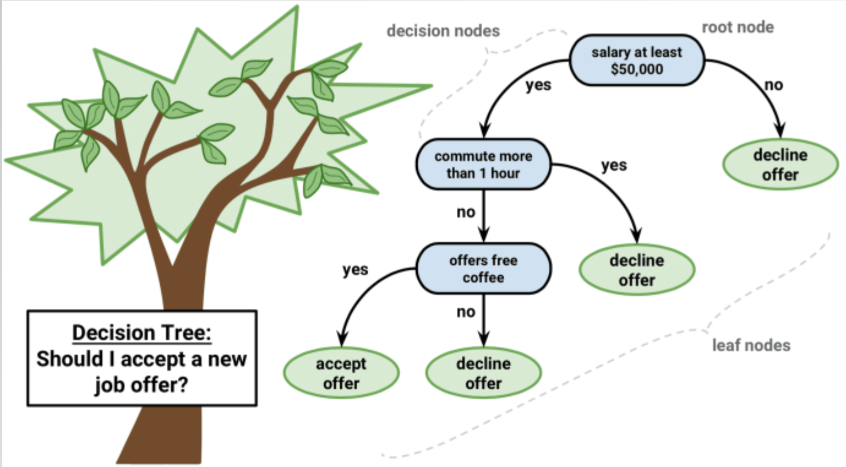
Decision Tree

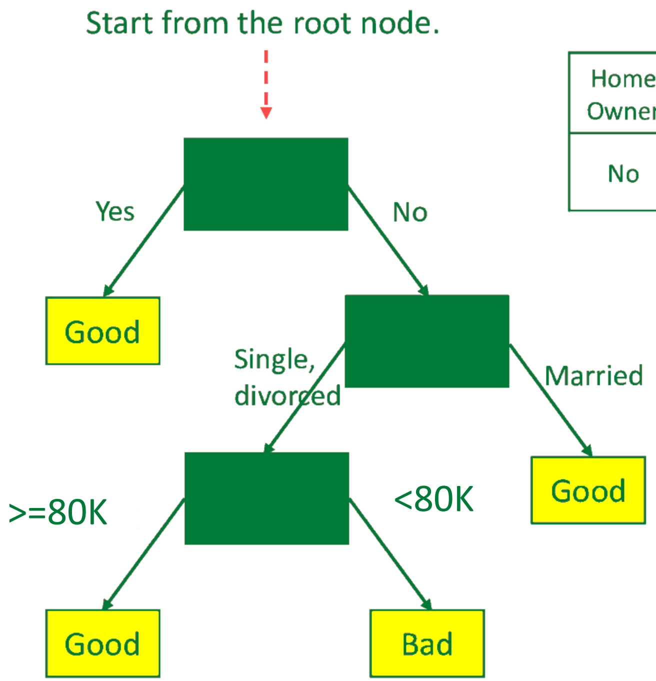
DR. SULTAN ALFARHOOD

What is a Decision Tree?

- A decision tree is a tree-based supervised learning method used to predict the output of a target variable.



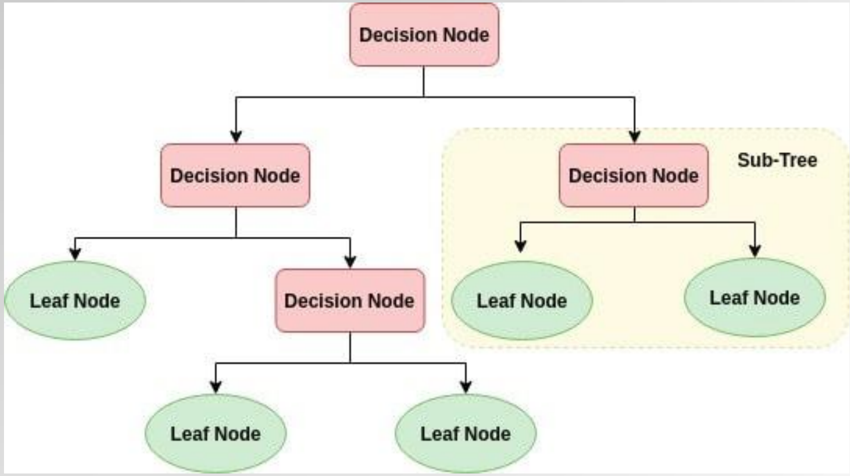
Decision Tree Example (Animated)



Home Owner	Marital Status	Annual Income	Credit Rating
No	Married	80K	?

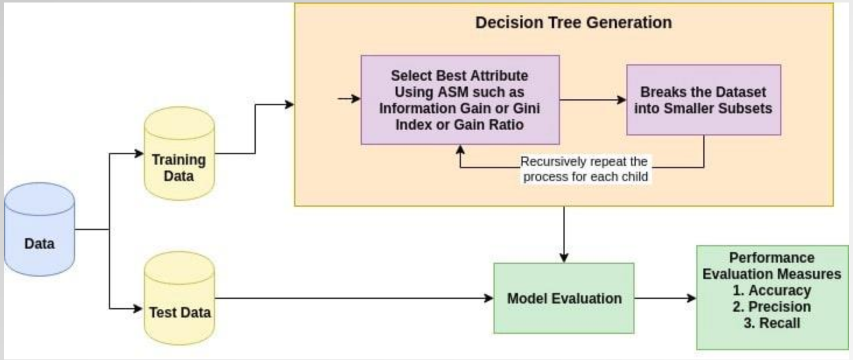
How the training process works

1. Select the **best attribute** to split the records (training data) based on Attribute Selection Measures (ASM).
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Starts tree building by repeating this process recursively for each child until one of the condition will match:
   - All the tuples belong to the same attribute value.
   - There are no more remaining attributes.
   - There are no more instances.



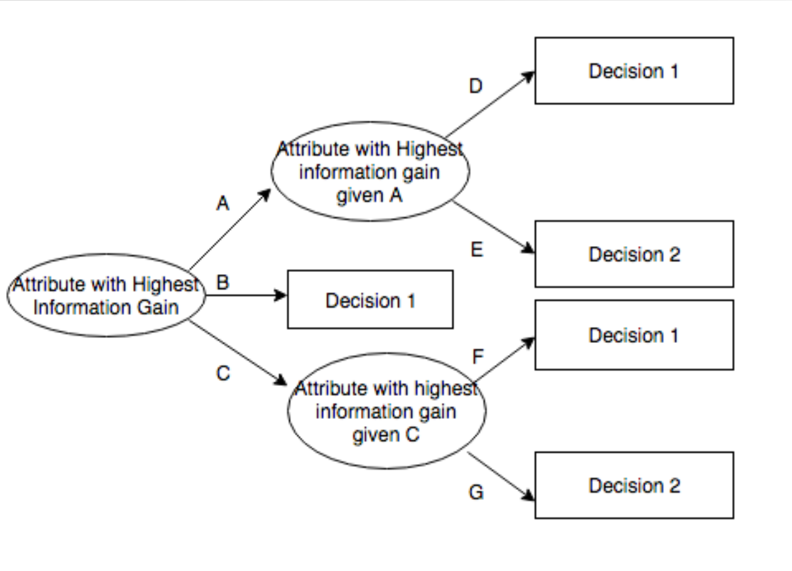
Attribute Selection Measures (ASM)

- **Attribute Selection Measures (ASM)** is a heuristic for selecting the splitting criterion that partition data into the best possible manner.
- **ASM** provides a rank to each feature (or attribute) by explaining the given dataset.
  - Best score attribute will be selected as a splitting attribute.
- There are various formulations of the decision tree learning algorithm; **ID3** (Iterative Dichotomiser 3) is a popular one.



ID3 Algorithm

- It is a classification algorithm that follows a greedy approach by selecting a **best attribute** that yields **maximum Information Gain** (IG).
- **ID3 Steps**
  1. Calculate the Information Gain of each feature.
  2. Split the dataset S into subsets using the feature for which the Information Gain is maximum (considering that all rows don't belong to the same class).
  3. Make a decision tree node using the feature with the maximum Information gain.
     - If all rows belong to the same class, make the current node as a leaf node with the class as its label.
  4. Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.



Information Gain

Chose the best feature

- Information Gain (IG) computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values

$$\boldsymbol{IG(S, A)} = \boldsymbol{Entropy(S)} - \sum_{v=1}^{n_v} \left(\left(\frac{|\boldsymbol{S}_{\mathrm{v}}|}{|\boldsymbol{S}|}\right) \times \boldsymbol{Entropy(S_{\mathrm{v}})} \right)$$

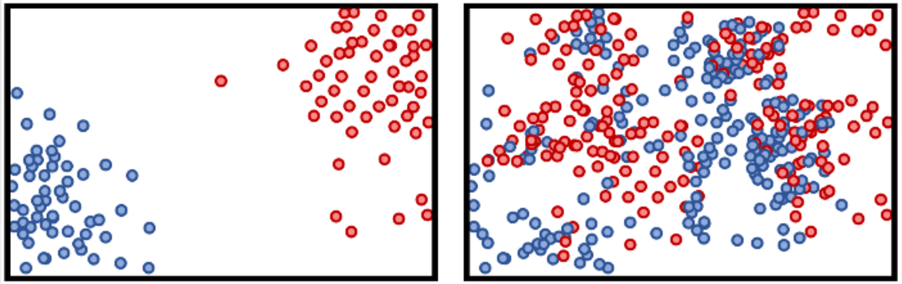
- Where:
  - $\boldsymbol{n_v}$ is the number of unique values for the feature column $\boldsymbol{A}$
  - $|\boldsymbol{S}|$ is the number of rows in $\boldsymbol{S}$
  - $\boldsymbol{S}_{\mathrm{v}}$ is the set of rows in $\boldsymbol{S}$ for which the feature column $\boldsymbol{A}$ has value **v**
  - $|\boldsymbol{S}_{\mathrm{v}}|$ is the number of rows in $\boldsymbol{S}_{\mathrm{v}}$

Entropy

- Entropy is a measure of the amount of uncertainty in the dataset S.
- Mathematical representation of Entropy is shown here:

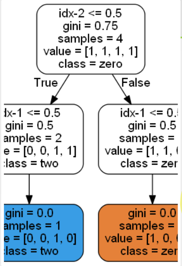
$$H(S) = -\sum_{c \in C} P(c)\, log_2\, P(c)$$

- Where:
  - S is the current dataset for which entropy is being calculated (changes every iteration of the ID3 algorithm).
  - C is the set of classes in S {i.e., C=(yes, no) }
  - $P(c)$: The proportion of the number of elements in class c to the number of elements in set S.



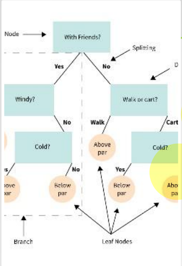
Low Entropy High Entropy

Decision Tree Advantages



Decision trees are to understand, interpret, and visualize

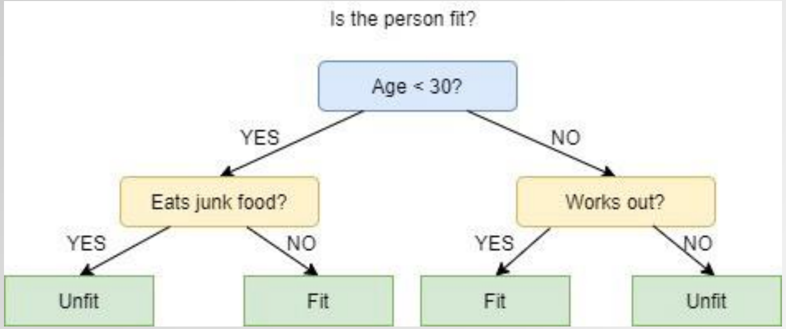
They can effectively handle both numerical and categorical data



They can determine the worst, best, and expected values for several scenarios

Decision Tree Disadvantages

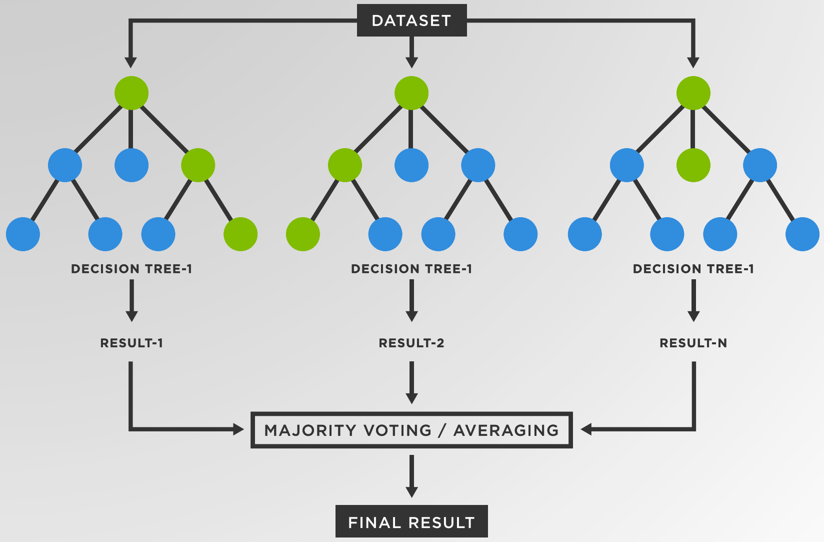
- **Overfitting**
  - It generally leads to overfitting of the data which ultimately leads to wrong predictions
- Affected by noise



Decision Tree based Ensample Approaches

- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM
- CatBoost

Random Forest



Thank you