Training Machine learning Classifers

Name : Sruthi Pisipati

Program : Masters in Analytics

E-mail: spisipati2@student.gsu.edu

***Task -1) What can you tell me about this data set? E.g. What is the size of the data set? How many descriptive attributes do you have in your file? Are the classes balanced*?**

Answer: The dataset "wineData.csv" is very small with only 118 rows. It has 14 columns. All of the descriptive attributes are of type numerical/floating/decimal. The following table shows the data types of all columns in the dataset.

```
=====================================
===== Data Frame Datatypes ========
Class                                object
Alcohol                             float64
Malic acid                          float64
Ash                                 float64
Alcalinity of ash                   float64
Magnesium                             int64
Total phenols                       float64
Flavanoids                          float64
Nonflavanoid phenols                float64
Proanthocyanins                     float64
Color intensity                     float64
Hue                                 float64
OD280/OD315 of diluted wines        float64
Proline                               int64
```

Fig. 1. Output showing the datatypes related to Descriptive attributes

The size of the dataset is 118 rows & 14 columns. The number of descriptive attributes is 14. The classes are balanced. This can be observed from the bar graph as below. The class which is shown as 1 in the below graph is the class which belongs to "Chardonnay" & Class which is shown as 0 in the below graph is "Merlot". As observed both classes are evenly balanced with the number of records – 59.

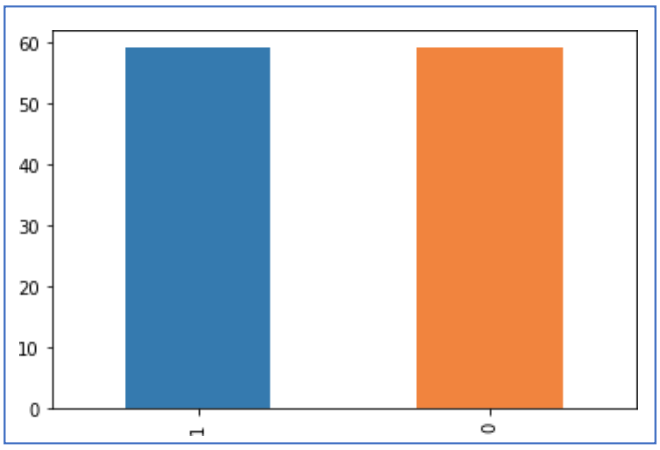

Fig. 2. Bar graph showing balanced class

Task-2) Which Decision Tree is the best in your opinion (provide sufficient information about your winning classifier, so I could re-produce it easily) and why? Show your score charts (generated for training and testing sets) to support your recommendation. Finally, visualize the winning tree and include it in your written report.

Answer: After normalization of the dataset, the dataset looks as under:

	Alcohol	Malic acid	...	Proline	Class
0	1.684211	0.383399	...	1.122682	0
1	1.142105	0.411067	...	1.101284	0
2	1.121053	0.640316	...	1.293866	0
3	1.757895	0.478261	...	1.714693	0
4	1.163158	0.731225	...	0.651926	0

Fig. 3. Output showing the normalized data between 0 & 2 and the class mapping to numericals.

The following results show the results of Decision trees built on two different criteria – Entropy & Gini index.

```
=============================================================
Decision Tree Classifier - With Criterion Entropy
=============================================================
    LevelLimit  Score For Training  Score for Testing
1          1.0            0.962025           0.923077
2          2.0            0.974684           0.923077
3          3.0            1.000000           0.948718
4          4.0            1.000000           0.948718
5          5.0            1.000000           0.948718
6          6.0            1.000000           0.948718
7          7.0            1.000000           0.948718
8          8.0            1.000000           0.948718
9          9.0            1.000000           0.948718
10        10.0            1.000000           0.948718
11        11.0            1.000000           0.948718
=============================================================
Decision Tree Classifier - With Criterion Gini
=============================================================
    LevelLimit  Score For Training  Score for Testing
1          1.0            0.962025           0.923077
2          2.0            0.974684           0.923077
3          3.0            1.000000           0.948718
4          4.0            1.000000           0.948718
5          5.0            1.000000           0.948718
6          6.0            1.000000           0.948718
7          7.0            1.000000           0.948718
8          8.0            1.000000           0.948718
9          9.0            1.000000           0.948718
10        10.0            1.000000           0.948718
11        11.0            1.000000           0.948718
```

Fig. 4. Score of Decision Tree classifier with varying depths & two types of criteria.

The decision tree with a depth of 2.0, Training score of 97.4% and testing score of 92.30 % is the best classifier. Both the decision trees with criteria "Entropy" & "Gini index" show the exact same result. If we choose any other decision tree which is greater than depth level of 3.0 and with either of the criteria – Entropy & Gini index, it seems that the score of training is 100% which is a clear case of overfitting. Hence the best decision tree chosen here is with depth level 2.0. The visualization of the best decision tree is shown in the following figures. Also we can observe that with a depth

greater than 3.0 & a training score of 100%, the testing score is constant at 94.87%. It does not get any better with increasing depths of the decision tree. This happens both in case of criteria taken as "Entropy" & criteria taken as "Gini index".

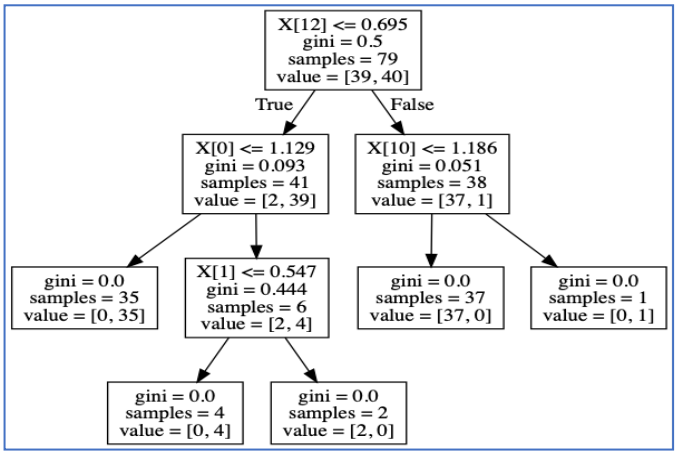

Fig.5 Decision tree with a depth level of 2.0 ,training score of 97.4% and testing score of 92.30 % is the best classifier. The criteria is Gini index.

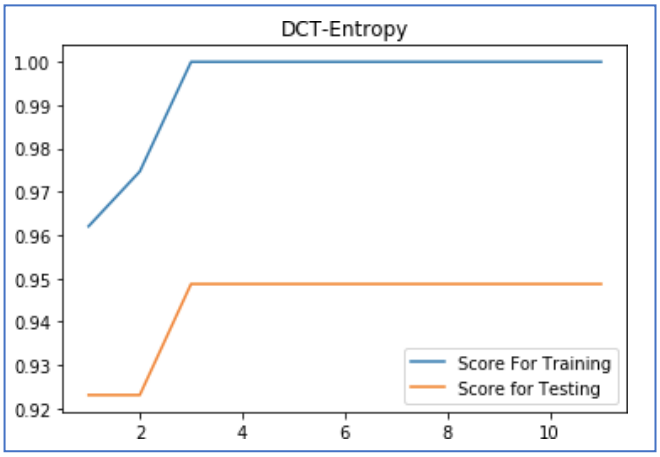

Fig.6 Graph showing variaton of scores for Training & Testing – Decision tree with Entropy as criteria

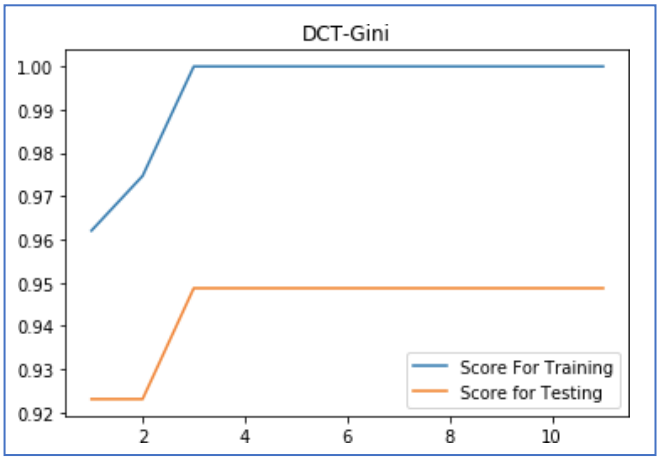

Fig.7 Graph showing variaton of scores for Training & Testing – Decision tree with Gini index as criteria

Task-3) Which KNN classifier of the ones you investigated is the best in your opinion (provide sufficient information about your winning classifier, so I could re-produce it easily) and why? Show your score charts (generated for training and testing sets) to support your recommendation.

***Answer*:** There seems to be zigzag & frequently varying trends across all the KNN classifiers generated with different neighbors count & different metrics used – Euclidean & Manhattan, along with uniform weights & weighted distances. This can be said because of the scores oscillating between 100% to 90% & back again to 100% as seen in the below charts. None of them can be termed as the best KNN classifier because of the uncertain behavior portrayed.

However, if we have to choose one, the KNN classifier with K = 2, Score for training = 96.20% & Score for testing = 94.87% can be chosen as the best KNN classifier. This is the highest optimal score generated which does not overfit. Other than this all others either overfit or are low scores than these.

```
============================================================
KNN Classifier - With Eucledian distance , Uniform Weights
============================================================
     KNN  Score For Training  Score for Testing
1    1.0             1.000000            0.948718
2    2.0             0.962025            0.948718
3    3.0             0.974684            1.000000
4    4.0             0.936709            0.948718
5    5.0             0.962025            0.974359
6    6.0             0.949367            0.974359
7    7.0             0.974684            1.000000
8    8.0             0.949367            0.974359
9    9.0             0.987342            0.974359
10  10.0             0.962025            0.974359
11  11.0             0.962025            1.000000
============================================================
KNN Classifier - With Eucledian distance , Weighted Distance
============================================================
     KNN  Score For Training  Score for Testing
1    1.0                 1.0            0.948718
2    2.0                 1.0            0.948718
3    3.0                 1.0            1.000000
4    4.0                 1.0            0.974359
5    5.0                 1.0            0.974359
6    6.0                 1.0            0.974359
7    7.0                 1.0            1.000000
8    8.0                 1.0            0.974359
9    9.0                 1.0            0.974359
10  10.0                 1.0            0.974359
11  11.0                 1.0            1.000000
```

Fig.8 Score of KNN classifier with varying depths, Eucledian distance, uniform weights & weighted distances.

```
============================================================
KNN Classifier - With Manhattan distance , Uniform Weights
============================================================
     KNN  Score For Training  Score for Testing
1    1.0             1.000000            0.974359
2    2.0             0.949367            0.923077
3    3.0             0.974684            1.000000
4    4.0             0.949367            0.948718
5    5.0             0.974684            0.974359
6    6.0             0.949367            0.974359
7    7.0             0.962025            1.000000
8    8.0             0.924051            1.000000
9    9.0             0.949367            1.000000
10  10.0             0.949367            1.000000
11  11.0             0.949367            1.000000
============================================================
KNN Classifier - With Manhattan distance , Weighted Distance
============================================================
     KNN  Score For Training  Score for Testing
1    1.0                 1.0            0.974359
2    2.0                 1.0            0.974359
3    3.0                 1.0            1.000000
4    4.0                 1.0            1.000000
5    5.0                 1.0            0.974359
6    6.0                 1.0            1.000000
7    7.0                 1.0            1.000000
8    8.0                 1.0            1.000000
9    9.0                 1.0            1.000000
10  10.0                 1.0            1.000000
11  11.0                 1.0            1.000000
```

Fig.9 Score of KNN classifier with varying depths, Manhattan distance, uniform weights & weighted distances.

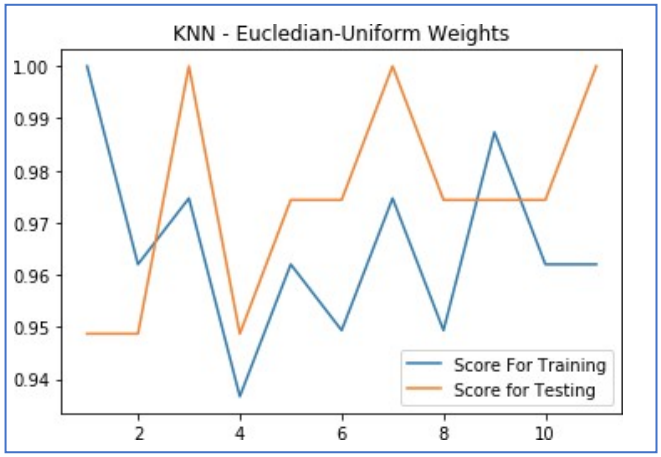

Fig.10 Graph showing variaton of scores for Training & Testing – KNN – Eucledian-Uniform weights

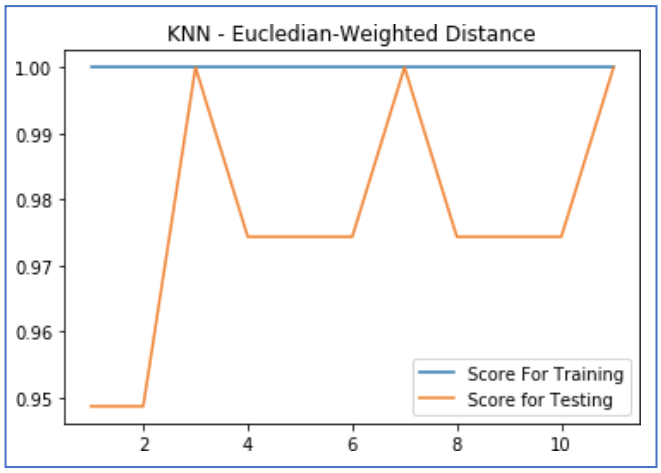

Fig.11 Graph showing variaton of scores for Training & Testing – KNN – Eucledian-Weighted distance

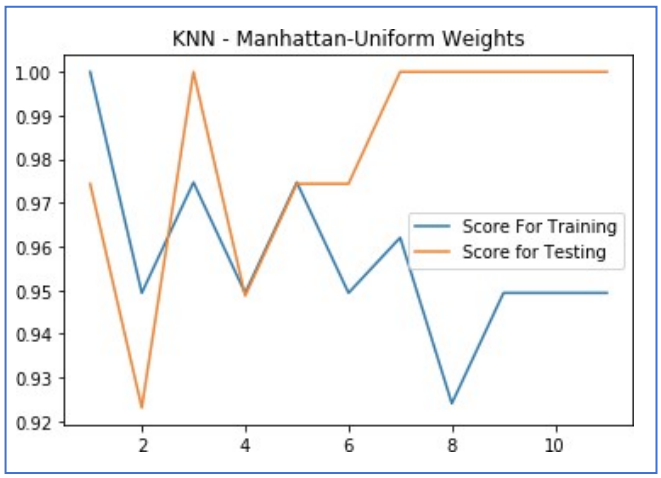

Fig.12 Graph showing variaton of scores for Training & Testing – KNN – Manhattan – Uniform weights

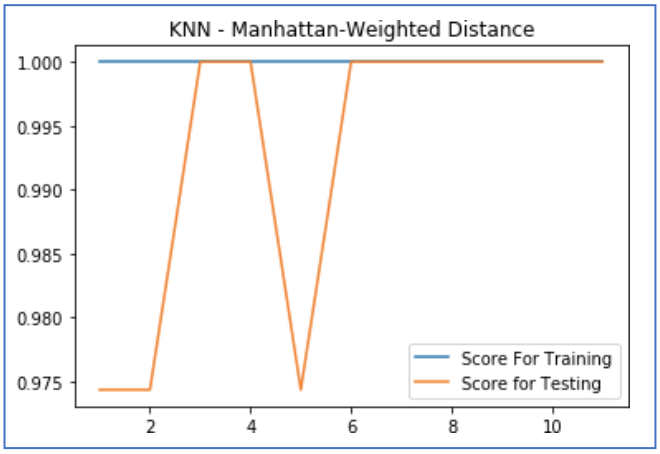

Fig.13 Graph showing variaton of scores for Training & Testing – KNN – Manhattan – Weighted distance

Task-4) Present the ranking you were able to generate in the report and discuss how you could use this knowledge to your advantage. Would it be fair to compare your Random Forest classifier against the winning classifiers from Task 2 and 3? Out of these three classifiers, which algorithm, when run on your data, has a better chance to win this comparison in your opinion and why? How would you implement a fair comparison?

Answer: The random forest algorithm generated produced the following ranking for the features

```
====================================
Ranking of Features by RandomForest
====================================
 1) Alcohol                        0.279086
 2) Malic acid                     0.197170
 3) Ash                            0.140328
 4) Alcalinity of ash              0.116354
 5) Magnesium                      0.073876
 6) Total phenols                  0.060100
 7) Flavanoids                     0.031585
 8) Nonflavanoid phenols           0.026852
 9) Proanthocyanins                0.023341
10) Color intensity                0.014991
11) Hue                            0.013428
12) OD280/OD315 of diluted wines   0.013286
13) Proline                        0.009603
```

Fig.14 Ranking of features produced by Random Forest algorithm

The ranking generated above can be used to discard the features which are not useful & keep only the features which are important. In this way an optimal machine learning model can be built. For example, in the above scores generated, the top 4 seems to be much useful than the rest.

It is not fair to compare random forest classifier to other two classifiers, since the samples used by random forest classifier are different than the samples used by other two classifiers – Decision Tree & KNN. Hence, we cannot compare these three classifiers. Out of these three classifiers if we have to choose one, KNN with K = 2 has a better chance ofwinning, since the score is greater than Decision Tree & it does not overfit as Random forest does for this particular dataset. The following shows the score of how random forest overfits the dataset

```
============================================================
Random Forest with different count of Trees
============================================================
   Count Of Trees  Score For Training  Score for Testing
1           491.0                 1.0                1.0
```

Fig.15 Score produced by Random Forest algorithm, which shows how it overfits the dataset given.

Random forest for a small dataset like this, is not an optimal algorithm as it tends to overfit. A fair comparison can be implemented by using the same dataset & same features to compare the classifier.

Task-5-) Were you able to beat the best results from the tasks above? Yes or no (state clearly), and why? Any ideas that you could investigate if you would have more time or more resources?

Answer: As part of improving the machine learning results, an additional 4 aggregate features are added. Ratios are not opted as derived features as few columns contain zeros as values & using ratios would lead to NaNs.

```
Alcohol
Malic acid
Ash
Alcalinity of ash
Magnesium
Total phenols
Flavanoids
Nonflavanoid phenols
Proanthocyanins
Color intensity
Hue
OD280/OD315 of diluted wines
Proline
Malic acid+Ash
Alcalinity of ash+Magnesium
Magnesium+Total phenols
Nonflavanoid phenols+Flavanoids
```

Fig.16.List of updated descriptive features along with the derived features

After deriving new features, ranking has been done for the features using random forest algorithm. The results are as under:

```
====================================
Ranking of Features by RandomForest
====================================
 1) Alcohol                          0.265208
 2) Malic acid                       0.191126
 3) Ash                              0.132942
 4) Alcalinity of ash                0.103199
 5) Magnesium                        0.077098
 6) Total phenols                    0.057722
 7) Flavanoids                       0.044892
 8) Nonflavanoid phenols             0.023224
 9) Proanthocyanins                  0.021361
10) Color intensity                  0.015513
11) Hue                              0.013775
12) OD280/OD315 of diluted wines     0.010568
13) Proline                          0.010374
14) Malic acid+Ash                   0.010028
15) Alcalinity of ash+Magnesium      0.008398
16) Magnesium+Total phenols          0.007700
17) Nonflavanoid phenols+Flavanoids 0.006873
```

Fig.17.Ranking of all features

From the above ranking presented above, only top 5 features have been selected to build all the classifiers.
The results generated is not greater than the best classifier in the classifiers built previously. To have a fair comparison same dataset was used as well. The following shows that the best classifier has a score for training = 96.20% & Score for testing = 94.87% , which is same as the best KNN classifier seen in Task – 3. Hence, by adding new features, applying ranking & selecting the best features did not improve the model. The following are the results showcasing the same.

```
============================================================
Decision Tree Classifier - With Criterion Entropy-Fair
============================================================
    LevelLimit  Score For Training  Score for Testing
1          1.0            0.924051           0.897436
2          2.0            0.974684           0.948718
3          3.0            1.000000           0.948718
4          4.0            1.000000           0.948718
5          5.0            1.000000           0.948718
6          6.0            1.000000           0.948718
7          7.0            1.000000           0.948718
8          8.0            1.000000           0.948718
9          9.0            1.000000           0.948718
10        10.0            1.000000           0.948718
11        11.0            1.000000           0.948718
```

Fig.18 Score of DCT classifier with varying depths, & criterion Entropy

As seen above, though the best score for Training – 97.46% , score for Testing remains at 94.87% which has been already achieved in Task-3 for KNN classifier. Hence the results did not improve.

Given time, more derived attributes such as Ratios can be used & all the NaNs generated can be replaced with mean of each descriptive feature. Once this task is done, ranking of features can be done again to evaluate which among them are the best features. All the classifiers can be run again to check if the scores have been improved.