**CSE 537 – Assignment 5**

Group Members –

Ankit Agrahari – 110946823 Tarun Lohani – 110921666

Prasad Marla – Rashmi Narvekar –

**Question 1 - Clickstream Mining with Decision Trees**

**1. For each value of threshold, what is your tree's accuracy and size (size equals number of internal nodes and leaves)? What do you observe? If all your accuracies are low, tell us what you have tried to improve the accuracies and what you suspect is failing.**

Following is the tree’s accuracy and size (number of nodes) for each p value –

1. **p-value – 0.01 -**

Accuracy - 0.70976

Size – 266

1. **p-value – 0.05 -**

Accuracy - 0. 7312

Size – 420

1. **p-value – 1 -**

Accuracy - 0. 7102

Size – 17119

We observe that the size of the tree increases as the p-value increases which is expected as lower is the p-value, higher is the chi-square pruning. Also, accuracy is best for the p-value 0.05.

The accuracy is more than 70% for all the p-values which is reasonably good. There were a few cases of ambiguity in building the tree. The tree structure given to us puts an array of size 5 initialized by value -1 for each TreeNode. While building the tree we replace them with the children but there were cases when a node didn’t have all the children. For those missing children, we have put a TreeNode value of “F” and considering it as a leaf value. To improve the accuracy, instead of considering the missing children as leaf of value “F”, we could copy the TreeNode of a smaller sibling to this child. This would help to improve the accuracy. Currently, if in the test data, a missing child is encountered, it returns a “F” value straightaway which results as a 0 in the output.

**2. Explain which options work well and why?**

In our case, p-value of 0.05 works best.