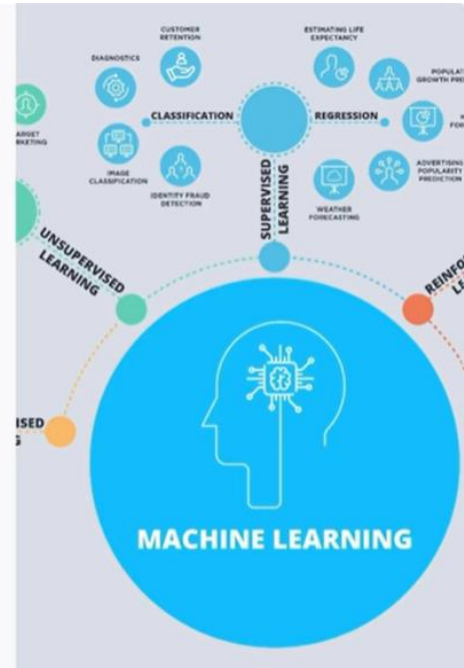


# Supervised learning (linear regression, logistic regression, decision trees, random forests, SVMs)

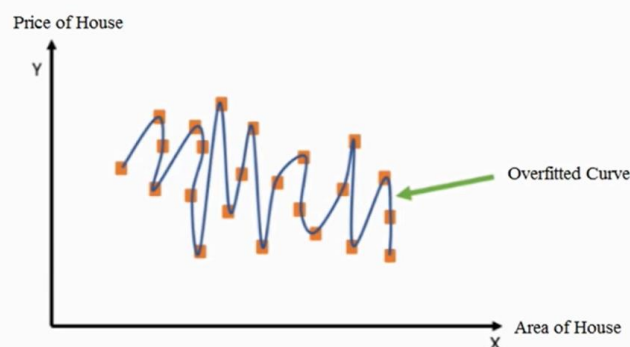


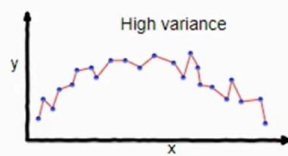
## Introduction to Supervised Learning

Supervised learning is a fundamental machine learning technique where an algorithm is trained on labeled data to make predictions on new, unseen data. This powerful approach has numerous applications, from spam filtering to medical diagnosis.

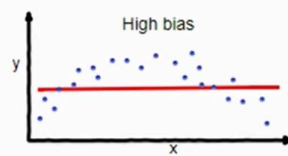


## Overfitting

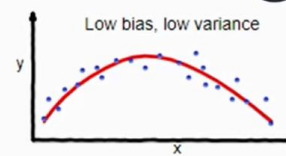




**overfitting**



**underfitting**



**Good balance**

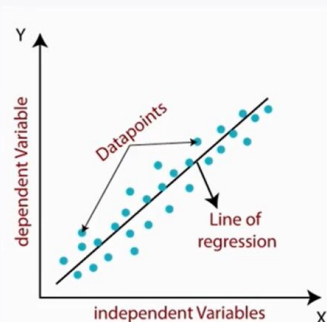
#### Reasons for Underfitting

- Too simple model
- Insufficient training
- Poor feature selection

#### Reasons for Overfitting

- Too complex model
- Too much training
- Limited training data

## Linear Regression: Understanding Continuous Relationships



#### Identifying Patterns

Uncover how variables are related

#### Predicting Outcomes

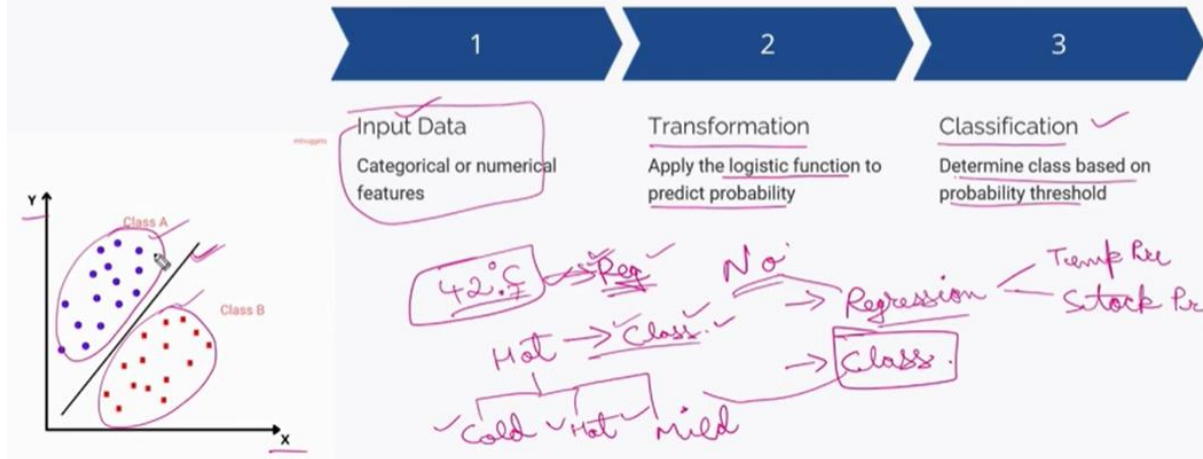
Forecast future values based on input data

#### Estimating Parameters

Determine the strength and direction of relationships

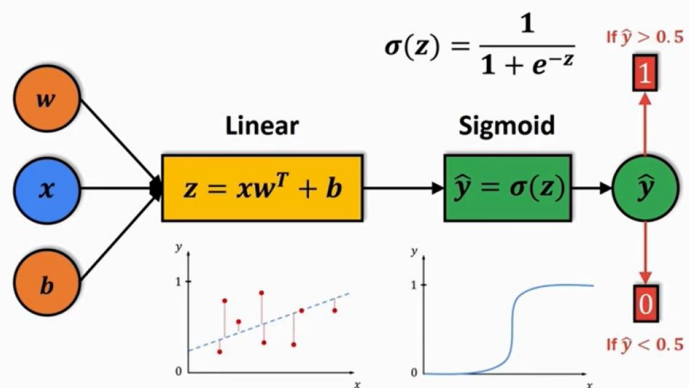
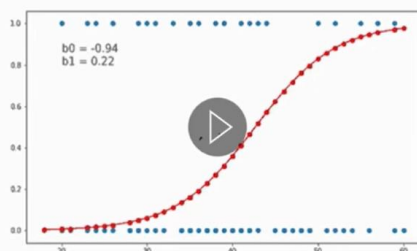
Linear regression is a powerful technique for understanding the continuous relationships between variables. By identifying patterns in data, linear regression allows you to build predictive models that can forecast future outcomes. The key is estimating the parameters that describe the strength and direction of the relationships between your input and output variables.

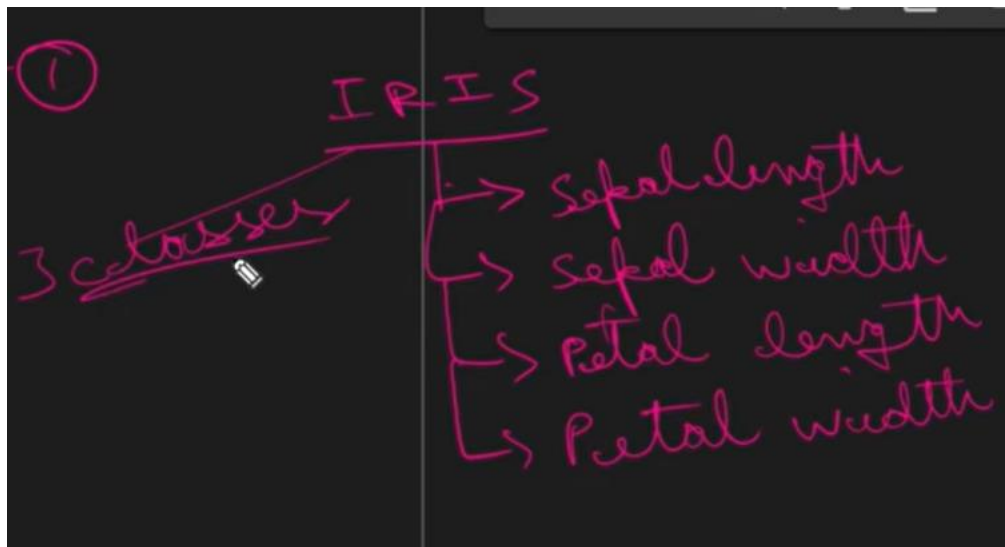
# Logistic Regression: Predicting Binary Outcomes



$$z = b_0 + b_1x_1 + \dots + b_rx_r$$

$$p(x) = \frac{1}{1 + e^{-z}}$$





## Decision Trees: Building a Hierarchical Model

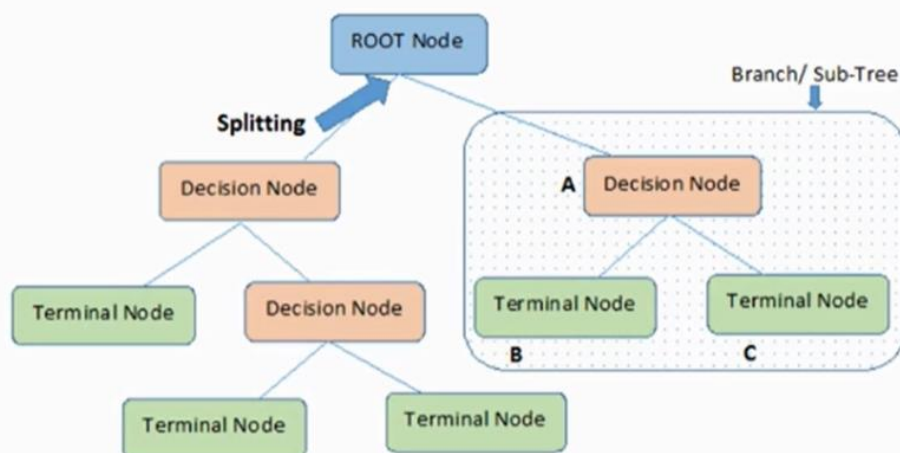
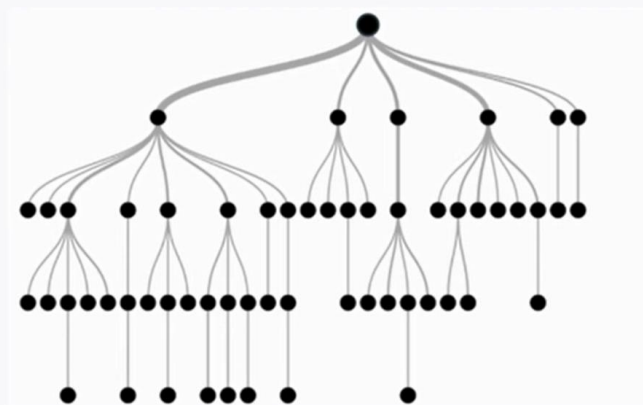


### Feature Selection

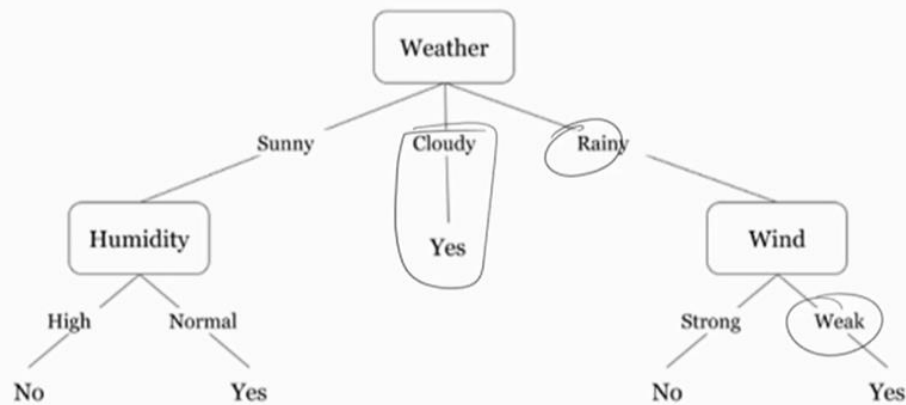
Identify the most important features that will drive the decision-making process. This is a crucial first step in constructing an effective decision tree.

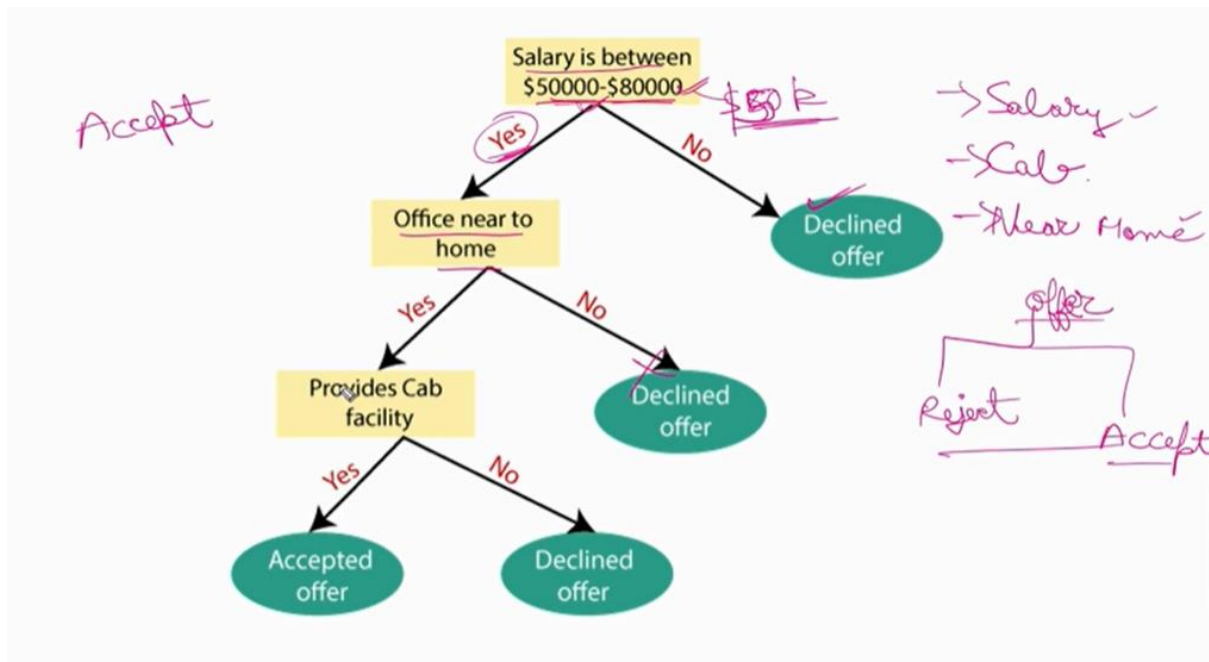
### Recursive Partitioning

The decision tree algorithm repeatedly splits the data based on the features, creating a hierarchical tree-like structure of decisions and outcomes.

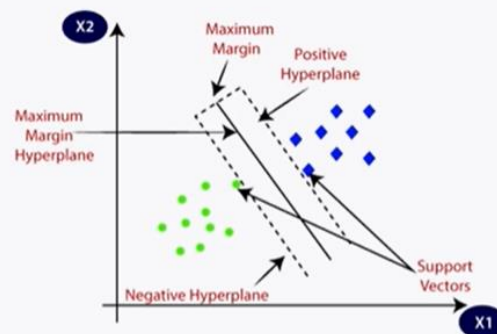


Day	Weather	Temperature	Humidity	Wind	Play?
1	Sunny	Hot	High	Weak	No
2	Cloudy	Hot	High	Weak	Yes
3	Sunny	Mild	Normal	Strong	Yes
4	Cloudy	Mild	High	Strong	Yes
5	Rainy	Mild	High	Strong	No
6	Rainy	Cool	Normal	Strong	No
7	Rainy	Mild	High	Weak	Yes
8	Sunny	Hot	High	Strong	No
9	Cloudy	Hot	Normal	Weak	Yes
10	Rainy	Mild	High	Strong	No





## Support Vector Machines (SVMs):



- 1** **Maximize Margin**  
 SVM identifies the optimal hyperplane that maximizes the distance between data points of different classes.
- 2** **Kernel Trick**  
 By using kernel functions, SVMs can efficiently handle non-linear problems in high dimensional spaces.
- 3** **Robust to Outliers**  
 SVMs are less sensitive to outliers compared to other classification algorithms.



# Comparing and Contrasting the Supervised Learning Algorithms



## Linear Regression

Ideal for modeling continuous, linear relationships between variables. Produces a straight line of best fit to predict a target variable.

1

2

## Logistic Regression

Specialized for binary classification problems, outputting probabilities of class membership. Fits a sigmoid curve to model non-linear relationships.

Class 1  
Class 2

## Decision Trees

Construct hierarchical models by recursively partitioning data based on feature importance. Provide interpretable, rule-based predictions.

3

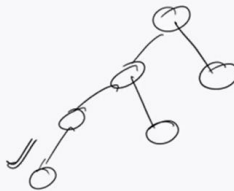
4

## Support Vector Machines (SVMs)

Find the optimal hyperplane to separate classes with maximum margin. Effective for high-dimensional, sparse data with non-linear patterns.



$x_1$   
 $x_2$   
 $x_3$   
 $x_4$



# Choosing the Right Algorithm for Your Problem

1

## Understand Your Data

Consider the structure, size, and characteristics of your dataset to determine which algorithm is best suited to handle it effectively.

$x \ y \rightarrow \text{linear}$

2

## Identify Your Objective

Clearly define whether you need to predict a continuous value (regression), a binary outcome (classification), or a more complex multi-class problem.

Reg.  
Class.

3

## Evaluate Model Performance

Use appropriate evaluation metrics like accuracy, precision, recall, or R-squared to assess how well each algorithm performs on your specific task.

4

## Consider Model Complexity

Simpler models like linear regression may be preferable if interpretability is important, while more complex algorithms like SVMs or random forests can handle nonlinear relationships.