Surviving the Titanic

Miguel, Nicholas, Quinton

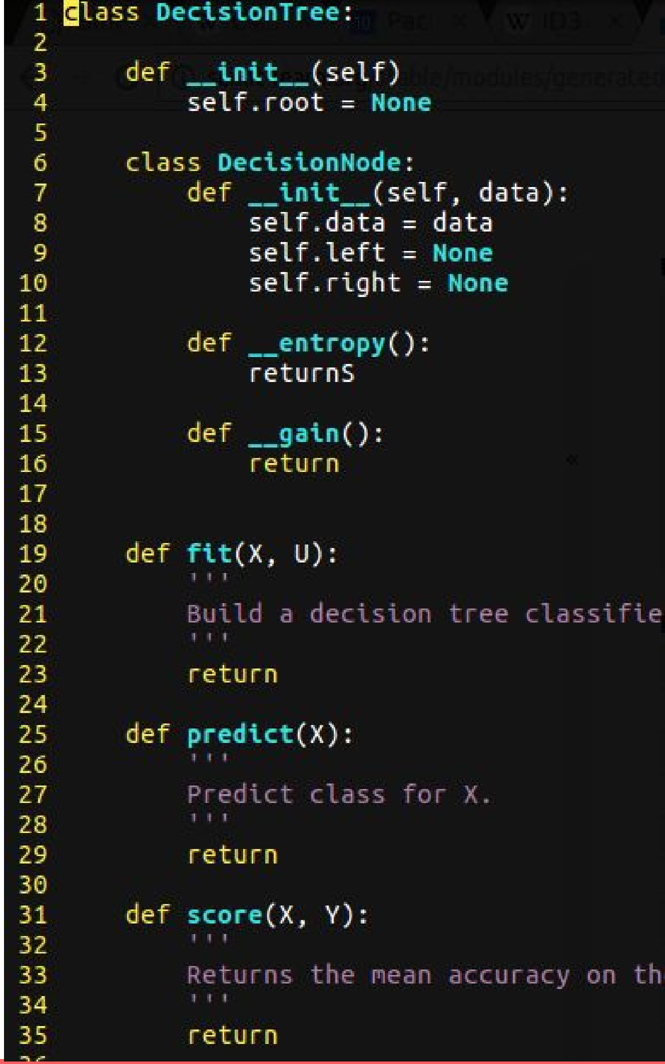
Introduction

- The RMS Titanic sank on 15 April 1912, after colliding with an iceberg during its maiden voyage from Southampton to New York.
- Out of 2224 passengers and crew, 1502 lost their lives.
- We would like to identify more factors that improved the likelihood of survival. Furthermore, we will map these factors to an appropriate feature space.
- We will implement and train a decision tree to analyze what sorts of people were likely to survive.
- We will compare the accuracy of our decision tree to the accuracy of scikit-learn’s bagging meta-estimator with a decision tree classifier as a base estimator.

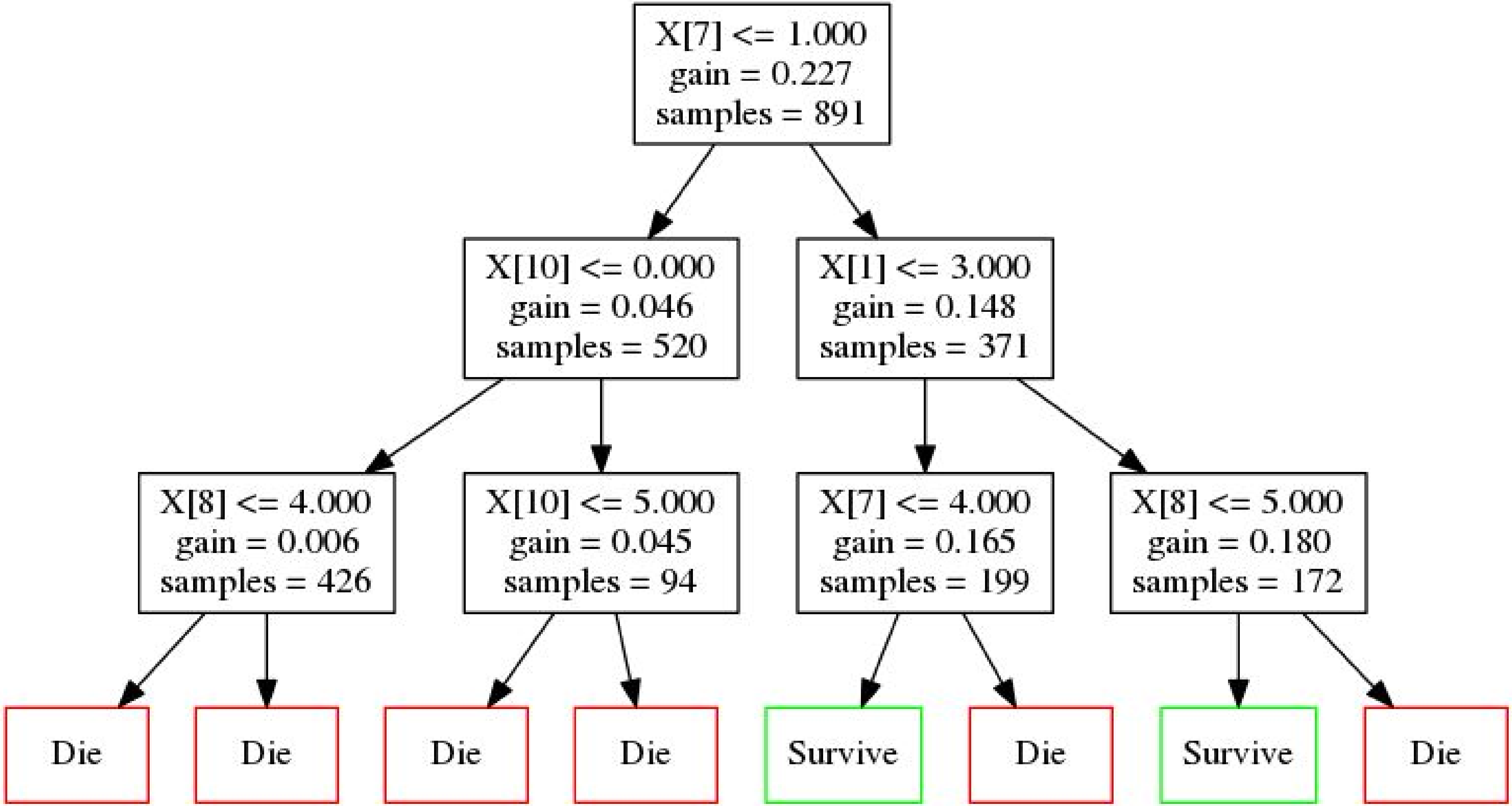
Design

Model

- Dataset
  - Loads .csv file
  - Cleans data
- Feature Space
  - Uses a Dataset object to create a 2-dimensional array that represents a feature space
- Decision Tree
  - CART algorithm
- Bagging Meta-Estimator
  - scikit-learn
  - Decision tree classifier as a base estimator.
  - Bagging is used to reduce the variance of a base estimator



```
class DecisionTree:

    def __init__(self)
        self.root = None

    class DecisionNode:
        def __init__(self, data):
            self.data = data
            self.left = None
            self.right = None

        def __entropy():
            returnS

        def __gain():
            return


    def fit(X, U):
        '''
        Build a decision tree classifie
        '''
        return

    def predict(X):
        '''
        Predict class for X.
        '''
        return

    def score(X, Y):
        '''
        Returns the mean accuracy on th
        '''
        return
```


X[7] <= 1.000
gain = 0.227
samples = 891
X[10] <= 0.000
gain = 0.046
samples = 520
X[1] <= 3.000
gain = 0.148
samples = 371
X[8] <= 4.000
gain = 0.006
samples = 426
X[10] <= 5.000
gain = 0.045
samples = 94
X[7] <= 4.000
gain = 0.165
samples = 199
X[8] <= 5.000
gain = 0.180
samples = 172
Die
Die
Die
Die
Survive
Die
Survive
Die

Features

Training Data

Contents

Variable	Definition	Key
passengerid	Passenger ID	
survival	Survival	0=No, 1=Yes
pclass	Ticket Class	1=1st, 2=2nd, 3=3rd
name	Name (with Title)	
sex	Sex	
age	Age in Years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket Number	
fare	Passenger Fare	
cabin	Cabin Number	
embarked	Port of Embarkation	C=Cherbourg, Q=Queenstown, S=Southampton

Feature Space

From Dataset

1. Pclass
2. Sex
3. Age
4. Sibsp
5. Parch
6. Embarked

Generated

1. Family size (parch + sibsp + 1)
2. Age Interval
   a. 1 if age < 10
   b. 0 if 10 <= age <= 60
   c. -1 if age > 60
3. Title (Mr. , Mrs. , Miss. , Dr. , etc.)
4. Deck Level (A, B, C,...)

Data Processing

Cleaned in Java

- Missing Information
  - Filled with -1
- Read in the .csv file
- Output the .json feature space and survival labels

Shuffled in Python

- Read in the .json feature space and survival labels
- Shuffled the indices
- Returned them for training

Interesting Anomalies

Title

- There is only one “Ms.” in the training dataset, and she was a widow
- The title “Jonkheer.”

Name

- Longest Name:
  - Penasco y Castellana, Mrs. Victor de Satode (Maria Josefa Perez de Soto y Vallejo)
  - Age: 17

Deck

- Mr. Stephen Weart Blackwell had a cabin on Deck T or Top Deck
  - Was prescribed travel by his doctor for his wellbeing
  - Closest to lifeboats
  - Didn’t survive

Training

Accuracy Metric

Let $f : \mathbb{N} \times \mathbb{N} \to \mathbb{Q} \cap [0, 1]$ be the function defined by

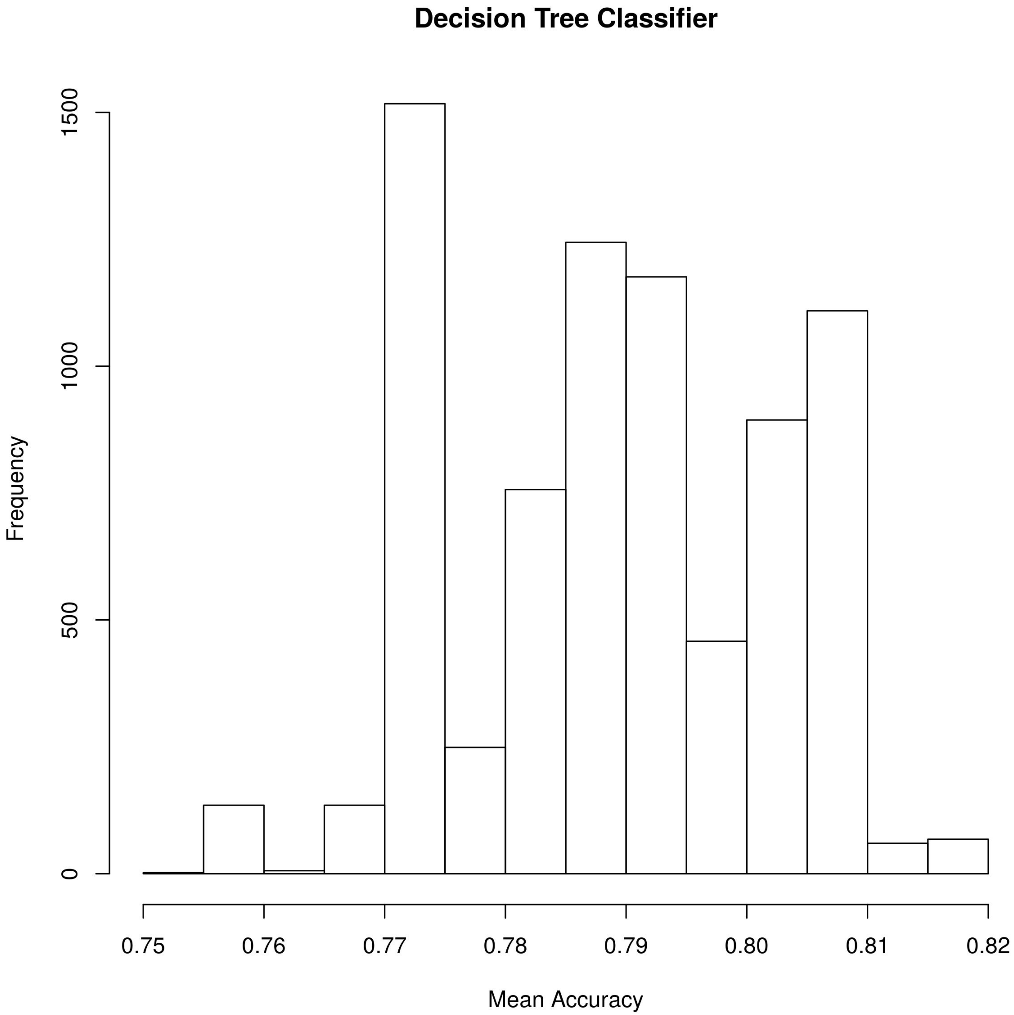
$$f[(a, b)] := \frac{a}{a + b}$$

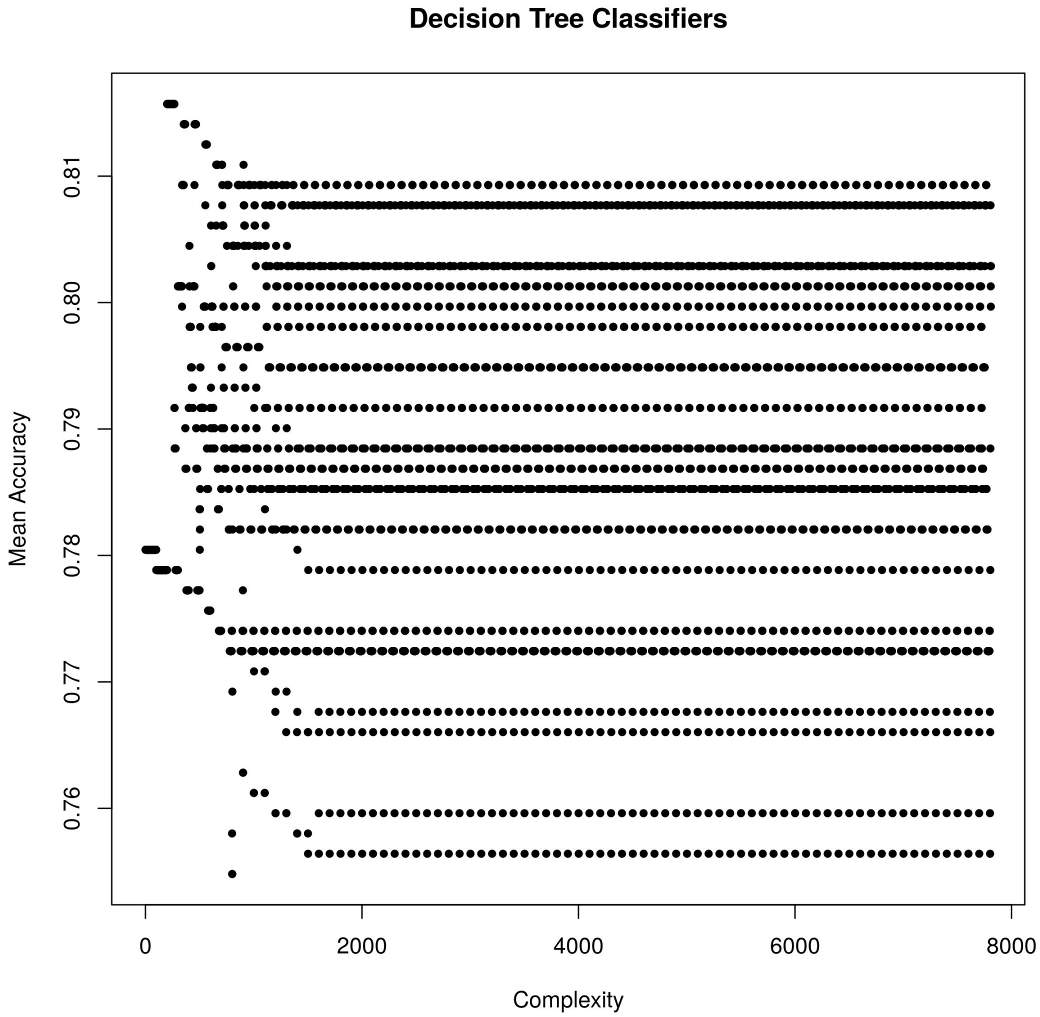
where *a* is the number of correct predictions and *b* is the number of incorrect predictions.

Decision Tree Classifier

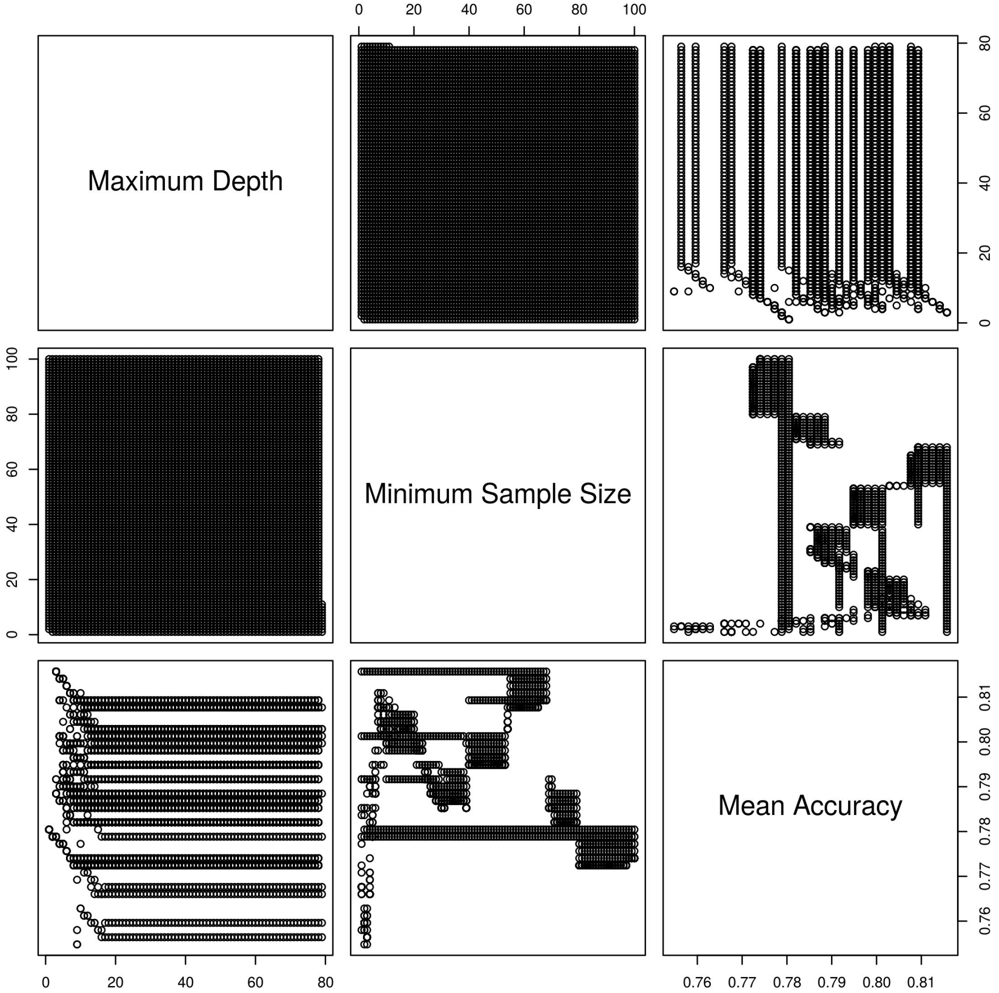
Parameters

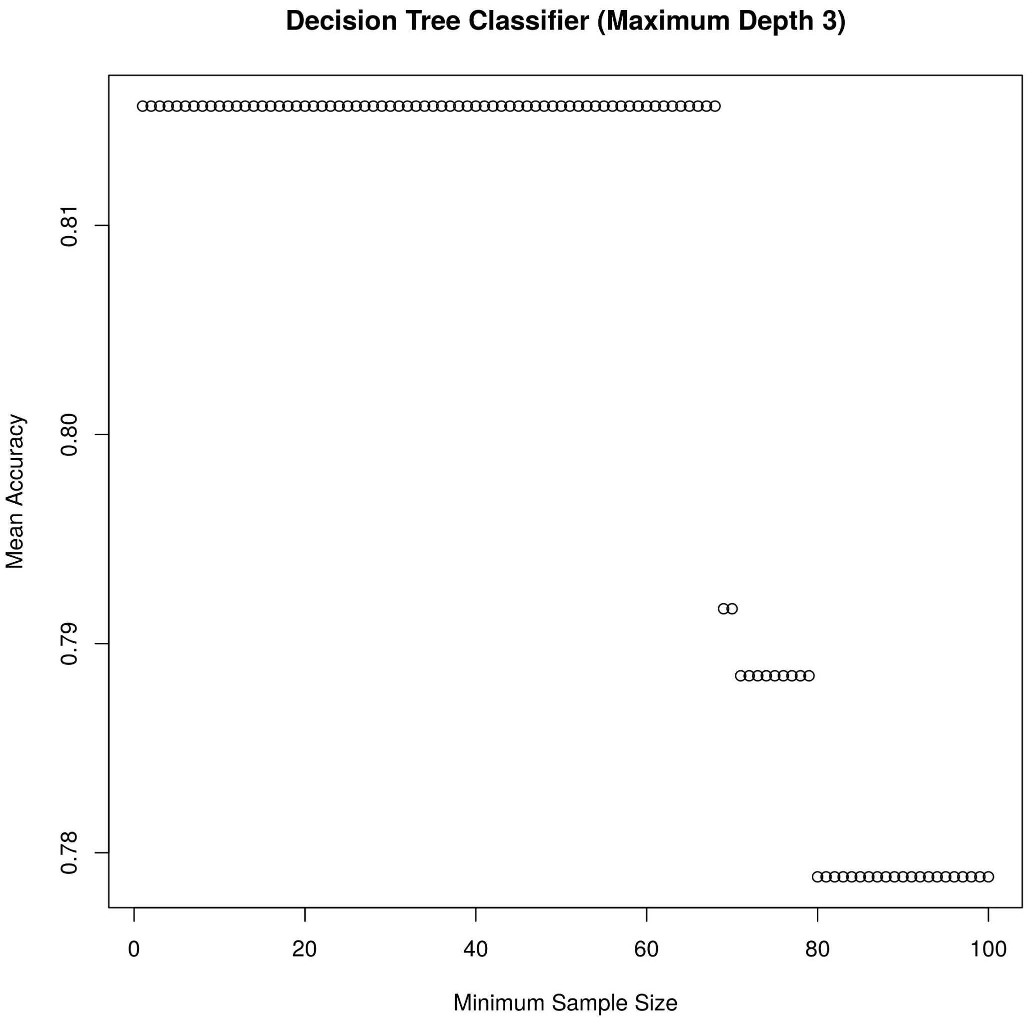
- Maximum depth: 1 - 100
- Minimum sample size: 1 - 100


Decision Tree Classifier
Frequency
0
500
1000
1500
0.75
0.76
0.77
0.78
0.79
0.80
0.81
0.82
Mean Accuracy



Performance Decreased with Complexity


Maximum Depth
Minimum Sample Size
Mean Accuracy
0
20
40
60
80
100
0.76
0.77
0.78
0.79
0.80
0.81



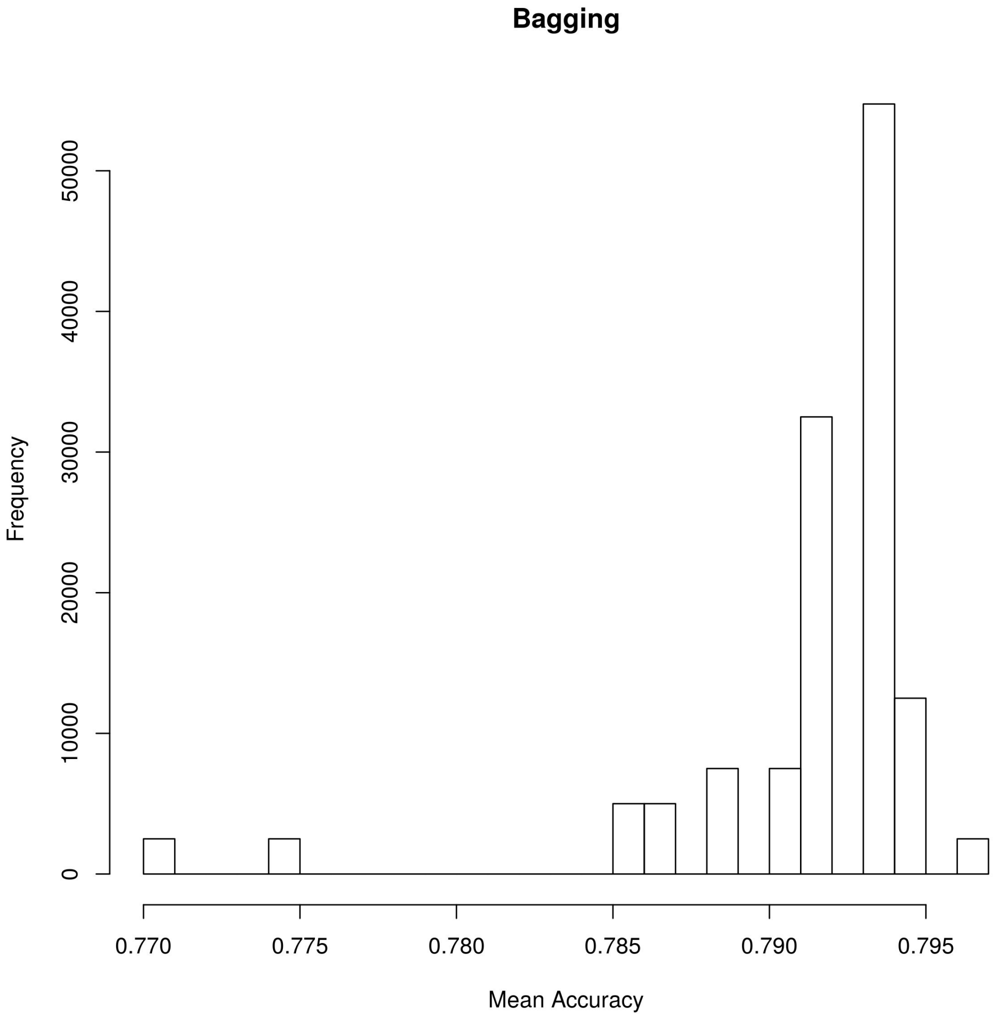
Best Models

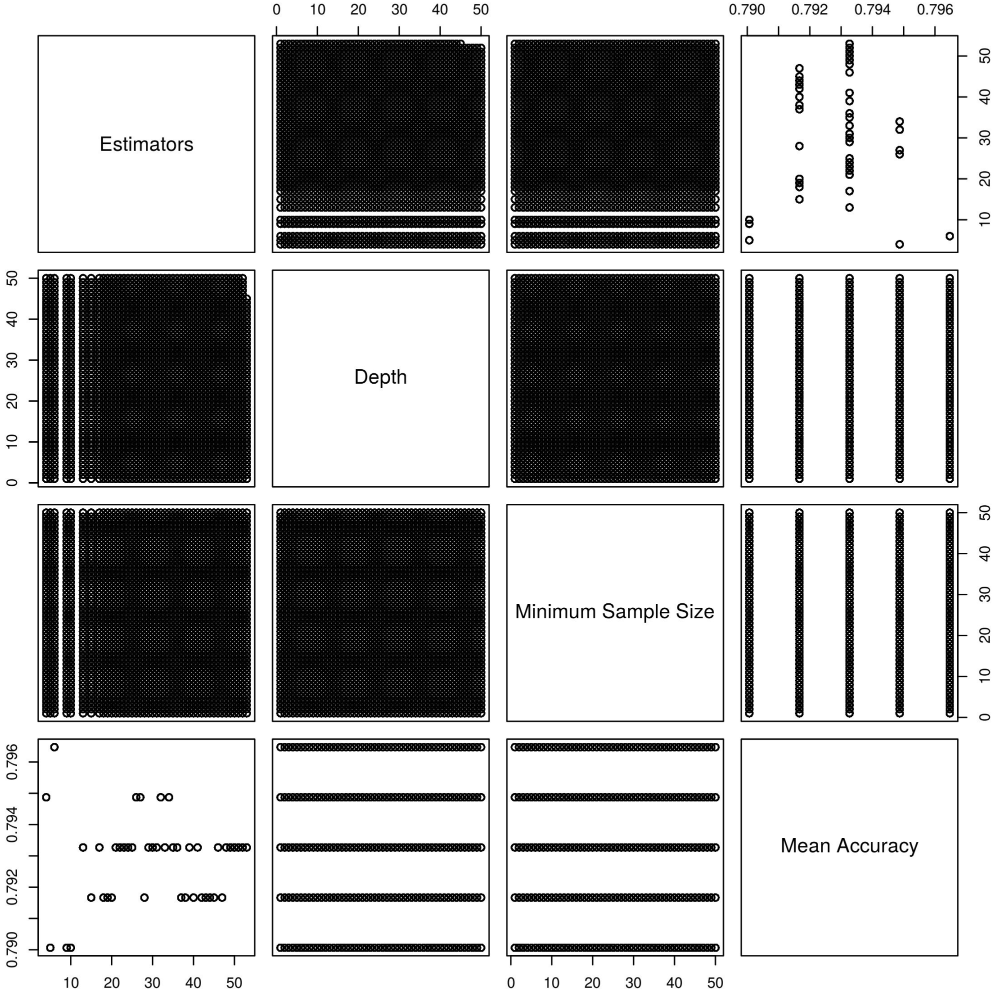
- Mean accuracy of at least 0.815
- Maximum tree depth of 3
- Minimum sample size less than 68

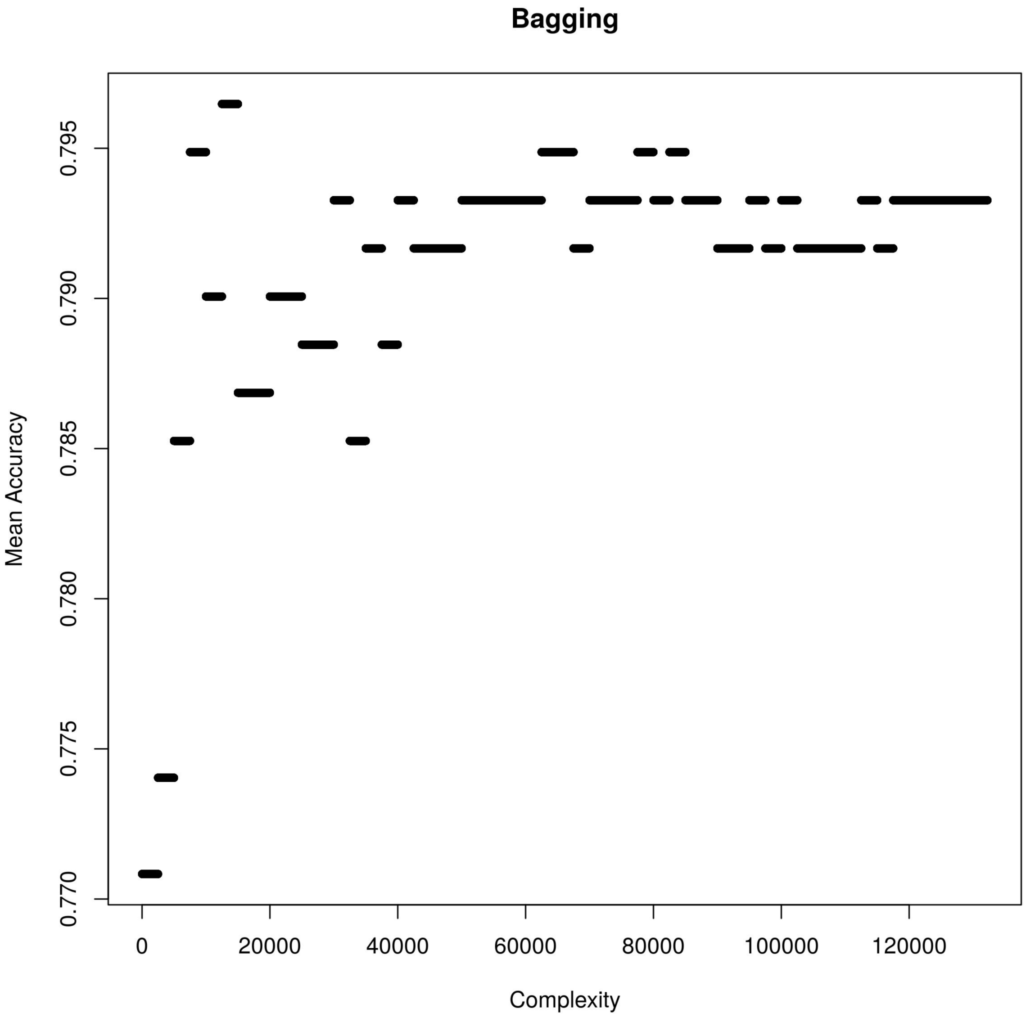
Bagging Classifier

Parameters

- Number of estimators: 53 out of 300
- Random state: 311
- Maximum tree depth: 50
- Minimum sample size: 50


Bagging
Frequency
0
10000
20000
30000
40000
50000
0.770
0.775
0.780
0.785
0.790
0.795
Mean Accuracy


Estimators
Depth
Minimum Sample Size
Mean Accuracy
0
10
20
30
40
50
0.790
0.792
0.794
0.796



Best Models

- 0.770 mean accuracy
- 6 estimators
- Maximum tree depth: 1-50
- Minimum sample size: 1-50

Testing

Decision Tree Classifiers

0.8157

Training

0.7940

Testing

Bagging Classifiers

0.7965

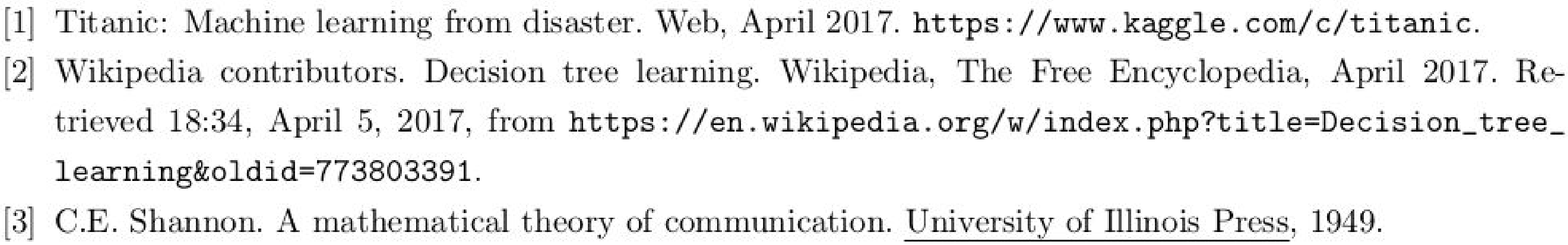
Training

0.8215

Testing

Demo

References



[1] Titanic: Machine learning from disaster. Web, April 2017. https://www.kaggle.com/c/titanic.

[2] Wikipedia contributors. Decision tree learning. Wikipedia, The Free Encyclopedia, April 2017. Retrieved 18:34, April 5, 2017, from https://en.wikipedia.org/w/index.php?title=Decision_tree_learning&oldid=773803391.

[3] C.E. Shannon. A mathematical theory of communication. University of Illinois Press, 1949.

Questions