ECT_HW2

2019

第一大題

第一大題

用 Weka 軟體對 contact-lenses.arff 建立 J48 決策樹，選擇 “Use training set” ，設定Attribute: contact-lenses 為 Output，在過程中對重要步驟截圖並加以說明，並回答以下問題：

第一大題(a)-題目

(a)在前處理部分，右下角選擇不同屬性作為Class，請解釋長條圖的數量、上方的數字以及不同顏色意義為何?(15%)

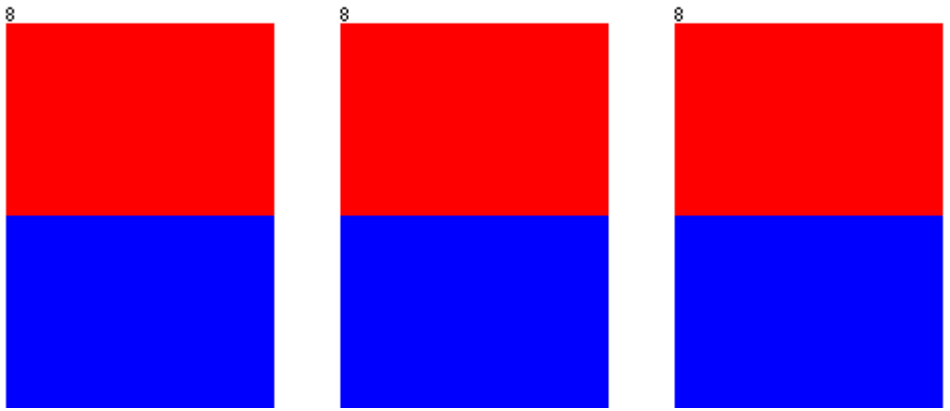
第一大題(a)-解答

Selected attribute

Name: age
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count	Weight
1	young	8	8.0
2	pre-presbyopic	8	8.0
3	presbyopic	8	8.0

Class: spectacle-prescrip (Nom) ▼ Visualize All

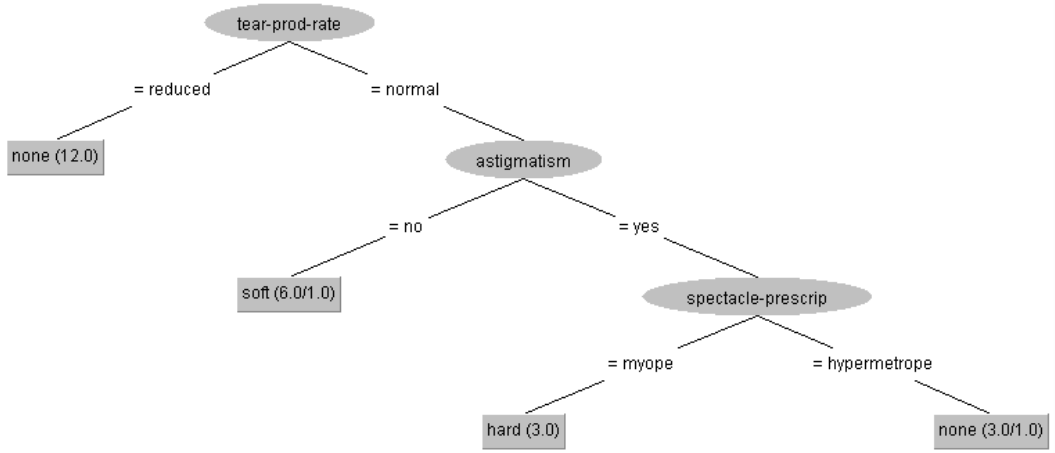


- 在每個屬性中，長條圖的數目=Distinct的數目
- 長條圖上方數字代表有多少instance為此屬性值，顏色代表在此屬性值中，選擇的class的分布

第一大題(b)-題目

(b)使用 Visualize Tree 或 Classifier Output 列出三格 Classification Rule並解釋。(20%)

第一大題(b)-解答




```
tear-prod-rate = reduced: none (12.0)
tear-prod-rate = normal
|   astigmatism = no: soft (6.0/1.0)
|   astigmatism = yes
|   |   spectacle-prescrip = myope: hard (3.0)
|   |   spectacle-prescrip = hypermetrope: none (3.0/1.0)
```

- 若tear-prod-rate = reduced 則class=none
- 若tear-prod-rate = normal，且astigmatism = no，則class=soft
- 若tear-prod-rate = normal，astigmatism = yes，且spectacle-prescrip = myope，則class=hard 。

第二大題

第二大題

請利用weka和python對 glass.csv 進行Supervised learning中的DecisionTree分析 ,並回答以下問題:

第二大題(a)-題目

(a)請運用python的train_test_split 對glass.csv 資料集，預測目標屬性為Type進行訓練集(66%)、測試集(34%)切分並進行訓練，請將重要程式碼截圖並說明(10%)

第二大題(a)-解答

```
#x:input
x=data.loc[:,['RI','Na','Mg','Al','Si','K','Ca','Ba','Fe']]
#y:output
y=data.loc[:,['Type']]
```

```
from sklearn import preprocessing
le = preprocessing.LabelEncoder()
y_encoded=le.fit_transform(y.Type)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(x, y_encoded, test_size=0.34)
```

```
from sklearn import tree
clf = tree.DecisionTreeClassifier(criterion = 'entropy',max_depth=10,max_leaf_nodes = 12)
glass_clf = clf.fit(X_train,y_train)
```

第二大題(b)-題目

(b)請利用參數(

criterion = ‘entropy’,max_depth=3, max_leaf_nodes = 4)

對切分出的訓練集進行訓練，並用metrics.accuracy_score( )計算出模型對於測試集的精準度，並與WEKA設定演算法J48 Percentage spilt 66%跑出的結果截圖說明並一起呈現比較。(20%)

第二大題(b)-解答

```python
from sklearn import metrics
# 績效
accuracy = metrics.accuracy_score(test_y_predicted, y_train)
print(accuracy)
```

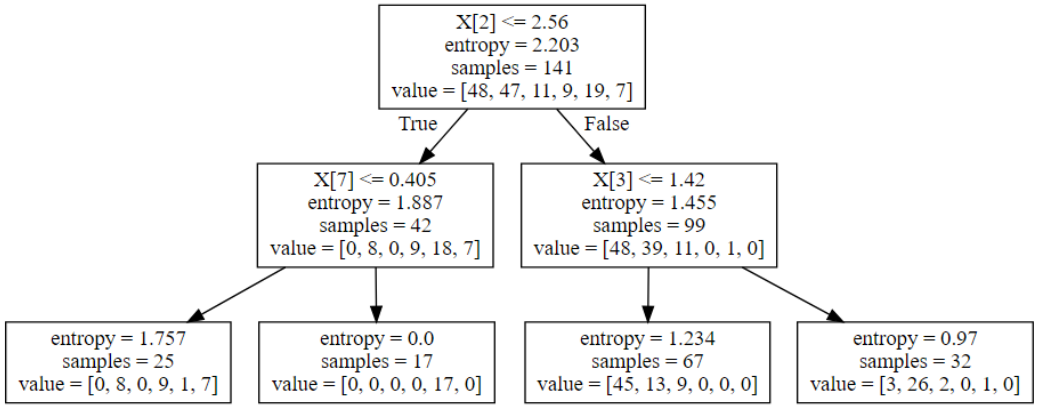
```
0.6879432624113475
```

```
=== Summary ===

Correctly Classified Instances          42               57.5342 %
Incorrectly Classified Instances        31               42.4658 %
Kappa statistic                          0.4259
Mean absolute error                      0.1246
Root mean squared error                  0.3287
Relative absolute error                 58.7442 %
Root relative squared error            101.8335 %
Total Number of Instances               73

=== Detailed Accuracy By Class ===
```

第二大題(c)-題目

(c)請利用graphviz套件跑出決策樹圖形截圖並加以說明當中X[?]、samples、及value各代表的訊息，另外運用WEKA visualize tree觀察決策樹圖形並說明WEKA中葉節點的標籤及括號內的數字代表的意義。(20%)

第二大題(c)-解答

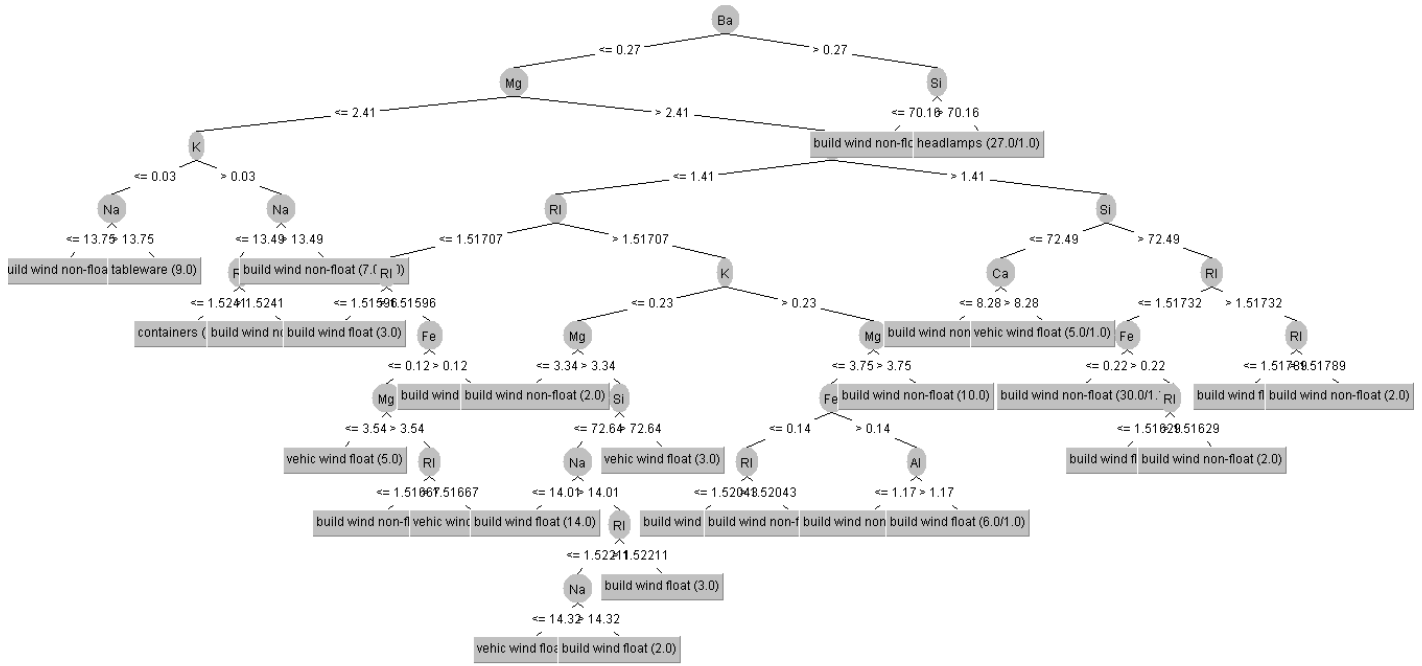


X[?]代表feature

Samples代表被分類的樣本總數

Value代表被分類的樣本分布

第二大題(c)-解答



- 各葉節點代表分類類別
- 括號中的數值代表(總分類數/錯誤樣本數)

第二大題(d)-題目

(d)請試著調整DecisionTreeClassifier 的參數，提升模型準確率，請截圖並附上每次測試的結果，觀察並說明準確率上升的原因，另外說明模型在訓練集的準確度通常較訓練集的準確度高的原因為何(15%)

第二大題(d)-解答

請同學自由作答

模型容易對訓練集overfitting