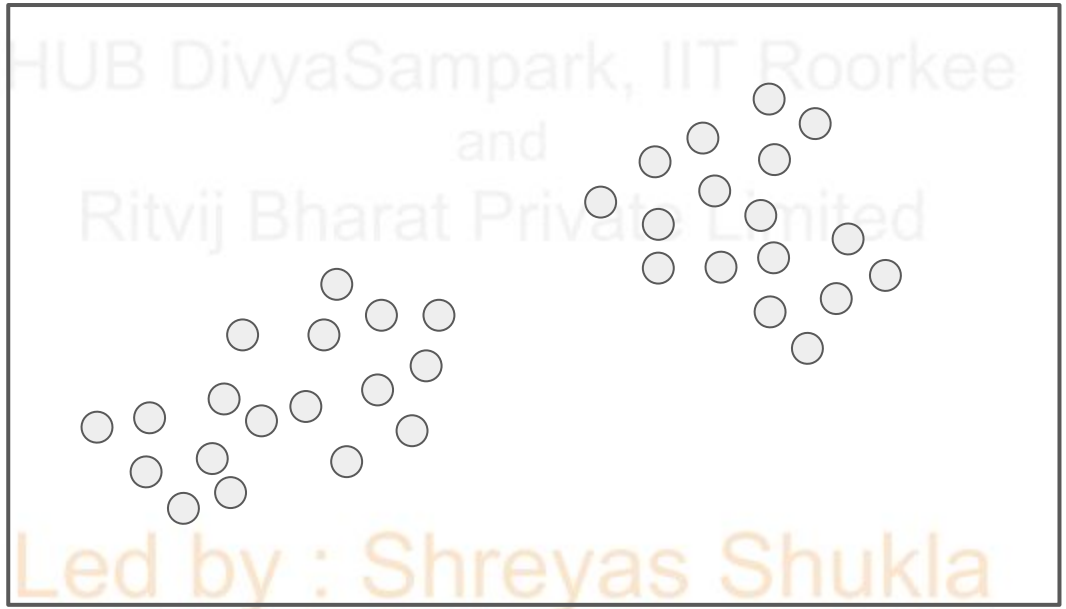
K Nearest Neighbors

One of the simplest machine learning algorithms.

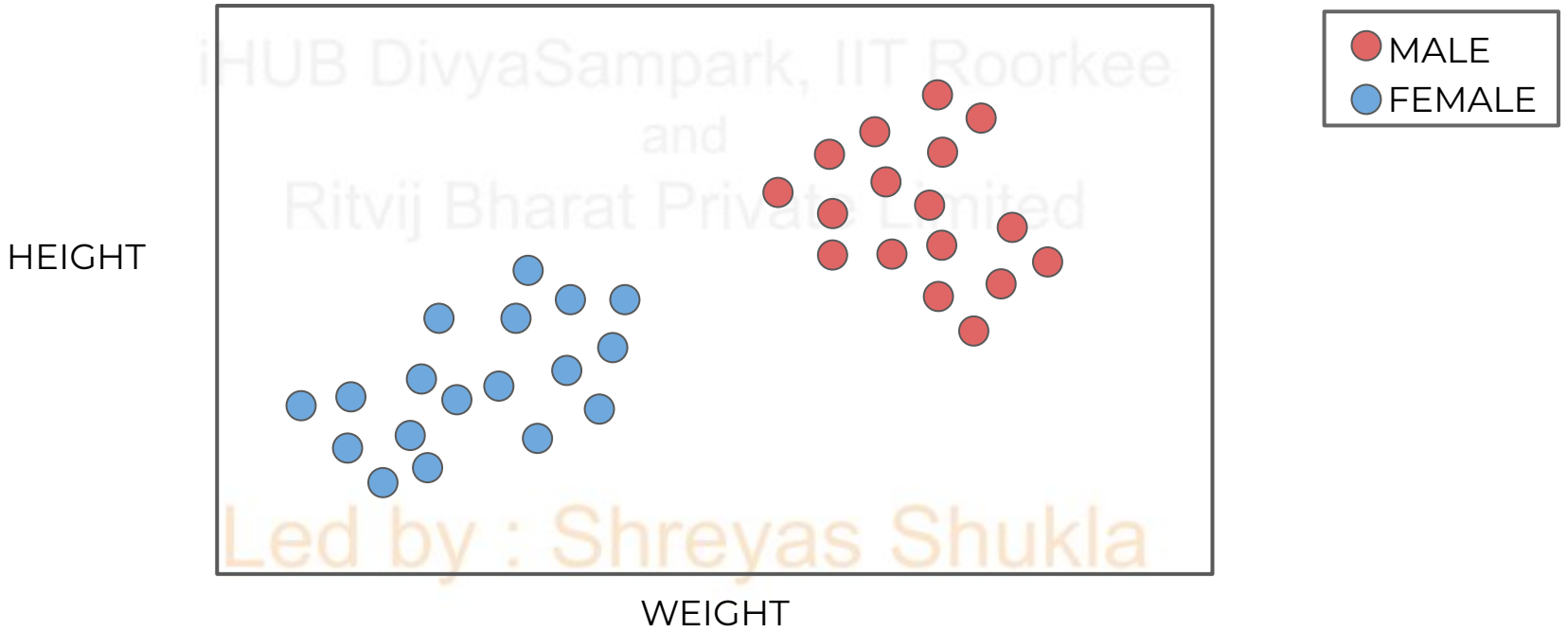
Assigns a label to new data based on the **distance** between the old data and new data.

Let’s imagine we have a dataset of baby chick heights and weights.

How could we train an algorithm to identify the sex of a new baby chick based on historical features?

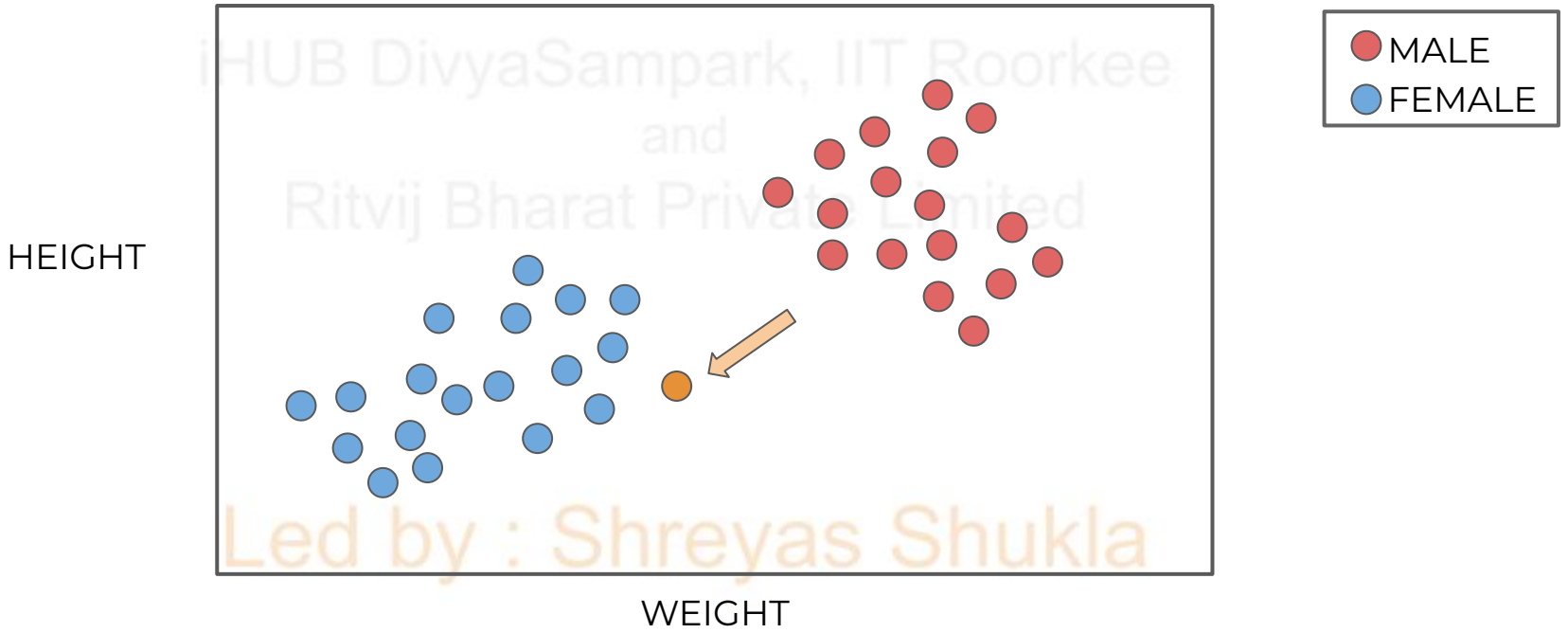


Historically, we know the sex of the chicks:

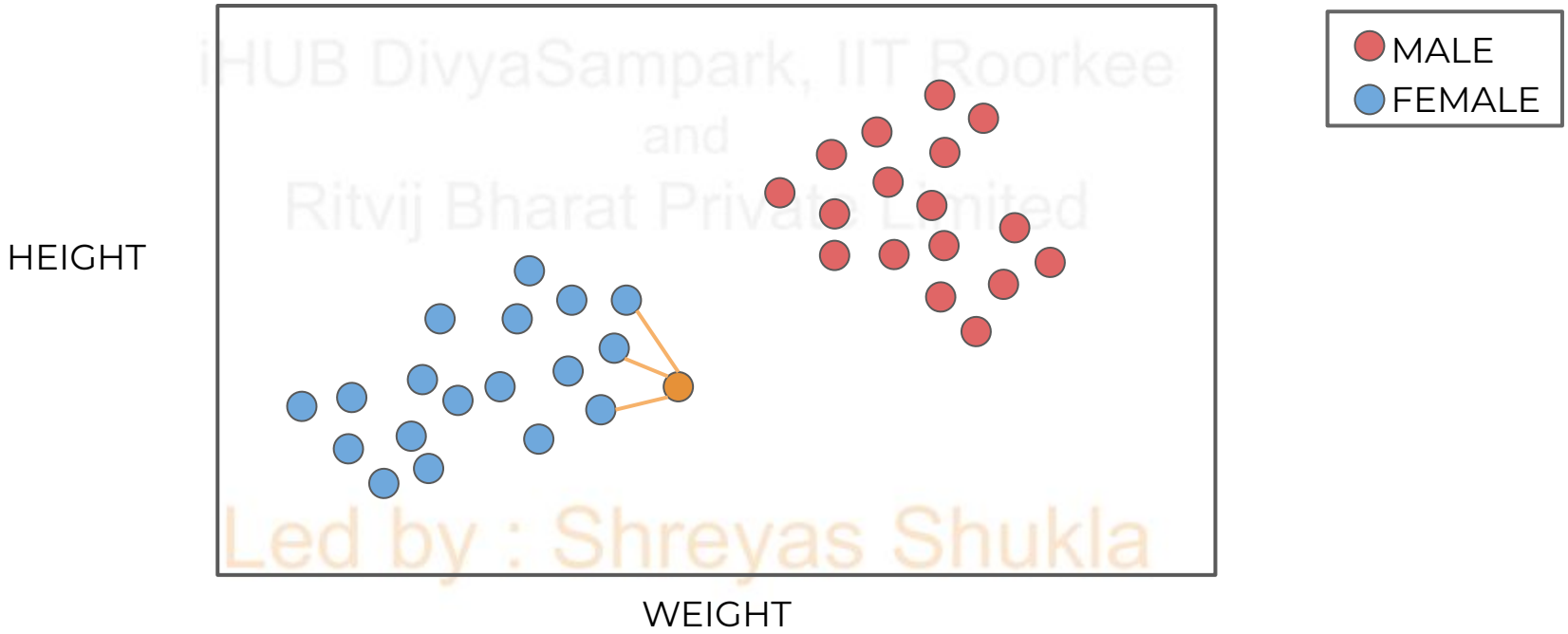


Say, We have a new data point

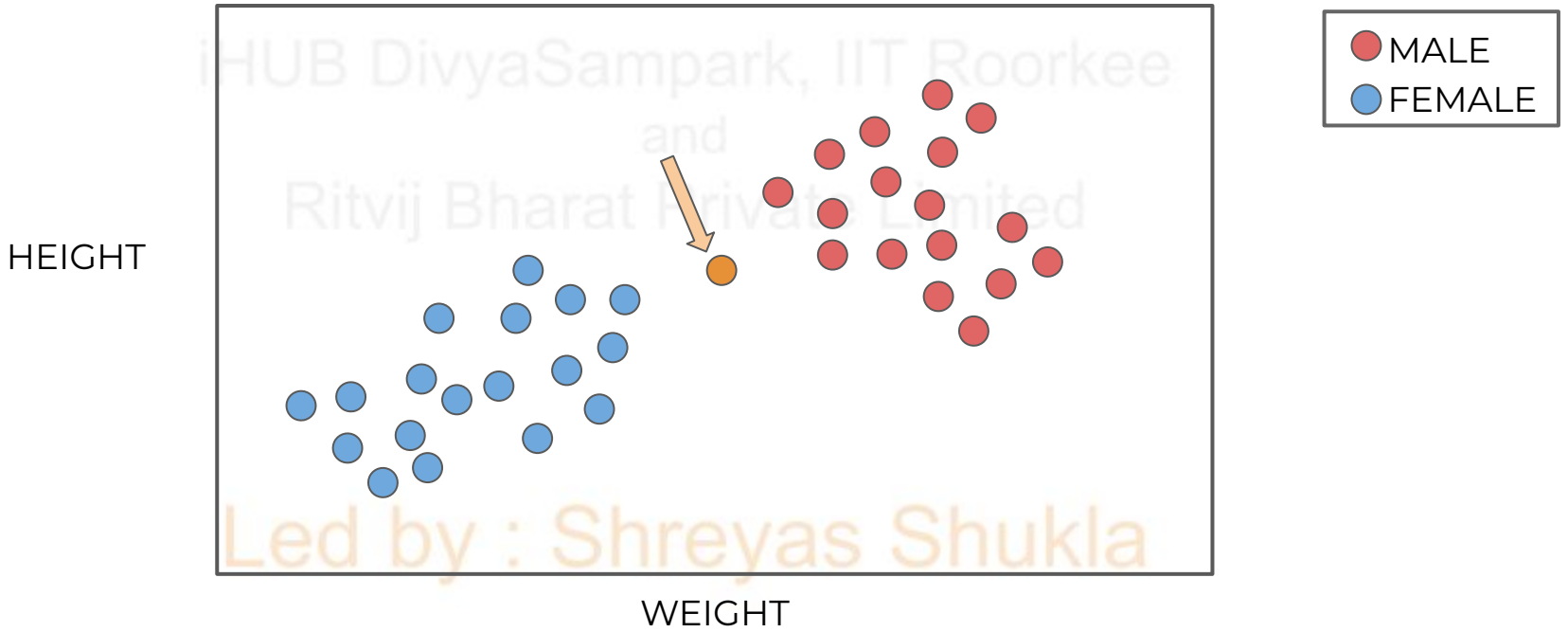
We intuitively “know” this is likely female.



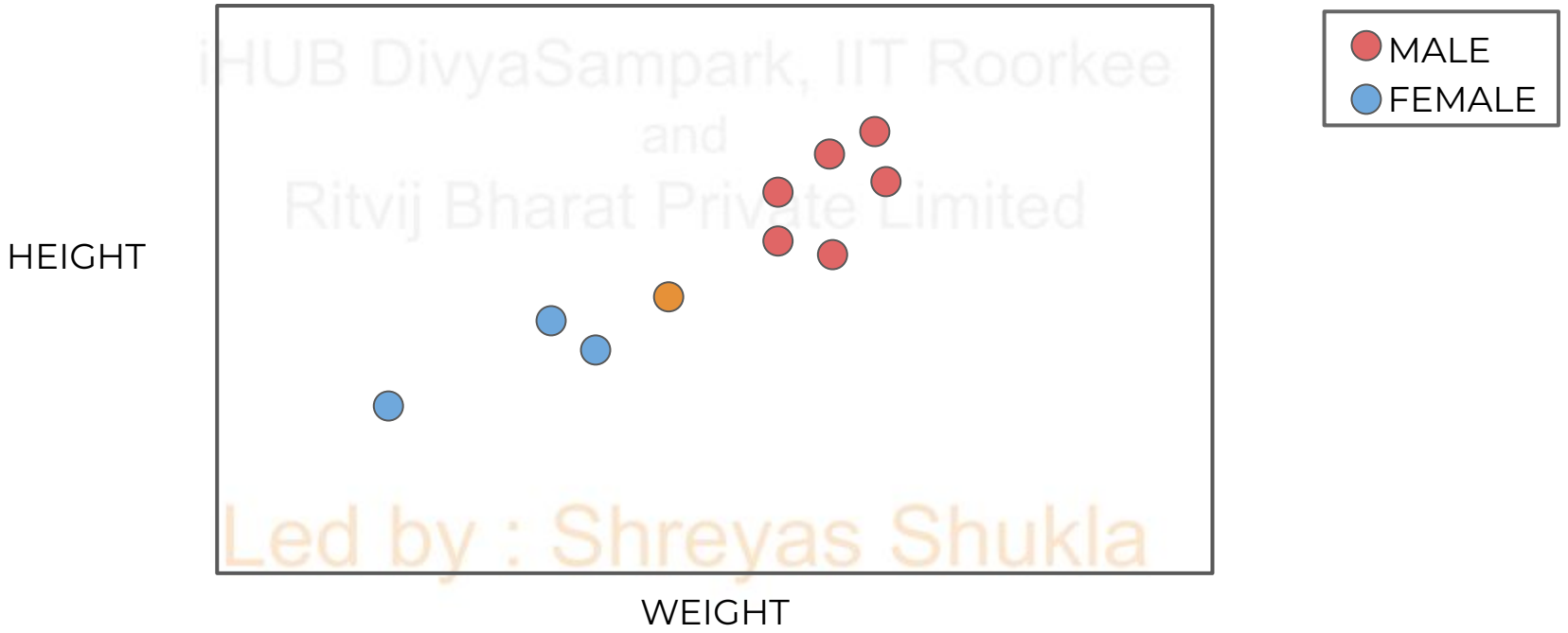
Our Intuition comes from **distance** to points!



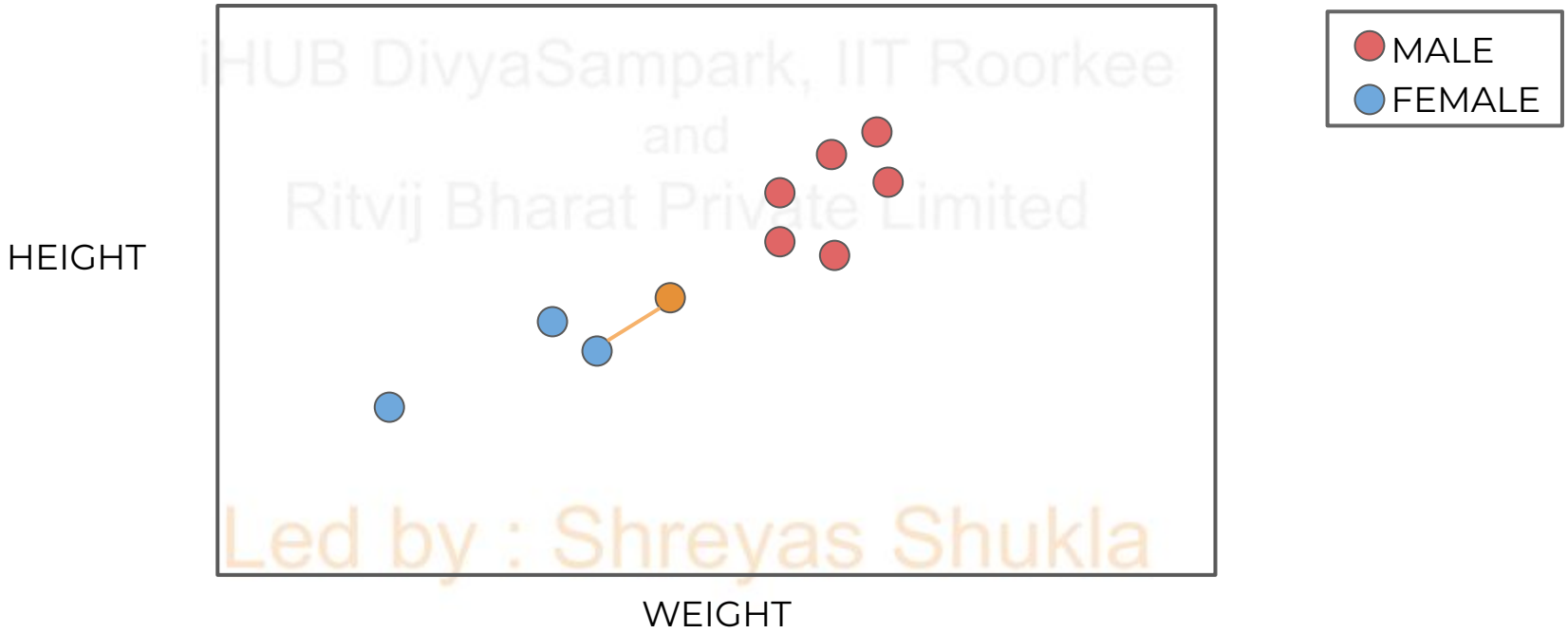
But what about a less obvious one?



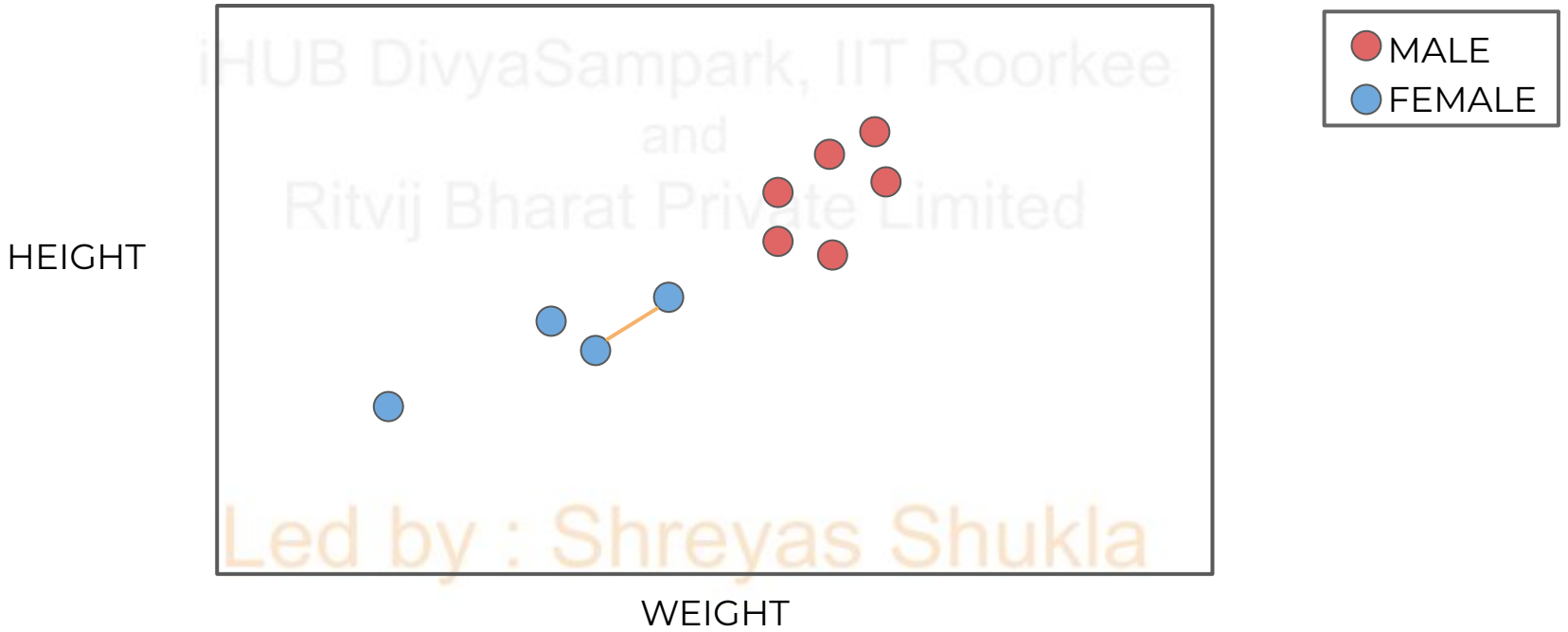
Let's imagine a situation like this:



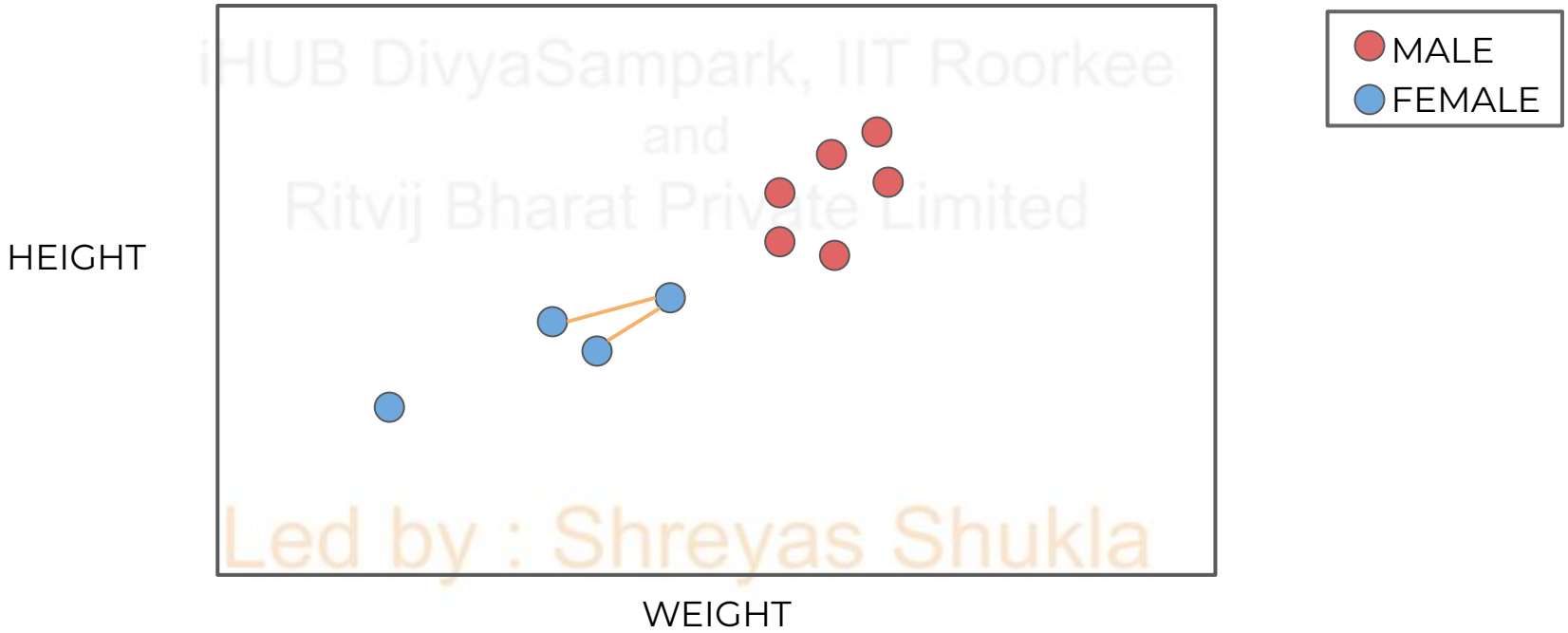
K=1



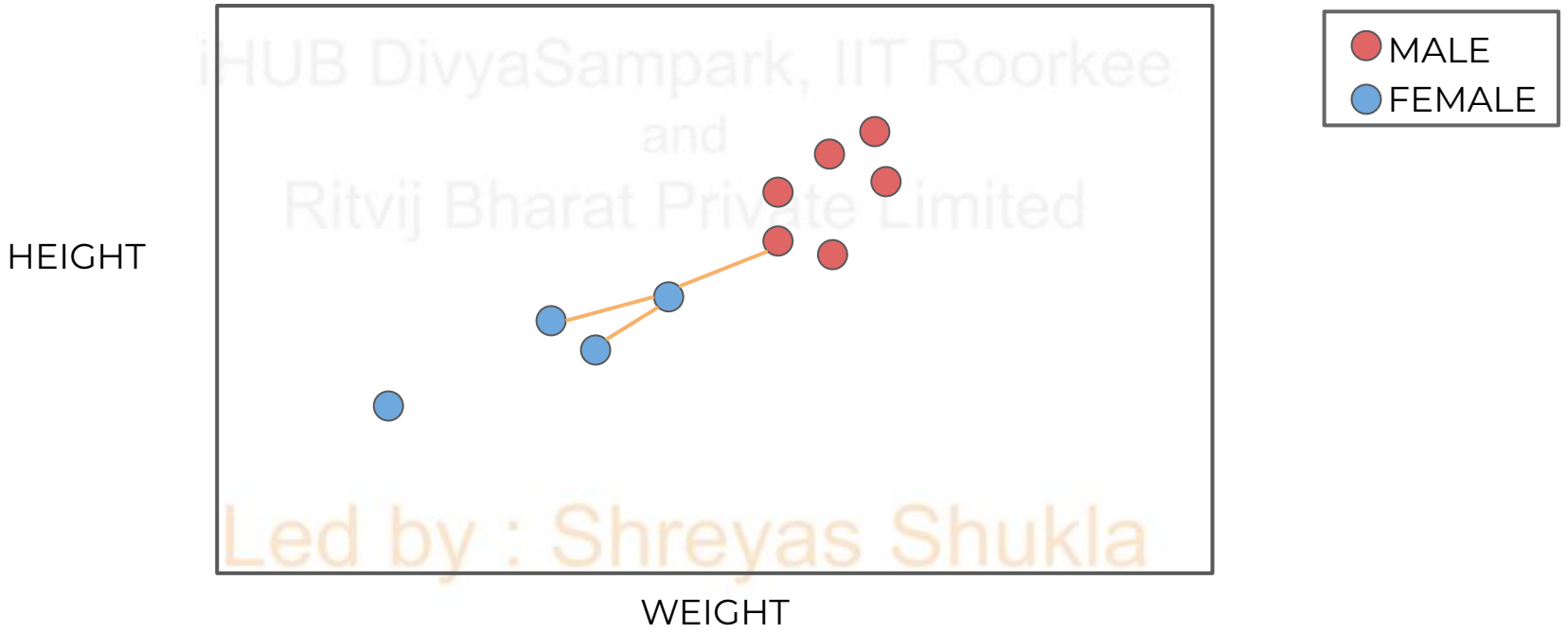
K=1



K=2

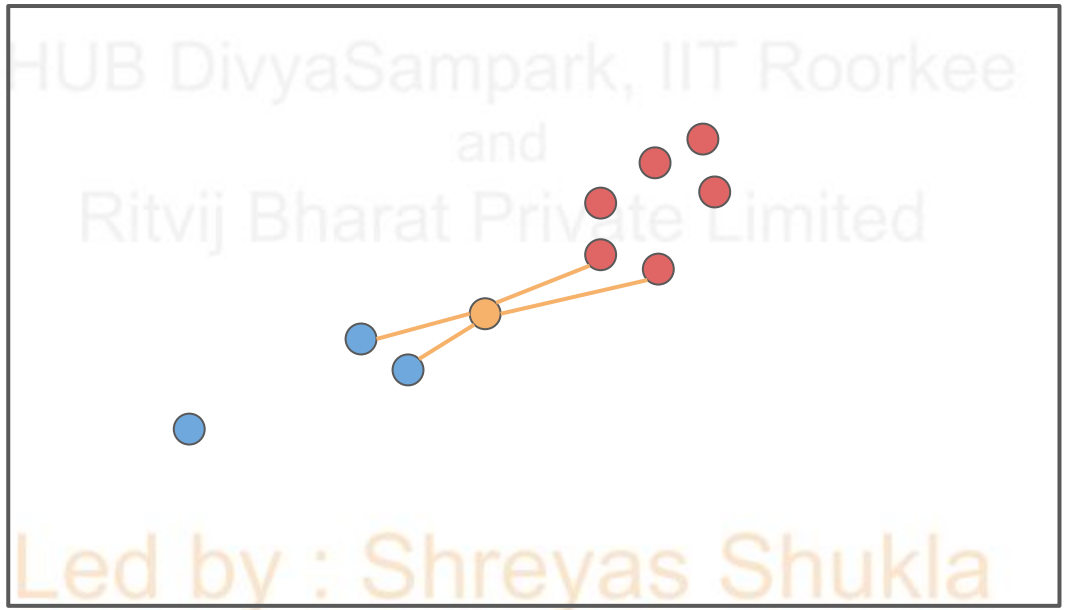


K=3



K=4

We have a TIE !!



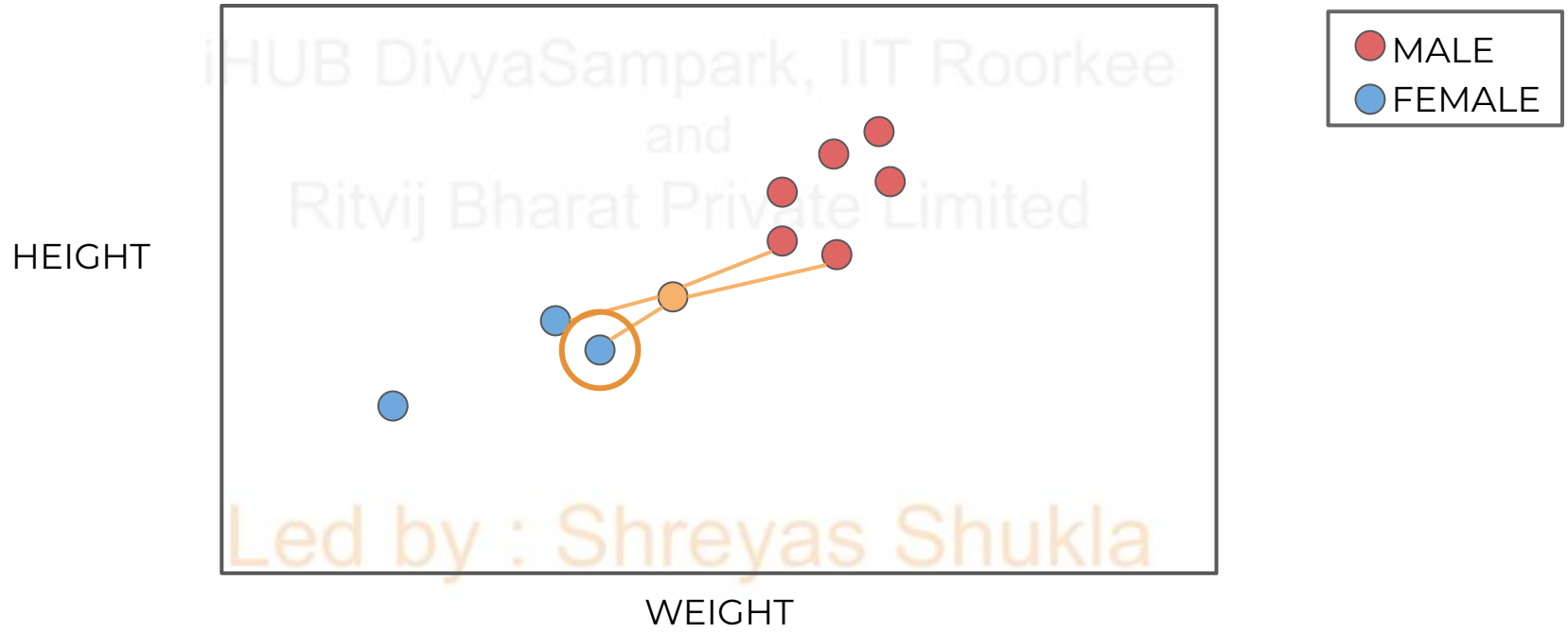
Tie considerations and options:

- Always choose an odd K.
- Reduce K by 1 until tie is broken.
- Randomly break tie.
- Choose nearest class point.

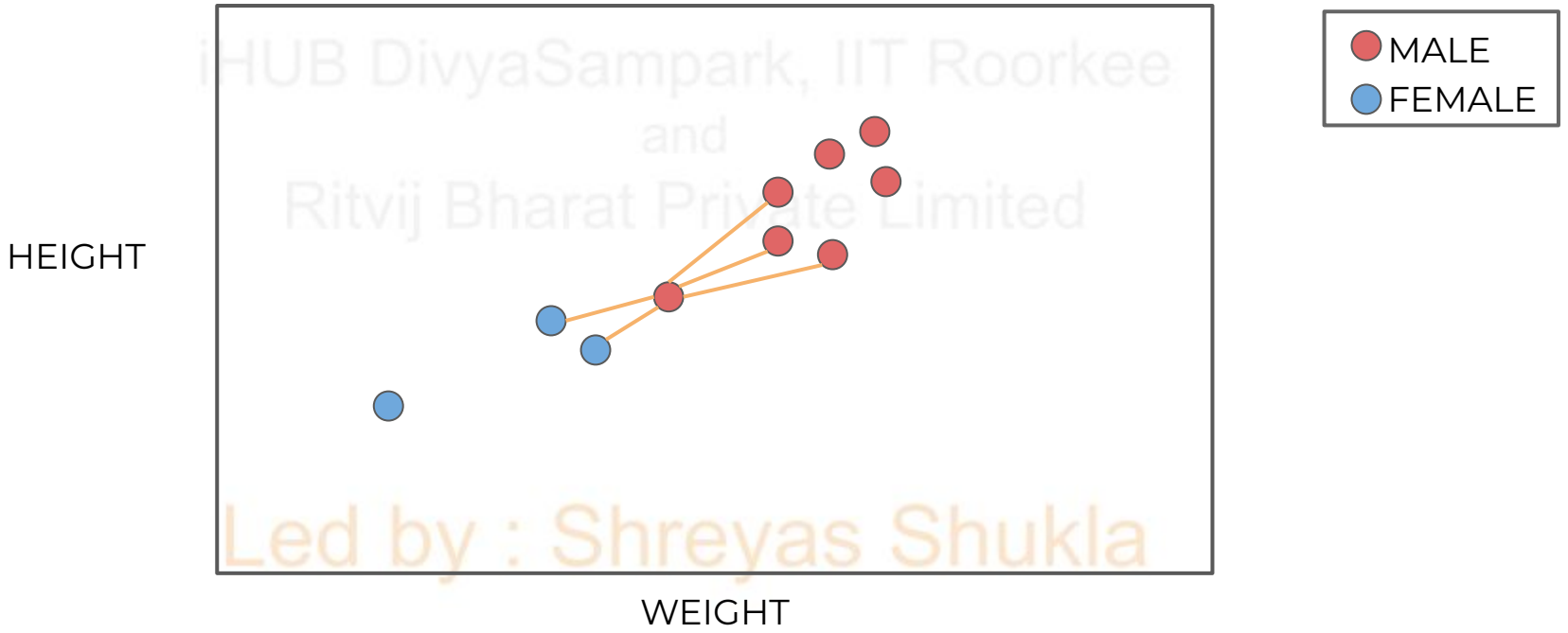
What does Scikit-Learn do here?

Warning: Regarding the Nearest Neighbors algorithms, if it is found that two neighbors, neighbor k+1 and k, have identical distances but different labels, the results will depend on the ordering of the training data.

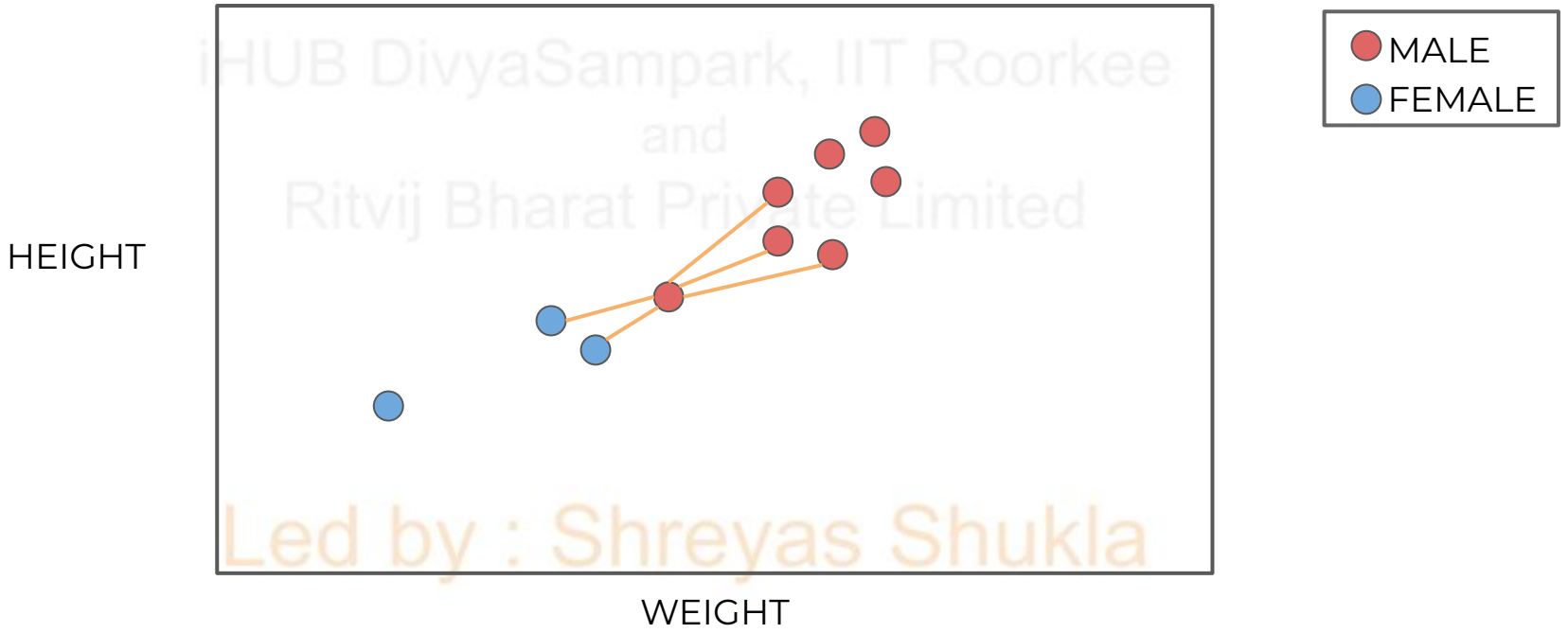
Choose closest K for K = 4



K=5 causes a switch from previous K values.



How to choose best K value?



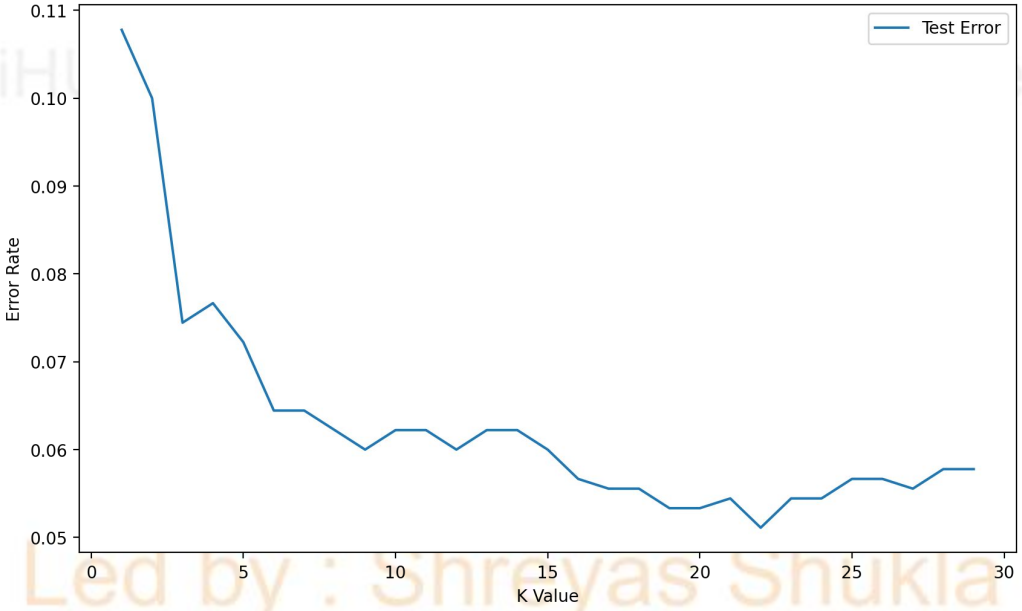
We want a K value that **minimizes** error:

- Error = 1 - Accuracy

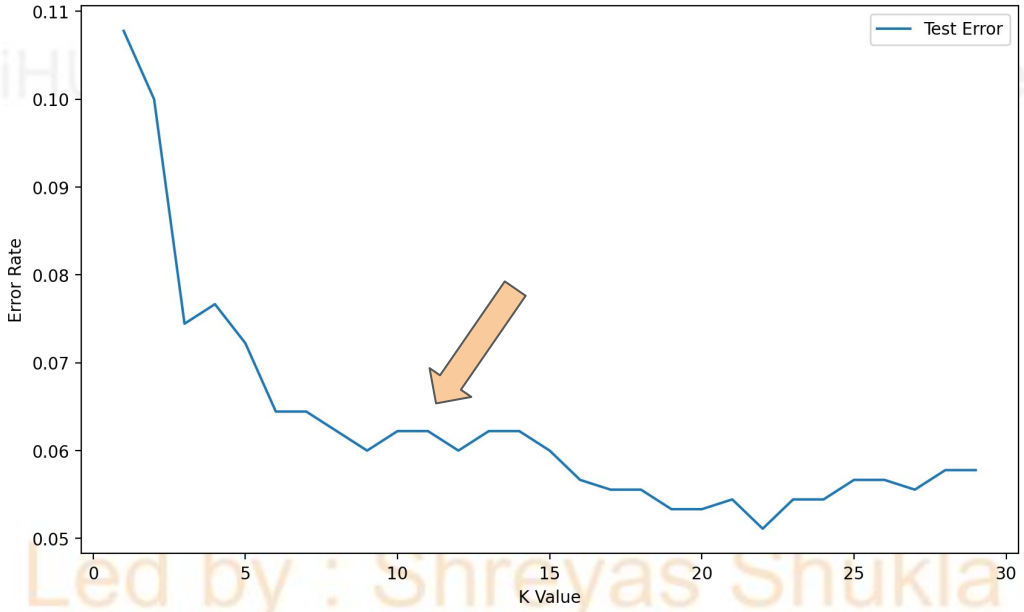
Two methods:

- Elbow method.
- Cross validate a grid search of multiple K values and choose K that results in lowest error or highest accuracy.

Elbow method:



Elbow method:



CV only takes into account the K value with the lowest error rate across multiple folds.

This could result in a more complex model

Consider the context of the problem to decide if larger K values are an issue.

KNN Algorithm

- Choose K value.
- Sort feature vectors (N dimensional space) by distance metric.
- Choose class based on K nearest feature vectors.

KNN Considerations:

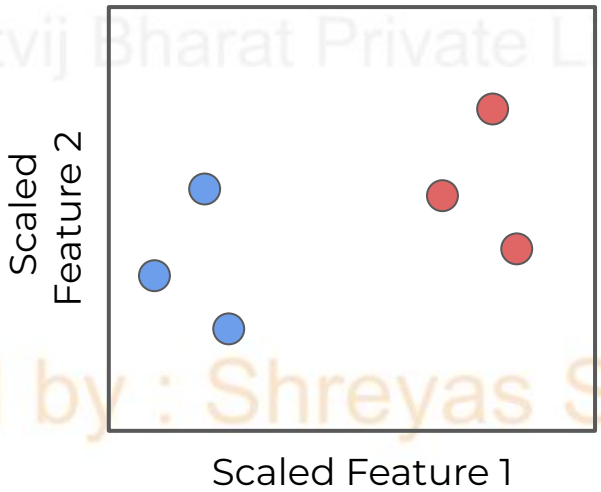
- Distance Metric
  - Many ways to measure distance:
    - Minkowski
    - Euclidean
    - Manhattan
    - Chebyshev

KNN Considerations:

- Scaling for Distance
  - Features could have vastly different value ranges!

KNN Considerations:

- Scaling is necessary for KNN.



Keep in mind the following considerations:

- Choosing the optimal K value.
- Scaling features.

Let’s explore how to perform KNN for classification!