Maggie Lin

mclin

1 Data Properties (13 points)

Q#	Nominal / Ordinal / Interval / Ratio	Binary / Discrete / Continuous	Examples
i)	Ratio	Discrete	Population of 100,000
ii)	Ratio	Discrete	Income of $30,000
iii)	Nominal	Binary	Cancer is present/absent (no order between the two answer choices)
iv)	Ordinal	Discrete	Patient experiencing severe pain
v)	Interval	Continuous	7.0 or 9.1 or 2.3
vi)	Nominal	Discrete	Hat, shirt, pants, shoes (no order to the clothing)
vii)	Ratio	Continuous	20 grams or 300 grams
viii)	Ordinal	Discrete	January 1st comes before January 2nd
ix)	Nominal	Binary	Yes / No (no order between the two answer choices)
x)	Ordinal	Discrete	5 - Excellent

1 b)

Answer: No, not all continuous attributes are ratios.
Explanation / Counterexample: Celsius or Fahrenheit are continuous variables since they can be represented by any real values within a certain logical range; however, both measurements of temperature lack an absolute zero since zero degrees celsius or fahrenheit does not represent the absence of temperature. The zero degree in Celsius and Fahrenheit is just a point on the temperature scale and without a meaningful zero value that indicates “none”, these two measurements of temperature are continuous but not ratio.

1 c)

Answer: No, not all ratio attributes are continuous.
Explanation / Counterexample: While ratio attributes have the addition and multiplication properties and contain a meaningful zero value, it does not necessarily have to be represented by continuous real values. For example, both population and income have a meaningful zero value which represents the absence of the subject but people cannot be split in half and money cannot have fractional values greater than the hundredth or else coins can no longer represent the money value.

1 d)

A**nswer:** Yes, all ordinal attributes are discrete.
Explanation / Counterexample: Ordinal attributes do not have the addition and multiplication properties meaning that even if numbers are used to order objects, precision of the number means nothing. Hence, ordinal attributes are usually measured on a scale with finite or countably infinite set of values to order items but not infinite numbers of values within a given range due to uneven or unmeasurable intervals.

2 Sampling (10 points)

2 a)

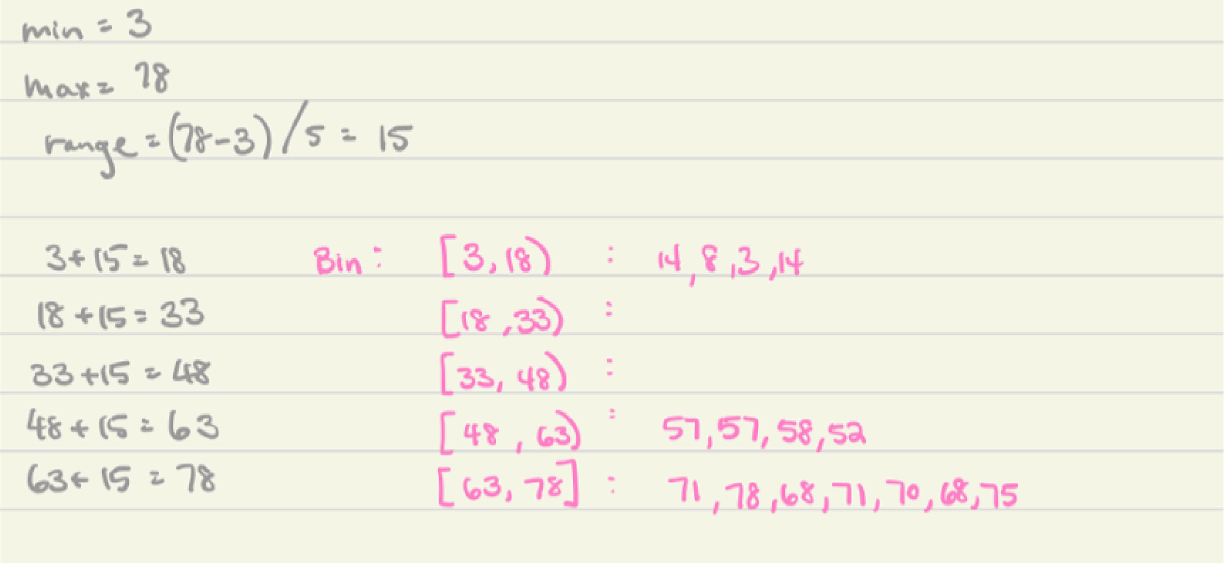
Q#	Simple Random Sample With Replacement/ Simple Random Sample Without Replacement / Stratified Sampling / Progressive Sampling	Reason
i)	Progressive/Adaptive Sampling	The sample will grow in size until it reaches 85% accuracy on the validation data set.
ii)	Stratified Sampling	There are three different groups of programming languages and all three sets have to have an equivalent population in your sample.
iii)	Simple Random Sample Without Replacement	Each computer science student will have the same probability of being selected and since a participant cannot participate more than once, there will be no replacement/repeat of the same study participant.

2 b)

Q#	Reason
i)	Stratified sampling is appropriate here because each state(group) of the United State(population) is represented in the sample. Depending on which type of stratified sampling is used, representation will vary.
ii)	1200 Participants /50 States = 24 Surveys
iii)	(53 California House Representatives / 435 House Representatives) * 1200 Participants = 146 Surveys
iv)	If we stratify by the Senate method, we can ensure all states are represented equally with every state having the same amount of votes. This method is also much more straightforward and simple to calculate. If we stratify by the House method, we can ensure that the representation is proportional to the population size; however, we will have to keep track of the census and population size since those numbers can change. While this method is a bit more tedious, it helps better represent the majority vote based on what the majority wants.

3 Discretization (12 points)

3 a)



min = 3

max = 78

range = (78-3) / 5 = 15

3+15 = 18 Bin: [3, 18) : 14, 8, 3, 14

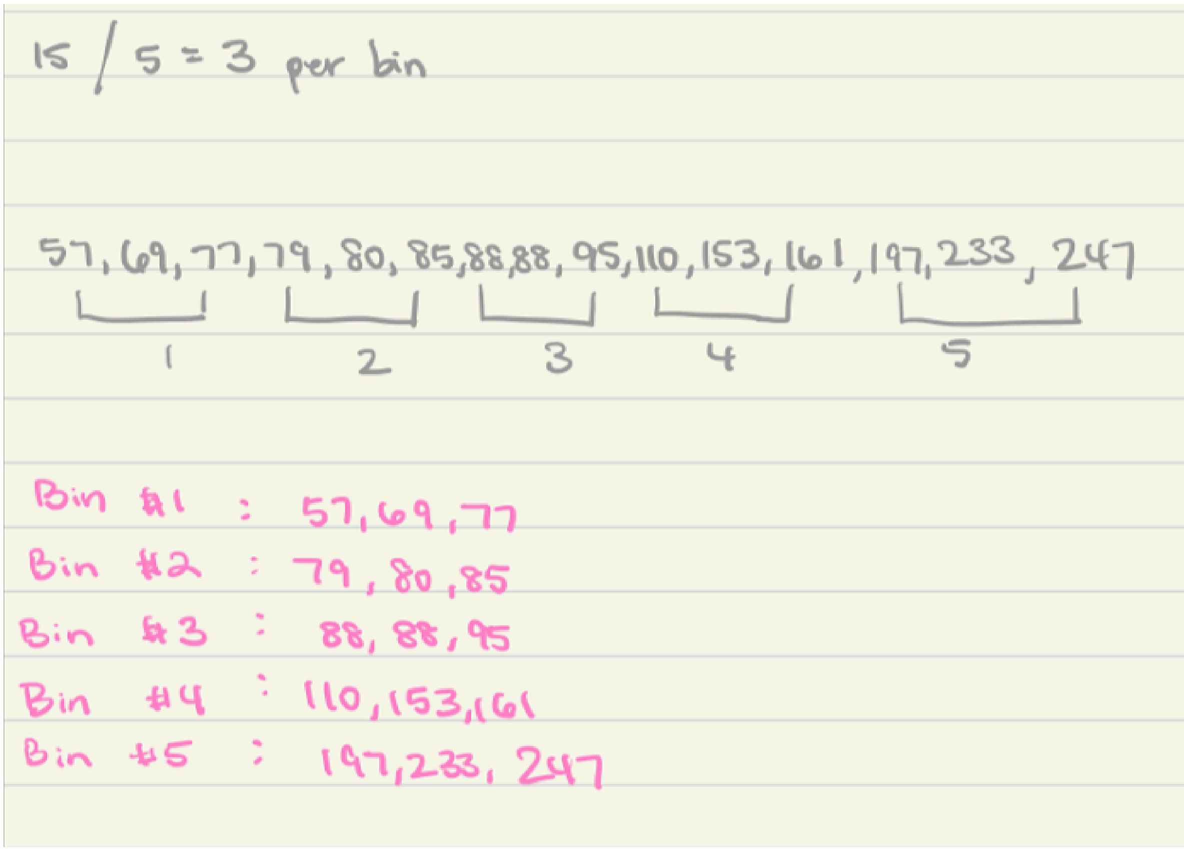
18 + 15 = 33 [18, 33) :

33 + 15 = 48 [33, 48) :

48 + 15 = 63 [48, 63) : 57, 57, 58, 52

63 + 15 = 78 [63, 78] : 71, 78, 68, 71, 70, 68, 75

3 b)



15 / 5 = 3 per bin

$$\underbrace{57, 69, 77}_{1}, \underbrace{79, 80, 85}_{2}, \underbrace{88, 88, 95}_{3}, \underbrace{110, 153, 161}_{4}, \underbrace{197, 233, 247}_{5}$$

Bin #1 : 57, 69, 77

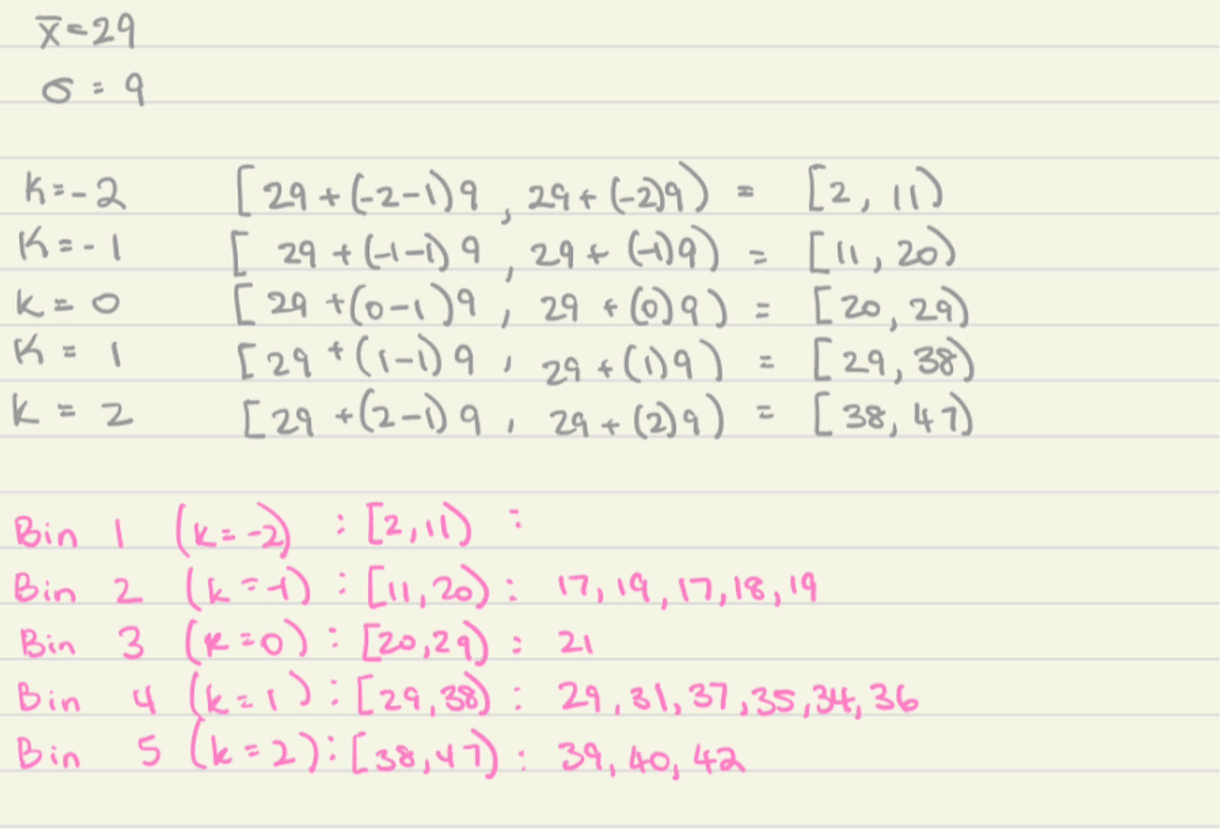
Bin #2 : 79, 80, 85

Bin #3 : 88, 88, 95

Bin #4 : 110, 153, 161

Bin #5 : 197, 233, 247

3 c)



$\bar{x} = 29$

$\sigma = 9$

$k = -2 \quad [29 + (-2-1)9,\ 29 + (-2)9) = [2, 11)$

$k = -1 \quad [29 + (-1-1)9,\ 29 + (-1)9) = [11, 20)$

$k = 0 \quad [29 + (0-1)9,\ 29 + (0)9) = [20, 29)$

$k = 1 \quad [29 + (1-1)9,\ 29 + (1)9) = [29, 38)$

$k = 2 \quad [29 + (2-1)9,\ 29 + (2)9) = [38, 47)$

Bin 1 $(k=-2)$: $[2,11)$:

Bin 2 $(k=-1)$: $[11,20)$: 17, 19, 17, 18, 19

Bin 3 $(k=0)$: $[20,29)$: 21

Bin 4 $(k=1)$: $[29,38)$: 29, 31, 37, 35, 34, 36

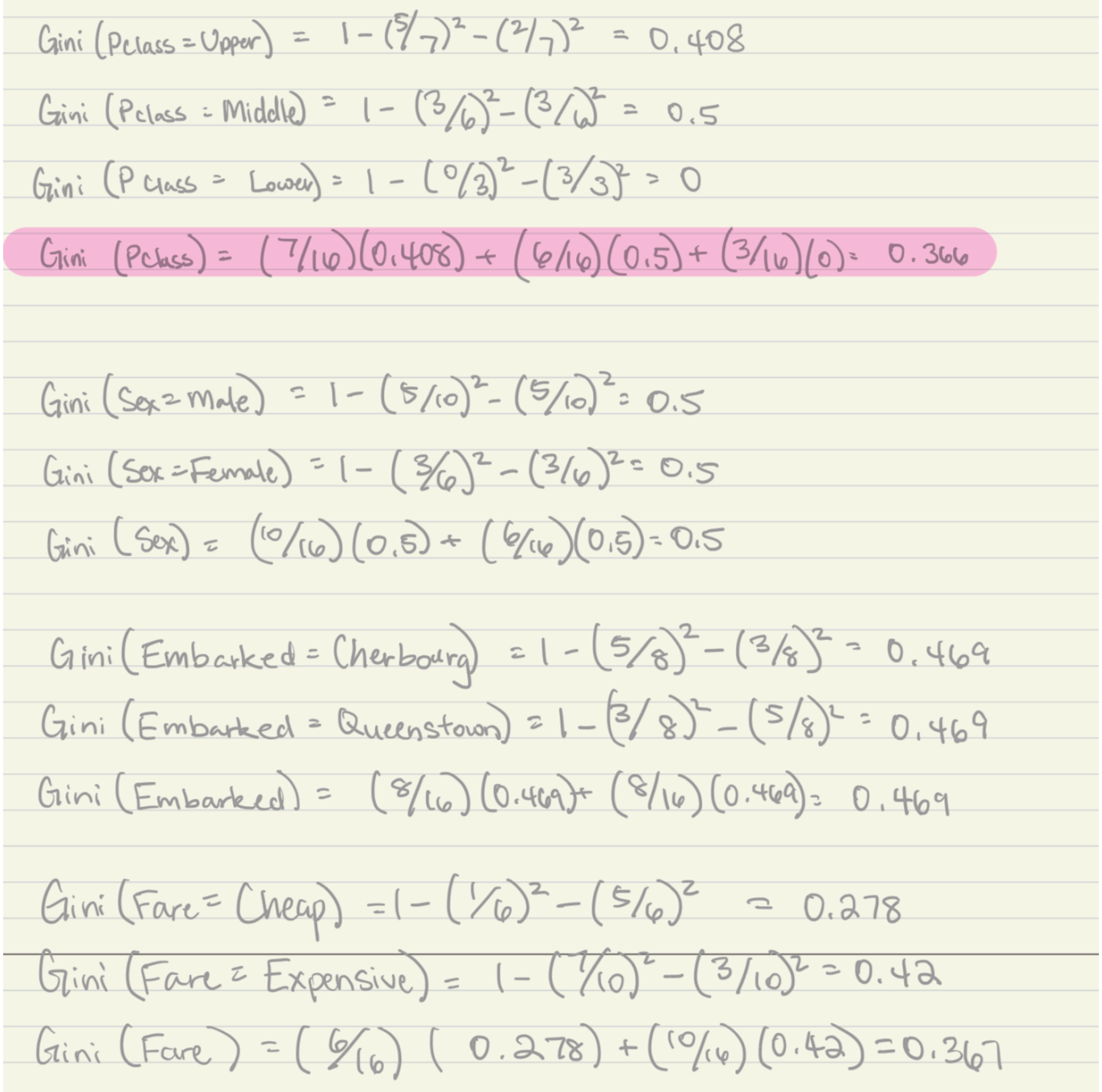
Bin 5 $(k=2)$: $[38,47)$: 39, 40, 42

3 d)

When analyzing continuous data such as grades by discretizing it into bins of equal width to represent the letter grade distribution, it is easier to analyze the data such as finding the skewed distribution of a population's grade or finding outliers. If you use equal frequency binning, it will be hard to see the skewed distribution and the outlier since each bin/interval has an equal number of values.

4 Decision Tree Construction (18 points)

4 a)



$$\text{Gini}(\text{Pclass} = \text{Upper}) = 1 - (5/7)^2 - (2/7)^2 = 0.408$$

$$\text{Gini}(\text{Pclass} = \text{Middle}) = 1 - (3/6)^2 - (3/6)^2 = 0.5$$

$$\text{Gini}(\text{P class} = \text{Lower}) = 1 - (0/3)^2 - (3/3)^2 = 0$$

$$\text{Gini}(\text{Pclass}) = (7/16)(0.408) + (6/16)(0.5) + (3/16)(0) = 0.366$$

$$\text{Gini}(\text{Sex} = \text{Male}) = 1 - (5/10)^2 - (5/10)^2 = 0.5$$

$$\text{Gini}(\text{Sex} = \text{Female}) = 1 - (3/6)^2 - (3/6)^2 = 0.5$$

$$\text{Gini}(\text{Sex}) = (10/16)(0.5) + (6/16)(0.5) = 0.5$$

$$\text{Gini}(\text{Embarked} = \text{Cherbourg}) = 1 - (5/8)^2 - (3/8)^2 = 0.469$$

$$\text{Gini}(\text{Embarked} = \text{Queenstown}) = 1 - (3/8)^2 - (5/8)^2 = 0.469$$

$$\text{Gini}(\text{Embarked}) = (8/16)(0.469) + (8/16)(0.469) = 0.469$$

$$\text{Gini}(\text{Fare} = \text{Cheap}) = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

$$\text{Gini}(\text{Fare} = \text{Expensive}) = 1 - (7/10)^2 - (3/10)^2 = 0.42$$

$$\text{Gini}(\text{Fare}) = (6/16)(0.278) + (10/16)(0.42) = 0.367$$

Gini Pclass = Upper

Gini (Sex = Male) = $1 - (3/5)^2 - (2/5)^2 = 0.48$

Gini (Sex = Female) = $1 - (2/2)^2 - (0/2)^2 = 0$

Gini (Sex) = $(\frac{5}{7})(0.48) + (\frac{2}{7})(0) = 0.343$

Gini (Embarked = Cherbourg) = $1 - (4/5)^2 - (1/5)^2 = 0.32$

Gini (Embarked = Queenstown) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini (Embarked) = $(\frac{5}{7})(0.32) + (2/7)(0.5) = 0.37$

Gini (Fare = Cheap) = $1 - (1/2)^2 - (1/2)^2 = 0.5$

Gini (Fare = Expensive) = $1 - (4/5)^2 - (1/5)^2 = 0.32$

Gini (Fare) = $(\frac{2}{7})(0.5) + (5/7)(0.32) = 0.37$

Gini Pclass = Middle

Gini (Sex = Male) = $1 - (2/3)^2 - (1/3)^2 = 0.44$

Gini (Sex = Female) = $1 - (1/3)^2 - (2/3)^2 = 0.44$

Gini (Sex) = $(3/6)(0.44) + (3/6)(0.44) = 0.44$

Gini (Embarked = Cherbourg) = $1 - (1/3)^2 - (2/3)^2 = 0.44$

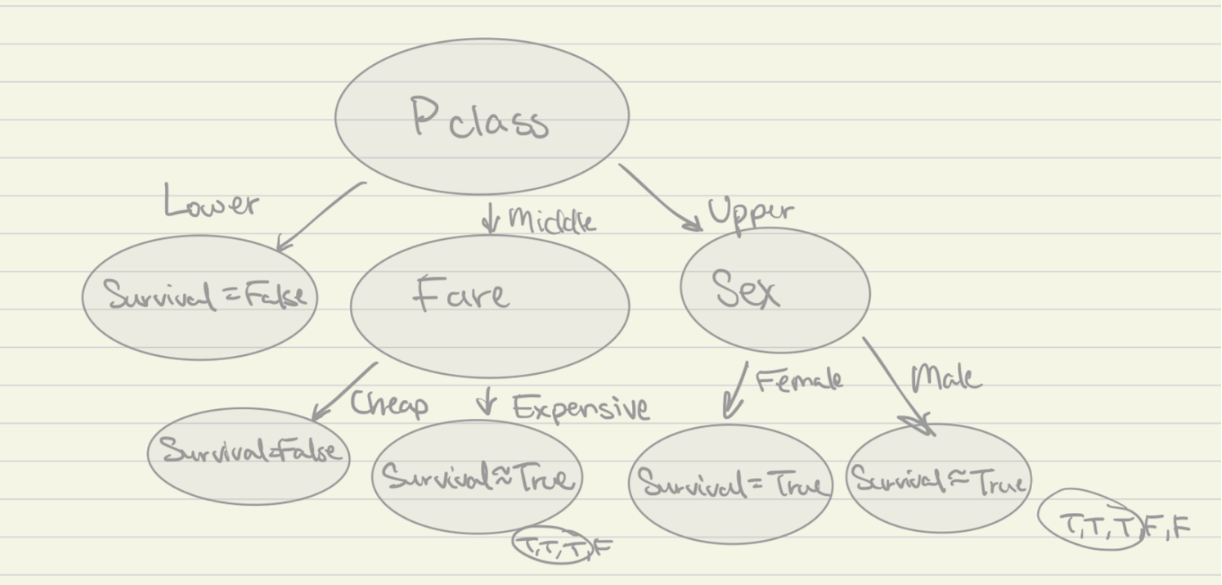
Gini (Embarked = Queenstown) = $1 - (2/3)^2 - (1/3)^2 = 0.44$

Gini (Embarked) = $(3/6)(0.44) + (3/6)(0.44) = 0.44$

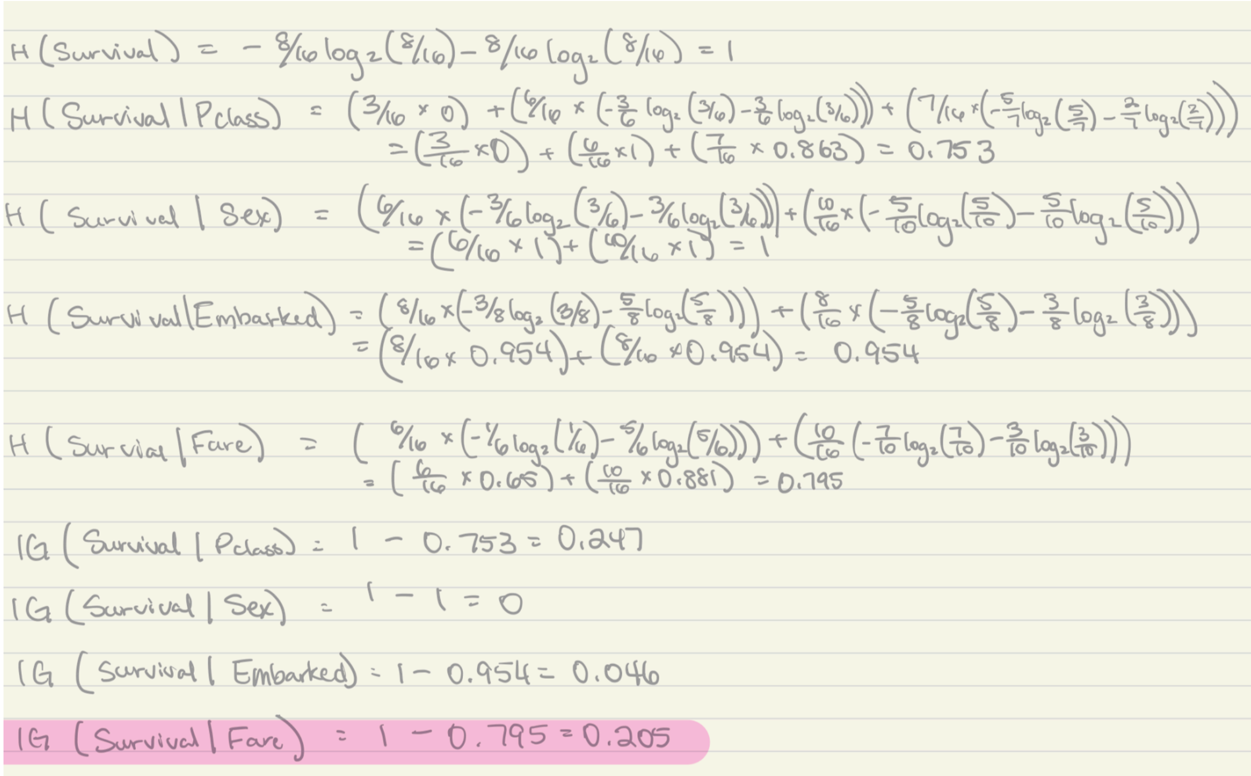
Gini (Fare = Cheap) = $1 - (0/2)^2 - (2/2)^2 = 0$

Gini (Fare = Expensive) = $1 - (3/4)^2 - (1/4)^2 = 0.375$

Gini (Fare) = $(2/6)(0) + (4/6)(0.375) = 0.25$



4 b)



$H(\text{Survival}) = -8/16 \log_2(8/16) - 8/16 \log_2(8/16) = 1$

$H(\text{Survival} \mid \text{Pclass}) = (3/16 \times 0) + (6/16 \times (-\frac{3}{6} \log_2(3/6) - \frac{3}{6} \log_2(3/6))) + (7/16 \times (-\frac{5}{7} \log_2(\frac{5}{7}) - \frac{2}{7} \log_2(\frac{2}{7})))$
$= (\frac{3}{16} \times 0) + (\frac{6}{16} \times 1) + (\frac{7}{16} \times 0.863) = 0.753$

$H(\text{Survival} \mid \text{Sex}) = (6/16 \times (-3/6 \log_2(3/6) - 3/6 \log_2(3/6))) + (\frac{10}{16} \times (-\frac{5}{10} \log_2(\frac{5}{10}) - \frac{5}{10} \log_2(\frac{5}{10})))$
$= (6/16 \times 1) + (10/16 \times 1) = 1$

$H(\text{Survival} \mid \text{Embarked}) = (8/16 \times (-3/8 \log_2(3/8) - \frac{5}{8} \log_2(\frac{5}{8}))) + (\frac{8}{16} \times (-\frac{5}{8} \log_2(\frac{5}{8}) - \frac{3}{8} \log_2(\frac{3}{8})))$
$= (8/16 \times 0.954) + (8/16 \times 0.954) = 0.954$

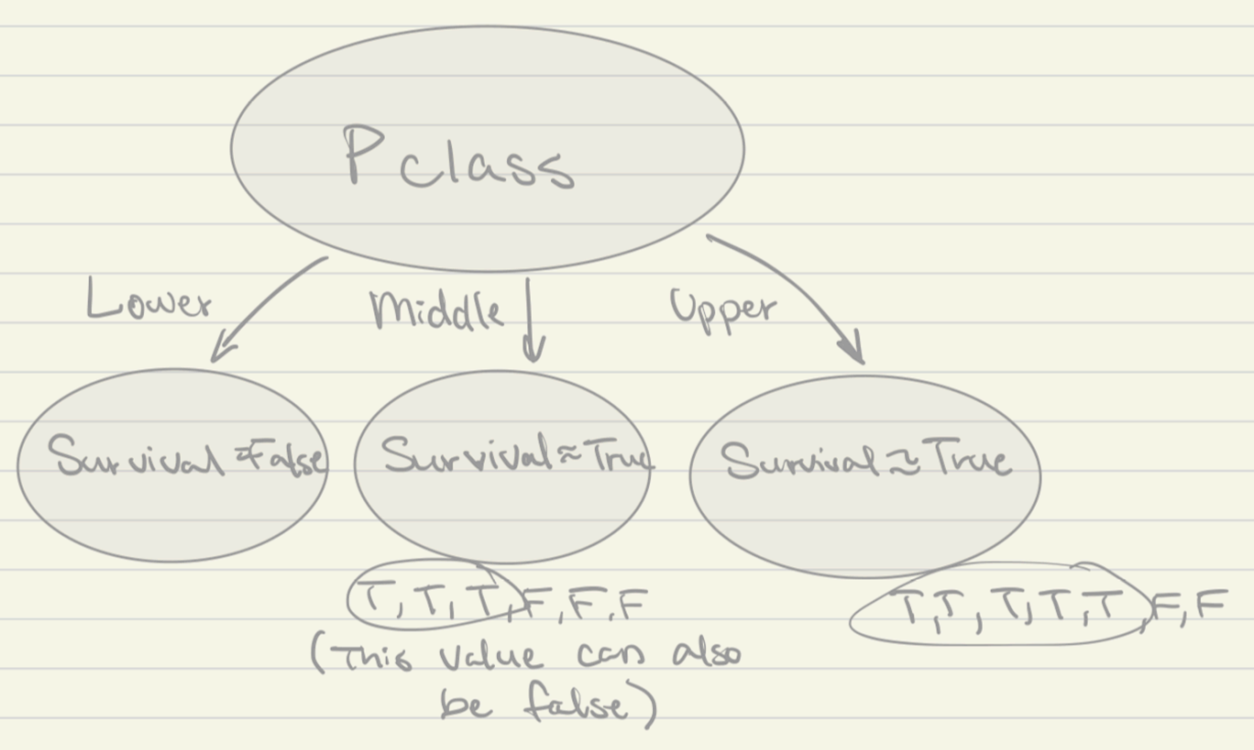
$H(\text{Survial} \mid \text{Fare}) = (6/16 \times (-1/6 \log_2(1/6) - 5/6 \log_2(5/6))) + (\frac{10}{16} (-\frac{7}{10} \log_2(\frac{7}{10}) - \frac{3}{10} \log_2(\frac{3}{10})))$
$= (\frac{6}{16} \times 0.65) + (\frac{10}{16} \times 0.881) = 0.795$

$IG(\text{Survival} \mid \text{Pclass}) = 1 - 0.753 = 0.247$

$IG(\text{Survival} \mid \text{Sex}) = 1 - 1 = 0$

$IG(\text{Survival} \mid \text{Embarked}) = 1 - 0.954 = 0.046$

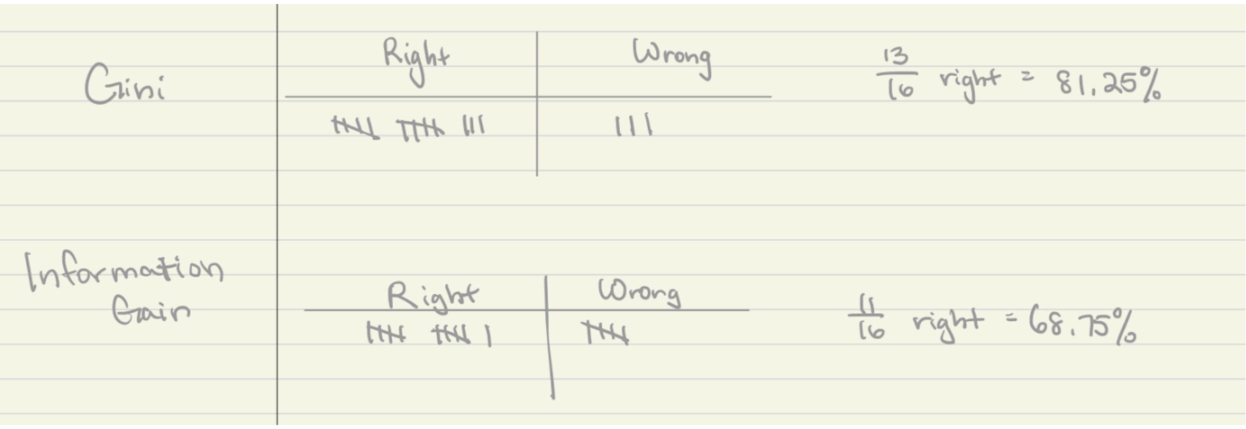
$IG(\text{Survival} \mid \text{Fare}) = 1 - 0.795 = 0.205$



4 c)

A passenger in the middle class who bought a cheap fare. In the Gini Decision Tree, this passenger does not survive. In the Information Gain Decision Tree, the passenger survives.

4 d)



	Right	Wrong	
Gini	𝍸 𝍸 𝍷𝍷𝍷	𝍷𝍷𝍷	$\frac{13}{16}$ right = 81.25%
Information Gain	𝍸 𝍸 𝍷	𝍸	$\frac{11}{16}$ right = 68.75%

The Gini decision tree performs better on the training set with an accuracy rate of 81.25% while the Information Gain decision tree only has an accuracy rate of 68.75%.

4 e)

We can predict the answer if the test dataset is similar to the training data set but not with complete certainty. Given the calculated accuracy of the two decision trees, it is predicted that the Gini decision tree will perform better on a test dataset. If the test dataset is not similar to the training dataset, it is unknown which decision tree will perform better.

5 Dimensionality Reduction (10 points)

5 a)

I think PCA is not useful because of how high the loading values are in the first dataset chart. PC_1, PC_2, and PC_3 all have loading values close to -1 and 1 which means that these components are largely influenced by one feature. This can possibly be due to the fact that the raw data has not been normalized so the variables with larger scales are disproportionately influencing the PCA values.

5 b)

It is reasonable to keep up to 3 principal components due to the fact that up to the third principal component, all their eigenvalues are more than 1. After the third component, the eigenvalues drop to less than one which is a good cutoff. The slope from the first to the second component is also really steep, but the slope from the second component to the fourth component is shallow. We keep the third component due to an eigenvalue of more than one, but the fourth component has a shallow slope along with an eigenvalue of less than one which seems to be a good choice to drop this component along with any components after it.

5 c)

A's value for PC2 will decrease. The loading value for PC2 Feat_4 is -0.5 while the loading value for PC2 Feat_5 is -0.22. Given that both Feat_4 and Feat_5 will be affected by 2, the larger loading value will prevail. -0.5 * 2 = -1 while -0.22 * -2 = 0.44 so the overall value for A will decrease.