A Review : Machine Learning Algorithms

Aditya Vishwakarma Shiv Kumar Tiwari Pinkal Jain

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

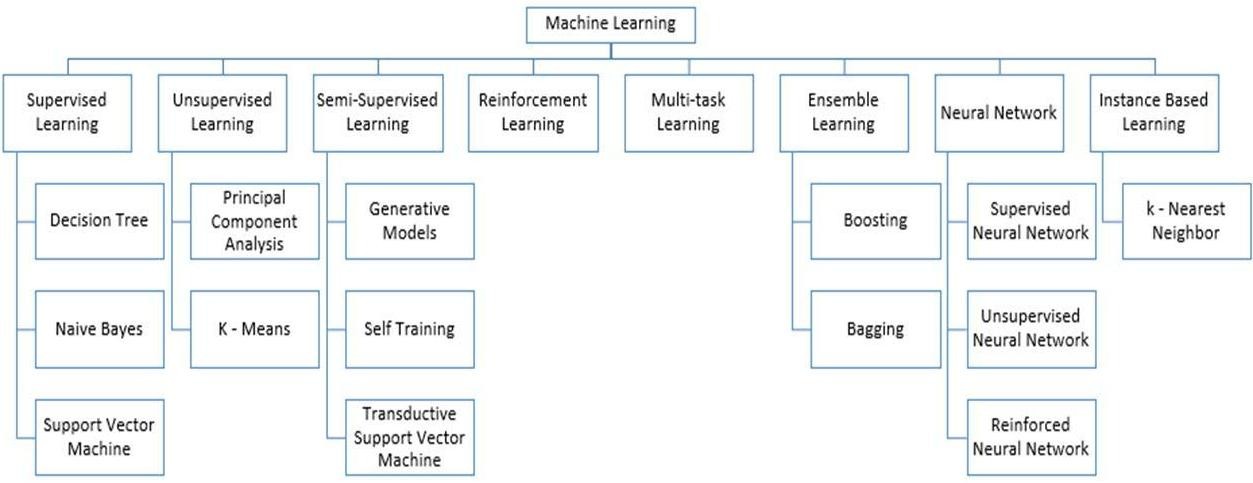
**Abstract: *In this paper, various machine learning methods have been described. Many jobs, such as data mining, image processing, predictive analytics, etc., utilise these algorithms. The main advantage of machine learning is the ability for an algorithm to work independently once it has mastered the utilisation of data.***

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

# Introduction

Machine learning is used to teach machines how to handle data more effectively. Even after viewing the data, there are times when we are unable to spot a pattern or make inferences from it. In that case, we employ machine learning [1]. The demand for machine learning has expanded as a result of the abundance 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.

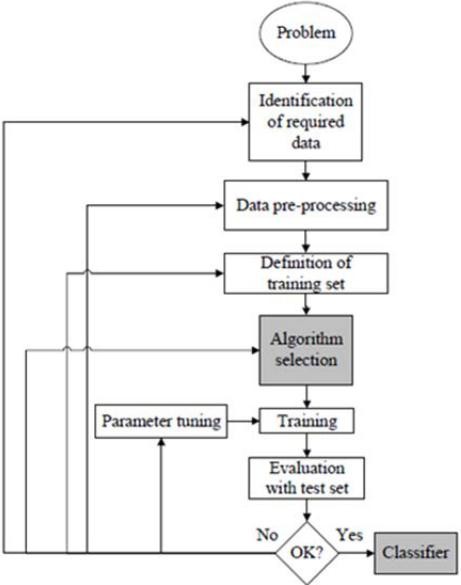
1. **Learning Methods**

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 Instruction supervised machine learning refers to algorithms that need assistance from outside sources. The dataset is divided into train and test datasets. The train dataset's output variable has to be predicted or categorized. Every algorithm uses some sort of pattern that is extracted from the training dataset to predict or categories the test 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 already been mapped to which outputs are shown here by the indicated data. To put it more precisely, 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. Assume that our input is a dataset of images of cats and dogs. As a result, we will first teach the computer to understand the images, teaching it things like the size and shape of a dog's tail, the shape, colour, and height of a cat's eyes (cats are shorter than dogs, 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, colour, eyes, ears, tail, and so forth, and conclude that it is a cat.

This is the process the computer employs in supervised learning to identify the items..

**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 labelled 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 utilised 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

Because it does not require supervision, unsupervised learning differs from the supervised learning approach. This is the definition of unsupervised machine learning, when the system makes output predictions without any human oversight after being trained on an unlabeled dataset.In unsupervised learning, the models are trained on data that has neither been categorised nor labelled, and they are then allowed to behave autonomously on that data.

### **The main objective of the unsupervised learning method is to categorise or group the unsorted dataset according to similarities, differences, and patterns. The machines are to find the hidden patterns in the input dataset.**

### **Let's use an example to help you better understand it. Suppose 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 result, the machine will now learn its patterns and distinctions, such as colour differences and form differences, and anticipate the outcome when tested with the test dataset.**

**Unsupervised Machine Learning Categories**

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

### Clustering

### Association

### 1) Clustering

To find the natural groupings in the data, we employ the clustering approach. It is a strategy for organising the items into groups so that those that are most similar to one another remain together and have little to no resemblance to those in other groups. By grouping clients based on their buying habits, the clustering method is presented. The popular clustering methods listed below include

* **K-Means Clustering algorithm**
* **Mean-shift algorithm**
* **DBSCAN Algorithm**
* **Principal Component Analysis**
* **Independent Component Analysis**

### 2) Association

In 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 maximise profit. Continuous production, web usage mining, market basket analysis, etc. are some applications of this technology.

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

**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 labelled 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:

**A machine learning method that falls between supervised and unsupervised learning is called semi-supervised learning. It uses a mixture of labelled and unlabeled datasets during the training phase and falls between supervised learning (with labelled training data) and unsupervised learning (without labelled training data) techniques.**

### **The majority of the data used in semi-supervised learning is unlabeled, despite being a halfway ground between supervised and unsupervised learning. It works with data that has a few labels in it. Even though labels are expensive, businesses may only require a small number of labels. Compared to supervised and unsupervised learning, which depend on the presence or lack of labels, it is entirely different.**

### **The idea of semi-supervised learning is proposed to address the shortcomings of supervised learning and unsupervised learning methods. Instead than depending mostly on labelled data like in supervised learning, the major goal of semi-supervised learning is to employ all available data as fully as feasible. Utilising an unsupervised learning technique, comparable data are initially clustered, aiding in the conversion of the unlabeled data into labelled data. This is because tagged data is more expensive to buy 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

## **Using reinforcement learning, a software component explores its surroundings on its own. It acts, learns from its mistakes, and becomes better with practise. Feedback is the cornerstone of reinforcement learning. A reinforcement learning agent's objective is to maximise rewards since it gets rewarded for every positive action and penalised for every bad one. In contrast to supervised learning, reinforcement learning is purely reliant on the experiences of the agents.The method of reinforcement learning is comparable to that of a human person; for instance, a youngster learns different things via encounters in his daily life. Playing a game in which an agent's actions establish states at each step and the environment serves as the game's environment is an example of reinforcement learning.**

## 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 utilise 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 organise resources to wait for various workloads.**

* **Robotics : In robotics applications, RL is commonly used. Robots are utilised 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**

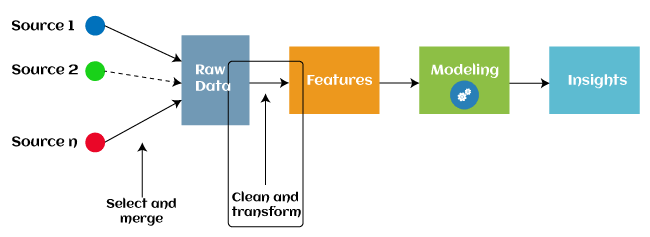
# Simple tasks are not appropriate for RL algorithms.

# 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

**Feature engineering, a pre-processing step in machine learning, is used to extract features from raw data that may be utilised to build a prediction model using either machine learning or statistical modelling. To enhance model performance, machine learning elements are created. We'll learn a lot about feature engineering in machine learning in this topic. So let's first identify the traits before getting into the specifics. So why is feature engineering necessary?**

The feature engineering stage of machine learning involves taking the raw data and extracting its features. It aids in the clarification of a significant issue for predictive models, increasing the model's accuracy for unobserved data. The feature engineering technique is used to build the model, which comprises of the most practical predictor variables and an outcome variable.

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. These processes are described as below:

1. **Feature creation:** Creating features is the process of selecting the most useful variables to include in a prediction model. 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:** The feature engineering transformation stage involves modifying the predictor variable 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. **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 modelling. Examples of feature extraction methods (PCA) include cluster analysis, text analytics, edge detection algorithms, and principal components analysis.
4. **Feature Selection:** The remaining variables in the dataset are either redundant or worthless; only a limited subset of them can be used 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.

Machine learning's feature selection technique is used to locate and select the most pertinent features from the data in order to eliminate irrelevant or unnecessary information. By eliminating the redundant, superfluous, or noisy characteristics from the initial feature set, feature selection is a technique for choosing the subset of the most crucial features. The following is a list of the benefits of feature selection in machine learning:

It aids in avoiding the dimensionality curse.

* It aids in the model's simplicity so that researchers may quickly comprehend it.
* It cuts down on training time.
* It improves generalisation by reducing over fitting.

## 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 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:?**
* **(ylog(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

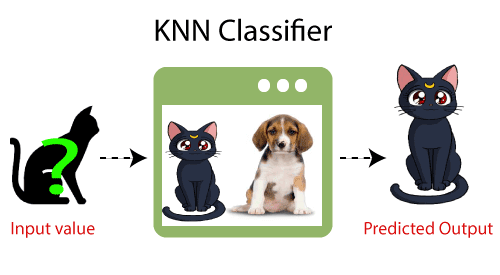
* One of the most well-known Machine Learning algorithms, under the Supervised Learning method, is logistic regression. With a predetermined set of independent factors, it is used to predict the categorical dependent variable.
* Logistic regression forecasts the results of a dependant variable with a 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

* K-Nearest Neighbor(KNN) Algorithm for Machine Learning \s• K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning approach.
* The K-NN algorithm places the new case in the category that is most similar to the available categories, assuming that the new case/data and the existing cases are similar.
* The K-NN algorithm saves all of the information that is available and categorizes new data based on similarity. This means that utilizing the K-NN method, fresh data may be quickly and accurately sorted into a suitable category.
* The K-NN algorithm can be used for both classification and regression problems, but it is most frequently used for classification issues.
* Since K-NN is a non-parametric method, it makes no assumptions about the underlying data.

• It is also known as a lazy learner algorithm since it saves the training dataset rather than learning from it immediately. Instead, it uses the dataset to execute an action when classifying data.

• The KNN algorithm simply saves the dataset during the training phase and subsequently classifies fresh data into a category that is quite similar to the new data.

• If we have a visual of a creature that resembles both a cat and a dog, but we are unsure of its identity. However, since the KNN method is based on a similarity metric, we may utilize it for this identification. Our KNN model will identify the shared characteristics in the fresh data.

**How does K-NN function?**

* The following algorithm may be used to describe how the K-NN works:
* Step-1: Choose the Kth neighbor's number.
* Compute the Euclidean distance between K neighbours in step two.
* Step 3: Based on the determined Euclidean distance, choose the K nearest neighbours.
* Step 4: Count the number of data points in each category among these k neighbours.
* Step 5: Put the new data points to the category where the neighbour count is highest.
* Step 6: Our model is complete.

**Conclusion**

In this work, numerous machine learning algorithms are surveyed. Today, whether intentionally or not, everyone uses machine learning. from updating images on social networking sites to receiving product recommendations when buying online. The vast majority of the well-known machine learning methods are introduced in this publication.

# 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,