PART1:
K-NEAREST NEIGHBORS: THE SIMPLEST CLASSIFICATION ALGORITHM

What is K-Nearest Neighbors?

K-Nearest neighbors (KNN) is a supervised ML classification algorithm which classifies a data point to a category based on its nearest neighbors.

The **neighbor** of a data point (A) is defined as the point (B) which is closer to A in terms of proximity or distance.

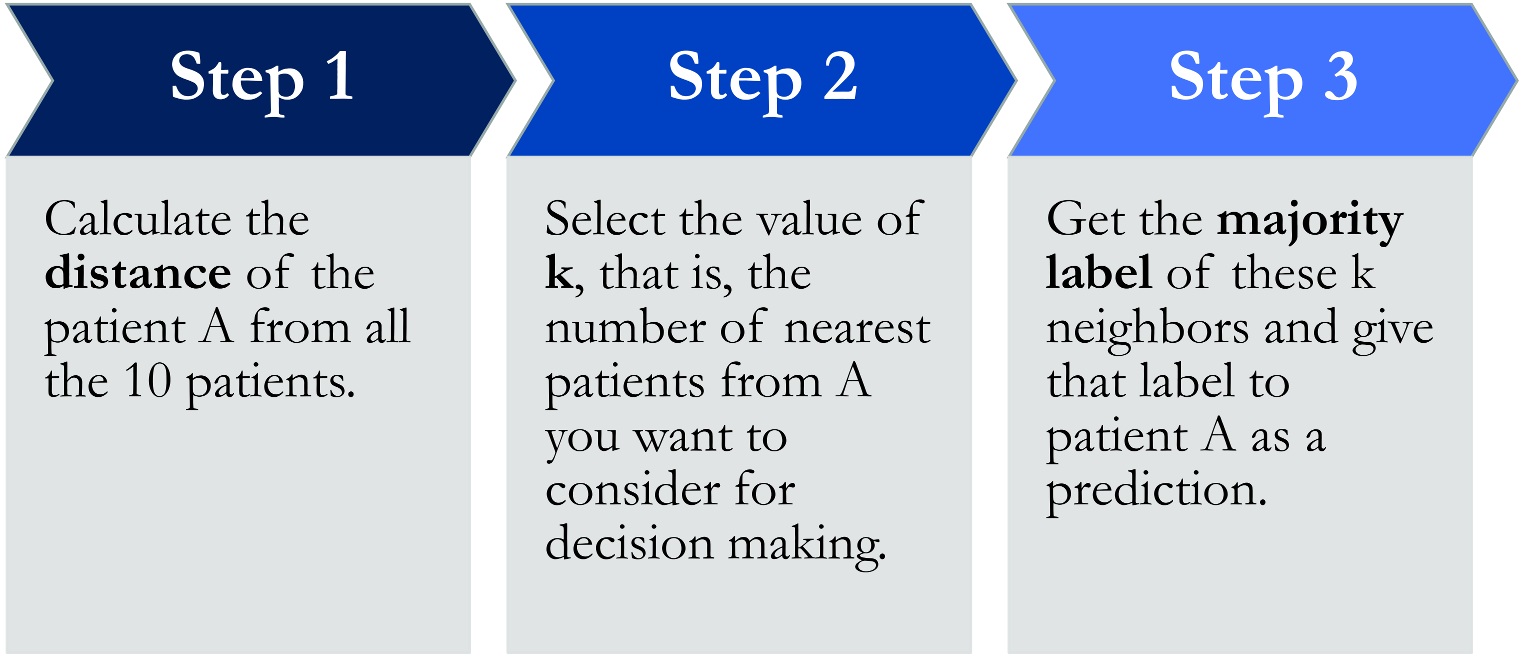
K here is the number of neighbors you want to look around A to assign it to a category. Preferrable choice of K should be an odd number to avoid any ties.

How to apply K-NN?

Problem Statement: Given the BMI (43.6) and age (40) of a person, you need to predict the person A is diabetic or not.

Data you have: Historical data of BMI, age, and the label that the patient is diabetic (1) or not (0).

Here, we have data of 10 patients. In real world, you'll get much more voluminous data.



Step 1

Calculate the **distance** of the patient A from all the 10 patients.

Step 2

Select the value of **k**, that is, the number of nearest patients from A you want to consider for decision making.

Step 3

Get the **majority label** of these k neighbors and give that label to patient A as a prediction.

Distance Calculation

Three most widely used distance metrics in K-NN

Euclidean Distance

- It measures the straight-line distance between two points in Euclidean space. For two points, A and B, in n-dimensional space, the Euclidean distance is calculated as:

$$\text{Euclidean Distance}(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \ldots + (n_1 - n_2)^2}$$

Manhattan Distance

- Manhattan distance, also known as L1 norm, measures the distance between two points as the sum of the absolute differences of their coordinates. The formula for Manhattan distance is:

$$\text{Manhattan Distance}(A, B) = |x_1 - x_2| + |y_1 - y_2| + \ldots + |n_1 - n_2|$$

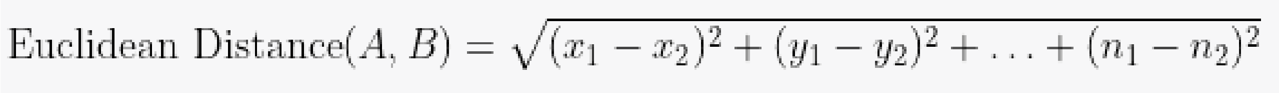
Minkowski Distance

- Minkowski distance is a generalization of both Euclidean and Manhattan distances. It is parameterized by a value, p, and can be adjusted to behave like either of those metrics. When p = 2, it is equivalent to the Euclidean distance, and when p = 1, it is equivalent to the Manhattan distance.

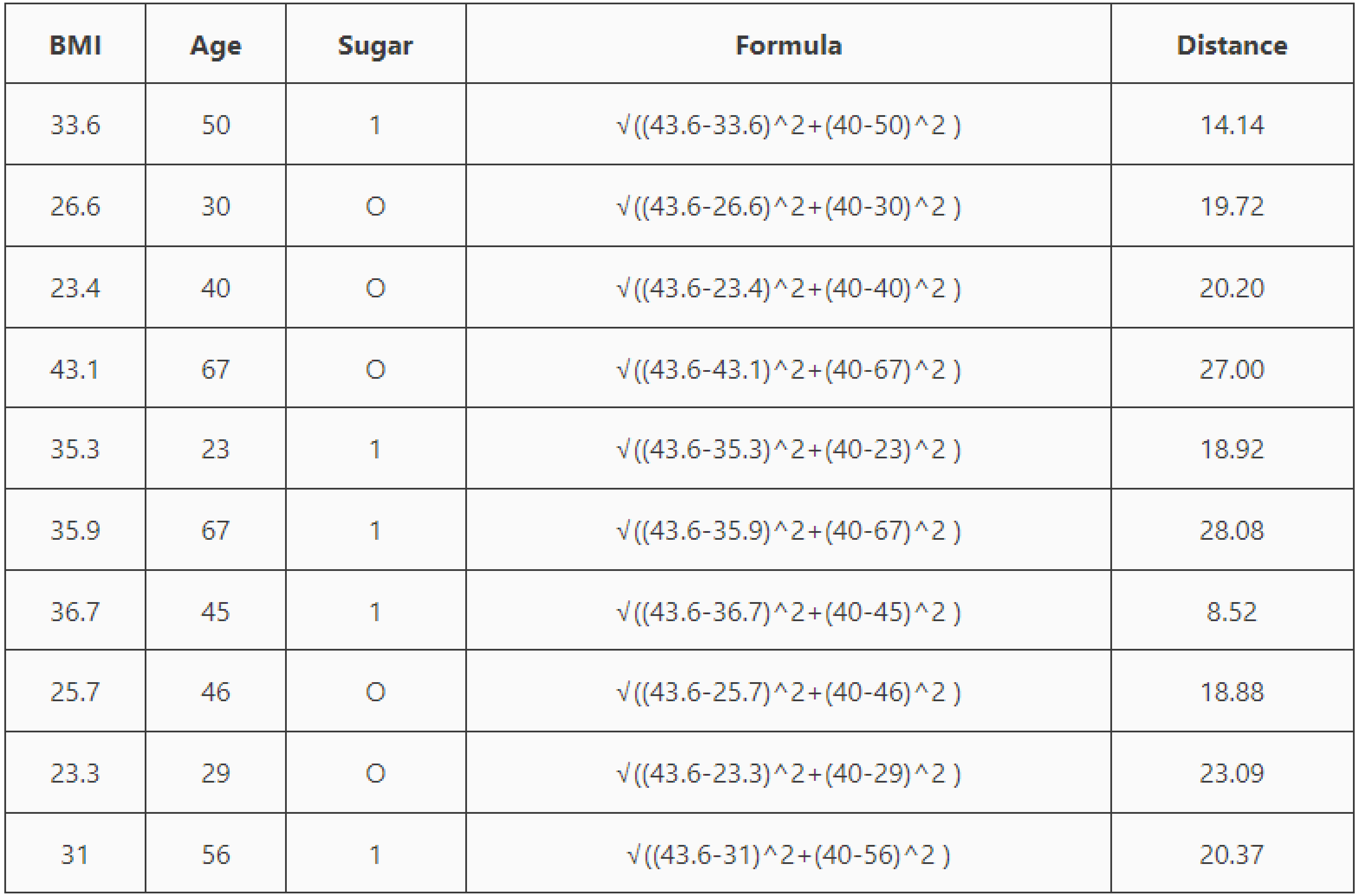
$$\text{Minkowski Distance}(A, B) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Distance calculation in example using Euclidean distance

Test Example BMI=43.6, Age=40



$$\text{Euclidean Distance}(A, B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + \ldots + (n_1 - n_2)^2}$$

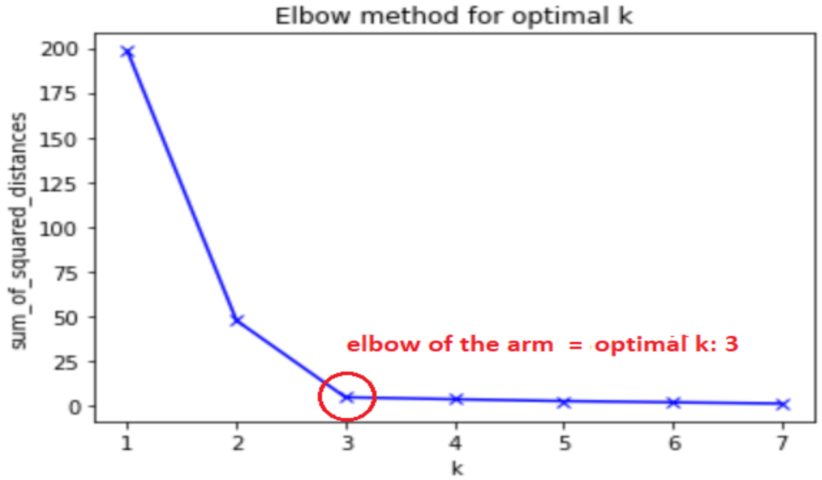


BMI	Age	Sugar	Formula	Distance
33.6	50	1	√((43.6-33.6)^2+(40-50)^2)	14.14
26.6	30	O	√((43.6-26.6)^2+(40-30)^2)	19.72
23.4	40	O	√((43.6-23.4)^2+(40-40)^2)	20.20
43.1	67	O	√((43.6-43.1)^2+(40-67)^2)	27.00
35.3	23	1	√((43.6-35.3)^2+(40-23)^2)	18.92
35.9	67	1	√((43.6-35.9)^2+(40-67)^2)	28.08
36.7	45	1	√((43.6-36.7)^2+(40-45)^2)	8.52
25.7	46	O	√((43.6-25.7)^2+(40-46)^2)	18.88
23.3	29	O	√((43.6-23.3)^2+(40-29)^2)	23.09
31	56	1	√((43.6-31)^2+(40-56)^2)	20.37

Select value of “k”

Experimentation: One of the simplest methods is to experiment with different values of k. Train and test your KNN model using a range of k values and evaluate the model's performance using cross-validation or a hold-out validation set. Choose the k value that results in the best performance metric for your specific problem, such as accuracy, F1-score, or AUC-ROC.

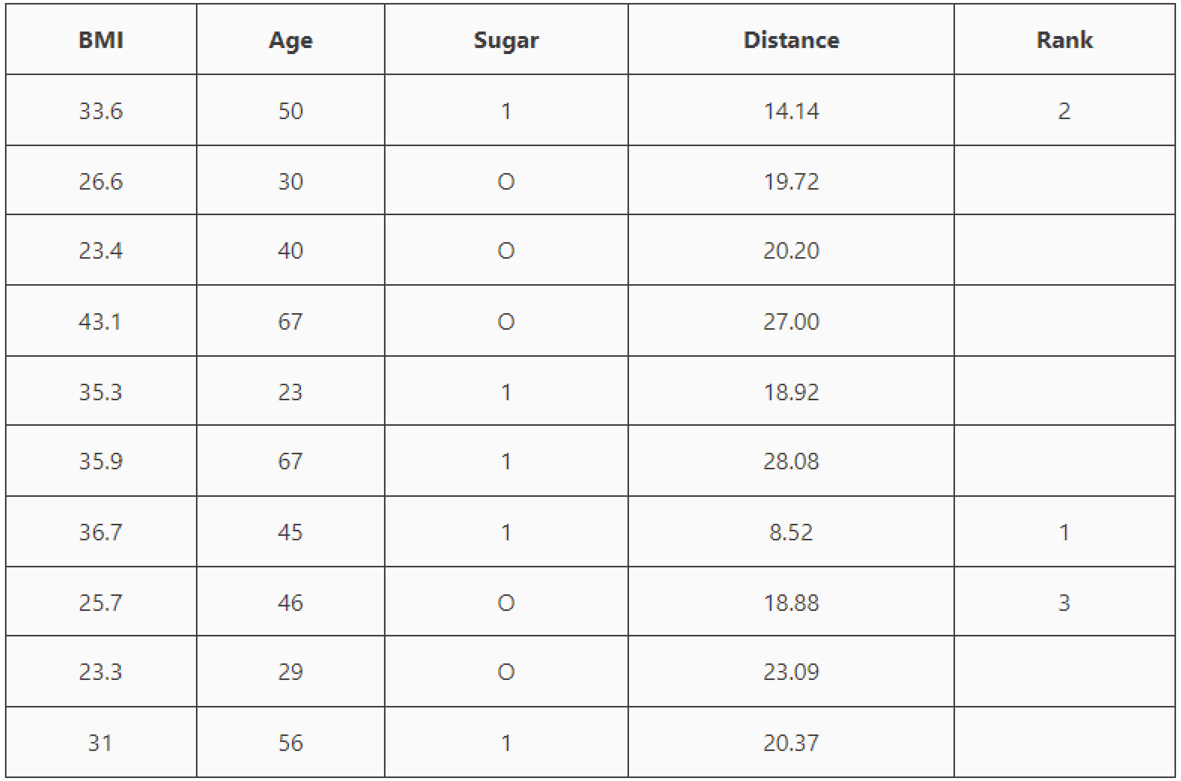
Elbow Method: The elbow method is a graphical technique for selecting k. Plot the performance metric (e.g., WSS, accuracy) on the y-axis against different k values on the x-axis. The point where the performance begins to stabilize or "elbow" is a good choice for k. Beyond this point, increasing k may not significantly improve the model.



Getting value of k and majority label in the example

Value of k: Here, we have taken by experimentation the value of k as 3. Hence, we will look for 3 nearest neighbors for test point BMI = 43.6 and Age = 40.

The 3 NN are as follows:



BMI	Age	Sugar	Distance	Rank
33.6	50	1	14.14	2
26.6	30	O	19.72	
23.4	40	O	20.20	
43.1	67	O	27.00	
35.3	23	1	18.92	
35.9	67	1	28.08	
36.7	45	1	8.52	1
25.7	46	O	18.88	3
23.3	29	O	23.09	
31	56	1	20.37	

Majority label: From the above table we can see that the three labels are 1, 1, and 0. The majority label out of these is 1 and hence, our prediction for the test point BMI = 43.6 and Age = 40 is 1 (Diabetic).

Applications of KNN in real life

Quality Control in Manufacturing: KNN can be used to monitor and maintain product quality in manufacturing processes. It compares measurements from newly produced items to those of known high-quality items, flagging any deviations.

Customer Segmentation: In marketing, KNN can help segment customers based on their behavior, preferences, or demographics. This information is valuable for targeted marketing campaigns and product recommendations.

Recommendation Systems: KNN is commonly used in recommendation systems to suggest products, movies, or music to users based on their preferences and the preferences of similar users. By measuring the similarity between users or items, KNN can provide personalized recommendations.

KNOWLEDGE CHECK

References and resources to study

- KNN Solved Example Diabetic Patient
- Determining the Optimal K for K-Means Algorithm
- KNN Algorithm - Finding Nearest Neighbors