Final Data mining project

Pooriya Babaei 943611043004

Data set Inf

This dataset describes 23 species of Agaricus and Lepiota mushroom families. Each specie is Edible, Poisonous, Not recommended or Unknown edibility. There is no simple rules to determine the type of a mushroom.

Attributes

1. cap-shape: bell=b,conical=c,convex=x,flat=f, knobbed=k,sunken=s

2. cap-surface: fibrous=f,grooves=g,scaly=y,smooth=s

3. cap-color: brown=n,buff=b,cinnamon=c,gray=g,green=r,
pink=p,purple=u,red=e,white=w,yellow=y

4. bruises?: bruises=t,no=f

5. odor: almond=a,anise=l,creosote=c,fishy=y,foul=f, musty=m,none=n,pungent=p,spicy=s

6. gill-attachment: attached=a,descending=d,free=f,notched=n

7. gill-spacing: close=c,crowded=w,distant=d

8. gill-size: broad=b,narrow=n

9. gill-color: black=k,brown=n,buff=b,chocolate=h,gray=g,
green=r,orange=o,pink=p,purple=u,red=e, white=w,yellow=y

10. stalk-shape: enlarging=e,tapering=t

11. stalk-root: bulbous=b,club=c,cup=u,equal=e, rhizomorphs=z,rooted=r,missing=?

12. stalk-surface-above-ring: fibrous=f,scaly=y,silky=k,smooth=s

13. stalk-surface-below-ring: fibrous=f,scaly=y,silky=k,smooth=s

14. stalk-color-above-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
pink=p,red=e,white=w,yellow=y

15. stalk-color-below-ring: brown=n,buff=b,cinnamon=c,gray=g,orange=o,
pink=p,red=e,white=w,yellow=y

16. veil-type: partial=p,universal=u

17. veil-color: brown=n,orange=o,white=w,yellow=y

18. ring-number: none=n,one=o,two=t

19. ring-type: cobwebby=c,evanescent=e,flaring=f,large=l,
none=n,pendant=p,sheathing=s,zone=z

20. spore-print-color: black=k,brown=n,buff=b,chocolate=h,green=r,
orange=o,purple=u,white=w,yellow=y

21. population: abundant=a,clustered=c,numerous=n, scattered=s,several=v,solitary=y

22. habitat: grasses=g,leaves=l,meadows=m,paths=p, urban=u,waste=w,woods=d

Data Set Characteristics:	Multivariate	**Number of Instances:**	8124	**Area:**	Life
Attribute Characteristics:	Categorical	**Number of Attributes:**	22		
Associated Tasks:	Classification	**Missing Values?**	Yes		

Implementation

- First we import the dataset with csv package
- Separate target class feature from other features
- Split dataset into TRAIN and TEST sets with KFOLD CROSS VALIDATION which is in SKLEARN.MODEL_SELECTION.
- To make the decision trees for both GINI and ID3, use DecisionTreeClassifier in SKLEARN.TREE. We put the max depth size equal to 20 and the min split equal to 2. The splitter itself chooses the best split point.
- Install GraphViz to save the graph in a pdf
- To predict the test set data, use predict function and for measuring the precision, recall and f-measure, use classification_report function

- The only difference between ID3 and GINI in this implementation is the criterion paramaeter in DecisionTreeClassifier. For gini is ‘gini’ and for id3 is ‘entropy’.

- For KNN method, use train_test_split func in SKLEARN.MODEL_SELECTION and put the test set size 1/5 of train test size.
- Then set K(neighbor numbers) in range of 1 to 40 to find the best results. KNeighborsClassifier helps us to set K and classify. To see the results you should run the ‘knn.py’ file.

Output samples:

KNN

Below is the results for K = 1

```
DataSet:
[1 1 1 ... 1 0 0]

K is  1

Classifier Predictions:
[1 1 1 ... 1 0 0]

acuuracy is 1.0

              precision    recall  f1-score   support

           0       1.00      1.00      1.00       698
           1       1.00      1.00      1.00       431

   micro avg       1.00      1.00      1.00      1129
   macro avg       1.00      1.00      1.00      1129
weighted avg       1.00      1.00      1.00      1129
```

ID3

```
Accuracy is  0.9464285714285714

              precision    recall  f1-score   support

           0       0.98      0.85      0.91       507
           1       0.94      0.99      0.96      1117

   micro avg       0.95      0.95      0.95      1624
   macro avg       0.96      0.92      0.94      1624
weighted avg       0.95      0.95      0.95      1624


Process finished with exit code 0
```

GINI

```
Accuracy is  0.9415024630541872

              precision    recall  f1-score   support

           0       0.96      0.85      0.90       507
           1       0.94      0.98      0.96      1117

   micro avg       0.94      0.94      0.94      1624
   macro avg       0.95      0.92      0.93      1624
weighted avg       0.94      0.94      0.94      1624
```

So as we see, results for gini and id3 are very close to eachother.

Note :

The exported rules from ID3 are in ID3Rules.txt file .

The exported rules from GINI are in GiniRules.txt file .

ID3 tree is in ID3 pdf and GINI tree is in GINI pdf.

Sorry to say this, but you have to hard code the path of Graphviz bin folder at line 15 of Gini.py and ID3.py