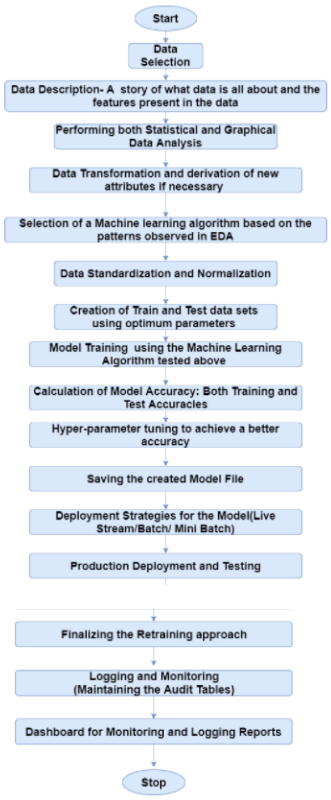
Application Flow

Decision Tree is one of the most fundamental algorithms for classification and regression in the Machine Learning world.

But before proceeding with the algorithm, let's first discuss the lifecycle of any machine learning model. This diagram explains the creation of a Machine Learning model from scratch and then taking the same model further with hyperparameter tuning to increase its accuracy, deciding the deployment strategies for that model and once deployed setting up the logging and monitoring frameworks to generate reports and dashboards based on the client requirements. A typical lifecycle diagram for a machine learning model looks like:



Decision Tree

Decision tree algorithm is one of the most versatile algorithms in machine learning which can perform both classification and regression analysis. It is very powerful and works great with complex datasets. Apart from that, it is very easy to understand and read. That makes it more popular to use. When coupled with ensemble techniques – which we will learn very soon- it performs even better. As the name suggests, this algorithm works by dividing the whole dataset into a tree-like structure based on some rules and conditions and then gives prediction based on those conditions. Let's understand the approach to decision tree with a basic scenario. Suppose it's Friday night and you are not able to decide if you should go out or stay at home. Let the decision tree decide it for you.

Although we may or may not use the decision tree for such decisions, this was a basic example to help you understand how a decision tree makes a decision. So how did it work?

So how did it work?

- It selects a root node based on a given condition, e.g. our root node was chosen as time >10 pm.
- Then, the root node was split into child notes based on the given condition. The right child node in the above figure fulfilled the condition, so no more questions were asked.
- The left child node didn't fulfil the condition, so again it was split based on a new condition.
- This process continues till all the conditions are met or if you have predefined the depth of your tree, e.g. the depth of our tree is 3, and it reached there when all the conditions were exhausted.

Tree Pruning

Tree pruning is the method of trimming down a full tree (obtained through the above process) to reduce the complexity and variance in the data. Just as we regularised linear regression, we can also regularise the decision tree model by adding a new term.

$$\sum_{m=1}^{|T|} \sum_{i:\ x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

Where, T is the subtree which is a subset of the full tree T0 And α is the non-negative tuning parameter which penalises the MSE with an increase in tree length. By using cross-validation, such values of α and T are selected for which our model gives the lowest test error rate. This is how the decision tree regression model works. Let's now see the working algorithm of doing classification using a decision tree. Greedy Algorithm As per Hands-on machine learning book "greedy algorithm greedily searches for an optimum split at the top level, then repeats the process at each level. It does not check whether or not the split will lead to the lowest possible impurity several levels down. A greedy algorithm often produces a reasonably good solution, but it is not guaranteed to be the optimal solution."

Post-pruning

Post-pruning, also known as backward pruning, is the process where the decision tree is generated first and then the non-significant branches are removed. Cross-validation set of data is used to check the effect of pruning and tests whether expanding a node will make an improvement or not. If any improvement is there then we continue by expanding that node else if there is reduction in accuracy then the node not be expanded and should be converted in a leaf node.

Pre-pruning

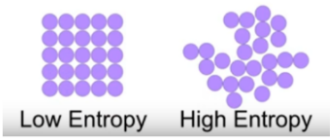
Pre-pruning, also known as forward pruning, stops the non-significant branches from generating. It uses a condition to decide when should it terminate splitting of some of the branches prematurely as the tree is generated.

Classification Trees

Regression trees are used for quantitative data. In the case of qualitative data or categorical data, we use classification trees. In regression trees, we split the nodes based on RSS criteria, but in classification, it is done using classification error rate, Gini impurity and entropy. Let's understand these terms in detail.

Entropy

Entropy is the measure of randomness in the data. In other words, it gives the impurity present in the dataset.



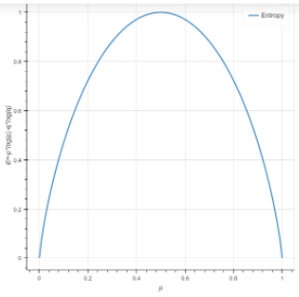
When we split our nodes into two regions and put different observations in both the regions, the main goal is to reduce the entropy i.e. reduce the randomness in the region and divide our data cleanly than it was in the previous node. If splitting the node doesn't lead into entropy reduction, we try to split based on a different condition, or we stop. A region is clean (low entropy) when it contains data with the same labels and random if there is a mixture of labels present (high entropy). Let's suppose there are 'm' observations and we need to classify them into categories 1 and 2. Let's say that category 1 has 'n' observations and category 2 has 'm-n' observations.

p= n/m and q = m-n/m = 1-p

then, entropy for the given set is:

entropy = Σ -p * log2(p)

When all the observations belong to category 1, then p = 1 and all observations belong to category 2, then p =0, int both cases E =0, as there is no randomness in the categories. If half of the observations are in category 1 and another half in category 2, then p =1/2 and q =1/2, and the entropy is maximum, E =1.



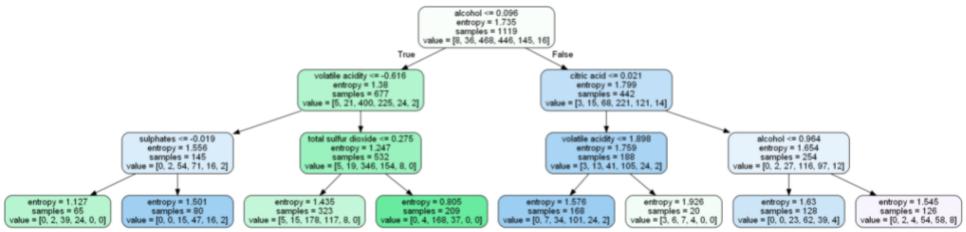
Information Gain

Information gain calculates the decrease in entropy after splitting a node. It is the difference between entropies before and after the split. The more the information gain, the more entropy is removed.

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

Where, T is the parent node before split and X is the split node from T.

A tree which is splitted on basis of entropy and information gain value looks like:

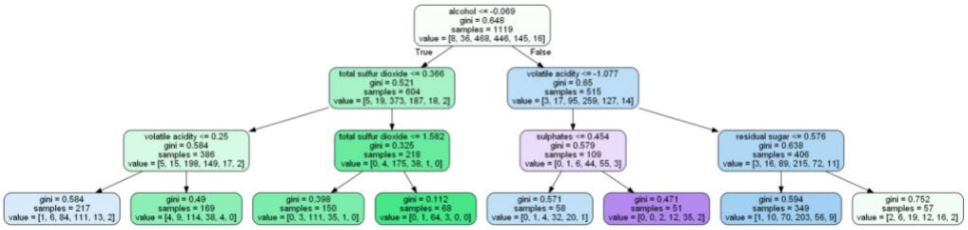


Ginni Impurity

According to wikipedia, 'Gini impurity is a measure of how often a randomly chosen element from the set would be incorrectly labelled if it was randomly labelled according to the distribution of labels in the subset.' It is calculated by multiplying the probability that a given observation is classified into the correct class and sum of all the probabilities when that particular observation is classified into the wrong class.

Ginni impurity value lies between 0 and 1, 0 being no impurity and 1 denoting random distribution. The node for which the Ginni impurity is least is selected as the root node to split.

A tree which is splitted on basis of ginni impurity value looks like:



Entropy and Information Gain

X feature

- Total ones in 'X' = 3
- Count of Label 'A' when X equal to 1 => 2
- Count of Label 'B' when X equal to 1 => 1

- Total zeros in 'X' = 1
- Count of Label 'A' when X equal to 0 => 0
- Count of Label 'B' when X equal to 0 => 1

entropy = Σ -p * log2(p)

Y feature

- Total ones in 'Y' = 2
- Count of Label 'A' when Y equal to 1 => 2
- Count of Label 'B' when Y equal to 1 => 0

- Total zeros in 'Y' = 2
- Count of Label 'A' when Y equal to 0 => 0
- Count of Label 'B' when Y equal to 0 => 2

Z feature

- Total ones in 'Z' = 2
- Count of Label 'A' when Z equal to 1 => 1
- Count of Label 'B' when Z equal to 1 => 1

- Total zeros in 'Z' = 2
- Count of Label 'A' when Z equal to 0 => 1
- Count of Label 'B' when Z equal to 0 => 1

Finding Information Gain using entropy

Inf.Gain = 1 - Σ sv/s * E

- s = Total records
- sv = Category counts (1/0)
- E = entropy

- entropy_X_1 = 0.9182958340544896
- entropy_X_0 = nan
- entropy_Y_1 = nan
- entropy_Y_0 = nan
- entropy_Z_1 = 1.0
- entropy_Z_0 = 1.0

Gini Indexing

Gini Impurity = 1 - Σ p2

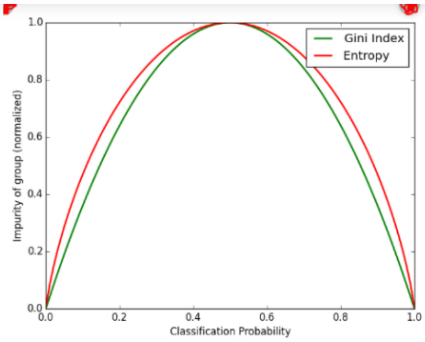
Different Algorithms for Decision Tree

- ID3 (Iterative Dichotomiser) : It is one of the algorithms used to construct decision tree for classification. It uses Information gain as the criteria for finding the root nodes and splitting them. It only accepts categorical attributes.
- C4.5 : It is an extension of ID3 algorithm, and better than ID3 as it deals both continuous and discreet values.It is also used for classfication purposes.

- Classfication and Regression Algorithm(CART) : It is the most popular algorithm used for constructing decison trees. It uses ginni impurity as the default calculation for selecting root nodes, however one can use "entropy" for criteria as well. This algorithm works on both regression as well as classfication problems. We will use this algorithm in our pyhton implementation.

Entropy and Ginni impurity can be used reversibly. It doesn't affects the result much. Although, ginni is easier to compute than entropy, since entropy has a log term calculation. That's why CART algorithm uses ginni as the default algorithm.

If we plot ginni vs entropy graph, we can see there is not much difference between them:



Advantages of Decision Tree:

- It can be used for both Regression and Classification problems.
- Decision Trees are very easy to grasp as the rules of splitting is clearly mentioned.
- Complex decision tree models are very simple when visualized. It can be understood just by visualising.
- Scaling and normalization are not needed.

Disadvantages of Decision Tree:

- A small change in data can cause instability in the model because of the greedy approach.
- Probability of overfitting is very high for Decision Trees.
- It takes more time to train a decision tree model than other classification algorithms.

Implementation in Python

we will use Sklearn module to implement decision tree algorithm. Sklearn uses CART (classification and Regression trees) algorithm and by default it uses Gini impurity as a criteria to split the nodes.

What are hyper parameters?

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
                       max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=None, splitter='best')
```

We can see above the decision tree classifier algorithm takes all those parameters which are also known as hyperparameters.

Let's see the most important ones of the parameters(as per sklearn documentation) :

Parameters

- criterion : string, optional (default="gini")

  ```
  The function to measure the quality of a split. Supported criteria are
  "gini" for the Gini impurity and "entropy" for the information gain.
  ```

- splitter : string, optional (default="best") The strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.
- max_depth : int or None, optional (default=None) The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.
- min_samples_split : int, float, optional (default=2) The minimum number of samples required to split an internal node:
  - If int, then consider `min_samples_split` as the minimum number.
  - If float, then `min_samples_split` is a fraction and `ceil(min_samples_split * n_samples)` are the minimum number of samples for each split.

  .. versionchanged:: 0.18 Added float values for fractions.
- min_samples_leaf : int, float, optional (default=1) The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least `min_samples_leaf` training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.
  - If int, then consider `min_samples_leaf` as the minimum number.
  - If float, then `min_samples_leaf` is a fraction and `ceil(min_samples_leaf * n_samples)` are the minimum number of samples for each node.
- max_features : int, float, string or None, optional (default=None) The number of features to consider when looking for the best split: