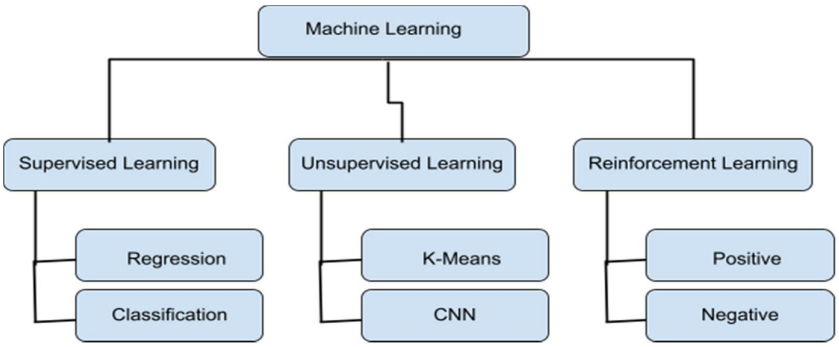


Fig. 1. Machine Learning Techniques

labeled training data that has been used to train the machines. The term "labeled data" refers to input data that has already been assigned the appropriate output. In supervised learning, the training data that is given to the computers serve as the supervisor, instructing them on how to correctly predict the output [5].

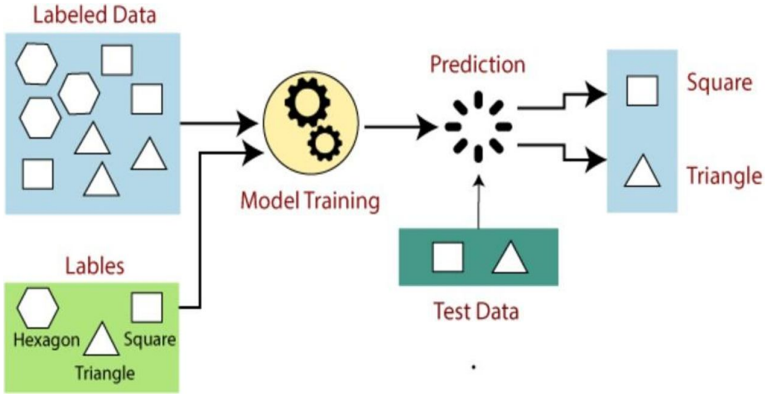


Fig. 2. Prediction Process of data

Supervised learning can be further divided into two types of problems:

A. Regression

Regression is a technique for figuring out how independent features or variables relate to a dependent feature or result. Once the link between the independent and dependent variables has been estimated, outcomes may then be predicted. Regression is a statistical study area that is essential to machine learning forecast models. It is useful for forecasting and predicting outcomes from data since it is used as a method to predict continuous outcomes in predictive modeling. Regression using machine learning often entails drawing a line of best fit through the data points. To obtain the best fit line, the distance between each point and the line is minimized. Regression algorithms are used if there is a relationship between the input variable and the output variable. It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.[6]
The Classification Learning algorithm is of the following types:

Linear regression: Predictive analysis is done using linear regression. The link between the criteria, or scalar response, and the several predictors, or explanatory factors, is modeled using a linear approach called linear regression. The conditional probability distribution of the response given the values of the predictors is the main focus of linear regression. Overfitting could occur in linear regression. The formula for linear regression [10] is:

$$Y = bX + a \qquad (1)$$

Polynomial regression: Polynomial regression is used for curvilinear data. Polynomial regression is fit with the method of least squares. The goal of regression analysis is to model the expected value of a dependent variable y regarding the independent variable x. The equation for polynomial regression is [6]:

$$\beta_0 + \beta_0 x_1 + \varepsilon \qquad (2)$$

Stepwise regression: To fit regression models to prediction models, stepwise regression is utilized. The process happens automatically. The variable is either added to or deleted from the set of explanatory variables at each phase. Forward selection, backward elimination, and bidirectional elimination are methods for stepwise regression. The formula for stepwise regression is [6].

$$b_{j.std} = b_j (S_x * S_y^{-1}) \quad (3)$$

Ridge regression: A method for examining data from multiple regression is ridge regression. Estimates based on least squares are unbiased in the presence of multicollinearity. Ridge regression lowers the standard errors because it adds some bias to the regression estimates. The ridge regression equation is [6]:

$$\beta = (X^T X + \delta * I)^{-1} X^T y \quad (4)$$

Lasso regression: The regression analysis technique known as lasso regression includes regularisation and variable selection. Soft thresholding is used in lasso regression. Only a portion of the available covariates is chosen for use in the final model through lasso regression. Using a lasso regression [6]:

$$N^{-1} \sum_{i=1}^{N} f(x_i, y_I, \propto, \beta) \quad (5)$$

ElasticNet regression: The lasso and ridge methods' penalties are combined linearly in ElasticNet regression, a regularised regression technique. Support vector machines, metric learning, and portfolio optimization all use elastic net regression. The penalty function is given by [6]:

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j| \quad (6)$$

B. Classification

When the output variable is categorical, meaning there are two classes, such as Yes-No, Male-Female, True-False, etc., classification methods are utilized. Classification Algorithms can be further divided into the Mainly two category:

Linear Models

Logistic Regression: The probability of an outcome, which can only have Boolean values, is predicted via logistic regression. Both numerical and categorical data are used to make the prediction. For two reasons, a linear regression cannot be used to predict the value of a binary variable. The values cannot be predicted using linear regression with an acceptable range. The residuals for dichotomous experiments won't be regularly distributed around the expected line because there is only one of two potential values for each experiment. However, a logistic curve that is restricted to values between 0 and 1 is produced via logistic regressions. Similar to a linear regression, a logistic regression builds its curve using the natural logarithm rather than the probability. Additionally, each group will have equal variances or the predictors' distribution will be atypical.[2]

Support Vector Machines: Information is classified using a Support Vector Machine (SVM) by locating the hyperplane that is maximized and utilized as the margin between the two classes.[2]

- Create various hyperplanes before deciding which one is appropriate.
- The kernel trick for non-linear hyperplane is utilized in SVM approach for misclassifications of the linear hyperplane.
- Optimize the hyperplane with the maximum margin between the classes.
- For a high-dimensional space, we rephrase the issue so that the data is implicitly mapped to this space.

Non-linear Models

K-Nearest Neighbours: K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition based on their nearest neighbors and it should be an odd number. It uses the distance factors like Euclidean, Manhattan, Minkowski, etc [6]. Algorithm K nearest neighbors measured by a distance function like Euclidean, Manhattan, Minkowski, etc. Most of the time we took K as an odd number only for making the decision. If K = 1, then simply assigned it to the class of its nearest neighbor. Or If K is an odd number, then assigned to the class of maximum votes of its nearest neighbor [2].

Naïve Bayes: Bayes' theorem is dependent on Naive Bayesian classifiers with independent assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Regardless of its simplicity, the Naive Bayesian classifier often outperforms more sophisticated classification methods and often does surprisingly well, and is widely used. Algorithm Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x),

and P(x|c). The naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.[2]

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (7)$$

Where,
$P(c|x)$ – Posterior Probability
$P(x|c)$ – Likelihood
$P(c)$ – Class Prior Probability
$P(x)$ – Predictor Prior Probability

II. Literature Review

Kapil Sethi, Ankit Gupta, Gaurav Gupta et. al. according to this machine learning gives the possibilities of computer learning to play games , Mathematical statistics, statistical computational principles, and develop algorithms used for text interpretation, and pattern recognition. This paper focuses on explaining the concept and evaluation of Machine learning. This paper also demonstrates the fields or areas where Machine learning is used.
Shagan Sah explains various machine learning approaches described in this study. Numerous applications, such as data classification, prediction, or pattern recognition, use these techniques. By using relevant data to build an algorithm, machine learning aims to automate human help. This article should also serve as a glossary of numerous machine learning terms for quick access.
Rabi Narayan Behera et. al. explain that, in contrast to Statistics, which was a branch of mathematics that only occasionally considered computational approaches, machine learning (ML) has developed into an independent research field that has not only established the foundation for statistical-computational principles of learning procedures but has also created several algorithms that are frequently used for text interpretation, pattern recognition, and other tasks**.** This paper aims to compare the three most prominent machine learning algorithms based on certain fundamental ideas, as well as to explain the concept and evolution of machine learning.
R. Vijaya Kumar Reddy et. al. use a categorical target or categorical class variable to predict related information is known as classification. This method can be applied to any kind of statistical data. These algorithms are employed for several tasks, including image classification, predictive modeling, data mining, etc. Building a straightforward and unambiguous model of how class labels are distributed about predictor features is the basic goal of supervised learning. The classifiers are then used to compare the value of the class label, which is unknown, to the class labels of the testing cases, where the values of the predictor characteristics are known. We present examples of various classification methods used in supervised machine learning in this study.
Nagaraju Kolla et. al. explain through supervised learning algorithms for classifications, including logistic regression, decision trees, support vector machines, the Bayes algorithm, and the K-nearest neighbours (KNN) algorithm, the current study investigates the topic of predicting brand loyalty. The data set contains factors such as loyalty status, gender, family size, age, frequency of purchase, and FMCG purchase from a sample of 265 FMCG customers. Python tools like Pandas (Data analysis), Numpy (Numerical calculation), Matplotlib (Visualization), and Sklearn have been used to analyze data (Modeling). Logistic regression has fared better than other methods for supervised classification.

III. Problem Statement

Analysis and compare regression supervised machine learning algorithms based on different datasets. Datasets can be accessed from UCI Machines Learning Repository Keggale, Github. We have taken different data sets from different sources. In this analysis, we have many parameters to analyze these data sets. In this paper, we analyze regression algorithms based on different parameters which are the size of training and testing data, accuracy, MSE (Mean Squared Error), and RMSE (Root Mean Squared Error). Our problem is that we have to use different training and test data sizes using a supervised machine learning technique that is regression and find the value of the above-mentioned parameters.

IV. Result and Discussion

Python was used in the regression and analysis of the regression machine learning algorithms using different sizes of training and testing datasets. Table 1 shows the dataset with 11 attributes where ‘Turnover’ is the dependent variable and remaining are independent variables.

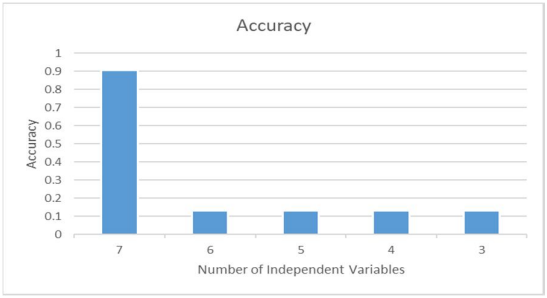
TABLE I. DATASET

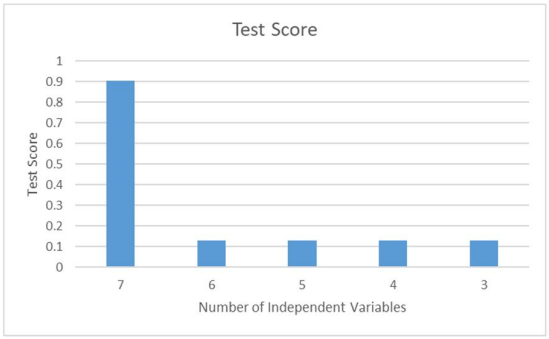
	Date	Symbol	Prev Close	Open	High	Low	Last	Close	VWAP	Volume	Turnover
0	1/3/2000	HDFCBANK	157.40	166.00	170.00	166.00	170.00	170.00	169.52	33259	5.638122e+11
1	1/4/2000	HDFCBANK	170.00	182.00	183.45	171.00	174.00	173.80	174.99	168710	2.952261e+12
2	1/5/2000	HDFCBANK	173.80	170.00	173.90	165.00	168.00	166.95	169.20	159820	2.704094e+12
3	1/6/2000	HDFCBANK	166.95	168.00	170.00	165.30	168.95	168.30	168.44	85026	1.432166e+12
4	1/7/2000	HDFCBANK	168.30	162.15	171.00	162.15	170.75	168.35	166.79	85144	1.420158e+12
...	...	...	...	...	...	...	...	...	...	...	...	...
5301	4/26/2021	HDFCBANK	1414.15	1413.00	1429.00	1402.75	1407.55	1404.80	1413.19	15085476	2.131861e+15
5302	4/27/2021	HDFCBANK	1404.80	1407.25	1442.00	1404.80	1435.05	1438.70	1430.40	10296453	1.472810e+15
5303	4/28/2021	HDFCBANK	1438.70	1436.25	1479.00	1431.00	1475.00	1476.80	1463.19	12051970	1.763438e+15
5304	4/29/2021	HDFCBANK	1476.80	1486.20	1503.65	1461.00	1471.65	1472.50	1481.15	12039276	1.783196e+15
5305	4/30/2021	HDFCBANK	1472.50	1445.00	1453.80	1407.50	1412.90	1412.30	1421.13	17616451	2.503529e+15

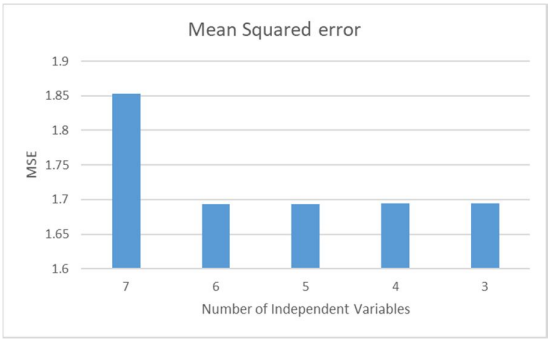
Table II: shows the results of different parameters based on different independent attributes where the training size is 80% and the test size is 20%. It shows that the number of independent variables is increasing and the results are better than previous.

TABLE II. TRAINING AND TEST SIZE DATA RATIO 80:20

Parameters	No. of Independent Variable				
	7	6	5	4	3
Accuracy	0.9049	0.1313	0.1312	0.1310	0.1305
Training Score	0.9246	0.0949	0.0942	0.0932	0.0922
Test Score	0.9049	0.1313	0.1312	0.1310	0.1305
MSE	1.8527	1.6934	1.6937	1.6941	1.6950
RMSE	13611	41152	41155	41159	41171







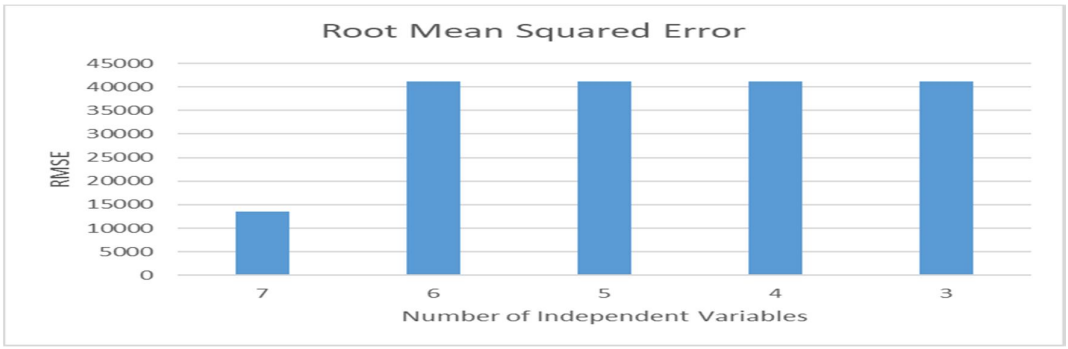
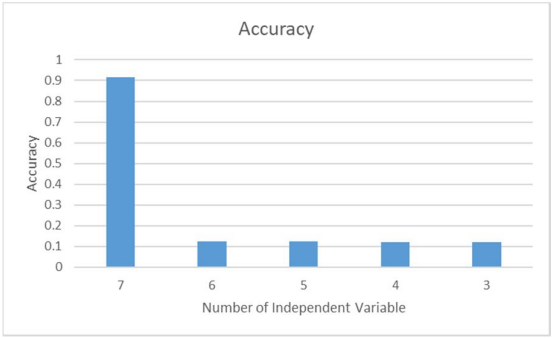


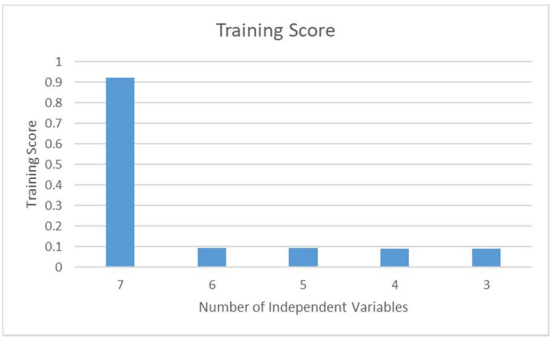
Fig. 3. Graphical Representation of table II

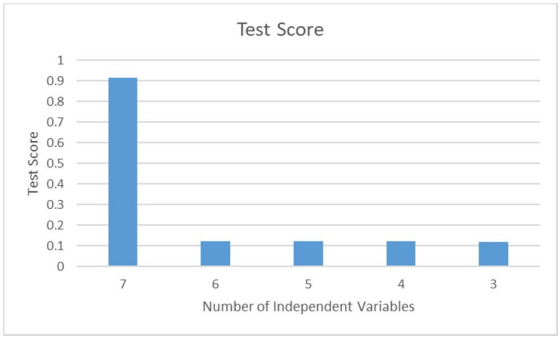
Table III: shows the results of different parameters based on different independent attributes where the training size is 70% and the test size is 30%.

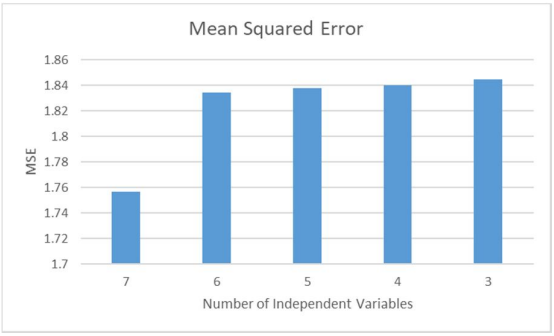
TABLE III. TRAINING AND TEST SIZE DATA RATIO 70:30

Parameters	No of Independent Variable				
	7	6	5	4	3
Accuracy	0.9161	0.1242	0.1228	0.1215	0.1194
Training Score	0.9199	0.0937	0.0931	0.0908	0.0907
Test Score	0.9161	0.1242	0.1228	0.1215	0.1194
MSE	1.7569	1.8346	1.8377	1.8403	1.8447
RMSE	13254	42833	42868	42899	42950









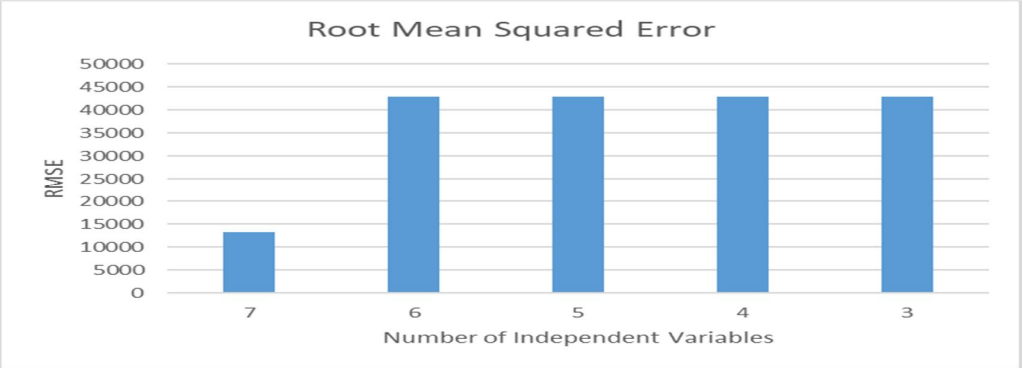


Fig. 4. Graphical Representation of table III

V. Conclusion

Machines have played a significant influence in our lives along with current technological advancements. Every aspect of our life generates a large amount of data, and this amount keeps growing. These data are utilized very effectively because of the machines. Despite the misconception that these devices are only utilized in the domains of engineering and computer science, they are present in all aspects of daily life.

This paper contains regression techniques used in supervised machine learning algorithms. A Regression is a method of predicting similar information from numerical datasets. Nowadays machine learning algorithms are becoming more popular for regression problems. This paper gives an introduction to supervised machine learning algorithms used for the regression of continuous variables to analyze the result.

References

[1] Kapil Sethi, Ankit Gupta, Gaurav Gupta, Varun Jaiswal, 2019 “Comparative Analysis of Machine Learning Algorithms on Different Datasets” , International Conference on Innovation in Computing, ICIC 2019.

[2] R. Vijaya Kumar Reddy, Dr. U. Ravi Babu, 2018, “ A Review on Classification Techniques in Machine Learning”, International Journal of Advance Research in Science and Engineering IJARSE, Volume no 7, Special issue no 3.

[3] Shagan Sah “ Machine Learning: A Review of Learning Types” , Preprint 2020.

[4] 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.

[5] Rabi Narayan Behera, Kajaree Das, “A Survey on Machine Learning: Concept, Algorithms, and Applications “ Vol. 5, Issue 2, February 2017.

[6] Aviral Gupta, Akshay Sharma, Dr. Amita Goel, “Review of Regression Analysis Models” International Journal of Engineering Research & Technology (IJERT), Vol. 6 Issue 08, August 2017.

[7] Özer Çelik, “ A Research on Machine Learning Methods and Its Applications”, Journal of Educational Technology and Online Learning · September 2018.

[8] Gulden Kaya Uyanil, Nese Guler,” A study on multiple linear regression analysis “, Procedia - Social and Behavioral Sciences 106 (2013) 234 – 240.

[9] Osisanwo F.Y., Akinsola J.E.T., Awodele O., Hinmikaiye J. O., Olakanmi O., Akinjobi J.,” Supervised Machine Learning Algorithms: Classification and Comparison, International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017.

[10] Khushbu Kumari, Suniti Yadav,” Linear Regression Analysis Study”, Journal of the Practice of Cardiovascular Sciences · January 2018.

[11] Jitendra Singh Kushwah et. al. “A Comprehensive System for Detecting Profound Tiredness for Automobile Drivers Using a CNN”, Advanced Computing and Intelligent Technologies Proceedings of ICACIT 2022 Part of the book series: Lecture Notes in Electrical Engineering (LNEE, volume 914), Print ISBN: 978-981-19-2979-3, Online ISBN: 978-981-19-2980-9, 31 August 2022, Springer, Singapore.

[12] Jitendra Singh Kushwah et. al. “Comparative study of regressor and classifier with decision tree using modern tools” in First International Conference on Design and Materials (ICDM)-2021 of “Materials Today: Proceedings” ISSN: 2214-7853, Volume 56, Part 6, 2022, Pg 3571-3576.

[13] Jitendra Singh Kushwah et. al. “Analysis and visualization of proxy caching using LRU, AVL tree and BST with supervised machine learning” in 1st International Conference on Computations in Materials and Applied Engineering–2020 of “Materials Today: Proceedings” DOI: https://doi.org/10.1016/j.matpr.2021.06.224 ISSN: 2214-7853.

[14] Jitendra Singh Kushwah, Aditya Vidyarthi et. al. “COMPARATIVE ANALYSIS OF LINEAR REGRESSION, RANDOM FOREST AND SUPPORT VECTOR MACHINE USING DATASET” in International Journal for Innovative Engineering and Management Research. Vol. 11 Issue 1 pg. 30-35, ISSN: 2456-5083, Jan 2022.
[15] Pooja Rao, Vinod Kumar Yadav, Aditya Vidyarthi “An Approach to Find Nearest Location Services through Skyline Query & amp; KNN”, Conference SCI-2K19.
[16] Aditya Vidyarthi “Wireless Communications and Mobile Computing” Wiley-Hindawi; Volume 2022, Article ID 1191492, 11 pages, https://doi.org/10.1155/2022/1191492.
[17] Deepti Gupta, Rajeev Kumar Singh “A Survey on Disease Detection in Plant - using Image Processing Techniques” International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177 Volume 7 Issue VI, June 2019.