Name	SBU ID
Apurva Kumar	109751733
Gaurav Gawde	110008637

AI Assignment 05 Clickstream Project

Aim:

Implementing ID3 decision tree learner in Python which uses chi-squared split stopping criterion with p-value threshold given as a parameter.

Use thresholds 0.05,0.01 and 1 where 1 corresponds to full tree.

Class Details:

Class Name	Description
Chi Class	This class is designed for calculating chi-squared values.
Clickstream Class	This class is the main class of the project which implements various methods to achieve the goal of the project.
DTreeNode Class	This class represents DTreeNode object.
PageView Class	This class represents PageView object.

Class and Function Details:

1.Chi Class Functions

Functions	• Poz() • ex() • pochisq() • critchi() • main()

2.Clickstream Class Functions

Functions	• computeAccuracy() • buildDataMaps() • learnTree() • chiSquare() • informationGain() • computeRange() • entropy() • predictTree()

	• main()

3.DTreeNode Class Functions

Functions	• DTreeNode() • getIndex() • getName() • getLabel() • getDefaultLabel() • getBranches()

4.PageView Class

Functions	• PageView() • getLabel() • getFeatures()

Statistics:

[A] p-value threshold = 0.01

Test-Data Prediciton Statistics

Tree size: 459
Matches: 18315
Accuracy of Data: 73.26%

[B] p-value threshold = 0.02

```
Test-Data Prediciton Statistics
-------------------------------
Tree size: 678
Matches: 18266
Accuracy of Data: 73.06%
```

[C] p-value threshold = 0.05

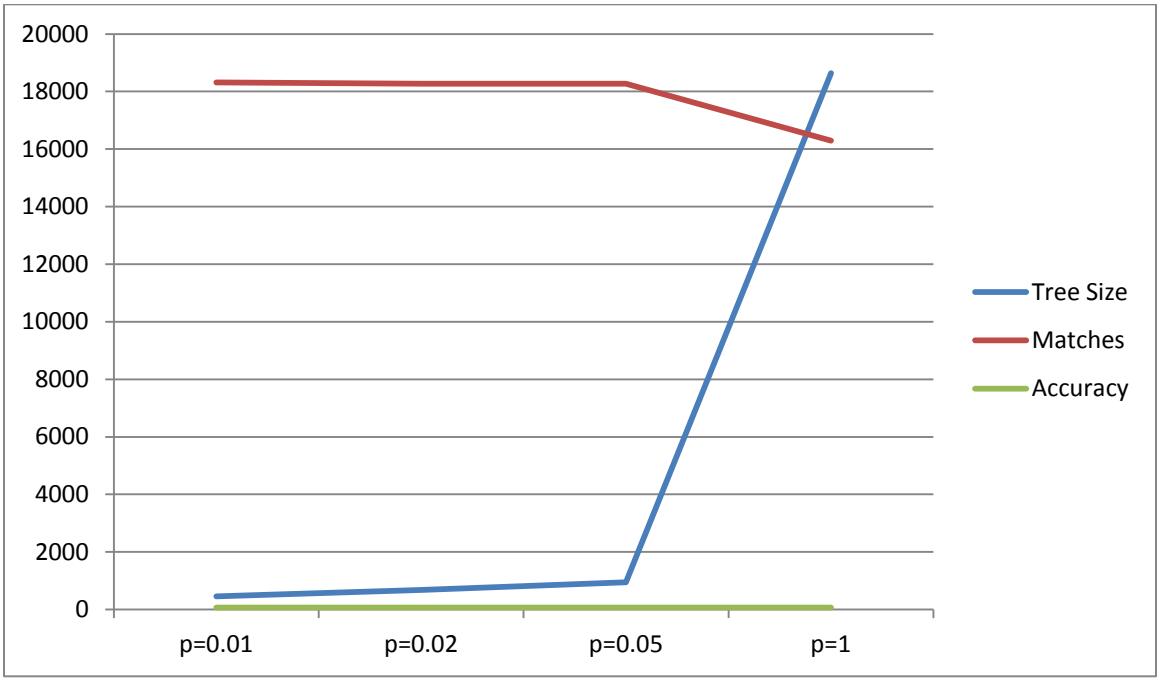
Test-Data Prediciton Statistics

Tree size: 948
Matches: 18266
Accuracy of Data: 73.06%

[D]p-value threshold = 1

```
Test-Data Prediciton Statistics
------------------------------
Tree size: 18643
Matches: 16293
Accuracy of Data: 65.17%
```

Diagrams:



Conclusion:

Thus we can conclude that as we go on increasing threshold value, number of matches goes on decreasing and hence the accuracy of the data reduces.