https://towardsdatascience.com/understanding-decision-trees-once-and-for-all-2d891b1be579

understanding_decision_trees.ipynb

Iris Dataset composed of:

4 features = Petal length, petal width, sepal length & sepal width

Target Variable: Iris species

↳ 3: 0 setosa, 1 versicolor, 2 virginica

DATA: 150 observations (Rows), 50 for each iris = Balanced DATA

For easier understanding work w/ 2 of petal width & sepal width

- Confusion Matrix: y-axis is target
  x-axis is species decision tree predicted
- 2 feature Decision tree 93% accurate
- How was the Decision Tree Built?
  - Split data into 2
    1. Training Set
    2. Testing Set
       ↳ used to measure performance

↳ Goal of Decision Tree is to split training set into homogenous areas where only 2 iris species is present according to given features

Node 0: Root Node

- First decision boundary of tree: Petal width = 0.8cm
  ↳ easily visible in graph
  ↳ decided by testing all possible decision splitting dataset
  → choose 1 that minimizes Gini Impurity of 2 splits

↳ measures probability from randomly chosen element to be incorrectly classified

⇒ probability choosing an element times probability misclassified

$$\sum_{i=1}^{J} p_i \sum_{k \neq i} p_k = \sum_{i=1}^{J} p_i(1-p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = \sum_{i=1}^{J} p_i - \sum_{i=1}^{J} p_i^2$$

all possible classes J

$$= 1 - \sum_{i=1}^{J} p_i^2$$

↳ used to perform Gini Test

- How does the algo split?
  - Tries all possible boundaries along all features
  - computes Gini impurity of 2 groups
  - Choose boundary w/ lowest Gini impurity for each Group

(only 5e's)
- left Node Gini = 0 so becomes a leaf & doesn't get split again
- Right node GI = ~0.5 cuz there about equal # of Ve & Vi
- creating group w/ 1 not always best cuz you might not have all of em
- Overfitting can happen if we don't limit size of Tree