Data Mining and Visualisation

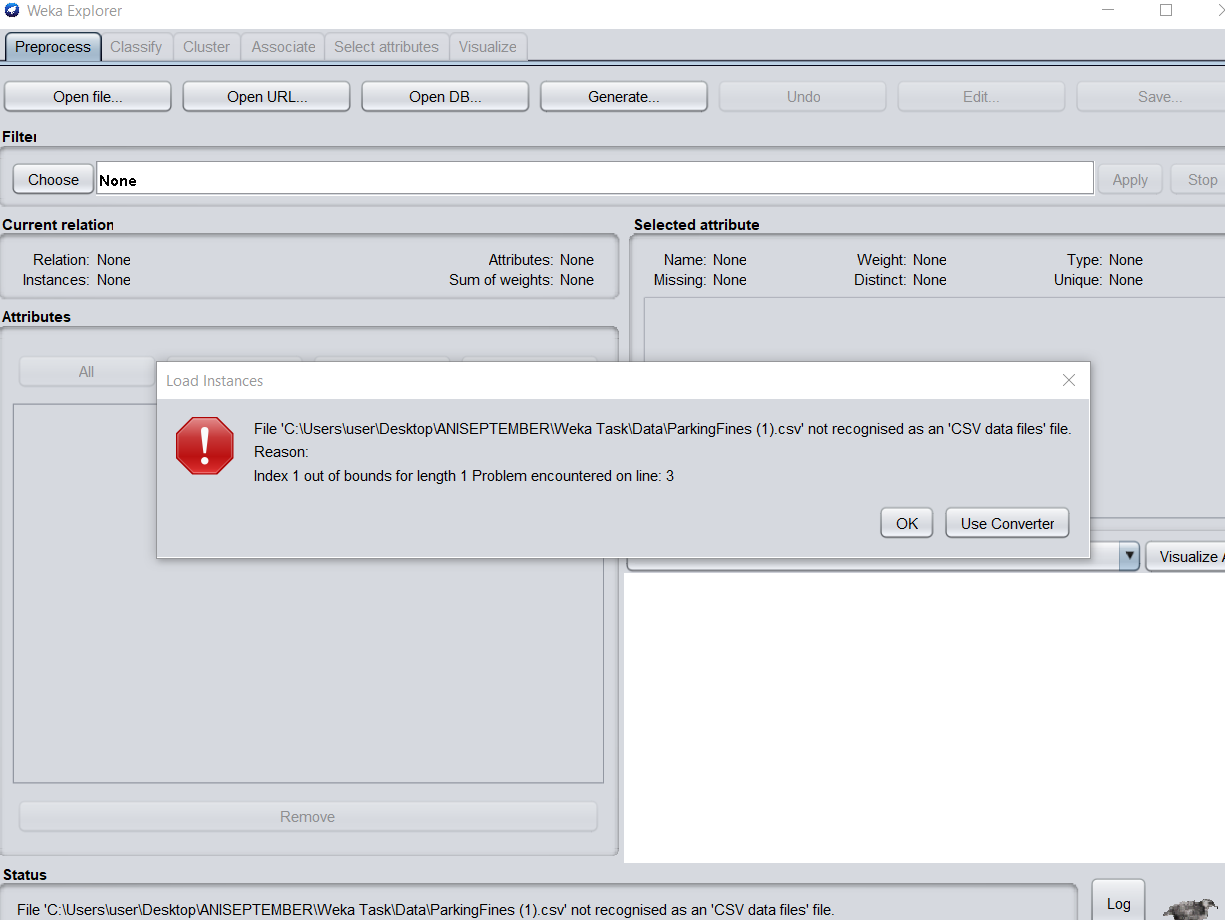
Name of the Student

Name of the University

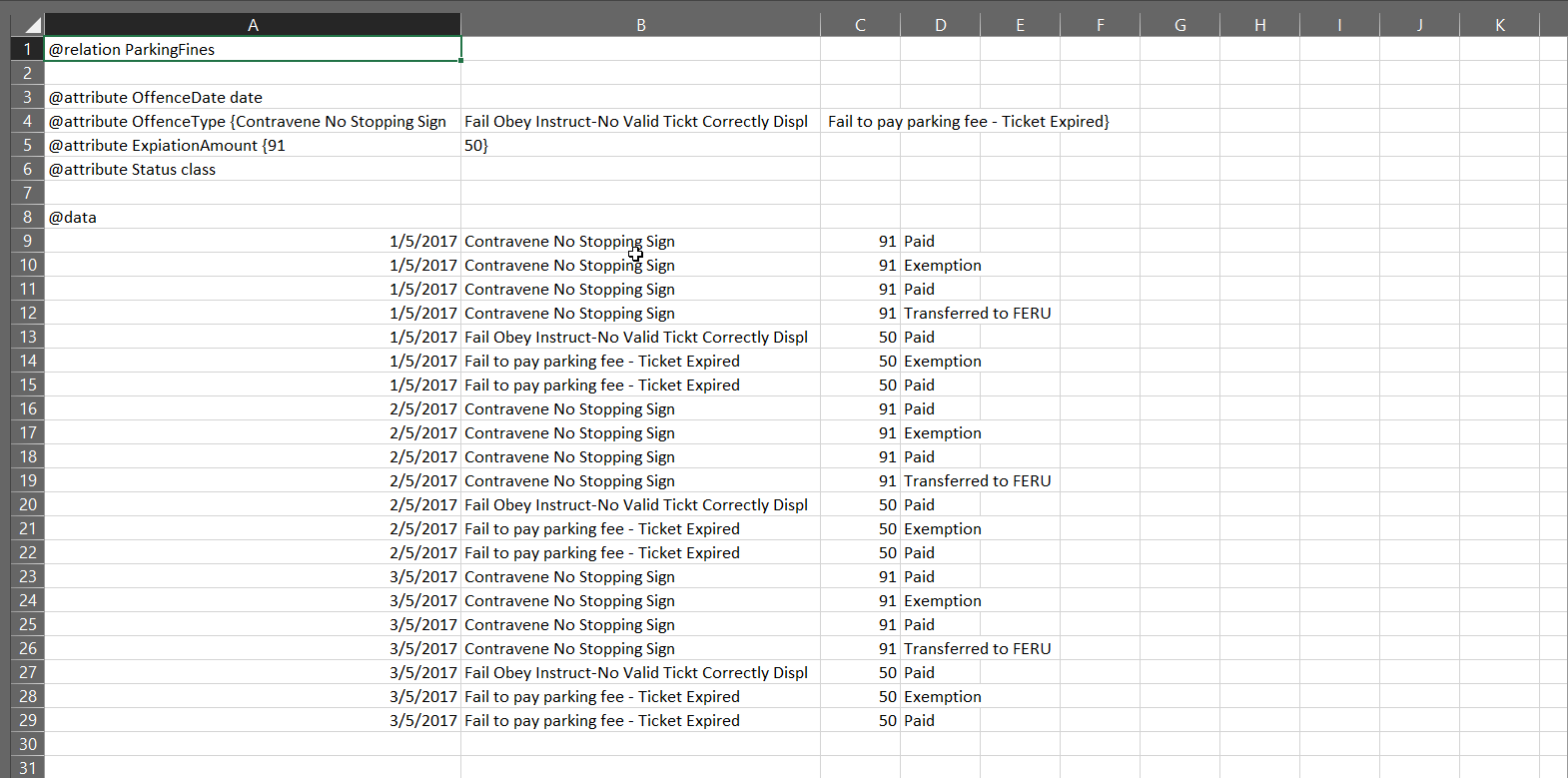
Authors note

# Task 1

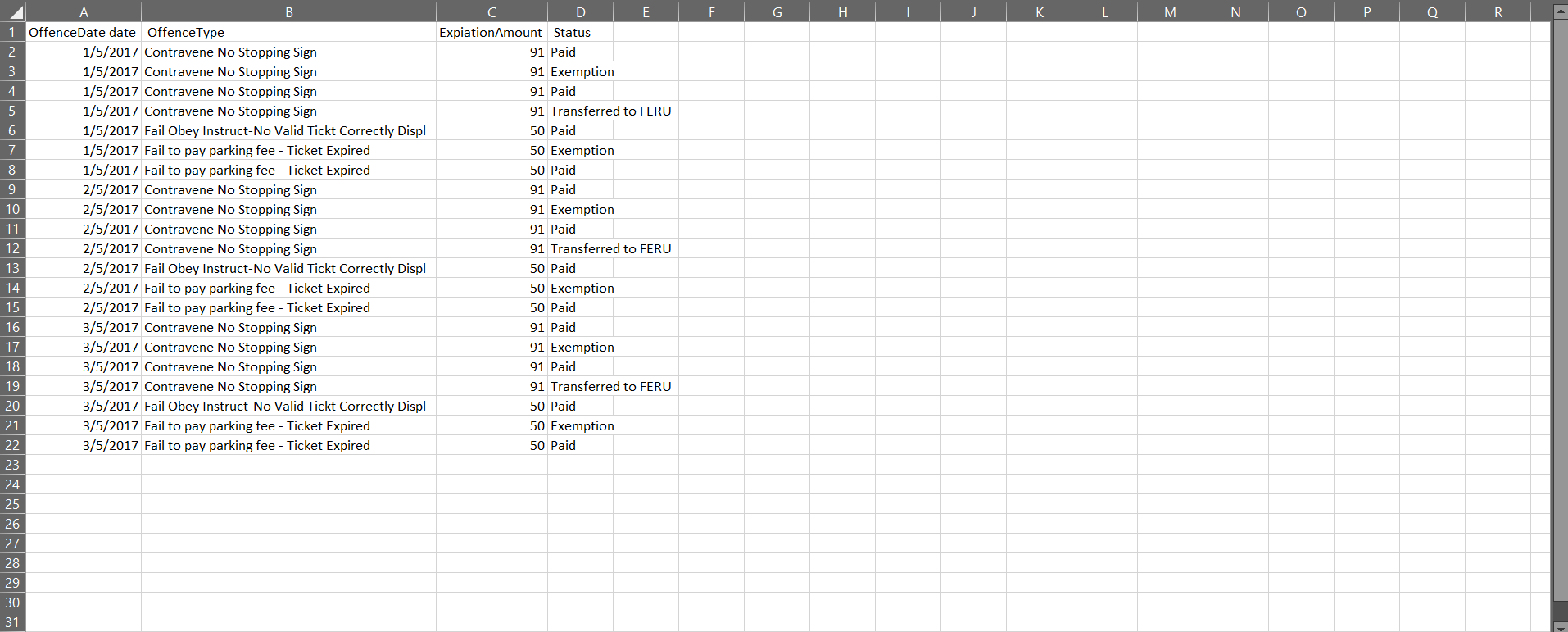
For the given data set when imported to the weka prompts the following error



This error occurs due to the header in the datafiles. When investigated it is found that the data file contains following format;



It can be observed that the file does not have appropriate headers for the data columns which are later transformed into the following format in order to start the analysis.



B

1.

Proportion of people who are accused of committing the offence of “Contravene No Stopping” actually paid is given by following tabular data;

Status = Paid

| OffenceType = Contravene No Stopping Sign : 91 (6/0)

| OffenceType = Fail Obey Instruct-No Valid Tickt Correctly Displ : 50 (3/0)

| OffenceType = Fail to pay parking fee - Ticket Expired : 50 (3/0)

Status = Exemption

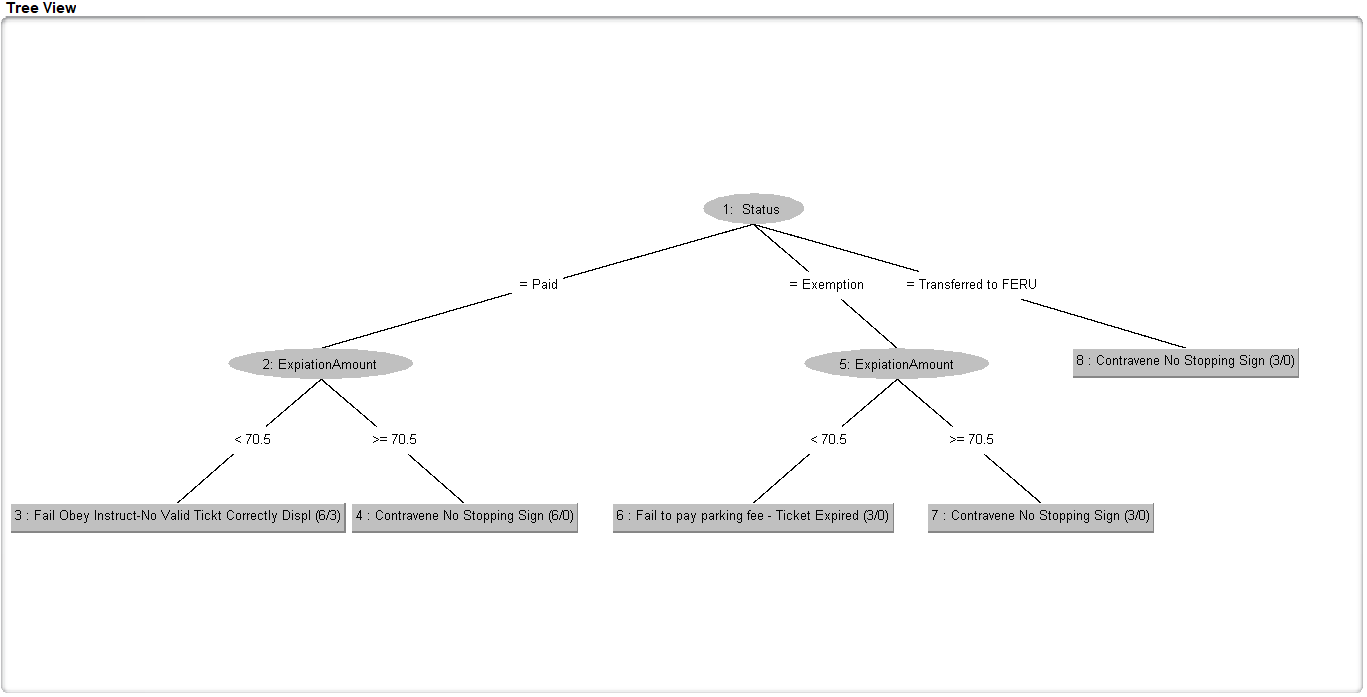
| OffenceType = Contravene No Stopping Sign : 91 (3/0)

| OffenceType = Fail Obey Instruct-No Valid Tickt Correctly Displ : 0 (0/0)

| OffenceType = Fail to pay parking fee - Ticket Expired : 50 (3/0)

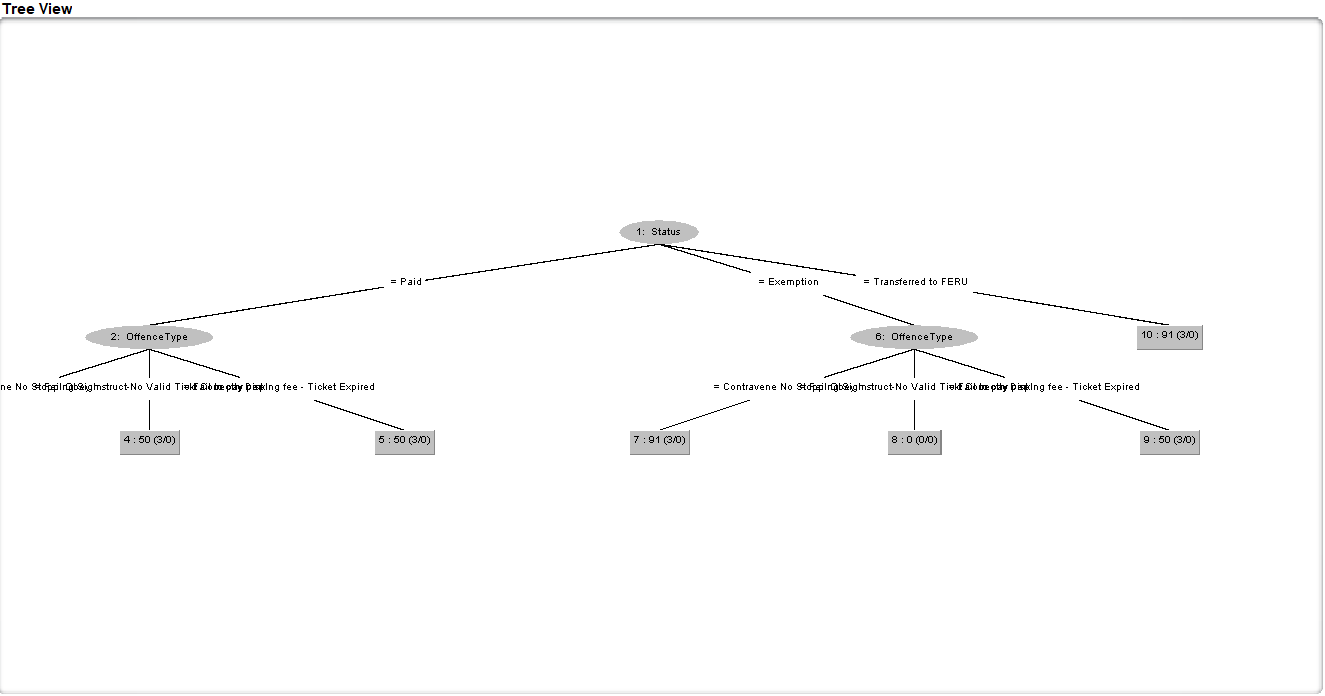
Status = Transferred to FERU : 91 (3/0)

Along with that the number of people who paid the fines is given by 6 as depicted in the following decision tree;



Here it can be seen that 6 people paid the fine for Contravene No stopping sign. Therefore, the proportion is= 6/21=0.2857

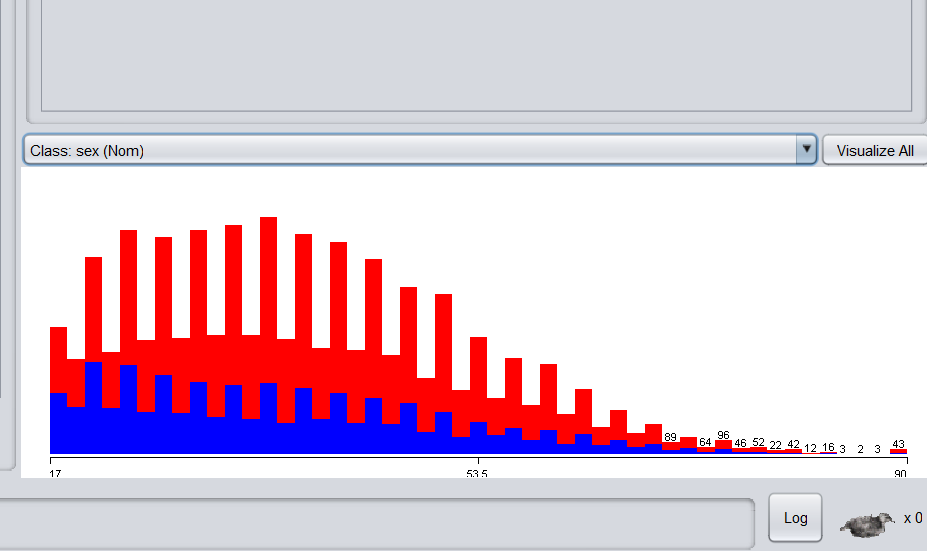
2. from the above data it can be seen that number of people who were expected from the fine of $50 is 3. Therefore, proportion of people who were fined $50 and exempted from paying the fine is given by; 3/21= .1428 following is decision tree that is used for getting the number of exempted cases.



# Task 2

A

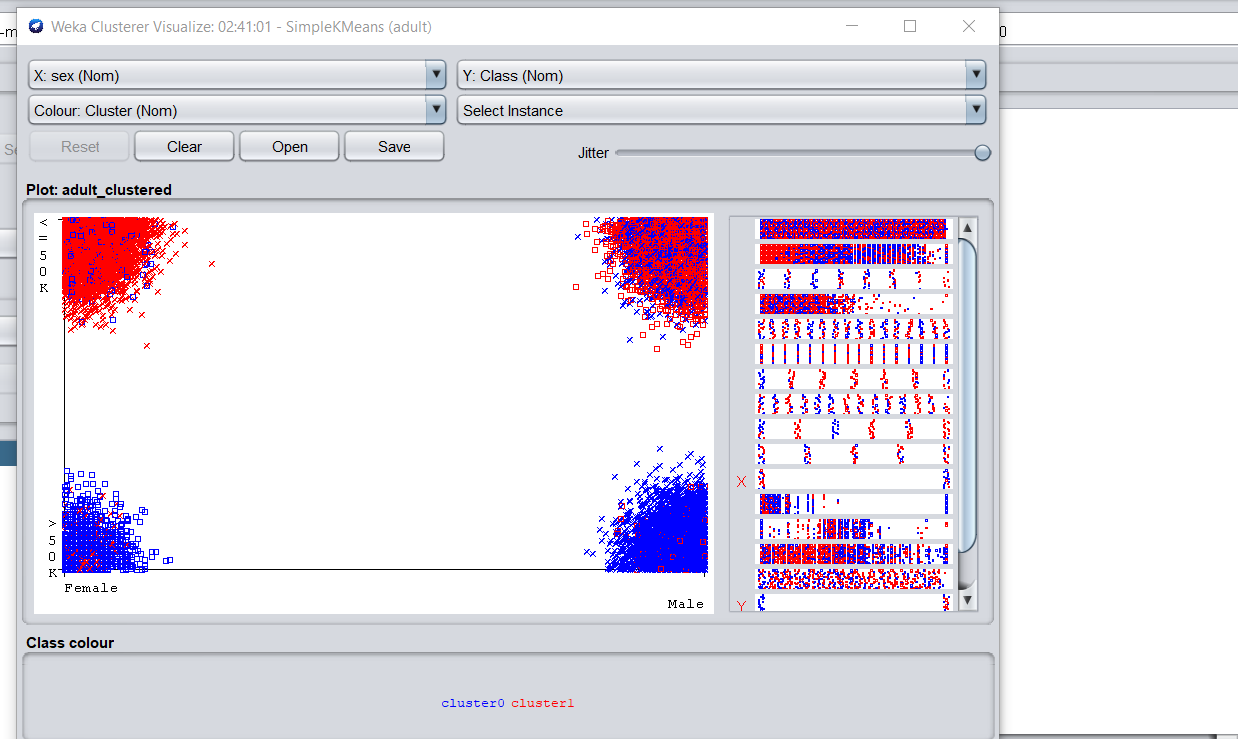
Most popular age bracket is 17-53



B. from the given plot it can be stated that for the age brackets 35 years to 50 the earning of the people whether male or female are same.

In case of age group 60-85 years the earning of the females are higher compared to male.

C. following is the cluster analysis and results given below that proves that there are more number of the men who are earning the less than $50000 compared to the number of females earning less that $50000.



Details are given by;

Class attribute: sex

Classes to Clusters:

0 1 <-- assigned to cluster

1750 9021 | Female

12075 9715 | Male

Where 0 for >50k and 1 for <=50k.

# Task 3

For the given table below;

|  |  |  |
| --- | --- | --- |
| A | B | Class Label |
| T | F | + |
| T | T | + |
| T | T | + |
| T | F | - |
| T | T | + |
| F | F | - |
| F | F | - |
| F | F | - |
| T | T | - |
| T | F | - |

No for the contingency table, we have

A=T A=F

+ 4 0

-3 3

Now for

B=T B=F

+ 3 1

-1 5

Therefore, from the above the entropy before split will be given by;

−(0.4 log 0.4) – (0.6 log 0.6 )

= 0.971

The information gain is given by;

-(4/7)log (4/7) –(3/7)log (3/7)

= .985

-(3/3)log (3/3) – (0/3) log (0/3) [both cancels each other]

=0

Finally the information gain is;

=.971-(4/10)\*.9852-(6/10)\*0

=.281

Here again carrying out the similar steps; we get the information gain as;

=.256

As the information gain for A is higher compared to B thus it can be stated that the decision tree will split depending upon and consequently other nodes of the tree too.

B.

**In case of using the Information gain** as the metric for split, it can be stated that it considered the amount of **info which is**  gained only after knowing value of considered attributes. The gain is actually difference in entropy of the distribution before and after split. Where as in case of gain ratio, it tries to reduce the amount of bias in case of the multi-valued attributes while considering the count and size of branches of the decision tree into account while trying to choose attribute. Therefore as there are only two attributes thus gain ratio would not make too much difference in the results.