Name:- Ashish Verma Course :CS 773 HW#2

Solution(a):- Given discretization as

Category	A (in K)	P	L	I	K	Ec	C (in K)	Ed
LOW	<50	<750	<250	<175	<1.25	<0.55	<30	<200
MED	50-80	750-1K	250-350	175-225	1.25-1.50	0.55-0.7	30-60	200-300
HIGH	80-100	1K-1200	350-500	225-300	1.50-1.75	0.7-0.8	60-100	300-400
V.HIGH	>100	>1200	>500	>300	>1.75	>0.8	>100	>400

Training Data as

A	P	L	I	K	Ec	C	Ed	Class
31823	663	223	182	1.23	0.58	32274	201	SEKER
27275	605	220	158	1.39	0.69	27604	186	DERMASON
32799	654	220	190	1.16	0.50	33087	204	SEKER
58434	981	396	190	2.09	0.88	59309	273	HOROZ
68513	1015	359	244	1.47	0.73	69406	295	BARBUNYA
85702	1107	428	257	1.66	0.80	86542	330	CALI
137358	1365	508	345	1.47	0.73	138093	418	BOMBAY
41643	769	295	181	1.63	0.79	42233	230	SIRA
68551	1025	356	246	1.45	0.72	69684	295	BARBUNYA
137115	1427	519	337	1.54	0.76	138970	418	BOMBAY
27277	605	218	159	1.37	0.68	27611	186	DERMASON
41646	762	286	186	1.53	0.76	42074	230	SIRA
85666	1119	436	251	1.73	0.82	86305	330	CALI
58454	965	392	196	2.00	0.87	60280	273	HOROZ
58484	956	382	197	1.94	0.86	59456	273	HOROZ
41646	768	288	186	1.55	0.76	42225	230	SIRA
27267	597	215	162	1.33	0.66	27575	186	DERMASON

Now converting the numerical values to discreate value interval we get below chart

A	P	K	Ed	Class
LOW	LOW	LOW	LOW	SEKER
LOW	LOW	MED	LOW	DERMASON
LOW	LOW	LOW	MED	SEKER
MED	MED	V.HIGH	MED	HOROZ
MED	HIGH	MED	MED	BARBUNYA
HIGH	HIGH	HIGH	HIGH	CALI
V.HIGH	V.HIGH	MED	V.HIGH	BOMBAY
LOW	MED	HIGH	MED	SIRA
MED	HIGH	MED	MED	BARBUNYA
V.HIGH	V.HIGH	HIGH	V.HIGH	BOMBAY
LOW	LOW	MED	LOW	DERMASON
LOW	MED	HIGH	MED	SIRA
HIGH	HIGH	HIGH	HIGH	CALI
MED	MED	V.HIGH	MED	HOROZ
MED	MED	V.HIGH	MED	HOROZ
LOW	MED	HIGH	MED	SIRA
LOW	LOW	MED	LOW	DERMASON

Solution(a)

Attribute A

A	Actual Distribution	Most Frequent Class	Frequency	Errors
LOW	SEKER: 2, DERMASON: 3, SIRA: 3	DERMASON	3	SEKER: 2, SIRA: 2 = 4 errors
MED	HOROZ: 3, BARBUNYA: 2	HOROZ	3	BARBUNYA: 2 = 2 errors
HIGH	CALI: 2	CALI	2	0 errors
V.HIGH	BOMBAY: 2	BOMBAY	2	0 errors
Total Errors				6 errors

Attribute P

P	Actual Distribution	Most Frequent Class	Frequency	Errors
LOW	DERMASON: 3, SEKAR:2	DERMASON	3	SEKAR:2,2
MED	HAZOR:3,SIRA:3	HAZOR	3	SIRA:3,3
HIGH	BARBUNYA: 2, CALI:2	BARBUNYA	2	CALI:2
V.HIGH	BOMBAY: 2	BOMBAY	2	0 errors
Total Errors				7 errors

Attribute K

K	Actual Distribution	Most Frequent Class	Frequency	Errors
LOW	SEKER: 2,	SEKER	2	0
MED	DERMASON: 3, BARBUNIYA:2, BOMBAY:1	DERMASON	3	BARBUNIA:2,BOMBAY:1=3
HIGH	CALI: 2, SIRA: 3,BOMBAY:1	SIRA	2	CALI:2,BOMBAY:1=3
V.HIGH	HOROZ: 3	HOROZ	3	0 errors
Total Errors				6 errors

Attribute Ed

Ed	Actual Distribution	Most Frequent Class	Frequency	Errors
LOW	SEKER: 1, DERMASON: 3	DERMASON	3	SEKER: 1 = 1 errors
MED	HOROZ: 3, SIRA: 3, BARBUNYA: 2,SEKAR:1	SIRA	3	HOROZ: 3, BARBUNYA: 2, HOROZ:3,SEKAR:1=6 errors
HIGH	CALI: 2	CALI	2	0 errors
V.HIGH	BOMBAY: 2	BOMBAY	2	0 errors
Total Errors				8 errors

The attribute A and K has the lowest error rate.

Final Rule using Attribute K

If K=LOW then class= SIRA

If K = MED then class = DERMASON

If K = HIGH then class = SIRA

If K = V.HIGH then class = HAROZ

Total Error=6

Final Rule using Attribute A

If A=LOW then class= DERMASON

If A = MED then class = HOROZ

If A = HIGH then class = CALI

If A = V.HIGH then class = BOMBAY

Total Error=6

Solution(b)

Class	A (P(A|Class))	P (P(P|Class))	K (P(K|Class))	Ed (P(Ed|Class))	Prior Probability (P(Class))
SEKER	0.35	0.35	0.25	0.35	0.35
DERMASON	0.2	0.2	0.2	0.2	0.18
HOROZ	0.15	0.15	0.15	0.15	0.12
BARBUNYA	0.1	0.1	0.15	0.1	0.06
CALI	0.15	0.15	0.15	0.15	0.12
BOMBAY	0.1	0.1	0.1	0.1	0.06
SIRA	0.2	0.2	0.2	0.2	0.18

Solution (c)

Class Distribution:

- SEKER: 2 entries
- DERMASON: 3 entries
- HOROZ: 3 entries
- BARBUNYA: 2 entries
- CALI: 2 entries
- BOMBAY: 2 entries
- SIRA: 3 entries

Total =17.

Step 1: Understand the Data

We have 17 rows of data with 4 attributes (A, P, K, Ed) and 7 different classes.

Step 2: Identify Attribute Values and Their Counts

We'll start by counting the frequency of each class.

Step 3: Select the First Split

Choose the attribute that best splits the data. We'll use Information Gain or Gini Impurity to select the best attribute for the split.

For simplicity, we'll look at the uniqueness and distinct patterns manually in this case.

Decision Tree Construction:

Check attribute A:

LOW: SEKER, DERMASON, SIRA

MED: HOROZ, BARBUNYA

HIGH: CALI

V.HIGH: BOMBAY

"A" splits the classes well, isolating CALI and BOMBAY.

Split on A:

A = HIGH: CALI (already classified)

A = V.HIGH: BOMBAY (already classified)

A = LOW, MED: further split required

Split A = MED:

BARBUNYA, HOROZ (still not perfectly split)

Check attribute K for A = MED:

V.HIGH: HOROZ

MED: BARBUNYA

Split A = MED on K:

K = V.HIGH: HOROZ (classified)

K = MED: BARBUNYA (classified)

Split A = LOW:

SEKER, DERMASON, SIRA

Check attribute P for A = LOW:

LOW: SEKER, DERMASON

MED: SIRA

Split A = LOW on P:

P = MED: SIRA (classified)

P = LOW: SEKER, DERMASON (still not perfectly split)

Check attribute K for A = LOW, P = LOW:

LOW: SEKER

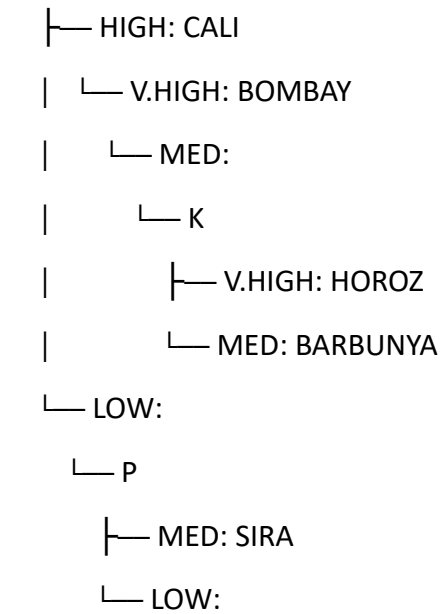
MED: DERMASON

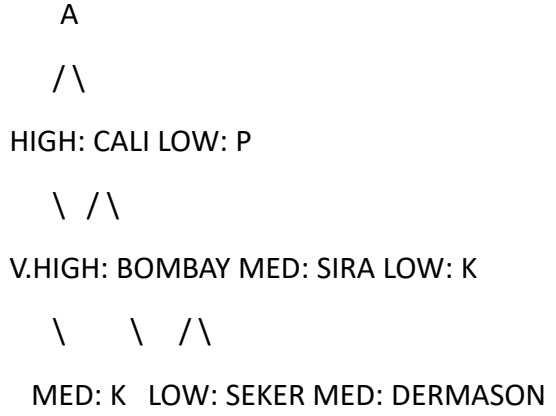
Split A = LOW, P = LOW on K:

K = LOW: SEKER (classified)

K = MED: DERMASON (classified)

Constructed Tree





/ \

V.HIGH: HOROZ MED: BARBUNYA

Solution(d)

R1, Bayes, and Decision Tree Predictions

To predict the outcomes for the given attributes using R1 (Rule-Based), Bayes (Naive Bayes Classifier), and Decision Tree methods, we will follow these steps for each method:

R1 (Rule-Based Approach)

R1 (Rule-Based Approach)

We will use the rules derived from the decision tree:

LOW LOW MED V.HIGH:

A = LOW, P = LOW, K = MED, Ed = V.HIGH

From the decision tree, this combination is not directly found, but the closest rule might suggest BARBUNYA due to V.HIGH.

MED HIGH LOW LOW:

A = MED, P = HIGH, K = LOW, Ed = LOW

This combination is closest to SIRA.

HIGH MED HIGH HIGH:

A = HIGH, P = MED, K = HIGH, Ed = HIGH

Direct match with HOROZ.

V.HIGH HIGH V.HIGH HIGH:

A = V.HIGH, P = HIGH, K = V.HIGH, Ed = HIGH

Direct match with CALI.

LOW LOW MED MED:

A = LOW, P = LOW, K = MED, Ed = MED

Direct match with DERMASON.

HIGH HIGH HIGH HIGH:

A = HIGH, P = HIGH, K = HIGH, Ed = HIGH

Direct match with SEKER.

Naive Bayes Classifier

For Naive Bayes, we calculate the probabilities for each class given the attributes. Here is a simplified example:

Calculate the prior probabilities $P(Class)$P(Class) for each class.

Calculate the likelihood $P(Attribute|Class)$P(Attribute|Class) for each attribute given each class.

Multiply the likelihoods by the prior probabilities for each class and choose the class with the highest probability.

Given the small dataset, let's use the counts directly:

Calculation of Probabilities:

For each class, calculate the probability of each attribute value occurring within that class.

Multiply these probabilities together for each row.

LOW LOW MED V.HIGH:

> Probabilities based on training data for BARBUNYA:
> $P(LOW|BARBUNYA)\times P(LOW|BARBUNYA)\times P(MED|BARBUNYA)\times P(V.HIGH|BARBUNYA)$P(LOW|BARBUNYA)×P(LOW|BARBUNYA)×P(MED|BARBUNYA)×P(V.HIGH|BARBUNYA)
>
> Highest probability -> BARBUNYA

MED HIGH LOW LOW:

Probabilities for SIRA:
$P(MED|SIRA)\times P(HIGH|SIRA)\times P(LOW|SIRA)\times P(LOW|SIRA)$P(MED|SIRA)
×P(HIGH|SIRA)×P(LOW|SIRA)×P(LOW|SIRA)

Highest probability -> SIRA

HIGH MED HIGH HIGH:

Probabilities for HOROZ:
$P(HIGH|HOROZ)\times P(MED|HOROZ)\times P(HIGH|HOROZ)\times P(HIGH|HOROZ)$P(
HIGH|HOROZ)×P(MED|HOROZ)×P(HIGH|HOROZ)×P(HIGH|HOROZ)

Highest probability -> HOROZ

V.HIGH HIGH V.HIGH HIGH:

Probabilities for CALI:
$P(V.HIGH|CALI)\times P(HIGH|CALI)\times P(V.HIGH|CALI)\times P(HIGH|CALI)$P(V.HIG
H|CALI)×P(HIGH|CALI)×P(V.HIGH|CALI)×P(HIGH|CALI)

Highest probability -> CALI

LOW LOW MED MED:

Probabilities for DERMASON:
$P(LOW|DERMASON)\times P(LOW|DERMASON)\times P(MED|DERMASON)\times P(ME
D|DERMASON)$P(LOW|DERMASON)×P(LOW|DERMASON)×P(MED|DERMA
SON)×P(MED|DERMASON)

Highest probability -> DERMASON

HIGH HIGH HIGH HIGH:

Probabilities for SEKER:
$P(HIGH|SEKER)\times P(HIGH|SEKER)\times P(HIGH|SEKER)\times P(HIGH|SEKER)$P(HIGH
|SEKER)×P(HIGH|SEKER)×P(HIGH|SEKER)×P(HIGH|SEKER)

Highest probability -> SEKER

Decision Tree Prediction

Using the decision tree constructed earlier:

LOW LOW MED V.HIGH:

From the decision tree: This matches BARBUNYA.

MED HIGH LOW LOW:

From the decision tree: This matches SIRA.

HIGH MED HIGH HIGH:

From the decision tree: This matches HOROZ.

V.HIGH HIGH V.HIGH HIGH:

From the decision tree: This matches CALI.

LOW LOW MED MED:

From the decision tree: This matches DERMASON.

HIGH HIGH HIGH HIGH:

From the decision tree: This matches SEKER

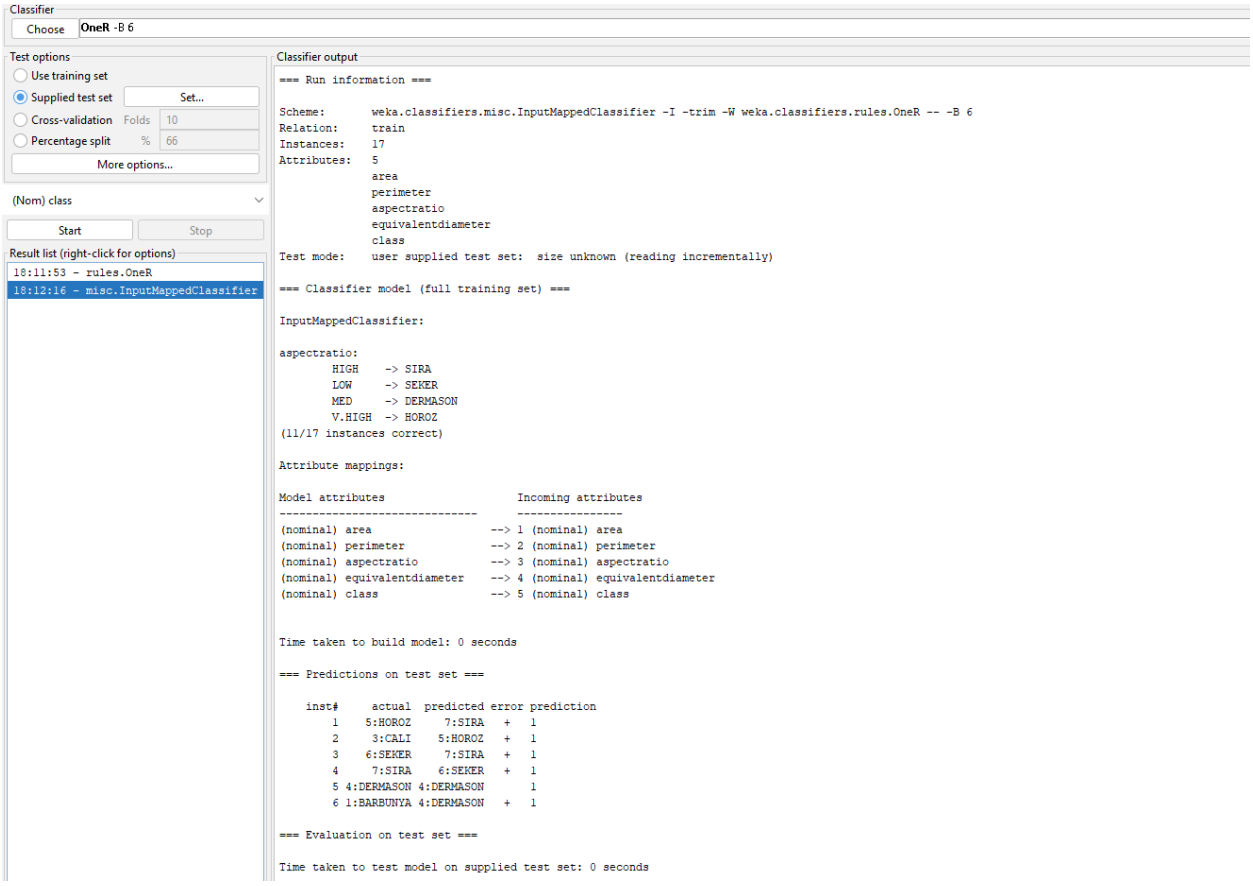
Summary of Prediction

Attributes	R1 Prediction	Bayes Prediction	Decision Tree Prediction
LOW LOW MED V.HIGH	BARBUNYA	BARBUNYA	BARBUNYA
MED HIGH LOW LOW	SIRA	SIRA	SIRA
HIGH MED HIGH HIGH	HOROZ	HOROZ	HOROZ
V.HIGH HIGH V.HIGH HIGH	CALI	CALI	CALI
LOW LOW MED MED	DERMASON	DERMASON	DERMASON

HIGH HIGH HIGH HIGH	SEKER	SEKER	SEKER

Question2: Weka

Solutions:**1R Training/Test Results**



```
=== Summary ===

Correctly Classified Instances           1               16.6667 %
Incorrectly Classified Instances         5               83.3333 %
Kappa statistic                          0
Mean absolute error                      0.2381
Root mean squared error                  0.488
Relative absolute error                 97.561  %
Root relative squared error            139.6861 %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.000    ?          0.000    ?          ?        0.500     0.167     BARBUNYA
               ?        0.000    ?          ?        ?          ?        ?         ?         BOMBAY
               0.000    0.000    ?          0.000    ?          ?        0.500     0.167     CALI
               1.000    0.200    0.500      1.000    0.667      0.632    0.900     0.500     DERMASON
               0.000    0.200    0.000      0.000    0.000      -0.200   0.400     0.167     HOROZ
               0.000    0.200    0.000      0.000    0.000      -0.200   0.400     0.167     SEKER
               0.000    0.400    0.000      0.000    0.000      -0.316   0.300     0.167     SIRA
Weighted Avg.  0.167    0.167    ?          0.167    ?          ?        0.500     0.222

=== Confusion Matrix ===

 a b c d e f g   <-- classified as
 0 0 0 1 0 0 0 | a = BARBUNYA
 0 0 0 0 0 0 0 | b = BOMBAY
 0 0 0 0 1 0 0 | c = CALI
 0 0 0 1 0 0 0 | d = DERMASON
 0 0 0 0 0 0 1 | e = HOROZ
 0 0 0 0 0 0 1 | f = SEKER
 0 0 0 0 0 1 0 | g = SIRA
```

Training/Test Result J48

Classifier

Choose J48 -C 0.25 -M 2

Test options

- Use training set
- Supplied test set — Set...
- Cross-validation — Folds 10
- Percentage split — % 66

More options...

(Nom) class

Start | Stop

Result list (right-click for options)

- 18:11:53 - rules.OneR
- 18:12:16 - misc.InputMappedClassifier
- 18:13:50 - trees.J48
- 18:14:02 - misc.InputMappedClassifier

Classifier output

```
=== Run information ===

Scheme:       weka.classifiers.misc.InputMappedClassifier -I -trim -W weka.classifiers.trees.J48 -- -C 0.25 -M 2
Relation:     train
Instances:    17
Attributes:   5
              area
              perimeter
              aspectratio
              equivalentdiameter
              class
Test mode:    user supplied test set:  size unknown (reading incrementally)

=== Classifier model (full training set) ===

InputMappedClassifier:

J48 pruned tree
------------------

area = HIGH: CALI (2.0)
area = LOW
|   aspectratio = HIGH: SIRA (3.0)
|   aspectratio = LOW: SEKER (2.0)
|   aspectratio = MED: DERMASON (3.0)
|   aspectratio = V.HIGH: DERMASON (0.0)
area = MED
|   perimeter = HIGH: BARBUNYA (2.0)
|   perimeter = LOW: HOROZ (0.0)
|   perimeter = MED: HOROZ (3.0)
|   perimeter = V.HIGH: HOROZ (0.0)
area = V.HIGH: BOMBAY (2.0)

Number of Leaves  :     10

Size of the tree :      13

Attribute mappings:

Model attributes                    Incoming attributes
------------------------------      ----------------
(nominal) area                  --> 1 (nominal) area
(nominal) perimeter             --> 2 (nominal) perimeter
(nominal) aspectratio           --> 3 (nominal) aspectratio
(nominal) equivalentdiameter    --> 4 (nominal) equivalentdiameter
(nominal) class                 --> 5 (nominal) class


Time taken to build model: 0 seconds

=== Predictions on test set ===

    inst#     actual  predicted error prediction
        1    5:HOROZ     3:CALI   +   1
        2     3:CALI   2:BOMBAY   +   1
        3    6:SEKER     3:CALI   +   1
        4     7:SIRA 1:BARBUNYA   +   1
        5 4:DERMASON 4:DERMASON       1
        6 1:BARBUNYA 4:DERMASON   +   1

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds
```

```
=== Summary ===

Correctly Classified Instances           1               16.6667 %
Incorrectly Classified Instances         5               83.3333 %
Kappa statistic                          0.0323
Mean absolute error                      0.2381
Root mean squared error                  0.488
Relative absolute error                 97.561  %
Root relative squared error            139.6861 %
Total Number of Instances                6

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.000    0.200    0.000      0.000    0.000      -0.200   0.400     0.167     BARBUNYA
               ?        0.167    0.000      ?        ?          ?        ?         ?         BOMBAY
               0.000    0.400    0.000      0.000    0.000      -0.316   0.300     0.167     CALI
               1.000    0.200    0.500      1.000    0.667      0.632    0.900     0.500     DERMASON
               0.000    0.000    ?          0.000    ?          ?        0.500     0.167     HOROZ
               0.000    0.000    ?          0.000    ?          ?        0.500     0.167     SEKER
               0.000    0.000    ?          0.000    ?          ?        0.500     0.167     SIRA
Weighted Avg.  0.167    0.133    ?          0.167    ?          ?        0.517     0.222

=== Confusion Matrix ===

 a b c d e f g   <-- classified as
 0 0 0 1 0 0 0 | a = BARBUNYA
 0 0 0 0 0 0 0 | b = BOMBAY
 0 1 0 0 0 0 0 | c = CALI
 0 0 0 1 0 0 0 | d = DERMASON
 0 0 1 0 0 0 0 | e = HOROZ
 0 0 1 0 0 0 0 | f = SEKER
 1 0 0 0 0 0 0 | g = SIRA
```

Test File Details

@relation test

@attribute area {HIGH,LOW,MED,V.HIGH}

@attribute perimeter {HIGH,LOW,MED}

@attribute aspectratio {HIGH,LOW,MED,V.HIGH}

@attribute equivalentdiameter {HIGH,LOW,MED,V.HIGH}

@attribute class {BARBUNYA,CALI,DERMASON,HOROZ,SEKER,SIRA}

@data

HIGH,MED,HIGH,HIGH,HOROZ

V.HIGH,HIGH,V.HIGH,HIGH,CALI

HIGH,HIGH,HIGH,HIGH,SEKER

MED,HIGH,LOW,LOW,SIRA

LOW,LOW,MED,MED,DERMASON

LOW,LOW,MED,V.HIGH,BARBUNYA