MOHAMMED SHOEBUDDIN HABEEB :: 002521737

HOMEWORK 3 FoDS

TASK 1:

From the WineData file provided, we get 13 descriptive features and a target class label, which identifies which category a wine belongs to. The initial description of data is given below in the table:

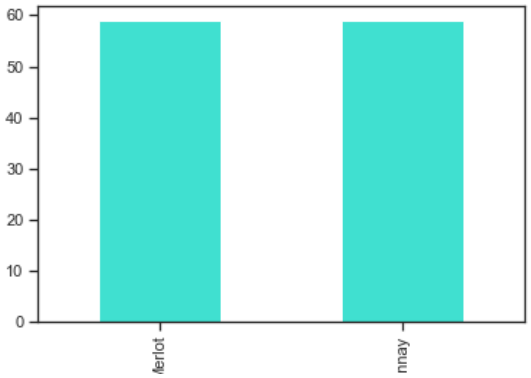
Statistical Summary of Data

	mean	std	min	max	count
Alcohol	13.0	0.9	11.0	14.8	118.0
Malic acid	2.0	0.9	0.7	5.8	118.0
Ash	2.4	0.3	1.4	3.2	118.0
Alcalinity	18.7	3.5	10.6	30.0	118.0
Magnesiu	100.3	14.0	78.0	151.0	118.0
Total pher	2.5	0.5	1.1	3.9	118.0
Flavanoid	2.5	0.7	0.6	5.1	118.0
Nonflavan	0.3	0.1	0.1	0.7	118.0
Proanthoc	1.7	0.5	0.4	3.6	118.0
Color inte	4.3	1.6	1.3	8.9	118.0
Hue	1.1	0.2	0.7	1.7	118.0
OD280/OD	3.0	0.5	1.6	4.0	118.0
Proline	815.8	356.0	278.0	1680.0	118.0

As we see, the size of all the descriptive features is 118 tuples.

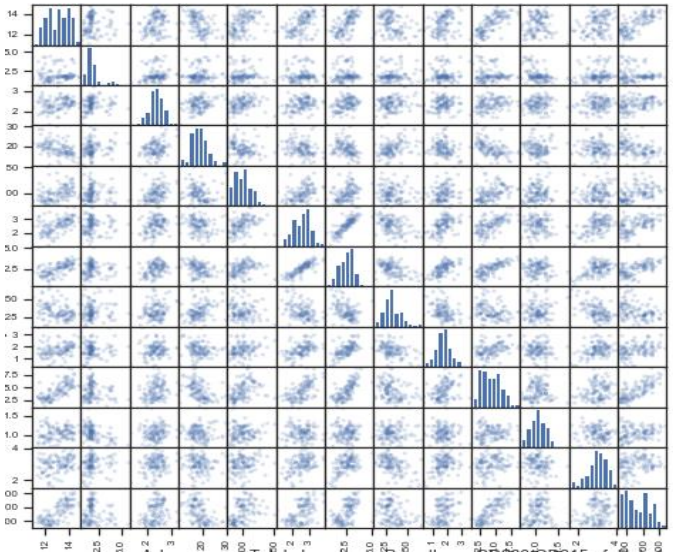
From the Data distribution bar plot of classes, we can notice that both classes are balanced with equal distribution for Chardonnay and merlot

Bar Plot for Class Distribution



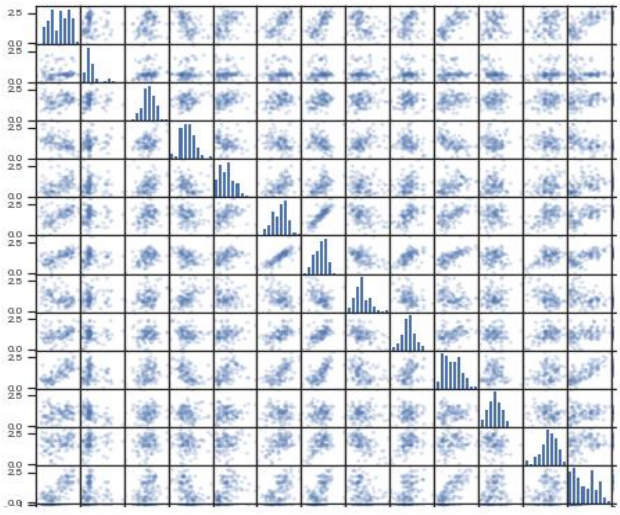
For further insight into the Data, we plot the SPLOM, which gives correlations between the data apart from The Shape of the histogram. Moreover, from the scatter plot matrix shown below, we observe that most of the descriptive features follow normal distribution apart from Proline and Malic Acid which possibly could follow exponential distributions. Furthermore, we also notice most of the attributes have almost no correlation with other features, which will be further investigated in task 5.

SPLOM for Descriptive Features



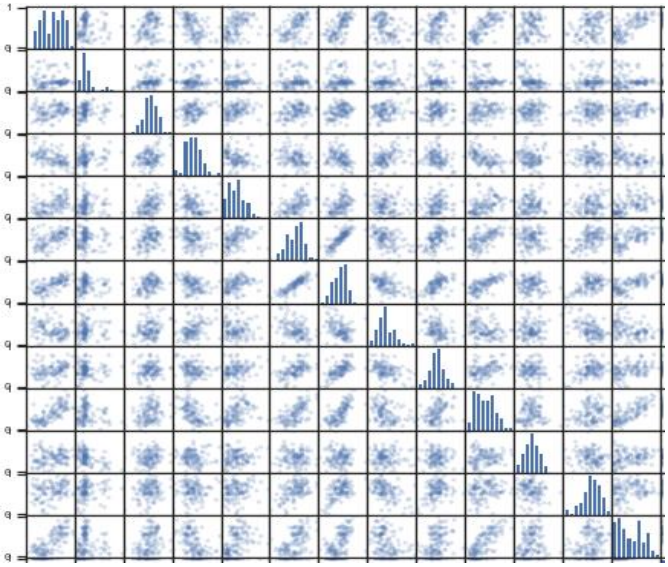
From the SPLOMS of normal distributions for the min-max range [0,1] and [0,3], we notice identical correlation graphs and no major changes. As this normalization respect the proportions between different data points, I don't believe we get much of difference on either scale. The only thing that could affect is the presence of outliers which this dataset doesn't suffer from. I believe the main reason for using a bigger scale is the presence of bigger ranges like proline and Magnesium which would get too compact if compressed in a very small space.

SPLOM after Normalization to [0,1]



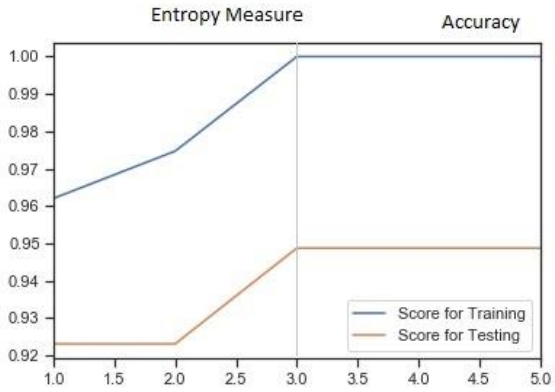
Comparing both scatter plots for ranges [0,1] and [0,3]

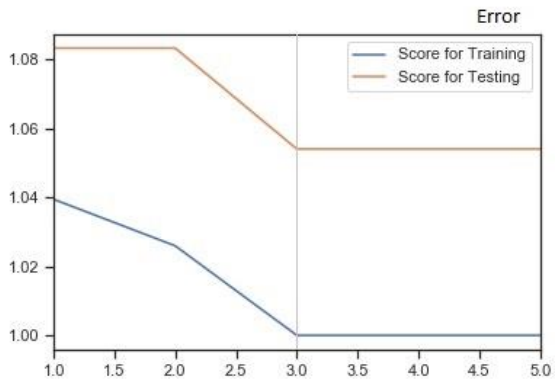
SPLOM after Normalization to [0,3]



TASK 2:

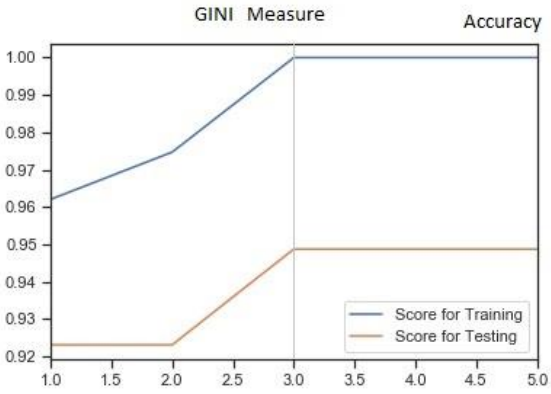
Entropy Accuracy Plot

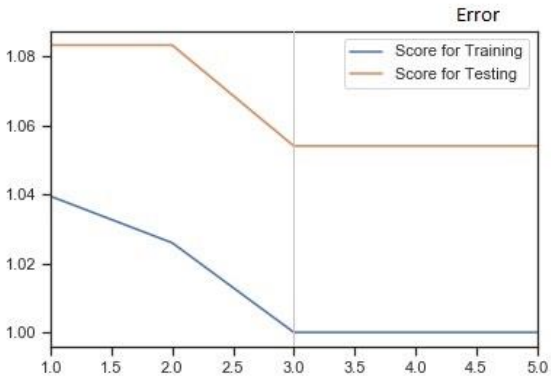




Level	Score for Training	Score for Testing
1	0.962	0.923
2	0.975	0.923
3	1.000	0.949
4	1.000	0.949
5	1.000	0.949

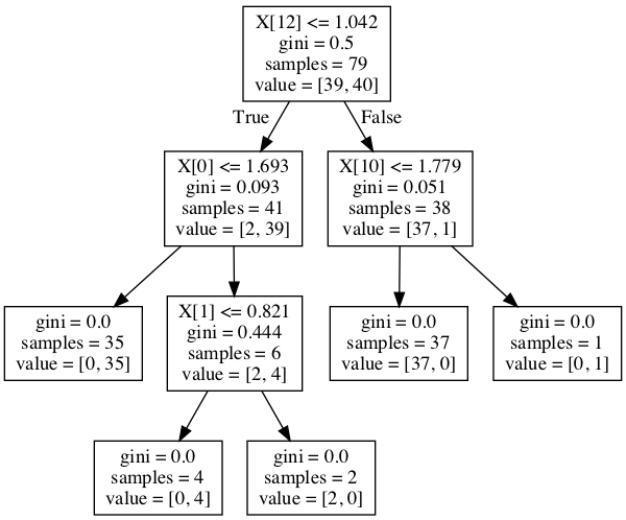
Gini Accuracy Plot





Level	Score for Training	Score for Testing
1	0.962	0.923
2	0.975	0.923
3	1.000	0.949
4	1.000	0.949
5	1.000	0.949

We know entropy as the average or expected uncertainty associated with this set of events, and Gini is the probability that how often a randomly chosen element from that set is incorrectly labeled.



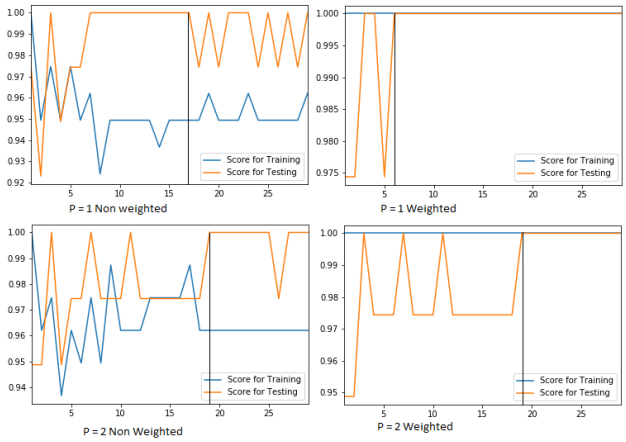
But for the given dataset, we get that both indexes give similar accuracy/error values. I believe that is the case because the biggest Information Gains we take will lead us to similar misclassifications. I believe our dataset could also affect the similarity problem. Using another cross-validation data apart from testing data could probably give more insights into the accuracy.

From the given information, we get our accuracy at 95% for both Gini and Entropy measures. We can prune the other nodes and leaves after node three as they don't give any more meaningful insights. It would be highly recommended to **use the Gini index for this dataset** as it would be **computationally cheaper** than Entropy, as it involves logarithmic functions, which could make it heavier.

TASK 3

KNN Classification

For the given Data set Features, we implement KNN classification and get the following results:



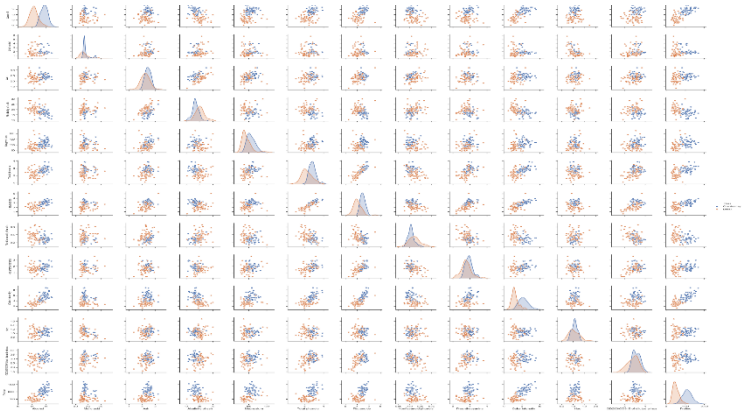
To investigate more about KNN, we will look first into a case where I am considering the possibility of 30 nearest neighbors. Since KNN thrives on more neighbors to produce a smooth boundary and overfits on lower K values, **we will always prefer a higher K value** for a set of similar test results.

From the figures given above, we notice that the accuracy graphs are wriggly; this is major because of the presence of overlapping classes shown in pair plot and an orthogonal dataset to feature set. From the pair plot, we notice that none of the features have circular or box distribution, which would help us decide Euclidian or Manhattan similarity approach. I would approach this dataset with a p-**value greater than two** or cosine similarity index. We tend to select a classifier to make sure that it remains stable or at least follows a similar trend. All the figures above give 100% accuracy which is an advantage over decision tree classifier. My winning classifier would be a Euclidian Classifier, which is weighted, i.e., at (2,2) in my figure above. I rejected other measures because they are not consistent with the training values and testing values.

The reason for not choosing smaller k values over larger k values is because smaller k values tend to generalize the data, and Larger values give us better results as they are more preferable. Moreover, since our classes are **balanced,** we don't necessarily have a problem where larger k value would give **imbalanced errors** in the classes. Hence, K value **can be taken as big as possible,** but for computational and feasible cases and considering all the reasoning above, I have decided K value will be 19 for the Euclidian Weighted KNN classifier.

I will work on choosing different P values and using different similarity measures like Cosine or Mahalanobis distance in Task 5 for better KNN classifier.

PAIRPLOT FOR THE WINE DATASET



K Value	Score for Training	Score for Testing
1	1	0.949
2	1	0.949
3	1	1.000
4	1	0.974
5	1	0.974
6	1	0.974
7	1	1.000
8	1	0.974
9	1	0.974
10	1	0.974
11	1	1.000
12	1	0.974
13	1	0.974
14	1	0.974
15	1	0.974
16	1	0.974
17	1	0.974
18	1	0.974
19	1	1.000
20	1	1.000

SCORE CHART FOR WINNING KNN CLASSIFIER

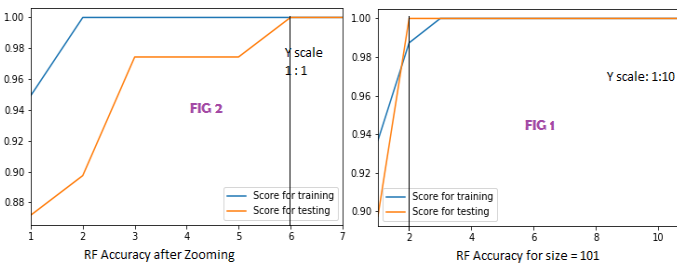
TASK 4

Feature_SORTED	Importance_Sorte
Proline	0.279
Alcohol	0.197
Color intensity	0.140
Flavanoids	0.116
Total phenols	0.074
Magnesium	0.060
Malic acid	0.032
Alcalinity of ash	0.027
Ash	0.023
OD280/OD315 of dilu	0.015
Proanthocyanins	0.013
Nonflavanoid phenol	0.013
Hue	0.010

This table gives an idea of feature importance of the feature, which shows heavy dependency on **Proline, Alcohol, and Color intensity,** whereas the lower rows of the tables show the lower dependency of these features. These can be used to modify our dataset to choose features selectively. The main advantage of using this important measure is it gives insights into the dataset and help us understand which features are driving our decision trees and the subsequent Random forest distribution. We can again classify this **importance based on subgroups, which are probabilities** of the importance of a subset of features.

The mean decrease in impurity importance of a feature is computed by measuring how effective the feature is at reducing uncertainty (classifiers) or variance (regressors) when creating decision trees within RFs. The problem is that this mechanism, while fast, does not always give an accurate picture of importance.

Using the dataset and after applying random Forest distribution, we get the following decision tree scores on testing set and a plot, which represents how our accuracy varies when we alter the size of trees from 1 to 102. It always preferable to have a **larger number of trees** for better accuracy, but that comes with a computational cost that should be balanced while making a final decision.



Accuracy Plot for Random Forest

As we can see from the Accuracy plot for the random forest, Fig 2 is the distribution when the size of trees if 102, and the maximum depth is 3 for our decision tree, whereas Fig 1 is the distribution when the size of trees is 13 at a similar depth. We can choose a larger size of the tree for better accuracy as the randomness increases which in turn increases the accuracy of our random forest classifier, but that **will incur a higher computational cost**. For this dataset, I am choosing Fig 1 as it gives the same result when the **size of the tree is 11**. This is equivalent to the score we got with the KNN classifier. This accuracy score is shown below:

Count of Trees	Score for training	Score for testing
1	0.949	0.872
3	1.000	0.897
5	1.000	0.974
7	1.000	0.974
9	1.000	0.974
11	1.000	1.000
13	1.000	1.000

It would not be fair to compete with the Random forest with the KNN classifier. KNN is robust to noisy training data and is **more effective** in case of a large number of training examples. The decision tree, on the other hand, can and should be compared to a random forest. Random Forest is nothing more than a bunch of randomized Decision Trees combined. They can handle categorical features very well. This algorithm **can handle high dimensional spaces** as well as a large number of training examples. Random Forests can almost **work out of the box,** and that is one reason why they are very popular.

Comparison of all the three classifiers i.e., Decision Tree(95%) , KNN (100% for k = 19) and Random Forest(100% for size of tree = 11) would show that we should **prefer Random forest** for this dataset mainly because it wins over decision tree as it is an **ensemble of better-randomized decision trees** , it would win over KNN as the RF is computationally easier and KNN gives very **unstable orthogonal splits** to the datasets which cannot be relied on. Moreover, as our dataset has only 118 balances, tuples KNN would be not a good choice as it would make more sense to use on bigger datasets.

Thus decision tree and Random forest could be compared on accuracy measure, whereas Random Forest and KNN can be compared on computational cost and size of the dataset. To bring them on a fairer scale, we should modify our KNN(better similarity measures) and decision trees (Better ensembles) to compete for the Random forest.

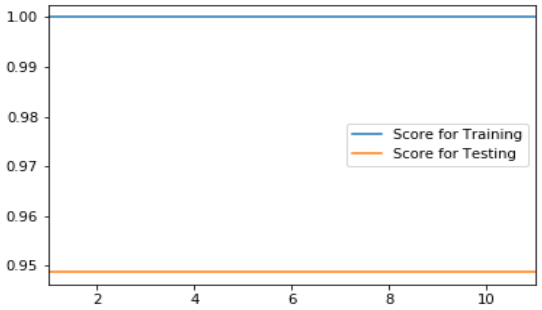
TASK 5

Modifications on Decision Trees :

ADABOOST

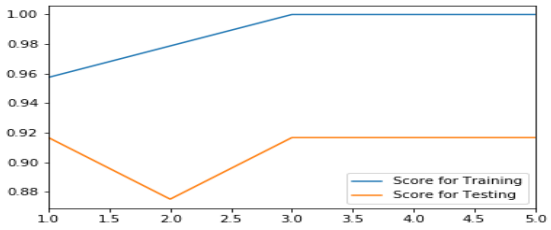
We implement the Adaboost algorithm ensemble with a max depth of trees as three we get a consistent result of 95% accuracy which is way off from the values we got in a random forest. We show the accuracy plot below for the same :

Count(Trees)	Score for Training	Score for Testing
1	1	0.94871794
11	1	0.94871794
21	1	0.94871794
31	1	0.948717949
41	1	0.948717949
51	1	0.948717949
61	1	0.948717949
71	1	0.948717949
81	1	0.948717949
91	1	0.948717949
101	1	0.948717949



Another approach I did was **reducing the size** of stratified testing and training data sets from previous (2/3:1/3) to (80:20). This might lead to **generalizations** occurring due to the presence of a bigger training set, but I believe testing data in our previous case can be reduced as there are not many tuples in question.

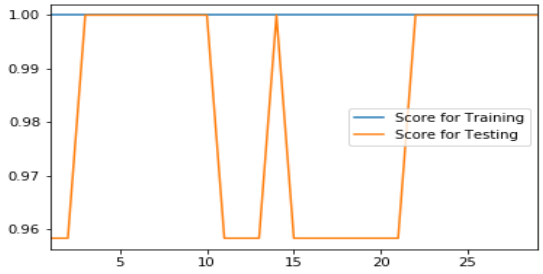
Below is the accuracy plot for the decision tree on a different test sample.



As we can see, this misclassifies heavily and cannot help much as the **accuracy is 92%** at its best.

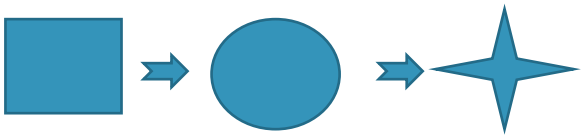
Modifications on KNN:

We implement the KNN classifier with the modified dataset (80:20) for the accuracy plot to show better results than the one we initially got from previous implementations.



I then tried KNN classifier on the original dataset but with **p = 4 and 5,** and interestingly the graph showed more promise as it was able to **dissect the orthogonal dataset** more consistently. I believe because the shape of our classification has become more of star-ish than euclidian or Minkowski which are circular or rectangular. The accuracy plot for p = 5 and p = 4 are shown below for comparison.

Probable shape changes when p-value increases:

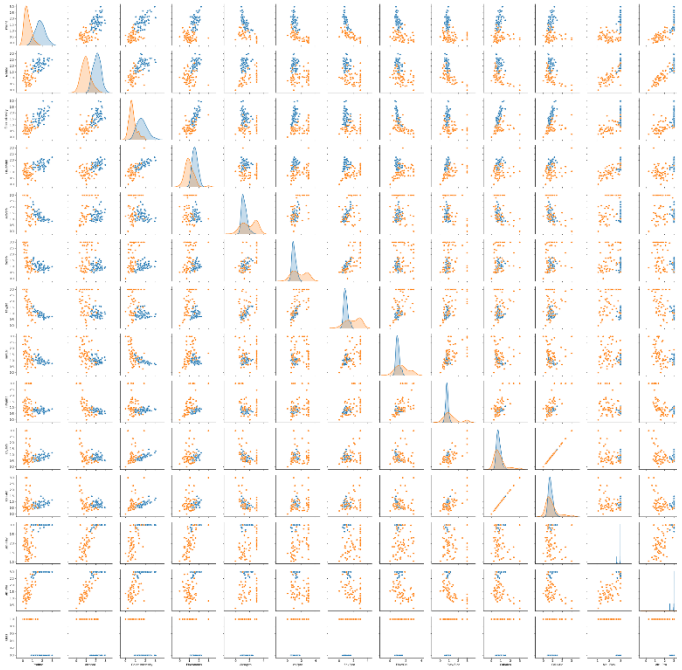


With these results, I would prefer choosing **p = 4 at k 19 which has 100% accuracy**, as it has more k values and comparable accuracy with the result from task 3.

Derived Attributes :

	min	max	mean	std
Proline	0	3	1.151	0.762
Alcohol	0	3	1.573	0.704
Color inte	0	3	1.201	0.642
Flavanoid	0	3	1.305	0.481
Alco/pro	0	3	1.630	0.820
col/pro	0	3	1.345	0.801
flav/pro	0	3	1.472	0.866
Flav/Col	0	3	1.288	0.651
Flav/Ã¡lco	0	3	0.970	0.586
Col/alco	0	3	0.833	0.474
Col+Alc	0	3	0.833	0.474
Alc+Flav	1.057895	3	2.482	0.642
Alc+Pro	0.276034	3	2.243	0.838

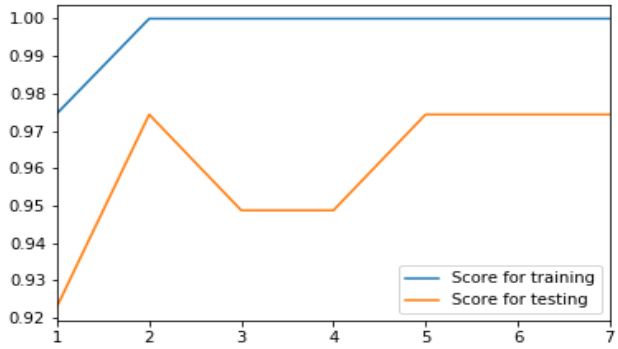
We use **division and aggregates to make new data** on our four most important sets of features from the importance plot as I believe modifying less important data **won't help much** going forward.



As we see from the pair plot and feature importance. Especially the diagonal of pair plot, we get a fairly **independent cluster of data** amongst the new nine derived attributes. I will drop all unnecessary derived features(**color except green**) as they don't perform any better than our previous original features and keep the all derived features above Flavanoids, as shown in the table below. **Feature importance** plays a massive role in my conclusions as I believe finding the strength of the impact of features on the class values could massively affect our analysis.

Feature	Importance
Alc+Pro	0.2162
Proline	0.2122
Alc+Flav	0.1454
Alcohol	0.1295
Color intensity	0.0883
Flavanoids	0.0746
flav/pro	0.0311
Flav/Col	0.0259
Alco/pro	0.0211
col/pro	0.0191
Flav/Ã¡lco	0.0174
Col/alco	0.0098
Col+Alc	0.0095
Col+Alc	0.0095

Now we plot to compare our winning Classifier from all the tasks, i.e., Random Forest with a new resultant random Forest classifier from the above modifications



As we can see, this **fails to beat** our original data Random Forest which goes on to explain that having a better feature to Class dependability is not necessarily implied with better classifier performance.

In the end, I would have preferred to perform NCA, LCA analysis to identify feature dependability, and also, if time and complexity would have permitted, I would have implemented equal frequency binning to understand the granular dependency furthermore and see how dimensionality changes accordingly.

REFERENCES

1. John D. Kelleher, Brian Mac Namee, and Aoife D'Arcy, "Fundamentals of Machine Learning for Predictive Data Analytics", London.

2. Slides by Dr. Rafal Angryk on Fundamentals of Data Science, Georgia State University

3. Slides by Tan, Steinbach, Kumar on Introduction to Data Mining, Indiana State University

4. Slides by Brendan Blake Camp on Data Mining, Georgia State University

5. A visual introduction to Machine Learning - http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

6. Neighbourhood Components Analysis Jacob Goldberger, Sam Roweis, Geoff Hinton, Ruslan Salakhutdinov Department of Computer Science, University of Toronto

7. Online resources include Wikipedia and Stack Overflow