**Machine Learning Specialization by deeplearning.ai**

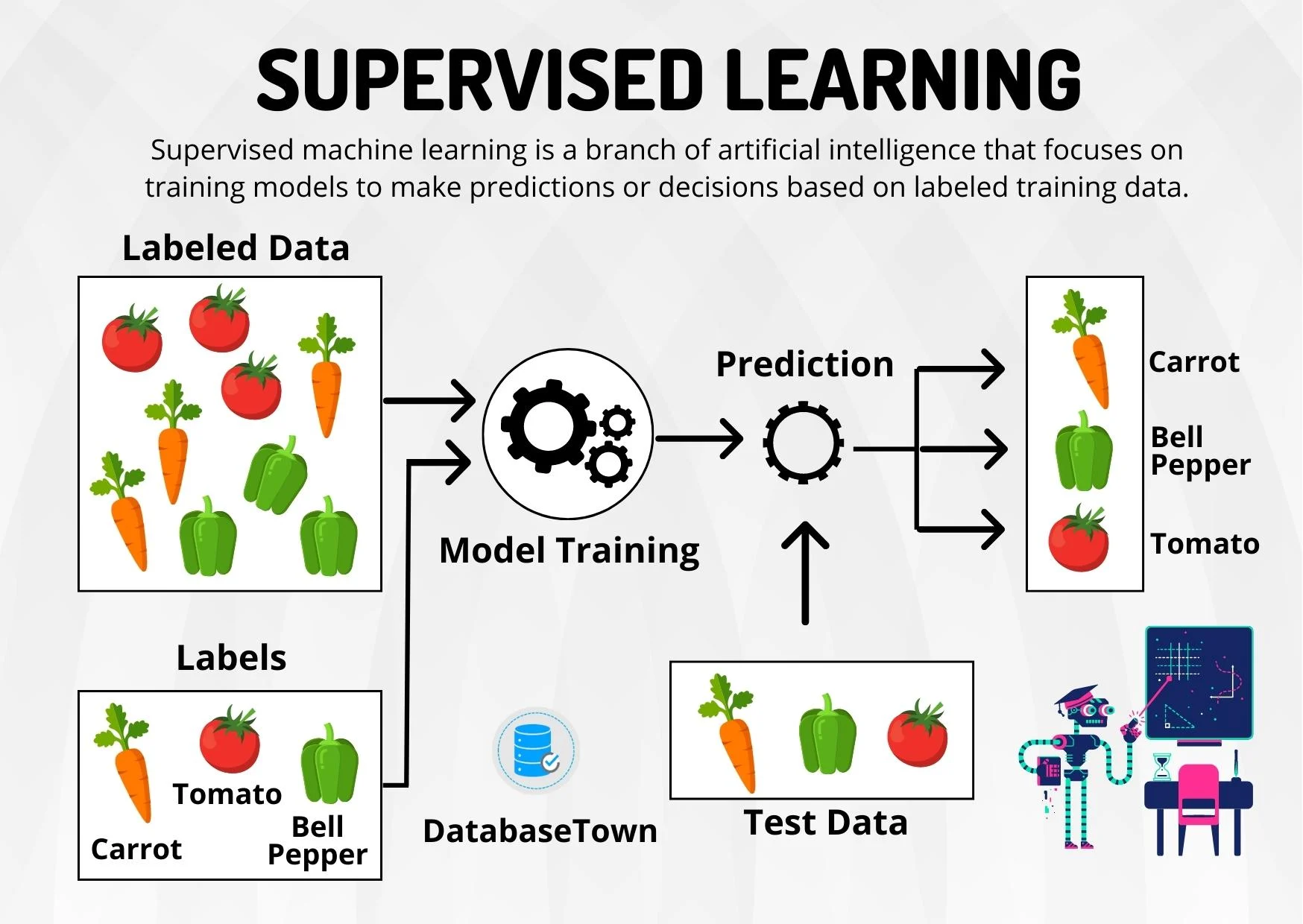
**What is Machine Learning?**

“Field of study that gives computers the ability to learn without being explicitly programmed.” - Arthur Samuel (1959)

**Machine learning algorithms**

* Supervised Learning - learn from data labeled with the “right answers”
* Unsupervised Learning - learn from unlabeled data by spotting the patterns, trends and structures.
  + Recommender systems
  + Reinforcement learning

**Supervised Learning**



Supervised learning is a type of machine learning paradigm where an algorithm is trained on a labeled dataset to make predictions or decisions based on input data. The term “supervised” refers to the process of providing the algorithm with input-output pairs during training, with the intention of learning a mapping function that can predict the correct output for new, unseen inputs.

The following steps take place in supervised learning:

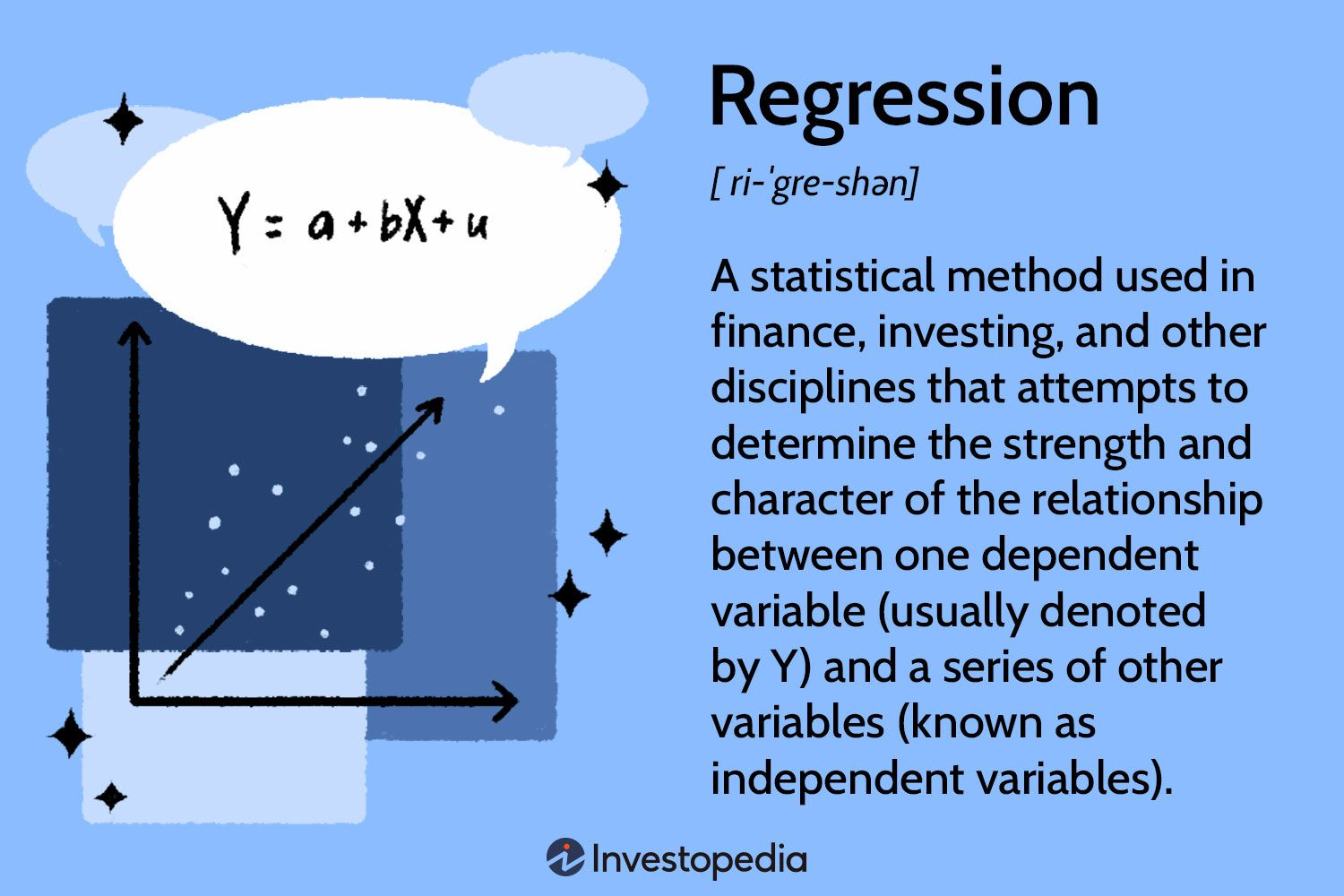
1. **Labeled Dataset:** The process starts with a labeled dataset, which consists of input data (also known as features) and corresponding output data (also known as labels or targets)
2. **Training Phase:** During the training phase, the supervised learning algorithm analyzes the labeled dataset to learn the underlying patterns and relationships between the input and output data. The goal is to create a model that can generalize the training data and make accurate predictions on new, unseen data.
3. **Model Creation:** The algorithm uses various mathematical and statistical techniques to create a model that maps the input data to the output labels. The model represents the learned patterns and can take new input data to produce predictions.
4. **Prediction Phase:** Once the model is trained, it is evaluated on a separate set of data called the test or validation set. This data was not used during training and is used to assess how well the model generalizes to new, unseen examples.
5. **Model Evaluation:** The performance of the model is measured using evaluation metrics appropriate for the specific task. For example, in classification tasks (where the goal is to assign inputs to discrete categories), metrics like accuracy, precision, recall and F1 scores are commonly used. In regression tasks (where the goal is to predict continuous values), metrics like mean squared error (MSE) or mean absolute error (MAE) are often used.
6. **Refinement and Iteration:** If the model’s performance is not satisfactory, the process can be iterated by adjusting various parameters, choosing different algorithms, or improving the quality of the training data.

Examples of supervised learning tasks include:

* **Image Classification:** Identifying objects or patterns in images and assigning them to predefined classes (e.g., recognizing cats and dogs in pictures).
* **Speech Recognition:** Converting spoken language into text.
* **Language Translation:** Translating text from one language to another.
* **Stock Market Prediction:** Predicting future stock prices based on historical data.
* **Email Spam Detection:** Classifying emails as spam or not spam based on their content.

**Regression**

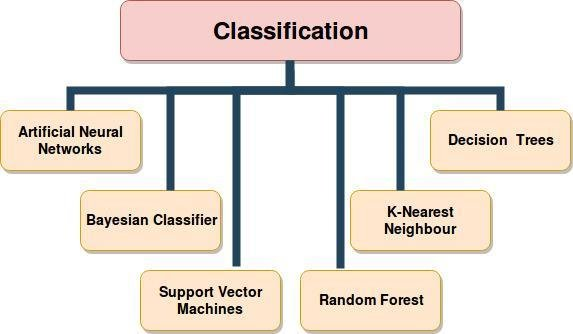
Regression is an example of supervised learning. It is a statistical method used for predicting numeric values based on the relationship between input features and the target variable.



Key Concepts of Regression:

1. **Target Variable:** In regression, there is a target variable (also known as the dependent variable or the response variable) that we want to predict. This is a continuous numeric value, such as stock prices, house prices, or temperature.
2. **Input features:** The input features (also known as independent variables or predictors) are the factors that may influence the target variable. There are numeric values or features that help the model make predictions. For example, in predicting house prices, input features could be the number of rooms, the size of the house and the location.
3. **Relationship:** Regression seeks to identify the mathematical relationship between the input features and the target variable. It tries to find the best-fitting line or curve that represents the patterns in the data.
4. **Best-Fit Line/Curve:** The goal of regression is to find the line (for simple linear regression) or curve (for multiple linear regression or polynomial regression) that minimizes the distance between the predicted values and actual values in the training data.
5. **Cost Function:** To find the best-fit line or curve, regression algorithms use a cost function (also known as the loss function). The cost function measures the error between predicted values and the actual values. The model’s objective is to minimize this error during training.
6. **Training:** During the training phase, the regression algorithm adjusts the model’s parameters (coefficients and intercepts for linear regression) to minimize the cost function. This process is typically done using optimization techniques like gradient descent.
7. **Prediction:** Once the model is trained, it can be used to make the predictions on new input data. The model takes the input features and calculates the predicted value of the target variable based on the learned relationship.
8. **Evaluation:** To assess the performance of the regression model, evaluation metrics such as Mean Squared Error (MSE) or Mean Absolute Error (MAE) are commonly used.

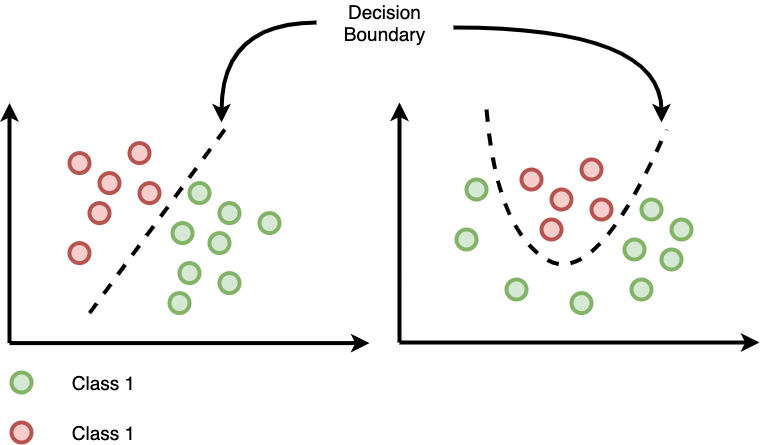
**Classification**

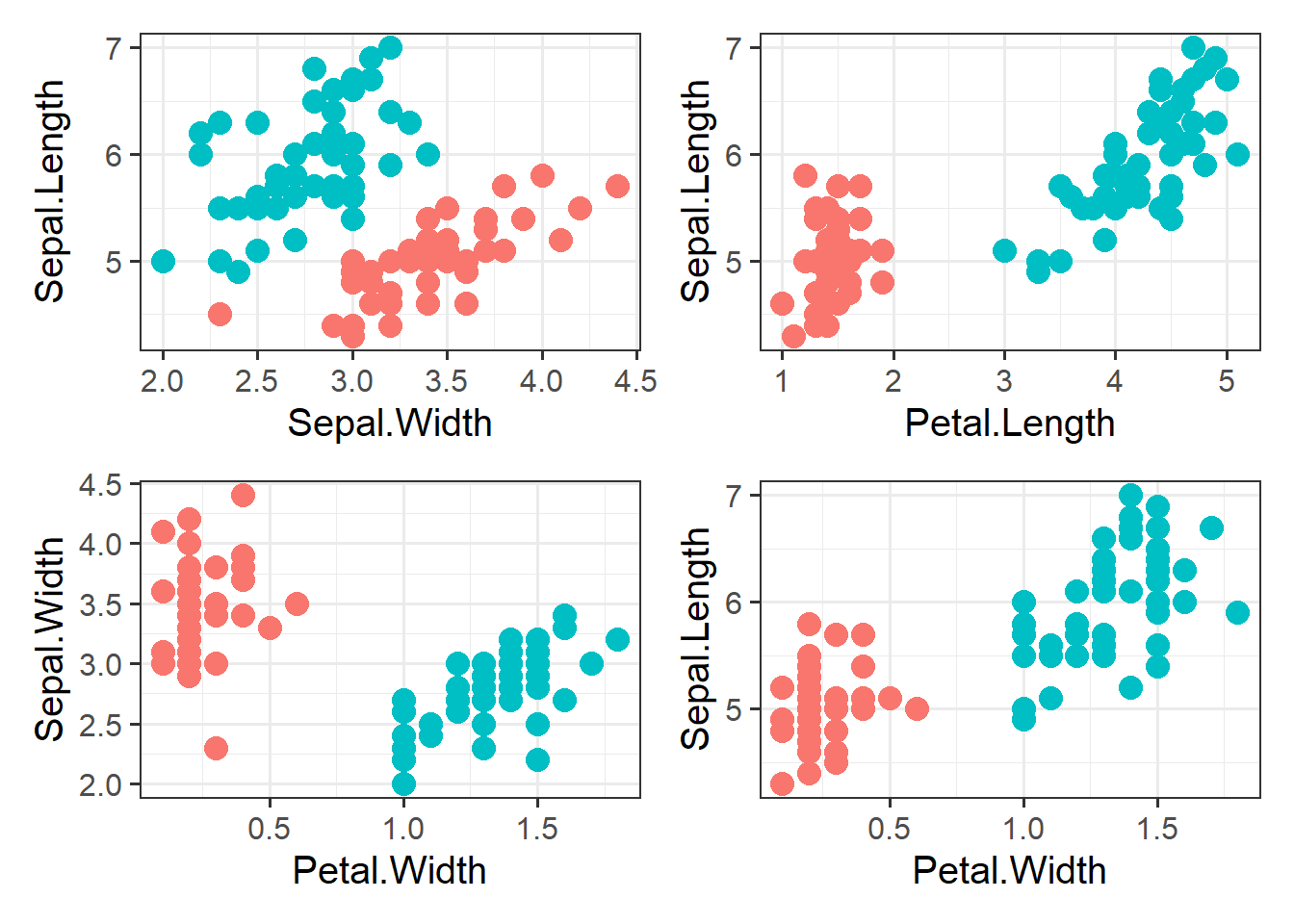


Classification is another example of supervised learning, and it is used to predict discrete categories or classes for given input data.

Key concepts of Classification:

1. **Target Labels:** In classification, the target variable (also known as the dependent variable or the class label) consists of discrete categories or classes. For example, it could be binary (e.g., “Yes” or “No”) or multiclass (e.g., “Cat”, “Dog”, “Bird”).
2. **Input Features:** Similar to regression, classification also involves input features (also known as independent variables or predictors) that represent the characteristics of the data. These features help the model make decisions about the class labels.
3. **Decision Boundary:** The goal of a classification algorithm is to create a decision boundary, which is a dividing line or surface that separates different classes in the feature space. The model learns this boundary from the training data.

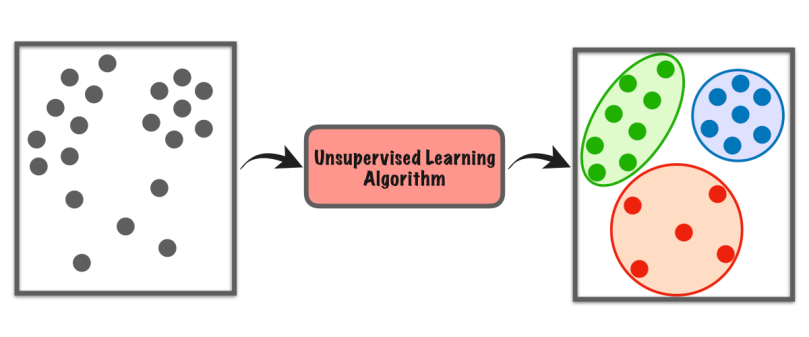


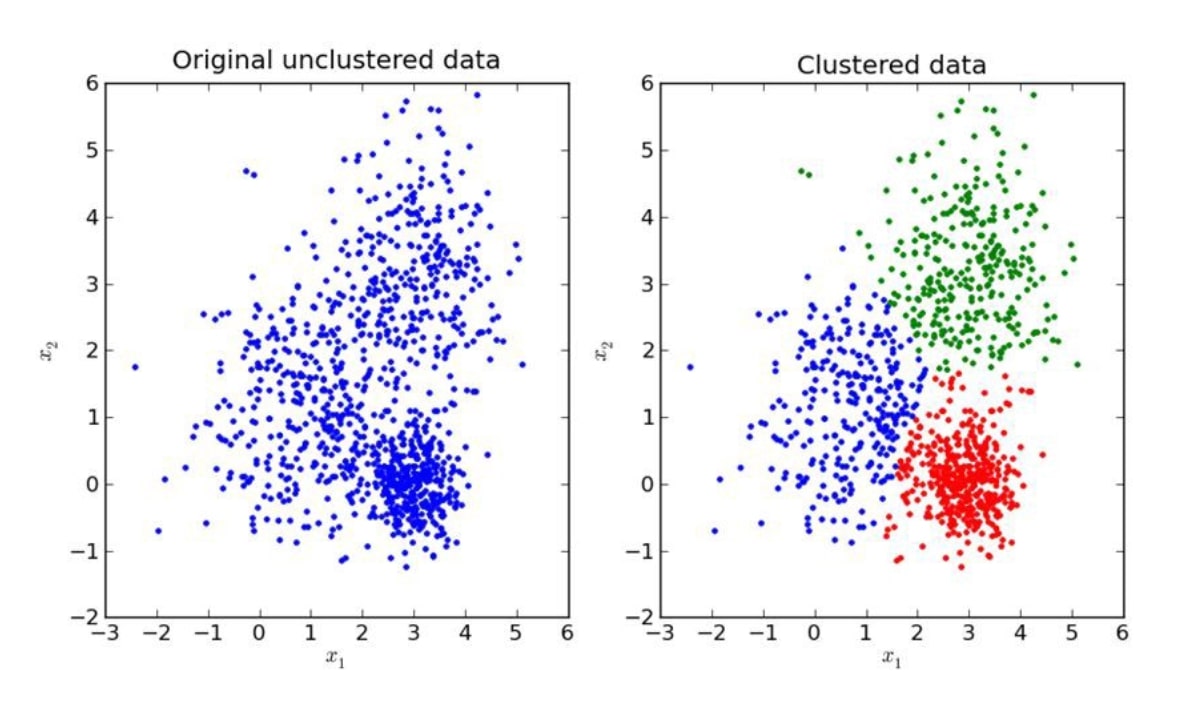


1. **Training Data:** The classification algorithm is trained on a labeled dataset, where each data point includes both the input features and the corresponding class.
2. **Learning Algorithm:** The classification algorithm analyzes the training data and learns the patterns and relationships between the input data and the correct class label.
3. **Prediction:** Once the model is trained and the decision boundary is established, the classification model can predict the class label of new, unseen data based on the input features. The model assigns the data point to the class on one side of the decision boundary.
4. **Evaluation:** To assess the performance of the classification model, various metrics are used, such as accuracy, precision, recall, F1 score, and confusion matrix. These metrics measure how well the model correctly predicts class labels compared to the actual labels.
5. **Binary and Multiclass Classification:** In binary classification, there are only two possible classes. In multiclass classification, there are more than two classes, and the model needs to distinguish between all of them.

**Unsupervised Learning**

Unsupervised Learning is a type of machine learning paradigm where an algorithm is used to find patterns, relationships, or structures in data without explicit guidance from labeled examples. Unlike supervised learning, unsupervised learning deals with unlabeled data, meaning the algorithm tries to identify inherent structures or grouping within the data without knowing the correct answers beforehand.





Key characteristics of Unsupervised Learning:

1. **Unlabeled Data:** In supervised learning, the dataset contains only input data (features) without corresponding output labels. The algorithm’s objective is to discover patterns or hidden structures within this data.
2. **Clustering:** One of the primary tasks in unsupervised learning is clustering, where the algorithm groups similar data points together based on their inherent similarities. Data points within the same cluster are more similar to each other than to data points in other clusters.
3. **Dimensionality Reduction:** Another common task in unsupervised learning is dimensionality reduction, where the algorithm reduces the number of input features while retaining the important information. This can be beneficial for data visualization and processing high-dimensional data.
4. **Density Estimation:** Unsupervised learning can also be used for density estimation, where the algorithm models the underlying probability distribution of the data.
5. **Feature Learning:** Some unsupervised learning algorithms can learn representations of the data, also known as feature learning. These learned representations can be useful for downstream tasks like supervised learning.
6. **No ground truth**: Unlike supervised learning, there is no ground truth or correct answers provided during the training phase. The algorithm explores the data structure on its own, making it more explanatory in nature.

Examples of Unsupervised Learning Tasks:

1. **Clustering:** Grouping customers based on their purchasing behavior of the market segmentation.
2. **Dimensionality Reduction:** Reducing the number of features in a dataset to visualize or compress the data while preserving important information.
3. **Anomaly Detection:** Identifying unusual or abnormal patterns in data that deviate significantly from the norm.
4. **Topic Modeling:** Automatically, discovering topics within a collection of documents.
5. **Generative Models:** Creating new samples that resemble the training data, such as generating realistic images.

**Clustering**

Key points about Clustering:

1. **Unsupervised Learning:** Clustering does not rely on labeled data; it works with only input features and does not have access to the correct class labels during training.
2. **Similarity Measure:** Clustering algorithms use a similarity measure (e.g. distance metrics) to assess how close or similar data points are to each other. Data points that are closer to each other are more likely to be grouped into the same cluster.
3. **Cluster Centers:** In some clustering algorithms, the center of a cluster (centroid) is computed, and data points are assigned to the nearest centroid, forming clusters.
4. **Number of Clusters:** The number of clusters is usually determined before clustering, but some algorithms can automatically find the optimal number of clusters based on data characteristics.
5. **Applications:** Clustering is widely used in various fields, such as customer segmentation in marketing, image segmentation in computer vision, anomaly detection, document clustering, and more.
6. **Algorithms:** There are several clustering algorithms, including k-means, hierarchical clustering, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), and Gaussian Mixture Models (GMMS).
7. **Evaluation:** Measuring the quality of clustering can be subjective, there are metrics like silhouette score and within-cluster sum of squares (WCSS) that help assess the clustering performance.