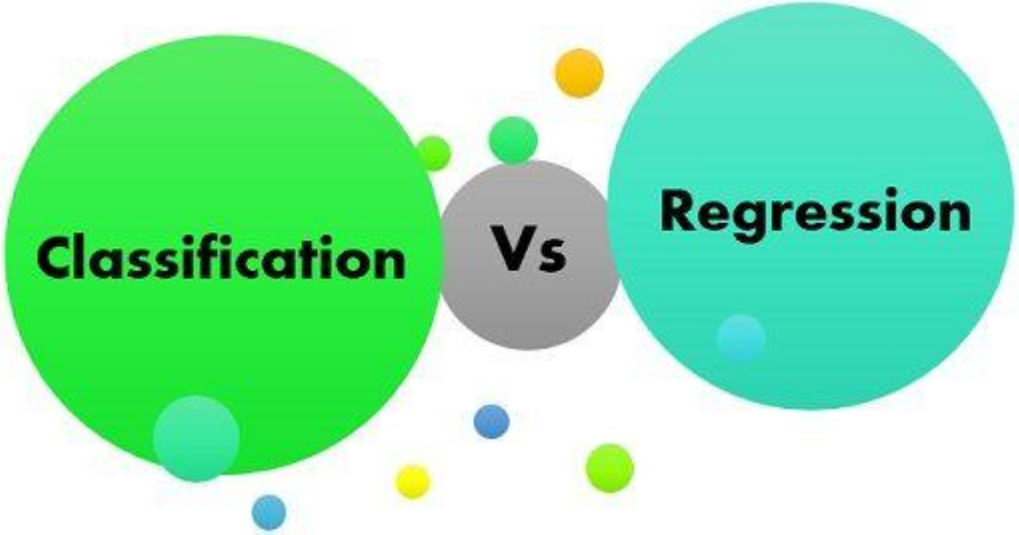
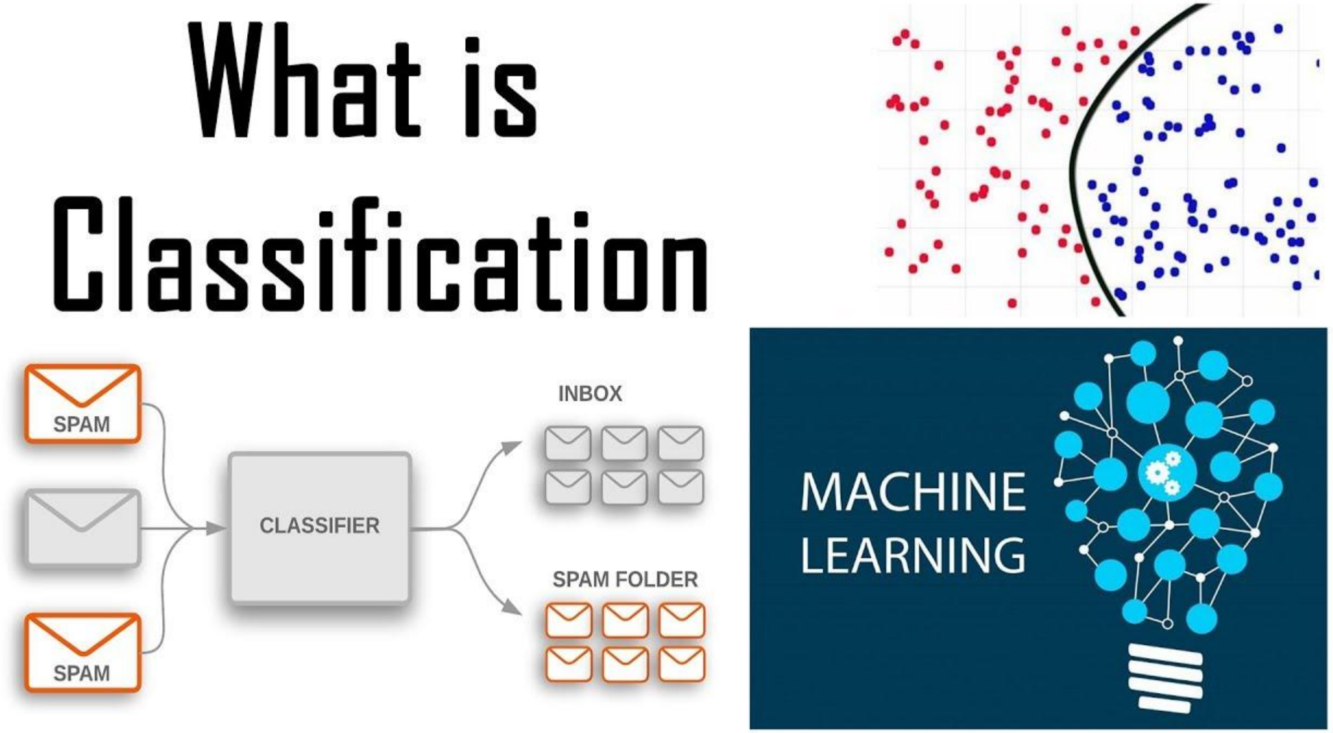


Classification Vs Regression

Classification is the process of dividing the datasets into different categories or groups by adding label.



K - Nearest Neighbour

K-Nearest Neighbour is a simple algorithm that stores all the variable cases and classifies the new data based on the similarity measure.

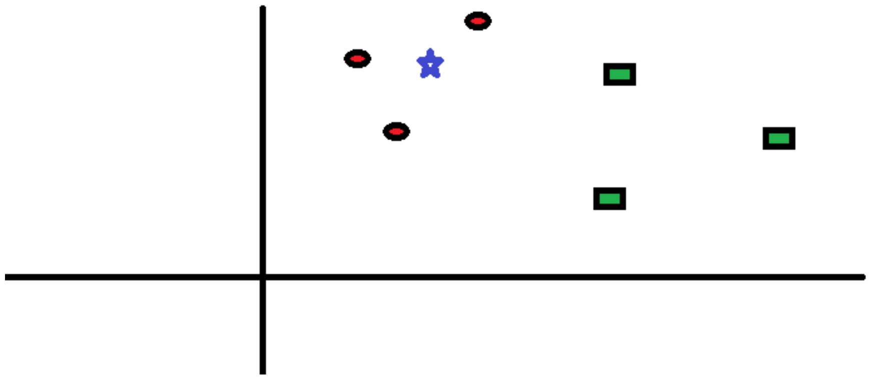
Knn is lazy learner:

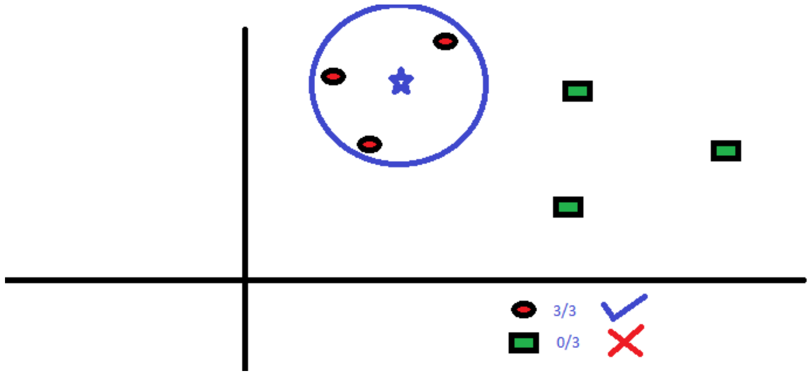
K-NN is a **lazy** learner because it doesn't learn a discriminative function from the training data but “memorizes” the training dataset instead.

(Discriminative function) statistical procedure that classifies unknown individuals and the probability of their classification into a certain group

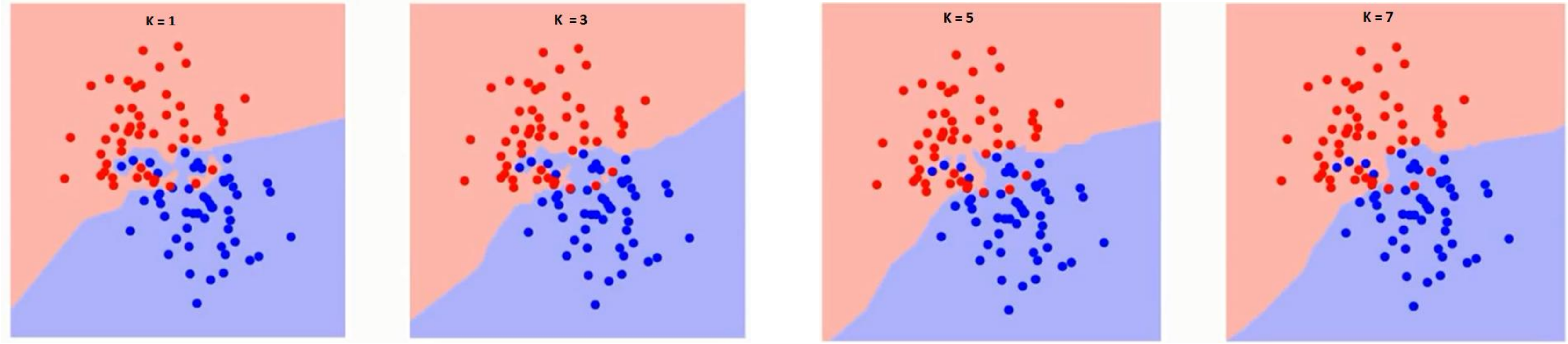
No Model is learned here

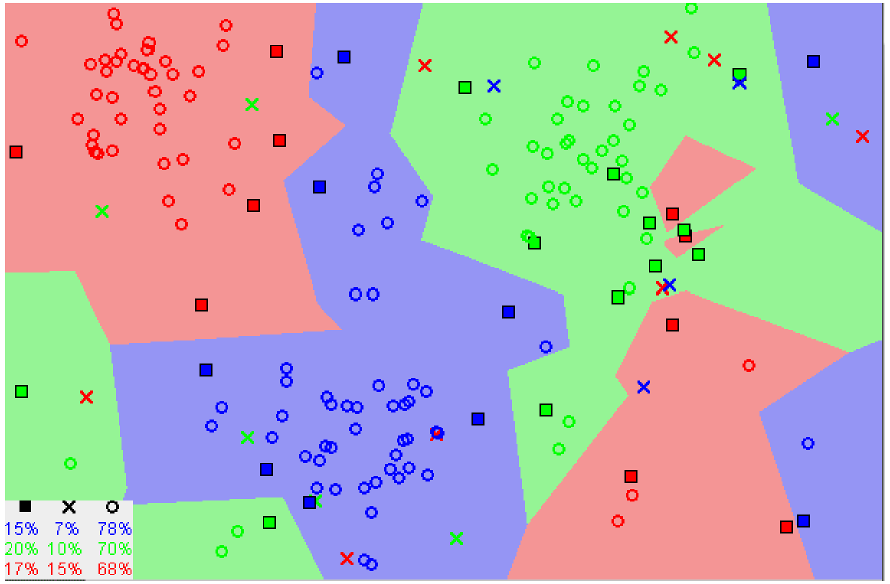
Amazon has a targeted marketing tool





How do we choose the factor K?



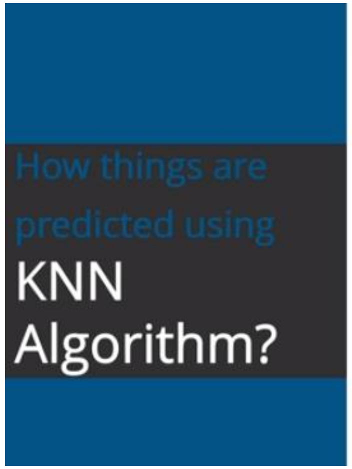


Distance functions

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

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

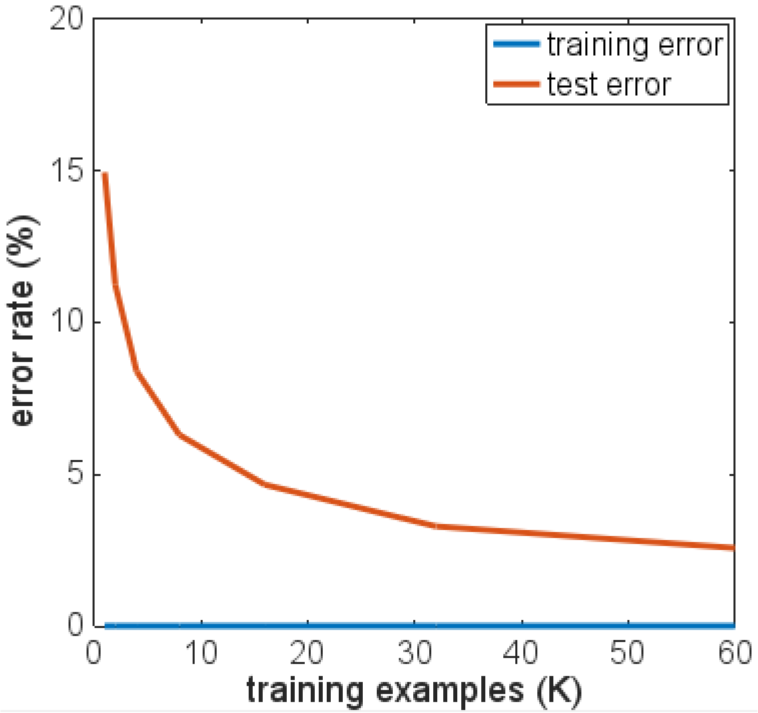
Minkowski $$\left(\sum_{i=1}^{k}\left(\left|x_i - y_i\right|\right)^q\right)^{1/q}$$



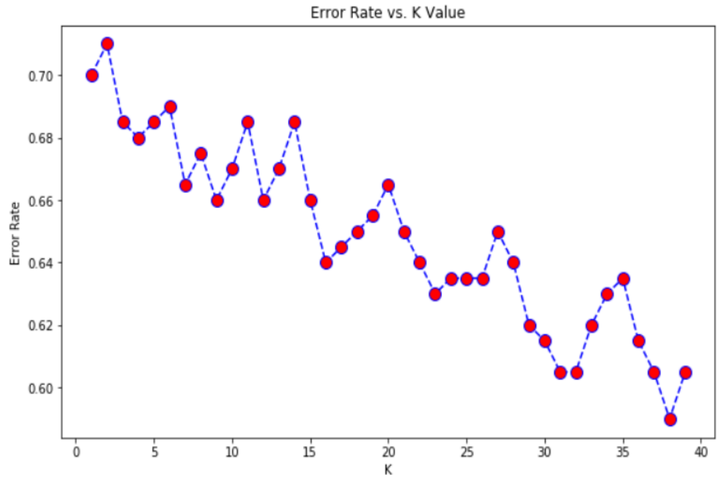
How things are predicted using KNN Algorithm?

Manhattan Distance
vs
Euclidean Distance

How to choose Number of k



```
Minimum error:- 0.59 at K = 37
```



The KNN Algorithm

```
1.Load the data
2.Initialize K to your chosen number of neighbours
3. For each example in the data
    3.1 Calculate the distance between the query example and the current example from the data.
    3.2 Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels
8. If classification, return the mode of the K labels
```

I LOVE MY NEIGHBORS
ONLY IF THEY
ARE NEAR TO ME!
imgflip.com

Advantage

- ➢ No Training Period
- ➢ Easy Implementation

Disadvantage

1.Does not work well with large dataset as calculating distances between each data instance would be very costly.
2.Does not work well with high dimensionality as this will complicate the distance calculating process to calculate distance for each dimension.
3.Sensitive to noisy and missing data
4.Feature Scaling- Data in all the dimension should be scaled (normalized and standardized) properly .