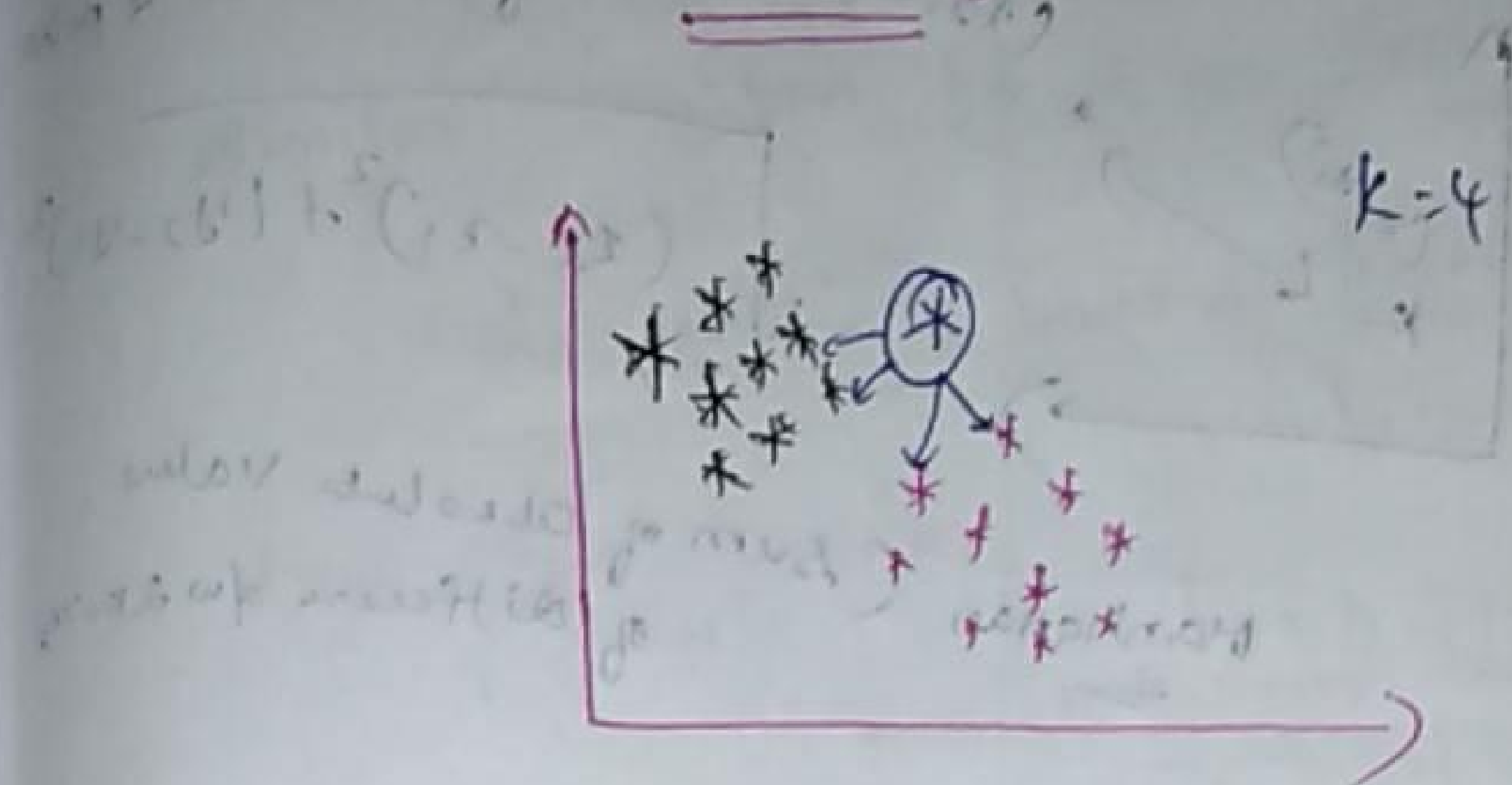


## K-nearest neighbour



- ⇒ K select value  $K=4$
- ⇒ calculate distance
- ⇒ select majority neighbour
- ⇒ now model is ready

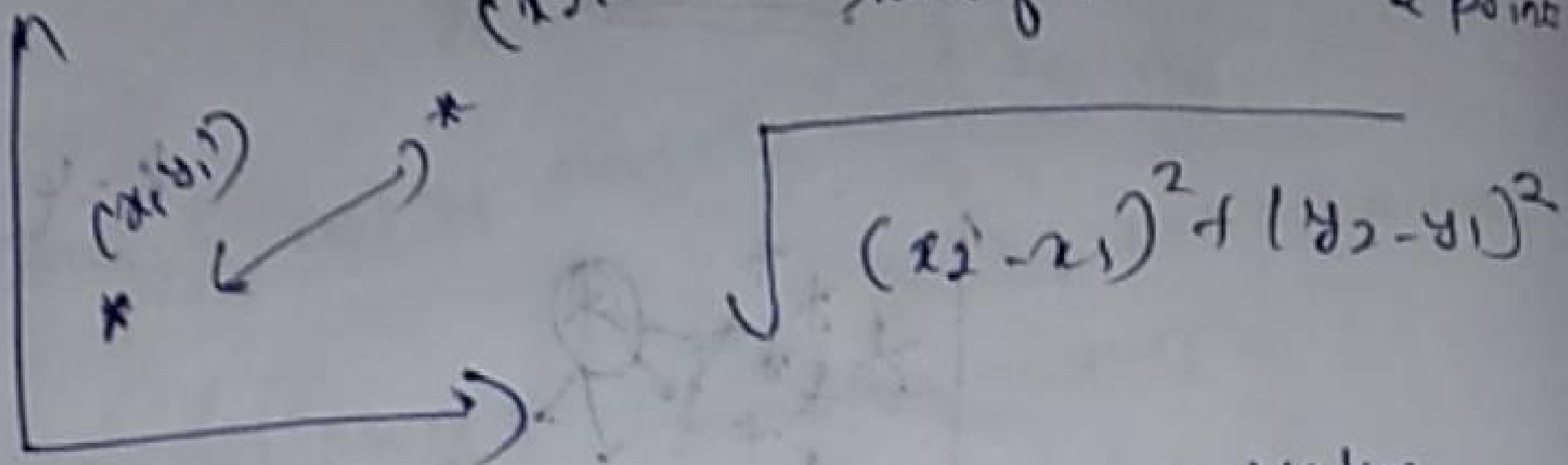
- ① K-value
- ② calculate the distance of the nearest neighbors.
- ③ How many nearest neighbors belongs to category 1 & category 2.

④ How calculate distance?

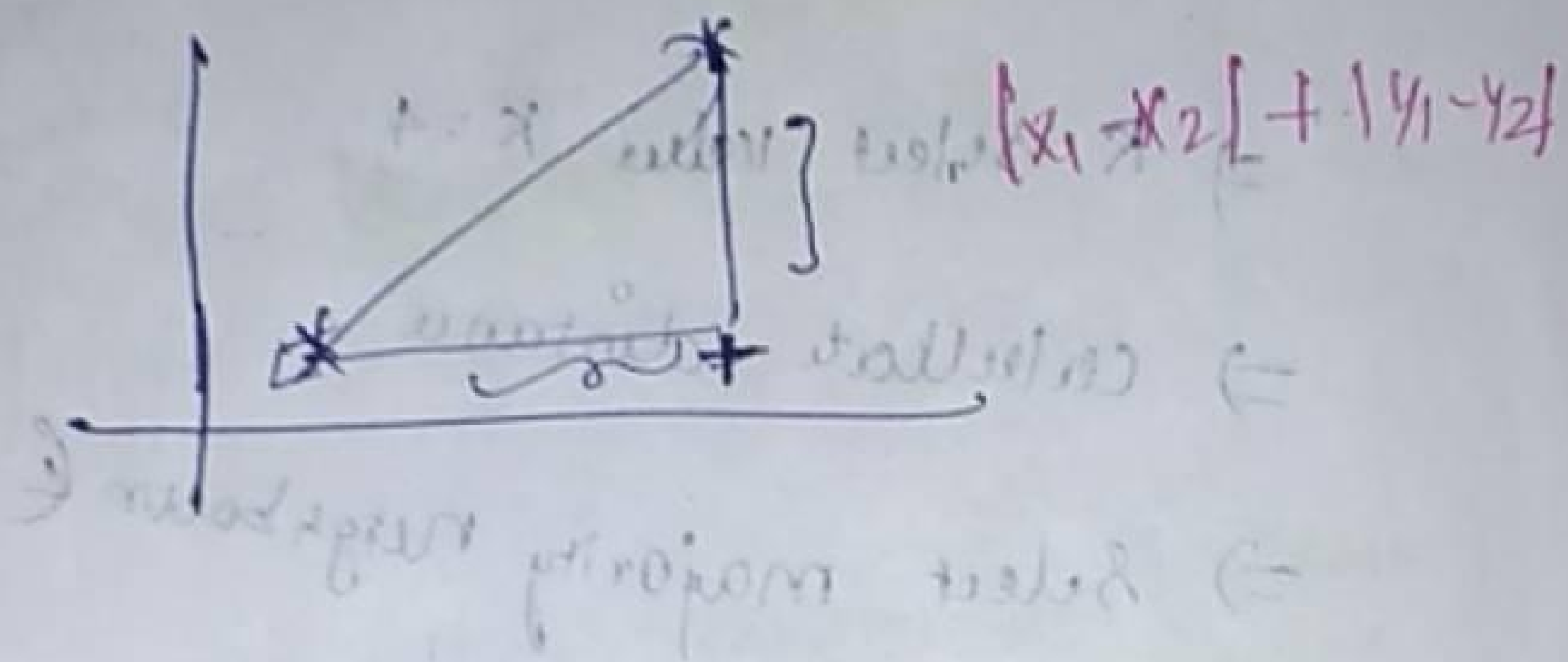
1. Euclidean
2. Manhattan

# Euclidean

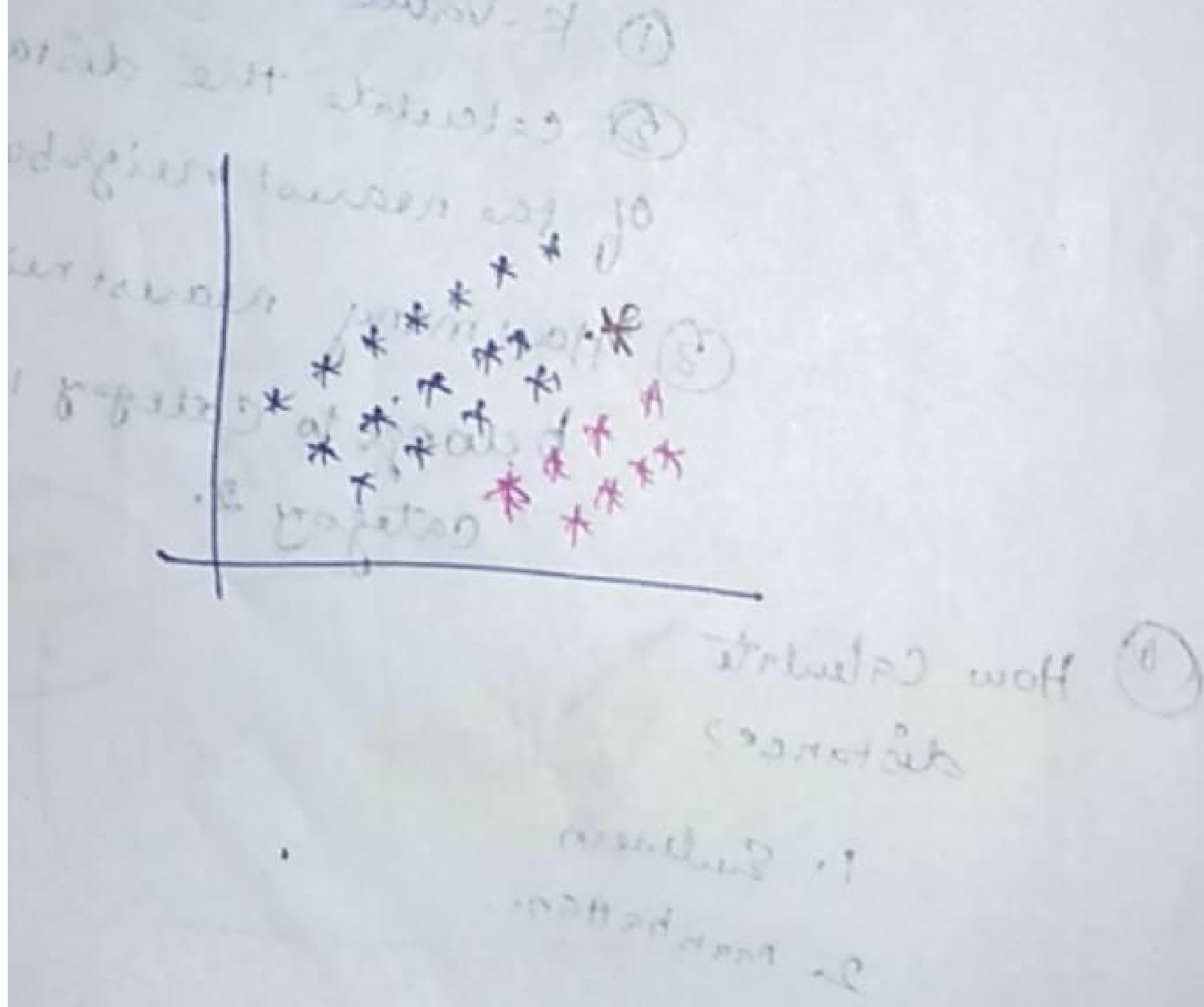
( $x_1, y_1$ ) Square root of the sum of the distance b/w 2 points



Manhattan (Sum of absolute value of difference b/w 2 points)



Dataset → 900 44  
100 no  
k-NN bias



$K$  is a number used to identify  
similar neighbour for a new data point.

The decision is based on feature similarity.

How chose  $K$  value?

Similar

to  $K$ -means cluster

Advantage:

⇒ used for both classification & regression.

⇒ Simple algorithm hence easy to

interpret the prediction

⇒ Training step faster

⇒ non-parametric, so makes no  
assumption about the underlying data pattern.

disadvantages

⇒ High memory requirements as  $K$  must have  
to store all data points.

⇒ Prediction stage costly.

⇒ Sensitive to outliers, impacted by  
noise or irrelevant data.



K-Nearest Neighbors (KNN) vs K-Means Clustering  
 Supervised vs Unsupervised

Classification  
 Regression

Unsupervised

Clustering

K in KNN is number  
 of nearest neighbors.

K in K-means is  
 no of cluster.