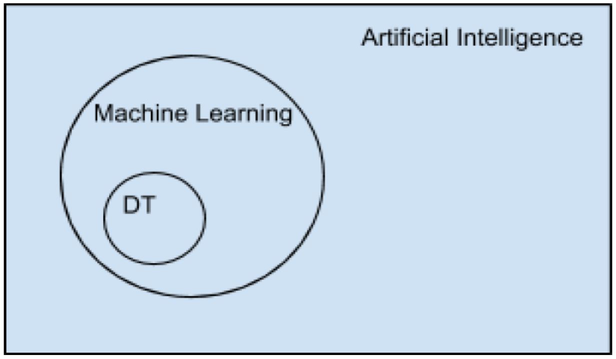
Decision Tree White Paper

Pre-requisites

1. What is Machine Learning?

Machine learning is a subfield of Artificial Intelligence in which Machine is taught to make decisions. Examples of Machine Learning: Email SPAM/HAM filtering, Facebook Image tagging, Google News etc.



2. Types of Machine Learning Technique?

There are 2 ways in which a machine can be taught to make decisions:

a. **Supervised learning:** Machine is fed with examples of labelled scenarios. Using these examples machine learns which decision it needs to take if it encounters similar scenarios

Eg: You give a child lots of images of Car and child learns that the car has 4 wheels, windows etc. Next time the child would recognize those images with above properties and would able to classify them as car

b. **Unsupervised learning:** Machine is fed with examples of unlabeled scenarios. Using these examples machine deduces properties associated with those scenarios to make decision

Eg: You give child lots of images of Car and Bike, but don't tell him that which one is car/bike. The child learns that images having 4 wheels are similar in nature while images having 2 wheels are similar. And although he doesn't know what exactly is car and bike, but can sort them separate.

3. What is Regression and Classification?

Regression: It is a supervised ML technique. It is used in the scenarios in which a continuous variable is predicted.

Eg: You have size of a house and its price. Now the algorithm will try to identify the relation between house size and it's pricing and based on the identified relation between the two it will predict that what should be the most probable price of a house with "x" sq. feet of size

Classification: It is an unsupervised ML technique. It is used in the scenarios in which a categorical variable is predicted.

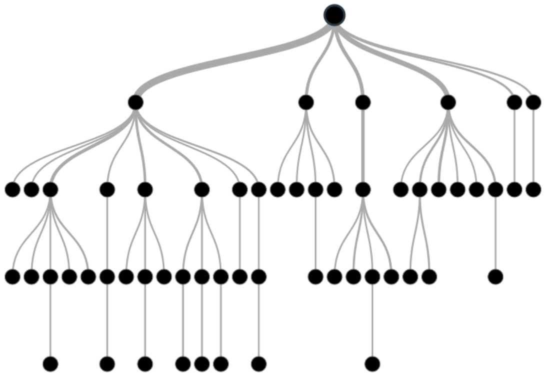
Eg: You have Marks of students and Pass/Fail indicator against it. Now the algorithm identifies that if the marks is above "X" then the student is "Pass"

Decision Tree

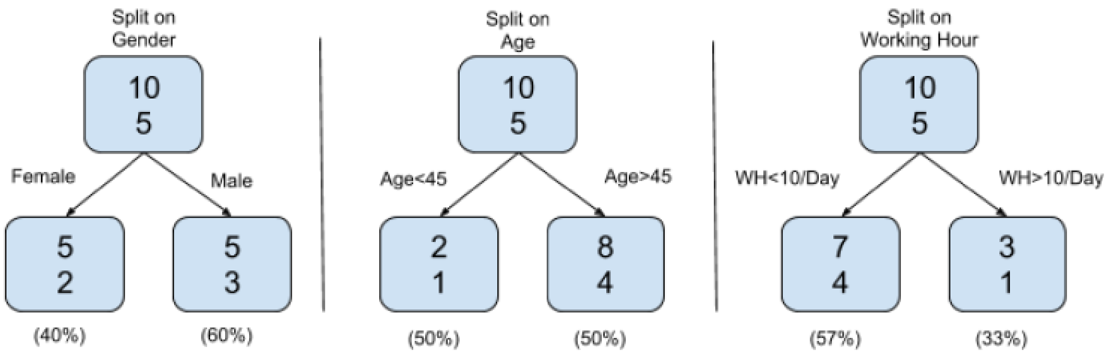
1. What is Decision Tree? How does it help in better decision making?
2. Types of decision tree
3. How does tree split into branches?
4. Case Study-Using Decision tree to Predict Heart disease in a patient
5. Random Forest-Extended Application and robust algorithm for Predictions
6. Model Comparison- Decision Tree VS Random Forest ,Model Evaluation criteria- Confusion Matrix, Recall/Precision, Lift Curve, concordance

Definition:

Decision tree is a supervised Machine Learning algorithm used for both Regression and Classification. Decision tree is created by splitting the dataset into different homogeneous subsets



Eg: There are 10 people with 3 variables - Age, Gender and Working Hour. 5 out of these 10 loves cooking. We need to segregate people who love cooking based on the input variable



As you can see in the above diagram that splitting the data set on Gender gives the best set to homogeneous groups giving 60% of the population who loves cooking and is the most significant variable.

Advantages and Applications:

Few of the advantages of decision tree are:

1. It is easy to understand as the output comes out in the form of tree and it's easy for users to interpret what the algorithm did
2. It is useful in data exploration as it gives the splitting based on significance of variables
3. It is not influenced by the outlier/Null values and hence requires less data cleaning
4. It can handle both continuous and categorical variables
5. Does not require any underlying assumptions in data. Works with both linearly and nonlinearly related variables.

Applications: Decision tree can be used for almost all the prediction related problems. A decision tree gives room to researchers to explore and devise a new approach on top of decision tree to give better predictions

Types of Decision Tree:

a.) Decision Tree for Classification Problem: When the task is to classify the data into different categories then, the decision tree used is Classification decision tree.

For ex : UHG Bank Pvt Ltd wants to predict whether a loan application should be granted loan or not based on applicant's demographics and credit history. Here, target/dependent variable is **Loan Approved** having 'Yes'or 'No' as two different categories.

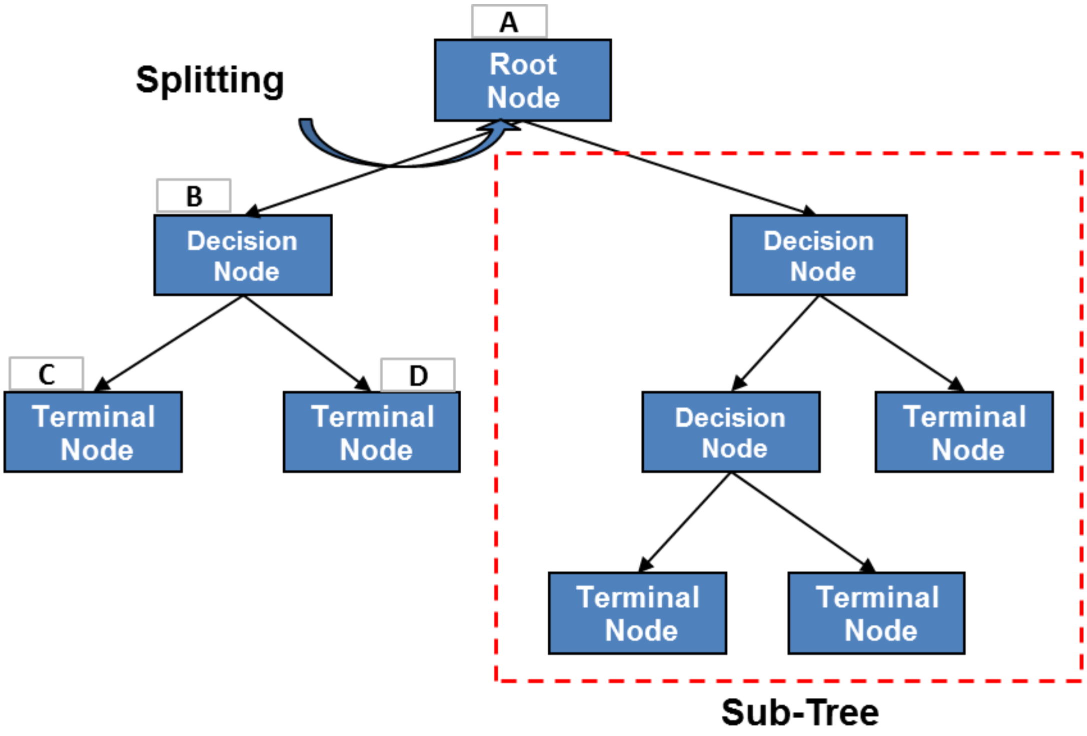
b.) Decision Tree for Regression Problem: When prediction is to be done for a continuous dependent variable then, the decision tree used is Regression decision tree.

For ex : UHG Housing Development Authority Ltd. wants to predict the housing price based on locality and house attributes (house size, number of rooms etc.).Here dependent variable is the House Price.

Different types of tree based algorithms are: CART, CHAID, Random Forest, and Boosting.

Common terminologies used in Decision tree:

1. **Root Node**: It represents entire population or sample which gets divided into two or more homogeneous sets (branches or sub-trees). Ex- Node **A**
2. **Child and Parent Node**: A node, which is divided into sub-nodes is called parent node of sub-nodes whereas sub-nodes are the child of parent node. **B** is parent node for child node **C** and **D**
3. **Branch / Sub-Tree:** A sub-section of entire tree (Highlighted in Red Border).



4. **Decision Node**: A sub-node which splits into further sub-nodes. Ex- Node **B**

5. **Leaf/ Terminal Node**: These nodes do not split further. There are 5 leaf/terminal nodes in above example.

6. **Splitting:** It is a process of dividing a node into two or more sub-nodes. Most widely used splitting criterions are Gini-Index and Information Gain (will be explained in next section).

7. **Pruning**: Process of removing the sub-nodes of a decision node. It is opposite of splitting process. This will help in tuning the tree to get less biased decision tree.

How does tree split into branches?

Splitting is the most important part while creating over a dataset as it heavily affects the accuracy of the decision tree. There are multiple algorithms which decision tree can use while splitting the dataset.

Gini Index: Gini Index says that if we select 2 items randomly from the dataset created after splitting then they must belong to the same class.

Formula for calculating GINI: **P^2 + Q^2** (Where P is probability of Success and Q is probability of failure)

Taking the above “Love for Cooking” Example:

1. If split is made on Gender:

a. GINI for Sub-node female: (0.4)(0.4) + (0.6)(0.6) = 0.52

b. GINI for Sub-node male: (0.6)(0.6) + (0.4)(0.4) = 0.52

c. Weighted GINI for Gender split: (5/10)*0.52 + (5/10)*0.52 = 0.52

2) If split is made on Age:

a. GINI for Sub-node Age>45: (0.5)(0.5) + (0.5)(0.5) = 0.5
b. GINI for Sub-node Age<45: (0.5)(0.5) + (0.5)(0.5) = 0.5
c. Weighted GINI for Age split: (2/10)*0.5 + (8/10)*0.5 = 0.5

3) If split is made on working hour:

a. GINI for Sub-node WH>10: (0.33)(0.33) + (0.67)(0.67) = 0.7789
b. GINI for Sub-node WH<=10: (0.57)(0.57) + (0.43)(0.43) = 0.5089
c. Weighted GINI for WH split: (7/10)*0.7789 + (3/10)*0.508 = 0.69

As we can see that GINI index is highest if split is made on working hour. Hence the root node will split on working hour.

Note: GINI will be 1 if data is 100% homogeneous. Resulting in pure terminal node.

Similarly subsequent splits are made and tree is created

Information Gain (IG): Higher the information gain more is the significance of the variable.

IG = 1 - Entropy

Entropy = - (P) Log2P - (Q) Log2Q

Classification Case study: Heart Disease Prediction

Objective: To Predict the presence of heart disease in a patient.

Predictors: There are 13 predictors including Age, Sex, Chol (a cholesterol measurement), Chest pain, blood sugar, resting blood pressure; and other heart and lung function measurements.

Dependent Variable: have 2 distinct values, where "Yes" indicates the presence of heart disease

Model Training and Validation: Dataset has information for 454 patients.

a.) Training Set: Information for 70% random patients is used to create decision tree rules. We call this data as Training dataset as Machine uses this data to learn the patterns.

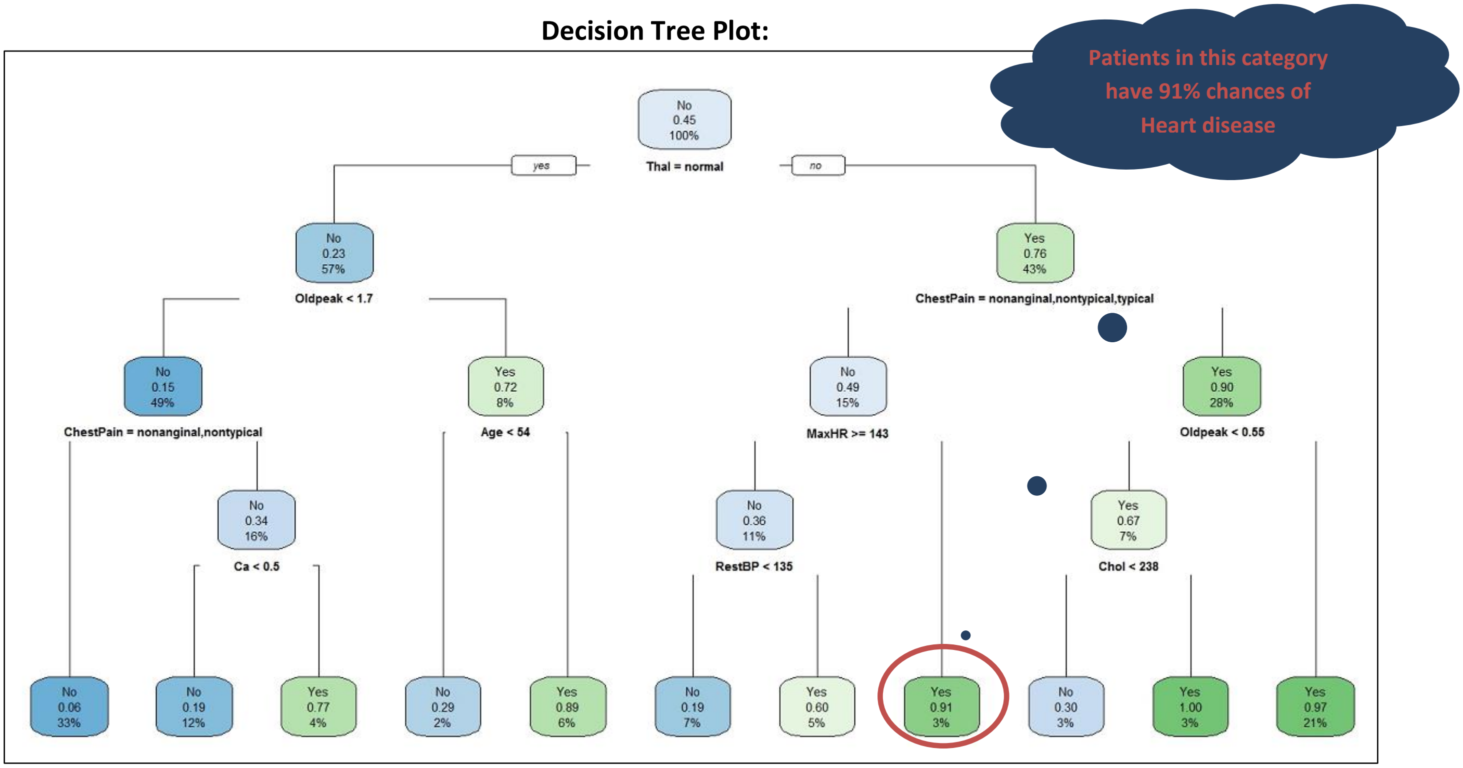
b.) Test Set: Remaining 30% is used to validate the decision tree accuracy. This data is called as Test dataset.

Event Rate:

Train -> 45% Out of 100, there are 45 patients with Heart disease.

Test -> 43% Out of 100, there are 43 patients with Heart disease. It also means that if we randomly pick 100 patients then around 43 patients will have heart disease

Decision Tree Plot:



Patients in this category have 91% chances of Heart disease

Interpretation: A Patient with

a.) **Thalassemia (thal) other than “normal”**

And

b.) **“**typical angina” or “atypical angina” or “non-angina” type of **chest pain (ChestPain)**

And

c.) **Maximum Heart Rate (MaxHr) less than 143**

Has **91%** chance of having heart disease

Decision Tree accuracy on Test Dataset:

Since, decision tree gives probability value(0-1) for a patient having heart disease we have to assign a cutoff value in order to determine whether this patient has heart disease or not.

For Ex- Patient ‘X’ has probability 0.67 of being heart patient. If cutoff value is 0.7 then, this patient is not considered as heart patient however; with cut off value of 0.5 we classify this patient with heart disease.

Cut-off Value considered - 0.50

Confusion Matrix		Predicted	
		No	Yes
Actual	No	**63**	**15**
	Yes	**8**	**51**

Accuracy= (63+51)/(63+8+15+51) = 83.2%

Which means 83.2% times, decision tree correctly predicts whether a patient has heart disease or not.

Recall = (51)/(51+8)=86.4% , It means, out of 100 patients with heart disease, our model is able to identify 86.4% heart disease patients.

Precision= 51/(15+51)=77.2%, Which means out of 100 patients ,which are classified as “Yes” for having heart disease by decision tree, 77.2% are correctly classified.

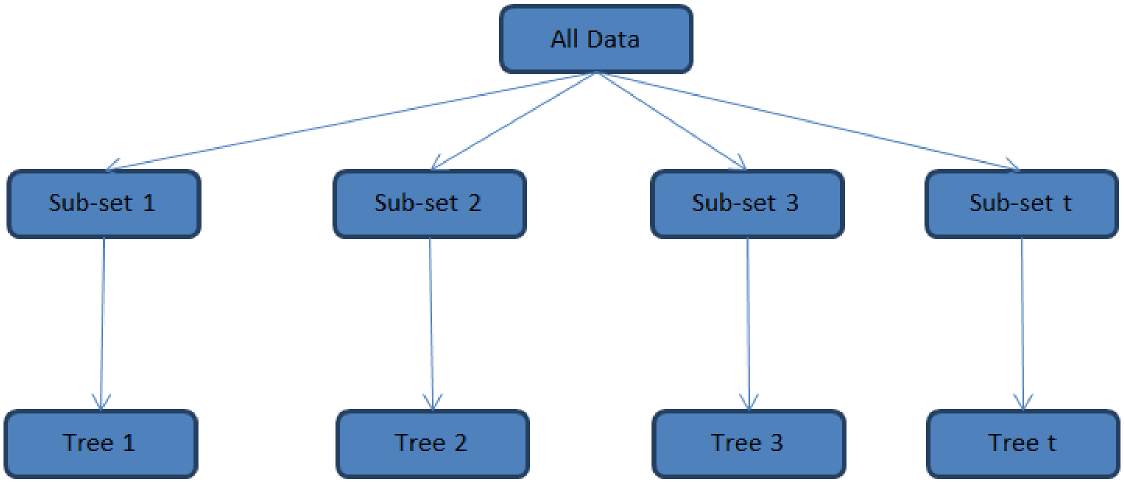
Random Forest

Random forest is a tree-based machine learning algorithm which involves building of several Decision trees and combines output from all the trees to improve decision making. It can be used for classification and regression. It takes votes from all the trees created and count of votes will be the predicted probability in-case of classification and averages of prediction in-case of regression.

For ex: - You and your 10 friends went together to buy a Laptop for you. You picked HP-Z01 Ideapad and asked your friend's opinion on whether to buy it or not. Eight (8) of them said, you should buy it. Since, the majority is in favor, you decide to buy it. This is how random forest works; instead of taking output from a single decision tree it takes votes from all the trees.

It uses the concept of bagging:

1) Let's say that the dataset has N rows and M columns
2) It will take n rows such that n<N and m columns such that m<M to create the decision tree. If you want to create 1000 tree, then the algorithm will repeat the step to create 1000 trees.
3) Each of these trees is fully grown
4) While validating the model, outputs from all the 1000 trees will be considered.



Heart disease prediction using Random Forest:

Accuracy on Test Dataset:

Probability Cut-off Value: 0.50

Confusion Matrix		Predicted	
		No	Yes
Actual	No	**70**	**08**
	Yes	**02**	**57**

Accuracy= (70+57)/ (70+08+02+57) = 92.7%

This means 92.7% times, random forest correctly predicts whether a patient has heart disease or not.

Recall = (57)/(57+02)=96.6% , It means, out of 100 patients with heart disease, our model is able to identify 96.6% heart disease patients.

Precision= 57/(57+08)=87.6%, Which means out of 100 patients ,which are classified as "Yes" for having heart disease by random forest model, 87.6% are correctly classified.

Scenario where above machine algorithms can be used:

1. Identification: Forex Bank Ltd wants to automate the loan process where a loan application submitted by the applicant would be approved or rejected at the time of submission to the bank portal. Using the historical applications data (Loan_approved, Sex, Salary, Job_type, past_loan etc.), a predictive model can be built to identify those application where loan should be granted to those who are more likely to pay the debt. Since, predictive model provides probability value a cutoff value is decided which gives high precision accuracy.

2. Prioritization: Mr. Austin, Who is Senior Manager in Preventive Health Management wing of UHC, wants to conduct a free of cost preventive health check-up camp for those members who are likely to have one of the chronic diseases in future. Having a budget constraint, camp can accommodate only 1000 patients. Early diagnosis of a chronic disease can save 1000$ on an average. But, what should Mr. Austin do in order to ensure the maximum throughput from the camp?

A Predictive model was built which gave a risk score to members based on their demographics and historical medical records. Out of thousands of members to be targeted for this campaign, only 1000 that are most likely to have a chronic disease were asked to join the health check-up camp.