

# ITDO6014

## AI AND DS-1

Module 6: Introduction to ML

# Introduction to Machine Learning

2

- Introduction to Machine Learning
- Machine learning is a field of artificial intelligence (AI) where computers are programmed to learn from data and improve their performance on a specific task without being explicitly programmed to do so. In traditional programming, humans write explicit instructions for a computer to follow. However, in machine learning, instead of coding explicit rules, we feed the computer lots of data and let it figure out the patterns on its own.
- In simpler terms, machine learning is like teaching a computer to learn from examples. Just like how we learn from experience, machines learn from data. They use algorithms to analyze data, find patterns, and make decisions or predictions based on that analysis.

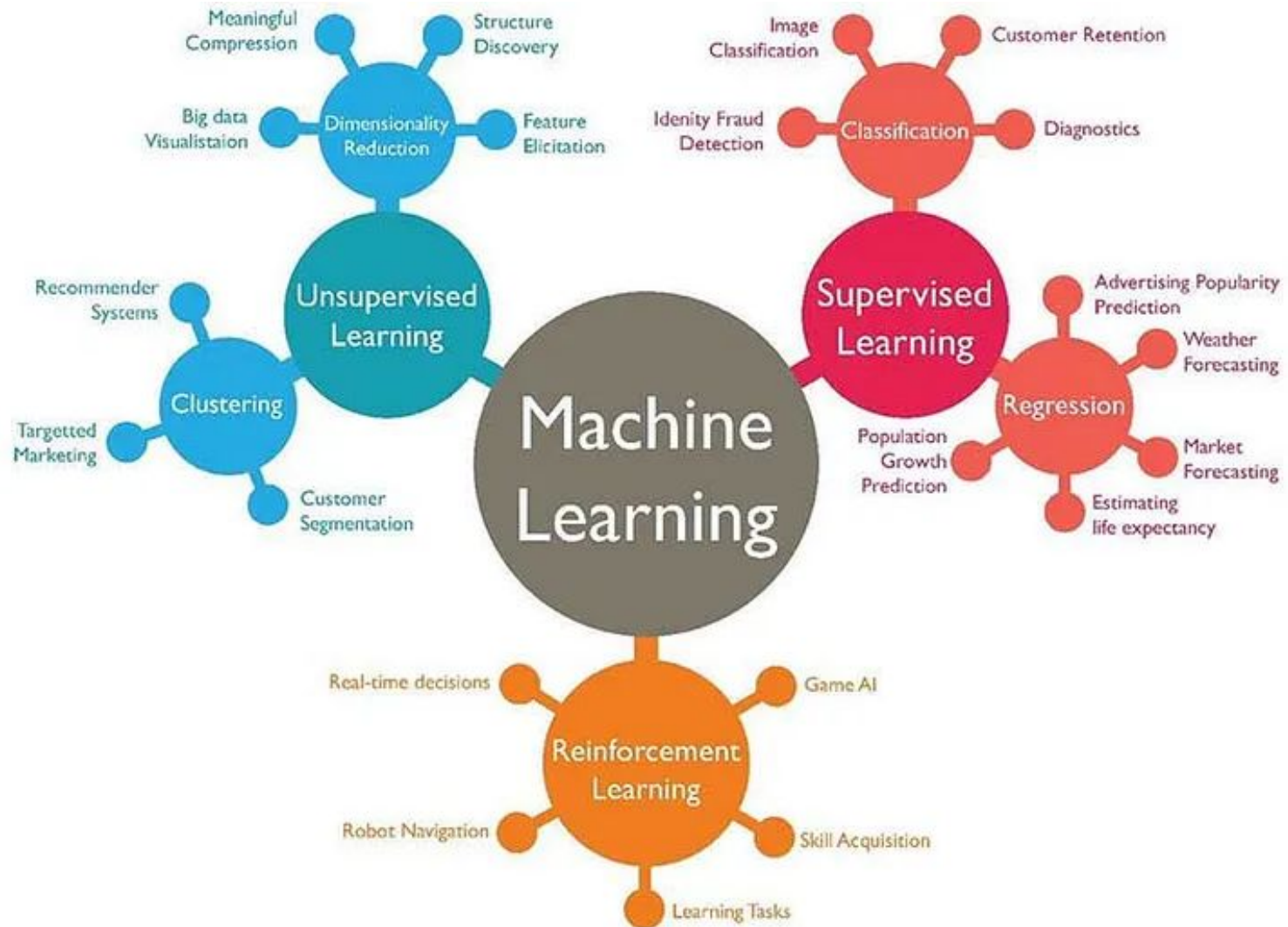
# Types of Machine Learning

3

- Machine learning can be classified into 3 types of algorithms.
- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

# Types of Machine Learning

4



# Issues in Machine Learning Algorithm

5

- Issues in Machine learning:
- **Data Quality and Quantity:** The performance of machine learning models heavily depends on the quality and quantity of the data used for training. Issues such as missing data, noisy data, biased data, and unrepresentative samples can significantly impact the performance and generalization ability of models.
- **Overfitting and Underfitting:** Overfitting occurs when a model learns to memorize the training data instead of generalizing from it, while underfitting occurs when a model is too simplistic to capture the underlying patterns in the data. Balancing between the two is crucial for developing models that perform well on unseen data.

# Issues in Machine Learning Algorithm

6

- Issues in Machine learning:
- **Interpretability and Explainability:** Many machine learning models, especially deep learning models, are often considered "black boxes" because they lack interpretability and explainability. Understanding how and why a model makes certain predictions is crucial, especially in sensitive domains like healthcare and finance.
- **Bias and Fairness:** Machine learning models can inherit biases present in the data, leading to unfair or discriminatory outcomes, especially towards certain demographic groups. Addressing biases and ensuring fairness in machine learning algorithms is essential for ethical and equitable deployment.

# Issues in Machine Learning Algorithm

7

- Issues in Machine learning:
- **Scalability and Efficiency:** As datasets and model complexities increase, scalability and efficiency become significant concerns. Developing algorithms that can handle large-scale datasets and run efficiently on various hardware platforms is crucial for practical deployment in real-world applications.
- **Generalization to Unseen Data:** Machine learning models should generalize well to unseen data to be useful in real-world scenarios. Ensuring that models can perform well on data distributions different from the training distribution is a challenging problem, especially in domains where data is constantly evolving.
- **Security and Privacy:** Machine learning models are vulnerable to various security threats, such as adversarial attacks, data poisoning, and model stealing. Moreover, the use of sensitive data in training models raises privacy concerns, necessitating techniques for privacy-preserving machine learning.

# Issues in Machine Learning Algorithm

8

- Issues in Machine learning:
- **Reproducibility and Replicability:** Reproducing and replicating results reported in machine learning research can be challenging due to factors like differences in data, code, and hyperparameters. Ensuring reproducibility and replicability is essential for building trust in machine learning research.
- **Domain Adaptation and Transfer Learning:** Deploying machine learning models in new domains or tasks where labeled data is scarce can be difficult. Techniques like domain adaptation and transfer learning aim to transfer knowledge from a source domain with ample labeled data to a target domain with limited labeled data.
- **Ethical and Social Implications:** Machine learning systems can have profound ethical and social implications, such as exacerbating existing inequalities, infringing on privacy rights, and influencing decision-making processes. Addressing these ethical and social implications requires interdisciplinary collaboration and thoughtful consideration.



# Application of Machine Learning Algorithm

9

- Application of Machine Learning:
- **Healthcare:** Machine learning is used for medical image analysis, diagnosis prediction, personalized treatment recommendation, drug discovery, and genomic data analysis. It helps in early detection of diseases, identifying patterns in patient data, and improving healthcare outcomes.
- **Finance:** In finance, machine learning is applied for fraud detection, credit scoring, algorithmic trading, risk management, customer segmentation, and personalized financial advice. It helps financial institutions make data-driven decisions, optimize portfolios, and mitigate risks.
- **E-commerce and Retail:** Machine learning powers recommendation systems, demand forecasting, customer segmentation, pricing optimization, supply chain optimization, and sentiment analysis in e-commerce and retail. It enhances user experience, increases sales, and improves operational efficiency.

# Application of Machine Learning Algorithm

10

- Application of Machine Learning:
- **Natural Language Processing (NLP):** NLP applications include sentiment analysis, text summarization, machine translation, chatbots, question answering systems, and information extraction. NLP enables machines to understand, interpret, and generate human language, facilitating communication and interaction between humans and computers.
- **Autonomous Vehicles:** Machine learning algorithms enable autonomous vehicles to perceive the environment, make decisions, and navigate safely. It involves tasks such as object detection, lane detection, path planning, and real-time decision-making based on sensor data from cameras, LiDAR, and radar.
- **Manufacturing and Industry 4.0:** Machine learning is used for predictive maintenance, quality control, process optimization, supply chain management, and robotics in manufacturing and Industry 4.0. It helps reduce downtime, improve product quality, and optimize production processes.

# Application of Machine Learning Algorithm

11

- Application of Machine Learning:
- **Cybersecurity:** Machine learning techniques are employed for anomaly detection, intrusion detection, malware classification, and threat intelligence in cybersecurity. It helps detect and mitigate security threats in real-time by analyzing vast amounts of network and system data.
- **Environmental Monitoring:** Machine learning is used for environmental monitoring, including climate modeling, pollution detection, biodiversity assessment, and disaster prediction. It helps analyze complex environmental data and support decision-making for sustainable resource management and conservation efforts.
- **Marketing and Advertising:** Machine learning enables targeted advertising, customer segmentation, campaign optimization, churn prediction, and personalized marketing strategies. It helps businesses better understand consumer behavior, improve marketing ROI, and enhance customer engagement.

# Application of Machine Learning Algorithm

12

- Application of Machine Learning:
- **Education:** Machine learning is applied in educational technologies for adaptive learning, intelligent tutoring systems, learning analytics, and automated grading. It personalizes learning experiences, provides feedback to students and educators, and supports data-driven decision-making in education.

# □ Steps in developing a Machine Learning Application

13

- Steps in developing a Machine Learning Application:
- **Problem Definition:**
  - ❖ Define the problem you want to solve and determine if it's suitable for a machine learning approach.
  - ❖ Clearly specify the goals and objectives of the project.
  - ❖ Define the success metrics to measure the performance of the machine learning model.
- **Data Collection and Preparation:**
  - ❖ Identify relevant data sources that contain information necessary for solving the problem.
  - ❖ Collect and gather the data, ensuring it's of sufficient quantity and quality.
  - ❖ Preprocess the data by cleaning, transforming, and formatting it for analysis.
  - ❖ Split the data into training, validation, and test sets for model development and evaluation.

# □ Steps in developing a Machine Learning Application

14

- Steps in developing a Machine Learning Application:
- **Exploratory Data Analysis (EDA):**
  - ❖ Perform exploratory data analysis to gain insights into the data.
  - ❖ Visualize the data using plots, histograms, and summary statistics to understand its distribution, correlations, and patterns.
  - ❖ Identify potential outliers, missing values, and anomalies that may affect the performance of the model.
- **Feature Engineering:**
  - ❖ Select relevant features or variables that are predictive of the target variable.
  - ❖ Create new features through transformations, scaling, encoding categorical variables, and feature extraction.
  - ❖ Apply dimensionality reduction techniques if necessary to reduce the complexity of the feature space.

# □ Steps in developing a Machine Learning Application

15

- Steps in developing a Machine Learning Application:
- **Model Selection and Training:**
  - ❖ Choose appropriate machine learning algorithms based on the problem type, data characteristics, and performance requirements.
  - ❖ Train multiple models using the training data and evaluate their performance using the validation set.
  - ❖ Tune hyperparameters and optimize model performance using techniques like cross-validation and grid search.
- **Model Evaluation:**
  - ❖ Evaluate the trained models using the test dataset to assess their generalization performance.
  - ❖ Measure performance metrics such as accuracy, precision, recall, F1-score, ROC AUC, or others depending on the problem.
  - ❖ Perform error analysis to understand model shortcomings and areas for improvement.

## □ Steps in developing a Machine Learning Application

16

- Steps in developing a Machine Learning Application:
- **Model Deployment:**
  - ❖ Deploy the trained model into production or integrate it into the target application.
  - ❖ Choose an appropriate deployment environment, such as on-premises servers, cloud platforms, or edge devices.
  - ❖ Implement monitoring and logging mechanisms to track model performance and behavior in real-time.
  - ❖ Continuously update and retrain the model with new data to maintain its effectiveness over time.



## □ Steps in developing a Machine Learning Application

17

- Steps in developing a Machine Learning Application:
- **Maintenance and Iteration:**
  - ◆ Monitor the deployed model's performance and collect feedback from users.
  - ◆ Fine-tune and update the model periodically to adapt to changes in the data distribution or user requirements.
  - ◆ Iterate on the development process based on lessons learned and new insights gained from the deployed application.

# Machine Learning Algorithms

	<u>Unsupervised</u>	<u>Supervised</u>
<u>Continuous</u>	<ul style="list-style-type: none"><li>• Clustering &amp; Dimensionality Reduction<ul style="list-style-type: none"><li>○ SVD</li><li>○ PCA</li><li>○ K-means</li></ul></li></ul>	<ul style="list-style-type: none"><li>• Regression<ul style="list-style-type: none"><li>○ Linear</li><li>○ Polynomial</li></ul></li><li>• Decision Trees</li><li>• Random Forests</li></ul>
<u>Categorical</u>	<ul style="list-style-type: none"><li>• Association Analysis<ul style="list-style-type: none"><li>○ Apriori</li><li>○ FP-Growth</li></ul></li><li>• Hidden Markov Model</li></ul>	<ul style="list-style-type: none"><li>• Classification<ul style="list-style-type: none"><li>○ KNN</li><li>○ Trees</li><li>○ Logistic Regression</li><li>○ Naive-Bayes</li><li>○ SVM</li></ul></li></ul>

# Supervised Learning Algorithm

19

- Supervised learning is a type of machine learning where the algorithm learns a mapping from input data to output labels based on examples provided in a dataset. In supervised learning, the algorithm is trained on a labeled dataset, meaning each input data point is associated with a corresponding output label. The goal of supervised learning is to learn a mapping or relationship between input data and output labels so that the algorithm can make accurate predictions or decisions when given new, unseen data.
- There are two main types of supervised learning tasks:

Supervised learning problems can be further grouped into regression and classification problems.

- **Classification:** A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”. Sometimes these categories are represented by numbers but their value carries no meaning. They are just labels.
- **Regression:** A regression problem is when the output variable is a real number value, such as “dollars” or “weight”.

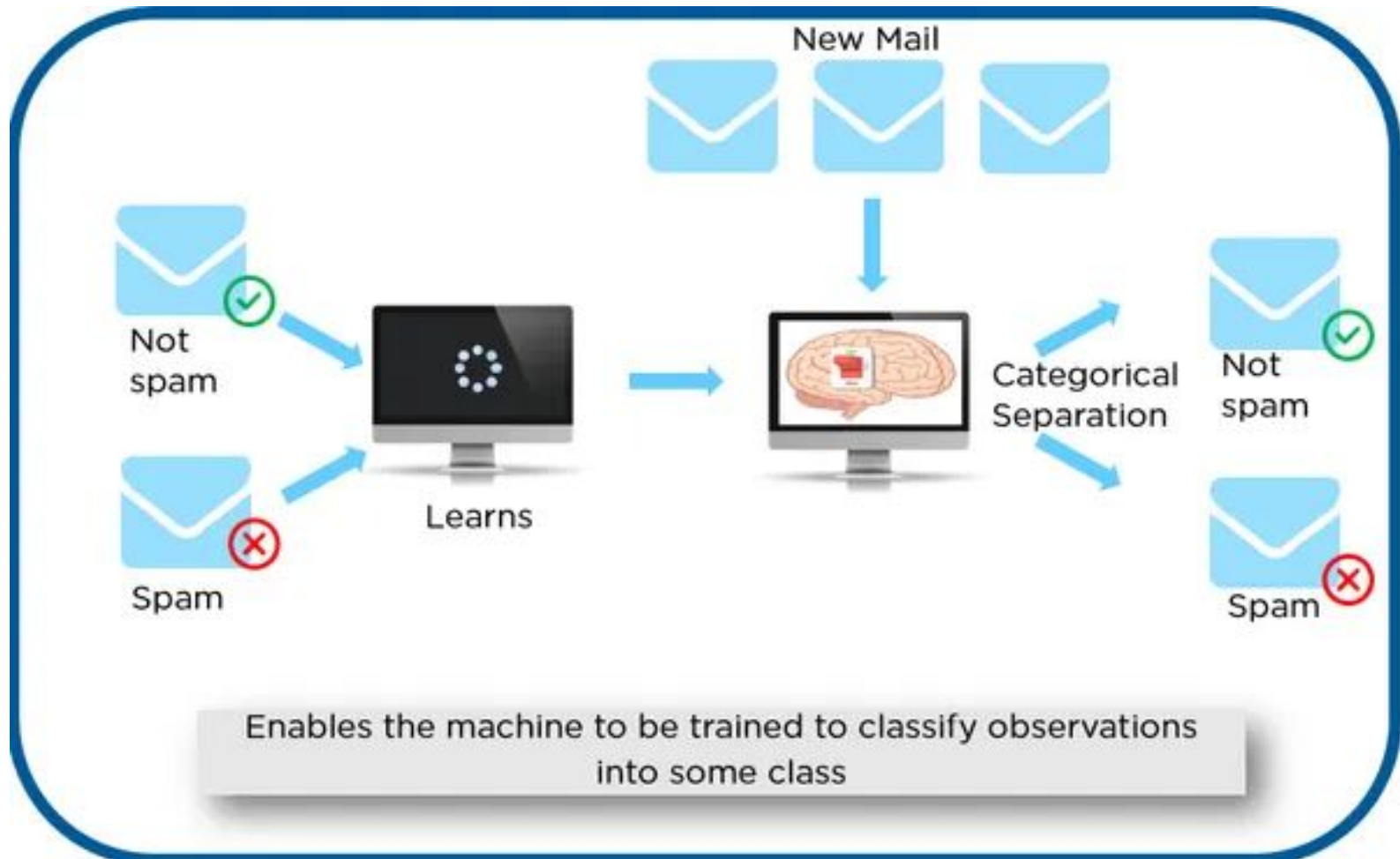
Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively.

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

# Supervised Learning Algorithm

21



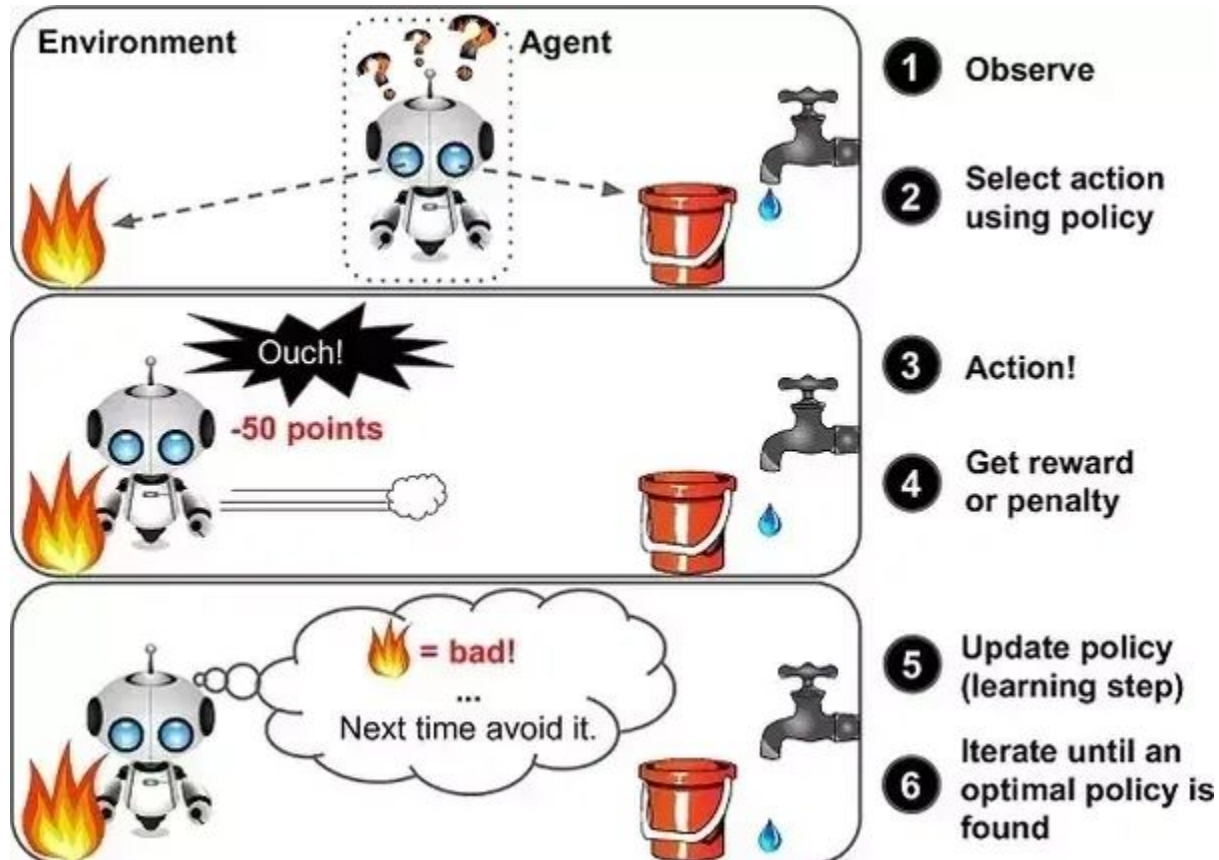
# Reinforcement Learning Algorithm

22

- A reinforcement learning algorithm, or agent, learns by interacting with its environment. The agent receives rewards by performing correctly and penalties for performing incorrectly. The agent learns without intervention from a human by maximizing its reward and minimizing its penalty. It is a type of dynamic programming that trains algorithms using a system of reward and punishment.

# Reinforcement Learning Algorithm

23



# Unsupervised Learning Algorithm

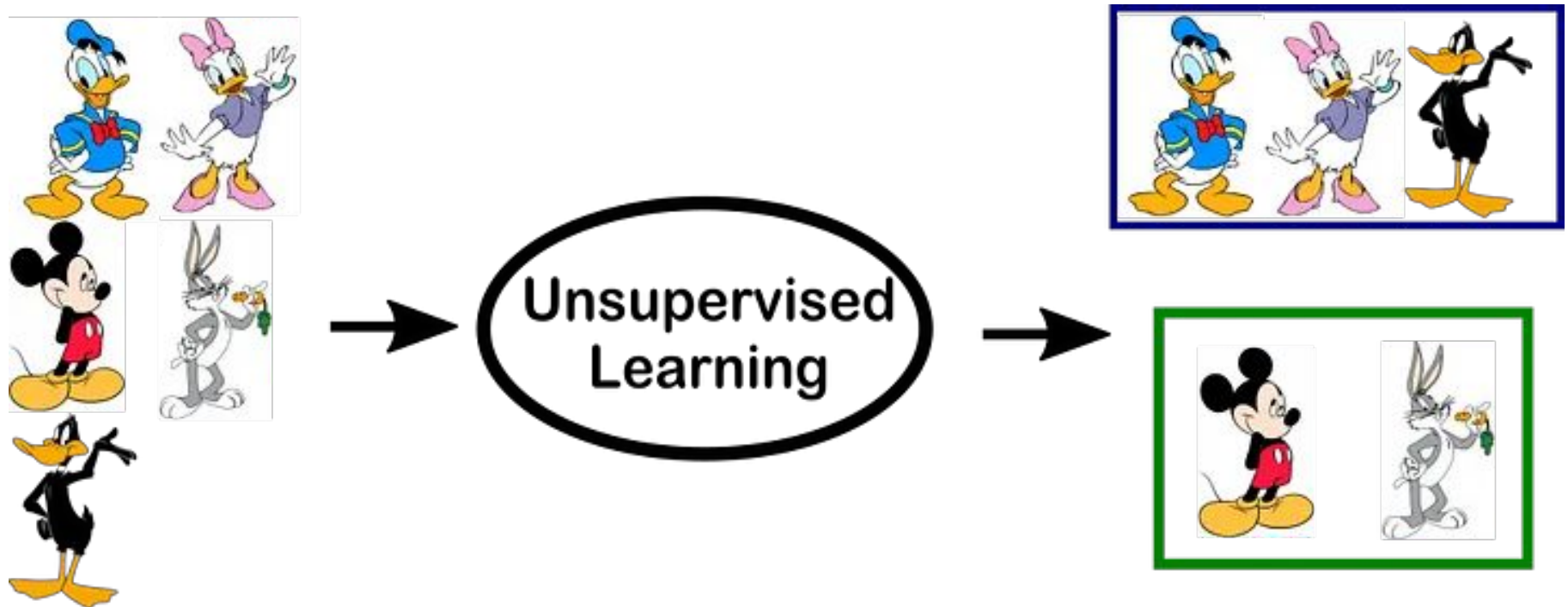
24

- Unsupervised learning is another type of machine learning where the algorithm learns patterns from unlabeled data, meaning the data does not have corresponding output labels. In unsupervised learning, the algorithm tries to find hidden structure or relationships in the input data without explicit guidance.
- The goal of unsupervised learning is often to discover the underlying structure of the data, such as grouping similar data points together or finding patterns in the data. Unlike supervised learning, where the algorithm learns to predict output labels based on input data, unsupervised learning focuses on understanding the inherent structure of the data itself.



# Unsupervised Learning Algorithm

25



# Unsupervised Learning Algorithm

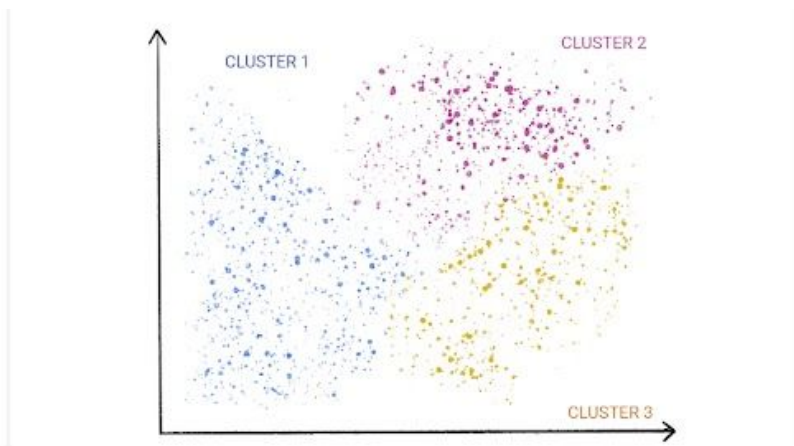
26

- Unsupervised learning algorithms are better suited for more complex processing tasks, such as organizing large datasets into clusters. They are useful for identifying previously undetected patterns in data and can help identify features useful for categorizing data.
- Imagine that you have a large dataset about weather. An unsupervised learning algorithm will go through the data and identify patterns in the data points. For instance, it might group data by temperature or similar weather patterns.

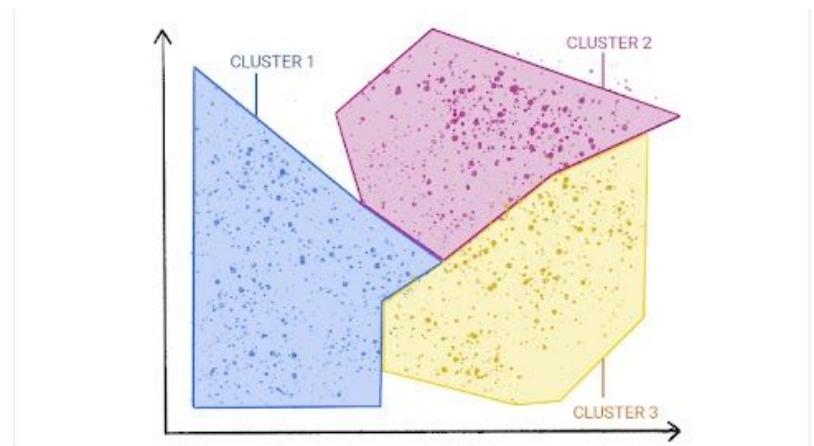
# Unsupervised Learning Algorithm

27

□ While the algorithm itself does not understand these patterns based on any previous information you provided, you can then go through the data groupings and attempt to classify them based on your understanding of the dataset. For instance, you might recognize that the different temperature groups represent all four seasons or that the weather patterns are separated into different types of weather, such as rain, sleet, or snow.



**Figure 1.** An ML model clustering similar data points.



**Figure 2.** Groups of clusters with natural demarcations.

# Unsupervised Learning Algorithm

28

- Unsupervised learning algorithms are better suited for more complex processing tasks, such as organizing large datasets into clusters. They are useful for identifying previously undetected patterns in data and can help identify features useful for categorizing data.
- Imagine that you have a large dataset about weather. An unsupervised learning algorithm will go through the data and identify patterns in the data points. For instance, it might group data by temperature or similar weather patterns.

# Unsupervised Learning Algorithm

29

- **Types of Unsupervised learning**
- Dimensionality reduction
- Dimensionality reduction is an unsupervised learning technique that reduces the number of features, or dimensions, in a dataset. More data is generally better for machine learning, but it can also make it more challenging to visualize the data.
- Dimensionality reduction extracts important features from the dataset, reducing the number of irrelevant or random features present. This method uses principle component analysis (PCA) and singular value decomposition (SVD) algorithms to reduce the number of data inputs without compromising the integrity of the properties in the original data.

# Unsupervised Learning Algorithm

30

- Real-world unsupervised learning examples
- **Anomaly detection:** Unsupervised clustering can process large datasets and discover data points that are atypical in a dataset.
- **Recommendation engines:** Using association rules, unsupervised machine learning can help explore transactional data to discover patterns or trends that can be used to drive personalized recommendations for online retailers.
- **Customer segmentation:** Unsupervised learning is also commonly used to generate buyer persona profiles by clustering customers' common traits or purchasing behaviors. These profiles can then be used to guide marketing and other business strategies.

# Unsupervised Learning Algorithm

31

- Real-world unsupervised learning examples
- **Fraud detection:** Unsupervised learning is useful for anomaly detection, revealing unusual data points in datasets. These insights can help uncover events or behaviors that deviate from normal patterns in the data, revealing fraudulent transactions or unusual behavior like bot activity.
- **Natural language processing (NLP):** Unsupervised learning is commonly used for various NLP applications, such as categorizing articles in news sections, text translation and classification, or speech recognition in conversational interfaces.
- **Genetic research:** Genetic clustering is another common unsupervised learning example. Hierarchical clustering algorithms are often used to analyze DNA patterns and reveal evolutionary relationships.

# Unsupervised Learning Algorithm

32

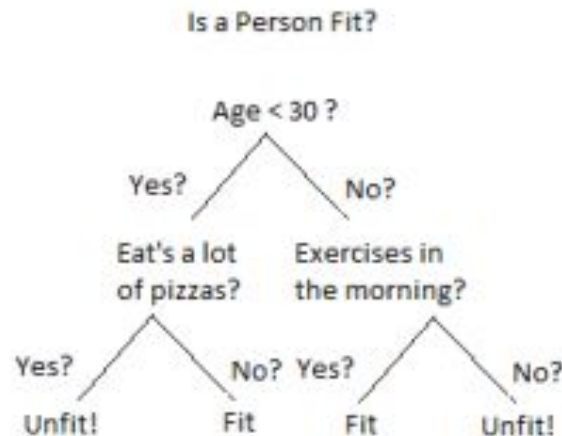
- Challenges of unsupervised learning:
- While unsupervised learning has many benefits, some challenges can occur when it allows machine learning models to execute without any human intervention. Some of these challenges can include:
- Computational complexity due to a high volume of training data
- Longer training times
- Higher risk of inaccurate results
- Human intervention to validate output variables
- Lack of transparency into the basis on which data was clustered



# Supervised Learning Algorithms

33

- 1. Decision Tree:
- Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split.



# Supervised Learning Algorithms

34

- ID3 Stands for **Iterative Dichotomiser 3**:
- Entropy:
- Entropy, also called as Shannon Entropy is denoted by  $H(S)$  for a finite set  $S$ , is the measure of the amount of uncertainty or randomness in data. Decision Trees modified

$$H(S) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)}$$

# Supervised Learning Algorithms

35

- Information Gain:
- Information gain is also called as Kullback-Leibler divergence denoted by  $IG(S,A)$  for a set  $S$  is the effective change in entropy after deciding on a particular attribute  $A$ . It measures the relative change in entropy with respect to the independent variables. Decision Trees modified  
Alternatively,

$$IG(S,A) = H(S) - \sum_{i=0}^n P(x) * H(x)$$

# Supervised Learning Algorithms

36

- Information Gain:
- Information gain is also called as Kullback-Leibler divergence denoted by  $IG(S,A)$  for a set  $S$  is the effective change in entropy after deciding on a particular attribute  $A$ . It measures the relative change in entropy with respect to the independent variables. Decision Trees modified Alternatively

$$IG(S,A) = H(S) - \sum_{i=0}^n P(x) * H(x)$$

- where  $IG(S, A)$  is the information gain by applying feature  $A$ .  $H(S)$  is the Entropy of the entire set, while the second term calculates the Entropy after applying the feature  $A$ , where  $P(x)$  is the probability of event  $x$ .

# Supervised Learning Algorithms

37

□ Example:

<https://www.xoriant.com/blog/decision-trees-for-classification-a-machine-learning-algorithm>

# Supervised Learning Algorithms

38

- 2. Random Forest Algorithm:
- Random Forest is a famous machine learning algorithm that uses supervised learning methods. You can apply it to both classification and regression problems. It is based on ensemble learning, which integrates multiple classifiers to solve a complex issue and increases the model's performance.
- In layman's terms, Random Forest is a classifier that contains several decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset. Instead of relying on a single decision tree, the random forest collects the result from each tree and expects the final output based on the majority votes of predictions.

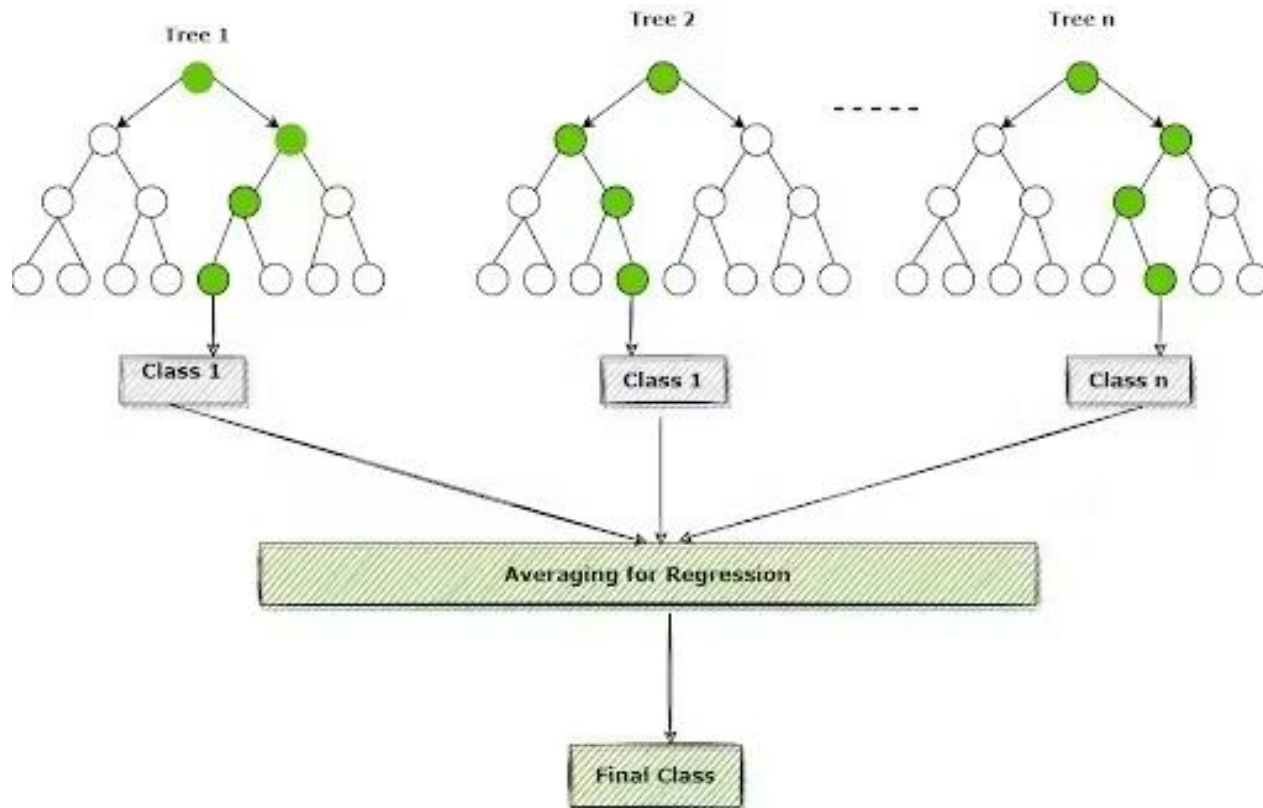
# Supervised Learning Algorithms

39

- The Working of the Random Forest Algorithm is quite intuitive. It is implemented in two phases: The first is to combine  $N$  decision trees with building the random forest, and the second is to make predictions for each tree created in the first phase.
- The following steps can be used to demonstrate the working process:
- Step 1: Pick  $M$  data points at random from the training set.
- Step 2: Create decision trees for your chosen data points (Subsets).
- Step 3: Each decision tree will produce a result. Analyze it.
- Step 4: For classification and regression, accordingly, the final output is based on Majority Voting or Averaging, accordingly.

# Supervised Learning Algorithms

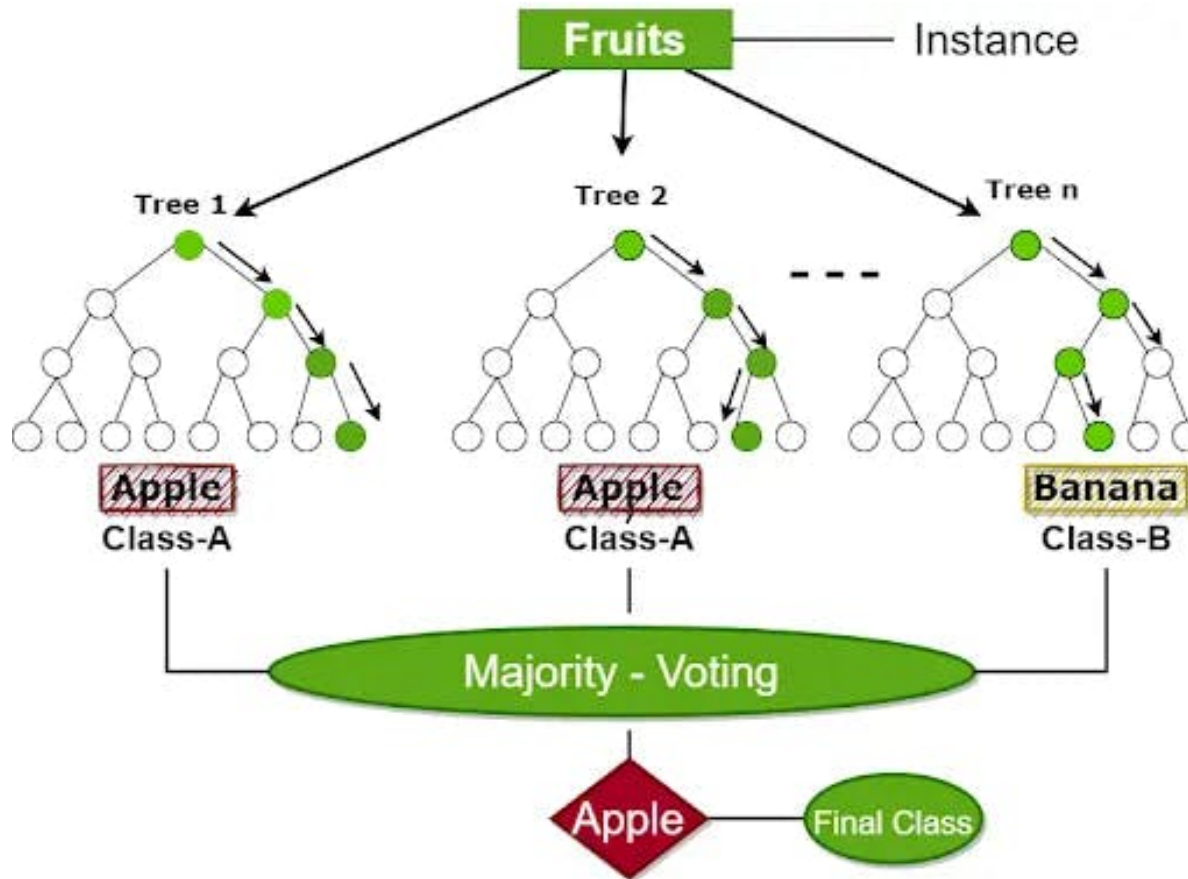
40





# Supervised Learning Algorithms

41



# Supervised Learning Algorithms

42

- 3. The K-Nearest Neighbors (K-NN) algorithm
- <https://www.freecodecamp.org/news/k-nearest-neighbors-algorithm-classifiers-and-model-example/>

# Supervised Learning Algorithms

43

- **4. Support Vector Machine Algorithm**
- <https://intellipaat.com/blog/tutorial/machine-learning-tutorial/svm-algorithm-in-python/>

# Unsupervised Learning Algorithm

44

- **Types of Unsupervised learning**
- Clustering
- Clustering is a technique for exploring raw, unlabeled data and breaking it down into groups (or clusters) based on similarities or differences. It is used in a variety of applications, including customer segmentation, fraud detection, and image analysis. Clustering algorithms split data into natural groups by finding similar structures or patterns in uncategorized data.
- Clustering is one of the most popular unsupervised machine learning approaches. There are several types of unsupervised learning algorithms that are used for clustering, which include exclusive, overlapping, hierarchical, and probabilistic.

# Unsupervised Learning Algorithm

45

- **Types of Unsupervised learning**
- **Exclusive clustering:** Data is grouped in a way where a single data point can only exist in one cluster. This is also referred to as “hard” clustering. A common example of exclusive clustering is the K-means clustering algorithm, which partitions data points into a user-defined number  $K$  of clusters.
- **Overlapping clustering:** Data is grouped in a way where a single data point can exist in two or more clusters with different degrees of membership. This is also referred to as “soft” clustering.

# Unsupervised Learning Algorithm

46

- **Types of Unsupervised learning**
- **K-means clustering** is a common example of an exclusive clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from each group's centroid. The data points closest to a given centroid will be clustered under the same category. A larger K value will be indicative of smaller groupings with more granularity whereas a smaller K value will have larger groupings and less granularity. K-means clustering is commonly used in market segmentation, document clustering, image segmentation, and image compression.
- <https://www.pinecone.io/learn/k-means-clustering/>
- <https://codinginfinite.com/k-means-clustering-explained-with-numeric-al-example/>

# Unsupervised Learning Algorithm

47

- **Types of Unsupervised learning**
- **Hierarchical clustering:** Data is divided into distinct clusters based on similarities, which are then repeatedly merged and organized based on their hierarchical relationships. There are two main types of hierarchical clustering: agglomerative and divisive clustering. This method is also referred to as HAC—hierarchical cluster analysis.
- **Probabilistic clustering:** Data is grouped into clusters based on the probability of each data point belonging to each cluster. This approach differs from the other methods, which group data points based on their similarities to others in a cluster.

# Unsupervised Learning Algorithm

48

- **Types of Unsupervised learning**
- Agglomerative clustering is considered a “bottoms-up approach.” Its data points are isolated as separate groupings initially, and then they are merged together iteratively on the basis of similarity until one cluster has been achieved. Four different methods are commonly used to measure similarity:
- **Ward’s linkage:** This method states that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
- **Average linkage:** This method is defined by the mean distance between two points in each cluster.



# Unsupervised Learning Algorithm

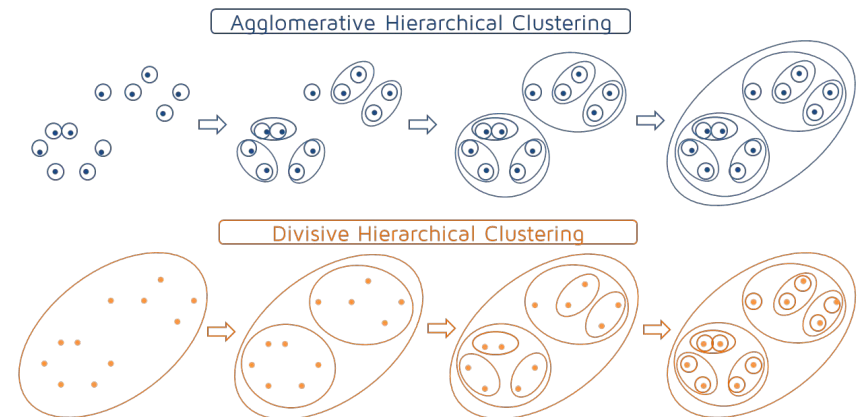
49

- **Types of Unsupervised learning**
- **Complete (or maximum) linkage:** This method is defined by the maximum distance between two points in each cluster.
- **Single (or minimum) linkage:** This method is defined by the minimum distance between two points in each cluster.
- Euclidean distance is the most common metric used to calculate these distances; however, other metrics, such as Manhattan distance, are also cited in clustering literature.

# Unsupervised Learning Algorithm

50

- **Types of Unsupervised learning**
- Divisive clustering can be defined as the opposite of agglomerative clustering; instead it takes a “top-down” approach. In this case, a single data cluster is divided based on the differences between data points. Divisive clustering is not commonly used, but it is still worth noting in the context of hierarchical clustering. These clustering processes are usually visualized using a dendrogram, a tree-like diagram that documents the merging or splitting of data points at each iteration.

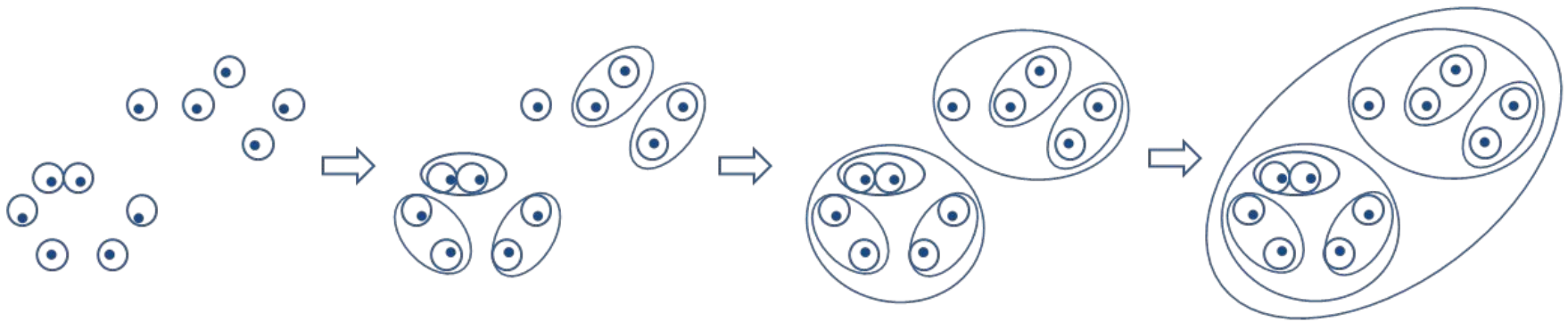


# Unsupervised Learning Algorithm

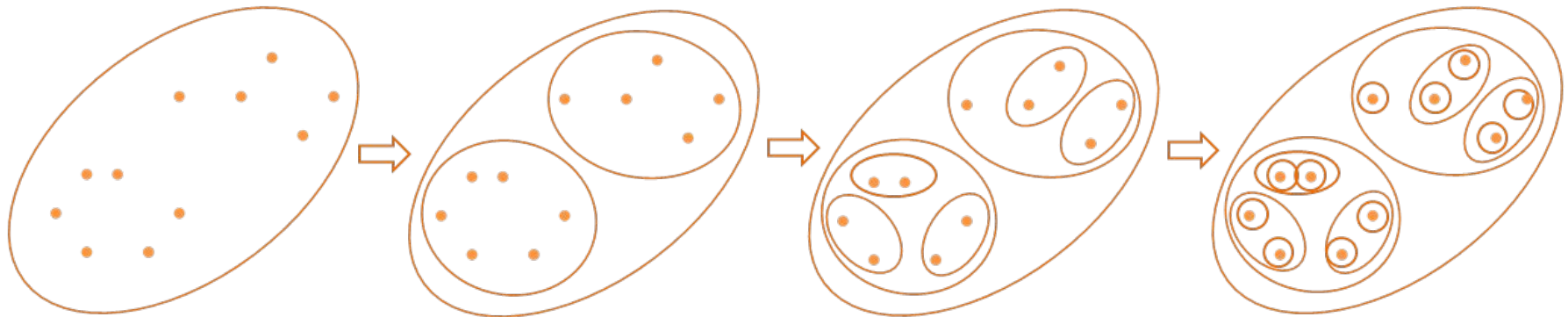
51

## Types of Unsupervised learning

Agglomerative Hierarchical Clustering



Divisive Hierarchical Clustering



# Unsupervised Learning Algorithm

52

- **Types of Unsupervised learning**
- Association
- Association rule mining is a rule-based approach to reveal interesting relationships between data points in large datasets. Unsupervised learning algorithms search for frequent if-then associations—also called rules—to discover correlations and co-occurrences within the data and the different connections between data objects.
- It is most commonly used to analyze retail baskets or transactional datasets to represent how often certain items are purchased together. These algorithms uncover customer purchasing patterns and previously hidden relationships between products that help inform recommendation engines or other cross-selling opportunities. You might be most familiar with these rules from the “Frequently bought together” and “People who bought this item also bought” sections on your favorite online retail

# Unsupervised Learning Algorithm

53

- **Types of Unsupervised learning**
- Association rules are also often used to organize medical datasets for clinical diagnoses. Using unsupervised machine learning and association rules can help doctors identify the probability of a specific diagnosis by comparing relationships between symptoms from past patient cases.
- Typically, Apriori algorithms are the most widely used for association rule learning to identify related collections of items or sets of items. However, other types are used, such as Eclat and FP-growth algorithms.

# Unsupervised Learning Algorithm

54

## Association Rules Exercise

### Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

# Unsupervised Learning Algorithm

55

## Association Rules Exercise

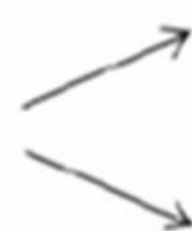
- Here are a dozen sales transactions.
- The objective is to use this transaction data to find affinities between products, that is, which products sell together often.
- The support level will be set at 33 percent; the confidence level will be set at 50 percent.

# Unsupervised Learning Algorithm

56

## Association Rules Exercise

*Rule :  $X \Rightarrow Y$*


$$\text{Support} = \frac{\text{freq}(X, Y)}{N}$$
$$\text{Confidence} = \frac{\text{freq}(X, Y)}{\text{freq}(X)}$$



# Unsupervised Learning Algorithm

57

## Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

1-item Sets	Frequency
Milk	9
Bread	10
Butter	10
Egg	3
Ketchup	3
Cookies	5

Frequent 1-item Sets	Frequency
Milk	9
Bread	10
Butter	10
Cookies	5

# Unsupervised Learning Algorithm

58

## Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

2-item Sets	Frequency
Milk, Bread	7
Milk, Butter	7
Milk, Cookies	3
Bread, Butter	9
Butter, Cookies	3
Bread, Cookies	4

Frequent 2-item Sets	Frequency
Milk, Bread	7
Milk, Butter	7
Bread, Butter	9
Bread, Cookies	4

# Unsupervised Learning Algorithm

59

## Transactions List

1	Milk	Egg	Bread	Butter
2	Milk	Butter	Egg	Ketchup
3	Bread	Butter	Ketchup	
4	Milk	Bread	Butter	
5	Bread	Butter	Cookies	
6	Milk	Bread	Butter	Cookies
7	Milk	Cookies		
8	Milk	Bread	Butter	
9	Bread	Butter	Egg	Cookies
10	Milk	Butter	Bread	
11	Milk	Bread	Butter	
12	Milk	Bread	Cookies	Ketchup

Milk, Bread, Butter, Cookies

3-item Sets	Frequency
Milk, Bread, Butter	6
Milk, Bread, Cookies	1
Bread, Butter, Cookies	3
Milk, Butter, Cookies	2

Frequent 3-item Sets	Frequency
Milk, Bread, Butter	6

# Unsupervised Learning Algorithm

60

## Association Rule Mining - Subset Creation

- Frequent 3-Item Set =  $I \Rightarrow \{\text{Milk, Bread, Butter}\}$
- Non-Empty subset are
  - $\{\{\text{Milk}\}, \{\text{Bread}\}, \{\text{Butter}\}, \{\text{Milk, Bread}\}, \{\text{Milk, Butter}\}, \{\text{Bread, Butter}\}\}$
- How to form Association Rule...?
  - For every non-empty subset  $S$  of  $I$ , the association rule is,
    - $S \rightarrow (I-S)$
    - If  $\text{support}(I) / \text{support}(S) \geq \text{min\_confidence}$

# Unsupervised Learning Algorithm

61

## Association Rule Mining - Subset Creation

- Non-Empty subset are
  - $\{\{\text{Milk}\}, \{\text{Bread}\}, \{\text{Butter}\}, \{\text{Milk, Bread}\}, \{\text{Milk, Butter}\}, \{\text{Bread, Butter}\}\}$
  - Min\_Support = 30% and Min\_Confidence = 60%
- Rule 1:  $\{\text{Milk}\} \rightarrow \{\text{Bread, Butter}\}$  {S=50%, C=66.67%}
  - Support =  $6/12 = 50\%$
  - Confidence =  $\text{Support}(\text{Milk, Bread, Butter}) / \text{Support}(\text{Milk}) = \frac{6/12}{9/12} = 6/9 = 66.67\% > 60\%$
  - Valid
- Rule 2:  $\{\text{Bread}\} \rightarrow \{\text{Milk, Butter}\}$  {S=50%, C=60%}
  - Support =  $6/12 = 50\%$
  - Confidence =  $\text{Support}(\text{Milk, Bread, Butter}) / \text{Support}(\text{Bread}) = 6/10 = 60\% \geq 60\%$
  - Valid



# Unsupervised Learning Algorithm

62

## Association Rule Mining - Subset Creation

- Non-Empty subset are
  - $\{\{\text{Milk}\}, \{\text{Bread}\}, \{\text{Butter}\}, \{\text{Milk, Bread}\}, \{\text{Milk, Butter}\}, \{\text{Bread, Butter}\}\}$
  - Min\_Support = 30% and Min\_Confidence = 60%
- Rule 5:  $\{\text{Milk, Butter}\} \rightarrow \{\text{Bread}\}$  {S=50%, C=85.7%}
  - Support =  $6/12 = 50\%$
  - Confidence =  $\text{Support}(\text{Milk, Bread, Butter}) / \text{Support}(\text{Milk, Butter}) = 6/7 = 85.7\% \geq 60\%$
  - Valid
- Rule 6:  $\{\text{Bread, Butter}\} \rightarrow \{\text{Milk}\}$  {S=50%, C=66.67%}
  - Support =  $6/12 = 50\%$
  - Confidence =  $\text{Support}(\text{Milk, Bread, Butter}) / \text{Support}(\text{Bread, Butter}) = 6/9 = 66.67\% \geq 60\%$
  - Valid