International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 03 | Mar 2019 www.irjet.net p-ISSN: 2395-0072

Comparison of Classification Algorithms Using Machine Learning

Ankta Pal [1], Neelesh Shrivastava[2], Pradeep Tripathi[3]

M.Tech Scholar, Department of Computer Science & Engineering, VITS Satna, (M.P), India, Email:ankitapal964@gmail.com [1]

Asst Prof, Department of Computer Science & Engineering, VITS Satna, (M.P)[2]

Asst Prof & Head, Department of Computer Science & Engineering, VITS Satna, (M.P)[3]

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 we have number of major component out of them some is very important to understanding about how machine learning explore and work efficiently.

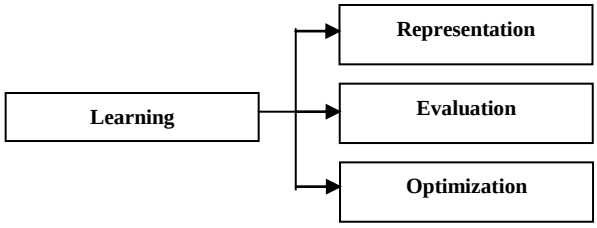


Figure 1: Evaluation of Machine Learning

Representation: A classifier can represent in such manner (means a definite language) so that a computer can understand easily.

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

1.1 Classification of Machine Learning

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

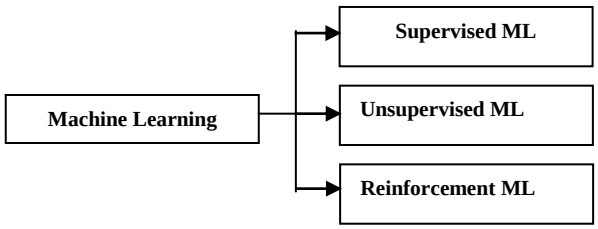


Figure 2: Classification of Machine Learning

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.

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.

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.

1.2 Machine Learning in Daily Life

Machine learning is using by us in day to day life in various form out of some names is given below with their working behavior.

Recommender systems: suggesting products or services that recommend products or services on the basis of previous choices are amongst the most widely recognized application of machine learning.

Organizing information: In search engines and spam filtering Machine learning also helps provide the results of queries entered in internet search engines, such as Google.

II REVIEW OF LITERATURE

According to the authors [2], 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.

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.

The work [4] presents that often the case that the matrix X*t*X is "close" to singular. This process is known as *multi colored* in a multiple regression model. In this phase we can find-out the OLS estimates, but they will likely have "bad" statistical properties. Slight variations in the statistical data (like adding or removing a few out puts) will lead to finding important changes in the coefficient estimates.

III DESCRIPTION OF USED ALGORITHM

Simple Linear Regression: Simple linear regression can be explained with the help of two variables.

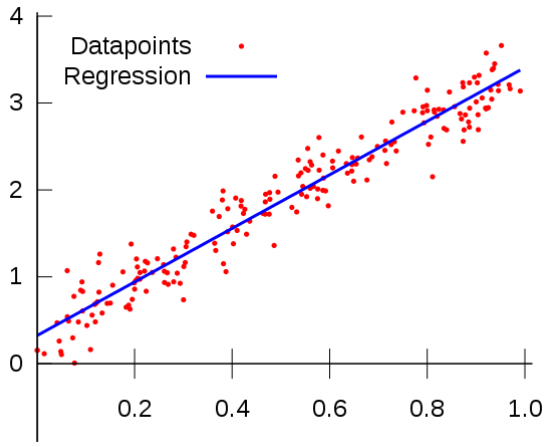


Figure 3: Model of Regression

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. It's quite useful to understand Linear Regression because of its simplicity and later on it will be useful to understand more modern approaches and the state of The art Algorithms such as Neural Networks and many more.

Support Vector Machine: Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is generally used in classification problems means to categorized problems into solution.

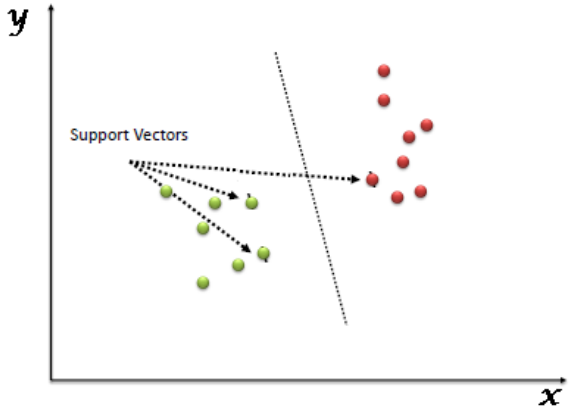


Figure 4: Model of Support Vector Machine

IV EXPERIMENTAL FRAMEWORK

Python is a prominent environment using by researcher to development or deployment of generated systems. It has vast set of libraries with number of modules, packages that supports programmer to attain in many ways to complete their work efficiently.

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.

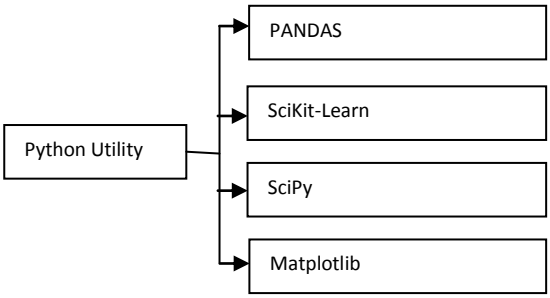


Figure 5: Libraries of Python

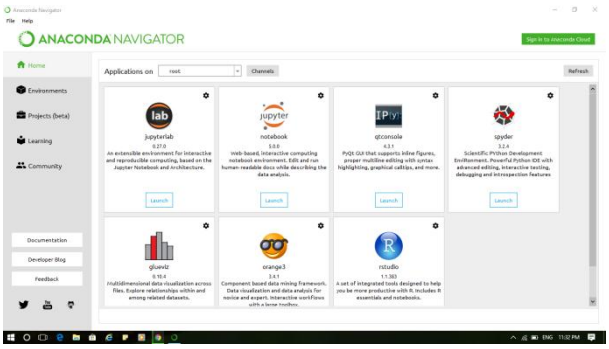


Figure 6: GUI Anaconda

Anaconda is a totally free Environment their source is really open to all for doing much.

V ALGORITHM

1. Input / Load data set
2. Apply feature extraction
3. Received Extracted data as output
4. Generate Training and Testing data set (By applying techniques:)

5. Apply Multiple Classification algorithm to training dataset (MLR)

6. Build the Reduction Explanatory Predictor

7. Building Model using different classifier

8. Perform / Obtain validity check

9. Utilize the "test" set predictions to calculate all the performance metrics (Measure Accuracy and other parameters)

10.Compare the Accuracy among different classification algorithm.

VI IMPLEMENTATION

The model employs filters for faster evaluation and lesser overall time. The pre-processing methods and application of filters affect a lot in final evaluation results of classifiers (ML based models). The feature extraction methods, conversion of nominal to binary and cleaning are few of those filters.

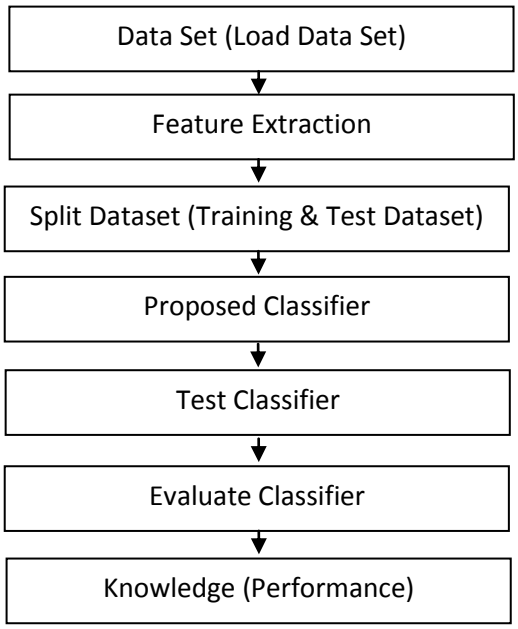
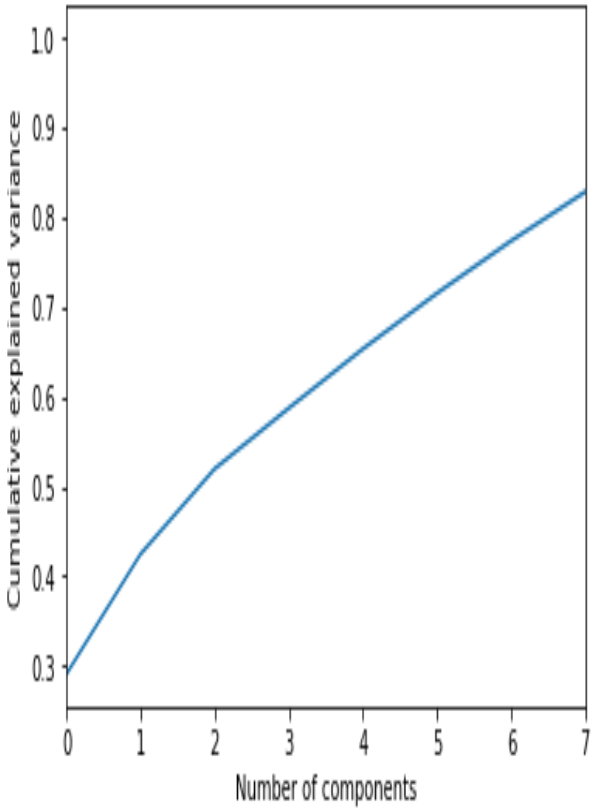


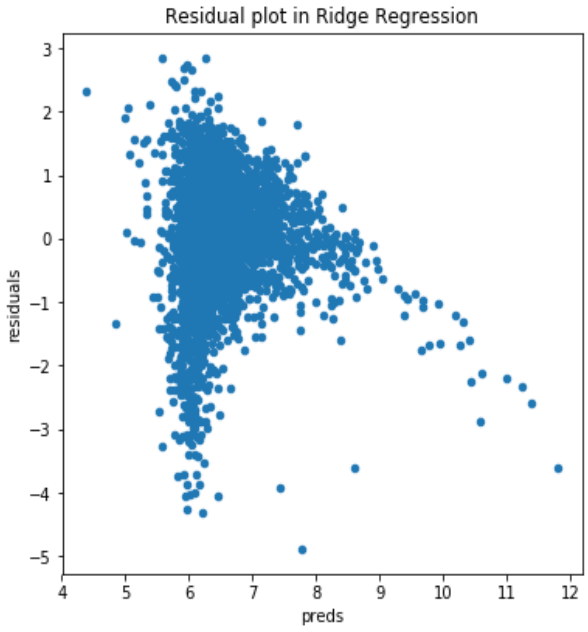
Figure 7: Proposed Data Mining Framework for Classification

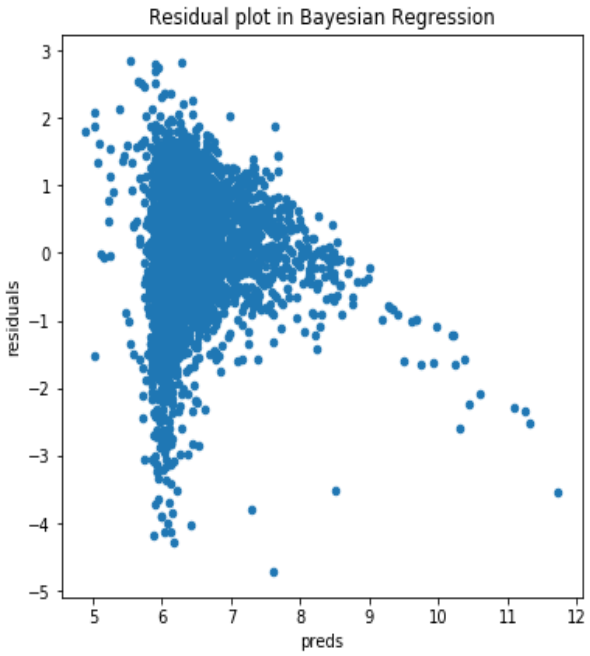
In this section we have shown the output of the regression algorithm with their residual plot, train error and test error.

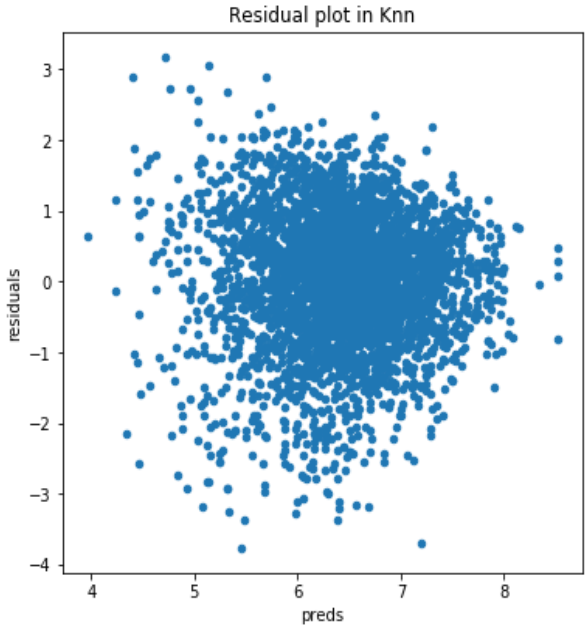
Algorithm	Test Error	Train Error
Ridge Regression	14.296076	12.729437
Knn	5.768323	12.492261
Bayesian Regression	0.131753	12.784852
Decision Tree	5.237878	14.264513
SVM	4.073167	5.772826
Elastic Net	14.274904	12.816194
Proposed Regression	0.131753	5.772826

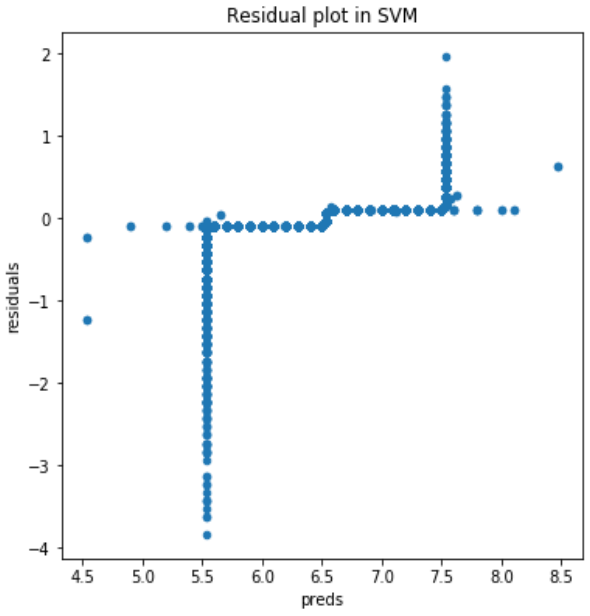
Table 1: Output Results

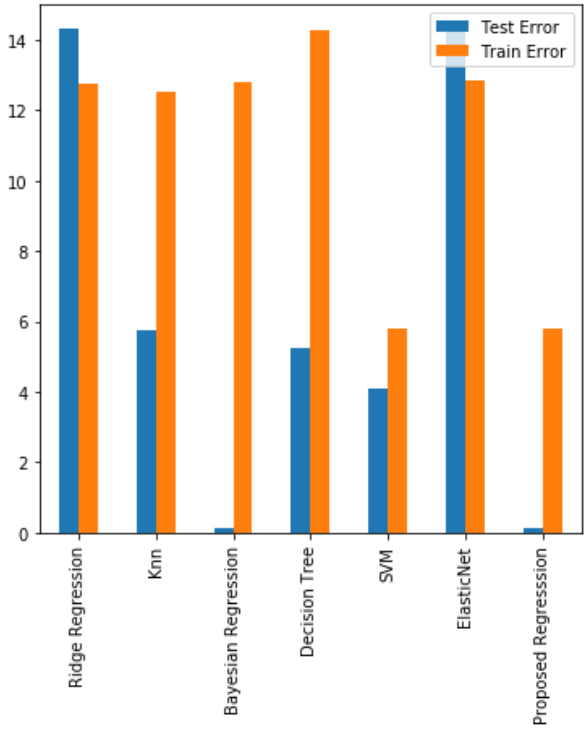



Residual plot in Ridge Regression
residuals
preds


Residual plot in Bayesian Regression
residuals
preds


Residual plot in Knn
residuals
preds


Residual plot in SVM
residuals
preds



VII CONCLUSION AND FUTURE WORK

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 shown they are giving good result in terms of test error and train error.

REFERENCES

[1] R. Bekkerman. The present and the future of the kdd cup competition: an outsider's perspective.

[2] R. Bekkerman, M. Bilenko, and J. Langford. Scaling Up Machine 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] L. Breiman. Random forests. Maching Learning, 45(1):5{32, Oct. 2001.

[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 Rank Challenge Overview. Journal of Machine Learning

[7] T. Chen, H. Li, Q. Yang, and Y. Yu. General functional matrix 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_cient second-order gradient boosting for conditional random _elds. In Proceeding of 18th Arti_cial Intelligence and Statistics Conference (AISTATS'15), volume 1, 2015.