# Simplified AI Lecture 5 Summary: Learning and Classification

## What is Learning in AI?

* Learning in AI is how artificial intelligence systems acquire knowledge or improve performance from data or experience.
* Instead of being explicitly programmed, AI systems learn from examples
* Machine learning models are the result of applying learning algorithms to data

## Important Machine Learning Terms

* **Machine Learning Algorithm**: A set of rules that guides how learning from data occurs (like a recipe)
* **Machine Learning Model**: The result of applying an algorithm to data (like the finished dish)
* Think of it this way:
  + Algorithm = cooking instructions
  + Data = raw ingredients
  + Model = final dish

## Types of Learning

### Supervised Learning => LABELED DATA

* The model learns from labeled data (input-output pairs)
* The AI is given both the question and the correct answer during training
* Examples: Email spam detection, image classification, disease diagnosis
* Algorithms: Linear regression, decision trees, support vector machines, neural networks

### Unsupervised Learning => UNLABELED

* The model finds patterns in unlabeled data on its own
* No correct answers are provided during training
* Examples: Customer segmentation, anomaly detection, topic modeling
* Algorithms: K-means clustering, PCA, autoencoders

### Semi-Supervised Learning => LABELED (SMALL AMOUNT) + UNLABELED (LARGE AMOUNT)

* Combines a small amount of labeled data with a large amount of unlabeled data
* Useful when labeling data is expensive or time-consuming.

### Reinforcement Learning

* An agent learns by interacting with an environment and receiving rewards or punishments
* The AI learns from its actions and consequences
* Examples: Game playing (like AlphaGo), robotics, self-driving cars
* Key terms: Agent, environment, reward, policy, value function

## Learning an Unknown Function

* **Objective**: Find a function that can predict outputs for new inputs.
* **Problem Setup**:
  + We have training data of input-output pairs: (x₁, y₁), (x₂, y₂), …, (xₙ, yₙ)
  + These pairs are produced by an unknown function: y = f(x)
* **Goal**: Discover a function f̂(x) that closely matches the true function f(x)
* **Key Concepts**:
  + Training Data: Used to build the model
  + Test Data: Used to evaluate the model’s performance on new data

## Types of Prediction Problems

### Classification

* The output variable is categorical (discrete classes/labels)
* The model places inputs into one of several categories
* Example: Classify emails as “Spam” or “Not Spam”
* Algorithms: Decision Trees, SVM, k-NN, Neural Networks

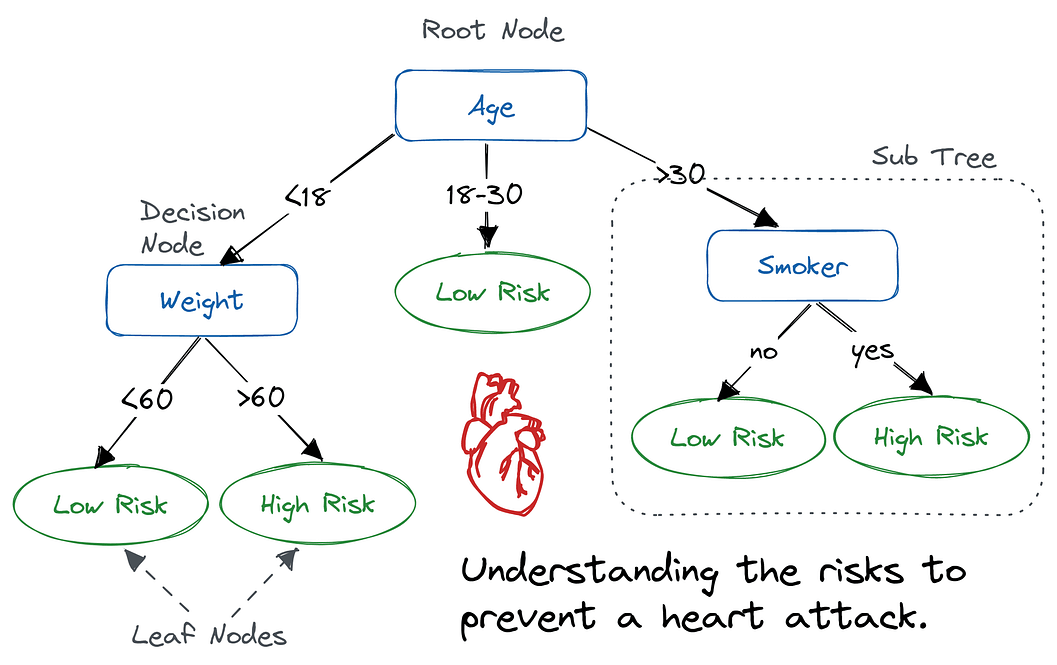
### Numeric Prediction (Regression)

* The output variable is continuous (real numbers)
* The model predicts numerical values
* Example: Predicting house prices, temperature, or stock prices
* Algorithms: Linear Regression, SVR, Random Forest Regression

## Classification Process

1. **Model Construction**:
   * Use training data with known class labels
   * Build a model using rules, decision trees, or mathematical formulas
2. **Model Usage**:
   * Apply the model to classify new, unseen data
   * Compare predictions with actual labels from test data
   * Calculate accuracy (% of correct predictions)
   * If accuracy is good, use the model in real-world applications

## Decision Trees



* A diagram representing possible solutions to a decision
* Shows different outcomes from a set of decisions
* Popular because they are easy to understand and interpret

### Structure of Decision Trees

* Takes input values (attributes) and gives one output (a decision)
* Starts at the root node and follows a path of yes/no questions
* At each step, it chooses a branch based on the answer
* Continues until it reaches a leaf node (the final decision)

### Learning a Decision Tree

* Goal: Find a tree with small error on the training data
* Process:
  1. Start with the root node
  2. At each step, split one of the leaves
  3. Repeat until a termination criterion is met

### Determining the Best Split

* We want child nodes to be more “pure” than the parent node
* Purity means nodes contain mostly examples of one class
* Methods to measure purity include:
  + Entropy: Measures uncertainty or disorder
  + Information Gain: Measures how much a split improves purity

### Advantages of Decision Trees

* Easy to understand and interpret
* Require minimal data preparation
* Can work with both numerical and categorical data
* Can be used with other decision-making tools
* Help visualize different possible outcomes

### Disadvantages of Decision Trees

* Can become too complex (overfitting)
* May create unrealistic models if based too much on expectations
* May focus too narrowly on certain decisions