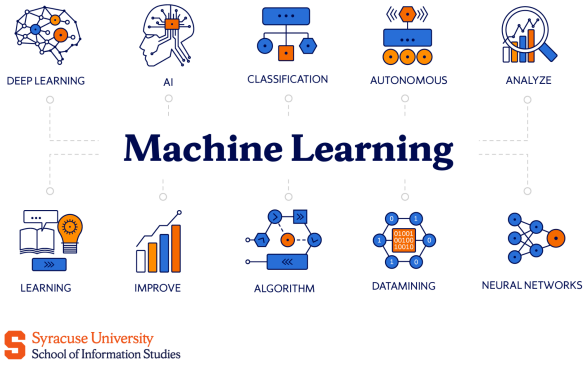
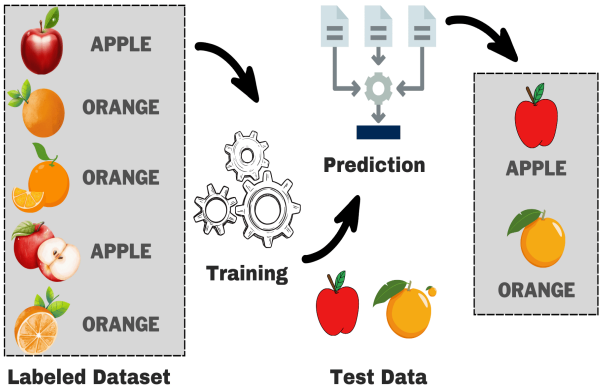
Types of Machine Learning Algorithms



Machine learning (ML) algorithms are a set of rules and procedures that are designed to learn from data and make predictions or decisions without being explicitly programmed. These algorithms are typically categorized into three main types based on the nature of the data and the learning process: **supervised learning**, **unsupervised learning**, and **reinforcement learning**.

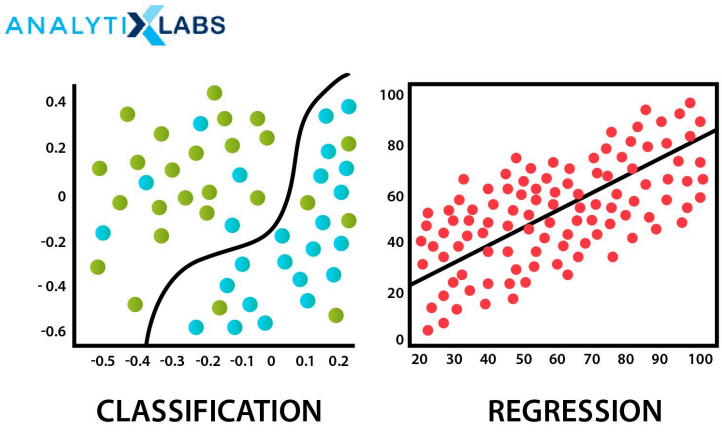
Supervised Learning

Supervised learning is a type of machine learning where the algorithm is trained on a **labeled dataset**. This means the training data includes both the input features and the corresponding correct output or "label." The algorithm's goal is to learn the mapping function from the input to the output so it can make accurate predictions on new, unseen data. Supervised learning is like a student learning from a teacher, where the teacher provides the questions and the correct answers.



Supervised learning problems can be further divided into two types:

- **Classification**: The output variable is a **categorical** value, such as "spam" or "not spam," "malignant" or "benign." The algorithm classifies data into predefined categories.
- **Regression**: The output variable is a **continuous** numerical value, such as a house's price, temperature, or a person's age. The algorithm predicts a real-valued output.

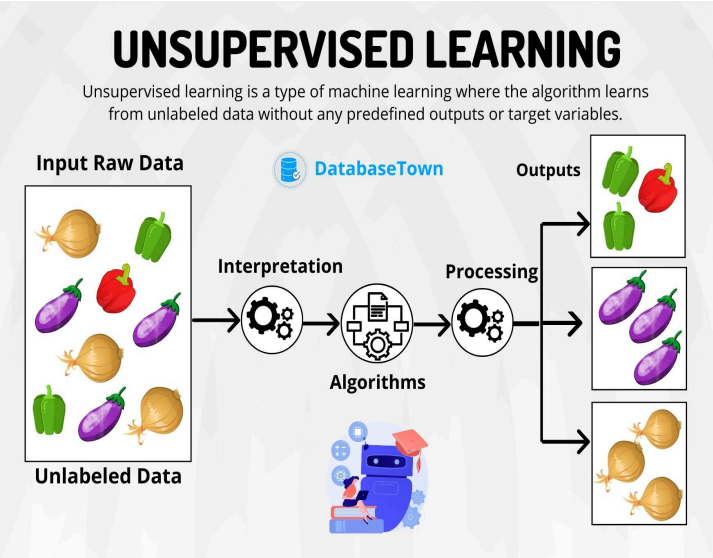


Common Algorithms:

- **Linear Regression**: Used for regression tasks, it models the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data. For example, **predicting house prices based on their size in square feet**.
- **Logistic Regression**: Used for classification tasks, it predicts the probability of an outcome that can only be one of two values. It's often used for binary classification. For example, **email spam or nor detection**.
- **Decision Trees**: These algorithms use a tree-like model of decisions and their possible consequences. Each internal node represents a feature, each branch represents a decision rule, and each leaf node represents the outcome. They can be used for both classification and regression. For example, **predicting whether a customer will purchase a product based on their characteristics.**
- **Random Forest**: An ensemble method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees. For example, **a self-driving car might use a random forest model to identify pedestrians and other vehicles on the road**.
- **Support Vector Machines (SVM)**: A powerful algorithm for classification and regression. SVMs find the optimal hyperplane that best separates different classes in a high-dimensional space. For example, **it can be used for classification, regression and outlier detection**.
- **k-Nearest Neighbors (k-NN)**: A simple, non-parametric algorithm used for both classification and regression. It classifies a data point based on the majority class or the average value of its **k** nearest neighbors. For example, **classifying a new fruit based on its color and weight**.

Unsupervised Learning

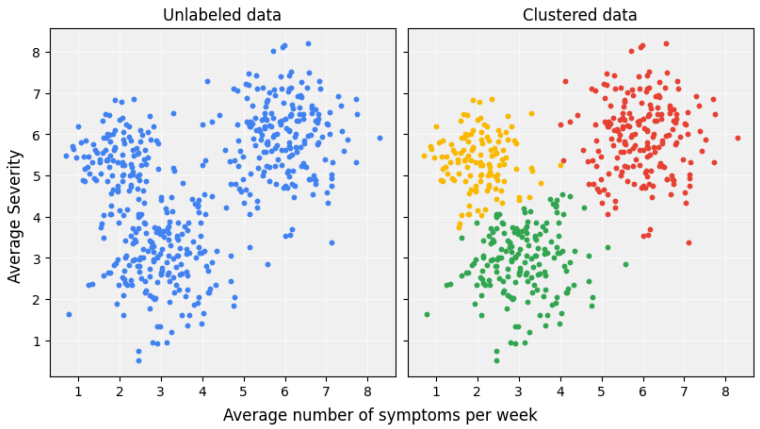
Unsupervised learning involves training algorithms on **unlabeled data**, meaning the dataset only contains input features and no corresponding output labels. The algorithm's goal is to discover hidden patterns, structures, and relationships within the data on its own. It's like a student learning on their own without a teacher, finding patterns and organizing information by themselves.



Unsupervised learning is commonly used for exploratory data analysis and to prepare data for other machine learning tasks.

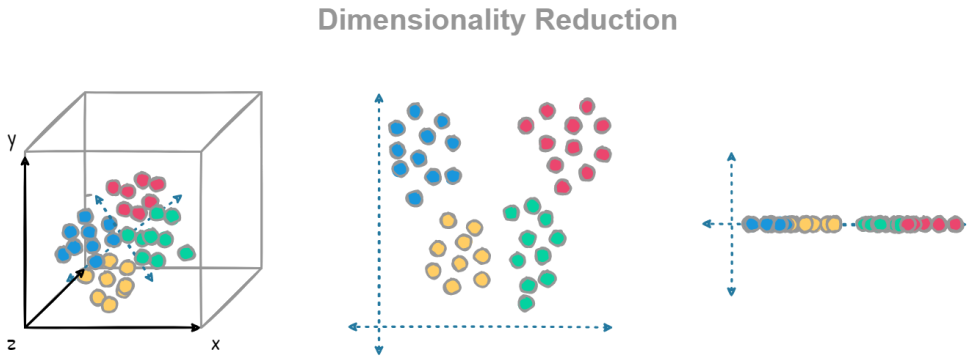
Common Algorithms:

- **Clustering**: This is the process of grouping a set of data points into clusters so that data points in the same cluster are more similar to each other than to those in other clusters. For example, **customer segmentation based on the user's preference**.



  - **K-Means Clustering**: Partitions data into **k** distinct clusters by finding the centroid of each cluster and assigning each data point to the nearest centroid.
  - **Hierarchical Clustering**: Builds a tree of clusters (a dendrogram) by either merging or splitting clusters based on their similarity.
  - **DBSCAN**: Finds clusters of varying shapes and sizes based on the density of data points.

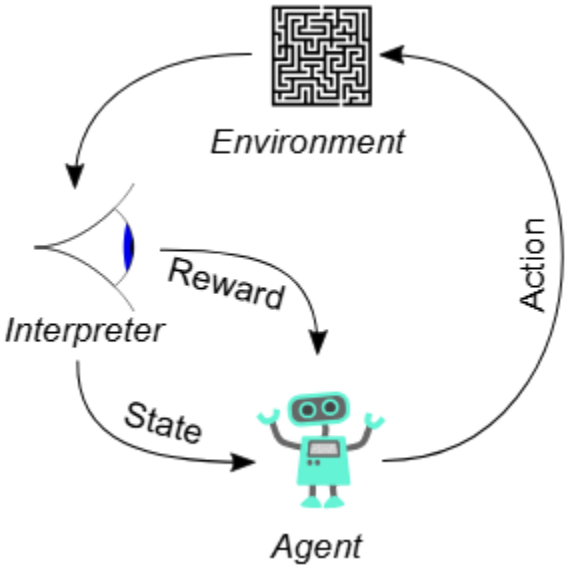
- **Dimensionality Reduction**: This technique reduces the number of features or variables in a dataset while retaining the most important information. This is useful for visualization and to prevent the "curse of dimensionality". For example, text categorization, Image retrieval, etc.



  - **Principal Component Analysis (PCA)**: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Reinforcement Learning

Reinforcement learning (RL) is a type of machine learning inspired by behavioral psychology. It involves an **agent** that learns to make decisions by performing **actions** in an **environment** to achieve a specific goal. The agent receives **rewards** for desirable actions and **penalties** for undesirable ones. The goal is to learn a policy—a set of rules—that maximizes the total cumulative reward over time.



RL is different from supervised and unsupervised learning because there's no fixed dataset. The agent learns through trial and error as it interacts with the environment.

Real-World Applications:

- **Robotics**: Training a robot to perform complex tasks like walking, grasping objects, or navigating a room.

- **Autonomous Vehicles**: A self-driving car learns to make decisions, such as braking, accelerating, or changing lanes, based on sensor data and the "rewards" of a safe trip.
- **Gaming**: AI agents have mastered complex games like Go (AlphaGo) and chess by playing against themselves and learning from the outcomes.
- **Resource Management**: Optimizing energy consumption in data centers by controlling heating, ventilation, and air conditioning (HVAC) systems.