iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** I **Month of publication:** January 2022

DOI: https://doi.org/10.22214/ijraset.2022.39980

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 10 Issue I Jan 2022- Available at www.ijraset.com

A Comparative Study on Supervised Machine Learning Algorithm

Monica Gupta[1], Dr. Sohil D. Pandya[2]
[1]*Department of Computer Engineering, Gujarat Technological University, India*
[2]*Assistant Professor, Department of Computer Applications, CMPICA, CHARUSAT, Changa, India*

Abstract: Machine learning enables computers to act and make data driven decisions rather than being explicitly programmed to carry out a certain task. It is a tool and technology which can answer the question from your data. These programs are designed to learn and improve over time when exposed to new data. ML is a subset or a current application of AI. It is based on an idea that we should be able to give machines access to data and let them learn from themselves. ML deals with extraction of patterns from dataset, this means that machines can not only find the rules for optimal behavior but also can adapt to the changes in the world. Many of the algorithms involved have been known for decades. In this paper various algorithms of machine learning have been discussed. Machine learning algorithms are used for various purposes but we can say that once the machine learning algorithm studies how to manage data, it can do its work accordingly by itself.
Keywords: Linear Regression, Logistic Regression, KNN, Naive Bayes, Decision Trees, SVM, Random Forest

I. INTRODUCTION

Machine learning is the science of getting computers to act by feeding them data and letting them learn few tricks on their own without being explicitly programmed According to Tom Mitchell "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." [1] So, it can be said that machine learning makes the decision based on the data and make prediction on various aspects. Machine learning counts on various algorithms to make the prediction with the help of huge data sets. According to the availability of types of algorithm and training data set one has to select the available techniques of "unsupervised learning", "supervised learning" and "reinforcement learning" for making the prediction or forecasting. Data scientists like to point out that there's no single one-size-fits-all type of algorithm that is best to solve a problem. [2] Depending on the dataset, types of parameters and what type of problem you want to work out, the type of algorithm or model is selected.

II. TYPES OF MACHINE LEARNING ALGORITHMS

In supervised learning the machine learns under the guidance. In supervised learning the machine learns by feeding them labeled data and explicitly telling them about the input and how the output must look.
Unsupervised learning means to perform without anyone guidance or supervision. In unsupervised learning the data is not labeled and the machine has to itself figure out from the data set and to figure out the hidden patterns to make the prediction about the output. Reinforcement learning establishes the pattern of behavior where input itself depends on the action we take. It follows the hit and trial concept where the machine learns from the experience from the given environment. An agent interacts with the environment and produces the rewards or punishment and once the agent gets trained it gets ready to predict from the new data.

III. SUPERVISED LEARNING TECHNIQUES

Under supervised learning there are two main categories of problem:

1) *Classification:* In classification prediction is done with label or class such as emails are classified into spam and non-spam mails classes
2) *Regression:* In regression prediction is done using continuous quantities or infinite possibilities like change in weather.

In this research paper we will discuss the most commonly used machine learning algorithms in machine learning.

a) Linear Regression
b) Logistic Regression
c) K - Nearest Neighbor(KNN)
d) Gaussian Naive Bayes

e) Decision Trees
f) Support Vector Machine (SVM)
g) Random Forest

A. Linear Regression

Regression is a relationship between dependent and independent variables. Linear regression is a statistical approach used for predictive study. It accomplishes the job of predicting a dependent variable (Y as an output) which is assumed from the independent variable (X as an input). It makes assumptions on continuous or real or numeric parameters. The connection between X and Y can be represented as follows

$$Y = mX + c$$

Here, m is the slope and c is the intercept. Based on the Equation, one can compute the result that will be through the connection shown between the independent (X) and the dependent parameters (Y).

Linear Regression has many advantages like easy implementation, overfitting can be decreased by using regularization, good fit if the relationship between predictor and response variable is linear, and focuses on data analysis. Linear regression has some of the disadvantages like deviation having negative consequences on the regression , it always presumes a straight line relationship between independent and dependent variable , not good fit to nonlinear relationship, the mean of dependent and independent variable relationship is also considered.

1) *Implementation Tools:* Python, R, MATLAB, and Excel.
2) *Methods/Techniques:* Simple linear regression, Ordinary least squares, Gradient descent, Regularization, Adam's method, Singular value decomposition. [1]

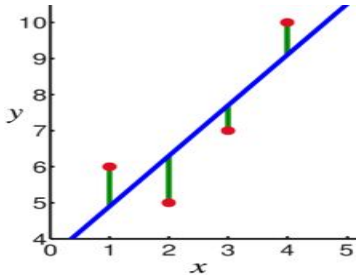


Fig. 1 Linear Regression model [3]

As shown in Fig 1 , The observations are marked in red and are the result of random deviations (marked in green) from the underlying relationship (marked in blue) between the independent variable(x) and the dependent variable (y).[3]

B. Logistic Regression

It is the most widely used classification algorithm when the dependent variable is in binary format, so we need to predict the outcome of a categorical dependent variable .So the outcome should be discrete or categorical in nature. Here discrete means values should be binary or it can have just two outcomes: either 0 or 1, either true or false, either yes or no, or either high or low. In logistic regression we don't need the value below 0 and above 1. In logistic regression, we have to predict categorical variables and solve classification problems.

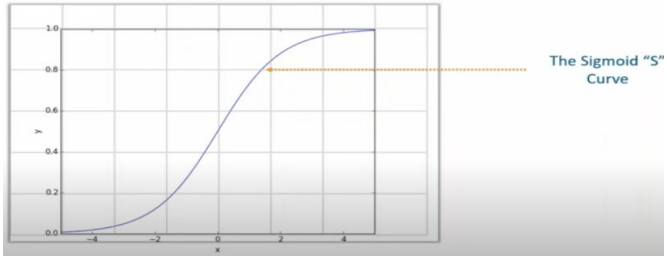


Fig. 2 Linear Regression model

In the above diagram there is a sigmoid curve which converts any value to negative infinity to infinity to binary values which a logistic regression needs.

The logistics Regression Equation is derived from the straight Line Equation

1) Equation of Straight Line

Y = C + B1X1 + B2X2 + … → Range will go from (infinity) to (infinity)

Logistic Regression Equation from Straight Line Equation

Y = C + B1X1 + B2X2 + … → Y can only be from 0 to 1

To get the range of Y between 0 and infinity,

Y Y = 0 then 0

1-Y Y = 1 then infinity

Now to get the range between (infinity) to (infinity)

$$\text{Log}\left(\frac{Y}{1-Y}\right)$$ → Y = C + C + B1X1 + B2X2 + … → Final Logistic Equation

Logistic Regression has many advantages like implementation in a simple way ,mathematical proficiency , proficiency in regulation ,efficient with respect of training , no scaling needed for input variables, works efficiently with large dataset , easily extend multiple classes , can overfit with multi scale which is controlled by the technique known as regularization , outputs are more calibrated than other models. Logistic Regression has many disadvantages like cannot solve nonlinear problem, complex relation is difficult to handle, for good prediction requires huge dataset, duplicity of data can lead to wrong training parameters. [2]

2) Implementation Tools: R, Python, Java, and MATLAB

C. K - Nearest Neighbor

KNN algorithms can be taken for both regression and classification problems but are most commonly used in classification problems. It is easy to understand and apply. Based on homogeneous data, the KNN learns the pattern available within. It is flexible and cheap to build the model but it is also a lazy learner which needs a huge dataset which results in complex calculations. In KNN the K is a parameter that counts the nearest neighbor which is to be involved in the voting process.

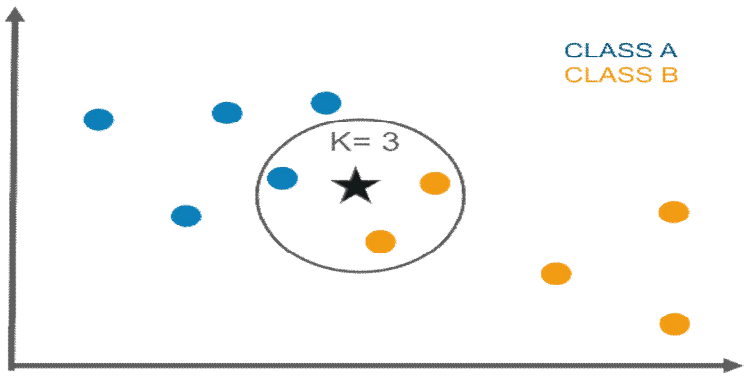


Fig. 3 K - Nearest Neighbor [8]

KNN has many advantages like its implementation is easy, can carry multi-class data sets, solves both classification and regression problems, and is good for nonlinear data. KNN has many disadvantages like calculation cost is higher, needs more storage for the dataset, computationally intensive, unrelated features can affect the accuracy.

1) Implementation Tools: R, WEKA, Scikit-learn of Python, KNIME, Orange [4]

D. Naive Bayes

A Naive Bayes classifier presumes that the existence of a specific property in a class is unrelated to the existence of any other property [9]. It is based on probabilistic logic which uses algorithms based on Bayes theorem. Bayes theorem is a mathematical probabilistic technique which helps to calculate the conditional probabilities of an event. Naive Bayes is a classification technique based on bayes with an assumption of independence among predictors. It is one of the simplest and efficient machine learning algorithms which can make fast predictions on the basis of the probability of the data. It is basically used in text classification and issues having numerous classes.

Here's the equation for Naive Bayes:

$$P(c|x) = P(x|c)\,P(c) / P(x)$$

$$P(c|x) = P(x1 \mid c) \times P(x2 \mid c) \times \ldots P(xn \mid c) \times P(c)$$

Here, P (c|x) is the posterior probability according to the predictor (x) for the class(c). P(c) is the prior probability of the class, P(x) is the prior probability of the predictor, and P(x|c) is the probability of the predictor for the particular class(c). [5]

Naïve Bayes (NB) have the following advantages like easy in implementing, less training data can provide good result, can manage both discrete and continuous data, good in solving prediction of multiclass problems, irrelevant feature does not affect the prediction, Naïve Bayes has the following disadvantages like assuming all features are independent which is not always applicable in real word cases, faces zero frequency problems, prediction done by NB is not always correct. [5][6]

1) Implementation Tools: WEKA, Python, R Studio and Mahout [4]

2) Methods: Gaussian Naive Bayes, Multinomial Naive Bayes , Bernoulli Naive Bayes

E. Decision Trees

Decision Tree works out on both classification and regression problems where data is being continually splitted based on some criteria or parameters. The tree has mainly two entities i.e. nodes and leaves. A decision tree employs a structure of nodes and branches. The data is splitted in the nodes and the decisions are in the leaves. The depth of a node is the minimum number of steps required to reach the node from the root. It is easy to handle the category and quantitative values in a decision tree but it is difficult to control the size of the tree

Each node shows the attributes in a group that is to be classified and each branch represents a value that the node can take [6]. An example of a decision tree is given in Fig. 3.

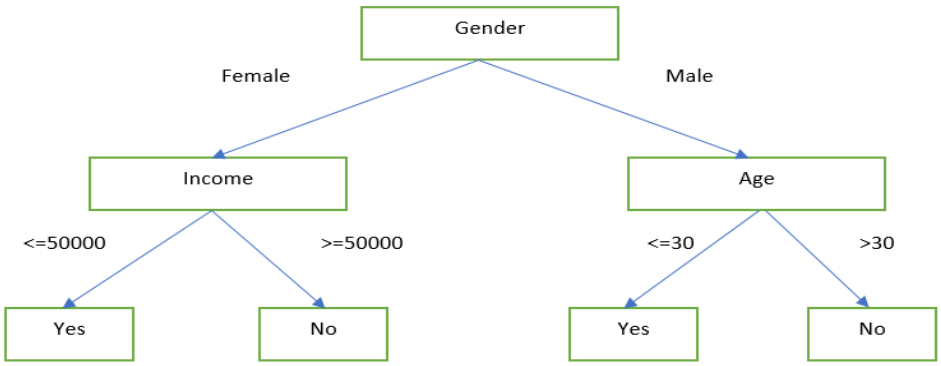


Fig. 3 Decision Tree

Decision Tree has many advantages like it needs less attempt in preparing the data at the time of preprocessing, does not needed normalization and scaling of data, interpretation and implementation is easy, easy understanding through visualization Decision Tree has many disadvantages like complexity increases with the increase in labels, small change can lead to different structure, may lead to more time to train the data, and tends to overfitting.

1) Implementation Tools: Weka, KNIME, Orange, Python (Scikit-learn) and R Studio [4]

F. Support Vector Machine Algorithm

SVM is the most widely and popular supervised machine learning algorithm which is mainly used for classification problems but also considered in regression models. It is a linear model which provides solutions for both linear and nonlinear problems. It works on the concept of margin calculation. Basically it separates the dataset into different classes and draws the hyperlane between them.

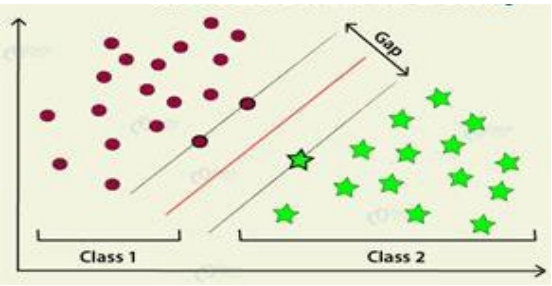


Fig. 4 Support Vector Machine [7]

SVM has many advantages like it function skillfully even with semi structured and unstructured data, the actual strength of SVM is kernel trick and furthermore it can deal with any complex problem with the suitable function , it can work well with high dimensional data , and due to generalization SVM has less risk of overfitting. SVM has disadvantages like it takes more time to train the model for a large dataset , selecting a proper kernel function is a difficult task , and does not work well with noisy data.

1) Implementation Tools: SVMlight with C, LibSVM with Python, MATLAB or Ruby, SAS, Kernlab, Scikit-learn, Weka [4]

G. Random Forest

This algorithm is also used for both regression & classification problems. The algorithm randomly creates a forest with several trees (so generally the more the trees in the forest the more robots the forest looks.). So in the random Forest classifier the higher the number of freezes in the forest, greater is the accuracy of the results. So in other ways we can say that a random forest builds collective decision trees called a forest and integrate them together to get a more precise and steady prediction and the forest it builds in the collection of decision trees and each decision tree is built only on a part of the dataset defined and it is trained with the bagging method. Random forest ensembles multiple decision trees to come up for the final decision. In the Fig. 5 according to random forest the entire dataset is split into three subset which leads to a particular decision tree and each decision tree will come with a certain outcome. Random forest will compile the results from all decision trees and it will result in the final outcome.

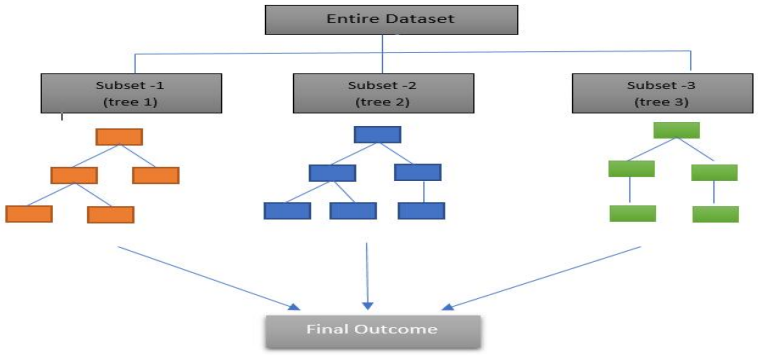


Fig. 5 Random Forest

Random Forest has many advantages like it automatizes lost values that are present in the data, it overcomes the overfitting problem that arises in the decision tree, and the large dataset is handled efficiently. Random Forest has disadvantages like it needs more computing and resources to bring out the output, requires time for the training as it integrates lots of decision trees.

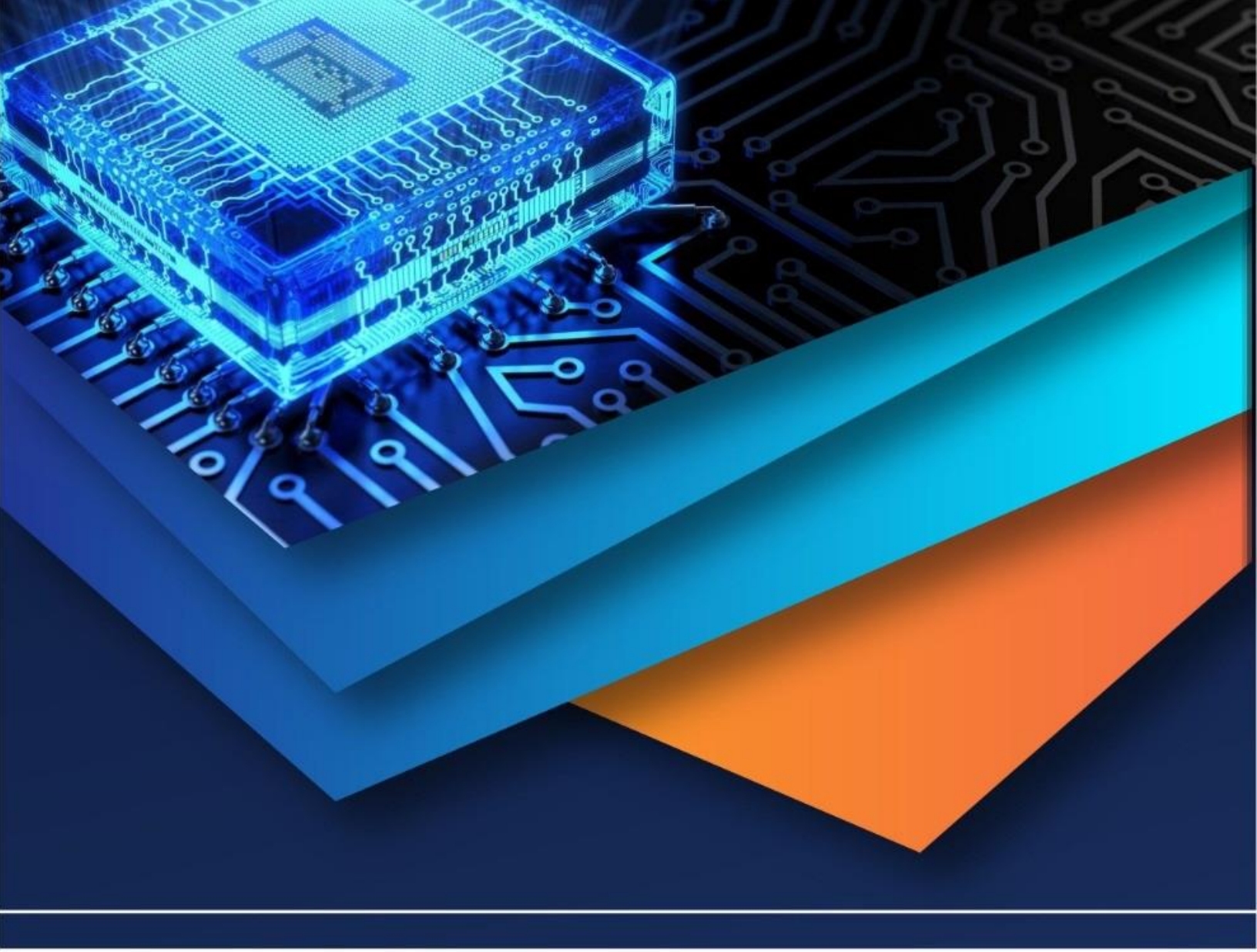
1) Implementation Tools: Python Scikit-Learn, R

IV. CONCLUSION

This paper reviews various machine learning algorithms with their advantages and disadvantages, methods and implementation tools. After analyzing all these different machine learning algorithms in this review paper, it is being found that each algorithm has distinct ways of getting all the data and information in machine learning.

REFERENCES

[1] https://learn.g2.com/linear-regression#linear-regression-types
[2] https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/
[3] A. Srivastava, S. Saini and D. Gupta, "Comparison of Various Machine Learning Techniques and Its Uses in Different Fields," 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), 2019, pp. 81-86, doi: 10.1109/ICECA.2019.8822068.
[4] O. Obulesu, M. Mahendra and M. ThrilokReddy, "Machine Learning Techniques and Tools: A Survey," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018, pp. 605-611, doi: 10.1109/ICIRCA.2018.8597302.
[5] https://www.upgrad.com/blog/naive-bayes-explained/
[6] https://www.kaggle.com/getting-started/225022
[7] Mahesh, Batta. (2019). Machine Learning Algorithms -A Review. 10.21275/ART20203995.
[8] https://www.edureka.co/blog/k-nearest-neighbors-algorithm/
[9] https://blog.floydhub.com/naive-bayes-for-machine-learning/
[10] https://dhirajkumarblog.medium.com/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY