Decision Tree

A **decision tree** is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

It is one way to display an algorithm that only contains conditional control statements.

A decision tree is a flowchart-like structure

- each internal node represents a “test” on an attribute ,
- each branch represents the outcome of the test
- and each leaf node represents a class label (decision taken after computing all attributes).
- The paths from root to leaf represent classification rules.

- Tree based learning algorithms are considered to be one of the best and mostly used supervised learning methods.
- Tree based methods empower predictive models with
- high accuracy,
- stability
- ease of interpretation.
- Unlike linear models, they map non-linear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression).
- Decision Tree algorithms are referred to as **CART (Classification and Regression Trees)**.

Common terms used in Decision Trees

- Root Node: It represents entire population or sample and this further gets divided into two or more homogeneous sets.
- Splitting: It is a process of dividing a node into two or more sub-nodes.
- Decision Node: When a sub-node splits into further sub-nodes, then it is called decision node.
- Leaf/ Terminal Node: Nodes do not split is called Leaf or Terminal node.
- Pruning: When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
- Branch / Sub-Tree: A sub section of entire tree is called branch or sub-tree.
- Parent and Child Node: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node.

Applications of Decision Trees

- Evaluation of brand expansion opportunities for a business using historical sales data
- Determination of likely buyers of a product using demographic data to enable targeting of limited advertisement budget
- Prediction of likelihood of default for applicant borrowers using predictive models generated from historical data
- Help with prioritization of emergency room patient treatment using a predictive model based on factors such as age, blood pressure, gender, location and severity of pain, and other measurements
- Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.
- Because of their simplicity, tree diagrams have been used in a broad range of industries and disciplines including civil planning, energy, financial, engineering, healthcare, pharmaceutical, education, law, and business.

Advantages

- **Easy to Understand**: Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
- **Useful in Data exploration:** Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. It can also be used in data exploration stage. For e.g., we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
- Decision trees implicitly perform variable screening or feature selection.
- Decision trees require relatively **little effort from users for data preparation**.
- **Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
- **Data type is not a constraint:** It can handle both numerical and categorical variables. Can also *handle multi-output problems.*
- **Non-Parametric Method:** Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.
- Non-linear relationships between parameters do not affect tree performance.
- The number of hyper-parameters to be tuned is almost null.

Prevent overfitting

- preventing overfitting is pivotal while modeling a decision tree and it can be done in 2 ways:
- Setting constraints on tree size
- Tree pruning

Pruning

- Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.
- Pruning reduces the complexity of the final classifier and hence improves predictive accuracy by the reduction of overfitting.