KNN - K-Nearest Neighbours

Consider the following dataset, for K=3 and test data (X, 35, 100) as (Person, Age, Salary($)) -> Predict the target.

Person	Age	Salary(K	Target	Distance	
A	18	50	N	52.8	
B	23	55	N	46.6	
C	24	70	N	31.9	(2)
D	41	60	Y	40.9	(3)
E	43	70	Y	30.1	(1)
F	38	40	Y	60.1	
X	35	100	?		

1) Distance = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

$d_1 = \sqrt{(35-18)^2+(100-50)^2} = 52.8$

$d_2 = \sqrt{(35-23)^2+(100-55)^2} = 46.6$

$d_3 = \sqrt{(35-24)^2+(100-70)^2} = 31.9$

$d_4 = \sqrt{(35-41)^2+(100-60)^2} = 40.4$

2) 1) E (30.1, Y)
2) C (31.9, N)
3) D (40.9, Y)

3) Majority -> Y : (2/3)

=) X, 35, 100, Y

For iris dataset:

- How to choose the K value?

1) Accuracy rate approach: We train the model with diff K values and calculate the accuracy for each K

2) Error Rate Approach: Error Rate = 1 - Accuracy

- A lower error rate indicates a better K value.

Demonstration of accuracy rate and error rate:

- small K value may lead to overfitting
- large K value may lead to underfitting

For Diabetes Dataset:

i) What is the purpose of feature scaling?

- Feature scaling is used to normalize the range of independent variables.

2) How to perform feature scaling?

→ Standardization $X_{scaled} = \frac{x - \mu}{\sigma}$

μ - mean, σ - standard deviation.

Used when data follows a normal distribution.