



SNS COLLEGE OF TECHNOLOGY

Coimbatore-35

An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade

Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai



DEPARTMENT OF MECHATRONICS ENGINEERING

UNIT 2 –DATA ANALYTICS

DECISION TREE



INTRODUCTION TO DATA ANALYTICS



A Decision Tree is a popular machine learning algorithm that is used for both classification and regression tasks. It works by recursively partitioning the data into subsets based on the values of input features, making decisions at each node in the tree. Here's an overview of Decision Trees:

Key Concepts:

1.Node:

1. A point in the tree where a decision is made.

2.Root Node:

1. The topmost node that represents the entire dataset and is split into subsets.

3.Leaf Node:

1. Terminal nodes that represent the final outcome or decision.

4.Splitting:

1. The process of dividing a node into two or more child nodes based on a chosen feature.

5.Decision Rule:

1. The condition used to split a node, typically based on a feature's value.

6.Entropy and Information Gain:

1. Entropy is a measure of impurity or disorder in a set of data. Information Gain is a metric used to measure the effectiveness of a decision rule in reducing entropy.

7.Gini Impurity:

1. Gini Impurity measures the likelihood of a misclassification if an element is randomly chosen. It is another criterion used for splitting nodes.

8.Pruning:

1. The process of removing branches or nodes from a tree to avoid overfitting.



INTRODUCTION TO DATA ANALYTICS



Types of Decision Trees:

1. Classification Trees:

1. Used for predicting categorical labels or classes.

2. Regression Trees:

1. Used for predicting continuous values.

Advantages:

1. Interpretability:

1. Decision Trees are easy to understand and interpret, making them suitable for explaining model decisions to non-technical stakeholders.

2. No Data Preprocessing:

1. Decision Trees can handle both numerical and categorical data without requiring extensive preprocessing.

Challenges:

1. Overfitting:

1. Decision Trees can easily overfit the training data, capturing noise in the dataset.

2. Instability:

1. Small variations in the data can result in different tree structures.

3. Biased Toward Dominant Classes:

1. In classification tasks, Decision Trees can be biased towards dominant classes.



INTRODUCTION TO DATA ANALYTICS



Decision Tree Construction Process:

1. Root Node:

1. Select the feature that provides the best split (highest Information Gain or lowest Gini Impurity) for the entire dataset.

2. Splitting:

1. Recursively split each node based on the selected feature until a stopping condition is met (e.g., maximum depth, minimum samples per leaf).

3. Leaf Nodes:

1. Assign a class label or regression value to each leaf node based on the majority class or average value of the subset.

Ensemble Methods:

To overcome some of the challenges associated with Decision Trees, ensemble methods like Random Forests and Gradient Boosting are often used. These methods combine multiple decision trees to create a more robust and accurate model.

Decision Trees are widely used in various domains, and their simplicity and interpretability make them a valuable tool in the field of machine learning.