K-Means Clustering and Evaluation Metrics

1 Overview of K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm used to partition a dataset into K distinct clusters. The goal is to assign each data point to the cluster with the nearest mean (centroid), effectively minimizing the variance within each cluster.

1.1 Steps in K-Means Clustering

1. **Initialization**: Randomly choose K data points as initial centroids.
2. **Assignment**: Assign each data point to the nearest centroid based on a distance metric (usually Euclidean distance).
3. **Update**: Calculate new centroids by taking the mean of all data points assigned to each cluster.
4. **Repeat**: Continue until convergence.

Mathematically, let C_k be the centroid of cluster k:

$$C_k = \frac{1}{N_k} \sum_{x_i \in S_k} x_i$$

2 Evaluation Metrics for Clustering

2.1 Confusion Matrix

A confusion matrix summarizes TP, TN, FP, and FN:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

2.2 Accuracy Calculation

Accuracy can be calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3 K-Medoids Clustering

K-Medoids uses actual data points as centroids, making it more robust to noise.