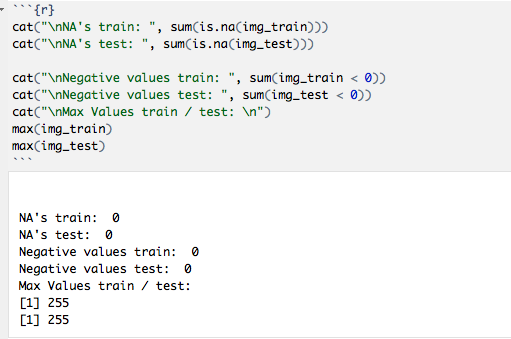
Homework 6 – Anthony Olivieri

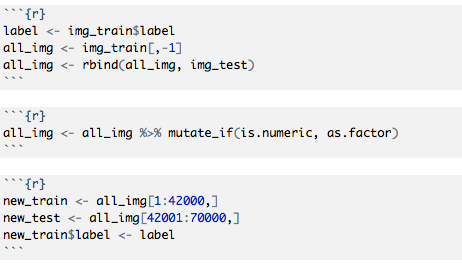
# Introduction | Data Preparation

Typical data preparation techniques were used, but unnecessary. For example, evaluating the dataset for missing or outlier values returned no results. The dataset is complete and the values were all within a reasonable range.



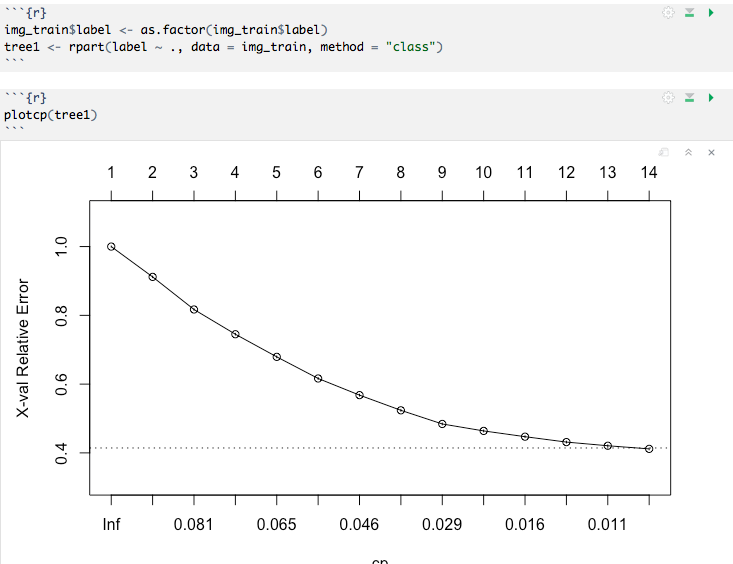
For the decision tree models, the label in the training dataset was converted into a factor and all other values were left as integers. This dataset (with non-discretized variables) was also run as a benchmark for the naïve Bayes model as a way to measure the performance gain by discretization.

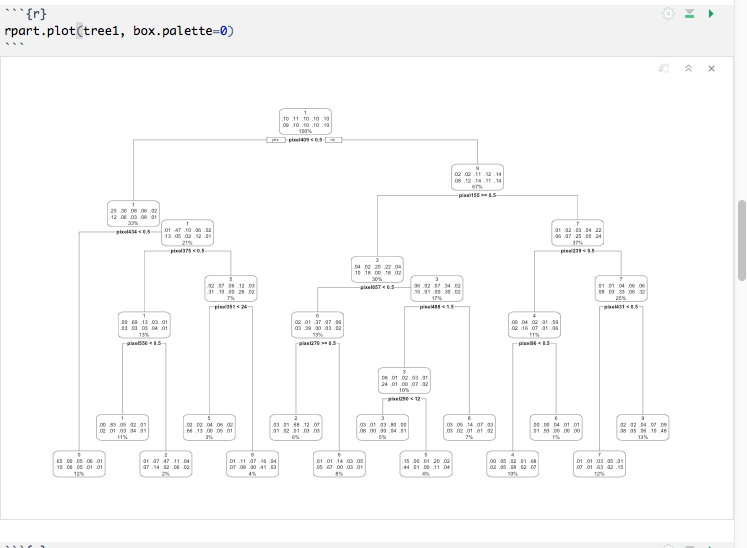
img\_train$label <- as.factor(img\_train$label)



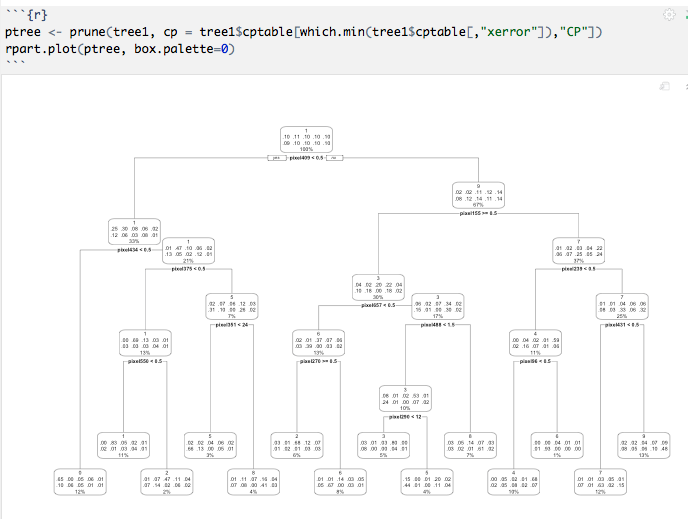
# Build and Tune Decision Tree Models

Several decision trees were created and evaluated for performance based on various complexity and bin parameters. No discernible elbow or considerable drop in errors was visible in the respective plots.





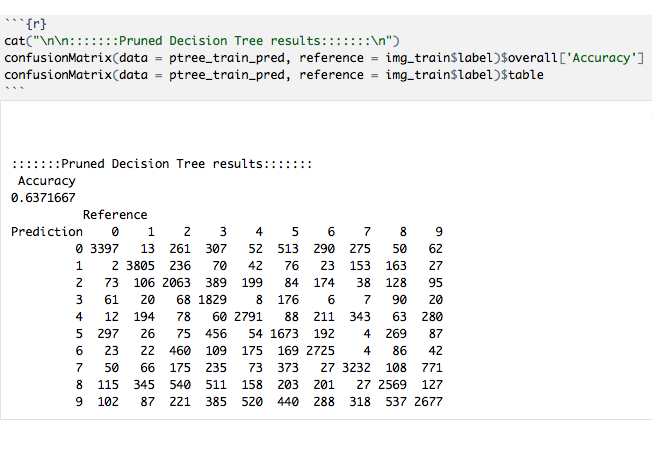
Since varying levels of cp and bin did not show marked improvement in errors, a pruned tree was created using the lowest possible error value from the original decision tree.



A prediction was created by running the model on the training data.

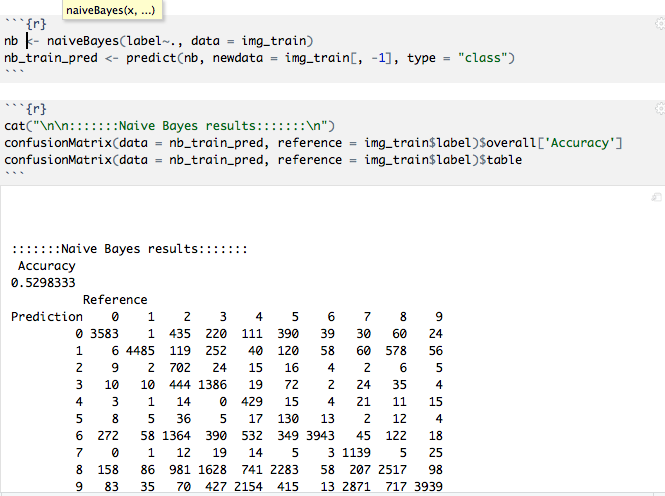


When a prediction was run against the training data the accuracy was ~64%.



# Naïve Bayes

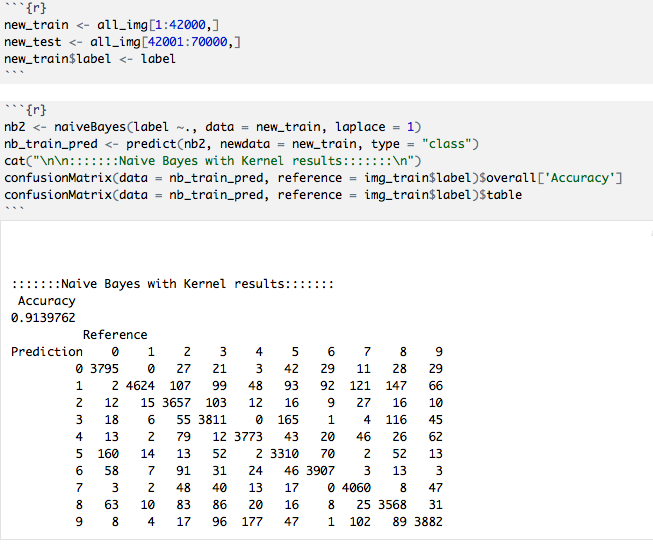
A benchmark naïve Bayes was run without discretizing the variables. As expected, this resulted in a very low prediction accuracy of around 53% when run against training data.



This low accuracy is because the naïve Bayes algorithm assumes a normal distribution of numeric values in the independent variables. To boost accuracy the data was discretized, all independent variables were re-typed as factors with levels ranging between the variable minimum and variable maximum: 0 to 255.

Discretization of the independent variables was a challenge. Discretizing training data and test data separately resulted in errors because a value (factor level) in the test data was not present in the training data; the resultant error was an out-of-bounds exception. To remove this error, all data was collected into one dataframe, then discretization performed, and the data split into train and test dataframes again.

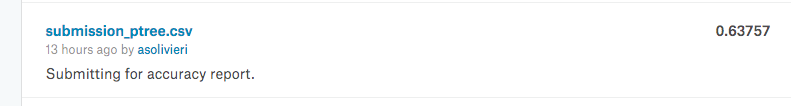
A new naïve Bayes model was constructed against the discretized data and resulted in an accuracy of ~91%.

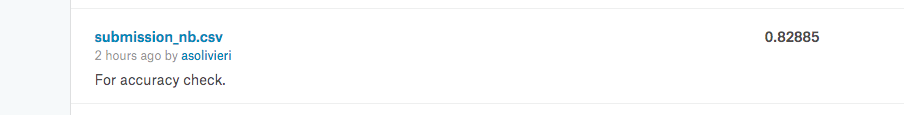


# Algorithm Performance

After discretization of the independent variables the naïve Bayes outperformed the pruned decision tree model. The naïve Bayes was computationally heavier, requiring around 11 minutes to perform on a Macbook pro laptop. The decision tree model was much faster, requiring only around 74 seconds on the same computer. Discretizing the independent variables on 70000 values also took a lot of human time (debugging errors as well as combining the training and test data then re-splitting the data) and processor time, ~ 30 seconds.

When checked against Kaggle for accuracy the pruned decision tree achieved ~64% accuracy. And the discretized naïve Bayes achieved ~83% accuracy.





The higher accuracy on training data for the naïve Bayes model might indicate overfitting: a drop of ~8%. The pruned decision tree did not show signs of overfitting, the training accuracy was almost exactly the same as the test accuracy.