**MIS 790- Directed Readings**

Analysis of Pilot Data sets

# TEAM INFORMATION

Team Member Names: Apoorva Reddy Panyala

Gaurav Gade

# About the data

**Independent variables** (features) for our analysis are: ***APL*** – Average payload packet length (Byte-based group); ***IOPR*** – Ratio between the number of incoming packets over the number of outgoing packets (Packet-based group); ***BS*** – Average bits per second (Time-based group); ***Duration*** (Behavior-based group).

**Dependent variable:** Bot packet or not (determined based on malicious IPs)

The raw data provided was divided into Training and Validation data set with only few variables like Source and Destination IP addresses etc. After performing the Extract, Transform and Load (ETL) process to convert/transform training data to derive values of the four variables (APL, IOPR, BS, Duration) from the captured packets (using apps: MS Message Analyzer, programming), 10,000 packets were extracted and derived about 600+ workflows for each of the training and testing datasets. However, all of these flows were botnet flows which makes it impossible to develop prediction models.

After extracting data flows from the entire datasets using Azure cloud system, from each of the training and testing datasets, 1000 data flows - 500 botnet and 500 non-botnet flows were extracted for a pilot test which included a total of 2000 flows.

# anlaytical models

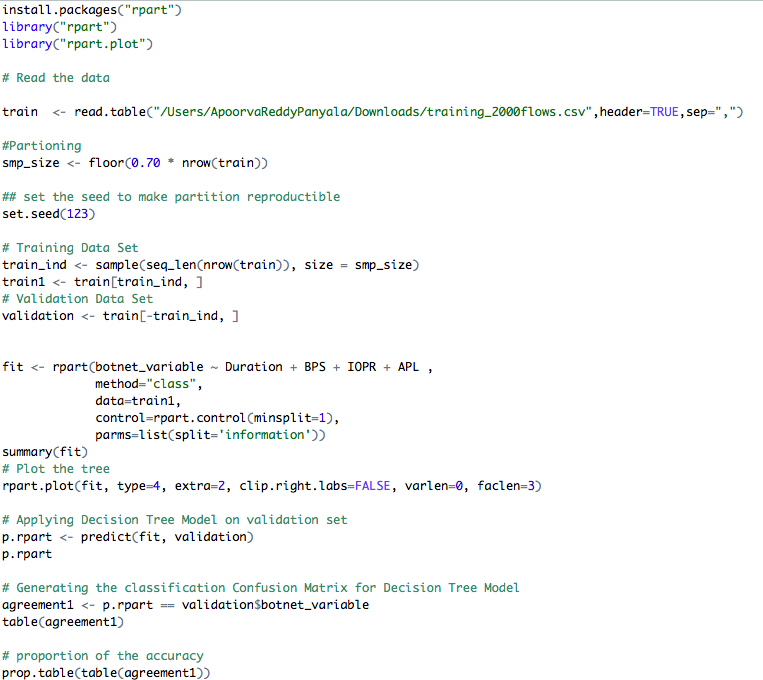
Our goal is to build a classification model on network packet traffic data, and classify incoming network packets into botnet and non-botnet data. Our aim is to build a model that has a higher accuracy than the model built by the team of security and data researchers from University of New Brunswick, by using better optimized models.

We started by building a Decision Tree Model and then used other models (KVSM, Naïve Bayes, Svm, Random Forest) to compare accuracy of results.

# R Coding

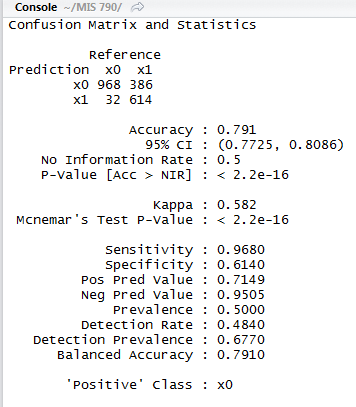


**Decision Tree Model:**



# results - summary of accuracy

**Confusion Matrix and Statistics for Decision Tree:**

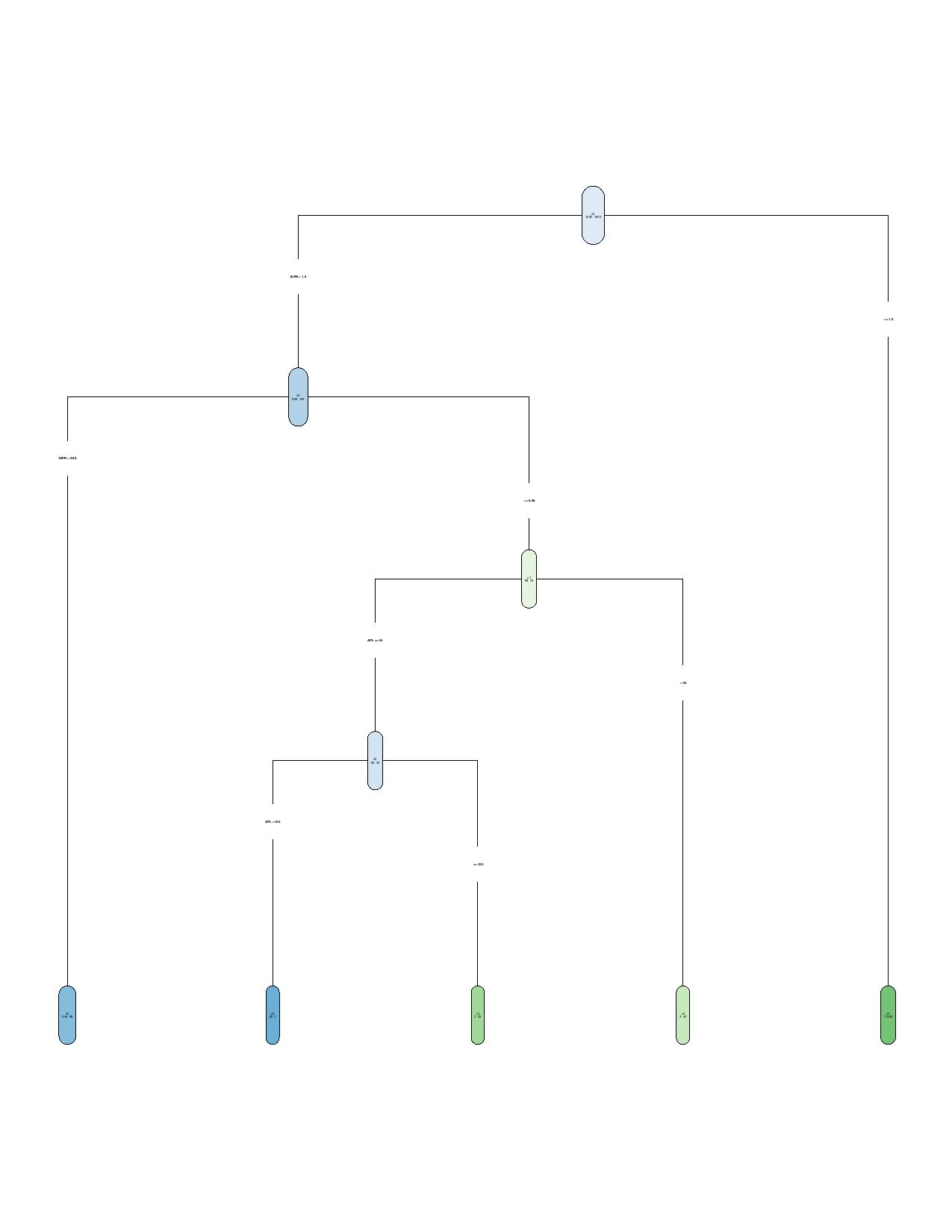


|  |  |
| --- | --- |
| FALSE | TRUE |
| 1192 | 8 |

|  |  |
| --- | --- |
| FALSE | TRUE |
| 0.993333333 | 0. 006666667 |

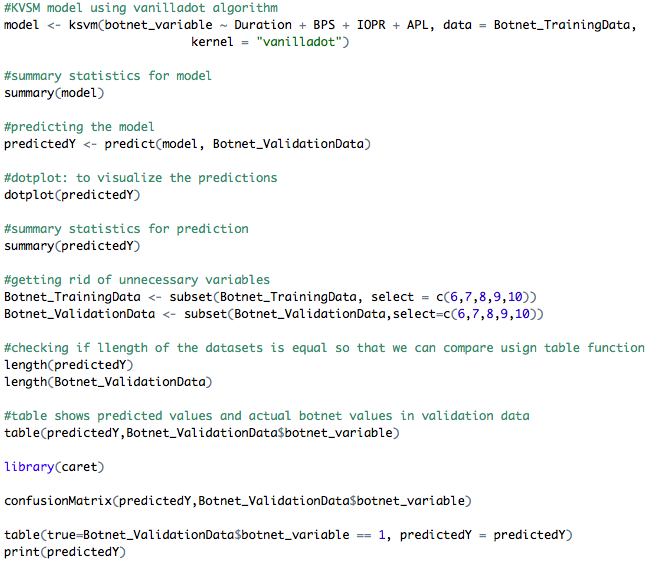
Our model predicted more True Negatives and showed a high detection rate of 99%. Our goal is to develop a model with more truthful detection ability of botnets so that we can predict and identify as many botnets as possible. We plan to achieve better results from a Decision Tree model before building an SVM.

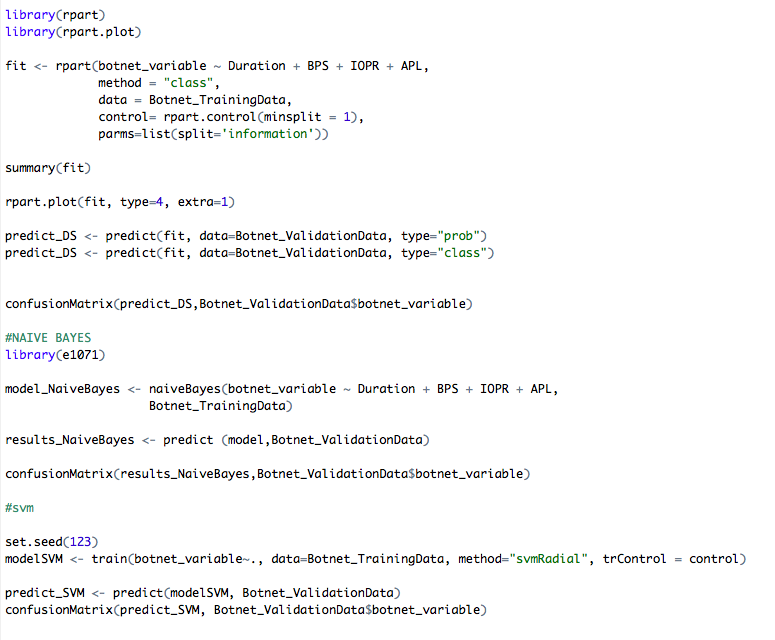
**Decision Tree:**

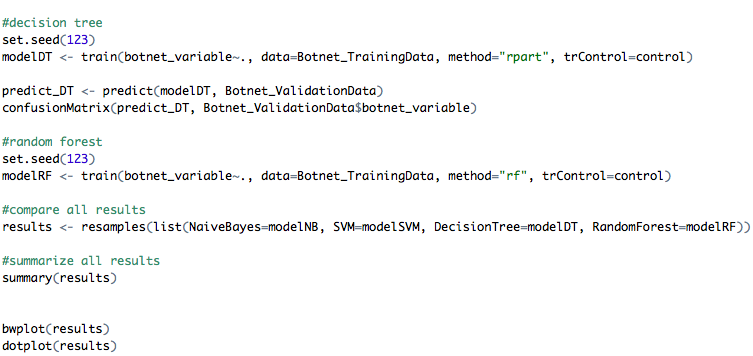


# Other anlaytical models for comparision

We built few other models in order to compare the accuracy of results. Below is the Rcode for these:

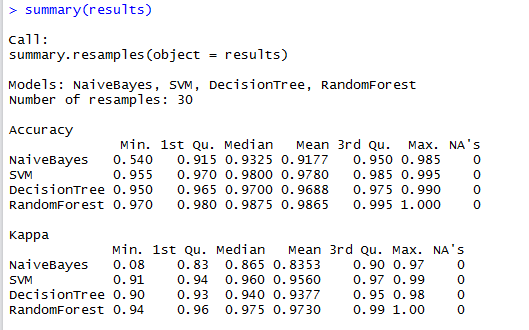




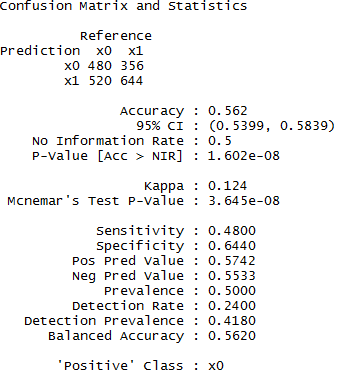


# results - summary of accuracy and graphs

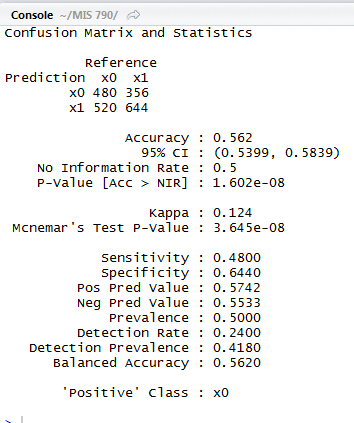
**Comparison Matrix of all models:**

****

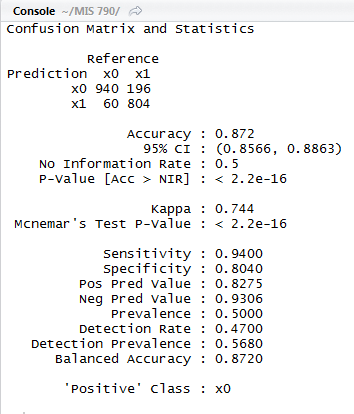
**Confusion Matrix and Statistics for KVSM:**

****

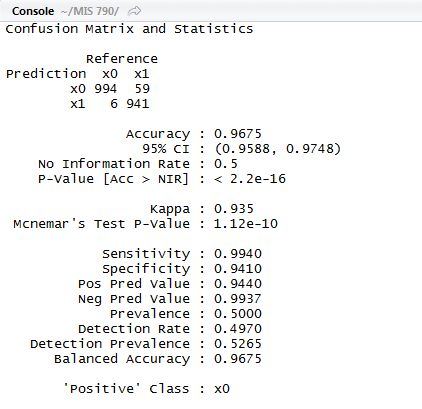
**Confusion Matrix and Statistics for Naïve Bayes:**

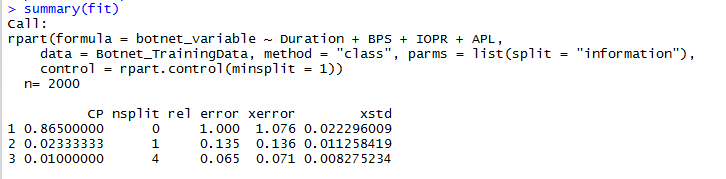
****

**Confusion Matrix and Statistics for SVM:**

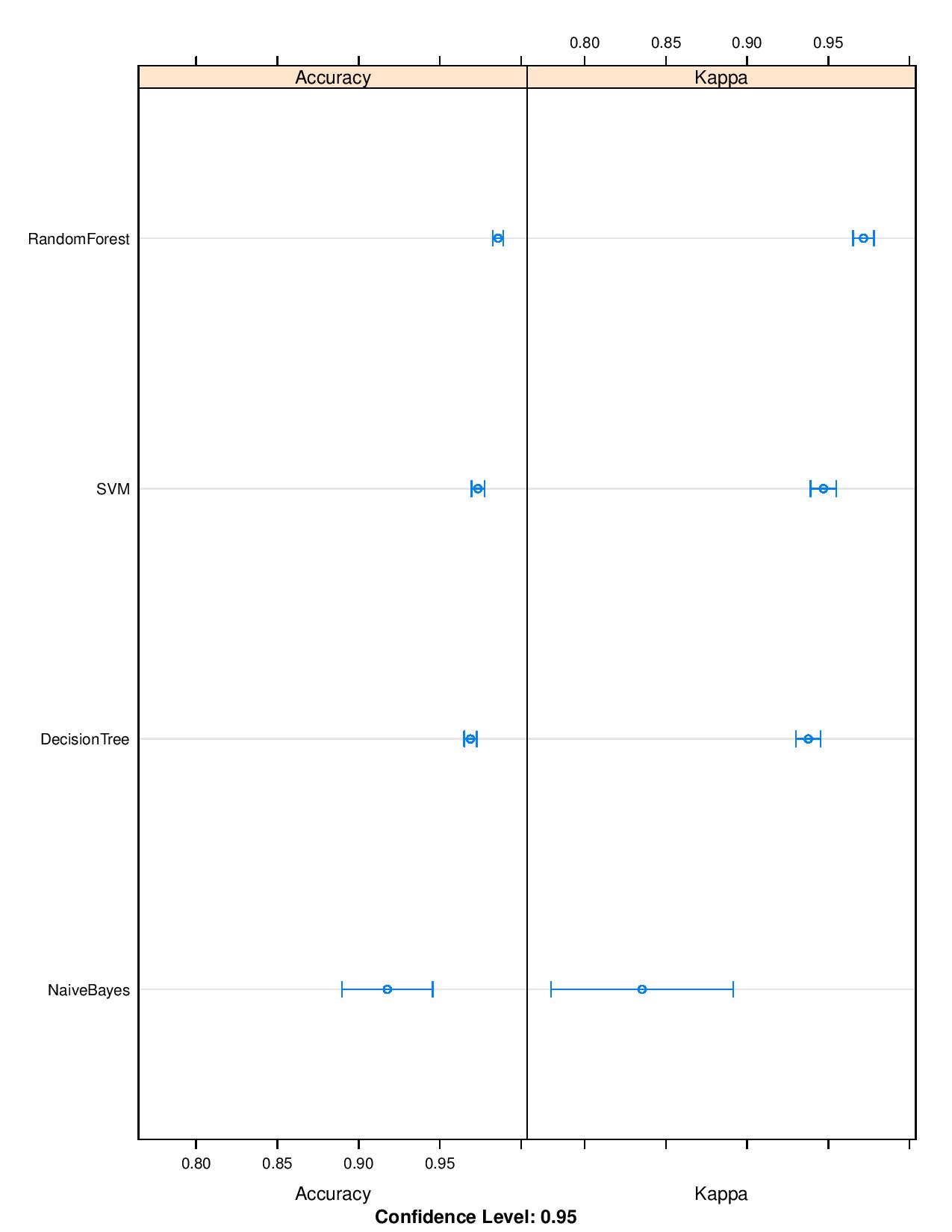
****

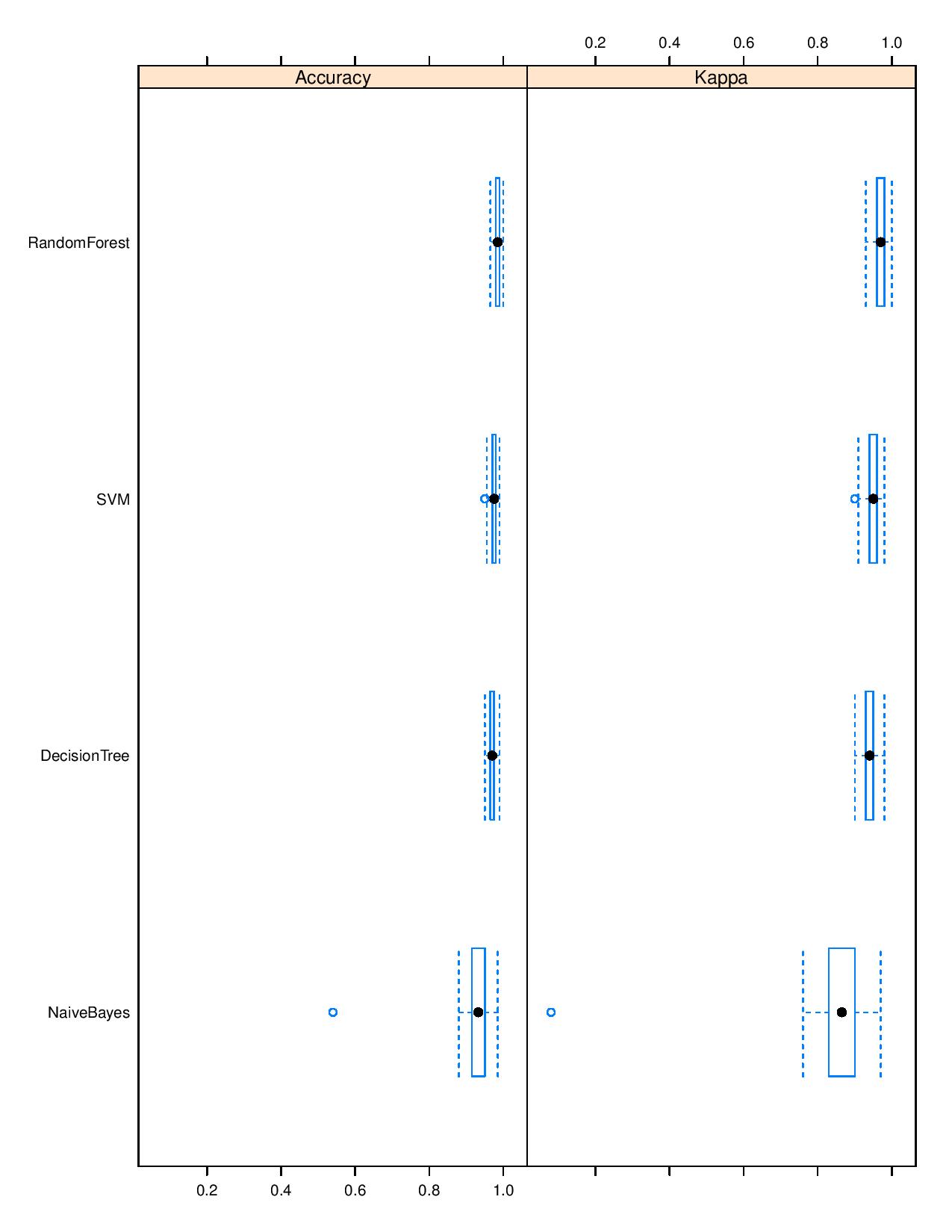
**Confusion Matrix and Statistics for rpart:**

****

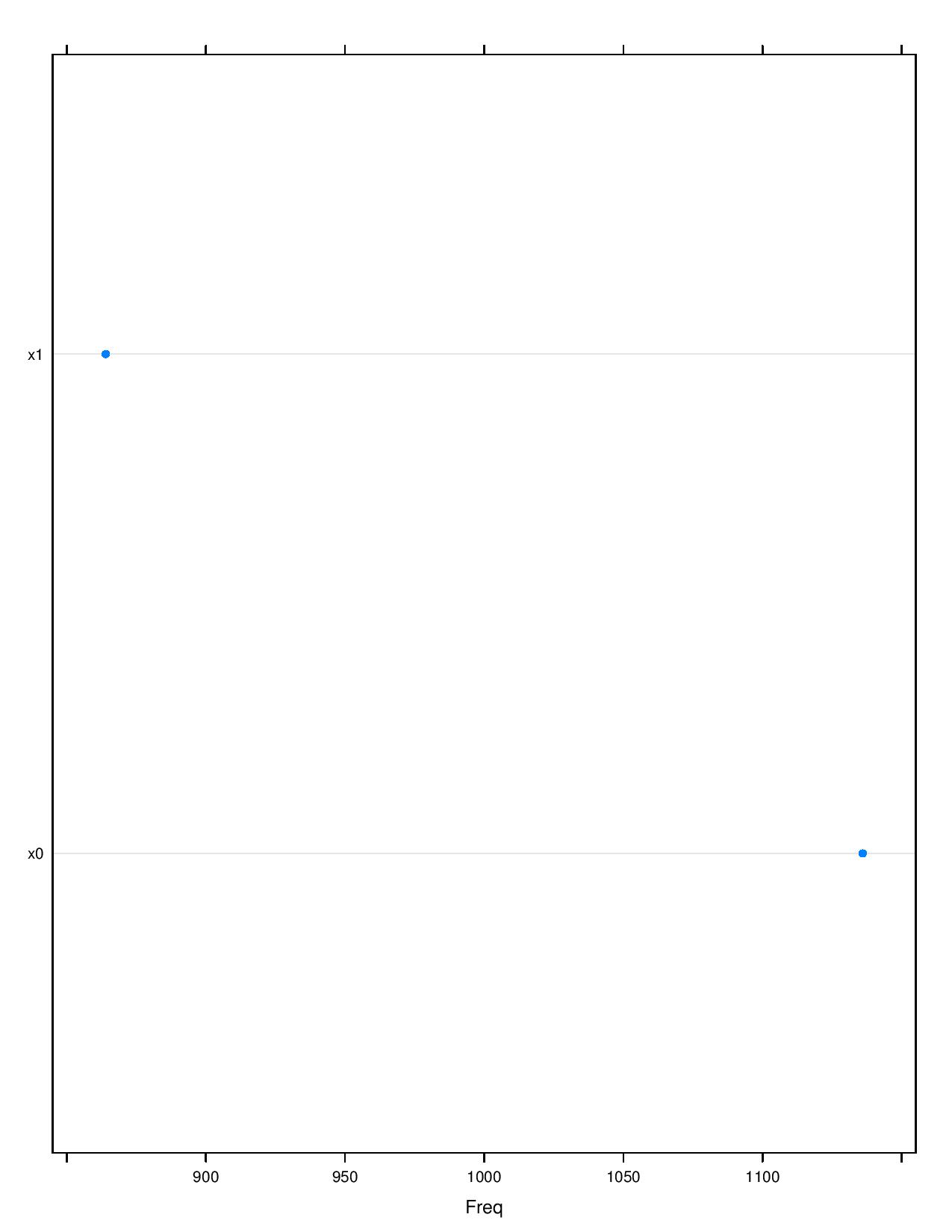
****

**Box plot for Comparison of Algorithms:**

****

****

**Frequency Chart:**

****