

Introduction to Machine Learning

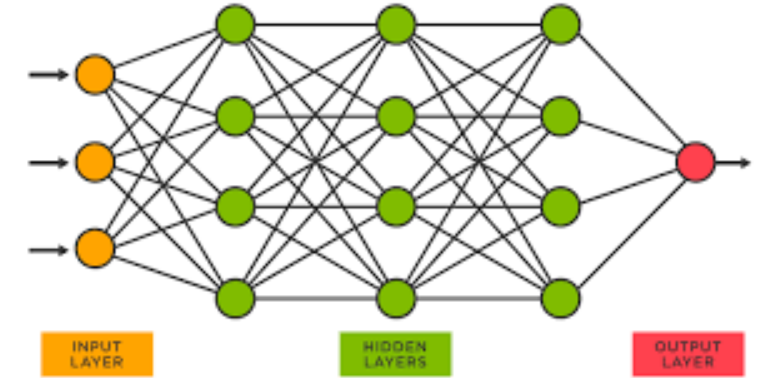
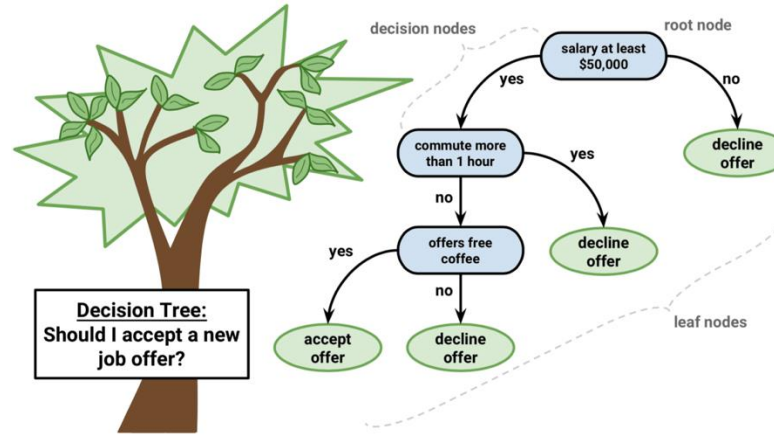
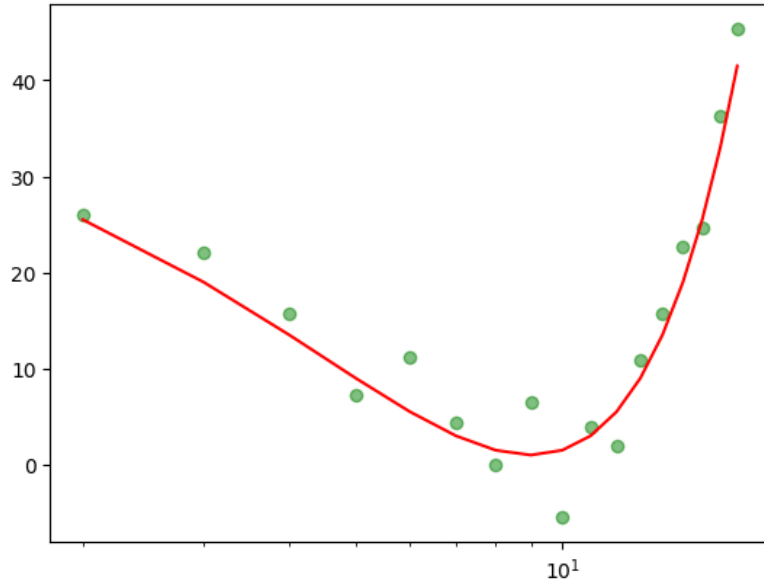
Gökhan Özşari

What is AI?

- “Artificial intelligence;
 - the capacity of computers or other machines to exhibit or simulate intelligent behaviour;
 - the field of study concerned with this.
 - In later use also: software used to perform tasks or produce output previously thought to require human intelligence, esp. by using machine learning to extrapolate from large collections of data.”
 - – Oxford English dictionary



This Photo by Unknown Author is licensed under CC BY-ND



Artificial Intelligence

Container term: can mean anything

Artificial intelligence: the capacity of computers or other machines to exhibit or simulate intelligent behaviour

Machine learning:

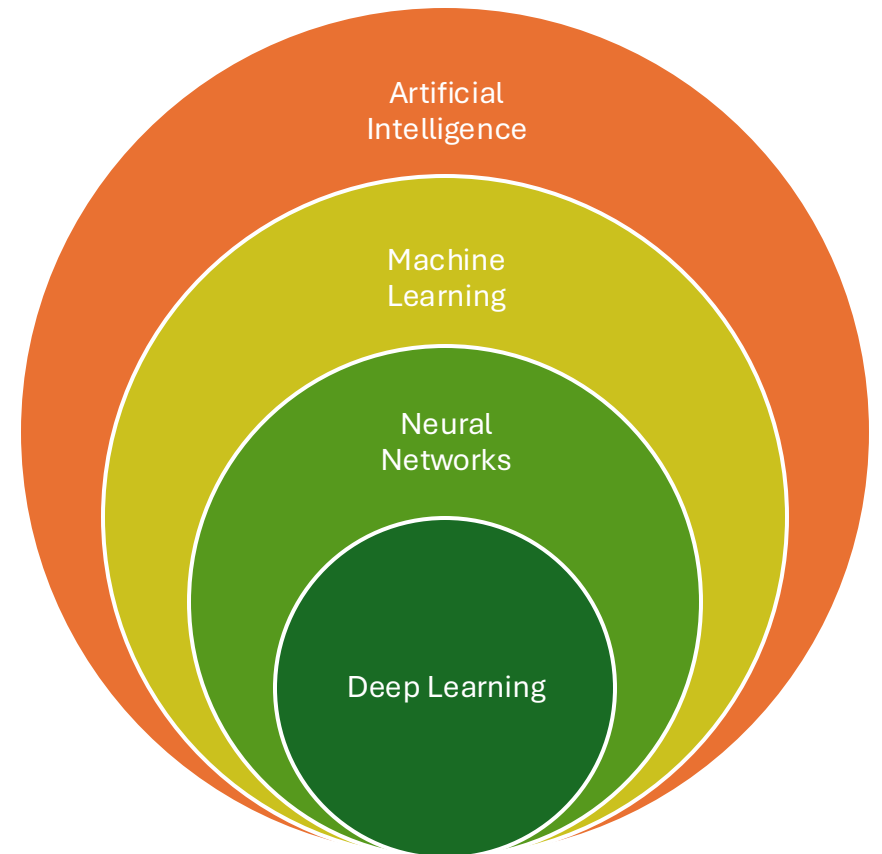
- Extrapolate from existing data
- More data better prediction

Neural networks:

- Complex form of machine learning
- Mimics neuron structure

Deep learning:

Complex neural networks, multiple hidden layers

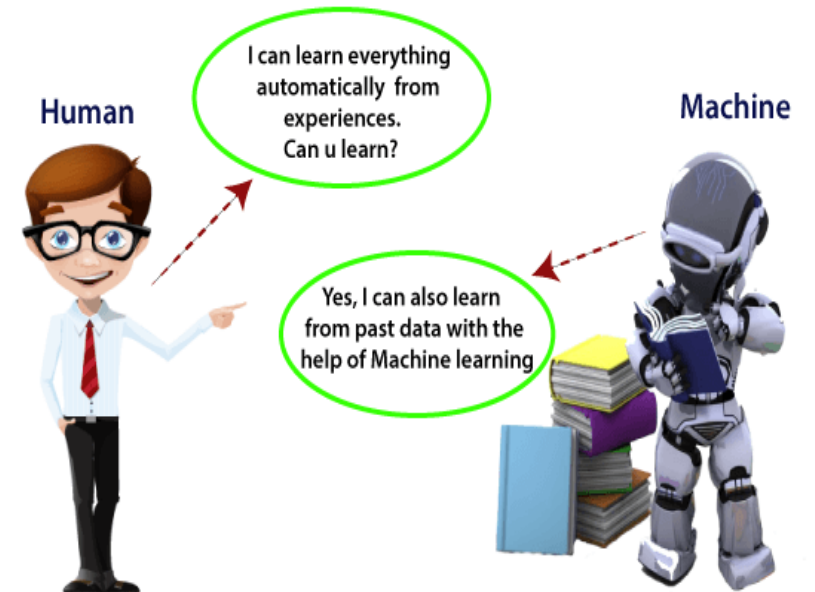


Machine Learning

- Machine Learning is a branch of Artificial Intelligence (AI) where computers are trained to learn from data and make decisions or predictions without being explicitly programmed.
- The goal of ML is to develop algorithms that can learn and improve over time. These algorithms are used to build models that can:
 - Make predictions (e.g., weather forecasts)
 - Classify data (e.g., spam detection)
 - Identify patterns (e.g., customer segmentation)
 - Optimize decisions (e.g., route planning)
 - And much more!





Key Characteristics:

- **Data-Driven:** ML relies on historical data to make predictions or decisions.
- **Self-Improving:** ML systems get better with experience as more variety of data becomes available.

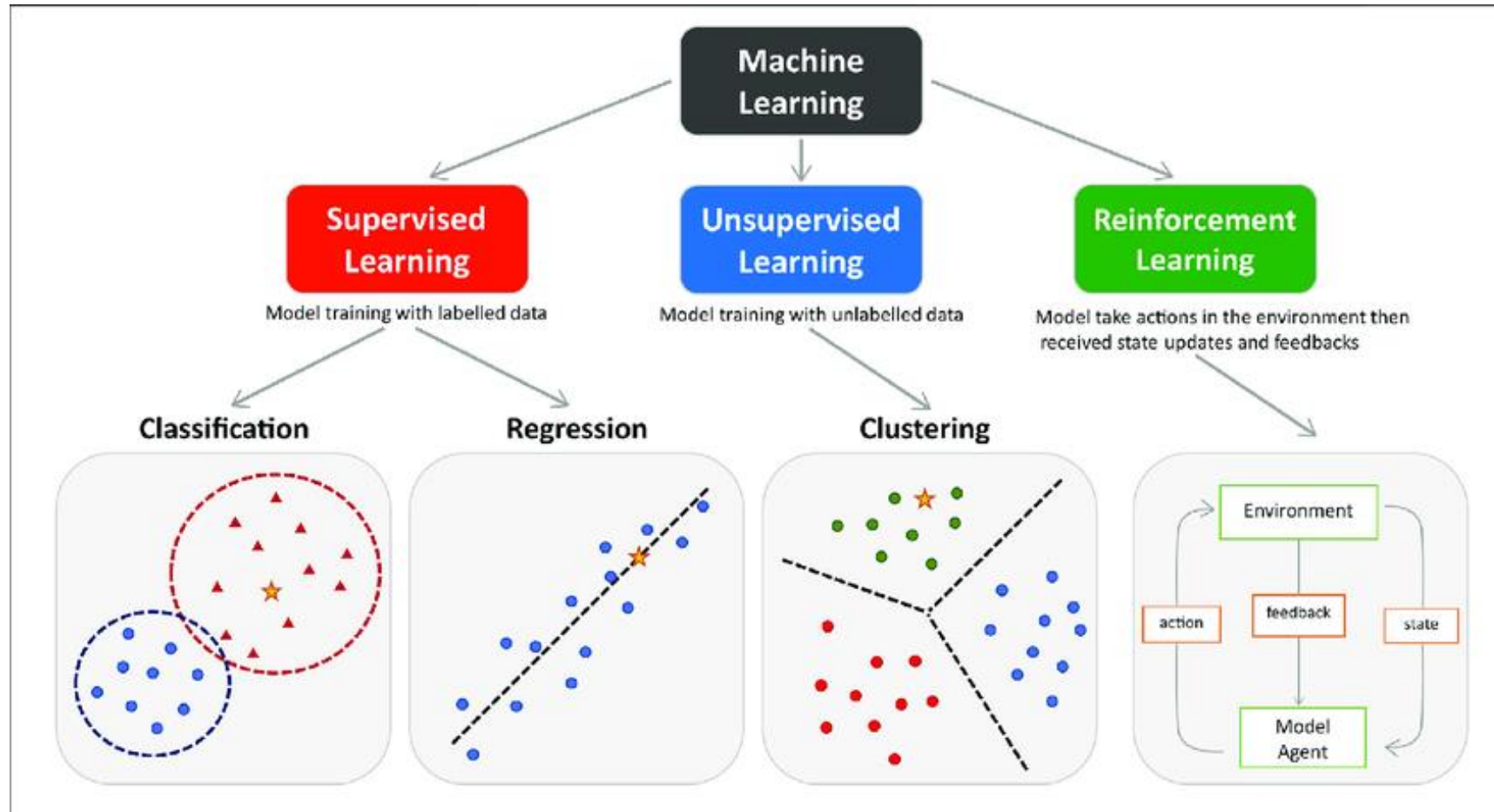


Why is ML Important?

ML powers some of the most impactful technologies in the world today:

- **LLMs:** ChatGPT, Gemini, etc.
- Netflix and YouTube recommendations 
- Google Translate 
- Fraud detection in banking 
- Self-driving cars 
- Medical diagnosis
- And many more!

Types of Machine Learning



Machine Learning Terminology

Dataset: A collection of data used for training and testing.

Feature: An individual property of the data.

Label: The target output in supervised learning.

age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	target
30	unemployed	married	primary	no	1787	no	no	cellular	19	oct	79	1	-1	0	unknown	no
33	services	married	secondary	no	4789	yes	yes	cellular	11	may	220	1	339	4	failure	no
35	management	single	tertiary	no	1350	yes	no	cellular	16	apr	185	1	330	1	failure	no
30	management	married	tertiary	no	1476	yes	yes	unknown	3	jun	199	4	-1	0	unknown	no
59	blue-collar	married	secondary	no	0	yes	no	unknown	5	may	226	1	-1	0	unknown	no
35	management	single	tertiary	no	747	no	no	cellular	23	feb	141	2	176	3	failure	no
36	self-employed	married	tertiary	no	307	yes	no	cellular	14	may	341	1	330	2	other	no
39	technician	married	secondary	no	147	yes	no	cellular	6	may	151	2	-1	0	unknown	no
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14	may	57	2	-1	0	unknown	no
43	services	married	primary	no	-88	yes	yes	cellular	17	apr	313	1	147	2	failure	no
39	services	married	secondary	no	9374	yes	no	unknown	20	may	273	1	-1	0	unknown	no
43	admin.	married	secondary	no	264	yes	no	cellular	17	apr	113	2	-1	0	unknown	no
36	technician	married	tertiary	no	1109	no	no	cellular	13	aug	328	2	-1	0	unknown	no
20	student	single	secondary	no	502	no	no	cellular	30	apr	261	1	-1	0	unknown	yes
31	blue-collar	married	secondary	no	360	yes	yes	cellular	29	jan	89	1	241	1	failure	no
40	management	married	tertiary	no	194	no	yes	cellular	29	aug	189	2	-1	0	unknown	no
56	technician	married	secondary	no	4073	no	no	cellular	27	aug	239	5	-1	0	unknown	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	20	apr	114	1	152	2	failure	no
25	blue-collar	single	primary	no	-221	yes	no	unknown	23	may	250	1	-1	0	unknown	no
31	services	married	secondary	no	132	no	no	cellular	7	jul	148	1	152	1	other	no

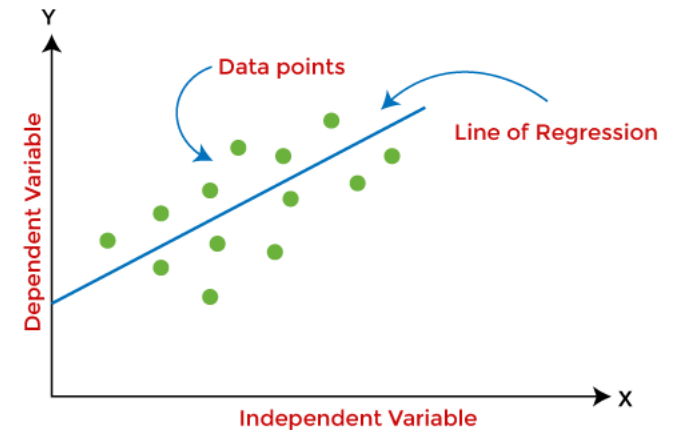
Machine Learning Terminology

Model: A mathematical representation of the data.

Training: The process of fitting a model to the data.

Testing: Evaluating a model on unseen data.

Prediction: An output generated by the model.



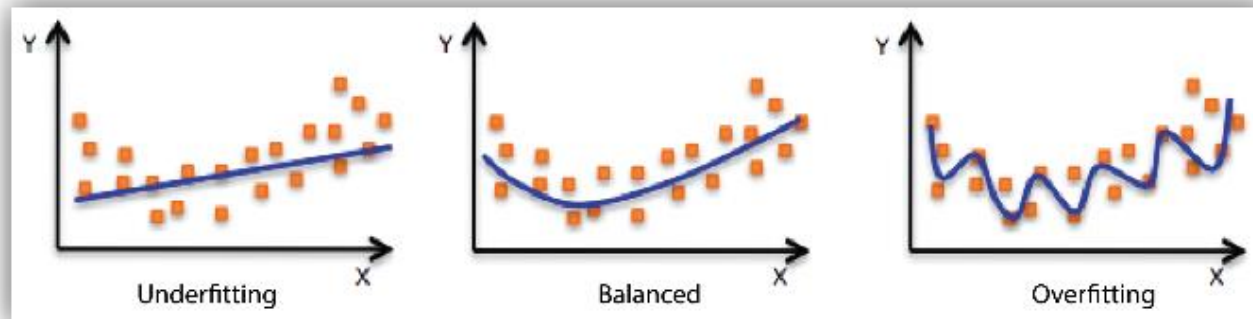
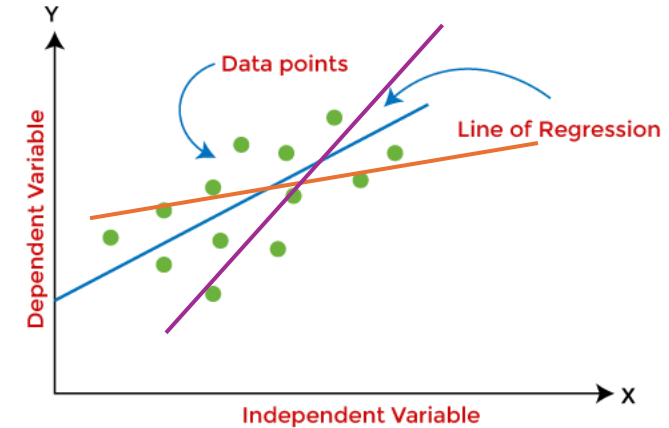
Machine Learning Terminology

Hyperparameters: Settings that control the learning process. They are set before training the model.

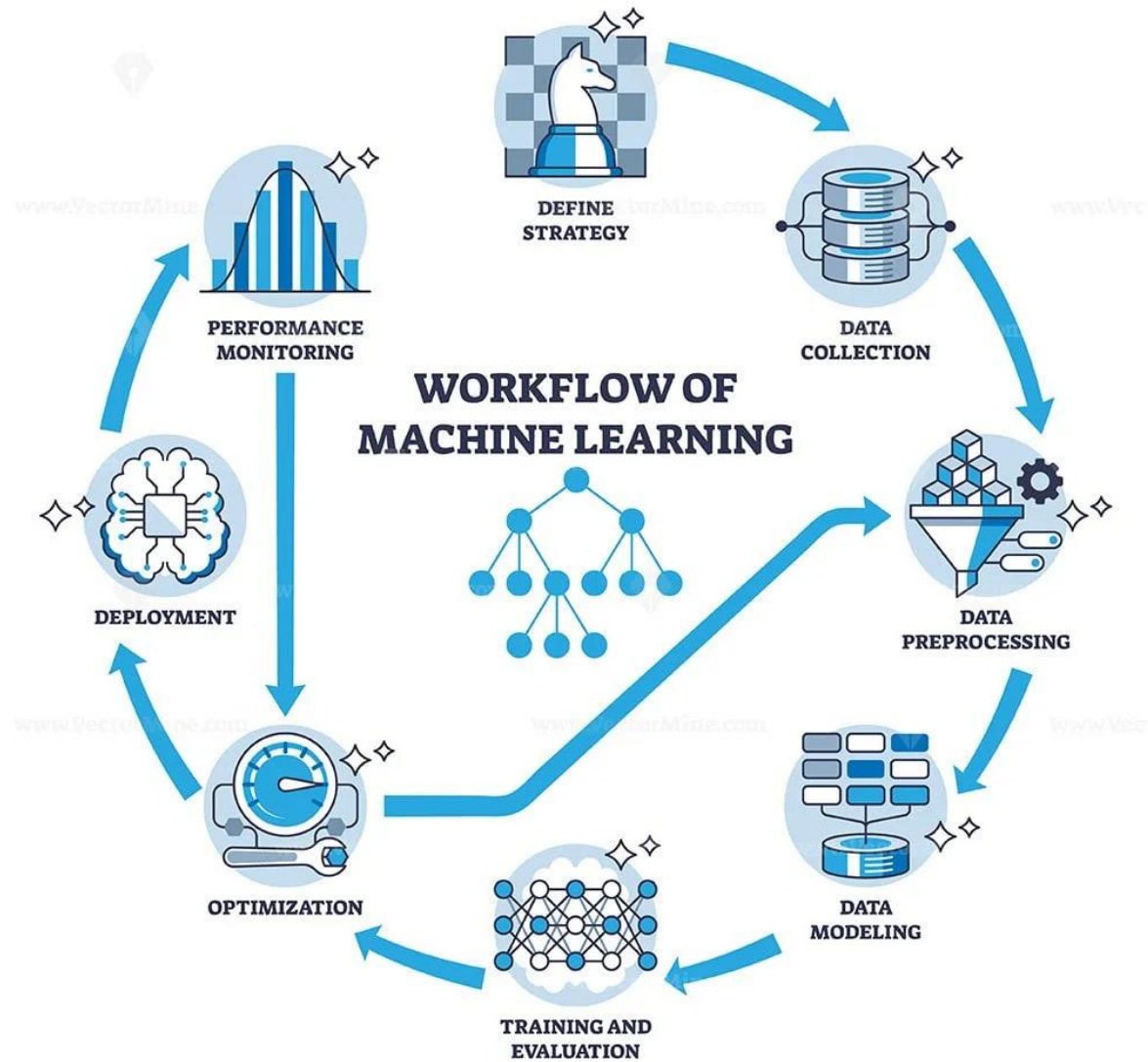
Evaluation Metric: A measure used to assess model performance.

Overfitting: When a model performs well on training data but poorly on unseen data.

Underfitting: When a model is too simple to capture the patterns in the data.



Machine Learning Workflow



Define Strategy & Data collection

Define the Problem

- Clearly define the objective.
- **Example:** Predicting whether a customer will churn based on their usage data.
- Specify whether it's a regression, classification, or clustering task.

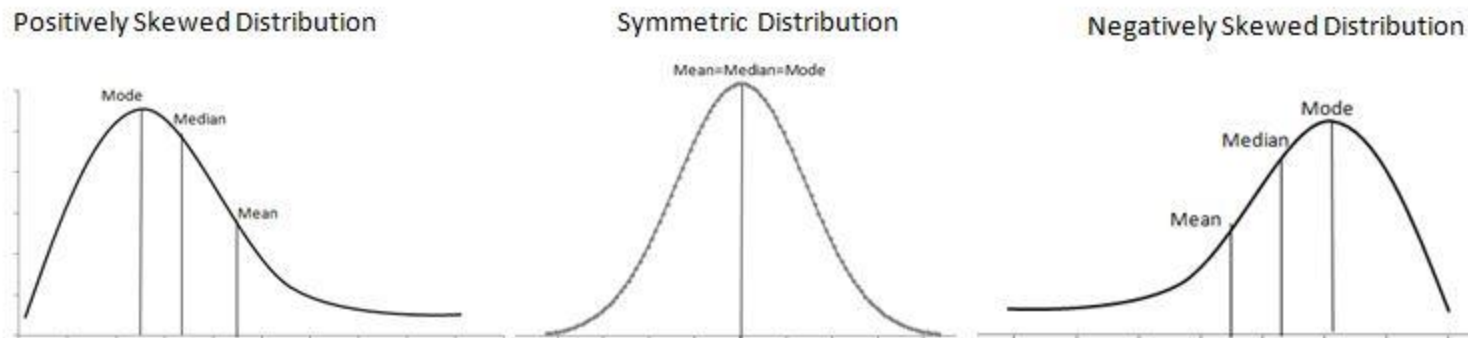
Collect and Clean Data

- Data Collection: Gather data from sources such as databases, APIs, or CSV files.
- Data Cleaning:
 - Handle missing values (e.g., replace with mean, median, or drop rows).
 - Remove duplicates.
- Address outliers that could distort results.

Explore visualize

Explore and Visualize Data

- Generate summary statistics (e.g., mean, standard deviation, and correlation).
- Visualize distributions and relationships using:
- Histograms
- Scatterplots
- Heatmaps
- **Example:** Plot the distribution of customer ages to check for skewness.



Feature Engineering

Feature engineering transforms raw data into meaningful features that improve model performance.

Feature Selection: Choose only the most relevant features.

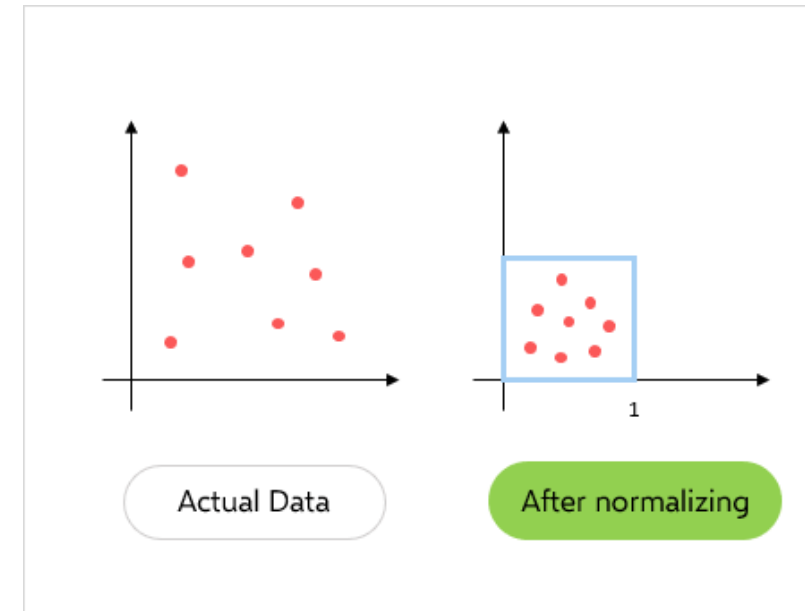
- **Example:** Removing highly correlated features to avoid redundancy.

Feature Transformation:

- Normalize numerical values (scaling data to 0-1 or z-score).
- Convert categorical features into numerical ones using one-hot encoding.

Feature Creation: Create new features from existing ones.

- **Example:** Extracting "month" from a "date" column.

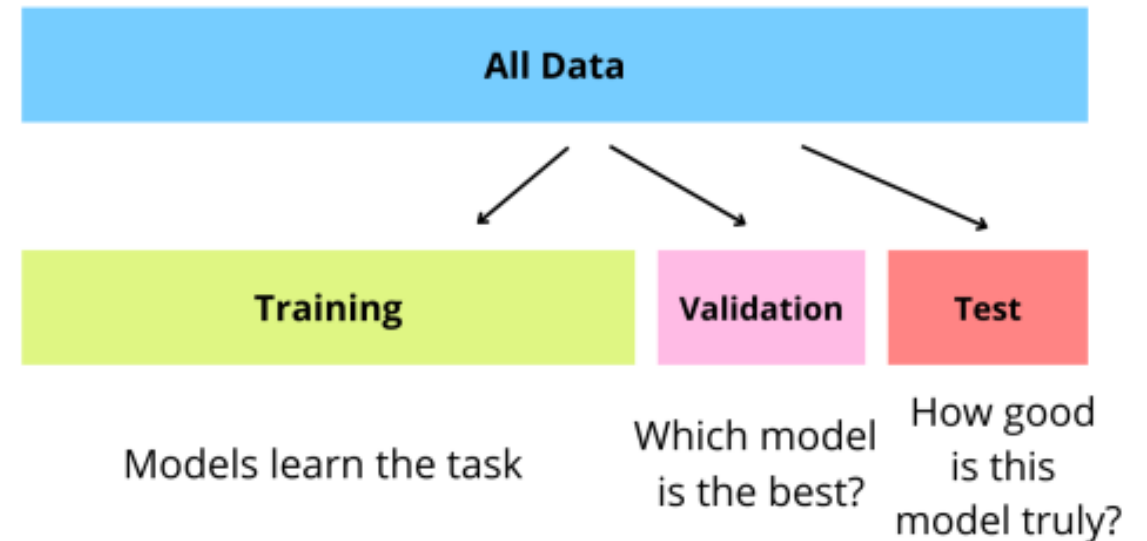


Split Data

Divide the dataset into:

- **Training Set:** For training the model.
- **Validation Set:** For hyperparameter tuning.
- **Test Set:** For final performance evaluation.

Typical split ratios: 70% training, 15% validation, 15% test.

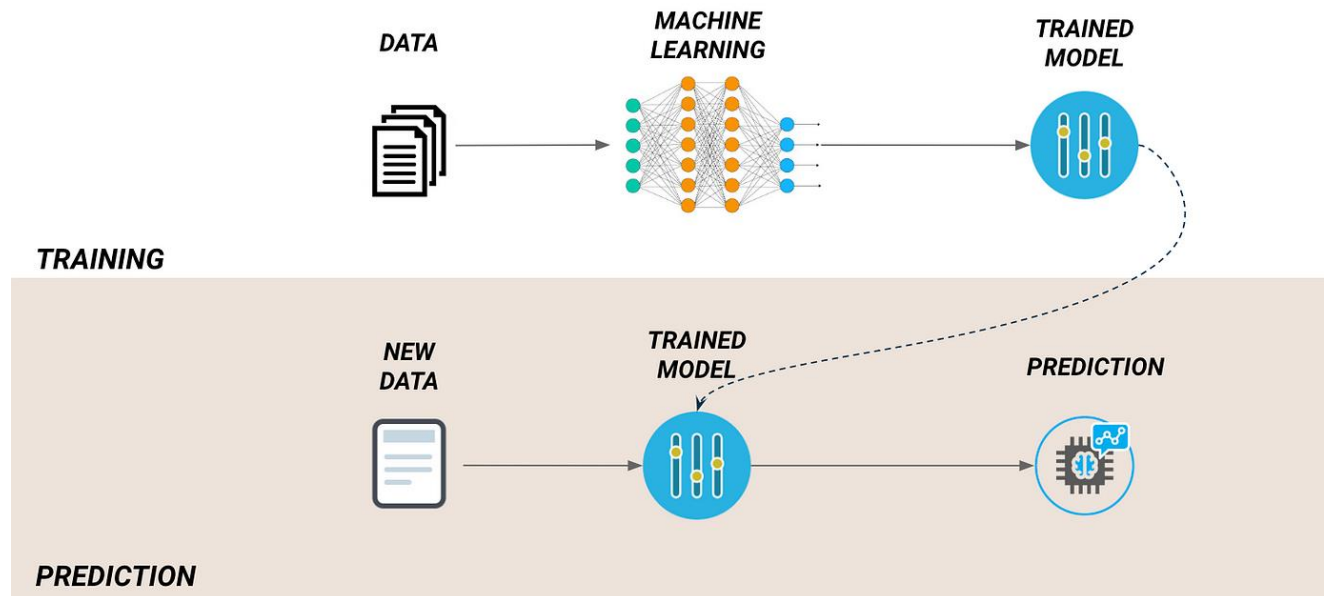


Choose and Train a Model

- Select an algorithm based on the task (e.g., regression for predicting numbers, classification for predicting categories).
- Train the model using the training dataset.

Example models:

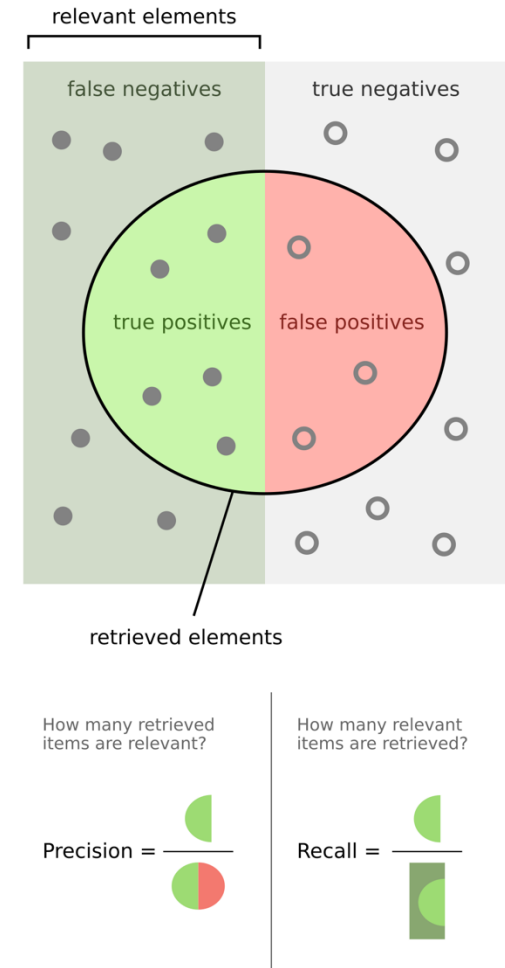
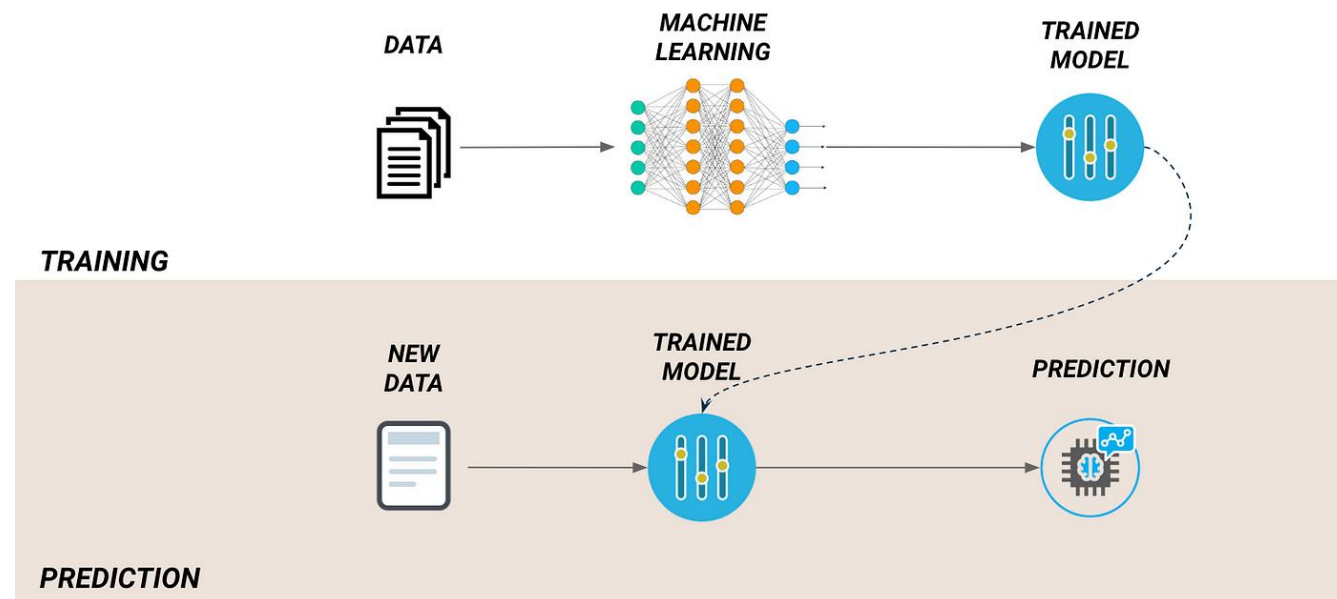
- Regression: Linear Regression
- Classification: Logistic Regression, Decision Trees
- Clustering: K-Means



Evaluate the Model

Use appropriate metrics to assess model performance:

- Regression: RMSE, MAE, R^2 .
- Classification: Accuracy, Precision, Recall, F1-score.
- Clustering: Silhouette Score.
- Visualize performance using confusion matrices or ROC curves.



Hyperparameter Optimization

Hyperparameter tuning helps to improve model performance by finding the best parameter configuration.

Grid Search:

Example: Trying multiple combinations of `max_depth` and `min_samples_split` in Decision Trees.

Random Search:

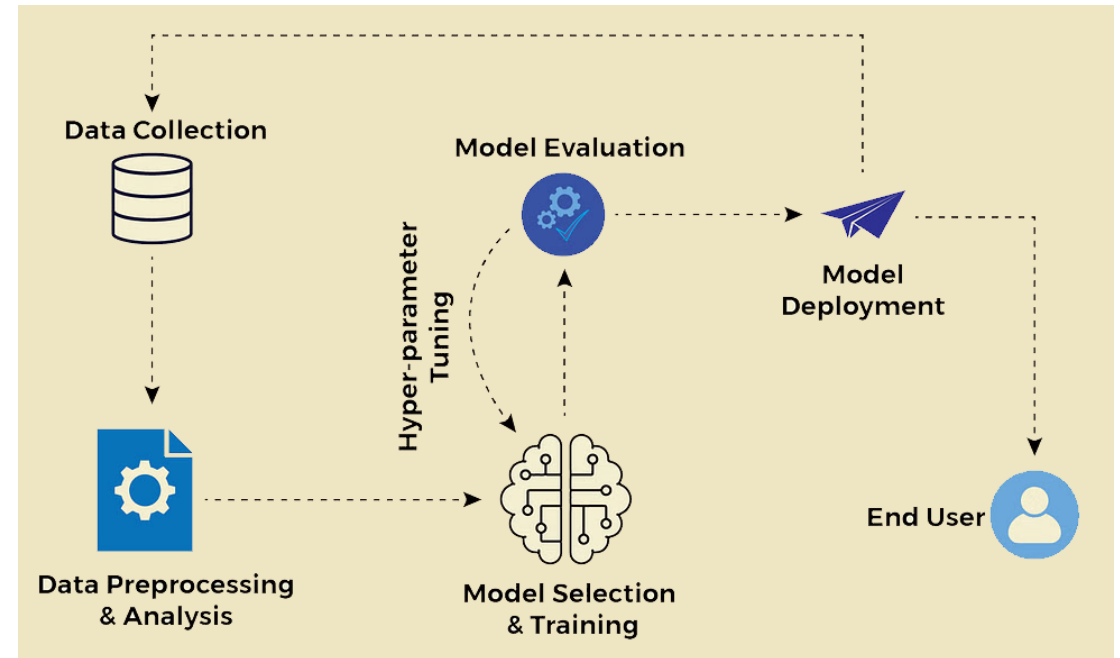
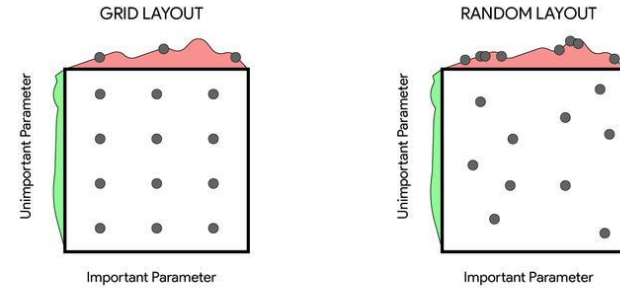
Randomly samples hyperparameter combinations to find the best results.

Automated Tools:

Libraries like Scikit-learn's **GridSearchCV** or **RandomizedSearchCV**.

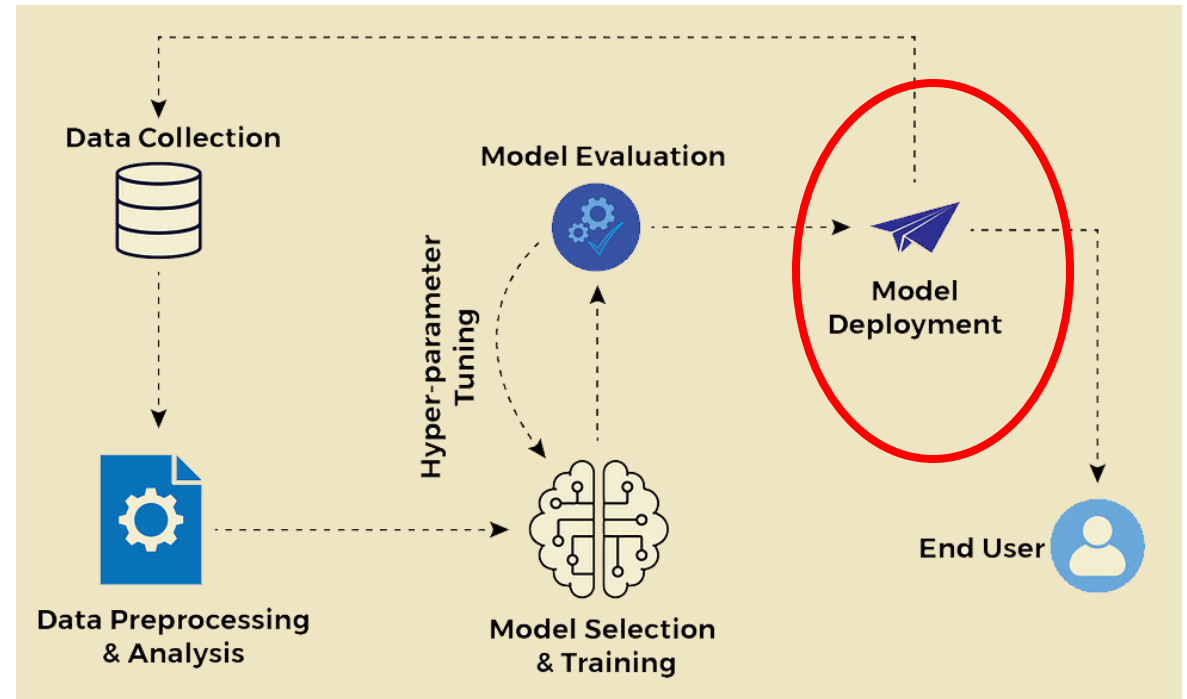
Example Parameters to Optimize:

- Learning rate for Gradient Boosting.
- Number of clusters in K-Means.



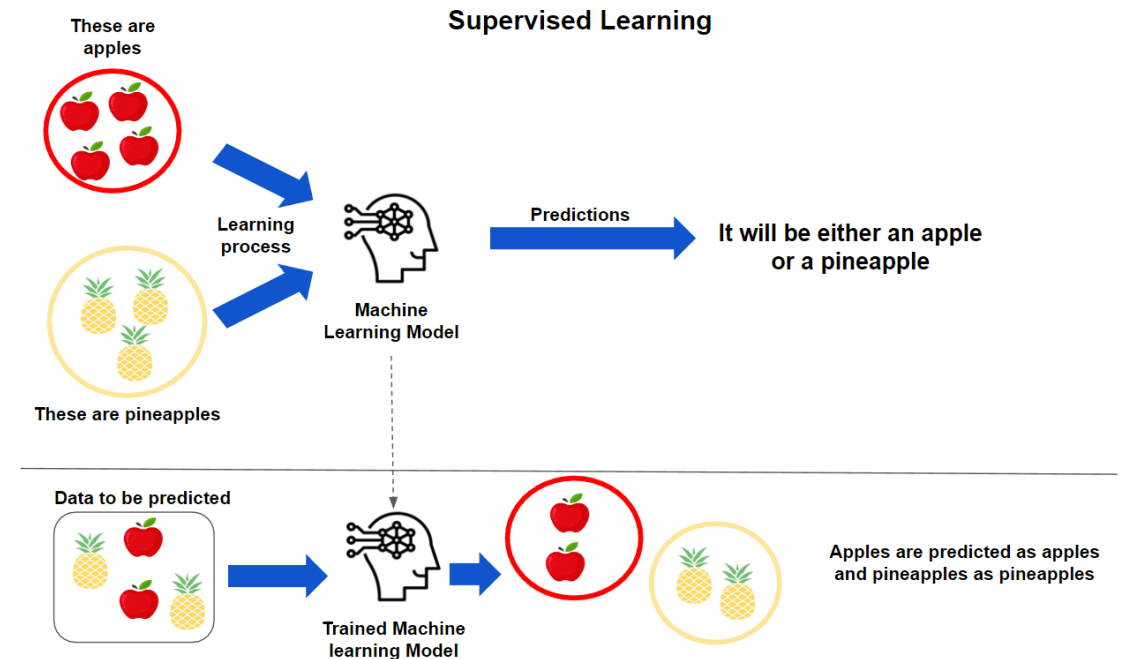
Deploy the Model

- Integrate the trained model into an application or business workflow.
- Monitor for model drift or performance degradation over time.



Supervised learning

- Supervised learning is the most common type of machine learning.
- It is the first type of learning that most people encounter.
- It is the type of learning that is used to train a model to predict an output based on an input.
- The input is called the **feature X** and the output is called the **label Y**.
- The model is trained on a dataset that contains both the features and the labels.
- The model learns the relationship between the features and the labels and uses this relationship to make predictions on new data.



Supervised learning

- Two types: Classification vs Regression

Regression



What will be the temperature tomorrow?

84°



Fahrenheit

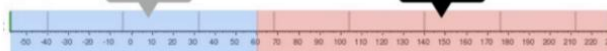
Classification



Will it be hot or cold tomorrow?

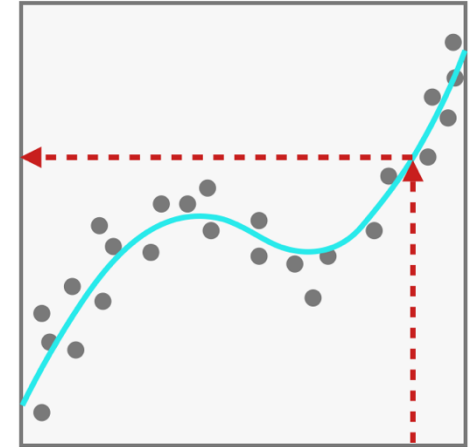
COLD

HOT



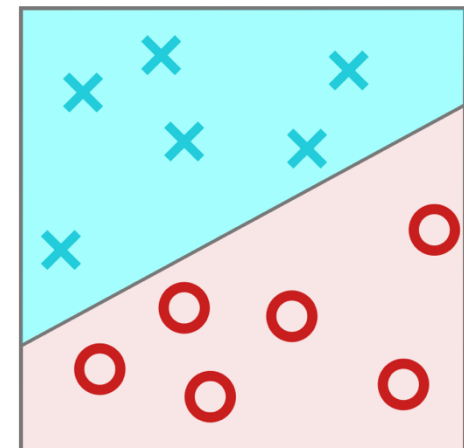
Fahrenheit

Regression predicts a numeric value



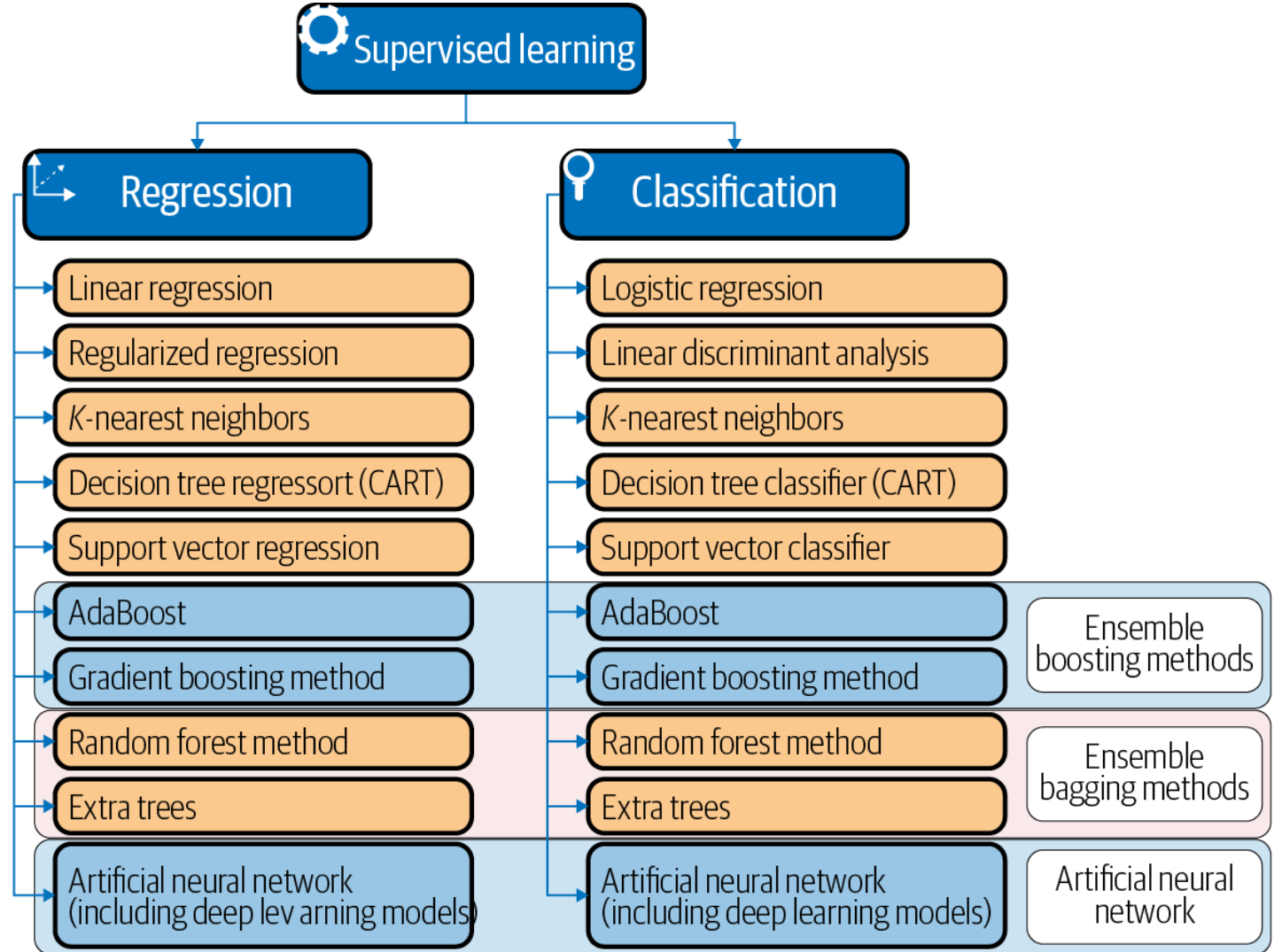
Here, the fitted line provides a predicted output, if we give it an input

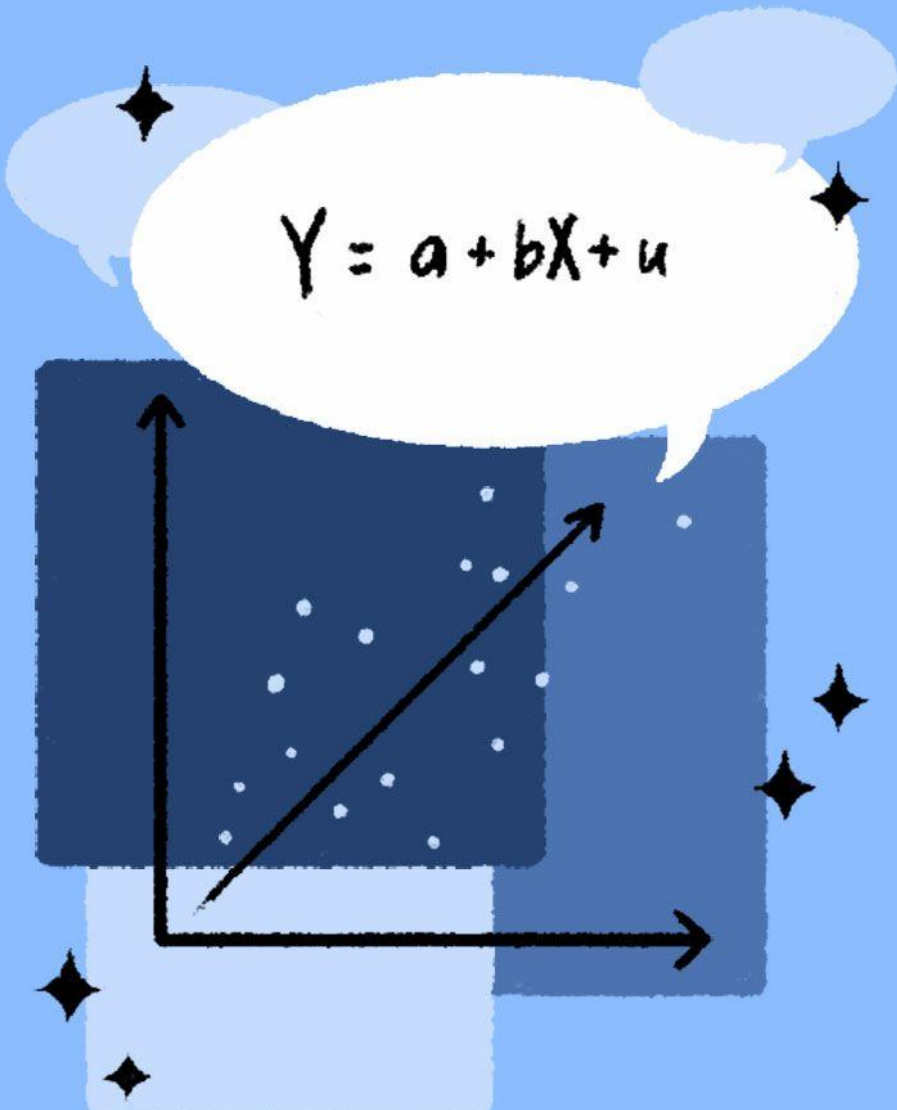
Classification Groups observations into "classes"



Here, the line classifies the observations into X's and O's

Supervised learning algorithms




$$Y = a + bX + u$$

Regression

[ri-'gre-shən]

A statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Evaluate Regression models

Importance of Metrics: Metrics help determine how well the regression model predicts the target values. A good model minimizes errors and explains the variance in the target variable.

Common Metrics:

Mean Absolute Error (MAE):

- Measures the average absolute difference between predicted and actual values.
- Lower MAE indicates better model performance.

Formula: $MAE = 1/n * \sum |y - \hat{y}|$ where (n) is the number of samples, (y) is the actual value, and (\hat{y}) is the predicted value.

Mean Squared Error (MSE):

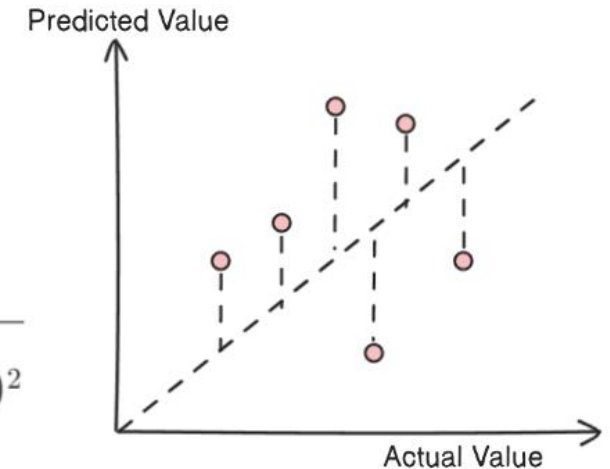
- Measures the average squared difference between predicted and actual values.
- Penalizes larger errors more heavily than MAE.
- Formula: $MSE = 1/n * \sum (y - \hat{y})^2$ where (n) is the number of samples, (y) is the actual value, and (\hat{y}) is the predicted value.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



Evaluate Regression models

Common Metrics:

Root Mean Squared Error (RMSE):

- The square root of MSE. Intuitive as it's in the same units as the target variable.
- Formula: $RMSE = \sqrt{MSE}$

R² Score (Coefficient of Determination):

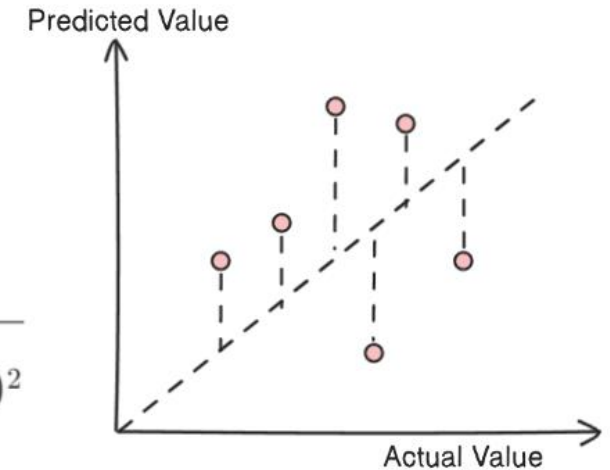
- Measures the proportion of variance in the target variable explained by the features.
- Values range from 0 (no explanation) to 1 (perfect fit).
- Formula: $R^2 = 1 - (\sum (y - \hat{y})^2 / \sum (y - \bar{y})^2)$
where (y) is the actual value, (\hat{y}) is the predicted value, and (\bar{y}) is the mean of the actual values.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



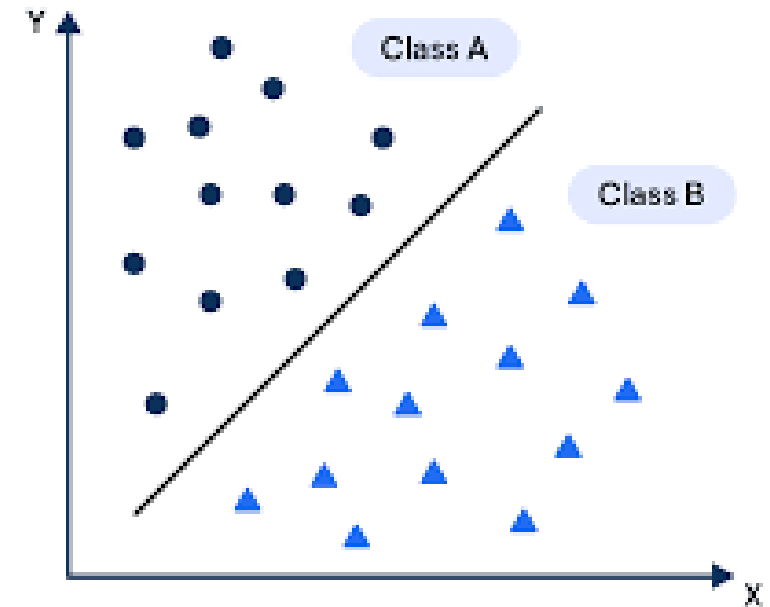
Classification

Classification is a supervised learning technique where the goal is to predict a discrete category (class) for an input.

Examples of Classification:

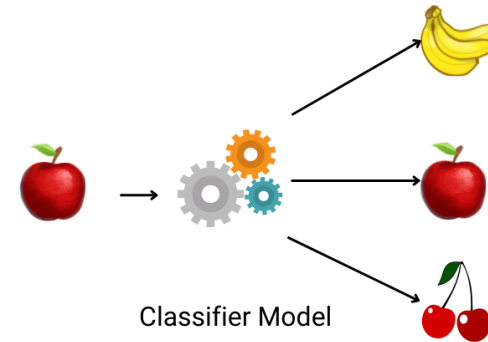
- Spam Detection:
 - Input: Email text
 - Output: Spam or Not Spam (binary classification).
- Disease Diagnosis:
 - Input: Patient symptoms.
 - Output: Disease type (multi-class classification).
- Image Recognition:
 - Input: Image pixels.
 - Output: Multiple objects in the image (multi-label classification).

Classification Algorithm



How Classification Works:

- Train the model on labeled data (inputs with corresponding class labels).
- Learn decision boundaries or probabilities for each class.
- Use the trained model to classify new, unseen inputs.



Predicted Probability Distribution		Predicted Probability Distribution
0.1		0
0.9	↔ Entropies measure this difference ↔	1
0.5		0

Classification types

Three Type of Classification Tasks

YAHOO!
JAPAN

Binary Classification



- Spam
- Not spam

Multiclass Classification



- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Multi-label Classification



- Dog
- Cat
- Horse
- Fish
- Bird
- ...

Evaluate Classification models

- True Positive (TP): Correctly predicted positive samples.
- False Positive (FP): Incorrectly predicted positive samples.
- True Negative (TN): Correctly predicted negative samples.
- False Negative (FN): Incorrectly predicted negative samples.

1 Accuracy:

Percentage of correctly predicted samples.

Limitation: Can be misleading for imbalanced datasets.

2 Precision:

Of all the predicted positives, how many are truly positive?

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Evaluate Classification models

3 Recall (Sensitivity):

Of all the actual positives, how many were correctly predicted?

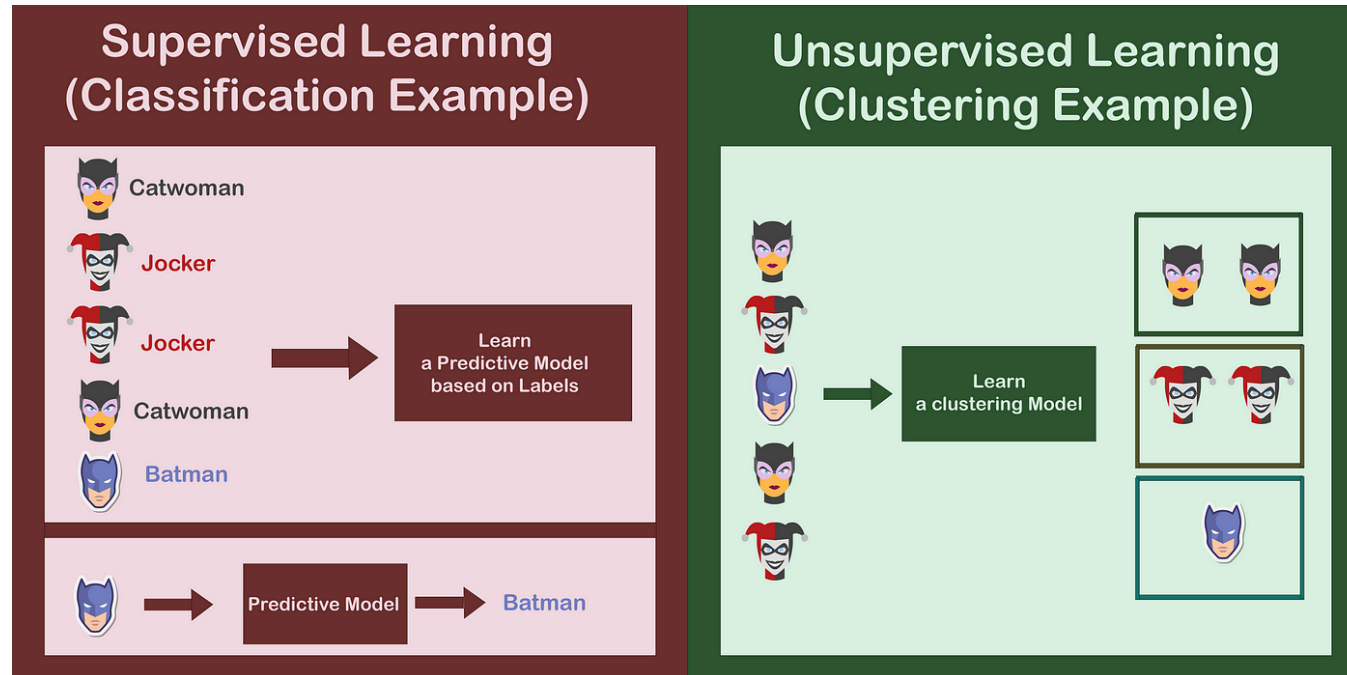
4 F1-Score:

Harmonic mean of precision and recall.
Useful when you want to balance both.

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

Unsupervised Learning

Unsupervised learning is a type of machine learning that looks for previously undetected patterns in a data set with **no pre-existing labels** and with a minimum of human supervision.

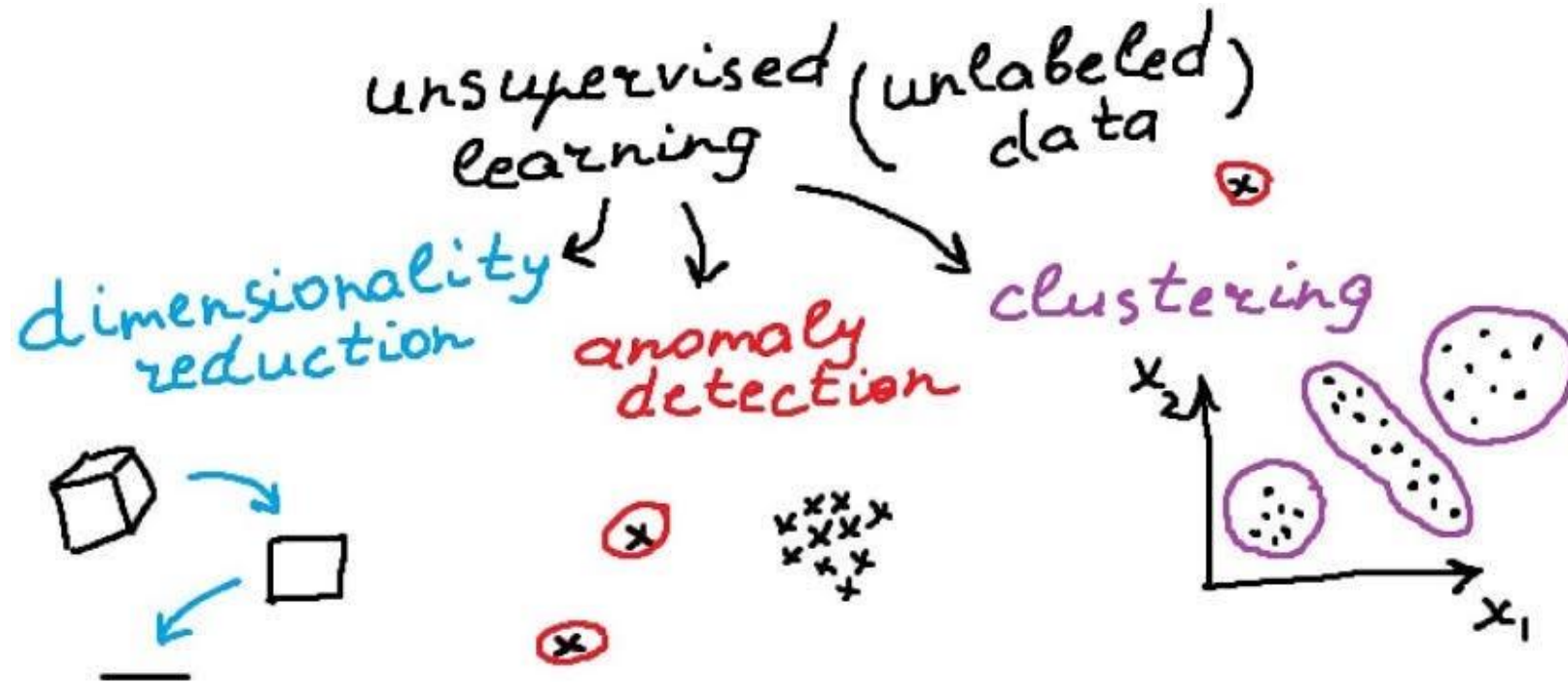


Types of Unsupervised Learning

Clustering is a type of unsupervised learning where the goal is to group data points into clusters based on their similarity.

Dimensionality reduction reduces the number of input features while retaining the essential patterns in the data.

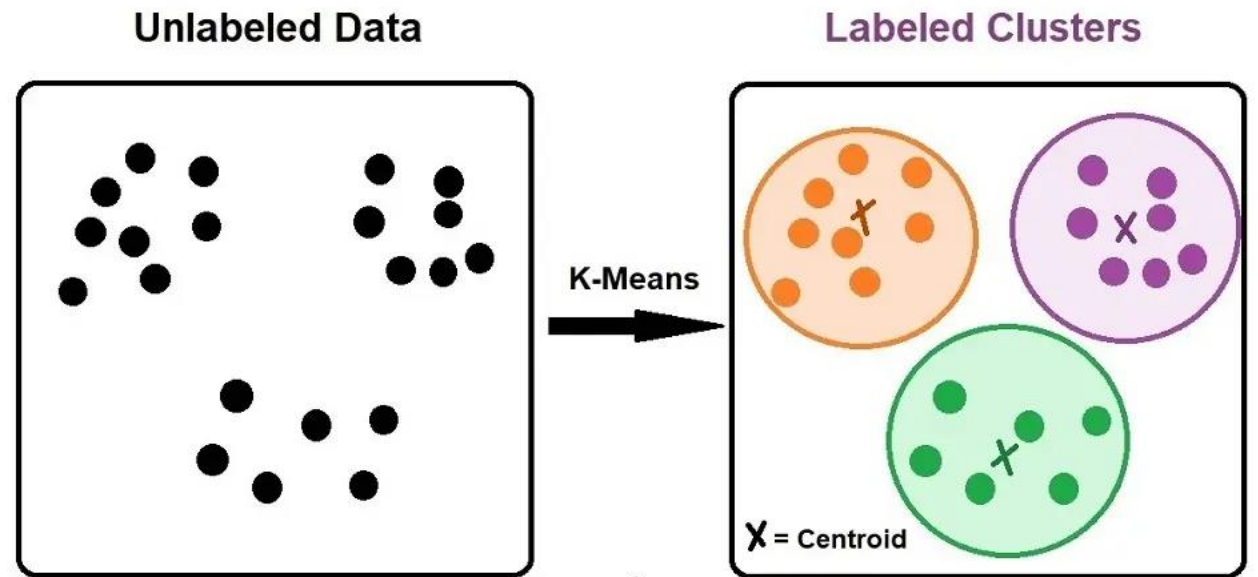
Anomaly detection identifies data points that deviate significantly from the normal patterns.



Clustering

How Clustering Works:

- Compute a similarity or distance measure (e.g., Euclidean distance, cosine similarity) between data points.
- Group similar data points into clusters.
- Evaluate the compactness and separation of clusters.



Types of Clustering

1 Partition-Based Clustering:

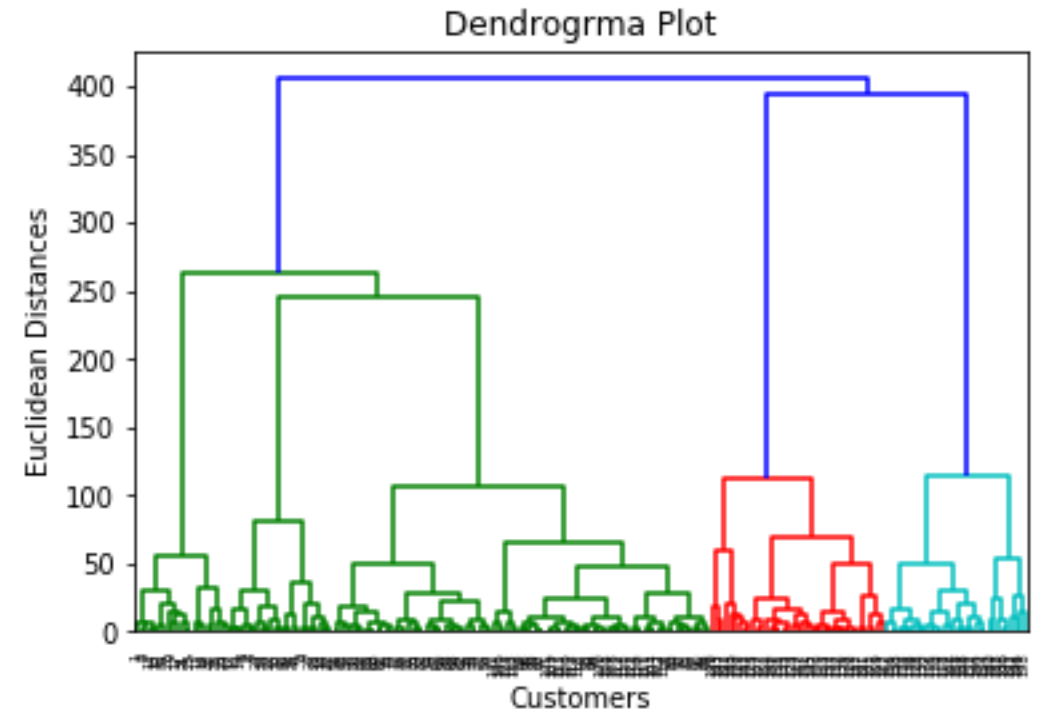
- Divides data into non-overlapping subsets.
- Example Algorithm: **K-Means** Clustering.
- Application: Customer segmentation.

2 Hierarchical Clustering:

- Builds a **tree-like structure of clusters (dendrogram)**.
- Can be agglomerative (bottom-up) or divisive (top-down).
- **Application:** Organizing documents into categories.

3 Density-Based Clustering:

- Groups data points based on dense regions.
- Example Algorithm: **DBSCAN**.
- **Application:** Detecting anomalies or outliers.



DBSCAN



k-means



Clustering algorithms

Algorithm	Handles Noise	Requires k	Suitable for Large Data	Cluster Shape Flexibility
K-Means	No	Yes	Moderate	Spherical
Mini-Batch K-Means	No	Yes	Yes	Spherical
Affinity Propagation	No	No	No	Arbitrary
Mean Shift	Yes	No	No	Arbitrary
Spectral Clustering	No	Yes	No	Complex
Agglomerative	No	No	Moderate	Arbitrary
DBSCAN	Yes	No	Moderate	Arbitrary
OPTICS	Yes	No	Moderate	Arbitrary
Birch	No	Yes	Yes	Spherical
Gaussian Mixture	No	Yes	Moderate	Gaussian

Evaluate Clustering Results

1 Silhouette Score:

- Measures how similar a point is to its cluster compared to other clusters.
- Ranges from -1 to 1:
 - 1: Perfect clustering.
 - 0: Overlapping clusters.
 - -1: Points assigned to the wrong cluster.

2 Inertia (Sum of Squared Distances):

- Measures the compactness of clusters.
- Lower inertia indicates tighter clusters.

3 Visualization:

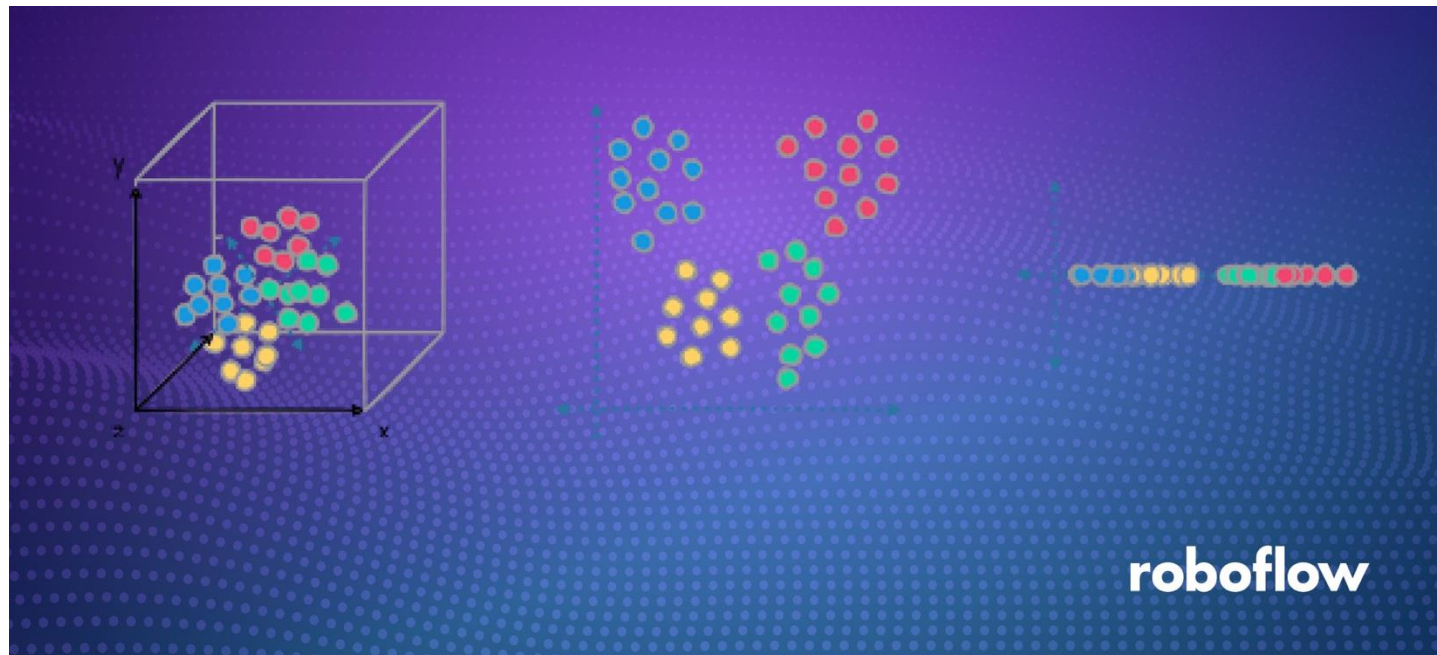
- Use scatterplots or dendrograms to visualize clusters and their separations.

Dimensionality Reduction

Dimensionality reduction reduces the number of input features while retaining the essential patterns in the data.

It is useful for:

- Simplifying datasets.
- Visualizing high-dimensional data.
- Speeding up computations.



Principal Component Analysis - PCA

PCA is one of the most popular dimensionality reduction techniques.

Key Steps:

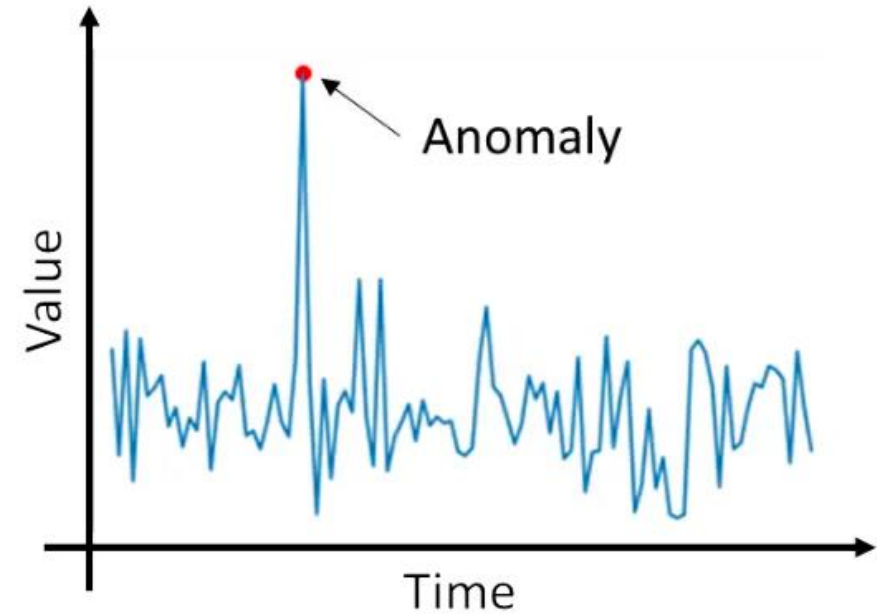
- Standardize the dataset (mean = 0, variance = 1).
- Compute the covariance matrix.
- Identify the principal components (eigenvectors of the covariance matrix).
- Project the data onto the new subspace defined by the top principal components.

Anomaly detection

Anomaly detection identifies data points that deviate significantly from the normal patterns.

It is often used for:

- Fraud detection.
- Monitoring network traffic for intrusions.
- Detecting manufacturing defects.



Anomaly detection



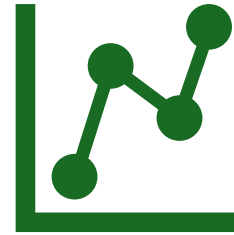
Techniques:

Density-Based Methods:

- DBSCAN (clustering-based).
- Isolation Forest (tree-based).

Statistical Methods:

- Z-scores: Data points with high Z-scores (e.g., > 3) are anomalies.



Applications:

Detecting fraudulent transactions in banking.

Identifying faulty sensors in IoT systems.

Finding outliers in customer behavior data.

Acknowledgements

Thanks to Leon Boschman for contributing his ideas, slides and feedback to this course material.