iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IX **Month of publication:** September 2021

DOI: https://doi.org/10.22214/ijraset.2021.37969

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: 7.429
Volume 9 Issue IX Sep 2021- Available at www.ijraset.com

A Survey on Machine Learning Algorithms

Ayan Gain[1], Ajeet Soni[2], Akash Guha[3]
[1, 2, 3]*B.Tech, Department of CS&IT, Sam Higginbottom University of Agriculture, Technology & Sciences, Prayagraj, Uttar Pradesh, India*

Abstract: In this paper, mainly three machine learning algorithms are discussed in detail and how they work is shown, their pseudo codes are shown and a comparison is done among them, and their performance is analyzed.
Keywords: Machine Learning, pseudo code, algorithms, data

I. INTRODUCTION

Machine Learning is used to find patterns in data that we humans cannot find quickly and manually, it is basically a method of data analysis that helps building automated models and is a part of artificial intelligence. With many datasets available from different sectors the demand for employing machine learning in various sectors have rapidly gone up in the past few years, various sectors from sports to medicine to defense have started using machine learning to achieve efficiency and better results.

There are various machine learning algorithms that are there for the use of programmers but it is important to note that an algorithm will only work as good as the data is filtered and abnormalities, outliers, etc. are removed accordingly for desired results. There are various ways mathematicians and programmers try to find the solution of a particular problem [1]. Fig.1 demonstrates the same.

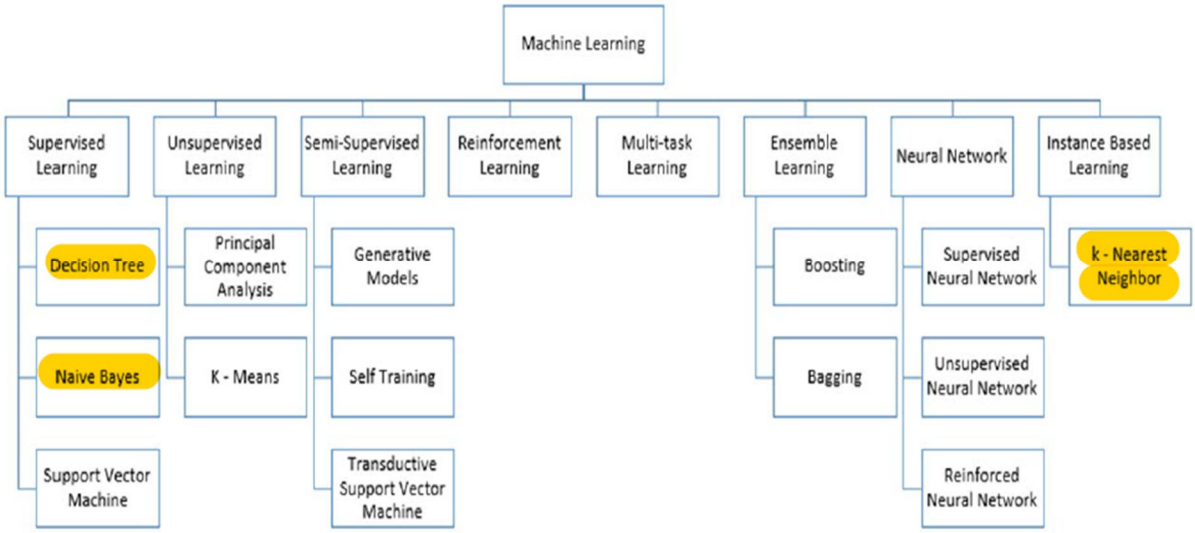


Fig. 1. Types of Learning [1]

II. TYPES OF LEARNING

A. *Supervised Learning*

In supervised learning the data already has predefined tags or labels such as True/False, Positive/Negative, Yes/No, etc. With these kinds of data, the algorithms need particular external assistance. In this kind of learning, we have input variables(x) and an output variable(y) and the we use an algorithm to find out the mapping function, i.e. y = f(x). It is known as supervised learning because the process of the algorithm learning from the training dataset can be envisioned as a teacher supervising the process. Usually new machine learning practitioner will begin supervised machine learning algorithms. These kinds of algorithms are designed to lead or learn by example. While the training is on, the algorithm will search for patterns to corelate with the desired outputs.

©IJRASET: All Rights are Reserved

357

1) *Decision Tree:* A decision tree is a tree like decision making structure which has branches and nodes and each outcome is represented by branches. It is mainly used for classification purposes. Every node represents attributes in a group that is to be classified and each branch represents a value that the node can take [1]. An example of decision tree is given in the Fig. 2.

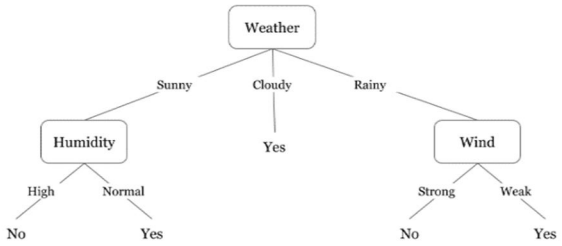


Fig. 2. Decision Tree [2]

A pseudo code for a decision tree on training data D is given in the Fig. 3.

As you can see the Fig. 2. A decision tree has been formed based on weather conditions such as sunny, cloudy, rainy and also various secondary features such as humidity and wind also these cumulatively will determine whether one can play tennis outside or not according to the given conditions on that given day. Now the dataset has various data entries upon which the decision tree algorithm is applied and the model is trained, after that it can give out decisions on new entries that were not previously present on the just trained upon dataset and give out predictions efficiently according how well the data was filtered or pre-processed, generally the accuracy score or r2score of the results if between 90-95% can be considered reliable and good to deploy and beta test in case of low risk models that doesn't involve life and death situations and where the stakes are low as one can say. The algorithm will work nicely on small to medium datasets and will be produce efficient outputs, in case of larger datasets and various attributes the algorithm may be generating various inefficient results and they can't be deployed in the market or even in low stake environment.

```
Input:   D = training dataset defined on attributes and with class values;
Output:  Decision tree <Tree_T0> capable of classifying the training dataset D;

Algorithm Build_Decision_Tree (D)
Begin/* Algorithm <Build_Decision_Tree> */
Tree_T0 = { } ; /* initialization */

/* If we have only one class (i.e., D is "pure") or a stopping criterion is invoked, then stop.
   We have reached a "leaf" node of the decision tree.
   Otherwise, determine the best attribute to split the remaining of the data.
   Proceed recursively until only "leaf" nodes are created and no more splits are possible */
IF (D is "pure" or a stopping criterion is invoked) THEN
        (Create a leaf node that corresponds to D and STOP); /* Tree_T0 has only this node */
ELSE
Begin /* else #1  */
        FOR each attribute in a ϵ D DO
        begin /* do-loop #1 */
                Compute the information-theoretic evaluative value if we split on a
        end /* for-loop #1 */
        aBEST =          Most promising attribute for a split;
        Tree_T0 =        Create a decision node based on attribute aBEST;
        Di =             Induced sub-datasets from D based on aBEST;
        FOR each sub-dataset Di DO
        begin /* do-loop #2 */
                /* Call recursively Algorithm <Build_Decision_Tree> with argument Di */
                Tree_Ti = Build_Decision_Tree (Di);
                Attach Tree_Ti to the corresponding branch of Tree_T0;
        end /* for-loop #2 */
End /* else #1  */
RETURN Tree_T0;
End /* Algorithm <Build_Decision_Tree> */
```

Fig. 3. Pseudo code of Decision Tree [3]

©IJRASET: All Rights are Reserved

358

2) *Naive Bayes:* It is mainly a classification algorithm which targets the text classification industry. It is uses conditional probability as a base for giving outcomes and uses the Bayes theorem as depicted in Fig. 5. as its base. It is also deployed to solve regression problems. It makes the assumption that a particular feature in a class is unrelated to the presence of any other feature [4].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig. 4. Conditional probability

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig. 5. Bayesian Rule

The Bayesian rule can be written in simple terms as:

Posterior = Prior * Likelihood
Evidence

Bayesian networks are used to represent conditional probability models in case of Naive Bayes classifier, it is basically a graphical model that is used for a variety of tasks such as prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction and decision making under the cover of uncertainty. The graphical model can be viewed as a set of variables and their conditional dependencies via a directed acyclic graph (DAG) [4]. A Bayesian Network can be seen in the Fig. 6.

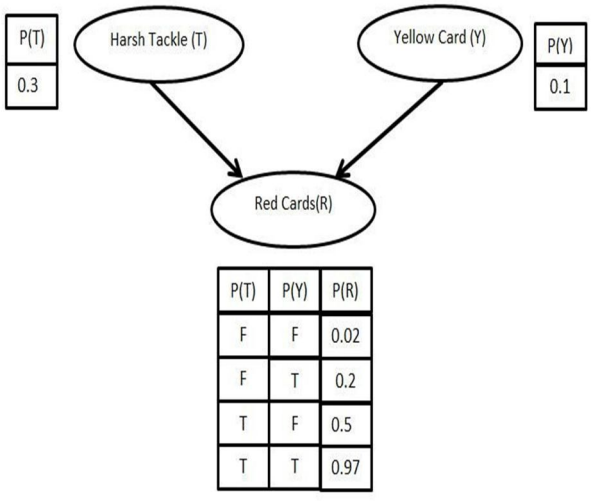


Fig. 6. An Example of Bayesian Network [5]

©IJRASET: All Rights are Reserved

The pseudo code for a Naïve Bayes classifier is given in the Fig. 7.

INPUT: training set T, hold-out set H, initial number of components k_0, and convergence thresholds δ_{EM} and δ_{Add}.

Initialize M with one component.
$k \leftarrow k_0$
repeat
  Add k new mixture components to M, initialized using k random examples from T.
  Remove the k initialization examples from T.
  repeat
    E-step: Fractionally assign examples in T to mixture components, using M.
    M-step: Compute maximum likelihood parameters for M, using the filled-in data.
    If $\log P(H|M)$ is best so far, save M in M_{best}.
    Every 5 cycles, prune low-weight components of M.
  until $\log P(H|M)$ fails to improve by ratio δ_{EM}.
  $M \leftarrow M_{best}$
  Prune low weight components of M.
  $k \leftarrow 2k$
until $\log P(H|M)$ fails to improve by ratio δ_{Add}.
Execute E-step and M-step twice more on M_{best}, using examples from both H and T.
Return M_{best}.

Fig. 7. Pseudo code of Naïve Bayes Classifier [6]

Some noteworthy and known uses of Naive Bayes Classifier are such as in text classifier such as spam classifier, sentimental analysis, information retrieval. Also used in image processing such as face detection softwares
but their efficiency in that field is an ongoing debate but their use in the medical field is under no debate as it can be used in disease detection, e.g. due to the COVID19 outbreak in India in 2020 the Government of India (GOI) through its health ministry launched a COVID19 detection app known as Corona Kavach which will tell the person according to his/her symptoms and their travel history and their current location the probability of them having the disease, these kind of apps use the Naive Bayes classifier to predict outcomes.

3) *Support Vector Machine:* Support Vector Machine (SVM) is the most powerful in the family of supervised algorithms and it is extensively used by developers as they produce higher efficiency in much lesser time and lower computation power than the other algorithms that it is bracketed with. It is extensively deployed in classification problems though it can be both deployed in regression and classification problems. The main goal of an SVM is to find out a hyperplane (a hyperplane is decision boundary separating two or more different classes of data) in an N-dimensional space that classifies the data points. an example of SVM is shown in the Fig. 8.

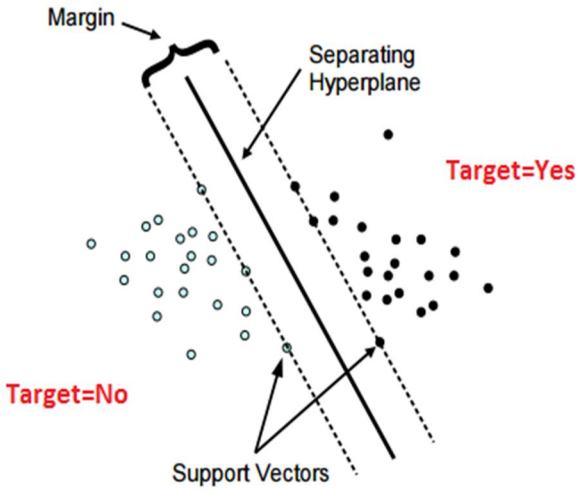


Fig. 8. Working of Support Vector Machine [7]

©IJRASET: All Rights are Reserved

Some main applications of SVM are shown in the Fig. 9.

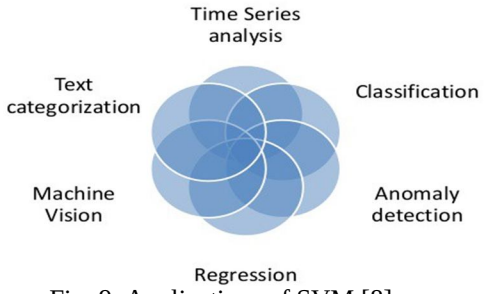


Fig. 9. Applications of SVM [8]

B. Unsupervised Learning

Unsupervised learning in a way is used to find out the hidden patterns in data, it is used in data with more or less no well-defined details and labels with minimum human help. It mainly uses previously gained experience to do the needful and recognize the class of data. It is deployed in clustering and feature reduction. There are issues with unsupervised learning such as: questions over its results, being more complex as compared to supervised learning. Despite of all of these questions over unsupervised learning is very much needed in case of large projects where the size and entries and in one may need to use clustering to gain some perspective on the data before constructing a classifier.

An example of unsupervised learning is shown in the Fig. 10.

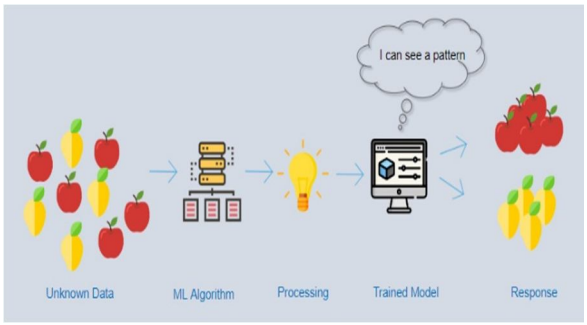


Fig. 10. Example of Unsupervised Learning [9]

1) K-Means Clustering: As discussed earlier unsupervised learning is mainly used for clustering. Now clustering is an algorithm in machine learning when acting upon a dataset creates groups in the data. The set of items that have similar or same characteristics are placed in the same cluster; this is known as k-means clustering as it creates k number of distinct clusters in the data. The center of a particular cluster is its mean. A clustered data is shown in the Fig. 11 and Fig. 12.

Fig.11. original unclustered data

Fig.12. Clustered data

©IJRASET: All Rights are Reserved

361

C. Semi-Supervised Learning

In the case of supervised learning the data that is trained upon well-defined and labeled data and in the case of unsupervised learning the data that is trained upon has no labels and it's the algorithms job to find out patterns in the data and give out various outcomes. Semi-Supervised learning can be categorized as then grey area between supervised and unsupervised learning where the algorithms works on mostly unlabeled data but with some labeled data, the best example of deployment of such algorithms are seen in text documents [10]. The working principle of semi-supervised learning is depicted in Fig. 13.

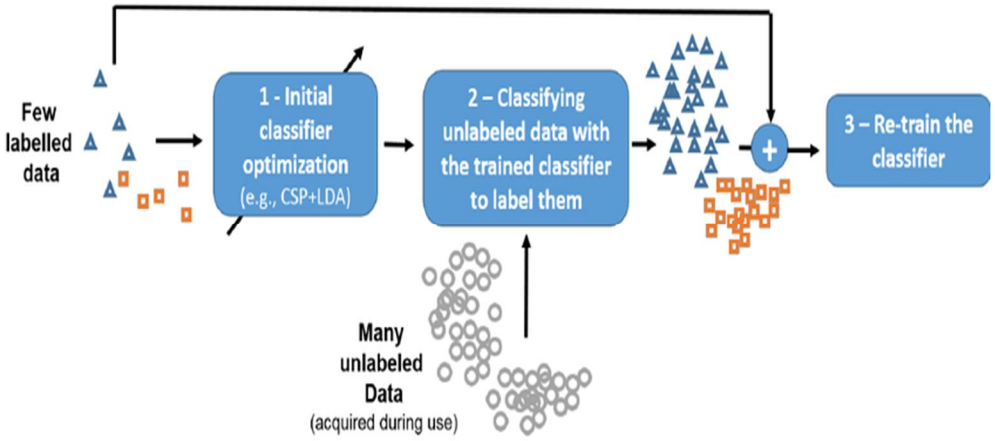


Fig.13. Working of Semi-Supervised Learning [11]

D. Reinforcement Learning

It is machine learning technique that helps an agent to learn in an interactive environment by the mans of trial and error techniques using feedback from the previous actions and gaining from previous experience. This technique uses a policy of rewarding and punishment instead of providing feedback for taking right steps in the case of supervised learning. It is to be noted that the learner has no previous knowledge of which steps to take until it faces a situation. The main application of reinforcement leaning is in the case of pure artificial intelligence where the next step take determines the outcomes of the future. The model for reinforcement learning is depicted in the Fig. 14.

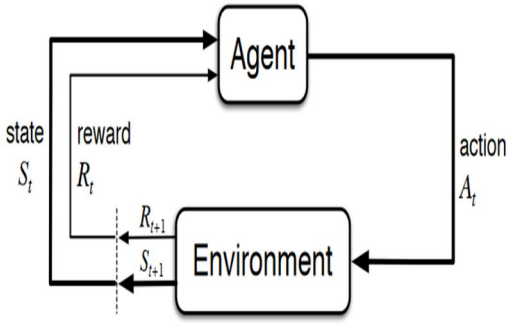


Fig.14. Reinforcement Learning Model [12]

E. Multitask Learning

Multitask learning uses generalization to solve and help others solve problems, in this the algorithm while solving tasks, the algorithm remembers the steps it had performed in that task, it later uses these experiences to solve other similar or same problems.

F. Ensemble Learning

Well this type of learning essentially means what ensemble means i.e. group of individuals viewed as a whole rather than an individual, that implies in this type of learning multiple learning models such as naïve Bayes, decision tress ,etc. can be combined and applied to a problem to achieve more accurate results. Two highly used Techniques are: Boosting and Bagging.

©IJRASET: All Rights are Reserved

362

G. Neural Network Learning

The neural network learning draws inspiration from the working biological neurons in the human body. The biological neuron mainly possesses three main working components namely dendrites, soma and axiom which are responsible for receiving electrical signals, processing of the signal and carrying the output respectively. Now in neural networks there are also three components namely input layer, hidden layer and output layer and their functions are like the dendrites, soma and axiom of the biological neurons respectively. The Fig. 15 and Fig. 16 will offer a good visualization of both.

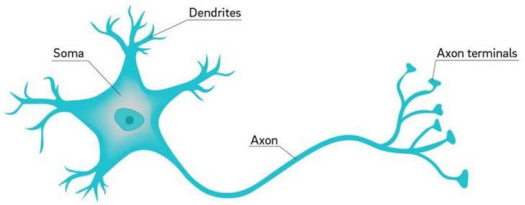


Fig.15. A Neuron [13]

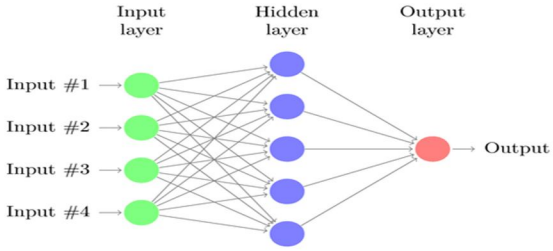


Fig.16. A Neural Network [14]

There are mainly three types of three types of neural network learning: supervised neural network, unsupervised neural network, reinforced neural network.

H. Instance-Based Learning: This type of learning is also known as memory based learning ,therein lies the real meaning , here what happens is that the algorithms when first trains a set of data (training data) it memorizes the results and the patterns and then when new set of instances comes it uses these previous trained instances, but the results totally depend upon the preprocessing of the data fed into the algorithm , if done correctly maximum efficiency can be achieved .

1) k-Nearest Neighbor: k-nearest neighbor (k-NN) takes a training data and a model is trained and when new data is introduced, it compares both the data, and k closest training examples are taken out. This type of algorithm works for both classification and regression problems. A depiction is shown in Fig. 17 is shown where the k number of instances are selected in the training data for a test data. k is basically an integer(1,2,3…) that needs to be chosen and then the distance of the test data to each row of the training data is calculated and they are arranged in ascending order and then top k rows are taken and categories are defined and the category which has more number of neighbors, the new data point is assigned to that category.

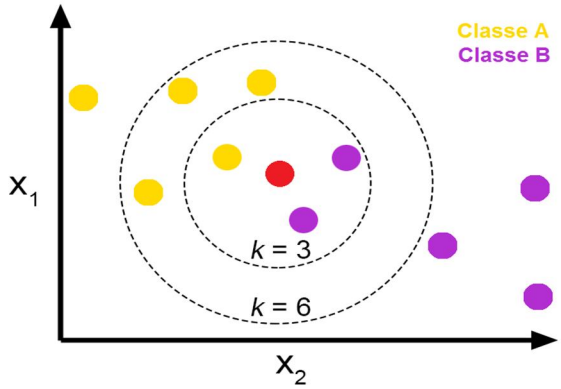


Fig.17. graph showing k neighbors [15]

©IJRASET: All Rights are Reserved

363

Now there is huge debate about what value of k should be chosen, whether a small value should be chosen, well that has its deficiencies and can lead to outliers being chosen and be noisy, well large values of k should be chosen as it reduces noise and outlier effects are not applicable here and good accuracy can be generated. The pseudo code for kNN is shown in the Fig.18.

Let $W = \{x_1, x_2, \cdots, x_n\}$ be a set of n labeled samples. The algorithm is as follows:

```
BEGIN
  Input y, of unknown classification.
  Set K, 1 ≤ K ≤ n.
  Initialize i = 1.
  DO UNTIL (K-nearest neighbors found)
    Compute distance from y to x_i.
    IF (i ≤ K) THEN
      Include x_i in the set of K-nearest neighbors
    ELSE IF (x_i is closer to y than any previous nearest
      neighbor) THEN
      Delete farthest in the set of K-nearest neighbors
      Include x_i in the set of K-nearest neighbors.
    END IF
    Increment i.

  END DO UNTIL
  Determine the majority class represented in the set of K-
  nearest neighbors.
  IF (a tie exists) THEN
    Compute sum of distances of neighbors in each class
    which tied.
    IF (no tie occurs) THEN
      Classify y in the class of minimum sum
    ELSE
      Classify y in the class of last minimum found.
    END IF
  ELSE
    Classify y in the majority class.
  END IF
END
```

Fig.18. Pseudo code for kNN [16]

III. ANALYSIS

This section will consist of the advantages and disadvantages of mainly Decision tree, Naive Bayes and kNN algorithms and comparative analysis will be done.

A. *Advantages of Decision Tree*

1) Normalization not required.
2) Less emphasis on the pre-processing phase.
3) The problem of missing field doesn't affect the tree that much.

B. *Disadvantages of Decision Tree*

1) Sometimes calculations do become complex.
2) Training time is much higher.
3) It is not well equipped to handle regression problems [17].

C. *Advantages of Naïve Bayes*

1) It's able to solve prediction problems.
2) It is much wise of it deploy incase the data has very high level of noise [18].
3) Involves linear scalability.

D. *Disadvantages of Naive Bayes*

1) Error percentage is high in classification problems.
2) Small amount of data can lead to misleading results.
3) The independent assumption may lead to a problem in regression problems.

©IJRASET: All Rights are Reserved

E. Advantages of k-Nearest Neighbors

1) Simple implementation.
2) No training periods.
3) Since no training period new data can be added easily.

F. Disadvantages of k-Nearest Neighbors

1) Noise in the data may hamper its efficiency.
2) Choosing the incorrect value of k may result in a decrease in efficiency and generate incorrect results.
3) Computation cost is too high

G. Decision Tree vs Naïve Bayes vs SVM

The performances of the above three algorithms were compared using a set of tweets with various labels, the data was taken from Sentimental140 dataset. It was preprocessed using python. A table comparing algorithms based on training time, prediction time and accuracy of prediction is shown in the Table. 1 [19].

Algorithm	Training Time (s)	Prediction Time (s)	Accura-cy
Naïve Bayes	2.708	0.328	0.692
SVM	6.485	2.054	0.6565
Decision Tree	454.609	0.063	0.69

Table. 1. Comparison between Naïve Bayes, Decision Tree, SVM [19]

When comparing kNN with all the others, are they are not so good and kNN has the upper edge upon all the other classification algorithms and its efficiency in case of large data is incomparable as shown in Fig. 19.

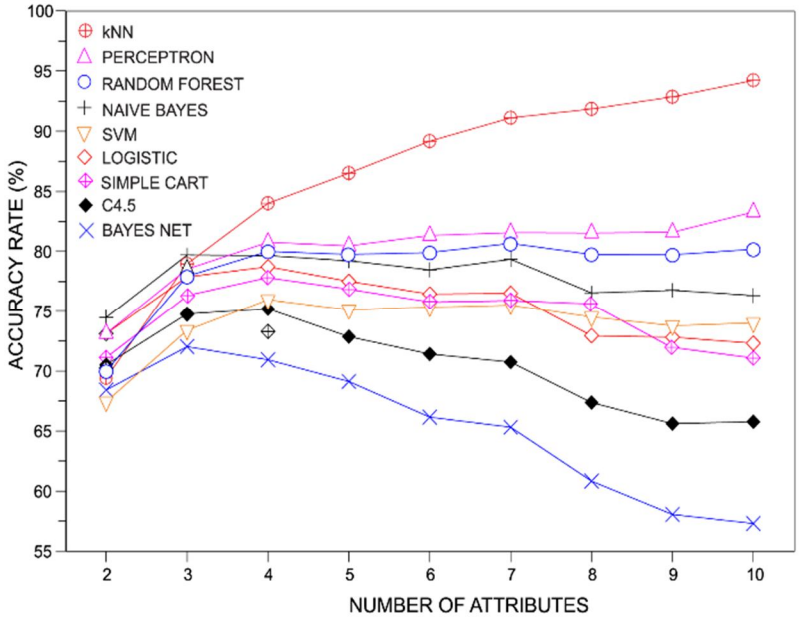


Fig.19. Graph showing comparison of accuracy rate of various ML algorithms [20]

©IJRASET: All Rights are Reserved

365

IV. CONCLUSION

This paper is a survey of various machine learning algorithms, with specific focus on three that are mainly used in classification and regression algorithms. The main aim was to help the reader to choose an algorithm with a free mind for his work by highlighting various good and bad aspects of the algorithms.

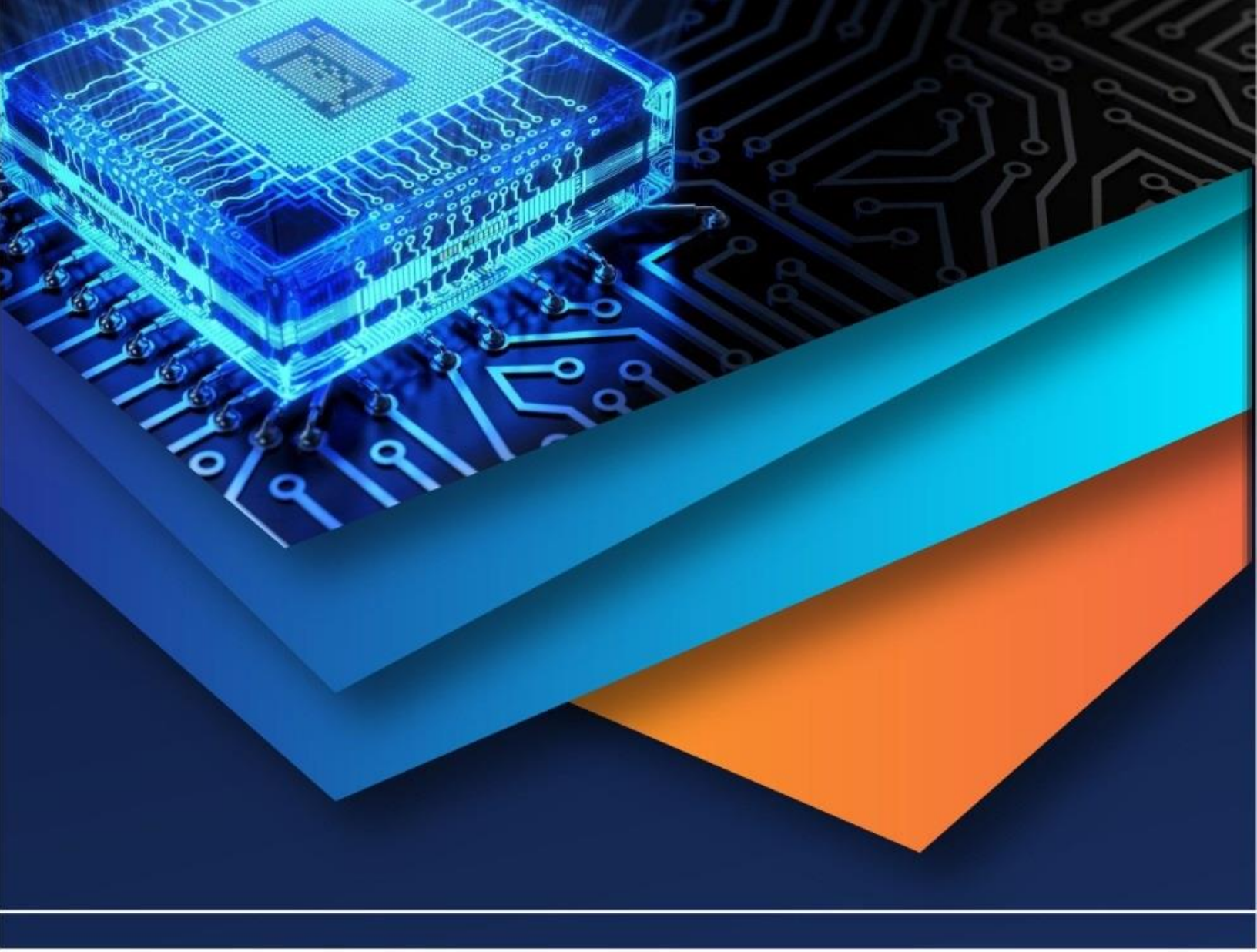
Well speaking of all the good about these algorithms is not all, every algorithm will work as good as the data is preprocessed if that is not done correctly the results maybe really disappointing. If everything is done correctly it will be wise to choose kNN in case of complex and large datasets and the Naive-Bayes in case of small and more or less common problems.

Machine Learning has already become an integral part of the digital world and will occupy more spaces in the time to come and the scope for studying this field will become more wider and field will become more and more powerful in the coming days. More and more techies are encouraged to take this wonderful field that has the capability to augment the human mind and perhaps outmatch it in the coming decades.

REFERENCES

[1] Ayon Dey, "Machine Learning Algorithms: A Review" International Journal of Computer Science and Information Technologies (IJCSIT), Vol. 7 (3), 2016, 1174-1179
[2] https://s3-ap-southeast-.amazonaws.c om/he-public-data/Fig%201-ze1a01b.png
[3] https://www.researchgate.net/profile /Evangelos_Triantaphyllou/publication/334533554
[4] https://www.kdnuggets.com/2019/04/naive-bayes-baseline-model-machine-learning-classification-performance.html
[5] https://miro.medium.com/max/1556/1*0BTzommpUoMfNBFDxvJ5sg.png
[6] D. Lowd, P. Domingos, "Naïve Bayes Models for Probability Estimation"
[7] https://static.wixstatic.com/media/8f929f_7ecacdcf69d2450087cb4a898ef90837~mv2.png
[8] https://miro.medium.com/max/1276/1*C1MId293xqoVio8av9Rc2g.jpeg
[9] https://www.simplilearn.com/ice9/free_resources_article_thumb/unsup.jpg
[10] X. Zhu, A. B. Goldberg, "Introduction to Semi – Supervised Learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 2009, Vol. 3, No. 1, Pages 1-130
[11] https://www.researchgate.net/profile/Fabien_LOTTE/publication/277605013/figure/fig4/AS:281234816159750@1444063013959/Principle-of-semi-supervised-learning-1-a-model-eg-CSP-LDA-classifier-is-first.png
[12] https://miro.medium.com/max/1400/1/1*-0G8EIeG24OYTbt5KZSalQ.png
[13] https://scx1.b-cdn.net/csz/news/800/2018/2-whyareneuron.jpg
[14] https://res.cloudinary.com/practicaldev/image/fetch/s--kakYFNCR--/c_limit%2Cf_auto%2Cfl_progressive%2Cq_auto%2Cw_880/https://thepracticaldev.s3.amazonaws.com/i/j4igfbcbeafcuwmthyoy.png
[15] https://miro.medium.com/max/753/0*jqxx3-dJqFjXD6FA
[16] J. M. Keller, M. R. Gray, J. A. Givens Jr., "A Fuzzy K-Nearest Neighbor Algorithm", IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-15, No. 4, August 1985
[17] https://medium.com/@dhiraj8899/top-5-advantages-and-disadvantages-of-decision-tree-algorithm-428ebd199d9a
[18] Sunpreet Kaur, Sonika Jindal, "A Survey on Machine Learning Algorithms", International Journal of Innovative Research in Advanced Engineering (IJIRAE), Vol. 3 (11), November 2016, 2349-2763
[19] Kajaree Das, Rabi Narayan Behera, "A Survey on Machine Learning: Concept, Algorithms and Applications", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017
[20] https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0094137

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)