Guibo Huo

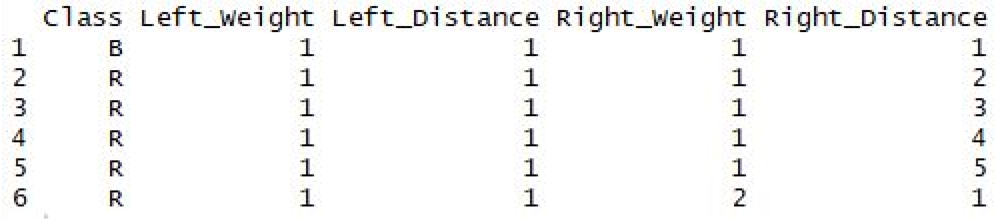
Machine Learning: Project 3

Summary:

Data on balance scale weight and distance were obtained from an online database[1]. We used two machine learning algorithms (K-nearest Neighbors and Classification and Regression Trees) to determine how accurately we can predict a balance scale tilt based on the scale's weight and distance. The factors weight of scale and distance of scale were used to determine the scale's tilted.

Dataset:

The data consists of 625 instances of balance scales tilt along with five attributes. There are four numeric attributes based on the left and right weight and the left and right distance. The last attribute is based on class name. Class name can be one of three possibilities: B, L and R (Balanced, Left tilted, Right Tilted respectively.) Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left-distance * left-weight) and (right-distance * right-weight). If they are equal, it is balanced.

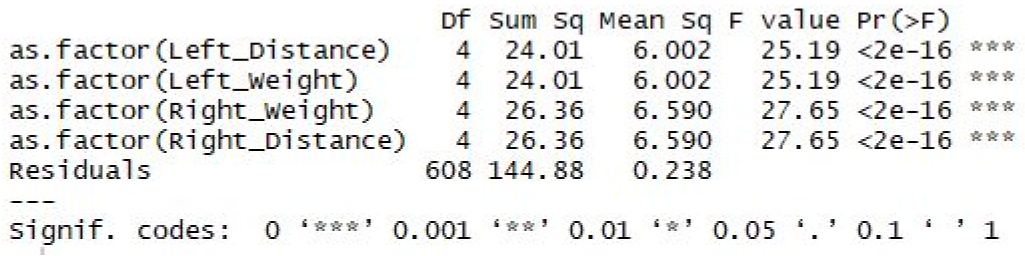


```
  Class Left_Weight Left_Distance Right_Weight Right_Distance
1     B           1             1            1              1
2     R           1             1            1              2
3     R           1             1            1              3
4     R           1             1            1              4
5     R           1             1            1              5
6     R           1             1            2              1
```

Figure.1: A brief example of the dataset. The first row was added later as the imported data did not have a header. The header containing the names of the attributes from the online database.

The dataset did not come with column names which made the data difficult to process. Therefore, column names were added based on the five attributes mentioned before. We used an Analysis of Variance Test (ANOVA) to test whether or not there is a significant difference between the three possible classes: Balanced, Left Tilt and Right Tilt. The result of the ANOVA can be seen below on page two, figure.2.

[1] Database: https://archive.ics.uci.edu/ml/datasets.html



```
                            Df Sum Sq Mean Sq F value Pr(>F)    
as.factor(Left_Distance)     4  24.01   6.002   25.19 <2e-16 ***
as.factor(Left_Weight)       4  24.01   6.002   25.19 <2e-16 ***
as.factor(Right_Weight)      4  26.36   6.590   27.65 <2e-16 ***
as.factor(Right_Distance)    4  26.36   6.590   27.65 <2e-16 ***
Residuals                  608 144.88   0.238                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure.2: The results of an anova test on the four numeric attributes of the dataset. We can see that every attribute has a significant p-value of less than two to the power of sixteen. Based off of this anova test we can assume that every one of the four numeric attributes are significant factors in predicting attribute Class.

Methodology:

We used the algorithms K-nearest Neighbors and Classification and Regression Trees (CART). Both are used as classification models in this experiment. In the K-nearest Neighbors method, we set the number of clusters to 3 (k = 3) since there were three possibilities for scale tilt. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In the CART method, we let the model decide how many branches would be the best predictor for Class. This is due to the model stops splitting when it detects no further gain in predictability can be made. Alternatively, the data are split as much as possible and then the tree is later pruned. The resulting CART tree can be seen below in figure 3.

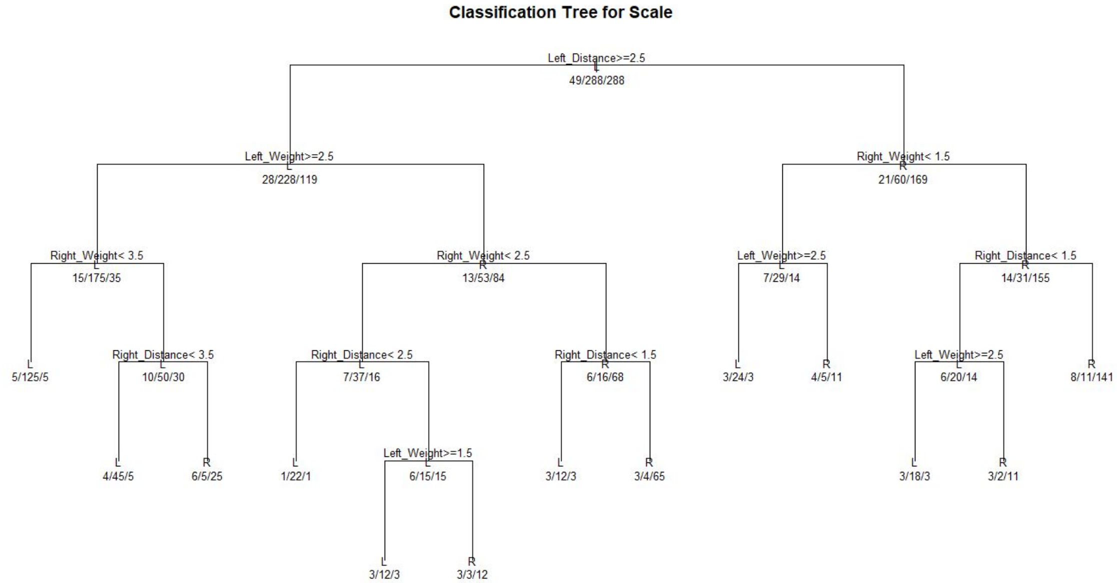


Figure 3. The resulting CART tree from the dataset. When a condition is true, the algorithm moves then test variable to the left of the branch. Based off of this tree, we see that the data has a skew to the left since there are more instances where the left weight is equal to or exceeds 2.5.

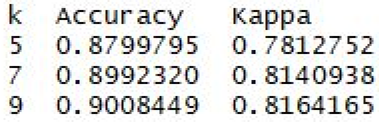
Results:

We tested these two algorithms using 10-folds cross-validation. The accuracy for the K-nearest Neighbors is 87.99 percent. The Kappa score is 78.12 percent. The accuracy means based on the given model we are able to correctly predict the tilt of the scale 87.99 percent of the time. The

Kappa score is high enough for us to accept this model since it has a strong level of agreement. For the Classification and Regression Trees (CART) model, using a complexity parameter of 0.062, we calculated an accuracy of 66.72 percent with a kappa score of 38.26 percent. For this model, both the accuracy score and the kappa score are too low to be accepted. The kappa score has a minimal level of agreement based on the kappa statistic standard for the level of agreement.

Discussion:

In the future, we would use more than two models to test determine how accurately we can predict a balance scale tilt. However, based on the current experiment we would use the K-nearest Neighbors algorithm since it provides us with a nearly 90 percent accuracy score and a close to 80 percent kappa score. After some more experimentation we saw both the kappa score and accuracy score rise when we increased the number of clusters.



```
k  Accuracy   Kappa
5  0.8799795  0.7812752
7  0.8992320  0.8140938
9  0.9008449  0.8164165
```

Figure 4: This table shows an increase in accuracy and kappa score as we increase the number of clusters(k). This is due to the algorithm being able to better determine the final outcome as we increase the number of clusters.