Decision Tree.

-it is a supervised ML algo that uses label data to classify a data point.

-It can be used to solve regression as well as classification problem.

-It is a graphical representation for getting all the possible solutions to problem/ decision based on given condition.

-It uses different nodes such as Root node, branch/decision node and leaf node.

-It tree like structured classifier , where internal nodes represent the features of a data set , branches represent the decision rules each leaf node represent the outcome.

On which basis DT select feature for further splitting?

sol 1) On the basis of impurity. DT select a feature with low impurity.

sol 2) INformation Gain.

How to calculate impurity?

1) Gini index - 1-p2-q2

where p is a prob of even will occure (like the movie) and q is the prob of event will not occure (not like the movie)

2) Entropy

Advantages of DT

-> Results of DT are easy to interpret.

-> DT are not affected by noisy data.

-> It can handle non linear data also.

-> IT can solve regression as well as classification problem.

Disadvantages of DT

->It is not suitable for large and high dimension datasets.

-> It is not flexible as it might lead to reconstruct DT.

->it always overfits. (IMP)

How to solve overfitting problem of DT?

--> use pruning techniques

1) max\_depth -: The maximum depth of the tree. If None, then nodes are expanded until

    all leaves are pure. Default value is None.

2) min\_sample\_leaf-:  The minimum number of samples required to be at a leaf node Default value -: 1

3) min\_sample\_split -:  The minimum number of samples required to split an internal node Default -: 2