*IOSR Journal of Computer Engineering (IOSR-JCE)*
*e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 22, Issue 1, Ser. IV (Jan - Feb 2020), PP 01-07*
*www.iosrjournals.org*

# A Survey on Machine Learning Algorithms

Baraskar Aashutosh Alankar[1], Gharade Abdul Hannan[2], Dubale Yash Nitin[3], Mohammed Ali[4]

[1](3rd Year, Computer Engineering, M.H. SabooSiddik Polytechnic, India)
[2](3rd Year, Computer Engineering, M.H. SabooSiddik Polytechnic, India)
[3](3rd Year, Computer Engineering, M.H. SabooSiddik Polytechnic, India)
[4](Lecturer, Computer Engineering, M.H. SabooSiddik Polytechnic, India)

***Abstract:***
*Machine Learning is a logical procedure where the PCs figure out how to take care of an issue, without unequivocally program them. Profound learning is at present driving the ML race fueled by better calculations, calculation force and huge information. Still ML old style calculations have their solid situation in the field.It is difficult to choose appropriate algorithm for a particular dataset which will help the developer to develop a productive approach. Our project will solve this difficulty by considering a few of the parameters like accuracy, time taken etc. of the algorithm for a particular dataset and compare each algorithm on the basis of these parameters and then predict that which out of these algorithms is the best used in the given scenario. We have done a comparative study over different machine learning supervised and unsupervised techniques like Linear Regression, Logistic Regression, K nearest neighbors and Decision Trees in this project. This project is implemented using python. Thus, we had used JupyterLab(Which is an IDE for Python) for implementing this project. In this project we have used different datasets such as pima-India-diabetes dataset for supervised learning, The Oxford-IIIT Pet Dataset for image classification algorithms etc. This paper center's around clarifying the idea and advancement of Machine Learning, a portion of the famous Machine Learning calculations and attempt to analyses most well-known calculations dependent on some essential ideas. In this project different machine learning algorithms are compared on the basis of some parameters such as Confusion matrix, accuracy, Area Under ROC Curve etc.*

***Key Words****: Machine learning, algorithms, datasets.*

---

Date of Submission: 18-02-2020 Date of Acceptance: 02-03-2020

---

## I. Introduction

Machine learning is a form of AI that enables a system to learn from data rather than through explicit programming. However, machine learning is not a simple process. As the algorithms ingest training data, it is then possible to produce more precise models based on that data. A machine-learning model is the output generated when you train your machine-learning algorithm with data. After training, when you provide a model with an input, you will be given an output. For example, a predictive algorithm will create a predictive model. Then, when you provide the predictive model with data, you will receive a prediction based on the data that trained the model[1].

Pattern recognition process and data classification are valuable for a long time. Humans have very strong skill for sensing the environment. They take action against what they perceive from environment [2]. Big data turns into Chunks due to multidisciplinary combined effort of machine learning, databases and statistics. Today, in medical sciences disease diagnostic test is a serious task. It is very important to understand the exact diagnosis of patients by clinical examination and assessment. For effective diagnosis and cost-effective management, decision support systems that are based upon computer may play a vital role. Health care field generates big data about clinical assessment, report regarding patient, cure, follow-ups, medication etc. It is complex to arrange in a suitable way. Quality of the data organization has been affected due to inappropriate management of the data. Enhancement in the amount of data needs some proper means to extract and process data effectively and efficiently [3]. Thus, ML algorithms are used to solve this problem.

So, in this project we are going to compare different types of machine learning algorithms on the basis of different datasets. Thus, it will help the beginners to choose correct ML algorithms.

## II. Material And Methods

This project is implemented using python. Thus, we had used JupyterLab(Which is an IDE for Python) for implementing this project. In this project we have used different datasets such as pima-India-diabetes dataset for supervised learning, The Oxford-IIIT Pet Dataset for image classification algorithms etc.

**JupyterLab:** JupyterLab is an interactive development environment for working with notebooks, code and data. Most importantly, JupyterLab has full support for Jupyter notebooks. Additionally, JupyterLab enables you to use text editors, terminals, data file viewers, and other custom components side by side with notebooks in a tabbed work area.JupyterLab understands many file formats (images, CSV, JSON, Markdown, PDF, Vega, Vega-Lite, etc.) and can also display rich kernel output in these formats[4].

**Pima-Indians-Diabetes-Dataset:** This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage [5].

The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and so on[5].

## III. Block Diagram

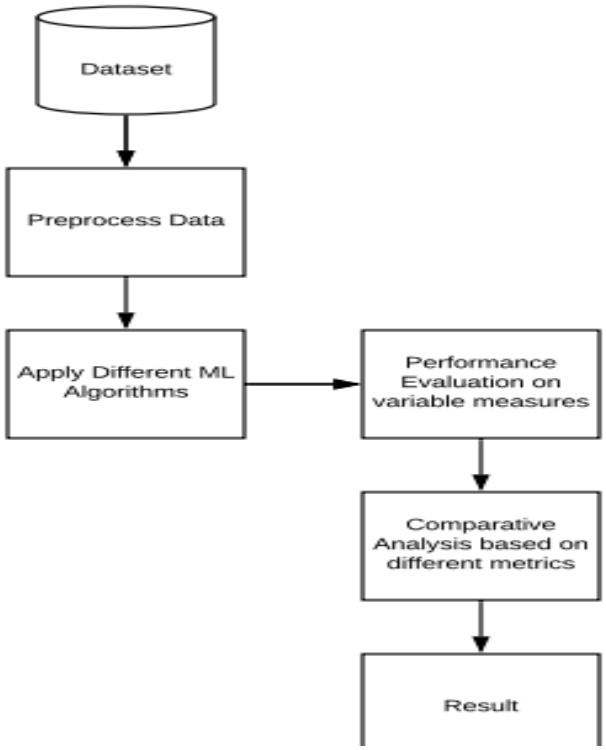



**Figure 1.** Block diagram of the proposed system

In this project, the above-mentioned datasets are used to compared various ML algorithms. We have directly taken preprocessed datasets from websites such as www.kaggle.com. On this dataset the algorithms are trained to give predictions. Then on the basis of this predictions the algorithms are compared. The comparison is then graphically represented using the Matplotlib(plotting library for the Python programming language).

## IV. Algorithms Which Are Compared In This Project

In this project various different algorithms are compared.They are as follows:

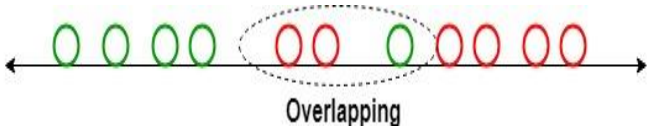
**1.Logistic Regression:**

Linear Regression [6], [7] is the most common predictivemodel to identify the relationship among the variables. Apartfrom univariate or multivariate data types the concept is linear.Linear regression can be either simple linear or multiple linearregression. The linear regression is described in Eq. 1.

$$y\ \beta\ += \varepsilon \quad (1)$$

In Eq. 1. y is the independent variable which can be eithercontinuous or categorical value, x is a dependent variable
which is always a continuous value. It is analyzed withprobability distribution and mainly focused on conditional
probability distribution with multivariate analysis.

**2.Linear Discriminant Analysis:**
Linear Discriminant Analysis or Normal Discriminant Analysis or Discriminant Function Analysis is a dimensionality reduction technique which is commonly used for the supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes. It is used to project the features in higher dimension space into a lower dimension space [8].
For example, we have two classes and we need to separate them efficiently. Classes can have multiple features. Using only a single feature to classify them may result in some overlapping as shown in the below figure. So, we will keep on increasing the number of features for proper classification [8].

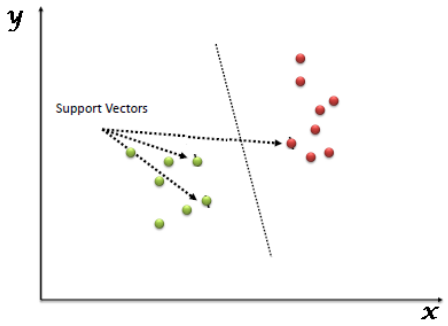



**3.K-Nearest Neighbours Algorithm:**
It is also simply known as KNN algorithm is one of the mostsimply understandable algorithms in the machine learningalgorithms. It is simple to understand and works incredibly well inapplication. It is a non-parametric learning algorithm that is it doesnot make any assumption on data distribution. Its training phaseis minimal and very fast. It stores all the training data in itsmemory as it does not have any type of generalization. In otherwords, all of the data that is used for training is used for the testingof the data. It is unlike the SVM in which all the non-supportvectors are discarded. The KNN works fine with the data beingeither scalar or multidimensional. The k value is given by the user.K means the number of nearest neighbors that are to be takeninto consideration. The distance between the new data and each ofthe data that is in the training data is calculated the distance can behamming distance, Manhattan distance or city block distance,Euclidean distance[9].

**4.Classification and Regression Trees:**
The representation for the CART model is a binary tree. This is your binary tree from algorithms and data structures, nothing too fancy. Each root node represents a single input variable (x) and a split point on that variable (assuming the variable is numeric).The leaf nodes of the tree contain an output variable (y) which is used to make a prediction [10].

**5.Naïve Bayes:**
It is a supervised classification methoddeveloped using Bayes' Theoremof conditional probability with a 'Naive'
assumption that every pair of features is mutually independent. That is, in simpler words, presence of a feature is not
effected by presence of another by any means. Irrespective of this over-simplified assumption, NB classifiers performed quite well in many practical situations, like in text classification and spam detection. Only a small amount oftraining data is needto estimate certain parameters. Besides, NB classifiershave considerably outperformed even
highly advanced classification techniques[11].

**6.SUPPORT VECTOR MACHINE:**
SVM is so popular a ML technique that it can be a group of its own. Ituses a separating hyperplane or a decision plane todemarcate decision boundaries among a set of data pointsclassified with different labels. It is a strictly supervised classification algorithm. In other words, the algorithm develops an optimal hyperplane utilising input data ortraining data and this decision plane in turnscategories new examples. Based on the kernel in use, SVM can performboth linear and nonlinear classification [11].




## V. Metrics Used In Evaluating Ml Algorithms

**1.Accuracy:**

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class. For example, consider that there are 98% samples of class A and 2% samples of class B in our training set. Then our model can easily get 98% training accuracy by simply predicting every training sample belonging to class A. When the same model is tested on a test set with 60% samples of class A and 40% samples of class B, then the test accuracy would drop down to 60%. Classification Accuracy is great, but gives us the false sense of achieving high accuracy. The real problem arises, when the cost of misclassification of the minor class samples are very high. If we deal with a rare but fatal disease, the cost of failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests [12].

**2. Negative Logarithmic Loss:**

Logarithmic Loss or Log Loss, works by penalizing the false classifications. It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples.

Log Loss uses negative log to provide an easy metric for comparison. It takes this approach because the positive log of numbers < 1 returns negative values, which is confusing to work with when comparing the performance of two models.Suppose, there are N samples belonging to M classes, then the Log Loss is calculated as below :

$$LogarithmicLoss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * \log(p_{ij})$$

where,

y_ij, indicates whether sample i belongs to class j or not

p_ij, indicates the probability of sample i belonging to class j

Log Loss has no upper bound and it exists on the range [0, ∞). Log Loss nearer to 0 indicates higher accuracy, whereas if the Log Loss is away from 0 then it indicates lower accuracy.In general, minimizing Log Loss gives greater accuracy for the classifier [12].

**3.Mean Absolute Error:**

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don't give us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data [12]. Mathematically, it is represented as :

$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|$$

**4.Mean Squared Error:**

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the square of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear

programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors [12].

$$MeanSquaredError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

**5.$R^2$ Metric:**

The $R^2$ (or R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. In statistical literature, this measure is called the coefficient of determination.This is a value between 0 and 1 for no-fit and perfect fit respectively [13].

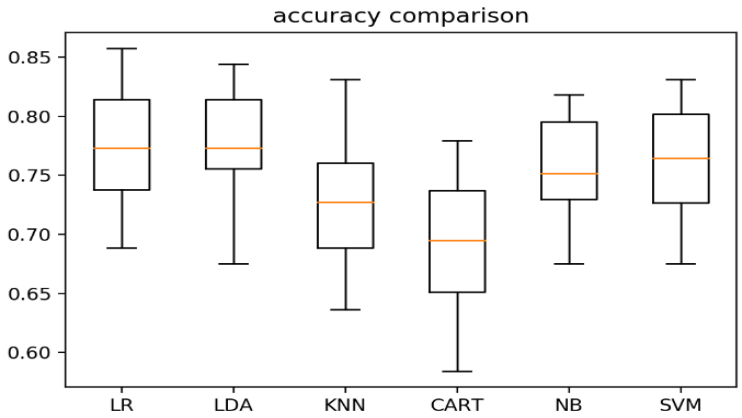
**6.Area Under ROC Curve:**

Area Under ROC Curve (or ROC AUC for short) is a performance metric for binary classification problems.The AUC represents a model's ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predictions perfectly. An area of 0.5 represents a model as good as random.

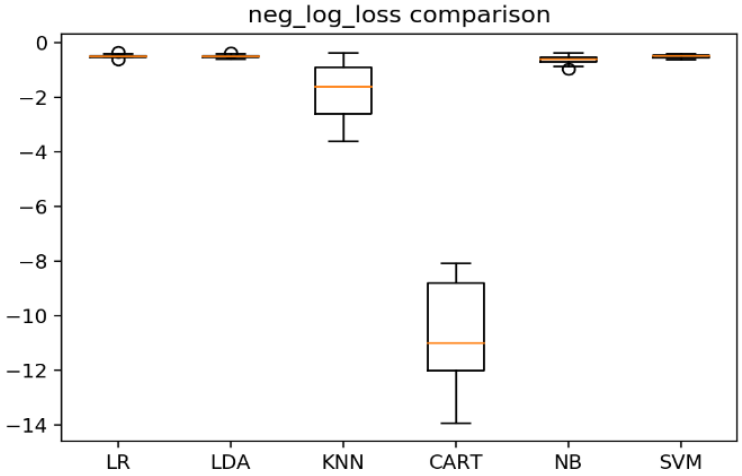
A ROC Curve is a plot of the true positive rate and the false positive rate for a given set of probability predictions at different thresholds used to map the probabilities to class labels. The area under the curve is then the approximate integral under the ROC Curve [13].

## VI. Result

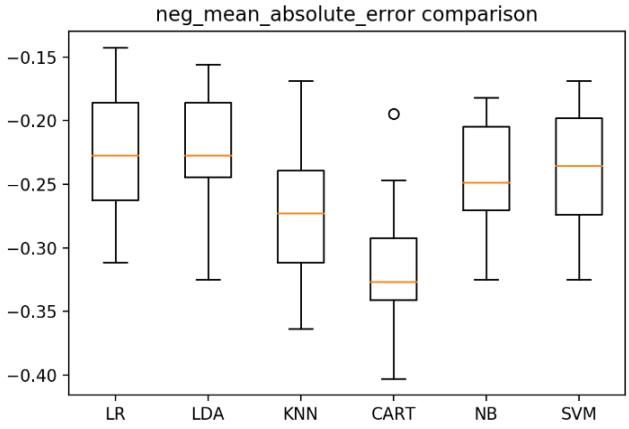
**1. Comparison on the basis of Accuracy:**



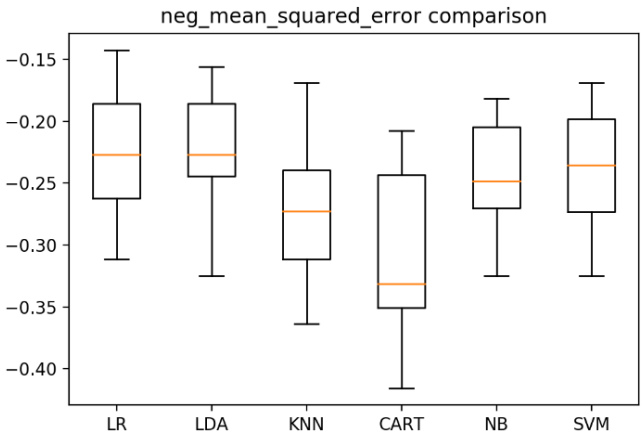

**2. Comparison on the basis of Logarithmic loss:**



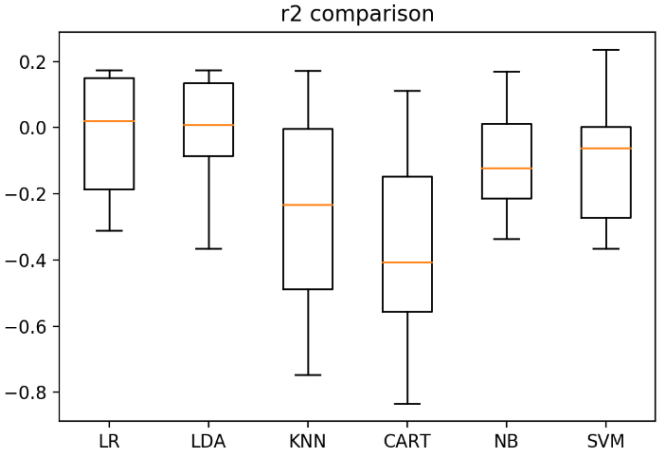
**3. Comparison on the basis of Mean Absolute Error:**



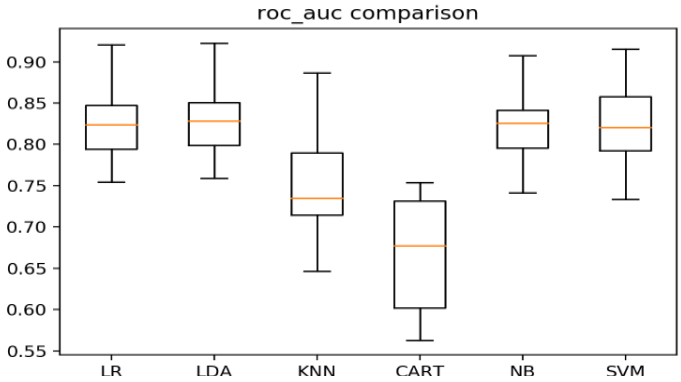

**4. Comparison on the basis of Mean Squared Error:**




**5. Comparison on the basis of $R^2$ Metrics:**



**6. Comparison on the basis of Area Under ROC Curve:**




## VII. Conclusion

In this paper, we have done a deeper understanding about the concept of machine learning. We took into consideration a few of the supervised machine learning algorithms like Logistic Regression, Linear Discriminant Analysis, KNN, Classification and Regression Tree and Naive Bayes. According to our research we found out that

Logistic Regression is better than other algorithms which were taken into consideration in this paper. It exceeded in the parameters like accuracy, logarithmic loss, mean squared error, mean absolute error and R 2 metrics. only in Areaof ROC Curve, Linear Discriminant Analysis Algorithm was better than LR with a minute difference. In future, wemay do the analysis on more different types of machine learning algorithms so it may be more helpful for datascientists.

## References

[1]. Judith Hurwitz, Daniel Kirsch, "Machine Learning for Dummies", IBM, 2018
[2]. Sharma, P. and Kaur, M. (2013) Classification in Pattern Recognition: A Review. International Journal of Advanced Research in Computer Science and Software E ngineering , 3, 298.
[3]. Rambhajani, M., Deepanker, W. and Pathak, N. (2015) A Survey on Implementation of Machine Learning Techniques for Dermatology Diseases Classification. I nternational Journal of Advances in Engineering &Technology , 8, 194-195. Daniele Miorandi, Sabrina Sicari, Francesco De Pellegrini, ImrichChlamtac, "Internet of things: Vision, applications and research challenges", Ad Hoc Networks, Int J, vol. 10, no. 7, pp. 1497–1516, April 2012.
[4]. https://blog.jupyter.org/
[5]. https://www.kaggle.com/uciml/pima-indians-diabetes-database
[6]. GAF Seber, AJ Lee, "Linear regression analysis", Wiley Seriesin Probability and Statistics,2012. foundation.Diabetes Care. 2008;31(4):811–822
[7]. DC Montgomery, EA Peck, GG Vining, "Introduction to linearregression analysis", Wiley Series in Probability and Statistics,2015.
[8]. https://www.geeksforgeeks.org/ml-linear-discriminant-analysis/
[9]. R.Ragupathy, Lakshmana PhaneendraMaguluri, "Comparative analysis of machine learning algorithmson social media test", International Journal of Engineering & Technology, 7 (2.8) (2018) 284-290.
[10]. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/
[11]. Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept,Algorithms and Applications", International Journal of Innovative Research in Computerand Communication Engineering, Vol. 5, Issue 2, February 2017
[12]. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234
[13]. https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Baraskar Aashutosh Alankar, etal. "A Survey on Machine Learning Algorithms." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 22.1 (2020), pp. 01-07.