KNN Algorithm

Abdul Haseeb

BS(AI)-III

Outline

KNN Algorithm Philosophy

How to choose Value of K?

Different types of Distances Used in KNN

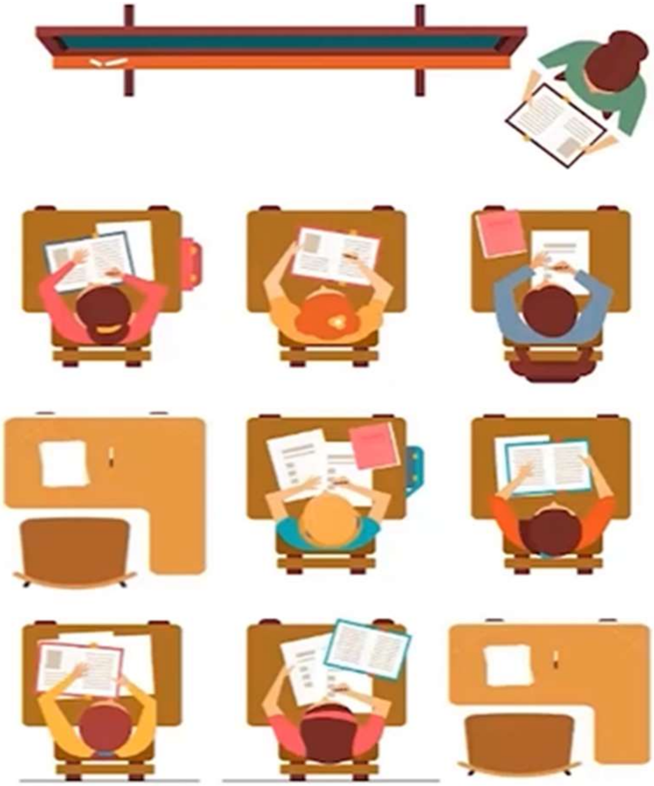
Solving a Problem using KNN Algorithm with Euclidean Distance

KNN Introduction

- K-Nearest Neighbors
- Lazy Learning Algorithm

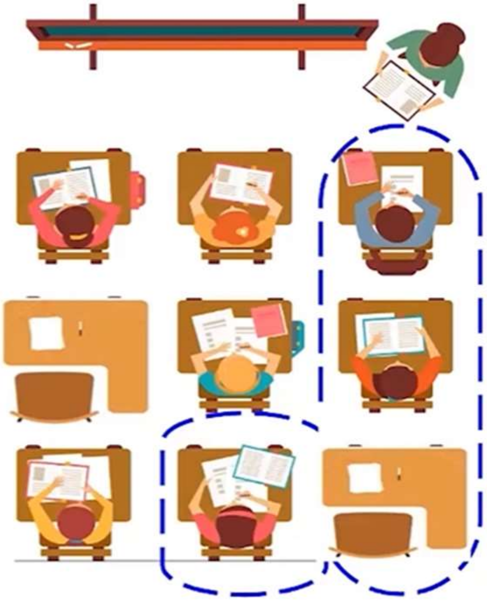
KNN Introduction

- Classroom with a strength of 9 students
- But at present 7 students enrolled



KNN Introduction

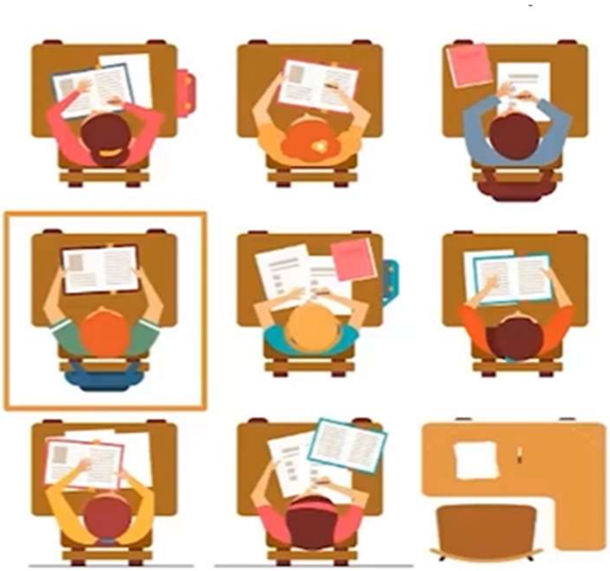
Interested in Studying Books, Novels etc



Interested in Sports etc

KNN Introduction

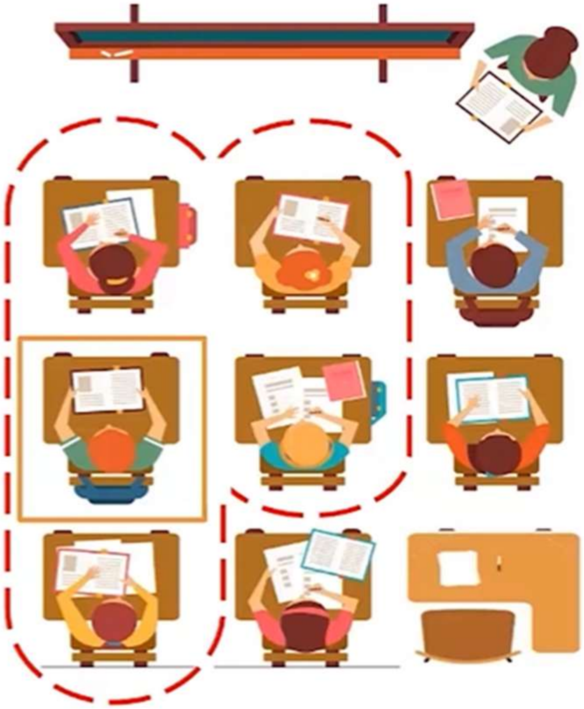
New Student, chooses to sit on the right most available desk



KNN Introduction

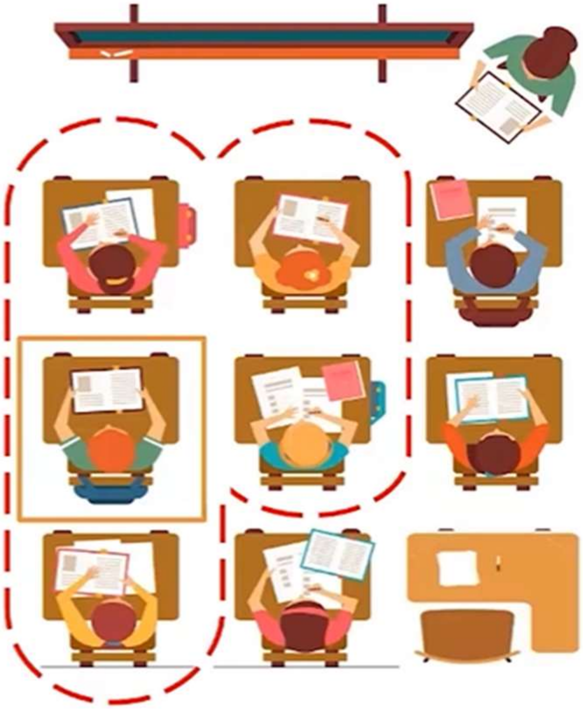
Principal wants to know interest of every student, but hasn’t enough time to interact with each one

From the group he sits, means from his neighbors, principal can infer that the student is one who loves study



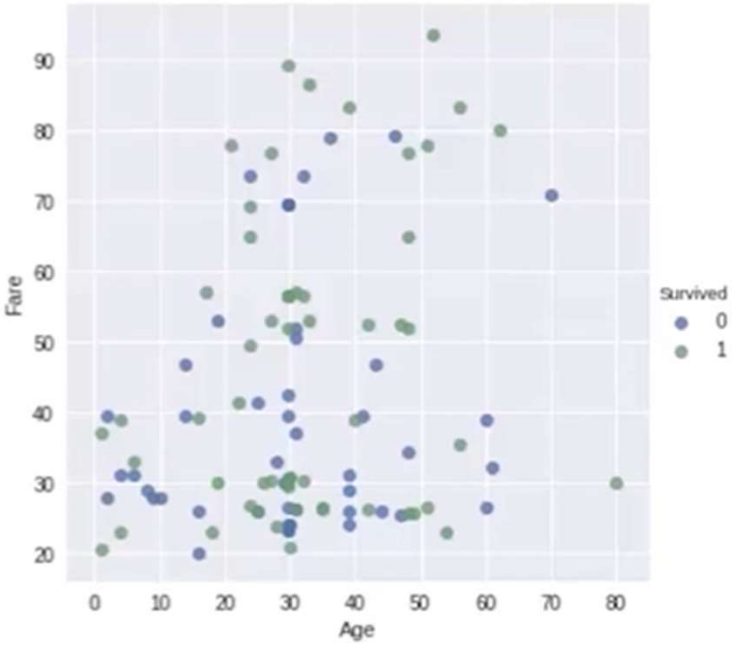
KNN Introduction

This is known as the lazy learning approach; we just inferred the output by observing the behaviors of nearest neighbors



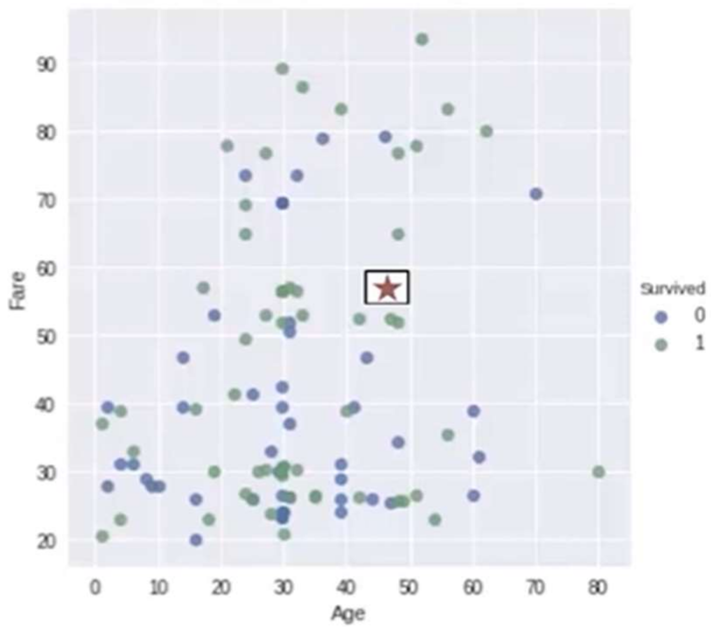
Building a KNN Model

- Plot the training dataset



Building a KNN Model

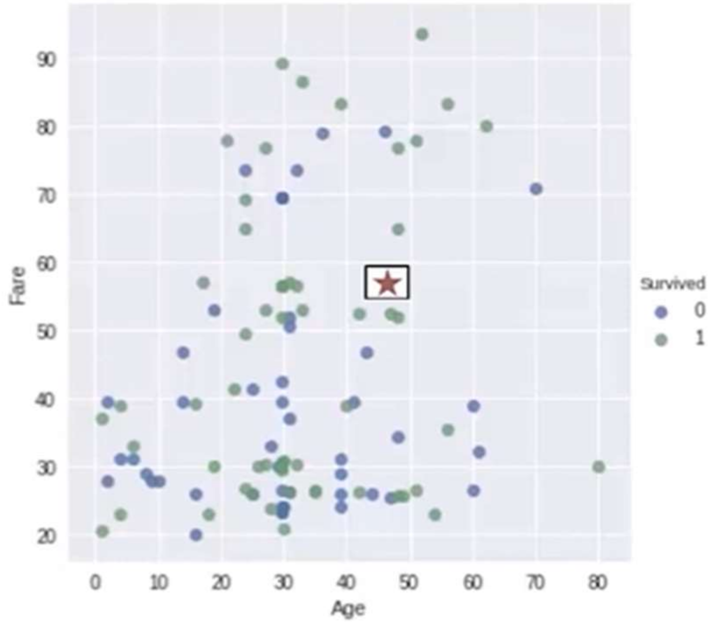
- Locate new “test” point/instance



Building a KNN Model

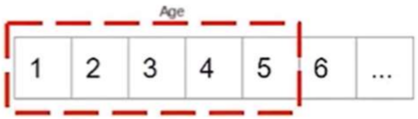
- Calculate distance from all training data points

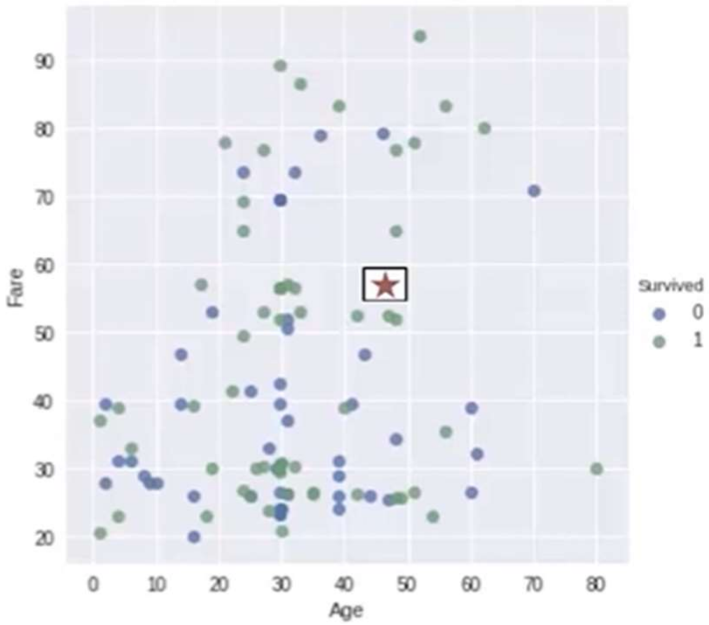
6	2	5	1	3	4	...



Building a KNN Model

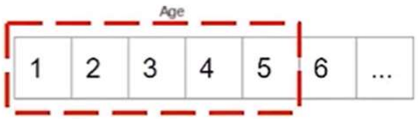
- Sort the list in ascending order





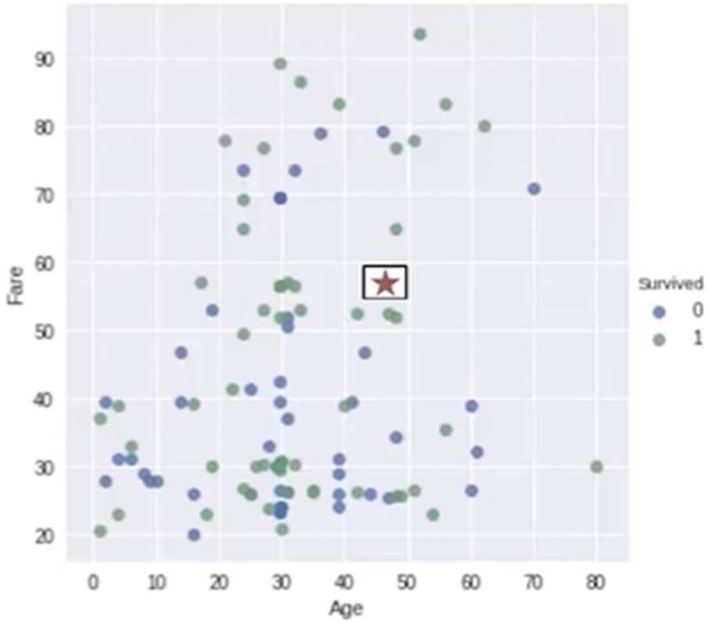
Building a KNN Model

- Choose first k distances from the sorted list



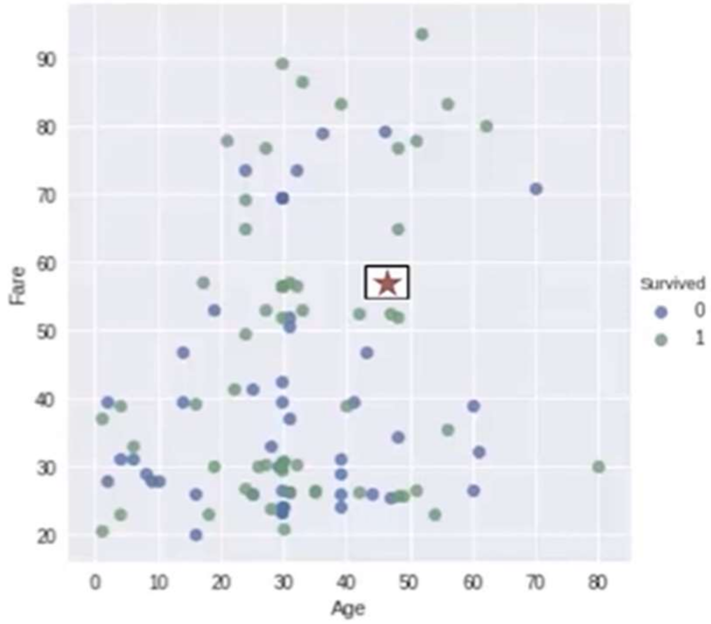
Age

1	2	3	4	5	6	...

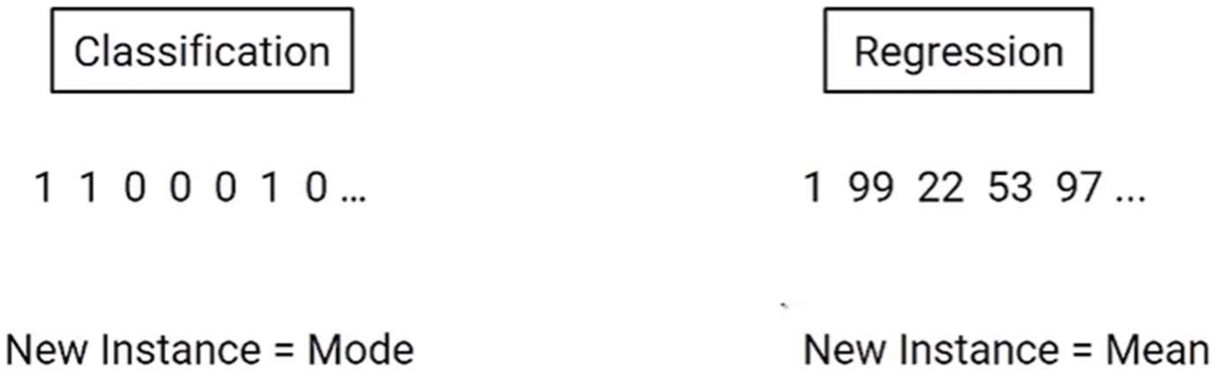


Building a KNN Model

- Now if you use k=5 what can we infer from the graph?
- Should we assign green(1=survive) or blue (0=Not Survived)
- We take mode(frequency of occurrence) of the k data points to calculate the answer
- Green Data Points are in abundance, so we will assign green to new data point



Building a KNN Model



Classification

1 1 0 0 0 1 0 ...

New Instance = Mode

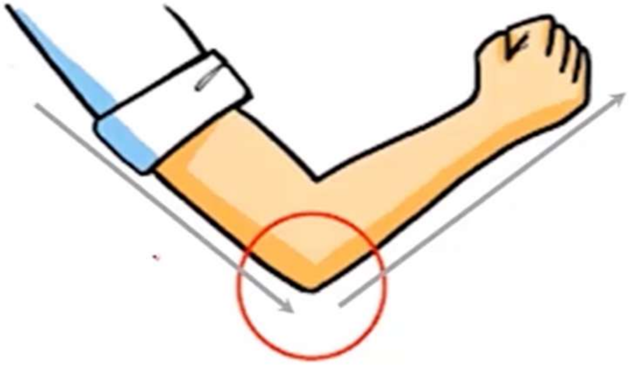
Regression

1 99 22 53 97 ...

New Instance = Mean

Determine the Right value of K

- Elbow Method:



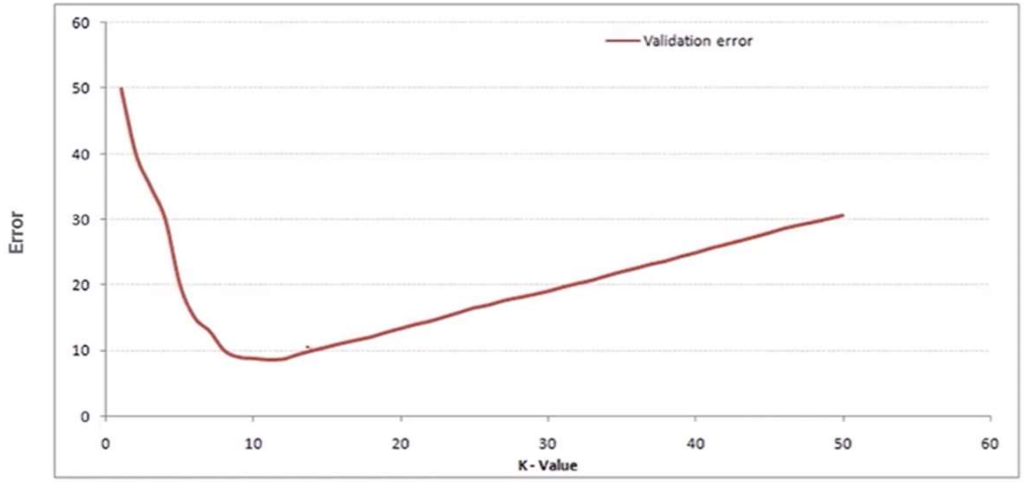
Determine the Right value of K

- Elbow Method:
    - Choose a range of k values:
        - Min 1 and Max number of data points in dataset
    - For each value of k we implement a KNN Model
    - Calculate error corresponding to each value of k and plot the error

Determine the Right value of K

- Elbow Method:

The Graph has a shape similar to the elbow, so we call it elbow method



We choose k=10, as it returns the minimum error

How to calculate Distance

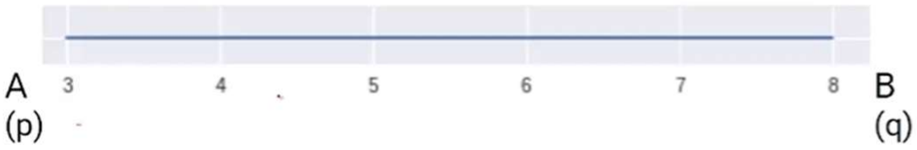
- Now after calculating the value of k, how to determine which points are the nearest

How to calculate Distance

- Manhattan Distance
- Euclidean Distance
- Minkowiski Distance
- Hamming Distance

Manhattan Distance

- Sum of absolute differences between the two points, across all the dimensions



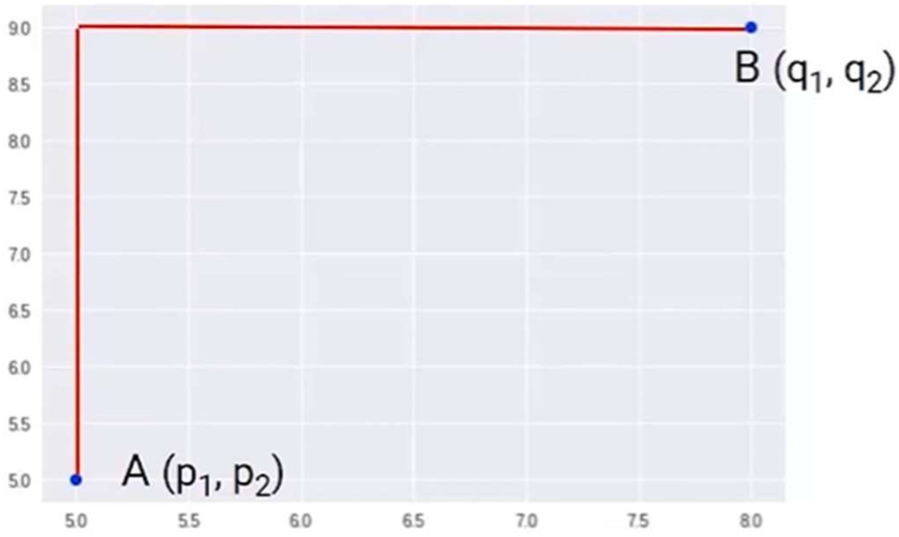
$d = |p - q|$

d=3-8
d=5

Manhattan Distance

Two Dimensions: $d = |p_1 - q_1| + |p_2 - q_2|$

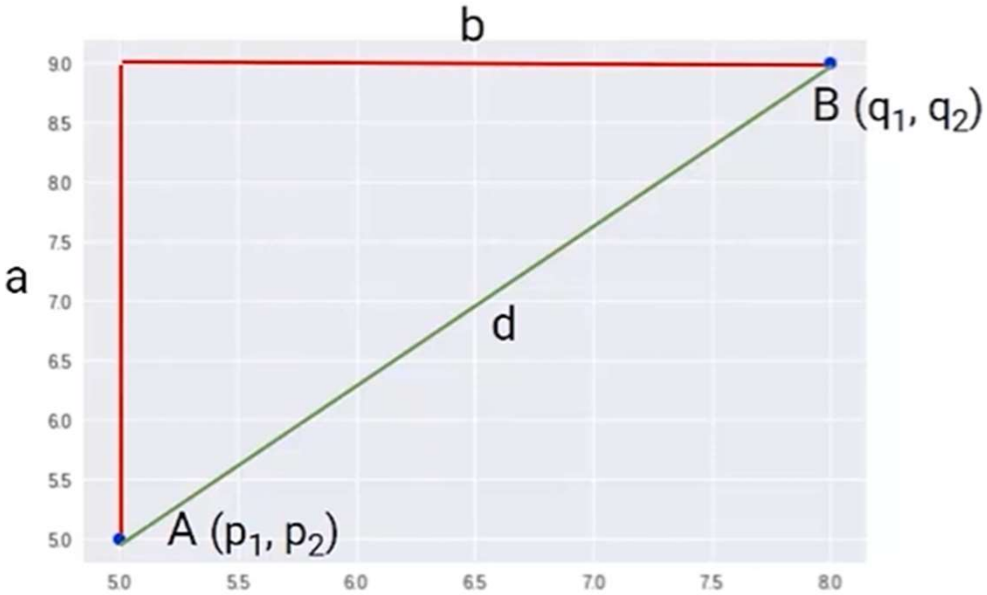
Let’s say p1=5 and p2=5
Let’s say q1=8 and q2=9
d=|5-8|+|5-9|
d=3+4
d=7



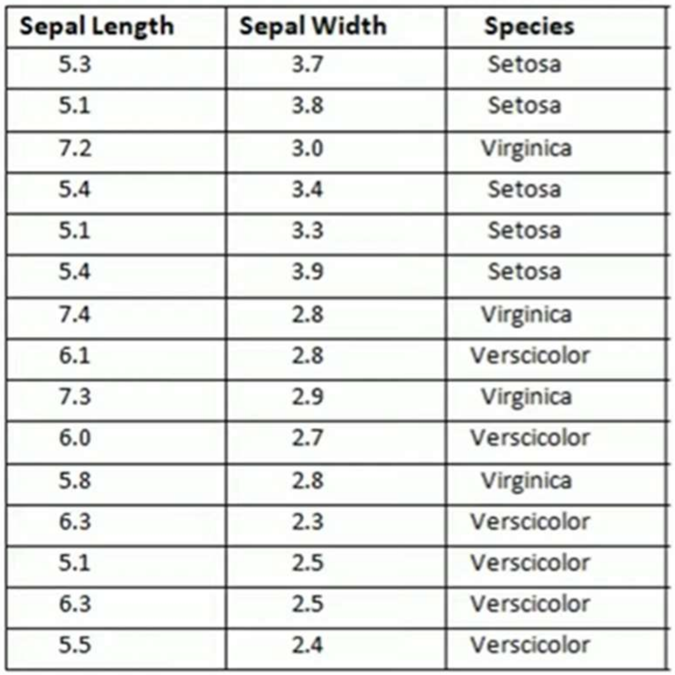
Euclidean Distance

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

Let’s say p1=5 and p2=5
Let’s say q1=8 and q2=9
$d=((5-8^{)2+(}5-9^{2}))^{1/2}$
$d=(9+16)^{1/2}$
d=5

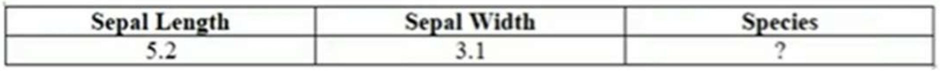


Problem



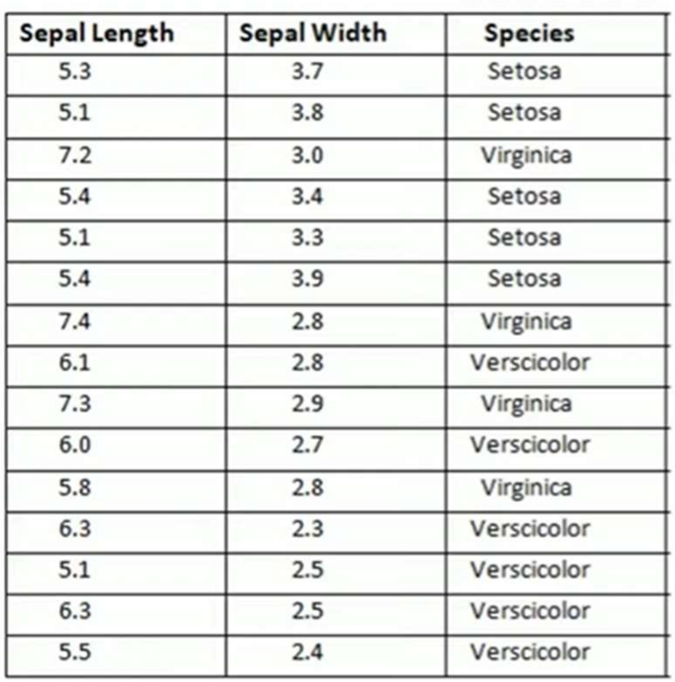
Sepal Length	Sepal Width	Species
5.3	3.7	Setosa
5.1	3.8	Setosa
7.2	3.0	Virginica
5.4	3.4	Setosa
5.1	3.3	Setosa
5.4	3.9	Setosa
7.4	2.8	Virginica
6.1	2.8	Verscicolor
7.3	2.9	Virginica
6.0	2.7	Verscicolor
5.8	2.8	Virginica
6.3	2.3	Verscicolor
5.1	2.5	Verscicolor
6.3	2.5	Verscicolor
5.5	2.4	Verscicolor

Given the value of k, classify the new example



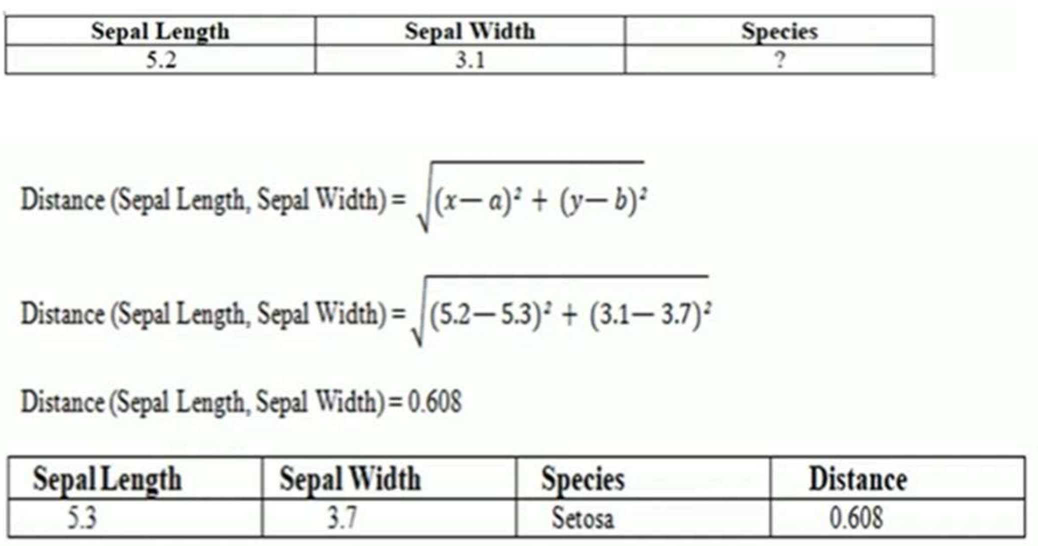
Sepal Length	Sepal Width	Species
5.2	3.1	?

Problem



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 01: Calculate the distance of this new data point with all the data points.



Sepal Length	Sepal Width	Species
5.2	3.1	?

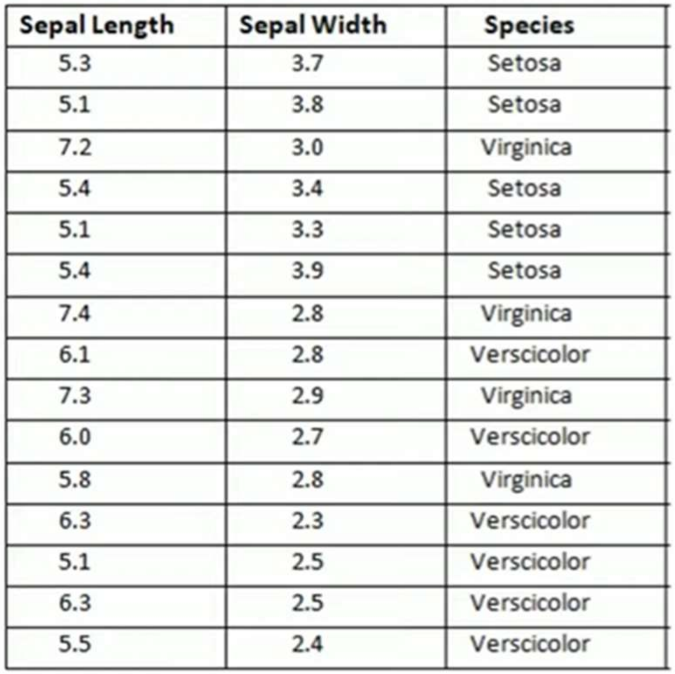
$$\text{Distance (Sepal Length, Sepal Width)} = \sqrt{(x-a)^2 + (y-b)^2}$$

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

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

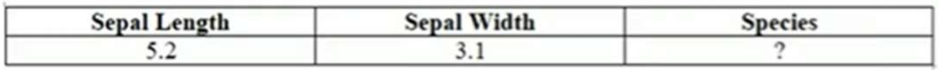
Sepal Length	Sepal Width	Species	Distance
5.3	3.7	Setosa	0.608

Problem



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

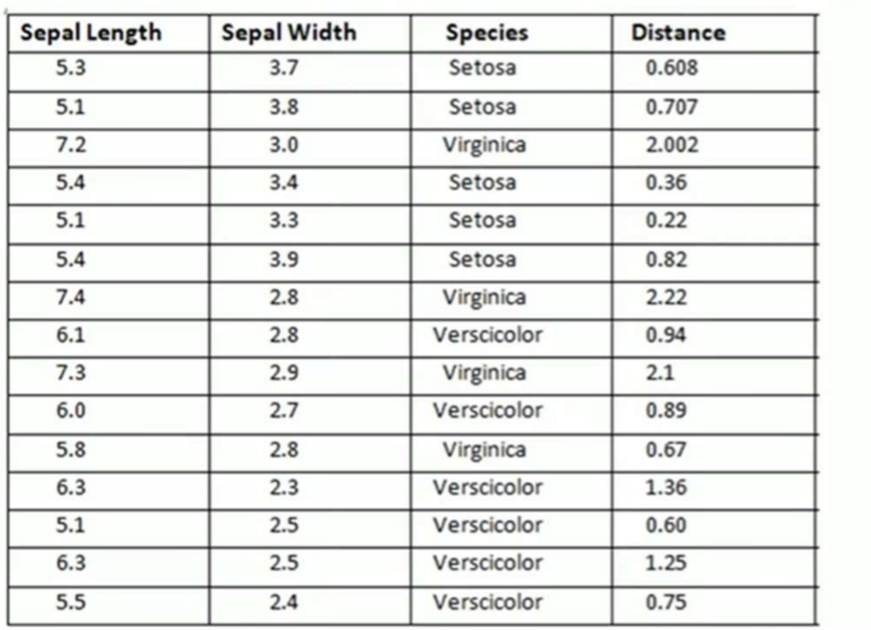
Calculate the distance with 5th and 7th example



Sepal Length	Sepal Width	Species
5.2	3.1	?

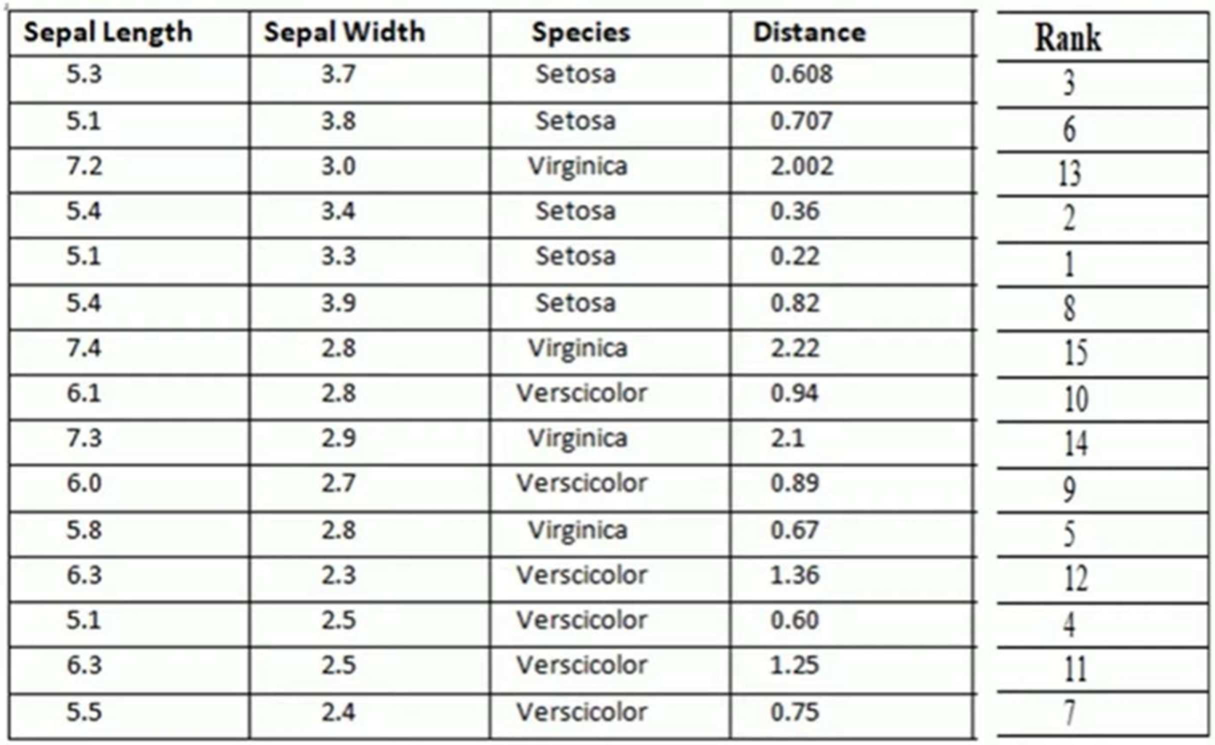
Problem

- All Distances:



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

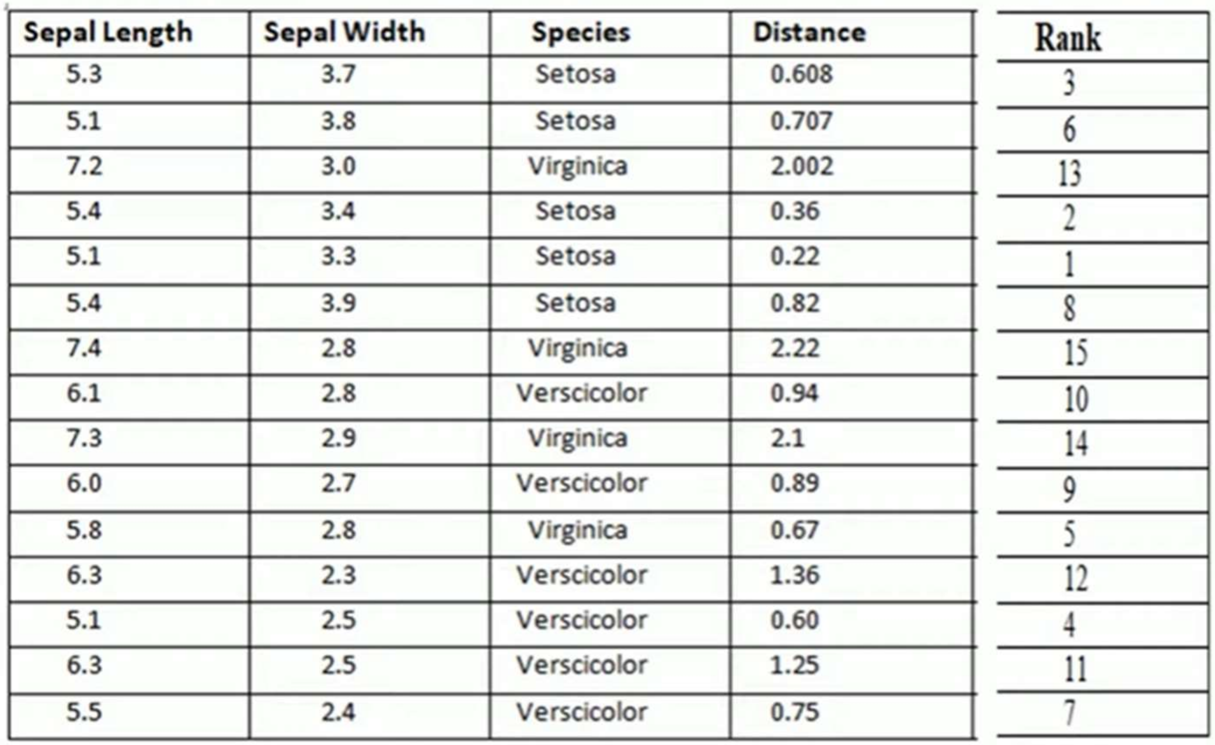
Problem



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

Assign Rank, smallest will be having the first rank and so on…

Problem



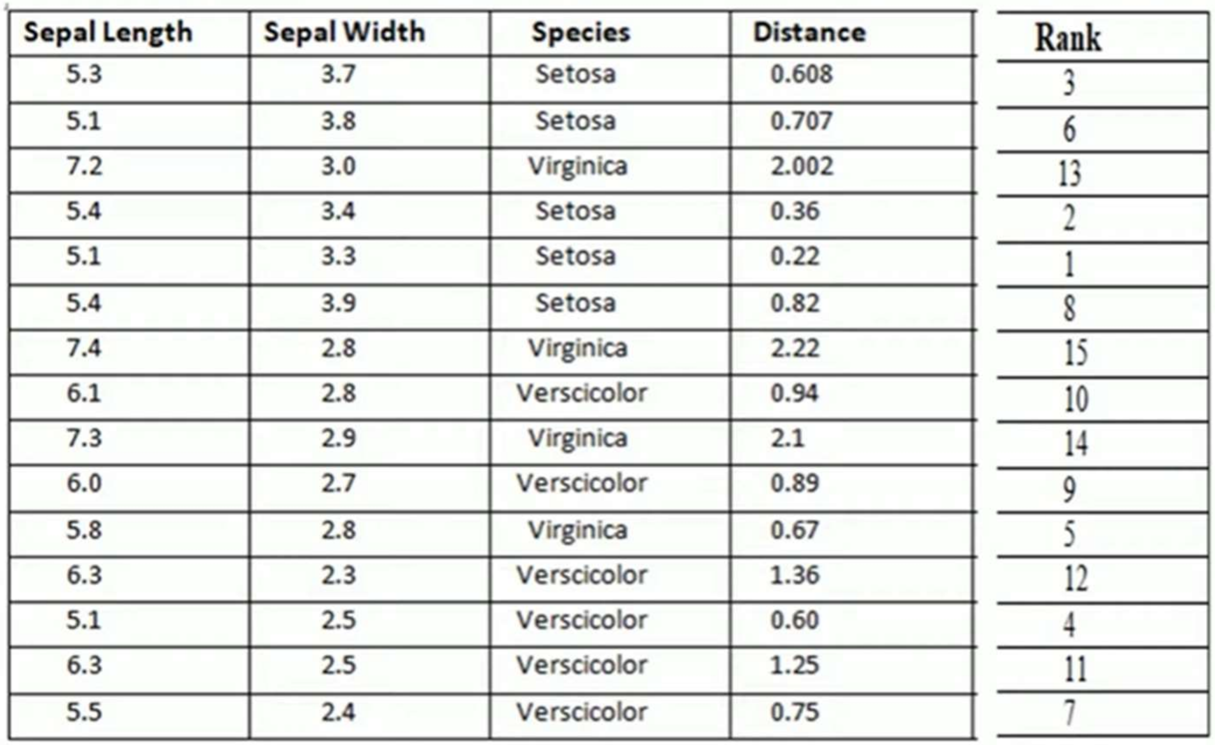
Sepal Length	Sepal Width	Species	Distance	Rank
5.3	3.7	Setosa	0.608	3
5.1	3.8	Setosa	0.707	6
7.2	3.0	Virginica	2.002	13
5.4	3.4	Setosa	0.36	2
5.1	3.3	Setosa	0.22	1
5.4	3.9	Setosa	0.82	8
7.4	2.8	Virginica	2.22	15
6.1	2.8	Verscicolor	0.94	10
7.3	2.9	Virginica	2.1	14
6.0	2.7	Verscicolor	0.89	9
5.8	2.8	Virginica	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 03: Given the value of K, Find the nearest neighbor

If k=1, pick the element with Rank#01

Element with Rank#01 has specie: setosa, so assign setosa to new example

Problem



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 03: Given the value of K, Find the nearest neighbor

What should be the label with K=5?

With K=15?