Fundamental Concepts:

Attribute: A characteristic or feature of a dataset.

Sample: An individual data point or instance within a dataset.

Similarity Calculations:

Mixed Types of Attributes: Handling different data types (e.g., numerical, categorical) in similarity calculations.

Normalization:

Min-Max: Scaling data between a specified range (typically 0 to 1).

Z-Score: Scaling data based on mean and standard deviation.

Decimal: Scaling data by shifting the decimal point.

Distance Calculation:

Manhattan: Sum of the absolute differences between corresponding attributes' values.

Chi-Square Test:

Statistical test to determine if there is a significant association between categorical variables in a contingency table.

Information Gain:

Measure of the effectiveness of a feature in classifying a dataset.

Gini Index:

Measure of impurity or disorder in a dataset's target variable.

Converting Ordinal Values to Numeric:

Assigning numerical values to ordinal categories while preserving their order.

Smoothing:

Mean or Median: Techniques to reduce noise or irregularities in data by replacing extreme values with the average or median.

Correlation from Scatter Plot:

Understanding the relationship between two variables based on the pattern observed in a scatter plot.

Histogram and Boxplot:

Histogram: A graphical representation of the distribution of numerical data.

Boxplot: A visual summary of data distribution through quartiles.

Cosine Similarity:

Measure of similarity between two non-zero vectors in an inner product space.

Define TP, TN, FP, FN:

TP (True Positive): Correctly predicted positive instances.

TN (True Negative): Correctly predicted negative instances.

FP (False Positive): Incorrectly predicted as positive.

FN (False Negative): Incorrectly predicted as negative.