DECISION TREES I

PROJECT VERSION

William Wang

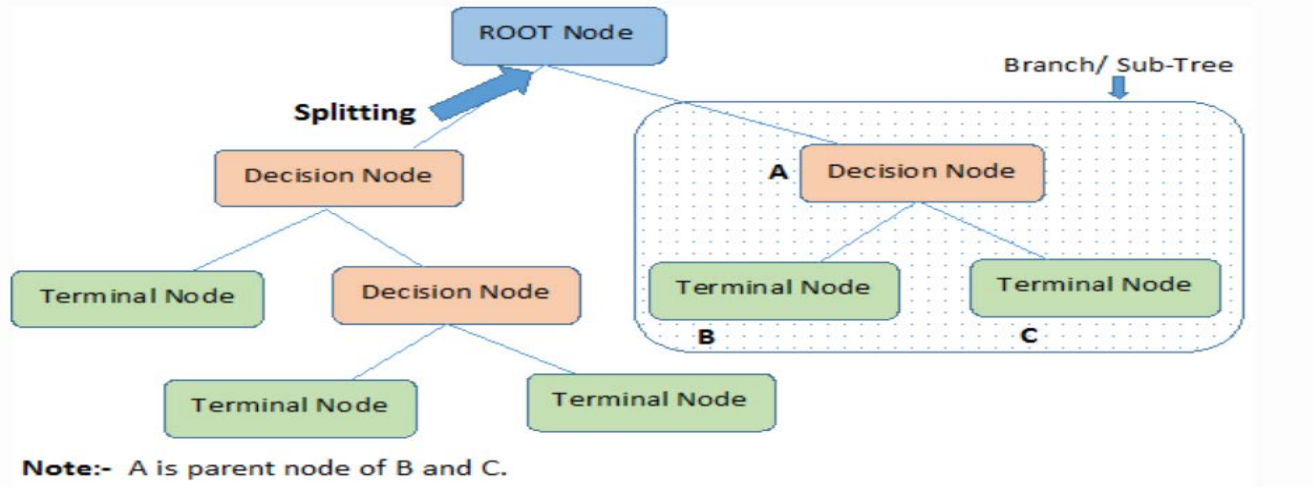
from ch52

The Concepts of Decision Trees

https://www.rdocumentation.org/

Decision Tree

- 決策樹是一種通過對歷史數據，進行估算實現對新數據進行分類和預測的算法。簡單來說決策樹算法就是通過對已有明確結果的歷史數據進行分析，尋找數據中的特徵。並以此為依據對新產生的數據結果進行預測與推估。
- 決策樹由3個主要部分組成，分別為決策節點，分支，和葉子節點。其中決策樹最頂部的決策節點是**根決策節點**。每一個分支都有一個新的決策節點。決策節點下面是葉子節點。每個決策節點表示一個待分類的數據類別或屬性，每個葉子節點表示一種**結果**。整個決策的過程從根決策節點開始，從上到下。根據數據的分類在每個決策節點給出不同的結果。

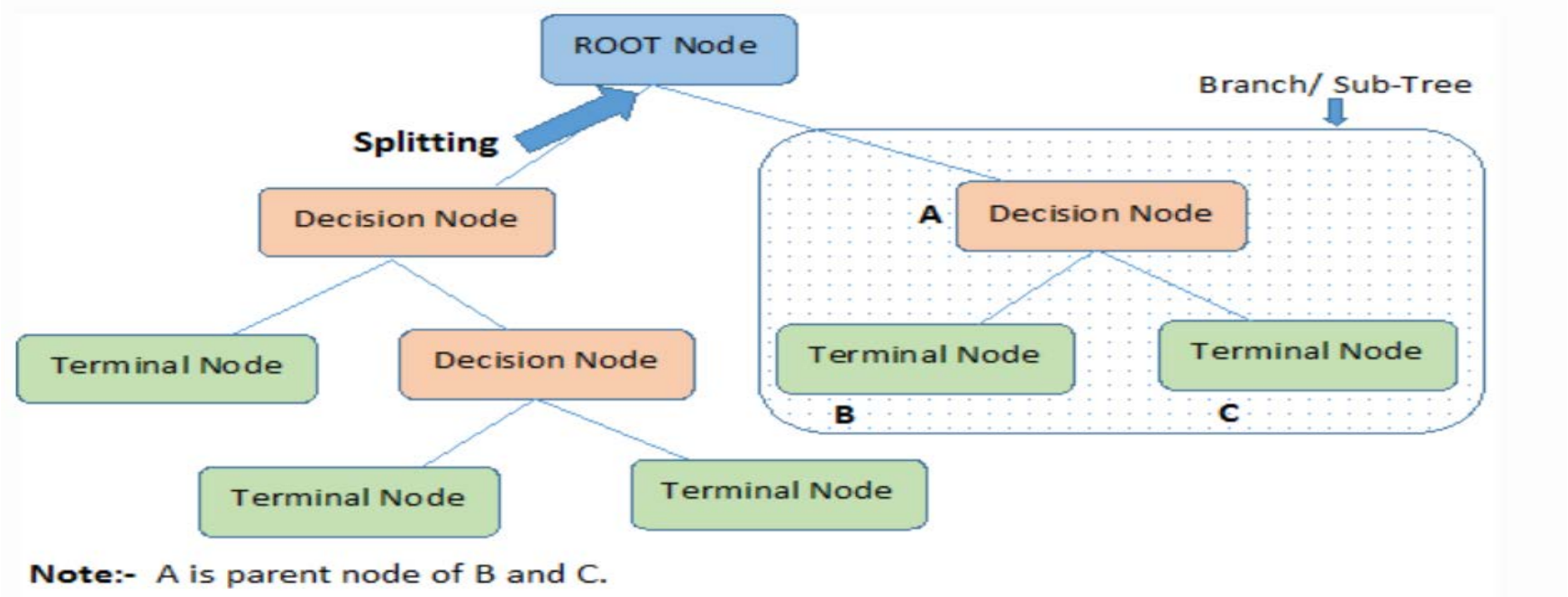


Important Terminology related to Decision Trees

- Let's look at the basic terminologies used with Decision trees:
  - **Root Node:** It represents entire population or sample and this further gets divided into two or more homogeneous sets.
  - **Splitting:** It is a process of dividing a node into two or more sub-nodes.
  - **Decision Node:** When a sub-node splits into further sub-nodes, then it is called decision node.
  - **Leaf/Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- That purity of the node increases with respect to the **target variable** (p.s. **goal variable/ property**). Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.
- The goal is to create a model that predicts the value of a target variable by decision rules inferred from the data features.

Important Terminology related to Decision Trees

- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say opposite process of splitting.
  - **Branch / Sub-Tree:** A sub section of entire tree is called branch or sub-tree.
  - **Parent and Child Node:** A node, which is divided into sub-nodes is called parent node of sub-nodes where as sub-nodes are the child of parent node.
- These are the terms commonly used for decision trees. As we know that every algorithm has advantages and disadvantages, below are the important factors which one should know.



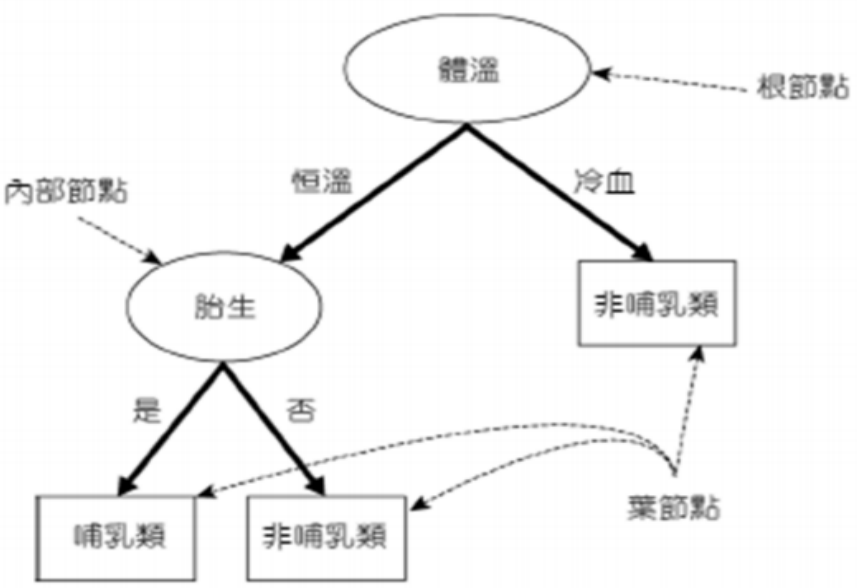
Decision Tree - retalk

決策樹

- 樹包含三個節點：
  - **根節點**：沒有任何進入的邊，而且有0 個或是輸出的邊
  - **內部節點**：每個節點都有一個輸入的邊，以及二個或多個輸出的邊
  - **葉節點**或是**終端節點**：每個節點都有一個輸入的邊，但沒有輸出的邊

每個葉節點都是一個類別標記

- 決策樹範例
  - 哺乳類動物分類的問題



生產管理上的應用

- 決策樹運用於LED產業製程異常分析
- 生產管理人員會評估在不同的生產管理方案
- 天氣或空氣品質預測會推估

Decision trees need historical outcome data, which are then trained to show the best decision path to take to achieve specified outcomes. The outcome variables are also known as the *target variables.*

生產管理上的應用

範例－決策樹問題（續）：報酬表

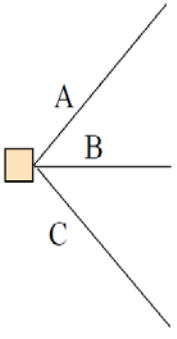
管理人員也會評估在不同的機率水準下選擇三個方案（A、B、C）所獲得的利潤，這些利潤（以千元計）如下表所示：

產量需求高中低的三個方案

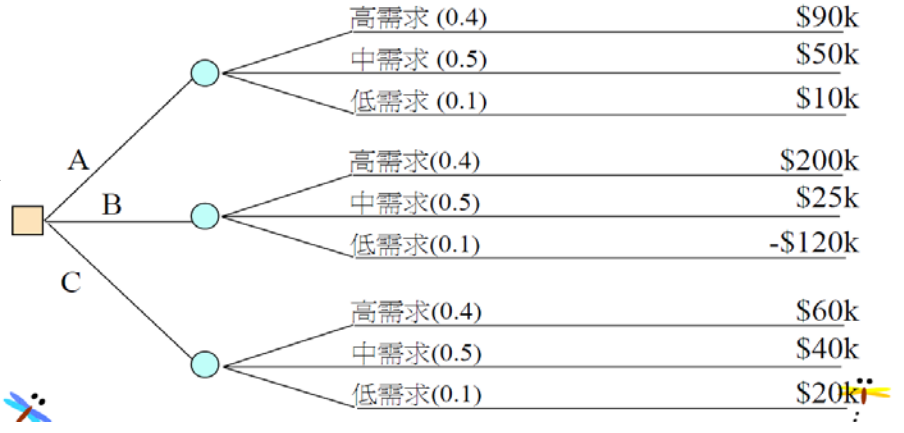
註:0.1, 0.5, 0.4為機率

	0.1	0.5	0.4
	低	中	高
A	10	50	90
B	-120	25	200
C	20	40	60

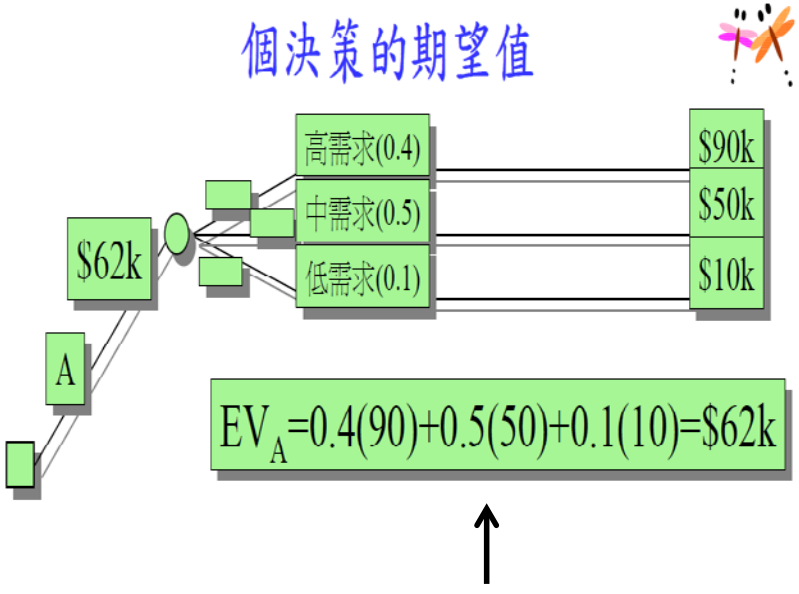
範例－決策樹問題（續）：步驟1. 我們開始畫出三個決策



範例－決策樹問題（續）：步驟2. 增加我們可能的自然狀態、機率及報酬

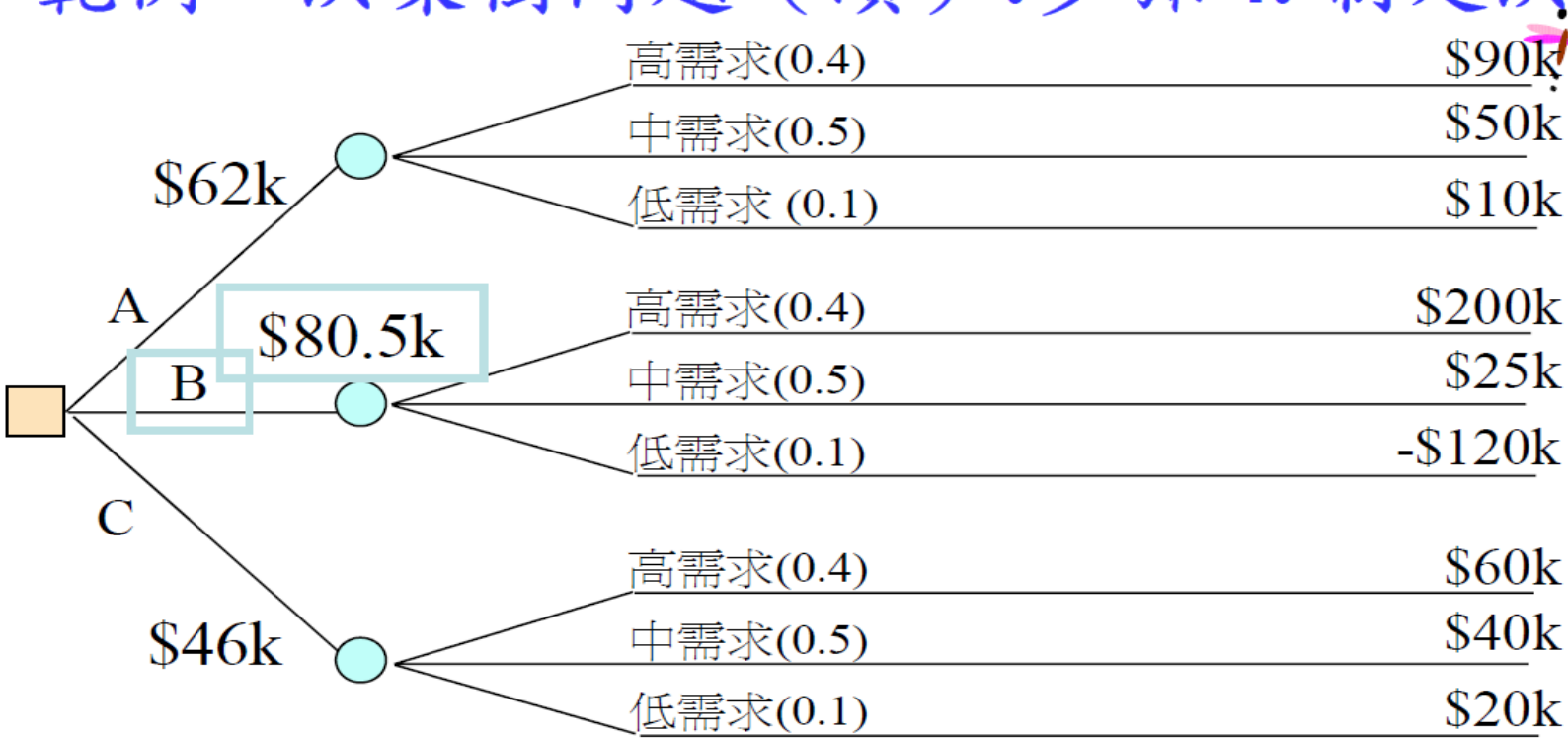


範例－決策樹問題（續）：步驟3. 確認每個決策的期望值



生產管理上的應用

範例－決策樹問題（續）:步驟 4. 制定決策



方案 **B** 產生最大的期望利潤，因此我們選擇 **B** 或是建構新設備。

Feature Selection for Decision Tree

- you need to divide given columns into two types of variables dependent(or target variable) and independent variable(or feature variables).
- In R,
  - variables dependent(or target variable):油耗
  - independent variable(or feature variables):價格, 產地, 可用性, 型態, 車重, 發動機功率, 淨馬力

Decision tree and Machine learning

- In the AI area, machine learning algorithms have become a hotspot in research and applications.
- At present, the two hottest algorithms for machine learning are neural network algorithms (CNN, RNN, LSTM, etc.) and tree algorithms (random forest, GBDT, XGBoost, etc.).
- The basis of tree algorithms is decision trees. Decision trees are widely used in statistics, data exploration, and machine learning because of their easy-to-understand, easy-to-build, and fast features. Therefore, learning the decision tree is an essential step in the road of machine learning.

Two types of Decision Trees

- ***Decision Trees can be categorized in two types based on its processes:***
  - ***Classification tree*** *analysis is when the predicted outcome is the class to which the data belongs.*
  - ***Regression tree*** *analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital).- ex.* CART Algorithm

- 有別於「分類」樹(classification tree)是用來找尋「最能區分標籤資料類別」的一系列變數，「迴歸」樹(regression tree)則是用來找尋「最能區分**目標連續變數**相近度」的一系列變數。迴歸樹投入變數可以式任何資料型態(與分類樹一樣)，唯一差別是**迴歸樹的目標變數(target variable/ goal property)是連續型變數**。

Two types of Decision Trees

- Decision tree is a type of supervised learning algorithm (having a pre-defined target variable) that is mostly used in classification problems. It works for both categorical and continuous input and output variables. In this technique, we split the population or sample into two or more homogeneous sets (or sub-populations) based on most significant splitter / differentiator in input variables.
- Depending on the type of a target variable to determine a decision tree is **Categorical Variable Decision Tree or Continuous Variable Decision Tree.**

Reminder: target variable/ goal variable/
target property/ goal property

- 對於決策樹來說，所有節點的分類或者回歸目標都要在根節點已經定義好了。如果決策樹的目標變數是離散的（序數型或者是列名型變數），則稱它為分類樹（Classification Tree）；如果目標變數是連續的（區間型變數），則稱它為回歸樹（Regression Tree）。
- 決策樹自上而下的循環分支學習（Recursive Regression）採用了貪心演算法。每個分支節點只關心自己的目標函數。具體來說，給定一個分支節點，以及落在該節點上對應樣本的觀測（包含自變數與目標變數），選擇某個（一次選擇一個變數的方法很常見）或某些預測變數，也許會經過一步對變數的離散化（對於連續自變數而言），經過搜索不同形式的分叉函數且得到一個最優解（最優的含義是特定準則下收益最高或損失最小）。
- ID3由Ross Quinlan在1986年提出。ID3決策樹可以有多個分支，但是不能處理特徵值為連續的情況。

常見的決策樹演算法比較

演算法	資料屬性	分割規則	修剪樹規則
ID3	離散型	Entropy, Gain Ratio	Predicted Error Rate
C4.5	離散型	Gain Ratio	Predicted Error Rate
CHAID	離散型	Chi-Square Test	No Pruning
CART	離散與連續型	Gini Index	Entire Error Rate (Training and Predicted)

演算法	分割規則	修剪規則
ID3	熵、Gain ratio	錯誤預估率
C4.5/ C5.0	Gain ratio	錯誤預估率
CART	Gain ratio	錯誤預估率
CHAID	卡方檢定	不用修剪

- 也就是首先決定樹根上以及樹節點是哪個變數呢？這些變數是從最重要到次重要依次排序的，那怎麼衡量這些變數的重要性呢？　**ID3演算法**用的是資訊增益，**C4.5演算法**用資訊增益率；**CART演算法**使用基尼係數。

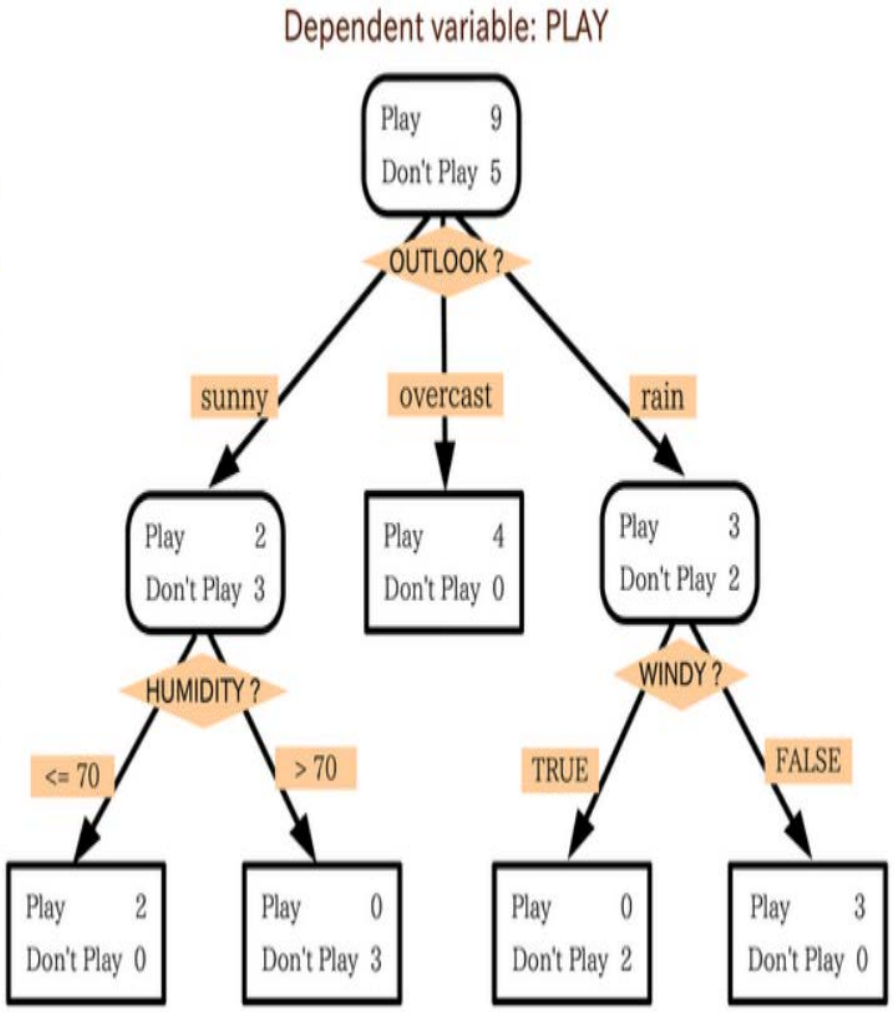
Example1

p.s. Setup "outlook" as a goal varable

- The manger would like to predict if people will come to play golf according to upcoming weather forecast. Goal propery?

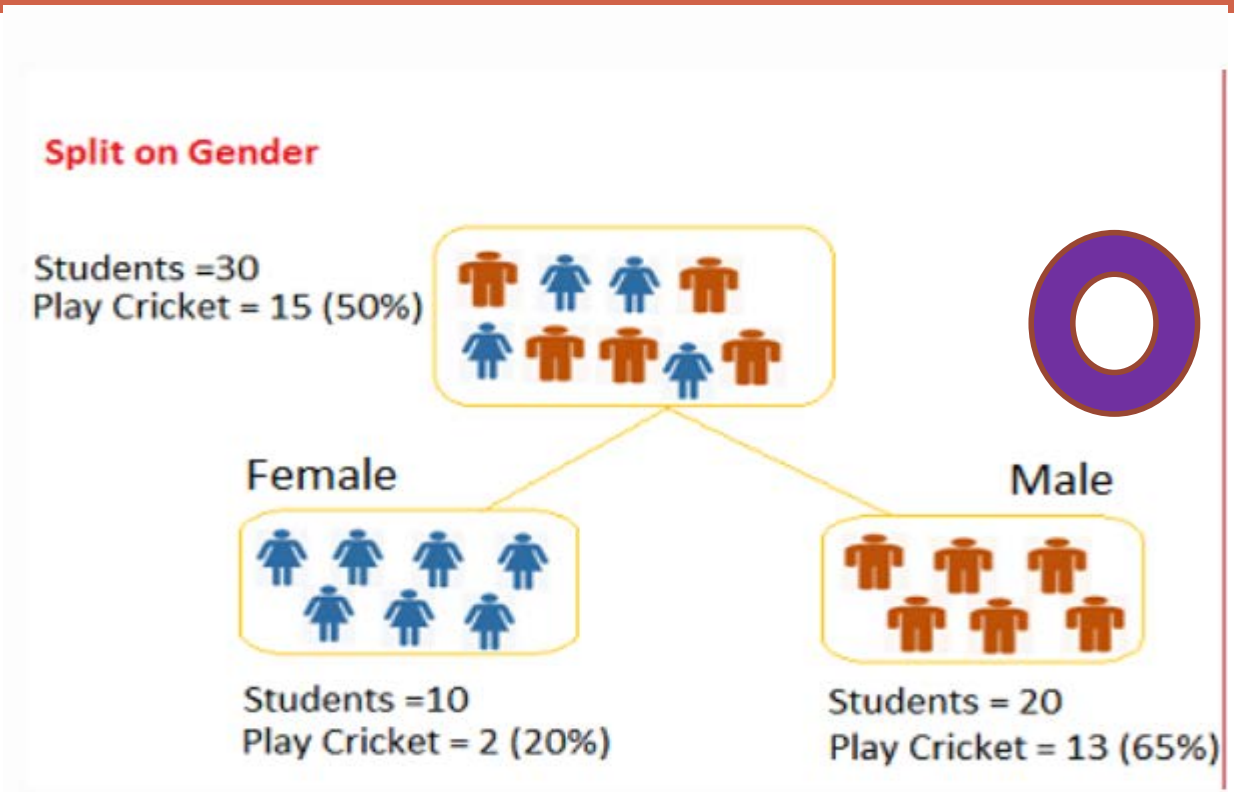
Play golf dataset

Independent variables				Dep. var
OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	FALSE	Don't Play
sunny	80	90	TRUE	Don't Play
overcast	83	78	FALSE	Play
rain	70	96	FALSE	Play
rain	68	80	FALSE	Play
rain	65	70	TRUE	Don't Play
overcast	64	65	TRUE	Play
sunny	72	95	FALSE	Don't Play
sunny	69	70	FALSE	Play
rain	75	80	FALSE	Play
sunny	75	70	TRUE	Play
overcast	72	90	TRUE	Play
overcast	81	75	FALSE	Play
rain	71	80	TRUE	Don't Play



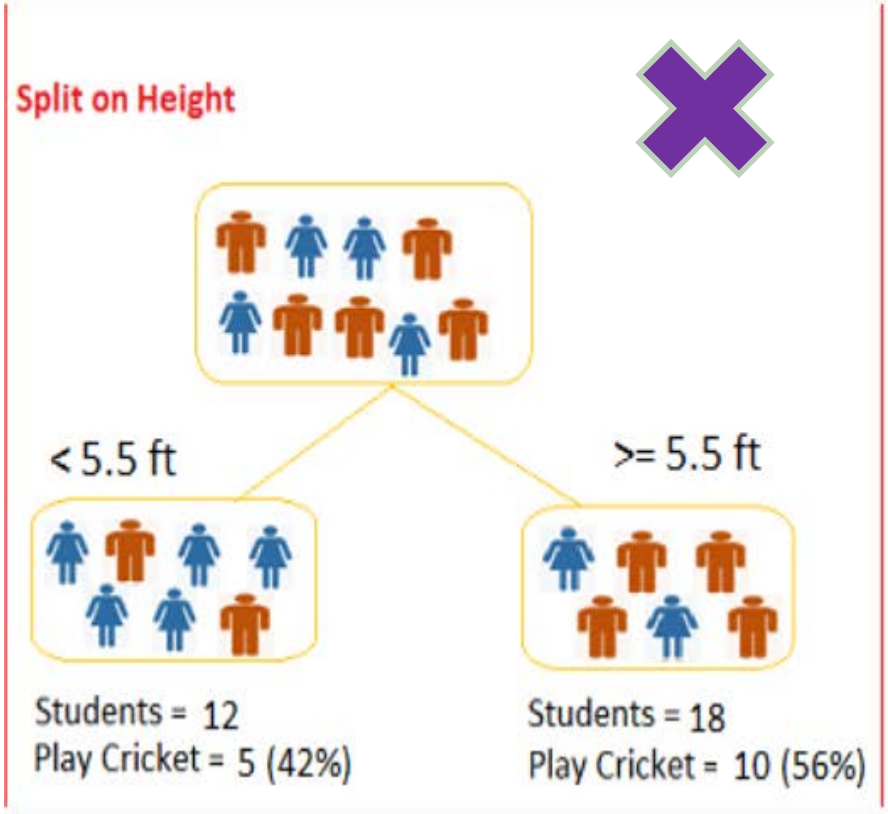
Example for categorical variable decision tree

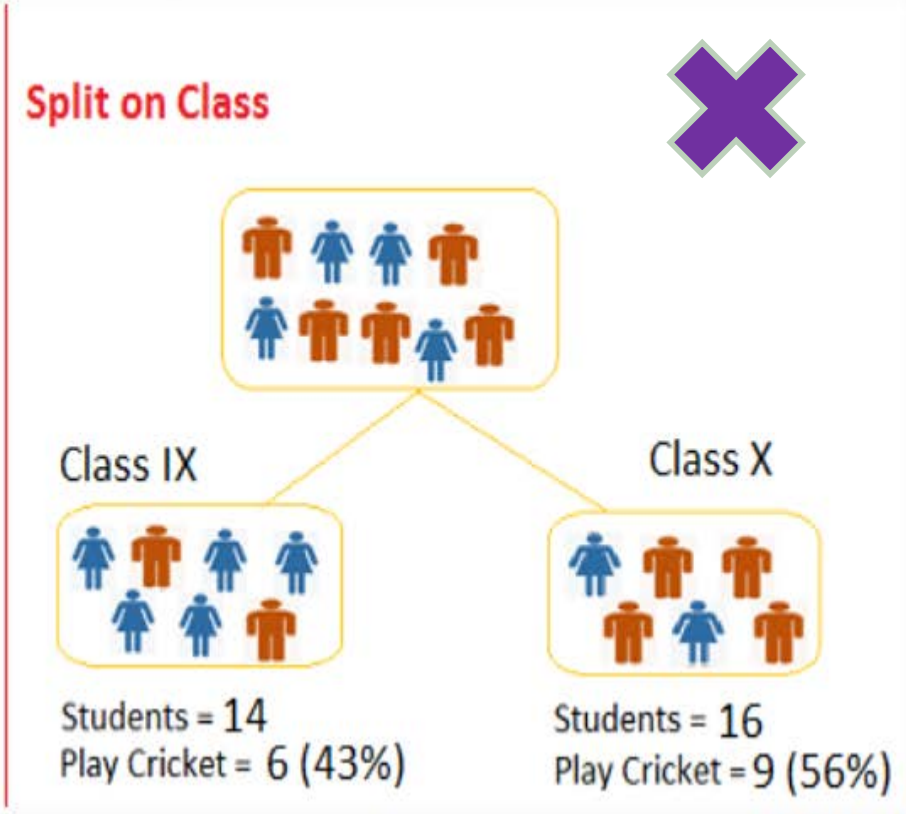
- Let's say we have a sample of 30 students with three variables *Gender* (Boy/ Girl), *Class* (IX/ X) and *Height* (5 to 6 ft). 15 out of these 30 play cricket in leisure time. Now, we want to create a model to predict who will play cricket during leisure period? In this problem, we need to segregate students who play cricket in their leisure time based on highly significant input variable among all three.
- This is where decision tree helps, it will segregate the students based on all values of three variable and identify the variable, which creates the best homogeneous sets of students (which are heterogeneous to each other). In the snapshot below, you can see that variable Gender is able to identify best homogeneous sets compared to the other two variables.



p.s.

1. Decision tree identifies the most significant variable and its value that gives best homogeneous sets of population. To identify the variable and the split, decision tree uses various algorithms.
2. Which one is a target variable? Is the root node? Or leaf nodes?
   Answer: The target variable was "Student will play cricket or not" i.e. YES or NO.
3. Do the results of the percentages of **playing cricket** come to the leaf nodes?
4. 3 varaibles only in the decision tree?


Split on Height
< 5.5 ft
>= 5.5 ft
Students = 12
Play Cricket = 5 (42%)
Students = 18
Play Cricket = 10 (56%)


Split on Class
Class IX
Class X
Students = 14
Play Cricket = 6 (43%)
Students = 16
Play Cricket = 9 (56%)

Example for Continuous Variable Decision Tree

- Let's say we have a problem to predict whether a customer will pay his renewal premium with an insurance company (yes/ no). Here we know that income of customer is **a significant variable** but insurance company does not have income details for all customers. Now, as we know this is an important variable, then we can build a decision tree to predict customer income based on occupation, age and various other variables. In this case, we are predicting values for continuous variable.

p.s.

1. Is the income a target variable? Does treating the target variable as continuous variable? Is the target variable a root node?
   Ans: Using "pay his renewal premium with an insurance company (yes/ no)" as a target variable is better than "income" since no existing information for customer income. However, we do have information for "pay his renewal premium with an insurance company (yes/ no)" in the past few years.
   The main idea is that we can create new variables / features in a decision tree that has better power to predict target variable.
2. Do the results of the percentages of **paying his renewal premium** come to the leaf nodes?

Advantages

- **Easy to Understand**: Decision tree output is very easy to understand even for people from non-analytical background. It does not require any statistical knowledge to read and interpret them. Its graphical representation is very intuitive and users can easily relate their hypothesis.
- **Useful in Data exploration:** Decision tree is one of the fastest way to identify most significant variables and relation between two or more variables. With the help of decision trees, we can create new variables / features that has better power to predict target variable. It can also be used in data exploration stage. For example, we are working on a problem where we have information available in hundreds of variables, there decision tree will help to identify most significant variable.
- **Less data cleaning required:** It requires less data cleaning compared to some other modeling techniques. It is not influenced by outliers and missing values to a fair degree.
- **Data type is not a constraint:** It can handle both numerical and categorical variables.
- **Non Parametric Method:** Decision tree is considered to be a non-parametric method. This means that decision trees have no assumptions about the space distribution and the classifier structure.

Disadvantages

- **Over fitting:** Over fitting is one of the most practical difficulty for decision tree models. This problem gets solved by setting constraints on model parameters and pruning
- **Not fit for continuous variables**: While working with continuous numerical variables, decision tree looses information when it categorizes variables in different categories.

Overfitting

- Overfitting is a significant practical difficulty for decision tree models and many other predictive models.
- Overfitting happens when the learning algorithm continues to develop hypotheses that reduce training set error at the cost of an increased test set error.
- Overfitting meaning your model is learning the noise from the data and its ability to generalize the results is very low. In this case you have a small training error but very large validation error. If you inspect (e.g. by plotting) the evolution of training and validation errors, you see that training error is always going down but validation error is goes up at some point. That is the point you need to stop training to avoid overfitting. I strongly recommend you to read this.

The Concepts

- Decision tree is a graph to represent choices and their results in form of a tree.
- The nodes in the graph represent an event or choice
- And the edges of the graph represent the decision rules or conditions. It is mostly used in Machine Learning and Data Mining applications using R.

The Concepts

- Examples of use of decision tress is − predicting an email as spam or not spam, predicting of a tumor is cancerous or predicting a loan as a good or bad credit risk based on the factors in each of these.
- Generally, a model is created with observed data also called training data. Then a set of validation data is used to verify and improve the model.
- R has packages which are used to create and visualize decision trees. For new set of predictor variable, we use this model to arrive at a decision on the category (yes/No, spam/not spam) of the data.
-

The End