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

# 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

## Classification Methods

### Decision Tree Classification

Even if one was not aware beforehand, by looking at the histograms for cylinders (Figure 2), model year (Figure 7), and origin (Figure 8), one can tell that these variables are multi-valued, but discrete. There are only 5 different values for cylinders, the date range for model year is restricted to between 1970 (70) and 1982 (82), and the origin is one of 3 values. Another interesting distribution is the acceleration (Figure 6) values. These values seem to take on a nice bell curve without any data processing/transformations. The other remaining histograms for MPG (Figure 1), displacement (Figure 3), horsepower (Figure 4), and weight (Figure 5) all seem to have a right-skewed distribution. No histogram was completed for car name since it was specified as a unique identifier.

### Naïve Bayes Classification

From my knowledge of cars, I chose to look at the relationship between mpg and the three variables displacement, horsepower, and weight. The best way to visualize the relationships was to use scatterplots. None of the three scatterplots for MPG vs displacement (Figure 9), MPG versus horsepower (Figure 10), as well as MPG versus weight (Figure 11) showed a linear or an exponential relationship, but all three scatterplots displayed a relationship of quadratic decay or a decreasing rate relative to the field value.

### Random Forest Classification

Since the three variables I compared with MPG (displacement, horsepower, and weight) all seemed to have similar relationships to MPG I wondered what their relationships between each other were as well. I decided the best way to view any possible relationship was to use a scatterplot matrix (Figure 12). The scatterplot matrix shows a close to a linear relationship between horsepower and weight, weight and displacement, as well as horsepower and displacement. Considering the similarities between not only the histograms, but the scatterplots as well, this was not very surprising and confirmed what was seen in the other graphs.

# RESULTS

## Analysis

### Decision Tree Classification

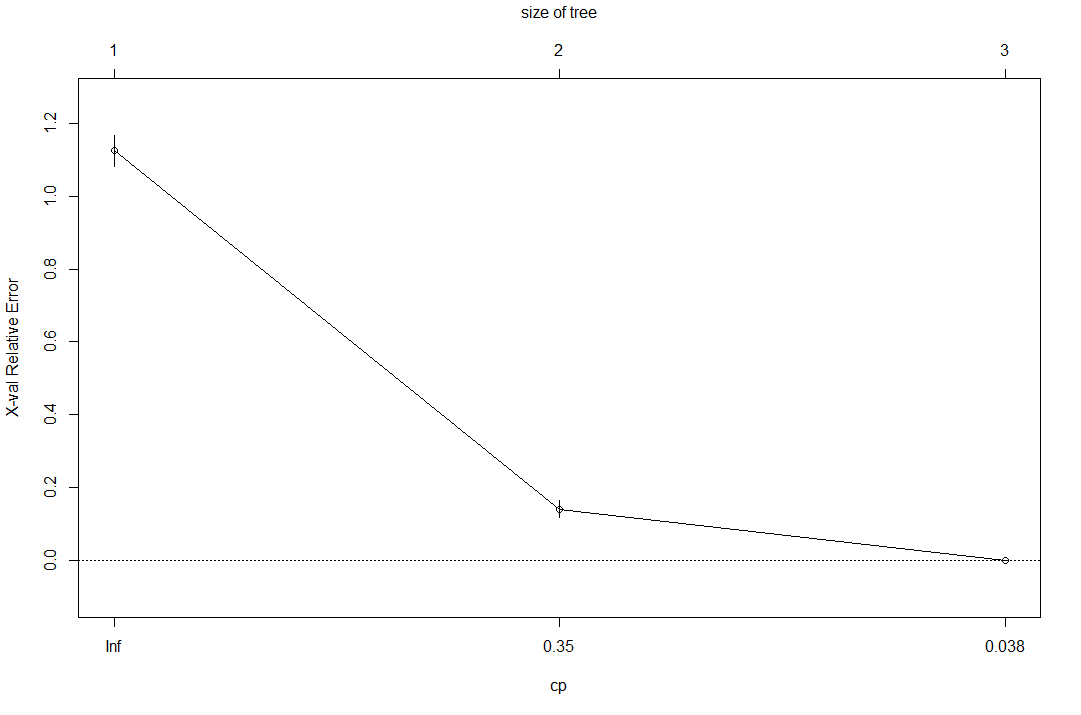


Figure 1. Error matrix

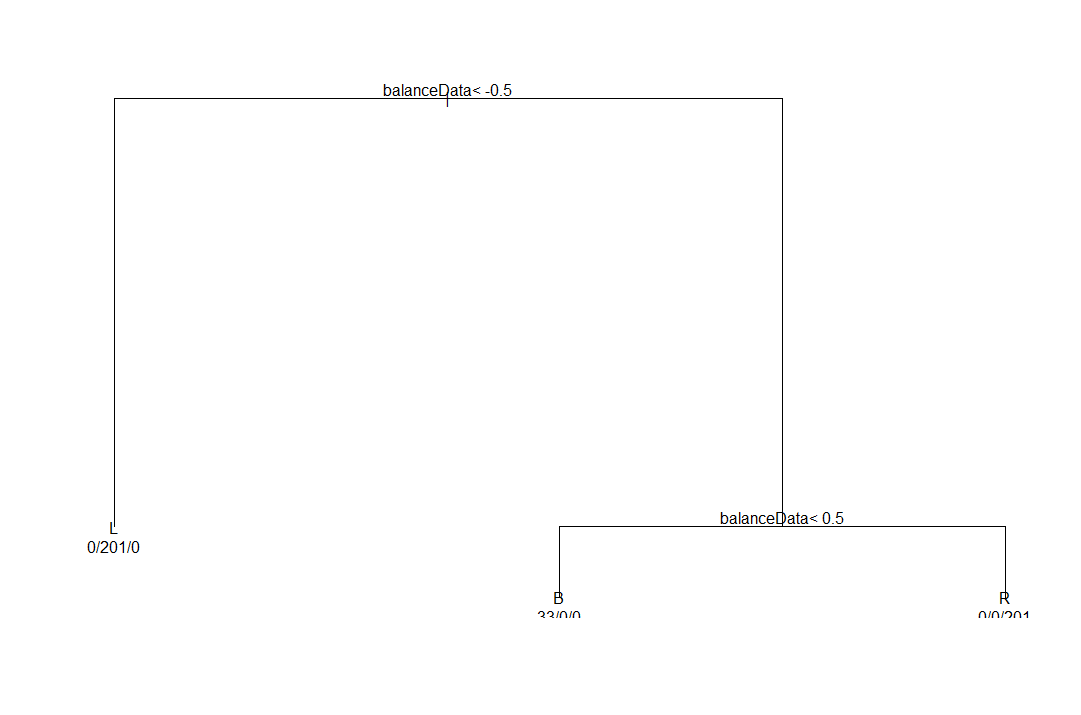
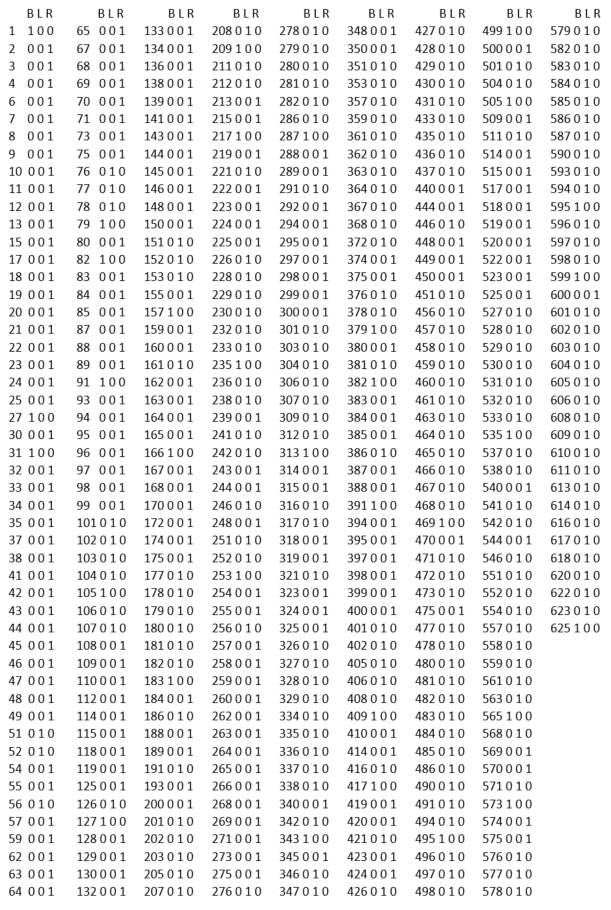


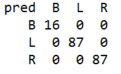
Figure 2. Decision Tree

Table . Decision Tree Classification Results



### Naïve Bayes Classification

Table . Naïve Bayes Classification Results



### Random Forest Classification

Table . Random Forest Classification Results

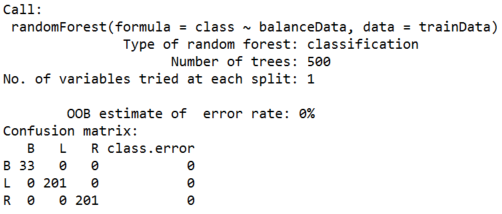


Table . Random Forest Classification Fit Importance



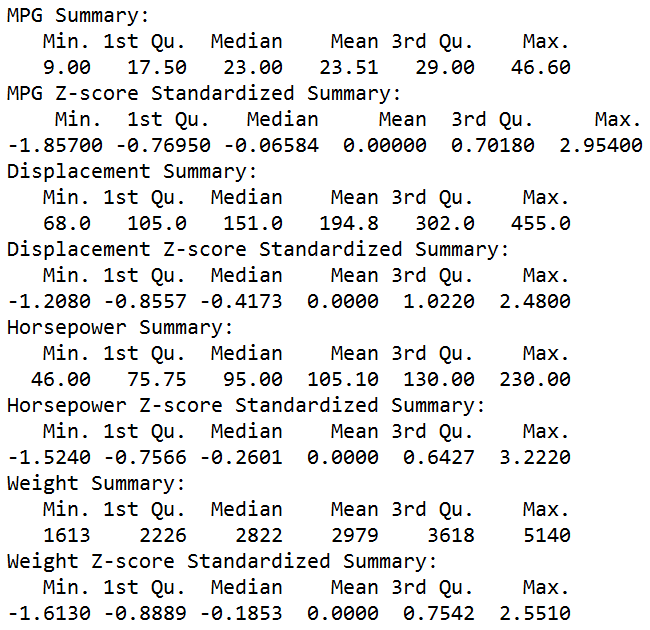
### Z-score Standardization

Z-score standardization takes the difference between the field value and the field value mean and scales the difference by the field’s standard deviation. The formula for Z-score standardization is as follows:

Field values below the mean will have a negative Z-score, field values above the mean will have a positive Z-score, and field values that fall on the mean will have a Z-score of 0 (zero).

For example, the horsepower has a mean of 105.10 and a standard deviation of 38.76868; Therefore, a value like 198, which is above the mean, would have a positive Z-score of 2.396709. Table 3 shows the summary of the original values as well as the Z-score standardized values for mpg, displacement, horsepower, and weight.

Table 5. Field Value and Z-score Standardized Summaries for MPG, Displacement, Horsepower, and Weight



### Decimal Scaling

Decimal Scaling ensures that normalized values lie between -1 and 1. The formula for decimal scaling is as follows:

where *d* is the number of digits in the data value with the largest absolute value.

For example, the weight data has largest absolute value of |5140|, which would make *d* the value 4; therefore, a value like 3090 would have a decimal scaled value of 0.309. Table 4 shows the summary of the original values as well as the decimal scaled values for mpg, displacement, horsepower, and weight.

# 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.

## Regression Fit

### MPG versus Horsepower

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.

### MPG versus Weight

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.