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)

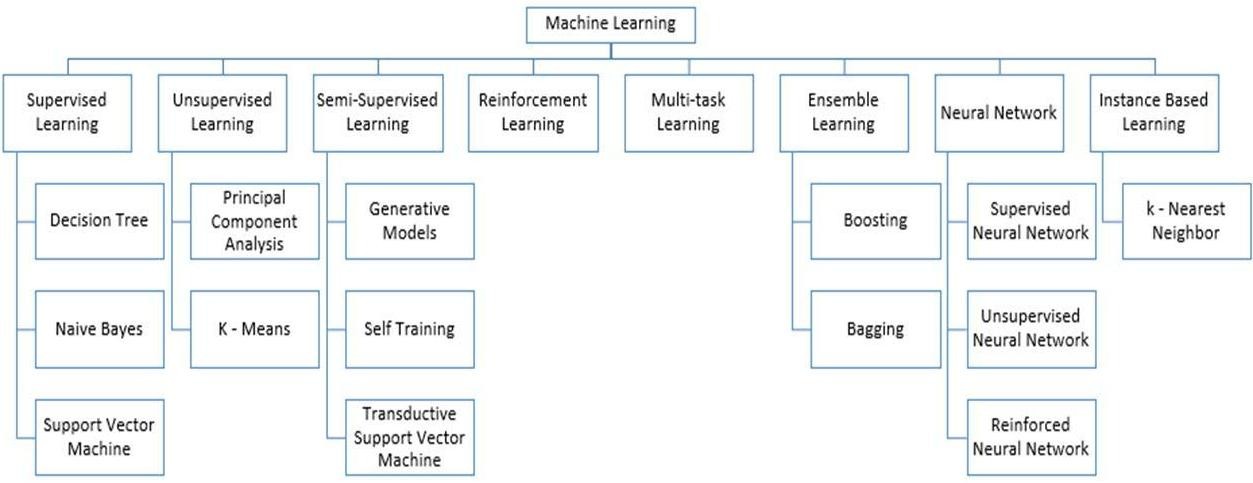
**Abstract: *In this paper, various machine learning methods have been described. Many jobs, such as data mining, image processing, predictive analytics, etc., utilise these algorithms. The main advantage of machine learning is the ability for an algorithm to work independently once it has mastered the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

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

# Introduction

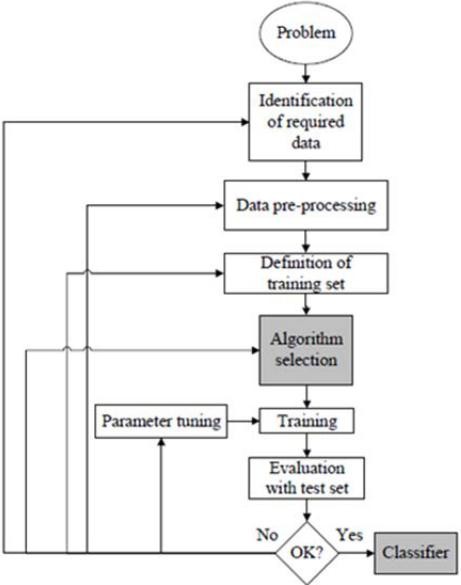
Machine learning is used to teach machines how to handle data more effectively. Even after viewing the data, there are times when we are unable to spot a pattern or make inferences from it. In that case, we employ machine learning [1]. The demand for machine learning has expanded as a result of the abundance 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. In this article, we examine three of the most popular supervised machine learning methods. Supervised machine learning is based on supervision, as its name suggests. With the supervised learning technique, this means that we train the machines using the "labelled" dataset, and then the machine predicts the output based on the training. Here, the labelled data indicates which inputs have already been mapped to which output. More precisely, we may state that after training the machine with input and related output, we ask it to predict the outcome using test dataset. Let's use an illustration to clarify supervised learning. Assume we have a dataset of photos of dogs and cats as our input. Thus, we will first train the machine to comprehend the photos, teaching it things like the size and shape of a dog's tail, the shape of a cat's eyes, their colour, and their height (dogs are taller than cats, for example). Following training, we input a cat image and ask the computer to recognise the object and forecast the outcome. Now that the machine is educated, it will examine every characteristic of the thing, including height, form, colour, eyes, ears, tail, and so on, and determine that it is a cat. As a result, it will be classified as a cat. This is the method the computer uses to recognise the objects in supervised learning.

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

The main goal of the supervised learning technique is to create a map between the input variable (x) and the output variable (y). In the actual world, supervised learning is used for risk assessment, 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**

When a classification problem has a categorical output variable, such as "Yes" or "No," Male or Female, Red or Blue, etc., classification methods are employed to solve the problem. The categories that are present in the dataset are predicted by the categorization algorithms. Spam detection, email filtering, and other examples of categorization systems in use today. Some popular classification algorithms are given below:

### ****Decision trees,****

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

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

1. **Regression**

Regression problems with a linear relationship between the input and output variables are solved using regression techniques. They are employed to forecast variables with continuous outputs, such as market trends, weather forecasts, 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****

## **In the medical field, supervised algorithms are frequently employed for diagnostic reasons. It is done utilizing historical data with labels for illness conditions and medical photos. The machine may diagnose a disease for new patients using such a procedure.**

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

## **Algorithms for supervised learning are used to identify fraudulent consumers, transactions, etc. In order to find the trends that might point to potential fraud, historical data is used.**

## ****Spam detection** - Classification algorithms are employed in spam detection and filtering. These algorithms determine whether an email is spam or not.**

## 2. Unsupervised Machine Learning

Unsupervised learning is distinct from the supervised learning method because, as its name implies, supervision is not required. In unsupervised machine learning, this means that the system is trained on an unlabeled dataset and makes output predictions without any human supervision.

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 unsupervised learning algorithm's primary goal is to classify or group the unsorted dataset based on commonalities, patterns, and differences. The hidden patterns in the input dataset are to be found by the machines.**

### **To better comprehend it, let's use an example. Suppose we feed the machine learning model photographs of a basket of fruit. The model has no prior knowledge of the photos, and its job is to identify patterns and groups of items.**

### **As a result, when the machine is tested with the test dataset, it will now learn its patterns and distinctions, such as colour differences and form differences, and anticipate the output.**

**Unsupervised Machine Learning Categories**

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

### Clustering

### Association

### 1) Clustering

While looking for the innate groupings in the data, we employ the clustering approach. It is a method of clustering the items such that those who have the most similarities stay in one group and share little to none in common with those in other groups. Grouping clients based on their purchase habits is an illustration of the clustering method. Some of the popular clustering algorithms are given below:

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

### 2) Association

### An unsupervised learning method called association rule learning identifies intriguing relationships between variables in a sizable dataset. This learning algorithm's primary goal is to identify the dependencies between data items and then map the variables in a way that maximises profit. This method is mostly used in continuous production, web usage mining, market basket analysis, etc.

### Apriori, Eclat, and FP-growth algorithms are a few of the well-known algorithms for learning association rules.

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

**Advantages:**

* These algorithms, as opposed to supervised ones, can be utilized for more challenging problems since they pirate on unlabeled datasets.
* Unsupervised algorithms are preferred for a variety of applications since it is simpler to get the unlabeled dataset than the labeled dataset.

**Disadvantages:**

### Because the dataset is not labelled and the algorithms are not trained using the exact output in advance, the output of an unsupervised algorithm may be less accurate.

### Working with unsupervised learning is more challenging because it uses a dataset that is unlabeled and does not map to the output.

## Uses for 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:

**Between supervised and unsupervised machine learning, there is a form of method known as semi-supervised learning. It employs a combination of labelled and unlabeled datasets during the training phase and stands in the between of supervised learning (with labelled training data) and unsupervised learning (without labelled training data) techniques.**

**While semi-supervised learning acts on data that has a few labels and is a middle ground between supervised and unsupervised learning, the majority of the data it uses is unlabeled. Labels are expensive, but for corporate needs, there might not be many labels. While supervised and unsupervised learning are dependent on the presence or lack of labels, it is entirely distinct from those methods.**

### **The idea of semi-supervised learning is presented in order to address the shortcomings of supervised learning and unsupervised learning methods. Instead of using solely labelled data as in supervised learning, the primary goal of semi-supervised learning is to make optimal use of all accessible data. An unsupervised learning technique is first used to cluster comparable data, and it also aids in labelling the unlabeled data into labelled data. The reason behind this is that labelled data is more expensive to acquire 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

## **With reinforcement learning, an AI agent (a software component) autonomously explores its surroundings by striking and trailing, acting, learning from experiences, and increasing performance. Reinforcement learning operates on a feedback-based method. The objective of a reinforcement learning agent is to maximise the rewards since the agent is rewarded for every good activity and penalised for every negative action. In contrast to supervised learning, reinforcement learning relies only 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: 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:** Real-time learning algorithms are widely used in game applications. It is utilized to perform at a superhuman level. The video games AlphaGO and AlphaGO Zero are examples of well-known RL algorithms.**

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

* **Robotics : RL is frequently applied in robotics applications. In the industrial and manufacturing sectors, robots are deployed, and reinforcement learning is used to increase their power. The development of intelligent robots utilizing AI and machine learning technologies is a goal shared by many sectors.**

### ****Data Mining:** One of the great uses of NLP is text mining.**

### Advantages and Disadvantages of Reinforcement Learning

**Advantages**

* It aids in the resolution of complicated real-world issues that are challenging to resolve using conventional methods.
* Because of the similarities between the RL learning model and human learning, the most accurate results may be obtained.
* Aids in obtaining long-term effects.

**Disadvantage**

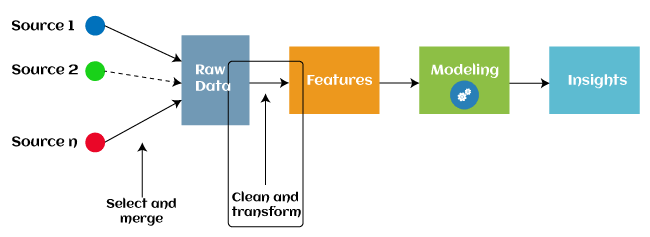
# For straightforward tasks, RL algorithms are not recommended.

# RL algorithms need enormous amounts of data and processing.

# An overflow of states brought on by excessive reinforcement learning may degrade the outcomes.

# Feature Engineering for Machine Learning

**The pre-processing stage in machine learning known as feature engineering is used to turn raw data into features that may be utilised to build a prediction model using either machine learning or statistical modelling. Machine learning feature engineering seeks to enhance model performance. We will study about feature engineering in machine learning in depth in this topic. So first, let's define features before getting into the specifics. And why is feature engineering necessary?**

Machine learning's pre-processing stage, feature engineering, pulls features from unprocessed data. It aids in better communicating a fundamental issue to predictive models, increasing the model's accuracy for unobserved data. The feature engineering method chooses the most practical predictor variables for the model, which is composed of 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:** Finding the most beneficial variables to include in a predictive model is known as feature creation. The procedure needs human ingenuity and intervention and is subjective. The addition, subtraction, and ration operations used to construct the new features provide them a tremendous deal of versatility.
2. **Transformations:** The feature engineering transformation stage entails modifying the predictor variable to raise the model's precision and effectiveness. By ensuring that all the variables are on the same scale and that the model is flexible enough to accept input from a range of data, for instance, it makes the model simpler to comprehend.
3. **Feature Extraction:** A feature engineering procedure that automatically creates new variables by removing existing ones from raw data is called feature extraction. This step's major goal is to decrease the amount of data so that it can be utilised and handled for data modelling more simply. Cluster analysis, text analytics, edge detection algorithms, and principle components analysis are examples of feature extraction techniques (PCA).
4. **Feature Selection:** Only a small subset of the dataset's variables may be used to create a machine learning model; the remainder are either redundant or useless. The overall performance and accuracy of the model may suffer if all these redundant and pointless information are included in the dataset.

In order to remove the unnecessary or less significant features from the data, it is crucial to discover and choose the most appropriate features from the data, which is accomplished with the aid of feature selection in machine learning. By eliminating the duplicate, irrelevant, or noisy characteristics from the original feature set, feature selection is a technique for choosing the subset of the most pertinent features. Below are some benefits of using 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

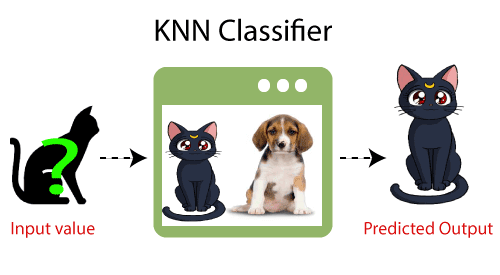
* One of the most well-known Machine Learning algorithms, under the Supervised Learning method, is logistic regression. With a predetermined set of independent factors, it is used to predict the categorical dependent variable.
* Logistic regression forecasts the results of a dependant variable with a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