Data Normalization

Min-Max Normalization

- The simplest method is rescaling the range of features to scale the range in [0, 1]. Selecting the target range depends on the nature of the data. The general formula is given as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

where x is an original value x' is the normalized value.

Z-score Normalization (Standardization)

- In machine learning, we can handle various types of data, e.g., audio signals and pixel values for image data, and this data can include multiple dimensions. Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance

$$x' = \frac{x - \bar{x}}{\sigma}$$

Where x is the original feature vector, x-bar is the mean of that feature vector, and sigma is its standard deviation.

Correlation Matrix

Correlation coefficient is a measure of the linear association between two variables X and Y.
It has a value between -1 and 1 where:

-1 indicates a perfectly negative linear correlation between two variables
0 indicates no linear correlation between two variables
1 indicates a perfectly positive linear correlation between two variables

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

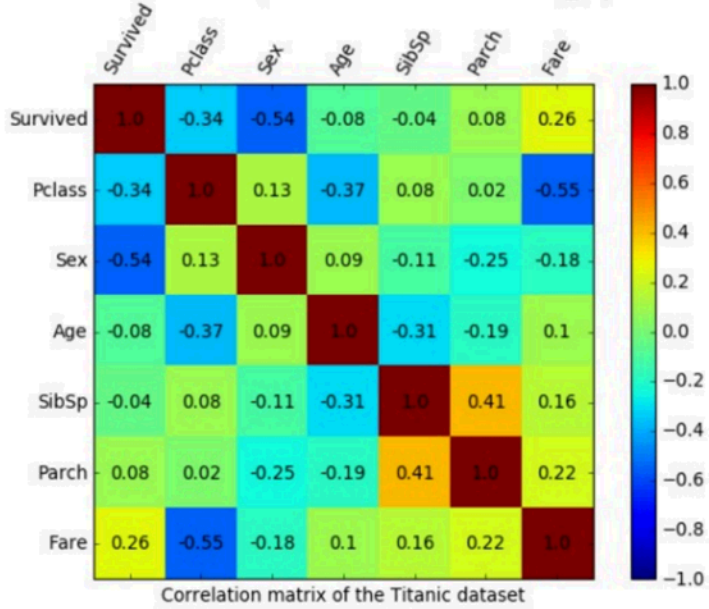
The primary purpose of a correlation matrix in data analysis to assess the strength and direction of relationships between variables.

The further away the correlation coefficient is from zero, the stronger the relationship between the two variables.

Notice that a correlation matrix is perfectly symmetrical. For example, the top right cell shows the exact same value as the bottom left cell.

Correlation Matrix

Given the correlation matrix below, answer the following questions.



Correlation matrix of the Titanic dataset

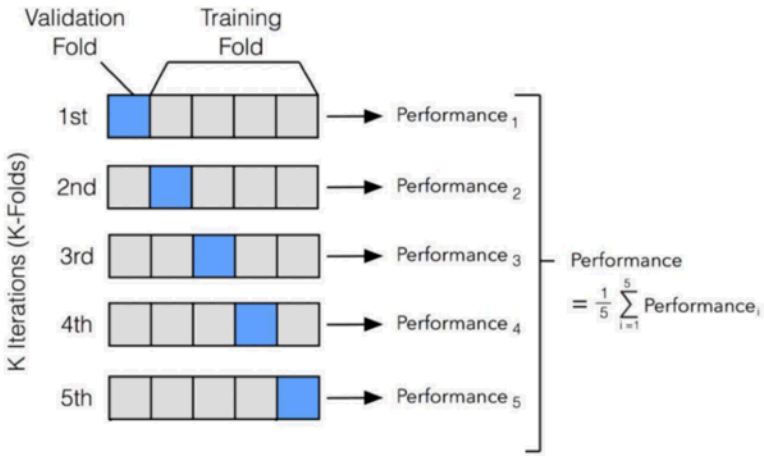
a. Which two variables have the strongest relationship?
b. Which two variables have the weakest relationship?
c. Which variable has the strongest relationship with “Survived”?
d. Which variables go down when “Fare” goes up?

Cross Validation

Cross-validation is a statistical method used in machine learning and data analysis to assess the performance and generalizability of a model. It involves splitting the data into subsets, training the model on some subsets, and testing it on others. This process helps evaluate how well a model performs on unseen data, reducing the risk of overfitting.

K-Fold Cross-Validation

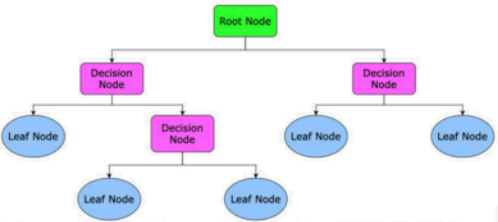
- The data is divided into k equal-sized folds (subsets).
- The model is trained k times, each time using k−1 folds for training and the remaining fold for testing.
- The results are averaged to produce a single performance metric.

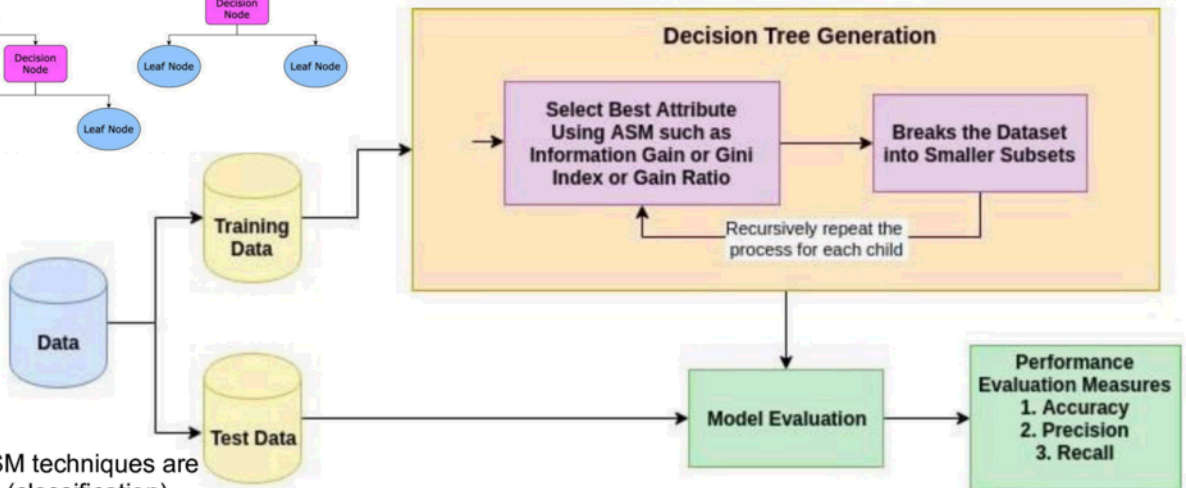


Decision Tree

Decision Trees algorithm is commonly used for classification tasks in machine learning

Limitation of decision trees - They are prone to overfitting.

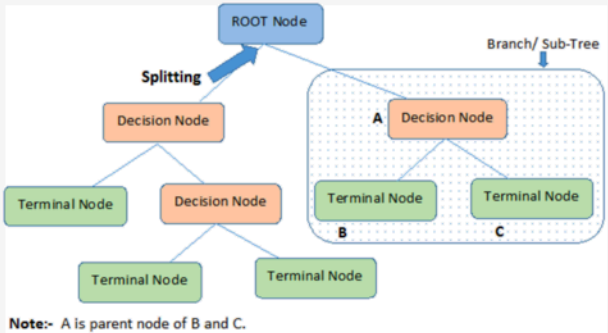




The main ASM techniques are

1. Gini index (classification)
2. Information Gain (classification)
3. Variance Reduction or Mean squared error (regression)

Decision Tree - Terminology



1.Root Node: This attribute is used for dividing the data into two or more sets. The feature attribute in this node is selected based on Attribute Selection Techniques.

2.Branch or Sub-Tree: A part of the entire decision tree is called a branch or sub-tree.

3.Splitting: Dividing a node into two or more sub-nodes based on if-else conditions.

4.Decision Node: After splitting the sub-nodes into further sub-nodes, then it is called the decision node.

5.Leaf or Terminal Node: This is the end of the decision tree where it cannot be split into further sub-nodes.

Ref: https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53

Decision Tree:

Attribute Selective Measure(ASM)

Gini Index is used for in Decision Trees to measure the "impurity" or "purity" of a dataset at a particular node.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

A feature with a lowest Gini index is chosen for split.

If a node has a Gini index of 0, the samples belong to only one class.

If a node has a Gini index of 0.5, the samples are completely mixed and classified evenly among classes.

A lower Gini index indicate that the split results in more homogeneity in the child nodes.

$$\textit{InformationGain(feature)} = \textit{Entropy(Dataset)} - \textit{Entropy(feature)}$$

$$\text{Entropy} = -\sum_{i=1}^{n} p_i * \text{Log}_2(p_i)$$

A feature with the largest information gain is chosen for split

Decision Tree

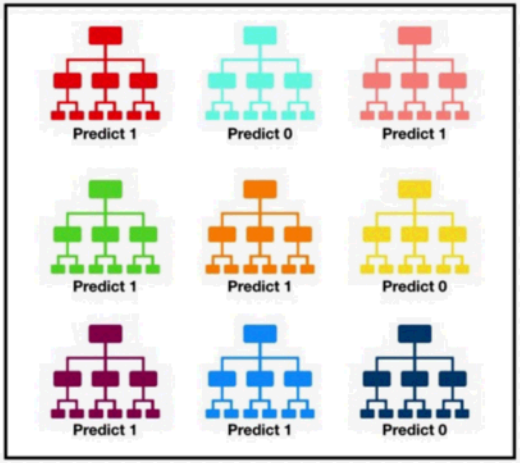
Given a dataset with three classes (A, B, C) and their proportions in a node are: P(A)=3/7, P(B)=2/7, P(C)=2/7, what is the Gini index?

Decision Tree

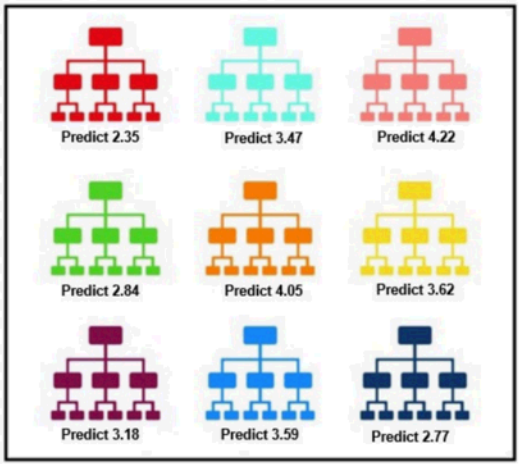
If a dataset contains 8 samples where 3 belong to Class A, 4 to Class B, and 1 to Class C, what is the entropy of this dataset?

Random Forest

Random forest is a ***Supervised Machine Learning Algorithm*** that is ***used in Classification and Regression problems***. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.



Classification: **Predict 1 = 6**
Predict 0 = 3
Model prediction = 1



Regression:
Model prediction = 3.43
(2.35+3.47+4.22+2.84+4.05+3.62+3.18+3.59+2.77)/9 = 3.43