Decision tree

OVERVIEW & PURPOSE

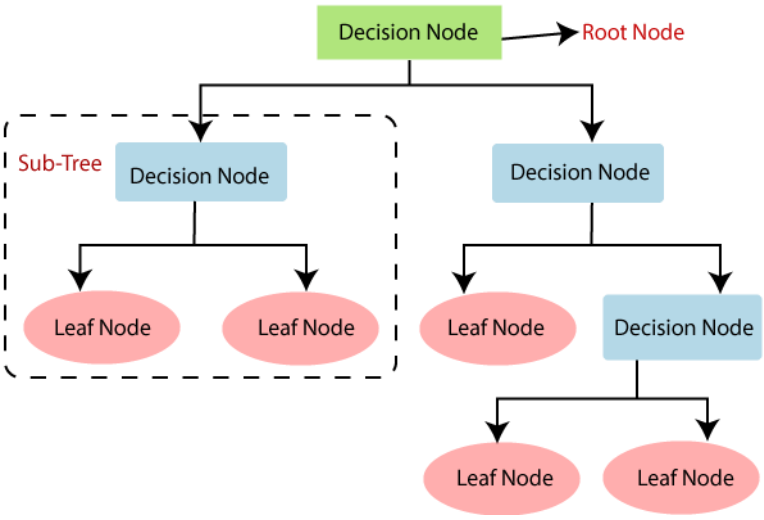
In this session, participants will learn the widely used techniques of decision tree in machine learning to predict the desired outcome

OBJECTIVE

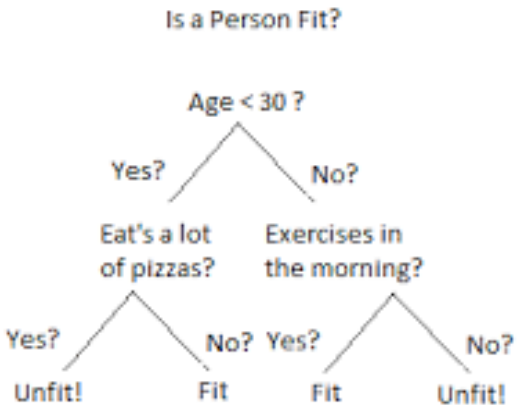
- We will be covering
  - Definition & Application of decision tree
  - Significance & Methodology of decision tree

What is a Decision Tree?

A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by recursively splitting the dataset into subsets based on the most significant attribute(s) at each level, making the decision at every node.



Imagine a tree where the root is the entire dataset, and the branches are the decisions or questions about the data that lead to leaves, which represent the output or the final decision.



Why is it used?

1. Interpretability: One of the main advantages of Decision Trees is their transparency and ease of interpretation. The rules derived from a trained decision tree can be visualized and understood, making them useful for decision-making processes.
2. Handles both numerical and categorical data: Unlike some algorithms which

require data to be of a specific type, Decision Trees can handle both numerical and categorical data.
3. Requires little data preprocessing: They don't require data normalization and can handle missing values.
4. Can be used for both classification and regression tasks: It's versatile.
5. Feature importance: They can be used to rank the importance of input features.

Important Steps in Building a Decision Tree:

1. **Feature Selection:** At every node, choose the best attribute to split the data. Measures like Information Gain, Gini Impurity, and Variance Reduction are used to select which feature to split on.
2. **Decision Making:** Based on the attribute's significance, the dataset is split into subsets. This happens recursively.
3. **Pruning:** This is the process of removing the branches from the tree that have little power in prediction, to prevent overfitting. Trees that are too deep tend to capture noise in the data.
4. **Stopping Criteria:** You don't always want to continue splitting until each leaf has a single data point (this would likely lead to overfitting). So, one would set criteria like the maximum depth of the tree, or the minimum samples required to make a new split.
5. **Tree Visualization:** This is more about understanding and representation, but visualizing the tree can give insights into how decisions are made.
6. **Prediction:** For a new data point, start at the root and navigate through the tree by following the splits until a leaf node is reached. The value or class at the leaf node is the prediction.

Limitations:

While Decision Trees are powerful, they have limitations:

- They are prone to overfitting, especially when the tree is deep.
- Small changes in the data can result in a different tree (high variance).
- They can become overly complex, which might not generalize the data well.
- Often, techniques like Random Forests or Gradient Boosted Trees are employed to overcome some of these limitations by using ensembles of trees.

Reading Material

- https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/
- https://www.geeksforgeeks.org/decision-tree-introduction-example/
- https://www.w3schools.com/python/python_ml_decision_tree.asp
- https://scikit-learn.org/stable/modules/tree.html

Resources

- https://www.youtube.com/watch?v=ZVR2Way4nwQ
- https://www.youtube.com/watch?v=PHxYNGo8NcI
- https://www.youtube.com/watch?v=_L39rN6gz7Y
- https://www.youtube.com/watch?v=RmajweUFKvM