COSC 757 Data Mining Assignment 2

Mary Snyder

Department of Computer & Information Sciences  
College of Science & Mathematics  
Towson University

msnyde8@students.towson.edu

**ABSTRACT**

In this paper, I will be exploring a dataset to become more familiar with data classification through the COSC 757 Data Mining Assignment 2.

**Categories and Subject Descriptors**

H.2.8 **[Database Management]** Database Applications – *Data mining*

**Keywords**

Classification; Multivariate; Categorical; Decision Tree Classification; Naïve Bayes Classification; Random Forest Classification; Training and Testing; Holdout Method; Cross-Validation; Bootstrap; Accuracy; Error Rate; Sensitivity; Specificity; Precision; Recall; F Measure;

# INTRODUCTION

## Dataset

I chose a dataset from the UCI Machine Learning Repository classified for the task of Classification. This Balance Scale Weight & Distance dataset was generated to model psychological experiment results. The dataset contains information regarding a scale either tipped to the right, tipped to the left, or balanced. There are 625 instances with a distribution of 49 balanced, 288 left tipped, and 288 right tipped classifications. The dataset has no missing values and contains 5 attributes: Class Name, Left-Weight, Left-Distance, Right-Weight, and Right-Distance.

## Objective of Analysis

Classification predicts categorical class labels. First a model is constructed to classify data based on a training set and the values (class labels) for a classifying attribute. The model can be represented as classification rules, decision trees, or even mathematical formulae. Second the model is used to classify new/unclassified data. A test set can be used, which is independent of the training set of data.

First, the classification algorithm examines the data set values for the predictor and the already classified target variables in the training set. This allows the algorithm to learn which values of the predictor variables are associated with values of the target variable. Now that the model has been built from the training set, the algorithm examines new records for which the target variable is unknown. Using the classifications learned from the training set, the algorithm classifies the new data.

# METHODOLOGY

Included below in Table 1, is a summary of the variables for the dataset including the minimum, maximum, mean, median, and standard deviation for each the field values. The mean and median of acceleration are extremely close to each other (median of 15.50 and mean of 15.52), which is an indicator of possible symmetry. By the same token, the mean and median of displacement (median of 151 and mean of 194.8), horsepower (median 95 and mean 105.08), and weight (median 2822 and mean 2979) are fairly far apart from each other indicating they are not symmetric.

## Preprocessing

The dataset description give the correct way to find the classification for the values as the great of (left-distance \* left-weight) and (right-distance \* right-weight), with equal values meaning it is balanced. I translated this into the following formula for use in classification:

(left-distance \* left-weight) – (right-distance \* right-weight)

where a result less than zero indicates left tripped scale, a result greater than zero indicates right tipped scale, and a result equal to zero indicates a balanced scale.

## Experiment Design

### Holdout Method

Given data is randomly partitioned into two independent sets:

Training set (e.g. 2/3) for model construction

Test set (e.g. 1/3) for accuracy estimation

## Classification Methods

### Decision Tree Classification

Decision Tree classification uses a flowchart-like tree structure for classification. Each tree node represents a test on an attribute, each tree branch represents an outcome of the attribute test, and each tree leaf node has a classification label.

The tree is constructed in a top-down recursive divide-and-conquer manner. The training examples are at the root at the start of the algorithm and partitioned reclusively based on provided selected attributes. Partitioning is stopped when all the samples for a given node belong to the same class, there are no remaining attributes for further partitioning, and/or there are no samples left.

### Naïve Bayes Classification

Naïve Bayes classification uses simple probabilistic classifiers based on Bayes’ theorem. It assumes there is a strong (naïve) independence between the attributes. In general terms, Bayes’ theorem describes the probability of an event based on an already observed event.

Bayes’ theorem is formally written as follows:

Given training data **X**, *posteriori probability of a hypothesis* H, P(H|**X**):

The theorem is used to determine the posteriori probability P(H|X) that the hypothesis holds given the observed data sample X, or in simpler terms the likelihood of the hypothesis given prior evidence, for each classification. The classification with the highest probability is assigned for that data.

### Random Forest Classification

Random Forest classification is similar to the decision tree classification. The algorithm takes each classifying attribute and generates a decision tree using a random selection of attributes at each node to determine the split. Each sample is fed through each of the decision trees to determine a result classification and each of the resulting classifications are tallied with the most popular classification being assigned.

More formally, each tree is constructed as follows:

Let N be the number of cases in the training set.

Let M be the number of input variables

For a number m < M (constant throughout the forest growing), select m variables at random out of the input variables for each node and use them to split the node.

# RESULTS

## Analysis

### Decision Tree Classification

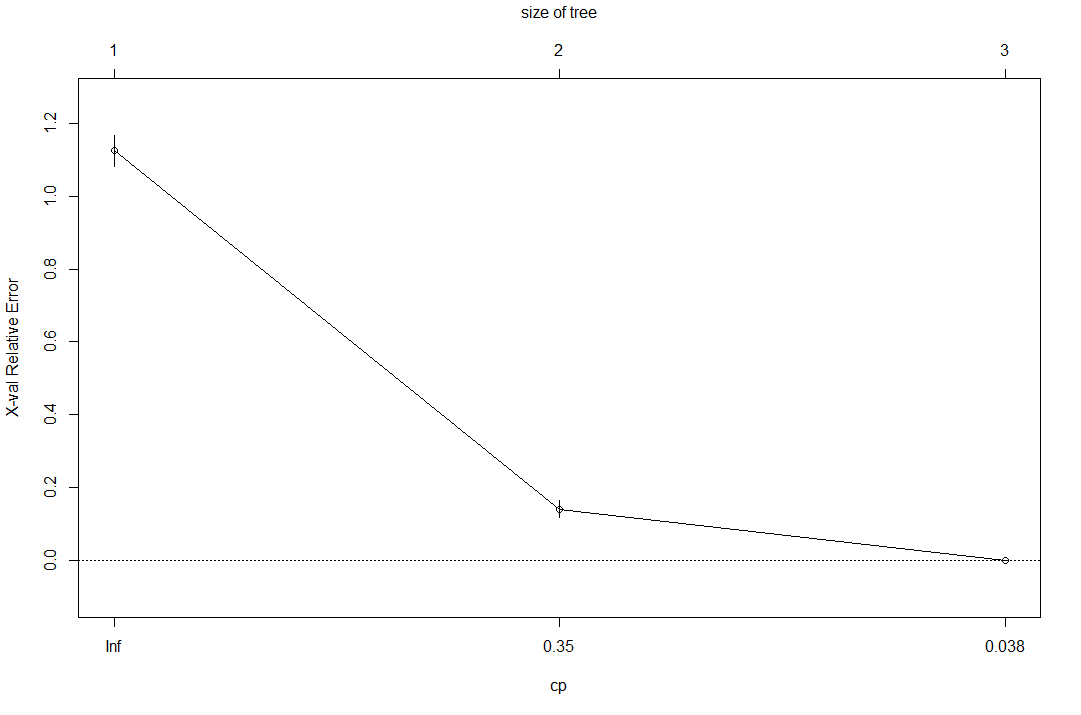


Figure . Error matrix

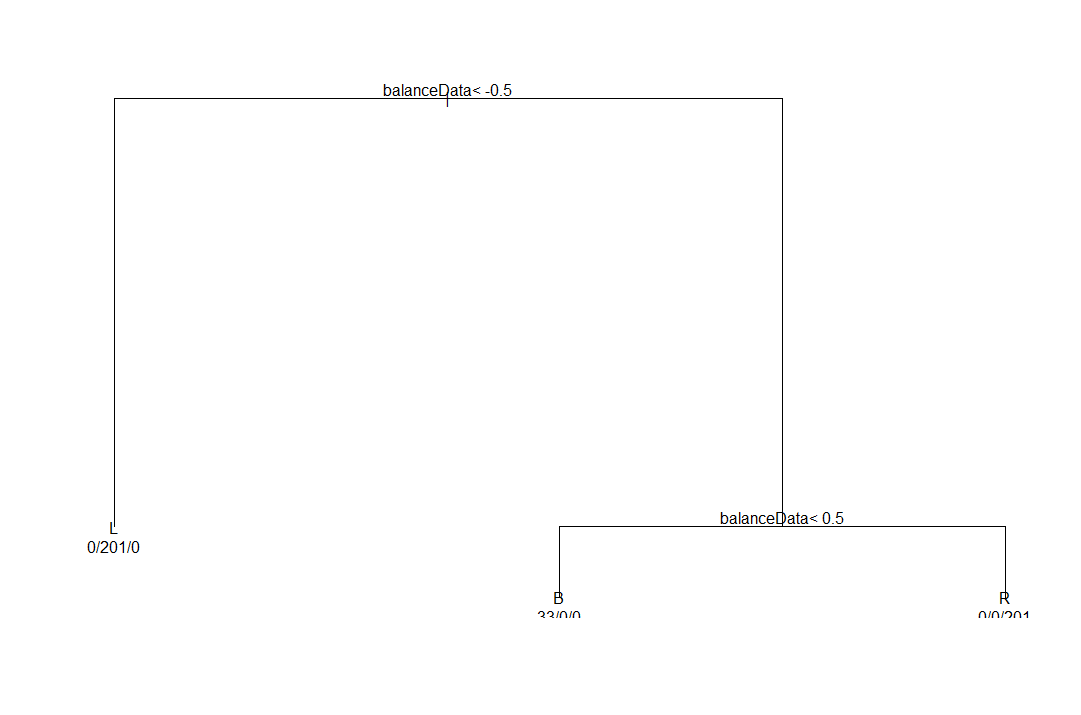
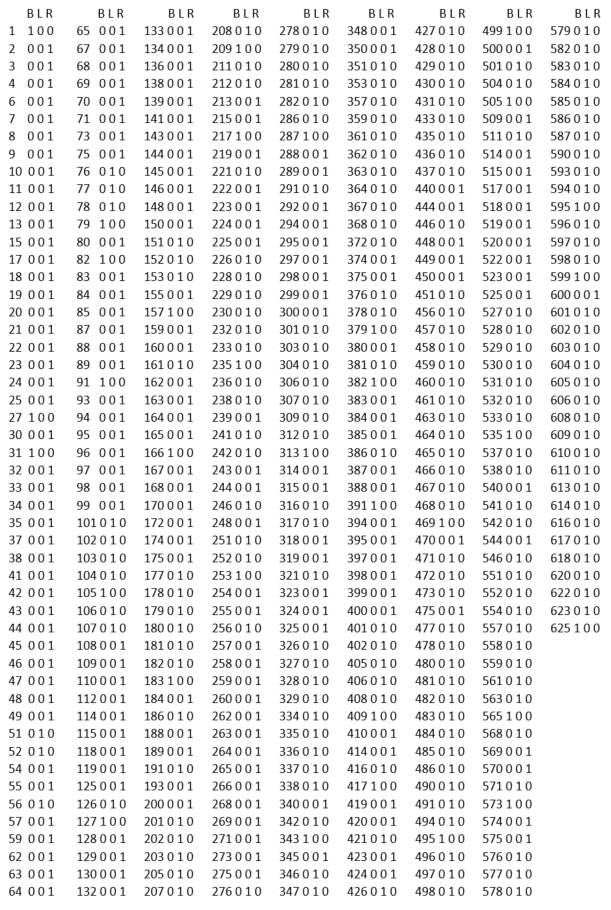


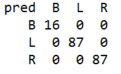
Figure . Decision Tree

Table . Decision Tree Classification Results



### Naïve Bayes Classification

Table . Naïve Bayes Classification Results



### Random Forest Classification

Table . Random Forest Classification Results

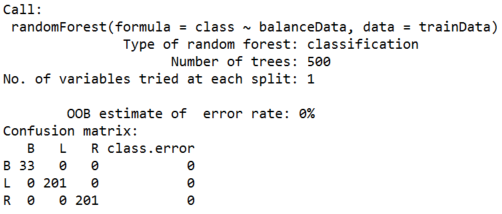


Table . Random Forest Classification Fit Importance



# CONCLUSIONS

From analyzing the data, there seemed to be an inverse relationship between horsepower and mpg as well as weight and mpg. In other words, as the horsepower increased the mpg decreased. Similarly, as weight increased the mpg seemed to decrease. For an even better understanding, regression analysis can be performed on the mpg, horsepower, and weight field values.

## Evaluation Metrics

### Accuracy and Error Rate

Accuracy is calculated as the percentage of test samples correctly calculated (TP is true positive, TN is true negative):

Error rate is calculated as the opposite, or 1- accuracy (FP is false positive, FN is false negative):

### Sensitivity and Specificity

Sensitivity is calculated as the true positive (TP) recognition rate:

Specificity is calculated as the true negative (TN) recognition rate:

Accuracy can be written as a function of both sensitivity and specificity:

### Precision and Recall

There is an inverse relationship between precision and recall.

Precision is measured as a percentage of the samples classified with a positive label that are actually positive, or exactness:

Recall is measured as a percentage of positive samples actually classified with a positive label, or completeness.

A score of 1.0 is a perfect score for either precision or recall.

### F-Measures

F-measure is a type of accuracy measurement which takes into account both precision and recall, with the resulting score assigned is between 0 and 1.

F-measure can also be a weighted measurement as follows:

## Decision Tree Classification Evaluation Metrics

First, I used the R regression fit for mpg and horsepower. The results, in Table 5, have an R-squared value of 0.6059 and adjusted R-squared value of 0.649. This is a higher R-squared value, which indicates the model fits the data fairly well.

## Naïve Bayes Classification Evaluation Metrics

Second, I used the R regression fit for mpg and weight. The results (Table 6) have an R-squared value of 0.6926 and adjusted R-squared value of 0.6918. This would be considered a higher R-squared value, which indicates a model that fits the data better.

## Random Forest Classification Evaluation Metrics

Second, I used the R regression fit for mpg and weight. The results (Table 6) have an R-squared value of 0.6926 and adjusted R-squared value of 0.6918. This would be considered a higher R-squared value, which indicates a model that fits the data better.

# REFERENCES

1. Larose, Daniel and Larose, Chantal D. 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. Wiley-Interscience.
2. Siegler, R.S. (1976). *Three Aspects of Cognitive Development.* Cognitive Psychology, 8, 481-520.
3. Klahr, D., & Sieger, R.S. (1978). *The Representation of Children’s Knowledge.* H.W. Reese & L. P. Lipsitt (Eds.), *Advances in Child Develpoment and Behavior*, pp. 61-116. New York: Academic Press
4. Langley, P. (1987). *A General Theory of Discrimination Learning.* D. Klahr, P. Langley, & R. Neches (Eds.), *Production System Models of Learning and Development*, pp. 99-161. Cambridge, MA: MIT Press
5. Newell, A. (1990). *Unified Theories of Cognition.* Cambridge, MA: Harvard Univeristy Press
6. McClelland, J.L. (1988). *Parallel Distributed Processing: Implication for Cognition and Development.* Technical Report AIP-47, Department of Psychology, Carnegie-Mellon University
7. Shultz, T., Mareschal, D., & Schmidt, W. (1994). *Modeling Cognitive Development on Balance Scale Phenomena.* Machine Learning, Vol. 16, pp. 59-88.