CSCI 720 Big Data Analytics

Assignment 4 – Trisha P Malhotra

**Apriori and FPGrowth**

**Preprocessing**:

* Converted continuous features: Age, Capital Gain, Capital Loss, Hours-per-week to categories using histogram. Made categories for Age, as instance, of teens, twenties, thirties, etc. for age ranges between 0-19, 20-29, 30-39, respectively.
* Similarly, for Capital Gain, Capital Loss, Hours-per-week

**Runtimes of Apriori and FPGrowth:**

* As per the plot showing Run time v/s Apriori and Run time v/s FPGrowth, it is clear that :
* Apriori takes longer to execute than FPGrowth :
* Apriori generates singletons, pairs, triplets, whereas FPGrowth inserts items sorted by frequency into a FPtree
* Runtime for Apriori is slow since candidate generation is extremely slow and execution time increases exponentially depending on the number of different items. On the other hand, for FPGrowth, runtime increases linearly, depending on the number of items and transactions.
* As per the plot, we can observe that Runtime and the MinSupport are inversely proportional.
* For minSupport of 2%, time taken by Apriori is the maximum at ~1500 seconds
* Whereas FPGrowth and its runtime increases for increasing minSupport values.
* For minSupport of 2%, time taken by FPGrowth is barely 6 seconds.

Below given are snapshots of the **plots** :









