iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71303

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: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

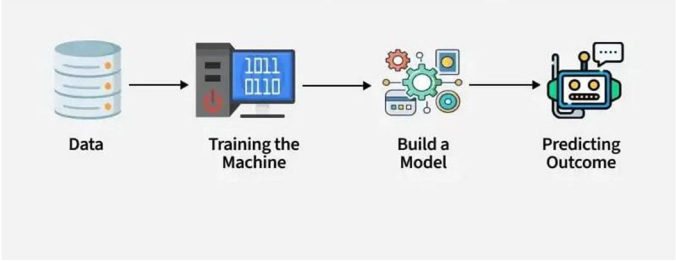
Smart Fraud Detection in E-Commerce Using Machine Learning

K. Dasarathi Shohi[1], R. Monika[2], K. Charumathi[3], M. Aarthi[4], K. Harini[5]
[1]*ME- Assistant professor Department of Computer Science and Engineering, M.I.E.T. Engineering College, Trichy, India*
[2, 3, 4, 5]*Student, Department of Computer Science and Engineering, M.I.E.T. Engineering College, Trichy, India*

Abstract: ***Machine Learning is an area of focus within Artificial Intelligence which has received considerable attention as part of the solutions to the digital transformation problem in the modern world. This paper attempts to review some of the most popular and frequently used machine learning algorithms. The author seeks to assist in decision making by highlighting the advantages and disadvantages of machine learning algorithms with respect to applications so that appropriate decisions can be made with respect to the chosen algorithm for the specific requirements of the application.***
Keywords: Artificial Neural Network, Back Propagation Algorithm, Bayesian Learning, Decision Tree, K Nearest Neighbor, Naïve Bayes, Support Vector Machine, and Logistic Regression.

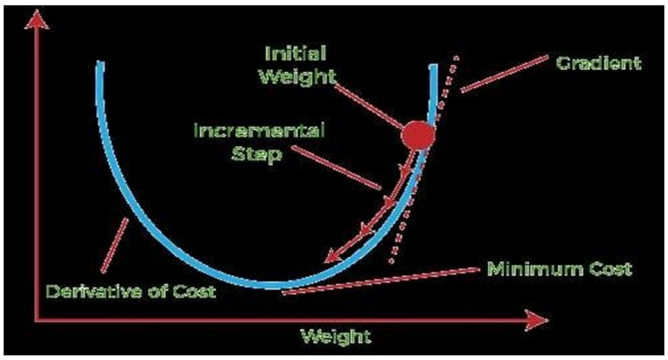
I. INTRODUCTION

As a start for this research paper, I believe it is best to explain the basics of Machine Learning. In Machine Learning, a computer program is given certain tasks to accomplish and it is claimed that the machine has learnt when it improves measurably in performing these tasks with increasing experience in executing them. Therefore, the machine is able to make decisions and do estimates / predictions based on the data. For example, consider a computer program that learns to detect / predict cancer from the medical investigation reports of a patient. It will get better at performing the tasks with increasing experience analyzing medical investigation reports of a larger population of patients. The measuring metric will be the number of successful predictions and detections of cancer cases made and validated by an experienced Oncologist. Machine Learning integrates with numerous domains such as: robotics, virtual personal assistants (e.g. Google), computer entertainment, pattern identification, natural languages, data mining, traffic forecasting, or in an online transport network like Uber estimating surge pricing during peak hours, product suggestion, stock market prediction, health diagnostics, online fraud analysis, agricultural consultation, refinement of return lists by a search engine (for instance Google), and Bots (chatterbots for custom online interaction).



II. GRADIENT DESCENT ALGORITHM

Gradient Descent is an iterative procedure aimed at decreasing a cost function. It is possible to calculate its partial derivative which is equivalent to the slope or gradient. In every iteration, the coefficients will be determined using the approach of taking the negative of the derivative and reducing the coefficients at each step by learning rate (also known as step size) they achieve local minima after several iterations. Therefore ultimately stop the iterations when achieve minimum cost function since there is no reduction observed after this. This method has three different variants: “Stochastic Gradient Descent” (SGD), “Batch Gradient Descent” (BGD), “Mini Batch Gradient Descent” (MBGD) In BGD the error will be computed for every example within the training dataset but the updates to the model will occur only after the evaluation of all training examples is completed. The primary benefit of the BGD algorithm is computational efficiency. It produces a stable error gradient and a stable convergence. But the algorithm has some issues. It may converge to a solution that is suboptimal for the problem at hand.Additionally, the algorithm must have access to the entire training dataset, which needs to be stored In SGD, error is cain memory.



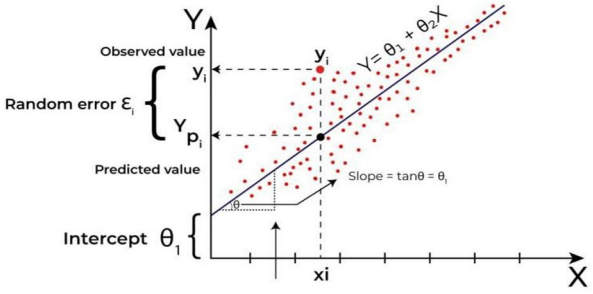
III. LINEAR REGRESSION ALGORITHM

Regression analysis assumes that there is a linear relationship between the independent variable X and the dependent variable Y. Such relationship can be useful in predicting the values of Y using different formulas depending on different contexts. Below are some examples for linear regression application: predicting the price of real-estate, forecasting sales, exam scores for students, and determining if a change in stock price will happen in the stock exchange.

It's referred to as supervised learning since in Regression, we have the labeled datasets and the output variable value is determined by input variable values, thereby making it supervised learning. Regression has several forms but linear regression is the simplest whereby an attempt is made to fit a straight line referred to as a straight hyperplane on the dataset, something that is achievable only when the relationship among variables within the dataset is linear

Although setting up Linear Regression may initially be complex and taxing, it offers the merit of being user friendly beyond that point. Also an advantage of Linear Regression is the speed of calculation, which offers insight into data learning, even with large volumes.

Another application of Linear Regression in scientific research includes acknowledging that the relationship between the variables being tested for is linear. Linear Regression is a simple approach best suited for research where simplicity, ease and fidelity are paramount.Still, it is not an advisable approach for a majority of practical scenarios as it oversimplifies real life Undesired traits.



IV. MULTIVARIATE REGRESSION ANALYSIS

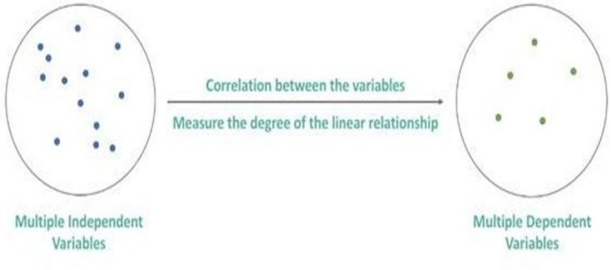
A simple linear regression model involves a dependent variable that is influenced by just one independent variable. But let's be real—life is rarely that straightforward. Typically, a dependent variable is affected by multiple factors. Take the price of a house, for instance; it can be influenced by a variety of elements such as the neighborhood, its size, the number of rooms, available amenities, and how far it is from the nearest station or shopping area.

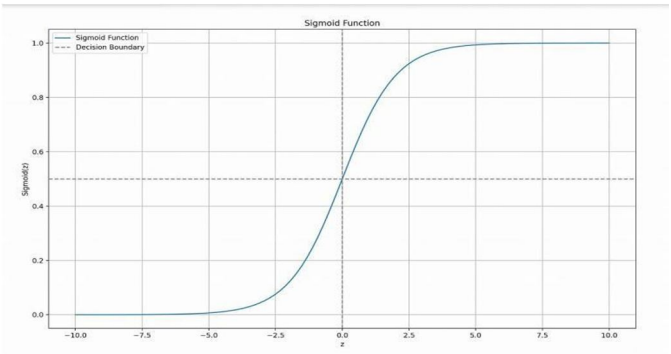
In essence, simple linear regression establishes a direct one-to- one relationship between the input and output variables. On the other hand, multiple linear regression showcases a many- to-one relationship, where several independent (input or predictor) variables relate to a single dependent (output or response) variable.

Just because you add more input variables doesn't guarantee that the regression will improve or yield better predictions. Both multiple and simple linear regression serve different purposes, and neither is inherently better than the other. In fact, sometimes adding more input variables can backfire, leading to overfitting. Plus, as you introduce more variables, they can start to interact with one another. This means that not only might the input variables be linked to the output variable, but they could also be interconnected, a phenomenon known as multicollinearity. Ideally, you want all input variables to correlate with the output variable, but not with each other.

Now, let's talk about the multivariate technique. It has some great advantages: it provides a deep understanding of the relationships between independent and dependent variables, as well as among the independent variables themselves. This is accomplished through methods like multiple regression, tabulation techniques, and partial correlation.

It effectively models the complexities of real-world problems in a practical and realistic manner. However, it's not all sunshine and rainbows. The complexity of this technique is quite high, requiring a solid grasp of statistical methods and modeling. Additionally, the sample size for statistical modeling needs to be substantial to yield reliable results.





V. LOGISTIC REGRESSION

Logistic regression is a handy tool for tackling classification problems. It provides a binomial outcome, essentially giving you the probability of whether an event will happen or not— think of it as a yes or no (0 or 1) based on the input variables. For instance, it can help predict whether a tumor is malignant or benign, or whether an email is spam.

These are classic examples of the binomial outcomes you can get from logistic regression. But it doesn't stop there; logistic regression can also handle multinomial outcomes. For example, it can predict your favorite type of cuisine—be it Chinese, Italian, or Mexican. And let's not forget about ordinal outcomes, like product ratings on a scale from 1 to 5.

Essentially, logistic regression is all about predicting categorical target variables, while linear regression focuses on predicting continuous values, such as estimating real estate prices over the next three years.

Now, what makes logistic regression appealing? For starters, it's simple to implement and computationally efficient. From a training perspective, it's quite effective, and you don't need to scale your input features. This algorithm is widely used in various industries. Since the output is a probability score, you'll need to set up customized performance metrics to determine a cutoff point for classifying your target.

Plus, logistic regression is robust against small noise in the data and multicollinearity.However, it does have its downsides. It struggles with non-linear problems because its decision surface is linear, and it can be prone to overfitting. It also requires that all independent variables be identified for it to work effectively.

Some real-world applications of logistic regression include predicting the risk of developing certain diseases, diagnosing cancer, assessing the mortality risk of injured patients, and even in engineering to estimate the probability of failure in processes, systems, or products.

VI. SUPPORT VECTOR MACHINE

Support Vector Machines (SVM) are versatile tools that can tackle both classification and regression tasks. At the heart of this method is the hyperplane, which acts as the decision boundary. When you have a collection of objects from different classes, you need a decision plane to effectively separate them. Sometimes, these objects can be tricky to separate linearly, and that's where complex mathematical functions known as kernels come into play. SVM's goal is to accurately classify these objects based on examples from the training dataset. Now, let's talk about the perks of using SVM: it can manage both semi-structured and structured data, and it can handle complex functions if you can derive the right kernel. Thanks to its focus on generalization, SVM has a lower chance of overfitting, and it scales well with high-dimensional data. Plus, it doesn't get trapped in local optima.On the flip side, there are some downsides to consider: SVM's performance can dip with larger datasets due to longer training times, and finding the right kernel function can be a challenge. It also struggles with noisy datasets and doesn't provide probability estimates, which can make understanding the final SVM model a bit tricky. In practical terms, Support Vector Machines are used in various fields like cancer diagnosis, credit card fraud detection, handwriting recognition, face detection, and text classification. So, when comparing Logistic Regression, Decision Trees, and SVM, it's often best to start with the logistic regression approach. If you want to see if there's a significant improvement, you can then try out decision trees (like Random Forests). When you have a high number of observations and features, SVM can be a great option to eposterior can actually serve as a new prior.

This approach is great for dealing with incomplete datasets through a Bayesian network. One of the benefits of this method is that it helps avoid overfitting the data, and you don't have to worry about removing contradictions from your dataset. However, there are some downsides to consider: choosing the right prior can be tricky, and the posterior distribution can be heavily influenced by the prior you select.

VII. NAIVE BAYES

This algorithm is simple and is based on conditional probability . In this approach there is a probability table which is the model and through training data it is updated. The "probability table" is based on its feature values where one needs to look up the class probabilities for predicting a new observation. The basic assumption is of conditional independence and that is why it is called "naive". In real world context the assumption that all input features are independent from one another can hardly hold true.

Naïve Bayes (NB) have the following advantage : implementation is easy, gives good performance , works with less training data, scales linearly with number of predictors and data points, handles continuous and discrete data, can handle binary and multi-class classification problems, make probabilistic predictions. It handles continuous and discrete data. It is not sensitive to irrelevant features.Naïve Bayes has the following disadvantages: Models which are trained and tuned properly often outperform NB models as they are too simple. If there is a need to have one of the feature as "continuous variable" (like time) then it is difficult to apply Naive Bayes directly, Even though one can make "buckets" for "continuous variables" it's not 100% correct. There is no true online variant for Naive Bayes, So all data need to be kept for retraining the model. It won't scale when the number of classes are too high, like > 100K. Even for prediction it takes more runtime memory compared to SVM or simple logistic regression. It is computationally intensive specially for models involving many variables. Naïve Bayes can be used in applications such as Recommendation System and forecasting of cancer relapse or progression after Radiotherapy.

VIII. K NEAREST NEIGHBOUR ALGORITHM

The Nearest Neighbor (KNN) Algorithm is a classification method that works by using a database filled with data points organized into various classes. Essentially, it takes a sample data point and tries to classify it as part of one of those classes. One of the cool things about KNN is that it doesn't make any assumptions about the underlying data distribution, which is why we call it non-parametric. Now, let's talk about the perks of using the KNN algorithm. First off, it's a straightforward technique that's easy to implement. Building the model doesn't break the bank, and it's incredibly flexible, making it a great fit for multi-modal classes where records can have multiple class labels. Plus, its error rate is at most double that of the Bayes error rate, and in some cases, it can even be the best method out there. For instance, KNN has shown to outperform SVM in predicting protein functions using expression profiles.On the flip side, there are some downsides to KNN. Classifying unknown records can get pretty pricey since it requires calculating the distance to the k-nearest neighbors. As the training set grows, the algorithm can become computationally heavy. Additionally, if there are noisy or irrelevant features, the accuracy can take a hit.

KNN is considered a lazy learner because it computes distances over k neighbors without generalizing the training data—it keeps all of it. While it can handle large datasets, that also means the calculations can be quite costly. When dealing with higher-dimensional data, you might see a drop in accuracy in certain regions. KNN has a variety of applications, including recommendation systems, diagnosing multiple diseases with similar symptoms, assessing credit ratings based on feature similarity, handwriting detection, financial institutions analyzing data before approving loans, video recognition, predicting votes for different political parties, and image recognition.

IX. K MEANS CLUSTERING ALGORITHM

The K Means Clustering Algorithm is a popular choice for tackling clustering problems. It falls under the category of unsupervised learning. One of its key advantages is that it's computationally more efficient than hierarchical clustering, especially when dealing with large datasets. When working with globular clusters and a small number of clusters (k), it tends to produce tighter groupings compared to hierarchical methods. Plus, it's relatively easy to implement and interpret the results, which makes it quite appealing. The algorithm has a complexity order of O(K*n*d), highlighting its efficiency.

However, there are some downsides to the K Means Clustering Algorithm. For starters, figuring out the right value for K can be tricky. Its performance can take a hit when the clusters are globular. Additionally, different initial partitions can lead to varying final clusters, which can affect the overall performance. If there's a significant difference in size and density among the clusters in the input data, that can also degrade performance. The uniform effect often results in clusters that are relatively uniform in size, even if the input data has clusters of different sizes. The spherical assumption—that the joint distribution of features within each cluster is spherical—can be hard to meet, especially when correlations between features come into play, which can unfairly weight correlated

features. The K value is often unknown, and the algorithm is sensitive to outliers and initial points. There's no unique solution for a given K value, so it's common to run K Means multiple times (anywhere from 20 to 100 times) and then select the result with the lowest J value.

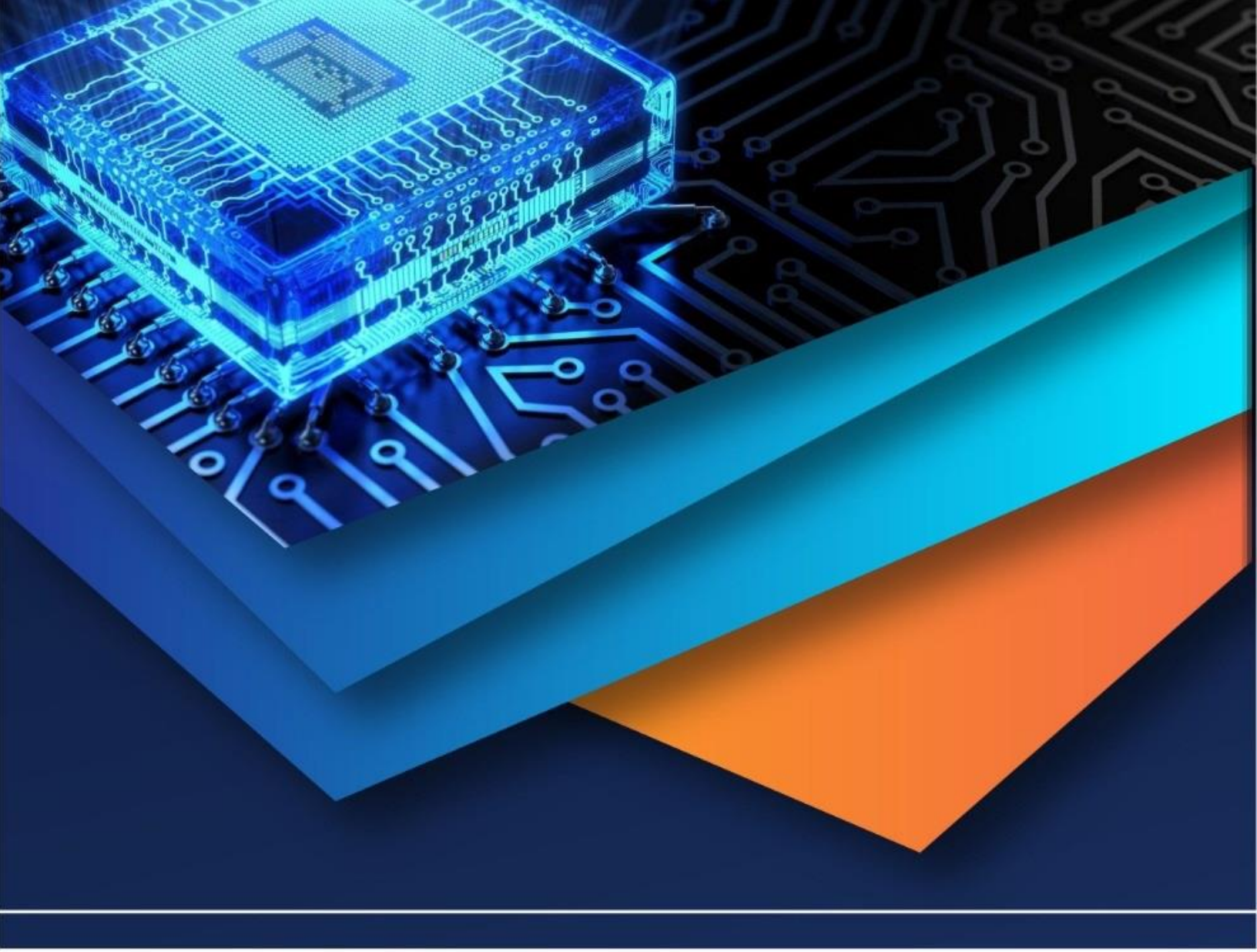
The K Means Clustering Algorithm can be applied in various scenarios, such as document classification, customer segmentation, rideshare data analysis, automatic clustering of IT alerts, analyzing call record details, and even detecting insurance fraud.

X. CONCLUSION

In this paper, we take a closer look at the most commonly used machine learning algorithms designed to tackle classification, regression, and clustering challenges. We dive into the pros and cons of these algorithms, comparing their performance and learning rates wherever possible. Additionally, we highlight real-world applications of these techniques. We also explore various types of machine learning methods, including supervised, unsupervised, and semi-supervised learning. Our goal is to provide readers with valuable insights that will help them make informed decisions when it comes to choosing the right machine learning algorithm for their specific problem- solving needs.

REFERENCES

[1] D. Pelleg, A. Moore (2000): "X-means: Extending K-means with Efficient Estimation of the Number of Clusters"; ICML'00 Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727-734.
[2] Rushika Ghadge, Juilee Kulkarni, Pooja More, Sachee Nene, Priya R , [3] "Prediction of Crop Yield using Machine Learning", International Research Journal of Engineering & Technology, Vol 5, Issue 2, Feb2018.
[3] C. Phua, V. Lee, K. Smith, R. Gayler (2010); "Comprehensive Survey of Data Mining-based Fraud Detection Research", ICICTA '10 Proceedings of the 2010 International Conference on Intelligent Computation Technology and Automation Volume 1, pp. 50-53.
[4] S. Cheng, J. Liu, X. Tang (2014); "Using unlabeled Data to Improve Inductive Models by Incorporating Transductive Models"; International Journal of Advanced Research in Artificial Intelligence, Volume 3 Number 2, pp. 33-38.
[5] Sonal S. Ambalkar, S. S. Thorat2, "Bone Tumor Detection from MRI Images using Machine Learning: A Review", International Research Journal of Engineering & Technology", Vol. 5, Issue 1, Jan -2018.
[6] Rajat Raina, Alexis Battele, Honglak Lee,Benjamin Packer, Andrew Y. Ng , "Self-taught Learning : Transfer of Learning from Unlabeled Data", Computer Science Department, Stanford University, CA, USA, Proceedings of 24th International Conference on Machine Learning Corvallis, OR, 2007.
[7] Jimmy Lin, Alek Kolcz, "Large-Scale Machine Learning at Twitter", Proceedings of SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.
[8] Dr. Rama Kishore, Taranjit Kaur, "Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition", International Journal of Scientific & Engineering Research, Volume 3, Issue 6, June-2012.
[9] Kedar Potdar, Rishab Kinnerkar, "A Comparative Study of Machine Algorithms applied to Predictive Breast Cancer Data", International Journal of Science & Research, Vol. 5, Issue 9, pp. 1550-1553, September 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY