BIG DATA ANALYTICS (CSCI -720)

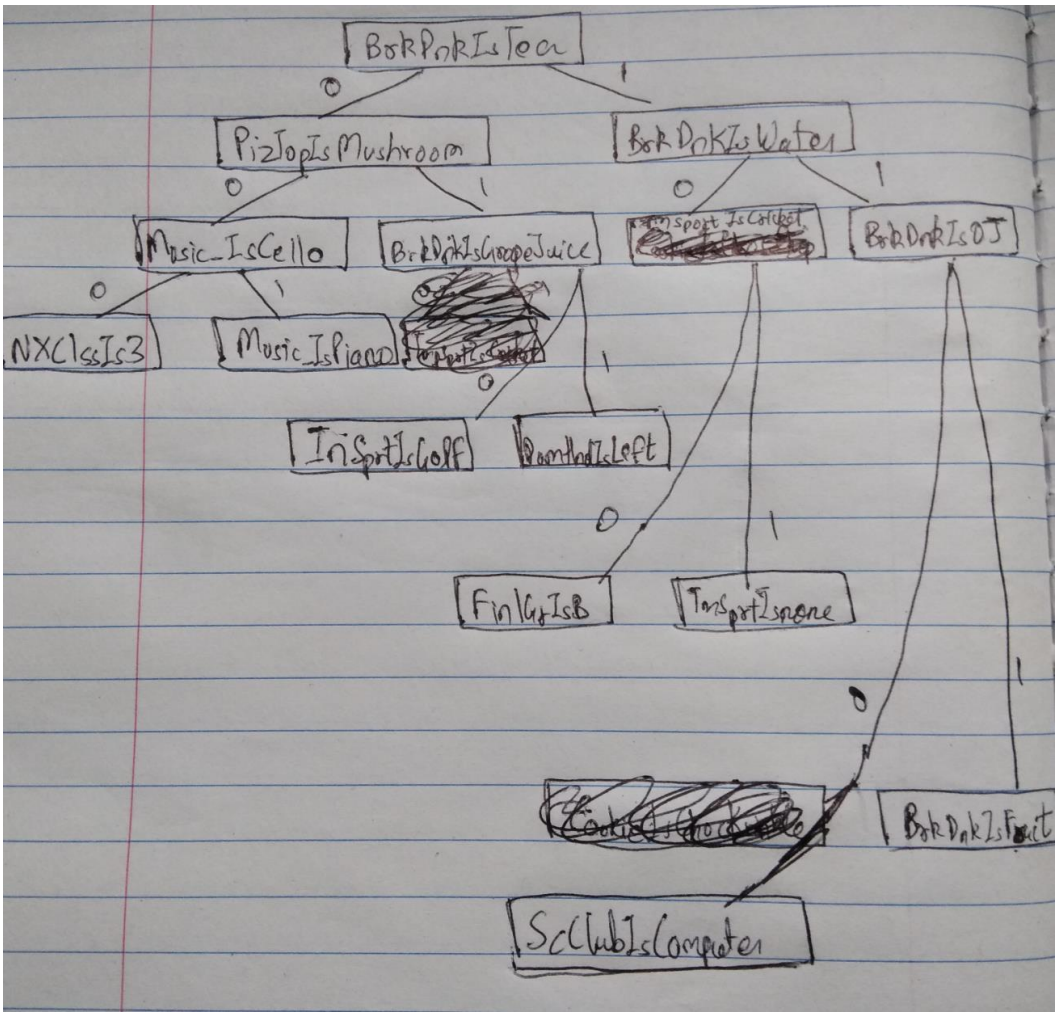
Homework-06 **Name: Rajkumar Lenin Pillai**

Question-a.) What structure did your final decision tree classifier have? What was the if-else tree you got?

Solution:

The attribute numbers in if-else statement is changed because certain attributes which had float numbers and attributes that were containing 0's or 1's are removed in the program since they do not provide any information and also cookie attributes are removed.

Structure of final decision tree classifier:



The if-else tree obtained:

```
if data_record[i][69] ==0:
  class =1
  if data_record[i][75] ==0:

   class =1


   if data_record[i][62] ==0:

    class =1


    if data_record[i][113] ==0:

     class =1



    else:
     if data_record[i][65] ==1:
      class =1
     else:
      class =1



  else:
   if data_record[i][64] ==1:
    class =1
   else:
    class =1


else:
 if data_record[i][63] ==1:
  class =1

 else:
  class =1
```

Question-b.) Run the original training data back through your classifier. What was the accuracy of your resulting classifier, on the training data?

Solution: Accuracy of classifier for validation data is 90.6%

Accuracy of classifier for training data is 91.27%

Output of classifier on training data:

```
Correct:  3651
Incorrect:  349
Accuracy:  91.27499999999999 %
```

Question-c.) What else did you learn along the way here?

Solution:

Bhattacharya coefficient serves as a good measure to understand the information provided by different attributes.

Splitting the attribute with minimum Bhattacharya coefficient yields a good accuracy of 90.6% on validation data which is good. Other functions like entropy , information gain still don't give difference between the attributes which has equal no of samples for each class but Bhattacharya coefficient does.

Question-d.) What can you conclude?

Solution:

We can say that ignoring the attributes which don't give much information is better since it creates confusion and effects the efficiency of the model. So bhattacharya coefficient serves as a good measure of comparing the differentiation power among the attributes