defour analytics

Decision Tree

Understanding Tree Analysis

Regression

- What is Decision tree?
- Why decision tree?
- Types of decision tree
- Constructing decision tree
- Random forest and CART (In detail)

Definition of decision tree

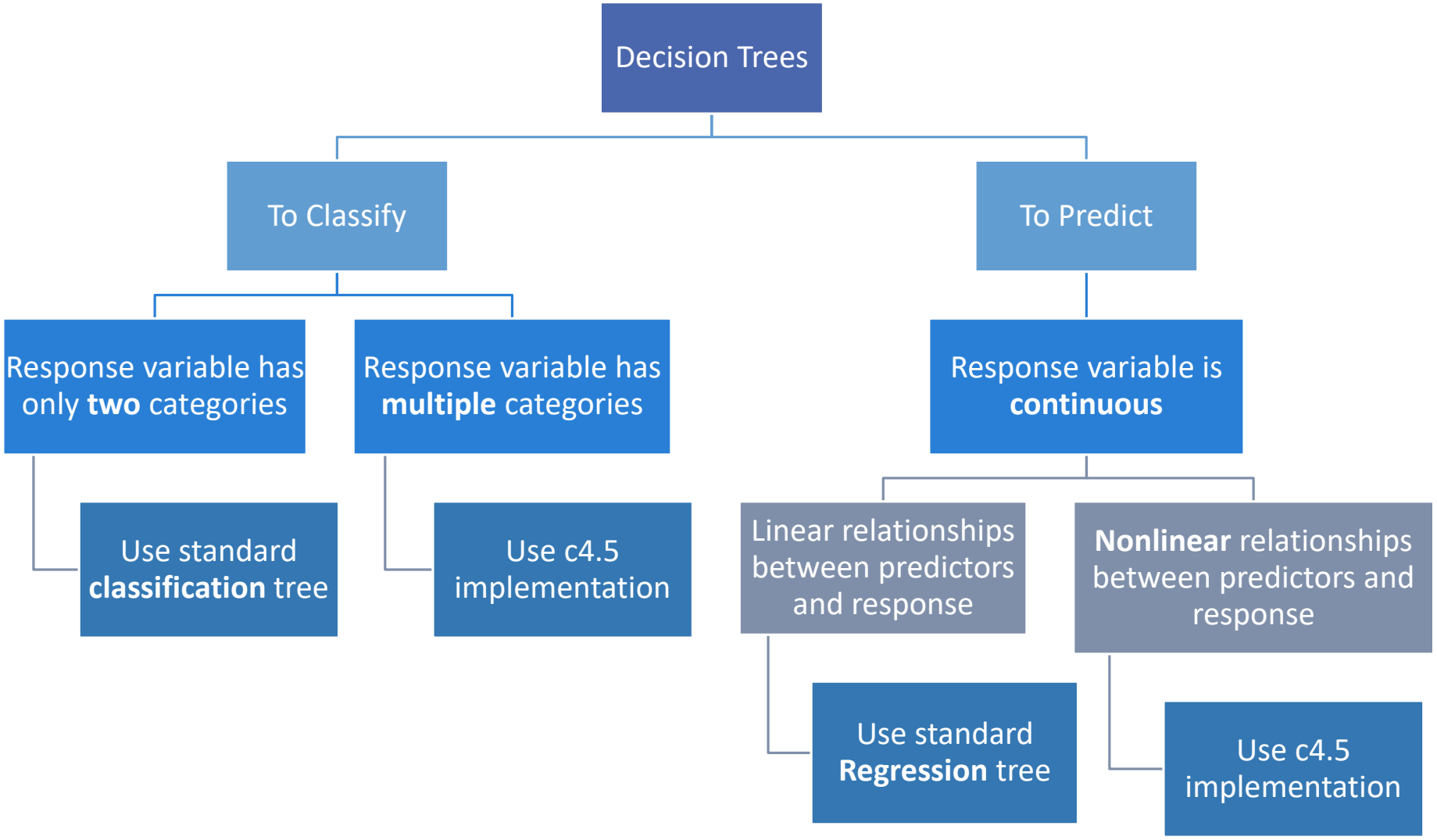
- A decision tree is a natural and simple way of inducing following kind of rules :

  If (Age is x) and (income is y) and (family size is z) and (credit card spending is p) then he will accept the loan.

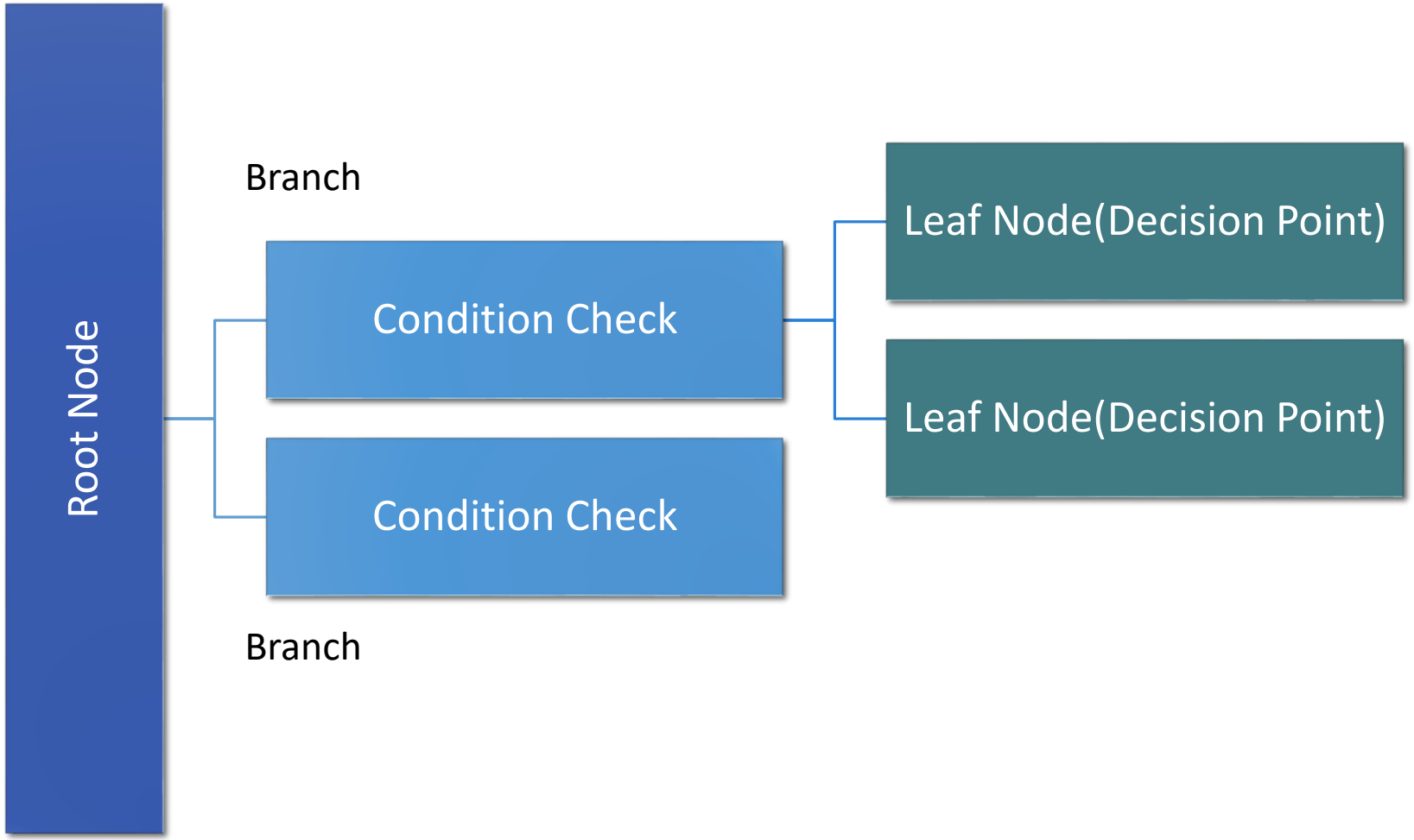
- It is powerful and perhaps most widely used modeling technique of all.

- Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance.

Why decision tree?



Decision tree terms



Industry related examples

- Decision trees can be helpful for analysing many types of decisions involving complex variables.

- Examples include
- Assessing the asset performance and asset behaviour
- Understanding customer complaint behaviour
  - deciding whether to replace wireline logs with logs acquired while drilling
  - evaluating water flood programs
  - optimizing workovers
  - choosing the best offshore platform topsides configuration.

Types of Decision Tree Algorithms

defour analytics

- Decision trees used in data mining are of two main types:
  - **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).

- Some techniques, often called **ensemble methods**, construct more than one decision tree:

Bagging decision trees - an early ensemble method, builds multiple decision trees by repeatedly resampling training data with replacement, and voting the trees for a consensus prediction.

Random Forest - uses a number of decision trees in order to improve the classification rate.

Boosted Trees - can be used for regression-type and classification-type problems.

Types of Decision Tree Algorithms

defour analytics

There are many specific decision-tree algorithms. Notable ones include:

1. ID3 (Iterative Dichotomiser 3)

2. C4.5 (successor of ID3)

3. CART (Classification And Regression Tree)

4. CHAID (Chi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.

5. MARS - extends decision trees to handle numerical data better.

Constructing a decision tree

defour analytics

Two Aspects

- Which attribute to choose?
  - Information Gain
  - Entropy
- Where to stop?
  - Termination criteria

Calculation of entropy

- Entropy is a measure of uncertainty in the data

$$\text{Entropy}(S) = \sum_{(i=1 \text{ to } l)} -|S_i|/|S| * \log_2(|S_i|/|S|)$$

  - S = set of examples
  - S_i = subset of S with value v_i under the target attribute
  - l = size of the range of the target attribute

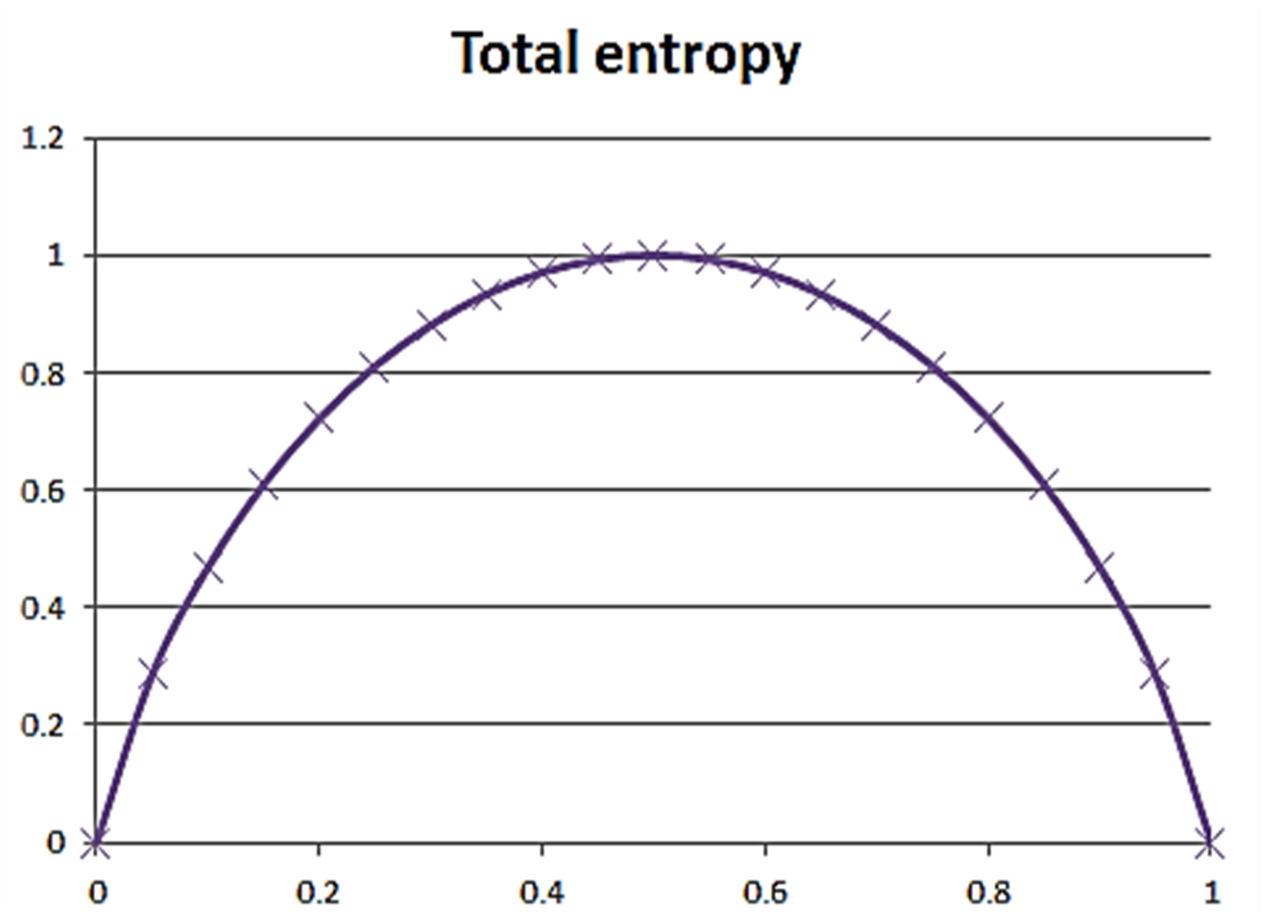
Entropy

- Let us say, I am considering an action like a coin toss. Say, I have five coins with probabilities for heads 0, 0.25, 0.5, 0.75 and 1. When I toss them which one has highest uncertainty and which one has the least?

$$H = - ipi \log2\ pi$$

- Information gain = Entropy of the system before split – Entropy of the system after split

Entropy: Measure of Randomness



Termination criteria

- All the records at the node belong to one class.
- A significant majority fraction of records belong to a single class.
- The segment contains only one or very small number of records.
- The improvement is not substantial enough to warrant making the split.

Pruning trees

- The decision trees can be grown deeply enough to perfectly classify the training examples which leads to overfitting when there is noise in the data.
- When the number of training examples is too small to produce a representative sample of the true target function.
- Practically, pruning is not important for classification.

Approaches to prune tree

- Three approaches
  - Stop growing the tree earlier, before it reaches the point where it perfectly classifies the training data,
  - Allow the tree to over fit the data, and then post-prune the tree.
  - Allow the tree to over fit the data, transform the tree to rules and then post-prune the rules.

Tree Pruning

defour analytics

- **Pessimistic pruning**

  Take the upper bound error at the node and sub-trees

$$e= [f+\frac{z^2}{2N}+z\sqrt{\frac{f}{N}-\frac{f^2}{N}+\frac{z^2}{4N^2}}]/[1+\frac{z^2}{N}]$$

- **Cost complexity pruning**

  J(Tree, S) = ErrorRate(Tree, S) + a |Tree|

  Play with several values a starting from 0

  Do a K-fold validation on all of them and find the best pruning α

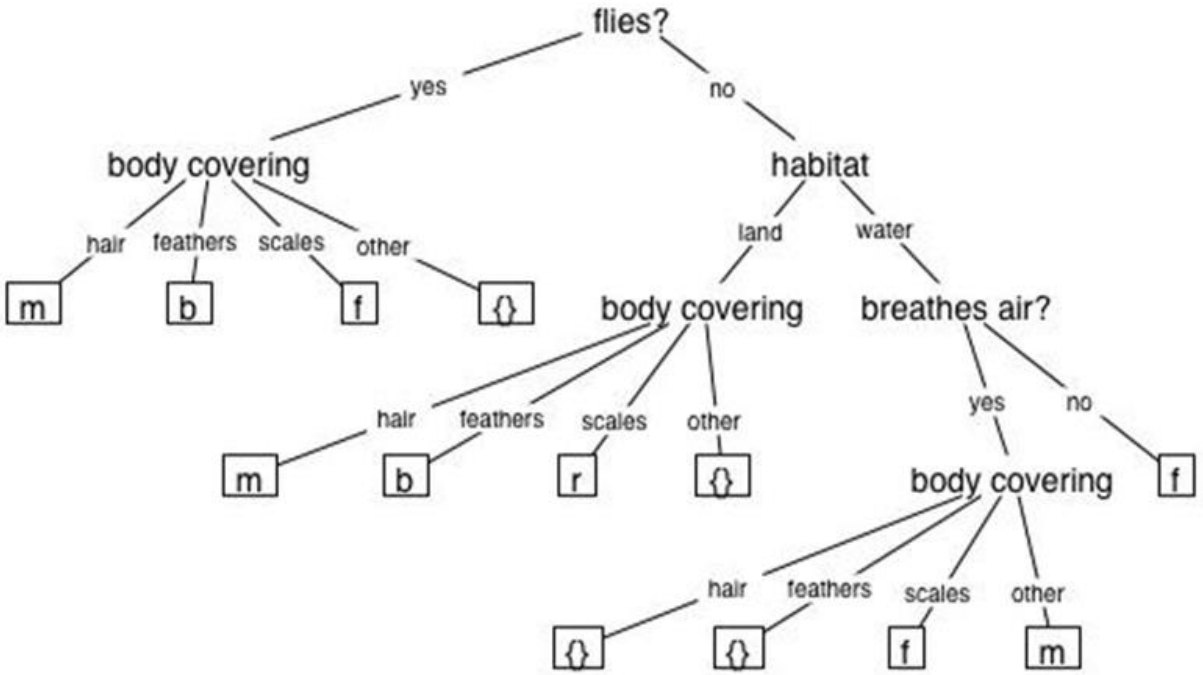
Two most popular decision tree algorithms

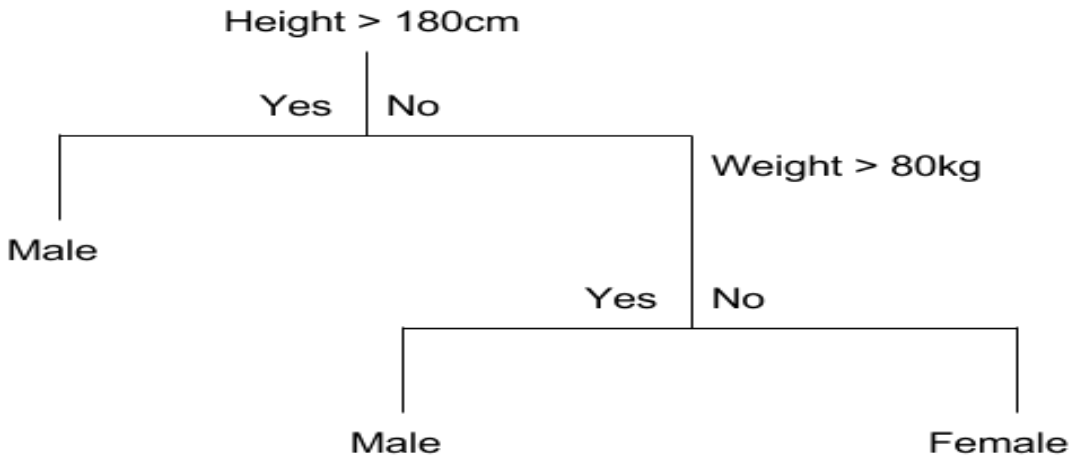
❑**Cart**

–Binary split

–Gini index

–Cost complexity pruning





❑**C4.5**

–Multi split

–Info gain

–pessimistic pruning

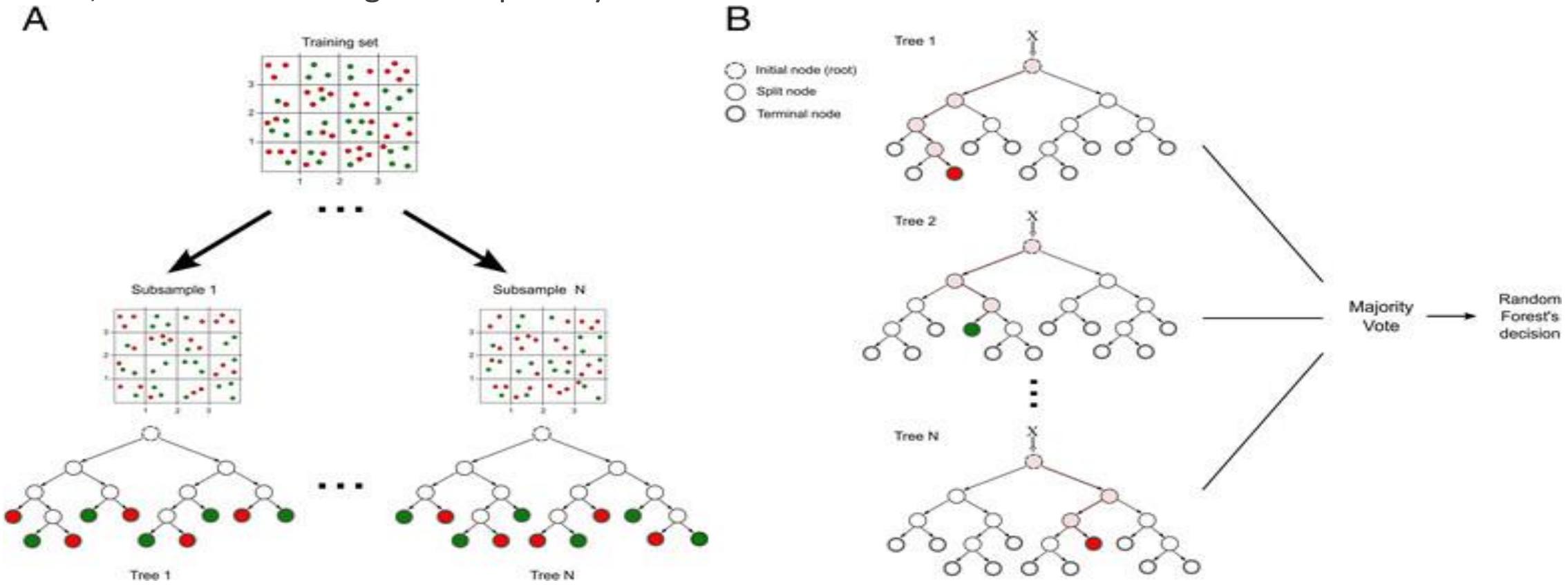
Random Forest

- Random Forest is considered to be a *panacea* of all data science problems. On a funny note, when you can't think of any algorithm (irrespective of situation), use random forest!

- Random Forest is a versatile machine learning method capable of performing both regression and classification tasks. It also undertakes dimensional reduction methods, treats missing values, outlier values and other essential steps of data exploration, and does a fairly good job. It is a type of ensemble learning method, where a group of weak models combine to form a powerful model.

Random Forest

- In Random Forest, we grow multiple trees as opposed to a single tree in CART model. To classify a new object based on attributes, each tree gives a classification and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest) and in case of regression, it takes the average of outputs by different trees.



Random Forest Algorithm

- It works in the following manner. Each tree is planted & grown as follows:
  - Assume number of cases in the training set is N. Then, sample of these N cases is taken at random but *with replacement*. This sample will be the training set for growing the tree.
  - If there are M input variables, a number m<M is specified such that at each node, m variables are selected at random out of the M. The best split on these m is used to split the node. The value of m is held constant while we grow the forest.
  - Each tree is grown to the largest extent possible and there is no pruning.
  - Predict new data by aggregating the predictions of the ntree trees (i.e., majority votes for classification, average for regression).

CART

- A non-parametric technique,using the methodology of tree building.
- Data mining tools like SAS E-Miner; R and SPSS has extended algorithm to handle nominal, ordinal and continuous dependent variables.
  - *One or more independent variables.* Predictor variables can be continuous, ordinal, or nominal.
  - *One dependent variable.* The target variable can be categorical or continuous.

CART Algorithms

- It makes use of *Recursive Partitioning*
- Take all of your data.
- Consider all possible values of all variables.
- Select the variable/value (X=t1) that produces the greatest "separation" in the target.

(X=t1) is called a "split".

- If X< t1 then send the data to the "left"; otherwise, send data point to the "right".
- Now repeat same process on these two "nodes". You get a "tree"

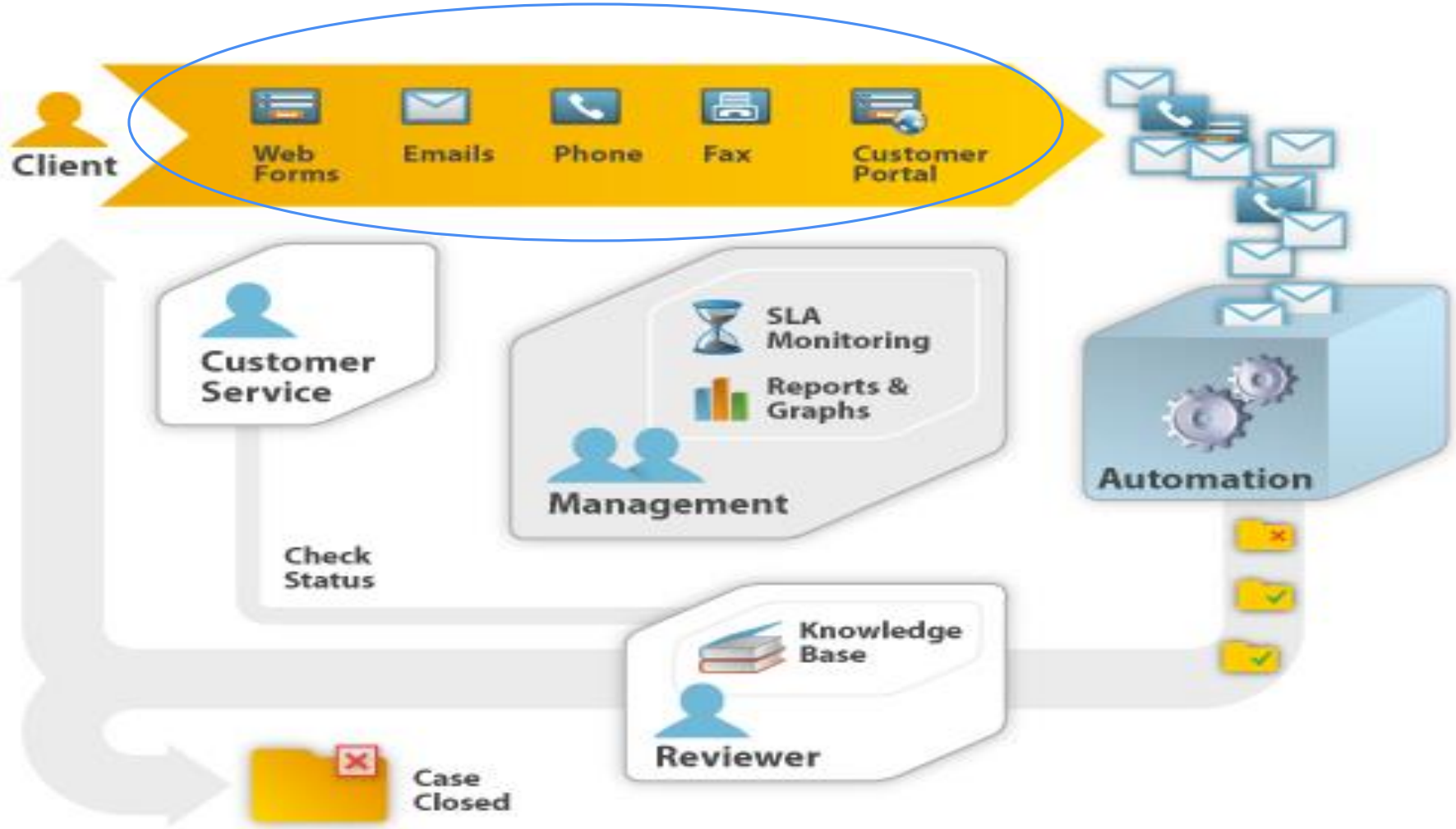
Note: CART only uses binary splits (Unlike CHAID which produces non-binary trees as well)

- "Separation" defined in many ways.
  - Regression Trees (continuous target): use sum of squared errors.
  - Classification Trees (categorical target): choice of entropy, Gini measure, "twoing" splitting rule.

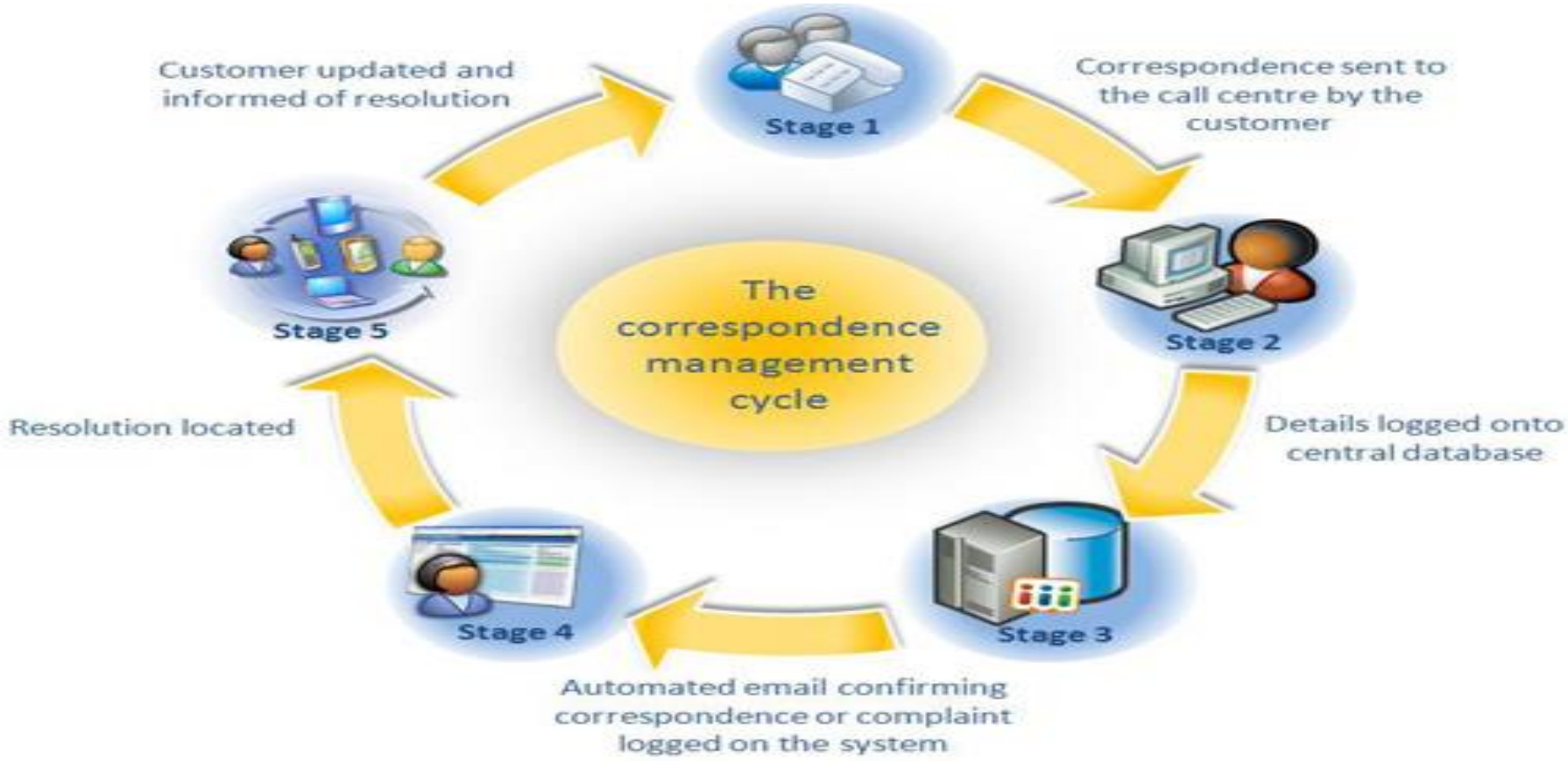
Customer Complaint Management

- It’s an online system where customer’s can log their complaints enabling them to see the resolution of their complaints.
- This system helps the customer satisfaction index.
- This system also measure the stipulated time to close the complaint.
- Through analysis of this system service provider can measure the cross functional department’s performance, SLA monitoring, identify hidden opportunities, spot the critical issues and optimize the process.
- Through which the binding factor for customer engagement is improved.

Types of complaint management system



Complaint Management Process



Problem Statement

- What are the chances that the complaint call time will get violated?
- Based upon the call closure time, is the performance good or bad?

i . e . If call is getting closed within 5 to 10 hours – performance is good.

If call takes more than 10 hours to close the complaint – performance is bad.

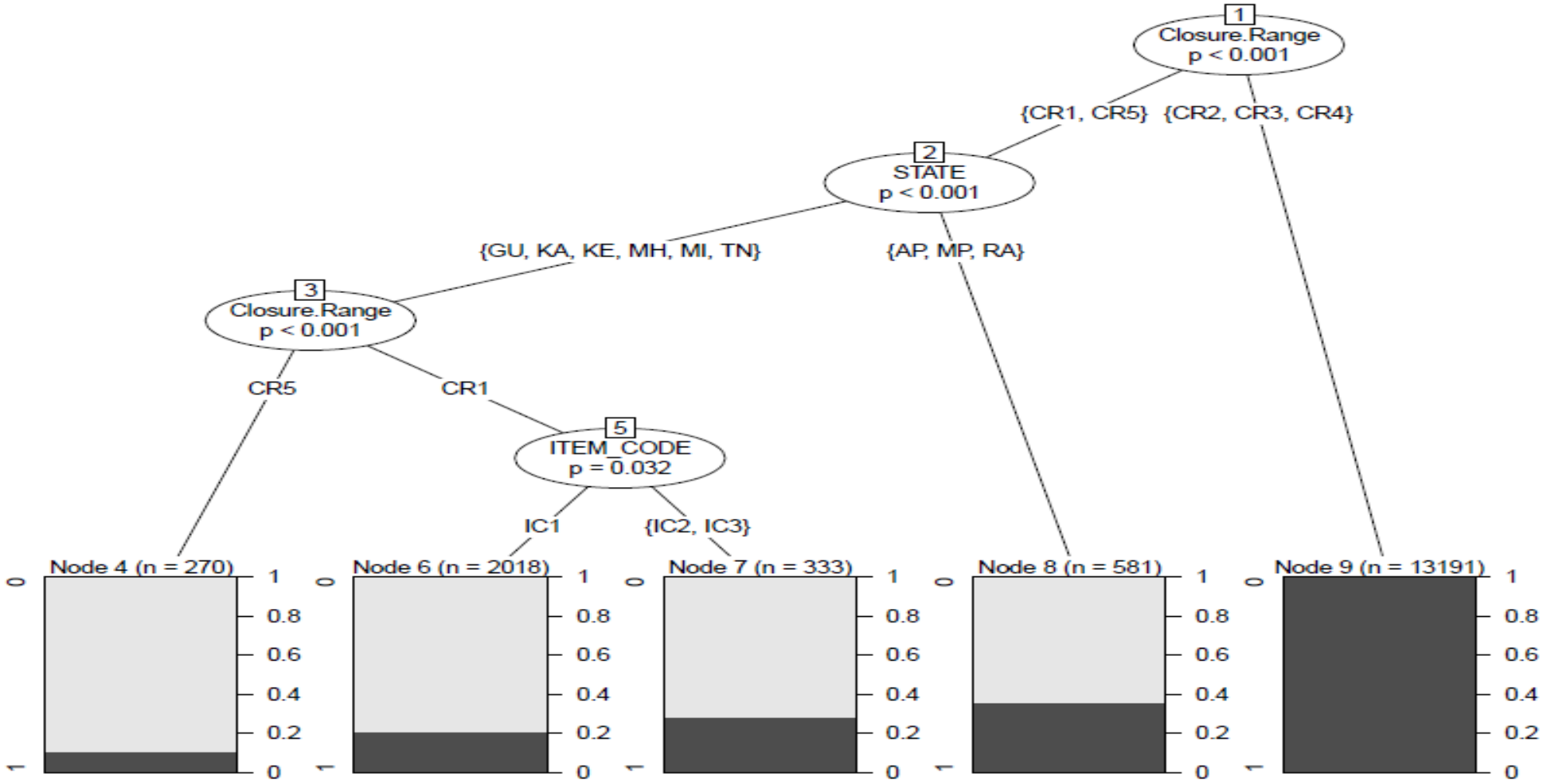
defour analytics

Stepwise description of actions to be taken on data

R script

```
library(data.table)
library(reshape2)
library(randomForest)
library(party)
setwd("C://R")
cm=read.csv("cm.csv")
cm$VIOLATED=factor(cm$VIOLATED)
cm$Closure.Range=factor(cm$Closure.Range)
cm$STATE=factor(cm$STATE)
cm$ITEM_CODE=factor(cm$ITEM_CODE)
fit=ctree(VIOLATED~Closure.Range+STATE+ITEM_CODE,data=cm)
plot(fit)
```

CHAID Decision Tree



Concluding statement

defour analytics

Design Thinking

Ask the participants what kind of problems related to their industry can be solved using this concept.

Question & Answers

defour analytics

THANK YOU