

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

K Nearest Neighbors

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

One of the simplest machine learning algorithms.

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

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

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

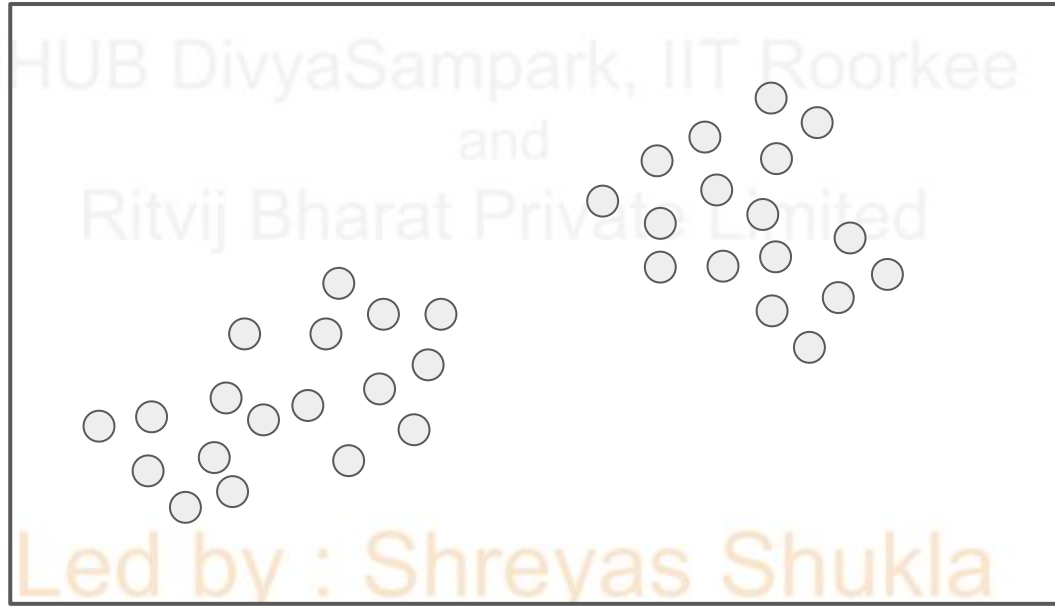
How could we train an algorithm to identify the sex of a new baby chick based on historical features?

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

HEIGHT



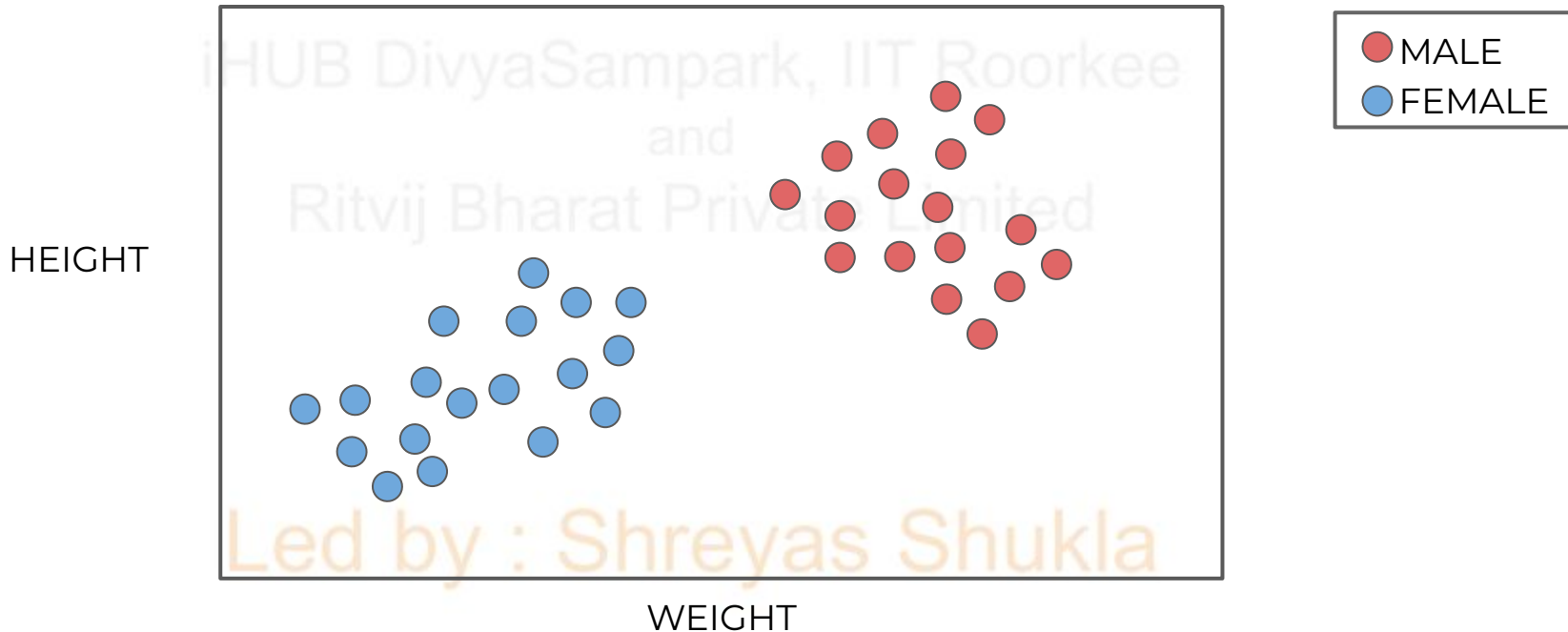
WEIGHT

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

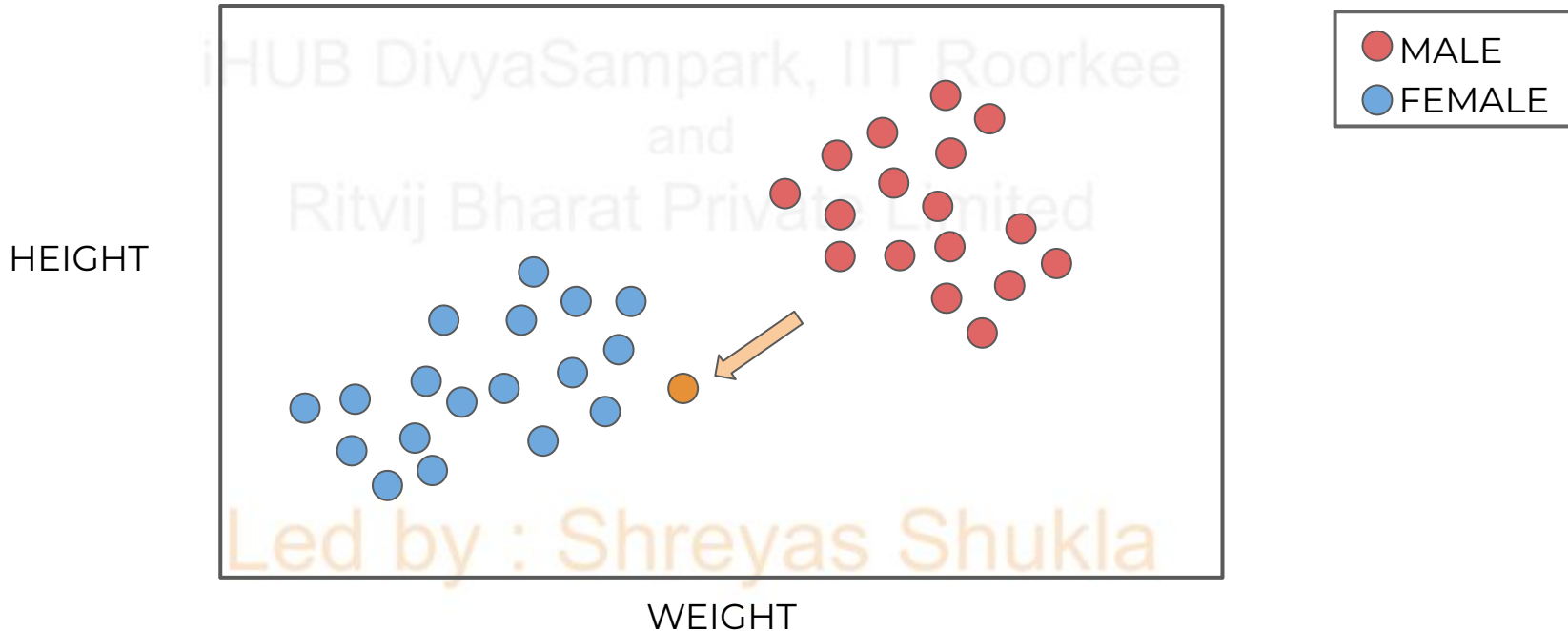
Historically, we know the sex of the chicks:



Say, We have a new data point

(27th Aug 2024 - 18th Oct 2024)

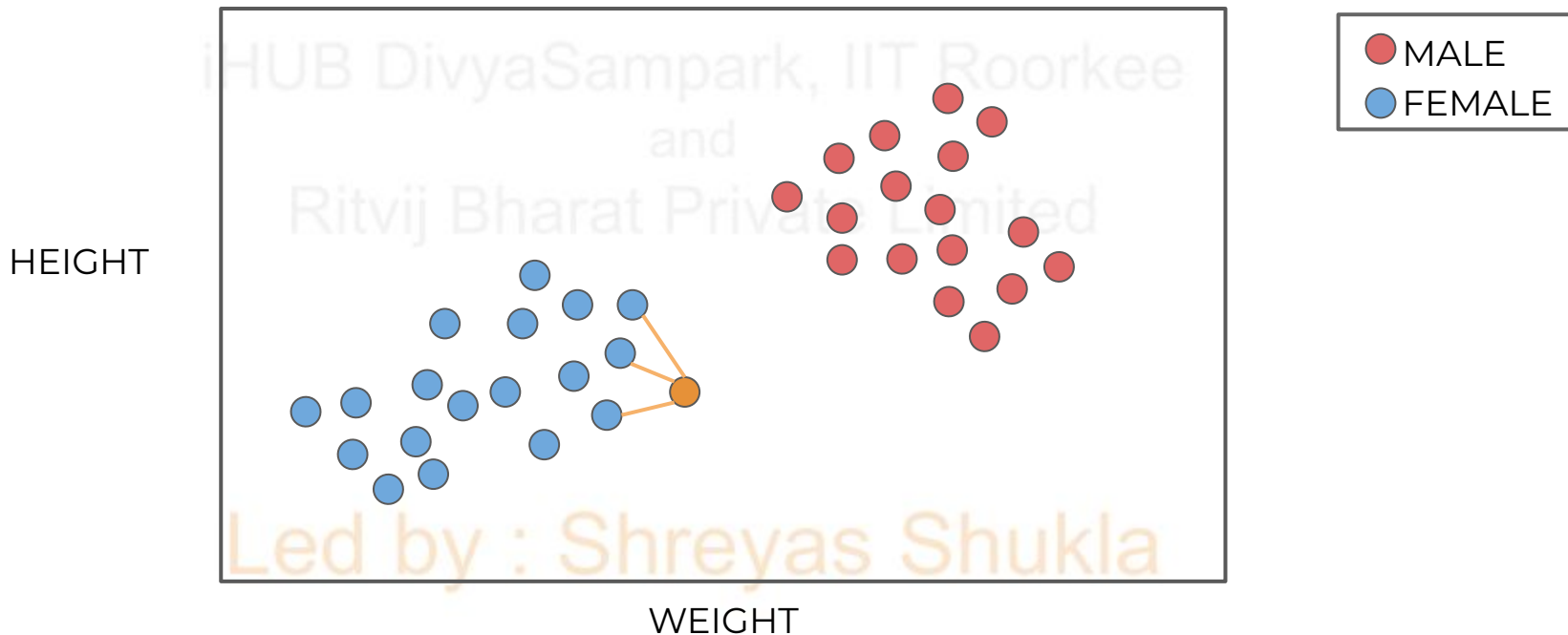
We intuitively “know” this is likely female.



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

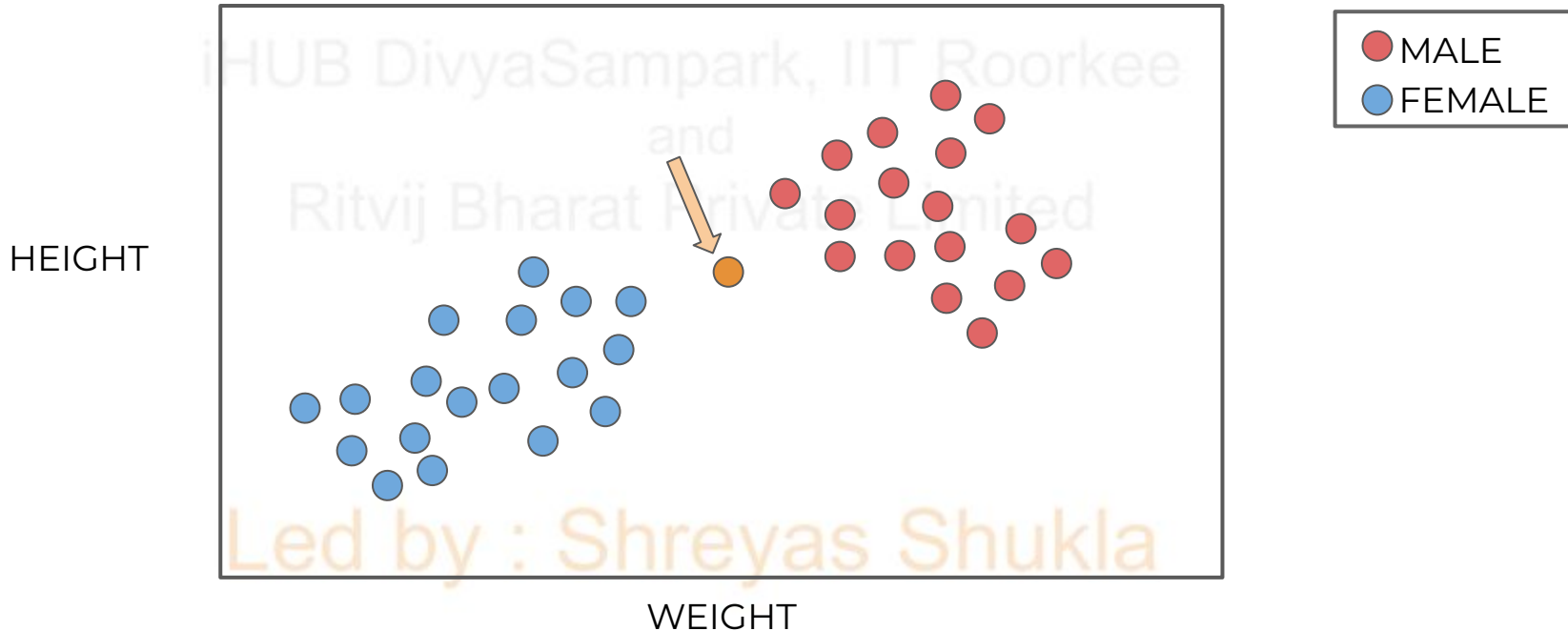
Our Intuition comes from **distance** to points!



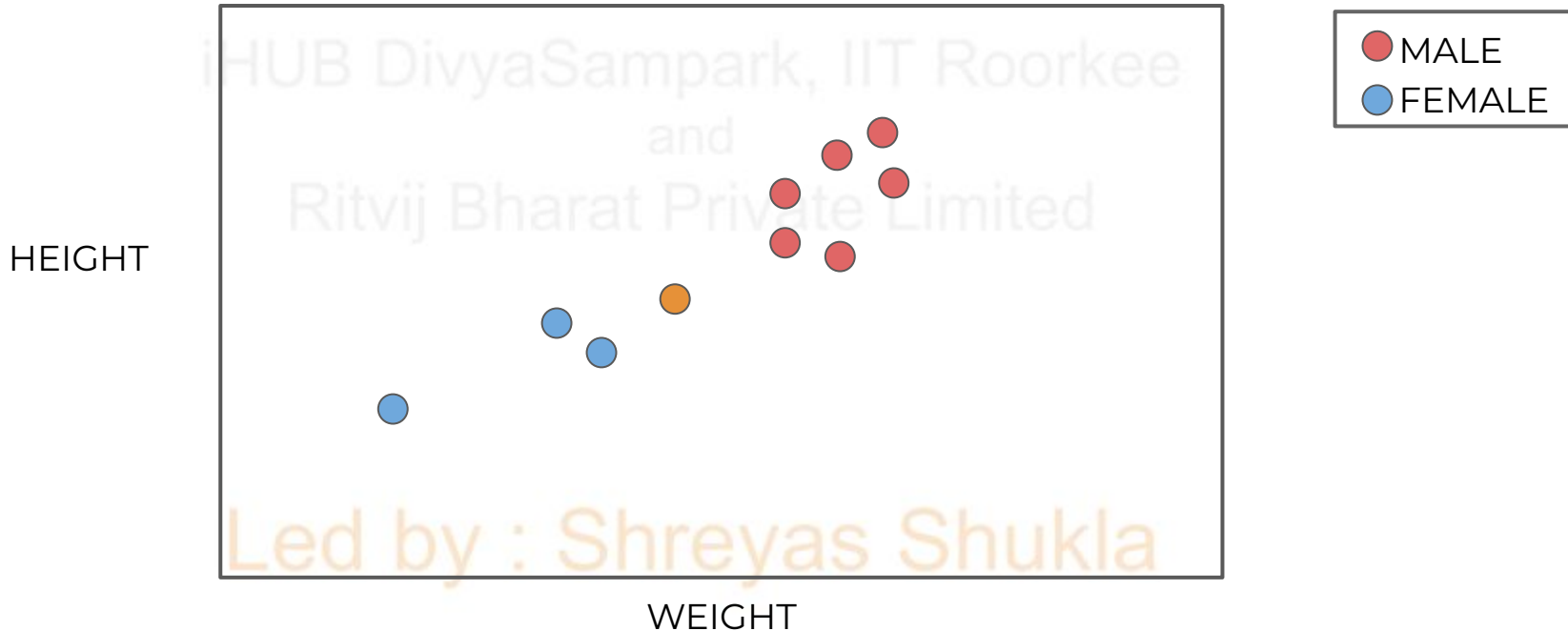
Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

But what about a less obvious one?

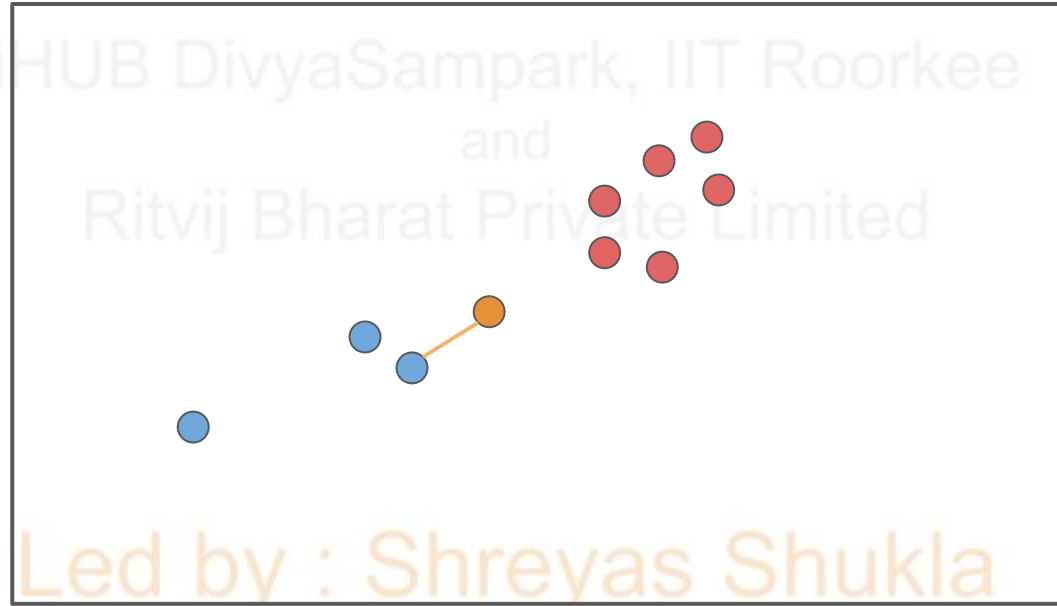


Let's imagine a situation like this:



K=1

HEIGHT



WEIGHT

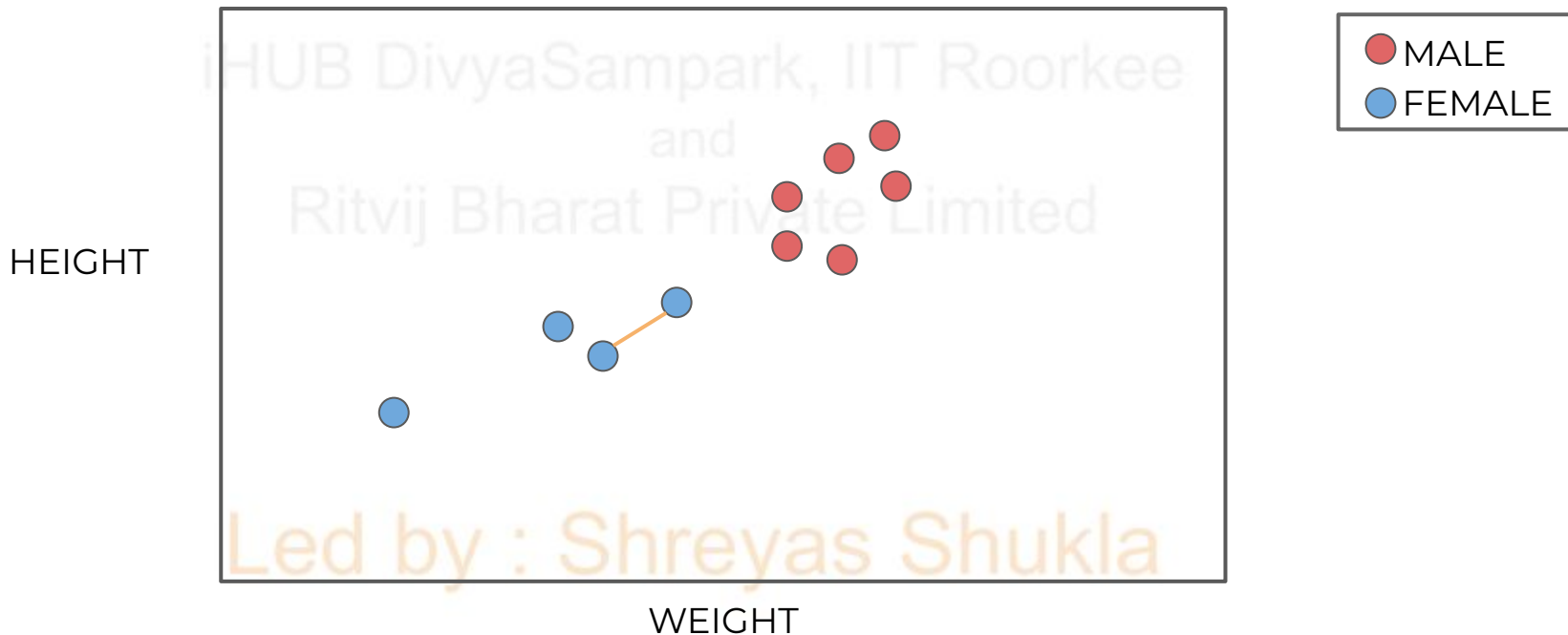
MALE
FEMALE

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

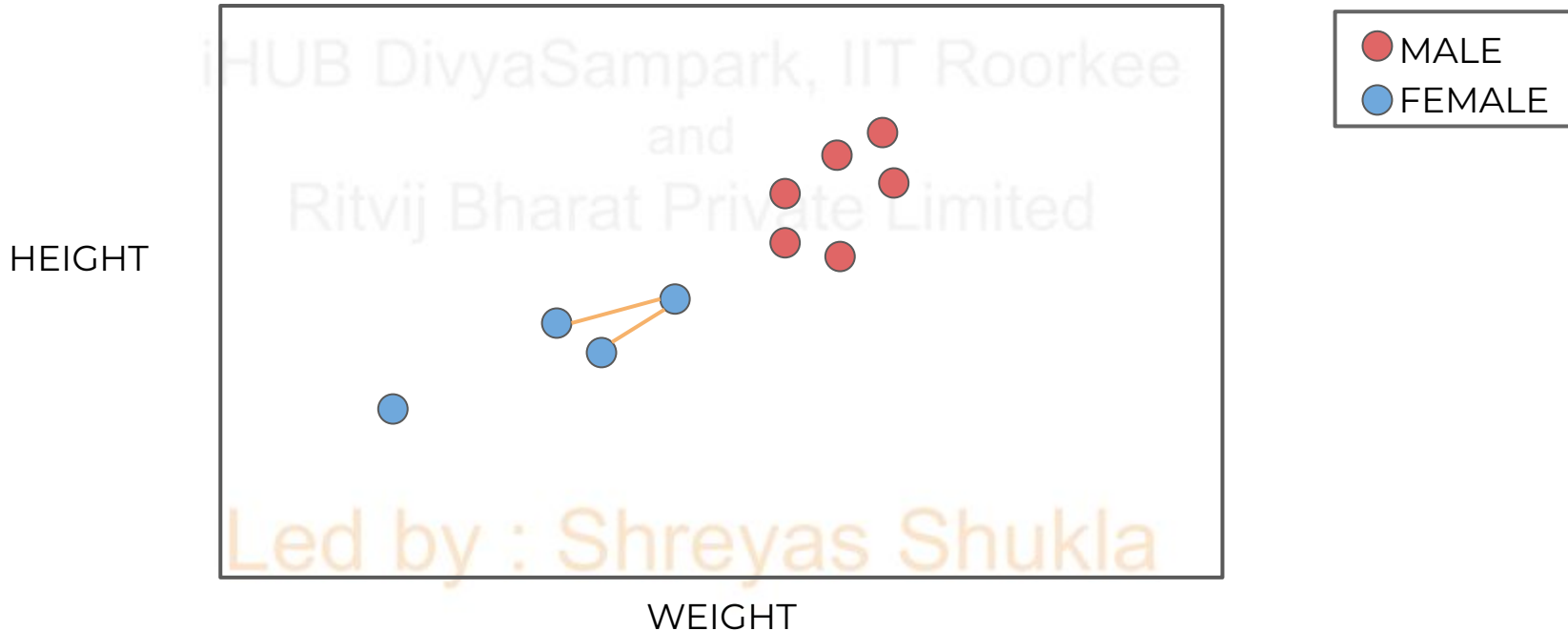
K=1



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

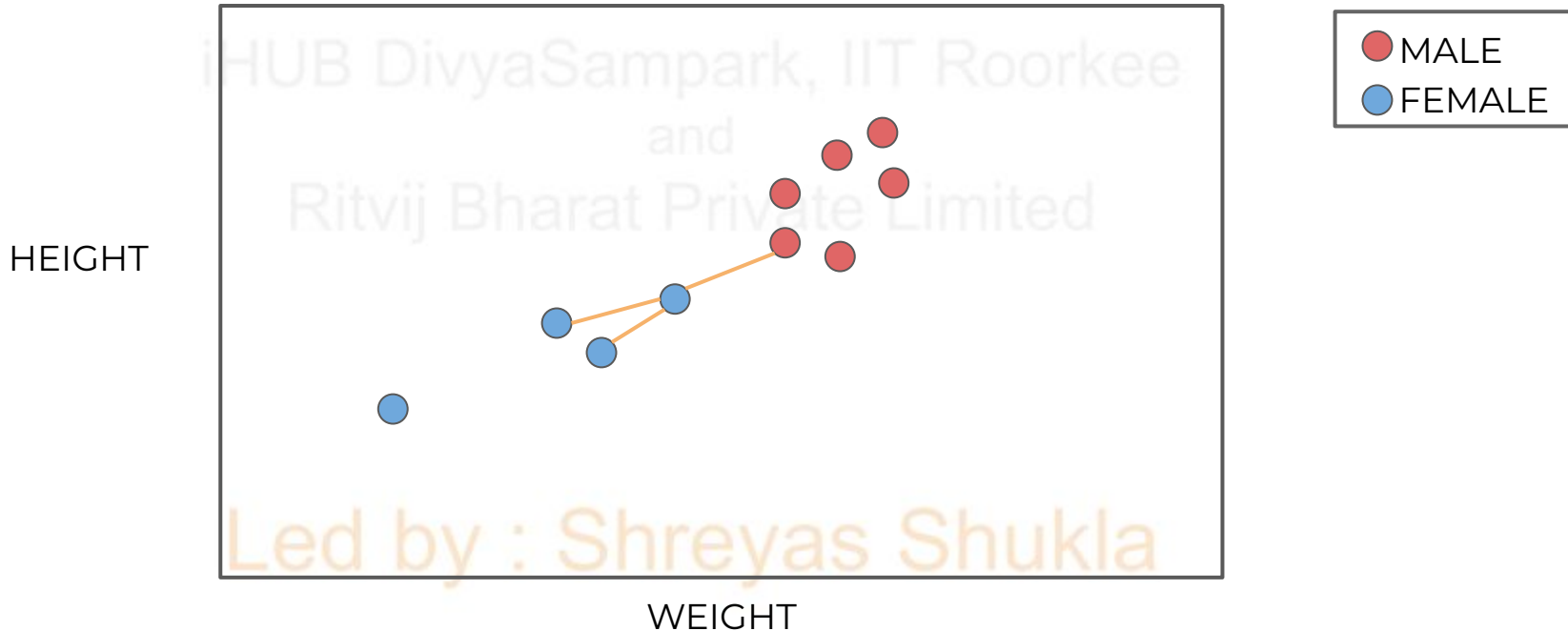
K=2



Mastering Machine Learning with Python

K=3

(27th Aug 2024 - 18th Oct 2024)

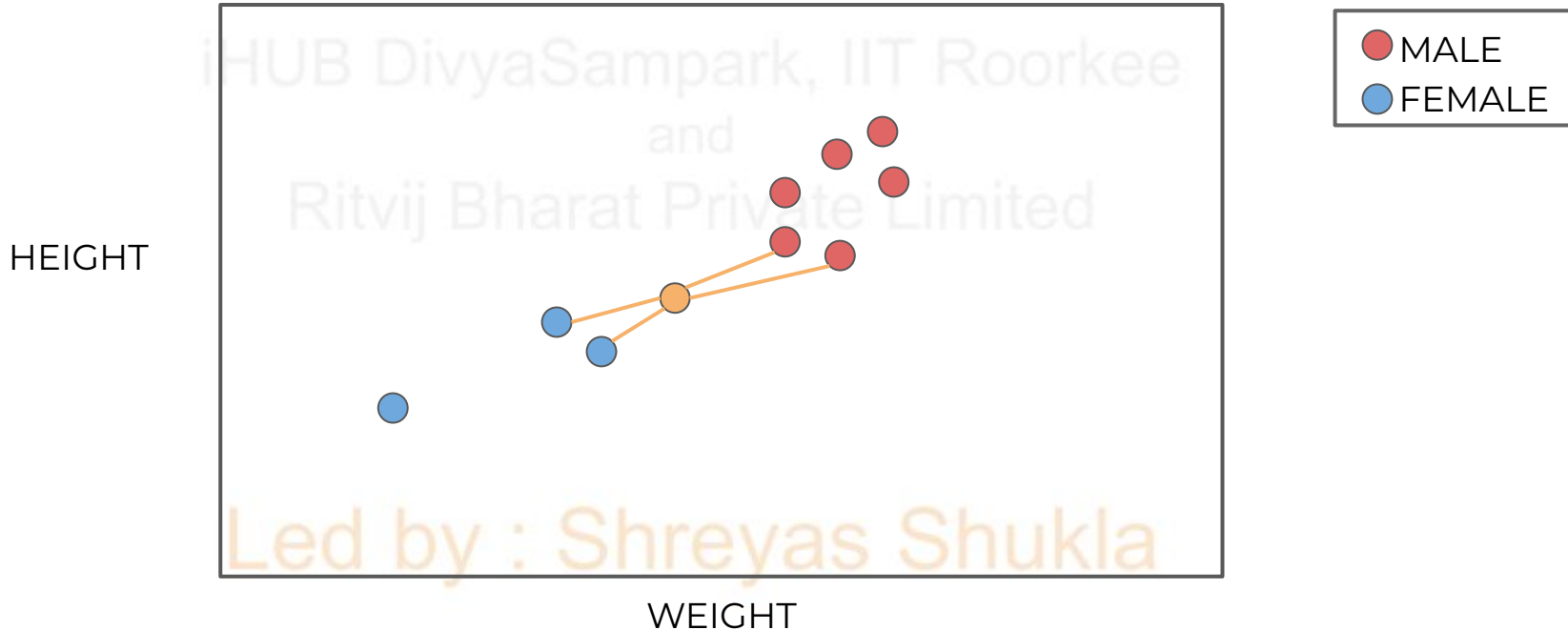


Mastering Machine Learning with Python

K=4

(27th Aug 2024 - 18th Oct 2024)

We have a TIE !!



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

Tie considerations and options:

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

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

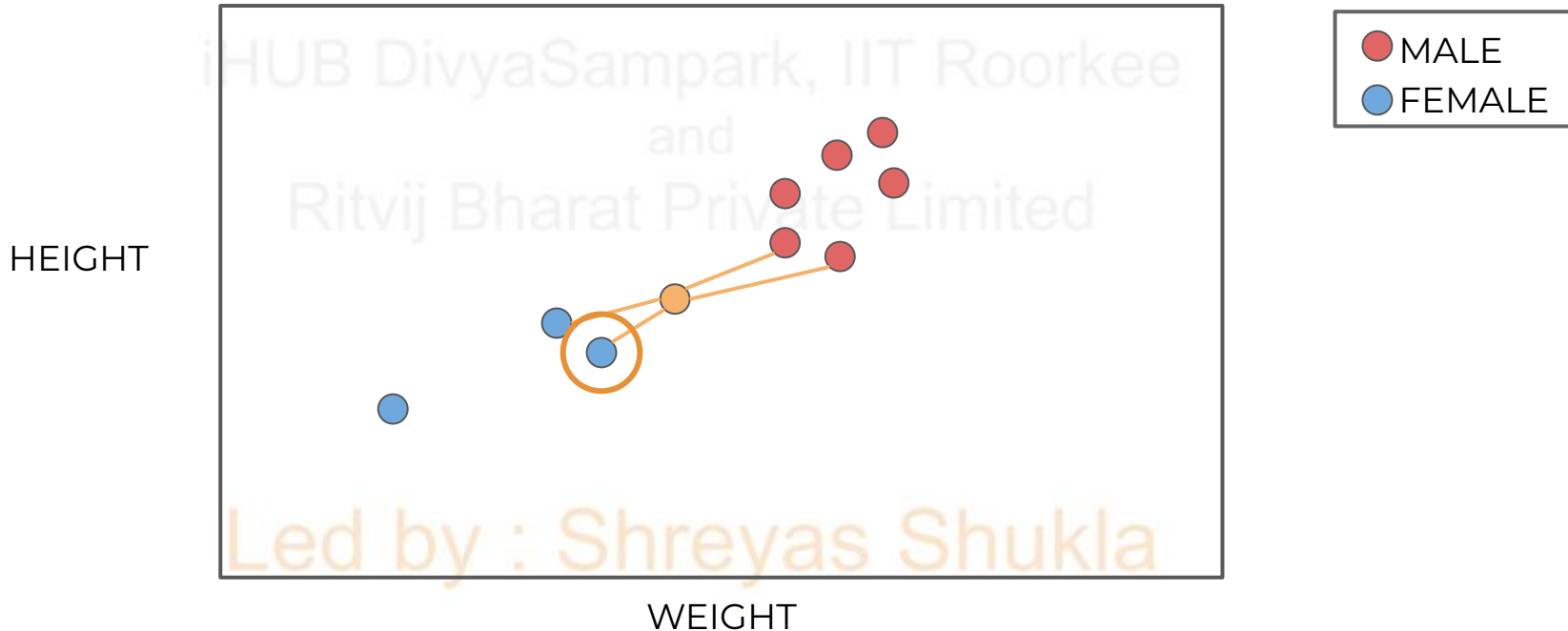
What does Scikit-Learn do here?

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.

Led by : Shreyas Shukla

Mastering Machine Learning with Python

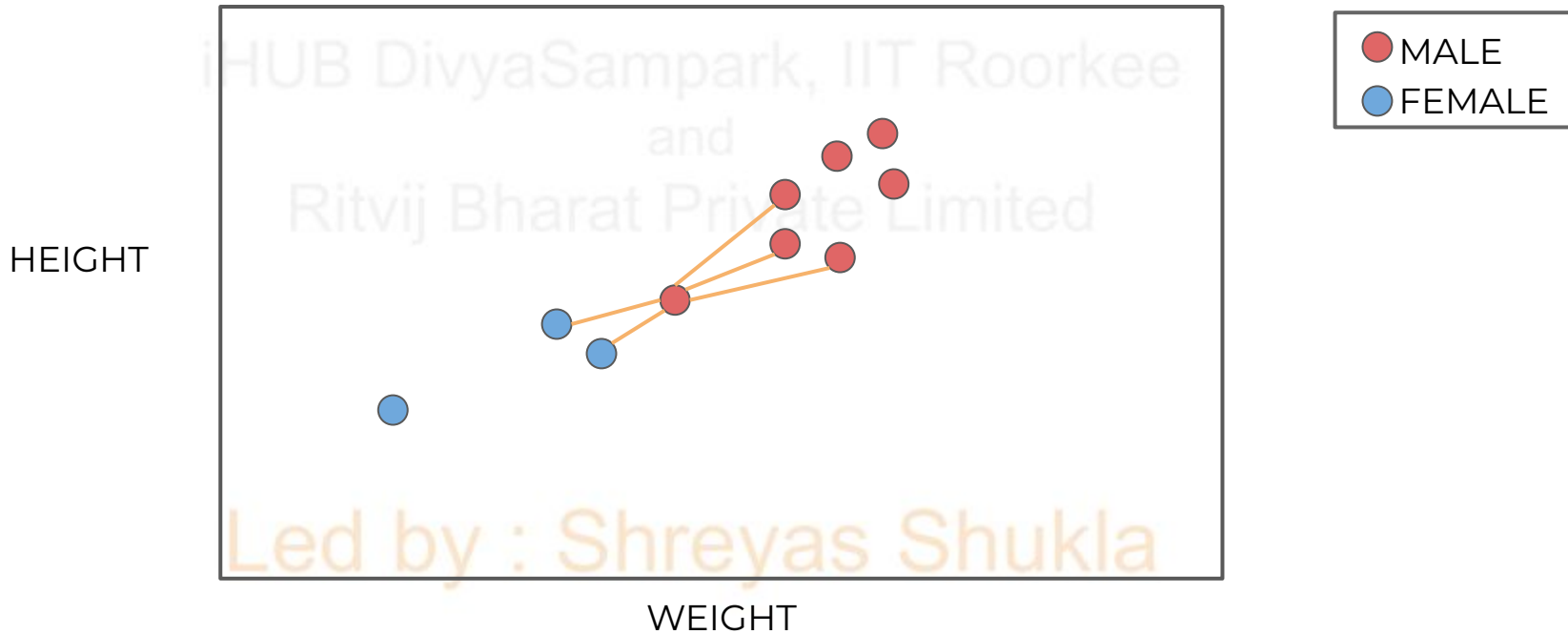
Choose closest K for K = 4



Mastering Machine Learning with Python

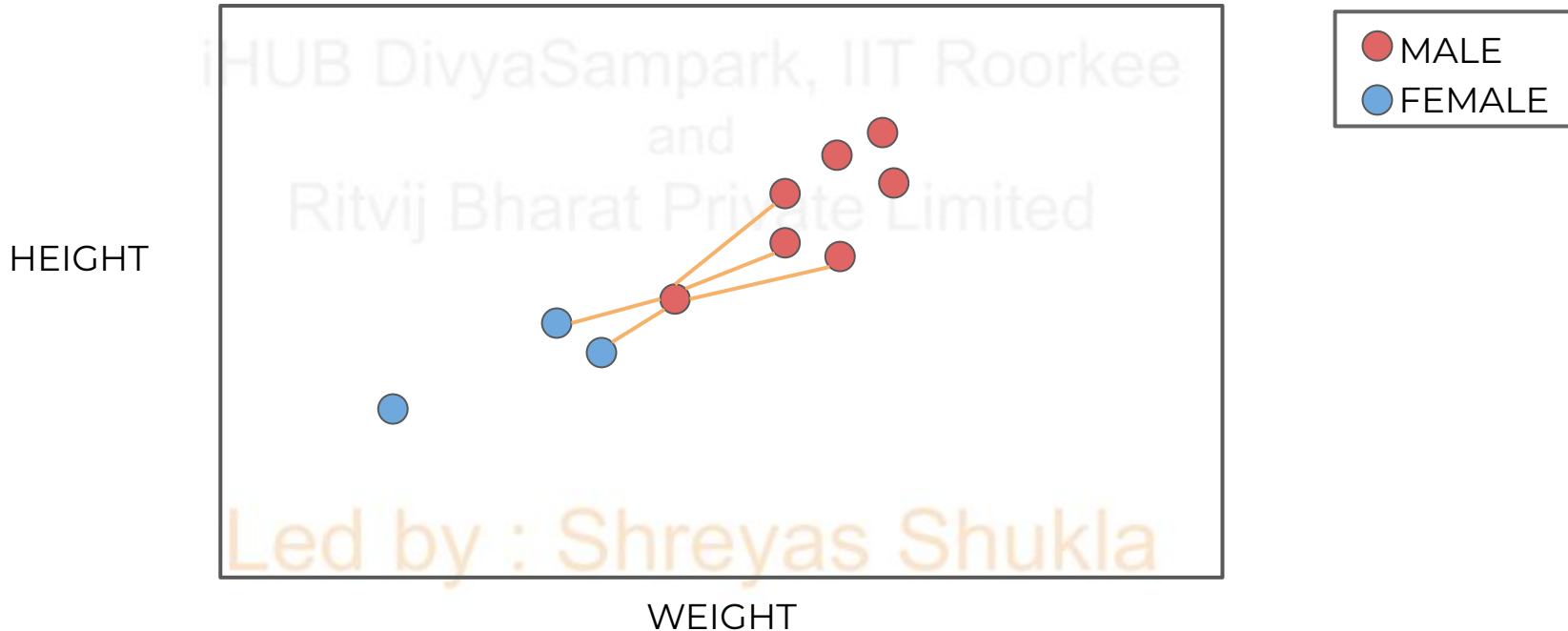
(27th Aug 2024 - 18th Oct 2024)

K=5 causes a switch from previous K values.



How to choose best K value?

(27th Aug 2024 - 18th Oct 2024)



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

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

- Error = 1 - Accuracy

Two methods:

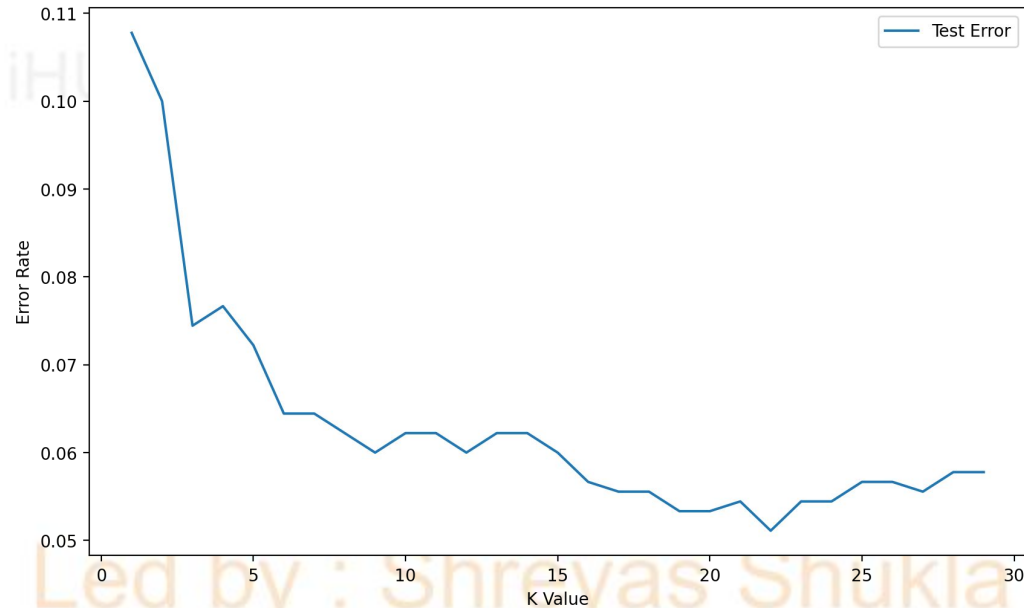
- Elbow method.
- Cross validate a grid search of multiple K values and choose K that results in lowest error or highest accuracy.

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

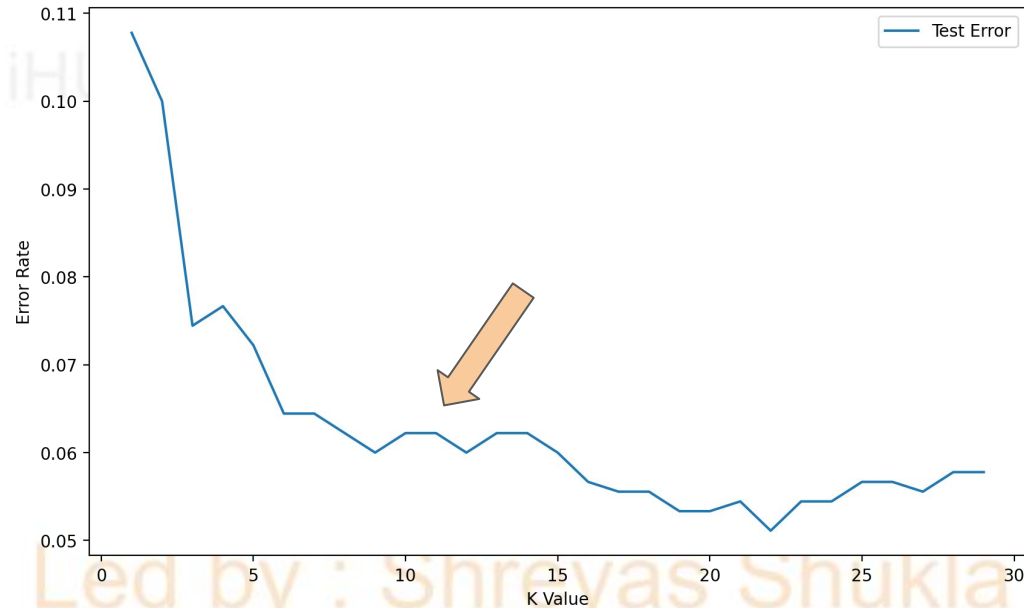
Elbow method:



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

Elbow method:



Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

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.

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

KNN Algorithm

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

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

KNN Considerations:

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

Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

KNN Considerations:

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



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

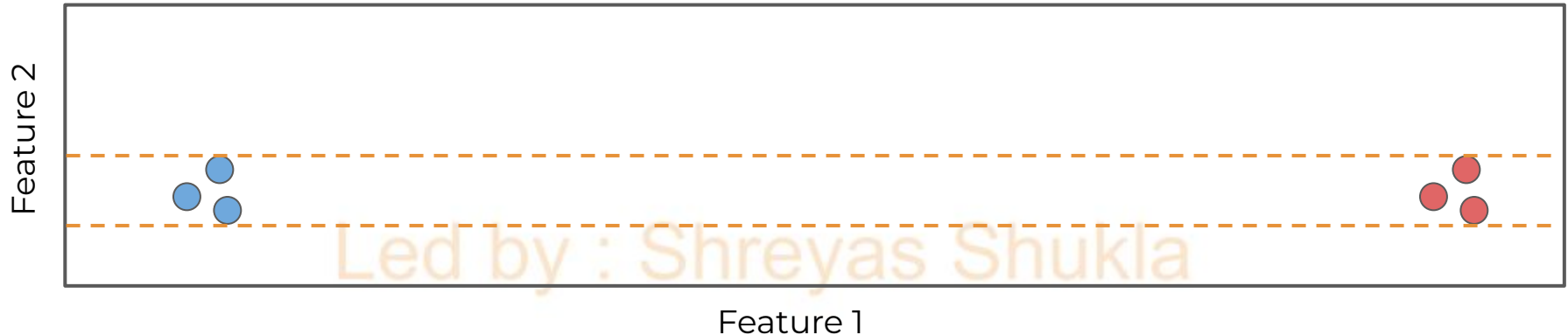
iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited



Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

iHUB DivyaSampark, IIT Roorkee
and
Ritvij Bharat Private Limited

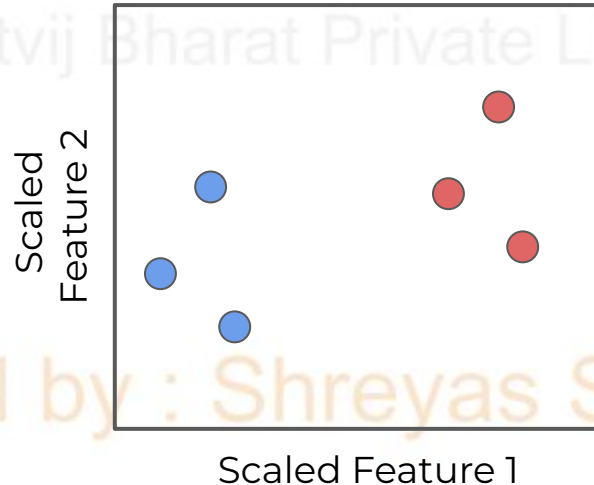


Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

KNN Considerations:

- Scaling is necessary for KNN.



Led by : Shreyas Shukla

Mastering Machine Learning with Python

(27th Aug 2024 - 18th Oct 2024)

Keep in mind the following considerations:

- Choosing the optimal K value.
- Scaling features.

Let's explore how to perform KNN for classification!

Led by : Shreyas Shukla