GENERAL ASSEMBLY

DATA SCIENCE

11 WEEK PART TIME COURSE

Week 7 - Decision Trees
Monday 2nd January 2016

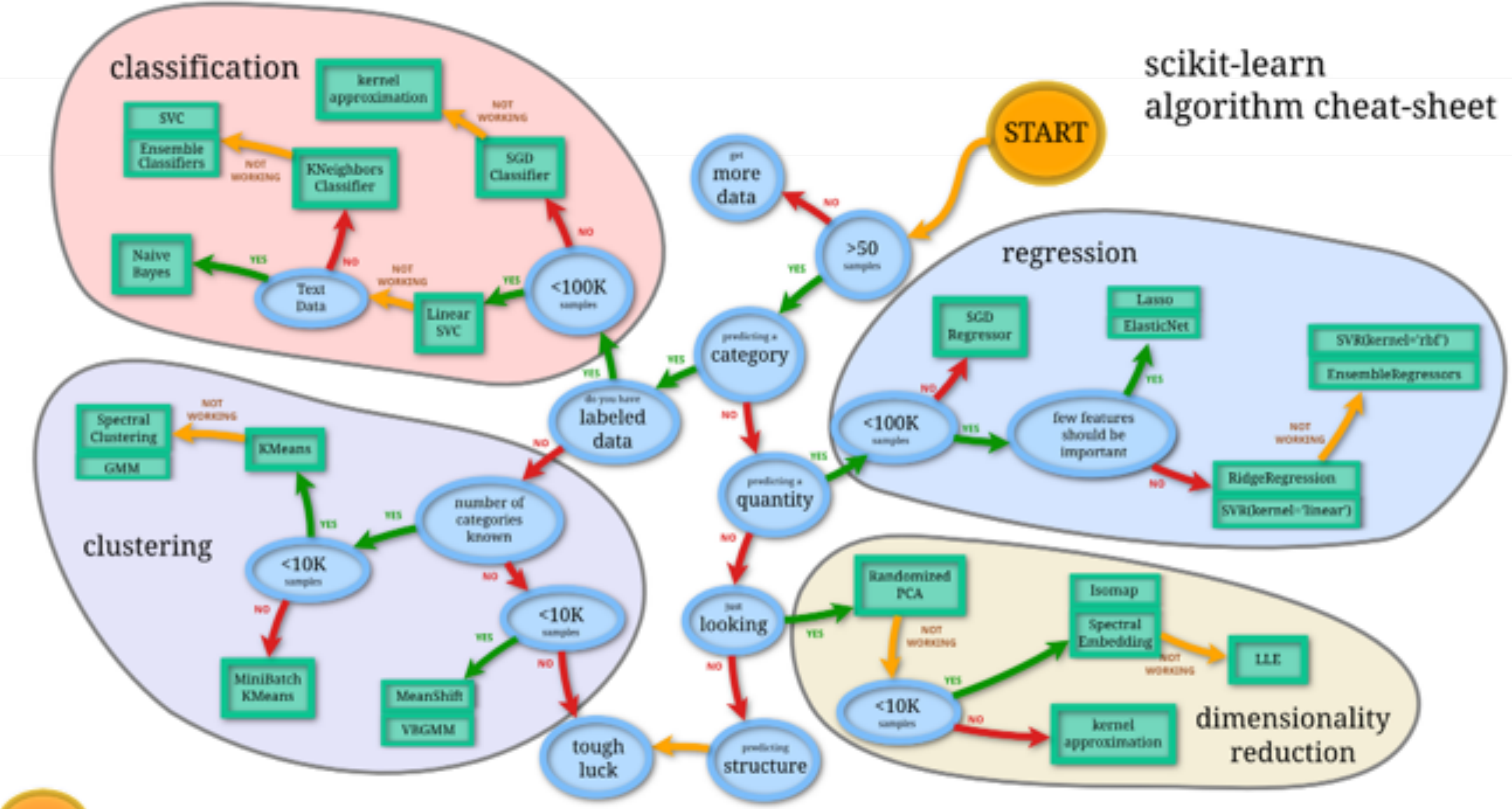
AGENDA

1. ..
2. What are decision trees?
3. How decision trees work
4. Visual example on Titanic dataset
5. Lab
6. Talks
7. Discussion

DATA SCIENCE PART TIME COURSE

DECISION TREES

scikit-learn algorithm cheat-sheet



START

>50 samples

get more data

predicting a category

do you have labeled data

predicting a quantity

just looking

predicting structure

tough luck

classification

<100K samples

Linear SVC

Text Data

Naive Bayes

KNeighbors Classifier

SVC

Ensemble Classifiers

SGD Classifier

kernel approximation

clustering

number of categories known

<10K samples

KMeans

Spectral Clustering

GMM

MiniBatch KMeans

<10K samples

MeanShift

VBGMM

regression

<100K samples

SGD Regressor

few features should be important

Lasso

ElasticNet

RidgeRegression

SVR(kernel='linear')

SVR(kernel='rbf')

EnsembleRegressors

dimensionality reduction

Randomized PCA

<10K samples

Isomap

Spectral Embedding

LLE

kernel approximation

Back

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.

WHAT ARE DECISION TREES?

- A supervised learning technique that can be used for classification or regression.
- Visually engaging and easy to interpret.
- Foundation for getting into very powerful techniques.
- Great for explaining to people!

DECISION TREES - SHORTCOMINGS

- Prone to overfitting.

DECISION TREES - SHORTCOMINGS

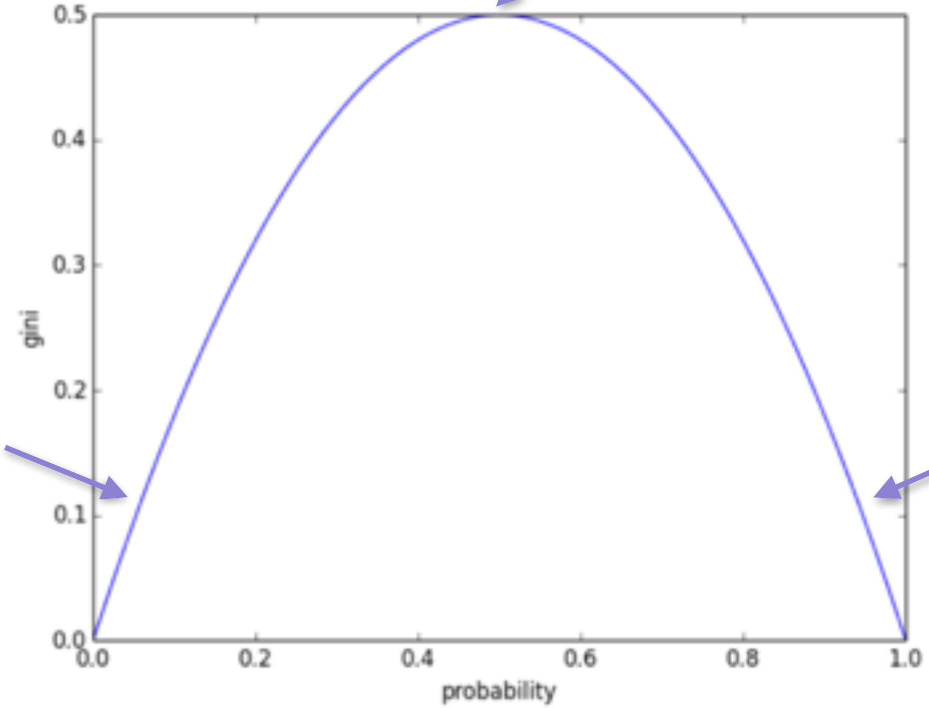
- Prone to overfitting.
- Predictive power is lower in comparison to many other modern techniques.

DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.

DECISION TREES – HOW IT WORKS

The Gini Index



Equal ratio of target classes 50:50

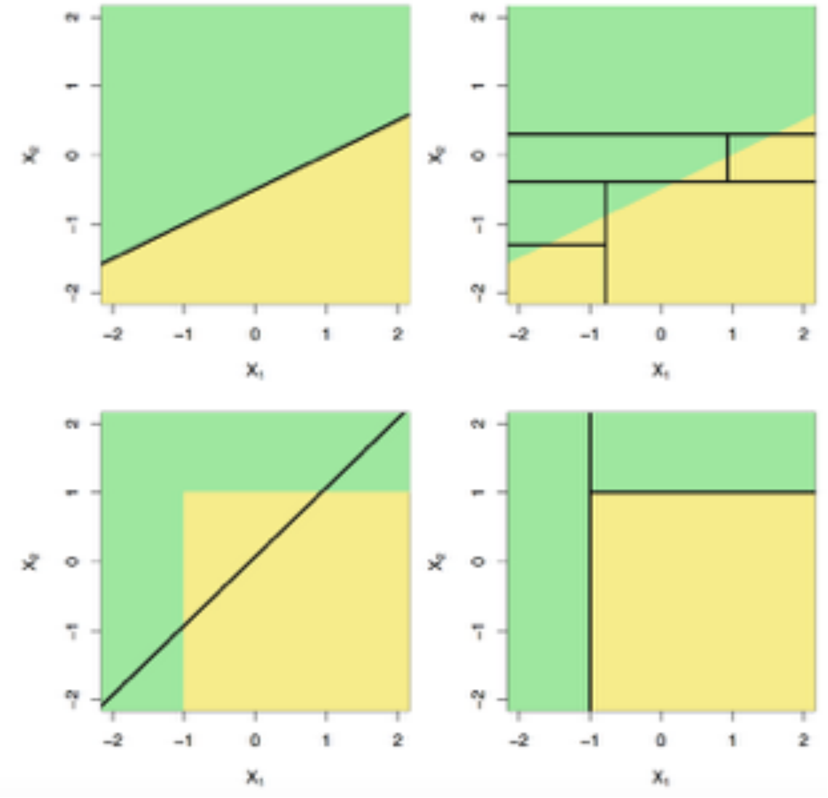
High purity of class 0

High purity of class 1

DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear.

DECISION TREES – HOW IT WORKS



Linear decision boundary

Non-linear decision boundary

DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits

DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.

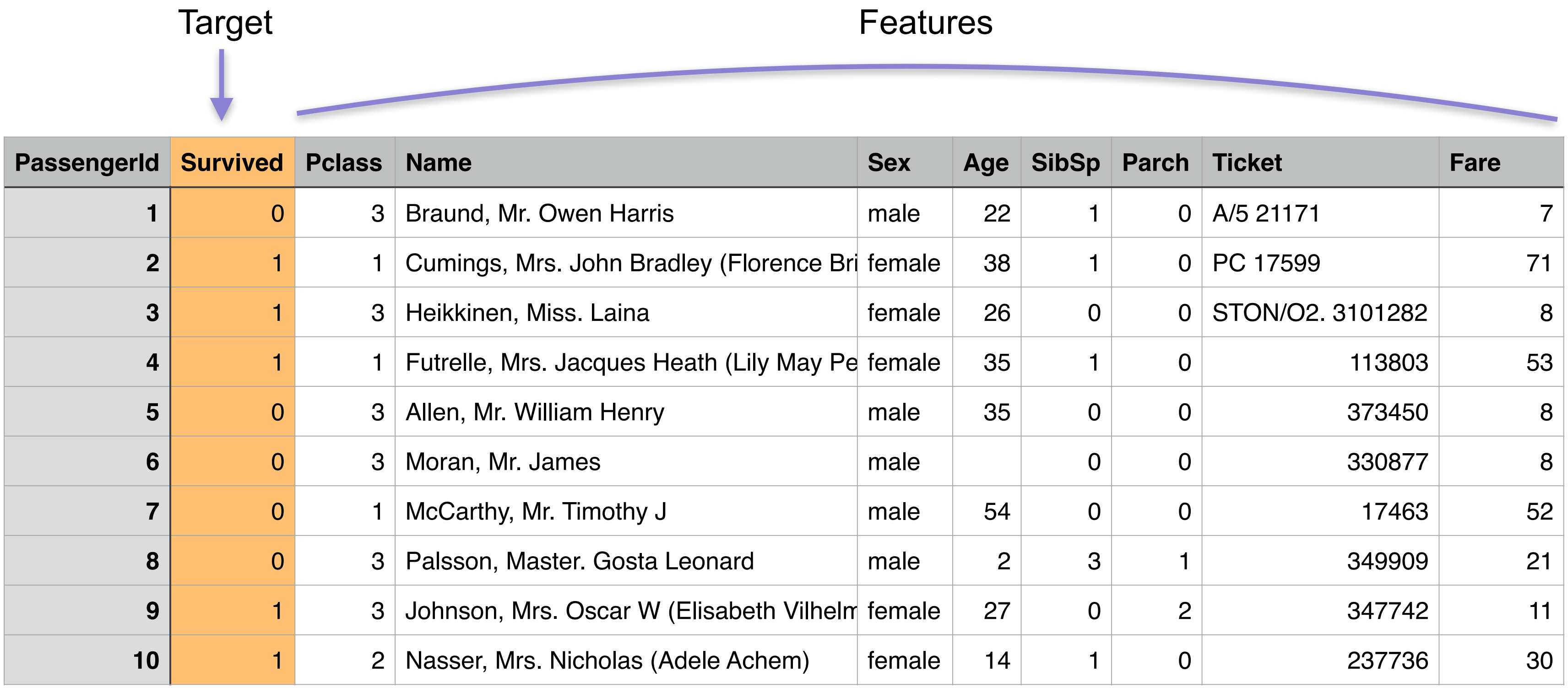
DECISION TREES - HOW IT WORKS

- Scans for a feature to split on that results in the greatest separation between classes in the resulting nodes.
- Non-linear
- Greedy process
- Splits within splits
- For a classification tree, we predict that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs.
- We naturally get combinations of features used for our prediction.

VISUALISING A DECISION TREE FIT

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

TITANIC DATA



Target

Features

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7
2	1	1	Cumings, Mrs. John Bradley (Florence Bri	female	38	1	0	PC 17599	71
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	8
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Pe	female	35	1	0	113803	53
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8
6	0	3	Moran, Mr. James	male		0	0	330877	8
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	52
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelm	female	27	0	2	347742	11
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30

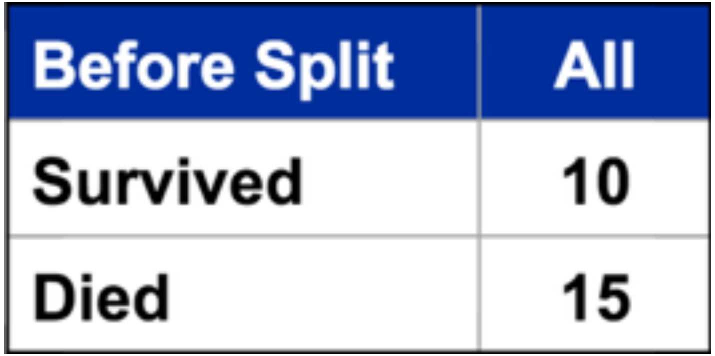
TITANIC DATA

In pairs, pick the two features from the titanic dataset that you believe will be the most predictive of survival.

TITANIC DATA

Variable	Description
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin

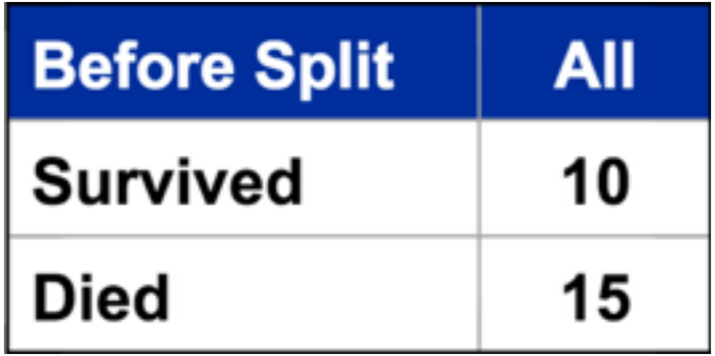
SPLITTING – USING GINI INDEX



Before Split	All
Survived	10
Died	15

$$1 - \sum \left(\frac{class_i}{total} \right)^2$$

SPLITTING - USING GINI INDEX

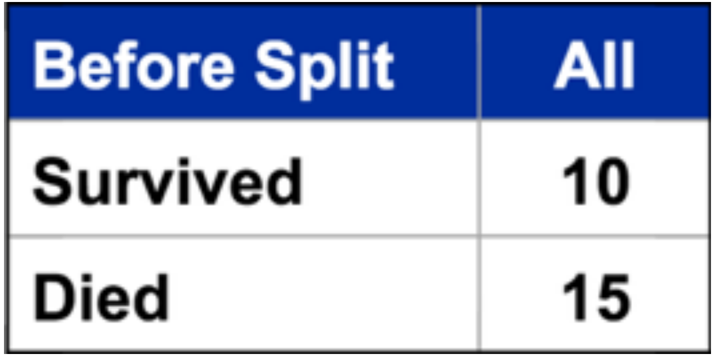


Before Split	All
Survived	10
Died	15

$$1 - \sum \left(\frac{class_i}{total} \right)^2$$

$$1 - \left(\frac{survived}{total} \right)^2 - \left(\frac{died}{total} \right)^2$$

SPLITTING – USING GINI INDEX



Before Split	All
Survived	10
Died	15

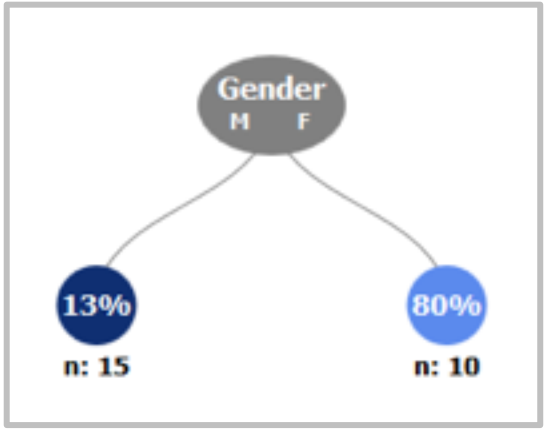
$$1 - \left(\frac{survived}{total}\right)^2 - \left(\frac{died}{total}\right)^2$$

$$1 - \left(\frac{10}{25}\right)^2 - \left(\frac{15}{25}\right)^2 = 0.48$$

SPLITTING – USING GINI INDEX

Gini_O

$= 0.48$



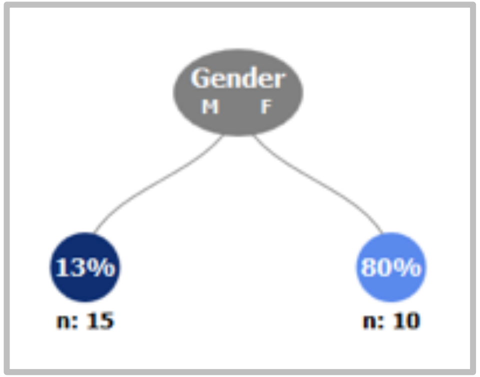
Gender	M
Survived	2
Died	13

Gender	F
Survived	8
Died	2

SPLITTING – USING GINI INDEX

$Gini_O$

$= 0.48$



$Gini_M$

$= 0.23$

$Gini_F$

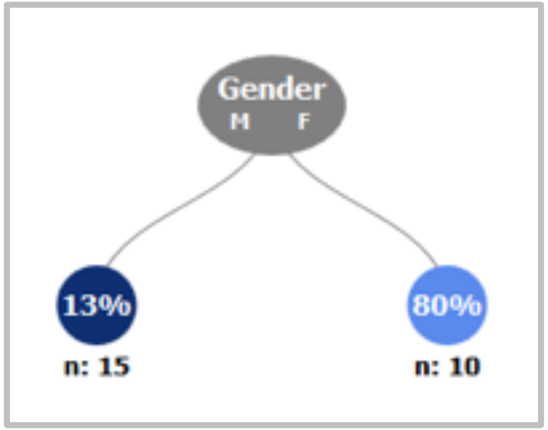
$= 0.32$

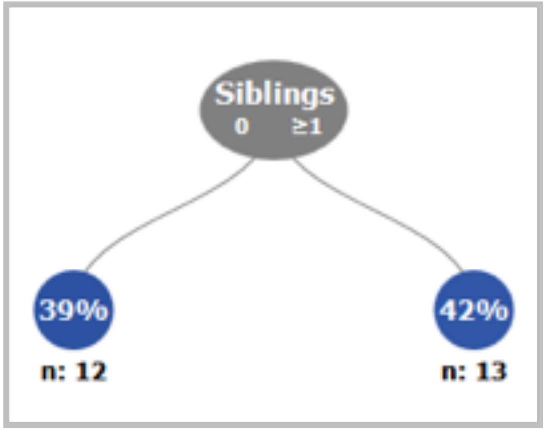
$Gini_C$

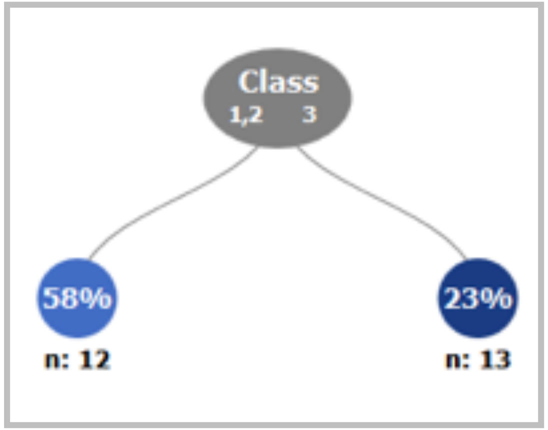
$$Gini_M\left(\frac{M}{M+F}\right) + Gini_F\left(\frac{F}{M+F}\right) =$$

$$0.23\left(\frac{15}{10+15}\right) + 0.32\left(\frac{10}{10+15}\right) = 0.27$$

SPLITTING – USING GINI INDEX







Gender	M	F
Survived	2	8
Died	13	2
Gini	0.27	

Siblings	0	≥1
Survived	5	5
Died	7	8
Gini	0.48	

Class	1,2	3
Survived	7	3
Died	5	10
Gini	0.42	

TITANIC

Using BigML to demonstrate a decision tree model on the Titanic dataset.

https://bigml.com/dashboard/datasets

BigML is a cloud based machine learning tool, designed to make machine learning more approachable.

LAB

SYNCHING A FORK

```
git remote -v
git remote add upstream https://github.com/ihansel/SYD_DAT_3.git
git remote -v
git fetch upstream
git checkout master
git merge upstream/master
OR git reset –hard upstream/master
```