University of Victoria

CSC - 503

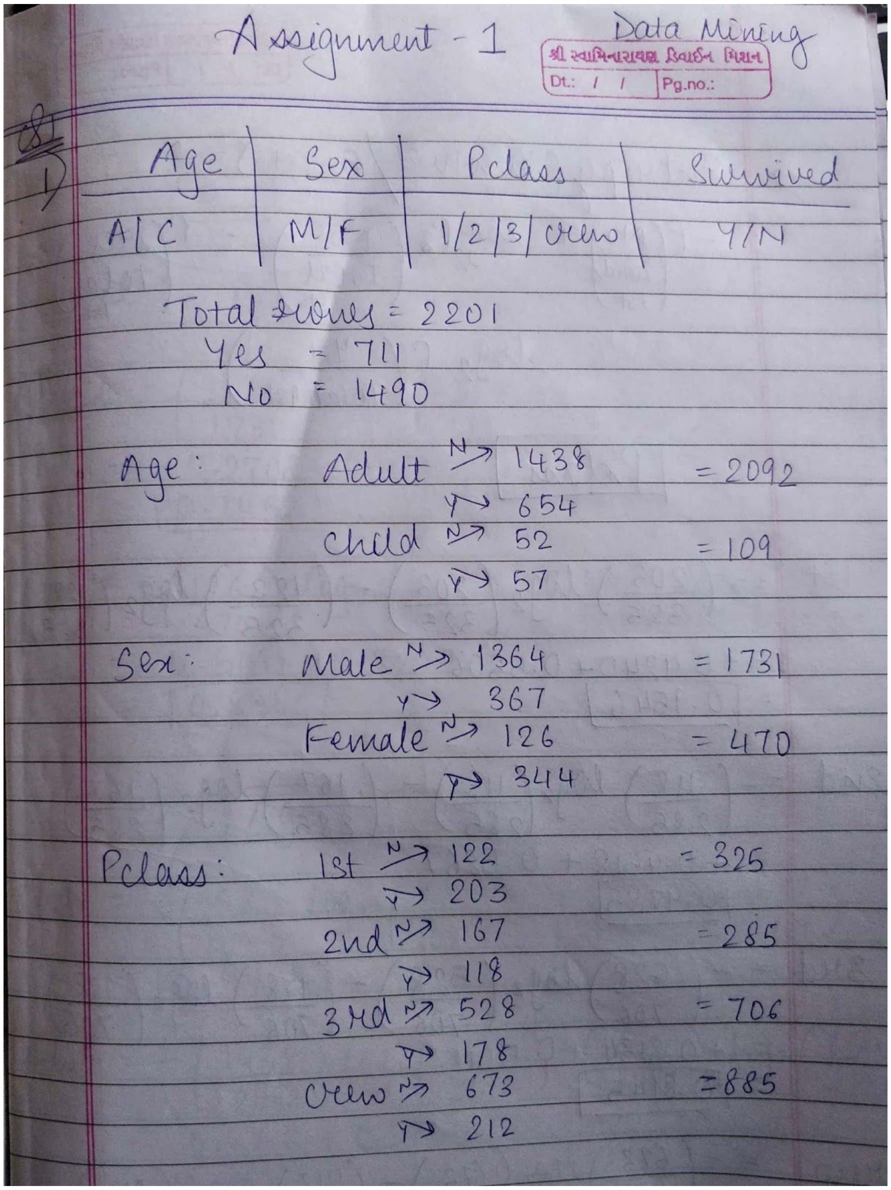
Report (Assignment - 1)

Submitted by :

Sannath Reddy Vemula - V00949217
Shaima Patel - V00949940

Submitted To : Alex Thomo
thomo@cs.uvic.ca

1) (4 points) Construct the root and the first level of a decision tree for the titanic dataset. Use the ID3 algorithm. Show the details of your construction (entropies calculated for each step). You can use a spreadsheet or SQL database to compute the counts.
Then, check your solution with Weka and submit a text file of your classifier output window.



Assignment - 1

Data Mining

Q1)

Age	Sex	Pclass	Survived
A/C	M/F	1/2/3/ crew	Y/N

Total rows = 2201
Yes = 711
No = 1490

Age: Adult $\xrightarrow{N}$ 1438 = 2092
$\xrightarrow{Y}$ 654
Child $\xrightarrow{N}$ 52 = 109
$\xrightarrow{Y}$ 57

Sex: Male $\xrightarrow{N}$ 1364 = 1731
$\xrightarrow{Y}$ 367
Female $\xrightarrow{N}$ 126 = 470
$\xrightarrow{Y}$ 344

Pclass: 1st $\xrightarrow{N}$ 122 = 325
$\xrightarrow{Y}$ 203
2nd $\xrightarrow{N}$ 167 = 285
$\xrightarrow{Y}$ 118
3rd $\xrightarrow{N}$ 528 = 706
$\xrightarrow{Y}$ 178
crew $\xrightarrow{N}$ 673 = 885
$\xrightarrow{Y}$ 212

→ Entropy (Survived / P = 1st) =

$$-P\left(\frac{No}{Total\ 1st}\right) * \log_2 P\left(\frac{No}{Total\ 1st}\right) - P\left(\frac{Yes}{Total\ 1st}\right) * \log_2 P\left(\frac{Yes}{Total\ 1st}\right)$$

$\boxed{Pclass}$

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

$= +0.4240 + 0.5306$

$= \boxed{0.9546}$

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

$= +0.4518 + 0.5267$

$= \boxed{0.9785}$

3rd $= -\left(\frac{528}{706}\right)\log_2\left(\frac{528}{706}\right) - \left(\frac{178}{706}\right)\log_2\left(\frac{178}{706}\right)$

$= +0.3134 + 0.5011$

$= \boxed{0.8145}$

crew $= -\left(\frac{673}{885}\right)\log_2\left(\frac{673}{885}\right) - \left(\frac{212}{885}\right)\log_2\left(\frac{212}{885}\right)$

$= + 0.3004 + 0.4938$

$= \boxed{0.7942}$

selected
↓

$\boxed{\text{Sex}}$

Male $= -\left(\frac{1364}{1731}\right)\log_2\left(\frac{1364}{1731}\right) - \left(\frac{367}{1731}\right)\log_2\left(\frac{367}{1731}\right)$

$= + 0.2708 + 0.4744$

$= \boxed{0.7452}$

Female $= -\left(\frac{126}{470}\right)\log_2\left(\frac{126}{470}\right) - \left(\frac{344}{470}\right)\log_2\left(\frac{344}{470}\right)$

$= 0.5091 + 0.3295$

$= \boxed{0.8386}$

$\boxed{\text{Age}}$

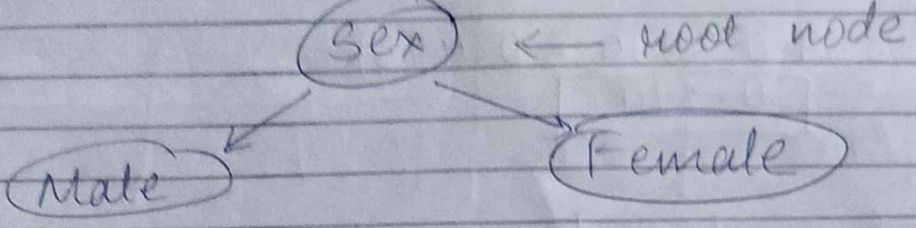
Child $= -\left(\frac{52}{109}\right)\log_2\left(\frac{52}{109}\right) - \left(\frac{57}{109}\right)\log_2\left(\frac{57}{109}\right)$

$= \boxed{0.9983}$

Adult $= -\left(\frac{1438}{2092}\right)\log_2\left(\frac{1438}{2092}\right) - \left(\frac{654}{2092}\right)\log_2\left(\frac{654}{2092}\right)$

$= 0.3713 + 0.5244$

$= \boxed{0.8967}$



Male		Female	
Pclass			
1st $\xrightarrow{y}$ 62	= 180	1st $\xrightarrow{y}$ 141	= 145
$\xrightarrow{N}$ 118		$\xrightarrow{N}$ 4	
2nd $\xrightarrow{y}$ 25	= 179	2nd $\xrightarrow{y}$ 93	= 106
$\xrightarrow{N}$ 154		$\xrightarrow{N}$ 13	
3rd $\xrightarrow{y}$ 88	= 510	3rd $\xrightarrow{y}$ 90	= 196
$\xrightarrow{N}$ 422		$\xrightarrow{N}$ 106	
crew $\xrightarrow{y}$ 192	= 862	crew $\xrightarrow{y}$ 20	= 23
$\xrightarrow{N}$ 670		$\xrightarrow{N}$ 3	
Age			
Male $\xrightarrow{y}$ 338	= 1667	Male $\xrightarrow{y}$ 316	= 425
$\xrightarrow{N}$ 1329		$\xrightarrow{N}$ 109	
Female $\xrightarrow{y}$ 29	= 64	Female $\xrightarrow{y}$ 28	= 45
$\xrightarrow{N}$ 35		$\xrightarrow{N}$ 17	

Male

Pclass

1St $= -\left(\frac{62}{180}\right)\log_2\left(\frac{62}{180}\right) - \left(\frac{118}{180}\right)\log_2\left(\frac{118}{180}\right)$

$= +0.5296 + 0.3993$

Selected $= \boxed{0.9289}$

↓

2nd $= -\left(\frac{25}{179}\right)\log_2\left(\frac{25}{179}\right) - \left(\frac{154}{179}\right)\log_2\left(\frac{154}{179}\right)$

$= 0.3966 + 0.1867$

$= \boxed{0.5833}$

3rd $= -\left(\frac{88}{510}\right)\log_2\left(\frac{88}{510}\right) - \left(\frac{422}{510}\right)\log_2\left(\frac{422}{510}\right)$

$= 0.4373 + 0.2261$

$= \boxed{0.6634}$

Crew $= -\left(\frac{192}{862}\right)\log_2\left(\frac{192}{862}\right) - \left(\frac{670}{862}\right)\log_2\left(\frac{670}{862}\right)$

$= 0.4825 + 0.2825$

$= \boxed{0.7650}$

Age

Adult $= -\left(\frac{338}{1667}\right)\log_2\left(\frac{338}{1667}\right) - \left(\frac{1329}{1667}\right)\log_2\left(\frac{1329}{1667}\right)$

$= 0.4667 + 0.2606$

$= \boxed{0.7273}$

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

$= 0.5174 + 0.4761$

$= \boxed{0.9935}$

Female

Pclass

selected ↓

1st $= -\left(\frac{141}{145}\right) \log_2 \left(\frac{141}{145}\right) - \left(\frac{4}{145}\right) \log_2 \left(\frac{4}{145}\right)$

$= 0.039 + 0.142$

$= \boxed{0.181}$

2nd $= -\left(\frac{93}{106}\right) \log_2 \left(\frac{93}{106}\right) - \left(\frac{13}{106}\right) \log_2 \left(\frac{13}{106}\right)$

$= 0.1656 + 0.3712$

$= \boxed{0.5368}$

3rd $= -\left(\frac{90}{196}\right) \log_2 \left(\frac{90}{196}\right) - \left(\frac{106}{196}\right) \log_2 \left(\frac{106}{196}\right)$

$= 0.5155 + 0.4795$

$= \boxed{0.995}$

crew $= -\left(\frac{20}{33}\right) \log_2 \left(\frac{20}{33}\right) - \left(\frac{3}{33}\right) \log_2 \left(\frac{3}{33}\right)$

$= 0.1753 + 0.3832$

$= \boxed{0.5585}$

~~Adult~~ Age

$$Adult = -\left(\frac{316}{425}\right)\log_2\left(\frac{316}{425}\right) - \left(\frac{109}{425}\right)\log_2\left(\frac{109}{425}\right)$$

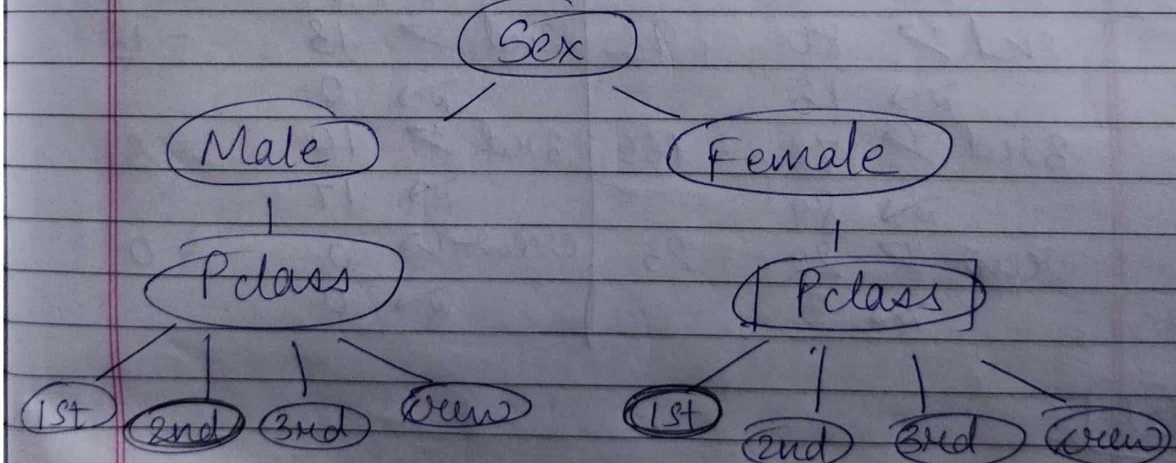
$$= +0.3178 + 0.5034$$

$$= \boxed{0.8212}$$

$$child = -\left(\frac{28}{45}\right)\log_2\left(\frac{28}{45}\right) - \left(\frac{17}{45}\right)\log_2\left(\frac{17}{45}\right)$$

$$= 0.4259 + 0.5305$$

$$= \boxed{0.9569}$$



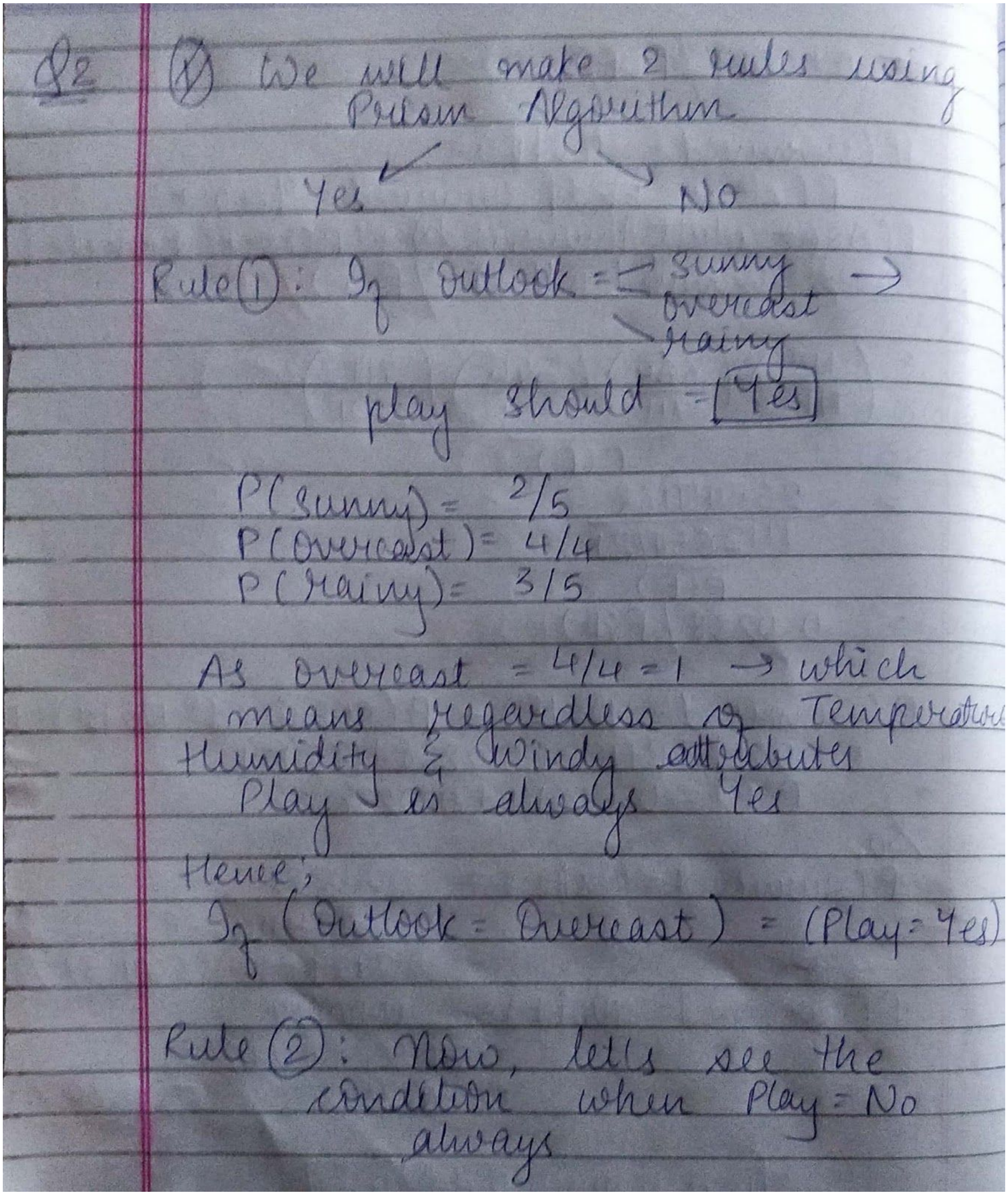
Male

→ Age

Adult

child

2) (4 points) Construct two rules using PRISM for the weather dataset. Show the details of your construction. Then, check your solution with Weka and submit a text file of your classifier output window.



Q2 (i) We will make 2 rules using Prism Algorithm

Yes ↙ ↘ No

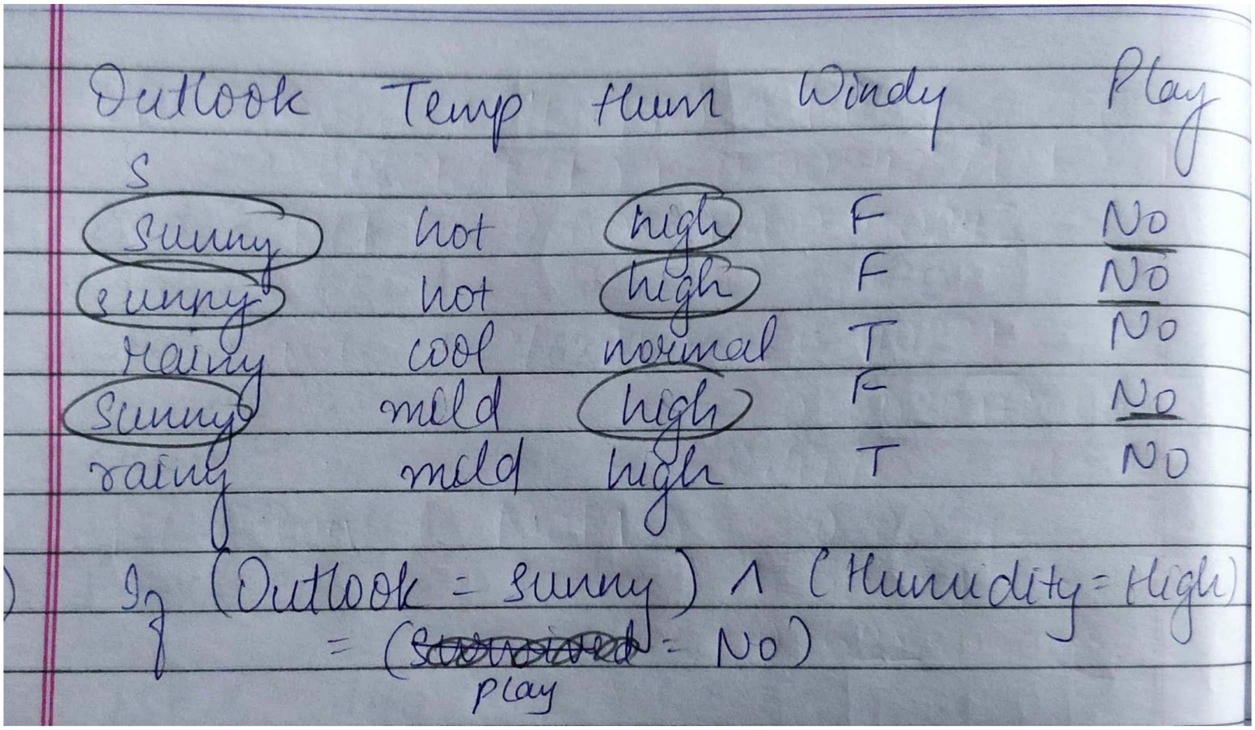
Rule (1): If Outlook = Sunny / Overcast / Rainy → play should = $\boxed{\text{Yes}}$

P(Sunny) = 2/5
P(Overcast) = 4/4
P(Rainy) = 3/5

As Overcast = 4/4 = 1 → which means regardless of Temperature Humidity & Windy attributes Play is always Yes

Hence;
If (Outlook = Overcast) = (Play = Yes)

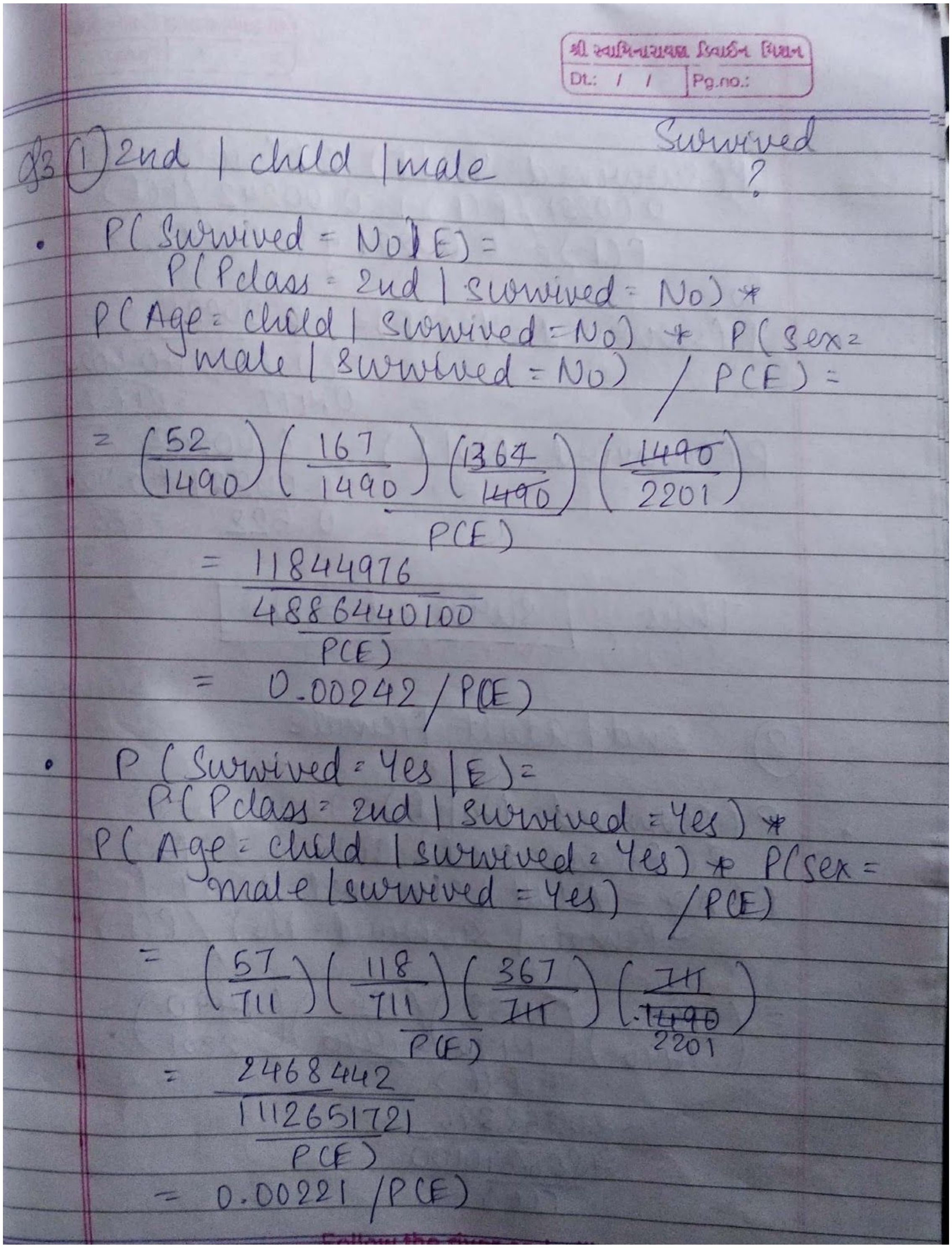
Rule (2): Now, lets see the condition when Play = No always



Outlook	Temp	Hum	Windy	Play
S				
Sunny	hot	high	F	No
sunny	hot	high	F	No
Rainy	cool	normal	T	No
Sunny	mild	high	F	No
rainy	mild	high	T	No

) If (Outlook = sunny) ∧ (Humidity = High)
= (~~survived~~ play = No)

3)(4 points) Classify using Naïve Bayes method on the titanic dataset the data items:
2nd child male ?
2nd adult female ?
Then, check your solution with Weka (the dataset is included with Weka).



Q3 ① 2nd | child | male Survived ?

- P(Survived = No | E) =
P(Pclass = 2nd | survived = No) *
P(Age = child | survived = No) * P(sex = male | survived = No) / P(E) =

$$= \left(\frac{52}{1490}\right)\left(\frac{167}{1490}\right)\left(\frac{1364}{\cancel{1490}}\right)\left(\frac{\cancel{1490}}{2201}\right)$$

P(E)

$$= \frac{11844976}{4886440100}$$

P(E)

= 0.00242 / P(E)

- P (Survived = Yes | E) =
P(Pclass = 2nd | survived = Yes) *
P(Age = child | survived = Yes) * P(sex = male | survived = Yes) / P(E)

$$= \left(\frac{57}{711}\right)\left(\frac{118}{711}\right)\left(\frac{367}{\cancel{711}}\right)\left(\frac{\cancel{711}}{\overset{\cancel{1490}}{2201}}\right)$$

P(E)

$$= \frac{2468442}{1112651721}$$

P(E)

= 0.00221 / P(E)

$P(\text{Survived} = \text{Yes} \mid E) + P(\text{survived} = \text{No} \mid$

$0.00221 / P(E) + 0.00242 / P(E) = 1$

$P(E) = 0.00221 + 0.00242$

So,

$$P(\text{survived} = \text{Yes} \mid E) = \frac{0.00221}{(0.00221 + 0.00242)}$$

$= 0.477 = \boxed{47.7\%}$

$$P(\text{survived} = \text{No} \mid E) = \frac{0.00242}{0.00221 + 0.00242}$$

$= 0.522 = \boxed{52.2\%}$

Thus, $\boxed{\text{Survived} = \text{No}}$

(2) 2nd | adult | female Survived ?

- $P(\text{Survived} = \text{No} \mid E) =$
$P(\text{Pclass} = \text{2nd} \mid \text{survived} = \text{No}) *$
$P(\text{Age} = \text{Adult} \mid \text{survived} = \text{No}) * P(\text{Sex} = \text{female} \mid \text{Survived} = \text{No}) / P(E)$

$$= \frac{\left(\frac{167}{1490}\right)\left(\frac{1438}{1490}\right)\left(\frac{126}{1490}\right)\left(\frac{1490}{2201}\right)}{P(E)}$$

$$= \frac{\frac{30258396}{4886440100}}{P(E)}$$

$= 0.00619 / P(E)$

- $P(\text{survived} = \text{Yes} \mid E) =$
  $P(\text{class} = \text{2nd} \mid \text{survived} = \text{Yes}) *$
  $P(\text{Age} = \text{adult} \mid \text{survived} = \text{Yes}) * P(\text{sex} = \text{Female} \mid$
  $\text{survived} = \text{Yes}) / P(E)$

$$= \frac{\left(\frac{118}{711}\right)\left(\frac{654}{711}\right)\left(\frac{344}{711}\right)\left(\frac{711}{2201}\right)}{P(E)}$$

$$= \frac{\frac{26547168}{1112651721}}{P(E)}$$

$= 0.0238 / P(E)$

$P(\text{Survived} = \text{Yes} \mid E) + P(\text{Survived} = \text{No} \mid E) = 1$

$0.00619 \mid P(E) + 0.0238 \mid P(E) = 1$

$P(E) = 0.00619 + 0.0238$

So,

$$P(\text{survived} = \text{Yes} \mid E) = \frac{0.0238}{0.00619 + 0.0238}$$

$$= 0.7959 = 79.59\%$$

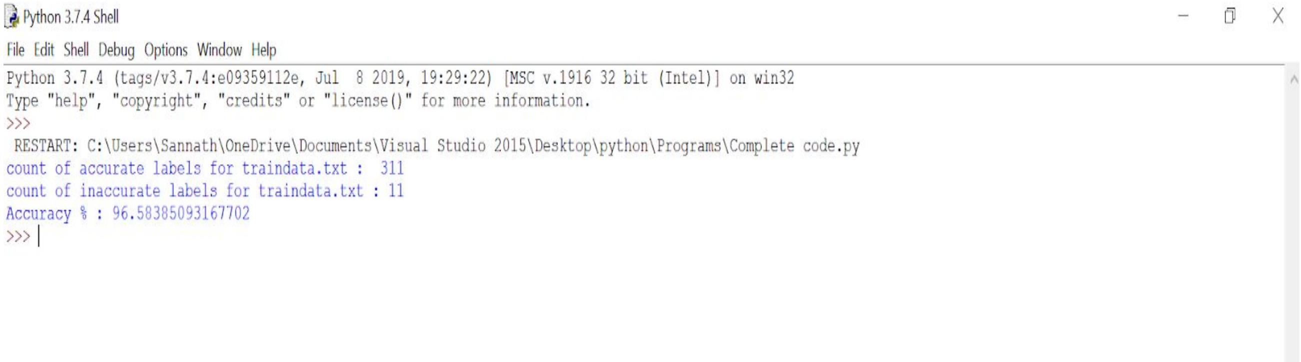
$$P(\text{survived} = \text{No} \mid E) = \frac{0.00619}{0.00619 + 0.0238}$$

$$= 0.2064 = 20.64\%$$

Thus, $\boxed{\text{survived} = \text{Yes}}$

4.1. Run your classifier by training on traindata.txt and trainlabels.txt then testing on traindata.txt and trainlabels.txt. Report the accuracy in results.txt (along with a comment saying what files you used for the training and testing data). In this situation, you are training and testing on the same data. This is a sanity check: your accuracy should be very high i.e. > 90%

(.py and .jpynb code are in separate file which is attached along with results.txt)



```
Python 3.7.4 (tags/v3.7.4:e09359112e, Jul  8 2019, 19:29:22) [MSC v.1916 32 bit (Intel)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
 RESTART: C:\Users\Sannath\OneDrive\Documents\Visual Studio 2015\Desktop\python\Programs\Complete code.py
count of accurate labels for traindata.txt :  311
count of inaccurate labels for traindata.txt : 11
Accuracy % : 96.58385093167702
>>> |
```

4.2. Run your classifier by training on traindata.txt and trainlabels.txt then testing on testdata.txt and testlabels.txt. Report the accuracy in results.txt (along with a comment saying what files you used for the training and testing data). We will not be letting you know beforehand what your performance on the test set should be.

(determined test labels of testdata.txt are saved into the file - ‘d_test_labels.txt’.)