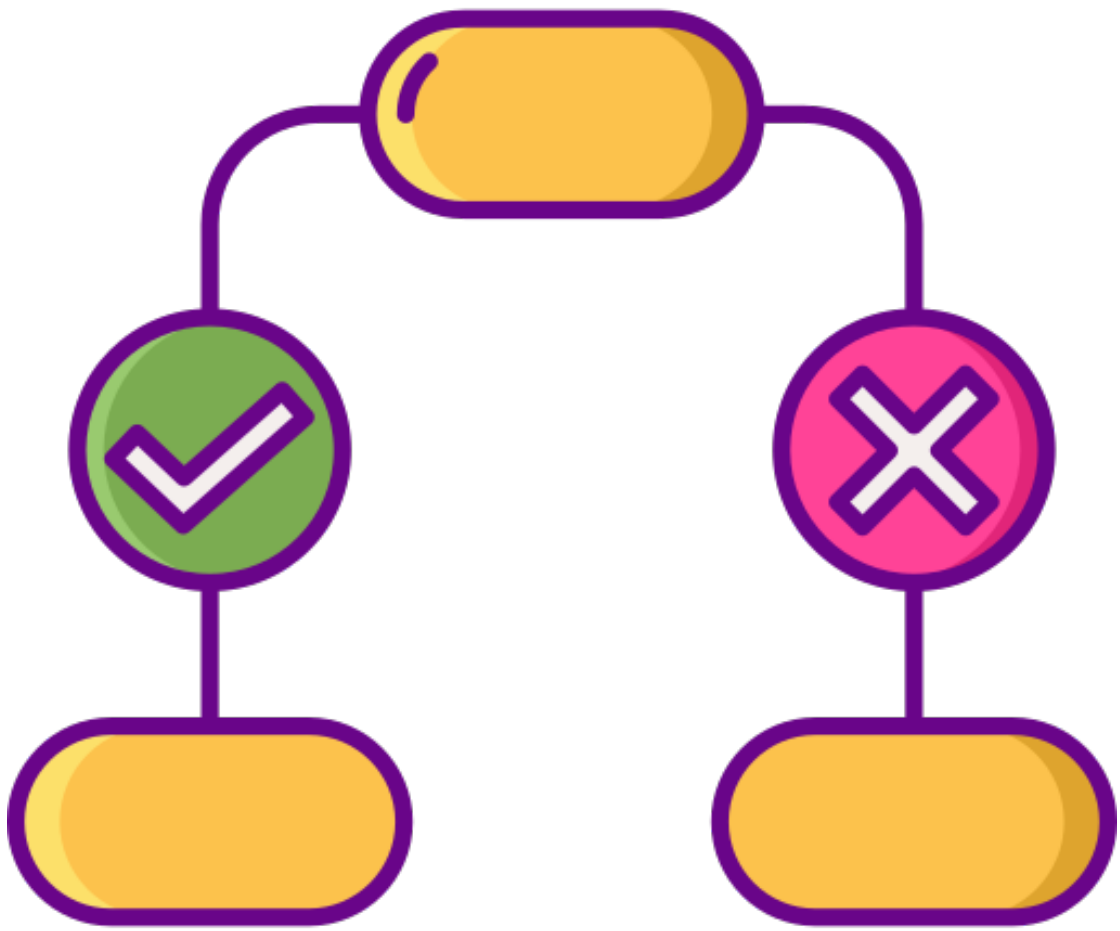
Decision Trees



DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

What is a Decision Tree?

Decision trees are a type of supervised learning algorithm used in machine learning. They are used for classification and regression analysis.

Decision trees use a tree-like structure of nodes, branches, and leafs to model decisions and their consequences.

Decision trees are non-parametric, which means they do not make any assumptions about the distribution of the data.

Key Terms

Root Node: It represents the entire population or sample, and this further gets divided into two or more homogeneous sets.

Leaf/ Terminal Node: A node that cannot be split further is called Leaf or Terminal node.

Decision Node: When a sub-node splits into further sub-nodes, then it is called a decision node.

Branch / Sub-Tree: A subsection of the entire tree is called a branch or sub-tree.

Key Terms

Parent and Child Node: A node, which is divided into sub-nodes, is called the parent node of sub-nodes, whereas sub-nodes are the child of the parent node.

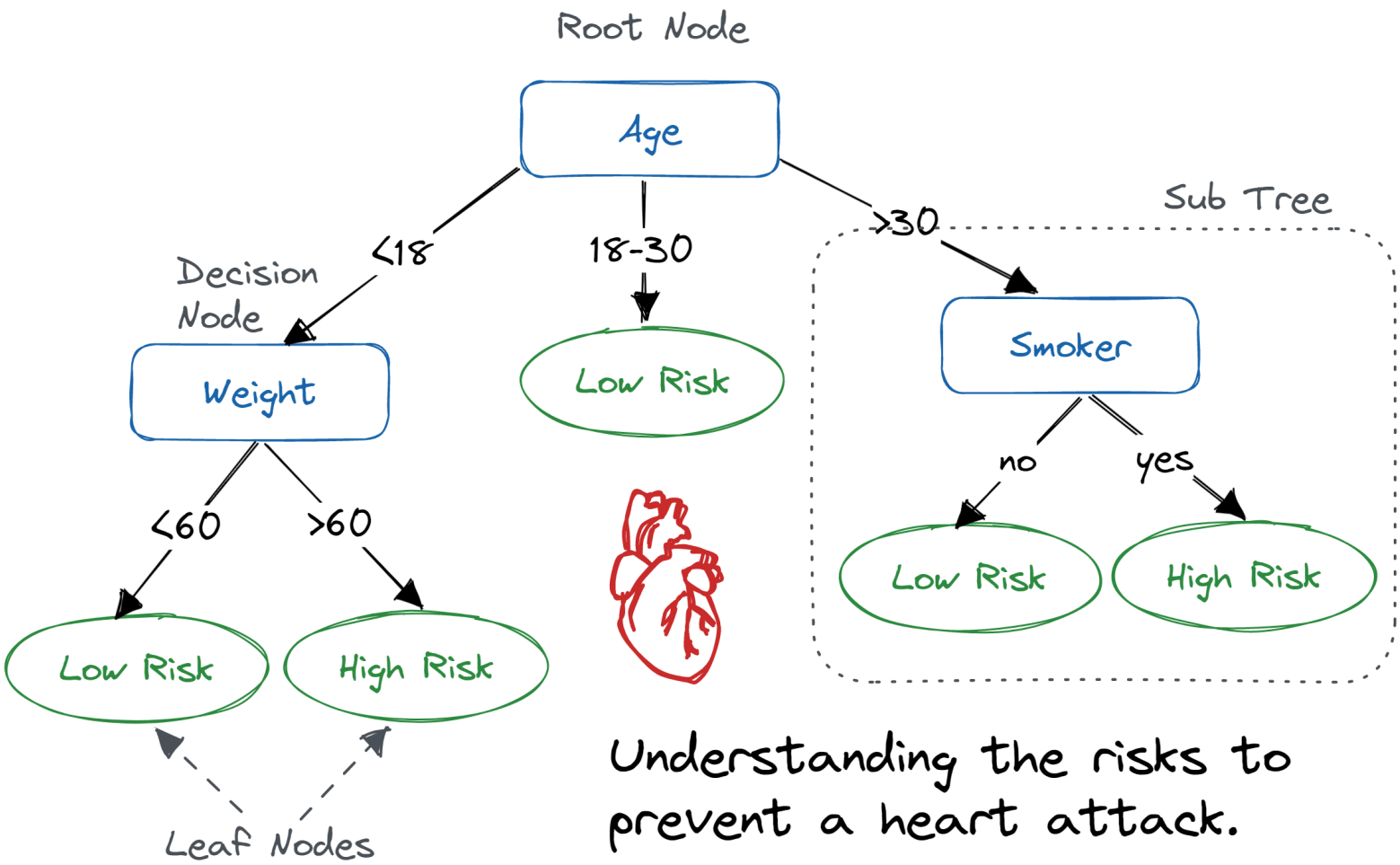
Splitting: It is a process of dividing a node into two or more sub-nodes.

Pruning: Pruning is when we selectively remove branches from a tree. The goal is to remove unwanted branches, improve the tree's structure, and direct new, healthy growth.

Key Terms

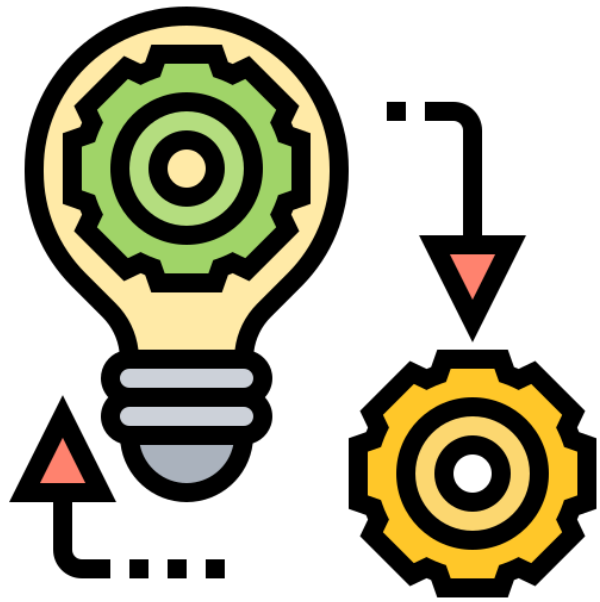
Impurity: In decision trees, impurity is a measure of the homogeneity of the labels at the node. The current implementation provides two impurity measures for classification (Gini impurity and entropy) and one impurity measure for regression (variance). The algorithm chooses the partition maximizing the purity of the split (i.e., minimizing the impurity).

Example



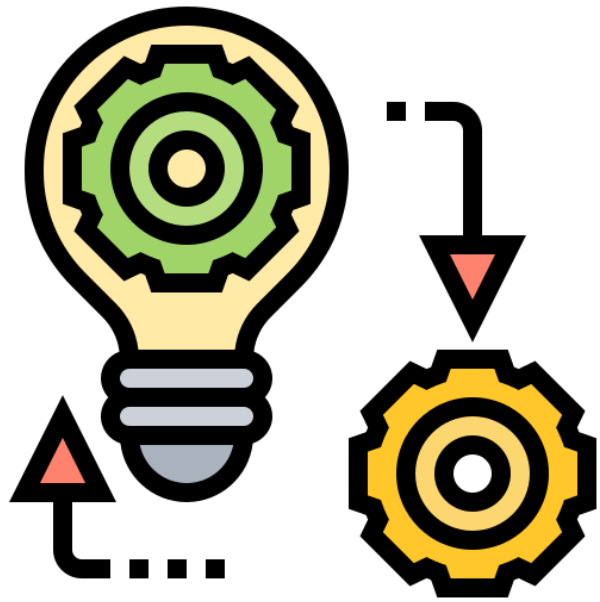
Understanding the risks to prevent a heart attack.

How decision trees work?



The decision tree algorithm works by recursively partitioning the data into subsets based on the values of different features.
At each node of the tree, the algorithm selects the feature that best separates the data into subsets that are most homogeneous with respect to the target variable.
This process is repeated until a stopping criterion is met, such as a maximum depth of the tree or a minimum number of data points in a leaf node.

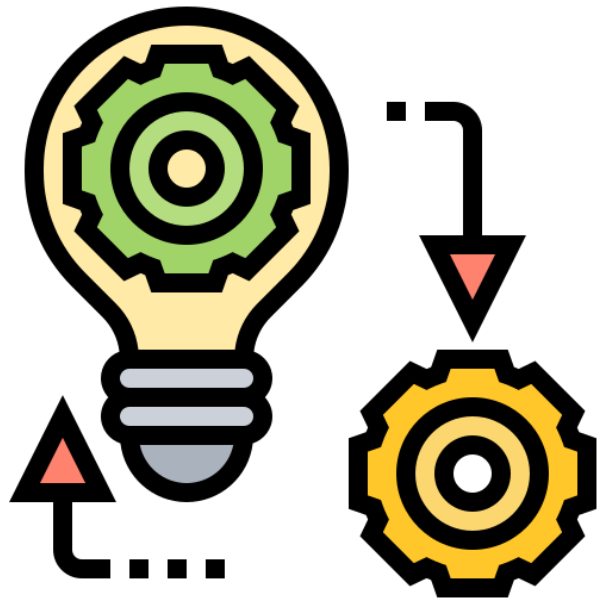
How decision trees work?



Step 1. Select the best feature to split the data: The algorithm looks at all the input features and selects the one that provides the best split in terms of maximizing the information gain or minimizing the impurity of the resulting subsets.

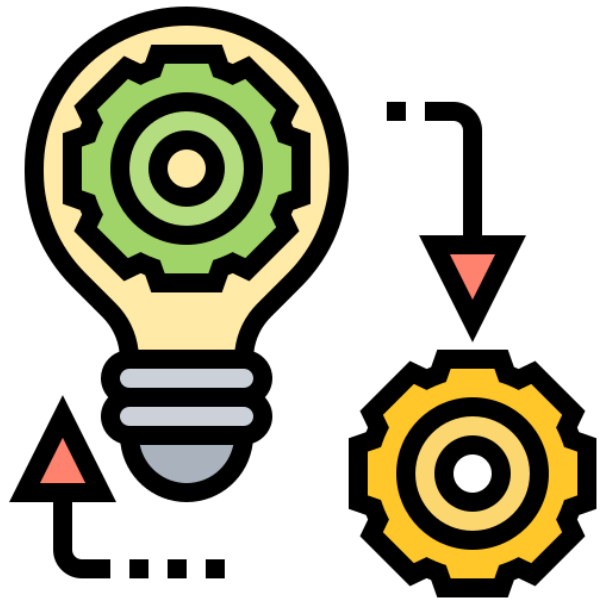
Step 2. Split the data: Once the best feature is selected, the algorithm splits the data into subsets based on the feature's values. Each subset represents a branch or child node of the tree.

How decision trees work?



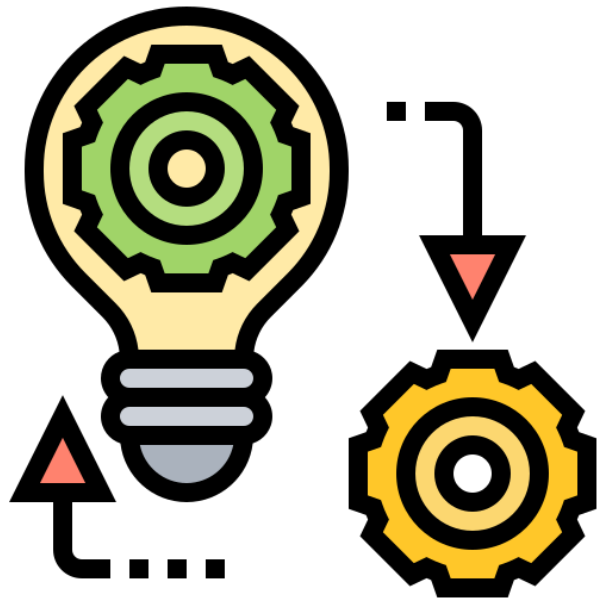
Step 3. Recurse on the child nodes: The algorithm recursively applies the above steps on each child node until a stopping criterion is met, such as reaching a maximum depth, having a minimum number of data points in a leaf node, or achieving a certain level of accuracy.

How decision trees work?



Step 4. Assign a class or regression value to each leaf node: Once the tree has been constructed, the algorithm assigns a class label or regression value to each leaf node based on the majority class of the data points or the average value of the target variable in that node.

How decision trees work?



Step 5. Make predictions: To make predictions on new data, the algorithm traverses the tree based on the feature values of the data point until it reaches a leaf node, and assigns the corresponding class label or regression value.

Advantages of Decision Trees

1. Easy to interpret: Decision trees provide a simple and intuitive representation of the decision-making process, making them easy to understand and interpret. This is especially useful for non-technical stakeholders who need to make decisions based on the model's predictions.

2. Can handle both categorical and numerical data: Decision trees can handle a mix of categorical and numerical data, making them suitable for a wide range of applications.

3. Able to capture complex interactions between variables: Decision trees can capture complex interactions between variables, including non-linear relationships, interactions, and dependencies, making them a good choice for data with many features.

Advantages of Decision Trees

4. Robust to noise: Decision trees are robust to noisy data and can handle missing values without the need for imputation.

5. Fast and efficient: Decision trees are relatively fast and efficient to train and can handle large datasets with millions of observations.

6. Can be combined with other algorithms: Decision trees can be used as building blocks in ensemble methods such as random forests and gradient boosting, which can further improve their performance.

Disadvantages of Decision Trees

1. Prone to overfitting: Decision trees can easily overfit the training data if not properly pruned or regularized, leading to poor generalization performance on new data.

2. Sensitive to small variations in the data: Decision trees are sensitive to small variations in the data, which can lead to different trees being generated for different training sets. This can make the model less robust and more difficult to interpret.

3. Can be biased towards features with many levels: Decision trees tend to be biased towards features with many levels, which can lead to overemphasis on these features at the expense of other important features.

Disadvantages of Decision Trees

4. May not handle continuous variables well: Decision trees are typically designed to handle discrete or categorical data, and may not perform as well on continuous variables without discretization.

5. May not be the most accurate algorithm for certain problems: While decision trees are a powerful and flexible algorithm, they may not always provide the best predictive accuracy compared to other algorithms like random forests or neural networks, especially for high-dimensional or complex data.

Follow **#DataRanch** on LinkedIn for more...

Data Analysis Steps

DATARANCH.org

Essential Chart Types

DATARANCH.org

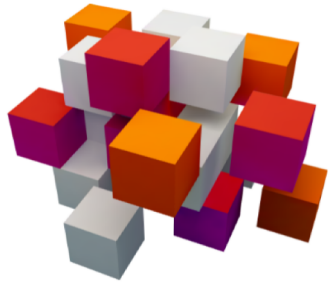
Data Cleaning Steps

DATARANCH.org

Common data fallacies to watch out for...

DATARANCH.org

Data Wrangling Steps



DATARANCH.org

What is Supervised Learning?

DATARANCH.org

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

Follow **#DataRanch** on LinkedIn for more...

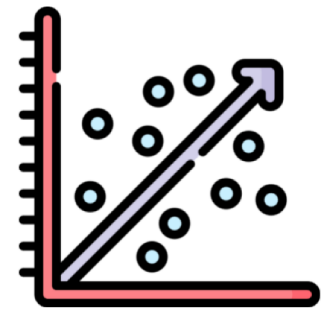
Logistic Regression

DATARANCH.org

What is Unsupervised Learning?

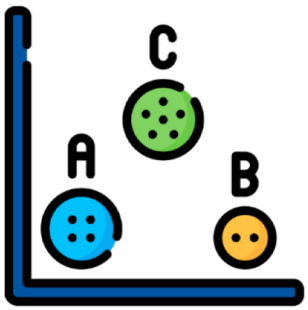
DATARANCH.org

Regression Analysis



DATARANCH.org

Clustering



DATARANCH.org

Principal Component Analysis

DATARANCH.org

t-Distributed Stochastic Neighbour Embedding (t-SNE)

DATARANCH.org

DATARANCH.org
VISUALIZE | ANALYZE | CAPITALIZE

DATARANCH.org

VISUALIZE | ANALYZE | CAPITALIZE

info@dataranch.org

linkedin.com/company/dataranch