Department of Computer Science & Engineering

Final Year B. Tech. (CSE) – I : 2021-22

4CS462 : PE2 - Data Mining Lab

Assignment No. 4

By DM21G03

(2018BTECS00082 : Hritik Belani , 2019BTECS00209 : Sailee Akim)

Date: 04/09/2021

- **Title**

  Data Analysis Tool

- **Objective**

Use / extend the data analysis tool (menu driven GUI) developed in Assignment No. 2 to perform the classification task.

- **Specification**

- Python 3.8.11

- Dataset

- **Introduction and Theory**

- **Entropy -** Entropy is the degree of uncertainty, impurity or disorder of a random variable, or a measure of purity. It characterizes the impurity of an arbitrary class of examples.*Entropy is the measurement of impurities or randomness in the data points.*Here, if all elements belong to a single class, then it is termed as "Pure", and if not then the distribution is named as "Impurity".It is computed between 0 and 1, however, heavily relying on the number of groups or classes present in the data set it can be more than 1 while depicting the same significance i.e. extreme level of disorder.In more simple terms, If a dataset contains homogeneous subsets of observations, then no impurity or randomness is there in the dataset, and if all the observations belong to one class, the entropy of that dataset becomes zero.
- **Information Gain -** Information gain computes the difference between entropy before and after split and specifies the impurity in class elements.

Information Gain = Entropy before splitting - Entropy after splitting

Generally, it is not preferred as it involves ‘log’ function that results in the computational complexity. Moreover;

1. Information gain is non-negative.
2. Information Gain is symmetric such that switching of the split variable and target variable, the same amount of information gain is obtained.
3. Information gain determines the reduction of the uncertainty after splitting the dataset on a particular feature such that if the value of information gain increases, that feature is most useful for classification.
4. The feature having the highest value of information gain is accounted for as the best feature to be chosen for split.

- **Gain Ratio -** Gain Ratio or Uncertainty Coefficient is used to normalize the information gain of an attribute against how much entropy that attribute has. Formula of gini ratio is given by

  Gain Ratio=Information Gain/Entropy

  From the above formula, it can be stated that if entropy is very small, then the gain ratio will be high and vice versa.Be selected as splitting criterion, Quinlan proposed following procedure, First, determine the information gain of all the attributes, and then compute the average information gain.Second, calculate the gain ratio of all the attributes whose calculated information gain is larger or equal to the computed average information gain, and then pick the attribute of higher gain ratio to split.

- **Gini Index -** The gini index, or gini coefficient, or gini impurity computes the degree of probability of a specific variable that is wrongly being classified when chosen randomly and a variation of gini coefficient. It works on categorical variables, provides outcomes either be “successful” or “failure” and hence conducts binary splitting only.The degree of gini index varies from 0 to 1, Where 0 depicts that all the elements be allied to a certain class, or only one class exists there. The gini index of value as 1 signifies that all the elements are randomly zdistributed across various classes, and. A value of 0.5 denotes the elements are uniformly distributed into some classes.

- **Procedure**

1. Implement the decision tree classifier using the following attribute selection measures and graphically show/visualize the tree:

   a. Information Gain

   b. Gain Ratio

   c. Gini Index

2. Tabulate the results in confusion matrix and evaluate the performance of above classifier using following metrics :

   a) Recognition rate

   b) Misclassification rate

   c) Sensitivity

   d) Specificity

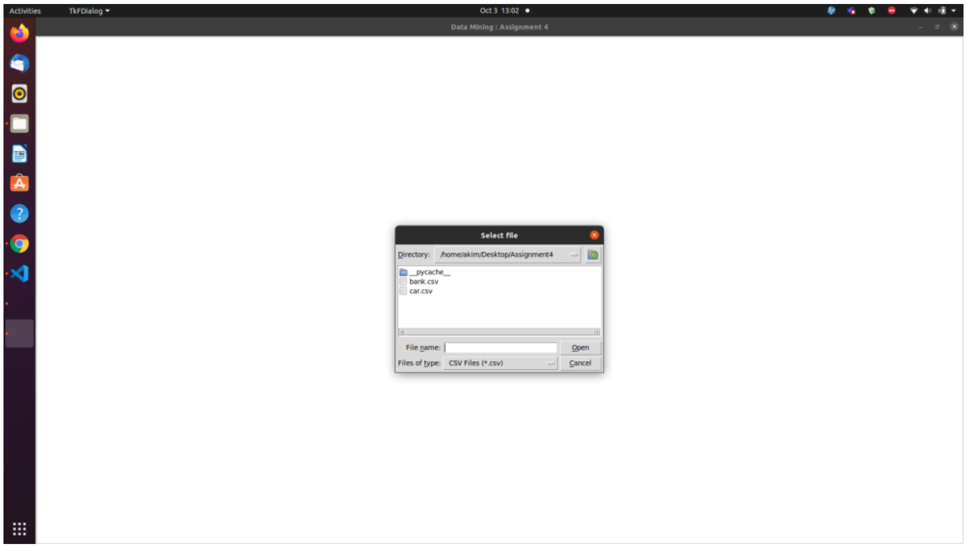
   e) Precision & Recall

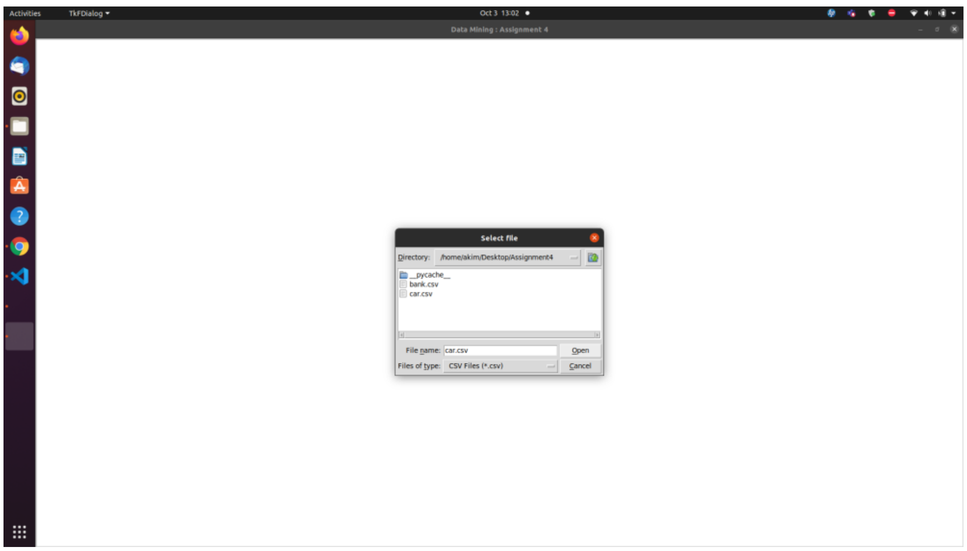
3. Use the following categorical data sets from UCI machine learning repository :

   a. Balance Scale data set

   b. Car evaluation data set

   c. Breast-cancer data set


Activities
TkFDialog
Oct 3 13:02
Data Mining : Assignment 4
Select file
Directory:
/home/akim/Desktop/Assignment4
__pycache__
bank.csv
car.csv
File name:
Open
Files of type:
CSV Files (*.csv)
Cancel


Activities
TkFDialog
Oct 3 13:02
Data Mining : Assignment 4
Select file
Directory:
/home/akim/Desktop/Assignment4
__pycache__
bank.csv
car.csv
File name:
car.csv
Open
Files of type:
CSV Files (*.csv)
Cancel


Activities
Tk
Oct 3 13:02
Data Mining : Assignment 4
****
*********** Final Tree with accuracy: 98.0% ************
****
Is persons == 2? id: 0 depth: 0
--> True:
Leaf id: 2 Predictions: {'unacc': 479} Label Class: unacc
--> False:
Is safety == low? id: 1 depth: 1
--> True:
Leaf id: 4 Predictions: {'unacc': 307} Label Class: unacc
--> False:
Is buying == vhigh? id: 3 depth: 2
--> True:
Is maint == vhigh? id: 8 depth: 3
--> True:
Leaf id: 18 Predictions: {'unacc': 40} Label Class: unacc
--> False:
Is maint == high? id: 17 depth: 4
--> True:
Leaf id: 36 Predictions: {'unacc': 38} Label Class: unacc
--> False:
Is lug_boot == small? id: 35 depth: 5
--> True:
Is safety == high? id: 72 depth: 6
--> True:
Is doors == 2? id: 146 depth: 7
--> True:
Is persons == 4? id: 294 depth: 8
--> True:
Leaf id: 590 Predictions: {'acc': 2} Label Class: acc
--> False:
Leaf id: 589 Predictions: {'unacc': 2} Label Class: unacc
--> False:
Leaf id: 293 Predictions: {'acc': 9} Label Class: acc
--> False:
Leaf id: 145 Predictions: {'unacc': 14} Label Class: unacc
--> False:
Is safety == high? id: 71 depth: 6
--> True:
Leaf id: 144 Predictions: {'acc': 27} Label Class: acc
--> False:
Is lug_boot == med? id: 143 depth: 7
--> True:
Is doors == 2? id: 288 depth: 8
--> True:
Leaf id: 578 Predictions: {'unacc': 2} Label Class: unacc
--> False:
Is doors == 3? id: 577 depth: 9
--> True:
Is persons == 4? id: 1156 depth: 10
--> True:
Leaf id: 2314 Predictions: {'unacc': 2} Label Class: unacc
--> False:
Leaf id: 2313 Predictions: {'acc': 2} Label Class: acc
--> False:
Leaf id: 1155 Predictions: {'acc': 6} Label Class: acc
--> False:

Activities Tk Oct 3 13:02 Data Mining : Assignment 4

```
          --> True:
            Leaf id: 538 Predictions: {'good': 3} Label Class: good
          --> False:
            Is doors == 3? id: 537 depth: 9
            --> True:
              Is persons == 4? id: 1076 depth: 10
              --> True:
                Leaf id: 2154 Predictions: {'good': 2} Label Class: good
              --> False:
                Leaf id: 2153 Predictions: {'vgood': 2} Label Class: vgood
            --> False:
              Leaf id: 1075 Predictions: {'vgood': 7} Label Class: vgood
        --> False:
          Leaf id: 267 Predictions: {'vgood': 14} Label Class: vgood
    --> False:
      Is maint == med? id: 65 depth: 6
      --> True:
        Is lug_boot == small? id: 132 depth: 7
        --> True:
          Is buying == low? id: 266 depth: 8
          --> True:
            Is doors == 2? id: 534 depth: 9
            --> True:
              Is persons == 4? id: 1070 depth: 10
              --> True:
                Leaf id: 2142 Predictions: {'good': 1} Label Class: good
              --> False:
                Leaf id: 2141 Predictions: {'unacc': 1} Label Class: unacc
            --> False:
              Leaf id: 1069 Predictions: {'good': 4} Label Class: good
          --> False:
            Is doors == 2? id: 533 depth: 9
            --> True:
              Is persons == 4? id: 1068 depth: 10
              --> True:
                Leaf id: 2138 Predictions: {'acc': 1} Label Class: acc
              --> False:
                Leaf id: 2137 Predictions: {'unacc': 1} Label Class: unacc
            --> False:
              Leaf id: 1067 Predictions: {'acc': 6} Label Class: acc
        --> False:
          Is lug_boot == med? id: 265 depth: 8
          --> True:
            Is doors == 2? id: 532 depth: 9
            --> True:
              Is persons == 4? id: 1066 depth: 10
              --> True:
                Leaf id: 2134 Predictions: {'acc': 1} Label Class: acc
              --> False:
                Leaf id: 2133 Predictions: {'good': 1} Label Class: good
            --> False:
              Is persons == 4? id: 1065 depth: 10
              --> True:
                Is doors == 3? id: 2132 depth: 11
                --> True:
                  Leaf id: 4266 Predictions: {'acc': 1} Label Class: acc
                --> False:
                  Leaf id: 4265 Predictions: {'vgood': 2} Label Class: vgood
              --> False:
```

Activities Tk Oct 3 13:02 Data Mining : Assignment 4

```
                Leaf id: 2131 Predictions: {'vgood': 6} Label Class: vgood
          --> False:
            Leaf id: 531 Predictions: {'vgood': 11} Label Class: vgood
      --> False:
        Is buying == med? id: 131 depth: 7
        --> True:
          Leaf id: 264 Predictions: {'acc': 36} Label Class: acc
        --> False:
          Is maint == high? id: 263 depth: 8
          --> True:
            Is lug_boot == small? id: 528 depth: 9
            --> True:
              Leaf id: 1058 Predictions: {'acc': 6} Label Class: acc
            --> False:
              Is lug_boot == med? id: 1057 depth: 10
              --> True:
                Is doors == 5more? id: 2116 depth: 11
                --> True:
                  Leaf id: 4234 Predictions: {'vgood': 2} Label Class: vgood
                --> False:
                  Is doors == 4? id: 4233 depth: 12
                  --> True:
                    Leaf id: 8468 Predictions: {'vgood': 1} Label Class: vgood
                  --> False:
                    Leaf id: 8467 Predictions: {'acc': 2} Label Class: acc
              --> False:
                Leaf id: 2115 Predictions: {'vgood': 5} Label Class: vgood
          --> False:
            Is doors == 2? id: 527 depth: 9
            --> True:
              Is lug_boot == big? id: 1056 depth: 10
              --> True:
                Leaf id: 2114 Predictions: {'acc': 2} Label Class: acc
              --> False:
                Is persons == 4? id: 2113 depth: 11
                --> True:
                  Leaf id: 4228 Predictions: {'acc': 1} Label Class: acc
                --> False:
                  Leaf id: 4227 Predictions: {'unacc': 1} Label Class: unacc
            --> False:
              Leaf id: 1055 Predictions: {'acc': 9} Label Class: acc
  --> False:
    Is lug_boot == small? id: 31 depth: 5
    --> True:
      Is maint == vhigh? id: 64 depth: 6
      --> True:
        Leaf id: 130 Predictions: {'unacc': 13} Label Class: unacc
      --> False:
        Is doors == 2? id: 129 depth: 7
        --> True:
          Is persons == 4? id: 260 depth: 8
          --> True:
            Is maint == high? id: 522 depth: 9
            --> True:
              Is buying == med? id: 1046 depth: 10
              --> True:
                Leaf id: 2094 Predictions: {'unacc': 1} Label Class: unacc
              --> False:
                Leaf id: 2093 Predictions: {'acc': 1} Label Class: acc
```

Activities Tk Oct 3 13:03 Data Mining : Assignment 4

```
              Leaf id: 1027 Predictions: {'good': 4} Label Class: good
        --> False:
          Is doors == 2? id: 255 depth: 8
          --> True:
            Is lug_boot == med? id: 512 depth: 9
            --> True:
              Is maint == vhigh? id: 1026 depth: 10
              --> True:
                Leaf id: 2054 Predictions: {'unacc': 3} Label Class: unacc
              --> False:
                Is buying == med? id: 2053 depth: 11
                --> True:
                  Leaf id: 4108 Predictions: {'unacc': 2} Label Class: unacc
                --> False:
                  Leaf id: 4107 Predictions: {'acc': 2} Label Class: acc
            --> False:
              Leaf id: 1025 Predictions: {'acc': 7} Label Class: acc
          --> False:
            Is doors == 3? id: 511 depth: 9
            --> True:
              Is persons == 4? id: 1024 depth: 10
              --> True:
                Is lug_boot == med? id: 2050 depth: 11
                --> True:
                  Is maint == high? id: 4102 depth: 12
                  --> True:
                    Leaf id: 8206 Predictions: {'acc': 1} Label Class: acc
                  --> False:
                    Leaf id: 8205 Predictions: {'unacc': 1} Label Class: unacc
                --> False:
                  Leaf id: 4101 Predictions: {'acc': 3} Label Class: acc
              --> False:
                Leaf id: 2049 Predictions: {'acc': 7} Label Class: acc
            --> False:
              Leaf id: 1023 Predictions: {'acc': 24} Label Class: acc
Confusion Matrix:
[[[36 27]
  [10 10]]

 [[66 13]
  [ 4  0]]

 [[19  8]
  [40 16]]

 [[72  8]
  [ 2  1]]]
              precision    recall  f1-score   support

         acc       0.27      0.50      0.35        20
        good       0.00      0.00      0.00         4
       unacc       0.67      0.29      0.40        56
       vgood       0.11      0.33      0.17         3

    accuracy                           0.33        83
   macro avg       0.26      0.28      0.23        83
weighted avg       0.52      0.33      0.36        83
```

❖ Conclusion

Thus implemented and dicovered various ways to build decision trees along with their metrics such as recall, precision, support etc.

❖ References

https://www.analyticssteps.com/blogs/what-gini-index-and-information-gain-decision-trees

https://datascience.stackexchange.com/questions/9325/python-library-that-can-compute-the-confusion-matrix-for-multi-label-classificat