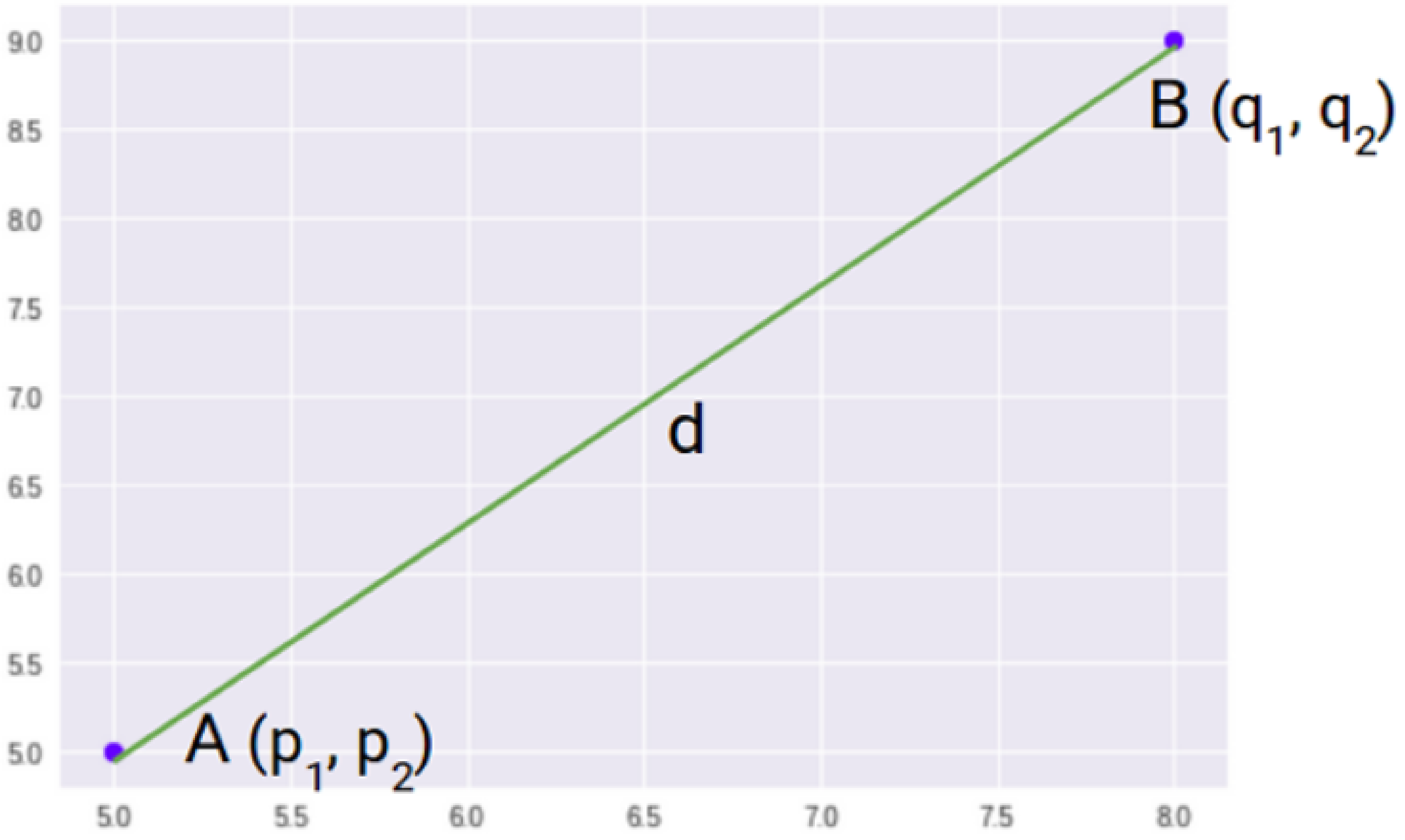
Part 1

Understanding Distance Metrics in Machine Learning



Let’s say you need to create clusters using **K-Means Clustering or KNN**.

How will you define the similarity between different clusters? This will happen if their features are similar, right?

When plotting, points closer to each other on the graph have smaller distances, making it easier to determine the distance between the points and establish their resemblance.

BUT...

How do we calculate this distance, and what are the different distance metrics in machine learning?

Also, Do the metrics vary depending on the specific problem at hand??

What Are Distance Metrics?

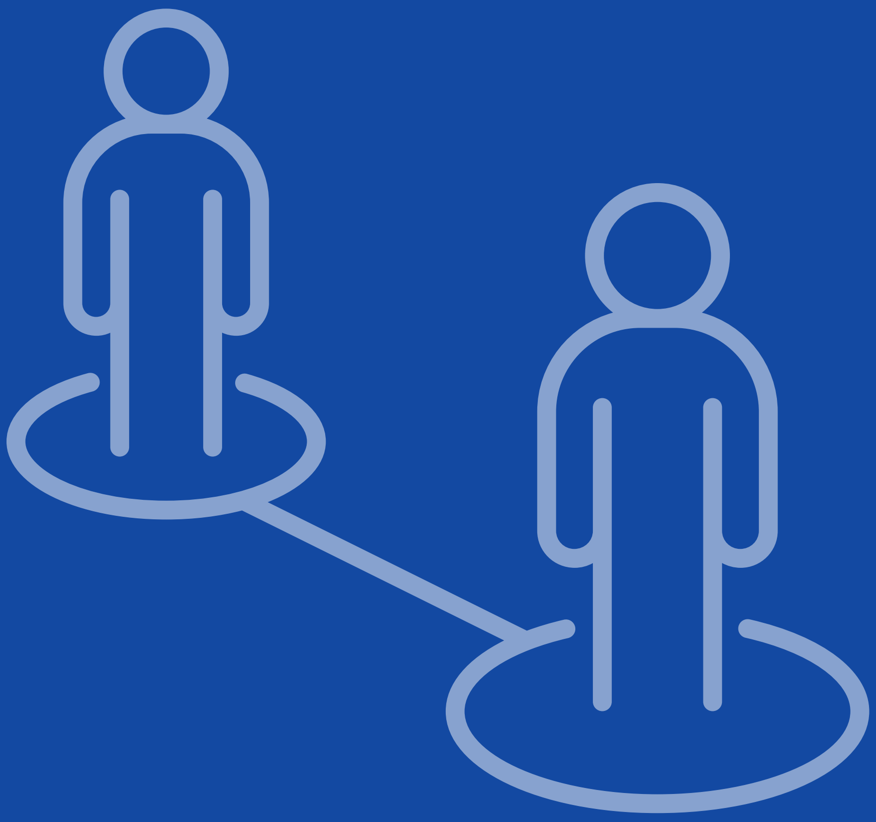
Distance metrics are a key part of several ML algorithms. These metrics are used in both supervised and unsupervised learning, generally to calculate the similarity between data points. An effective distance metric improves the performance of our machine learning model.

Types of Distance Metrics in Machine Learning

- **Euclidean Distance**
- Manhattan Distance
- Minkowski Distance
- Hamming Distance

Let's focus on Euclidean Distance now...

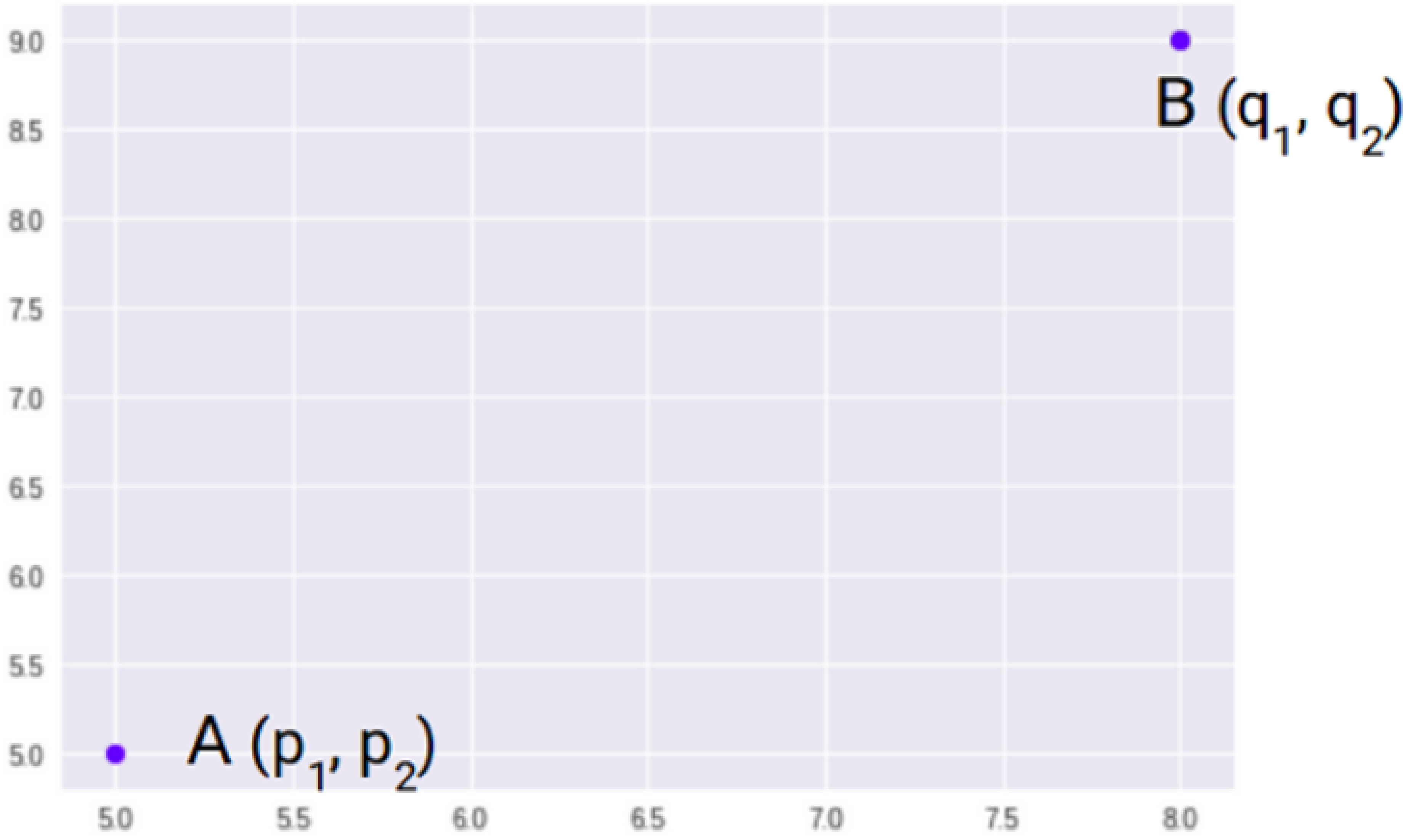
Euclidean Distance



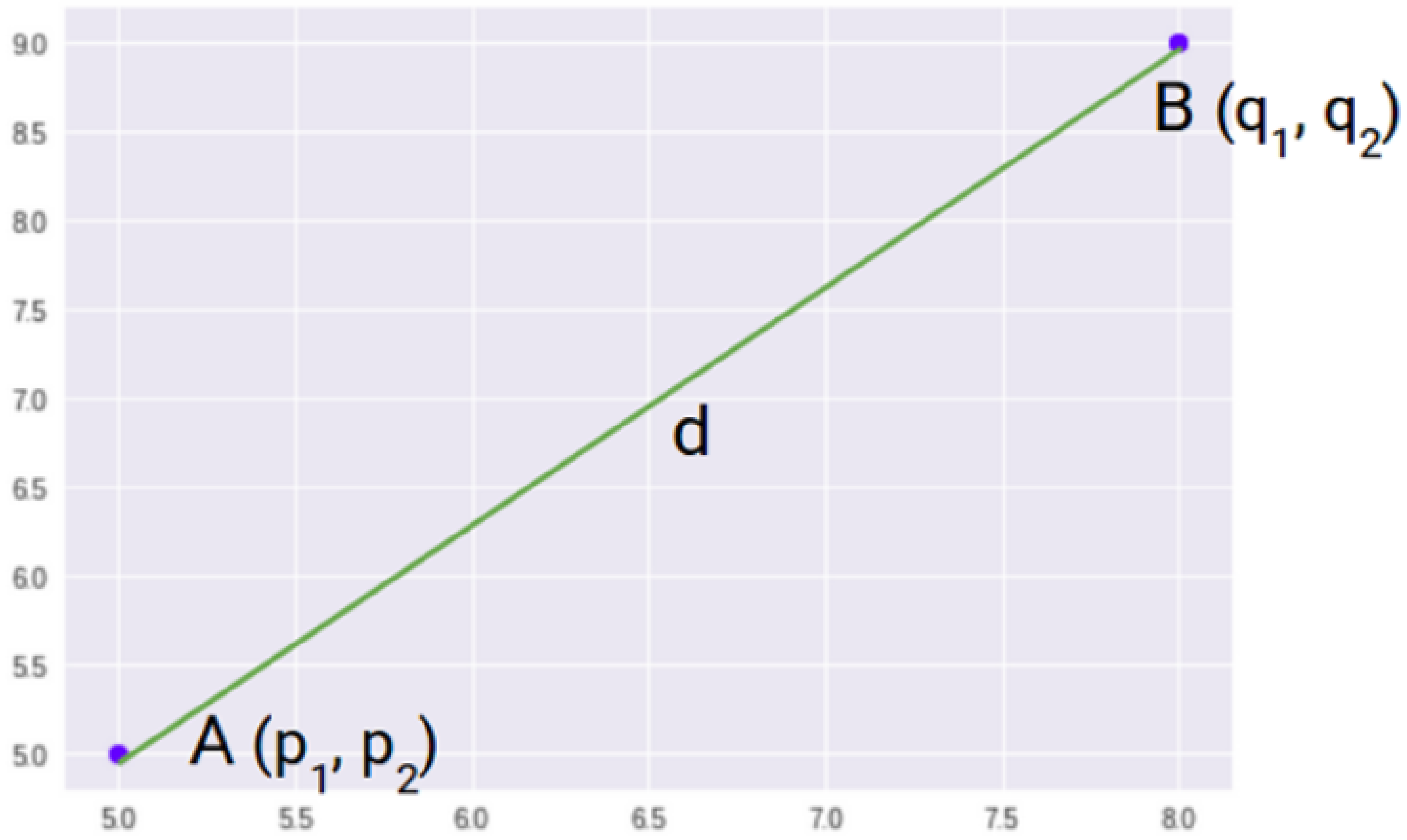
It represents the shortest distance between two vectors. It is the square root of the sum of squares of differences between corresponding coordinates of two data points.

Calculation

Let’s say we have two points, as shown below:



So, the Euclidean Distance between these two points, A and B, will be:



Formula for Euclidean Distance

$$d = ((p_1 - q_1)^2 + (p_2 - q_2)^2)^{1/2}$$

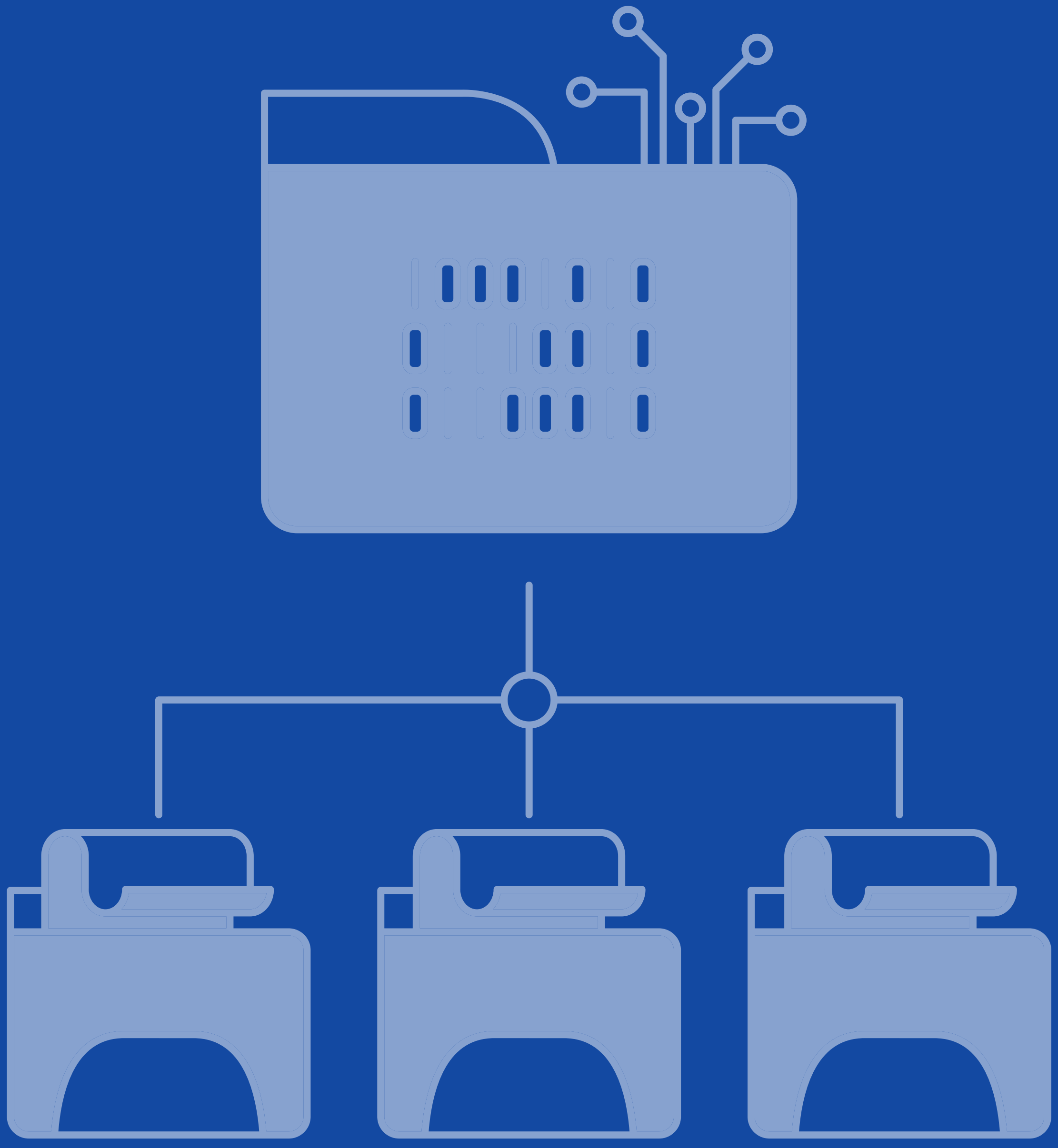
Where,

n = number of dimensions

pi, qi = data points

Analytics Vidhya

Use cases



1. **K-Means Clustering:** It helps determine the similarity between data points and assign them to clusters based on their proximity to the cluster centroids.

2. **Nearest Neighbor Classification:** It identifies the k nearest neighbors to a data point and assigns the majority class label of these neighbors for classification.

3. **Anomaly Detection:** Data points that have a significantly larger distance from the centroid or neighboring points can be considered as potential anomalies.

4. Dimensionality Reduction: PCA identifies directions of maximum variance in data using Euclidean distance and reduces data dimensionality by projecting onto these directions while preserving important information.

5. Recommendation Systems: It helps identify users or items with similar preferences by measuring the distance between their rating vectors or feature vectors.

6. **Image Processing:** By computing the distance between feature vectors extracted from images, similarity or dissimilarity between images can be measured