Name: Janmejay Mohanty

CS-559-B

Homework 3 Assignment

Solution 1: (1)

Entropy $(t) = \sum_j p(j|t) \log_2 p(j|t)$

Entropy$(Class) = -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$

$= (-0.44 * -1.18) - (0.56 * -0.84)$

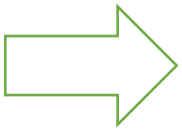
$= 0.51997 + 0.47111$

$= 0.99107$

$= 0.991$

Starting with Attribute 1 (a_1). Putting all instances into a table:

a_1	$+$	$-$
T	3	1
F	1	4



a_1	$+$	$-$
T	$\frac{3}{4}$	$\frac{1}{4}$
F	$\frac{1}{5}$	$\frac{4}{5}$

Entropy$(a_1) = \frac{4}{9}\left(-\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) + \frac{5}{9}\left(-\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right)\right)$

$= 0.44[(-0.75 * -0.4150) - (0.25 * -2)] + 0.5556[(-0.2 * -2.3219) - (0.8 * -0.3219)]$

$= 0.44[0.31125 + 0.5] + 0.5556[0.46438 + 0.25752]$

$= 0.44 * 0.81125 + 0.5556 * 0.7219$

$= 0.35695 + 0.4010$

$= 0.75795$

Information Gain$(a_1) = 0.991 - 0.75795 = 0.23305 = 0.23$

Values of Attribute 2(a_2), lies in the following range of $[1.0, 8.0]$. After doing sorting on the Attribute 2(a_2) we now split positions midway between neighboring values.

Instances	Attribute 2 (a_2)	Class
1	1.0	+
6	3.0	−
4	4.0	+
3	5.0	−
9	5.0	−
2	6.0	+
5	7.0	−
8	7.0	+
7	8.0	−

Split 1:

0.5	≤	>
+	0	4
−	0	5
E/IG	0.991	0

Entropy$(t) = \sum_j p(j|t) \log_2 p(j|t)$

$$= -\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$$

$$= -[(-0.51997) + (-0.47111)]$$

$$= 0.99107$$

$$= 0.991$$

Information Gain: $\Delta = 0.991 - 0.991 = 0$

Split 2:

2.0	≤	>
+	1	3
−	0	5
E/IG	**0.8484**	**0.143**

$\leq$ Entropy $= -\left[\left(\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right) + (0) * \log_2(0)\right]$

$$= 0$$

$>$ Entropy $= -\left[\left(\frac{3}{8}\right) * \log_2\left(\frac{3}{8}\right) + (0) * \log_2(0)\right]$

$$= -[(-0.53064) + (-0.42379)]$$

$$= 0.95443$$

Weighted Average $= \left[\left(\frac{1}{9}\right) * 0\right] + \left[\left(\frac{8}{9}\right) * 0.95443\right]$

$= 0.84839$

$= 0.8484$

Information Gain: $\Delta = 0.991 - 0.8484 = 0.1426 = 0.143$

3.5	<=	>
+	1	3
−	1	4
E/IG	0.9885	0.0024

Split 3:

$\leq$ Entropy $= -\left[\left(\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right) + \left(\frac{1}{2}\right) * \log_2\left(\frac{1}{2}\right)\right]$

$= -[(-0.5 - 0.5)]$

$= 1$

$>$ Entropy $= -\left[\left(\frac{3}{7}\right) * \log_2\left(\frac{3}{7}\right) + \left(\frac{4}{7}\right) * \log_2\left(\frac{4}{7}\right)\right]$

$= -[(-0.52388) + (-0.46135)]$

$= 0.98523$

Weighted Average $= \left[\left(\frac{2}{9}\right) * 1\right] + \left[\left(\frac{7}{9}\right) * 0.95443\right]$

$= 0.988512$

Information Gain: $\Delta = 0.991 - 0.988512 = 0.002488 = 0.0024$

Split 4:

4.5	<=	>
+	2	2
−	1	4
E/IG	0.9183	0.0727

$\leq$ Entropy $= -\left[\left(\frac{2}{3}\right) * \log_2\left(\frac{2}{3}\right) + \left(\frac{1}{3}\right) * \log_2\left(\frac{1}{3}\right)\right]$

$= -[(-0.38998 - 0.52832)]$

$= 0.9183$

$>$ Entropy $= -\left[\left(\frac{2}{6}\right) * \log_2\left(\frac{2}{6}\right) + \left(\frac{4}{6}\right) * \log_2\left(\frac{4}{6}\right)\right]$

$$= -[(-0.5283) + (-0.38998)]$$

$$= 0.9183$$

Weighted Average $= \left[\left(\frac{3}{9}\right) * 0.9183\right] + \left[\left(\frac{6}{9}\right) * 0.9183\right]$

$$= 0.9183$$

Information Gain: $\Delta = 0.991 - 0.9183 = 0.0727$

Split 5:

5.5	<=	>
+	2	2
−	3	2
E/IG	0.9838	0.0072

$\leq$ Entropy $= -\left[\left(\frac{2}{5}\right) * \log_2\left(\frac{2}{5}\right) + \left(\frac{3}{5}\right) * \log_2\left(\frac{3}{5}\right)\right]$

$$= -[(-0.52877 - 0.44218)]$$

$$= 0.97095$$

$>$ Entropy $= -\left[\left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right) + \left(\frac{2}{4}\right) * \log_2\left(\frac{2}{4}\right)\right]$

$$= -[(-0.5) + (-0.5)]$$

$$= 1$$

Weighted Average $= \left[\left(\frac{5}{9}\right) * 0.97095\right] + \left[\left(\frac{4}{9}\right) * 1\right]$

$$= 0.5394 + 0.4444$$

$$= 0.9838$$

Information Gain: $\Delta = 0.991 - 0.9838 = 0.0072$

Split 6:

6.5	<=	>
+	3	1
−	3	2
E/IG	0.9728	0.0184

$\leq$ Entropy $= -\left[\left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right) + \left(\frac{3}{6}\right) * \log_2\left(\frac{3}{6}\right)\right]$

$$= -[-0.5 - 0.5]$$

$$= 1$$

$$> \text{Entropy} = -\left[\left(\frac{1}{3}\right) * \log_2\left(\frac{1}{3}\right) + \left(\frac{2}{3}\right) * \log_2\left(\frac{2}{3}\right)\right]$$

$$= -[(-0.52832) + (-0.38998)]$$

$$= 0.9183$$

$$\text{Weighted Average} = \left[\left(\frac{6}{9}\right) * 1\right] + \left[\left(\frac{3}{9}\right) * 0.9183\right]$$

$$= 0.6667 + 0.3061$$

$$= 0.9728$$

Information Gain: $\Delta = 0.991 - 0.9728 = 0.0184$

Split 7:

7.5	<=	>
+	4	0
−	4	1
E/IG	0.8889	0.1021

$$\leq \text{Entropy} = -\left[\left(\frac{4}{8}\right) * \log_2\left(\frac{4}{8}\right) + \left(\frac{4}{8}\right) * \log_2\left(\frac{4}{8}\right)\right]$$

$$= -[-0.5 - 0.5]$$

$$= 1$$

$$> \text{Entropy} = -\left[\left(\frac{0}{1}\right) * \log_2\left(\frac{0}{1}\right) + \left(\frac{1}{1}\right) * \log_2\left(\frac{1}{1}\right)\right]$$

$$= 0$$

$$\text{Weighted Average} = \left[\left(\frac{8}{9}\right) * 1\right] + \left[\left(\frac{1}{9}\right) * 0\right]$$

$$= 0.8889$$

Information Gain: $\Delta = 0.991 - 0.8889 = 0.1021$

Split 8:

8.5	<=	>
+	4	0
−	5	0
E/IG	0.991	0

$$\leq \text{Entropy} = -\left[\left(\frac{4}{9}\right) * \log_2\left(\frac{4}{9}\right) + \left(\frac{5}{9}\right) * \log_2\left(\frac{5}{9}\right)\right]$$

$$= -[-0.51997 - 0.47111]$$

$$= 0.99108$$

$$= 0.991$$

$>$ Entropy $= -\left[\left(\frac{0}{0}\right) * \log_2\left(\frac{0}{0}\right) + \left(\frac{0}{0}\right) * \log_2\left(\frac{0}{0}\right)\right]$

$$= 0$$

Weighted Average $= \left[\left(\frac{9}{9}\right) * 0.991\right] + \left[\left(\frac{0}{9}\right) * 0\right]$

$$= 0.991 + 0$$

$$= 0.991$$

Information Gain: $\Delta = 0.991 - 0.991 = 0$

From the above calculation, we can conclude that the split position with maximum information gain at $\boldsymbol{a_2 = 2.0}$ **with Entropy** $\boldsymbol{= 0.8484}$ **and Information Gain** $\boldsymbol{= 0.143.}$

Information Gain $(a_1) = 0.23 >$ Information Gain $(a_2) = 0.143$

Therefore, according to the information gain, the best first splitting for decision tree is $\boldsymbol{a_1}$ due to its higher information gain in comparison to $\boldsymbol{a_2}$.

(2)

If we use "Instances" as new attribute then attribute "Instances" has no predictive power since new rows are assigned to new Instances values.

The Information gain for each Instances value is 0. Therefore, the overall Information gain for Instances is 0.

Therefore, the instances are not suitable attribute for a decision in the tree.

Solution 2: (1)

A	B	Class Label	
		$+$	$-$
T	T	0	20
T	F	20	10
F	T	15	0
F	F	0	35

The contingency tables after splitting on Attributes A and B are:

A	+	−
T	20	30
F	15	35

B	+	−
T	15	20
F	20	45

$$Gini(t) = 1 - \sum_j [p(j|t)]^2$$

$$Gini(parent) = 1 - \left(\frac{35}{100}\right)^2 - \left(\frac{65}{100}\right)^2$$

$$= 1 - 0.35^2 - 0.65^2$$

$$= 1 - 0.1225 - 0.4225$$

$$= 0.455$$

The gain in Gini after splitting on A is:

$$Gini(A = T) = 1 - \left(\frac{20}{50}\right)^2 - \left(\frac{30}{50}\right)^2$$

$$= 1 - 0.4^2 - 0.6^2$$

$$= 1 - 0.16 - 0.36$$

$$= 1 - 0.52$$

$$= 0.48$$

$$Gini(A = F) = 1 - \left(\frac{15}{50}\right)^2 - \left(\frac{35}{50}\right)^2$$

$$= 1 - 0.3^2 - 0.7^2$$

$$= 1 - 0.09 - 0.49$$

$$= 1 - 0.58$$

$$= 0.42$$

$$Gini(A) = \frac{50}{100} Gini(A = T) + \frac{50}{100} Gini(A = F)$$

$$= \frac{1}{2} \times 0.48 + \frac{1}{2} \times 0.42$$

$$= 0.24 + 0.21$$

$$= 0.45$$

The gain in Gini after splitting on B is:

$$Gini(B = T) = 1 - \left(\frac{15}{35}\right)^2 - \left(\frac{20}{35}\right)^2$$

$$= 1 - 0.43^2 - 0.57^2$$

$$= 1 - 0.1849 - 0.3249$$

$$= 1 - 0.5098$$

$$= 0.4902$$

$$Gini(B = F) = 1 - \left(\frac{20}{65}\right)^2 - \left(\frac{45}{65}\right)^2$$

$$= 1 - 0.31^2 - 0.69^2$$

$$= 1 - 0.0961 - 0.4761$$

$$= 1 - 0.5722$$

$$= 0.4278$$

$$Gini(B) = \frac{35}{100} Gini(B = T) + \frac{65}{100} Gini(B = F)$$

$$= 0.35 \times 0.4902 + 0.65 \times 0.4278$$

$$= 0.17157 + 0.27807$$

$$= 0.44964$$

As, $\boldsymbol{Gini(A) > Gini(B)}$.

Therefore, Attribute A will be chosen to split the node.

(2)

Cost Matrix	Attribute Value		
Actual Class		T	F
	$+$	-1	100
	$-$	0	-10

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{-1+0}{-1+0+100-10} = \frac{-1}{89} = -0.011$

Precision $= \frac{TP}{TP+FP} = \frac{-1}{-1+100} = \frac{-1}{99} = -0.010$

Recall $= \frac{TP}{TP+FN} = \frac{-1}{-1-10} = 0.91$

F1-Score $= \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*-0.010*0.91}{-0.010+0.91} = \frac{-0.0182}{0.9} = -0.020$

A	+	−
T	20	30
F	15	35

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{20+30}{20+30+15+35} = \frac{50}{100} = 0.5$

Precision $= \frac{TP}{TP+FP} = \frac{20}{20+15} = \frac{20}{35} = 0.571$

Recall $= \frac{20}{20+35} = 0.364$

F1-Score $= \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.571*0.364}{0.571+0.364} = \frac{0.416}{0.935} = 0.445$

B	+	−
T	15	20
F	20	45

Accuracy $= \frac{TP+TN}{TP+TN+FP+FN} = \frac{15+20}{15+20+20+45} = \frac{35}{100} = 0.35$

Precision $= \frac{TP}{TP+FP} = \frac{15}{15+20} = \frac{15}{35} = 0.429$

Recall $= \frac{15}{15+45} = 0.25$

F1-Score $= \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.429*0.25}{0.429+0.25} = \frac{0.2145}{0.679} = 0.316$

From the above calculation, the best first splitting will be from Attribute A.

Solution 3: (1)

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	−1	−1	−1	−1	−1	1	1

$$D_1 = \frac{1}{N}$$

$$= \frac{1}{10}$$

$$= 0.1$$

Considering Hypotheses, $H1$.

$$H1: if\ X \leq 0.35 \rightarrow Y = 1, else\ Y = -1$$

H_1

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	−1	−1	−1	−1	−1	1	1

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

$$\epsilon_t = P_{t\sim D^t}[h_t(x_i) \neq y_i] = \sum_{i=1}^{N} D_t(i)\,\delta(h_t(x_i) \neq y_i)$$

$$err_1 = 0.1 \times 2 = 0.2$$

$$\alpha_1 = \frac{1}{2} \times \ln\left(\frac{1-0.2}{0.2}\right)$$

$$= \frac{1}{2} \times \ln(4)$$

$$= 0.693$$

$$D_t(i) = \frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t}$$

For those who are not classified correctly: $D(H1) = 0.1 \times e^{0.69\times 1} = 0.1994$

For those who are classified correctly: $D(H1) = 0.1 \times e^{0.69\times(-1)} = 0.0502$

Considering Hypotheses, $H2$.

$$H2: if\ X \leq 0.75 \rightarrow Y = -1, else\ Y = 1$$

							H_2			
ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	−1	−1	−1	−1	−1	1	1

$$err_2 = 0.1 \times 4 = 0.4$$

$$\alpha_2 = \frac{1}{2} \times \ln\left(\frac{1-0.4}{0.4}\right)$$

$$= \frac{1}{2} \times \ln(1.5)$$

$$= 0.203$$

$$D_t(i) = \frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t}$$

For those who are not classified correctly: $D(H2) = 0.1 \times e^{0.20\times 1} = 0.1221$

For those who are classified correctly: $D(H2) = 0.1 \times e^{0.20\times(-1)} = 0.0819$

Considering Hypotheses, $H3$.

$$H3: if\ X \leq 0.3\ or\ X \geq 0.95 \rightarrow Y = 1, else\ Y = -1$$

H_3

ID	1	2	3	4	5	6	7	8	9	10
X	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Y	1	1	1	−1	−1	−1	−1	−1	1	1

$$err_3 = 0.1 \times 2 = 0.2$$

$$\alpha_3 = \frac{1}{2} \times \ln\left(\frac{1-0.2}{0.2}\right)$$

$$= \frac{1}{2} \times \ln(4)$$

$$= 0.693$$

$$D_t(i) = \frac{D_t(i)e^{(-\alpha_t y_i h_t(x_i))}}{Z_t}$$

For those who are not classified correctly: $D(H3) = 0.1 \times e^{0.69\times 1} = 0.1994$

For those who are classified correctly: $D(H3) = 0.1 \times e^{0.69\times(-1)} = 0.0502$

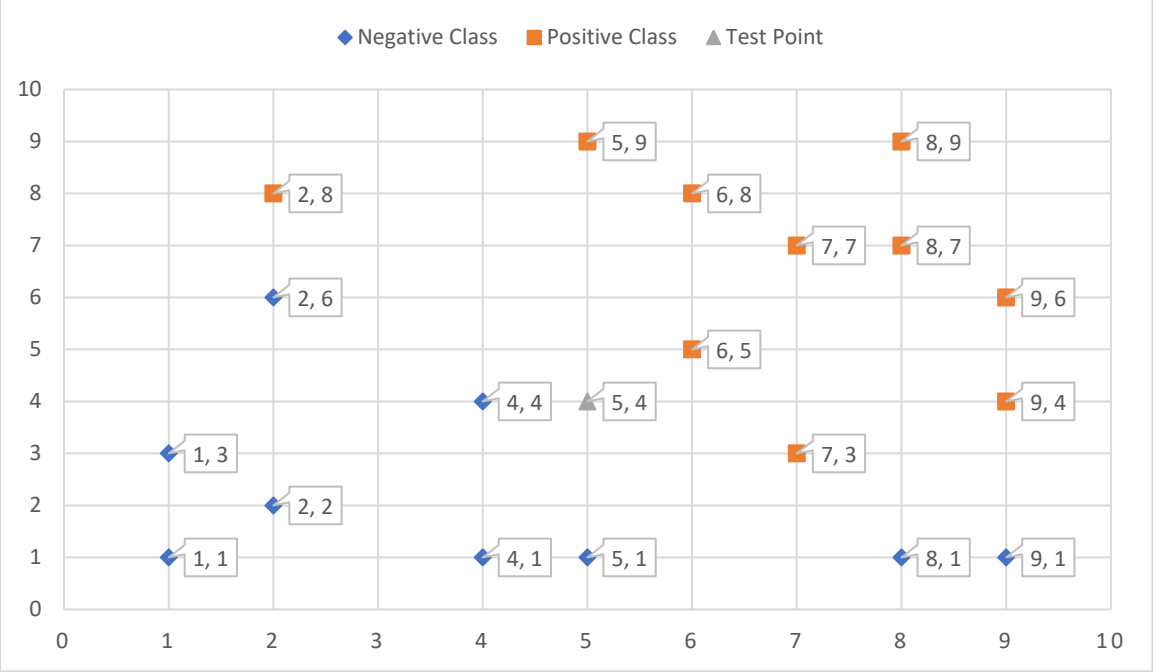
(2)

i	1	2	3
err_i	0.2	0.4	0.2
α_i	0.693	0.203	0.693
$Not\ Classified\ Correctly, D(Hi)$	0.1994	0.1221	0.1994
$Classified\ Correctly, D(Hi)$	0.0502	0.0819	0.0502

From the above table, instance 2 needs to be reweighted after first iteration. Because its error rate $err_2 = 0.4$ is larger than the other two instances error rate.

Solution 4: Part I

(1)



The test point (triangle), $TP = (5,4)$.

Taking random points which appears to be nearest to the test point (TP):

Point 1, $P1 = (4,4)$ $[Negative\ Class]$

Using Euclidean Distance:

$$D1 = \sqrt{(5-4)^2 + (4-4)^2}$$

$$= \sqrt{1}$$

$$= 1$$

Point 2, $P2 = (6,5)$ $[Positive\ Class]$

$$D2 = \sqrt{(5-6)^2 + (4-5)^2}$$

$$= \sqrt{1+1}$$

$$= \sqrt{2}$$

$$= 1.414$$

Point 3, $P3 = (7,3)$ $[Positive\ Class]$

$$D3 = \sqrt{(5-7)^2 + (4-3)^2}$$

$$= \sqrt{4+1}$$

$$= \sqrt{5}$$

$$= 2.236$$

Point 4, $P4 = (5,1)$ $[Negative\ Class]$

$$D4 = \sqrt{(5-5)^2 + (4-1)^2}$$

$$= \sqrt{9}$$

$$= 3$$

Point 5, $P5 = (4,1)$ $[Negative\ Class]$

$$D5 = \sqrt{(5-4)^2 + (4-1)^2}$$

$$= \sqrt{1+9}$$

$$= \sqrt{10}$$

$$= 3.162$$

Point 6, $P6 = (7,7)$ $[Positive\ Class]$

$$D6 = \sqrt{(5-7)^2 + (4-7)^2}$$

$$= \sqrt{4+9}$$

$$= \sqrt{13}$$

$$= 3.606$$

Point 7, $P7 = (2,6)$ $[Negative\ Class]$

$$D7 = \sqrt{(5-2)^2 + (4-6)^2}$$

$$= \sqrt{9+4}$$

$$= \sqrt{13}$$

$$= 3.606$$

Point 8, $P8 = (8,1)$ $[Negative\ Class]$

$$D8 = \sqrt{(5-8)^2 + (4-1)^2}$$

$$= \sqrt{9+9}$$

$$= \sqrt{18}$$

$$= 4.243$$

As the distances keeps increases and also, we had calculated 8 Euclidean distances till now. So, there is no further need to calculate Euclidean distances for remaining points.

Above 8 Euclidean Distances, we can say that:

$$D1 < D2 < D3 < D4 < D5 < D6 = D7 < D8$$

So, the 5 nearest neighbors according to the Euclidean Distances are,

3 Negative Class points $P1 = (4,4),\ P4 = (5,1), P5 = (4,1)$

2 Positive Class points $P2 = (6,5),\ P3 = (7,3)$

(2)

Manhattan Distance, $d(x_i, x_j) = \sum_{m=1}^{D} |x_{im} - x_{jm}|$

Manhattan Distance Weighted, $mdw = \frac{1}{d^2}$

Taking again the test point (triangle), $TP = (5,4)$.

Point 1, $P1 = (4,4)$ $[Negative\ Class]$

Using Manhattan Distance:

$$d1 = |5 - 4| + |4 - 4|$$
$$= 1 + 0$$
$$= 1$$

Manhattan Distance Weighted, $\boldsymbol{mdw1 = \frac{1}{1^2} = 1}$

Point 2, $P2 = (6,5)$ $[Positive\ Class]$

Using Manhattan Distance:

$$d1 = |5 - 6| + |4 - 5|$$
$$= 1 + 1$$
$$= 2$$

Manhattan Distance Weighted, $\boldsymbol{mdw2 = \frac{1}{2^2} = \frac{1}{4} = 0.25}$

Point 3, $P3 = (7,3)$ $[Positive\ Class]$

Using Manhattan Distance:

$$d1 = |5 - 7| + |4 - 3|$$
$$= 2 + 1$$
$$= 3$$

Manhattan Distance Weighted, $\boldsymbol{mdw3 = \frac{1}{3^2} = \frac{1}{9} = 0.111}$

The Manhattan Distance Weighted 3- nearest neighbor are 1, 0.25, 0.111.

Part II

```
#Author : Janmejay Mohanty
import numpy as np
import pandas as pd
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)
testdata = pd.read_csv('test.csv')
testdata.head()
traindata = pd.read_csv('train.csv')
traindata.head()
X = testdata.drop(['actual-class','ID'],axis=1,)
y = testdata['actual-class']
y.head()
X.head()
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.30)
from sklearn.neighbors import KNeighborsClassifier
KNN = KNeighborsClassifier(n_neighbors=3)
KNN.fit(X_train,y_train)
predict = KNN.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
print(confusion_matrix(y_test,predict))
print(classification_report(y_test,predict))
def euclidean_dist(A, B):                    #Euclidean Distance function
   return np.sqrt(sum(np.square(A-B)))
def euclidean_weight(d):                     #Euclidean Weighted Function
   return 1/d**2
X = traindata.drop(['class'],axis=1,)
y = traindata['class']
y.head()
```

```
X.head()

X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.30)

KNN = KNeighborsClassifier(n_neighbors=3)

KNN.fit(X_train,y_train)

predict = KNN.predict(X_test)

print(confusion_matrix(y_test,predict))

print(classification_report(y_test,predict))

#test.csv has more accuracy as compared to train.csv
```