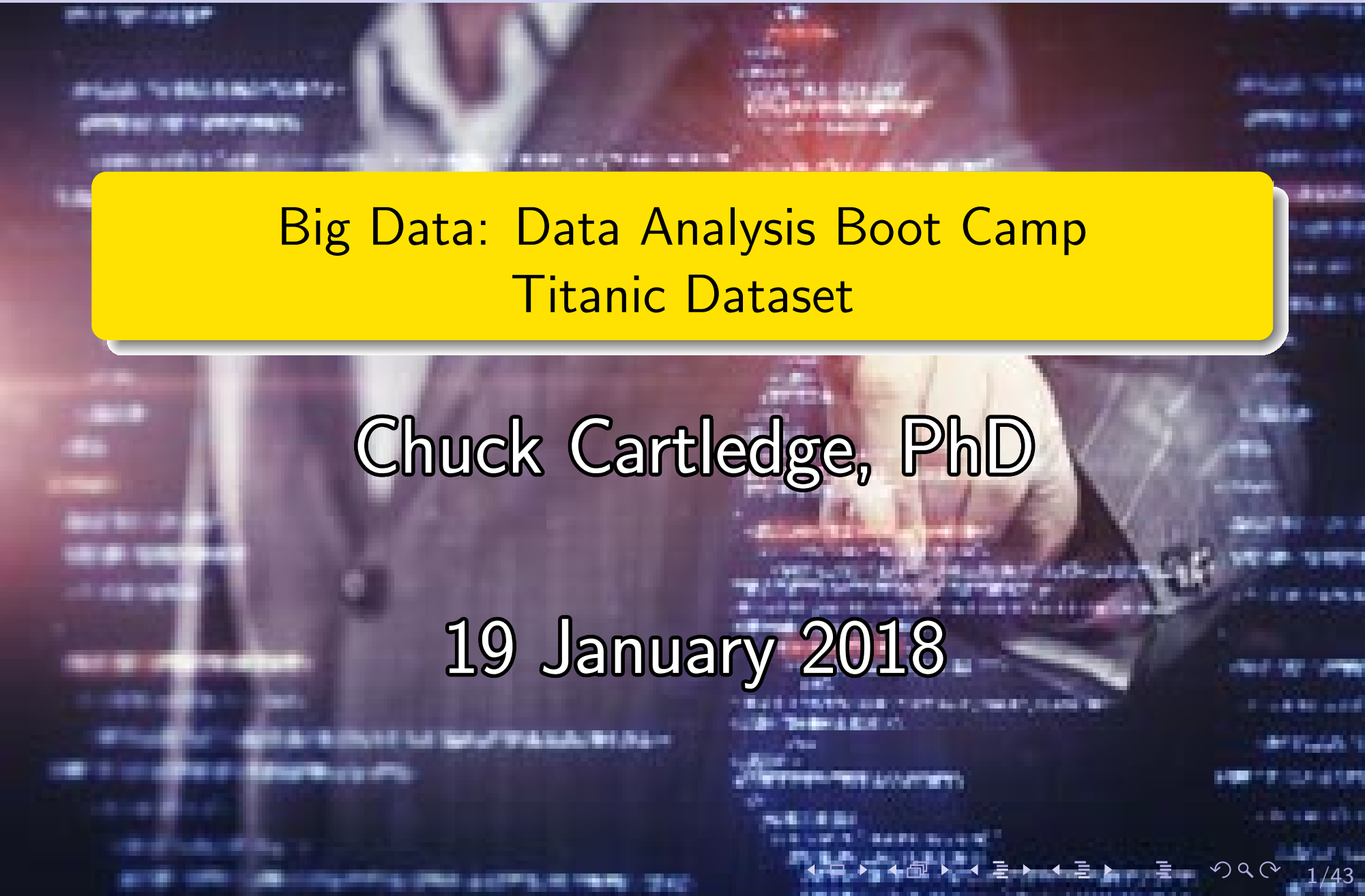


Big Data: Data Analysis Boot Camp
Titanic Dataset

Chuck Cartledge, PhD

19 January 2018

Table of contents (1 of 1)

1. Intro.
2. Background
   - "Well" settled data
   - Data from diverse places
3. Classification problem
   - What is it?
   - Training and testing
   - Types of errors
4. Techniques
   - Results
5. Hands-on
6. Q & A
7. Conclusion
8. References
9. Files

What are we going to cover?

We're going to talk about:

- R's RMS Titanic dataset.
- Other Titanic datasets that contain different data.
- Modeling the datasets to see who will live and who will die.

Basic information

- Ordered: 17 Sep. 1908
- Completed: 2 Apr. 1912
- Maiden voyage: 10 Apr. 1912
- Sank: 14 Apr. 1912

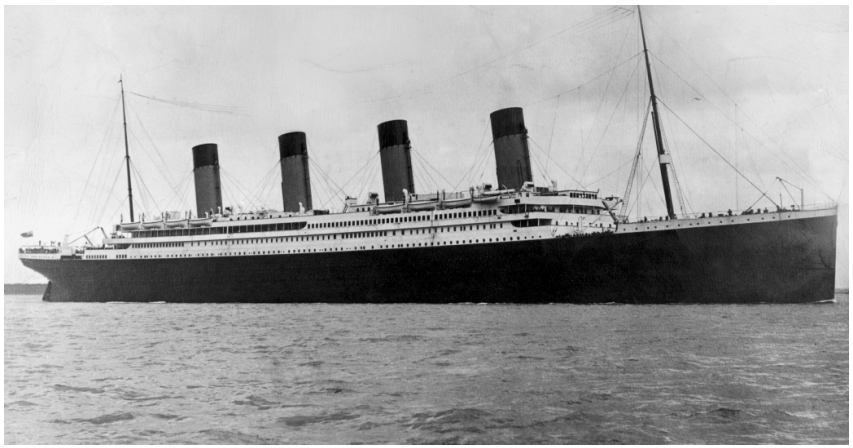


Image from [7].

Where she was damaged

- Red are water tight bulkheads
- Green is where the iceberg hit

As the bow settled, water overflowed the bulkheads

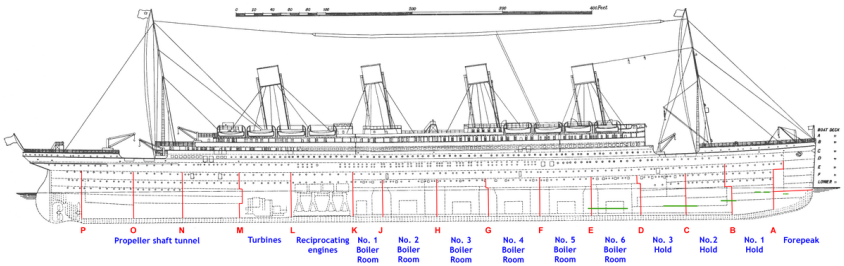


Image from [12].

Same image.

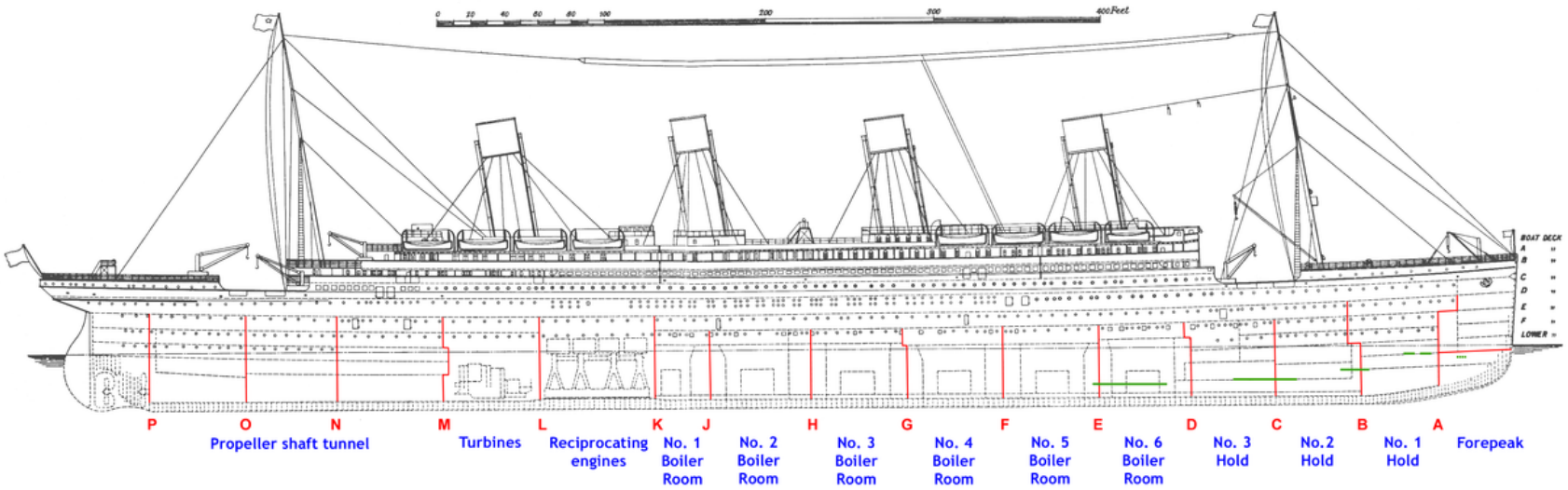


Image from [12].

How many died and why?

- Sailing capacity (passengers and crew): 3,372
- Lifeboat capacity: 1,178
- Number of people on board (accounts vary): 2,201
- Number of people who survived: ~706 - 712 (R thinks 711)

Passengers, crew, builder's men, and others.

Image from [9].

Expected first class passengers

Lots of lists of 1st class passengers. Even, some of 2nd, and 3rd class passengers[11].
Lists of non-passengers (ship's crew, and builder's technicians) are more challenging[6].
R has a built-in Titanic dataset: `Titanic`

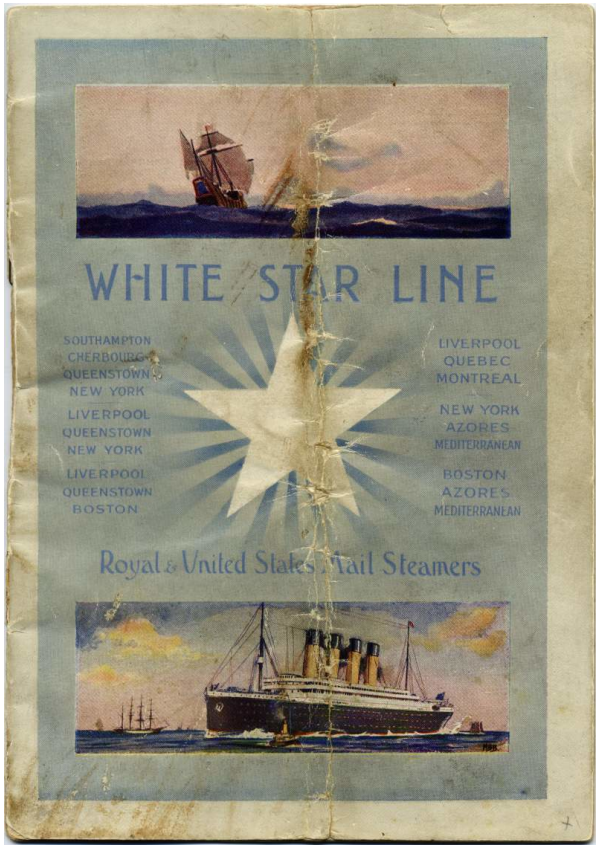


Image from [8].

A crew list

A reasonable collection of crew and builder's representatives is available.

- Name, job, status (lost or not)
- Age, place of birth

A

Abbott, Ernest Owen Pantryman *Lost*
98, Northumberland Road, Nicholstown, Southampton
Age: 21 *Place of Birth:* Hants, Southampton

Abraham, C
see Abrams, William

Abrams, H
see Abrams, William

Abrams, William Fireman *Lost*
3 or 11, Charles Street, Chapel, Southampton
Age: 35 Place of Birth: Northwich

Adams, R Fireman *Lost*
168, Romsey Road, Shirley, Southampton
Age: 26 *Place of Birth:* Hants
Crew Agreement has 168 Pound Tree Road

Ahier, Percy Snowden Steward *Lost*
136, Northumberland Road, Nicholstown, Southampton
Age: 20 Place of Birth: Jersey

Akerman, Albert Steward *Lost*
25, Rochester Street, Northam, Southampton
Age: 28 Place of Birth: Salisbury, Wiltshire

Akerman, Joseph Francis Assistant Pantryman *Lost*
25, Rochester Street, Northam, Southampton
Age: 35 Place of Birth: Salisbury, Wiltshire

Image from [10].

titanic3 dataset from PASWR[15]

- Part of the PASWR library
- Thomas Cason of UVa has greatly updated and improved the titanic data frame using the *Encyclopedia Titanic*.
- Focuses and expands the passenger data.

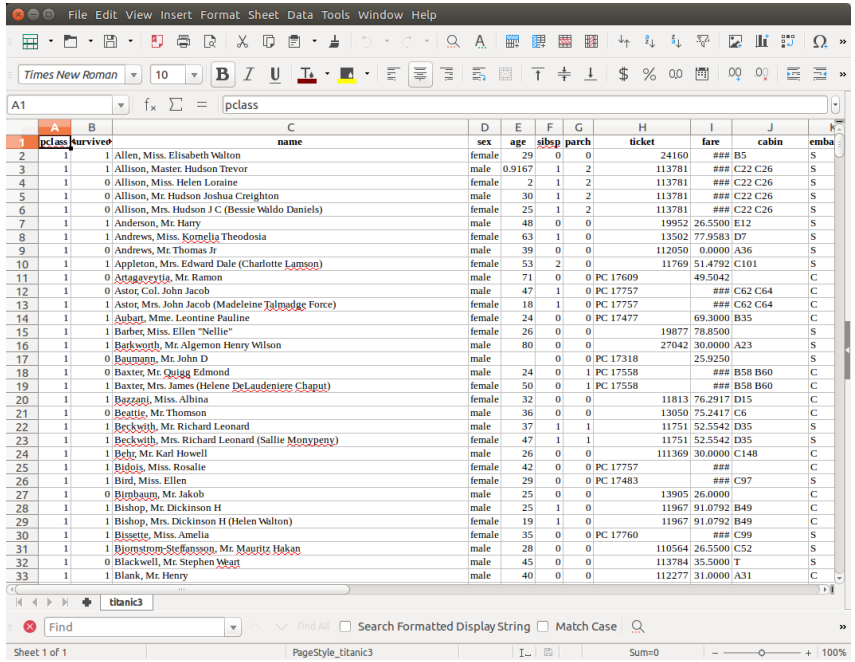


Image from [2].

titanic3 attributes/variables

Name	Explanation
Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Bringing the pieces together

Combining:

- Passenger data from `titanic3`
- Crew data from Southampton
- Not all data in both datasets

Get a reasonable estimation of who survived, or not when the RMS Titanic went down.

A definition

> *"Classification is the task of learning a* **target function** *f that maps each attribute set* **x** *to one of the predefined class labels* **y***."*
>
> *Tan, et al. [13]*

As a picture

1. A collection of correctly labeled data (training data) is available.
2. The supervised data is processed by some sort of machine learning algorithm (there are many) to create a model (or classifier).
3. Unlabeled (test or new) data, is processed by the model and predictions are made.

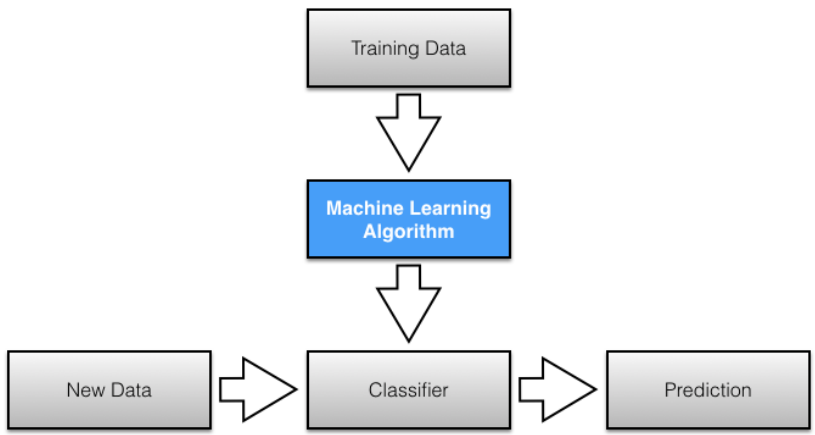


Image from [5].

Same image.

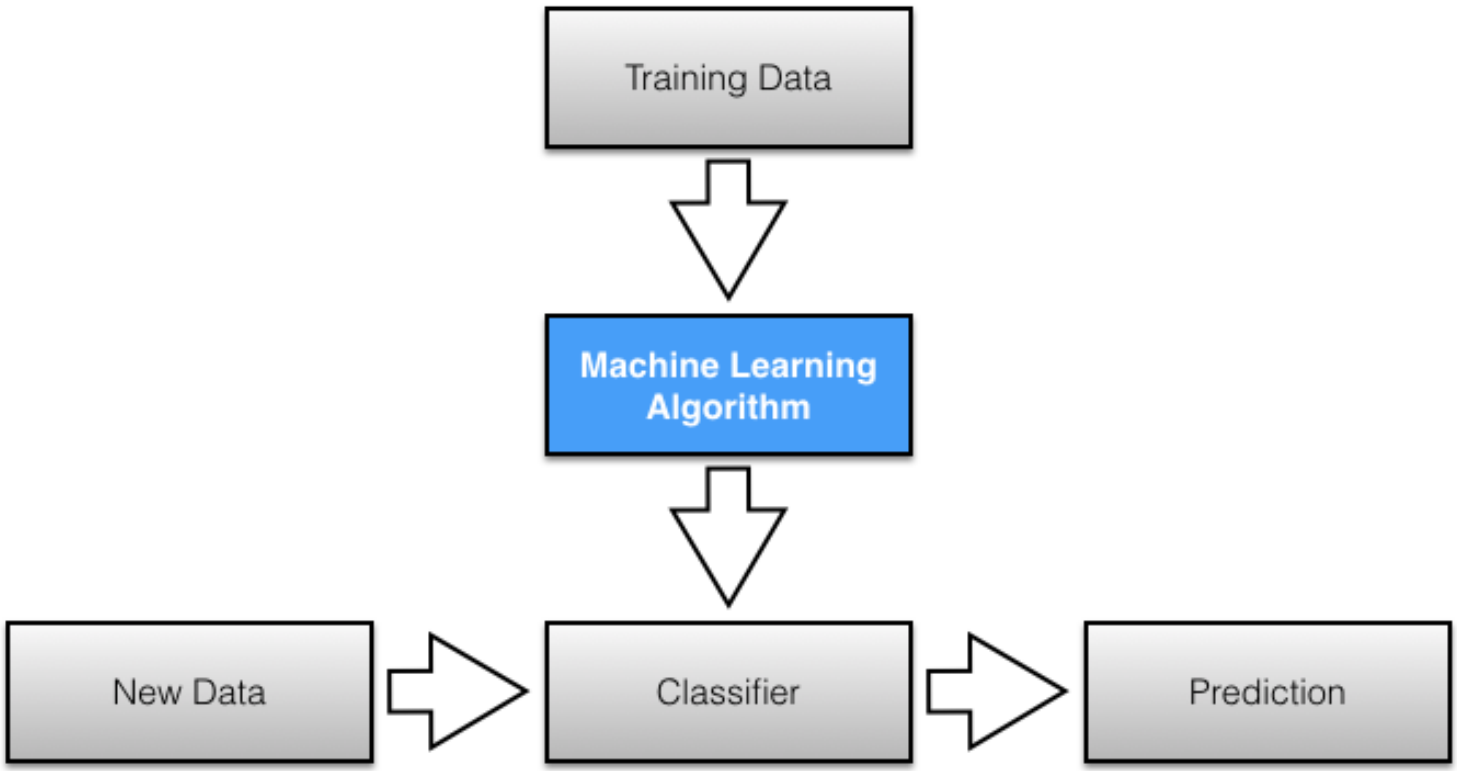


Image from [5].

Supervised vs. Unsupervised learning

- **Supervised learning**

  A training dataset with correct answers (labels) is "mined" to create a model

- **Unsupervised learning**

  Data are provided with no apriori knowledge of labels or patterns. The goal is to discover labels and patterns.

- **Semi-supervised learning**

  Knowledge from one dataset is applied to another dataset to help with mining, analysis, classification, and interpretation.

Supervised vs. Unsupervised learning techniques

	Supervised Learning	*Unsupervised Learning*
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

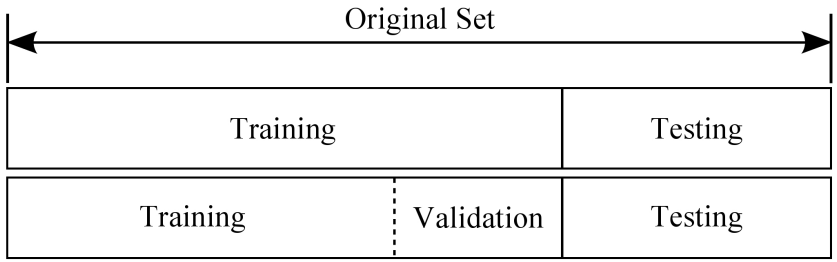
With the Titanic dataset, we will be focusing on classification.

Working with data

Supervised learning requires:

- Training data – usually about 70% of available data
- Testing data – usually about 30% of available data

Training data can also be partitioned into validation data.



Lots of different things can be done with training data

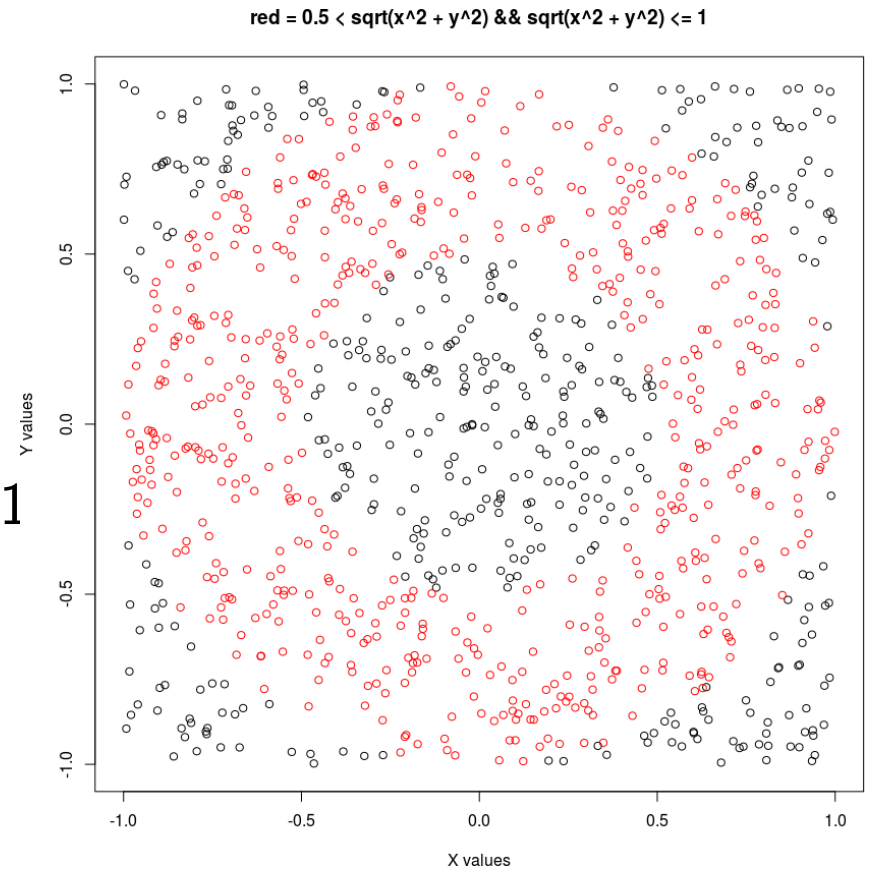
- Use as one monolithic entity
- Randomly sample data (with and without replacement)
- Divide original training data into training and validation subsets to create multiple models
- With multiple models:
  - Choose best one,
  - Use all and vote on the outcome

Sample problem space

- 1,000 data points between +/- 1
- Two classes of data points

$$color = \begin{cases} red, & \text{if } 0.5 \leq \sqrt{x^2 + y^2} \leq 1 \\ black, & \text{otherwise} \end{cases}$$

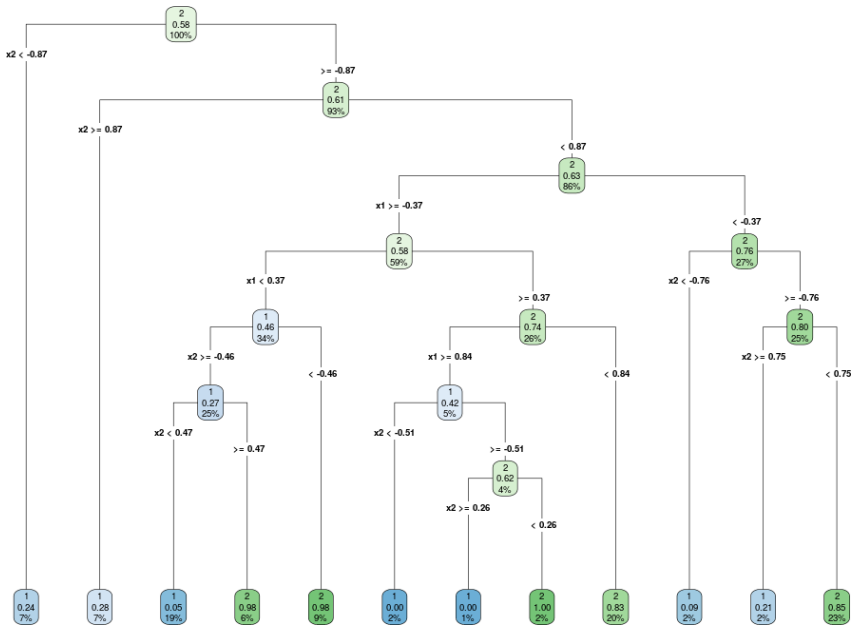
(See attached file.)



A decision tree based on sample data

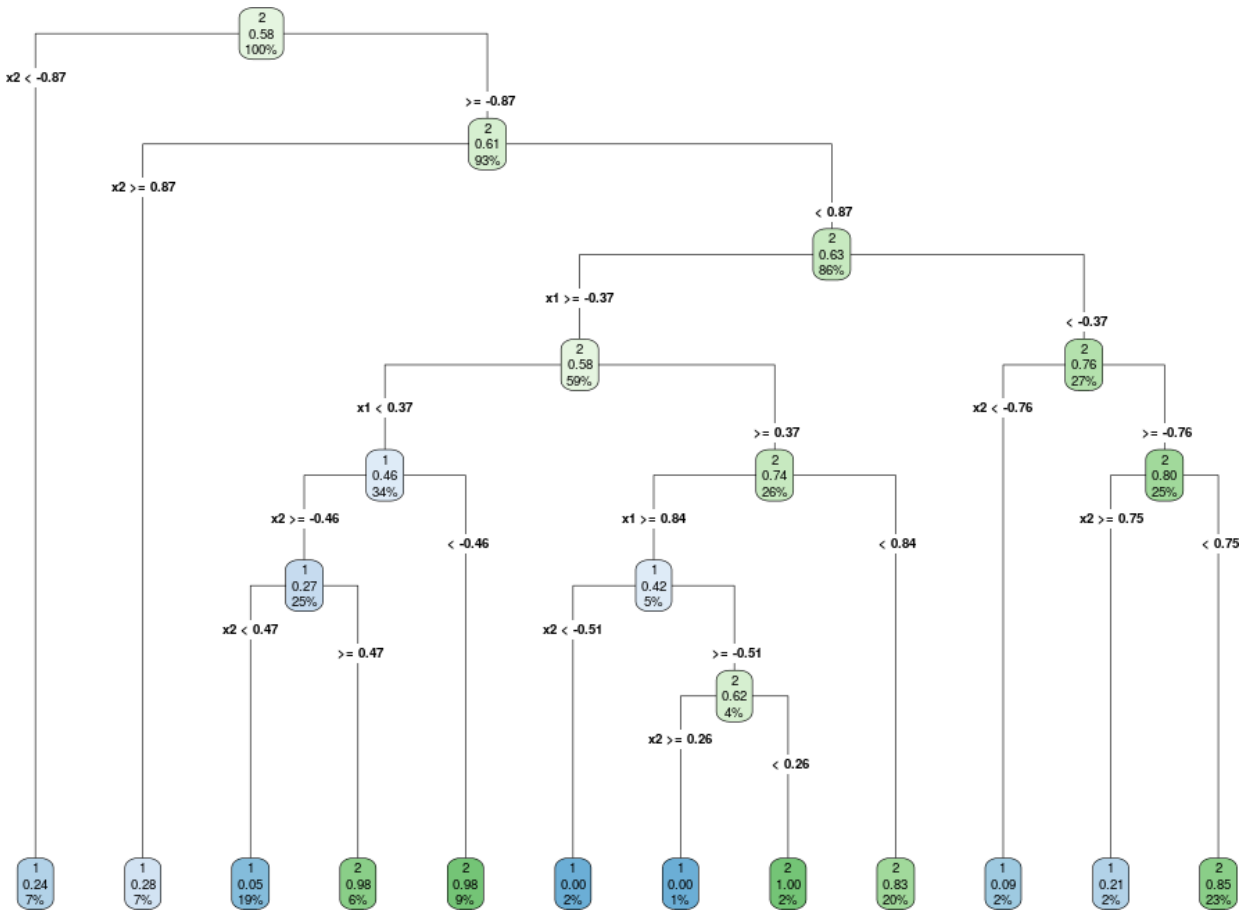
A decision tree to classify the circular data problem.

- All nodes are labeled.
- Each mode shows the percentage of the problem space they address.



Attached file.

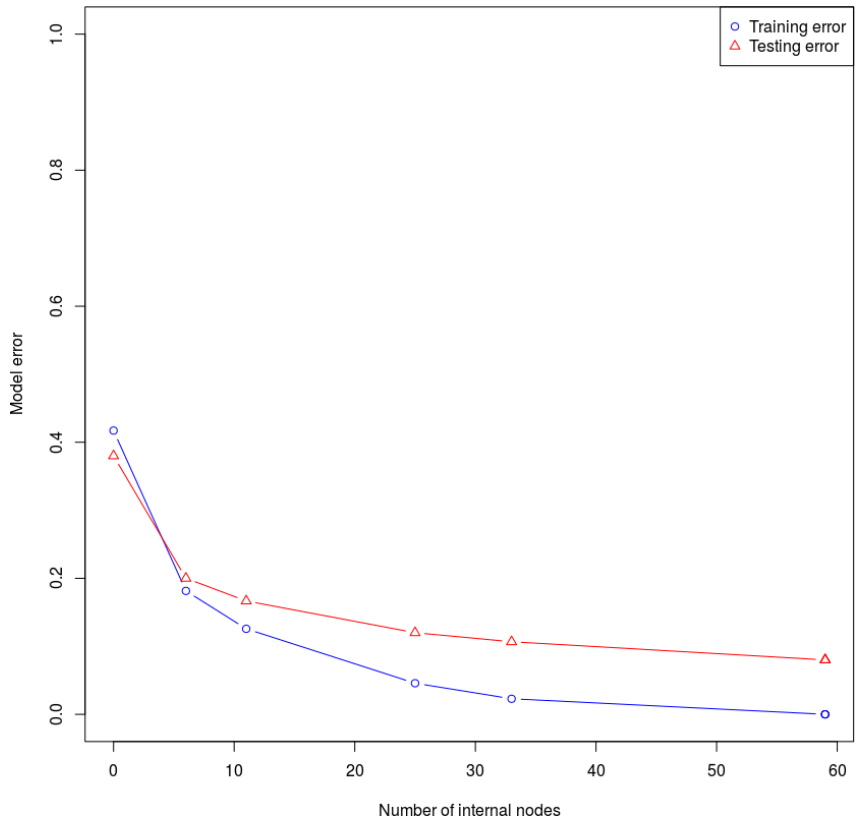
Same image.



Attached file.

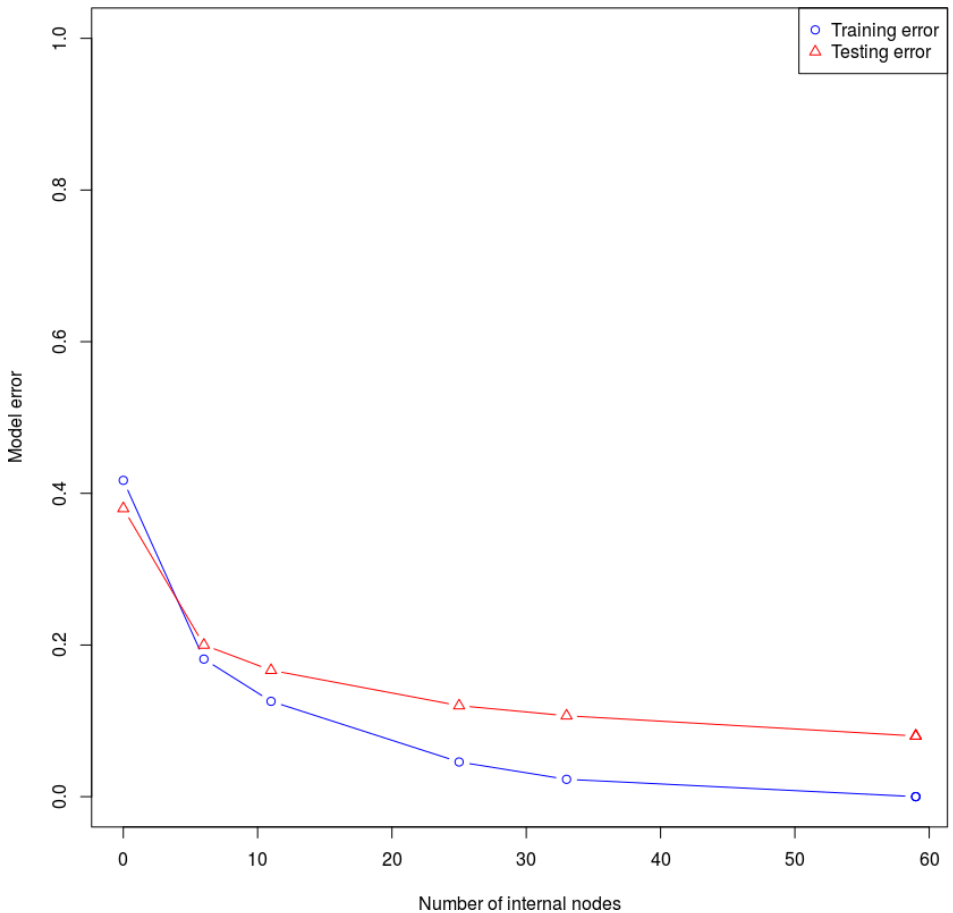
Errors in machine learning

- Total sample divided into training (70%) and testing (30%) datasets
- Training dataset was partitioned into different sized decision trees (models)
- Training and testing datasets were classified using each model
- Results were compared to the original data
- Initially models **under-fitted** until around 6 nodes
- Finally models **over-fitted** beyond 25 nodes



Training and testing errors

Same image.



Training and testing errors

A collection of decision tree techniques

rpart from the **rpart** library. "Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone." [14]

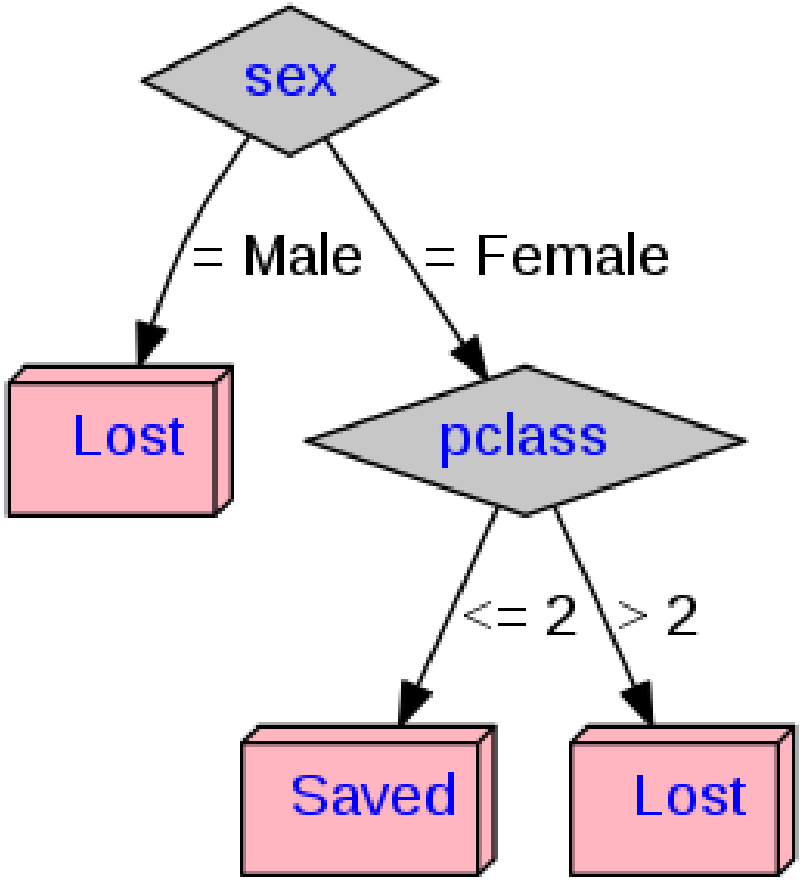
C50 from the **C50** library. "C5.0 decision trees and rule-based models for pattern recognition." [4]

Random Forest from the **randomForest** library. "Classification and regression based on a forest of trees using random inputs." [1]

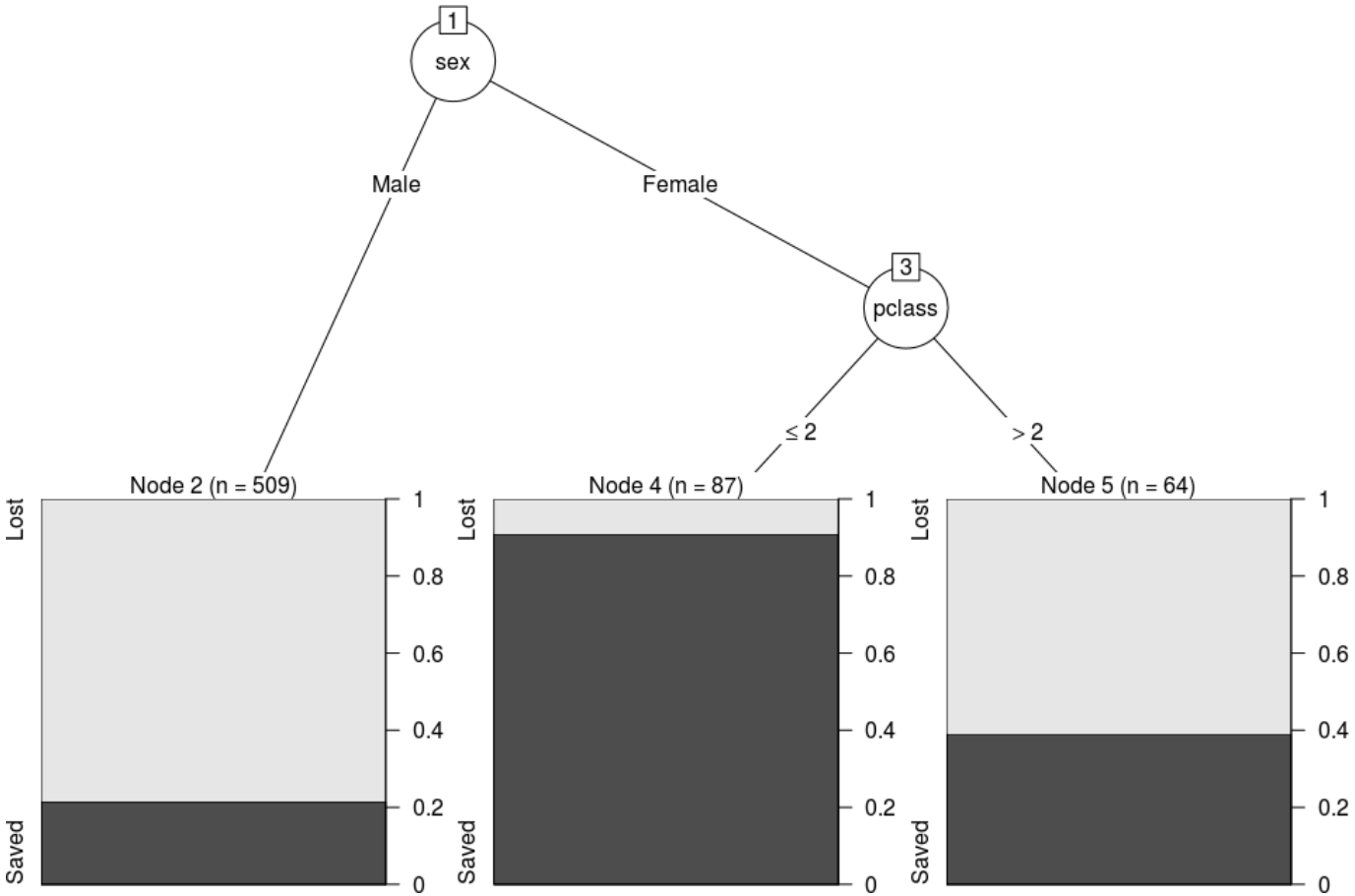
J48 from the **RWeka** library. "An R interface to Weka (Version 3.9.1). Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization." [3] The J48 algorithm is run in a pruned and unpruned mode.

These and additional techniques to be covered in detail later.

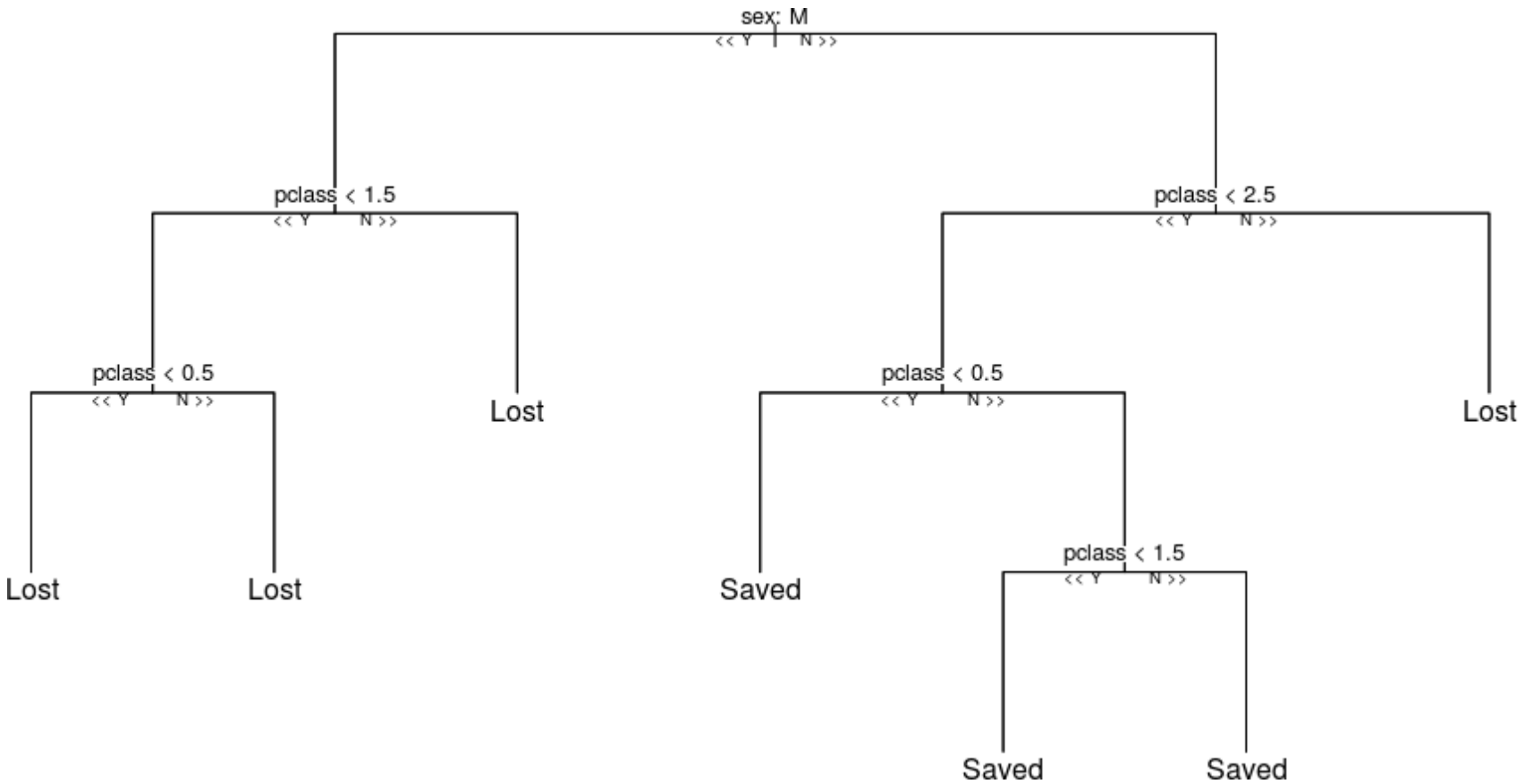
C50 decision tree



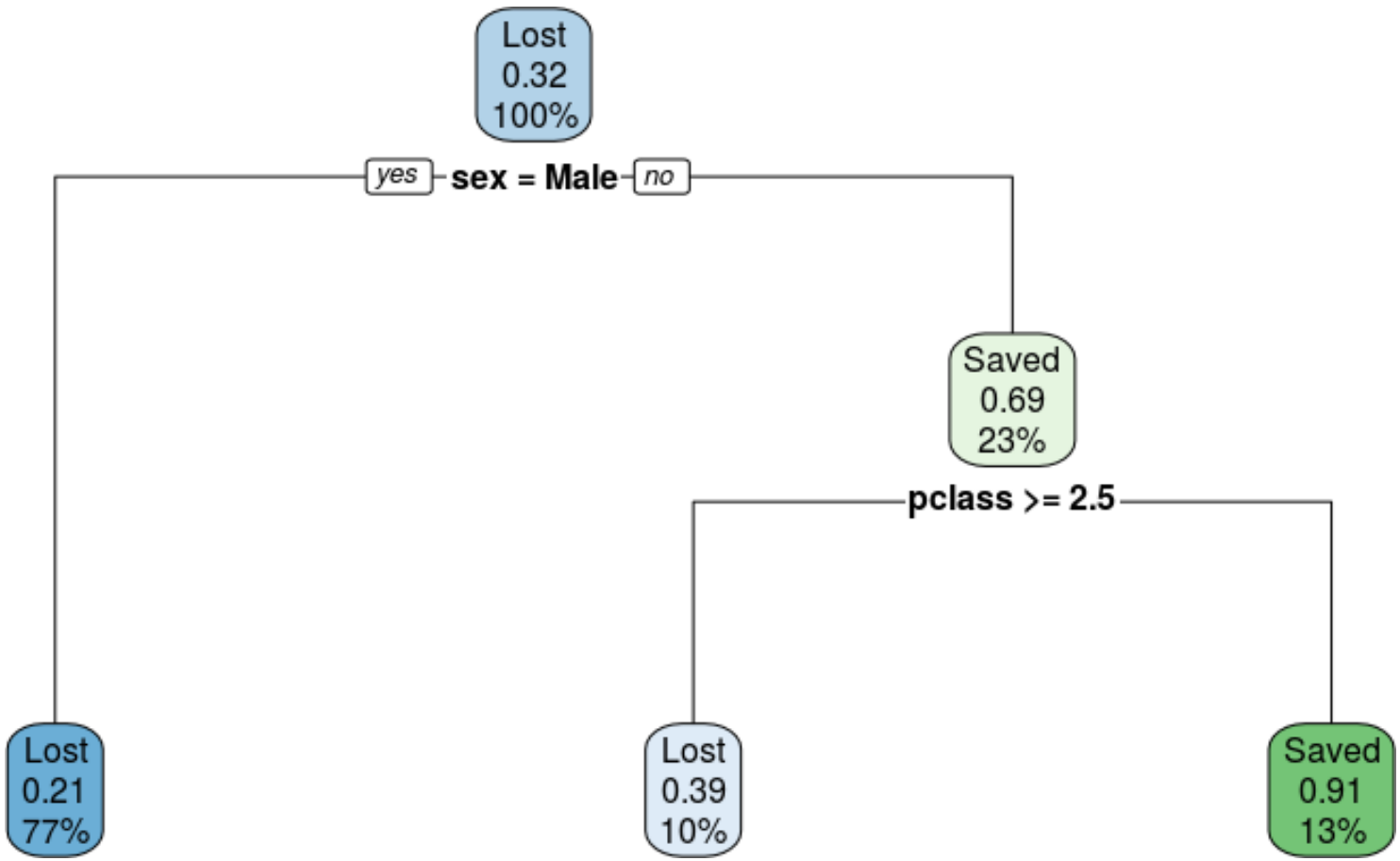
Random Forest decision tree



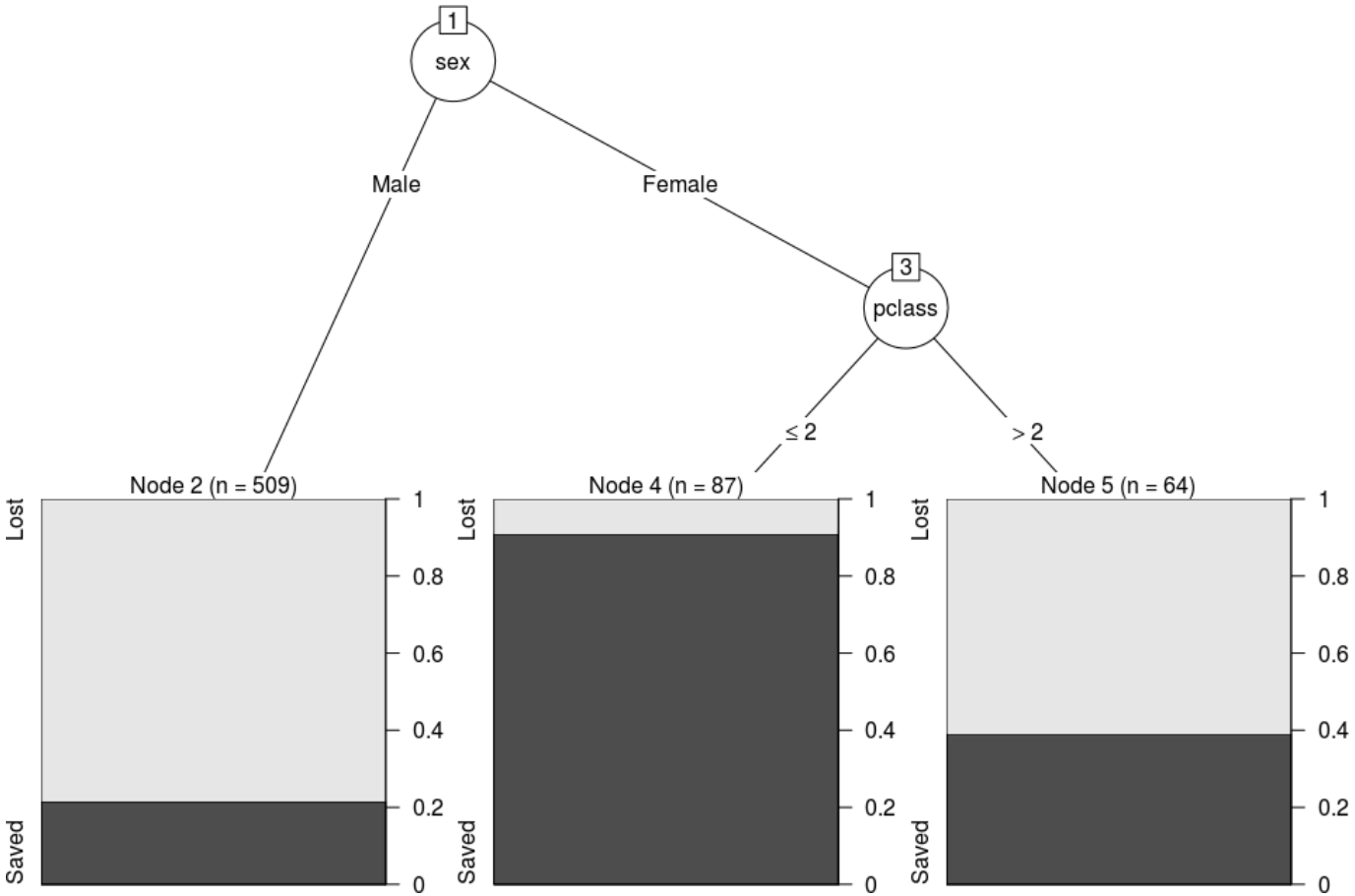
rpart decision tree



J48 (unpruned) decision tree

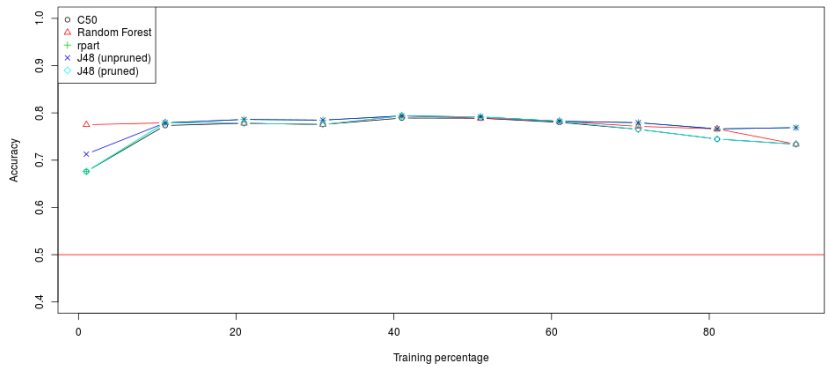


J48 (pruned) decision tree



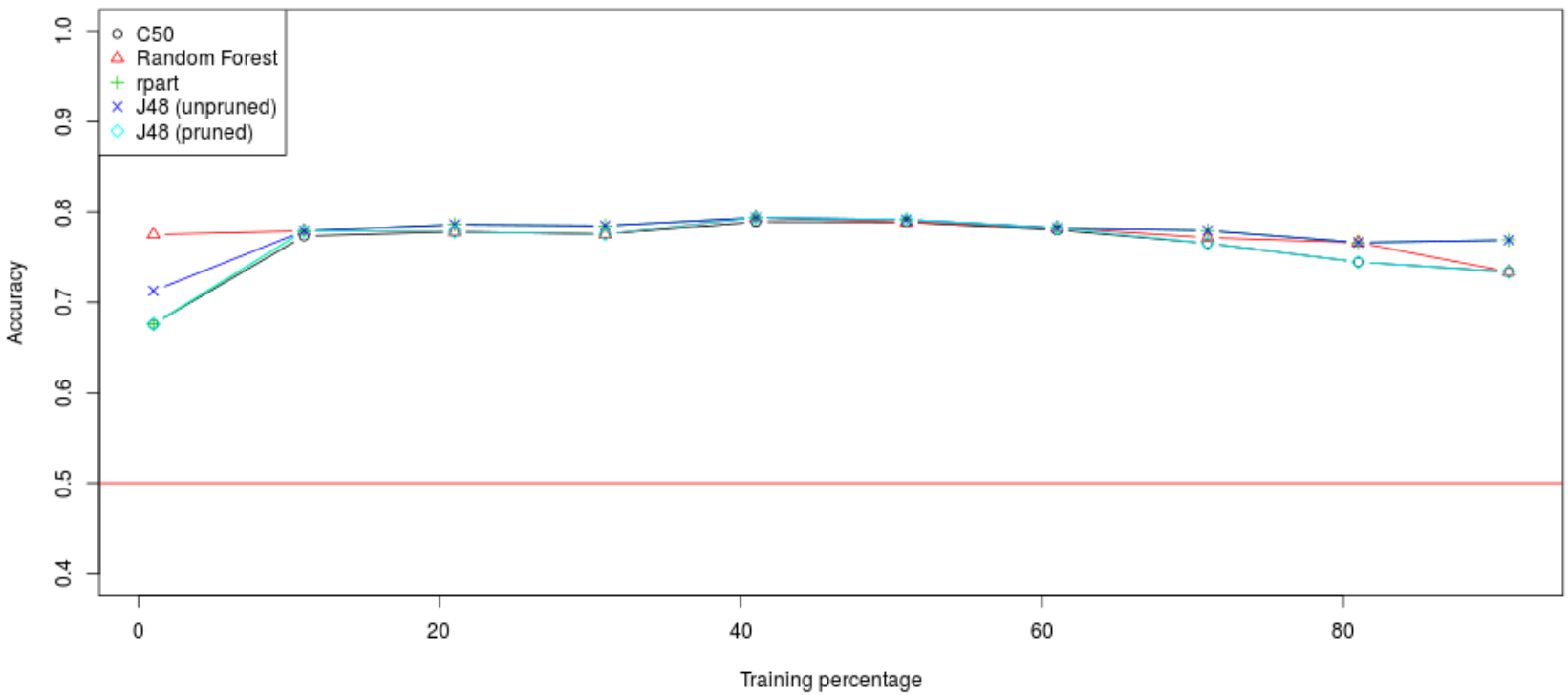
Accuracy based on training percentage

The horizontal line at 50% represents the accuracy that would be achieved based on using an unbiased coin to decide the likelihood of survival.



Using training percentages from about 10 to 60 result in all algorithms having nearly identical accuracies. Below 10%, the Random Forest approach appears best.

Same image.

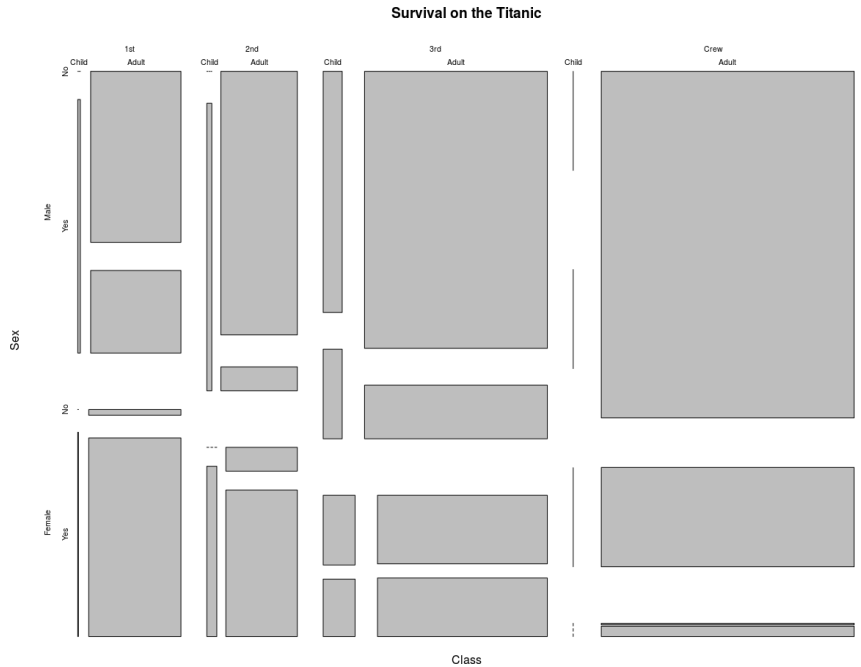


Using training percentages from about 10 to 60 result in all algorithms having nearly identical accuracies. Below 10%, the Random Forest approach appears best.

Simple mosaic from the `titanic` package

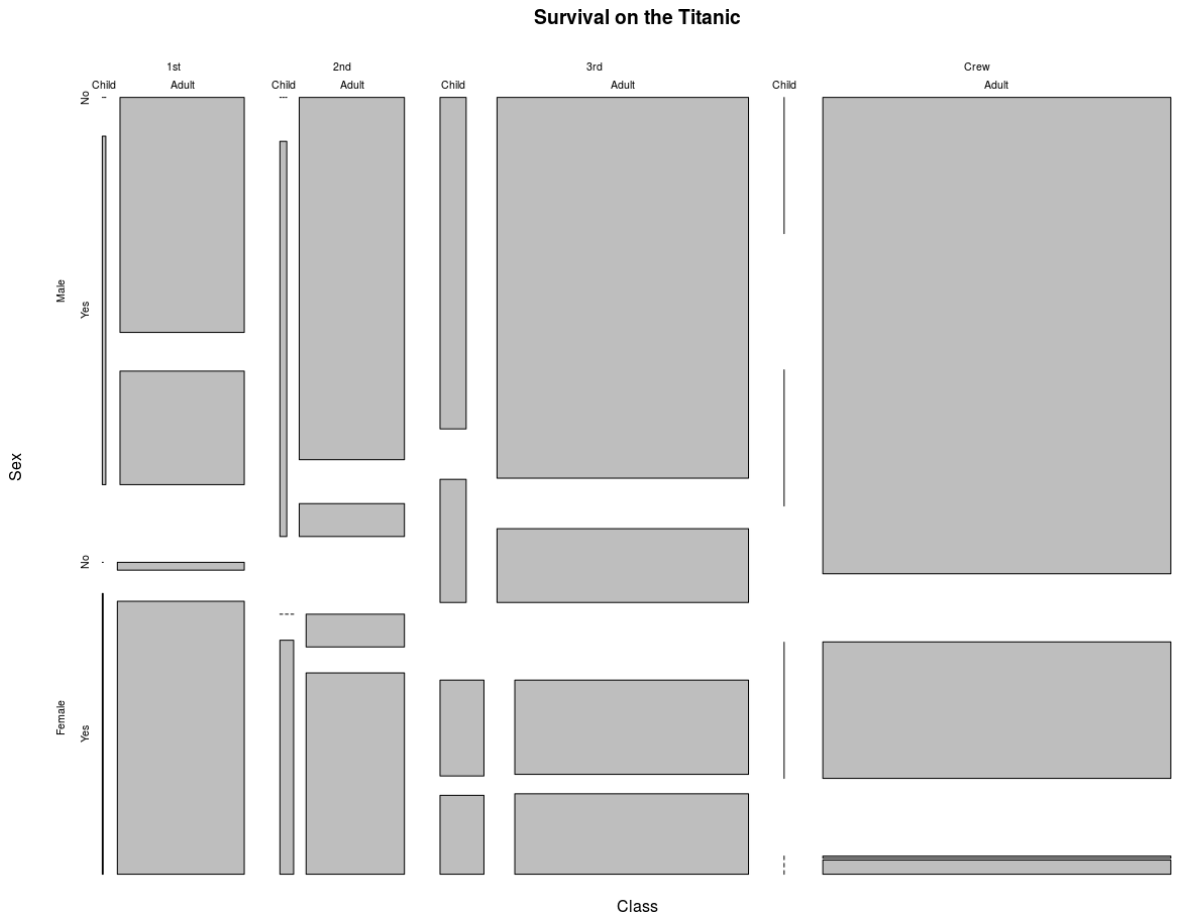
Sometimes you don't need a lot of code.

```
library(titanic)
library(graphics)
mosaicplot(Titanic, main =
"Survival on the Titanic")
```



A summary of all personnel on the RMS Titanic broken down by gender, by survival or not, and class. It is interesting to look at the data and consider the adage: "women and children first."

Same image.



A summary of all personnel on the RMS Titanic broken down by gender, by survival or not, and class. It is interesting to look at the data and consider the adage: "women and children first."

Some simple exercises to get familiar with data analysis

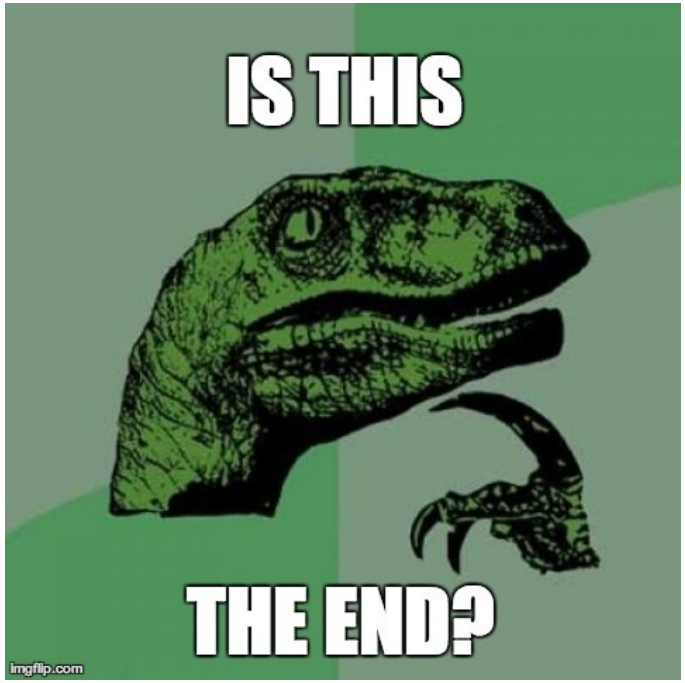
1. Using the Titanic report as a guide, create a recursive partition decision tree modeling survival based on sex and number of siblings
2. Create a recursive partition decision tree modeling survival based on all available data

Q & A time.

Q: How many Harvard MBA's does it take to screw in a light bulb?
A: Just one. He grasps it firmly and the universe revolves around him.

What have we covered?

- All of the decision tree algorithms tested had comparable results (~76% accuracy) when the training dataset was between 10 and 60% of the entire dataset.
- Random forest performed most consistently over the widest range of training percentages of all tested algorithms.



Next: LPAR Chapter 2, basic data visualization

References (1 of 5)

[1] Leo Breiman, randomforest: Breiman and cutlers random forests for classification and regression, http://stat-www.berkeley.edu/users/breiman/RandomForests, 2006.

[2] Jr. Frank E. Harrell, Titanic Data, http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html, 2002.

[3] K Hornik, A Zeileis, T Hothorn, and C Buchta, RWeka: an R interface to Weka, R package version 0.4-32 (2017).

[4] M Kuhn, S Weston, N Coulter, M Culp, and R Quinlan, C5.0 Decision Trees and Rule-Based Models, R Package Version 0.1. 0 **24** (2015).

References (2 of 5)

[5] Sebastian Raschka, Predictive modeling, supervised machine learning, and pattern classification, http://sebastianraschka.com/Articles/2014_intro_supervised_learning.html, 2014.

[6] Encyclopedia Titanica Staff, Encyclopedia Titanica, Titanic Facts, History and Biography, https://www.encyclopedia-titanica.org/, 2017.

[7] History Staff, Titanic photo galleries, http://www.history.com/topics/titanic/pictures/titanic-before-and-after/rms-sailing-from-southampton, 2017.

References (3 of 5)

[8] ISM Staff, Titanic Survivor, Titanic Passenger List Booklet, http://www.phillyseaport.org/web_exhibits/mini_exhibits/titanic_passenger_list/titanic_passenger_list-object-passenger_list.html, 2017.

[9] OceanGate Staff, Titanic Survey Expedition: 2018, http://www.oceangate.com/expeditions/titanic-survey-2018.html, 2017.

[10] Southampton Staff, Titanic crew list, http://www.plimsoll.org/Southampton/Titanic/titaniccrewlist/Default.asp, 2017.

References (4 of 5)

[11] Titanic Facts Staff, Titanic Passenger List, http://www.titanic-facts.com/titanic-passenger-list.html, 2017.

[12] Wikipedia Staff, Sinking of the RMS Titanic, https://en.wikipedia.org/wiki/Sinking_of_the_RMS_Titanic, 2017.

[13] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Introduction to Data Mining, Pearson Education India, 2006.

[14] Terry Therneau, Beth Atkinson, and Brian Ripley, rpart, Available at CRAN. R-project. org/package= rpart. Accessed May (2015).

References (5 of 5)

[15] Maria Dolores Ugarte, Ana F Militino, and Alan T Arnholt, Probability and Statistics with R, CRC Press, 2008.

Files of interest

1. A short report about the Titanic
2. titanic3 data
3. titanic3 meta data
4. Example circular binary classifier