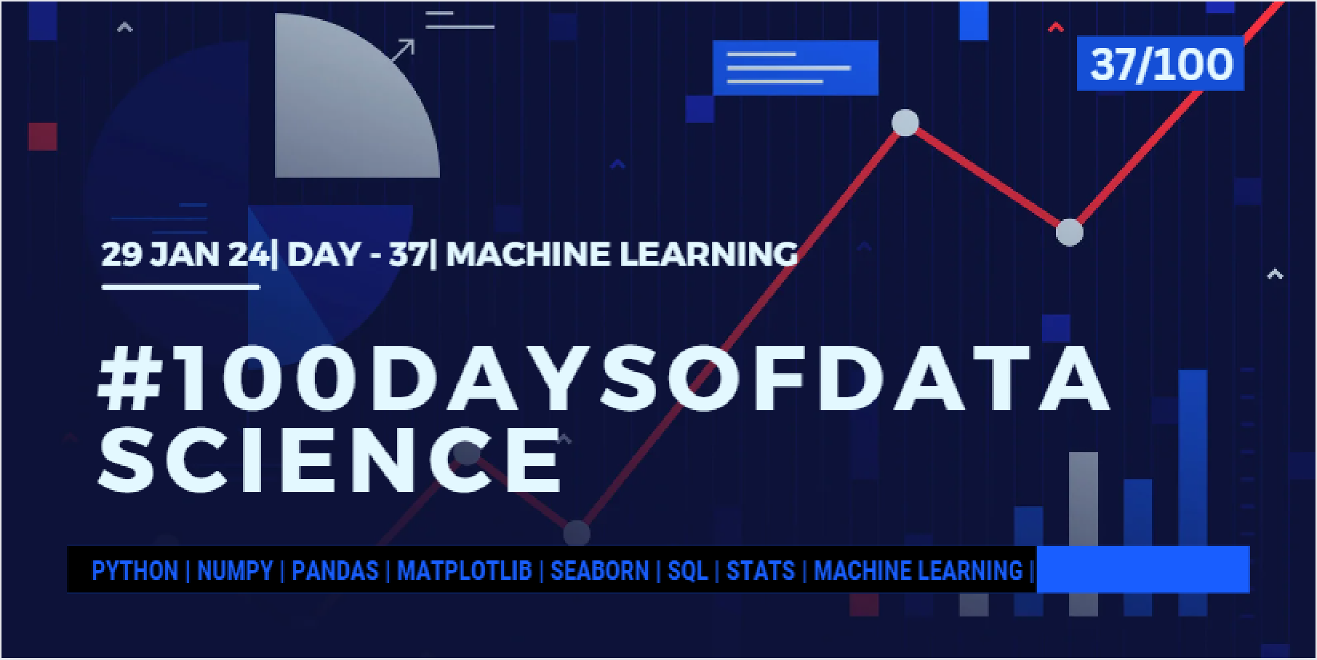


37/100

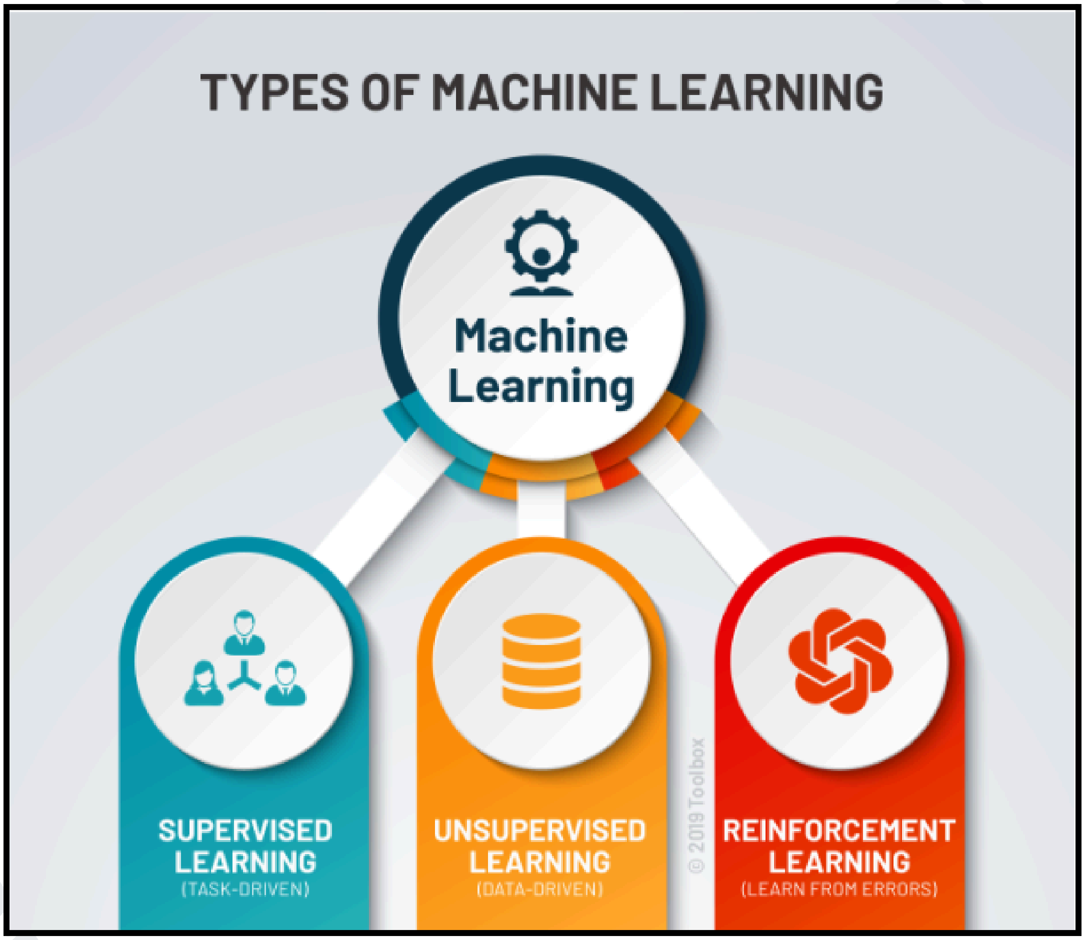
29 JAN 24| DAY - 37| MACHINE LEARNING

#100DAYSOFDATA SCIENCE

PYTHON | NUMPY | PANDAS | MATPLOTLIB | SEABORN | SQL | STATS | MACHINE LEARNING |

Types of Machine Learning

Machine learning is a subset of AI, which enables the machine to automatically learn from data, improve performance from past experiences, and make predictions. Machine learning contains a set of algorithms that work on a huge amount of data. Data is fed to these algorithms to train them, and on the basis of training, they build the model & perform a specific task.



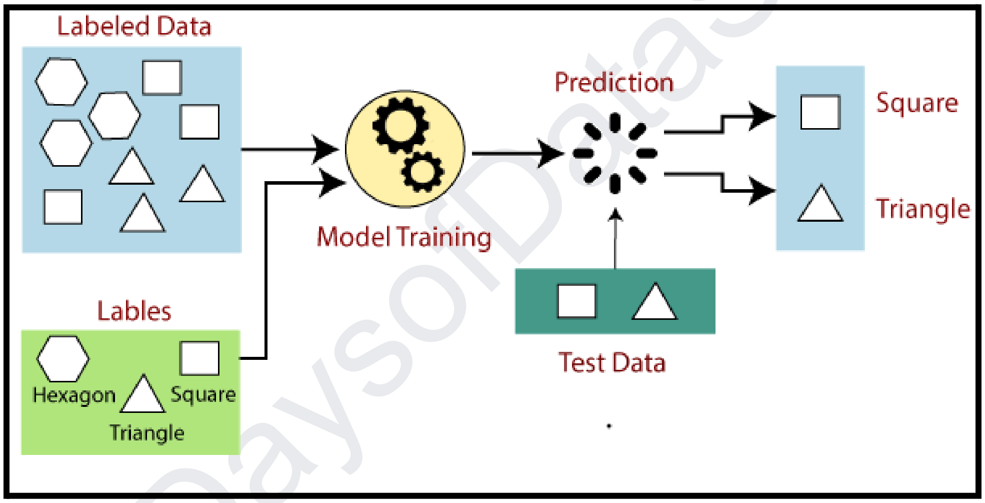
There are several types of machine learning, each with special characteristics and applications. Some of the main types of machine learning algorithms are as follows:

- Supervised Machine Learning
- Unsupervised Machine Learning
- Reinforcement Learning

. Supervised learning .

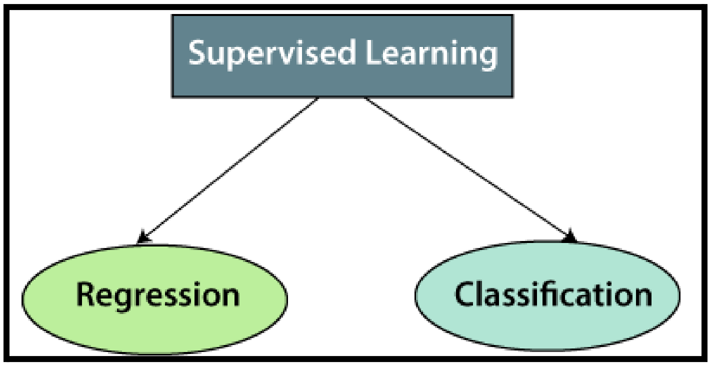
Supervised learning is the most common type of machine learning. In this approach, the model is trained on a labeled dataset. In other words, the data is accompanied by a label that the model is trying to predict. This could be anything from a category label to a real-valued number. The model learns a mapping between the input (features) and the output (label) during the training process. Once trained, the model can predict the output for new, unseen data.

For example, consider an input dataset of parrot and crow images. Initially, the machine is trained to understand the pictures, including the parrot and crow's color, eyes, shape, and size. Post-training, an input picture of a parrot is provided, and the machine is expected to identify the object and predict the output. The trained machine checks for the various features of the object, such as color, eyes, shape, etc., in the input picture, to make a final prediction. This is the process of object identification in supervised machine learning.



Supervised machine learning can be classified into two types of problems, which are given below:

- Classification
- Regression



a) Classification

Classification algorithms are used to solve the classification problems in which the output variable is categorical, such as "Yes" or No, Male or Female, Red or Blue, etc. The classification algorithms predict the categories present in the dataset. Some real-world examples of classification algorithms are Spam Detection, Email filtering, etc.

Some popular classification algorithms are given below:

- Random Forest Algorithm
- Decision Tree Algorithm
- Logistic Regression Algorithm
- Support Vector Machine Algorithm

b) Regression

Regression algorithms are used to solve regression problems in which there is a linear relationship between input and output variables. These are used to predict continuous output variables, such as market trends, weather prediction, etc.

Some popular Regression algorithms are given below:

- Simple Linear Regression Algorithm
- Multivariate Regression Algorithm
- Decision Tree Algorithm
- Lasso Regression

Advantages and **Disadvantages** of Supervised Learning

1. Advantages:

- Since supervised learning works with the labeled dataset so we can have an exact idea about the classes of objects.
- These algorithms are helpful in predicting the output on the basis of prior experience.

2. Disadvantages:

- These algorithms are not able to solve complex tasks.
- It may predict the wrong output if the test data is different from the training data.
- It requires lots of computational time to train the algorithm.

Applications of Supervised Learning

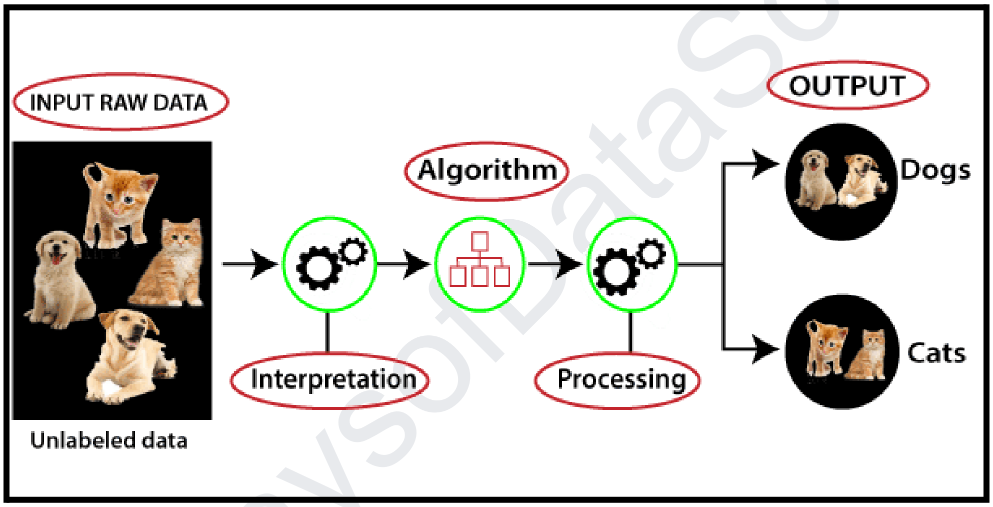
Some common applications of Supervised Learning are given below:

- Image Segmentation
- Medical Diagnosis
- Fraud Detection
- Spam detection
- Speech Recognition

. Unsupervised learning .

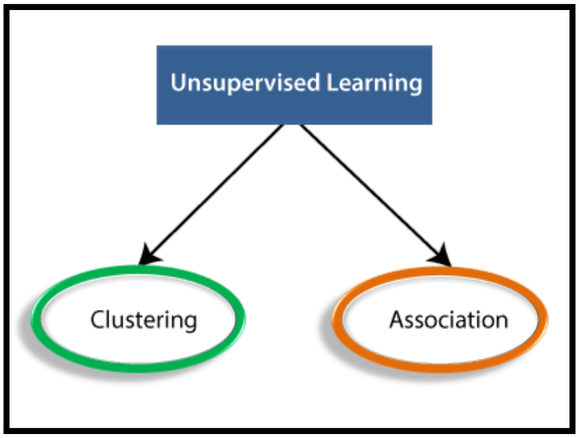
Unsupervised learning, on the other hand, involves training the model on an unlabeled dataset. The model is left to find patterns and relationships in the data on its own.
This type of learning is often used for clustering and dimensionality reduction. Clustering involves grouping similar data points together, while dimensionality reduction involves reducing the number of random variables under consideration by obtaining a set of principal variables.

For example, consider an input dataset of images of a fruit-filled container. Here, the images are not known to the machine learning model. When we input the dataset into the ML model, the task of the model is to identify the pattern of objects, such as color, shape, or differences seen in the input images and categorize them. Upon categorization, the machine then predicts the output as it gets tested with a test dataset.



There are two main categories of unsupervised learning that are mentioned below:

- Clustering
- Association



a) Clustering

Clustering is the process of grouping data points into clusters based on their similarity. This technique is useful for identifying patterns and relationships in data without the need for labeled examples.

Here are some clustering algorithms:

- K-Means Clustering algorithm
- DBSCAN Algorithm
- Principal Component Analysis

b) Association

Association rule learning is a technique for discovering relationships between items in a dataset. It identifies rules that indicate the presence of one item implies the presence of another item with a specific probability.

Here are some association rule learning algorithms:

- Apriori Algorithm
- FP-growth Algorithm

Advantages and **Disadvantages** of Unsupervised Learning

1. Advantages::

- It helps to discover hidden patterns and various relationships between the data.
- Used for tasks such as customer segmentation, anomaly detection, and data exploration.
- It does not require labeled data and reduces the effort of data labeling.

2. Disadvantages:

- Without using labels, it may be difficult to predict the quality of the model's output.
- Cluster Interpretability may not be clear and may not have meaningful interpretations.
- It has techniques such as autoencoders and dimensionality reduction that can be used to extract meaningful features from raw data.

Applications of Unsupervised Learning

Here are some common applications of unsupervised learning:

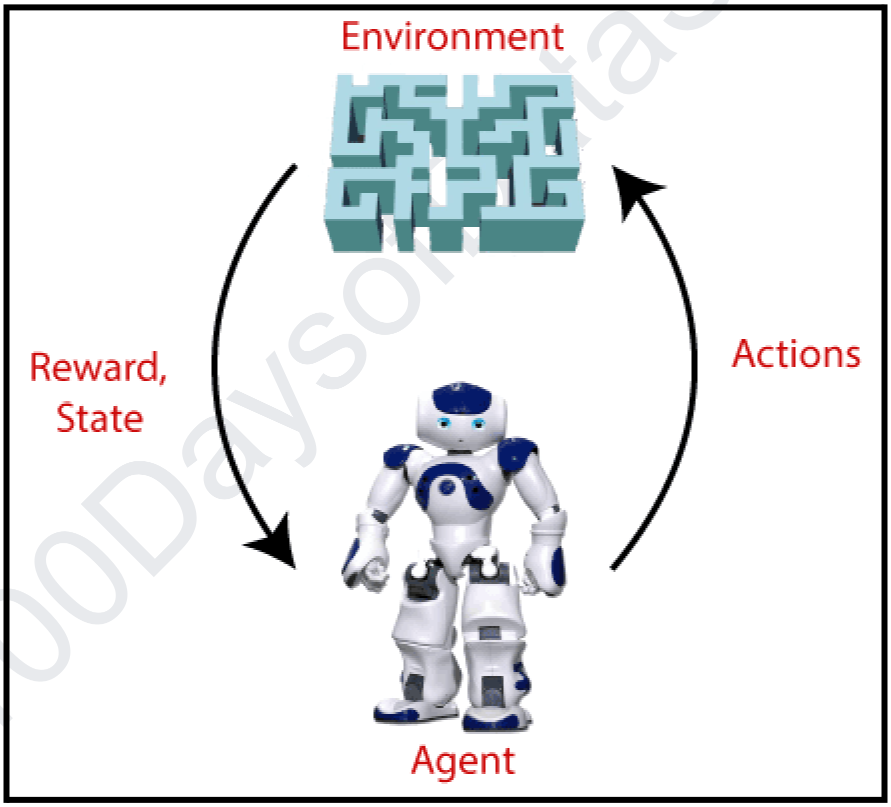
- Clustering
- Anomaly detection
- Dimensionality reduction
- Recommendation systems
- Topic modeling.
- Data preprocessing

. Reinforcement learning .

Reinforcement learning is a type of machine learning where an agent learns to make decisions by interacting with its environment. The agent is rewarded or penalized (with points) for the actions it takes, and its goal is to maximize the total reward.

Unlike supervised and unsupervised learning, reinforcement learning is particularly suited to problems where the data is sequential, and the decision made at each step can affect future outcomes.

Unlike supervised learning, reinforcement learning lacks labeled data, and the agents learn via experiences only. Consider video games. Here, the game specifies the environment, and each move of the reinforcement agent defines its state. The agent is entitled to receive feedback via punishment and rewards, thereby affecting the overall game score. The ultimate goal of the agent is to achieve a high score.



Reinforcement learning is categorized mainly into two types of methods/algorithms:

- Positive Reinforcement Learning
- Negative Reinforcement Learning:

a) Positive Reinforcement Learning

Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.

b) Negative Reinforcement Learning

Negative reinforcement learning works exactly opposite to the positive RL. It increases the tendency that the specific behaviour would occur again by avoiding the negative condition.

Some of the popular clustering algorithms are given below:

- Q-learning
- SARSA (State-Action-Reward-State-Action
- Deep Q-learning

Advantages and **Disadvantages** of Reinforcement Learning

Advantages:

- It helps in solving complex real-world problems which are difficult to be solved by general techniques.
- The learning model of RL is similar to the learning of human beings; hence most accurate results can be found.
- Helps in achieving long term results.

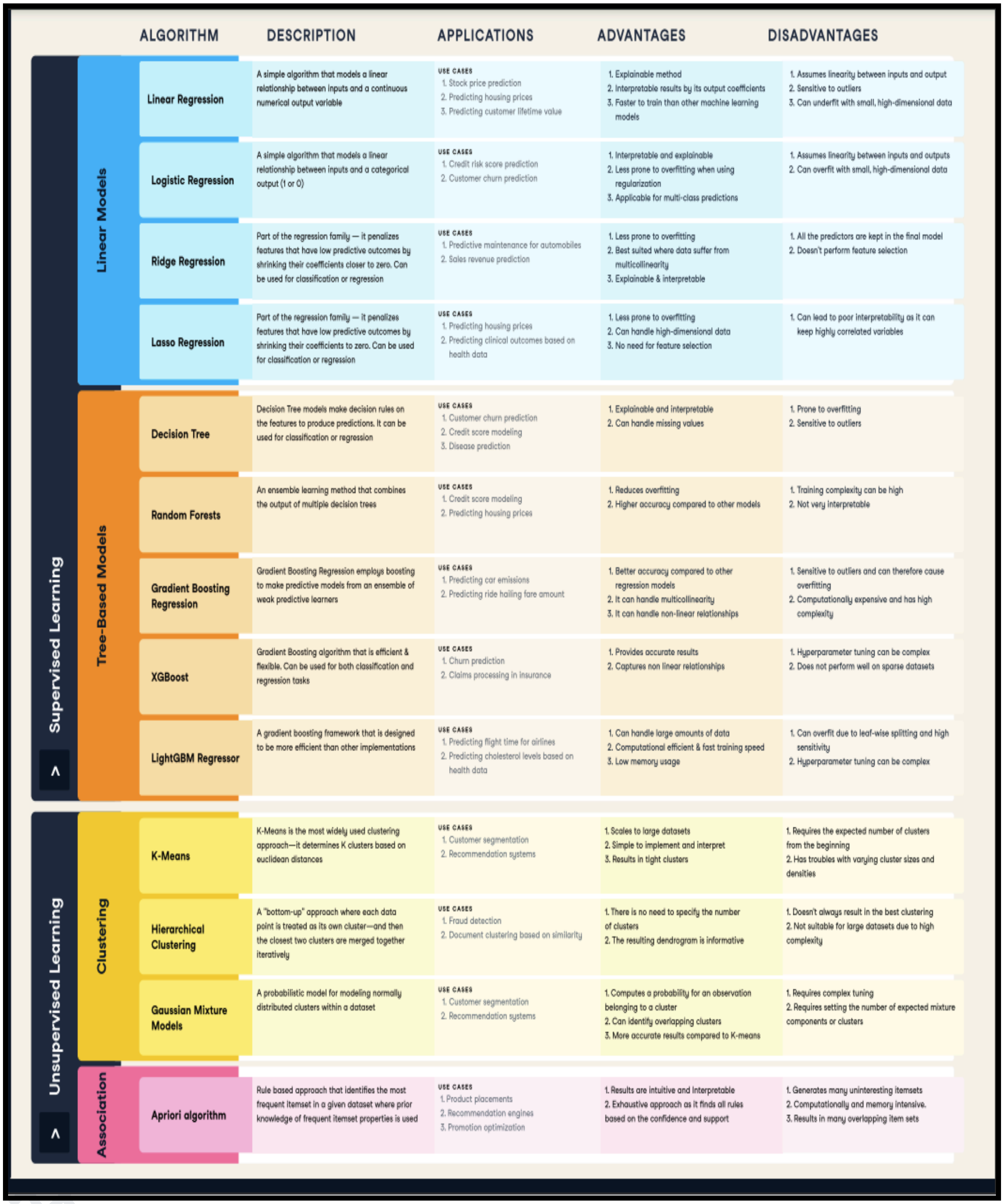
Disadvantages:

- RL algorithms are not preferred for simple problems.
- RL algorithms require huge data and computations.
- Too much reinforcement learning can lead to an overload of states which can weaken the results.

Applications of Reinforcement Learning

Here are some common applications of Reinforcement Learning

- Video Games
- Resource Management
- Robotics
- Text Mining



		ALGORITHM	DESCRIPTION	APPLICATIONS	ADVANTAGES	DISADVANTAGES
Supervised Learning ^	Linear Models	Linear Regression	A simple algorithm that models a linear relationship between inputs and a continuous numerical output variable	USE CASES 1. Stock price prediction 2. Predicting housing prices 3. Predicting customer lifetime value	1. Explainable method 2. Interpretable results by its output coefficients 3. Faster to train than other machine learning models	1. Assumes linearity between inputs and output 2. Sensitive to outliers 3. Can underfit with small, high-dimensional data
		Logistic Regression	A simple algorithm that models a linear relationship between inputs and a categorical output (1 or 0)	USE CASES 1. Credit risk score prediction 2. Customer churn prediction	1. Interpretable and explainable 2. Less prone to overfitting when using regularization 3. Applicable for multi-class predictions	1. Assumes linearity between inputs and outputs 2. Can overfit with small, high-dimensional data
		Ridge Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients closer to zero. Can be used for classification or regression	USE CASES 1. Predictive maintenance for automobiles 2. Sales revenue prediction	1. Less prone to overfitting 2. Best suited where data suffer from multicollinearity 3. Explainable & interpretable	1. All the predictors are kept in the final model 2. Doesn't perform feature selection
		Lasso Regression	Part of the regression family — it penalizes features that have low predictive outcomes by shrinking their coefficients to zero. Can be used for classification or regression	USE CASES 1. Predicting housing prices 2. Predicting clinical outcomes based on health data	1. Less prone to overfitting 2. Can handle high-dimensional data 3. No need for feature selection	1. Can lead to poor interpretability as it can keep highly correlated variables
	Tree-Based Models	Decision Tree	Decision Tree models make decision rules on the features to produce predictions. It can be used for classification or regression	USE CASES 1. Customer churn prediction 2. Credit score modeling 3. Disease prediction	1. Explainable and interpretable 2. Can handle missing values	1. Prone to overfitting 2. Sensitive to outliers
		Random Forests	An ensemble learning method that combines the output of multiple decision trees	USE CASES 1. Credit score modeling 2. Predicting housing prices	1. Reduces overfitting 2. Higher accuracy compared to other models	1. Training complexity can be high 2. Not very interpretable
		Gradient Boosting Regression	Gradient Boosting Regression employs boosting to make predictive models from an ensemble of weak predictive learners	USE CASES 1. Predicting car emissions 2. Predicting ride hailing fare amount	1. Better accuracy compared to other regression models 2. It can handle multicollinearity 3. It can handle non-linear relationships	1. Sensitive to outliers and can therefore cause overfitting 2. Computationally expensive and has high complexity
		XGBoost	Gradient Boosting algorithm that is efficient & flexible. Can be used for both classification and regression tasks	USE CASES 1. Churn prediction 2. Claims processing in insurance	1. Provides accurate results 2. Captures non linear relationships	1. Hyperparameter tuning can be complex 2. Does not perform well on sparse datasets
		LightGBM Regressor	A gradient boosting framework that is designed to be more efficient than other implementations	USE CASES 1. Predicting flight time for airlines 2. Predicting cholesterol levels based on health data	1. Can handle large amounts of data 2. Computational efficient & fast training speed 3. Low memory usage	1. Can overfit due to leaf-wise splitting and high sensitivity 2. Hyperparameter tuning can be complex
Unsupervised Learning ^	Clustering	K-Means	K-Means is the most widely used clustering approach—it determines K clusters based on euclidean distances	USE CASES 1. Customer segmentation 2. Recommendation systems	1. Scales to large datasets 2. Simple to implement and interpret 3. Results in tight clusters	1. Requires the expected number of clusters from the beginning 2. Has troubles with varying cluster sizes and densities
		Hierarchical Clustering	A "bottom-up" approach where each data point is treated as its own cluster—and then the closest two clusters are merged together iteratively	USE CASES 1. Fraud detection 2. Document clustering based on similarity	1. There is no need to specify the number of clusters 2. The resulting dendrogram is informative	1. Doesn't always result in the best clustering 2. Not suitable for large datasets due to high complexity
		Gaussian Mixture Models	A probabilistic model for modeling normally distributed clusters within a dataset	USE CASES 1. Customer segmentation 2. Recommendation systems	1. Computes a probability for an observation belonging to a cluster 2. Can identify overlapping clusters 3. More accurate results compared to K-means	1. Requires complex tuning 2. Requires setting the number of expected mixture components or clusters
	Association	Apriori algorithm	Rule based approach that identifies the most frequent itemset in a given dataset where prior knowledge of frequent itemset properties is used	USE CASES 1. Product placements 2. Recommendation engines 3. Promotion optimization	1. Results are intuitive and Interpretable 2. Exhaustive approach as it finds all rules based on the confidence and support	1. Generates many uninteresting itemsets 2. Computationally and memory intensive. 3. Results in many overlapping item sets