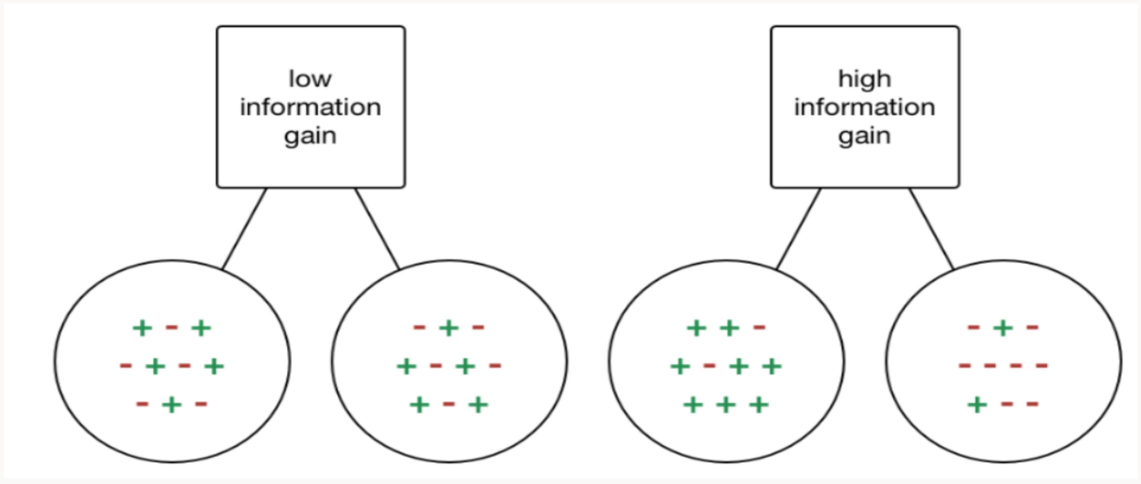
ML Report Lab07

Name: Raman Kumar
Reg NO: BL.EN.P2DSC22009
Course: Mtech Data Science

Q.1 Justify why information gain is a suitable measure to select an attribute for DT construction. Refer to the class notes and the formula below.
Which classifier would you choose for your classification problem? Provide justifications for your choice.

Entropy assists us in determining a *node's impurity*. It ranges between 0 and 1. Entropy = 0 denotes the pure node, which does not require further splitting, whereas entropy = 1 represents the most impure node.



Consider a binary decision tree where the classification is based on "Yes" or "No" to grasp impurity better.

The nodes are now divided one by one until only pure nodes or leaf nodes remain. Pure nodes indicate that we receive a clear Yes or No response and that additional splitting is not possible. This entropy is currently calculated for just one node.

What if, however, we wanted to know what the average entropy of the entire tree was, as well as which combination if chosen, would result in the best classification. The idea of Information Gain is necessary for it.

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Conclusion: Instead of calculating the entropy of a single node, Information Gain provides the decrease or drop in Entropy of the entire tree. Information gain allows you to utilize entropy to calculate how a modification to the dataset affects its purity.

KNN vs SVM: SVM wins

- SVM takes better care of outliers than KNN.
- If training data << large features, then SVM gives better results.

KNN vs Neural: KNN wins

- Neural needs large training data compared to KNN to achieve sufficient accuracy.
- Neural needs a lot of parameter tuning.

SVM vs NN: SVM wins

- SVM has a convex parameter fun.
- Whereas, NN will be stuck in local minima only
- SVM performs better when less data there, but for NN we need a lot more of data.
- But NN can do multiclass classification better in comparison to SVM. SVM needs multiple models to classify.

So. I think SVM is actually better in terms of accuracy but if we consider the time taken to model it is not that good.
So if time constraints or limits will be there, we can always go with the KNN classifier.

Q2. If a case of equal value for the criterion measure happens at a node, what would be your approach to resolve the conflict?

Even while the two nodes' entropies are equal, that doesn't always mean that their information gains are as well.

- Choose the node between those two where the disagreement is developing that has a greater IG. The opposite is also true; If both have the same IG, then the entropy may not be equal, therefore we may choose the one with less entropy to go on with.

Another approach is to examine the entropy values of the nodes underneath these two conflicting nodes. Choose the node with the lowest entropy of the nodes below it.