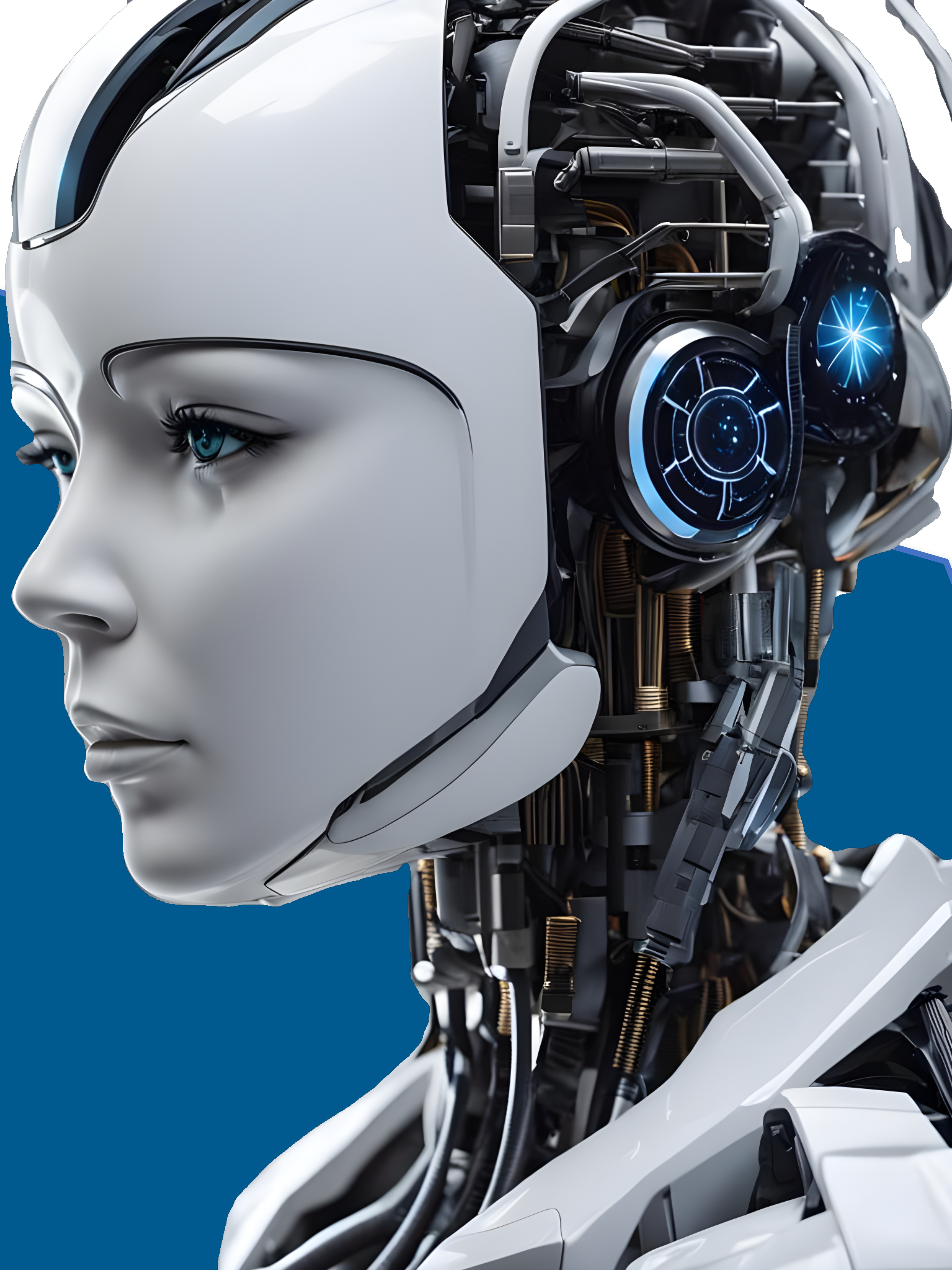
Tishk International University
IT Department
Course Code: IT-344/A

Introduction to Machine Learning

Classifications (k-NN)

Lecture 4



Spring 2024
Hemin Ibrahim, PhD
hemin.ibrahim@tiu.edu.iq

Outline

- Learning
- Supervised Learning
- Classification
- *K*-Nearest Neighbors classification
- Distance Metrics

Objectives

- To grasp the concept of supervised learning, where an algorithm learns from labeled data and makes predictions or decisions based on that learning.
- To understand classification in machine learning: categorizing data into predefined classes based on features.
- To understand the K-Nearest Neighbors (KNN) algorithm and its principle of classification based on the majority vote of its k-nearest neighbors in the feature space
- To understand and apply distance metrics like Euclidean, Manhattan, and Minkowski distances to measure the similarity or dissimilarity between data points

Learning

- AI to solve some problems
- Give no explicit instruction to the computer
- Give data to computer to learn what to do.

Different form of learning

- Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning

Supervised Learning: Supervised learning involves training using "***labelled***" training dataset, and enabling machines to predict outputs based on the provided training data.

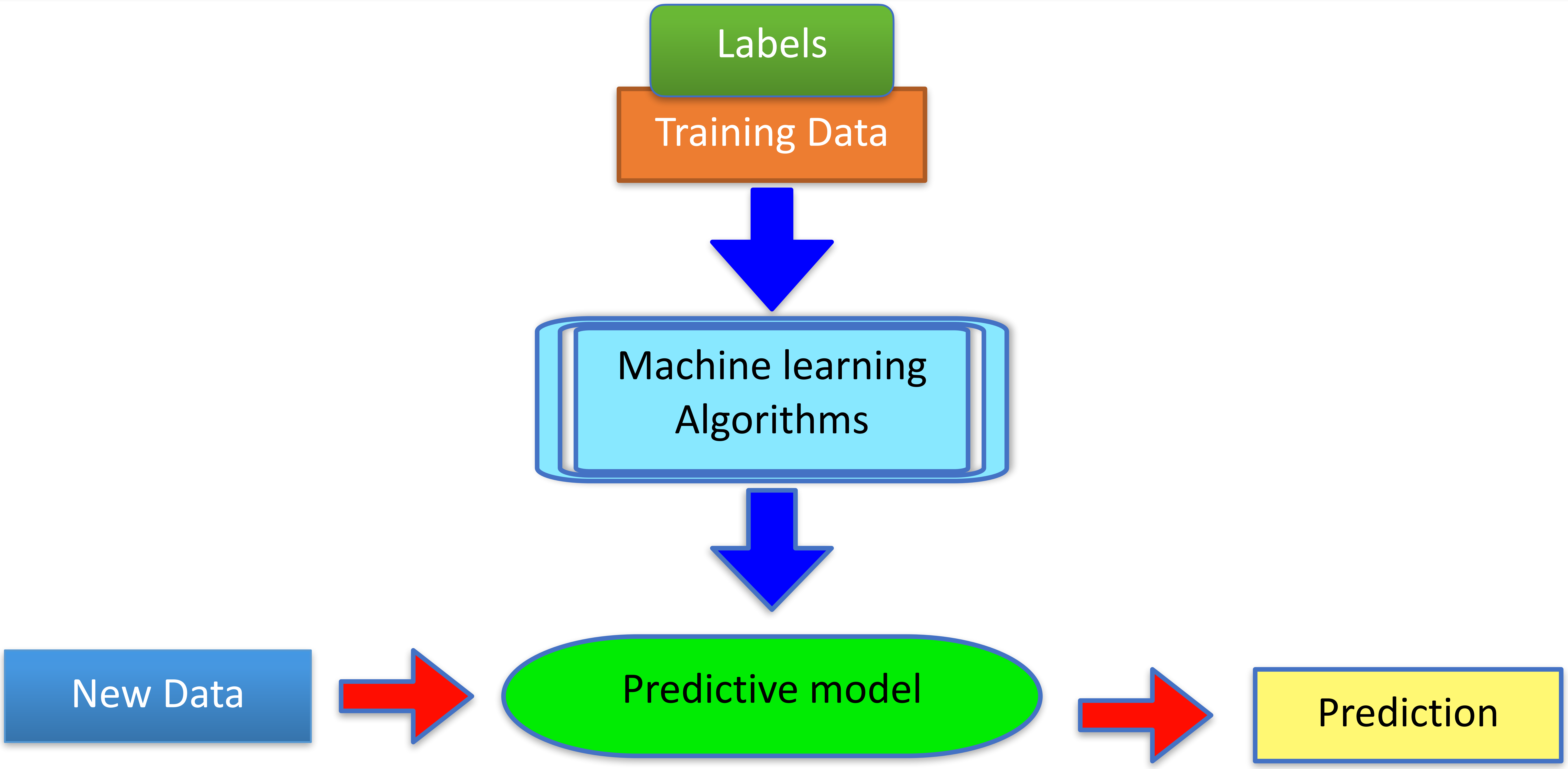
Supervised Learning

Classifications: is a supervised learning approach where the goal is to predict discrete output values (represent **categories** or **classes**).

Examples:

- Email Spam Detection
- Handwritten Digit Recognition
- Image Classification
- Raining or Not

Classification



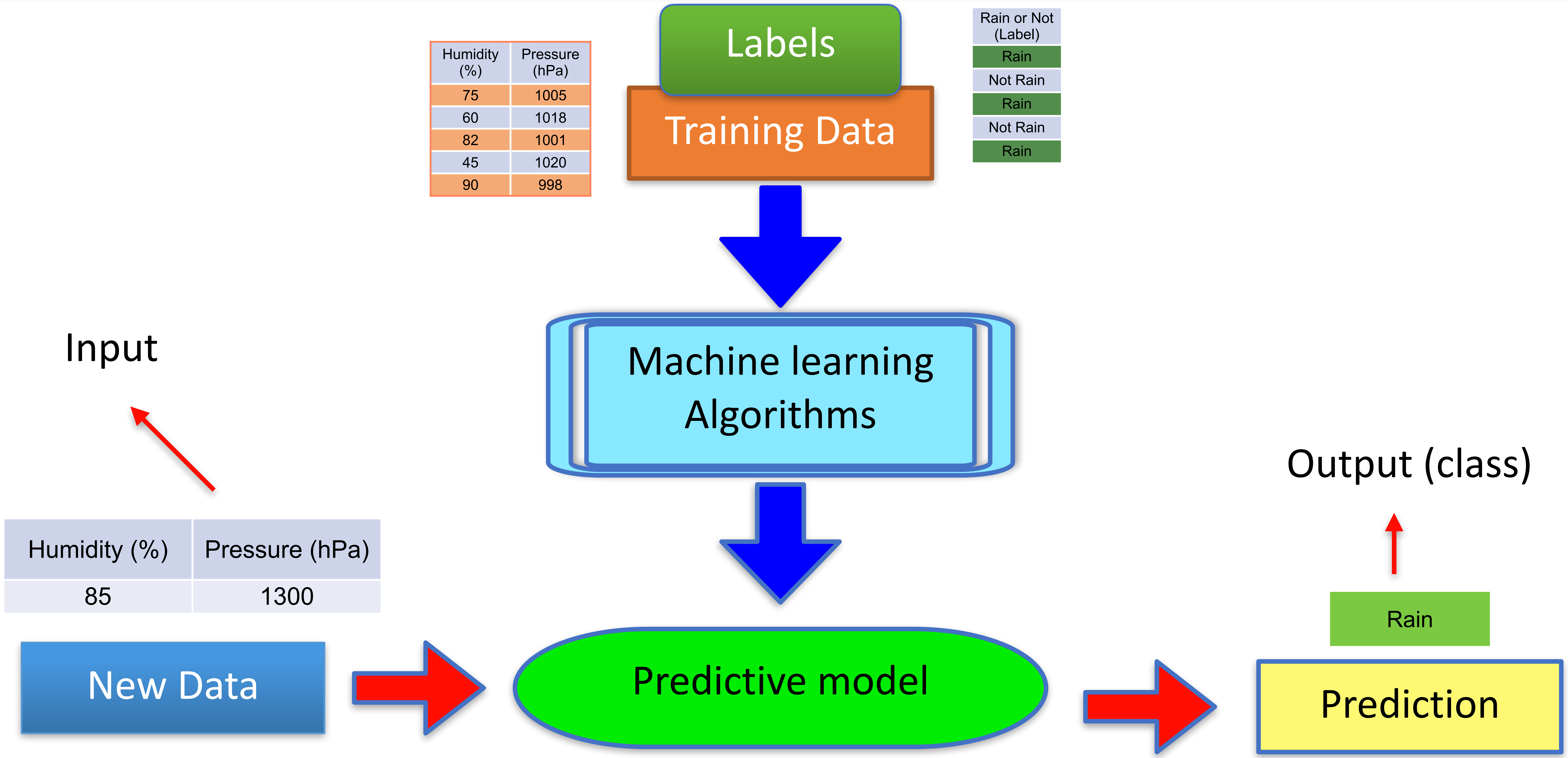
Classification

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain

Input

Output (class)

Classification



Humidity (%)	Pressure (hPa)
75	1005
60	1018
82	1001
45	1020
90	998

Rain or Not (Label)
Rain
Not Rain
Rain
Not Rain
Rain

Humidity (%)	Pressure (hPa)
85	1300

Classification

f(humidity, pressure) = Rain or No Rain

$f(78, 1004) = \text{No Rain}$

$f(99, 1400) = \text{Rain}$

$f(87, 1100) = \text{Rain}$

$f(65, 975) = \text{No Rain}$

Classification

features

labels

data points

Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

Classification

features | labels | data points

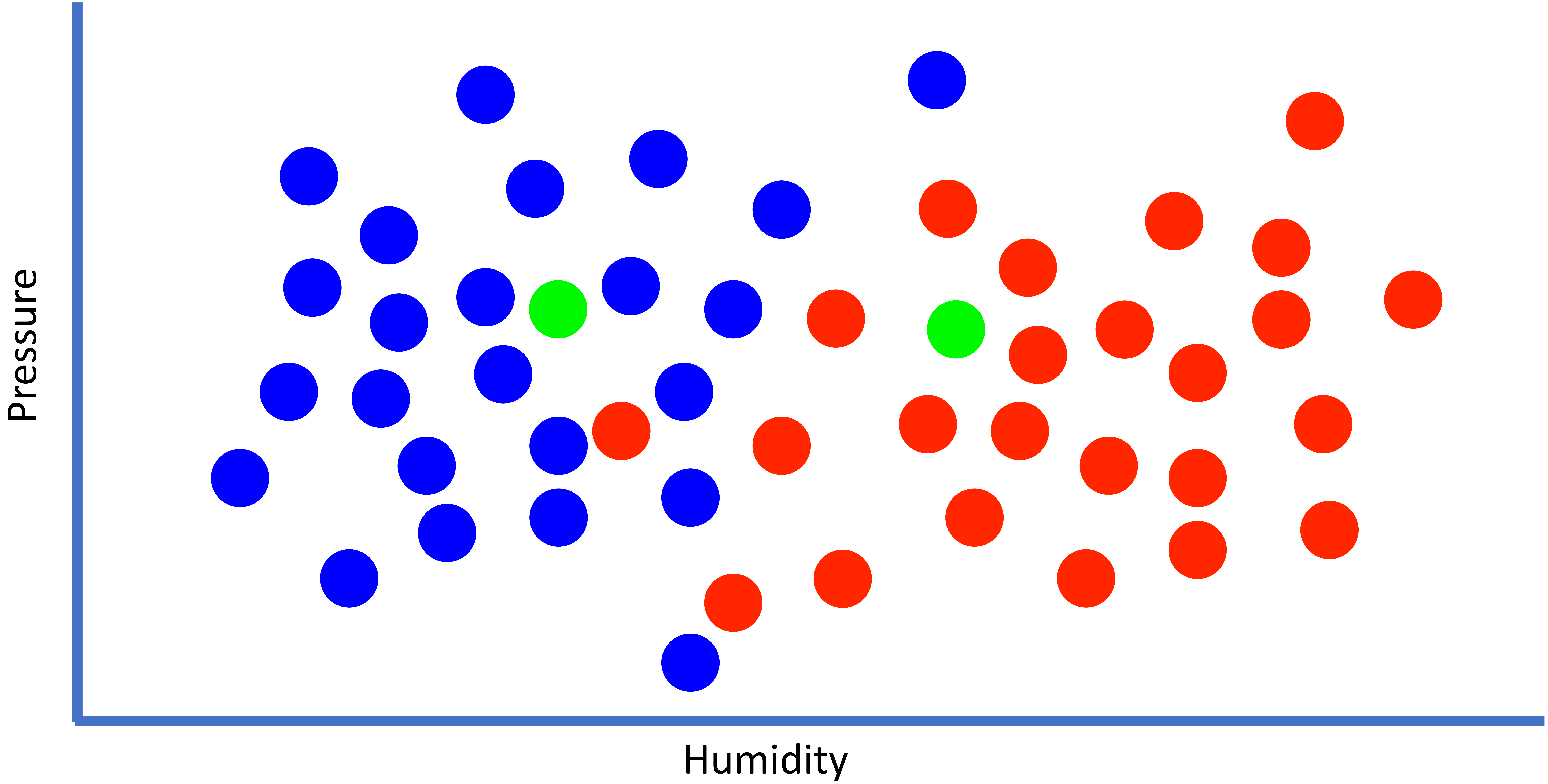
Family History	Resting Blood Pressure	Cholesterol	Heart Disease?
Yes	Low	Normal	No
No	Medium	Normal	No
No	Low	Abnormal	Yes
Yes	Medium	Normal	Yes
Yes	High	Abnormal	Yes

$f(\quad ? \quad) = ?$

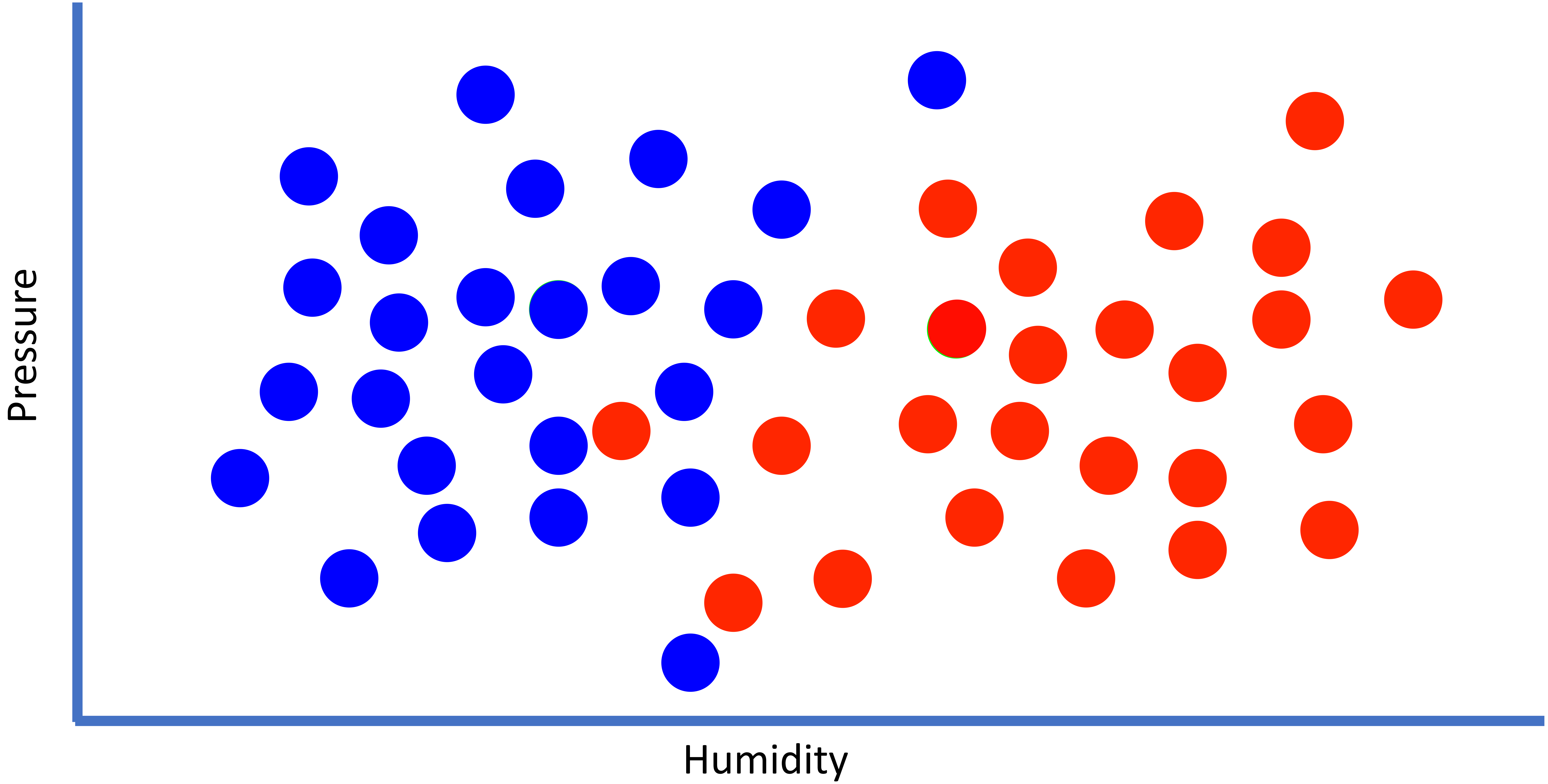
Classification

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain

Classification



Classification



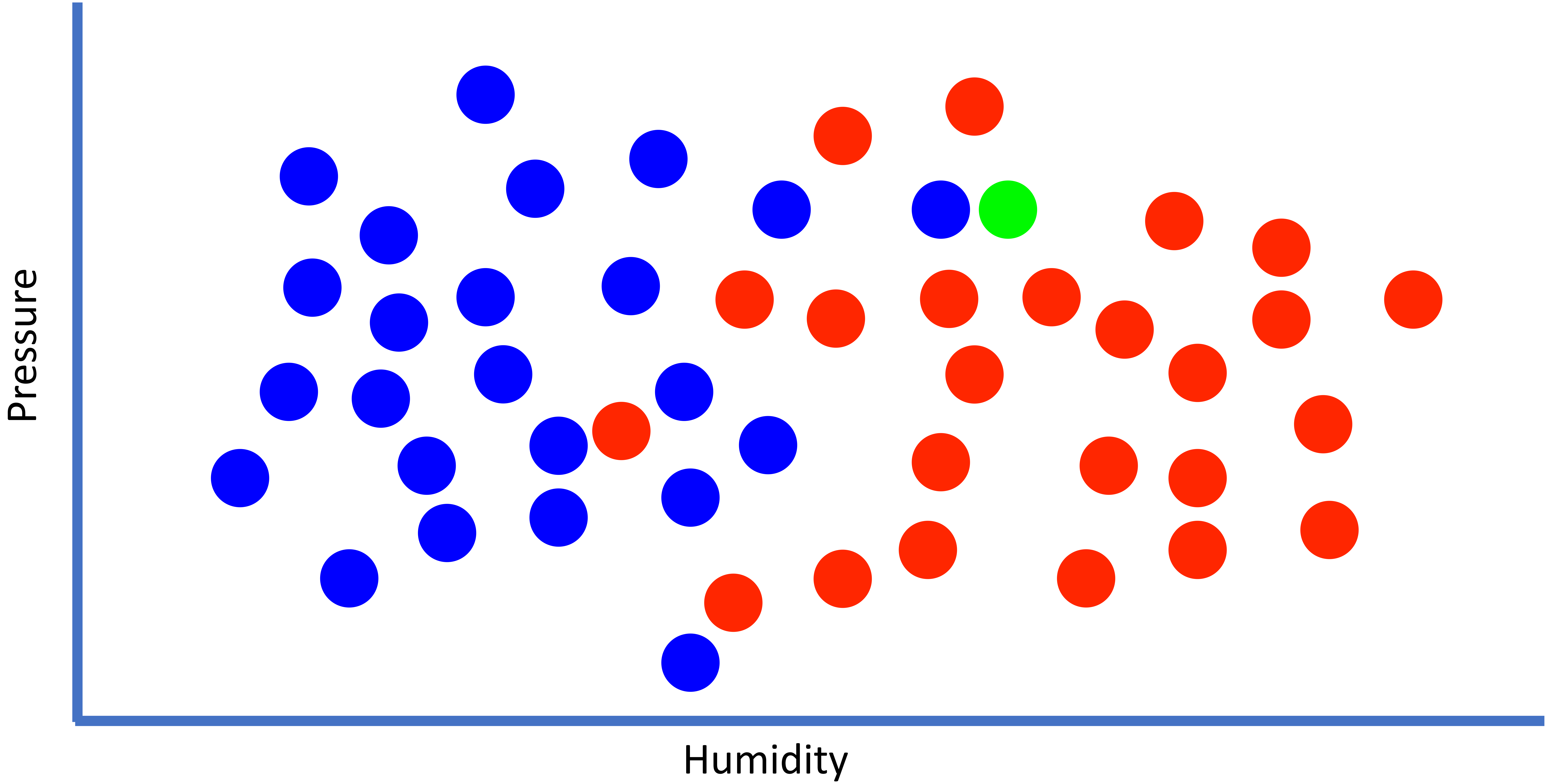
Nearest Neighbor Classification

Classification

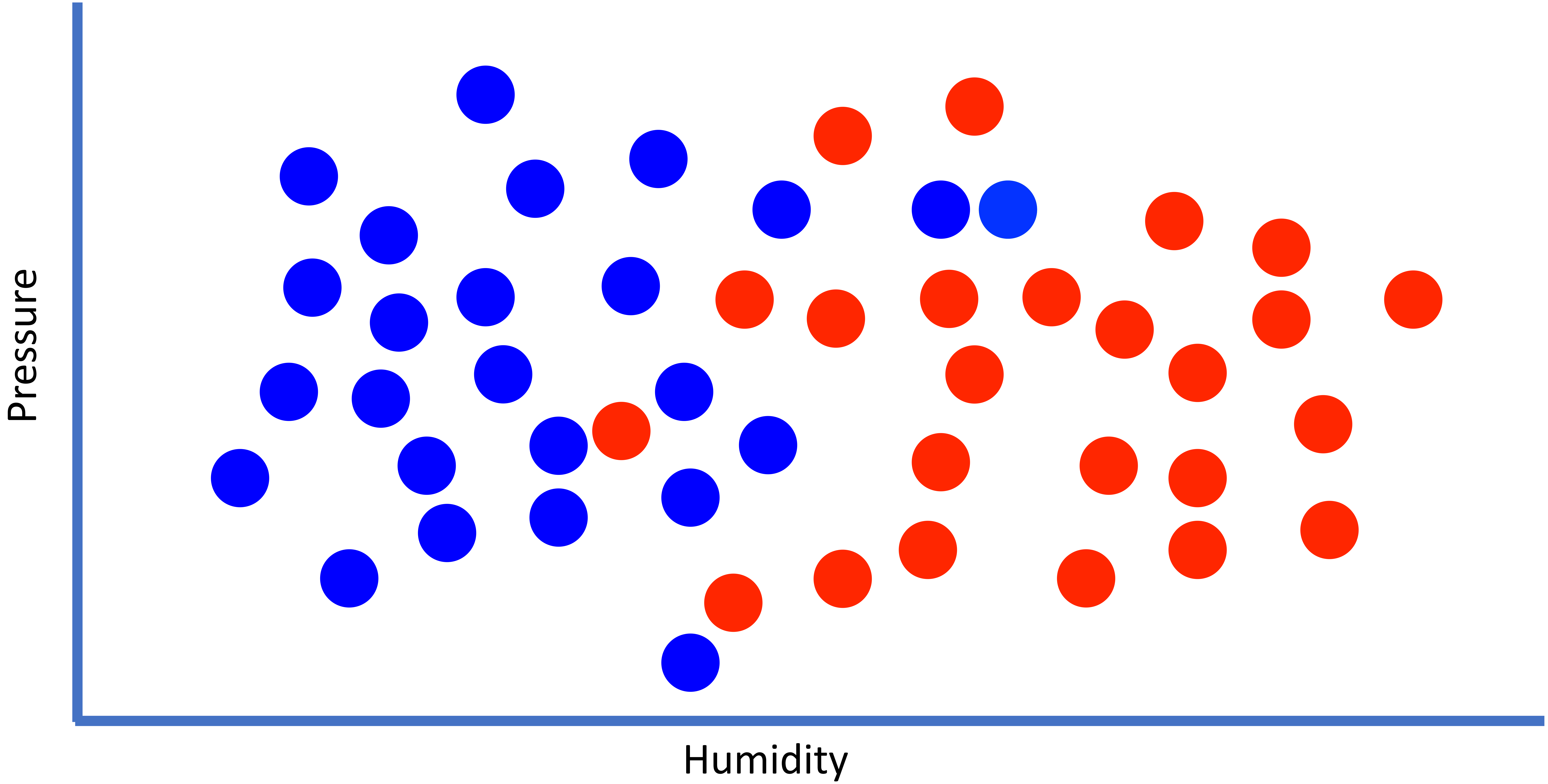
Nearest Neighbor Classification: also known as the Nearest Neighbor algorithm, is one of the simplest and intuitive methods for classification tasks in machine learning.

- When presented with an input, designate the class corresponding to the nearest data point to that input.

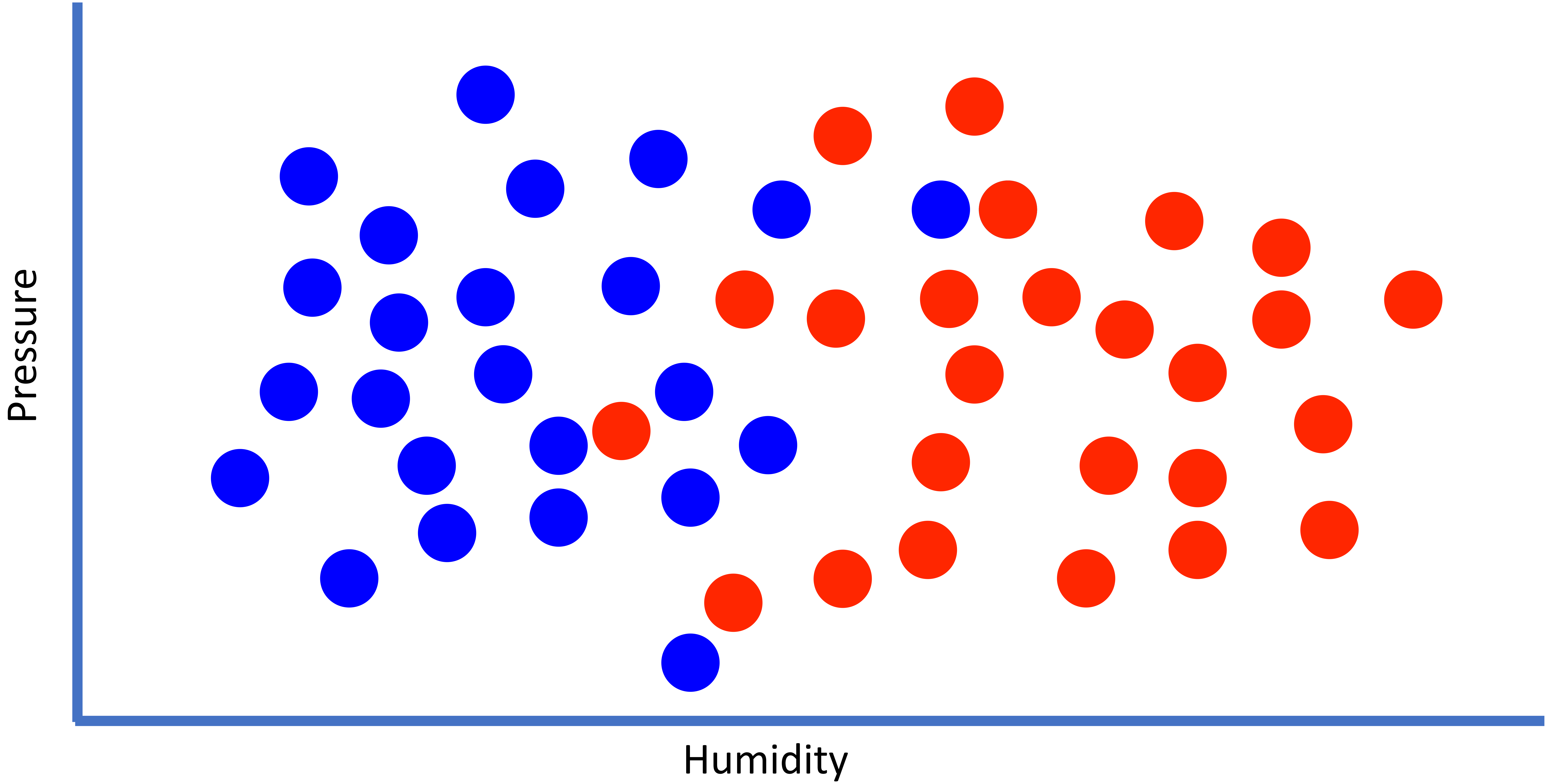
Classification



Classification



Classification



Nearest Neighbor Classification

k-Nearest Neighbor Classification

1-NN

2-NN

3-NN

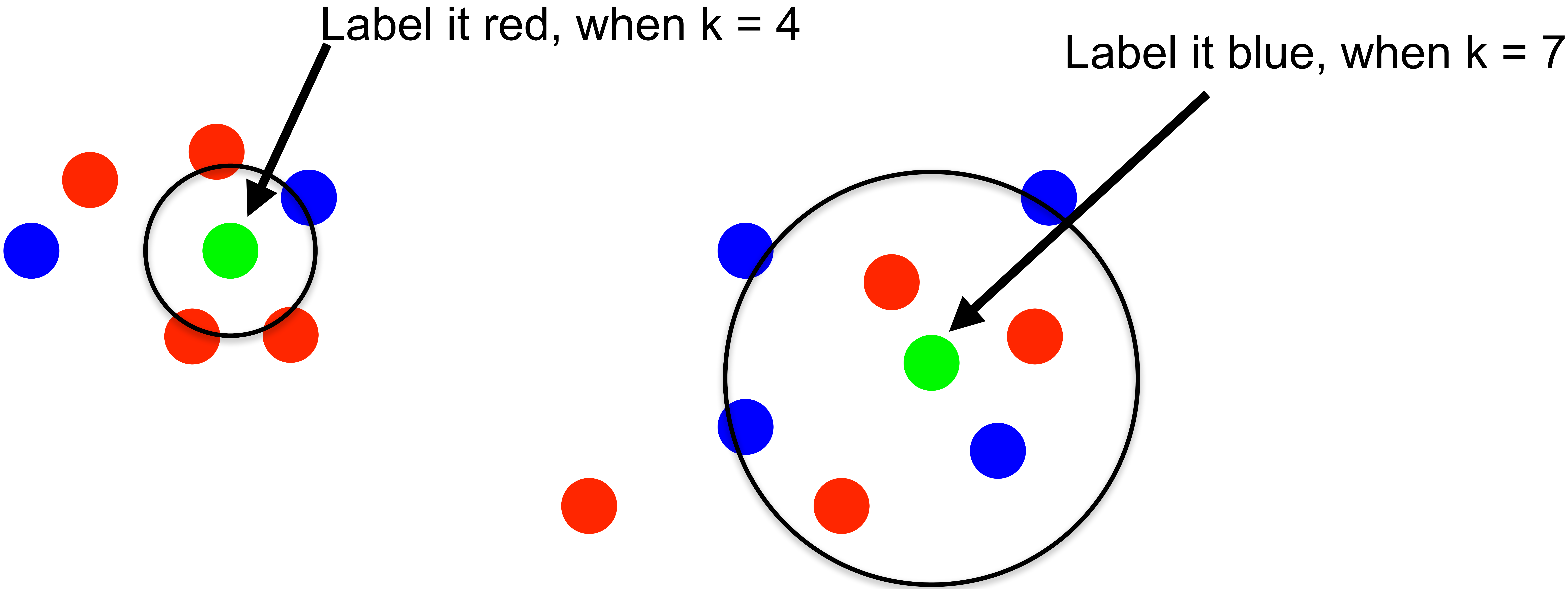
.
.
.

k-Nearest Neighbors Classification

The k-Nearest Neighbors (k-NN) algorithm is a supervised learning method used for classification and regression tasks.

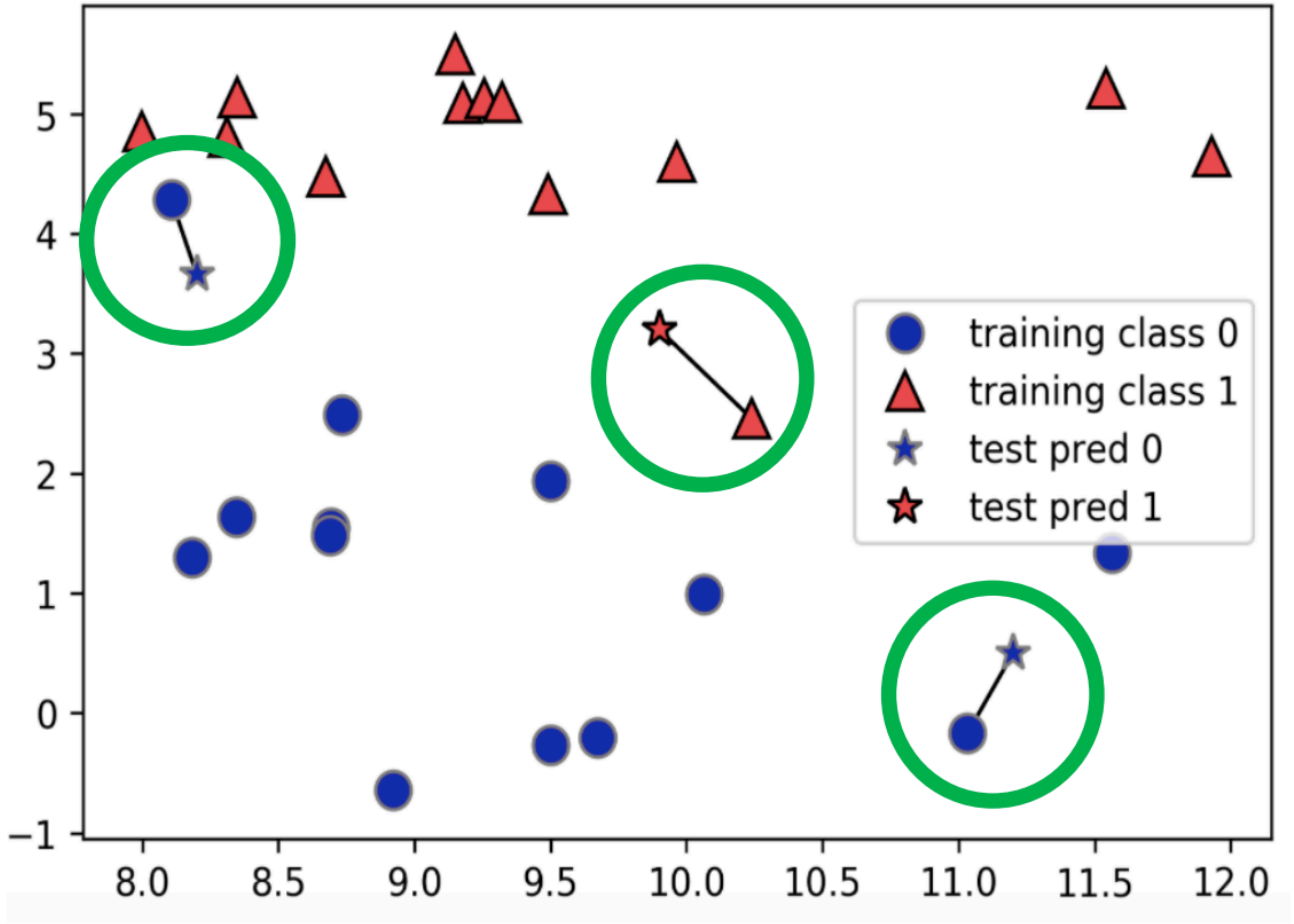
- In k-NN classification, the class of a new data point is determined by the majority class among its k nearest neighbors in the feature space.
- When presented with an input, designate the class corresponding to the ***k*** nearest data point to that input.
- A method for classify cases based on similarity to other cases.

k-Nearest Neighbors Classification

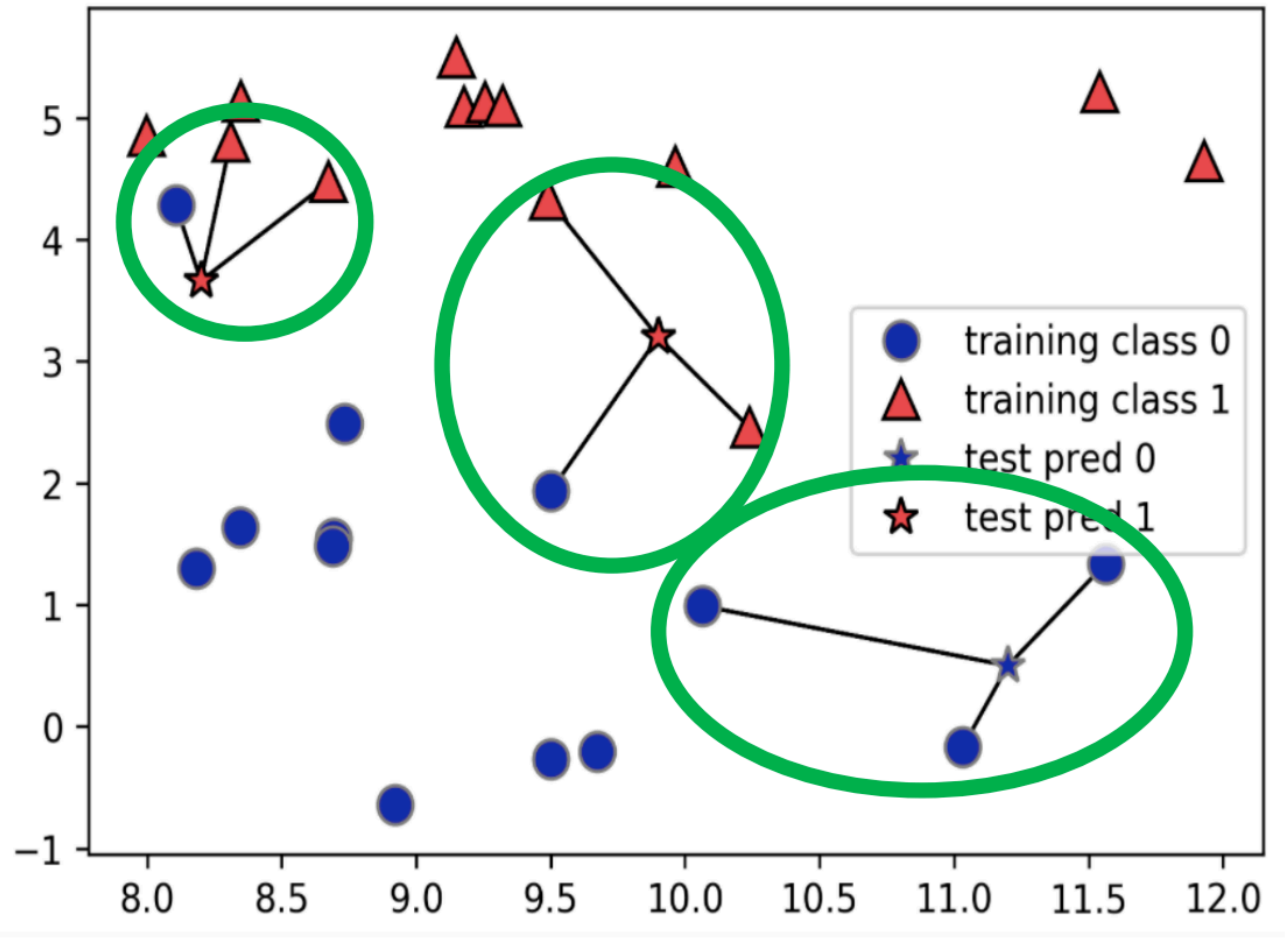


k-Nearest Neighbors Classification

When K=1:



When K=3:



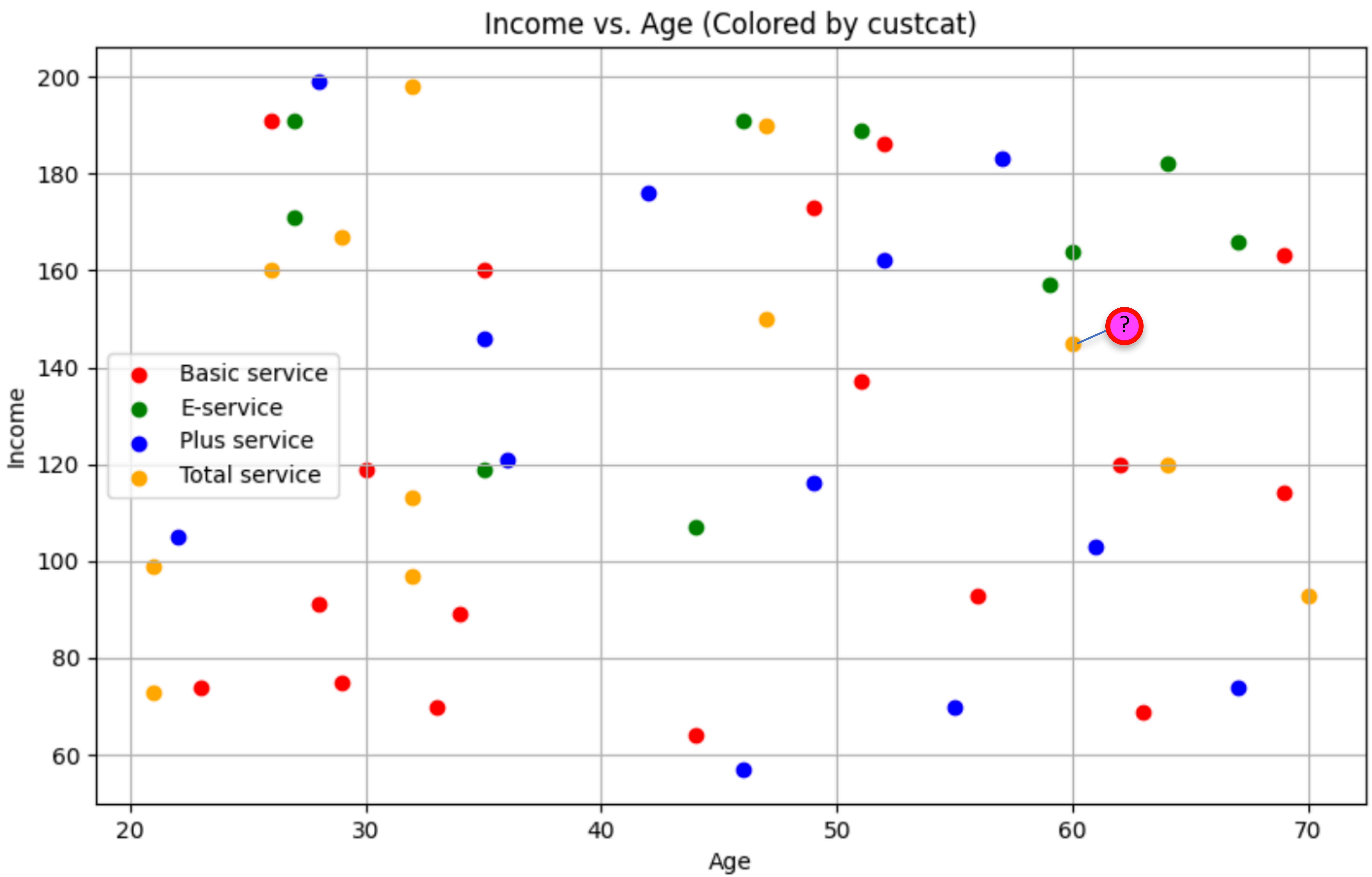
k-Nearest Neighbor - Example

region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
2	44	1	9	64	4	5	0	0	2	1
3	33	1	7	136	5	5	0	0	6	4
3	52	1	24	116	1	29	0	1	2	3
2	33	0	12	33	2	0	0	1	1	1
2	30	1	9	30	1	2	0	0	4	3
2	39	0	17	78	2	16	0	1	1	3
3	22	1	2	19	2	4	0	1	5	2
2	35	0	5	76	2	10	0	0	3	4
3	63	1	7	145	4	31	0	0	5	?

Value	Label
1	Basic service
2	E-Service
3	Plus Service
4	Total service

KNN: A method for classify cases based on similarity to other cases.

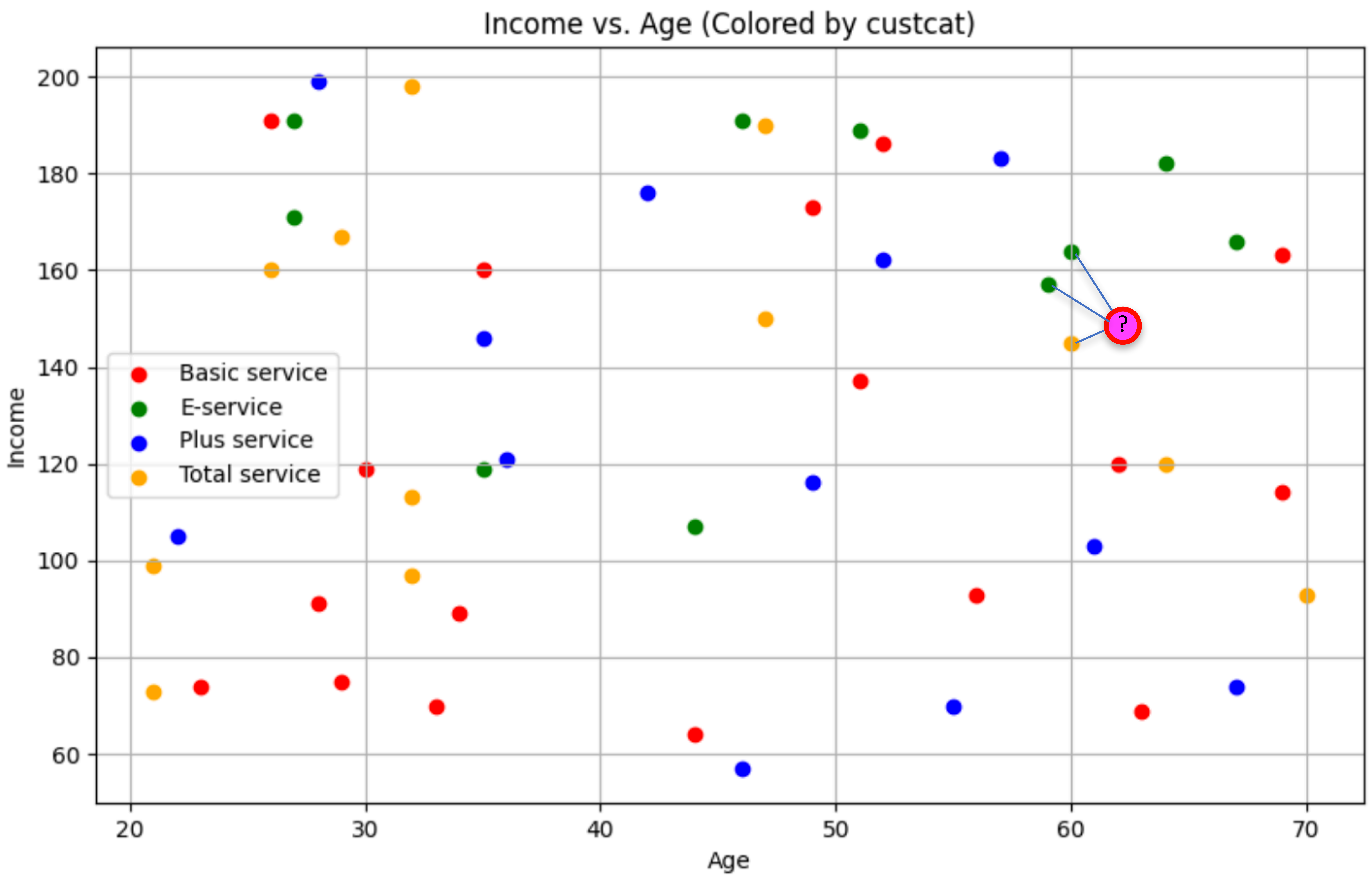
k-Nearest Neighbors - Example



When k = 1

Total Service

k-Nearest Neighbors - Example



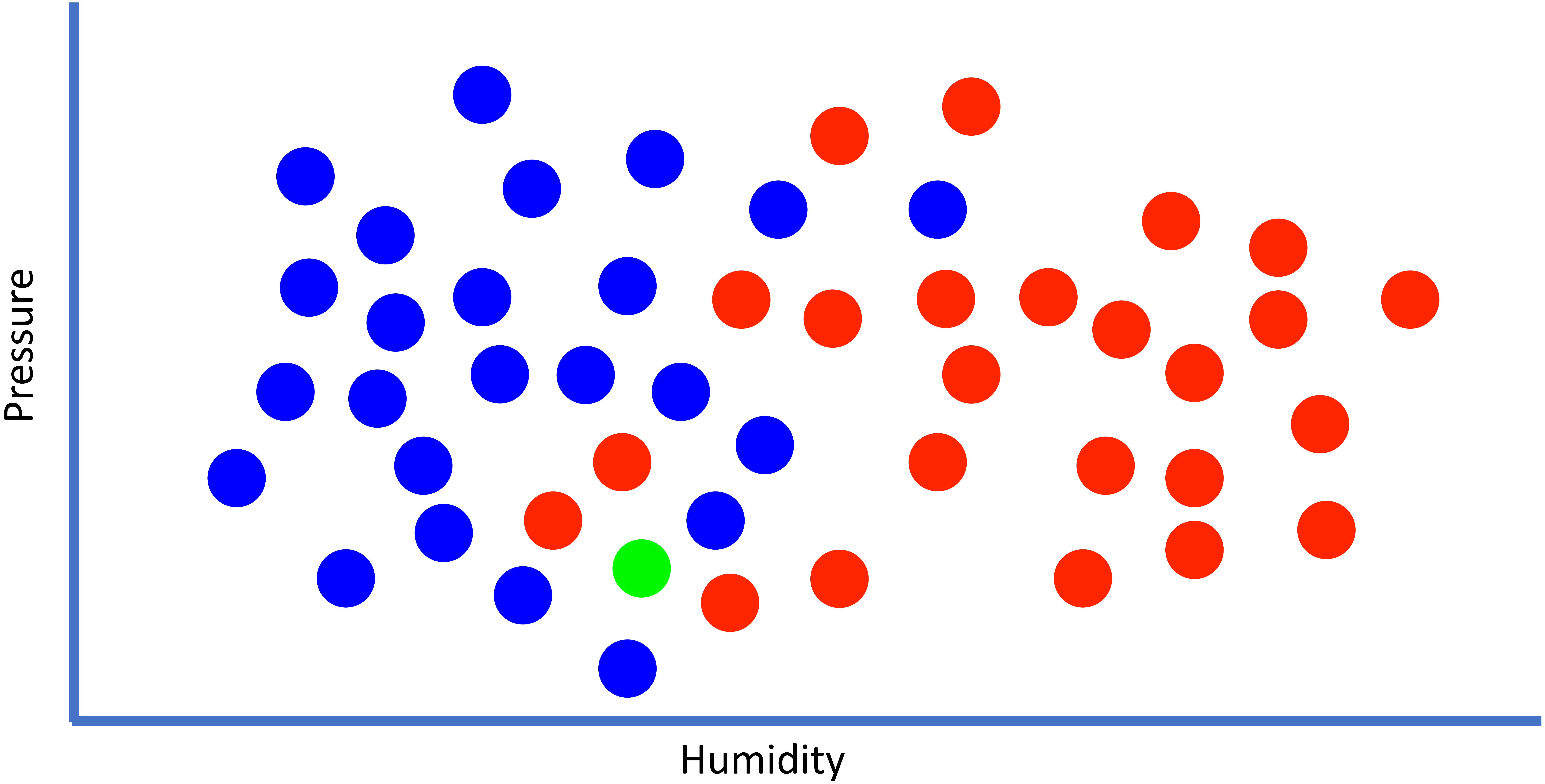
When k = 3

E-Service

k-Nearest Neighbors - Voting

- **Majority voting:** Refers to the process of deciding the class label of a data point based on the majority class among its nearest neighbors.
- In simple majority voting, each neighbor gets equal weight in the voting process. The class with the most votes wins.

Example for discussion



k-Nearest Neighbors - Distance Metrics

- **Distance Metrics**: Refers to the method used to quantify the distance between data points, which is crucial for identifying the nearest neighbors and making predictions in KNN.
- **Example**: Common distance metrics include Euclidean distance, Manhattan distance, and Minkowski distance.

Distance Metrics

- **Euclidean Distance:** is the most common distance metric used in KNN.
- It calculates the straight-line distance between two points in Euclidean space.
- Mathematically, the Euclidean distance between points x1 and x2 in a *d*-dimensional space is given by:

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n} (x2_i - x1_i)^2}$$

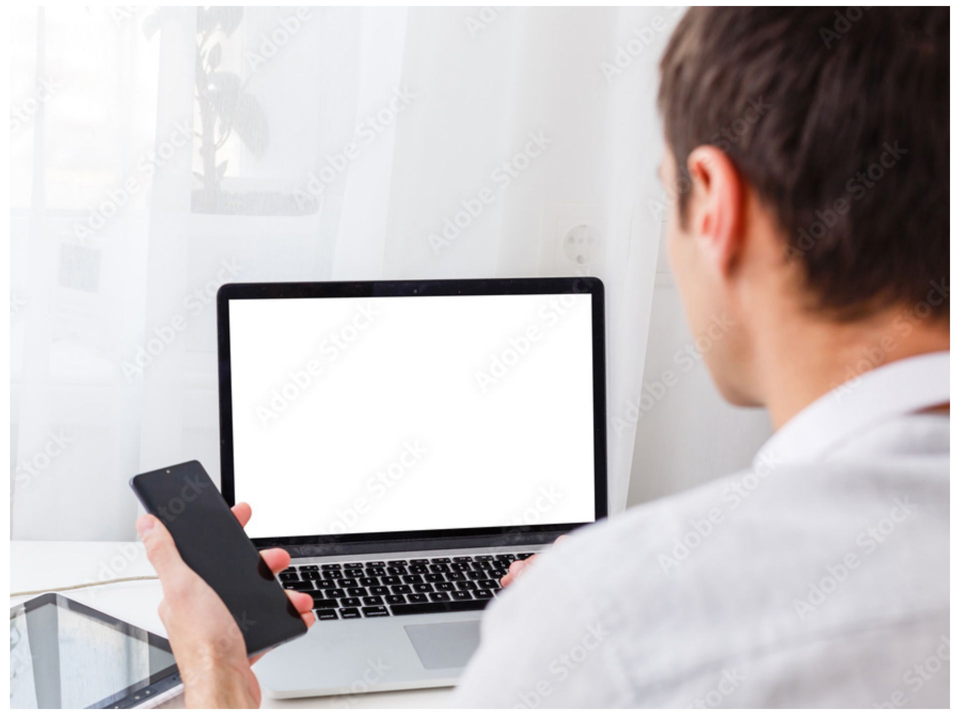
Steps

- **Choose the value of k:** Determine the appropriate number of nearest neighbors (k) to consider for classification.
- **Compute the distance from the unknown case to all cases:** Calculate the distance between the unknown data point and all data points in the dataset using a chosen distance metric (e.g., Euclidean distance).
- **Select the k-nearest neighbors:** Identify the k observations in the training dataset that have the shortest distances to the unknown data point.
- **Predict the response of the unknown data point:** Determine the class label or response value of the unknown data point by considering the most common response value among its k-nearest neighbors.

Example #1

Employee 1

Employee 2



Age
30

Age
45

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n} (x2_i - x1_i)^2}$$

$$d(x1,x2) = \sqrt{(45-30)^2} \Rightarrow d(x1,x2) = \sqrt{(15)^2} \Rightarrow d(x1,x2) = \sqrt{225} = 15$$

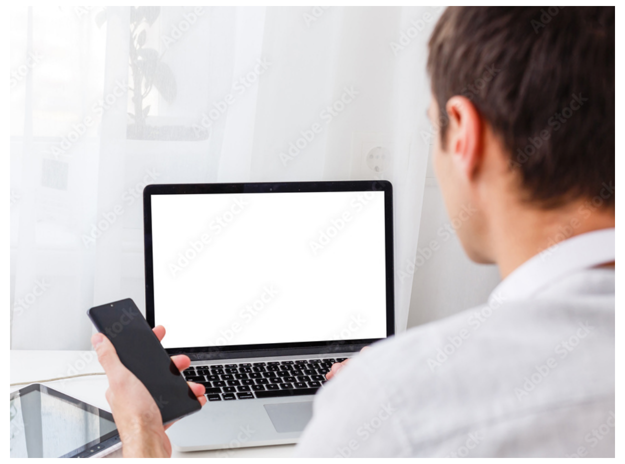
Euclidean distance is often used as a measure of **dissimilarity** between data points.

Example #1

Employee 1

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n} (x2_i - x1_i)^2}$$

Employee 2



Age	Salary
30	1500

Age	Salary
45	1000

Multi-Dimensional vector

$$d(x1,x2) = \sqrt{(45-30)^2 + (1000-1500)^2}$$

$$d(x1,x2) = \sqrt{(15)^2 + (-500)^2}$$

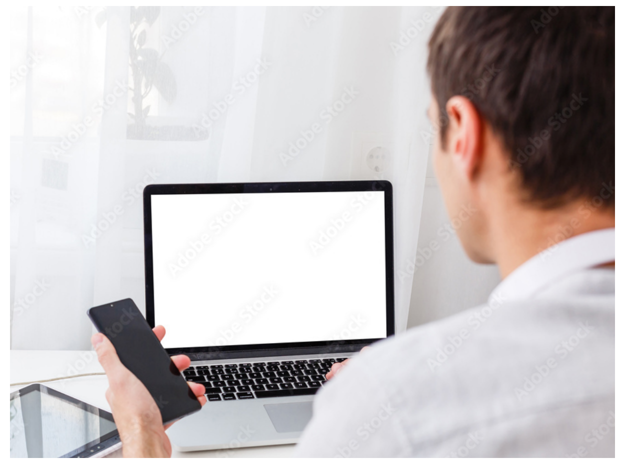
$$d(x1,x2) = 500.224$$

Example #1

Employee 1

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n} (x2_i - x1_i)^2}$$

Employee 2



Age	Salary	Education
30	1500	5

Age	Salary	Education
45	1000	3

$$d(x1,x2) = \sqrt{(45-30)^2 + (1000-1500)^2 + (3-5)^2}$$

$$d(x1,x2) = \sqrt{(15)^2 + (-500)^2 + (-2)^2}$$

$$d(x1,x2) = 500.228$$

Example #2

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain
2024-02-26	70	1008	???

Example #2

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain
2024-02-26	70	1008	???

$$d(x1,x2) = \sqrt{\sum_{i=1}^{n} (x2_i - x1_i)^2}$$

$$d(x1,x2) = \sqrt{(x2_h - x1_h)^2 + (x2_p - x1_p)^2}$$

1. Distance to 2023-02-26:
   $$d(x_1, x_2) = \sqrt{(75-70)^2 + (1005-1008)^2}$$
2. Distance to 2023-02-27:
   $$d(x_1, x_2) = \sqrt{(60-70)^2 + (1018-1008)^2}$$
3. Distance to 2023-02-28:
   $$d(x_1, x_2) = \sqrt{(82-70)^2 + (1001-1008)^2}$$
4. Distance to 2023-03-01:
   $$d(x_1, x_2) = \sqrt{(45-70)^2 + (1020-1008)^2}$$
5. Distance to 2023-03-02:
   $$d(x_1, x_2) = \sqrt{(90-70)^2 + (998-1008)^2}$$

Example #2

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain
2024-02-26	70	1008	???

$$d(x1,x2) = \sqrt{(x2_h - x1_h)^2 + (x2_p - x1_p)^2}$$

1. Distance to 2023-02-26:

$$d(x_1, x_2) = \sqrt{(75-70)^2 + (1005-1008)^2}$$
$$d(x_1, x_2) = \sqrt{5^2 + (-3)^2}$$
$$d(x_1, x_2) = \sqrt{25+9}$$
$$d(x_1, x_2) = \sqrt{34} \approx 5.83$$

2. Distance to 2023-02-27:

$$d(x_1, x_2) = \sqrt{(60-70)^2 + (1018-1008)^2}$$
$$d(x_1, x_2) = \sqrt{(-10)^2 + 10^2}$$
$$d(x_1, x_2) = \sqrt{100+100}$$
$$d(x_1, x_2) = \sqrt{200} \approx 14.14$$

3. Distance to 2023-02-28:

$$d(x_1, x_2) = \sqrt{(82-70)^2 + (1001-1008)^2}$$
$$d(x_1, x_2) = \sqrt{12^2 + (-7)^2}$$
$$d(x_1, x_2) = \sqrt{144+49}$$
$$d(x_1, x_2) = \sqrt{193} \approx 13.89$$

4. Distance to 2023-03-01:

$$d(x_1, x_2) = \sqrt{(45-70)^2 + (1020-1008)^2}$$
$$d(x_1, x_2) = \sqrt{(-25)^2 + 12^2}$$
$$d(x_1, x_2) = \sqrt{625+144}$$
$$d(x_1, x_2) = \sqrt{769} \approx 27.73$$

5. Distance to 2023-03-02:

$$d(x_1, x_2) = \sqrt{(90-70)^2 + (998-1008)^2}$$
$$d(x_1, x_2) = \sqrt{20^2 + (-10)^2}$$
$$d(x_1, x_2) = \sqrt{400+100}$$
$$d(x_1, x_2) = \sqrt{500} \approx 22.36$$

Example #2

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain
2024-02-26	70	1008	???

$$d(x1,x2) = \sqrt{(x2_h - x1_h)^2 + (x2_p - x1_p)^2}$$

1. Distance to 2023-02-26:

$$d(x_1,x_2) = \sqrt{(75-70)^2 + (1005-1008)^2}$$
$$d(x_1,x_2) = \sqrt{5^2 + (-3)^2}$$
$$d(x_1,x_2) = \sqrt{25+9}$$
$$d(x_1,x_2) = \sqrt{34} \approx 5.83$$

2. Distance to 2023-02-27:

$$d(x_1,x_2) = \sqrt{(60-70)^2 + (1018-1008)^2}$$
$$d(x_1,x_2) = \sqrt{(-10)^2 + 10^2}$$
$$d(x_1,x_2) = \sqrt{100+100}$$
$$d(x_1,x_2) = \sqrt{200} \approx 14.14$$

3. Distance to 2023-02-28:

$$d(x_1,x_2) = \sqrt{(82-70)^2 + (1001-1008)^2}$$
$$d(x_1,x_2) = \sqrt{12^2 + (-7)^2}$$
$$d(x_1,x_2) = \sqrt{144+49}$$
$$d(x_1,x_2) = \sqrt{193} \approx 13.89$$

4. Distance to 2023-03-01:

$$d(x_1,x_2) = \sqrt{(45-70)^2 + (1020-1008)^2}$$
$$d(x_1,x_2) = \sqrt{(-25)^2 + 12^2}$$
$$d(x_1,x_2) = \sqrt{625+144}$$
$$d(x_1,x_2) = \sqrt{769} \approx 27.73$$

5. Distance to 2023-03-02:

$$d(x_1,x_2) = \sqrt{(90-70)^2 + (998-1008)^2}$$
$$d(x_1,x_2) = \sqrt{20^2 + (-10)^2}$$
$$d(x_1,x_2) = \sqrt{400+100}$$
$$d(x_1,x_2) = \sqrt{500} \approx 22.36$$

1. Distance to 2023-02-26: $d \approx 5.83$ (Rain)
2. Distance to 2023-02-27: $d \approx 14.14$ (Not Rain)
3. Distance to 2023-02-28: $d \approx 13.89$ (Rain)
4. Distance to 2023-03-01: $d \approx 27.73$ (Not Rain)
5. Distance to 2023-03-02: $d \approx 22.36$ (Rain)

3-NN

Example #2

Date	Humidity (%)	Pressure (hPa)	Rain or Not (Label)
2023-02-26	75	1005	Rain
2023-02-27	60	1018	Not Rain
2023-02-28	82	1001	Rain
2023-03-01	45	1020	Not Rain
2023-03-02	90	998	Rain
2024-02-26	70	1008	???

$$d(x1,x2) = \sqrt{(x2_h - x1_h)^2 + (x2_p - x1_p)^2}$$

1. Distance to 2023-02-26:

$$d(x_1,x_2) = \sqrt{(75-70)^2 + (1005-1008)^2}$$
$$d(x_1,x_2) = \sqrt{5^2 + (-3)^2}$$
$$d(x_1,x_2) = \sqrt{25+9}$$
$$d(x_1,x_2) = \sqrt{34} \approx 5.83$$

2. Distance to 2023-02-27:

$$d(x_1,x_2) = \sqrt{(60-70)^2 + (1018-1008)^2}$$
$$d(x_1,x_2) = \sqrt{(-10)^2 + 10^2}$$
$$d(x_1,x_2) = \sqrt{100+100}$$
$$d(x_1,x_2) = \sqrt{200} \approx 14.14$$

3. Distance to 2023-02-28:

$$d(x_1,x_2) = \sqrt{(82-70)^2 + (1001-1008)^2}$$
$$d(x_1,x_2) = \sqrt{12^2 + (-7)^2}$$
$$d(x_1,x_2) = \sqrt{144+49}$$
$$d(x_1,x_2) = \sqrt{193} \approx 13.89$$

4. Distance to 2023-03-01:

$$d(x_1,x_2) = \sqrt{(45-70)^2 + (1020-1008)^2}$$
$$d(x_1,x_2) = \sqrt{(-25)^2 + 12^2}$$
$$d(x_1,x_2) = \sqrt{625+144}$$
$$d(x_1,x_2) = \sqrt{769} \approx 27.73$$

5. Distance to 2023-03-02:

$$d(x_1,x_2) = \sqrt{(90-70)^2 + (998-1008)^2}$$
$$d(x_1,x_2) = \sqrt{20^2 + (-10)^2}$$
$$d(x_1,x_2) = \sqrt{400+100}$$
$$d(x_1,x_2) = \sqrt{500} \approx 22.36$$

1. Distance to 2023-02-26: $d \approx 5.83$ (Rain)
2. Distance to 2023-02-27: $d \approx 14.14$ (Not Rain)
3. Distance to 2023-02-28: $d \approx 13.89$ (Rain)

Among these neighbors, two are classified as "Rain", and one is classified as "Not Rain". Therefore, we classify the unknown data point as "Rain" based on **majority voting.**

Other Distance Metrics

Manhattan Distance (L1 norm): It measures the distance between two points by summing the absolute differences between their corresponding coordinates.

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Minkowski Distance: Minkowski distance is a generalized distance metric that includes both Manhattan and Euclidean distances. It is defined as:

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

Strengths & Weaknesses of KNN

Thank You