Big Data Analytics.
Shubham Saxena
Ss4017@rit.edu
Homework 04 – Decision Trees.

Question 1. 3 hrs

Question 2. Files:

a. HW04_Saxena_Shubham_Trainer.java
b. HW04_Saxena_Shubham_Classifier.java
c. HW04_Saxena_Shubham_MyClassifications.txt
d. Data2.java
e. AttributeClassPair.java
f. scatter.m

Question 3. Write-UP: HW04_Saxena_Shubham_

a. Gini Index. (Weighted), entropy.
b. Gini Index. (Weighted).
c. It was easy to implement and did the job.
a. Entropy and gini Index.
d. I used less than equal to when comparing the values so I used the latest minimum. It shouldn't matter which one we pick if the two attributes have equal gini index.
e. No.
f. As advised in the program I used the assumption that this code is to work only for this specific data. So the model knows that it has to work with only 3 attribute columns + 1 class column.
g. I could not locate any noise that needed cleaning.
(It does have noise in it.)
h.

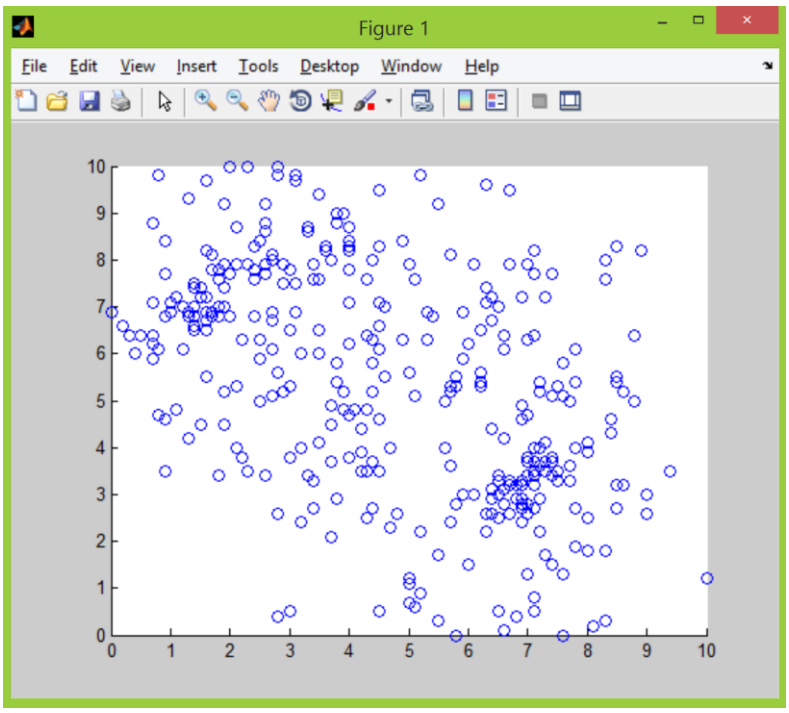
	Attribute Number Used	Threshold Used
a	1	4.2
b	2	3.3
c	3	2.0

i. What was the accuracy of your resulting classifier, on the training data? (This is what you are trying to maximize.)
j. I generated the program.
k. Debugging the giant gini index formula because of the double and int values. Though this is primarily a coding practice issue more than the assignment's issue.
l. Did anything go wrong: Yes. I couldn't find the time to do the plotting API in java. The sorting of the columns had to be done individually and keeping the class values associated with the columns in the sorting.

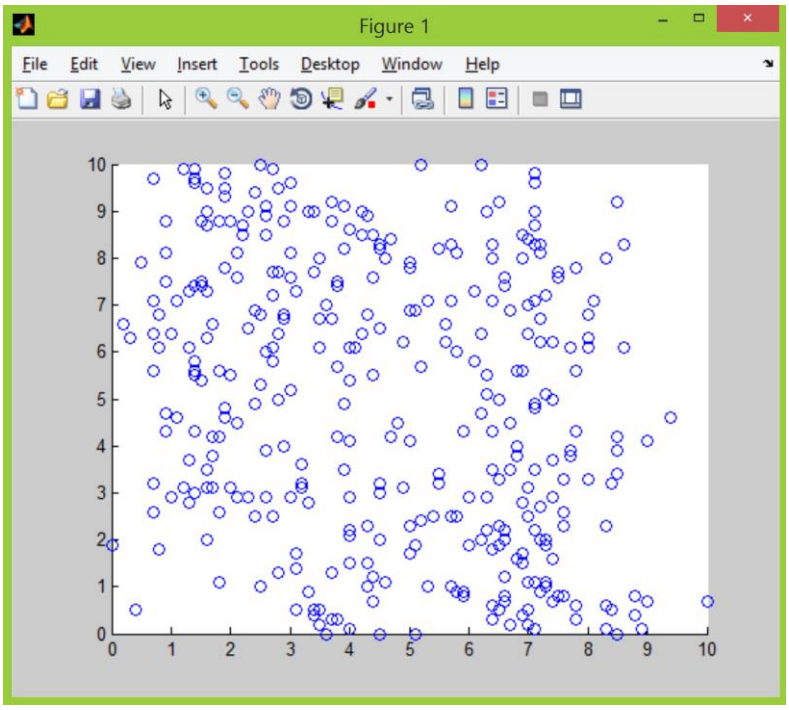
Question 4. 10 hrs.

Big Data Analytics.
Shubham Saxena
Ss4017@rit.edu
Homework 04 – Decision Trees.
Question 5. (1 pts) BONUS:

Attribute 1 vs 2.



Attribute 1 vs 3.



Big Data Analytics.
Shubham Saxena
Ss4017@rit.edu
Homework 04 – Decision Trees.

Attribute 2 vs 3.