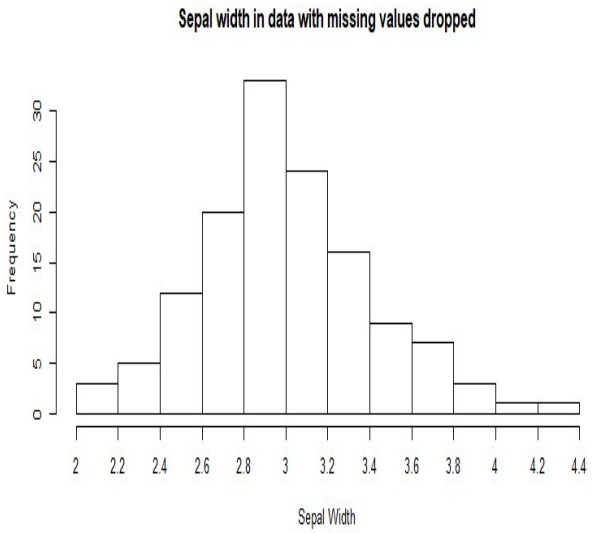
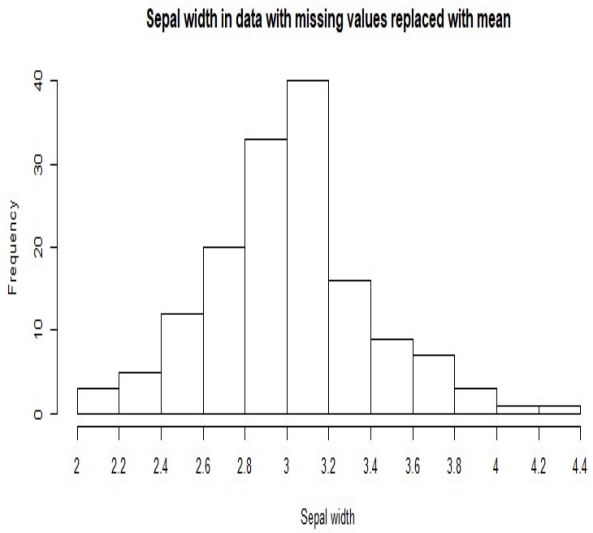
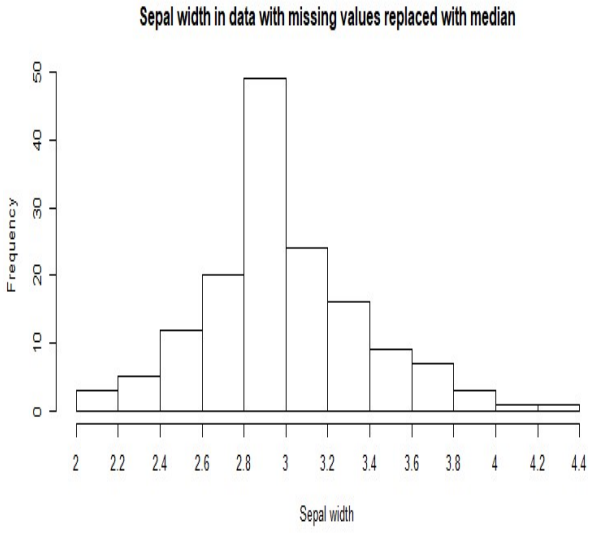
DATA MINING

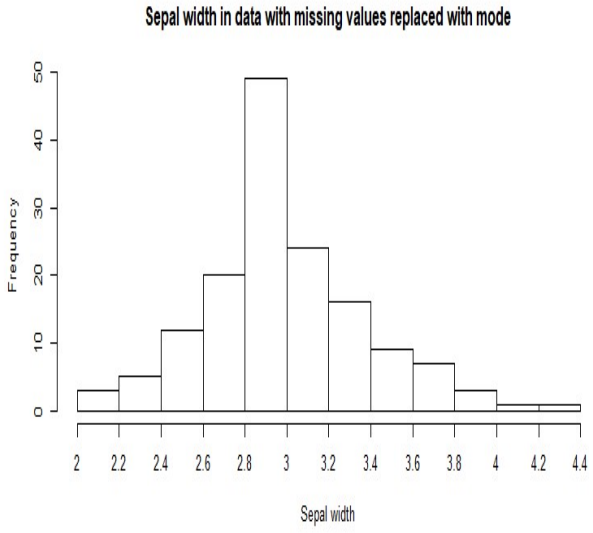
Deepak kumar(1793606)

Part B Question 4: Histograms









Comment on Histograms:

It can be clearly observed from the set of histograms that :

a)When the Sepal width rows with missing values is dropped and histogram is plotted then,its observed that values between range 2.8 to 3 have maximum frequency in the data i.e about 34,values between range 3 to 3.2 have second highest frequency i.e 23,and the values with the third highest frequency are

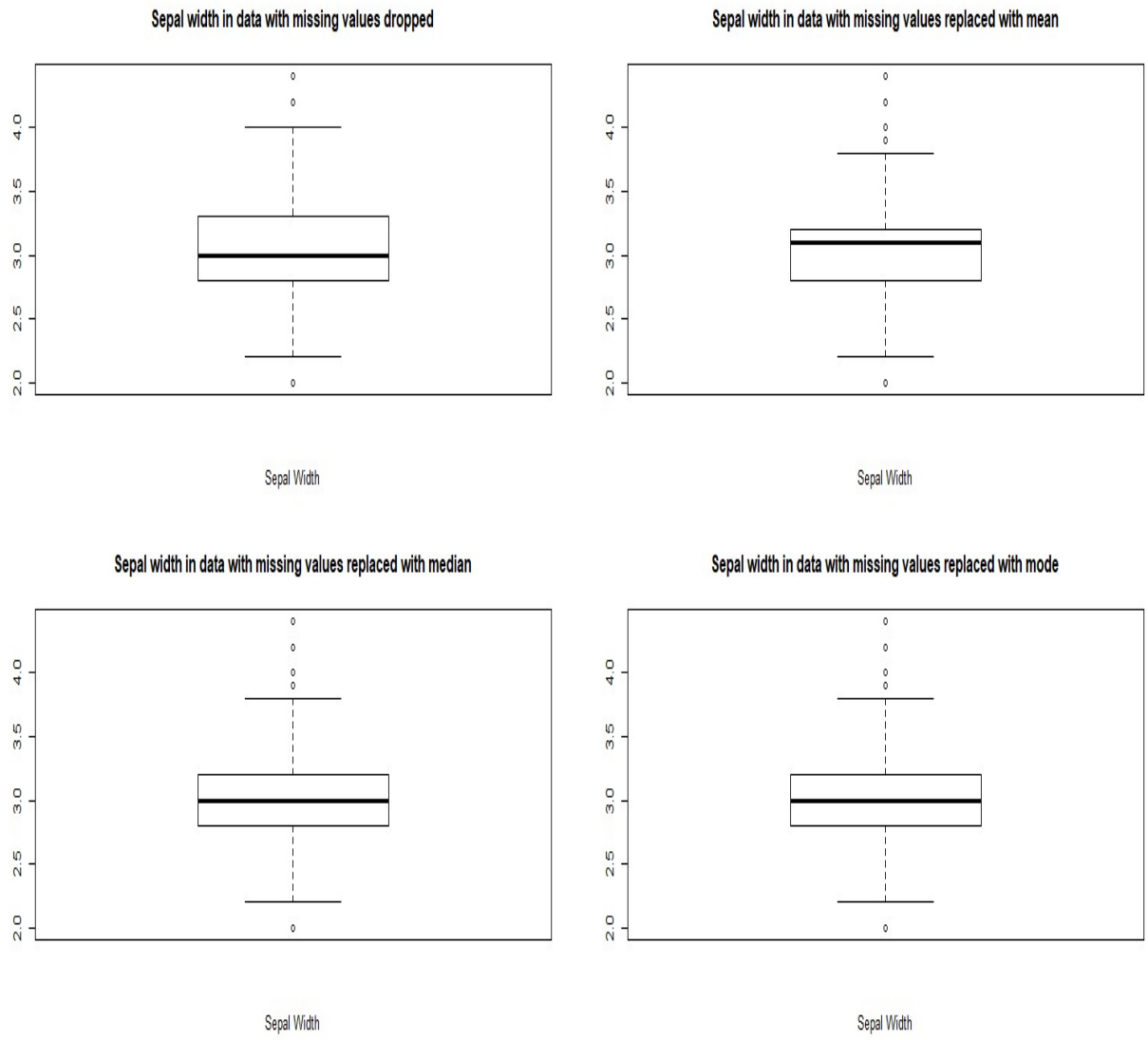
those between range 2.6 to 2.8.Values with lowest frequency are those with their value between 4 to 4.4 as their frequency is hardly 2.

b)When the Sepal width rows with missing values is replaced with mean and a histogram is plotted then,its observed that values between range 3 to 3.2 have maximum frequency in the data i.e about 40,values between range 2.8 to 3.0 have second highest frequency i.e 33,and the values with the third highest frequency(20) are those between range 2.6 to 2.8.Values with lowest frequency are those with their value between 4 to 4.4 as their frequency is hardly 2.

c)When the Sepal width rows with missing values is replaced with median and a histogram is plotted then,its observed that values between range 2.8 to 3.0 have maximum frequency in the data i.e about 48,values between range 3 to 3.2 have second highest frequency i.e around 23,and the values with the third highest frequency(20) are those between range 2.6 to 2.8.Values with lowest frequency are those with their value between 4 to 4.4 as their frequency is just 1.

d)When the Sepal width rows with missing values is replaced with mode and a histogram is plotted then,its observed that the plot is almost completely similar to that of histogram plotted with missing values is replaced with median.

Part B Question 4: Boxplot



Comment on boxplot:

It can be clearly observed from the set of boxplots that:

a) When the Sepal width rows with missing values is dropped and boxplot is plotted then, its observed that 1st quartile here is around 2.8,2nd quartile or median of the data is 3.0 and the 3rd quartile is around 3.3. The minimum value is around 2.2 and the maximum value is around 4.0.

b) When the Sepal width rows with missing values is replaced with mean and a boxplot is plotted then, its observed that 1st quartile here is around 2.8 i.e. same as previous boxplot,2nd quartile or median of the data is 3.1 and the 3rd quartile is around 3.2. The minimum value is around 2.2 and the maximum value is around 3.75.

c)When the Sepal width rows with missing values is replaced with median and a boxplot is plotted then, its observed that 1st quartile here is around 2.8,2nd quartile or median of the data is 3.0 and the 3rd quartile is around 3.2. The minimum value is around 2.2 and the maximum value is around 3.75.

d)When the Sepal width rows with missing values is replaced with mode and a boxplot is plotted then, its observed that the boxplot is almost completely like that of boxplot plotted with missing values is replaced with median.

Part C question 2.

Class	Sex	Age	Survived	Freq
1st	Male	Child	No	0
2nd	Male	Child	No	0
3rd	Male	Child	No	35
1st	Female	Child	No	0
2nd	Female	Child	No	0
3rd	Female	Child	No	17
1st	Male	Adult	No	118
2nd	Male	Adult	No	154
3rd	Male	Adult	No	387
1st	Female	Adult	No	4
2nd	Female	Adult	No	13
3rd	Female	Adult	No	89
1st	Male	Child	Yes	5
2nd	Male	Child	Yes	11
3rd	Male	Child	Yes	13
1st	Female	Child	Yes	1
2nd	Female	Child	Yes	13
3rd	Female	Child	Yes	14
1st	Male	Adult	Yes	57
2nd	Male	Adult	Yes	14
3rd	Male	Adult	Yes	75
1st	Female	Adult	Yes	140
2nd	Female	Adult	Yes	80
3rd	Female	Adult	Yes	76

For the above given dataset we have survived as the class attribute.

Here, as there are 12 yes for survived and 12 no as well so there are 12 positive(P) cases and 12 negative(N) cases also.

Entropy for Class:

$$= \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \frac{N}{P+N}\log_2\left(\frac{N}{P+N}\right)$$

Information gain for each attribute:

$$I(P_i, N_i) = \frac{-P}{P+N}\log_2\left(\frac{P}{P+N}\right) - \left(\frac{N}{P+N}\right)\log_2\left(\frac{N}{P+N}\right)$$

Entropy (Attribute):

$$E = \frac{\Sigma P_i + N_i}{P+N} \left(I(P_i, N_i) \right)$$

Gain:

E (Class) − E (Attribute)

Now, we will find the entropy for the class ie survived

$$E(\text{Survived}) = -\frac{12}{12+12}\log_2\left(\frac{12}{12+12}\right) - \frac{12}{12+12}\log_2\left(\frac{12}{12+12}\right)$$

$$= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)$$

$$= +$$

E(Survived) given that from 1316 instances of data provided in titanic survival data 817 didn't survive and 499 survived.

$$E(\text{Survived}) = -\frac{203}{325}\log_2\left(\frac{203}{325}\right) - \frac{122}{325}\log_2\left(\frac{122}{325}\right)$$

$$= 0.9574$$

Now, Calculating the information gain for Class.

	P_i	N_i	$I(P_i, N_i)$
1st	203	122	0.96
2nd	118	167	0.97
3rd	178	528	0.81

For 1st Class $= -\frac{203}{325}\log_2\left(\frac{203}{325}\right) - \frac{122}{325}\log_2\left(\frac{122}{325}\right)$

$$= (-0.625 X - 0.69)(-0.375 X - 1.42)$$

$$= 0.43 + 0.53$$

$$= 0.96 \text{ (0.9547 accurate)}$$

2nd Class: $-\frac{118}{285}\log_2\left(\frac{118}{285}\right) - \frac{167}{285}\log_2\left(\frac{167}{285}\right)$

$$= -0.41 X - 1.28 - 0.58 X - 0.78$$

$$= 0.9785$$

Information gain for 3rd class:

$$-\frac{178}{706}\log_2\left(\frac{178}{706}\right) - \frac{528}{706}\log_2\left(\frac{528}{706}\right)$$

$$= 0.8146$$

Now, if we calculate the entropy for class:

$$E(\text{class}) = \frac{203+122}{1316} \times 0.96 + \frac{167}{1316} \times 0.97 + \frac{178+528}{1316} \times 0.87$$

$$= 0.8847$$

Information gain for Sex

	Pi	Ni	I(Pi, Ni)
male	175	694	0.72
female	324	123	0.85

Information gain for male:

$$\frac{175}{324} - \frac{175}{(175+694)}\log_2\left(\frac{175}{(175+694)}\right) - \frac{324}{(324+123)}\log_2\left(\frac{324}{324+123}\right)$$

$$= 0.7246$$

Information gain for female:

$$= -\frac{324}{324+123} \times \log_2\left(\frac{324}{324+123}\right) - \frac{123}{324+123}\log_2\left(\frac{123}{447}\right)$$

$$= -0.72\log_2(0.72) - 0.27\log_2(0.28)$$

$$= 0.8487$$

Entropy for Sex is given by:

$$\frac{175+694}{1316} \times 0.72 + \frac{324+123}{1316} \times 0.85$$

$$= 0.7668$$

Considering the Age attribute

	P_i	N_i	$I(P_i, N_i)$
Child	57	52	0.9984
Adult	442	765	0.9477

Information gain for Child:

$$E(\text{Child}) = -\frac{57}{109}\log_2\left(\frac{57}{109}\right) - \frac{52}{109}\log_2\left(\frac{52}{109}\right)$$

$$= 0.9984$$

$$E(\text{Adult}) = -\frac{765}{1207}\log_2\left(\frac{765}{1207}\right) - \frac{442}{1207}\log_2\left(\frac{442}{1207}\right)$$

$$= 0.9477$$

Entropy of Age is given as:

$$\left(\frac{57+52}{1316}\right) \times (0.9984) + \frac{442+765}{1316} \times (0.95)$$

$$= 0.082 + 0.87$$

$$= 0.9519$$

Final Information gain for Class:

$$= 0.9574 - 0.8847$$

$$= 0.0727$$

Final Information gain for Sex:

$$= 0.9574 - 0.7668$$

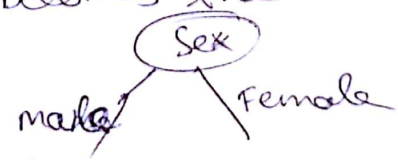
$$= 0.1906$$

Final Information gain for Age:

$$= 0.9574 - 0.9519$$

$$= 0.0055$$

As the information gain is maximum for Sex so, it becomes the root node



Now we need to categorize the tree as per the male and female members:

Entropy (Female) as calculated before is 0.8487

Now, we calculate the entropy of class for female members:

	P_i	N_i	$I(P_i, N_i)$
1st	141	4	0.1821
2nd	93	13	0.5364
3rd	90	106	0.9951

Information gain

Entropy of 1st class:

$$-\frac{141}{145}\log_2\left(\frac{141}{145}\right) - \frac{4}{145}\log_2\left(\frac{4}{145}\right)$$

$$= 0.1821$$

Information gain for 2nd class:

$$-\frac{93}{106}\log_2\left(\frac{93}{106}\right) - \frac{13}{106}\log_2\left(\frac{13}{106}\right)$$

$$= 0.5364$$

Information gain for 3rd class:

$$-\frac{90}{196}\log_2\left(\frac{90}{196}\right) - \frac{106}{196}\log_2\left(\frac{106}{196}\right)$$

$$= 0.9951$$

Entropy of the attribute Class is given as: for female sex

$$E(\text{Class}) = \left(\frac{145}{447}\right)\times(0.18) + \left(\frac{106}{447}\right)\times(0.53) + \left(\frac{196}{447}\right)\times(0.49)$$

$$= 0.6227$$

Now, we consider the attribute age:

Information gain for adult is calculated as

$$-\frac{296}{402}\log_2\left(\frac{296}{402}\right) - \frac{106}{402}\log_2\left(\frac{106}{402}\right)$$

E(Adult) = 0.8322

Information gain for child is calculated as

$$E(\text{Child}) = -\frac{17}{45}\log_2\left(\frac{17}{45}\right) - \frac{28}{45}\log_2\left(\frac{28}{45}\right)$$

$$= 0.9564$$

Entropy for the attribute age given sex is female is given as:

$$E(\text{Age}) = \frac{402}{447} \times 0.83 + \frac{45}{447} \times 0.9564$$

$$= 0.8447$$

Final gain of attribute class given members are female:

$$0.8487 - 0.6227$$

$$= 0.2259$$

Final gain of attribute age given members are female:

$$0.8487 - 0.8444$$

$$= 0.0039$$

Now we consider calculation of gain when the sex is male.

Information gain calculated for male in part I was 0.7247.

Information gain for class 1st when sex is male

$$= -\frac{62}{180}\log_2\left(\frac{62}{180}\right) - \frac{118}{180}\log_2\left(\frac{118}{180}\right)$$

$$= 0.9290$$

Information gain for class 2nd when sex is female

$$= \frac{-25}{179}\log_2\left(\frac{25}{179}\right) - \frac{154}{179}\log_2\left(\frac{154}{179}\right)$$

$$= 0.5833$$

Information gain for class 3rd when sex is male:

$$= \frac{-88}{510}\log_2\left(\frac{88}{510}\right) - \frac{422}{510}\log_2\left(\frac{422}{510}\right)$$

$$= 0.6635$$

Entropy of the class when sex is male is given as:

$$= \frac{180}{869} \times (0.92) + \frac{179}{869} \times 0.58 + \frac{510}{869} \times 0.66$$

$$= 0.7019$$

Now we consider attribute age:

Information gain for Adult is given as (when sex is male)

~~E(Adult) = $\frac{659}{805}\log_2$~~

$$= -\frac{146}{805}\log_2\left(\frac{146}{805}\right) - \frac{659}{805}\log_2\left(\frac{659}{805}\right)$$

$$= 0.6830$$

Information gain for Children with male sex:

$$\frac{-35}{64}\log_2\left(\frac{35}{64}\right) - \frac{29}{64}\log_2\left(\frac{29}{64}\right)$$

$$= 0.9936$$

Entropy of attribute age when the sex is male is

$$= \frac{805}{869} \times 0.68 + \frac{64}{869} \times 0.99$$

$$= 0.7059$$

Total Information gain of Class when sex is male is

$$0.7947 - 0.7019$$

$$= 0.0228$$

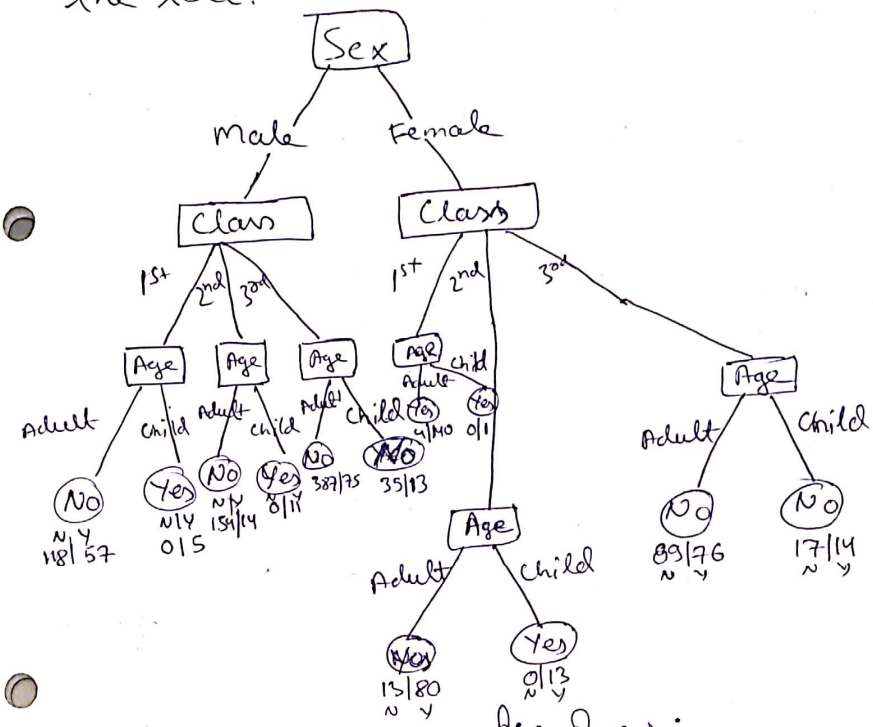
Total Information gain of Age when the sex is male is:

= 0.7247 - 0.7059

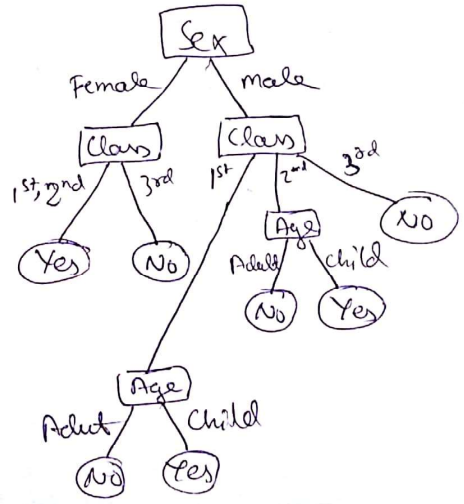
= 0.0187

Information gain of Age for attributes

So on the basis of gains found, when the Sex is male and when sex is female, we split the tree:



Above tree can be generalised as:



Rules obtained from the tree:

a) IF Sex = Female $\wedge$ (Class = 1st $\vee$ Class = 2nd)
THEN Survived = Yes

b) IF Sex = Male $\wedge$ (Class = 1st $\vee$ Class = 2nd)
$\wedge$ Age = Adult THEN Survived = No

c) IF Sex = Male $\wedge$ (Class = 1st $\vee$ Class = 2nd)
$\wedge$ Age = Child THEN Survived = Yes

d) IF (Sex = Male $\vee$ Sex = Female) $\wedge$
Class = 3rd THEN Survived = No

PART D Question 1:

Question : 4(a)

4a)

If we draw tables for the cluster provided for Hair Eye color data

When Sex = Male

Hair \ Eye	Brown	Blue	Hazel	Green
Black	32	11	10	3
Brown	53	50	25	15
Red	10	10	7	7
Blond	3	30	5	8

When Sex = Female

Hair \ Eye	Brown	Blue	Hazel	Green
Black	36	9	5	2
Brown	66	34	29	14
Red	16	7	7	7
Blond	4	64	5	8

Drawing an individual table for Hair with Male and Female columns we have (We add males and females with different coloured hair.)

Hair	Male	Female
Black	56	52
Brown	143	143
Red	34	37
Blond	46	81
Total:	279	313

Similarly we make another table for Eye with Male and female columns by adding the males and females with ~~hair~~ eyes of different colours:

Eye	Male	Female
Brown	98	122
Blue	101	114
Hazel	47	46
Green	33	31
Total:	279	313

Now we will use the secondary tables that we have created to do the probability computations

Prior probability:

Prior Probability of Males =

$$= \frac{\text{Total number of males}}{(\text{Total number of males} + \text{Total Number of females})}$$

$$= \frac{279}{279+313} = 0.4712$$

Prior probability of Females =

$$\frac{\text{Total number of females}}{\text{Total males} + \text{Total females}}$$

$$= \frac{313}{279+313} = 0.5287$$

Now we calculate the liklihood function for hair:

~~P(Male|Black) = 56/279~~

$P(\underset{\text{Hair}}{\text{Black}} | \text{Male}) = \frac{56}{279} = 0.20$

$P(\underset{\text{Hair}}{\text{Black}} | \text{Female}) = \frac{52}{313} = 0.17$

$P(\underset{\text{Hair}}{\text{Brown}} | \text{Male}) = \frac{143}{279} = 0.51$

$P(\underset{\text{Hair}}{\text{Brown}} | \text{Female}) = \frac{143}{313} = 0.46$

$P(\underset{\text{Hair}}{\text{Red}} | \text{Male}) = \frac{34}{279} = 0.12$

$P(\underset{\text{Hair}}{\text{Red}} | \text{Female}) = \frac{37}{313} = 0.12$

$P(\underset{\text{Hair}}{\text{Blond}} | \text{Male}) = \frac{46}{279} = 0.16$

$P(\underset{\text{Hair}}{\text{Blond}} | \text{Female}) = \frac{81}{313} = 0.26$

Further, we calculate the likelihood function for Eye color:

$P(\underset{\text{Eye}}{\text{Brown}} \mid \text{Male}) = \frac{98}{279} = 0.35$

$P(\underset{\text{Eye}}{\text{Brown}} \mid \text{Female}) = \frac{122}{313} = 0.39$

$P(\underset{\text{Eye}}{\text{Blue}} \mid \text{Male}) = \frac{101}{279} = 0.36$

$P(\underset{\text{Eye}}{\text{Blue}} \mid \text{Female}) = \frac{114}{313} = 0.36$

$P(\underset{\text{Eye}}{\text{Hazel}} \mid \text{Male}) = \frac{47}{279} = 0.17$

$P(\underset{\text{Eye}}{\text{Hazel}} \mid \text{Female}) = \frac{46}{313} = 0.15$

$P(\underset{\text{Eye}}{\text{Green}} \mid \text{Male}) = \frac{33}{279} = 0.12$

$P(\underset{\text{Eye}}{\text{Green}} \mid \text{Female}) = \frac{31}{313} = 0.10$

PART D Question 3:

Part D – Question 3.

If the number of attribute instances is low or zero, the probabilities estimated from frequency of the attribute will yeild bad results, ie about 0 probability.

Other probabilities in the model will not even matter if one probability is 0, as the end product obtained after multiplication of probabilities will be 0.

Usually laplacian correction is applied to fix this kind of issue. Since in Naive bayes products of probabilities of the features is evaluated during the training of the model and we clearly don't want it to evaluate to zero. So, to get rid of this, we have to assign some non-zero probabilities to words (features) which do not occur in the particular sample. This is what Laplace smoothing does.

3) If one of the features like if there were no red haired men in the dataset, then the probability for that feature would be 0.0

i.e $P(\text{Red Hair} \mid \text{Male}) = \frac{0}{279} = 0$

Laplacian correction is a solution for this kind of issue.

Now if $P(\text{Red Hair} \mid \text{Male}) = 0$

then in the Naive bayes function we will check the attribute for which the probability is 0, and then loop through the attribute to add 1 to every instance. The outcome of this discussion is discussed on next page

Let us assume,
If the probability initially was

$$P(\text{Red hair} \mid \text{Male}) = \frac{0}{279} = 0$$

$$P(\text{Black hair} \mid \text{Male}) = \frac{56}{279} = 0.20$$

$$P(\text{Brown hair} \mid \text{Male}) = \frac{143}{279} = 0.51$$

$$P(\text{Blond hair} \mid \text{Male}) = \frac{46}{279} = 0.16$$

The after the laplacian correction the probability should look like:

$$P(\text{Red hair} \mid \text{Male}) = \frac{1}{283}$$

$$P(\text{Black hair} \mid \text{Male}) = \frac{57}{283}$$

$$P(\text{Brown hair} \mid \text{Male}) = \frac{144}{283}$$

$$P(\text{Blond hair} \mid \text{Male}) = \frac{47}{283}$$