Automated Learning and Data Analysis

Team G40

TEAM MEMBER 1 Sujith Tumma,stumma2

TEAM MEMBER 2 Vishnu Challa,vchalla2

TEAM MEMBER 3 Indu Chenchala,ichench

February 10, 2022

1 DATA PROPERTIES

(a) Classify the following attributes as *nominal, ordinal, interval* or *ratio*. Also classify them as *binary*[1], *discreet* or *continuous*. If necessary, give a few examples of values that might appear for this attribute to justify your answer. If you make any assumptions in your answer, you must state them explicitly.

i. **Ratio and Discrete** - As the population count data follows the properties like distinctiveness, ordering and can be used to perform operations like addition, multiplication etc. And also it is discrete as it is a point measure.
Example: 1,173,225 is the population of wake county in North Carolina, This data follows the above mentioned properties.

ii. **Ratio and Continuous** - Annual income can be zero for some people like students who are unemployed and there is no fixed range of these values. The incomes of two employees can be compared by performing operations like addition and multiplication. These values are continuous as they are not a point measure.
Example: 100,275.6 might be an annual salary of a person which follows the above properties.

iii. **Nominal and Binary Classification**- The answer for this problem is either a 'yes' or 'no' which is unique and exists only between those two values.
Example: The answer for the question 'if a patient has cancerous tumours' is either a yes or no.

iv. **Ordinal and Discrete** - These values are unique and can be ordered based on the severity. And also these values are fixed to five types so these are discrete.
Example: Level of pain during injury diagnosis for a patient is 3. It indicates the level of severity.

v. **Interval and Continuous** - pH values always range from 0-14 and are fixed within a given range and is continuous as it is not a point measure.
Example: pH value of a liquid is 5.6 , which says it is less acidic and pH value greater than 7 says it is acidic.

vi. **Nominal and Discrete** - Because the values are unique and are certain about the characteristics of a product.

[1] binary attributes are a special case of discreet attributes

Example: Given examples in the question are distinct and discrete which indicate if a particular cloth is a hat, shirt or pants etc.

vii. **Ratio and Continuous** - An interval variable is a one where the difference between two values is meaningful. A ratio variable has all the properties of an interval variable, but also has a clear definition of 0.0. Variables like height, weight, enzyme activity are ratio variables. Example: The sugar content in a liquid is 5.6, which indicates the quantity as well as the quality.

viii. **Interval and Discrete** - Because the dates in the calendar have a particular range and they end up as a set of fixed values in a given range without any timestamps in them in a calendar. Example: 9-02-2021, this value follows the above mentioned properties.

ix. **Nominal and Binary Classification** - The response for this question is unique and always either a yes or not varying between these two values.

x. **Ordinal and Discrete** - The values always range in between the given 5 types and can be ordered based on the quality of the attribute.
Example: 5 - Excellent , This shows some order of satisfaction and in this particular case , it is discrete.Because we have either of values in 1,2,3,4,5.

(b) Are all continuous attributes ratio? If so, explain why, and if not, give a counterexample?

i. **No**, The pH of water that has been measured using a pH scale is the best example.Because, This is a value on a linear calibrated scale, but it is not relative to a true zero point in time or space. Also, time of the day with a timestamp is continuous and interval due to same above reason.

(c) Are all ratio attributes continuous? If so, explain why, and if not, give a counterexample?

i. **No**, the given statement is false. Because , the Population counts in counties of North Carolina is the best example which is an example for the ratio attributes and is not continuous but it is discrete

(d) Are all ordinal attributes discrete? If so, explain why, and if not, give a counterexample?

i. **No**, the given statement is wrong. The best counter example is Student GPA , which is a ordinal attribute and is continuous.

2 SAMPLING

(a) State the sampling method used in the following scenarios and give a reason for your answer. Choose from the following options: simple random sample with replacement, simple random sample without replacement, stratified sampling, progressive/adaptive sampling.
.

i. **Progressive sampling** : Here, we are continuing the sampling until we reach some percentage of the accuracy this is progressive sampling

ii. **Stratified Sampling**: Here,the number of students in each group is different but the sample should have equal population irrespective of the population size.So this is proportionate stratified sampling

iii. **Random sampling without replacement**: Here, the sample is Random and each participant should not participate more than once

(b) The U.S. Congress is made up of 2 chambers: 1) a Senate of 100 members, with 2 members from each state, and 2) a House of Representatives of 435 members, with members from each state proportional to that state's population. For example, Alaska has 2 Senators and 1 House representative, while California has 2 Senators and 53 House representatives. Both the Senate and

the House are conducting surveys of their constituents, which they want to reflect the makeup of each chamber. You suggest that they use stratified sampling for this survey, sending surveys to a certain number of people from each state. Each survey will be sent to 1200 participants

i. If we take random sampling it may not include people from all the states . So stratified would make more sense for this case.

ii. As mentioned in the above question description each survey is being sent to 1200 participants. That means each person receives a survey resulting in 1200 surveys in total all over the country. With that said Alaska proportion of population is 2/100 which is 1/50. So, on calculating (1/5 * 1200) = 24, in total we will be sending **24 surveys** in Alaska state.

iii. From the interpretation explained in the question (ii) the proportion of California is 53/435. So, on calculating (53/435 * 1200) = 146, in total we will be sending **146 surveys** in California state.

iv. Senate approach is proportionate stratified sampling,with this approach we were able to reach equal proportion of participants. For senate chamber makeup, this approach would make more sense Because it is independent of population.
House approach is disproportionate stratified sampling , with this approach we were able to reach equal number of participants. Since population count affects house chamber this approach would be a better choice.

3 DISCRETIZATION

(a) Discretize the attribute AGE by binning it into 5 equal-width intervals (the range of each interval should be the same, and they should collectively span from the minimum to maximum values). Show your work by writing intervals for each bin

Formulae : binsize = (max - min)/bins (78 - 3)/5 = 15.
So each bin should have an interval of 15. We should be starting with minimum as 3 in the first interval. ANS : Intervals :[3-18), [18-33), [33-48), [48-63), [63-78)

BINS :

Bin1:[3, 8, 14, 14]

Bin2:[52, 57, 57, 58]

Bin3:[68, 68, 70, 71, 71, 75]

Bin4:[78]

(b) Discretize the attribute GLUCOSE LEVEL by binning it into 5 equal-depth intervals (the number of items in each interval should be the same). Show your work.
Formulae : binsize = samplesize/bins = (15/5) = 3
ANS :

BINS :

Bin1:[57, 69, 77]

Bin2:[79, 80, 85]

Bin3:[88, 88, 95]

Bin4:[110, 153, 161]

bin5:[197, 233, 247]

(c) Consider the following new approach to discretizing a numeric attribute: Given the mean ($\bar{x}$) and the standard deviation () of the attribute values, bin the attribute values into the following intervals: [$\bar{x}$ + (k - 1), $\bar{x}$ + k), for all integer values k, i.e. k = . . . - 4, -3, -2, -1, 0, 1, 2 . . . Assume that the mean of the attribute BMI above is $\bar{x}$ = 29 and that the standard deviation = 9. Discretize BMI using this new approach. Show your work

ANS:

k=-2 = [2,11)

k=-1 = [11,20)

k=0 = [20,29)

k=1 = [29,38)

k=2 = [39,47)

For k = -3 and -4 any BMI values does not fall in that range and its the same with k = 3 and k=4.Hence only 4 bins are used for this data.

BINS :

Bin1:[7, 17, 18, 19, 19]

Bin2:[21]

Bin3:[29, 31, 34, 35, 36, 37]

Bin4:[39, 40, 42]

(d) Give an example of a situation where you would want to use equal-width binning, rather than equal-frequency

ANS:

When the data is evenly spread or distributed and is not sparse and if there are not many outliers (outliers make wastage of useless bins) or missing value , In this case it is good to use equal width binning rather than equal frequency.

4 DECISION TREE CONSTRUCTION

(a) Construct the decision tree manually, using Gini index to select the best attribute to split on. The maximum depth of your tree should be 2 (count the root node as depth 0), meaning that any node at depth 2 will automatically be a leaf node, even if it has objects with different classes.

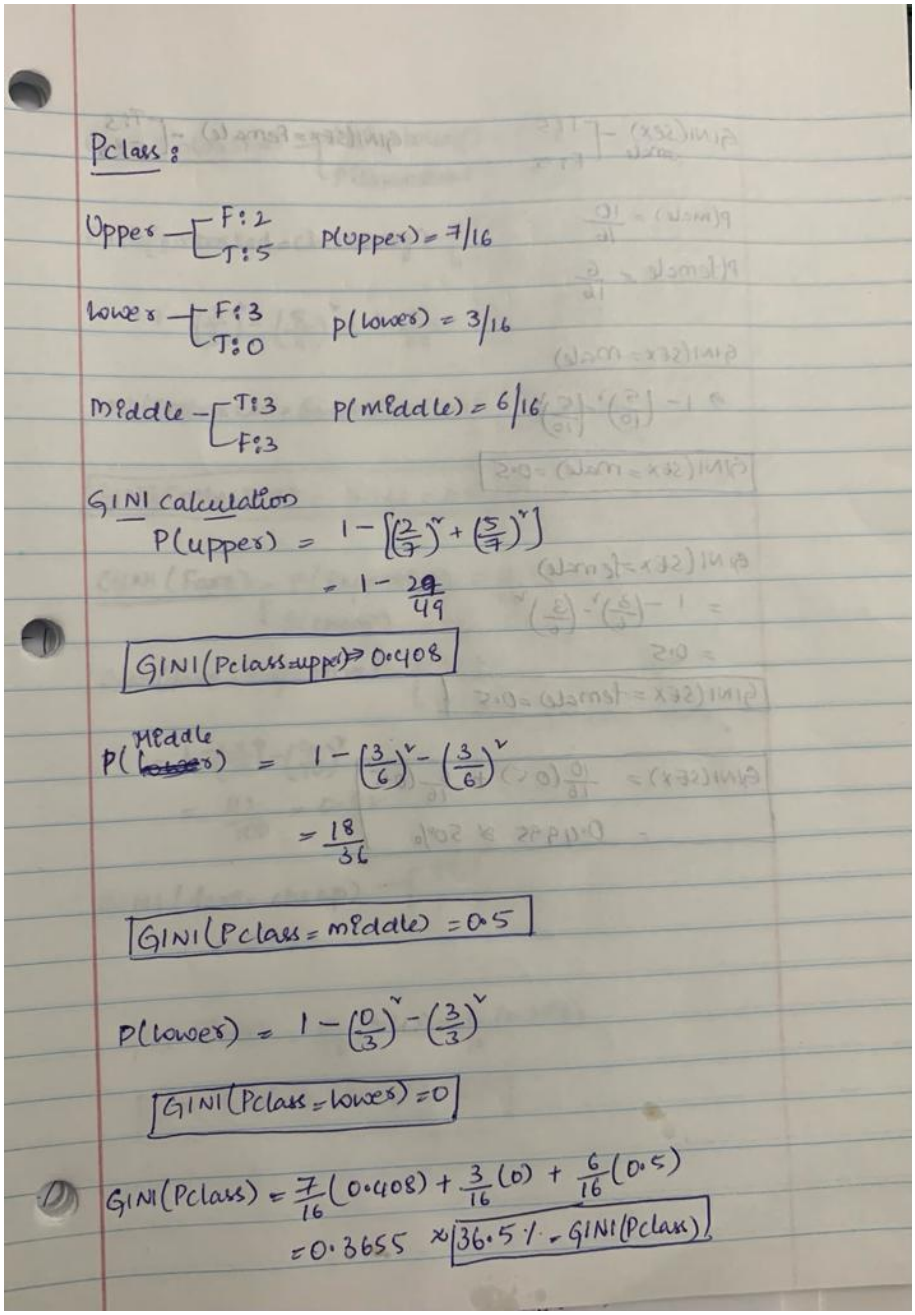

Figure 1: GINI INDEX CALCULATION FOR ROOT NODE SPLIT

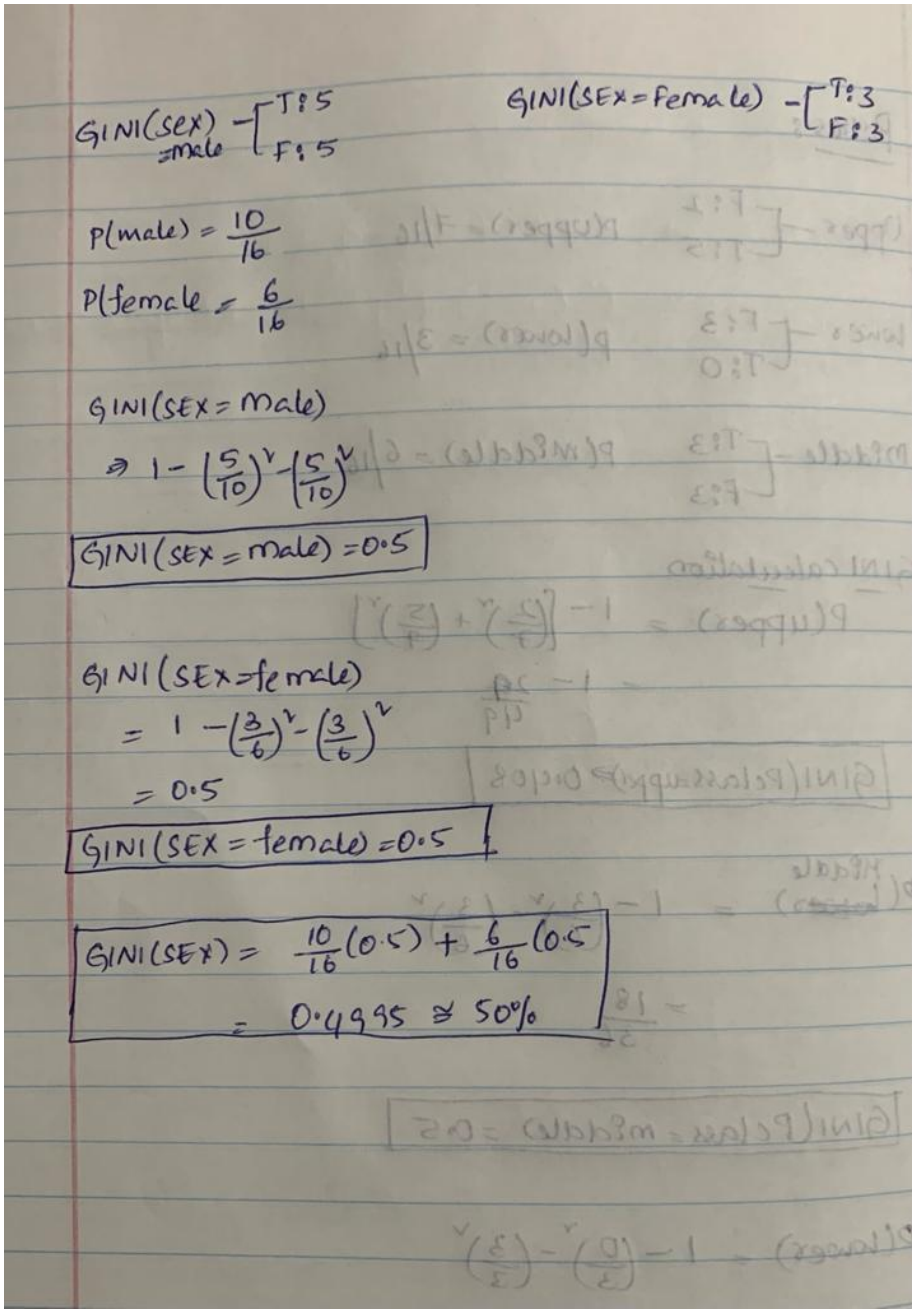

Figure 2: GINI INDEX CALCULATION FOR ROOT NODE SPLIT

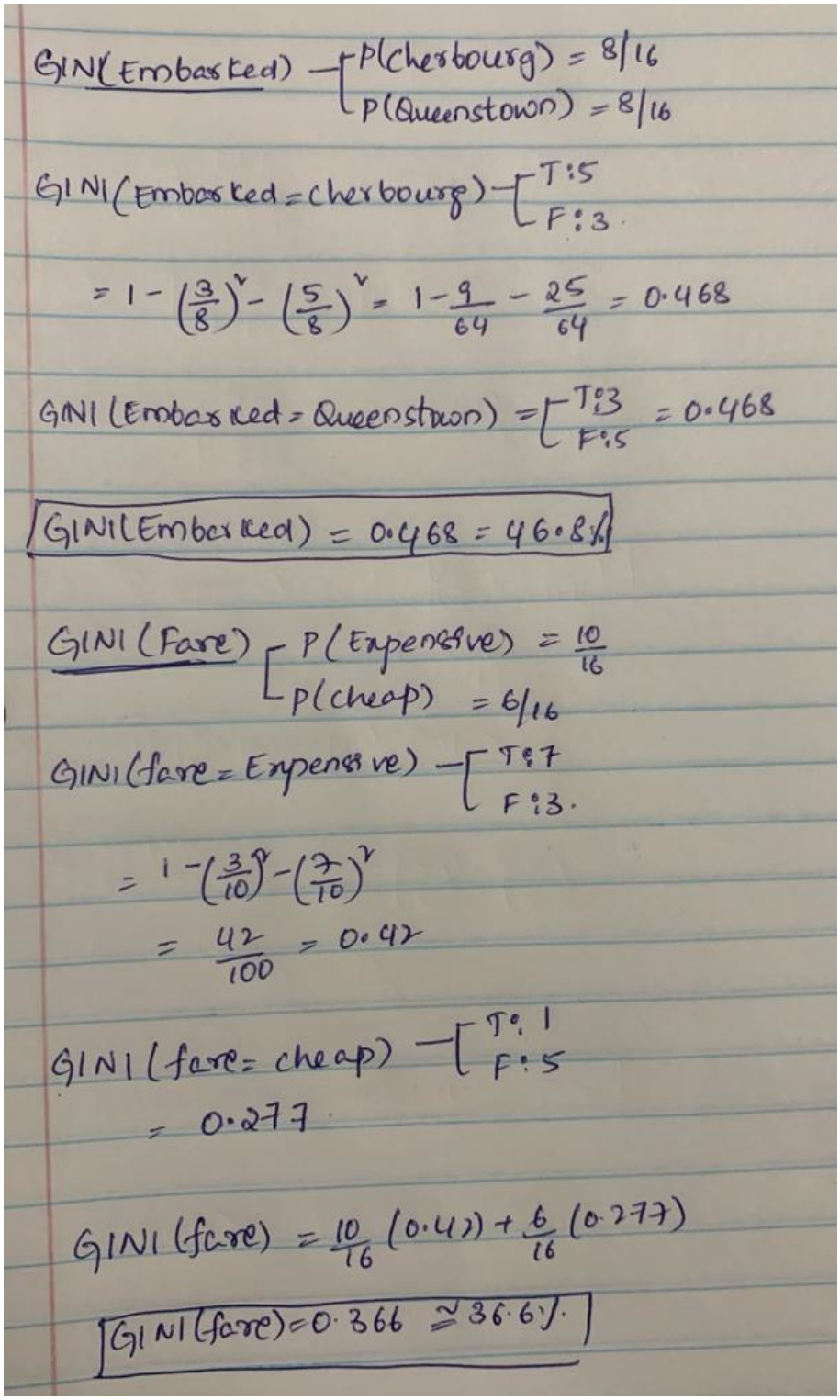

Figure 3: GINI INDEX CALCULATION FOR ROOT NODE SPLIT

1.From the above calculations for the Gini Index we have to select attribute which has less Gini.So we need to choose the Pclass which has less Gini

2.The same procedure applies to the split at the depth node 1

3.On computing all the GINI index ,final decision tree is shown in the below Fig

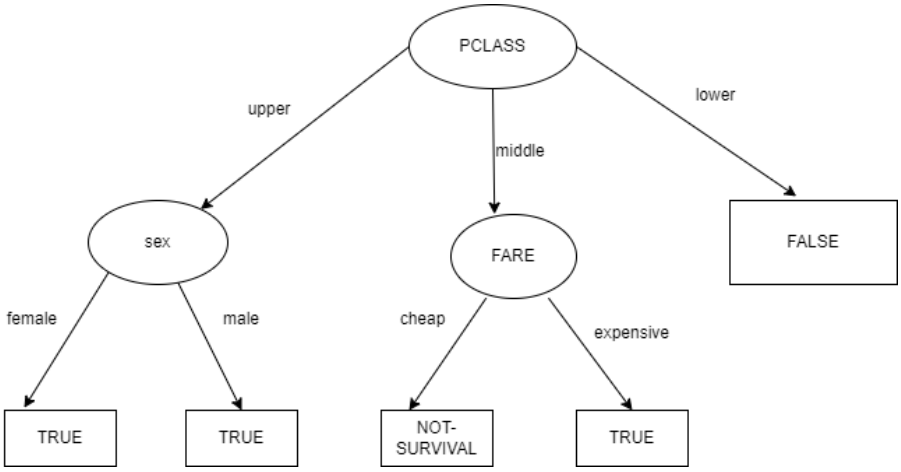

Figure 4: `Final decision tree using GINI INDEX`

(b) Construct the tree manually using Information Gain. The maximum depth of the tree should be 1.

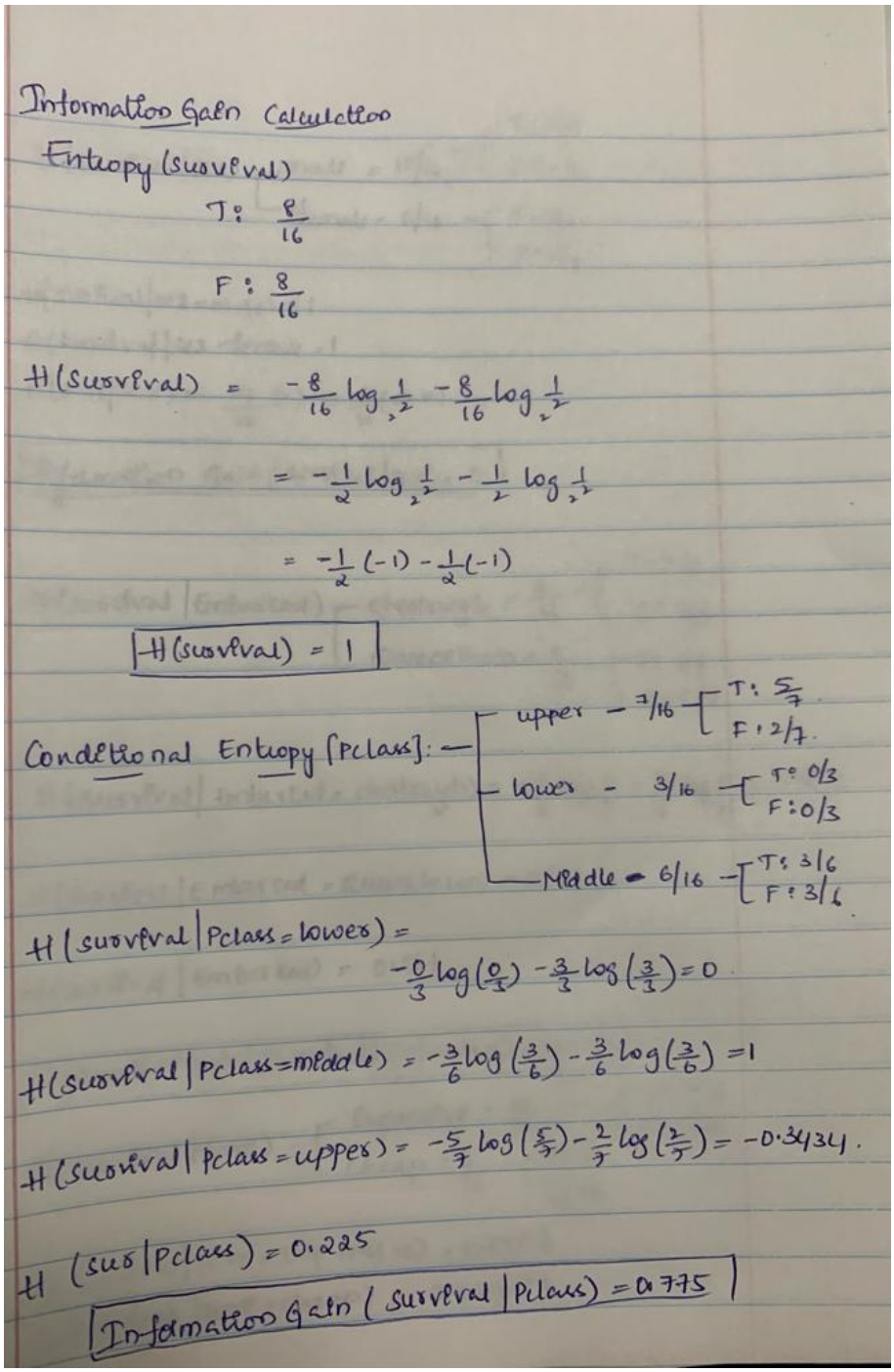

Figure 5: `IG calculation for root node split`

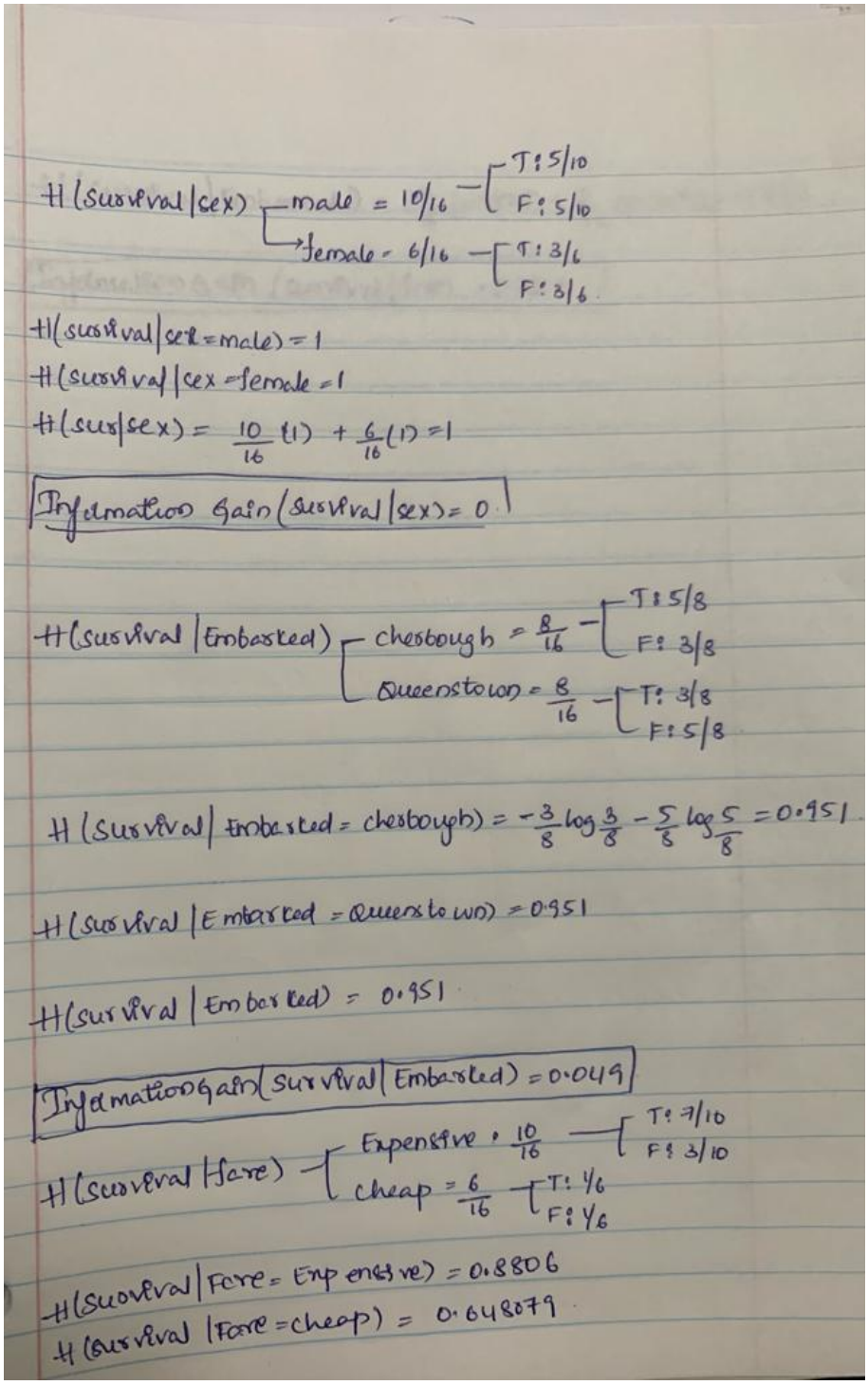

Figure 6: `IG calculation for root node split`

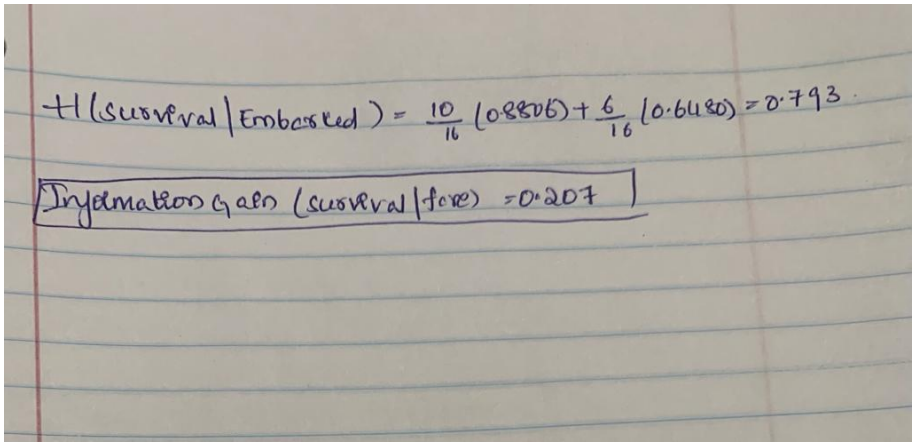

Figure 7: `IG calculation for root node split`

1. From the above calculation we need to select the attribute which has highest Information Gain
2. So we got the attribute PCLASS has Highest Information gain

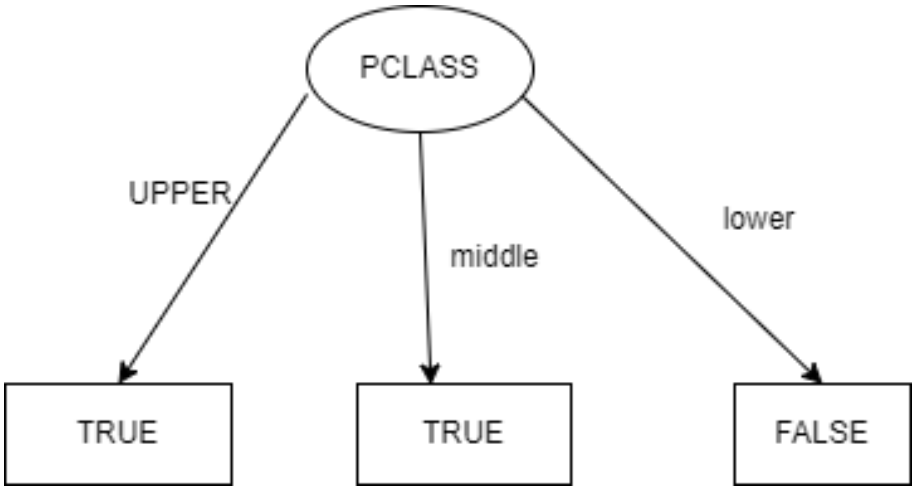

Figure 8: `Final decision tree using Information Gain`

(c) Give an example of a data object (either in the training dataset or not) that would be classified differently by the two trees.

ANS:

FOR Decision tree using GINI:Target value is True for below data object

pclass	**Sex**	**Embarked**	**fare**	**Survival**
Middle	Male	Cherbourg	Cheap	False

FOR Decision tree using Information Gain: Target value is False for given data object

pclass	Sex	Embarked	fare	Survival
Middle	Male	Cherbourg	Cheap	True

(d) Use the training dataset (hw1 dt.csv) to calculate accuracy for each tree (correctly classified). Which decision tree will perform better on the training dataset?

ANS: On Calculating Confusion matrix the values of Confusion matrix we get

For Gini:
TP=8
TN=0
FP=3
FN=5
ACCURACY= (8+5)/16 = 81 percentage

Recall(GINI)=1 For IG:
TP=11
TN=0
FP=2
FN=3
ACCURACY=(11+3)/16= 87.5 percentage

As Accuracy of decision tree using IG is more so the classification using IG performs better

(e) Which will perform better on a test dataset? Can we know the answer?

Recall(GINI)=1

Recall(IG)=1

The tree with the highest recall does the best . Because , When the person survives the trip and we predict that he/she does not survive, then it's not a big loss . But, When the person does not survive and we predict that he/she survives then it's a big loss. So the False positive should be less. So , recall must be very high . For this problem, the statement recall metric would be a better option because of the false positive rate.Also ,F1 score is higher for IG(91%) comparing GINI(84%). The complexity and depth of the tree for GINI is more. So , IG is prefered over GINI.

5 Dimensionality Reduction

(a) Based on the table and figure in Figure 1, do you think that performing PCA was useful? Why or why not? If not, what properties of the dataset caused PCA to be less useful?

By performing PCA we were able to reduce the dimension from 6 to 3 in total based on the idea of considering the features which have eigen values above 1. Performing PCA was useful because, only with 3 dimensions we were able to explain 100% of variance. But we cannot ensure the same behaviour for large dataset as the data might contain larger values as it is not normalized. answer

(b) In Figure 2, what is the most reasonable number of principal components to retain for dimensionality reduction? Briefly justify your choice. Hint: There may be more than one reasonable answer.

Based on the eigen values from the eigen-scree plot above, we can choose only 3 principal components based on rule of thumb eigen value greater than 1. An another interpretation would be, based on the steepness of the slope we can consider all the 6 values as principal components.

(c) Consider 1 instance/row in the dataset, called A, and the PCs in Figure 2 (after normalization). If we increased A's value for Feat 4 by 2 and decreased its value for Feat 5 by 2, after normalization, how would you expect A's value for PC2 to change? Would it increase, decrease, or stay the same? Briefly justify your answer.

The A's value for PC2 might decreases. If we add some value to a feature and subtract the value of same amount in another feature ,the percentage variance described by PC2 might become less significant.