Project 7 – Random Forest

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

Objective

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

Random forest 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.

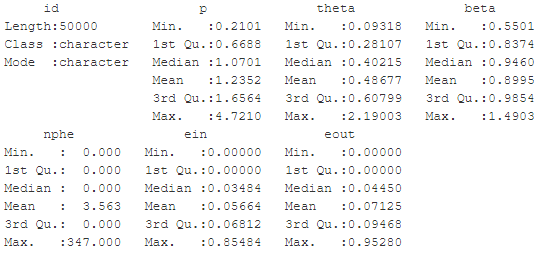
Step 4: Train data

Step 5: Classify data on defaults.

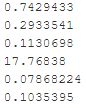
Step 6: Use the loop (train function) from the caret package to optimize kappa with respect to mtry.

Step 7: Create loop to optimize accuracy with respect to mtry (same as index).

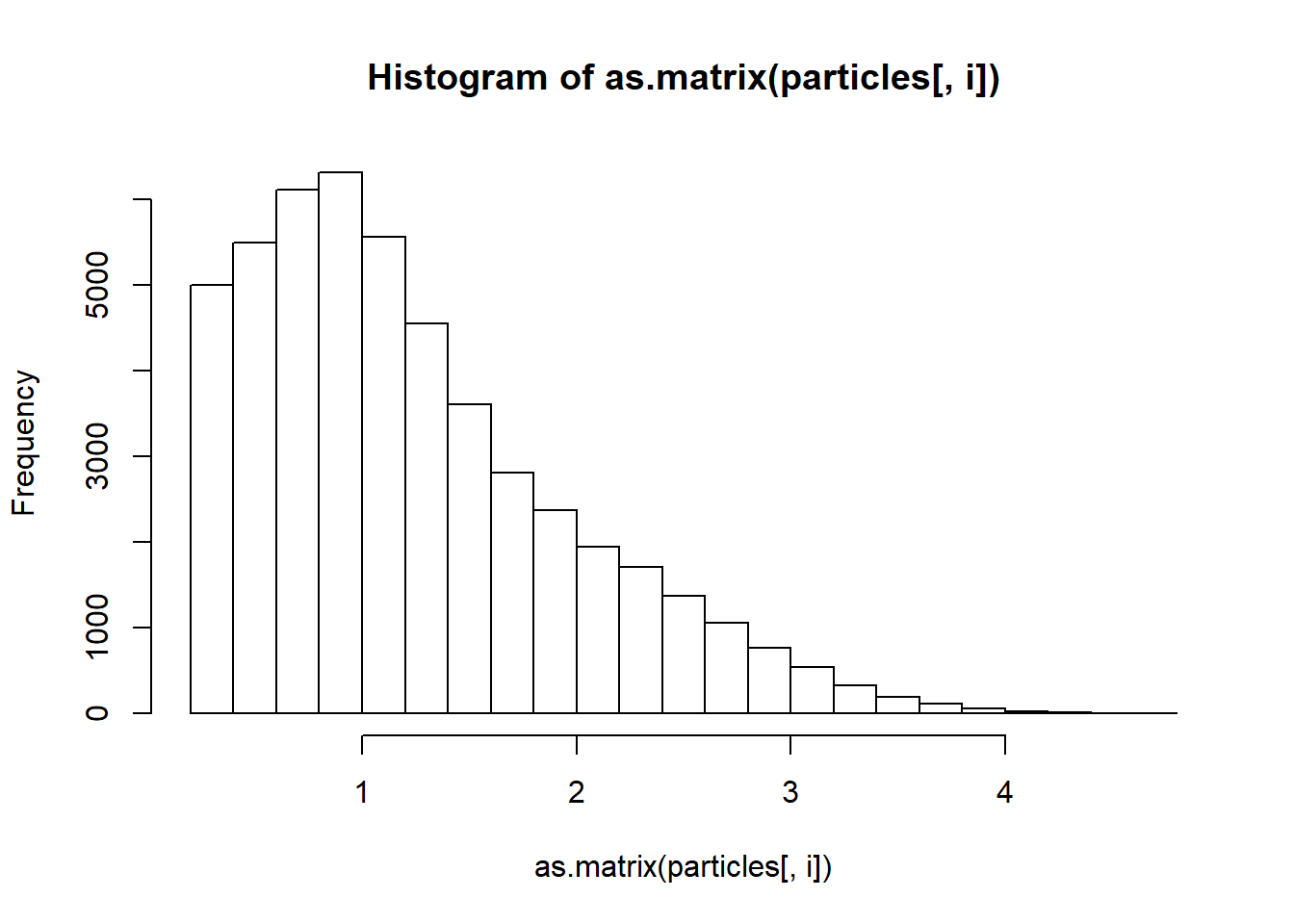
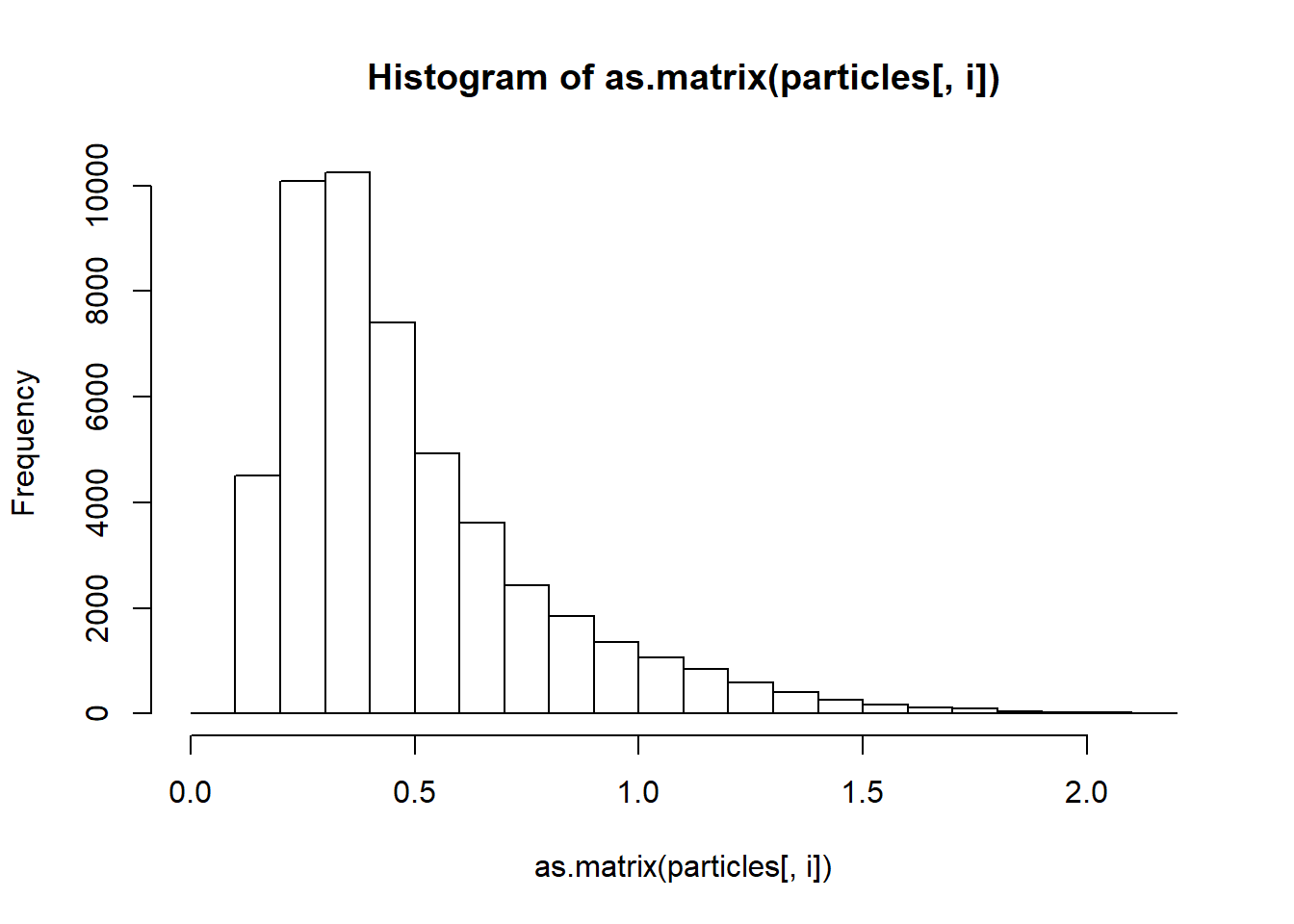
Step 8: Plot accuracy of the index and write confusion matrix for the index value with max accuracy.

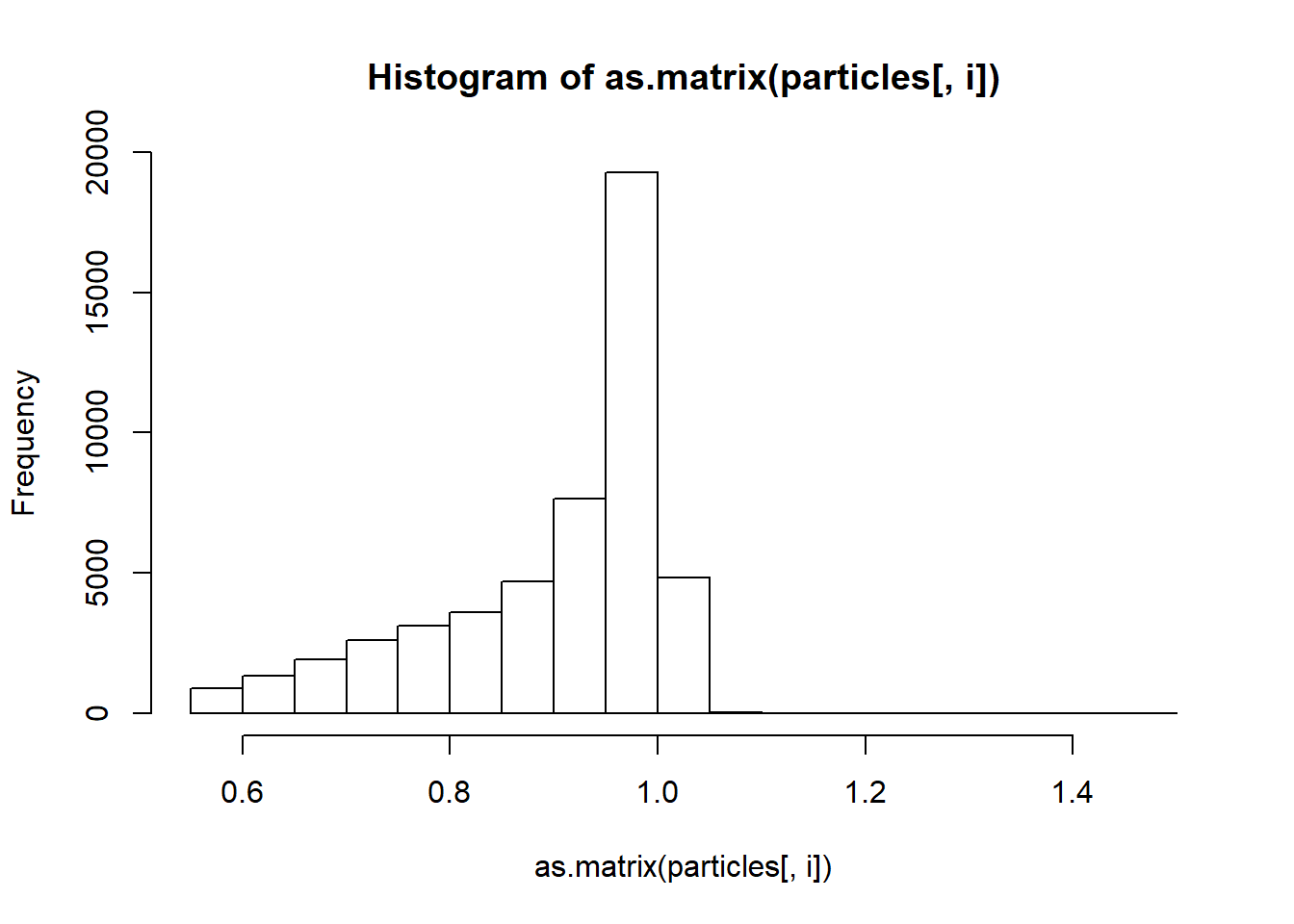
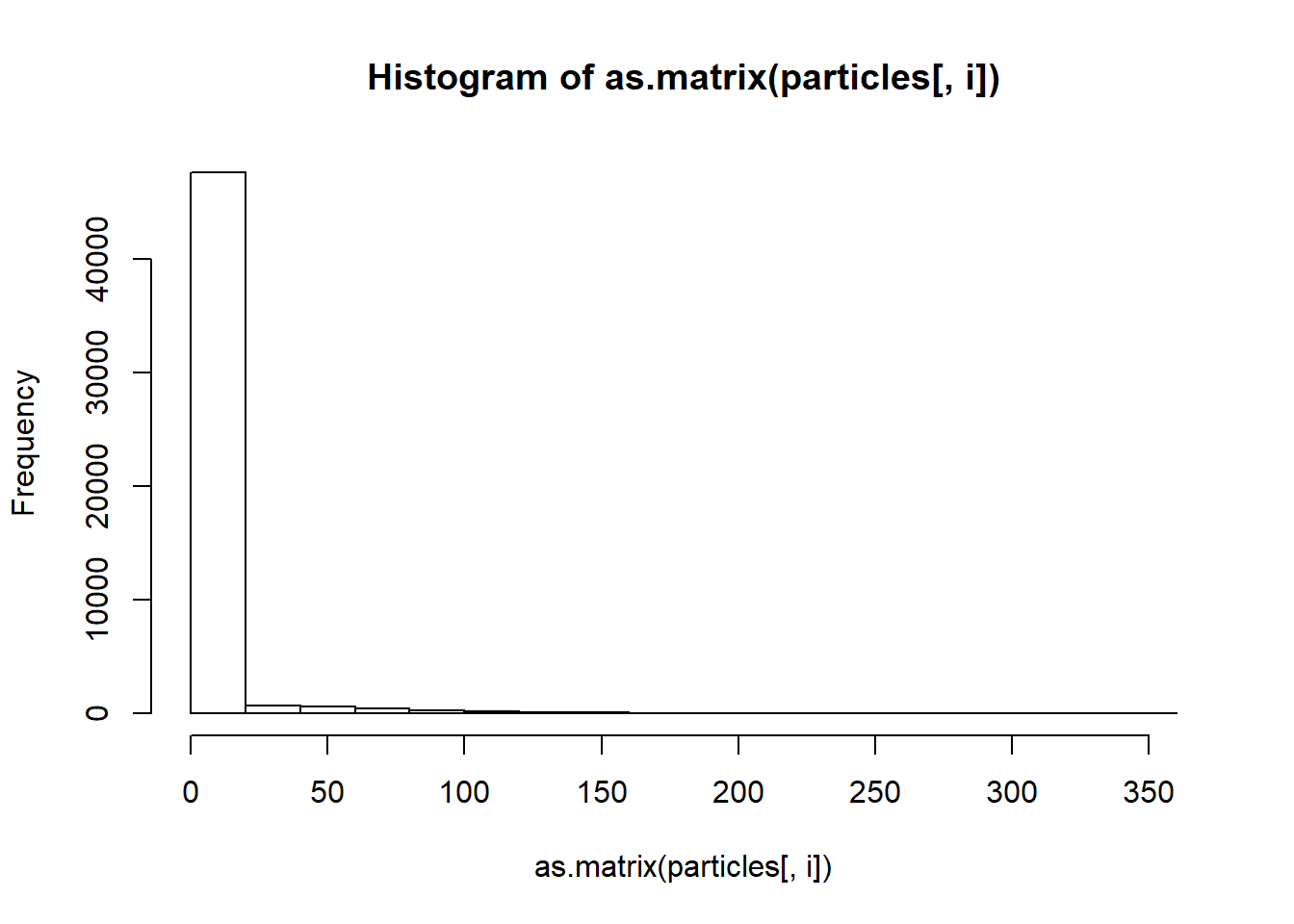


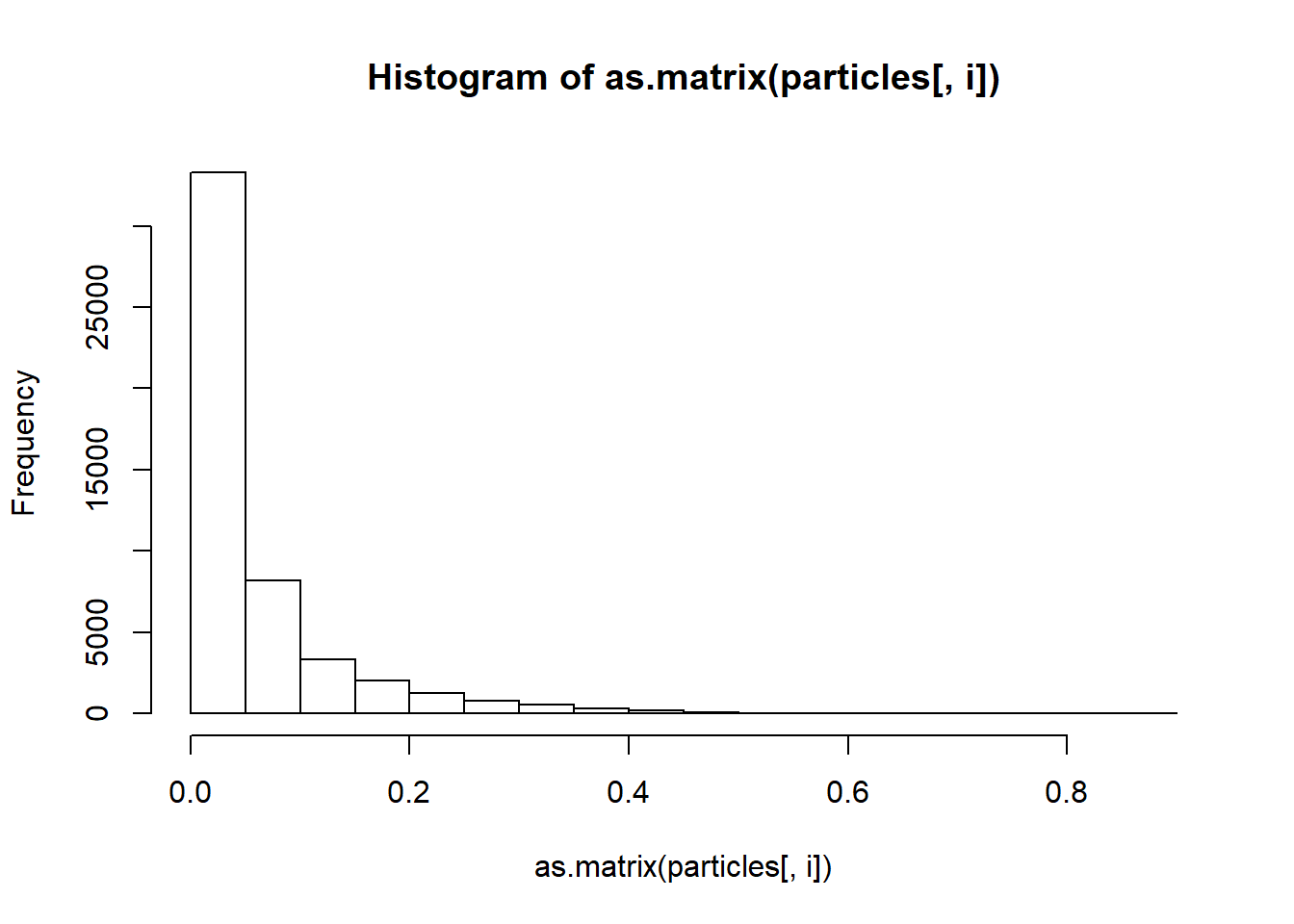
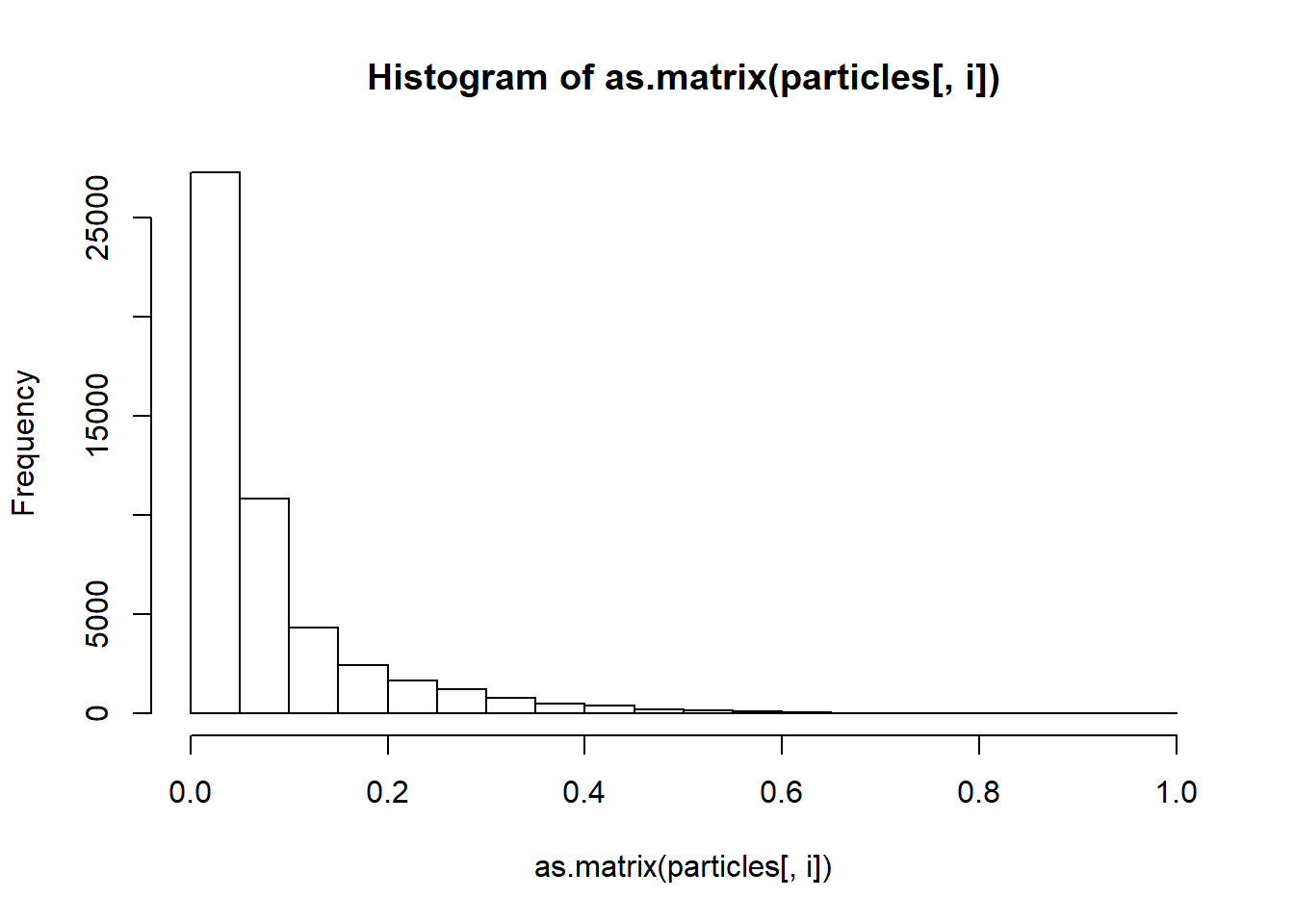
Here are the standard deviations of the variables in order of above left to right then top down.



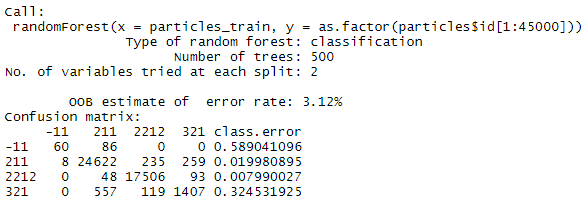
Here are the corresponding histograms in the aforementioned order.

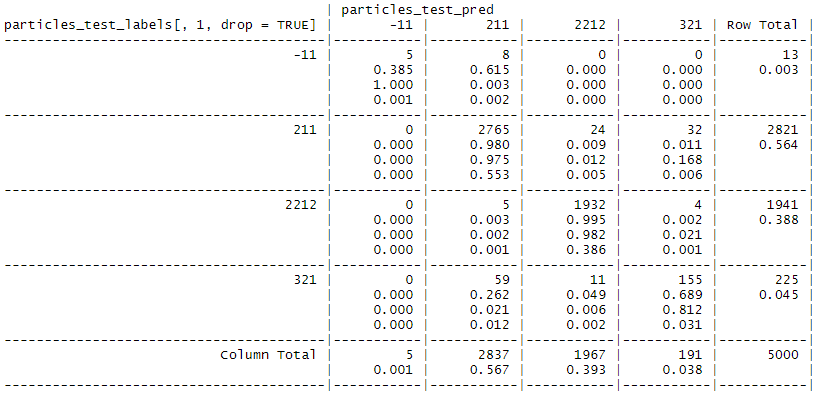
 

Results

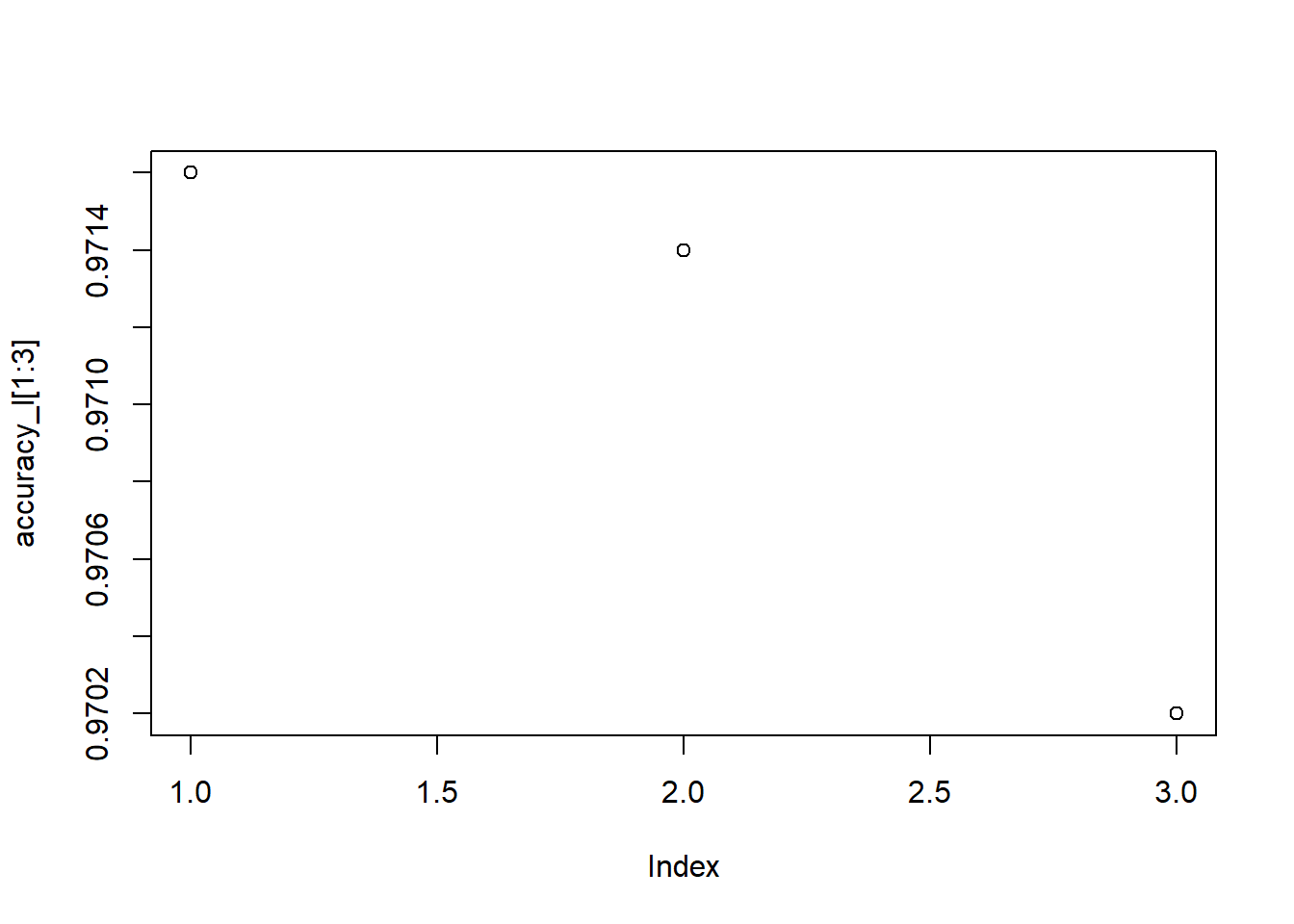


The accuracy is 0.9714



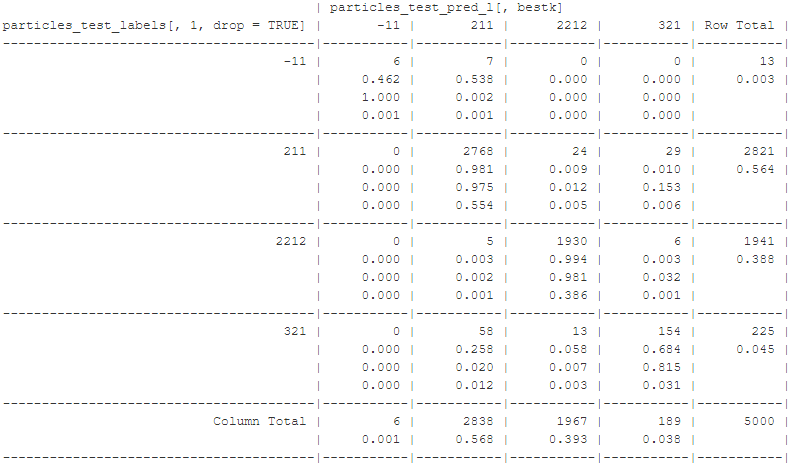
^Columns are predicted values and rows are actual values

After optimization:



^ the mtry corresponds to the index plus one, so the max accuracy occurs at mtry = 2

"The best accuracy is 0.9716"



^Columns are predicted values and rows are actual values

Interpretation of the Results

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

Note 2: With parameters at default (mtry = square root of 6), the accuracy is already very good.

Note 3: The accuracy for default mtry and optimized mtry are the roughly same, though optimized is slight better.

Note 4: Although accuracy is only slightly better, the results for the difficult to classify particle -11 were much more significantly better (than that of the accuracy).