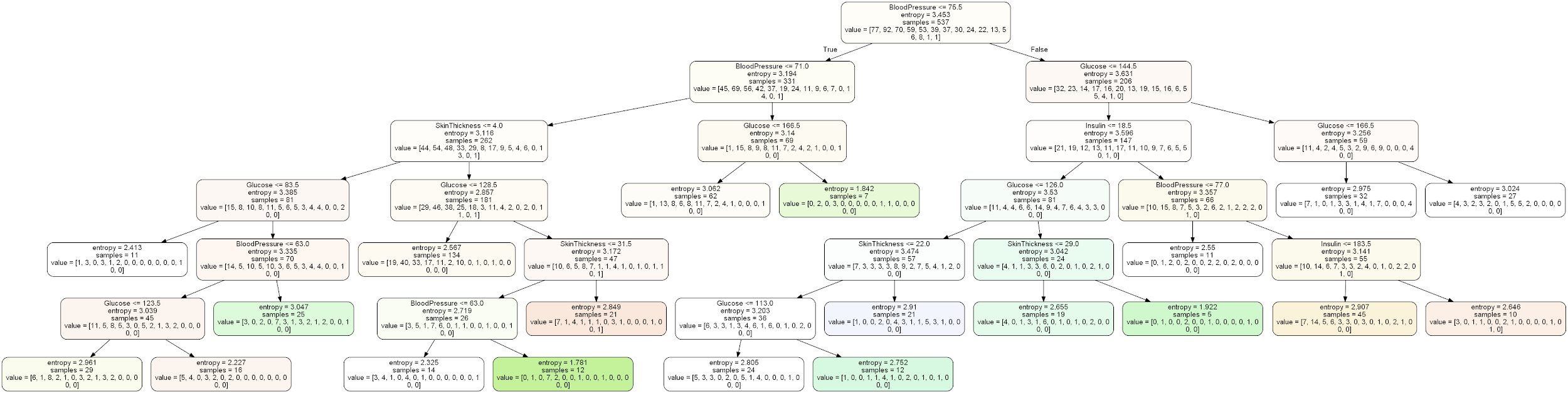
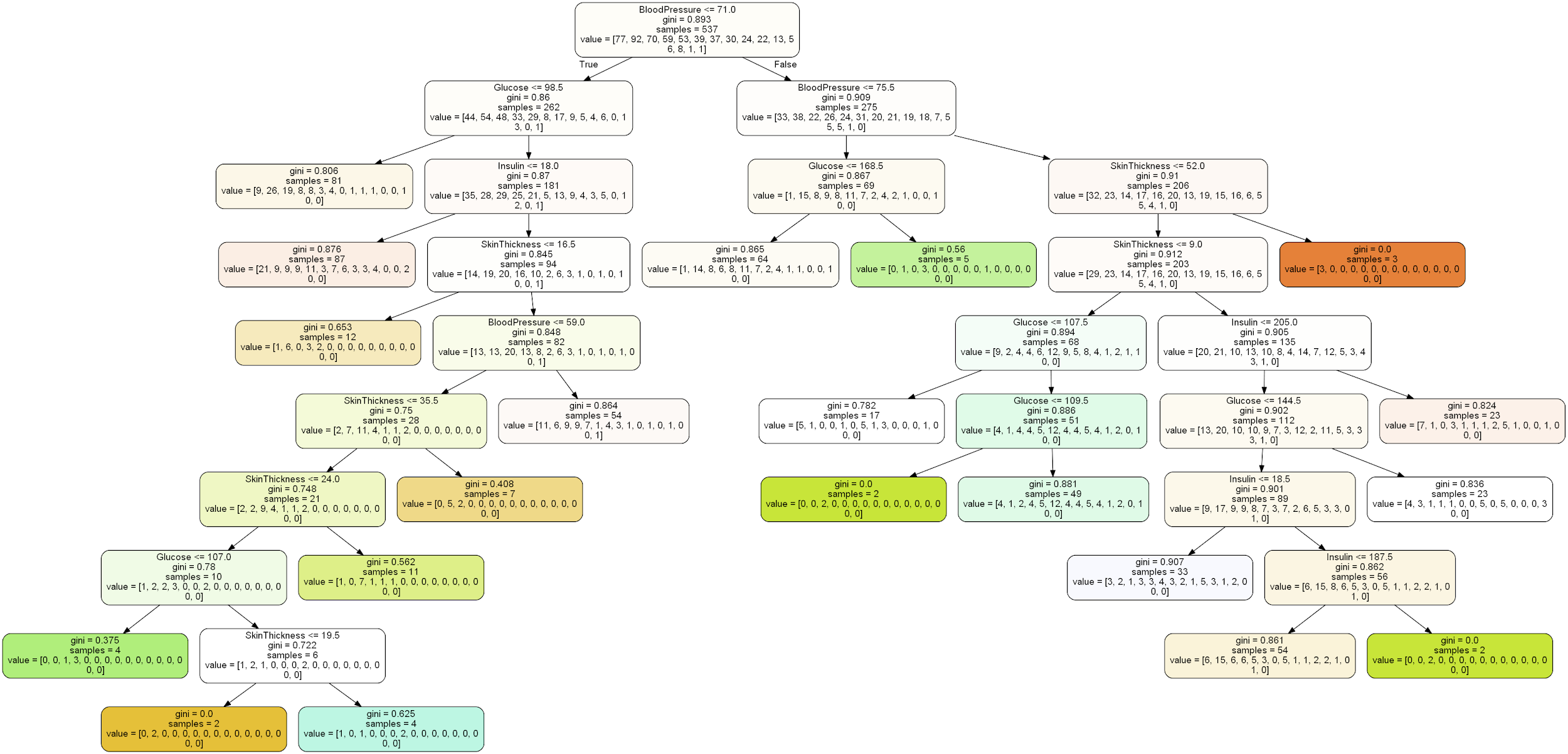
**Homework 2**

Create 3 decisions tree model with 20 or less nodes using the cleaned dataset (the one in hw1).

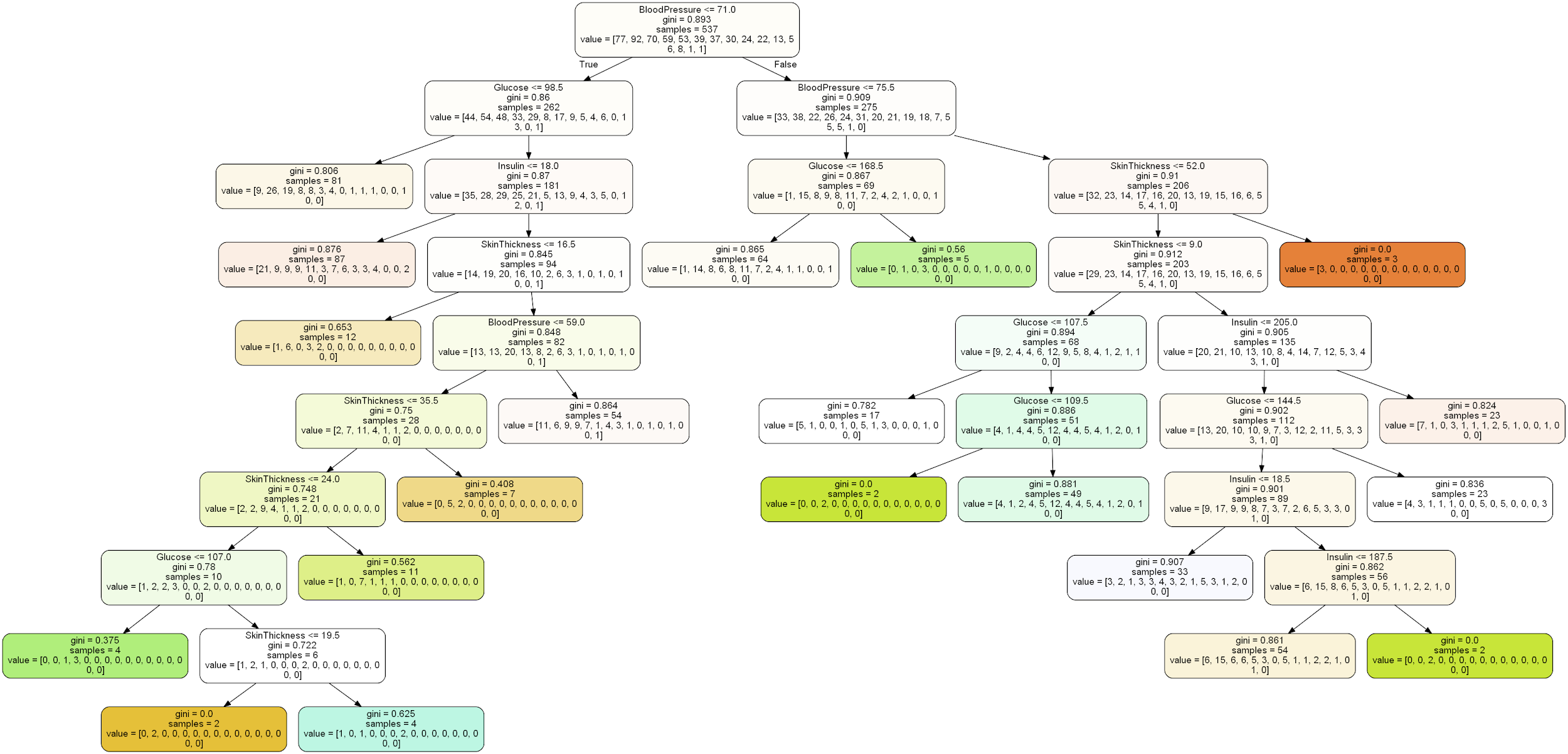
Entropy



Gini



Random Forest



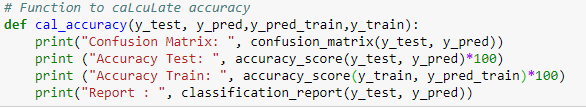
Using the 3 decision trees above. Explain how these decision trees were obtained. Report the training accuracy and the testing accuracy for each decision tree. Interpret the learnt decision tree. What do they tell you about the importance of the 8 continuous attributes for the classification problem?

Gini works by measuring the impurity of the node. The gini is calculated using by summing the probability of mistakenly categorizing a data item. The gini reaches zero if all the data are in one category. Gini will put target variable in two bin a 1 or a 0 and make a binary split. Higher gini value means higher homogeneity.

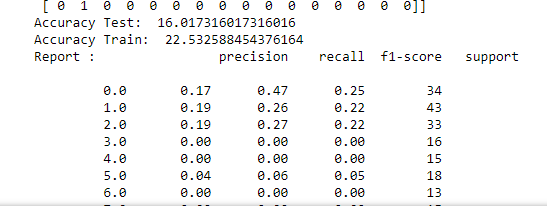
The entropy classifier tries to find the information gained by taking entropy of parent - weighted average\* entropy of children. The parent node is calculated and the individual nodes are calculated. The node is split where is has the least entropy compared to the parent node.

The random forest works will make multiple trees opposed to single trees like CART. This works by training and sample data, which is trained on a growing tree. The best split in chosen and then held constant as multiple trees grow. Each tree is grown to the largest without pruning.

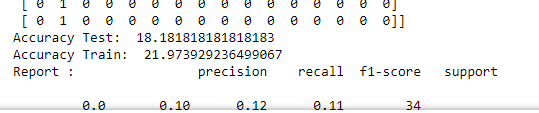
New data is then added to all the trees will take the tree with the majority votes or classification and average for regression.



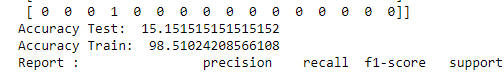
GINI



ENTROPY



RANDOM FOREST



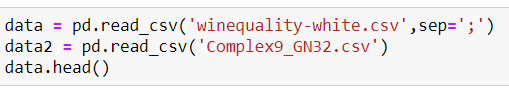
The decision trees were obtained using sklearn library for gini, entropy and

random forest.

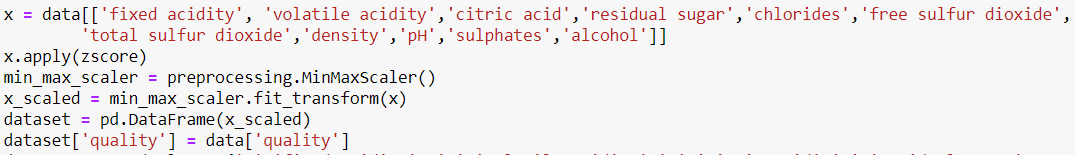
Interpret the learnt decision tree. What do they tell you about the importance of the 8 continuous attributes for the classification problem?

All the decision trees has blood pressure as their starting node and it looks like there is where the most information could be uncovered. All the other 8 attributes play an important role in making the decision tree. Selecting the right attribute will greatly affect the performance of the decision tree in time complexity. Have a complicated tree will increase the time to predict a new data set.

1. Download the original White Wine Quality dataset and the Complex9\_RN32 dataset.



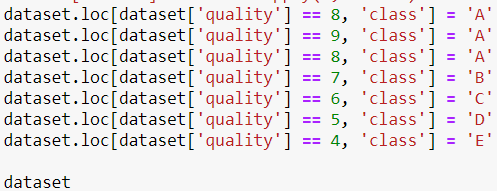
1. In the White Wine Quality dataset, normalize the first through 11th attribute into z scores



1. In the White Wine Quality dataset, keep the 12th attribute “quality” as is.



1. In the White Wine Quality dataset, introduce a new ordinal attribute called class, which will be attribute 13, based on the value of the 12th attribute “quality” as follow: 10 through 8 = A, 7 = B, 6 = C, 5 = D, 4 = E. Keep in mind, when clustering the dataset only the first 11 attributes will be used; attributes 12 and 13 will be used to evaluate the quality of a clustering result.

****

For the following 3 python functions, the following test cases must be used

1st case Let a, b, c be the following vectors:

a=(0,1,1,1,1,2,2,3)

b=(A,A,A,E,E,D,D,C)

c=(8,8,8,4,4,5,5, 6)

entropy(a,b)=4/7\*H(0.5,0,0,0,0.5)+ 2/7\*0 + 1/7\*0=4/7

ordinal-variatio+n(a,b)=4/7\*(16/6)+0+0=32/21

variance(a,c)=4/7\*(16/3)+0+0=64/21 note: 4/7\*variance(8,8,4,4)+0+0

2nd case

a=(1,1,1,0,0,2,2,2)

b=(A,A,A,E,E,D,D,C))

c=(8,8,8,4,4,5,5,6))

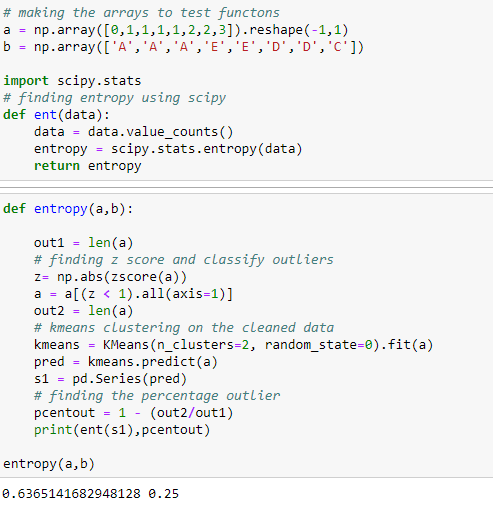
entropy(a,b)= ½\*0+1/2\*H(0,0,1/3,2/3,0)=1.37/3=0.456

ordinal-variation(a,b)= ½\*0 +1/2\*(2/3)=1/3

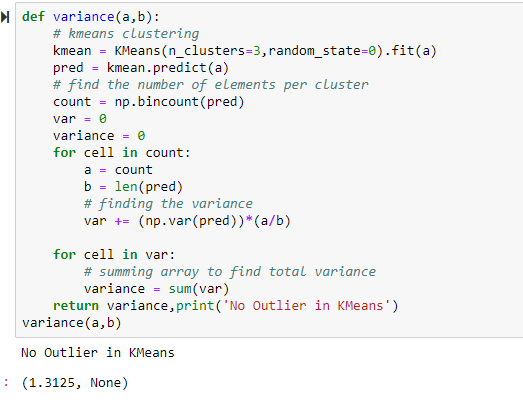
variance(a,c)= ½\*0+1/2\*((2\*(1/3)\*\*2)+(2/3)\*\*2)/2)=1/6 note: 0+1/2\*variance(5,5,6)

1. Write an Python-function entropy(a,b) that computes the entropy and the pecentage of outliers of a clustering result based on an apriori given set of class lables, where a gives the assignment of objects in O to clusters, and b contains the class labels of the examples in O. The entropy function H is defined as follows:

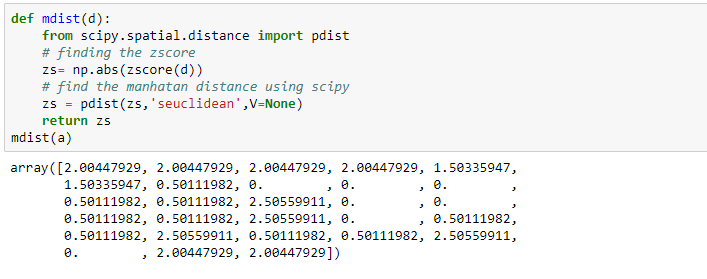
Assume we have m classes in our clustering problem; for each cluster Ci we have proportions pi=(pi1,…,pim) of examples belonging to the m different classes (for cluster numbers i=1,..,k); the entropy of a cluster Ci is computed as follows: H(pi)= Σj=1 (pij\*log2(1/pij)) (H is called the entropy function) Moreover, if pij=0, pij\*log2(1/pij) is defined to be 0. The entropy of a clustering X is the size-weighted sum of the entropies on the individual clusters: H(X)= Σr=1 (|Cr|/|Σp|Cp|)\*H(pr) In the above formulas ”|…|” represents the set cadinality function4 . Moreover, we assume that X={C1,…,Ck} is a clustering with k clusters C1,…,Ck, You can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters; therefore, you should ignore cluster 0 and its instances when computing H(X). The entropy function returns a vector: (<entropy>, <percentage\_of\_outliers); For example, if the function returns (0.11, 0.2) this would indicate that the entropy is 0.11, but 20% of the objects in the dataset O have been classified as outliers.

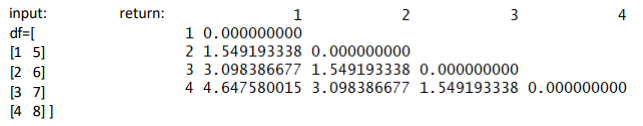
****

1. Write an Python-function ordinal-variation (a,b) that computes the original agreement of a bag b of ordinal classes associated with the instances of clusters given by a—the orignial classes are named A, B, C, D, and E in the WWQ dataset. It is defined as follows: Let φ be definied as follows: φ(A)=4, φ(B)=3, φ(C)=2, φ(D)=1, φ(E)=0, If o is an object in the WWQ dataset, o.class denotes the value of the 13th attribute of o (which takes values A, B, C, D or E) Let C be a cluster of WWQ objects, then the ordinal agreement in C5 is defined as follows: Ordinal-variation(C) = (Σc,c’∈C and c≠c’ |φ(c.class)−φ(c’.class)|)/(|C|\*\*2-|C|)) If |C|=1 then Ordinal\_variation(C)=0 In the above formulas ”|…|” represents the set cadinality function. Moreover, assuming X={C1,…,Ck} is a clustering consisting of k clusters C1,…,Ck, Ordinal-variation(X) is the number of instances weighted sum of Ordinalvariation(C1),…,Ordinal-variation(Ck); that is: Ordinal-variation(X)= Σr=1 (|Cr|/|Σp|Cp|)\*Ordinal-variation(Cr) However, we give X in the form of (a,b) where a gives the assignment of objects in O to clusters, and b is class variable assoicated with each object in O. Again, ignore all instances of cluster 0 from ordinal agreement computations, as those examples represent outliers.
2. Write an Python-function variance(a,b) which computes the variance of the clustering result X based on an apriori given set of numerical observations—one numerical observation is associated with with each object, where a gives the assignment of objects in O to clusters, and b is the numerical observation associated with each object in O. The variance of a clustering is the weighted sum of the variance6 observed in each cluster with respect to the numerical variable. The observed cluster variance is weighted by number\_of\_example\_in\_the cluster/total number of examples in all clusters; the same way how variance is assessed by regression tree learning algorithms. In general, the function variance returns a vector: (<variance>, <percentage\_of\_outliers). If the used clustering algorithm supports outliers, outliers should be ignored in variance computations; You can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters. For example if the function variance returns (2.8, 0.3) this would indicate that the variance of the evaluated clustering is 2.8 and that 30% of the objects in the clustered dataset are outliers. If cluster 0 does not exist, assume that there are no outliers. (Use numpy.var)

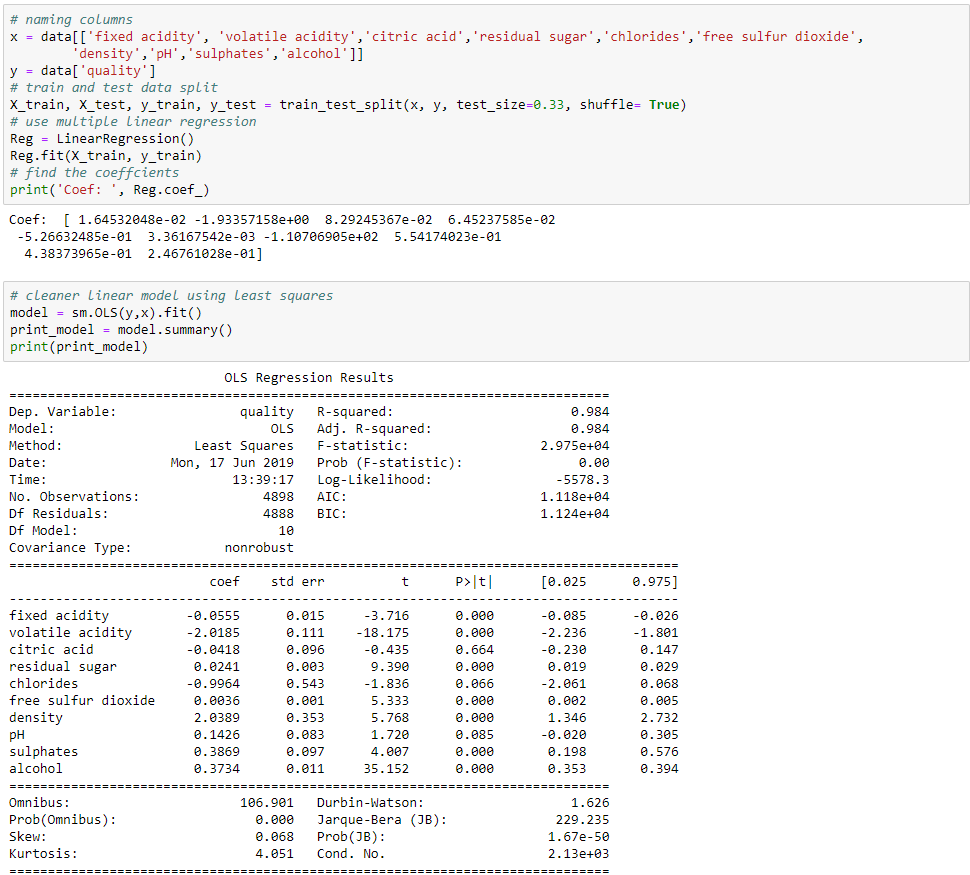


1. Write an Python-function mdist(d) that takes a dataframe d containing only continous attributes as its input, transforms the attribute values in d into z-scores, and then returns a distance matrix7 of the Manhattan distances of the objects in the z-scored dataframe as its result.



****

1. Learn a linear model that predicts the 12th attribute using the first 11 attributes for the Wine Quality data set. (Use linear\_model from sklearn)



1. Interpret the obtained coefficients of the linear model obtained in question 9 and access its quality of the obtained regression function and the importance of the 8 attributes. Also, compare this task’s finding with the findings of the previous task.

From the linear model the volatile acidity has the most impact on the coefficient in the negative direction meaning that it will go down in response of the predictor while holding everything else constant. In the positive direction the density will change the most in response to the predictor variable while holding everything else the same. The importance of the 8 variable is assigned the quality of a wine and from my observation the acidity and chlorides are the most impactful in the negative direction meaning these values will be lower for better wine.