Machine Learning Algorithms

ISE4132 : AI Application System

AMIN AL (아민알)
Integrated System Engineering (ISE)
010-6853-6648
alaminanik@inha.ac.kr

Types of Machine Learning Algorithms

Supervised Learning

- Relies on labeled data
- Learns mapping from features → target

Unsupervised Learning

- Works on unlabeled data
- Finds hidden patterns or groupings

Supervised Learning

Two main tasks:

Classification:

- Categorical outputs
  Example: fraud / not fraud, Loan paid back → *Yes / No*

Regression:

- Continuous outputs
  Example: Predicting income in dollars

Supervised Learning

Supervised algorithms rely on human knowledge to complete their tasks. For example, we have a dataset related to loan repayment that contains several demographic indicators, as well as whether a loan was paid back or not:

Income	Age	Marital Status	Location	Savings	Paid
$90,000	28	Married	Austin, Texas	$50,000	y

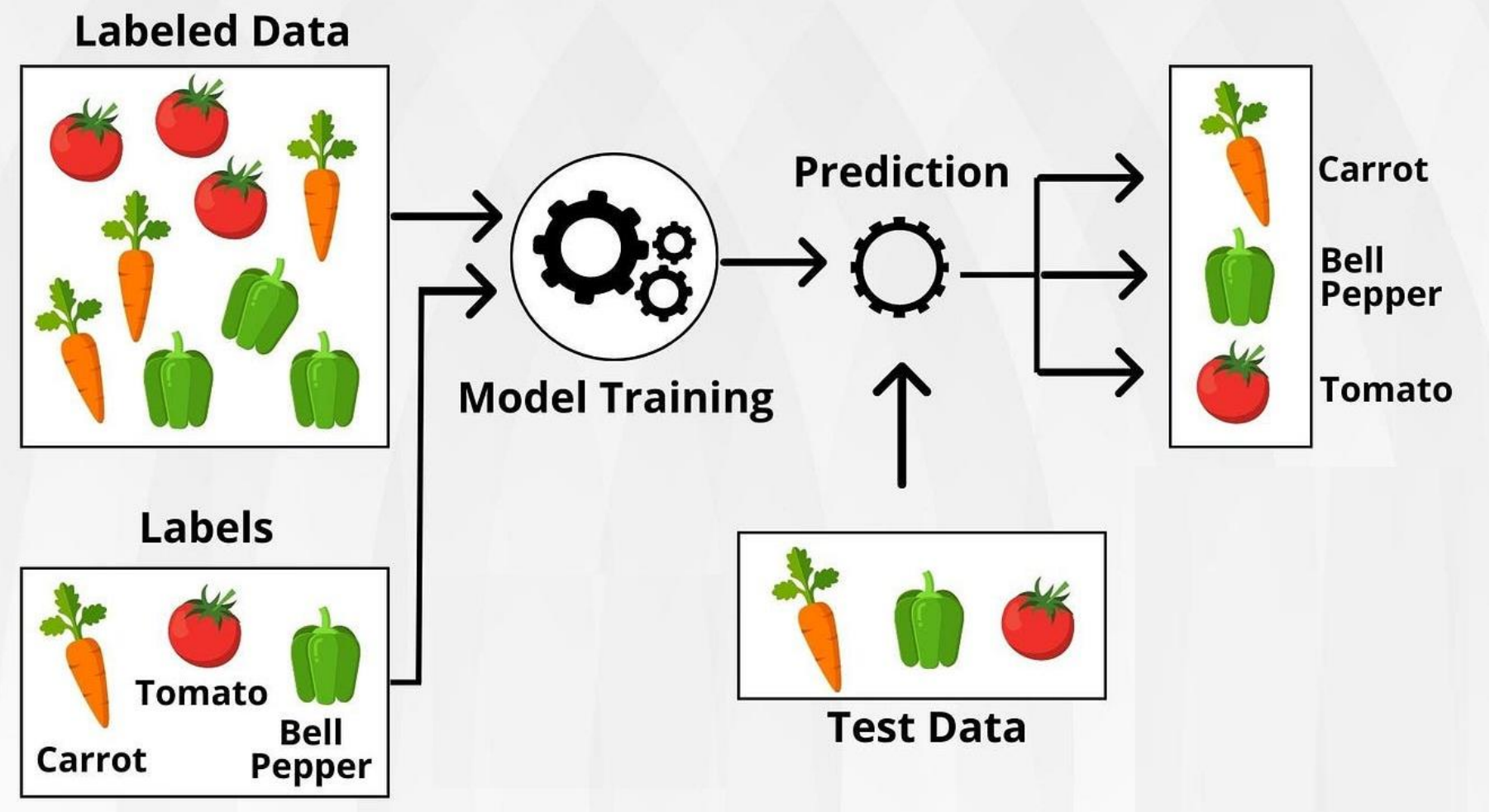
Supervised Learning

In supervised learning, algorithms learn to predict the target based on the features, or in other words, what indicators give a high probability that an applicant will pay back a loan or not?

$$y = f(x) + \epsilon$$

Here, label y is a function of the input features x, plus some amount of error ϵ that it caused naturally by the dataset

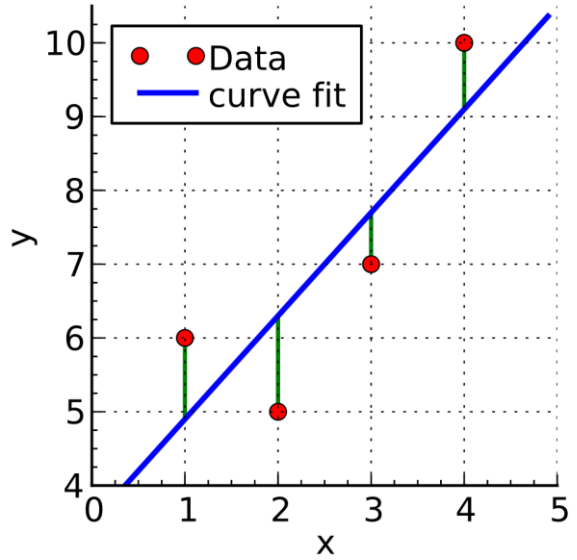
Supervised Learning



Linear Regression

Linear regression is a statistical method that models the linear relationship between a dependent variable and one or more independent variables, allowing for predictions and analysis of how variables influence each other.

- Fit a straight line (or plane) to data points.
- Predicting house prices from size, number of rooms, etc.



$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon$$

y : Predicted output

x_i : Features

β_i : model coefficients

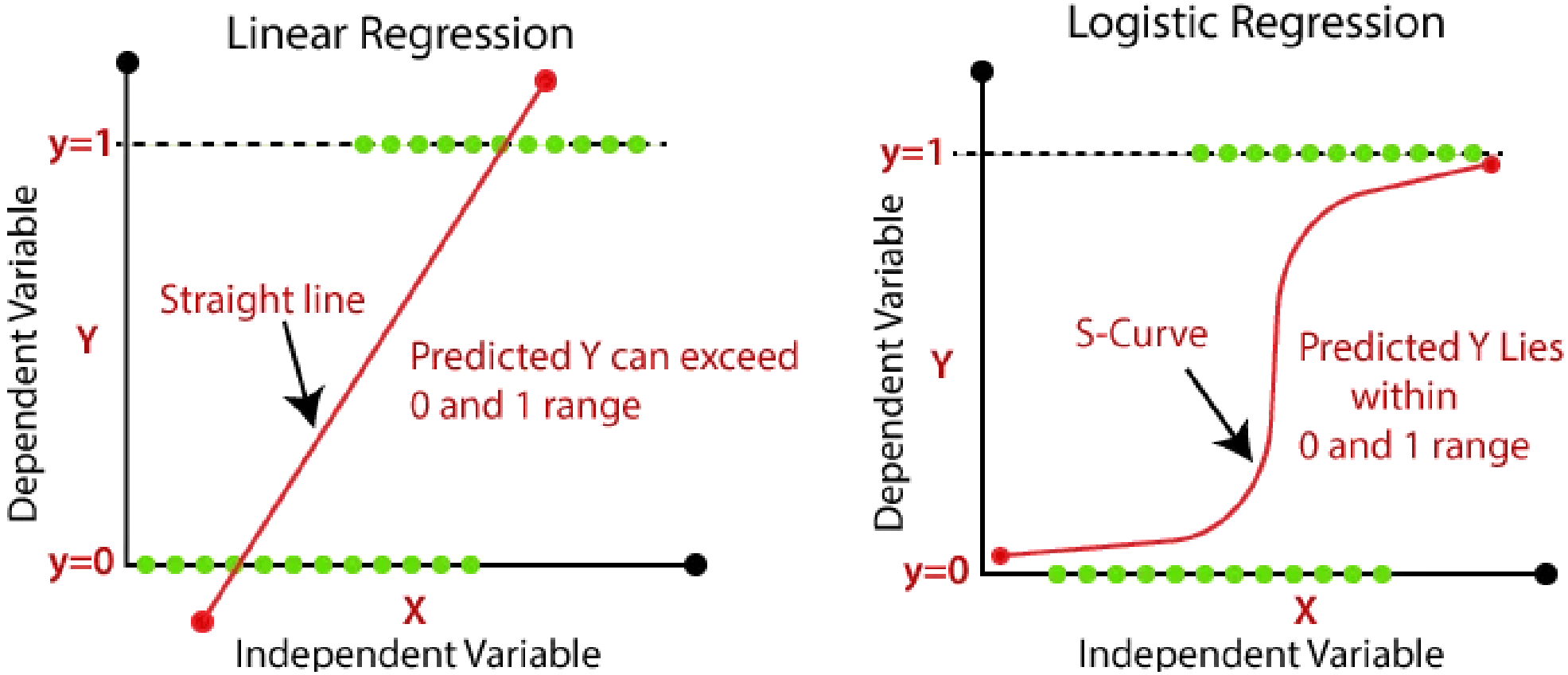
ϵ : error

parameters (numbers) within a statistical or machine learning model that quantify the relationship between input variables (features) and the model's output (target variable)

Logistic Regression

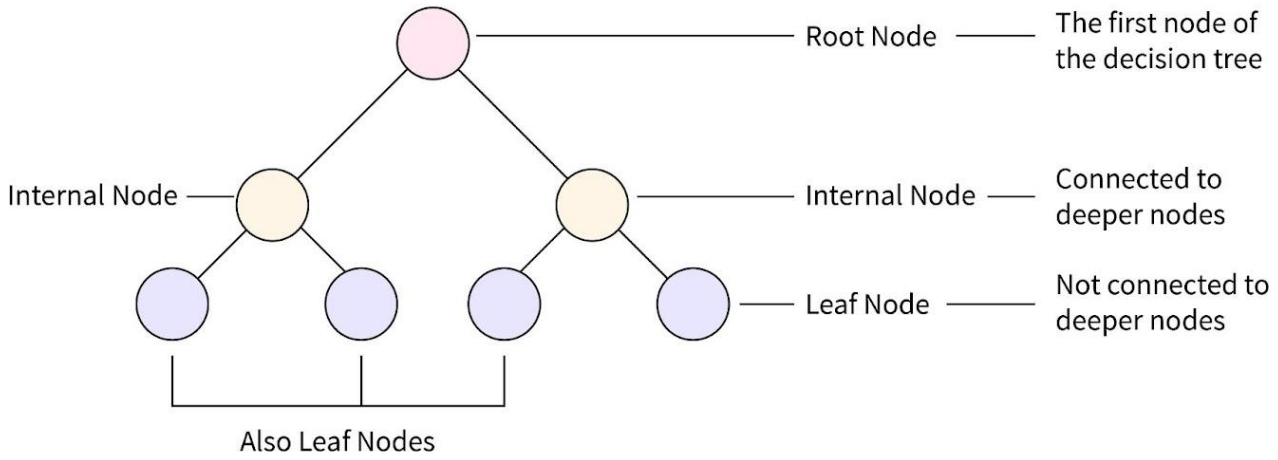
- Logistic regression is a supervised machine learning algorithm used for classification problems.
- Despite the name, it’s for classification, not regression.
- Predicts probability of belonging to a class.
- Example: Will a student pass the exam? (Yes/No).
- Output probability is between 0 and 1

Linear Regression vs Logistic Regression

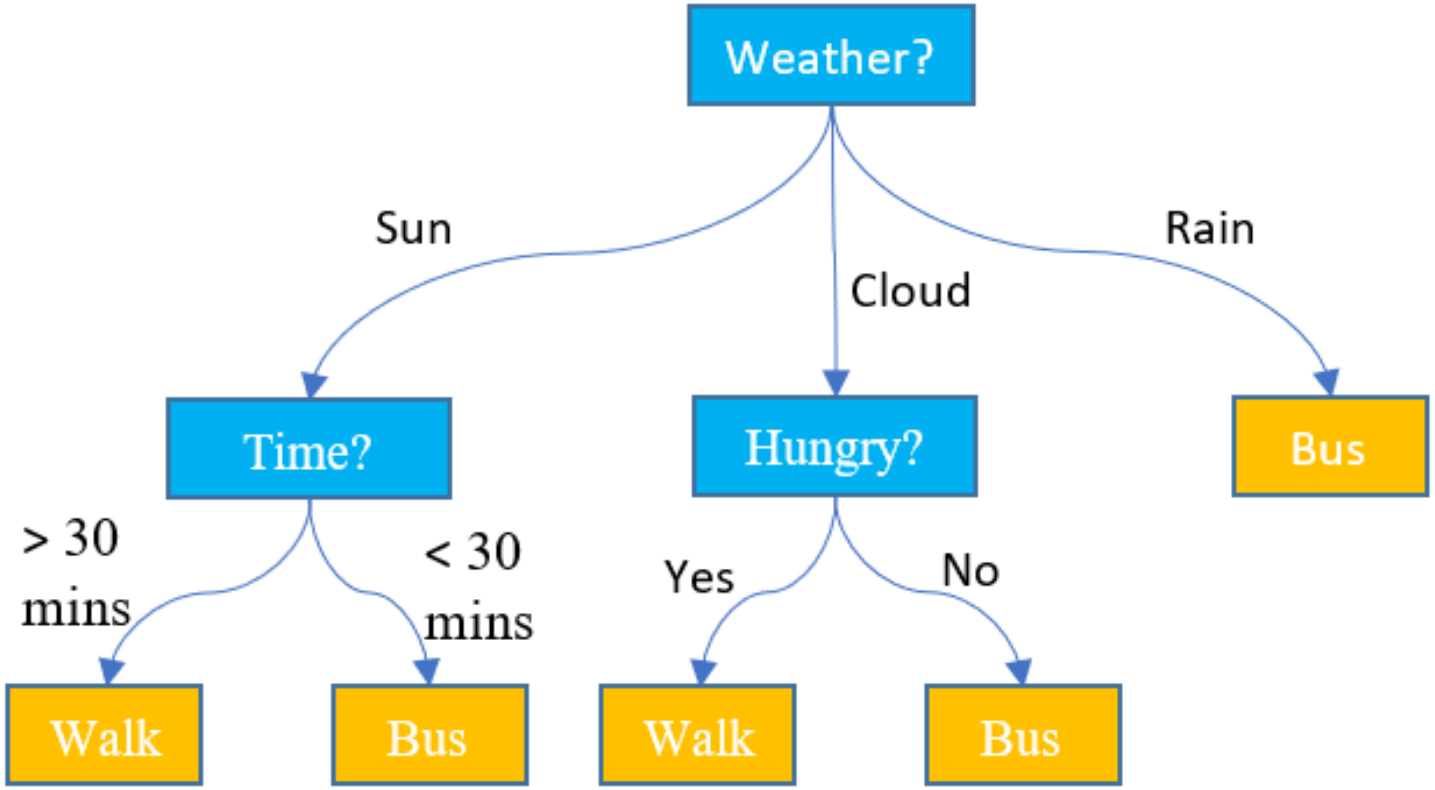


Decision Trees

- Tree-structured model for classification/regression.
- Model works like a flowchart (questions → answers).
- Example: Predict whether someone will buy a car:
  - Is income > $50k?
  - Is age < 30?
  - … then predict YES/NO.



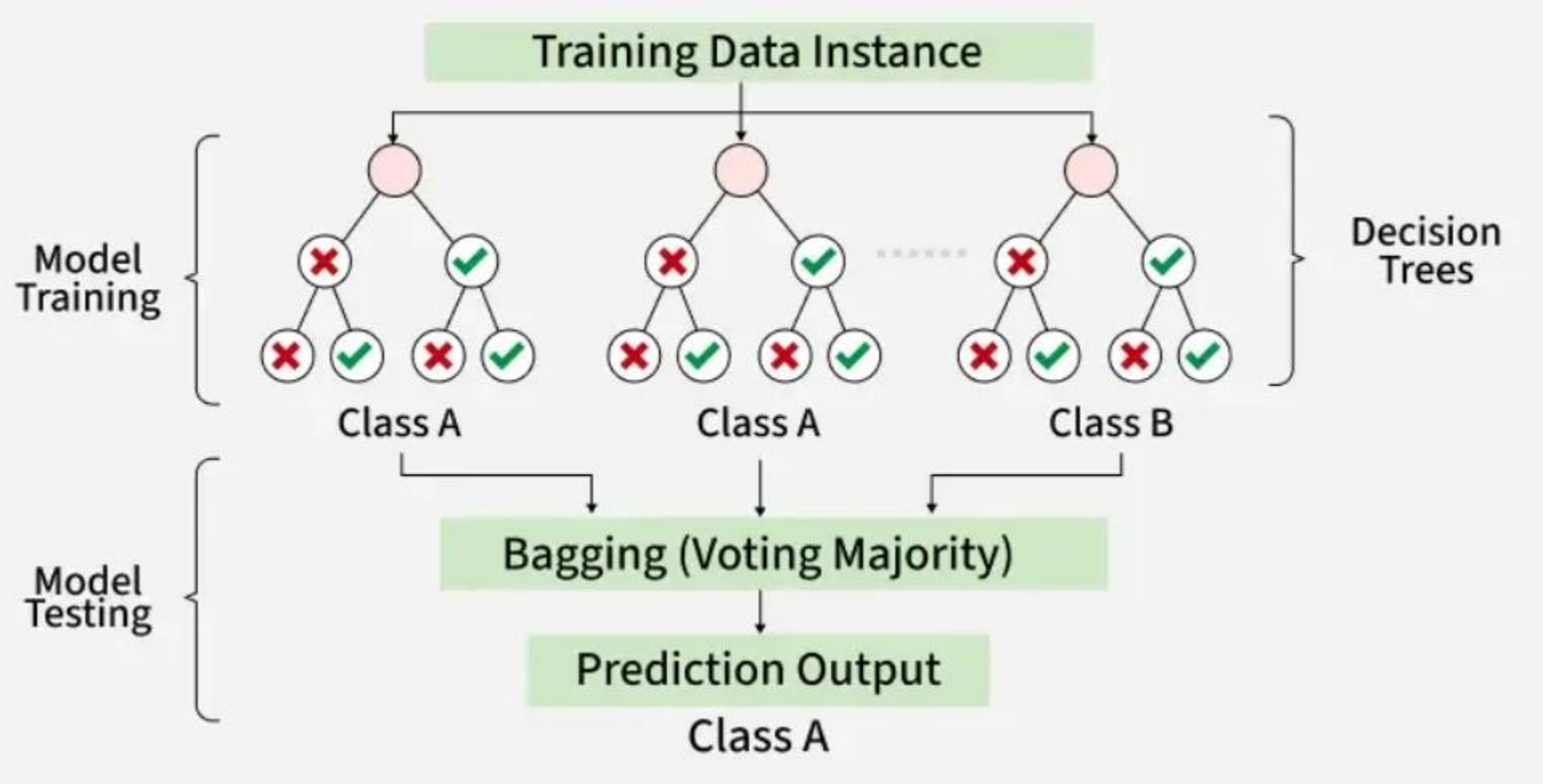
Decision Trees



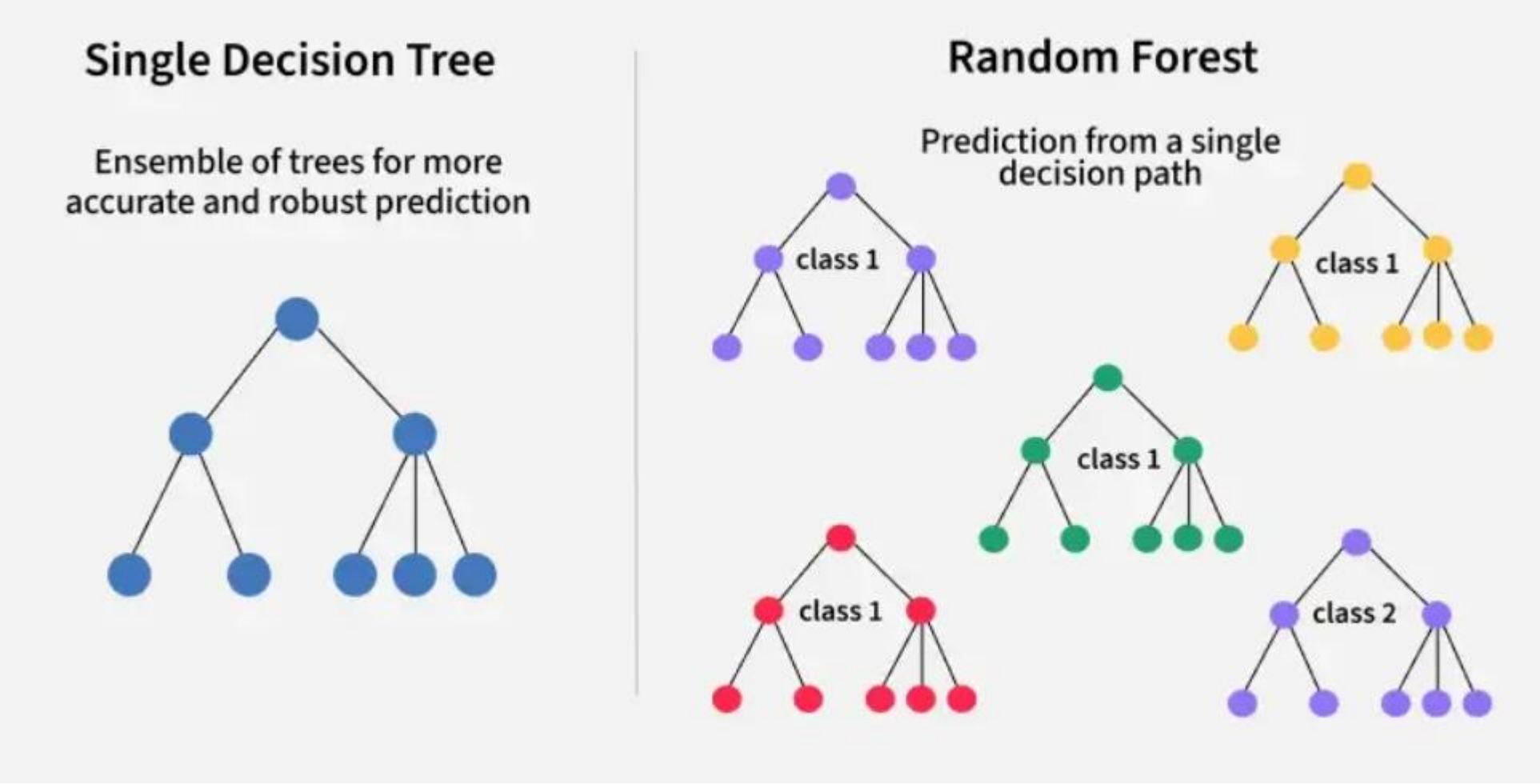
Random forests

- Build many decision trees and combine them.
- Each tree is trained on random samples and random features.
- Final prediction:
  - **Classification** → majority vote.
  - **Regression** → average of predictions.
- Example: Multiple doctors giving opinions → majority wins.
- More accurate and stable than a single decision tree.

Random forests



Decision Tree vs Random forests



Comparison of Supervised Learning Models

Model	Type of Output	Best For	Strengths	Weaknesses
Linear Regression	Continuous values (numbers)	Predicting house prices, sales, growth trends	Simple, easy to interpret, fast	Only works well for linear relationships
Logistic Regression	Categories (Yes/No, Multi-class)	Pass/Fail prediction, disease detection	Probabilistic output, easy to train	Not great for complex, non-linear data
Naive Bayes	Categories	Spam filtering, text classification	Very fast, works with small data, good for text	Assumes independence of features
Support Vector Machine (SVM)	Categories or continuous	Image recognition, bioinformatics	Handles complex data, works in high dimensions	Slow with large datasets
Decision Tree	Categories or continuous	Customer decision making, risk analysis	Easy to understand, visual, interpretable	Overfits easily, unstable with small changes
Random Forest	Categories or continuous	Credit scoring, fraud detection, medical diagnosis	Accurate, reduces overfitting, robust	Harder to interpret, more computationally heavy

Classification and Regression Models

These predict **labels or categories** (e.g., Yes/No, Red/Blue/Green).

- Logistic Regression

- Naive Bayes

- Support Vector Machine (SVM) (classification version)

- Decision Tree (classification version)

- Random Forest (classification version)

These predict **numerical values** (e.g., salary, temperature, sales).

Linear Regression

Support Vector Machine (SVM) (regression version → SVR)

Decision Tree (regression version)

Random Forest (regression version)

Unsupervised Learning

- Computer learns patterns from data without labels (no “right answers”).

- Goal: Discover hidden structure in data.

Two main tasks:

Clustering → Group similar data points.
Dimensionality Reduction → Simplify data while keeping important info.

Example:

Netflix grouping users by viewing habits.
Google News clustering similar articles.

Clustering

Definition: Clustering is an unsupervised machine learning algorithm that organizes and classifies different objects, data points, or observations into groups or clusters.

K-Means Clustering:

- Divide data into k groups.
- Simple and widely used.

Clustering Examples

Document Classification
Cluster documents in multiple categories based on tags, topics, and the content of the document. this is a very standard classification problem and k-means is a highly suitable algorithm for this purpose.

Customer Segmentation
Clustering helps marketers improve their customer base, work on target areas, and segment customers based on purchase history, interests, or activity monitoring.

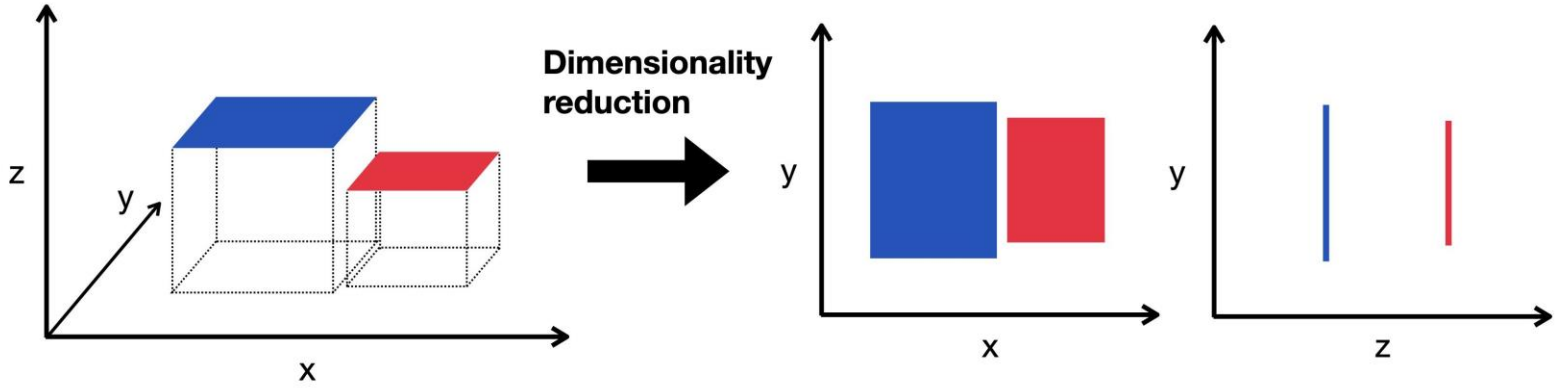
Insurance Fraud Detection
Machine learning has a critical role to play in fraud detection and has numerous applications in automobile, healthcare, and insurance fraud detection. utilizing past historical data on fraudulent claims, it is possible to isolate new claims based on its proximity to clusters that indicate fraudulent patterns. since insurance fraud can potentially have a multi-million dollar impact on a company, the ability to detect frauds is crucial.

Dimensionality Reduction

Definition: Reduce number of features while keeping patterns.

Example:
Instead of 100 exam questions, find 2–3 main “skills” that explain performance.

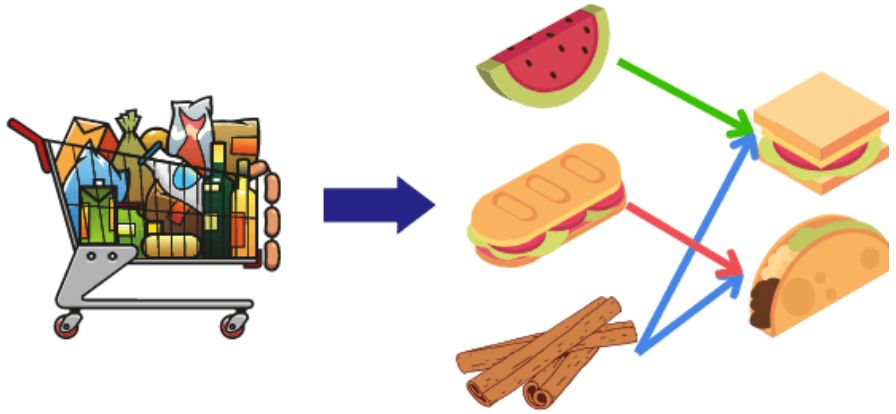
Common method: Principal Component Analysis (PCA)
- Creates new features (principal components).
- Captures maximum variance in fewer dimensions.
- Helps in visualization and reduces computational cost.



Association Rule Learning

Definition: Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases.

Example:
Market Basket Analysis (Amazon, Walmart): If a customer buys bread, they often buy butter.



"93% of people who purchased item A also purchased item B"

Advantages and Limitations of Unsupervised Learning

Advantages:

- Works with unlabeled data (cheap and available).
- Useful for exploring data and finding hidden patterns.
- Helps in recommendation systems, market segmentation, anomaly detection.

Limitations:

- Harder to evaluate accuracy (no labels to compare).
- Results may be less interpretable.
- Choice of algorithm and parameters (like number of clusters) can strongly affect results.

Supervised vs Unsupervised Learning

Feature	Supervised Learning	Unsupervised Learning
Data	Data is **labeled** → each input has a correct output (X → Y)	Data is **unlabeled** → only inputs (X), no outputs (Y)
Goal	Learn a function to **predict outcomes** on new data	Find **patterns, groups, or structures** in the data
Output Type	Classification (Yes/No, categories) OR Regression (numbers)	Clusters (groups) OR Reduced features (simpler representation)
Examples	Predict: will a loan be paid back? (Yes/No) → classification. Predict house price → regression.	Group customers by shopping style. Detect unusual network activity.

Class Activity

Discussion

Create features (inputs) and possible labels (outputs) for each scenario.

1. Predicting student exam scores → Regression

2. Predicting if an email is spam → Classification

3. Predicting house prices → Regression

4. Predicting if a patient has a disease → Classification

Class Activity

Decision Tree Flowchart Game - draw a decision tree

Should I carry an umbrella today?
Features: Weather (Sunny, Rainy, Cloudy), Forecast (High/Low chance of rain), Temperature, Season (Summer/Winter)

Should I go shopping this weekend?
Features: Budget, Time available, Weather, Sale ongoing (Yes/No)

Should I go jogging this morning?
Features: Weather (Sunny/Rainy/Cloudy), Temperature, Time available (Yes/No), Mood (Energetic/Tired)

Class Activity

Discussion

- Why might Random Forests be better than a single decision tree?

- When would you use clustering instead of classification?

Class Activity

Discussion

1. Look at the Age, Income, and Spending Score columns.
2. Try to group customers manually into 2–3 clusters based on similarity.

Example clusters:
- *Young, low-income, high spending*
- *Middle-aged, medium-income, medium spending*
- *Older, high-income, low spending*

3. Explain why you grouped them in that way.

Customer ID	Age	Income (in $1000s)	Spending Score (1–100)	Favorite Product Category
1	22	25	80	Electronics
2	45	75	40	Furniture
3	19	15	90	Clothing
4	34	55	60	Electronics
5	50	80	30	Furniture
6	28	40	70	Clothing
7	41	65	50	Electronics
8	23	30	85	Clothing
9	36	60	55	Electronics
10	48	78	35	Furniture
11	26	35	75	Electronics
12	20	20	95	Clothing