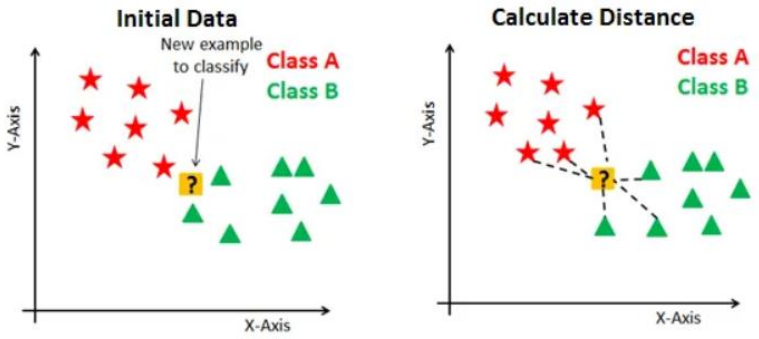
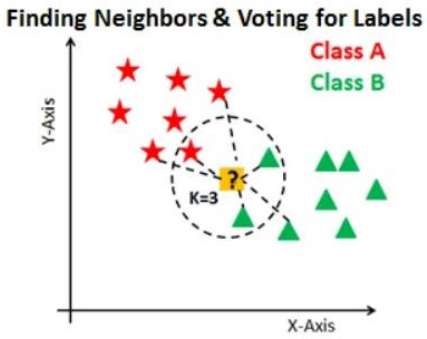
K-Nearest Neighbor (KNN)

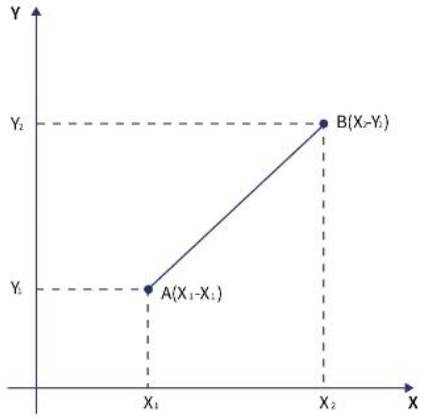
- K-Nearest Neighbor (K-NN) is a simple **supervised machine learning** algorithm.
- K-NN assumes **similarity** between **new data** and **existing data**.
- K-NN classifies new data based on its **similarity to existing data**.
- K-NN can be used for both **regression and classification**, but is more commonly used for classification.
- K-NN is a **non-parametric algorithm**, meaning it doesn't make assumptions about the underlying data.
- K-NN is a **lazy learner algorithm**, meaning it doesn't learn from the training set immediately.
- K-NN stores the training set and performs classification when new data arrives.

How does K-NN work?

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.
- **Step-6:** Our model is ready.







Euclidean Distance A_1 and $B_2 = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}$

Applications for k-NN

- Medical Diagnosis
- Recommendation Systems
- Financial Risk Analysis
- Image Recognition
- Pattern Recognition
- Stock Market Forecasting
- Agriculture
- Sentiment Analysis
- Intrusion Detection
- Handwriting Detection
- Gene Expression Analysis
- Real Estate Valuation

Advantages & Disadvantages of KNN Algorithm

Advantages

- It is very easy to understand and implement
- It is an instance-based learning(lazy learning) algorithm.
- KNN does not learn during the training phase hence new data points can be added with affecting the performance of the algorithm.
- It is well suited for small datasets.

Disadvantages

- It fails when variables have different scales.
- It is difficult to choose K-value.
- It leads to ambiguous interpretations.
- It is sensitive to outliers and missing values.
- Does not work well with large datasets.
- It does not work well with high dimensions