International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

# Analysis of Classification Algorithms using Machine Learning

**M. Tech. Scholar Soniya Rathore** soniya.rathore39@gmail.com
**Asst. Prof. Ankur Taneja** ankurtaneja5@gmail.com
**Asst. Prof. Mahesh Patidar**
Department of Computer Science & Engineering
SAMCET
Bhopal, M.P, India

**Abstract - In this work our main focus is on regression which is one of the most important methods in machine learning algorithm. Regression is a statistical approach that is used to find the relationship between variables. It is basically used to predict the outcome from the given dataset. In this work we will discuss the regression algorithms which are available in machine learning algorithm and propose one algorithm that will have less train error and test error as compared to other existing algorithm. The accuracy measure will be in the form of train and test error.**

***Keywords*- Classification, Data Mining, Linear Regression, Machine Learning techniques, python.**

## I. INTRODUCTION

Machine learning systems itself grasp programs or plan from data. This is generally a very impressive alternative to making or substitute constructing them and in the last some past years the utilizing of machine learning has increase rapidly in computer science. Machine learning is used in Web search i.e Query search, Network filters, recommending in many systems, for placing ad, To find-out credit scoring, fraud detection, In stock trading, drug design in medical fields, and many other applications. A recent report from the many big and Global Institute like McKinsey asserts that machine learning (a.k.a. data mining or find-out future analysis) will be the next generation technology for society and market where we are keeping abundant amount of data [16]. So many machine learning projects extends their time to process the given data to give better results in many domains. By developing this technology knowledge is fairly easy to communicate for business requirement. In Machine Leaning major component is given below.

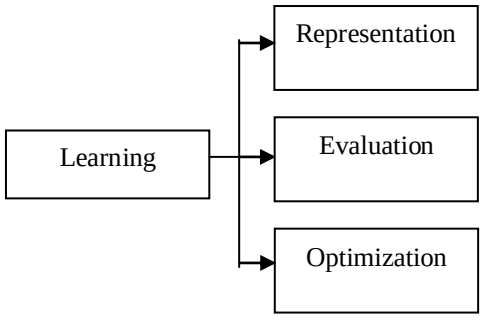


Figure 1 Evaluation of Machine Learning.

**1. Representation-** A classifier can represent in such manner (means a definite language) so that a computer can understand easily. If a classifier is not in the hypothesis space, it cannot be learned. Selection of classifier for any individual's problem is important.

**2. Evaluation-**It is like function which decides which classifier is bad and which one is good. This is also called objective function.

**3. Optimization-**To finding optimization technique is very important and is key to the learner, and also helps to find-out the classifier.

**4. Classification of Machine Learning**

There are 3 branches of machine learning we can understand this classification in details with sketch diagram.

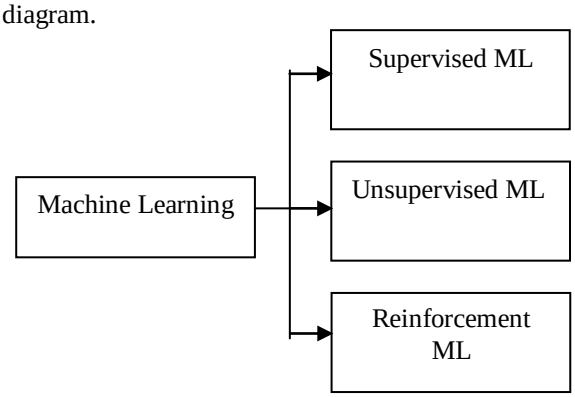


Figure 2 Classification of Machine Learning.

**4.1 Supervised Learning**: In supervised machine learning, a system is trained with data that has been labeled. The labels categories each data point into one or more groups, such as 'apples' or 'oranges'. The system learns how this data known as training data is structured, and uses this to predict the categories of new or 'test' data.

**4.2 Unsupervised Learning:** In this learning, learning is without labels. It aims to detect the characteristics that make data points more or less similar to each other, for example by creating clusters and assigning data to these clusters.

**4.3 Reinforcement Learning:** In this learning focuses on learning from experience, and lies between unsupervised and supervised learning. In a typical reinforcement

learning setting, an agent interacts with its environment, and is given a reward function that it tries to optimize, for example the system might be rewarded for winning a game. The goal of the agent is to learn the consequences of its decisions, such as which moves were important in winning a game, and to use this learning to find strategies that maximize its rewards.

**2. Machine Learning in Daily Life**

Machine learning is using by us in day to day life in various form out of few names is given below

- Online shopping.
- Customer support
- Virtual Assistants
- Email Spam and Malware Filtering.

## II. RELATED WORK

In this papers Author's Explained How Researchers have already explored advanced regression techniques for forest variables estimation and recent literature provides examples which show their suitability in comparison with MLR, Neural networks, SVM and decision trees are the admired schemes for classification.

In this paper [3] three techniques are compared by applying ML techniques on KDD CUP'99 data set. The techniques are supposed to be good for identifying the anomalies detection, but the performance may differ in terms of different algorithms.

In this paper [4] the author explain by the use of movie trailer uploaded on most popular video platform YouTube can predict the success and failure of a movie before it is released by help of these the filmmakers, In this paper they used various regression algorithms for prediction but in this paper they used only linear regression, Polynomial Regression, Gradient Boosted Tree and Simple Regression Tree they can also use SVM which can give them better performance.

After reading we realize that gradient tree boosting algorithms in this part. The Explanation follows from the same idea in existing literatures in gradient boosting. Specifically, the second order method is originated from Friedman et al. [12]. We make minor improvements in the regularized objective, which may get helpful in implementation or using.

## III. PROPOSED WORK

**1. Simple Linear Regression-** It is statistical way so that it allows us to remember and study relationships between two continuous variables.

**Note:** Linear Regression might be old but it's still useful, but there's a drawback of using linear regression because it's made on assumptions that our data have linear relationships while in many real world scenarios that not true.

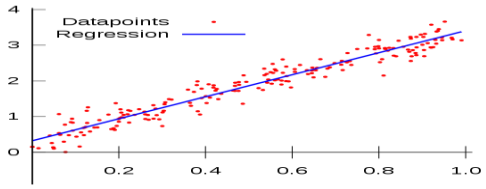


Figure 3 Model of Regression.

**2. Support Vector Machine-** Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges.

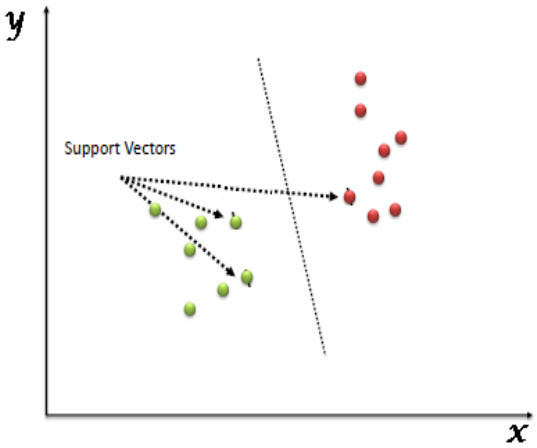


Figure 4 Model of Support Vector Machine.

## IV. REQURED FRAMEWORK

Python and its libraries are using in data science and data analysis very efficiently. They are also largely used for creating expandable machine learning algorithms. Python can apply various machine learning techniques such as Classification, Regression, and Clustering. Python offers to researcher ready-to-Implement Environment for doing or performing data mining tasks on huge volumes and a variety of data effectively in less time.

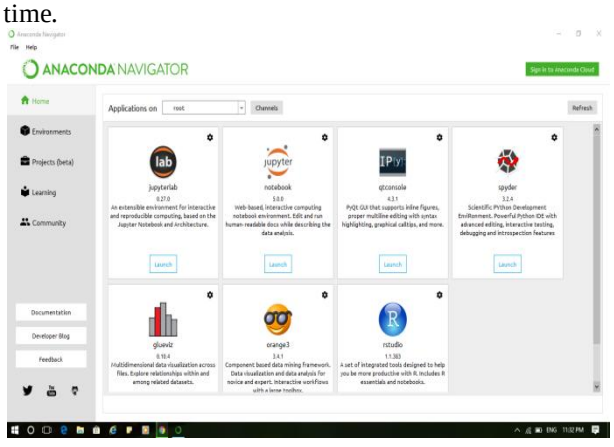


Figure 5 GUI Anaconda.

International Journal of Scientific Research & Engineering Trends
Volume 5, Issue 2, Mar-Apr-2019, ISSN (Online): 2395-566X

## V. FLOW CHART OF PROPOSED ALGORITHM

The flowchart employs filters for faster evaluation and lesser overall time.

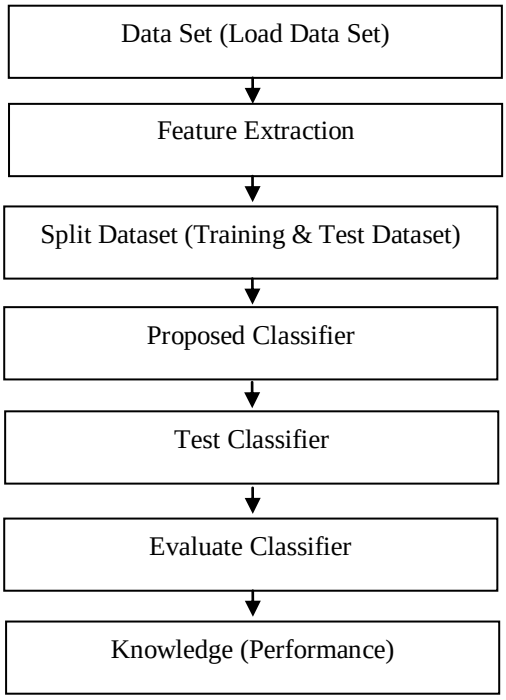


Figure 6 Proposed Data Mining Framework for Classification.

In the above flow we represent how our algorithm works. For better understanding we are giving pictorial representation of above flow.

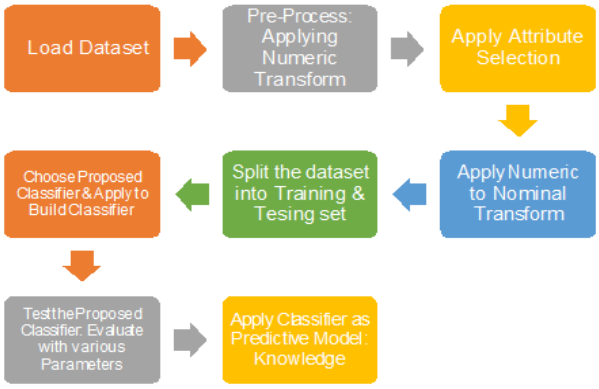


Figure 7 Dataset & Features.

Machine learning data is usually described in a matrix called dataset. This matrix is structured in a way that corresponds to each row an observation (example) data set and each column represents a feature (also variable or attribute) that describes the data. Data values can take many representations. Data can be numerical (integer or real numbers) or nominal data, where values are differentiated by name. Nominal data is type of Categorical data type of that, as its name indicates, the data only can have a fixed set of nominal values (or categories).

## VI. RESULTS

In this paper we wanted to identify which is the best algorithm for the movies data set. For the experiment purpose, we used Anaconda, an open source distribution for Python. As the data set contain numeric value, so it is possible to apply supervised learning. In this approach we also applied different regression techniques on the models. We divided the movie dataset into 70:30 portion to find the best technique. 70% data were used for training purpose and 30% data were used for validation. The above Result is showing in our Editor for better understanding we are converting it into tabular form.

Table 1 Result.

| Algorithm | Test Error | Train Error |
|---|---|---|
| Ridge Regression | 14.296076 | 12.729437 |
| Knn | 5.768323 | 12.492261 |
| Bayesian Regression | 0.131753 | 12.784852 |
| Decision Tree | 5.237878 | 14.264513 |
| SVM | 4.073167 | 5.772826 |
| Elastic Net | 14.274904 | 12.816194 |
| Proposed Regression | 0.131753 | 5.772826 |

The proposed work is also shown in the graph that is being shown below:

## VII. CONCLUSION AND FUTURE WORK

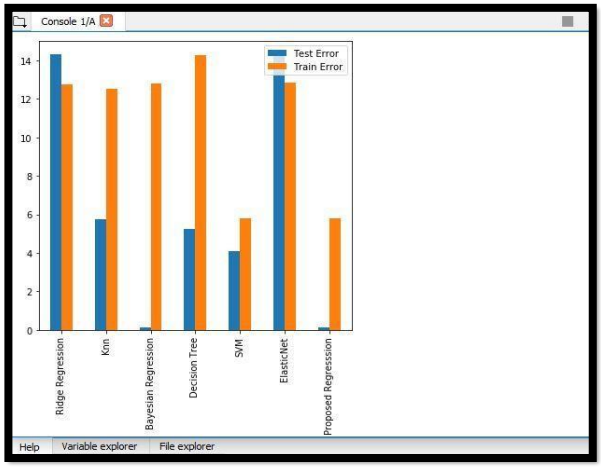


Figure 8 Result.

Our observation different terms and condition. This represents our work in a new approach. In our work we have tried to minimize the train and test error. So we have already discussed about the regression algorithms and all have their own computation strategy. Out of these regression algorithms we have observed that Bayesian regression and SVM is performing better in terms of test error and train error respectively. So our approach is basically to combine the features of Bayesian and regression, so that we get combine output of both. After implementing the combine algorithm of Bayesian and SVM we have show they are giving good result in terms of test error and train error.

The future works focus on applying some other techniques to improving the performances of these methods for up to (99.99%). Another concept that can be implemented Deep learning in place of machine learning technology. The reason behind this is best and efficient techniques using nowadays. Deep learning is also introduced nowadays which is becoming more popular for classification purpose. So we can also implement deep learning in future work also.

## REFERENCES

[1] R. Bekkerman. The present and the future of the kdd cupcompetition: an outsider's perspective.
[2] R. Bekkerman, M. Bilenko, and J. Langford. Scaling UpMachine Learning: Parallel and Distributed Approaches.Cambridge University Press, New York, NY, USA, 2011.
[3] J. Bennett and S. Lanning. The netix prize. In Proceedings of the KDD Cup Workshop 2007, pages 3{6,New York, Aug. 2007.
[4] Mining Trailers Data From YouTube for Predicting Gross Income of Movies by Md Shamsur Rahim, AZM Ehsetma chowdhury, Md.Asiful Islam, Mir Riyanul Islam in 2017 IEEE Region 10 Humanitarian TechnologyConference(R10-HTC)21 - 23 Dec 2017, Dhaka, Bangladesh
[5] C. Burges. From ranknet to lambdarank to lambdamart:An overview. Learning, 11:23{581, 2010.
[6] O. Chapelle and Y. Chang. Yahoo! Learning to RankChallenge Overview. Journal of Machine Learning
[7] T. Chen, H. Li, Q. Yang, and Y. Yu. General functionalmatrix factorization using gradient boosting. In Proceeding of 30th International Conference on Machine Learning (ICML'13), volume 1, pages 436{444, 2013.
[8] T. Chen, S. Singh, B. Taskar, and C. Guestrin. E_cientsecond-order gradient boosting for conditional random_elds. In Proceeding of 18th Arti_cial Intelligence andStatistics Conference (AISTATS'15), volume 1, 2015.
[9] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, andC.-J. Lin. LIBLINEAR: A library for large linear classi_cation. Journal of Machine Learning Research,
[10] J. Friedman. Greedy function approximation: a gradientboosting machine. Annals of Statistics, 29(5):1189{1232,2001.
[11] X. Yan, X. G. Su, Regression Analysis: Theory and Computing, World Scientific Publishing Co. Pte. Ltd., 2009.
[12] M. H. Kutner, C. J. Nachtsheim, J. Neter, Applied Linear Regression Models, 4th ed., McGraw-Hill Irwin, 2004.
[13] T. Hastie, R. Tibshirani, M. Wainwright, Statistical Learning with Sparsity. The Lasso and Generalizations, Chapman & Hall, 2015.