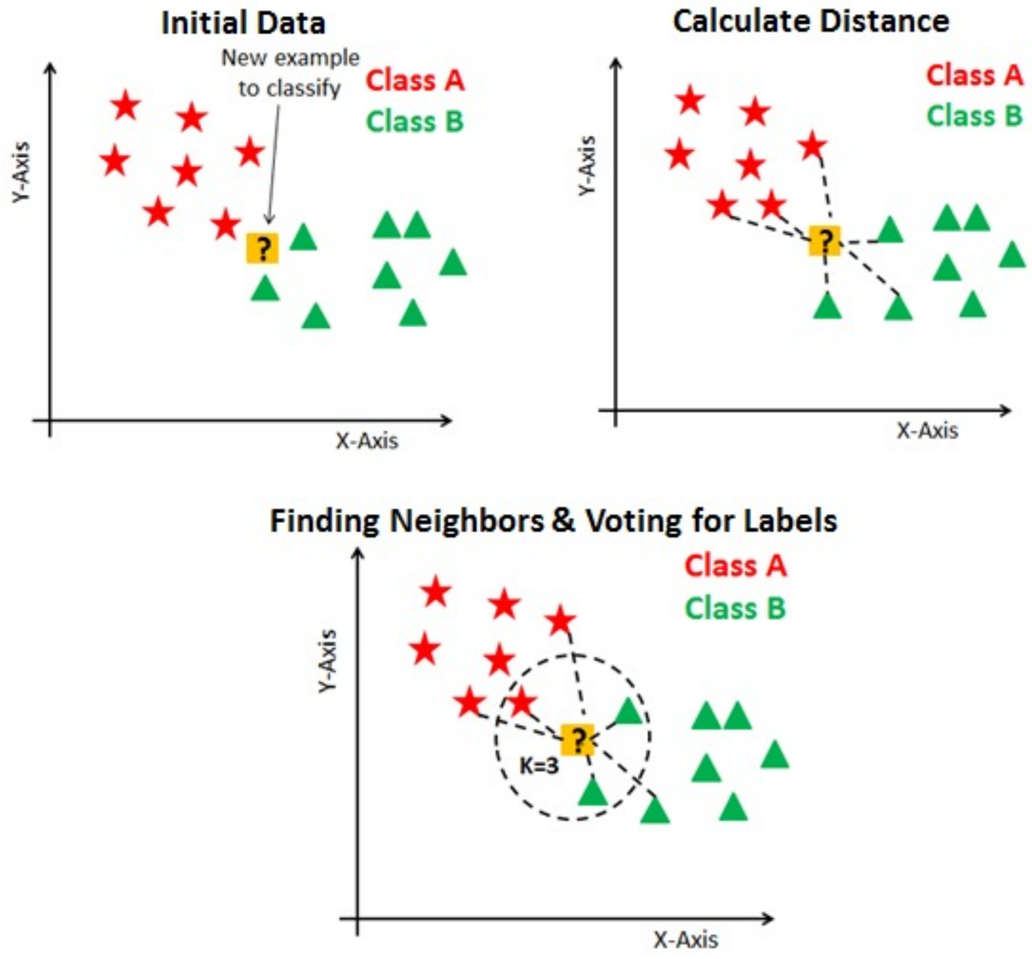
K-Nearest Neighbors (K-NN)

K Nearest Neighbors



Initial Data

Calculate Distance

Finding Neighbors & Voting for Labels

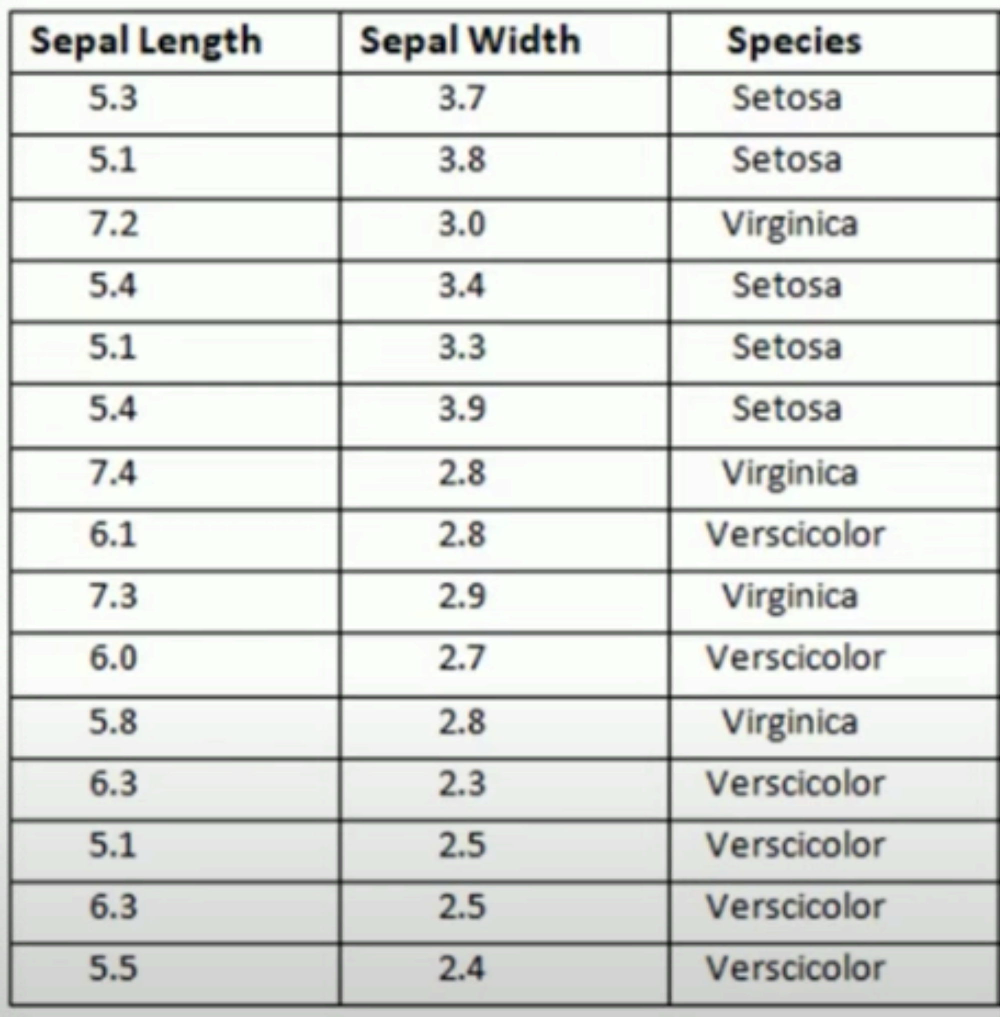
Steps:

- **Step 1: Find the distance.** We need to find the distance from this example to all examples (the 15 data rows given). We will be using Euclidian distance formula (any other distance formulas can be used, e.g., Manhattan distance formula).
- **Step 2: Find Rank.** Calculate the rank of each data point with respect to the new data point given. So, the minimum distance will have the first rank, the next distance will have the second rank, etc.
- **Step 3: Find the Nearest Neighbor.** Given the value of K (if K =1 for instance), we need to identify the first rank example.

Example 1:

- Given the value of K , now we want to classify this new example with the help of KNN classifier:

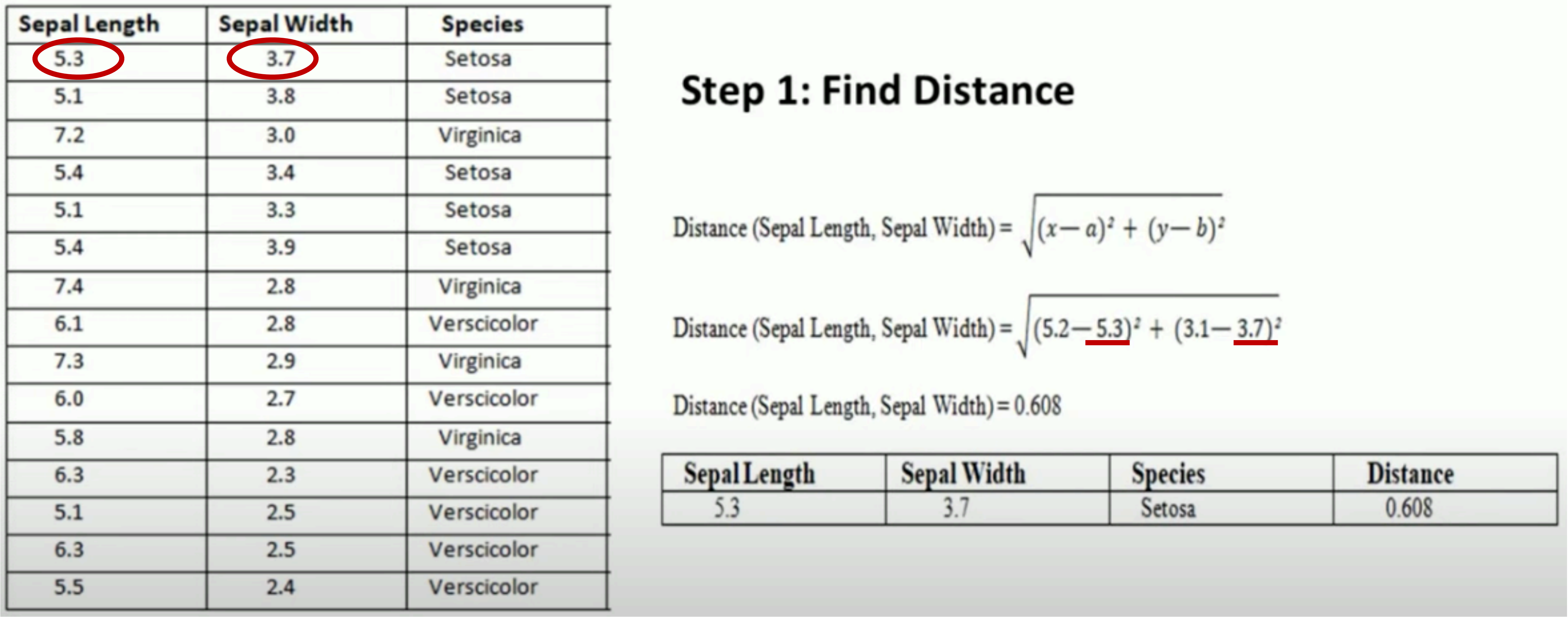
we would like to predict the species based on these two given features: Sepal Length = 5.2, Sepal Width = 3.1.



Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

Example 1

Calculate the distance between the given example and each example in the table.



Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

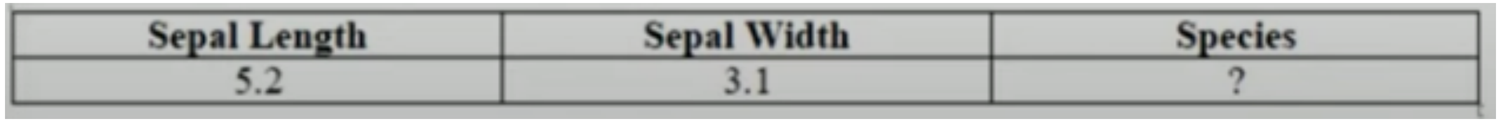
Step 1: Find Distance

$$\text{Distance (Sepal Length, Sepal Width)} = \sqrt{(x-a)^2 + (y-b)^2}$$

$$\text{Distance (Sepal Length, Sepal Width)} = \sqrt{(5.2-\underline{5.3})^2 + (3.1-\underline{3.7})^2}$$

$$\text{Distance (Sepal Length, Sepal Width)} = 0.608$$

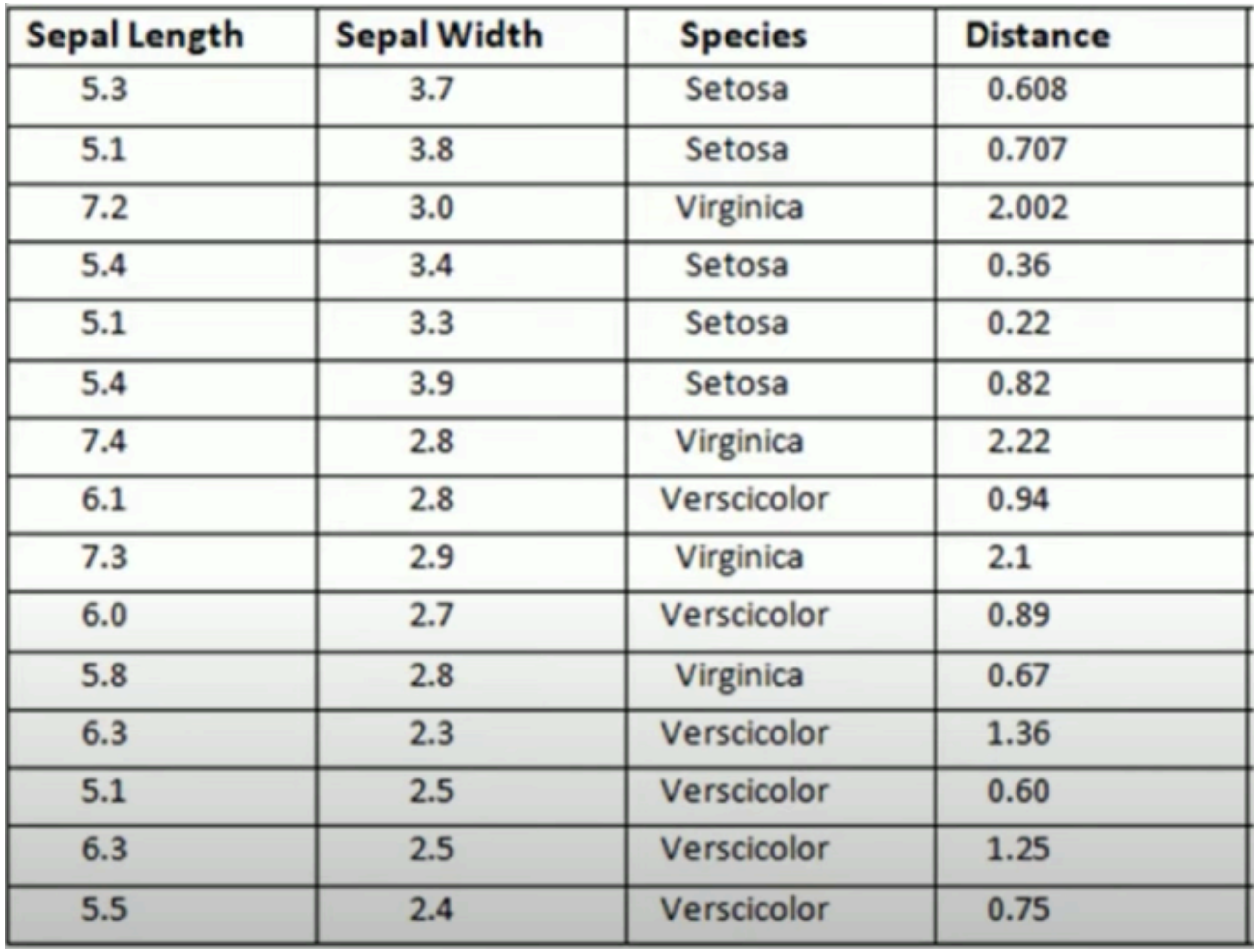
Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608



Sepal Length	Sepal Width	Species
5.2	3.1	?

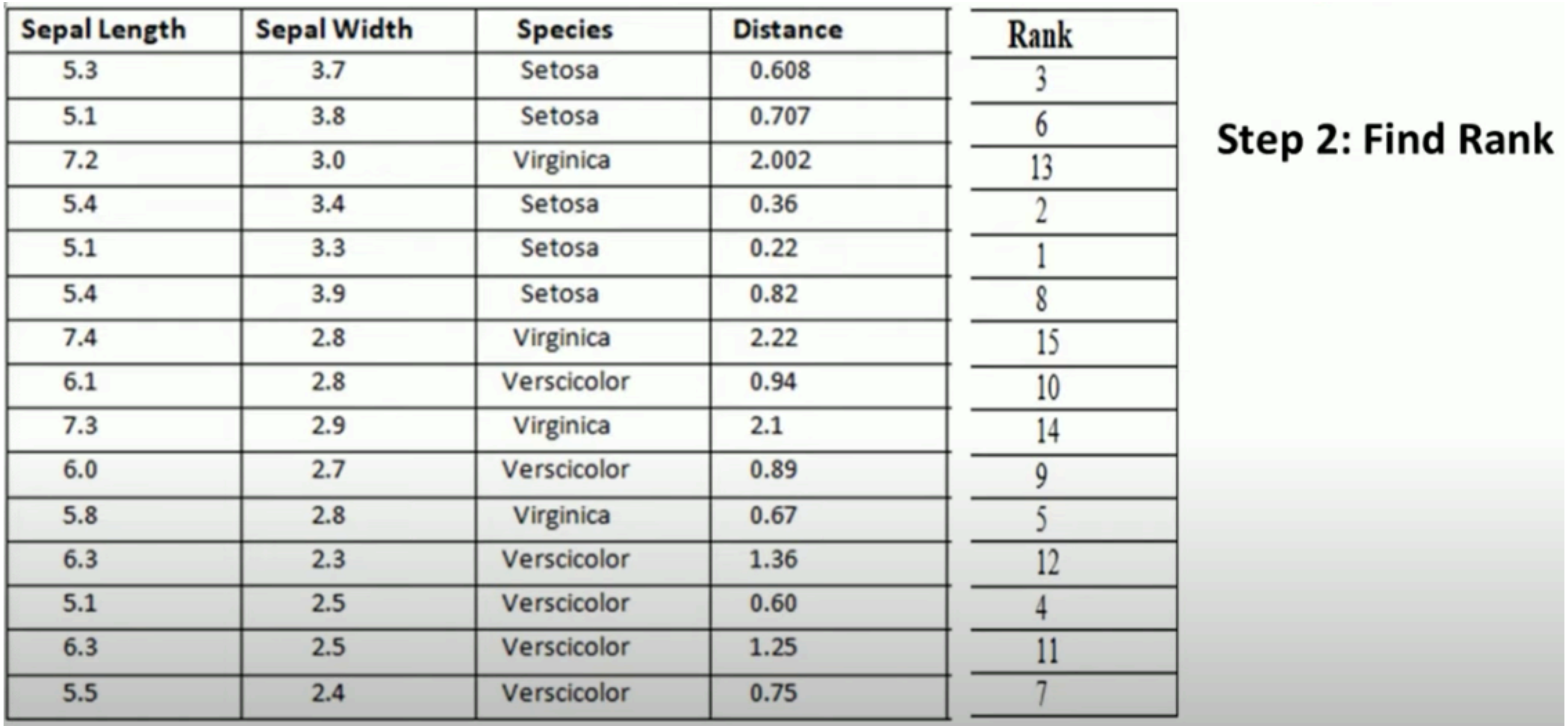
Example 1

- **The total distances from the new data point to each data point in our dataset:**



Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608
5.1	3.8	Setosa	0.707
7.2	3.0	Virginica	2.002
5.4	3.4	Setosa	0.36
5.1	3.3	Setosa	0.22
5.4	3.9	Setosa	0.82
7.4	2.8	Virginica	2.22
6.1	2.8	Verscicolor	0.94
7.3	2.9	Virginica	2.1
6.0	2.7	Verscicolor	0.89
5.8	2.8	Virginica	0.67
6.3	2.3	Verscicolor	1.36
5.1	2.5	Verscicolor	0.60
6.3	2.5	Verscicolor	1.25
5.5	2.4	Verscicolor	0.75

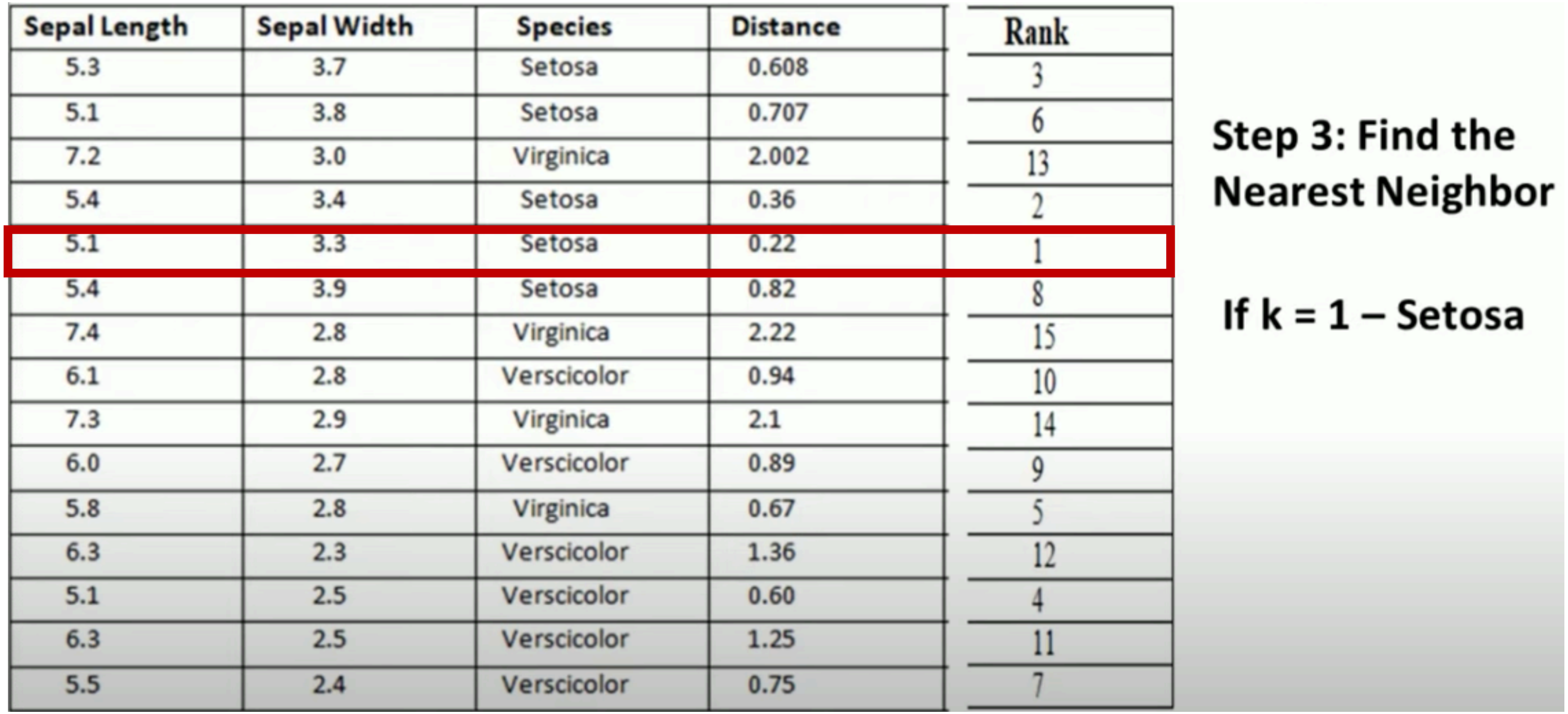
Example 1



Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 2: Find Rank

Example 1



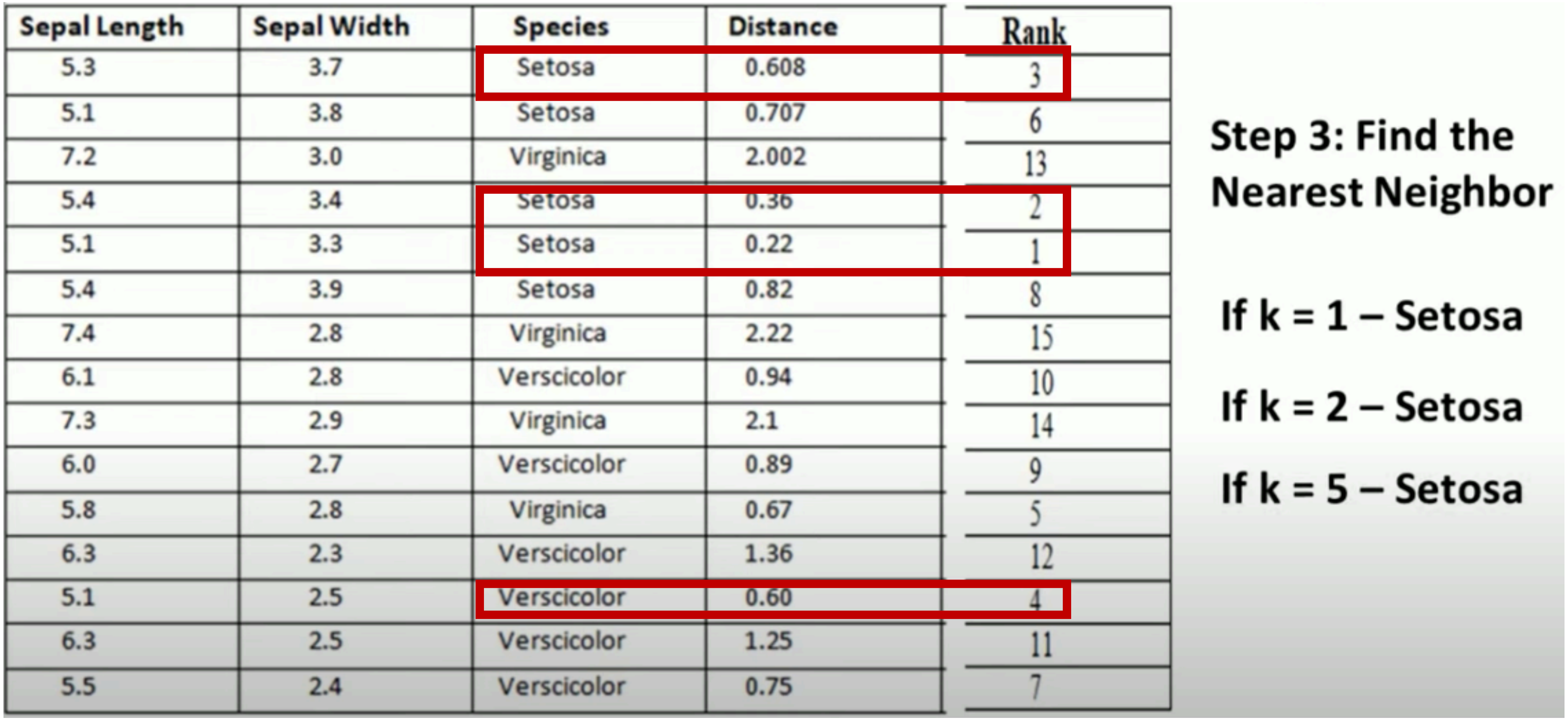
Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 3: Find the Nearest Neighbor

If k = 1 – Setosa

Example 1

If K = 5, then we can identify the new data point as Setosa, since the first three ranks are identified as Setosa, 4th rank is Verscicolor, and 5th rank is Virinica. (3 out 5 ranks are Setosa, so will pick the majority class)



Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	0.67	5
6.3	2.3	Verscicolor	1.36	12
5.1	2.5	Verscicolor	0.60	4
6.3	2.5	Verscicolor	1.25	11
5.5	2.4	Verscicolor	0.75	7

Step 3: Find the Nearest Neighbor

If k = 1 – Setosa

If k = 2 – Setosa

If k = 5 – Setosa

How to identify the appropriate value for K?

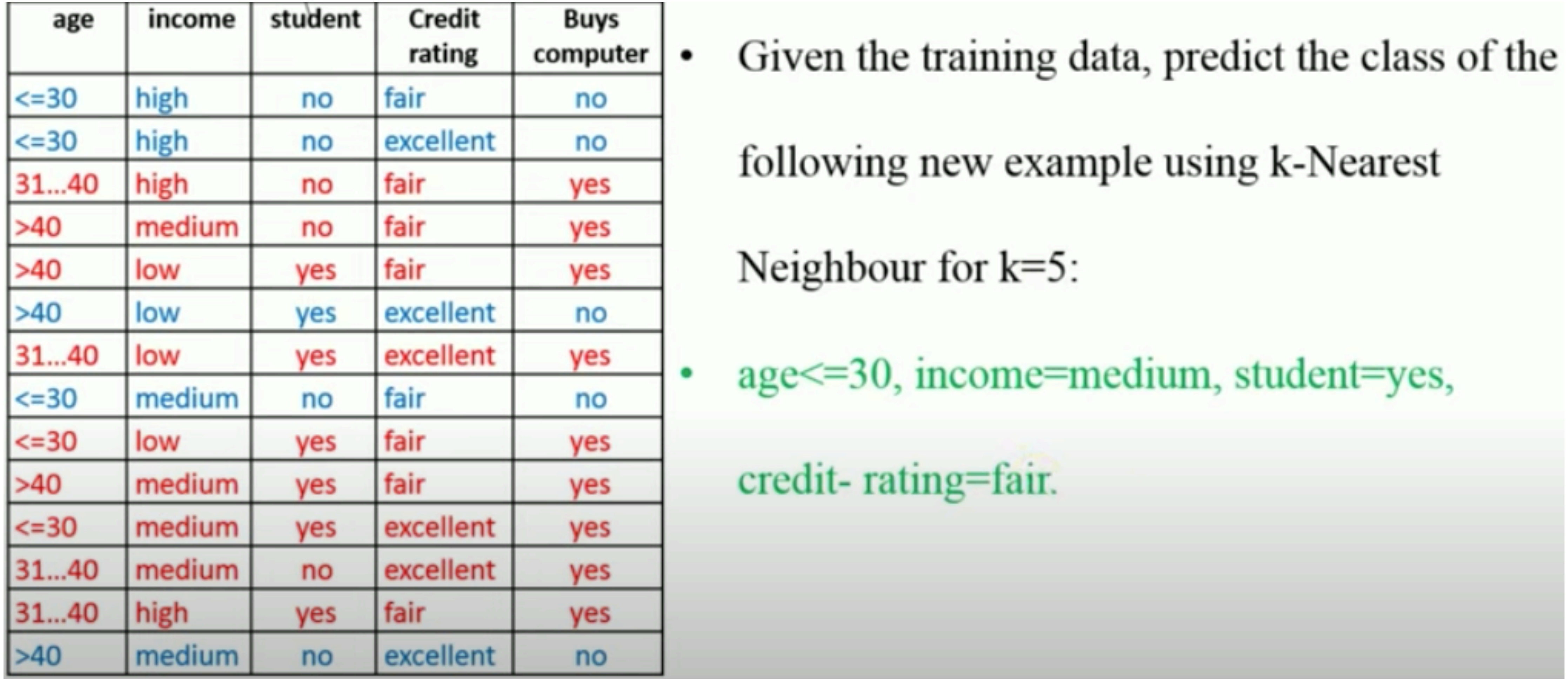
- Choosing the value of K:

- low values for K(like k=1 or k=2), can be noisy and subject to the effects of outliers. .

- large values for K smooth over things (we usually go with k=5), but if K is too large, neighborhood may include points from other classes.

-try different K values and pick the one that gives you higher performance. As a starting point you can take the square root of N (size of the dataset).

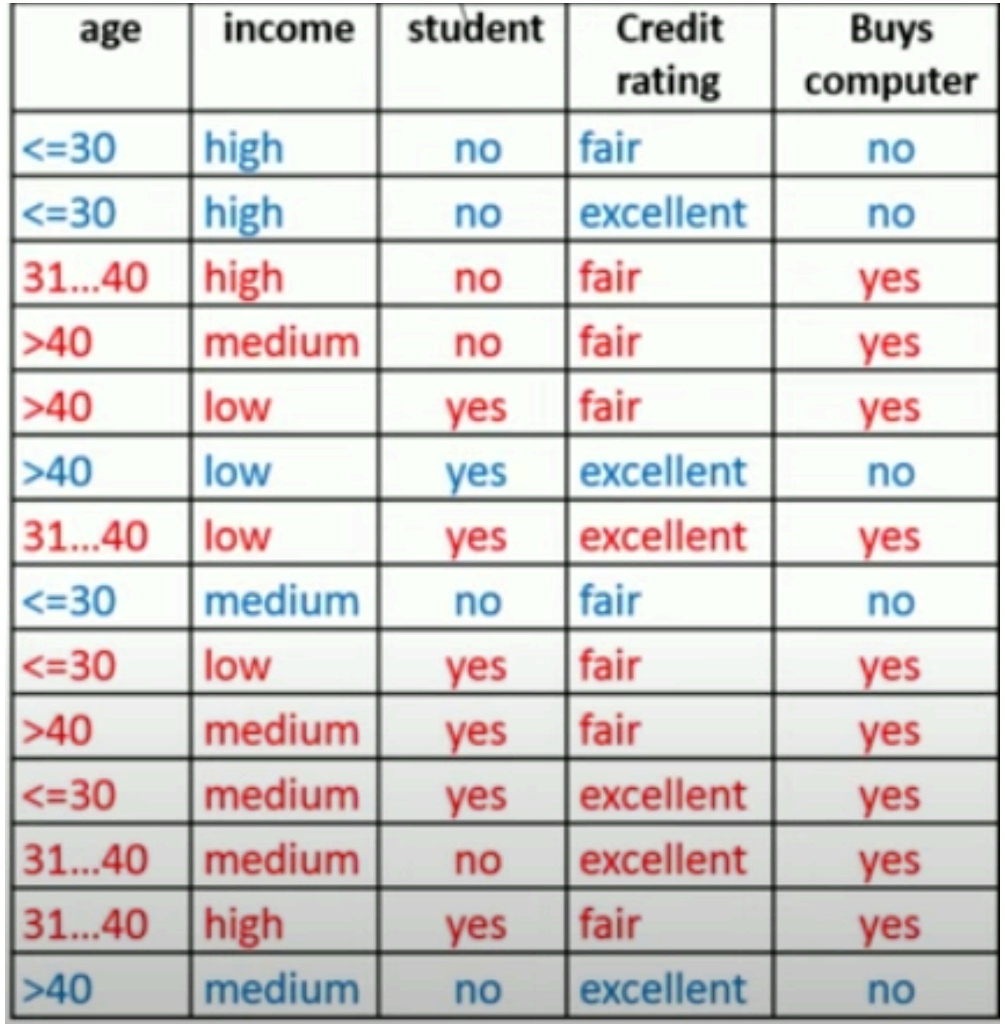
Using Similarity Measure to find the Nearest Neighbors



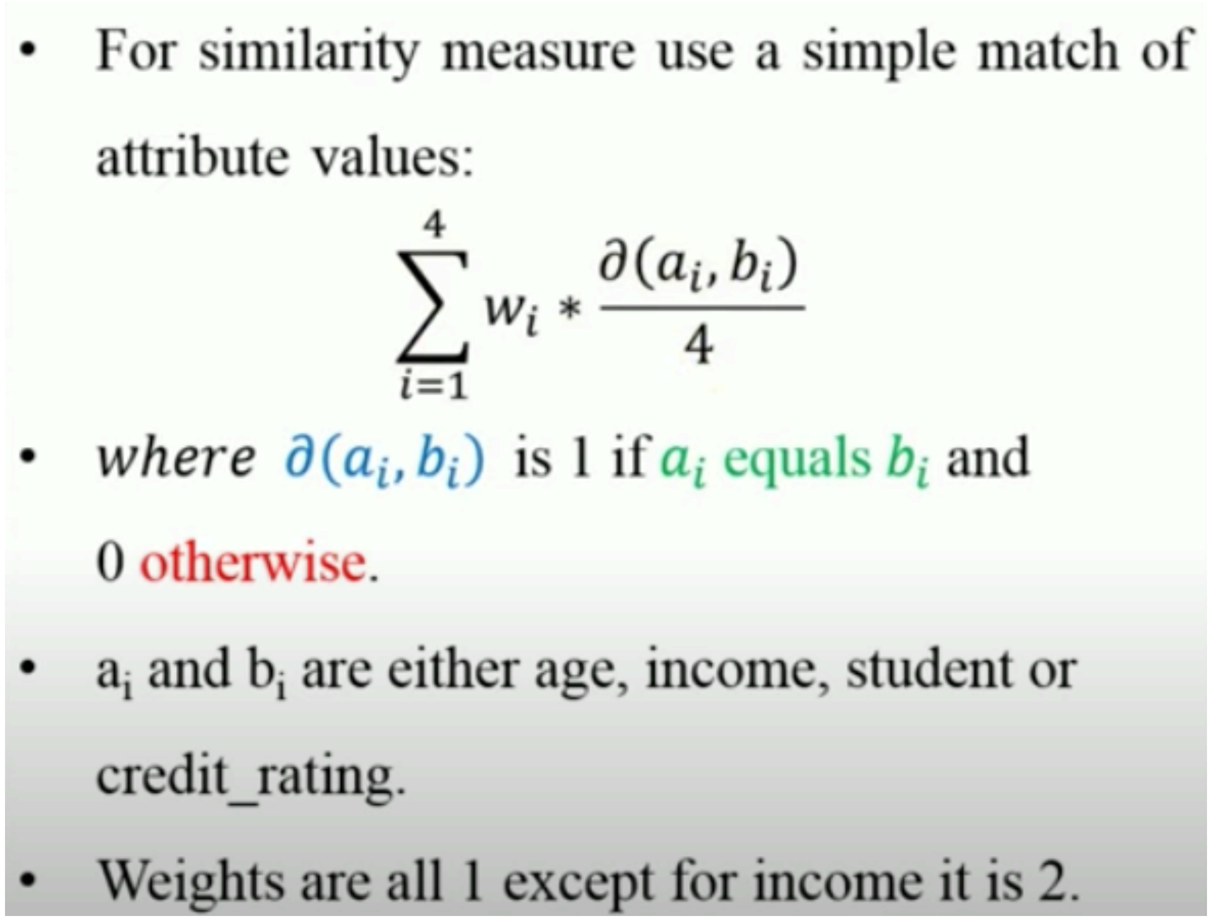
age	income	student	Credit rating	Buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

- Given the training data, predict the class of the following new example using k-Nearest Neighbour for k=5:
- age<=30, income=medium, student=yes, credit- rating=fair.

Using Similarity Measure to find the Nearest Neighbors



age	income	student	Credit rating	Buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

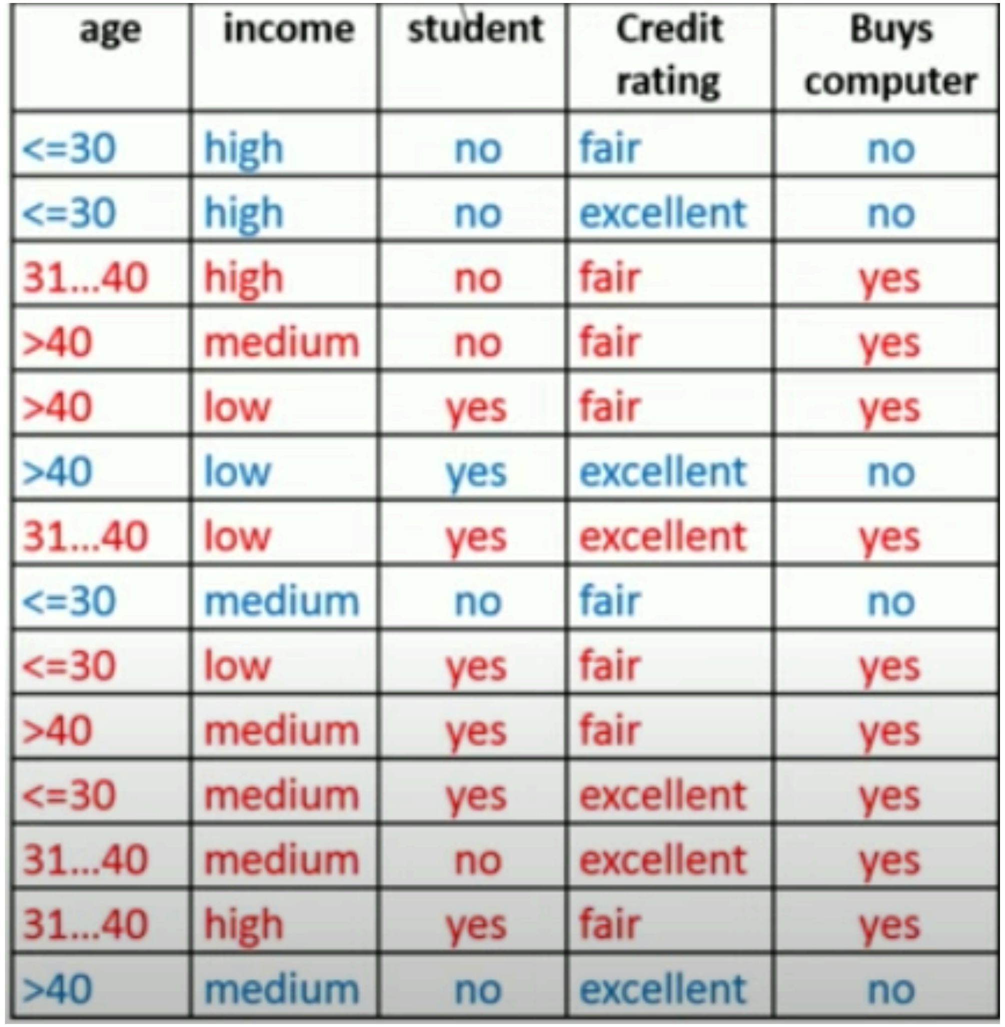


- For similarity measure use a simple match of attribute values:

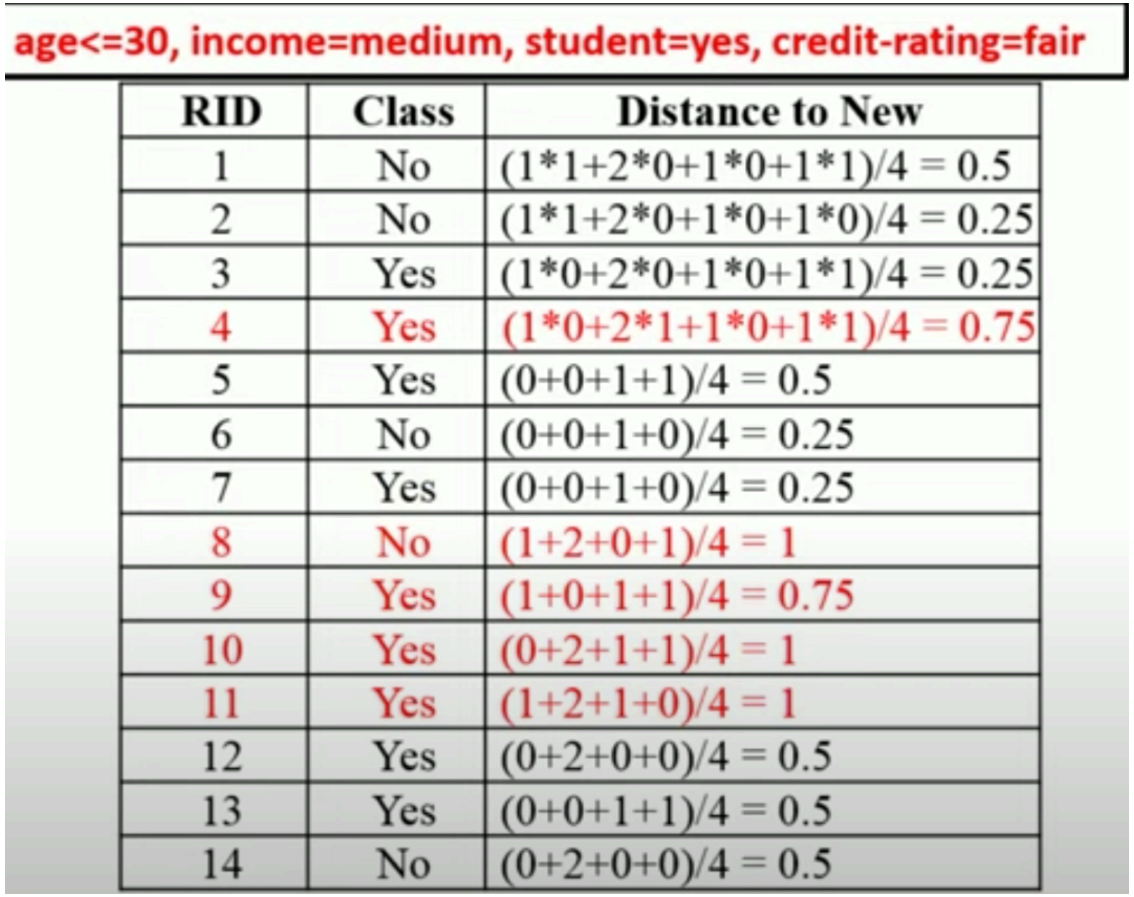
$$\sum_{i=1}^{4} w_i * \frac{\partial(a_i, b_i)}{4}$$

- *where* $\partial(a_i, b_i)$ is 1 if a_i equals b_i and 0 otherwise.
- a_i and b_i are either age, income, student or credit_rating.
- Weights are all 1 except for income it is 2.

Using Similarity Measure to find the Nearest Neighbors



age	income	student	Credit rating	Buys computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



age<=30, income=medium, student=yes, credit-rating=fair

RID	Class	Distance to New
1	No	(1*1+2*0+1*0+1*1)/4 = 0.5
2	No	(1*1+2*0+1*0+1*0)/4 = 0.25
3	Yes	(1*0+2*0+1*0+1*1)/4 = 0.25
4	Yes	(1*0+2*1+1*0+1*1)/4 = 0.75
5	Yes	(0+0+1+1)/4 = 0.5
6	No	(0+0+1+0)/4 = 0.25
7	Yes	(0+0+1+0)/4 = 0.25
8	No	(1+2+0+1)/4 = 1
9	Yes	(1+0+1+1)/4 = 0.75
10	Yes	(0+2+1+1)/4 = 1
11	Yes	(1+2+1+0)/4 = 1
12	Yes	(0+2+0+0)/4 = 0.5
13	Yes	(0+0+1+1)/4 = 0.5
14	No	(0+2+0+0)/4 = 0.5

Among the five nearest neighbors, we have 4 yes and 1 no, then the algorithm will predict a yes (this student Will buy a computer).