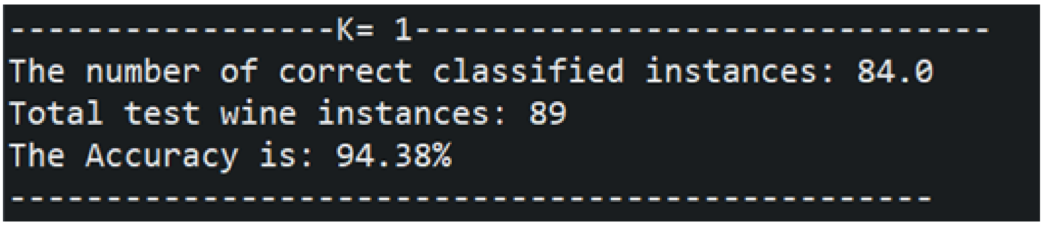
Yun Zhou, 300442776

Name : Yun Zhou
ID: 300442776

This report contains part1, part2, part3, all parts. You can either read this, or read separately inside the corresponding part folder.

Part 1: KNN

1. Report the class labels of each instance in the test set predicted by the basic nearest neighbour method (where k=1), and the classification accuracy on the test set of the basic nearest neighbour method. Make sure you keep the same order as in the test set file.



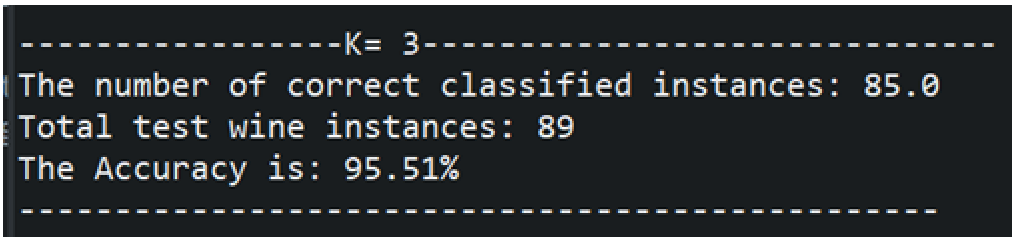
```
-----------------K= 1------------------------------
The number of correct classified instances: 84.0
Total test wine instances: 89
The Accuracy is: 94.38%
------------------------------------------------
```

The above screenshot is from my application part 1.

As we can see that in the case of the basic nearest neighbour which is when K = 1, there are 84 correct classified instances out of 89 test wine instances, the accuracy is 94.38%.

(p.s. The full log that reports the guess class label of each test Wine instance, whether each test Wine has been labelled correctly or not is at the end of this report. Also, you can see it in **sampleoutput_part1.txt**)

2. Report the classification accuracy on the test set of the k-nearest neighbour method where k=3, and compare and comment on the performance of the two classifiers (k=1 and k=3).



```
-----------------K= 3------------------------------
The number of correct classified instances: 85.0
Total test wine instances: 89
The Accuracy is: 95.51%
------------------------------------------------
```

As we can see, when the number of K = 3, the accuracy of my KNN is 95.51%(85/89), the correct classified instances are 85 out of 89.

Compare to the performance from basic nearest neighbour(K =1), we can see that: the correct classified instance is increased from 84 to 85, so that the accuracy is creasing from 94.38% to 95.51% as well.

Although the correct classified number only increase by 1, the accuracy is still increasing not decreasing, we have reason to believe that my KNN can also run pretty well when the k is increasing,Wine test set and the accuracy is around 95%.

3. Discuss the main advantages and disadvantages of k-Nearest Neighbour method.

The KNN is the classification method which depends on the similarity/distance of the features, when do the regression which is classify which class is this test instance belongs to, just sort the train list then find K's number of nearest neighbors, pretty easy to use.

Therefore, the main **advantage** is that: it is pretty *easy to use as well as it is pretty easy to be constructed*, the logic and the principle can be easy to understand. Also, the accuracy of KNN is also pretty high, part 1 result proves that *the KNN achieve good results in many cases,* these advantages means the KNN method can be rely on.

The main **disadvantage** is pretty hurt as well. First, due to the regression need to sort the whole training list, so *the effiency will be very low* if the train list is huge, it can run a pretty long time until it reaches the end of the world.

Except KNN is not effiency, also, KNN depends on the particular feature. In our part 1, we divide each feature by their range, I think it is assigning the weight to each feature, if each feature is not be assigned the weight preperly, the accuracy result may be very low.

4. Assuming that you are asked to apply the k-fold cross validation method for the above problem with k=5, what would you do? State the major steps.

1. Merge the training Wine set and testing Wine set into one big set, let's call it the whole set.

2. Chop the whole set into 5 equals subsets(k=5), which means the size of 5 subsets are equal and if add them together, it is the Whole set.

3. For each subsets,

   (1) Treat it as the test set

   (2) Treat the rest 4 subsets (i.e.k-1) as the training set

   (3) Train classifier using the training set, apply it to the test set

The step 3 will be repeated 5 times which is the number of the K(the folds), which means the training/test process will be repeated 5 times with each of the 5 subsets used exactly once as the test set. For each fold, the test results can be added together in order to get the average single estimation.

In this case, this k-fold cross validation can be applied by the above steps, and can be used to measure the performance of my KNN.

5. In the above problem, assuming that there were actually no class labels available. Which method would you use to group the examples in the data set? State the major steps.

No class label provided means it is the Unsupervised learning, I will use the **K-means clustering method**, since it is used for unlabelled data and can partition m instances into k clusters.

To be able to implement K-means, I will:

1. Initialize k initial “means” randomly from the data set.

2. Create k clusters by assigning every instance to the nearest cluster: based on the nearest mean according to the distance measure.

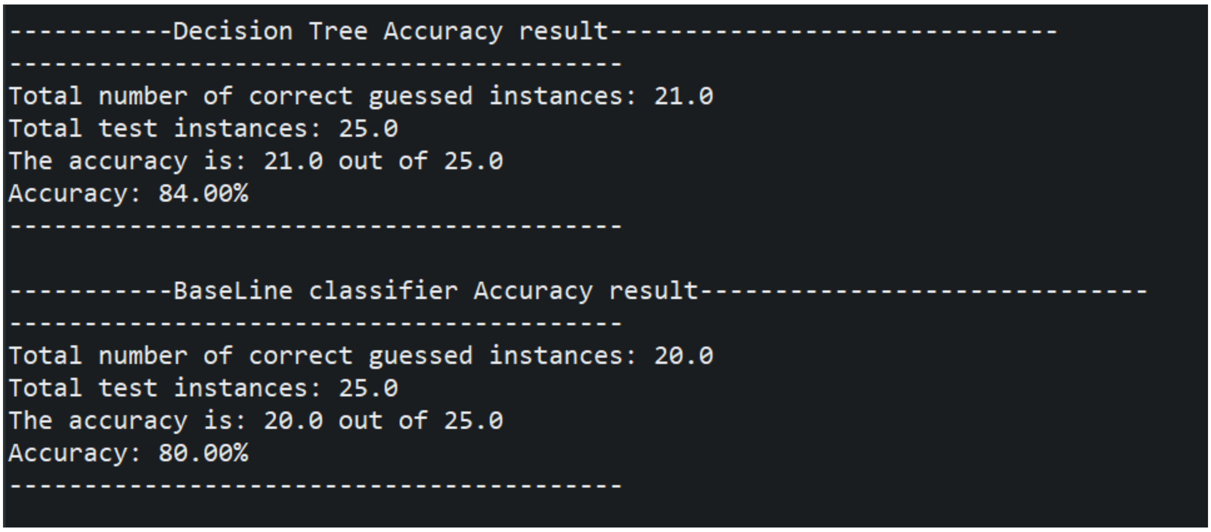
3. Update the centroid, which will replace the old means with the centroid (mean) of each cluster

Repeat the above two steps until convergence (no change in each cluster centroid).(p.s. Centroid is not an instance)

Yun Zhou, 300442776

PART 2: Decision Tree

Q1: You should first apply your program to the hepatitis-training and hepatitis-test files and report the classification accuracy in terms of the fraction of the test instances that it classified correctly. Report the constructed decision tree classifier printed by your program. Compare the accuracy of your decision tree program to the baseline classifier (which always predicts the most frequent class in the training set), and comment on any difference.



```
-----------Decision Tree Accuracy result-----------------------------
----------------------------------------
Total number of correct guessed instances: 21.0
Total test instances: 25.0
The accuracy is: 21.0 out of 25.0
Accuracy: 84.00%
----------------------------------------

-----------BaseLine classifier Accuracy result-----------------------------
----------------------------------------
Total number of correct guessed instances: 20.0
Total test instances: 25.0
The accuracy is: 20.0 out of 25.0
Accuracy: 80.00%
----------------------------------------
```

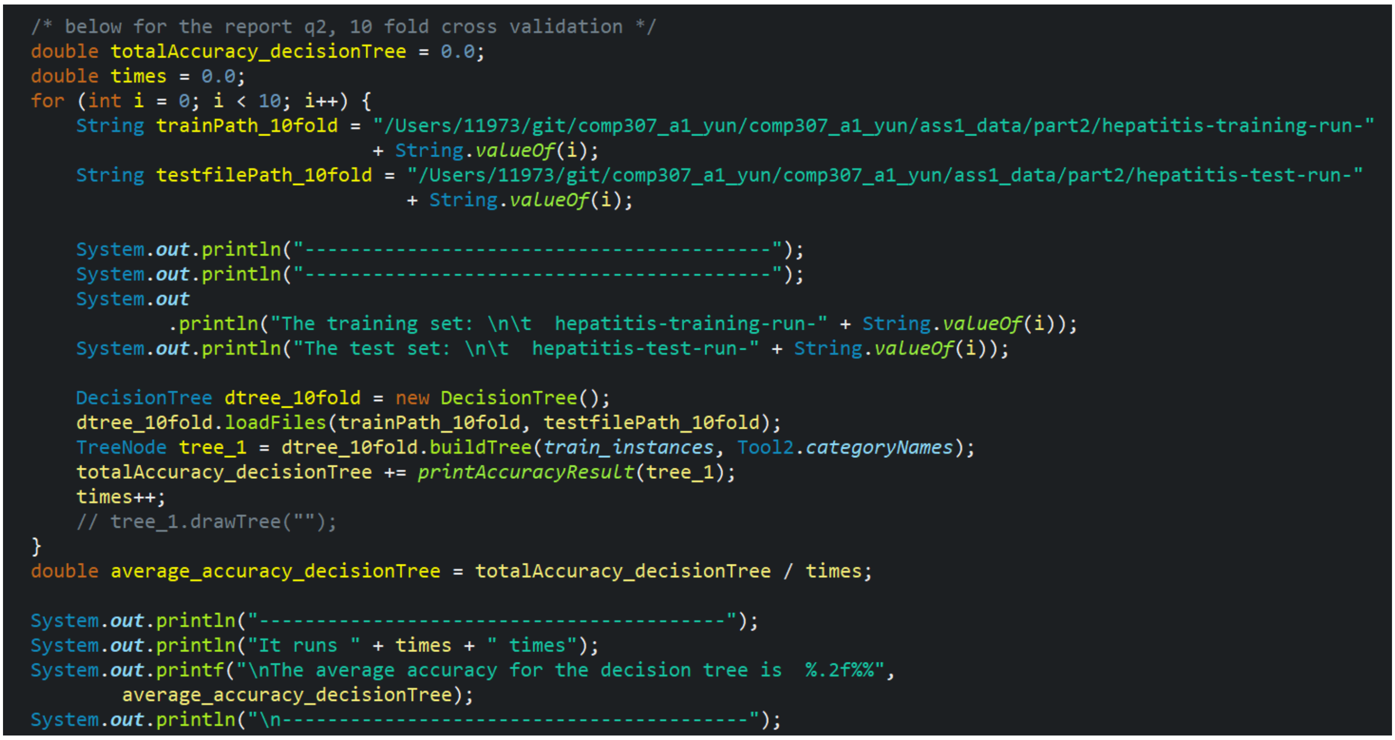
The full log result **including the constructed Decision Tree** can be **seen** in the **Q1_sampleoutput.txt** which is inside the **part2 folder,** also, you can see the tree and accuracy result if you run my program in the terminal.

From the output accuracy result shown above, we can see that, after the decision tree is made by the training, it can get 21 out of 25 correct guessed instances, meanwhile, the baseline classifier get 20 out of 25.
Therefore, I can get that my decision tree algorithm model can perform better prediction on the 25 instances test set. This is not surprise since the baseline classifier is a stupid and simple approach which always predicts the most frequent class in the training set, and the constructed decision tree will predict the result by making the decision up to each attribute value, which means more accurate and smart.

Yun Zhou, 300442776

2. You should then apply 10-fold cross-validation to evaluate the robustness of your algorithm. We have provided files for the split training and test sets. The files are named as hepatitis-training-run-*, and hepatitis-test-run-*. Each training set has 107 instances and each test set has the remaining 30 instances. You should train and test your classifier on each pair, and calculate the average accuracy of the classifiers across the 10 folds (show your working).

By following this question instruction, in order to evaluate the robustness of my decision tree algorithm, I go back to add a block of the code, which run the 10-fold cross-validation.



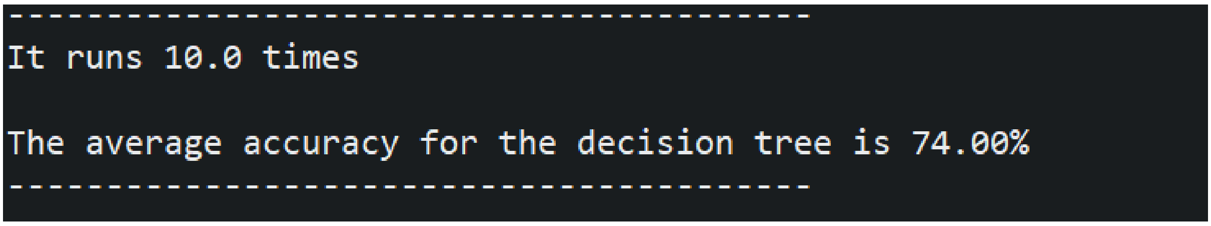
```java
/* below for the report q2, 10 fold cross validation */
double totalAccuracy_decisionTree = 0.0;
double times = 0.0;
for (int i = 0; i < 10; i++) {
    String trainPath_10fold = "/Users/11973/git/comp307_a1_yun/comp307_a1_yun/ass1_data/part2/hepatitis-training-run-"
                            + String.valueOf(i);
    String testfilePath_10fold = "/Users/11973/git/comp307_a1_yun/comp307_a1_yun/ass1_data/part2/hepatitis-test-run-"
                               + String.valueOf(i);

    System.out.println("----------------------------------------");
    System.out.println("----------------------------------------");
    System.out
            .println("The training set: \n\t  hepatitis-training-run-" + String.valueOf(i));
    System.out.println("The test set: \n\t  hepatitis-test-run-" + String.valueOf(i));

    DecisionTree dtree_10fold = new DecisionTree();
    dtree_10fold.loadFiles(trainPath_10fold, testfilePath_10fold);
    TreeNode tree_1 = dtree_10fold.buildTree(train_instances, Tool2.categoryNames);
    totalAccuracy_decisionTree += printAccuracyResult(tree_1);
    times++;
    // tree_1.drawTree("");
}
double average_accuracy_decisionTree = totalAccuracy_decisionTree / times;

System.out.println("----------------------------------------");
System.out.println("It runs " + times + " times");
System.out.printf("\nThe average accuracy for the decision tree is  %.2f%%",
        average_accuracy_decisionTree);
System.out.println("\n----------------------------------------");
```

As you can see that, I use the for loop to iterate through each pair of the train/test set, add each of the accuracy results together and finally divided by running times.



```
-----------------------------------------
It runs 10.0 times

The average accuracy for the decision tree is 74.00%
-----------------------------------------
```

Here you can see the result, as we can see that, the average accuracy of the classifiers across the 10 folds is 74%, which proves the robustness of my decision tree algorithm.

(p.s. The full log can be seen in **Q2_sampleoutput.txt** which is inside the **part2 folder**)

3. “Pruning” (removing) some of leaves of the decision tree will always make the decision tree less accurate on the training set. Explain:

(a) how you could prune leaves from the decision tree;

There are lots of approaches that can prune some of leaves of the decision tree, pruning will result less accurate on the training set, but can run better on testing set which avoid the overfitting.
From the tutorial website below, I can use the approach: error complexity pruning. This approach will generate a series of trees, and each one is made by pruning the full tree by different amounts, and finally select one of these by assessing its performance with an independent data set.

https://alanjeffares.wordpress.com/tutorials/decision-tree/

(b) why it reduces accuracy on the training set;

The pruning can avoid the case of overfitting. In the case of overfitting, the decision tree algorithm is constantly dividing nodes in order to classify the training samples as accurately as possible, which will lead the result of it can run well on training, but not on testing.
Pruning can make sure that the algorithm is not too biased on training, so that it can run better on testing, but it will reduce the accuracy on the training set.

and (c) why it might improve accuracy on the test set.

The purpose of the Pruning is to avoid the case of the overfitting, which means the decision tree runs well on training, but not on testing. The overfitting is caused due to the decision tree algorithm is constantly dividing the nodes in order to classify the training samples as accurately as possible.
During the learning process, this will cause the whole tree to have too many branches, which leads to overfitting.
The pruning can avoid the case of the overfitting, less branch means it’s not too biased on training, which might improve the accuracy on the test set.

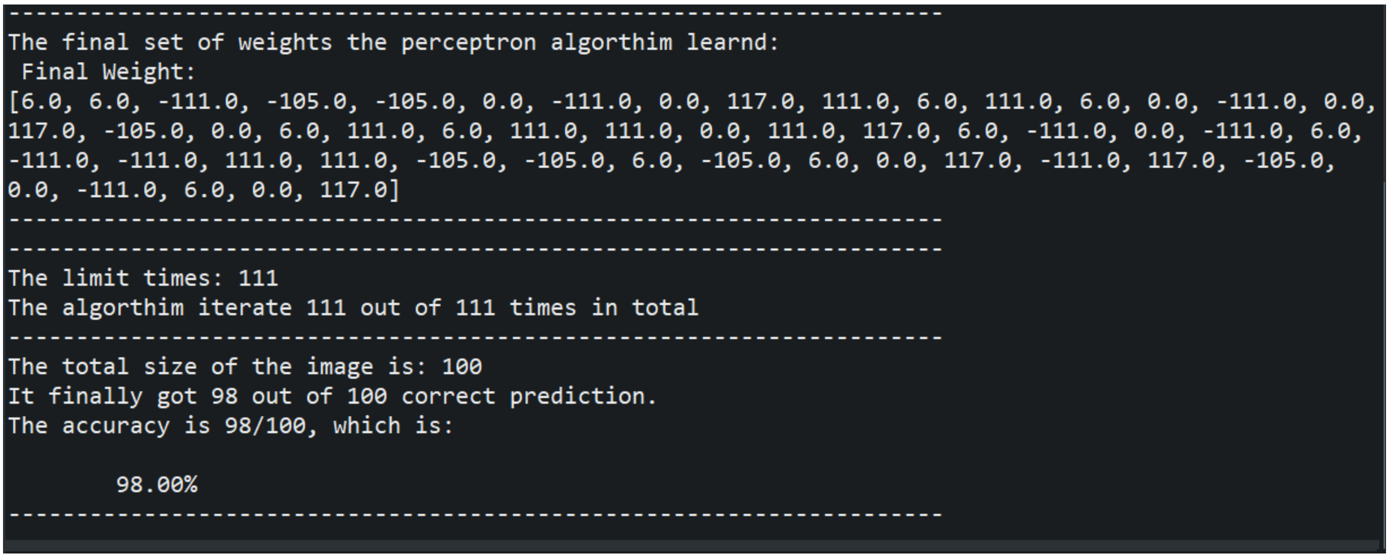
4. Explain why the impurity measure (from lectures) is not an appropriate measure to use if there are three or more classes in the dataset.

The impurity measure from the lecture works well on the set that has only 2 classes since 2 class labels are easy to calculate, only pure and impure two options, so that it can get a smooth graph.
However, in the real world, three and more classes are common. If the impurity measure is still in use, the more the classes, the less the accuracy, the graph is not smooth anymore, so that the impurity measure doesn’t make any scene. For more classes, we need to use Gini impurity or the Entropy.

Yun Zhou, 300442776

PART 3: Perceptron

1. Report on the accuracy of your perceptron. For example, did it find a correct set of weights? Did its performance change much between different runs?



```
--------------------------------------------------------------------
The final set of weights the perceptron algorthim learnd:
 Final Weight:
[6.0, 6.0, -111.0, -105.0, -105.0, 0.0, -111.0, 0.0, 117.0, 111.0, 6.0, 111.0, 6.0, 0.0, -111.0, 0.0,
117.0, -105.0, 0.0, 6.0, 111.0, 6.0, 111.0, 111.0, 0.0, 111.0, 117.0, 6.0, -111.0, 0.0, -111.0, 6.0,
-111.0, -111.0, 111.0, 111.0, -105.0, -105.0, 6.0, -105.0, 6.0, 0.0, 117.0, -111.0, 117.0, -105.0,
0.0, -111.0, 6.0, 0.0, 117.0]
--------------------------------------------------------------------
--------------------------------------------------------------------
The limit times: 111
The algorthim iterate 111 out of 111 times in total
--------------------------------------------------------------------
The total size of the image is: 100
It finally got 98 out of 100 correct prediction.
The accuracy is 98/100, which is:

        98.00%
--------------------------------------------------------------------
```

The screenshot above is the result, due to the feature is random generated each time, so this result is unique.

As we can see that, my perceptron got 98% accuracy. I think it find the correct set of weight, because the initialization of the weight values are all set to 0, and most of the final weights values are changed.

2. Explain why evaluating the perceptron's performance on the training data is not a good measure of its effectiveness. For an A+, you should create additional data to get a better measure (e.g. using MakeImage.java). If you do, report on the perceptron's performance on this additional data.

Run on the training data will change the associated feature weight in order to get better accuracy results on the training set, therefore the weights are only for the training set, not mention that the features are random.

Yun Zhou, 300442776

Full Log, for Question 1:

(Can also be seen in the sampleoutput_part1.txt, which is inside the part1 folder)

```
============     Part 1     ===========

-----------------K= 1-----------------------------

----------------------------------------------

For the 1th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 2th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 3th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 4th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 5th test wine:

The real class label is 1

The guess class label by my knn is 1
```

```
which is: true

----------------------------------------------

For the 6th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 7th test wine:

The real class label is 2

The guess class label by my knn is 1

which is: false

----------------------------------------------

For the 8th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 9th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 10th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 11th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 12th test wine:

The real class label is 2

The guess class label by my knn is 3

which is: false

----------------------------------------------

For the 13th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 14th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 15th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 16th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 17th test wine:

The real class label is 3

The guess class label by my knn is 3
```

```
which is: true

----------------------------------------------

For the 18th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 19th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 20th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 21th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 22th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 23th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 24th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 25th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 26th test wine:

The real class label is 2

The guess class label by my knn is 3

which is: false

----------------------------------------------

For the 27th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 28th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 29th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 30th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 31th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 32th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 33th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 34th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 35th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 36th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 37th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 38th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 39th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 40th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 41th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 42th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 43th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 44th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 45th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 46th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 47th test wine:

The real class label is 3

The guess class label by my knn is 3
```

```
which is: true

----------------------------------------------

For the 48th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 49th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 50th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 51th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 52th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 53th test wine:

The real class label is 1

The guess class label by my knn is 1
```

```
which is: true

----------------------------------------------

For the 54th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 55th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 56th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 57th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 58th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 59th test wine:

The real class label is 1

The guess class label by my knn is 1
```

```
which is: true

---------------------------------------------

For the 60th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

---------------------------------------------

For the 61th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

---------------------------------------------

For the 62th test wine:

The real class label is 2

The guess class label by my knn is 1

which is: false

---------------------------------------------

For the 63th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

---------------------------------------------

For the 64th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

---------------------------------------------

For the 65th test wine:

The real class label is 2

The guess class label by my knn is 2
```

```
which is: true

----------------------------------------------

For the 66th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 67th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

----------------------------------------------

For the 68th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 69th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

----------------------------------------------

For the 70th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

----------------------------------------------

For the 71th test wine:

The real class label is 2

The guess class label by my knn is 1
```

```
which is: false

---------------------------------------------

For the 72th test wine:

The real class label is 3

The guess class label by my knn is 3

which is: true

---------------------------------------------

For the 73th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

---------------------------------------------

For the 74th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

---------------------------------------------

For the 75th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

---------------------------------------------

For the 76th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

---------------------------------------------

For the 77th test wine:

The real class label is 1

The guess class label by my knn is 1
```

```
which is: true
---------------------------------------------
For the 78th test wine:
The real class label is 3
The guess class label by my knn is 3
which is: true
---------------------------------------------
For the 79th test wine:
The real class label is 1
The guess class label by my knn is 1
which is: true
---------------------------------------------
For the 80th test wine:
The real class label is 1
The guess class label by my knn is 1
which is: true
---------------------------------------------
For the 81th test wine:
The real class label is 2
The guess class label by my knn is 2
which is: true
---------------------------------------------
For the 82th test wine:
The real class label is 2
The guess class label by my knn is 2
which is: true
---------------------------------------------
For the 83th test wine:
The real class label is 3
The guess class label by my knn is 3
```

```
which is: true

---------------------------------------------

For the 84th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

---------------------------------------------

For the 85th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

---------------------------------------------

For the 86th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

---------------------------------------------

For the 87th test wine:

The real class label is 1

The guess class label by my knn is 1

which is: true

---------------------------------------------

For the 88th test wine:

The real class label is 2

The guess class label by my knn is 2

which is: true

---------------------------------------------

For the 89th test wine:

The real class label is 1

The guess class label by my knn is 1
```

```
which is: true

------------------------------------------------

-----------------K= 1------------------------------

The number of correct classified instances: 84.0

Total test wine instances: 89

The Accuracy is: 94.38%

------------------------------------------------
```