Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

Part1

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-Nearest Neighbour result:
1) k =1  Original: 3 Predict: 3  Match
2) k =1  Original: 3 Predict: 3  Match
3) k =1  Original: 3 Predict: 3  Match
4) k =1  Original: 1 Predict: 1  Match
5) k =1  Original: 1 Predict: 1  Match
6) k =1  Original: 1 Predict: 1  Match
7) k =1  Original: 2 Predict: 1  Mismatch
8) k =1  Original: 2 Predict: 2  Match
9) k =1  Original: 1 Predict: 1  Match
10) k =1  Original: 2 Predict: 2  Match
11) k =1  Original: 2 Predict: 2  Match
12) k =1  Original: 2 Predict: 3  Mismatch
13) k =1  Original: 3 Predict: 3  Match
14) k =1  Original: 3 Predict: 3  Match
15) k =1  Original: 1 Predict: 1  Match
16) k =1  Original: 2 Predict: 2  Match
17) k =1  Original: 3 Predict: 3  Match
18) k =1  Original: 3 Predict: 3  Match
19) k =1  Original: 1 Predict: 1  Match
20) k =1  Original: 1 Predict: 1  Match
21) k =1  Original: 3 Predict: 3  Match
22) k =1  Original: 2 Predict: 2  Match
23) k =1  Original: 2 Predict: 2  Match
24) k =1  Original: 3 Predict: 3  Match
25) k =1  Original: 2 Predict: 2  Match
26) k =1  Original: 2 Predict: 3  Mismatch
27) k =1  Original: 2 Predict: 2  Match
28) k =1  Original: 3 Predict: 3  Match
29) k =1  Original: 2 Predict: 2  Match
30) k =1  Original: 1 Predict: 1  Match
31) k =1  Original: 2 Predict: 2  Match
32) k =1  Original: 1 Predict: 1  Match
33) k =1  Original: 2 Predict: 2  Match
34) k =1  Original: 1 Predict: 1  Match
35) k =1  Original: 2 Predict: 2  Match
36) k =1  Original: 2 Predict: 2  Match
37) k =1  Original: 2 Predict: 2  Match
38) k =1  Original: 2 Predict: 2  Match
39) k =1  Original: 2 Predict: 2  Match
40) k =1  Original: 1 Predict: 1  Match
41) k =1  Original: 2 Predict: 2  Match
42) k =1  Original: 2 Predict: 2  Match
43) k =1  Original: 3 Predict: 3  Match
44) k =1  Original: 1 Predict: 1  Match
45) k =1  Original: 2 Predict: 2  Match
46) k =1  Original: 1 Predict: 1  Match
47) k =1  Original: 3 Predict: 3  Match
48) k =1  Original: 2 Predict: 2  Match
49) k =1  Original: 2 Predict: 2  Match
50) k =1  Original: 1 Predict: 1  Match
51) k =1  Original: 3 Predict: 3  Match
52) k =1  Original: 1 Predict: 1  Match
53) k =1  Original: 1 Predict: 1  Match
54) k =1  Original: 3 Predict: 3  Match
55) k =1  Original: 3 Predict: 3  Match
56) k =1  Original: 1 Predict: 1  Match
57) k =1  Original: 1 Predict: 1  Match
58) k =1  Original: 3 Predict: 3  Match
59) k =1  Original: 1 Predict: 1  Match
60) k =1  Original: 3 Predict: 3  Match
61) k =1  Original: 3 Predict: 3  Match
62) k =1  Original: 2 Predict: 1  Mismatch
63) k =1  Original: 2 Predict: 2  Match
64) k =1  Original: 3 Predict: 3  Match
65) k =1  Original: 2 Predict: 2  Match
66) k =1  Original: 3 Predict: 3  Match
67) k =1  Original: 3 Predict: 3  Match
68) k =1  Original: 1 Predict: 1  Match
69) k =1  Original: 1 Predict: 1  Match
70) k =1  Original: 2 Predict: 2  Match
71) k =1  Original: 2 Predict: 1  Mismatch
72) k =1  Original: 3 Predict: 3  Match
73) k =1  Original: 2 Predict: 2  Match
74) k =1  Original: 2 Predict: 2  Match
75) k =1  Original: 1 Predict: 1  Match
76) k =1  Original: 1 Predict: 1  Match
77) k =1  Original: 1 Predict: 1  Match
78) k =1  Original: 3 Predict: 3  Match
79) k =1  Original: 1 Predict: 1  Match
80) k =1  Original: 1 Predict: 1  Match
81) k =1  Original: 2 Predict: 2  Match
82) k =1  Original: 2 Predict: 2  Match
83) k =1  Original: 3 Predict: 3  Match
84) k =1  Original: 1 Predict: 1  Match
85) k =1  Original: 2 Predict: 2  Match
86) k =1  Original: 1 Predict: 1  Match
87) k =1  Original: 1 Predict: 1  Match
88) k =1  Original: 2 Predict: 2  Match
89) k =1  Original: 1 Predict: 1  Match
Accuracy is 94.38% for k =1
```

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

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).

==========

Accuracy is 94.38% for k =1

Accuracy is 95.51% for k =3

K = 3 has higher performance than K = 1. The higher K value demonstrates that more closest neighbours are being compared to each other. The result is that when K = 3 has a “higher resolution” when it predicts. K = 1 has a higher risk of the data being over-fitted to the training dataset so it does not give more of general classification that K = 3 would have allowed. This means more correctly classified instances occur. However, this does not mean that increasing K to a bigger number gives a reading with more accuracy. What happens instead is that the data is under fitted so it becomes more generalized.

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

Advantages

- Works very well with big training data sets
- It runs simply without complications when dealing training data sets
- Unlike some other AI classifications KNN algorithm is easier to understand
- Using Euclidean distance helps to measure the distance correctly

Disadvantages

- When the training data set is big it will be difficult to calculate the distance for each training instance because it you need to consider not only the classification but also the performance of the run time.
- Instances branch out to couple of features and features branch out to dimensions. So, when you are dealing with a massive data set it will be more complicated to find the closest nearest neighbour.
- There are many ways to calculate the distance between the neighbours. So its hard to choose a specific method for this classification because we can consider taking the weight of the distance into account or doing the inverse square of the distance.
- KNN classification is sensitive to outliers therefore this means the accuracy of the result will be changed.

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

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.

Assuming there are 100 instances this means that 50 instances for the training data set and 50 for the test data set. Assuming we are using only 50 for test data set only.

- The total is 50 we want 5 sets, which means each set has 10 instances.
- Take the first set for the test data set and then the rest for the training data set
- Use the training data set for doing the classification and consider the test set as well.
- Then report the output
- Repeat the process recursively for other sets
- Finally, to get an estimation we get the average of the overall results.

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.

K-means Clustering

- K=3 refers to 3 clusters which are chosen randomly from the data.
- To determine K clusters, we consider the mean of the nearest neighbours when calculating distance.
- Replace the previous mean with the centroid of each cluster. This means that each time you replace the mean, it moves to the centre of the element.
- Repeat the process until it converges.

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

Part2

1. 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.

```
BILIRUBIN = True:
        BIGLIVER = True:
                ANOREXIA = True:
                        ANOREXIA = True:
                                Class = live, prob = 1.00
                        ANOREXIA = False:
                                ANOREXIA = True:
                                        ANOREXIA = True:
                                                Class = live, prob = 1.00
                                        ANOREXIA = False:
                                                ANOREXIA = True:
                                                        ANOREXIA = True:
                                                                Class = live, prob = 1.00
                                                        ANOREXIA = False:
                                                                ANOREXIA = True:
                                                                        Class = live, prob = 1.00
                                                                ANOREXIA = False:
                                                                        ANOREXIA = True:
                                                                                Class = die, prob = 1.00
                                                                        ANOREXIA = False:
                                                                                Class = live, prob = 1.00
                                                ANOREXIA = False:
                                                        Class = live, prob = 1.00
                                ANOREXIA = False:
                                        ANOREXIA = True:
                                                Class = live, prob = 1.00
                                        ANOREXIA = False:
                                                Class = die, prob = 1.00
                ANOREXIA = False:
                        Class = die, prob = 1.00
        BIGLIVER = False:
                BIGLIVER = True:
                        SGOT = True:
                                SGOT = True:
                                        Class = live, prob = 1.00
                                SGOT = False:
                                        Class = die, prob = 1.00
                        SGOT = False:
                                Class = live, prob = 1.00
                BIGLIVER = False:
                        BIGLIVER = True:
                                BIGLIVER = True:
                                        Class = live, prob = 1.00
                                BIGLIVER = False:
                                        Class = die, prob = 1.00
                        BIGLIVER = False:
                                Class = live, prob = 1.00
BILIRUBIN = False:
        BILIRUBIN = True:
                BILIRUBIN = True:
                        BILIRUBIN = True:
                                BILIRUBIN = True:
                                        Class = die, prob = 1.00
                                BILIRUBIN = False:
                                        BILIRUBIN = True:
                                                Class = live, prob = 1.00
                                        BILIRUBIN = False:
                                                Class = die, prob = 1.00
                        BILIRUBIN = False:
                                Class = live, prob = 1.00
                BILIRUBIN = False:
                        Class = die, prob = 1.00
        BILIRUBIN = False:
                Class = live, prob = 1.00
1)  Result: live   True value result: live
--------------------
2)  Result: die   True value result: die
--------------------
3)  Result: die   True value result: live
--------------------
4)  Result: live   True value result: live
--------------------
5)  Result: live   True value result: live
--------------------
6)  Result: live   True value result: live
--------------------
7)  Result: live   True value result: die
--------------------
8)  Result: live   True value result: live
--------------------
9)  Result: live   True value result: live
--------------------
10)  Result: live   True value result: live
--------------------
11)  Result: live   True value result: live
--------------------
12)  Result: live   True value result: live
--------------------
13)  Result: live   True value result: live
--------------------
14)  Result: live   True value result: live
--------------------
15)  Result: live   True value result: die
--------------------
16)  Result: live   True value result: die
--------------------
17)  Result: live   True value result: live
--------------------
18)  Result: live   True value result: live
--------------------
19)  Result: live   True value result: die
--------------------
20)  Result: live   True value result: live
--------------------
21)  Result: live   True value result: live
--------------------
22)  Result: live   True value result: live
--------------------
23)  Result: live   True value result: live
--------------------
24)  Result: live   True value result: live
--------------------
25)  Result: live   True value result: live
--------------------
Final result is 20 out of 25
```

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

```
Final result is 20 out of 25 (80.0% )
========================================
Baseline ( live ) is 81.25%
========================================
```

Based on the output, the decision tree is “live” skewed. This means the baseline classifier shows greater accuracy for “live” compared to the overall accuracy of this decision tree. This simply due to insufficient “die” data in the training data set whereby it is difficult for decision tree to classify the “die” values. Overall, it shows the overall accuracy is quite low of this decision tree as the number of “live” and “die” are not even.

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).

```
int countValue = 0;
for (Instance instanceTest: hepatitisTest) {
    String value = recursiveDecision(tree, instanceTest);
    if(value.equals(instanceTest.getCategory())) {
        countValue++;
    }
}
accuracyFoldTen.add((double)countValue/(double) hepatitisTest.size());
```

```
for(int i = 0; i < 10; i++) {
    crossValidations( trainingFile: "part2/hepatitis-training-run-" + i, testFile: "part2/hepatitis-test-run-" + i +"\n"+ dividerLine);
}
double totalValue = 0.0;
for(double accuracy: accuracyFoldTen) { totalValue += accuracy; }

System.out.println("K-fold for 10 average value is " + String.format("%.2f",(totalValue/(double) accuracyFoldTen.size())*100) + "%");
```

```
K-fold for 10 average value is 80.13%
```

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

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

Pruning is done inversely splitting which means when pairs have an impurity they will be eliminated. Then the parent node will swap to the child node which finally reduces the size of the tree.

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

Because pruning eliminates overfitting as it is not effective for training other data sets but rather useful to get the accurate result for its own training data when making the decision tree.

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

Because pruning mitigates overfitting as it has no overfitting. This helps to get accurate results for other data sets as well.

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 measures by considering the number of occurrences of binary value either 0 or 1 of two classes which shows the purity. This means, splitting binary value to more than 2 classes would be complicated as this requires fraction values to measure.

Comp307 Assignment 1 Report
Name: Elgene Menon Leo Anthony
StudentID:300492604

Part3

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?

Based on the output, it shows weights are classified about 100 percent of the images where it prints the success weights only and terminates the program. The learning rate was a constant of 0.2 and set maximum epochs of 1000.The failure rate is when the number of epochs is greater than 100 percent. Based on the result for testing the perceptron, it shows as the number of features increases, the number of epochs reduces the success rate.

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.

This is because of some possible overfitting. The weights for training this data set may not work for other training data sets as perceptron is trained using this training data set. Therefore, the weights are measured specifically for this purpose. To do perceptron for the other training data set requires different weights to do so. Since this is the result it shows the perceptron is 100 percent accuracy for this data set. This means it is not good enough to use the weights for classifying other data sets.