Project 3 – Classification trees or rules

*Joseph Park - x172917 Date: 4/10/20*

Objective

The assignment is an illustration of classification based on the classification trees or rules model using particle collision data from a particle accelerator to identify particles.

Classification trees or rules model

Step 1: Download the data set from Kaggle, look for missing data and look at summary, size and shape.

Step 2: Since the data set is too large to process, randomly sample.

Step 3: Partition data into training vs testing and target vs classified data.

Step 4: Train data

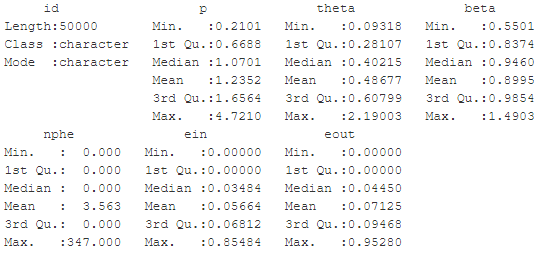
Step 5: Classify data.

Step 6: Create a loop to optimize classification metric of choice, which is accuracy (for particles classification), with respect to number of trials (for performance of course as opposed to just using the default or some specific number of trials).

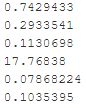
Step 7: Plot accuracy of the index (of the loop iteration which is the number of trials) and write confusion matrix for the index value with max accuracy.

Step 8: Repeat steps 4-7 but with rules this time and compare results.

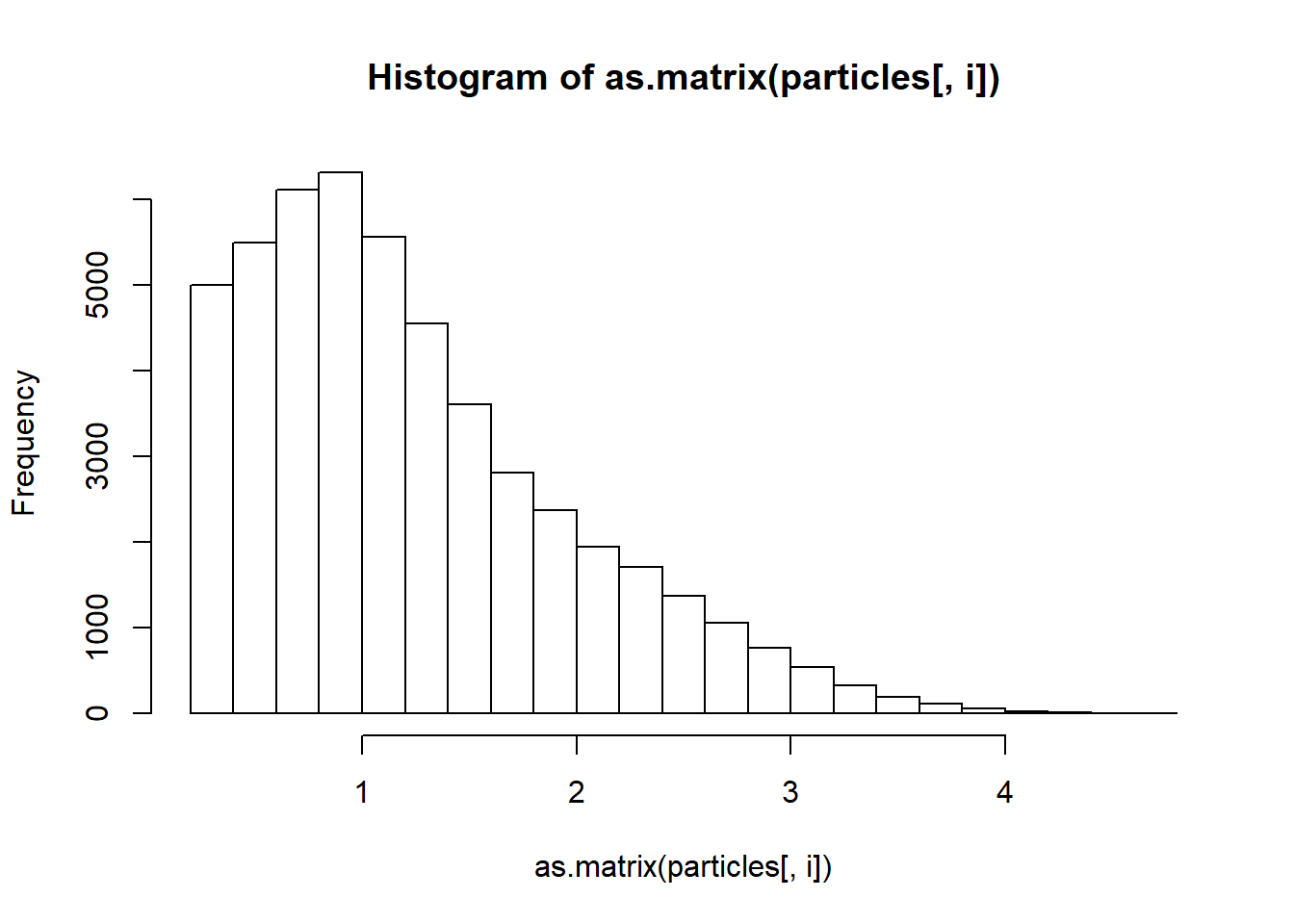
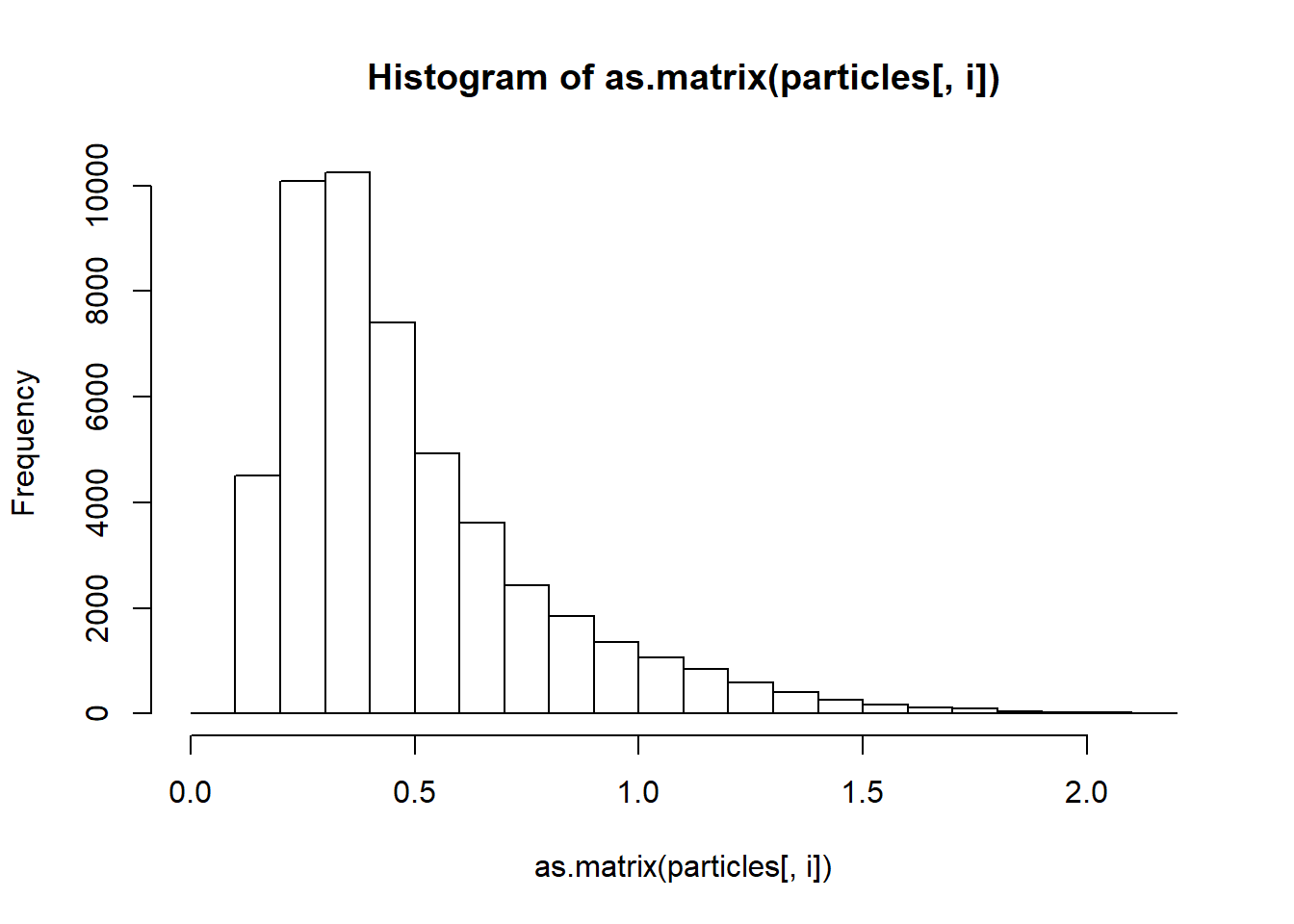
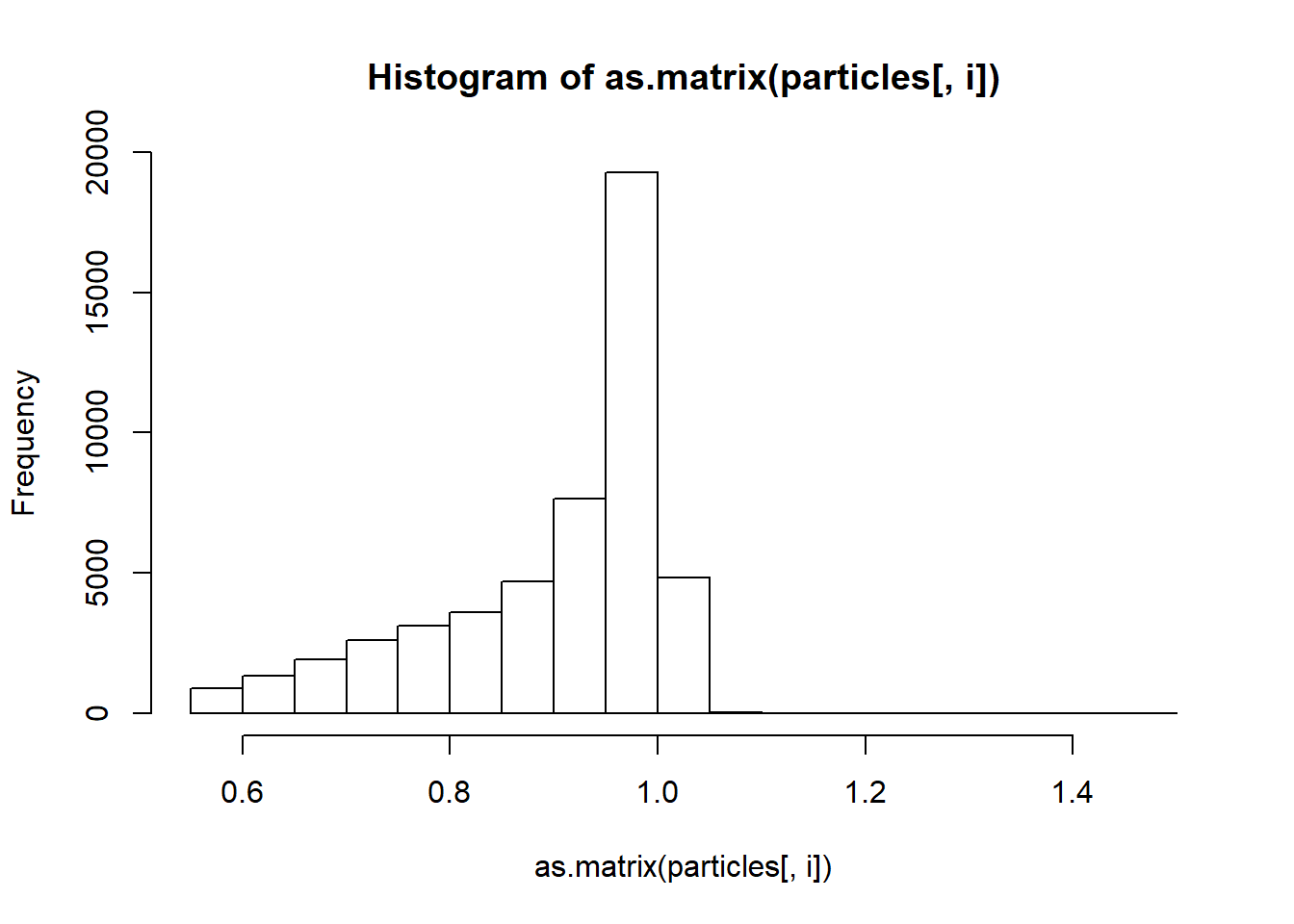
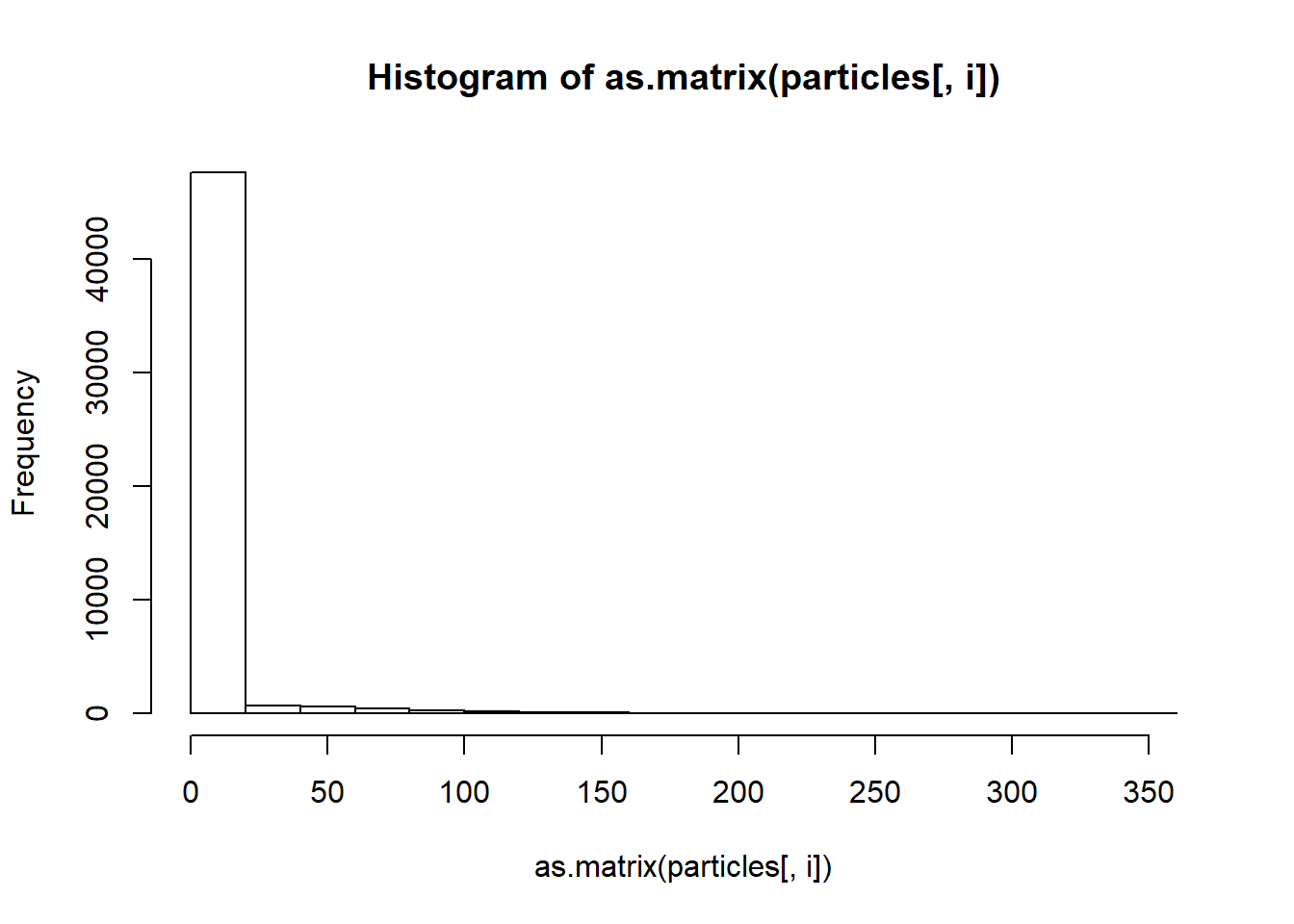
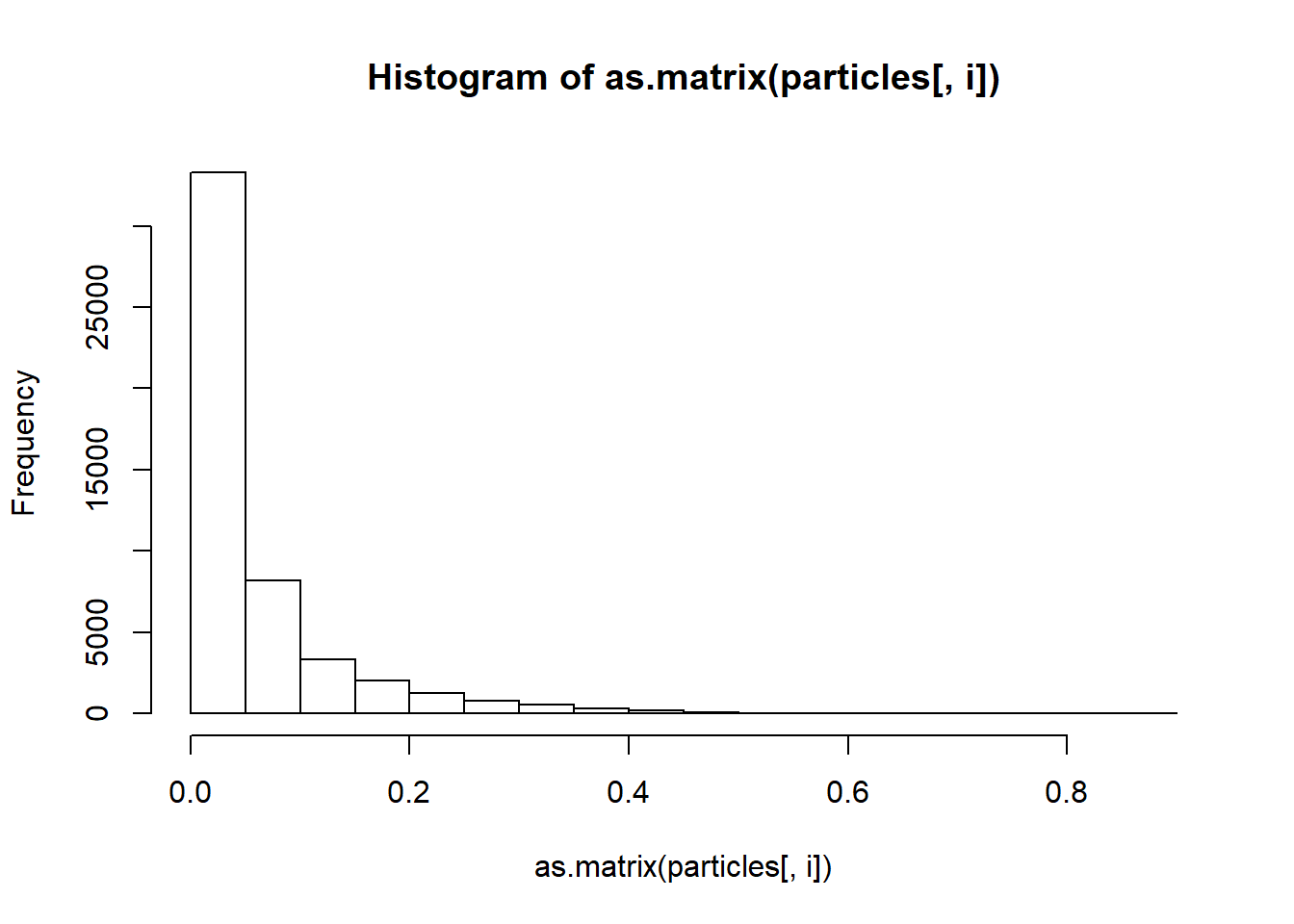
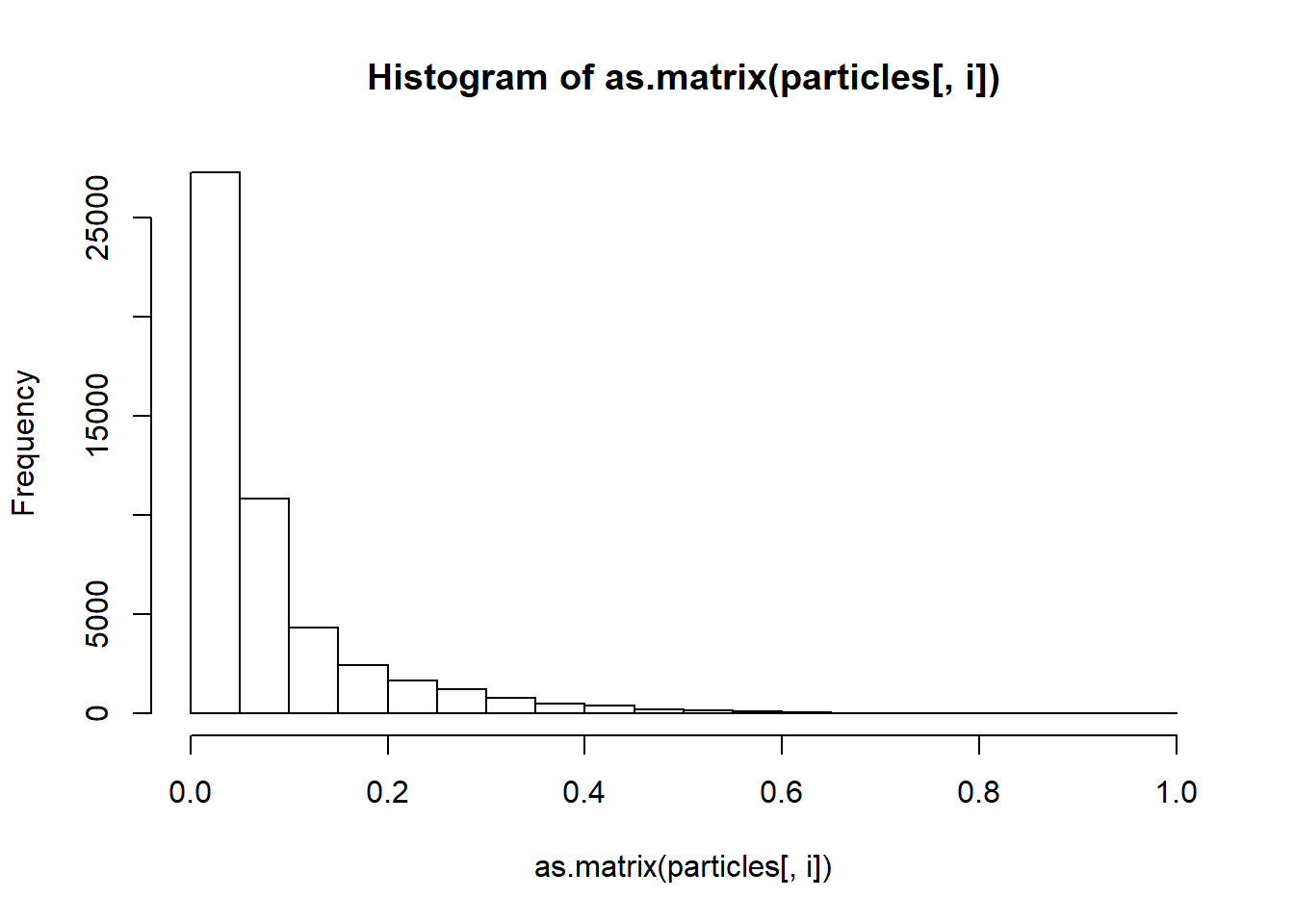
Step 9: Repeat steps 4-7 but with cost particular to particle -11



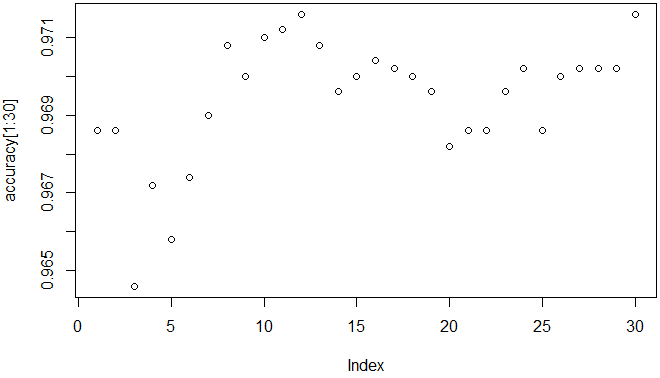
Here are the standard deviations of the variables in order of above left to right then top down.



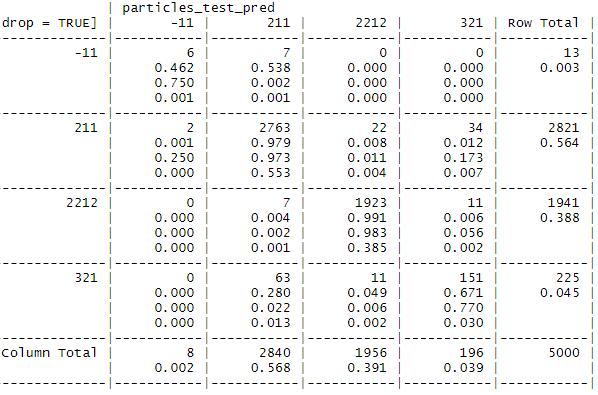
Here the corresponding histograms and the aforementioned order.

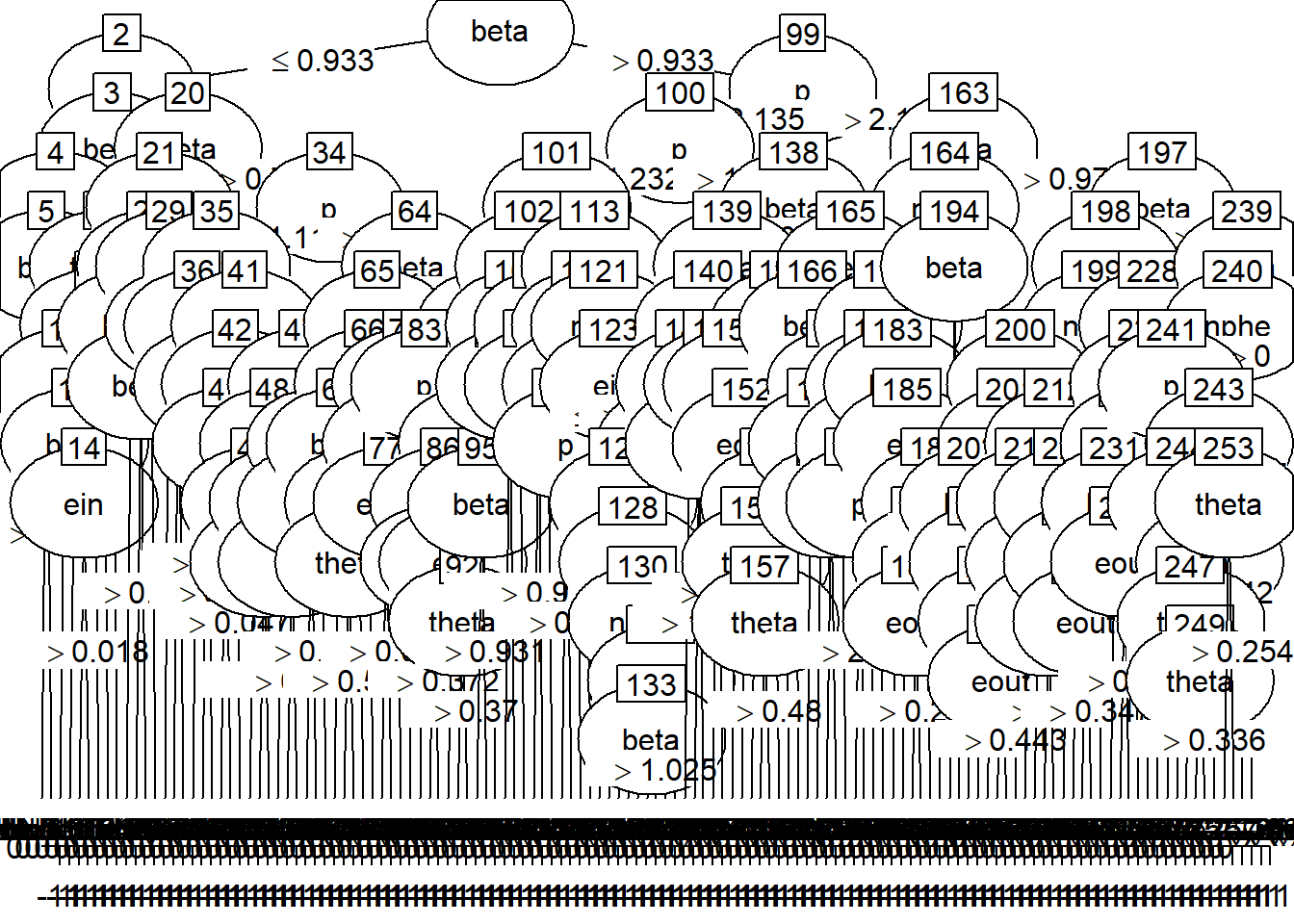
Results

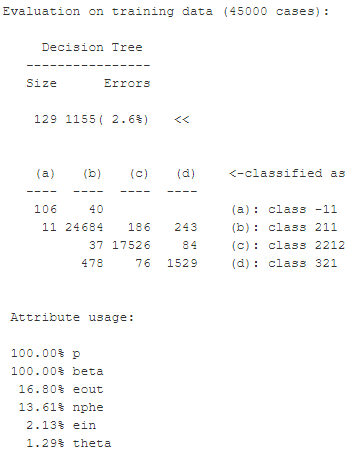


"The best accuracy is 0.9716"

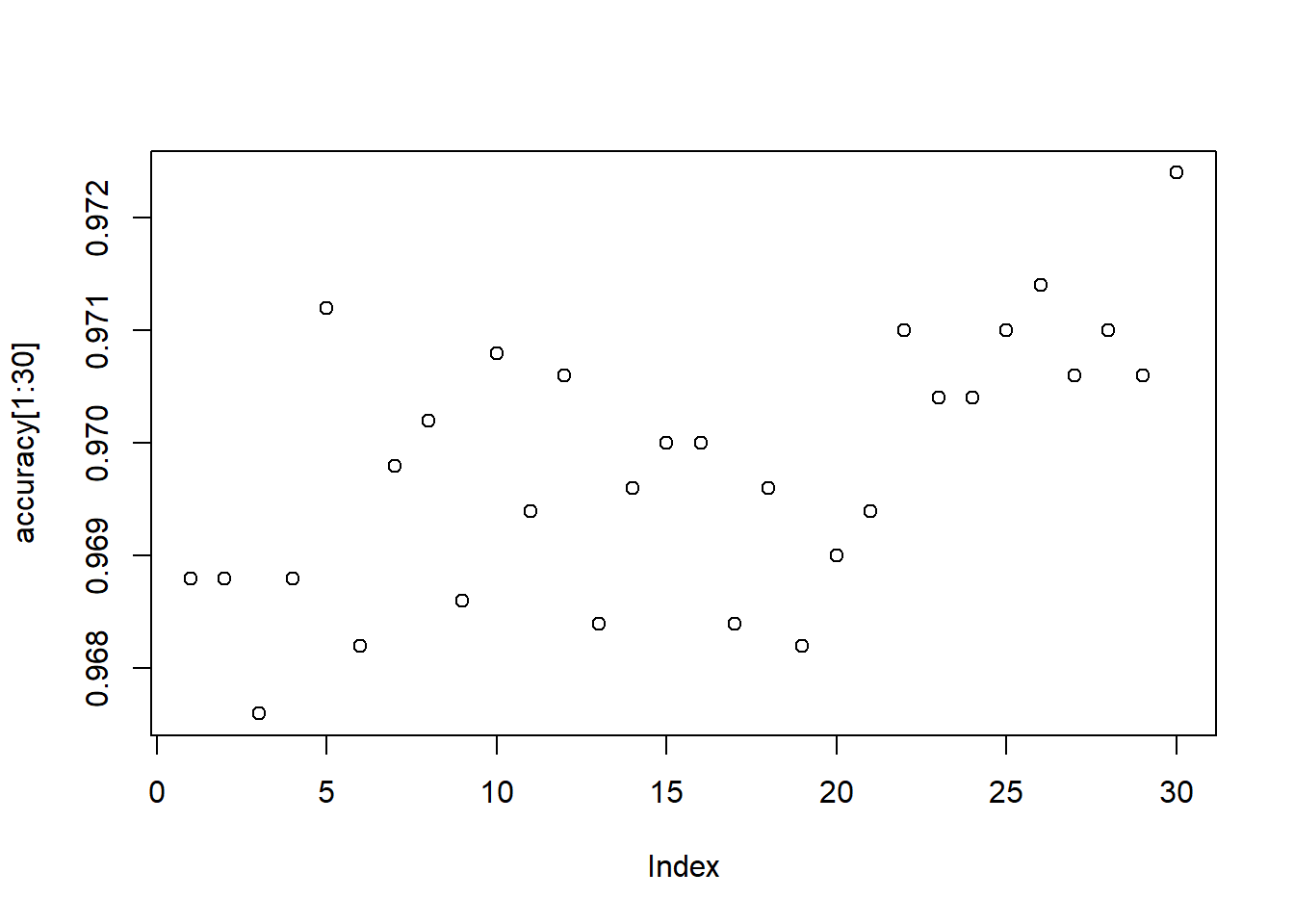


^Columns are predicted values and rows are actual values

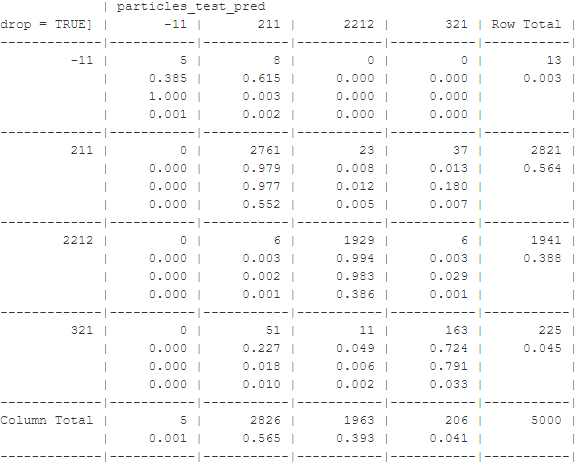




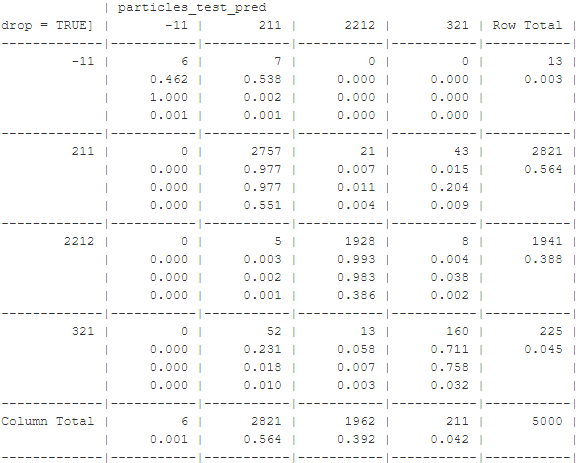
Now with rules:



"The best accuracy is 0.9724"



With high cost on particle -11:



Interpretation of the Results

Note 1: If particle -11 was of particular interest, a supercomputer (and thus ability to process a sample size significantly more 5000) would be of use, as 13 data points is not reliable.

Note 2: The accuracy is very good for classification trees and rules and all attempts to improve performance.

Note 3: As with kNN and naïve Bayes, classification of particle -11 and particle 321 is significantly worse than the other two particles, but the difference between the particles much than kNN and naïve Bayes for classification trees and rules and all attempts to improve performance.

Note 4: My interpretation from the association rules project about the variable p or beta being a good place to start for classification trees or rules is reflected in the plot of the tree.

Note 5: But, the trees are large, though that is not unexpected.

Note 6: The two trial numbers (12 and 30) with the clearly best accuracy are approximately equal.

Note 7: Using trial number 12 (with C5.0 without rules or cost), we see from the confusion matrix that particle -11 is much better classified (than with kNN or Naïve Bayes) this time but still not great.

Note 8: Using rules with C5.0 gave a slightly better accuracy, but this time trial number 30 was alone had clearly the best accuracy and particle -11 did not do as well as without the rules.

Note 9: Using a heavy cost for all misclassifications of and as particle -11 did not make much difference.