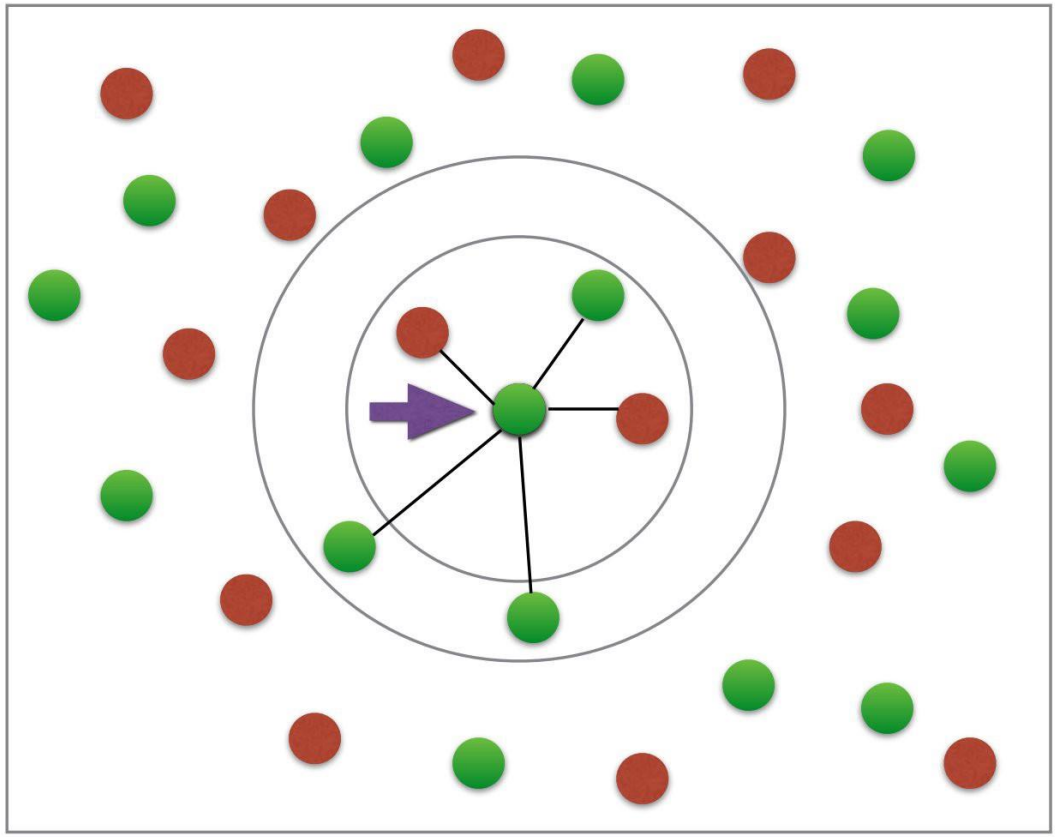
K-Nearest Neighbours

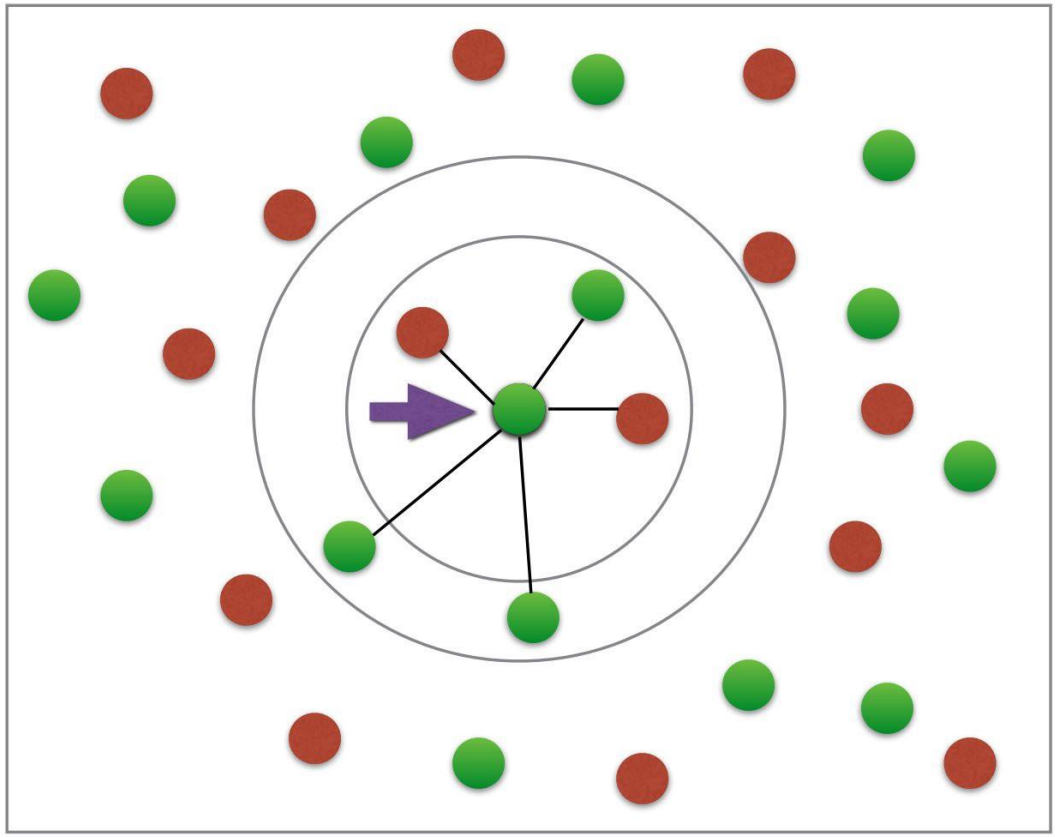


1*vbjGslZW1d4NGco7608o_g.jpeg (1238×988) (medium.com)

Objective

- Nearest Neighbours
- Telecome customer dataset
- Inference
- Implementation of KNN
- Feature Normalization
- Identify value of K

K-Nearest Neighbours



Telecom Customer Dataset

- In the given Telecommunication dataset, with predefined labels, we need to build a model which is used to predict the class of a new or unknown case.
- The example focuses on using demographic data to predict usage patterns.

Independent Variable

Dependent variable

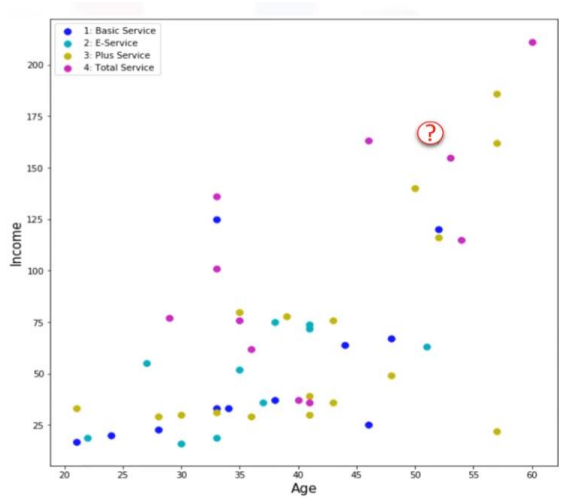
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

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

Telecom Customer Dataset

- In this example, our **objective is to build a classifier, for example using the rows 0 to 7, to predict the class of row 8.**
- We will use a specific type of classification called K-nearest neighbor.
- Just for sake of demonstration, let's use only two fields as predictors - specifically, Age and Income, and then plot the customers based on their group membership.

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?



Telecom Customer Dataset

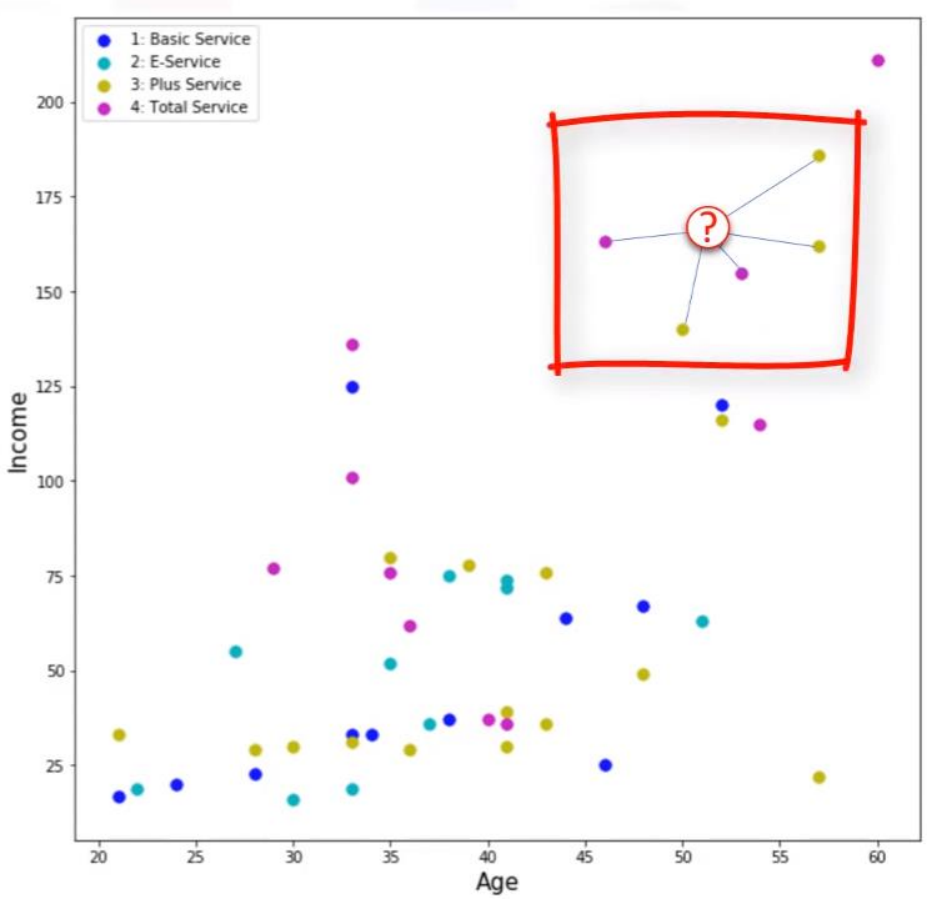
- **How can we find the class of new customer, available at record number 8 with a known age and income?**

- Can we say that the class of our new customer is most probably group 4 because its nearest neighbour is also of class 4?

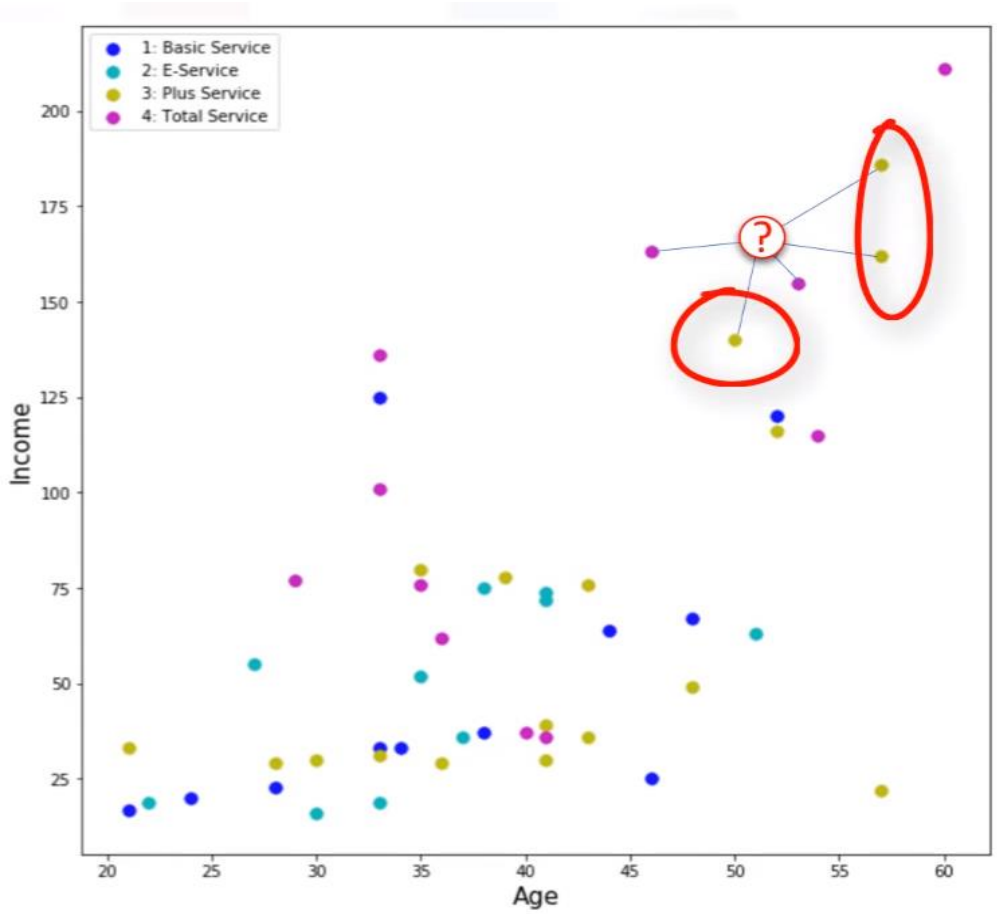
- Yes, we can say so!

Inference

- Now, the question is, “To what extent can we trust our judgment, which is based on the first nearest neighbor?”
- It might be a poor judgment, especially if the first nearest neighbor is a very specific case, or an outlier !
- What if we chose the five nearest neighbors, and did a majority vote among them ?



Inference

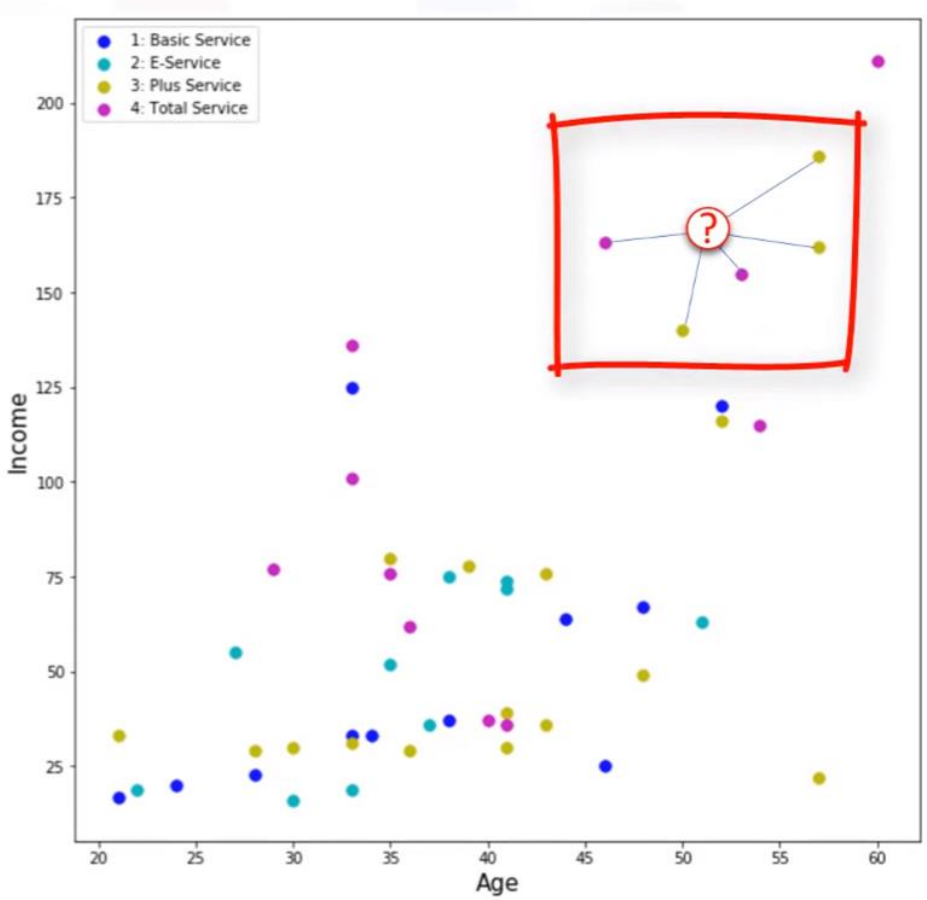


- Does this make more sense?
- Yes !
- In this case, the value of K in the k-nearest neighbours' algorithm is 5.
- This example highlights the intuition behind the k-nearest neighbours' algorithm.

K Nearest Neighbour

- The k-nearest-neighbors algorithm is a classification algorithm that takes a bunch of labelled “neighbor” points and uses them to learn how to label other points.

- Thus, the distance between two cases is a measure of their dissimilarity.



K-Nearest Neighbour Implementation steps

- In a classification problem, the k-nearest neighbors algorithm is implemented using following steps:

1. Pick a value for K.
2. Calculate the distance of unknown case from all cases.
3. Search for the K observations in the training data that are ‘nearest’ to the measurements of the unknown data point.
4. Predict the response of the unknown data point using the most popular response value from the K nearest neighbors.

Calculating Similarity between 2 Data Points

- How can we calculate the similarity between two data points?
- Assume that we have two customers, customer 1 and customer 2 who have only one feature, **Age**.
- We can easily use a specific type of Minkowski distance to calculate the distance of these 2 customers.
- It is called as Euclidian distance.

$$Dis(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

Calculating Similarity between 2 Data Points

- How can we calculate the similarity between two data points?
- Assume that we have two customers, customer 1 and customer 2 who have only one feature, **Age**.
- We can easily use a specific type of Minkowski distance to calculate the distance of these 2 customers.
- It is called as Euclidian distance.

$$Dis(x, y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2}$$

Calculating Similarity between 2 Data Points

- Age of customer 1 = 54 and
- Age of customer 2 = 50,
- Distance between both customer 1 & customer 2 “age” feature are :

  $$Dis(x,y)=\sqrt{((54-50)^2)}=4$$

- **If we have both income and age feature of both customer**, use the same formula.
- Age of customer 1 = 54 and income = 250
- Age of customer 2 = 50 and income = 240
- Distance between Customer 1 “age” & “income” features and between Customer 2 “age” and “income” features, would be calculated as:

  $$Dis(x,y)=10.77$$

Feature Normalization

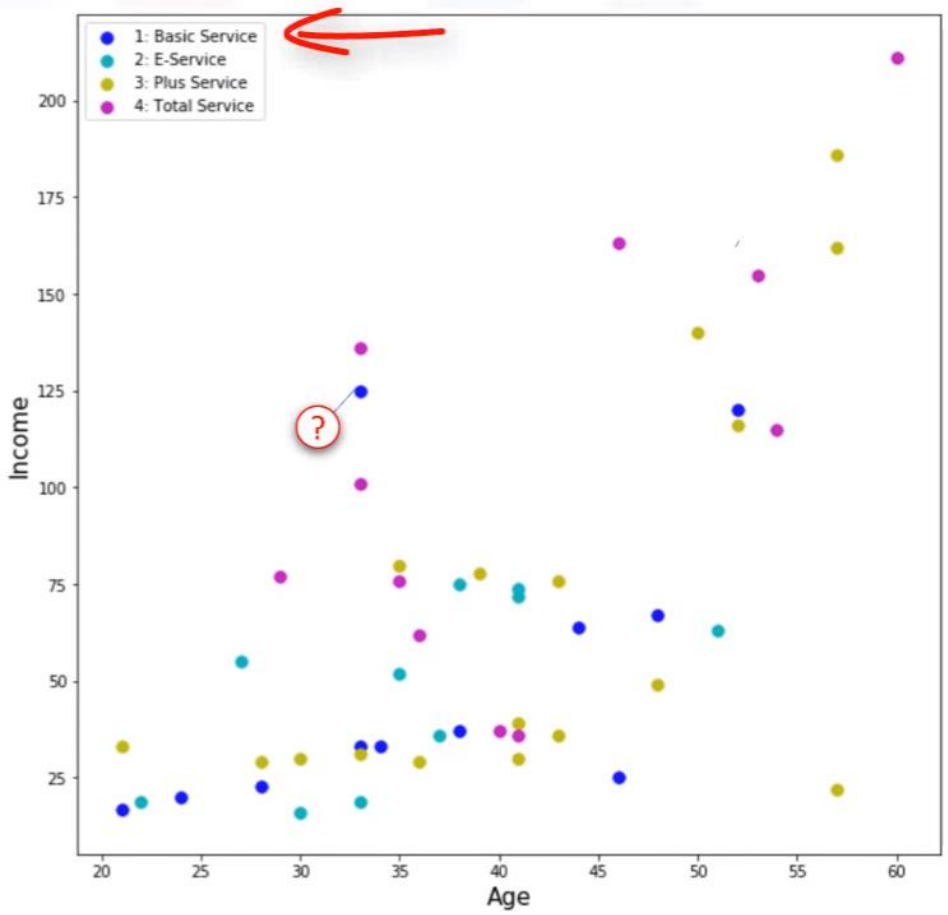
- We can also use the same distance matrix for multi-dimensional vectors.

- **We have to normalize our feature set to get the accurate dissimilarity measure.**

Identifying correct K value

- Assume that we want to find the class of the customer noted as question mark on the chart.

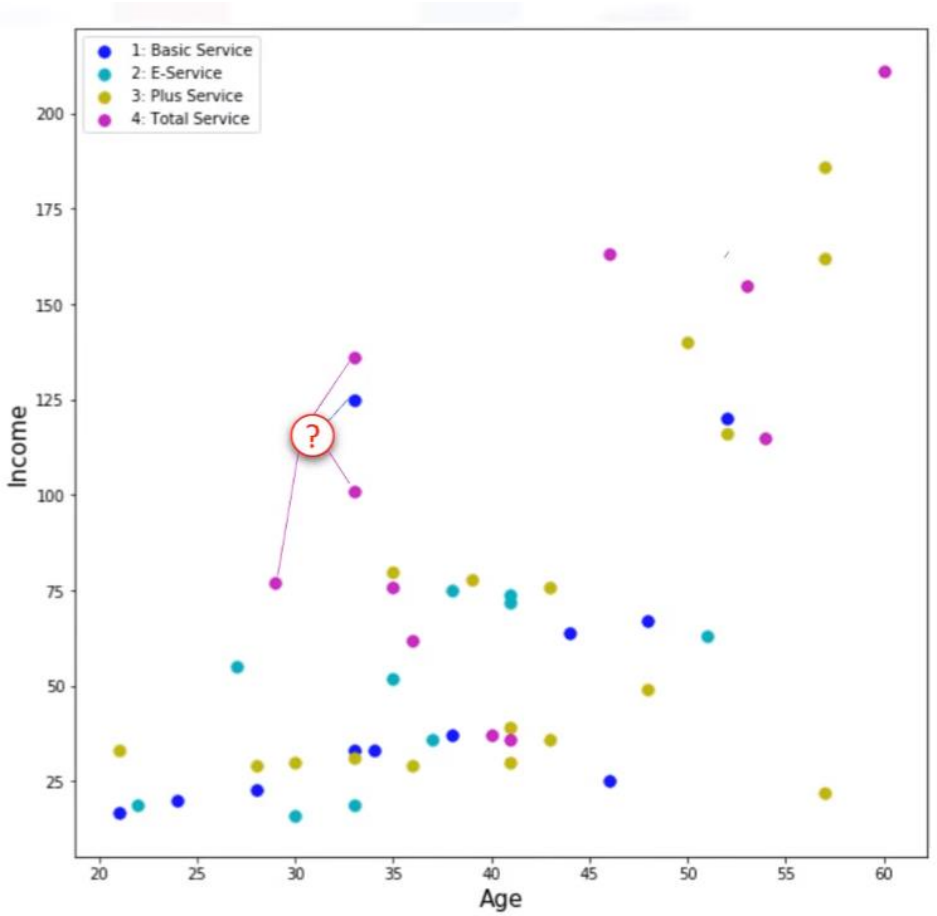
- What happens if we choose a very low value of K, let's say, k=1?



Value of K ?

- A low value of K causes a highly complex model, which might result in over-fitting of the model.

- It means the prediction process is not generalized enough to be used for out-of-sample cases.



Optimizing K?

- So, how we can find the best value for K?
- Calculate the accuracy of the model by choosing K=1 using all samples in your test set.
- Repeat this process, increasing the k, and see which k is best for your model.
- In this example, K=4 gives the best accuracy!

ML Hands-on Session

- A telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups.

- If demographic data can be used to predict group membership, the company can customize offers for individual prospective customers.

- As an Analyst, your job is to analyze the dataset, build a model to be used to predict usage pattern of a new or unknown customer.

Hands On