**Name: SUN RUI**

**Student ID: 18083229g**

**Answer 1:**

**a.**

I(Lifestyle, Street, Polarized) = \* + \* + \* ≈ 1.5656

Entropy for TPR (Tear Production Rate):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TPR | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Reduced | 2 | 3 | 2 | 1.5567 |
| Normal | 4 | 1 | 3 | 1.4056 |

E(TPR) = \* 1.5567 + \* 1.4056 ≈ 1.4761

Information\_Gain (TPR) = 1.5656 – 1.4761 = 0.0895

Entropy for Sex:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| M | 4 | 1 | 3 | 1.4056 |
| F | 2 | 3 | 2 | 1.5567 |

E(Sex) = \* 1.4056 + \* 1.5567 ≈ 1.4761

Information\_Gain (Sex) = 1.5656 – 1.4761 = 0.0895

Entropy for Age:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TPR | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Young | 3 | 0 | 2 | 0.971 |
| Old | 2 | 2 | 2 | 1.585 |
| Middle | 1 | 2 | 1 | 1.5 |

E(Age) = \* 0.971 + \* 1.585 + \* 1.5 ≈ 1.3577

Information\_Gain (Age) = 1.5656 – 1.3577 = 0.2079

Entropy for SP (Spectacle Prescription):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SP | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Myope | 2 | 4 | 1 | 1.3788 |
| Hypermetrope | 4 | 0 | 4 | 1 |

E(SP) = \* 1.3788 + \* 1 ≈ 1.1768

Information\_Gain (SP) = 1.5656 – 1.1768 = 0.3888

Entropy for Astigmatism:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Astigmatism | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Yes | 5 | 0 | 3 | 0.9544 |
| No | 1 | 4 | 2 | 1.3788 |

E(Astigmatism) = \* 0.9544 + \* 1.3788 ≈ 1.1525

Information\_Gain (Astigmatism) = 1.5656 – 1.1525 = 0.4131

According above tables, we can find that the Information Gain of Astigmatism is larger than others, so we can set the **Astigmatism as root node**.

Next, we continue choosing its child nodes when Astigmatism is Yes and No respectively:  
**Yes:** I(Lifestyle, Street, Polarized) = \* + \* + \* ≈ 0.9544

Entropy for TPR (Tear Production Rate):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TPR | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Reduced | 2 | 0 | 2 | 1 |
| Normal | 3 | 0 | 1 | 0.8113 |

E(TPR) = \* 1 + \* 0.8113 ≈ 0.9057

Information\_Gain (TPR) = 0.9544 – 0.9057 = 0.0487

Entropy for Sex:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| M | 3 | 0 | 1 | 0.8113 |
| F | 2 | 0 | 2 | 1 |

E(Sex) = \* 0.8113 + \* 1 ≈ 0.9057

Information\_Gain (Sex) = 0.9544 – 0.9057 = 0.0487

Entropy for Age:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TPR | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Young | 2 | 0 | 0 | 0 |
| Old | 2 | 0 | 2 | 1 |
| Middle | 1 | 0 | 1 | 1 |

E(Age) = \* 0 + \* 1 + \* 1 ≈ 0.75

Information\_Gain (Age) = 0.9544 – 0.75 = 0.2044

Entropy for SP (Spectacle Prescription):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SP | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Myope | 1 | 0 | 1 | 1 |
| Hypermetrope | 4 | 0 | 2 | 0.9183 |

E(SP) = \* 1 + \* 0. 9183 ≈ 0.9387

Information\_Gain (SP) = 0. 9544 – 0.9387 = 0.0157

According above three tables, we can find the Information Gain of Age is the largest one, so we choose **Age as one of child nodes** of Astigmatism when Astigmatism is Yes.

**No:** I(Lifestyle, Street, Polarized) = \* + \* + \* ≈ 1.3788

Entropy for TPR (Tear Production Rate):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| TPR | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Reduced | 0 | 3 | 0 | 0 |
| Normal | 1 | 1 | 2 | 1.5 |

E(TPR) = \* 0 + \* 1.5 ≈ 0.8571

Information\_Gain (TPR) = 1.3788 – 0.8571 = 0.5217

Entropy for Sex:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sex | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| M | 1 | 1 | 2 | 1.5 |
| F | 0 | 3 | 0 | 0 |

E(Sex) = \* 1.5 + \* 0 ≈ 0.8571

Information\_Gain (Sex) = 1.3788 – 0.8571 = 0.5217

Entropy for Age:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Young | 1 | 0 | 2 | 0.9183 |
| Old | 0 | 2 | 0 | 0 |
| Middle | 0 | 2 | 0 | 0 |

E(Age) = \* 0.9183 + \* 0 + \* 1 ≈ 0.3936

Information\_Gain (Age) = 1.3788 – 0.3936 = 0.9852

Entropy for SP (Spectacle Prescription):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SP | Lifestyle | Street | Polarized | I(c1, c2, c3) |
| Myope | 1 | 4 | 0 | 0.7219 |
| Hypermetrope | 0 | 0 | 2 | 0 |

E(SP) = \* 0.7219 + \* 0 ≈ 0.5156

Information\_Gain (SP) = 1.3788 – 0.5156 = 0.8632

According above three tables, we can find the Information Gain of Age is the largest one, so we choose **Age as another child nodes** of Astigmatism when Astigmatism is No.

Now, analyze present situation, we can conclude that:

**IF Astigmatism = Yes AND Age = Young THEN Recommendation = Lifestyle**

**IF Astigmatism = No AND Age = Middle THEN Recommendation = Street**

**IF Astigmatism = No AND Age = Old THEN Recommendation = Street**

Because of this, we can ensure three leaf nodes. In additional, when Astigmatism is No, the Age only have one attribute Young which is not ensured, however the attribute SP can help us to divide Recommendations in this branch if Age is Young, so we can draw a conclusion:

**IF Astigmatism = No AND Age = Young AND SPC = Myope THEN Recommendation = Lifestyle**

**IF Astigmatism = No AND Age = Young AND SPC = Hypermetrope THEN Recommendation = Polarized**

Until now, there is no more child branches when Astigmatism is No, we start consider that Astigmatism is Yes:

Firstly, when Astigmatism is Yes and Age is Old, we can find attribute TPR (Tear Production Rate) can divide this child branch perfectly:

**IF Astigmatism = Yes AND Age = Old AND TPR = Normal THEN Recommendation = Lifestyle**

**IF Astigmatism = Yes AND Age = Old AND TPR = Reduced THEN Recommendation = Polarized**

Secondly, when Astigmatism is Yes and Age is Middle, we can find attribute Sex can divide this child branch perfectly:

**IF Astigmatism = Yes AND Age = Middle AND SEX = F THEN Recommendation = Lifestyle**

**IF Astigmatism = Yes AND Age = Middle AND SEX = M THEN Recommendation = Polarized**

Finally, the decision tree:

Astigmatism

No

Yes

Age

Age

Middle

Old

Young

Young

Old

Middle

TPR

Sex

Lifestyle

SP

Street

Street

Hypermetrope

Myope

F

M

Reduced

Normal

Polarized

Lifestyle

Polarized

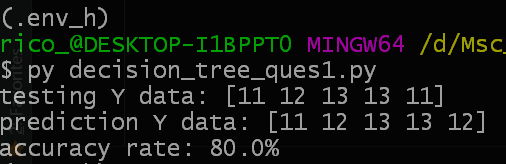
Polarized

Lifestyle

Lifestyle

**b.** The testing accuracy rate **80%**, the data in last row of Testing Data Set is not matching in my decision tree. The result from my decision tree is Street, actually the true result is Lifestyle.

**c. the result by Scikit-Learn:**



The accuracy rate is 80% by Scikit-Learn, the fault result is also one, the program also gives the prediction is 12(Street), actually the result should be Lifestyle. In conclusion, Scikit-Learn has the similar prediction with my decision.

Reference Python code:

*# -\*- encoding:utf-8 -\*-*import pandas as pd  
from sklearn import tree  
  
df\_train = pd.read\_csv("training\_data\_set.csv")  
df\_train.columns = ["Tear Production Rate", "Sex", "Age", "Spectacle Prescription", "Astigmatism", "Recommendation"]  
train\_X = df\_train.values[:, 0:5]  
train\_Y = df\_train.values[:, 5]  
  
df\_test = pd.read\_csv("test\_data\_set.csv")  
df\_test.columns = ["Tear Production Rate", "Sex", "Age", "Spectacle Prescription", "Astigmatism", "Recommendation"]  
test\_X = df\_test.values[:, 0:5]  
test\_Y = df\_test.values[:, 5]  
  
dtree = tree.DecisionTreeClassifier(criterion="gini")  
dtree.fit(train\_X, train\_Y)  
  
Y\_pred = dtree.predict(test\_X)  
print("testing Y data: {0}".format(test\_Y))  
print("prediction Y data: {0}".format(Y\_pred))  
print("accuracy rate: {0}%".format(dtree.score(test\_X, test\_Y)\*100))