International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

||Volume 8, Issue 7, July 2020||

# A Survey on Machine Learning Techniques, Applications and Tools

**Satyendra Bunkar**[1], **Dr. Jitendra Sheetlani**[2]

Research Scholar, Department of Computer Applications, SSSUTMS, Sehore, Madhya Pradesh, India[1]

Associate Professor, Department of Computer Applications, SSSUTMS, Sehore, Madhya Pradesh, India[2]

**ABSTRACT**: Machine learning is a reasonably novel discipline within Computer Science which is the sub-module of artificial intelligence that makes available anassortment of data analysis techniques. Some of these techniques are based on well-established statistical methods (e.g. logistic regression and principal component analysis) while many others are not. The machine learning is basically used for the predictive analysis and classification. The machine learning (ML) techniques also used in many applications like networking, weather forecasting, health care, multimedia etc. In this paper, we present the survey on various machine learning techniques or algorithm with their advantages and disadvantages. We also present the area of applications of machine learning in various sectors and tools used for the simulation of machine learning programming.

**KEYWORDS:** Machine Learning, Artificial Intelligence, Principal Component Analysis (PCA), Logistic Regression, Health care.

## I.INTRODUCTION

Machine learning is a paradigm that may refer to learning from past experience (which in this case is previous data) to improve future performance. The sole focus of this field is automatic learning methods.[1] Learning refers to modification or improvement of algorithm based on past "experiences" automatically without any external assistance from human.

While designing a machine (a software system), the programmer always has a specific purpose in mind. For instance, consider J. K. Rowling's Harry Potter Series and Robert Galbraith's Cormoran Strike Series. To confirm the claim that it was indeed Rowling who had written those books under the name Galbraith, two experts were engaged by The London Sunday Times and using Forensic Machine Learning they were able to prove that the claim was true.[2,3] They develop a machine learning algorithm and "trained" it with Rowling's as well as other writers writing examples to seek and learn the underlying patterns and then "test" the books by Galbraith. The algorithm concluded that Rowling's and Galbraith's writing matched the most in several aspects. So instead of designing an algorithm to address the problem directly, using Machine Learning, a researcher seek an approach through which the machine, i.e., the algorithm will come up with its own solution based on the example or training data set provided to it initially. Machine learning and algorithms are used in recommender systems, Web search, spam filters, ad placement, credit scoring, decision support systems, fraud detection, and many other applications.[21, 22]

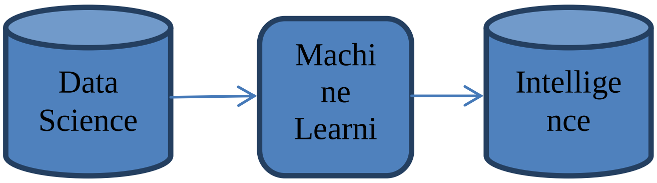

Fig.1: Steps of Machine Learning

This paper presents the survey on machine learning techniques, its applications and tools used for the simulation work. The organization of rest paper is done as follows:

Section II presents the classification of various techniques and algorithm of machine learning. In section III, briefly describe the applications of machine learning in various sectors. Section IV present the various machine learning tools which is used for simulation purpose. And last section briefly present the overall conclusion of research work and their future aspects.

## II.CLASSIFICATION OF MACHINE LEARNING

The machine learning techniques is classified into three categories namely supervised learning, unsupervised learning and reinforcement learning and the machine learning also uses some algorithm to implement these techniques such as regression algorithm, logistic regression, Multivariate adaptive regression etc.

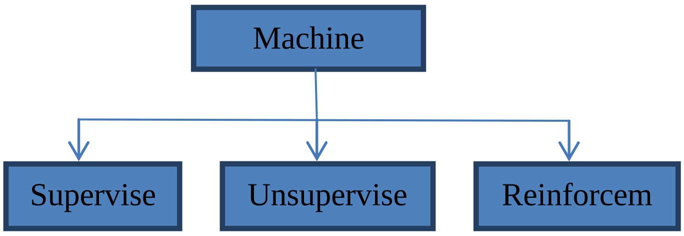

Fig.2: Classification of Machine Learning

### 2.1 Supervised Learning

It is a task driven machine learning technique which predict the next value. If the data are totally labeled, the problem is called supervised learning and mostly the task is to discover a function or model that explains the data.[4] There are many supervised learning techniques such as decision tree, random forest, support vector machine, naïve bayes classifier, linear regression, logistic regression etc.

**Challenges in Supervised learning**

Here, are challenges faced in supervised machine learning:

- Irrelevant input feature present training data could give inaccurate results
- Data preparation and pre-processing is always a challenge.
- Accuracy suffers when impossible, unlikely, and incomplete values have been inputted as training data
- If the concerned expert is not available, then the other approach is "brute-force." It means you need to think that the right features (input variables) to train the machine on. It could be inaccurate.

**Advantages of Supervised Learning:**

- Supervised learning allows you to collect data or produce a data output from the previous experience
- Helps you to optimize performance criteria using experience
- Supervised machine learning helps you to solve various types of real-world computation problems.

**Disadvantages of Supervised Learning**

- Decision boundary might be overtrained if your training set which doesn't have examples that you want to have in a class
- You need to select lots of good examples from each class while you are training the classifier.
- Classifying big data can be a real challenge.
- Training for supervised learning needs a lot of computation time.

#### 2.1.1 Decision Tree

A decision tree is a tree-like structure that has leaves, which represent classifications and branches, which in turn represent the conjunctions of features that lead to those classifications. An exemplar is labeled (classified) by testing its feature (attribute) values against the nodes of the decision tree. The best known methods for automatically building decision trees are the ID3 [5] and C4.5 [6] algorithms. Both algorithms build decision trees from a set of training data using the concept of information entropy. When building the decision tree, at each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of examples into subsets. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then performs recursion on the smaller subsets until all the training examples have been classified. The advantages of decision trees are intuitive knowledge expression, high classification accuracy, and simple implementation. The main disadvantage is that for data including categorical variables with a different number of levels, information gain values are biased in favor of features with more levels. The decision tree is built by maximizing the information gain at each variable split, resulting in a natural variable ranking or feature selection. Small trees (such as the one depicted in Fig. 4) have an intuitive knowledge expression for experts in a given domain because it is easy to extract rules from those trees just by examining them. For deeper and wider trees, it is much more difficult to extract the rules and thus the larger the tree, the less intuitive its knowledge expression. Smaller trees are obtained from bigger ones by pruning. Larger trees often have high classification accuracy but not very good generalization capabilities. By pruning larger trees, smaller trees are obtained that often have better generalization capabilities (they avoid over-fitting). Decision tree building algorithms (e.g., C4.5) are relatively simpler than more complex algorithms such as SVMs.

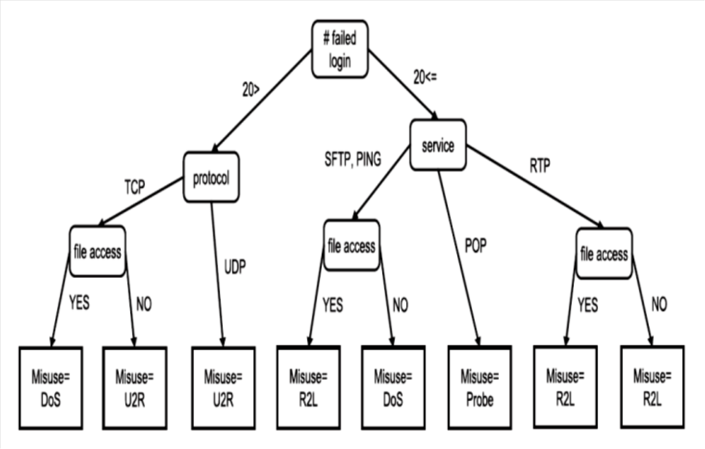

Fig. 3: Decision Tree [4]

### 2.1.2 Random Forest

A Random Forest is a classifier consisting of a collection of tree-structured classifiers {h(x, Θk) k=1, 2, ….}, where the {Θk } are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x [9]. Random Forest generates an ensemble of decision trees. To generate each single tree in Random Forest, Breiman followed following steps: If the number of records in the training set is N, then N records are sampled at random but with replacement, from the original data; this is bootstrap sample. This sample will be the training set for growing the tree. If there are M input variables, a number m << M is selected such that at each node, m variables are selected at Random out of M and the best split on these m attributes is used to split the node. The value of m is held constant during forest growing

In this way, multiple trees are induced in the forest; the number of trees is pre-decided by the parameter Ntree. The number of variables (m) selected at each node is also referred to as mtry or k in the literature. The depth of the tree can be controlled by a parameter node size (i.e. number of instances in the leaf node) which is usually set to one. Once the forest is trained or built as explained above, to classify a new instance, it is run across all the trees grown in the forest. Each tree gives classification for the new instance which is recorded as a vote. The votes from all trees are combined and the class for which maximum votes are counted (majority voting) is declared as classification of the new instance.

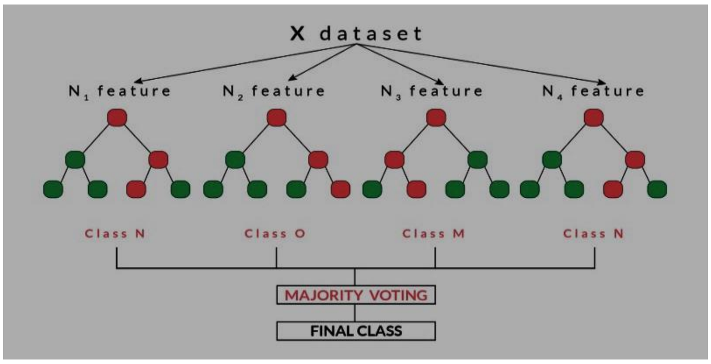

Fig.5: Random Forest

2.1.3 Naïve Bayes Classifier

We assume that a data set contains n instances (or cases) $x_i$ , i = 1..n, which consist of p attributes, i.e., $x_i = (x_{i1}, x_{i2}, ..., x_{ip})$. Each instance is assumed to belong to one (and only one) class $y \in \{y_1, y_2, ..., y_c\}$. Most predictive models in machine learning generate a numeric score s for each instance $x_i$. This score quantifies the degree of class membership of that case in class $y_j$. If the data set contains only positive and negative instances, y ∈ {0, 1}, then a predictive model can either be used as a ranker or as a classifier. The ranker uses the scores to order the instances from the most to the least likely to be positive. By setting a threshold t on the ranking score, s(x), such that {s(x) ≥ t} = 1, the ranker becomes a (crisp) classifier [10]. Naive Bayes learning refers to the construction of a Bayesian probabilistic model that assigns a posterior class probability to an instance: $P(Y = y_j |X = x_i)$. [11] The simple naive Bayes classifier uses these probabilities to assign an instance to a class. Applying Bayes' theorem (Eq. $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, and simplifying the notation a little, we obtain

$$P(y_j|x_i) = \frac{P(x_i|y_j)P(y_j)}{P(x_i)}$$

## 2.2 Unsupervised Learning

Social network analysis, genes clustering and market research are among the most successful applications of unsupervised learning methods. In the case of unsupervised learning the training dataset consists only of a set of input vectors x. While unsupervised learning can address different tasks, clustering or cluster analysis is the most common. Clustering is the process of grouping data so that the intracluster similarity is high, while the inter-cluster similarity is low. The similarity is typically expressed as a distance function, which depends on the type of data. There exists a variety of clustering approaches. Here, we focus on two algorithms; k-means and Gaussian mixture model as examples of partitioning approaches and model-based approaches, respectively, given their wide area of applicability. The reader is referred to [7] for a comprehensive overview of cluster analysis.Here, are prime reasons for using Unsupervised Learning:

- Unsupervised machine learning finds all kind of unknown patterns in data.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

**Unsupervised Learning Issues:**

- Unsupervised Learning is harder as compared to Supervised Learning tasks.
- How do we know if results are meaningful since no answer labels are available?
- Let the expert look at the results (external evaluation)
- Define an objective function on clustering (internal evaluation)

**Advantages:**

- Less complexity in comparison with supervised learning. Unlike in supervised algorithms, in unsupervised learning, no one is required to understand and then to label the data inputs. This makes unsupervised learning fewer complexes and explains why many people prefer unsupervised techniques.
- Takes place in real time such that all the input data to be analyzed and labeled in the presence of learners. This helps them to understand very well different models of learning and sorting of raw data.
- It is often easier to get unlabeled data — from a computer than labeled data, which need person intervention. This is also a key difference between supervised and unsupervised learning.

**Disadvantages of Unsupervised Learning**

- You cannot get precise information regarding data sorting, and the output as data used in unsupervised learning is labeled and not known
- Less accuracy of the results is because the input data is not known and not labeled by people in advance. This means that the machine requires to do this itself.
- The spectral classes do not always correspond to informational classes.
- The user needs to spend time interpreting and label the classes which follow that classification.
- Spectral properties of classes can also change over time so you can't have the same class information while moving from one image to another.

### 2.2.1 K-means Clustering

K-means clustering is considered to be a very famous and widely used partitioning based approach. It is a numerical, non-deterministic, an unsupervised and an iterative approach. In K-means clustering, the average value (mean value) of objects in the group represents individual cluster. The main objective of the K-means clustering algorithm is to acquire the stable number of clusters that reduces the Euclidean distances between data objects and center of clusters. [12] Let D={di}, i=1,2,……,n is a dataset of n-objects, and let K-is the numbers of clusters to be formed. Cj is the center of the cluster where j=1,2,…..,k. The distance between centroid and data object is measured by using Euclidean distance formula. The Euclidean distance is calculated by using sum of squares of distances [14]. Let A and B are two objects in a dataset. The Euclidean distance between A and B is calculated by using the below formula. Euclidean Distance (A, B) = $(|A_1\text{-}B_1|^2 + |A_2\text{-}B_2|^2 + \ldots\ldots\ldots\ldots +|A_{n\text{-}1}\text{-}B_{n\text{-}1}|^2 +|A_n\text{-}B_n|^2)^{½}$. It can be summarized as; Euclidean distance $=\sqrt{}\Sigma|A_i\text{-}B_i|^2$, i=1,2,…….n-1,n. Each cluster is filled with the objects which are near to the center of the cluster in the data space of cluster nodes. Later, center of the cluster updated by itself as the number of objects increases. This process is repeated until it reaches all the objects of cluster. [13]For various real-world applications, the k-means clustering algorithm has been shown to be efficient in generating excellent cluster outcomes. Generally, k-means clustering algorithm's required time is proportional to the number of clusters per every iteration and to the number of patterns generated. This method is particularly very expensive for big datasets. [13]

### 2.2.2 Principal Component Analysis

Principal Component Analysis (PCA) is an important method in machine learning due to its twofold nature. PCA reduces the dimensionality of the dataset, which takes the dimensions that encode the most important information and removes the dimensions that encode the least important information. By reducing the number of dimensions, the data utilizes less space, thus allowing classification on larger datasets in less time. Further, by taking only the salient dimensions, PCA projects the dataset onto dimensions that hold the most meaning, thus drawing out patterns in the dataset [15]. PCA is a useful statistical technique that has found application in fields such as face recognition and image compression and is a common technique for finding patterns in data of high dimension. But a major problem in mining scientific data sets is that the data is often high dimensional. When the number of dimensions reaches hundreds or even thousands, the computational time for the pattern recognition algorithms can become prohibitive. In many cases there are a large number of features representing the object. One problem is that the computational time for the pattern recognition [16].Principal Component Analysis (PCA, also called Karhunen-Loeve transform) is used for dimensionality reduction techniques of data analysis and compression. It is based on transforming a relatively large number of variables into a smaller number of uncorrelated variables by finding a few orthogonal linear combinations of the original variables with the largest variance. The first principal component of the transformation is the linear combination of the original variables with the largest variance; the second principal component is the linear combination of the original variables with the second largest variance and orthogonal to the first principal component and so on.

### 2.2.3 Association Rule

Association rules or association analysis is also an important topic in data mining. This is an unsupervised method, so we start with an unlabeled dataset.[16] An **unlabeled dataset** is a dataset without a variable that gives us the right answer. Association analysis attempts to find relationships between different entities. The classic example of association rules is market basket analysis. This means using a database of transactions in a supermarket to find items that are bought together. For example, a person who buys potatoes and burgers usually buys beer. This insight could be used to optimize the supermarket layout.

Online stores are also a good example of association analysis. They usually suggest to you a new item based on the items you have bought. They analyze online transactions to find patterns in the buyer's behavior.

These algorithms assume all variables are categorical; they perform poorly with numeric variables. Association methods need a lot of time to be completed; they use a lot of CPU and memory. Remember that Rattle runs on R and the R engine loads all data into RAM memory.

Suppose we have a dataset such as the following:

Our objective is to discover items that are purchased together. We'll create rules and we'll represent these rules like this:

Chicken, Potatoes → Clothes

This rule means that when a customer buys **Chicken** and **Potatoes**, he tends to buy **Clothes**.

As we'll see, the output of the model will be a set of rules. We need a way to evaluate the quality or interest of a rule. There are different measures, but we'll use only a few of them. Rattle provides three measures:

- **Support**
- **Confidence**
- **Lift**

**Support** indicates how often the rule appears in the whole dataset. In our dataset, the rule Chicken, Potatoes → Clothes has a support of 48.57 percent (3 occurrences / 7 transactions).

**Confidence** measures how strong rules or associations are between items. In this dataset, the rule Chicken, Potatoes → Clothes has a confidence of **1**. The items Chicken and Potatoes appear three times in the dataset and the items Chicken, Potatoes, and Clothes appear three times in the dataset; and 3/3 = 1. A confidence close to **1** indicates a strong association.

### 2.3 Reinforcement Learning

The computer works in a dynamic environment in which it has to complete a goal without the computer being explicitly told if the goal is reached (Fig. 4). This can be seen in unmanned vehicles.[8]

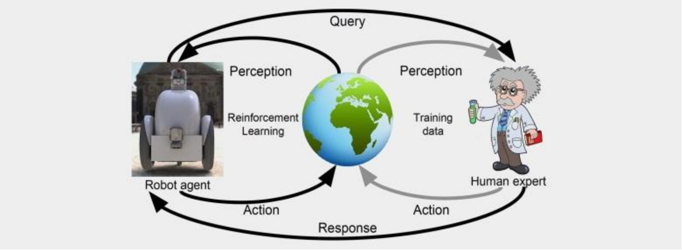

Fig. 4 Reinforcement Learning [8]

**Advantages**

- This learning model is very similar to the learning of human beings. Hence, it is close to achieving perfection.
- The model can correct the errors occurred during the training process.
- Once an error is corrected by the model, the chances of occurring the same error are very less.
- It can create the perfect model to solve a particular problem.
- Robots can implement reinforcement learning algorithms to learn how to walk.

**Disadvantages**

- Reinforcement learning as a framework is wrong in many different ways, but it is precisely this quality that makes it useful.
- Too much reinforcement learning can lead to an overload of states which can diminish the results.
- Reinforcement learning is not preferable to use for solving simple problems.
- The curse of dimensionality limits reinforcement learning heavily for real physical systems.

**Challenges of Reinforcement Learning**

Here are the major challenges you will face while doing Reinforcement earning:

- Feature/reward design which should be very involved
- Parameters may affect the speed of learning.
- Realistic environments can have partial observability.
- Too much Reinforcement may lead to an overload of states which can diminish the results.
- Realistic environments can be non-stationary.

Two kinds of reinforcement learning methods are:

#### 2.3.1 Positive Reinforcement

It is defined as an event, which occurs because of specific behavior. It increases the strength and the frequency of the behavior and impacts positively on the action taken by the agent. This type of Reinforcement helps you to maximize performance and sustain change for a more extended period. However, too much Reinforcement may lead to over-optimization of state, which can affect the results.

#### 2.3.2 Negative Reinforcement

Negative Reinforcement is defined as strengthening of behavior that occurs because of a negative condition which should have stopped or avoided. It helps you to define the minimum stand of performance. However, the drawback of this method is that it provides enough to meet up the minimum behavior.

**1) Q-Learning**

Q learning is a value-based method of supplying information to inform which action an agent should take.
Let's understand this method by the following example:

- There are five rooms in a building which are connected by doors.
- Each room is numbered 0 to 4

- The outside of the building can be one big outside area (5)
- Doors number 1 and 4 lead into the building from room 5

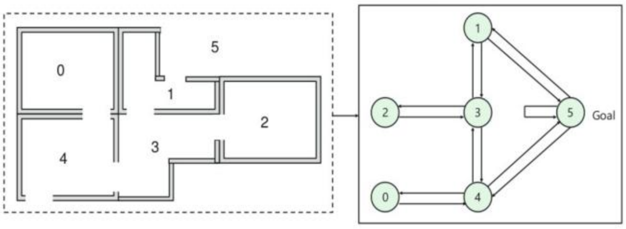

Fig. 5 Building example of Q Learning

Next, you need to associate a reward value to each door:

- Doors which lead directly to the goal have a reward of 100
- Doors which is not directly connected to the target room gives zero reward
- As doors are two-way, and two arrows are assigned for each room
- Every arrow in the above image contains an instant reward value

## III.APPLICATIONS OF MACHINE LEARNING

The value of machine learning technology has been recognized by companies across several industries that deal with huge volumes of data. By leveraging insights obtained from this data, companies are able work in an efficient manner to control costs as well as get an edge over their competitors. [17,23] This is how some sectors / domains are implementing machine learning -

- **Financial Services**
  Companies in the financial sector are able to identify key insights in financial data as well as prevent any occurrences of financial fraud, with the help of machine learning technology. The technology is also used to identify opportunities for investments and trade. Usage of cyber surveillance helps in identifying those individuals or institutions which are prone to financial risk, and take necessary actions in time to prevent fraud.
- **Marketing and Sales**
  Companies are using machine learning technology to analyze the purchase history of their customers and make personalized product recommendations for their next purchase. This ability to capture, analyze, and use customer data to provide a personalized shopping experience is the future of sales and marketing.
- **Government**
  Government agencies like utilities and public safety have a specific need FOR Ml, as they have multiple data sources, which can be mined for identifying useful patterns and insights. For example sensor data can be analyzed to identify ways to minimize costs and increase efficiency. Furthermore, ML can also be used to minimize identity thefts and detect fraud.[17]
- **Healthcare**
  With the advent of wearable sensors and devices that use data to access health of a patient in real time, ML is becoming a fast-growing trend in healthcare. Sensors in wearable provide real-time patient information, such as overall health condition, heartbeat, blood pressure and other vital parameters. Doctors and medical experts can use this information to analyze the health condition of an individual, draw a pattern from the patient history, and predict the occurrence of any ailments in the future. The technology also empowers medical experts to analyze data to identify trends that facilitate better diagnoses and treatment.
- **Transportation**
  Based on the travel history and pattern of traveling across various routes, machine learning can help transportation companies predict potential problems that could arise on certain routes, and accordingly advise their customers to opt for a different route. Transportation firms and delivery organizations are increasingly using machine learning technology to carry out data analysis and data modeling to make informed decisions and help their customers make smart decisions when they travel.[17]
- **Oil and Gas**
  This is perhaps the industry that needs the application of machine learning the most. Right from analyzing underground minerals and finding new energy sources to streaming oil distribution, ML applications for this industry are vast and are still expanding.

## IV. MACHINE LEARNING TOOLS

There is a variety of machine learning tools such as Scikit-learn, PyTorch, WEKA, TensorFlow, RapidMiner, R Studio etc., and the explanation about these method are given below:

### 4.1 Scikit-learn

Scikit-learn is for machine learning development in python. It provides a library for the Python programming language.[18]

Features:

- It helps in data mining and data analysis.
- It provides models and algorithms for Classification, Regression, Clustering, Dimensional reduction, Model selection, and Pre-processing.

Advantages:

- Easily understandable documentation is provided.
- Parameters for any specific algorithm can be changed while calling objects.

### 4.2 PyTorch

PyTorch is a Torch based, Python machine learning library. The torch is a Lua based computing structure, scripting verbal communication, and machine learning library.[122]

Features:

- It helps in building neural networks through Auto grade Module.
- It provides an assortment of optimization algorithms for structure neural networks.
- PyTorch can be used on cloud platforms.
- It provides distributed training, various tools, and libraries.

Advantages:

- It helps in creating computational graphs.
- Ease of use because of the hybrid front-end.

### 4.3 WEKA

These machine learning algorithms help in data mining.[122]

Features:

- Data preparation
- Classification
- Regression
- Clustering
- Visualization and
- Association rules mining.

Advantages:

- Provides online courses for training.
- Easy to comprehend algorithms.
- It is good for students as well.

Disadvantages:

Not much documentation and online maintain are accessible.

### 4.4 TensorFlow

Tensor Flow provides a JavaScript records which helps in machine learning. APIs will help you to build and train the models.

Description:

Helps in training and construction your models.

You can run your existing models with the facilitate of Tensor Flow.js which is a model converter. It helps in the neural network.

Pros:

- You can use it in two behaviour, i.e. by characters tags or by installing through NPM.
- It can smooth help for human pose inference.

Advantages:

- It is difficult to learn.

### 4.5 Rapid Miner

Rapid Miner provides a platform for machine learning, deep learning, data research, text mining, and extrapolative analytics. It can be used for research, learning and application development.

Features:

- Through GUI, it helps in designing and implementing logical workflows.
- It helps with data training.
- Result Visualization.
- Model validation and optimization.

Advantages:

- Extensible during plugins.
- Easy to utilize.
- No programming skills are required.

Disadvantages:

- The tool is expensive.

### 4.6 R Studio

**R Studio** is an integrated development environment (IDE) for R, a programming language for statistical computing and graphics. It is available in two formats: RStudio Desktop is a regular desktop application while RStudio Server runs on a remote server and allows accessing R Studio using a web browser.

R Studio Desktop and R Studio Server are together available in free and fee-based (commercial) editions. OS sustain depends on the format/edition of the IDE. Prepackaged distributions of RStudio Desktop are accessible for Windows, macOS, and Linux. RStudio Server and Server Pro run on Debian, Ubuntu, Red Hat Linux, CentOS, openSUSE and SLES.[20]

Features of R Programming

- Open-source. R is an open-source software environment. ...
- Strong Graphical Capabilities. ...
- Highly Active Community. ...
- A Wide Selection of Packages. ...
- Comprehensive Environment. ...
- Can Perform Complex Statistical Calculations. ...
- Distributed Computing. ...
- Running Code Without a Compiler.

Advantages:

- Interactive
- Easy Debugging
- Auto-complete feature
- Friendly environment for Package Development
- Codes and work can be save as a project (Much organized way)

Disadvantages:

- Sometime R Studio generate crisis in Viewing data.

It does not show all the fields (reason I establish that it’s an HTML based boundary so not all feature of R works here.)[19]

## V. CONCLUSION

Nowadays machine learning (ML) is a widely used technology for the analysis and prediction in various applications. This paper describes the classification of ML such as supervised, unsupervised and reinforcement learning. We also discuss the advantages, disadvantages of these learning techniques with their design issues and challenges. Machine learning is used in the various sectors such as health, transportation, government etc. which is discussed is this paper. In this paper we also discusses the different implementation tools of the machine learning such as R studio, Rapid Miner and WEKA tool with their features, advantages and disadvantages. In future work, need to develop the ensemble approach of machine learning for the effective analysis and prediction of the services.

## REFERENCES

[1] W. Richert, L. P. Coelho, "Building Machine Learning Systems with Python", Packt Publishing Ltd., ISBN 978-1-78216-140-0
[2] M. Welling, "A First Encounter with Machine Learning"
[3] M. Bowles, "Machine Learning in Python: Essential Techniques for Predictive Analytics", John Wiley & Sons Inc., ISBN: 978-1-118- 96174-2.
[4] Anna L. Buczak and Erhan Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection", IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 18, NO. 2, SECOND QUARTER 2016
[5] R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, 1986.
[6] R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[7] J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.
[8] Soubhik Das and Manisha J. Nene, "A Survey on Types of Machine Learning Techniques in Intrusion Prevention Systems", IEEE WiSPNET 2017 conference.
[9] Vrushali Y Kulkarni, Pradeep K Sinha, "Effective Learning and Classification using Random Forest Algorithm", International Journal of Engineering and Innovative Technology (IJEIT) Volume 3, Issue 11, May 2014.
[10] D. Berrar, An empirical evaluation of ranking measures with respect to robustness to noise, Journal of Artificial Intelligence Research 49 (2014) 241–267.
[11] Daniel Berrar, "Bayes' Theorem and Naive Bayes Classifier", Encyclopedia of Bioinformatics and Computational Biology, Volume 1, Elsevier, pp. 403-412.
[12] Ali, H. H., &Kadhum, L. E. (2017). "K- Means Clustering Algorithm Applications in Data Mining and Pattern Recognition". International Journal of Science and Research (IJSR), Vol.6 Issue.8, pp.1577–1584. https://doi.org/10.21275/ART20176024 9.
[13] Singh,S., & Gill, N.S. (2013). "Analysis and Study of K-Means Clustering Algorithm". International Journal of Engineering Research & Technology. Vol.2 Issu.7, pp. 2546-2551.
[14] Shaik, I., Hiwarkar, T., &Nalla, S. (2019). "Customer Segmentation Analysis of E- Commerce Big Data". International Journal of Engineering and Advanced Technology. Vol.8 Issue.5, pp. 582–587.
[15] Peter Wei , A Study of Principal Component Analysis on Classifiers Using Histogram of Gradients Features, final report available at: http://www.contrib.andrew.cmu.edu/~pwei/papers/FinalReport. pdf. [16] Witten, I. H., and E. Frank. 1999., "Data Mining: Practical Machine Learning tools and techniques with Java implementations" , Morgan Kaufman.
[16] https://hub.packtpub.com/clustering-and-other-unsupervised-learning-methods/
[17] https://www.outsource2india.com/software/articles/machine-learning-applications-how-it-works-who-uses-it.asp.
[18]-https://machinelearningmastery.com/machine-learning-tools/
[19] -https://www.quora.com/What-are-the-pros-and-cons-of-RStudio
[20]-"RStudio". rstudio.com. Retrieved 2 December 2016.
[21] VaseemNaiyer, Jitendra Sheetlani, Harsh Pratap Singh, "Software Quality Prediction Using Machine Learning Application", Smart Intelligent Computing and Applications, Springer, 2020.pp 319-327.
[22] Sheeraz Ahmad Peerzada, Jitendra Seethalani, "Machine Learning and Its Implications on Educational Data Base (U-DISE)", Smart Intelligent Computing and Applications, 2020, Springer.
[23] Sheeraz Ahmad Peerzada, NeelamadhabPadhy, Jitendra Sheetlani, Gh Hassan, "Predict the Performance of Students and School on Educational Database (U-DISE)", International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020. pp 1-6.