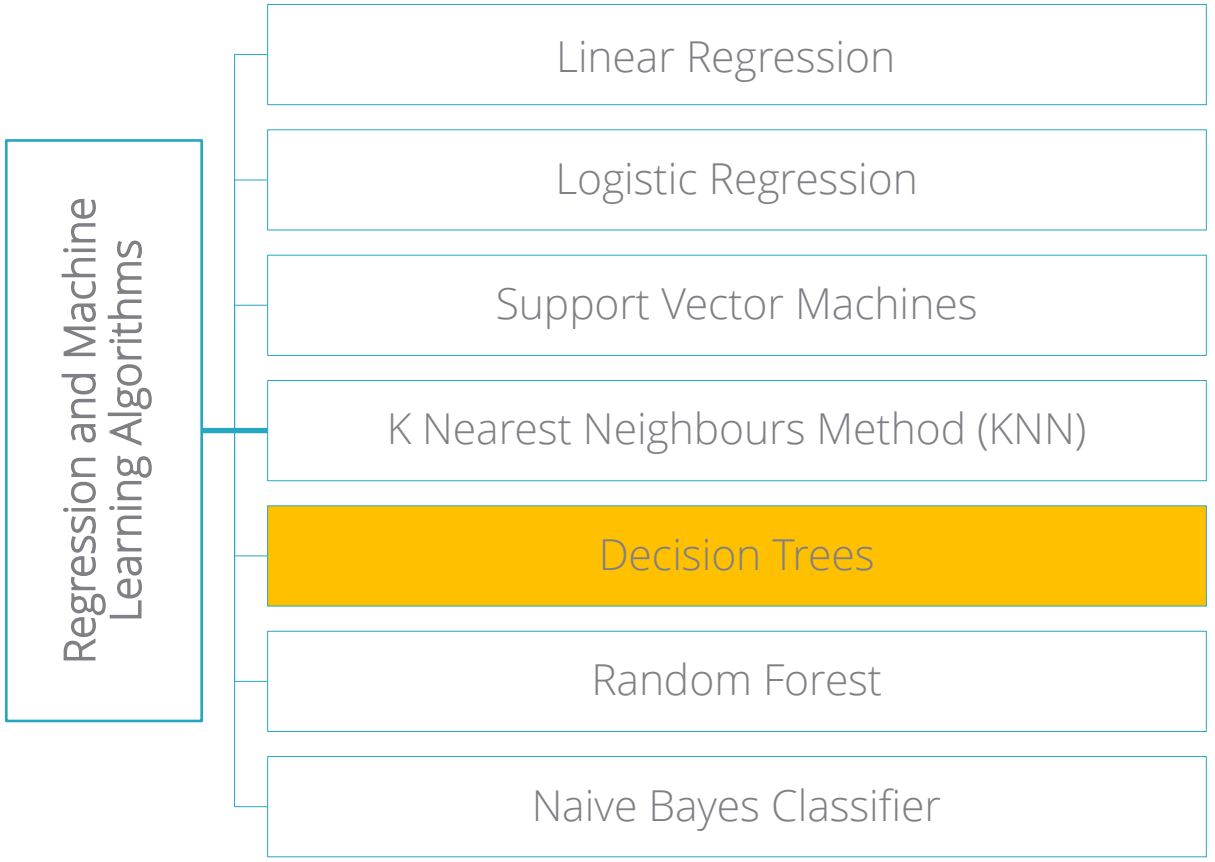
Decision Tree Method
ML ALGORITHM

MACHINE LEARNING METHODS

There is lot of overlapping between statistical modeling and machine learning. The Regression Models are used extensively in ML applications.



DATA SCIENCE INSTITUTE

Introduction to Decision Tree

- Decision Tree learning is one of the predictive modeling techniques.
  - Classification tree - when the predicted outcome is the class to which the data belongs.
  - Regression tree - 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).

- Decision Tree breaks down a data set into smaller subsets and presents association between target variable (dependent) and independent variables as a tree structure. A final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches and leaf node represents a classification or decision.

- The term Classification And Regression Tree (CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al.

Types of Decision Tree

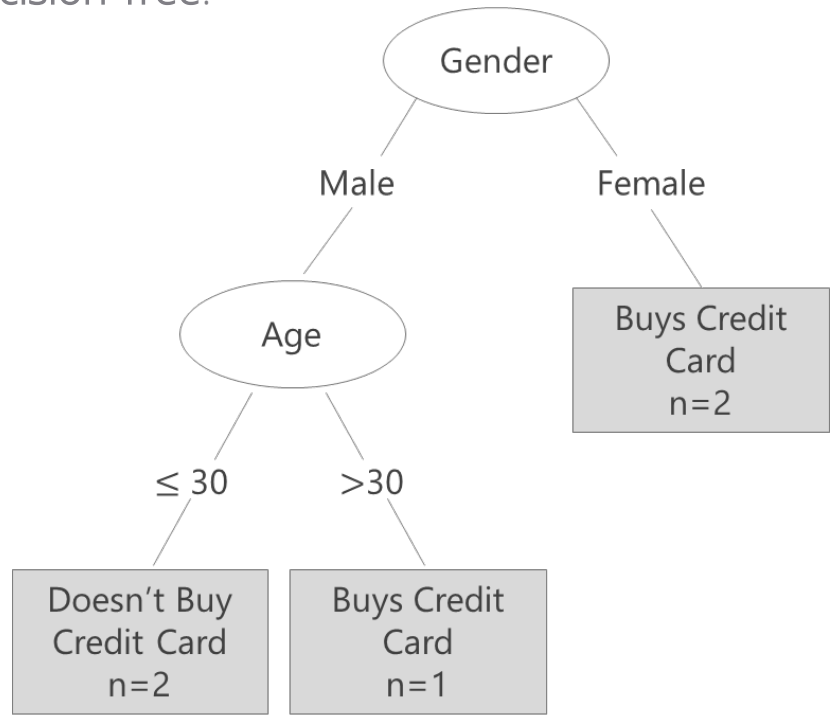
- Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.

- There are many specific decision-tree algorithms. Notable ones include:
  - ID3 (Iterative Dichotomiser 3)
  - CHAID (CHi-squared Automatic Interaction Detector). Performs multi-level splits when computing classification trees.
  - Conditional Inference Trees: Statistics-based approach that uses non-parametric tests as splitting criteria, corrected for multiple testing to avoid overfitting. This approach results in unbiased predictor selection and does not require pruning.

Decision Tree – Basic Framework

Suppose we have information about 5 customers and their decision about buying a credit card. This data can be represented as a Decision Tree:

Customer No.	Gender	Age	Occupation	Buys Credit Card
01	Male	≤30	Student	No
02	Male	≤30	Professional	No
03	Female	≤30	Business	Yes
04	Male	>30	Business	Yes
05	Female	>30	Professional	Yes



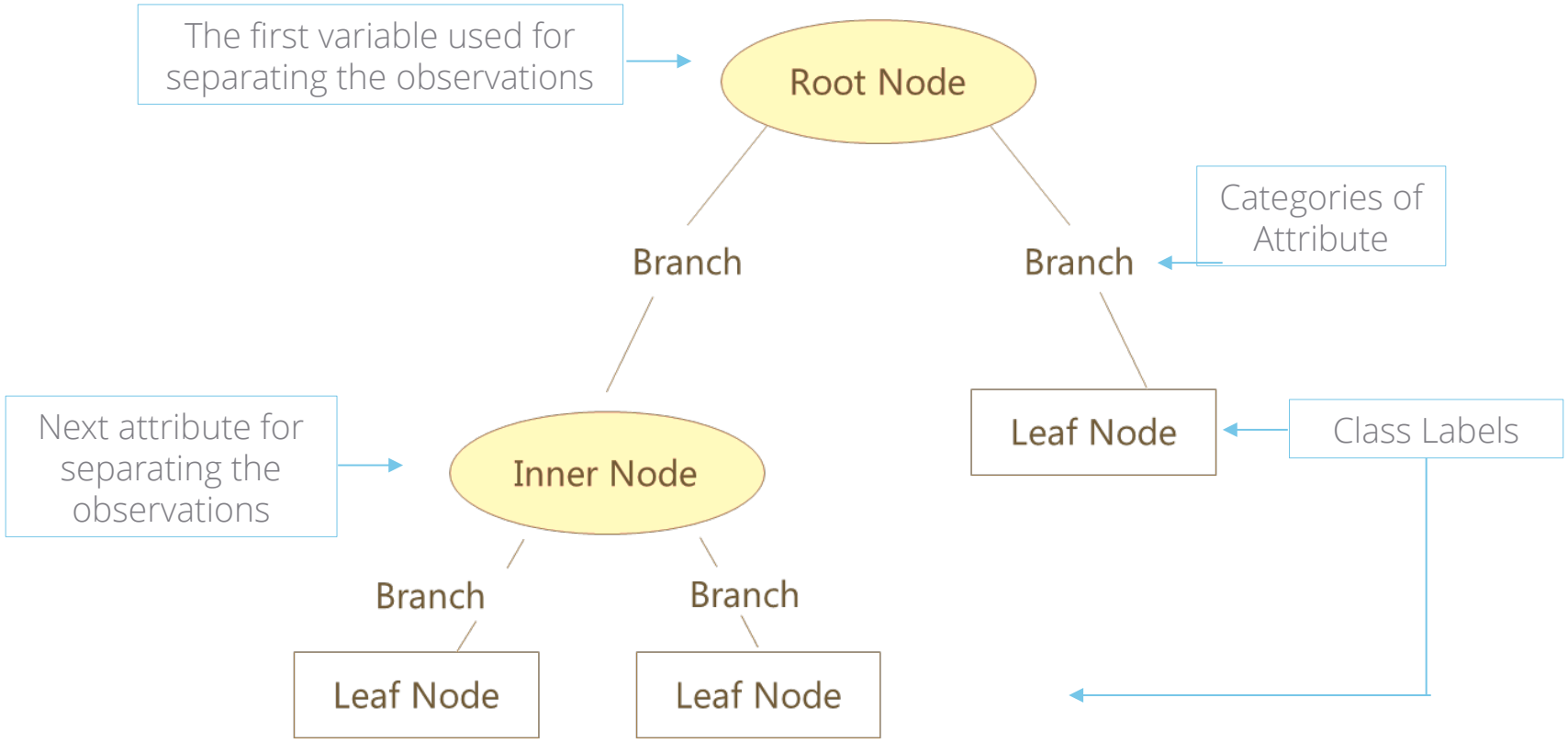
- First subset is based on Gender. All females opt for credit cards.
- Males, however, can be split further, based on Age. Male customers of age 30 years do not buy a credit card, whereas those 30 do buy cards.

DATA SCIENCE INSTITUTE

Decision Tree – Basic Components

Component	Description	Alternate terms
Root node	Has no incoming edges and zero or more outgoing edges	Parent node
Inner node	Each has exactly one incoming edge and two or more outgoing edges	Decision nodes / Child nodes
Leaf node	Each has exactly one incoming edges and no outgoing edges	Terminal nodes
Branches	Categories of attributes	Edges

Decision Tree – Basic Components



* Class labels show observations belong to which class. The leaf node also shows Number of observations and Error rate (Actual classification vs classification given by the tree)

DATA SCIENCE INSTITUTE

Binary and Non Binary Trees

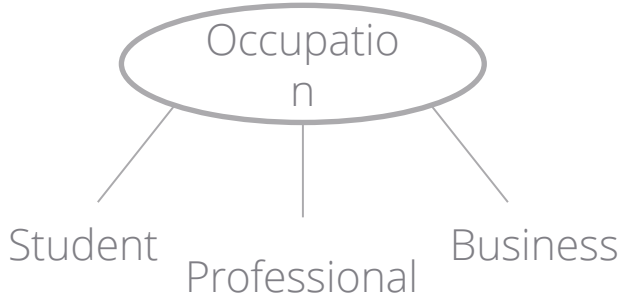
Consider the same example illustrated earlier.

Occupation is the nominal attribute under consideration, with three distinct categories.

Customer No.	Occupation	Buys Credit Card
01	Student	No
02	Professional	No
03	Business	Yes
04	Business	Yes
05	Professional	Yes

There are two ways in which a decision node can be split into further branches:

Binary Tree



Non-Binary Tree

Introduction to CHAID

- CHAID is a type of decision tree technique used for classification problem. The CHAID algorithm was developed in South Africa and was published in 1980 by Gordon V. Kass, who had completed a PhD thesis on this topic.

- The acronym CHAID stands for **Chi** Square **A**utomatic **I**nteraction **D**etection.

- It builds **Non-binary trees** (trees where more than 2 branches can attach to the root node or any node).

- chi-square test of independence is used at each step to determine the strength of relationship between dependent and independent variables.

- CHAID can be effectively applied in case of large data sets.

- The independent variables should be categorical so if data contains continuous predictors then they must be converted to categorical variables.

DATA SCIENCE INSTITUTE

Case Study – Employee Churn Model

Background

- A company has comprehensive database of its past and present workforce, with information on their demographics, education, experience and hiring background as well as their work profile. The management wishes to see if this data can be used for predictive analysis, to control attrition levels.

Objective

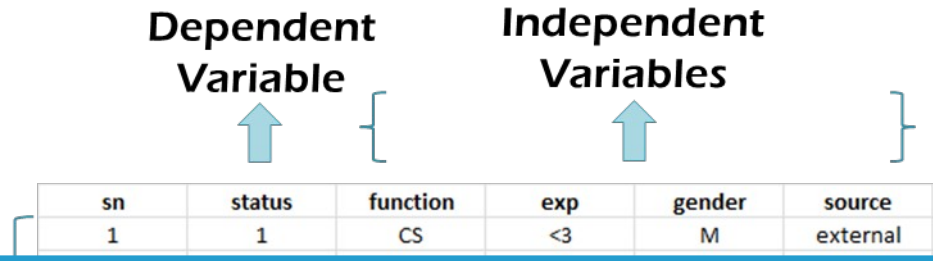
- To develop an Employee Churn model via Decision Tree

Available Information

- Sample size is 83
- **Gender, Experience Level** (<3, 3-5 and >5 years), **Function** (Marketing, Finance, Client Servicing (CS)) **and Source** (Internal or External) are independent variables
- **Status** is the dependent variable (=1 if employee left within 18 months from joining date)

Data Snapshot

EMPLOYEE CHURN DATA



Columns	Description	Type	Measurement	Possible values
sn	Serial Number	-	-	-
status	= 1 If the Employee Left Within 18 Months of Joining = 0 Otherwise	Binary	1,0	2
function	Employee Job Profile	Categorical	CS, FINANCE, MARKETING	3
exp	Experience in Years	Categorical	<3,3-5,>5	3
gender	Gender of the Employee	Categorical	M,F	2
source	Whether the Employee was Appointed via Internal or External	categorical	external, internal	2

DATA SCIENCE INSTITUTE

Case Study: Employee Churn Model

CHAID can be used to develop employee churn model.

Independent variables can be:

- Gender
- Experience Level (<3, 3-5 and >5 years)
- Function (Marketing, Finance, Client Servicing)
- Source (Internal or External)

Dependent variables is "status" (=1 if employee left within 18 months from joining date)

CHAID Algorithm

- CHAID algorithm includes following steps:
  - Preparing predictors
  - Merging categories
  - Selecting the split variable
- Preparing predictors
  - All continuous predictor variables are converted to categorical variables. Example: Experience Level (<3, 3-5 and >5 years).
  - The categorical variable are considered with the natural categories

    Example: Function (Marketing, Finance, Client Servicing).

CHAID Algorithm

- Merging categories
  - In this step categories of predictors are assessed for possible merging.
  - For every predictor all possible pairs of categories are considered.

    Example: For predictor "function", Marketing- Finance, Marketing-CS and Finance-CS are assessed pairwise.
  - The chi-square test of independence is used to decide whether categories in the pair are to be combined or not.
  - The level of significance value used in the testing is named as alpha-to-merge.
  - If the adjusted p-value (Bonferroni) for a particular pair is greater then alpha-to-merge ,then the categories are merged together .

    This merging is continued until no categories can be merged further.

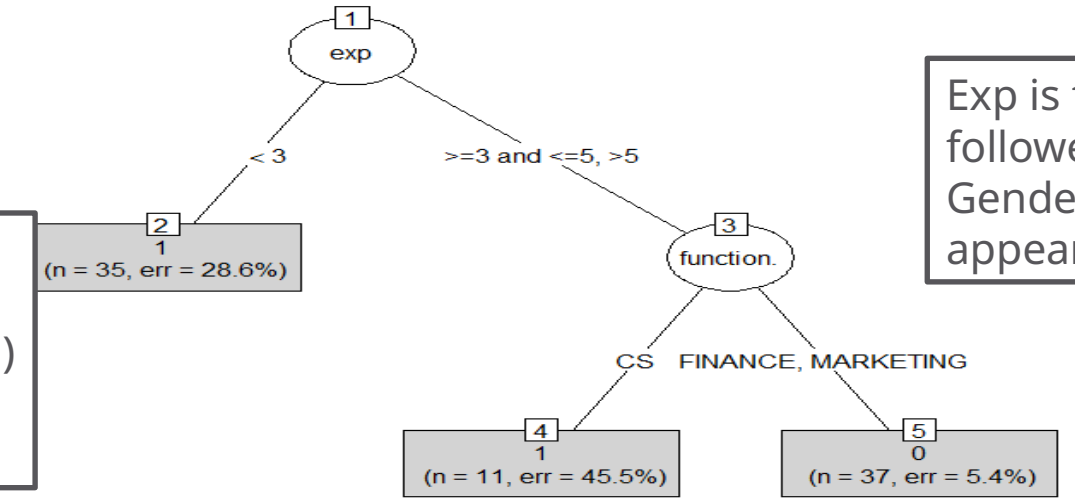
CHAID Algorithm

- Selecting the split variable
  - The Chi-square test of independence between dependent variable and each of predictors is performed.
  - Predictor variable with the smallest adjusted *p*-value (Bonferroni) is considered as a first split variable and appears at first node.
  - If there is tie between 2 predictors then the predictor with largest chi-square statistics value is considered as the split variable.
  - The level of significance value is named as alpha-to-split is decided apriori.
  - The process is repeated at each node using subset of data defined by the node.
  - If the smallest (Bonferroni) adjusted *p*-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed, and the respective node is a terminal node.

- Continue this process until no further splits can be performed .

Decision Tree in R
CHAID Implementation in "CHAID"

```
empdata<-read.csv(file.choose(),heade=T,stringsAsFactors =T)
install.packages("partykit")
install.packages("CHAID", repos = "http://R-Forge.R-project.org", type = "source")
library(CHAID)
empdata$status<-as.factor(empdata$status) # # classification problem
tree<-chaid(formula=status~function.+exp+gender+source,data=empdata)
plot(tree,type="simple")
```



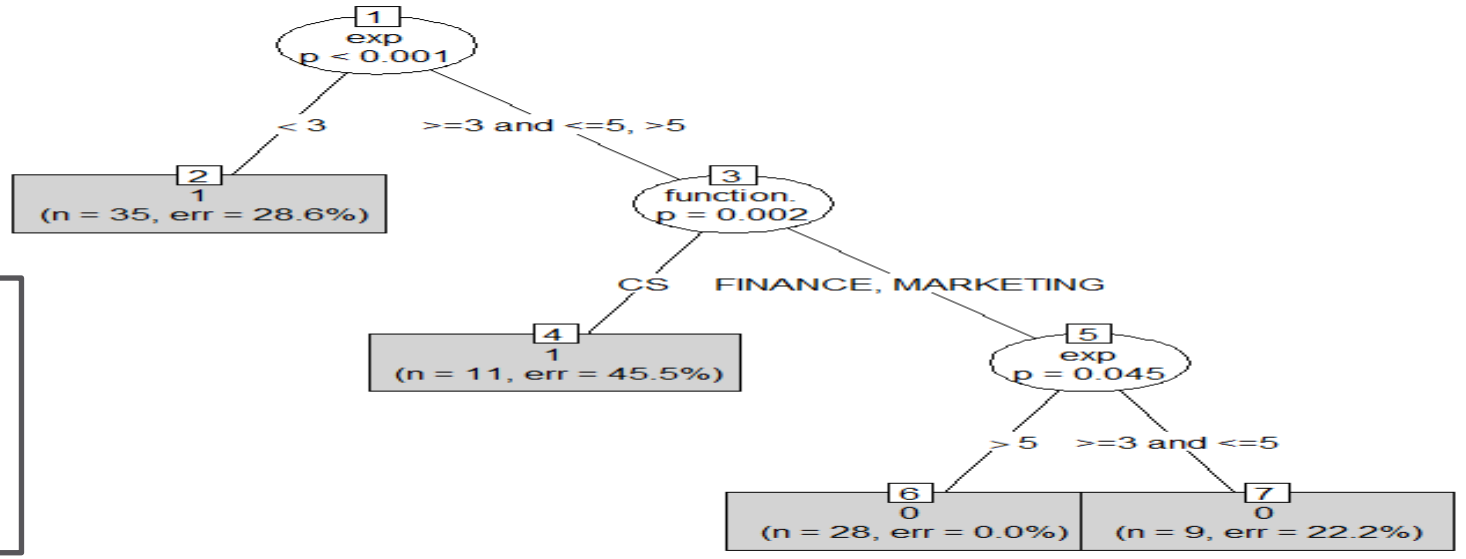
35 employees with exp<3.Out of which 25 Have status=1 (majority=1) 25/35=0.714. Err=28.6%

Exp is the first split variable followed by function. Gender and Source are not appearing in the decision tree.

DATA SCIENCE INSTITUTE

Decision Tree in R....
CHAID-like Implementation in "partykit"

```
empdata<-read.csv(file.choose(),heade=T,stringsAsFactors =T)
install.packages("partykit")
library(partykit)
empdata$status<-as.factor(empdata$status) # classification problem
ctree<-partykit::ctree(formula=status~function.+exp+gender+source,
                       data=empdata)
plot(ctree,type="simple")
```

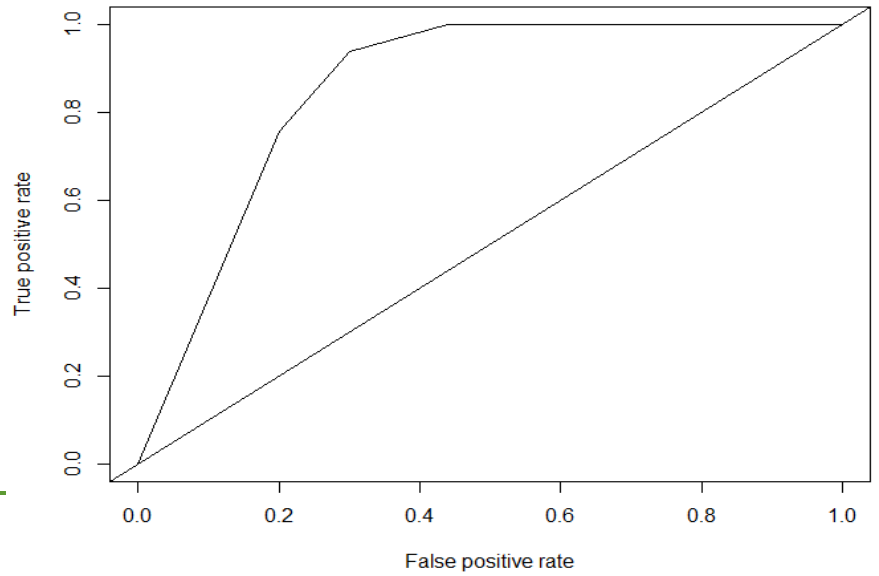


35 employees with exp<3.
Out of which 25 have Y=1 and 10 have Y=0
25/35=0.714

Decision Tree in R....
ROC Curve for "partykit" tree

```
predtree<-predict(ctree,empdata,type="prob")
library(ROCR)
pred<-prediction(predtree[,2],empdata$status)
perf<-performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)

## Area under ROC Curve in R (AUC)
auc<-performance(pred,"auc")
auc@y.values
```




```
[[1]]
[1] 0.8563636
```

DATA SCIENCE INSTITUTE

Note on R Packages

- Package CHAID in R has implemented Chaid algorithm as it was developed. The CHAID package is not available on CRAN. It can be downloaded from R-forge.
- The package "partykit" has a function called 'ctree' which implements Chaid algorithm to some extent. This package is available on CRAN.
- ctree function in partykit package can also handle continuous predictors.
- The package "rpart" handles classification as well as regression problems using decision tree. (CART: Classification and Regression Trees)
- library(rpart)
- tree <- rpart(Y ~ X1 + X2 + X3, method="class", data=...)
- tree <- rpart(Y ~ X1 + X2 + X3, method="anova", data=...)

THANK YOU!!

DATA SCIENCE
INSTITUTE