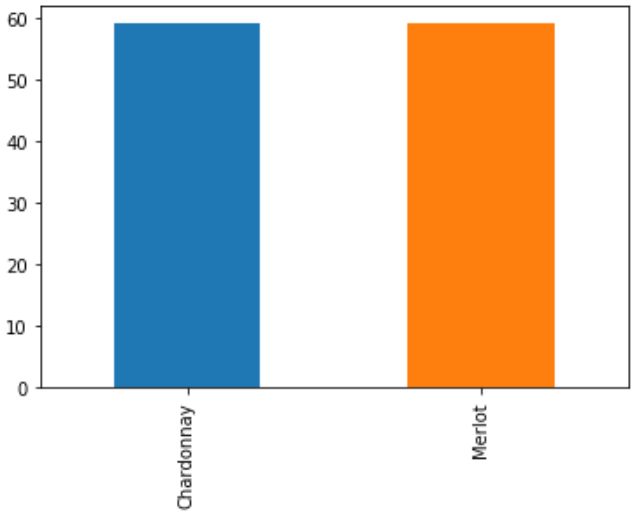
HW3

Name: Yukang Xu

Panther ID: 002462280

I. TASK 1

We have 118 instances and 13 descriptive features in this dataset. Descriptive feature Class is binary feature while the other features are numeric features.
After we map the dataset, we get the graph below. Because the number of instances in two group are the same, our data are balanced.



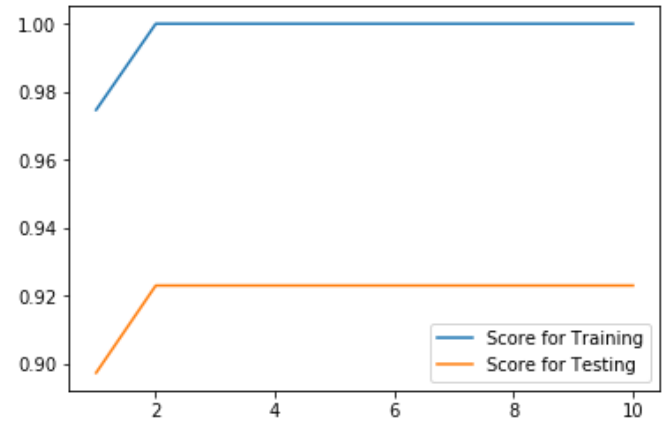
For data visualization, range may be important. But for machine learning procedure, the range of normalization is not that significant.

II. TASK 2

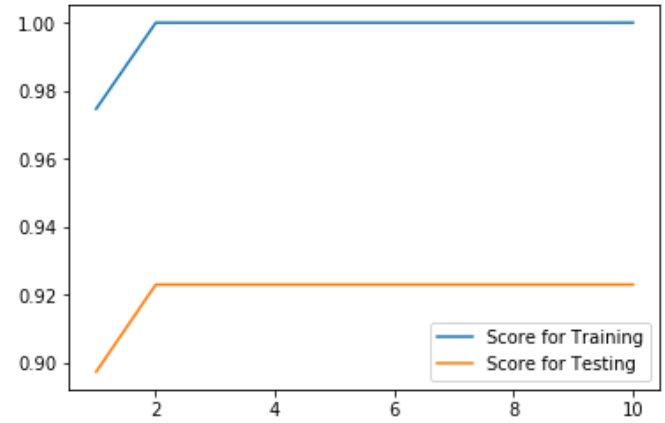
Firstly, I split the dataset into training data and testing data. I provide their information below.

dataset	proportion	instances
train data	0.666667	79
test data	0.333333	39

For decision tree algorithm, we apply entropy and Gini index with tree level 1 to 10, the score charts as following:

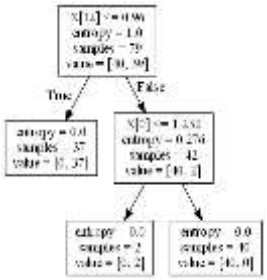


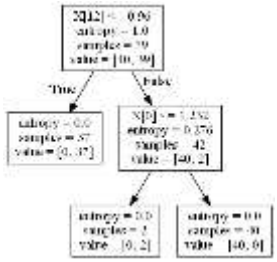
(Using Gini index)



(Using entropy)

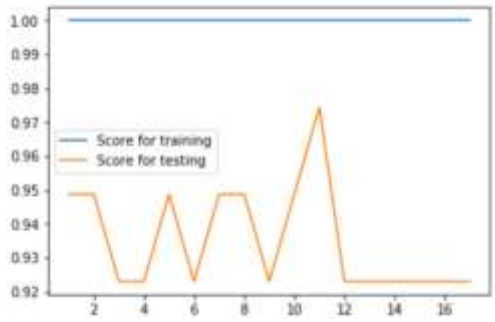
We find that entropy with tree level equal to 1 and Gini index with tree level to 1 give us the best classified accuracy (92.3%) toward testing dataset. The tree plot as following:



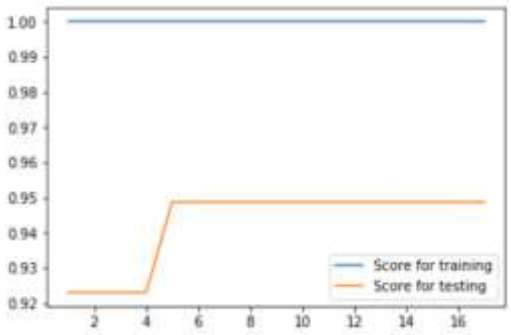


III. TASK 3

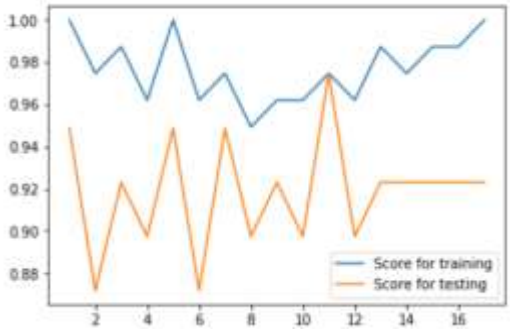
For KNN algorithm, we use Manhattan and Euclidean as distance measure with uniform weight or distance weight, and try k = 1 ~18 to get the score charts as following:



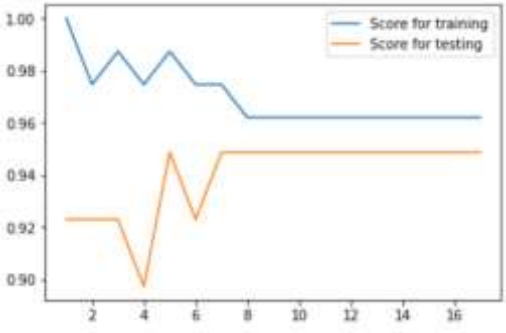
(Manhattan distance with distance weights)



(Euclidean distance with distance weights)



(Manhattan distance with uniform weights)

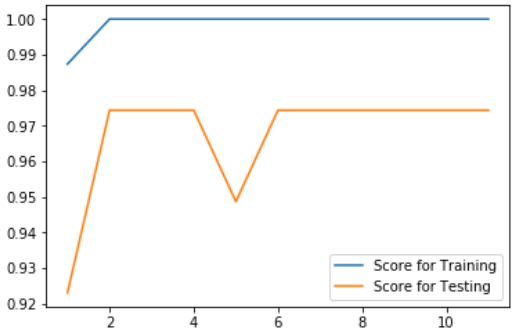


(Euclidean distance with uniform weights)

We find that Manhattan distance with uniform weight, k =11 and Manhattan distance with distance weight, k = 11 give us the best classified accuracy (97.4%) toward testing dataset.

IV. TASK 4

For the random forest algorithm, we try to use number of estimator = 1 ~ 102 to get the score chart as following:



We find when number of estimator =11, we get the best classified accuracy (97.4%) toward testing dataset.

And the feature ranking table as following:

(1)Alcohol	0.246878
(2)Malic acid	0.191926
(3)Ash	0.143436
(4)Alcalinity of ash	0.139236
(5)Magnesium	0.103365
(6)Total phenols	0.083723
(7)Flavanoids	0.023395
(8)Nonflavanoid phenols	0.023029
(9)Proanthocyanins	0.013248
(10)Color intensity	0.010798
(11)Hue	0.010252
(12)OD280/OD315 of diluted wines	0.005805
(13)Proline	0.004910

According to this table, we can know that which descriptive features might be more importance when we do classification. It provide a reference as we want to implement our algorithm.

Comparison

It's fair to compare random forest classifier against the winning classifiers from task 2 and task 3. Both tree algorithm use the same training dataset to make classification. Although they have different way to learn the dataset and make predicted class, it's clear when we see the predicted accuracy toward testing dataset and know which algorithm is more fit to this data.

According to all the classified result in task 2, task 3 and task 4, we can find that KNN(with Manhattan distance, k = 11, uniform or distance weight) and Random Forest has the best result (97.4%). Also, when we observe the score charts, we can

find that KNN algorithm and Random Forest actually performs better than decision tree. As a result, we think KNN and Random Forest has better chance to win this comparison.

I. TASK 5

A. Drop features

From Task 4 we have an importance ranking. Features which have less than 1% importance are dropped. So we drop Proanthocyanins, Color intensity, Hue, OD280 and Proline. Here are the results after change.

Algorithm	Before	After
Decision tree (entropy)	92.3%	95.0%
Decision tree (Gini index)	92.3%	95.0%
KNN (Manhattan, uniform weight)	97.4%	92.5%
KNN (Manhattan, distance weight)	97.4%	92.5%
KNN (Euclidean, uniform weight)	94.9%	90.0%
KNN (Euclidean, distance weight)	94.9%	90.0%
Random Forest	95.0%	95.0%

According to the table, we can see that Decision Tree and Random Forest gets the best result after implement.

B. Generate new features

In wine dataset, Ash and Alcalinity of Ash, Total phenols and Nonflavanoid phenols are related to each other. So we create two new columns which are called Ash+Alca of Ash and Total Phenols+Non Phenols. Here are the results after change.

Algorithm	Before	After
Decision tree (entropy)	95.0%	95.0%
Decision tree (Gini index)	95.0%	95.0%
KNN (Manhattan, uniform weight)	92.5%	87.5%
KNN (Manhattan, distance weight)	92.5%	87.5%
KNN (Euclidean, uniform weight)	90.0%	87.5%
KNN (Euclidean, distance weight)	90.0%	87.5%
Random Forest	95.0%	100%

when k = 9, the accuracy of random forest reaches 100%. Therefore, we beat the best results from the tasks above (KNN, k = 11, Manhattan distance, uniform or distance weight : 97.4%).

Idea

For deriving new descriptive features, actually we try a lot of possible combinations, but Random Forest gives us better results, that's why we apply it as new descriptive feature. If we know more about this dataset, we may give better algorithm to improve it.