A Review : Machine Learning Algorithms

Aditya Vishwakarma Shiv Kumar Tiwari Pinkal Jain

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

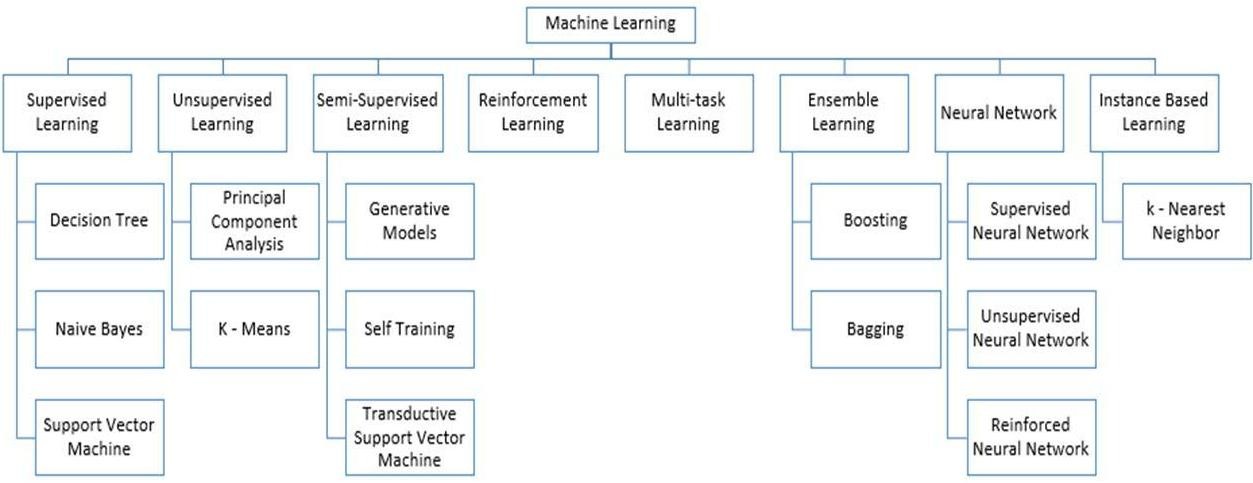
**Abstract: *In this study, a variety of machine learning approaches have been addressed. Numerous jobs, such as data mining, image processing, predictive analytics, etc., use these methods. The main advantage of machine learning is the ability of an algorithm to function independently once it has learned the utilization of data.***

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

# Introduction

To educate machines how to handle data more efficiently, machine learning is utilized. There are instances when we are unable to detect a pattern or draw conclusions from the data, even after viewing it. We use machine learning in that situation [1]. The large number of datasets has increased demand for machine learning. Machine learning is utilized in many sectors, including the military and healthcare, to gather relevant data. Learning from data is the goal of machine learning. Various studies have attempted to teach robots how to learn on their own [2]. [3]. Numerous mathematicians and programmers use a variety of techniques to deal with this problem. Fig. 1 shows several of them in action. Section 2 describes each machine learning technique.

1. **Learning Methods**

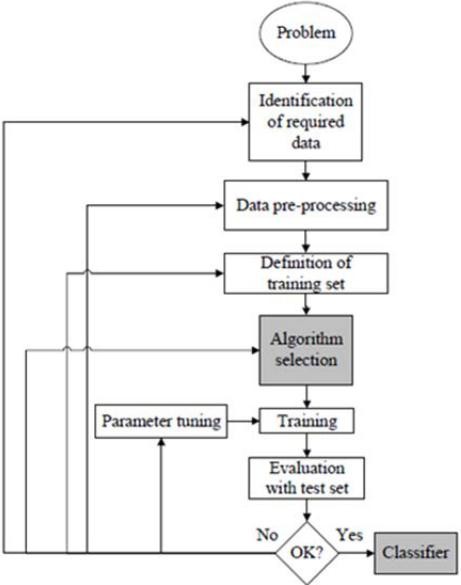
Trees that aggregate qualities by sorting them according to their values are known as decision trees. Decision trees are mostly used for classification. In every tree, there are nodes and branches. Each node in a graph represents an attribute.

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

## Supervised Machine Learning

Supervised machine learning algorithms are those that need external assistance. For the dataset, there are train and test datasets. The output variable from the train dataset has to be predicted or categorized. To predict or categories the test dataset, every algorithm uses some form of pattern that is taken from the training dataset [4]. Fig. 2 depicts the process for supervised machine learning algorithms. This article examines three of the most popular supervised machine learning methods. As the name implies, supervised machine learning is based on supervision. The "labeled" dataset is used to train the machines using the supervised learning approach, and following training, the computer predicts the outcome. Which inputs have already been translated to which outputs can be seen in the provided data here. To put it more simply, we may say that we ask the machine to anticipate the results using test datasets after training it with input and related output. Let's explain supervised learning with an example. Consider that the dataset we are using as our input contains images of both cats and dogs. After training, we enter a picture of a cat and ask the computer to identify it and predict the result.

After gaining knowledge, the machine will analyze the object's height, form, color, eyes, ears, tail, and other features to determine that it is a cat.

This is the method the computer uses to recognize the items during supervised learning.

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

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



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

### 1.Supervised Machine Learning Categories

### Problems with supervised machine learning can be divided into two categories, which are listed below:

### Classification

### Regression

1. **Classification**

### Classification techniques are used to handle problems where 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 techniques 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 following list of popular regression algorithms includes:

* Decision Tree Algorithm
* Multivariate Regression Algorithm
* Simple Linear Regression Algorithm
* Lashing Regression

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

* Since supervised learning uses a labeled dataset, we can precisely identify the object classes.
* These algorithms are useful for anticipating the outcome based on past performance.

**Disadvantages:**

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

### If the test data differs from the training data, it could anticipate the incorrect outcome.

### Training the algorithm involves a lot of computing time.

### Applications of Supervised Learning

Some common applications of Supervised Learning are given below:

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

## **Algorithms based on Supervised Learning are utilized for image segmentation. With the help of pre-established labels, image classification is carried out in this process on various image data.**

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

## **Supervised algorithms are commonly used for diagnostic purposes in the medical field. 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 could 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 because it doesn't call for monitoring. Unsupervised machine learning is when a system is trained on an unlabeled dataset and then makes output predictions without any human intervention. Unsupervised learning allows the models to operate independently on data that hasn't been categorised or tagged after they've been trained on it.

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

**To make things easier for you to comprehend, let's use an illustration. Let's say we provide the machine learning model with photos of a fruit basket. The model's job is to recognise groups of items and patterns in an image without any prior knowledge of the images.**

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

**Unsupervised Machine Learning Categories**

### Unsupervised learning may be further divided into the following two categories:

### Clustering

### Association

### 1) Clustering

* We use the clustering approach to identify the logical groups in the data. It's a method for grouping objects so that those that are most similar to one another stay together and have little to no resemblance to objects in other groups. The clustering method is described, which groups clients according to their purchasing patterns. The following list of well-liked clustering techniques includes
* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### 2) Association

An unsupervised learning method known as association rule learning reveals remarkable relationships between variables in a large dataset. Finding the relationships between data points and then mapping the variables to maximum profit are the major goals of this learning strategy. Applications of this technique include continuous production, web usage mining, market basket research, etc.

Apriority, Éclat, and FP-growth are a few well-known algorithms for learning association rules.

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

**Advantages:**

* Since 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 because the dataset is not labelled and the algorithms are not trained using the exact output in advance.

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

## Applications of Unsupervised Learning

## Network System : Document network analysis of text data for scholarly articles uses unsupervised learning to detect plagiarism and copyright.

## Recommendation Systems: Recommendation systems frequently construct recommendation applications for various web applications and e-commerce websites using unsupervised learning techniques.

## Anomaly Detection: A common use of unsupervised learning, anomaly detection can find out-of-the-ordinary data points in the dataset. It is employed to find erroneous transactions.

## Singular Value Decomposition: This technique, often known as SVD, is used to extract specific data from databases. Taking information on each user who is present in a specific place, for instance.

## Semi-Supervised Learning:

### **Semi-supervised learning is a machine learning technique that is in the middle between supervised and unsupervised learning. It lies between supervised learning (with labelled training data) and unsupervised learning (without labelled training data) techniques and uses a combination of labelled and unlabeled datasets during the training phase.**

### **Semi-supervised learning uses mostly unlabeled data, which is a middle ground between supervised and unsupervised learning. It works with data that has a few labels in it. Although labels are expensive, businesses may only require a few. Compared to supervised and unsupervised learning, which depend on the presence or lack of labels, it is entirely different.**

### **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 utilise all available data as fully as is practical rather than relying primarily on tagged data as in supervised learning. Comparable data are first clustered using an unsupervised learning algorithm, which helps turn the unlabeled data into labelled data. This is due to the fact that tagged data costs more to purchase than unlabeled data.**

### **We can visualise these algorithms with an example. An instructor is keeping an eye on a student while they are participating in supervised learning, both at home and at school. Unsupervised learning is when a student conducts independent research on a subject without the teacher's guidance.**

### 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 element uses reinforcement learning to investigate its environment on its own. It takes action, makes mistakes and grows stronger with practise. The basis of reinforcement learning is feedback. The goal of a reinforcement learning agent is to maximise rewards because it is rewarded for every successful action and punished for every unsuccessful one. Reinforcement learning is entirely dependent on the experiences of the agents, in contrast to supervised learning. The process of reinforcement learning is similar to how humans learn; for instance, a young child picks up new information from experiences in his daily life. Reinforcement learning is exemplified by playing a game in which an agent's actions produce states at each step and the environment acts as the game's environment.**

## Categories of Reinforcement Learning

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

### Positive Reinforcement Learning: The practise of adding something to the required behaviour to increase the likelihood that it will occur again is known as positive reinforcement learning. It strengthens the conduct of the agent and has a positive impact on it.

### Negative Reinforcement Learning: This approach to learning runs completely counter to constructive RL. By avoiding the unwanted situation, it increases the likelihood that the specific conduct would occur again.

### 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.**

### ****Management of Resources:** In order to reduce the average job slowness, the "Resource Management with Deep Reinforcement Learning" research demonstrated how to apply RL in computers to automatically train and organize resources to wait for various workloads.**

* **Robotics : In robotics applications, RL is commonly used. Robots are utilized in the manufacturing and industrial sectors, and reinforcement learning is used to boost their performance. Many industries share the objective of creating intelligent robots using AI and machine learning technology.**
* **Data Mining: One of the great uses of NLP is text mining.**

### Advantages and Disadvantages of Reinforcement Learning

**Advantages**

• Due to the similarities between the RL learning model and human learning, the most accurate results may be produced.

• It assists in the resolution of complex real-world challenges that are difficult to handle using traditional approaches.

• Assists in achieving lasting results.

**Disadvantage**

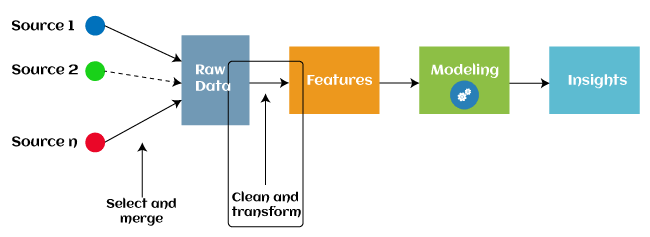
# Simple tasks are not appropriate for RL algorithms.

# RL algorithms demand enormous computer and data resources.

# An overabundance of states brought on by excessive reinforcement learning may undermine the outcomes.

# Feature Engineering for Machine Learning

**To extract features from raw data that may be used to create a prediction model using either machine learning or statistical modelling, feature engineering is a pre-processing phase in machine learning. The creation of machine learning components improves model performance. In this topic, we'll discover a lot about feature engineering in machine learning. So before we dive into the details, let's first define the qualities. Therefore, why is feature engineering required?**

**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.**

Since 2016, some machine learning programmes that aid in automatically extracting features from raw data have also adopted automated feature engineering. Four operations make up the majority of feature engineering in machine learning: feature creation, transformations, feature extraction, and feature selection. As follows is a description of these procedures:

1. Finding the most useful variables to use in a predictive model is referred to as feature generation. The process is arbitrary and dependent on human inventiveness and interference. They are incredibly versatile thanks to the addition, subtraction, and ration operations that were employed to build the additional features.

2.Transformations: During the feature engineering transformation step, the predictor variable is changed to increase the model's efficacy and accuracy. By ensuring that all the variables are on the same scale and that the model is flexible enough to receive input from a number of sources, for instance, it makes the model simpler to grasp.

3. Engineering approach known as "feature extraction" that substitutes new variables for the ones that are currently present. Feature Extraction: A feature engineering technique known as feature extraction automatically adds new variables by eliminating the ones that already exist from raw data. 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 modelling. Examples of feature extraction methods (PCA) include cluster analysis, text analytics, edge detection algorithms, and principal components analysis.

4.Feature Selection: Only a tiny portion of the dataset's variables can be used to create a machine learning model; the rest are either redundant or useless. The overall effectiveness and accuracy of the model may suffer if all of these pointless and redundant pieces of data are included to the dataset. It is essential to identify and pick the most appropriate 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 machine feature selection are listed below.

## 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 Bays

## 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 used to assess the effectiveness of a classifier, the result of which is a probability value between**
* **The value of log loss for a decent binary classification model should be close to 0.**
* **If the anticipated value differs from the actual value, the value of log loss rises.**
* **The lower the log loss, the more accurate the model is.**
* **Cross-entropy for binary classification may be computed as:?**
* **(log(p)+(1?y)log(1?p))**
* **Where p = projected output and y = actual output**
  + 1. **Confusion Matrix:**

**The confusion matrix describes the performance of the model and gives us a matrix or table as an output.**

* **The error matrix is another name for it.**
* **The matrix includes the results of the forecasts in a summary manner, including the total number of accurate and inaccurate 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

* One of the most well-known Machine Learning algorithms, under the Supervised Learning method, is logistic regression. With a predetermined set of independent factors, it is used to predict the categorical dependent variable.
* Logistic regression forecasts the results of a dependant variable with a categorical component. As a result, the result must be a discrete or categorical value. Rather of providing the precise values of 0 and 1, it provides the probabilistic values that fall between 0 and 1. It can be either Yes or No, 0 or 1, true or false, etc.
* The main difference between linear regression and logistic regression is how they are used. Whereas logistic regression is used to solve regression issues, linear regression is used to solve classification difficulties.
* In logistic regression, we fit a "S" shaped logistic function, which predicts two maximum values, rather than a regression line (0 or 1).
* The logistic function's curve shows the possibility of several things, like whether or not the cells are malignant, whether or not a mouse is fat depending on its weight, etc.
* Since it can classify fresh data using both continuous and discrete datasets, logistic regression is a key machine learning approach.
* Logistic regression may be used to categories observations using a variety of data formats and can quickly identify the factors that will work best for the classification.

# 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 Neighbors.

• 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.

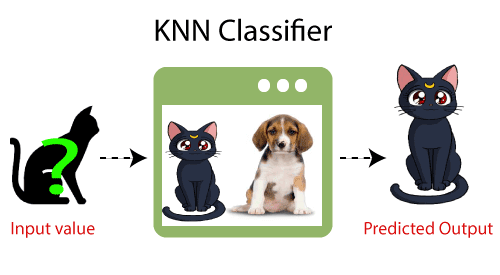
• The K-NN algorithm categorizes new data based on similarity and preserves all of the existing data. This indicates that new data can be reliably and quickly categorized using the K-NN approach.

• Although the K-NN technique can be applied to classification and regression problems, classification problems are where it is most frequently used.

• K-NN does not make any assumptions about the underlying data because it is a non-parametric approach.

• 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.

• If we see something that looks like both a dog and a cat, but we're not sure what it is. The KNN approach can be used for this identification, though, as it is based on a similarity measure. Our KNN model will locate the commonalities in the new data.



**How does K-NN function?**

* The K-NN operates according to the following algorithm:
* Choose the Kith neighbor's phone number in step one.
* In step two, get the Euclidean distance between K neighbors.
* Step 3: Select 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 group with the highest neighbor count.
* Step 6: The model is finished.

**Conclusion**

This study surveys a wide range of machine learning algorithms. Nowadays, everyone employs machine learning, whether consciously or unconsciously. By changing pictures on social networking sites to getting product suggestions when shopping online. This publication introduces the vast majority of the well-known machine learning techniques.

# 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,