# Reliable Algorithms for Machine Learning Models: Implementation Research in Data Science

**Kajal Singh, Anukriti Mukherjee**

***Abstract:** Machine Learning generates programs that make predictions and informed decisions about complex problems in an efficient and reliable way. These ML programs autonomously test solutions against the dataset to find the best fit for the problem. The paper aims to review the ML algorithms that develop prediction models by utilizing training dataset and known output. The paper also focuses on ML principles, algorithms, approaches, and applications for Supervised, Unsupervised, and Reinforcement learning that can perform tasks without being explicitly programmed for it. Completely opposite to rule-based programming, the machine learning paradigm uses examples of real data sets and pre-process it before providing the desired outputs based on these examples. In the case of more involved and complex tasks, it can be challenging for humans to explicitly program the models. On the other hand, it can be more effective to help the machines develop the algorithms for advanced tasks. This paper will also present the trending real-world applications of Machine Learning in Image Recognition and Biomedicine. Additionally, it will provide a background analysis of machine learning and related fields of data science.*

***Index Terms:** Machine Learning (ML), Supervised Learning, Unsupervised Learning, Reinforcement Learning, KNN, K-Means Clustering.*

## I. INTRODUCTION

Machine Learning is an emerging technology that deploys various mathematical algorithms to learn automatically from the provided dataset. As a result, the ML models not only provide predicted output but also improvise the programs based on previous experiences without being explicitly programmed to do so [1]. Humans are bound to learn from their experiences and make decisions based on these learnings [2]. Similarly, ML makes machines behave like humans and learn from the data to make accurate and effective predictive models [3]. In a data-driven world, it is impossible for humans to manually process data and generate predictions. This is where machine learning is applied to solve complex problems [4]. It even identifies hidden patterns in the data extracting meaningful insights and making effective predictions for the users [5]. ML can be broadly classified into three categories: Supervised Learning, Unsupervised Learning, and Reinforcement Learning.

Based on the provided training dataset, the predicted output by Supervised Learning can be of two types: Regression and Classification. Unsupervised Learning is a method that trains the model without any sort of supervision [6]. Based on the predicted output by unsupervised learning, it can be of two types: Clustering and Association. Reinforcement is a type of learning that tends to keep on improving the model accuracy by interacting with the environment constantly [7]. There are several real-world applications of ML that are vigorously being implemented in business analytics, biomedicine, image recognition, product recommendation, self-driving cars, Google translators, Amazon Alexa, and Apple Siri voice assistants. In the data-driven modern world, the data is being produced at a massive rate that signifies an emerging scope of machine learning in various other fields like the automotive industry, eCommerce, cloud computing, robotics, and quantum computing [8].

## II. BACKGROUND

In the modern-day, the question "Can a machine think?" has been replaced by "Can machine performs tasks that we can do?" which signifies that a fundamentally operational definition has been replaced by a cognitive one. The term Machine Learning was first coined by Arthur Samuel who was a pioneer in the field of artificial intelligence and computer gaming. It was interchangeably used with terms like self-teaching computer that was majorly used for pattern recognition and classification. Now, machine learning has two objectives, one is to classify the data as per the applied model and algorithm, the other one is to predict the future outcome based on these models.

- *Artificial Intelligence Vs Machine Learning*

Around the 1970s, machine learning was a part of the AI revolution but later, it branched off where it evolved on its own. AI is essentially a system that is based on human-made thinking power that does not require to be pre-programmed. They utilize algorithms that can work with their own intelligence. ML is a type of AI that utilizes supervised or unsupervised data to make predictions or calculated decisions without explicitly programming it. ML enables the system to continue learning and improving based on the provided mathematical data models.

- *Data Mining Vs Machine Learning*

ML and Data Mining are closely related as they employ similar methodologies. Data Mining is focused on the analysis of the unknown properties of the data where it employs ML models but with different purposes. On the other hand, ML focuses on prediction based on the historical data models where it can utilize the data mining techniques as unsupervised algorithms to improve the model's accuracy.

Manuscript received on February 16, 2022.
Revised Manuscript received on February 21, 2022.
Manuscript published on March 30, 2022.
***** Correspondence Author

**Kajal Singh***, Business Intelligence Analyst, Schneider Electric, Bangalore, India. Email: kjlsingh65@gmail.com

**Anukriti Mukherjee**, Programmer Analyst, Cognizant Technology Solutions, Kolkata, India. Email: anukriti98@gmail.com

- *Statistics Vs Machine Learning*

Statistics and ML are closely related fields in terms of the methodologies but different in their principle output. Where Statistical methods are more focused on providing the population inferences from the provided set of samples, ML models are utilized to provide generalized predictive patterns. Michael Jordan suggested that the idea of ML methods have been used for statistical calculations based on which he suggested to term the overall field as Data Science.

## III. ALGORITHMS

Machine Learning Algorithms develop predictive models to showcase interesting patterns such as trends or anomalies from complex data. These algorithms have the potential to create significant value out of the dataset that can be implemented for various projects. There are many ML algorithms that can be broadly classified into three categories:

### *A. Supervised Learning*

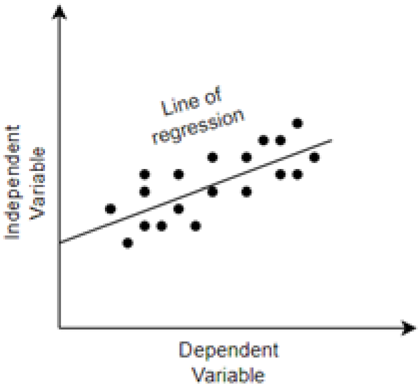
A Supervised Learning algorithm generates the model that will provide predictions based on the training dataset and known output. This learning has been utilized in various real-world projects to predict accurate results such as detecting anomalies from a pattern. The model keeps on correcting over time until the error is minimized sufficiently. In short, this algorithm uses a training data set that includes input values X and output values Y where the mapping function will keep on learning based on the variables.

Y = f(X)

The aim is to train the algorithm so well that when we supply input value X, it should be able to predict the output value Y accurately. This is called Supervised Learning as the model trains and supervises the learning process. Supervised Learning Algorithms can be further classified into 2 categories:

1. Regression

Under this type, the model predicts the output as continuous numerical values by establishing the relationship between the dependent and independent variables. Regression is a type of supervised algorithm as it finds the linear correlation for target variables and independent ones. Most common regression problems include predicting the employee’s salary, house price, etc.



**Graph 1.1**

Graph 1.1 shows the linear relationship between the independent and dependent variables where data points are following the line of regression. The Regression algorithm is simpler to implement and less complex as compared to other ML algorithms. On the other hand, if there is an anomaly in the provided training data set then it can affect the algorithm badly giving unexpected results. Additionally, it assumes the linear relation between all the variables hence is rarely used for real-world cases.

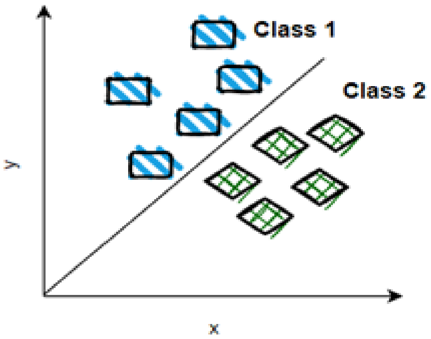
2. Classification

The Classification algorithm generates a predictive model that segregates the input data into specific class labels based on supervised learning. Here, the training data set must sufficiently cover all possible scenarios to accommodate the class labels by including examples of each case. Classification algorithms are suitable for categorical problems based on the training dataset where they learn from the provided observations and further classify the new observations into various sections. Some classification problems can cover cases like yes/no and 1/0. In regression, the model deals with the continuous values but under classification, it segregates the data into categories.

The classification algorithm is based on the below model function:

y = f(x)

Here, the discrete output y is mapped to the input value x. As a result, we can get a predicted output belonging to a specific category.



**Graph 1.2**

Graph 1.2 shows the dataset being categorized into class 1 and class 2. Machine learning algorithm utilizes the pre-categorized training dataset to classify the future data into learned categories. The algorithm works on the concept of determining the likelihood of a particular type of training dataset falling into a predetermined category. A classification algorithm is a type of pattern recognition based on which a similar pattern is determined for the input data. One of the common use cases for classification is to filter the mails into “spam” or “non-spam” categories.

Principle

1. Supervised algorithm generates a model based on the determined patterns of the training dataset. This learning experience is further used to predict the output for the future input dataset.
2. The learning experience also enables the models to improve their accuracy that optimizes the overall performance of the algorithm.
3. The efficiency of supervised learning makes it the most favorable algorithm to be utilized in real-world applications.

The commonly used supervised algorithm is discussed below:

- KNN Clustering: K Nearest neighbor or KNN is a type of unsupervised learning producing output that is easy to interpret and consumes lesser calculation time. KNN algorithm considers the variable value "K" to choose the nearest neighbors. This value of K affects the class boundaries that segregate the dataset. For any training sample, choosing K=1 will always give the error rate as 0.

### B. *Unsupervised Learning*

An Unsupervised learning algorithm generates the model that detects the pattern from the data on its own. In comparison to supervised algorithms, it is computationally complex and less accurate. The algorithm deals with un-labeled dataset and recognizes unknown patterns that determine the answers for the input dataset. It can be useful in finding anomalies and outliers in the data by self-organizing and determining the pattern based on the probability density.

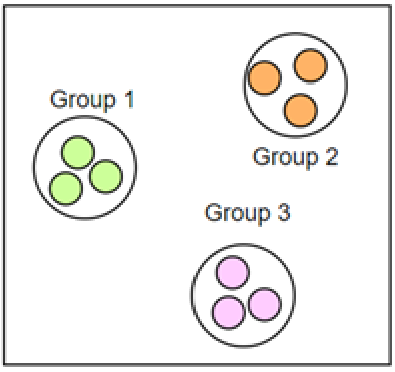
Principle

1. Unsupervised Algorithm works with un-labeled data and finds unknown patterns through self-learning models.
2. The determined patterns are useful to predict the association, clustering, and categorization of the input dataset.
3. This can also help in detecting the anomalies and defects in the data based on which corrective measures can be taken.

Unsupervised learning can be further split up into 2 major categories i.e. Clustering and Association.

1. Clustering

Clustering is a type of unsupervised learning where it determines the pattern in the data with the purpose of generating groups or clusters. These groups can be based on various factors like shape, size, color, etc. The model keeps on learning from the data and identifies similarities to pinpoint the hidden patterns. In this way, the algorithm will classify the data into clusters and generate the output based on the clustered data.



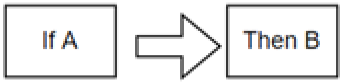
**Graph 2.1**

The above graph 2.1 shows the classification of the dataset into 3 classes: Group 1, Group 2, and Group 3. The pattern is generated using unsupervised learning by segregating the data points based on the similarities. In the graph, the groups are divided based on the color pattern. The commonly used clustering algorithms are described below:

- Hierarchical Clustering: Under this, the data points are grouped together based on various features. The self-learning capability keeps on grouping the data until the hierarchy between the data points is created.
- K- Means Clustering: This method is focused on determining the centroid of the clusters and makes sure that the distance between the data points and the centroid is as less as possible. This type of unsupervised learning keeps on improving the performance of the model while also maintaining the distance between the clusters. Further, these clusters can be named as required for predicting the output for input data points.

2. Association

Association is a type of unsupervised learning where it determines the relationship between data items. This dependency is leveraged to predict the output for new input data items. Unlike other prescriptive models, the association rule is descriptive in nature that determines the data points dependency in terms of rules. The association rule works on the logic of the If and Else clause whereas if we have A then we will get B.



**Graph 2.2**

In graph 2.2, the "If" statement is known as Antecedent, and "Then" statement is known as Consequent. We can find the relation between two items and if the number of items increases then the association rules also increase accordingly. Following are the metrics to measure the relation between the data items:

a. Support: It is defined as the frequency of the Y occurrence in relation to the total transactions. If there are Y datasets with total transactions as T, then the Support can be calculated as:

$$\text{Support (X)} = \frac{\text{Frequency (Y)}}{\text{T}}$$

b. Confidence: It is defined as the frequency of occurrence of X and Y together in relation to the frequency of X alone. In short, it determines how frequently the association rule is found to be true. If there are X and Y datasets, then Confidence can be calculated as:

$$\text{Confidence} = \frac{\text{Frequency (X,Y)}}{\text{Frequency (X)}}$$

c. Lift: It is defined as the observed support for X and Y in relation to individual expected support for X and Y. In short, it determines the strength of the association rule.

If Lift = 1, then there is no association between the data points. If Lift > 1, then the association between the data points is positive. If Lift < 1, then the association between the data points is negative.

$$\text{Lift} = \frac{\text{Support }(X,Y)}{\text{Support }(X) \times \text{Support }(Y)}$$

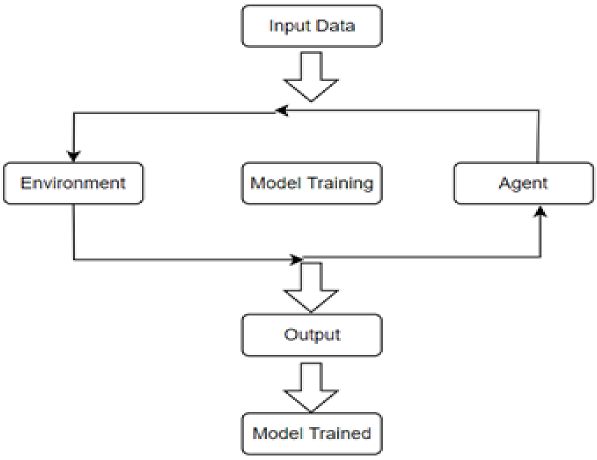
Association learning can broadly be classified into three main algorithms:

1. Apriori Algorithm: This algorithm generates the dependency rules based on the frequent datasets. It generally helps to understand the data that can be brought together.
2. Eclat Algorithm: Equivalence Class Transformation algorithm deals with the frequent datasets but engages in faster operation and execution than Apriori algorithm.
3. F-P Growth Algorithm: Frequent Pattern Growth or F-P Growth Algorithm extracts the most frequent patterns from the dataset. It is the improved version of Apriori that generates the patterns and showcases them in the form of a tree structure.

### *C. Reinforcement Learning*

Reinforcement learning is a type of hit and trial learning where the machine keeps on learning from its successes and failures. The machine learns from the environment based on which the learning agent performs actions to improve the accuracy of the model. As per the process, if the algorithm predicts the correct output for any provided dataset then the reinforcement signal is sent to further improve the model.

Reinforcement learning is commonly used in software or machines to predict the best possible option in a specific situation. This is like supervised learning, but the reinforcement model keeps on learning from the environment unlike supervised where the answers are provided with the training dataset and the model predicts the output based on it.



**Graph 3.1**

Under reinforcement learning, the model learns from the experience and environment even in the absence of the training dataset. Industrial automation and data processing are some of the common practical applications for Reinforcement learning.

## IV. APPLICATIONS

Supervised, Unsupervised and Reinforcement Machine learning algorithms are utilized in various real-world applications and have capability to further grow overtime. Some of the common applications are discussed below:

a. Image Recognition: It is the most common real-world application of machine learning. Image recognition identifies the objects in the digital image based on the supplied dataset. It is frequently used for facial recognition in law enforcement to match the image from the database of people.

There are different methods that have evolved overtime from machine learning and artificial intelligence, but the core focus of image recognition is to differentiate the objects in images. Using supervised algorithms, image recognition deals with object detection from the training dataset to create perceptions of how classes should look. This can be used to identify the hidden patterns, color, places, shapes, and many more from the input images. One of the common machine learning models used for image recognition is defined below: Support Vector Machines: SVM (Support Vector Machines) model works on the concept of generating the histograms that show good and bad examples of images (supervised learning) and then these histograms are trained with this learning. When the input image is supplied, the model matches the various parts of the image with the trained histogram model.

b. Digital Image Processing: With the digitization of medical data from electronic health records to clinical trial results, machine learning algorithms can help predict, analyze, and categorize large biomedical datasets. Machine Learning can have a revolutionary impact on healthcare in terms of computer aided diagnosis, personalized medicine, drug discovery, multimodal image processing and image segmentation [11].

In recent studies, researchers have applied digital image processing techniques to analyze tumors using Magnetic Resonance Imaging [12]. Computer assisted diagnosis of brain tumor classification is one of the most upcoming applications of Machine Learning methods in medical imaging. Evangelia I. Zacharaki et al. used pattern classification, feature extraction and region of interest (ROI) definition for differentiating different grades of gliomas. They observed that the SVM algorithm produced promising results in computer-based brain tumor evaluation [13]. Jainy Sachdeva et al. used a dual neural network essemble to assist radiologists in successfully identifying primary brain tumors such as astrocytoma and glioblastoma multiforme [14]. Many ML can also be used for predicting the recurrence of cancer. YucanXu et al. proposed a study to predict the recurrence of cancer in Stage IV colorectal cancer patients. They used Logistic Regression, Decision Tree, Gradient Boosting and LightGBM, observing that the gradient boosting and LightGBM models gave more accurate predictions than the other two algorithms [15].

c. Drug development: It is a complex process, comprising multiple variables. Machine Learning can be used to improve identification of prognostic biomarkers and high-quality pathological data. ML approaches applied to data collected during drug discovery can aid medical decision making and reduce the failure rate of drug development [16]. Moe Elbadawi et al. describes the challenges in conventional ML algorithms and the use of advanced techniques to automate the process of drug discovery, using Reinforcement Learning and Bayesian Neural Networks [17].

With the advancement in AI and development of continual learning models, ML can be used in healthcare to manage patient data and deliver diagnostic and prognostic results. A continual learning ML model can be safely implemented in diagnostic testing where the trained data would make a diagnostic call every time new patient data is available [18]. ML can also be applied to Electronic Health Records (EHRs) to personalize therapeutic treatments by using patient response in EHR to train machine learning methods. Leveraging EHR observational data from patients that have received the treatment vs those who haven't can individualize treatment recommendations. Standard Supervised Learning algorithms can be used for causal inference of these two datasets, supported by Balancing Neural Networks and Regression Trees [19].

## V. CONCLUSION

Machine Learning is a remarkable technology where it has already proposed solutions for extremely complex problems with supervised, unsupervised, and reinforcement learning models. In the future, the improved machine learning models will introduce innovative procedures for industrial applications. These models will deal with a large set of data to generate the algorithms simplifying the complex set of rules. Hence, the machine learning models will surely become more versatile with continued innovation to handle complex data modeling.

## REFERENCES

1. M. Zhang and Z. Zhou, "A Review on Multi-Label Learning Algorithms" in IEEE Transactions on Knowledge & Data Engineering, vol. 26, no. 08, pp. 1819-1837, 2014. doi: 10.1109/TKDE.2013.39
2. Trends in extreme learning machines: a review, by Huang, G., Huang, G., Song, S., & You, K. (2015). Neural Networks, (HIC: 0 , CV: 0)
3. A survey of multiple classifier systems as hybrid systems , by Corchado, E., Graña, M., & Wozniak, M. (2014). Information Fusion, 16, 3-17. ( HIC: 1 , CV: 22)
4. solo-learn: A Library of Self-supervised Methods for Visual Representation Learning Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, Elisa Ricci, 2022.
5. DoubleML - An Object-Oriented Implementation of Double Machine Learning in Python Philipp Bach, Victor Chernozhukov, Malte S. Kurz, Martin Spindler; 23(53):1−6, 2022
6. htps://www.sciencedirect.com/topics/engineering/machine-learning-algorithm
7. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/
8. https://en.wikipedia.org/wiki/Machine_learning
9. Approximate Information State for Approximate Planning and Reinforcement Learning in Partially Observed Systems Jayakumar Subramanian, Amit Sinha, Raihan Seraj, Aditya Mahajan, 2022
10. http://places.csail.mit.edu/places_NIPS14.pdf
11. K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp.910-914, doi:10.1109/ICECA.2018.8474918
12. Park, C., Took, C.C. & Seong, JK. Machine learning in biomedical engineering. Biomed. Eng. Lett. 8, 1–3 (2018). https://doi.org/10.1007/s13534-018-0058-3
13. Evangelia I. Zacharaki; Sumei Wang; Sanjeev Chawla; Dong Soo Yoo; Ronald Wolf; Elias R. Melhem; Christos Davatzikos (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. 62(6), 1609–1618
14. Sachdeva, Jainy, et al. "A dual neural network ensemble approach for multiclass brain tumor classification." International journal for numerical methods in biomedical engineering 28.11 (2012): 1107-1120
15. Xu, Y., Ju, L., Tong, J. et al. Machine Learning Algorithms for Predicting the Recurrence of Stage IV Colorectal Cancer After Tumor Resection. Sci Rep 10, 2519 (2020). https://doi.org/10.1038/s41598-020-59115-y
16. Vamathevan, J., Clark, D., Czodrowski, P. et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18, 463–477 (2019). https://doi.org/10.1038/s41573-019-0024-5
17. Moe Elbadawi, Simon Gaisford, Abdul W. Basit, Advanced machine-learning techniques in drug discovery, Drug Discovery Today, Volume 26, Issue 3, 2021, Pages 769-777,ISSN1359-6446, https://doi.org/10.1016/j.drudis.2020.12.003
18. Lee, Cecilia S., and Aaron Y. Lee. "Clinical applications of continual learning machine learning." The Lancet Digital Health 2.6 (2020): e279-e281
19. Bica, Ioana, et al. "From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges." Clinical Pharmacology & Therapeutics 109.1 (2021): 87-100

## AUTHOR PROFILE

**Kajal Singh** is a Business Intelligence Analyst, technologist, and researcher at Schneider Electric, Bangalore, India. She holds a Bachelor of Technology (B.Tech) degree in Electronics and Communication Engineering Department from Vellore Institute of Technology, Tamil Nadu, India (2020 graduated). She has gained experience in data modeling, mapping, and project management and further holds technical expertise in Oracle Business Intelligence Enterprise Edition (OBIEE), Oracle Transactional Business Intelligence, Oracle Analytics Cloud (OAC), Oracle Data Integrator (ODI) tools and techniques. Kajal is an Oracle certified Business Analytics Expert, Autonomous DatabaseSpecialist, and government certified fitness trainer. She is passionate about OAC Machine Learning, Artificial Intelligence, Operations Management, and Workflow Analysis. She aspires to work in the field of data-driven product management and is active on JIRA and Confluence tools to adopt the relevant agile frameworks like Kanban for her team projects.

**Anukriti Mukherjee** is a Programmer Analyst at Cognizant Technology Solutions, Kolkata, India. She holds a Bachelor of Technology (B. Tech.) in Biomedical Engineering from Vellore Institute of Technology, Tamil Nadu, India (2020 graduated). She has experience in pharmaceutical data management (Oracle, MySQL), Laboratory Information Management Systems (LABWARE LIMS) and technical expertise in biomedical equipment design, disease modelling and biosignal processing (MATLAB). Anukriti has done multiple certification courses on Data Science, Digital Signal Processing and Systems View in Patient Safety. She aspires to be a clinical engineer, developing innovative and advanced biomedical devices in the field of cardiovascular technology with a focus on AI based diagnostic cardiac devices.