Artificial Intelligence Masterclass

K – Nearest Neighbors (KNN)

H.M. Samadhi Chathuranga Rathnayake

M.Sc in CS (SU), PG.Dip in SML (Othm), PG.Dip in HRM (LRN), B.Sc (Hons) in IS (UOC), B.Eng (Hons) in SE (LMU), P. Dip EP & SBO (ABE), Dip SE, Dip IT, Dip IT & E-Com, Dip B.Mgt, Dip HRM, Dip Eng

K- Nearest Neighbors

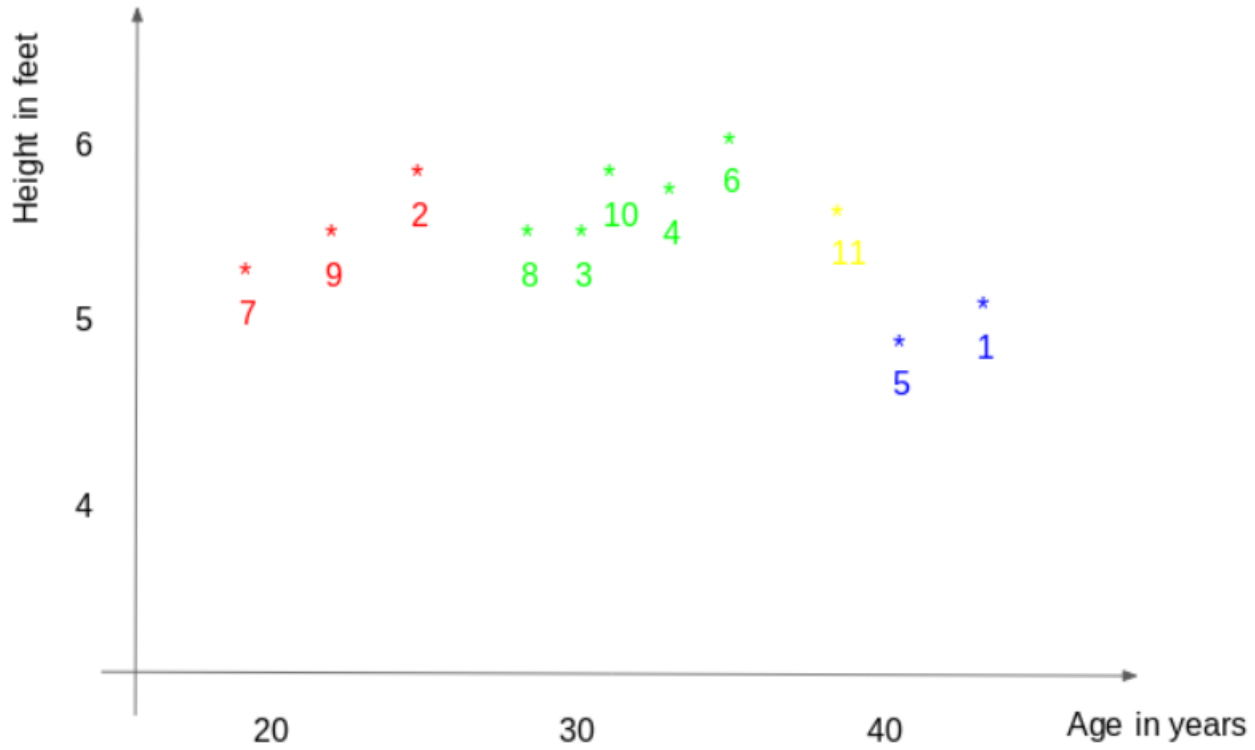
Consider the following example. Think that we are going to fit a model using Height and Age for the response variable Weight.

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

Think that we need to find the Weight for observation 11 using the given data.

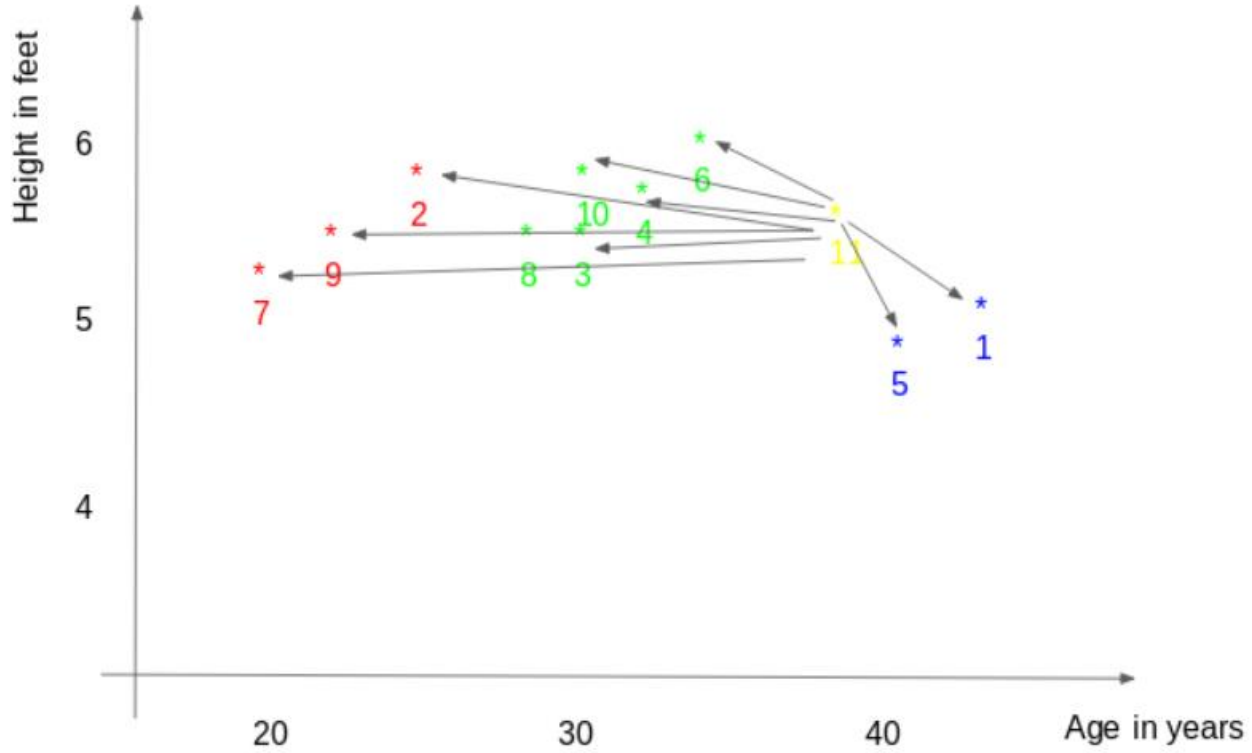
K- Nearest Neighbors

If we plot this,



K- Nearest Neighbors

First, the distance between the new point and each training point is calculated.



K- Nearest Neighbors

For this distance calculations we can use 3 types of distances.

- Euclidean Distance: Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

- Manhattan Distance: This is the distance between real vectors using the sum of their absolute difference.

$$\sum_{i=1}^{k}|x_i - y_i|$$

- Hamming Distance: It is used for categorical variables. If the value (x) and the value (y) are the same, the distance D will be equal to 0 . Otherwise D=1.

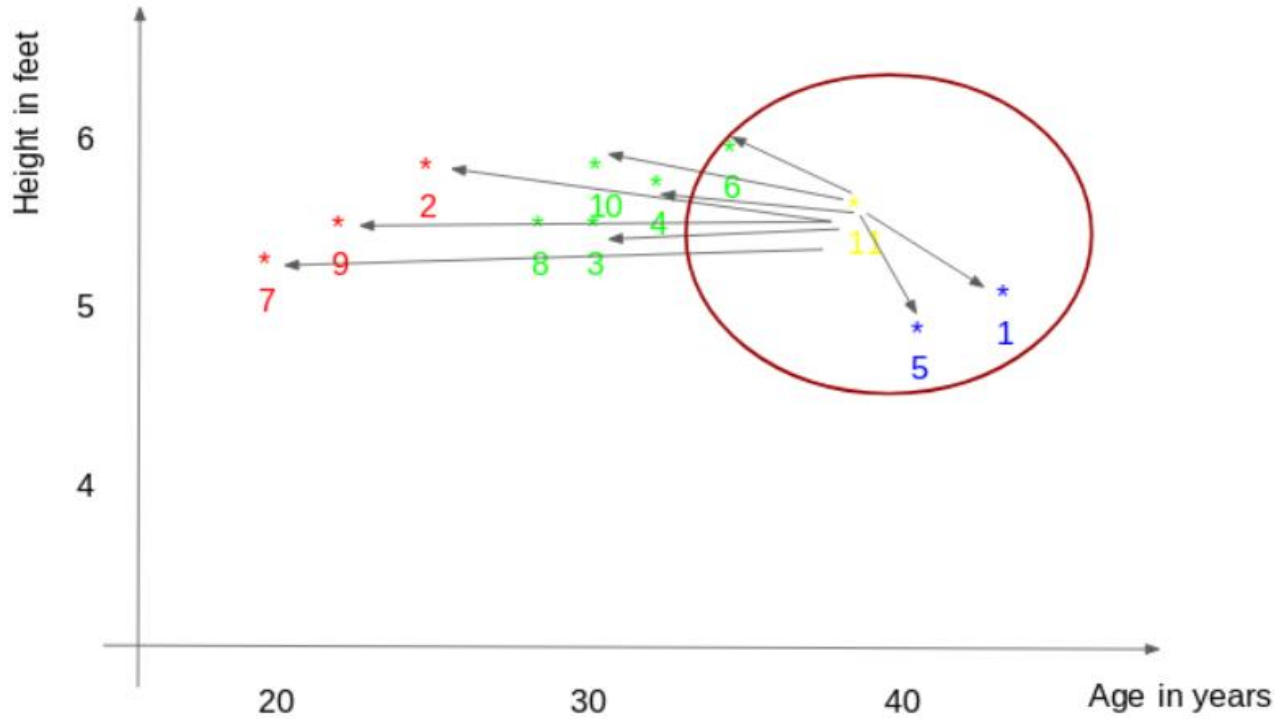
$$D_H = \sum_{i=1}^{k}|x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

K- Nearest Neighbors

The closest k data points are selected (based on the distance). Consider here k=3, so in this example, points 1, 5, 6 will be selected if the value of k is 3.



Then the average of these data points is the final prediction for the new point. Here, we have weight of ID11 is (77+72+60)/3 = 69.66 kg. If we have a classification case, the most frequent category will be chosen to assign.

K- Nearest Neighbors

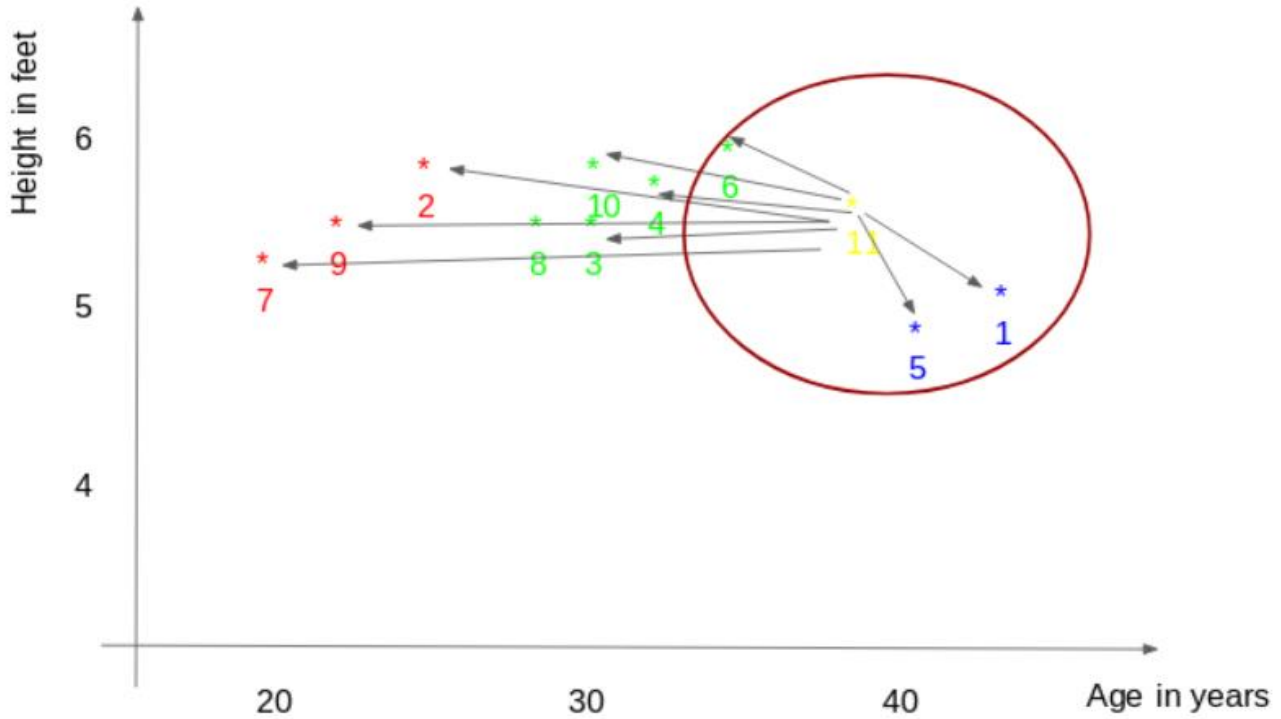
Consider the following example. Think that we are going to fit a model using Height and Age for the response variable Weight.

ID	Height	Age	Weight	Class
1	5	45	77	A
2	5.11	26	47	A
3	5.6	30	55	B
4	5.9	34	59	C
5	4.8	40	72	B
6	5.8	36	60	A
7	5.3	19	40	C
8	5.8	28	60	A
9	5.5	23	45	B
10	5.6	32	58	C
11	5.5	38	?	

Think that we need to find the Class for observation 11 using the given data.

K- Nearest Neighbors

The closest k data points are selected (based on the distance). Consider here k=3, so in this example, points 1, 5, 6 will be selected if the value of k is 3.



If we have a classification case, the most frequent category will be chosen to assign. 11th observation belongs to Class A.

K- Nearest Neighbors

To select the optimum k for the algorithm, we can use the Validation Set or Cross Validation approaches. Consider following graphs. Training & validation errors have been measured for several k values.