A Review : Machine Learning Algorithms

Aditya Vishwakarma Shiv Kumar Tiwari Pinkal Jain

M.tech, SRIT,Jabalpur ASST. Prof. GGCT, Jbp. ASST. Prof. GGCT, Jbp.

[Aadityavish2186@gmail.com](mailto:Aadityavish2186@gmail.com) [sshiv.tiwari22@gmail.com](mailto:sshiv.tiwari22@gmail.com) [pinku029jain@gmail.com](mailto:pinku029jain@gmail.com)

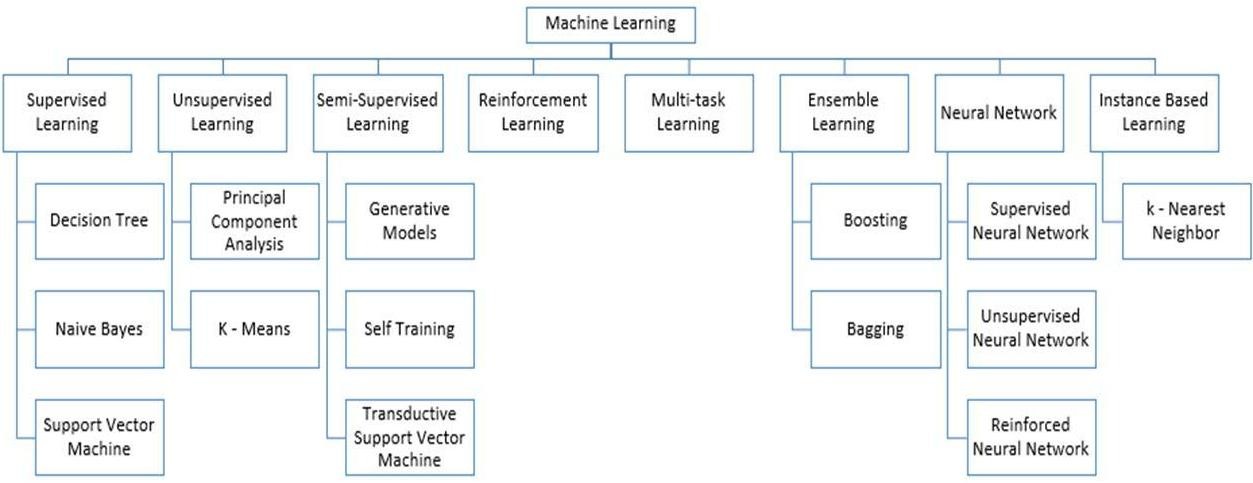
**Abstract:** In this article, the various machine learning techniques are covered. These techniques are used in several professions, including data mining, image processing, predictive analytics, etc. The major benefit of machine learning is that once an algorithm has learnt how to use data, it can work autonomously***.***

**Keywords:** Machine learning, algorithms,SVM.

# Introduction

Machine learning is used to teach machines how to handle data more effectively. Even after seeing the data, there are times when we are unable to see a pattern or make inferences from it. In this instance, machine learning is used [1]. The need for machine learning has grown as a result of the size of datasets. To obtain pertinent data, machine learning is used in numerous industries, including the military and healthcare. Machine learning aims to learn from the data. Trying to educate robots to learn on their own has been the topic of various studies [2] [3]. Many mathematicians and programmers employ a range of methods to address this issue. Several of them are demonstrated in Fig. 1.Section 2 details each machine learning technique.

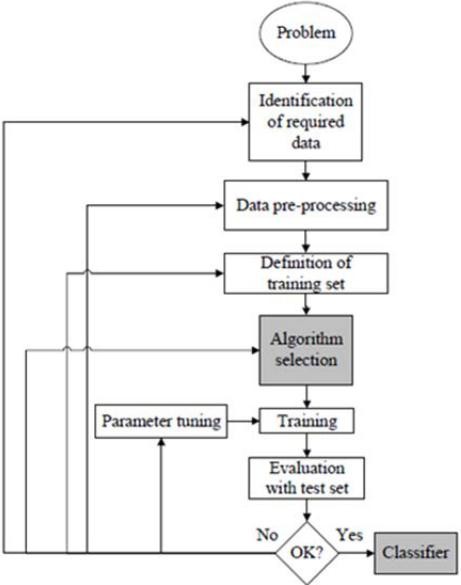
**2.Learning Methods**

Trees that aggregate qualities by sorting them according to their values are known as decision trees. Decision trees are mostly used for classification. In every tree, there are nodes and branches. Each node in a graph represents an attribute.

**Figure 1:** Types of Learning [2] [3]

## Supervised Machine Learning

Supervised Instruction supervised machine learning refers to algorithms that need assistance from outside sources. There are train and test datasets for the dataset. The output variable from the train dataset has to be predicted or classified. To predict or classify the test dataset, every algorithm employs a pattern formed from the training dataset [4].The procedure for supervised machine learning algorithms is shown in Fig. 2. Three of the most well-liked supervised machine learning techniques are examined in this article. As the name implies, supervised machine learning is based on supervision. The "labelled" dataset is used to train the machines using the supervised learning method, and after training, the computer predicts the output. Which inputs have previously been mapped to which outputs are shown here by the labeled data. To express it more accurately, we may say that we ask the machine to anticipate the results using test datasets after training it with input and associated output. Let's explain supervised learning with an example. Imagine that our input is a dataset of images of cats and dogs. Thus, we will first program the computer to understand the images, teaching it things like the size and form of a dog's tail and the color, shape, and height of a cat's eyes (dogs are taller than cats, for example). After training, we input a picture of a cat and ask the computer to identify it and predict the outcome. Now that the machine has acquired knowledge, it will evaluate every aspect of the object, including height, form, color, eyes, ears, tail, and so on, and conclude that it is a cat. It will thus be categorized as a cat. This is the method the computer uses to recognize the objects in supervised learning.



**Figure 2:** Workflow of supervised machine learning algorithm [4]

Making a map between the input variable (x) and the output variable is the primary objective of the supervised learning approach (y). Real-world applications of supervised learning include risk analysis, fraud detection, and spam filtering.



**Figure 3:** Decision Tree [5]

### 1.Supervised Machine Learning Categories

### There are two sorts of supervised machine learning issues, which are outlined below:

### Classification

### Regression

1. **Classification**

### Classification techniques are used to handle problems when the output variable is categorical, such as "Yes" or "No," Male or Female, Red or Blue, etc. The classification algorithms forecast the categories that are present in the dataset. Currently used categorization systems include spam detection, email filtering, and others. The following list of well-liked classification algorithms includes:****Decision trees,****

### ****Random Forest algorithm****

### ****Logistic Regression algorithm****

1. **Regression**

Regression methods are used to handle regression issues where the input and output variables have a linear relationship. They are used to predict variables with constant results, including market trends and weather predictions, etc.

The popular regression algorithms in the list below include:

* Multivariate Regression Method
* Decision Tree Algorithm
* Lashing Regression
* Basic Linear Regression Method

**The benefits and drawbacks of supervised learning:**

These techniques are excellent for predicting the outcome based on prior performance since supervised learning employs a labeled dataset, allowing us to properly identify the object classes.

**Disadvantages:**

### These algorithms can't handle challenging problems.

### • It could predict the wrong result if the test data differ from the training data.

### • The algorithm training takes a lot of computer time.

### Applications of Supervised Learning

These are a few typical uses for supervised learning.:

## ****Image Segmentation:****

## **Algorithms based on supervised learning are utilized for image segmentation. Using pre-established labels, this method classifies images from a range of picture data.**

## ****The medical diagnosis****

## **Supervised algorithms are commonly used for diagnostic purposes in the medical profession. It is carried out using historical data that includes labels for medical images and descriptions of illnesses. Such a process might be used by the machine to diagnose a disease in new patients.**

## ****Fraud Detection****

## **The application of supervised learning algorithms allows for the detection of fraudulent customers, transactions, etc. Historical data is utilized to identify trends that might indicate probable fraud.**

## ****Spam detection** - Spam detection and filtering use classification techniques. These formulas decide if an email is spam or not.**

## 2. Unsupervised Machine Learning

### Unsupervised learning is different from the supervised learning strategy since it doesn't call for monitoring. Unsupervised machine learning is defined as the process by which a computer program makes output predictions without any human intervention after training on an unlabeled dataset.

### After being trained on data, unsupervised learning enables the models to work freely on data that hasn't been labeled or classified.

### The primary goal of the unsupervised learning approach is to classify or organize the unsorted dataset into groups based on patterns, similarities, and differences. In the input dataset, the machines are tasked with uncovering hidden patterns.

**Let's give an example to help you better understand it. Assume we give images of a fruit basket to the machine learning model. The model's task is to spot patterns and collections of objects without having any prior information of the images.**

**As a consequence, the machine will now understand its patterns and distinctions, such as color differences and shape differences, and predict the outcome when tested with the test dataset.**

**Unsupervised Machine Learning Categories**

### The following two categories can be used to further categorize unsupervised learning:

### • Clustering\s

### • Association

### 1) Clustering

### We use the clustering technique to search for the intrinsic groups in the data. It is a technique for grouping the objects so that those that are most similar to one another stay together and have little to no resemblance to those in other groups. The clustering approach is demonstrated by categorizing customers according to their purchasing patterns. The following list of well-liked clustering algorithms includes

### Algorithm for K-Means Clustering

### Mean-shifting algorithm

### Principal Component Analysis and the DBSCAN algorithm

### Unified Component Analysis

### 2) Association

With a large dataset, an unsupervised learning technique called association rule learning uncovers fascinating connections between variables. The main objective of this learning method is to find the connections between data points and then map the variables to maximize profit. Continuous manufacturing, web usage mining, market basket research, etc. are some applications of this technology.

Other well-known algorithms for learning association rules include Apriori, Eclat, and FP-growth.

**The benefits and drawbacks of using an unsupervised learning algorithm**

**Advantages:**

* As they operate on unlabeled datasets, these algorithms, as opposed to supervised ones, can be applied to more difficult issues.
* For many applications, unsupervised methods are chosen since obtaining the unlabeled dataset is easier than obtaining the labeled dataset.

**Disadvantages:**

## The output of an unsupervised algorithm may be less precise since the dataset is not labeled and the algorithms are not trained using the exact result in advance.

## Why Since unsupervised learning employs a dataset that is unlabeled and does not correspond to the output, working with it is more difficult.

## Uses for Unsupervised Learning

## Network System: Unsupervised learning is used to identify plagiarism and copyright violations in text data for scholarly papers using document network analysis.

## Recommendation Systems: Using unsupervised learning approaches, recommendation systems typically create suggestion applications for a variety of online apps and e-commerce websites.

## Anomaly Detection: A popular use of unsupervised learning, anomaly detection identifies unusual data points within a collection. It is used to track down fraudulent transactions.

## Singular Value Decomposition (often referred to as SVD): This method is employed to extract certain data from databases. capturing data about each user who is present in a certain location, for example

## 3.Semi-Supervised Learning:

### **Semi-supervised learning is a type of machine learning technique that sits in between supervised and unsupervised learning. It falls between supervised learning (with labeled training data) and unsupervised learning (without labeled training data) approaches and uses a combination of labeled and unlabeled datasets during the training phase.**

### **Although semi-supervised learning is a middle ground between supervised and unsupervised learning, the majority of the data it utilizes is unlabeled. It operates on data that contains a few labels. Although labels are pricey, there might not be many labels needed for corporate needs. It is completely different from supervised and unsupervised learning, which depend on the presence or absence of labels.**

### **To overcome the drawbacks of supervised learning and unsupervised learning techniques, the concept of semi-supervised learning is put forth. The main objective of semi-supervised learning is to utilize all available data to the fullest extent possible rather than relying primarily on labelled data as in supervised learning. Comparable data are initially clustered using an unsupervised learning approach, which also helps classify the unlabeled data into labelled data. This is due to the fact that labeled data costs more to purchase than unlabeled data.**

### **Through an example, we may visualise these algorithms. While a student is engaged in supervised learning, both at home and in school, an instructor is watching over them. Also, unsupervised learning is when a pupil analyses a subject on their own without assistance from the teacher.**

### Advantages and disadvantages of Semi-supervised Learning

## ****The method is straightforward and simple to grasp, yet it is also quite effective.****

## **It is used to address issues with algorithms for supervised and unsupervised learning.**

## **The outcomes of iterations may not be stable.**

## **We are unable to use these techniques on data at the network level.**

## **Low accuracy**.****

## Reinforcement Learning

## **A software component that using reinforcement learning strikes and trails, acts, learns from experiences, and improves performance to autonomously explore its environment. The basis for reinforcement learning is feedback. As a reinforcement learning agent is rewarded for every positive behaviour and punished for every negative one, its goal is to maximize rewards. Reinforcement learning differs from supervised learning in that it is solely based on the experiences of the agents.**

## **The process of reinforcement learning is similar to how humans learn; for instance, a young child picks up new information through experiences in his everyday life. Playing a game in which the environment acts as the game's environment and an agent's activities determine states at each step is an example of reinforcement.**

## Categories of Reinforcement Learning

## There are primarily two categories of approaches or algorithms used in reinforcement learning:

### Positive Reinforcement Learning: Positive reinforcement learning refers to the process of adding something to the needed behaviour to make it more likely that it will happen again. It strengthens the agent's behaviour and has a favourable effect on it.

### Negative Reinforcement Learning: This method of learning operates in direct opposition to positive RL. By avoiding the undesirable circumstance, it makes it more likely that the particular behaviour would recur.

### Real-world Use cases of Reinforcement Learning

### ****• Video games:** Applications for games frequently utilize real-time learning techniques. It's used to achieve superhuman performance. Examples of well-known RL algorithms are the video games AlphaGO and AlphaGO Zero**.****

### ****• Resource Management:** The "Resource Management with Deep Reinforcement Learning" study showed how to use RL in computers to automatically train and organize resources to wait for different workloads in order to decrease average job latency.**

### ****• Robotics:** The use of RL in robotics applications is common. Robots are utilized in the manufacturing and industrial sectors, and reinforcement learning is used to boost their performance. Several industries share the objective of creating intelligent robots using AI and machine learning technology.**

### ****•** Text mining is one of the main applications of NLP.**

### Advantages and Disadvantages of Reinforcement Learning

**Advantages**

• Because the RL learning model and human learning are comparable, the most accurate results may be attained. • It assists in the resolution of complex real-world challenges that are difficult to handle using traditional approaches.

• Assists in achieving lasting results.

# ****Disadvantage****

# **RL algorithms are not suited for simple jobs.**

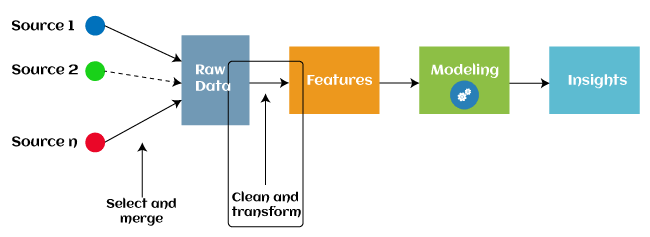
# **RL algorithms require vast quantities of computing and data.**

# **An abundance of states resulting from excessive reinforcement learning may lead to subpar results.**.

# Feature Engineering for Machine Learning

**To create features from raw data that may be utilized to create a prediction model using either machine learning or statistical modelling, feature engineering, a pre-processing stage in machine learning, is performed. Engineering machine learning features aims to improve model performance. In this topic, we'll learn a lot about feature engineering in machine learning. Hence, before diving into the details, let's first identify features. Why is feature engineering required, then?**

The pre-processing phase of machine learning, called feature engineering, extracts features from raw data. It helps to clarify a key problem for predictive models, improving the model's accuracy for unseen data. The model, which consists of predictor variables and an outcome variable, is built using the feature engineering approach, which selects the most useful predictor variables.



Automated feature engineering has been used by various machine learning programs since 2016 to assist in automatically extracting features from raw data. The majority of feature engineering in machine learning is composed of four operations: feature generation, feature transformations, feature extraction, and feature selection. As follows is a description of these procedures:

1. **Feature creation** :The process of creating features for a predictive model is called feature discovery. The process is arbitrary and dependent on human inventiveness and interference. The new features are quite versatile because they were built using addition, subtraction, and ration operations..
2. **Transformations:** To improve the model's accuracy and efficacy, the feature engineering transformation stage comprises changing the predictor variable. It makes the model easier to understand by, for example, ensuring that all the variables are on the same scale and that the model is adaptable enough to receive input from a variety of sources.
3. **Feature Extraction:** Feature extraction is a feature engineering technique that transforms raw data into new variables by automatically deleting the old ones. The main objective of this stage is to reduce the amount of data in order to make it easier to use and manage for data modeling. Examples of feature extraction methods include cluster analysis, text analytics, edge detection algorithms, and principal components analysis (PCA).
4. **Feature Selection:** The other variables in the dataset are either redundant or worthless; only a limited subset of them may be utilized to build a machine learning model. If all of these redundant and meaningless pieces of information are included in the dataset, the overall performance and accuracy of the model may deteriorate.

## It is essential to identify and pick the most relevant features from the data in order to remove the superfluous or less important information, which is performed with the help of feature selection in machine learning. Feature selection is a method for selecting the subset of the most important features by removing the redundant, unnecessary, or noisy qualities from the initial feature set. The advantages of feature selection in machine learning are listed below:

## • It helps the model be simple so that researchers can rapidly understand it; • It helps avoid the dimensionality plague.

## • It reduces training time and enhances generalization by minimizing overfitting.

## Types of ML Classification Algorithms:

## The following two categories can be used to further split classification algorithms:

## model linear

## logarithmic regression

## Support Vector Machines

## Non-Linear Models

## K-Nearest Neighbours

## Kernel SVM

## Naive Bayes

## Decision Tree Classification

## Random Forest Classification

## are some examples of machine learning techniques.

## Evaluating a Classification model:

When our model is finished, we must assess its performance to determine if it is a classification model or a regression model. Thus, we have the following options for assessing a classification model:

**1. Cross-Entropy Loss or Log Loss**

* **It is utilized to evaluate a classifier's performance, with the resulting probability value falling between**
* **A good binary classification model should have a log loss value that is near to 0.**
* **The value of log loss increases if the predicted value and the actual value are different.**
* **The accuracy of the model increases with decreasing log loss.**
* **How can the cross-entropy for binary classification be calculated?**
* **(ylog(p)+(1?y)log(1?p))**
* **Where y = actual output and p = predicted output**

**Confusion Matrix:**

**The confusion matrix outputs a matrix or table that summarizes the model's performance.**

**Another term for it is the error matrix.**

**• The outcomes of the forecasts are summarized in the matrix, along with the overall number of accurate and incorrect predictions.**

**Accuracy = (TP+TN)/Total Population**

* + 1. **AUC-ROC curve:**

**AUC stands for the area under the curve, while ROC stands for the receiver operating characteristics curve.**

* **It is a graph that displays the classification model's performance at various thresholds.**
* **The AUC-ROC Curve is used to show how well the multi-class classification model is doing.**
* **The TPR and FPR are used to draw the ROC curve, with the True Positive Rate (TPR) on the Y-axis and the FPR (False Positive Rate) on the X-axis.**

**Use cases of Classification Algorithms**

# Several situations call for the usage of classification methods. These are a few frequent applications for classification algorithms:

# Voice Recognition o Spam Email Detection

# Tumour cell identifications in cancer.

# Classification of drugs, biometric identification, etc

# Logistic Regression in Machine Learning

* • Logistic regression is one of the most well-known Machine Learning algorithms used with the Supervised Learning approach. It is used to forecast the categorical dependent variable using a specified set of independent variables.
* The outcomes of a dependent variable with a categorical component are predicted by logistic regression. The outcome must thus be a discrete or categorical value. It offers the probabilistic values that lie between 0 and 1 rather than the exact values between 0 and 1. It can be either True or False, 0 or 1, or Yes or No.

# The method of application is the primary distinction between logistic regression and linear regression. Regression problems are resolved using logistic regression, whereas classification problems are resolved using linear regression.

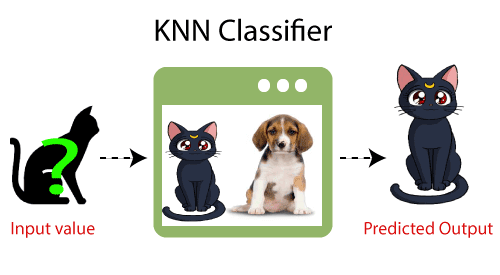
# In logistic regression, rather than fitting a regression line, we fit a "S" shaped logistic function that anticipates two maximum values (0 or 1).

# Depending on a mouse's weight, the logistic function's curve can indicate a number of possibilities, including whether or not the cells are cancerous and whether or not a mouse is obese.

# Logistic regression is a crucial machine learning technique because it can categorize new data using both continuous and discrete datasets.

# Observations may be categorized using a number of data formats using logistic regression, and it can quickly pinpoint the variables.

# K-Nearest Neighbor(KNN) Algorithm for Machine Learning

* • Machine Learning K-Nearest Neighbor (KNN) Algorithm One of the simplest machine learning algorithms, based on the supervised learning method, is K-Nearest Neighbor.
* • Assuming that the new case/data and the previous cases are comparable, the K-NN method places the new example in the category that is most similar to the other categories that are available.
* • The K-NN algorithm categorizes new data based on similarity and preserves all of the existing data. This indicates that new data may be reliably and rapidly categorized using the K-NN approach.
* Although the K-NN technique may be used to classification and regression problems, classification problems are where it is most often utilized.
* K-NN, also known as a lazy learner algorithm, is a non-parametric approach that makes no assumptions about the underlying data. It stores the training dataset rather than learning from it right away. Instead, it performs an action while categorizing data by using the dataset. During the training phase, the KNN algorithm simply saves the dataset and then classifies new data into a category that is quite similar to the new data.
* When we see a creature that resembles both a dog and a cat but whose identification we are dubious about. The KNN approach may be used for this identification, though, as it is based on a similarity measure. Our KNN model will determine the commonalities in the

**What's the working of K-NN?**

**•** The K-NN operates according to the following algorithm:

Choosing the Kth neighbor's number is the first step.

In step two, determine the Euclidean distance between K neighbors.

• Step 3: Choose the K closest neighbors based on the calculated Euclidean distance.

Step 4: Among these k neighbors, count the number of data points in each category.

Step 5: Assign the fresh data points to the category with the highest neighbor count.

• Step 6: The model is finished.

**Conclusion**

This study examines a variety of machine learning techniques. Currently, machine learning is used by everyone, whether they are aware of it or not. updating your profile image on social networking sites to receiving product recommendations when you purchase online. The vast majority of the well-known machine learning techniques are introduced in this article.

# 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. S.B. Kotsiantis, “*Supervised Machine Learning: A Review of Classification Techniques*”, Informatica 31 (2007) 249-268
5. L. Rokach, O. Maimon, “*Top – Down Induction of Decision Trees Classifiers – A Survey*”, IEEE Transactions on Systems, D. Lowd, P. Domingos, “Naïve Bayes Models for Probability Estimation” https://webdocs.cs.ualberta.ca/~greiner/C- 651/Homework2\_Fall2008.html
6. D. Meyer, “Support Vector Machines – The Interface to libsvm in package e1071”, August 2015
7. S. S. Shwartz, Y. Singer, N. Srebro, “*Pegasos: Primal Estimated sub - Gradient Solver for SVM*”, Proceedings of the 24th International Conference on Machine Learning, Corvallis, OR, 2007.

<http://www.simplilearn.com/what-is-machine-learning-> and-why-it-matters-article

1. P. Harrington, “*Machine Learning in action*”, Manning Publications Co., Shelter Island, New York, 2012 <http://pypr.sourceforge.net/kmeans.html>
2. K. Alsabati, S. Ranaka, V. Singh, “*An efficient k- means clustering algorithm*”, Electrical Engineering and Computer Science, 1997
3. M. Andrecut, “*Parallel GPU Implementation of Iterative PCA Algorithms*”, Institute of Biocomplexity and Informatics, University of Calgary, Canada, 2008
4. 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
5. X. Zhu, “*Semi-Supervised Learning Literature Survey*”, Computer Sciences, University of Wisconsin-

Madison, No. 1530, 2005

1. R. S. Sutton, “*Introduction: The Challenge of Reinforcement Learning*”, Machine Learning, 8, Page 225-227, Kluwer Academic Publishers, Boston, 1992
2. L. P. Kaelbing, M. L. Littman, A. W. Moore, “*Reinforcement Learning: A Survey*”, Journal of Artificial Intelligence Research, 4, Page 237-285, 1996
3. R. Caruana, “*Multitask Learning*”, Machine Learning, 28, 41-75, Kluwer Academic Publishers, 1997
4. D. Opitz, R. Maclin, “*Popular Ensemble Methods: An Empirical Study*”, Journal of Artificial Intelligence Research, 11, Pages 169-198, 1999
5. Z. H. Zhou, “*Ensemble Learning*”, National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China https://en.wikipedia.org/wiki/Boosting\_(machine\_learn ing) https://en.wikipedia.org/wiki/Bootstrap\_aggregating
6. V. Sharma, S. Rai, A. Dev, “*A Comprehensive Study of Artificial Neural Networks*”, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN 2277128X, Volume 2, Issue 10,

October 2012

1. S. B. Hiregoudar, K. Manjunath, K. S. Patil, “*A Survey: Research Summary on Neural Networks*”, International Journal of Research in Engineering and Technology, ISSN: 2319 1163, Volume 03, Special Issue 03, pages 385-389, May, 2014 https://en.wikipedia.org/wiki/Instance-based\_learning
2. P. Harrington,“ Machine Learning in Action”, Manning Publications Co., Shelter Island, New York,

ISBN 9781617290183, 2012

1. 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.

[24] Tiwari, S.K., Neogi, S.G., Mishra, A. (2023). Design and Implementation of Security Enhancement for Trusted Cloud Computing. In: Kumar, A., Mozar, S., Haase, J. (eds) Advances in Cognitive Science and Communications. ICCCE 2022. Cognitive Science and Technology. Springer, Singapore.

[25 Tiwari,S.K. Sao,N., Kushwaha, A.,Book Title “[Fundamental of Machine Learning](https://scholar.google.com/citations?view_op=view_citation&hl=en&user=gpoUlO4AAAAJ&citation_for_view=gpoUlO4AAAAJ:eQOLeE2rZwMC)”,lambert Publication, ISBN: 9789391265335, Oct. 2022,