A Study on Machine Learning Algorithms and its Applications

Mrs S.Subhashini , Mrs.Y.sharmila Begam, Dr.P.Umamaheswari
Research Scholar , Anna university Regional campus, Madurai,
Research Scholar , Anna university Regional campus, Madurai,
Assistant Professor, Dept of computer science, Anna university Regional campus, Madurai
E-mail :subhacsmsn@gmail.com,, sharmila.yusuf@gmail.com ,dharshukiran@gmail.com

Abstract

Machine learning is one of the rapidly developing fields of technology . It is growing very rapidly day by day . Applications of machine learning are vast in our daily life. Recently machine learning techniques are used in Google Maps, Google assistant, Alexa, Cortana**,** Siri etc. Machine learning's face detection and recognition algorithm are used in facebook for **a**utomatic friend tagging suggestion**.** ML algorithms for speech recognition is used in search by voice in google maps which shows the correct shortest route and predicts the traffic conditions. ML algorithms are also used for product recommendation to the user in Amazon**,** Netflix . Tesla, the car manufacturing company uses ML algorithms in manufacturing self driving cars .Machine learning algorithms such as multi-Layer Perceptron**,** Decision tree**,** and Naïve Bayes classifier are used for email spam filtering and malware detection. Voice instructions such as Play music, call someone, Open an email, Scheduling an appointment, etc. are given in virtual assistants using machine learning algorithms . Machine learning makes online transaction safe and secure by detecting fraud transaction, fake accounts**,** fake id**s**, and steal money in the middle of a transaction. Feed Forward Neural network algorithms checks whether it is a genuine transaction or a fraud transaction. Machine learning's long short term memory neural network is used for the prediction of stock market trends. Machine learning is used for disease diagnoses and Google's GNMT (Google Neural Machine Translation) which converts the text into our known languages. In this article, ML applications and various steps involved in ML life cycle are discussed . This article presents a study about various types of ML algorithms , ML in data processing, ML in data cleaning and challenges of ML ..

Keywords: Machine learning , supervised learning, unsupervised learning , Data cleaning, Model training.

1. Introduction

Machine learning is a subset of artificial intelligence which makes the computer systems to learn automatically without being explicitly programmed. It focuses primarily on the creation of algorithms that enable a computer to independently learn from data and previous experiences. Arthur Samuel first coined the term "machine learning" in 1959. Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed. It creates a mathematical model without being explicitly programmed which aids in making predictions or decisions with the assistance of sample historical data or training data. Machine learning uses statistics and computer science for developing predictive machine learning models. Algorithms that learn from historical data are either constructed or utilized in machine learning. The performance of the ML algorithm depends on the quantity of information we provide. A machine can learn when it is given more data to improve its performance. Some of the features of machine learning are as follows

- Detect various patterns in a given dataset.
- It learns from past data and improve automatically.
- It is a data-driven technology.
- Machine learning is much similar to data mining as it also deals with the huge amount of the data.
- Rapid increment in the production of data
- Solving complex problems which are difficult for a human
- Decision making in various sector including finance

Finding hidden patterns and extracting useful information from

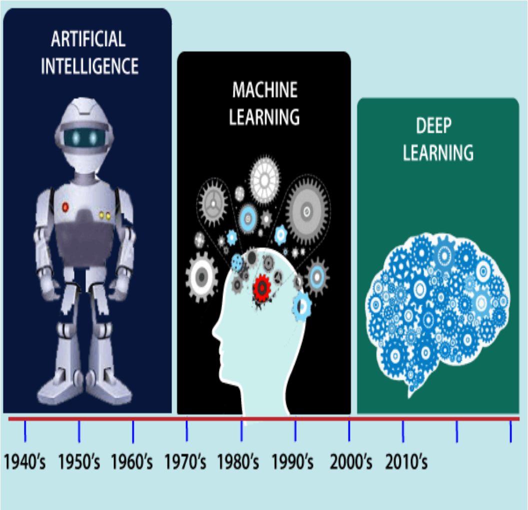


Fig 1 : ML historical data

Year	**Description**
1834	Charles Babbage , the father of the computer conceived a device that could be programmed with punch cards. All modern computers rely on its logical structure in these days.
1936	Alan Turing developed a theory that how a machine can determine and execute a set of instructions.
1940	The first manually operated computer, “ENIAC” was invented. It was the first electronic general-purpose computer. After that stored program computer such as EDSAC in 1949 and EDVAC in 1951 were invented.
1943	A human neural network was modeled with an electrical circuit in 1943. Some scientists started applying their idea to work and analyzed how human neurons might work in 1950.
1950	Alan Turing published a seminal paper entitled “Computer Machinery and Intelligence,” In this paper, “Can machines think?” is it possible to make machine think and learn were discussed.
1952	Arthur Samuel, the pioneer of machine learning, created a program to play a checkers game in IBM computer. It performed better more this machine already played.
1959	• Machine Learning” was first coined by Arthur Samuel.

	• First neural network was applied to a real-world problem
1974 to 1980	Failure of machine translation occurred
1985	Terry Sejnowski and Charles Rosenberg invented a neural network NETtalk able to teach itself how to correctly pronounce 20,000 words in one week.
1997	Deep blue intelligent computer won the chess game against the chess expert Garry Kasparov, and it became the first computer which had beaten a human chess expert.
2006	Elastic Compute Cloud (EC2) was launched by Amazon to provide scalable computing resources that made it easier to create and implement machine learning models.
2007	Accuracy of Netflix's recommendation algorithm
2008	Google delivered the cloud based google Forecast Programming interface that integrates AI into their applications
2010	The ImageNet Huge Scope Visual Acknowledgment Challenge (ILSVRC) was presented, driving progressions in PC vision, and prompting the advancement of profound convolutional brain organizations (CNNs).
2011	IBM's Watson demonstrates the potential of question-answering systems and natural language processing
2012	• AlexNet, a profound CNN created by Alex Krizhevsky is a predominant methodology in PC vision. • Google's Cerebrum project by Andrew Ng and Jeff Dignitary utilized profound methodology to perceive felines from unlabeled YouTube recordings.
2013	• Generative adversarial networks (GANs) which made it possible to create realistic synthetic data. • Google's DeepMind Technologies focused on deep learning and artificial intelligence.
2014	○ Facebook presented the DeepFace framework for

	facial acknowledgment. ○ AlphaGo a program created by DeepMind at Google defeated a world champion Go player .It demonstrated the potential of reinforcement learning in challenging games.
2015	• Microsoft developed an open-source profound learning library called mental toolbox (or CNTK). • The performance of sequence-to-sequence models in tasks like machine translation was enhanced by the introduction of the idea of attention mechanisms.
2016	• AI which focuses on making machine learning models easier to understand received some attention. • Google's DeepMind created AlphaGo Zero with Go abilities to play without human information.
2017	• pretrained models to be utilized for different errands with restricted information.
At Present	• Machine learning technology used in self-driving cars, Amazon Alexa, Chatboats, and the recommender system. • Present day AI models can be utilized for making different climate expectation, sickness forecast, financial exchange examination, and so on.

Table 1 . History of ML

2. Machine learning Life cycle

Machine learning life cycle is a cyclic process to build an efficient machine learning project. The main purpose of the life cycle is to find a solution to the problem or project. Seven major steps in machine learning life cycle are as follows:

- ○ Gathering Data
- ○ Data preparation
- ○ Data Wrangling
- ○ Analyse Data
- ○ Train the model

- Test the model
- Deployment

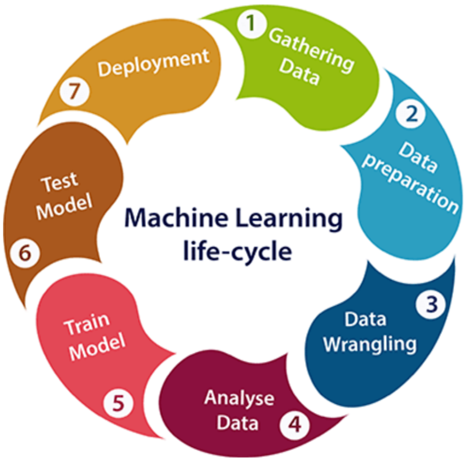


Fig 2. Machine learning cycle

Gathering Data:

Data Gathering is the first step of the machine learning life cycle. The goal of this step is to identify and obtain all data-related problems. First step is to identify the different data sources, as data can be collected from various sources such as files**,** database**,** internet**,** or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more will be the data, the more accurate will be the prediction.This step includes the below tasks:

- Identify various data sources
- Collect data
- Integrate the data obtained from different sources

By performing the above task, a coherent set of data also called as a dataset is obtained. It will be used in further steps.

Data preparation

After collecting the data, we need to prepare it for further steps. Data preparation is a step where we put our data into a suitable place and prepare it to use in our machine learning training.In this step, first, we put all data together, and then randomize the ordering of data.This step can be further divided into two processes:Data exploration is used to understand the nature of data that we have to work with. We need to understand the characteristics, format, and quality of data.

A better understanding of data leads to an effective outcome. In this, we find Correlations, general trends and outliers . The next step is preprocessing of data for its analysis.

Data Wrangling

Data wrangling is the process of cleaning and converting raw data into a useable format. It is the process of cleaning the data, selecting the variable to use, and transforming the data in a proper format to make it more suitable for analysis in the next step. It is one of the most important steps of the complete process. Cleaning of data is required to address the quality issues. It is not necessary that data collected is always useful as some of the data may not be useful. In real-world applications, collected data may have various issues, including:

- Missing Values
- Duplicate data
- Invalid data
- Noise

Various filtering techniques are used to clean the data. It is mandatory to detect and remove the above issues because it can negatively affect the quality of the outcome.

Data Analysis

The cleaned and prepared data is passed on to the analysis step. This step involves:

- Selection of analytical techniques
- Building models

- Review the result

The aim of this step is to build a machine learning model to analyze the data using various analytical techniques and review the outcome. It starts with the determination of the type of the problems, where we select the machine learning techniques such as Classification**,** Regression**,** Cluster
analysis**,** Association, etc. then build the model using prepared data and evaluate the model

Train Model

The model is trained to improve its performance for better outcome of the problem. Datasets are used to train the model using various machine learning algorithms. Training a model is required so that it can understand the various patterns, rules, and, features.

Test Model

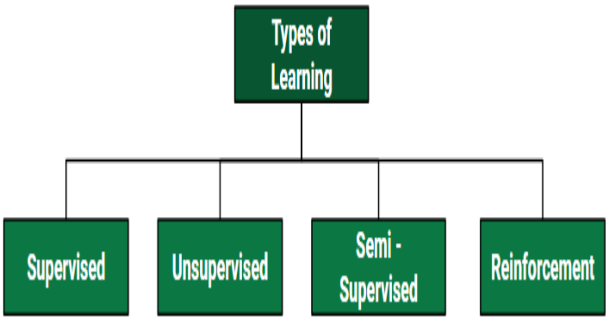
Once the machine learning model has been trained on a given dataset, then test the model. In this testing, check for the accuracy of our model by providing a test dataset to it. Testing the model determines the percentage accuracy of the model as per the requirement of project or problem.

Deployment

The last step of machine learning life cycle is deployment. Deploy the model in the real-world system.

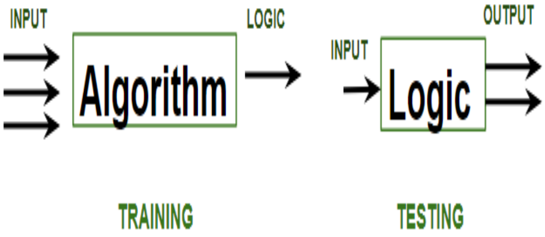
3. Types of Machine learning algorithms

A machine is said to be learning from past Experiences(data feed-in) with respect to some class of tasks if its Performance in a given Task improves with the Experience. For example, assume that a machine has to predict whether a customer will buy a specific product let's say "Antivirus" this year or not. The machine will do it by looking at the previous knowledge/past experiences i.e. the data of products that the customer had bought every year and if he buys an Antivirus every year, then there is a high probability that the customer is going to buy an antivirus this year as well. This is how machine learning works at the basic conceptual level.

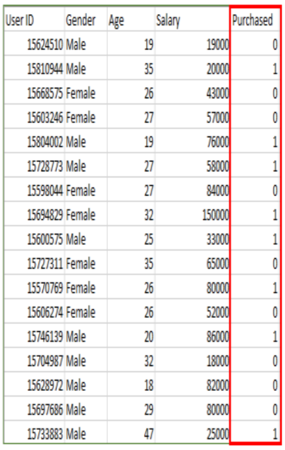


4. Supervised Machine Learning

Supervised learning is a machine learning technique that is widely used in various fields such as finance, healthcare, marketing, and more. It is a form of machine learning in which the algorithm is trained on labeled data to make predictions or decisions based on the data inputs. In supervised learning, the algorithm learns a mapping between the input and output data. This mapping is learned from a labeled dataset, which consists of pairs of input and output data. The algorithm tries to learn the relationship between the input and output data so that it can make accurate predictions on new, unseen data.

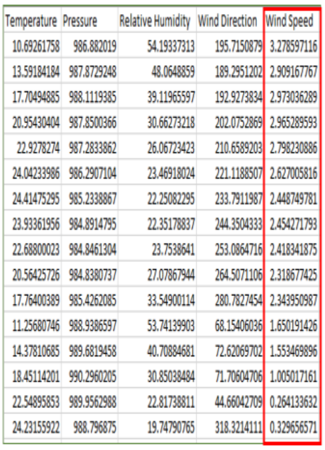


Supervised learning is where the model is trained on a labelled dataset. A labelled dataset is one that has both input and output parameters. In this type of learning both training and validation, datasets are labelled as shown in the figures below. The labeled dataset used in supervised learning consists of input features and corresponding output labels. The input features are the attributes or characteristics of the data that are used to make predictions, while the output labels are the desired outcomes or targets that the algorithm tries to predict.



User ID	Gender	Age	Salary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	1
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	1
15728773	Male	27	58000	1
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	1
15727311	Female	35	65000	0
15570769	Female	26	80000	1
15606274	Female	26	52000	0
15746139	Male	20	86000	1
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1

Figure A: CLASSIFICATION



Temperature	Pressure	Relative Humidity	Wind Direction	Wind Speed
10.69261758	986.882019	54.19337313	195.7150879	3.278597116
13.59184184	987.8729248	48.0648859	189.2951202	2.909167767
17.70494885	988.1119385	39.11965597	192.9273834	2.973036289
20.95430404	987.8500366	30.66273218	202.0752869	2.965289593
22.9278274	987.2833862	26.06723423	210.6589203	2.798230886
24.04233986	986.2907104	23.46918024	221.1188507	2.627005816
24.41475295	985.2338867	22.25082295	233.7911987	2.448749781
23.93361956	984.8914795	22.35178837	244.3504333	2.454271793
22.68800023	984.8461304	23.7538641	253.0864716	2.418341875
20.56425726	984.8380737	27.07867944	264.5071106	2.318677425
17.76400389	985.4262085	33.54900114	280.7827454	2.343950987
11.25680746	988.9386597	53.74139903	68.15406036	1.650191426
14.37810685	989.6819458	40.70884681	72.62069702	1.553469896
18.45114201	990.2960205	30.85038484	71.70604706	1.005017161
22.54895853	989.9562988	22.81738811	44.66042709	0.264133632
24.23155922	988.796875	19.74790765	318.3214111	0.329656571

Figure B: REGRESSION

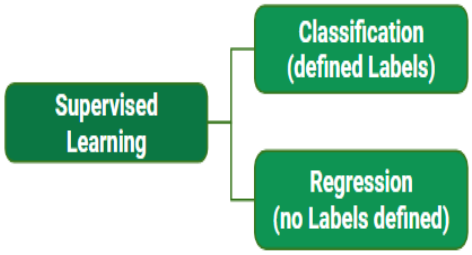
Both the above figures have labelled data set as follows:

- Figure A**:** It is a dataset of a shopping store that is useful in predicting whether a customer will purchase a particular product under consideration or not based on his/ her gender, age, salary.
  Input: Gender,age,Salary
  Output: Purchased i.e. 0 or 1; 1 means yes the customer will purchase and 0 means that the customer won't purchase it.
- Figure B: It is a Meteorological dataset that serves the purpose of predicting wind speed based on different parameters.
  Input: Dew Point, Temperature, Pressure, Relative Humidity, Wind Direction
  Output: Wind Speed

Training the system: While training the model, data is usually split in the ratio of 80:20 i.e. 80% as training data and the rest as testing data. In training data, we feed input as well as output for 80% of data. The model learns from training data only. We use different machine learning algorithms(which we will discuss in detail in the next articles) to build our model. Learning means that the model will build some logic of its own. Once the model is ready then it is good to be tested. At the time of testing, the input is fed from the remaining 20% of data that the model has never seen before, the model will predict some value and we will compare it with the actual output and calculate the accuracy.

5. Types of Supervised Learning Algorithm

Supervised learning is typically divided into two main categories: regression and classification. In regression, the algorithm learns to predict a continuous output value, such as the price of a house or the temperature of a city. In classification, the algorithm learns to predict a categorical output variable or class label, such as whether a customer is likely to purchase a product or not.One of the primary advantages of supervised learning is that it allows for the creation of complex models that can make accurate predictions on new data. However, supervised learning requires large amounts of labeled training data to be effective. Additionally, the quality and representativeness of the training data can have a significant impact on the accuracy of the model.Supervised learning can be further classified into two categories:


Supervised Learning
Classification (defined Labels)
Regression (no Labels defined)

Regression

Regression is a supervised learning technique used to predict continuous numerical values based on input features. It aims to establish a functional relationship between independent variables and a dependent variable, such as predicting house prices based on features like size, bedrooms, and location.The goal is to minimize the difference between predicted and actual values using algorithms like Linear Regression, Decision Trees, or Neural Networks, ensuring the model captures underlying patterns in the data.

Classification

Classification is a type of supervised learning that categorizes input data into predefined labels. It involves training a model on labeled examples to learn patterns between input features and output classes. In classification, the target variable is a categorical value. For example, classifying emails as spam or not.The model's goal is to generalize this learning to make accurate predictions on new, unseen data. Algorithms like Decision Trees, Support Vector Machines, and Neural Networks are commonly used for classification tasks.

Other Supervised Machine Learning Algorithm

Supervised learning can be further divided into several different types, each with its own unique characteristics and applications. Here are some of the most common types of supervised learning algorithms:

- Linear Regression: Linear regression is a type of regression algorithm that is used to predict a continuous output value. It is one of the simplest and most widely used algorithms in supervised learning. In linear regression, the algorithm tries to find a linear relationship between the input features and the output value. The output value is predicted based on the weighted sum of the input features.
- Logistic Regression: Logistic regression is a type of classification algorithm that is used to predict a binary output variable. It is commonly used in machine learning applications where the output variable is either true or false, such as in fraud detection or spam filtering. In logistic regression, the algorithm tries to find a linear relationship between the input features and the output variable. The output variable is then transformed using a logistic function to produce a probability value between 0 and 1.
- Decision Trees: Decision tree is a tree-like structure that is used to model decisions and their possible consequences. Each internal node in the tree represents a decision, while each leaf node represents a possible outcome. Decision trees can be used to model complex relationships between input features and output variables.A decision tree is a type of algorithm that is used for both classification and regression tasks.
    - Decision Trees Regression: Decision Trees can be utilized for regression tasks by predicting the value linked with a leaf node.
    - Decision Trees Classification: Random Forest is a machine learning algorithm that uses multiple decision trees to improve classification and prevent overfitting.
- Random Forests: Random forests are made up of multiple decision trees that work together to make predictions. Each tree in the forest is trained on a different subset of the input features and data. The final prediction is made by aggregating the predictions of all the trees in the forest.

  Random forests are an ensemble learning technique that is used for both classification and regression tasks.
    - Random Forest Regression: It combines multiple decision trees to reduce overfitting and improve prediction accuracy.
    - Random Forest Classifier: Combines several decision trees to improve the accuracy of classification while minimizing overfitting.
- Support Vector Machine(SVM): The SVM algorithm creates a hyperplane to segregate n-dimensional space into classes and identify the correct category of new data points. The extreme cases that help create the hyperplane are called support vectors, hence the name Support Vector Machine.

  A Support Vector Machine is a type of algorithm that is used for both classification and regression tasks

- Support Vector Regression**:** It is a extension of Support Vector Machines (SVM) used for predicting continuous values.
- Support Vector Classifier**:** It aims to find the best hyperplane that maximizes the margin between data points of different classes.

- K-Nearest Neighbors (KNN)**:** KNN works by finding k training examples closest to a given input and then predicts the class or value based on the majority class or average value of these neighbors. The performance of KNN can be influenced by the choice of k and the distance metric used to measure proximity. However, it is intuitive but can be sensitive to noisy data and requires careful selection of k for optimal results. A K-Nearest Neighbors (KNN) is a type of algorithm that is used for both classification and regression tasks.
  - K-Nearest Neighbors Regression: It predicts continuous values by averaging the outputs of the k closest neighbors.
  - K-Nearest Neighbors Classification**:** Data points are classified based on the majority class of their k closest neighbors.
- Gradient Boosting**:** Gradient Boosting combines weak learners, like decision trees, to create a strong model. It iteratively builds new models that correct errors made by previous ones. Each new model is trained to minimize residual errors, resulting in a powerful predictor capable of handling complex data relationships. A Gradient Boosting is a type of algorithm that is used for both classification and regression tasks.
  - Gradient Boosting Regression**:** It builds an ensemble of weak learners to improve prediction accuracy through iterative training.
  - Gradient Boosting Classification**:** Creates a group of classifiers to continually enhance the accuracy of predictions through iterations

Advantages of Supervised Learning

1. Labeled training data benefits supervised learning by enabling models to accurately learn patterns and relationships between inputs and outputs.
2. Supervised learning models can accurately predict and classify new data.
3. Supervised learning has a wide range of applications, including classification, regression, and even more complex problems like image recognition and natural language processing.
4. Well-established evaluation metrics, including accuracy, precision, recall, and F1-score, facilitate the assessment of supervised learning model performance.

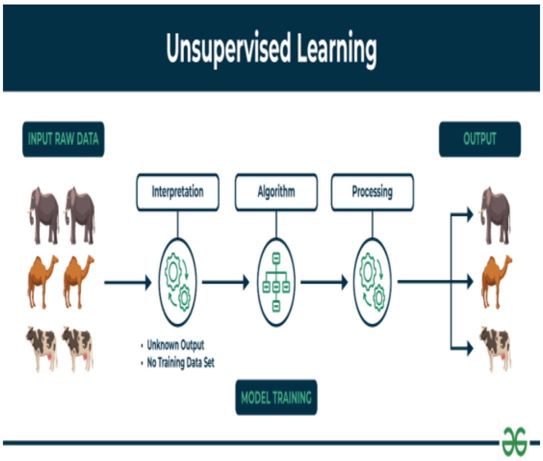
Disadvantages of Supervised Learning

. 1.Overfitting: Models can overfit training data, which leads to poor performance on new, unseen data due to the capture of noise.

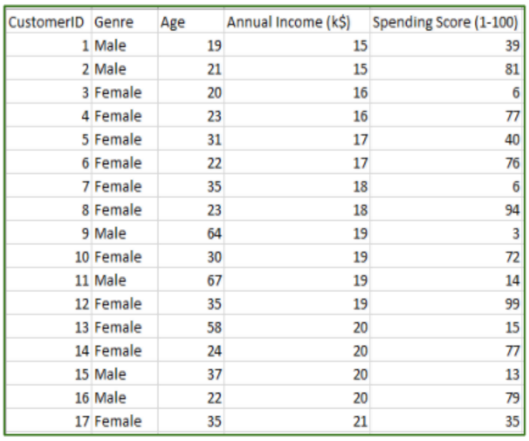
2. Feature Engineering: Extracting relevant features from raw data is crucial for model performance, but this process can be time-consuming and may require domain expertise.
3. Bias in Models: Training data biases can lead to unfair predictions.
4. Supervised learning heavily depends on labeled training data, which can be costly, time-consuming, and may require domain expertise.

6.Unsupervised Learning

Unsupervised learning is a branch of machine learning that deals with unlabeled data. Unlike supervised learning, where the data is labeled with a specific category or outcome, unsupervised learning algorithms are tasked with finding patterns and relationships within the data without any prior knowledge of the data's meaning. In artificial intelligence, machine learning that takes place in the absence of human supervision is known as unsupervised machine learning. Unsupervised machine learning models, in contrast to supervised learning, are given unlabeled data and allow discover patterns and insights on their own—without explicit direction or instruction. Unsupervised machine learning analyzes and clusters unlabeled datasets using machine learning algorithms. These algorithms find hidden patterns and data without any human intervention, i.e., we don't give output to our model. The training model has only input parameter values and discovers the groups or patterns on its own.



Unsupervised learning works by analyzing unlabeled data to identify patterns and relationships. The data is not labeled with any predefined categories or outcomes, so the algorithm must find these patterns and relationships on its own. This can be a challenging task, but it can also be very rewarding, as it can reveal insights into the data that would not be apparent from a labeled dataset.Data-set in Figure A is Mall data that contains information about its clients that subscribe to them. Once subscribed they are provided a membership card and the mall has complete information about the customer and his/her every purchase. Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.



CustomerID	Genre	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35

Figure A

The input to the unsupervised learning models is as follows:

- Unstructured data**:** May contain noisy(meaningless) data, missing values, or unknown data
- Unlabeled data: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.

There are mainly 3 types of Algorithms which are used for Unsupervised dataset.

- Clustering
- Association Rule Learning
- Dimensionality Reduction

Clustering in unsupervised machine learning is the process of grouping unlabeled data into clusters based on their similarities. The goal of clustering is to identify patterns and relationships in the data without any prior knowledge of the data's meaning. Broadly this technique is applied to group data based on different patterns, such as similarities or differences, our machine model finds. These algorithms are used to process raw, unclassified data objects into groups. For example, in the above figure, we have not given output parameter values, so this technique will be used to group clients based on the input parameters provided by our data.

Some common clustering algorithms

- K-means Clustering: Partitioning Data into K Clusters
- Hierarchical Clustering: Building a Hierarchical Structure of Clusters
- Density-Based Clustering (DBSCAN): Identifying Clusters Based on Density
- Mean-Shift Clustering: Finding Clusters Based on Mode Seeking
- Spectral Clustering: Utilizing Spectral Graph Theory for Clustering

Association rule learning is also known as association rule mining is a common technique used to discover associations in unsupervised machine learning. This technique is a rule-based ML technique that finds out some very useful relations between parameters of a large data set. This technique is basically used for market basket analysis that helps to better understand the relationship between different products. For e.g. shopping stores use algorithms based on this technique to find out the relationship between the sale of one product w.r.t to another's sales based on customer behavior. Like if a customer buys milk, then he may also buy bread, eggs, or butter. Once trained well, such models can be used to increase their sales by planning different offers.

- Apriori Algorithm: A Classic Method for Rule Induction
- FP-Growth Algorithm: An Efficient Alternative to Apriori
- Eclat Algorithm: Exploiting Closed Itemsets for Efficient Rule Mining
- Efficient Tree-based Algorithms: Handling Large Datasets with Scalability

Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much information as possible. This technique is useful for improving the performance of machine learning algorithms and for data visualization. Examples of dimensionality reduction algorithms includeDimensionality reduction is the process of reducing the number of features in a dataset while preserving as much information as possible.

- Principal Component Analysis (PCA): Linear Transformation for Reduced Dimensions
- Linear Discriminant Analysis (LDA): Dimensionality Reduction for Discrimination
- Non-negative Matrix Factorization (NMF): Decomposing Data into Non-negative Components
- Locally Linear Embedding (LLE): Preserving Local Geometry in Reduced Dimensions
- Isomap: Capturing Global Relationships in Reduced Dimensions

Advantages of Unsupervised learning

- **No labeled data required:** Unlike supervised learning, unsupervised learning does not require labeled data, which can be expensive and time-consuming to collect.
- **Can uncover hidden patterns:** Unsupervised learning algorithms can identify patterns and relationships in data that may not be obvious to humans.
- **Can be used for a variety of tasks:** Unsupervised learning can be used for a variety of tasks, such as clustering, dimensionality reduction, and anomaly detection.
- **Can be used to explore new data:** Unsupervised learning can be used to explore new data and gain insights that may not be possible with other methods.

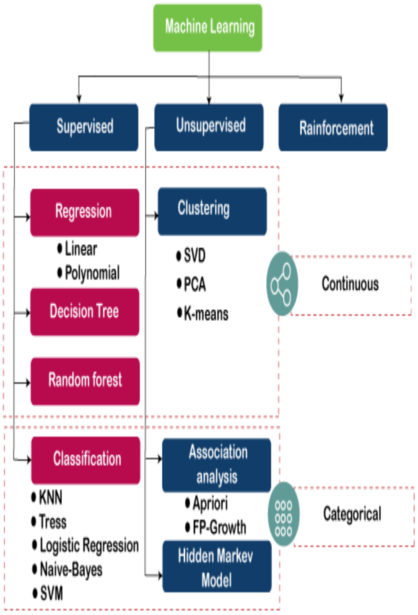
Disadvantages of Unsupervised learning

- **Difficult to evaluate:** It can be difficult to evaluate the performance of unsupervised learning algorithms, as there are no predefined labels or categories against which to compare results.
- **Can be difficult to interpret:** It can be difficult to understand the decision-making process of unsupervised learning models.

- **Can be sensitive to the quality of the data:** Unsupervised learning algorithms can be sensitive to the quality of the input data. Noisy or incomplete data can lead to misleading or inaccurate results.
- **Can be computationally expensive:** Some unsupervised learning algorithms, particularly those dealing with high-dimensional data or large datasets, can be computationally expensive

Applications of Unsupervised learning

- **Customer segmentation:** Unsupervised learning can be used to segment customers into groups based on their demographics, behavior, or preferences. This can help businesses to better understand their customers and target them with more relevant marketing campaigns.
- **Fraud detection:** Unsupervised learning can be used to detect fraud in financial data by identifying transactions that deviate from the expected patterns. This can help to prevent fraud by flagging these transactions for further investigation.
- **Recommendation systems:** Unsupervised learning can be used to recommend items to users based on their past behavior or preferences. For example, a recommendation system might use unsupervised learning to identify users who have similar taste in movies, and then recommend movies that those users have enjoyed.
- **Natural language processing (NLP):** Unsupervised learning is used in a variety of NLP tasks, including topic modeling, document clustering, and part-of-speech tagging.
- **Image analysis:** Unsupervised learning is used in a variety of image analysis tasks, including image segmentation, object detection, and image pattern recognition.



Supervised and unsupervised ML algorithms

7. Conclusion

The power of supervised learning lies in its ability to accurately predict patterns and make data-driven decisions across a variety of applications. Although supervised learning methods have benefits, their limitations require careful consideration during problem formulation, data collection, model selection, and evaluation. Advantages of Unsupervised learning are no labeled data required. Unlike supervised learning, unsupervised learning does not require labeled data, which can be expensive and time-consuming to collect. It Can uncover hidden patterns. Unsupervised learning algorithms can identify patterns and relationships in data that may not be obvious to humans. Unsupervised learning can be used for a variety of tasks such as clustering, dimensionality reduction, and anomaly detection. Unsupervised learning can be used to explore new data and gain insights that may not be possible with other methods. The key challenges of unsupervised learning are Evaluation, Interpretability, Overfitting, It is concluded that performance of unsupervised ML algorithms is better than supervised ML algorithms.

References

1.Ankerst M, Breunig MM, Kriegel H-P, Sander J. Optics: ordering points to identify the clustering structure. ACM Sigmod Record. 1999;28(2):49–60.
2.Anzai Y. Pattern recognition and machine learning. Elsevier; 2012.

3.. Ardabili SF, Mosavi A, Ghamisi P, Ferdinand F, Varkonyi-Koczy AR, Reuter U, Rabczuk T, Atkinson PM. Covid-19 outbreak prediction with machine learning. Algorithms. 2020;13(10):249.

4. Baldi P. Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, 2012; 37–49 .

5.Balducci F, Impedovo D, Pirlo G. Machine learning applications on agricultural datasets for smart farm enhancement. Machines. 2018;6(3):38.

6.Boukerche A, Wang J. Machine learning-based trafc prediction models for intelligent transportation systems. Comput Netw. 2020;181

7.Breiman L. Bagging predictors. Mach Lear n. 1996;24(2):123–40. 19. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

8.Breiman L, Friedman J, Stone CJ, Olshen RA. Classifcation and regression trees. CRC Press; 1984

9. Essien A, Petrounias I, Sampaio P, Sampaio S. A deep-learning model for urban trafc fow prediction with trafc events mined from twitter. In: World Wide Web, 2020: 1–24 .

10.Liu B, HsuW, Ma Y. Integrating classifcation and association rule mining. In: Proceedings of the fourth international conference on knowledge discovery and data mining, 1998.

11.Mahdavinejad MS, Rezvan M, Barekatain M, Adibi P, Barnaghi P, Sheth AP. Machine learning for internet of things data analysis: a survey. Digit Commun Netw. 2018;4(3):161–75.

12. Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016.

13. Fatima M, Pasha M, et al. Survey of machine learning algorithms for disease diagnostic. J Intell Learn Syst Appl. 2017;9(01):1

14. Quinlan JR. C4.5: programs for machine learning. Mach Learn. 1993

15.Safdar S, Zafar S, Zafar N, Khan NF. Machine learning based decision support systems (dss) for heart disease diagnosis: a review. Artif Intell Rev. 2018;50(4):597–623