CS 422

DATA MINING

HOME WORK – 1

SUBMITTED BY :

SACHIN KRISHNA MURTHY

CWID : A20354077

Summary :

The following things has been done, analyzed, learnt and enclosed in this report.

- Definitions and working procedure of the classifiers and other options used in weka.
- Attributes and maximum, minimum, mean and Standard deviation for all the data sets.
- Obtaining classification accuracy by using two decision tree algorithms :
    - Simple cart classifier
    - Decision Stump classifier
- Analyzing and determining Classification accuracy by :
    - Changing the parameters for each of the classifiers
    - Using percentage split option and analyzing the accuracy for different percentage of training and test sets.
- Calculation of Classification accuracy by introducing Noise in two ways :
    - Adding some of the missing values to each data set and analyzing the results by using two decision tree algorithms.
    - Introducing missing values i,e misclassifying some of the values in the data set and analyzing the results by using two decision tree algorithms.
- All the results obtained from the classifiers by accomplishing the above mentioned steps has been analyzed and inferred.
- Class Distribution for each of the data set.
    - Size of training and test set.
    - Analysis of the same in cross validation.

Introduction : Definitions and Functionalities

Simple Cart Classifier :

- Simple Cart is one of the decision tree algorithms and it makes use of decision tree as a predictive model which tries to map observations about the object to the resolutions about the object's target value.
- The steps involved are :
    - Initial formation of Binary Tree
    - Minimization of error on each leaf node of the tree.
- This classifier will not make any decisions, but they only delineate the data and also the resulting classification tree from this classifier can be used as input for making decisions.

Parameters of Simple Cart Classifier :

- debug : Using this parameter, Classifier may output additional info to the console when set to true.
- heuristic : Heuristic search is used for binary split for nominal attributes in multi-class problems when set to yes (default yes).
- minNumObj : It is possible to determine the minimal number of observations at the terminal nodes (default 2).
- numFoldsPruning : This parameter gets you the details of the number of folds in the internal cross-validation (default 5).
- seed :The random number seed to be used.
- sizePer :The percentage of the training set size (0-1, 0 not included).
- useOneSE : Use the 1SE rule to make pruning decisoin.
- usePrune : Use minimal cost-complexity pruning (default yes).

Decision Stump Classifier :

- Decision Stump is another decision tree algorithm that consists of only a single internal node.
- Hence it makes predictions using the value of a single attribute.

Parameters of Decision Stump Classifier :

- debug : Using this parameter, Classifier may output additional info to the console when set to true.

10 Fold Cross Validation :

Steps involved in the process of 10 fold cross validation :

- Initially the data set is divided into 10 parts.
- Among which 9 of them are used for training, the rest 1 is used for testing and the result is determined.
- Again this process is repeated by considering another 9 for training, the remaining 1 for testing and the result is calculated.
- Finally all the 10 results obtained from the above process will be averaged.

Percentage Split :

- Percentage split option helps you divide the data set into training and test set manually based on the proportion given by the user.
- For example, consider a data set of 100 instances and you are specifying the percentage of 70%. Then 70% of the data in the data set will be used for training and the remaining 30% is used as test data.

Dataset : Iris

- **Attributes :**
  - Number of Attributes : 5

- **Minimum, Maximum, Mean and Standard Deviation of the Attributes :**

Sl No.	Attribute	Minimum Value	Maximum Value	Mean	Standard Deviation
1.	Sepal Length	4.3	7.9	5.843	0.828
2.	Sepal Width	2	4.4	3.054	0.434
3.	Petal Length	1	6.9	3.759	1.764
4.	Petal Width	0.1	2.5	1.199	0.763
5.	Class				

- **Class Attribute :**

Sl No.	Label	Count
1.	Iris-setosa	50
2.	Iris-versicolor	50
3.	Iris-virginica	50

- **Decision Tree Algorithms**

- **Using Simple Cart** :

- **Test Mode : 10 fold cross validation**

```
CART Decision Tree

petallength < 2.45: Iris-setosa(50.0/0.0)
petallength >= 2.45
|  petalwidth < 1.75
|  |  petallength < 4.95: Iris-versicolor(47.0/1.0)
|  |  petallength >= 4.95
|  |  |  petalwidth < 1.55: Iris-virginica(3.0/0.0)
|  |  |  petalwidth >= 1.55: Iris-versicolor(2.0/1.0)
|  petalwidth >= 1.75: Iris-virginica(45.0/1.0)
```

Fig 2.1 : Decision Tree for Simple Cart Algorithm

- Number of Leaf Nodes : 5
- Size of the Tree : 9

➢ **Classification Accuracy :**

- Total Number of Instances : 150
- Correctly Classified Instances : 143
- Incorrectly Classified Instances : 7
- Percentage of Correctly Classified Instances : 95.3333
- Percentage of Incorrectly Classified Instances : 4.6667

➢ **Detailed Accuracy :**

➢ **Accuracy Calculation:**

- True Positive Rate = True Positive / (True Positive + False Negative)
- False Positive Rate = False Positive / (False Positive + True Negative)
- Precision = True Positive / (True Positive + False Positive)
- Recall = True Positive / (True Positive + False Negative)
- F-Measure = 2* (Precision * Recall) / (Precision + Recall)
- Accuracy = (Number of Correctly classified Instances) / (Total number of Instances)

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Iris-setosa
                 0.94      0.04       0.922     0.94      0.931      0.944    Iris-versicolor
                 0.92      0.03       0.939     0.92      0.929      0.949    Iris-virginica
Weighted Avg.    0.953     0.023      0.953     0.953     0.953      0.964
```

Fig 2.2: Detailed Accuracy by Class for Iris (Using Simple Cart)

- Addition to classification accuracy there are several measures to determine the accuracy of the classifier as mentioned in the above figure.

➢ **Confusion Matrix :**

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  4 46 |  c = Iris-virginica
```

Fig 2.3: Confusion Matrix for Iris (Using Simple Cart)

➢ **Analysis :**

- From the above figure of confusion matrix we will be able to determine the values of True positive, true negative, false positive and false negative values.
- True Positive (TP) : For any class ‘x’, the number of instances of ‘x’ that are classified as ‘x’.
- False Positive (FP) : For any class ‘x’, the number of instances that does not belong to ‘x’ and are classified as ‘x’.
- False Negative (FN) : For any class ‘x’, the number of instances that belong to ‘x’ and are not classified as ‘x’.
- True Negative (TN) : For any class ‘x’, the number of instances that does not belong to ‘x’ and are not classified as ‘x’.

Considering above Confusion Matrix (Fig 2.3) :

- For class Iris-setosa all the 50 instances are classified as Iris-setosa.
- For class Iris-versicolor, out of 50 instances, 47 instances has been classified as Iris-versicolor and the remaining 3 is misclassified as Iris-virginica.

- For class Iris-virginica, out of 50 instances, 46 instances has been classified as Iris-virginica and the remaining 4 is misclassified as Iris-versicolor.

➢ **Using Decision Stump :**

➢ **Test Mode : 10 fold cross validation**

- **Classification for Iris :**

```
Classifications

petallength <= 2.45 : Iris-setosa
petallength > 2.45 : Iris-versicolor
petallength is missing : Iris-setosa
```

Fig 2.4: Classification for Iris (using decision stump)

➢ **Classification Accuracy :**

- Total Number of Instances : 150
- Correctly Classified Instances : 100
- Incorrectly Classified Instances : 50
- Percentage of Correctly Classified Instances : 66.6667
- Percentage of Incorrectly Classified Instances : 33.3333

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        Iris-setosa
                 1         0.5        0.5       1         0.667      0.75     Iris-versicolor
                 0         0          0         0         0          0.75     Iris-virginica
Weighted Avg.    0.667     0.167      0.5       0.667     0.556      0.833
```

Fig 2.5: Details Accuracy by class for Iris (Using decision Stump)

- By looking at the above results, the classification accuracy is approximately equal to 67% which is very less as one third of the data set is misclassified.

➢ **Confusion Matrix :**

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
 50  0  0 |  a = Iris-setosa
  0 50  0 |  b = Iris-versicolor
  0 50  0 |  c = Iris-virginica
```

Fig 2.6 : Confusion Matrix for Iris (Using Decision Stump)

- **Observations:**

- For class Iris-setosa all the 50 instances are classified as Iris-setosa.
- For class Iris-versicolor all the 50 instances are classified as Iris-versicolor.
- But for class Iris-virginica, out of 50 instances, none of them has been classified correctly as the classification algorithm does not have any conditions for prediction of Iris-virginica.

- **Analysis :**

- Classification Accuracy denotes the number of correctly classified instances by total number of instances.
- It is however subjected to vary upon addition of noise and missing values.
- In the case of Iris, using simple cart classifier, the classification accuracy is significant with very less misprediction.
- For Iris data set, Classification accuracy using simple cart classifier is 95.3333% while it is 66.6667% using decision stump classifier.
- From the decision tree (Fig 2.1) and classification (Fig 2.4) it can be noticed that decision stump uses only a single attribute for classification whereas simple cart uses two.
- For the above reason simple cart classifier creates decision tree for the classification of all the three classes whereas decision stump classifier has it for only two classes.
- It can inferred that the accuracy for decision stump is less as it does not contain algorithm for a class.
- Thus, in this case the performance of simple cart classifier is better when compared with decision stump classifier.

- **Classification Accuracy for Different Training and Test Sets**

- **Using Simple Cart :**

Simple Cart	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	120	90	75	60	30
Correctly Classified Instances	111	86	71	55	29
Incorrectly Classified Instances	9	4	4	5	1

Percentage of Correctly Classified Instances	92.5	95.5556	94.6667	91.6667	96.6667
Percentage of Incorrectly Classified Instances	7.5	4.4444	5.3333	8.3333	3.3333

- **Using Decision Stump**

Decision Stump	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	120	90	75	60	30
Correctly Classified Instances	78	58	48	38	20
Incorrectly Classified Instances	42	32	27	22	10
Percentage of Correctly Classified Instances	65	64.4444	64	63.3333	66.6667
Percentage of Incorrectly Classified Instances	35	35.5556	36	36.6667	33.3333

- **Analysis :**

- The accuracy mostly depends on the complexity and parameters of the algorithm.
- From the above tables it can noticed that for 80 vs 20 set the percentage of classification accuracy is higher.
- Using simple cart we have observed the classification accuracy of 96.6667% and using decision stump it is 66.6667%.
- Hence we can infer that higher the number of training set, higher is the classification accuracy.
- There is no change in the number of leaf nodes and the size of the tree even when done with different percentage of training and test sets.

- **By Changing the parameters of the Classifiers :**

- **Using Simple Cart :**

- Debug

Debug value	True	False
Correctly Classified Instances	143	143
Classification Accuracy	95.3333	95.3333
Number of Leaf nodes	5	5
Size of the tree	9	9

- Heuristic Value

heuristic value	True	False
Correctly Classified Instances	143	143
Classification Accuracy	95.3333	95.3333
Number of Leaf nodes	5	5
Size of the tree	9	9

- minNumObject

minNumObject value	0.0	1.0	2.0	3.0	4.0	5.0	10.0
Correctly Classified Instances	138	138	143	141	140	136	136
Classification Accuracy	92	92	95.3333	94	93.3333	90.6667	90.6667
Number of Leaf nodes	3	3	5	5	4	4	3
Size of the tree	5	5	9	9	7	7	5

- numFoldsPruning

numFoldsPruning value	2	4	5	6	8	10
Correctly Classified Instances	141	143	143	142	142	144
Classification Accuracy	94	95.3333	95.3333	94.6667	94.6667	96
Number of Leaf nodes	4	5	5	5	5	5
Size of the tree	7	9	9	9	9	9

- seed Value

seed value	0	1	2	4	5	10
Correctly Classified Instances	142	143	142	142	142	142
Classification Accuracy	94.6667	95.3333	94.6667	94.6667	94.6667	94.6667
Number of Leaf nodes	4	5	3	5	5	5
Size of the tree	7	9	5	9	9	9

- sizePer

sizePer value	0.1	0.3	0.5	0.7	0.9	1.0
Correctly Classified Instances	137	135	140	141	142	143
Classification Accuracy	91.3333	90	93.3333	94	94.6667	95.3333
Number of Leaf nodes	3	3	3	3	3	5
Size of the tree	5	5	5	5	5	9

- useOneSE value

useOneSE value	True	False
Correctly Classified Instances	141	143
Classification Accuracy	94	95.3333
Number of Leaf nodes	3	5
Size of the tree	5	9

- usePrune

usePrune value	True	False
Correctly Classified Instances	143	142
Classification Accuracy	95.3333	94.6667
Number of Leaf nodes	5	6
Size of the tree	9	11

➢ **Analysis :**

- As the parameter debug is used to get additional information in the console it does not have any effect on the accuracy.
- Changing the heuristic value does not lead to any change in the accuracy.
- For the minNumObject, numFoldsPruning and seed the classification accuracy is higher for the value of 2.0, 5 and 1. Also as the value goes away from 2.0, 5 and 1 the accuracy tends to decrease forming a bell curve.
- But for numFoldsPruning at the value of 10 and above the accuracy gets higher.
- As the value for sizePer decreases the accuracy also decreases and the accuracy is higher for the value of 1.0.
- It is observed that the accuracy is higher when the value of useOneSE is said to false and usePrune is said to true.
- Also we can see the changes in size of the tree and number of nodes in the above tables for different parameters. This tells us that the parameters impact decision tree creation.

➢ **Using Decision Stump**

- debug

Debug value	True	False
Correctly Classified Instances	100	100
Classification Accuracy	66.6667	66.6667

- Changing the value of debug does not have any effect on accuracy or no the decision tree.

➢ **Noise :**

➢ **Adding Missing Values :**

➢ **Using Simple Cart :**

Relation: iris

No.	sepallength Numeric	sepalwidth Numeric	petallength Numeric	petalwidth Numeric	class Nominal
1	5.1	3.5	1.4	0.2	Iris-...
2	4.9	3.0	1.4	0.2	Iris-...
3	4.7		1.3	0.2	Iris-...
4	4.6	3.1	1.5	0.2	Iris-...
75	6.4	2.9	4.3	1.3	Iris-...
76	6.6	3.0	4.4	1.4	Iris-...
77		2.8	4.8	1.4	Iris-...
78	6.7	3.0	5.0	1.7	Iris-...
79	6.0	2.9	4.5	1.5	Iris-...
80	5.7	2.6	3.5	1.0	Iris-...

19	5.7	3.8	1.7	0.3	Iris-...
20	5.1	3.8	1.5	0.3	Iris-...
21	5.4	3.4		0.2	Iris-...
22	5.1	3.7	1.5	0.4	Iris-...
23	4.6	3.6	1.0	0.2	Iris-...
24	5.1	3.3	1.7	0.5	Iris-...
121	6.9	3.2	5.7	2.3	Iris-...
122	5.6	2.8	4.9	2.0	Iris-...
123	7.7	2.8	6.7	2.0	Iris-...
124	6.3		4.9	1.8	Iris-...
125	6.7	3.3	5.7	2.1	Iris-...
126	7.2	3.2	6.0	1.8	Iris-...

Fig 2.7 : Iris Dataset after introducing missing values

➢ **Classification Accuracy (After introducing missing values) :**

- Total Number of Instances : 150
- Correctly Classified Instances : 141
- Incorrectly Classified Instances : 09
- Percentage of Correctly Classified Instances : 94
- Percentage of Incorrectly Classified Instances : 6

➢ **Misclassifying some of the examples :**

2	4.9	3.0	1.4	0.2	Iris-...
3	4.7	3.2	1.3	0.2	I...
4	4.6	3.1	1.5	0.2	
5	5.0	3.6	1.4	0.2	
6	5.4	3.9	1.7	0.4	Iris-se
7	4.6	3.4	1.4	0.3	Iris-ve
8	5.0	3.4	1.5	0.2	Iris-vir
9	4.4	2.9	1.4	0.2	
10	4.9	3.1	1.5	0.1	Iris-...

16	5.7	4.4	1.5	0.4	Iris-...
17	5.4	3.9	1.3	0.4	I...
18	5.1	3.5	1.4	0.3	
19	5.7	3.8	1.7	0.3	
20	5.1	3.8	1.5	0.3	Iris-se
21	5.4	3.4	1.7	0.2	Iris-ve
22	5.1	3.7	1.5	0.4	Iris-vir
23	4.6	3.6	1.0	0.2	
24	5.1	3.3	1.7	0.5	Iris-...

52	6.4	3.2	4.5	1.5	Iris-...
53	6.9	3.1	4.9	1.5	I...
54	5.5	2.3	4.0	1.3	
55	6.5	2.8	4.6	1.5	
56	5.7	2.8	4.5	1.3	Iris-se
57	6.3	3.3	4.7	1.6	Iris-ve
58	4.9	2.4	3.3	1.0	Iris-vir
59	6.6	2.9	4.6	1.3	
60	5.2	2.7	3.9	1.4	Iris-...

106	7.6	3.0	6.6	2.1	Iris-...
107	4.9	2.5	4.5	1.7	I...
108	7.3	2.9	6.3	1.8	
109	6.7	2.5	5.8	1.8	
110	7.2	3.6	6.1	2.5	Iris-se
111	6.5	3.2	5.1	2.0	Iris-ve
112	6.4	2.7	5.3	1.9	Iris-vir
113	6.8	3.0	5.5	2.1	
114	5.7	2.5	5.0	2.0	Iris-...

Fig 2.8 : Noise after Misclassifying

➢ **Using Simple Cart:**

➢ **Classification Accuracy (After misclassifying)**

- Total Number of Instances : 150

- Correctly Classified Instances : 137
- Incorrectly Classified Instances : 13
- Percentage of Correctly Classified Instances : 91.3333
- Percentage of Incorrectly Classified Instances : 8.6667

➢ **Analysis :**

- Missing values were introduced in four places as shown in fig 2.7.
- When missing values are introduced to the data set, the accuracy dropped down to 94%, with a margin of 1.0% from the original accuracy of the data set.
- The data was misclassified in four places as shown in fig 2.8.
- When misclassified some of the examples, the accuracy dropped down to 91.3333%, with a margin of 4% from the original accuracy of the data set.

➢ **Using Decision Stump :**

➢ **Classification Accuracy (After introducing missing values) :**

- Total Number of Instances : 150
- Correctly Classified Instances : 100
- Incorrectly Classified Instances : 50
- Percentage of Correctly Classified Instances : 66.6667
- Percentage of Incorrectly Classified Instances : 33.3333

➢ **Classification Accuracy (After misclassifying) :**

- Total Number of Instances : 150
- Correctly Classified Instances : 96
- Incorrectly Classified Instances : 54
- Percentage of Correctly Classified Instances : 64
- Percentage of Incorrectly Classified Instances : 36

➢ **Analysis :**

- When missing values were introduced there was no change in the accuracy.
- When some of the data was misclassified we could observe a 2.0% drop in the accuracy when compared with the accuracy of original data set.
- Missing values in the data set does not make much difference in the classification accuracy.
- But misclassifying some of the examples in the data set can reduce the accuracy.
- If the number of missing values raises then we might observe a slight drop in accuracy.
- We can infer that missing values are not as harmful as misclassifying.

- **Class Distribution :**

- Class Distribution for this Iris dataset is 1/3 : 1/3 : 1/3.
- The observations are evenly distributed among the three classes.
- Yes, it matters for our experiment. Accuracy is minimized when the distribution is uniform.

- For this Iris data set, the total number of instances is 150 of which 90% will be training set and the rest 10% will be test set for each of the 10 iterations.
- Number of Training set : 135
- Number of Test set : 15

- When cross validation is done the data is randomly divided into training and test set.
- Both the Training and Test set does not retain the same Class Distribution as that of full data set.
- No, it does not matter. If class distribution across all folds is same, then cross-validated accuracy estimate might be biased and not really estimate generalization accuracy.

Dataset : Vote

- **Attributes :**
- Number of Attributes : 17

Sl No.	Attribute Name	Label	Count
1.	Handicapped-infants	n	236
		y	187
2.	Water-project-cost-sharing	n	192
		y	195
3.	Adoption-of-the-budget-resolution	n	171
		y	253
4.	Physician-fee-freeze	n	247
		y	177
5.	El-salvador-aid	n	208
		y	212
6.	Religious-groups-in-schools	n	152
		y	272
7.	Anti-satellite-test-ban	n	182
		y	239

8.	Aid-to-nicaraquan-contras	n	178
		y	242
9.	mx-missile	n	206
		y	207
10.	Immigration	n	212
		y	216
11.	Synfuels-corporation-cutback	n	264
		y	150
12.	Education-spending	n	233
		y	171
13.	Superfund-right-to-sue	n	201
		y	209
14.	Crime	n	170
		y	248
15.	Duty-free-exports	n	233
		y	174
16.	Export-administration-act-south-africa	n	62
		y	269
17.	Class	democrat	267
		republican	168

- **Decision Tree Algorithms**

- **Using Simple Cart** :

  - Test Mode : 10 fold cross validation

```
CART Decision Tree

physician-fee-freeze=(y)
|  synfuels-corporation-cutback=(n): republican(141.7/4.0)
|  synfuels-corporation-cutback!=(n)
|  |  mx-missile=(n)
|  |  |  adoption-of-the-budget-resolution=(n): republican(19.28/3.31)
|  |  |  adoption-of-the-budget-resolution!=(n)
|  |  |  |  anti-satellite-test-ban=(y): republican(2.2/0.0)
|  |  |  |  anti-satellite-test-ban!=(y): democrat(5.01/0.02)
|  |  mx-missile!=(n): democrat(4.99/1.02)
physician-fee-freeze!=(y): democrat(249.66/3.74)
```

Fig 3.1 : Decision Tree for Simple Cart Classifier

- Number of leaf nodes : 6
- Size of the tree : 11

➢ **Classification Accuracy :**

- Total Number of Instances : 435
- Correctly Classified Instances : 415
- Incorrectly Classified Instances : 20
- Percentage of Correctly Classified Instances : 95.4023
- Percentage of Incorrectly Classified Instances : 4.5977

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.955     0.048      0.97      0.955     0.962      0.967    democrat
                 0.952     0.045      0.93      0.952     0.941      0.967    republican
Weighted Avg.    0.954     0.047      0.954     0.954     0.954      0.967
```

Fig 3.2 : Detailed Accuracy by Class for Vote (Using Simple Cart)

- Here the classification accuracy is around 95% which is remarkable.

➢ **Confusion Matrix :**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 255  12 |   a = democrat
   8 160 |   b = republican
```

Fig 3.3 : Confusion Matrix for Vote Dataset (Using Simple Cart)

- **Observation :**

- For class democrat out of 267 instances, 255 instances are classified as democrat and the rest 12 are classified as republican.
- For class republican, out of 168 instances, 8 has been misclassified as democrat.

- **Using Decision Stump :**

- **Classifications for Vote dataset :**

```
Classifications

physician-fee-freeze = n : democrat
physician-fee-freeze != n : republican
physician-fee-freeze is missing : democrat
```

Fig 3.4 : Classifications for Vote (Using Decision Stump)

- **Classification Accuracy :**
  - Total Number of Instances : 435
  - Correctly Classified Instances : 416
  - Incorrectly Classified Instances : 19
  - Percentage of Correctly Classified Instances : 95.6322
  - Percentage of Incorrectly Classified Instances : 4.3678

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.948     0.03       0.981     0.948     0.964      0.951    democrat
                 0.97      0.052      0.921     0.97      0.945      0.951    republican
Weighted Avg.    0.956     0.039      0.958     0.956     0.957      0.951
```

Fig 3.5 : Detailed accuracy by class for Vote (Using Decision Stump)

- The classification accuracy in this case is 95.6322 which is also remarkable.

➢ **Confusion Matrix :**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 253  14 |   a = democrat
   5 163 |   b = republican
```

Fig 3.6 : Confusion Matrix for Vote(Using decision Stump)

➢ **Observation :**

- For class democrat out of 267 instances, 253 instances are classified as democrat and the rest 14 are classified as republican.
- For class republican, out of 168 instances, 5 has been misclassified as democrat.

➢ **Analysis :**

- For Vote data set, Classification accuracy using simple cart classifier is 95.4023 % while it is 95.6322 % using decision stump classifier.
- The difference in accuracies of both the classifiers vary by a very slight margin of 0.23 which is negligible.

➢ **Classification Accuracy for Different Training and Test Sets**

➢ **Using Simple Cart :**

Simple Cart	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	348	261	217	174	87
Correctly Classified Instances	334	250	210	171	82
Incorrectly Classified Instances	14	11	7	3	5
Percentage of Correctly Classified Instances	95.977	95.7854	96.7742	98.2759	94.2529
Percentage of Incorrectly Classified Instances	4.023	4.2146	3.2258	1.7241	5.7471

➢ **Using Decision Stump :**

Decision Stump	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	348	261	217	174	87
Correctly Classified Instances	334	250	210	169	82

Incorrectly Classified Instances	14	11	7	5	5
Percentage of Correctly Classified Instances	95.977	95.7854	96.7742	97.1264	94.2529
Percentage of Incorrectly Classified Instances	4.2023	4.2146	3.2258	2.8736	5.7471

- **Analysis :**

- From the above tables it can noticed that for 60 vs 40 set the percentage of classification accuracy is higher followed by 50 vs 50.
- Using simple cart we have observed the classification accuracy of 98.2759% and using decision stump it is 97.1264%.
- In this we can infer that higher number of training set as well as its ratio with number of testing set nearer to 1, results in higher classification accuracy.
- There is no change in the number of leaf nodes and the size of the tree even when done with different percentage of training and test sets.

- **By Changing the parameters of the Classifiers :**

- **Using Simple Cart :**

- Debug :

Debug value	True	False
Correctly Classified Instances	415	415
Classification Accuracy	95.4023	95.4023
Number of Leaf nodes	66	6
Size of the tree	11	11

- Heuristic values :

heuristic value	True	False
Correctly Classified Instances	415	415
Classification Accuracy	95.4023	95.4023
Number of Leaf nodes	6	6
Size of the tree	11	11

- minNumObject :

minNumObject value	0.0	1.0	2.0	3.0	4.0	5.0	10.0
Correctly Classified Instances	415	415	415	413	417	416	414
Classification Accuracy	95.4023	95.4023	95.4023	94.9425	95.8621	95.6322	95.1724

Number of Leaf nodes	8	8	8	11	7	5	5
Size of the tree	15	15	15	21	13	9	9

- numFoldsPruning :

numFoldsPruning value	2	4	5	6	8	10
Correctly Classified Instances	416	415	415	414	415	419
Classification Accuracy	95.6322	95.4023	95.4023	95.1724	95.4023	96.3218
Number of Leaf nodes	2	2	6	4	6	6
Size of the tree	3	3	11	11	11	11

- seed :

seed value	0	1	2	4	5	10
Correctly Classified Instances	415	415	416	417	413	415
Classification Accuracy	95.4023	95.4023	95.6322	95.8621	94.9424	95.4023
Number of Leaf nodes	6	6	6	6	6	2
Size of the tree	11	11	11	11	11	13

- sizePer values :

sizePer value	0.1	0.3	0.5	0.7	0.9	1
Correctly Classified Instances	416	415	416	416	416	415
Classification Accuracy	95.6322	95.4023	95.6322	95.6322	95.6322	95.4023
Number of Leaf nodes	2	5	2	2	2	6
Size of the tree	3	9	3	3	3	11

- useOneSE values :

useOneSE value	True	False
Correctly Classified Instances	416	415
Classification Accuracy	95.6322	95.4023
Number of Leaf nodes	5	6
Size of the tree	9	11

- usePrune :

usePrune value	True	False
Correctly Classified Instances	415	413
Classification Accuracy	95.4023	94.9425
Number of Leaf nodes	6	51
Size of the tree	11	101

- **Analysis :**

- Debug and heuristic parameters does not have any effect on accuracy.
- The parameters minNumObject and sizePer forms uneven accuracy curve for different values.
- Also the highest accuracy for minNumObject is at the value 4.0 and there is a very slight difference in accuracy for different values for the parameter sizePer.
- The parameter numFoldsPruning forms an inverse bell curve for different values and has the lowest accuracy at value 6.
- useOneSE has the highest accuracy when its value is set to true and usePrune has the highest accuracy when its value is said to false.
- The parameter seed forms a bell curve for different values and has the highest accuracy at value 4.

- **Using decision Stump :**

- debug

Debug value	True	False
Correctly Classified Instances	416	416
Classification Accuracy	95.6322	95.6322

- Change in debug value does not affect the accuracy.

- **Noise :**

- **After Adding Missing values :**

- **Using Simple Cart :**

Relation: vote

No.	handicapped-infants Nominal	water-project-cost-sharing Nominal	adoption-of-the-budget-resolution Nominal
1	n	y	n
2	n	y	n
3		y	
4	n		y
5		y	y
6	n	y	y
7	n		n
8	n	y	n

- **Classification Accuracy :**

Correct instances	414
Incorrect instances	21
Correct%	95.1724
Incorrect%	4.8276

Number of leaf nodes	6
Size of tree	11

- **After Misclassifying :**

- **Change in the value of attributes after misclassifying :**

Attribute	**Label**	**Count**
Adoption rate budget	n	169
	y	255
Anti satellite test ban	n	179
	y	242
Physician fee freeze	n	250
	y	174

- **Classification Accuracy :**

Correct instances	410
Incorrect instances	25
Correct%	94.2529
Incorrect%	5.7471
Number of leaf nodes	13
Size of tree	25

- **Using Decision Stump (After Adding Missing Values) :**

  - **Classification Accuracy :**

Correct instances	416
Incorrect instances	19
Correct%	95.6322
Incorrect%	4.3678

  - **After Misclassifying :**

  - **Classification Accuracy :**

Correct instances	413
Incorrect instances	22
Correct%	94.9425
Incorrect%	5.0575

- **Analysis :**

- For this vote dataset, for simple cart classifier the accuracy reduces from 95.4023 in the original to 95.1724 when missing values are introduced, but reduces to 94.2529 when data is misclassified.
- For decision Stump classifier the accuracy of 95.6322 in the original remains same when missing values are introduced, but reduces to 94.9425 when data is misclassified.
- As already discussed for Iris dataset, missing values have very less or no impact on the accuracy of the dataset, while noise reduces the accuracy of the dataset.
- In this case also, there is same result as in Iris. As the number of missing values increase, the accuracy might further decrease.
- For this dataset, the original accuracy might have increased had there been less or missing values in the original dataset.

- **Class Distribution :**

- Class Distribution for this Vote dataset is 0.61 : 0.39.
- Here, the observations are unevenly distributed among the two classes.
- Yes, Class Distribution matters for our experiment. Accuracy is minimized when the distribution is uniform and uneven classes tends to have more accuracy.

- For this Vote data set, the total number of instances is 435 of which 90% will be training set and the rest 10% will be test set for each of the 10 iterations.
- Number of Training set : 392
- Number of Test set : 43

- Data is randomly divided into training and test set when the cross validation is done.
- Class Distribution will be not the same as that of full data set in Training and Test set.
- No, it does not matter. If class distribution across all folds is same, then cross-validated accuracy estimate might be biased and not really estimate generalization accuracy.

Dataset : Labor

- **Attributes :**

- Number of attributes : 17

Sl No.	Attribute Name	Minimum Value	Maximum Value	Mean	Standard Deviation	Label	Count
1.	Duration	1	3	2.161	0.708	-	-

2.	Wage-increase-first-year	2	7	3.804	1.371	-	-
3.	Wage-increase-second-year	2	7	3.972	1.164	-	-
4.	Wage-increase-third-year	2	5.1	3.913	1.304	-	-
5.	Cost-of-living-adjustment	-	-	-	-	None	22
						Tcf	8
						Ts	7
6.	Working-hours	27	40	38.039	2.506	-	-
7.	Pension	-	-	-	-	None	11
						Ret-allw	4
						Empl-contr	12
8.	Standby-pay	2	14	7.444	5.028	-	-
9.	Shift-differential	0	25	4.871	4.544	-	-
10.	Education-allowance	-	-	-	-	Yes	10
						No	12
11.	Statutory-holidays	9	15	11.094	1.26	-	-
12.	Vacation	-	-	-	-	Below-average	18
						Average	17
						generous	16
13.	Long-term-disability-assistance	-	-	-	-	Yes	20
						No	8

14.	Contribution-to-dental-plan	-	-	-	-	None	9
						Half	15
						full	13
15.	Bereavement-assitance	-	-	-	-	Yes	27
						No	3
16.	Contribution-to-health-plan	-	-	-	-	None	8
						Half	9
						full	20
17.	Class	-	-	-	-	Bad	20
						Good	37

- **Decision Tree Algorithms :**

- **Using Simple Cart :**

- Test Mode : 10 fold cross validation

```
CART Decision Tree

wage-increase-first-year < 2.65: bad(13.0/2.26)
wage-increase-first-year >= 2.65: good(34.73/7.0)
```

Fig 4.1 : Decision Tree for Labor(Using Simple Cart)

- Number of leaf nodes : 2
- Size of the tree : 3

- **Classification Accuracy:**

- Total Number of Instances : 57
- Correctly Classified Instances : 45
- Incorrectly Classified Instances : 12
- Percentage of Correctly Classified Instances : 78.9474
- Percentage of Incorrectly Classified Instances : 21.0526

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.75      0.189      0.682     0.75      0.714      0.748    bad
                 0.811     0.25       0.857     0.811     0.833      0.748    good
Weighted Avg.    0.789     0.229      0.796     0.789     0.792      0.748
```

Fig 4.2 : Decision accuracy by class for Labor (Using Simple Cart)

- **Confusion Matrix :**

```
=== Confusion Matrix ===

  a  b   <-- classified as
 15  5 |  a = bad
  7 30 |  b = good
```

Fig 4.3 : Confusion Matrix for Labor (Using Simple Cart)

- **Observation :**

- For class bad out of 20 instances, 15 of them are correctly classified and the rest 5 of them are misclassified as good.
- For class good, out of 37 instances, 7 has been misclassified as bad.

- **Using Decision Stump:**

- **Classifications for Labor :**

```
Classifications

pension = none : bad
pension != none : good
pension is missing : good
```

Fig 4.4 : Classifications for Labor (Using Decision Stump)

- **Classification Accuracy :**

- Total Number of Instances : 57
- Correctly Classified Instances : 46

- Incorrectly Classified Instances : 11
- Percentage of Correctly Classified Instances : 80.7018
- Percentage of Incorrectly Classified Instances : 19.2982

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.55      0.054      0.846     0.55      0.667      0.835    bad
                 0.946     0.45       0.795     0.946     0.864      0.835    good
Weighted Avg.    0.807     0.311      0.813     0.807     0.795      0.835
```

Fig 4.5 : Detailed accuracy by class for Labor (Using decision Stump)

- **Confusion Matrix :**

```
=== Confusion Matrix ===

  a  b   <-- classified as
 11  9 |  a = bad
  2 35 |  b = good
```

Fig 4.6 : Confusion matrix for labor (Using decision Stump)

- **Observation :**

- For class bad out of 20 instances, 11 of them are correctly classified as bad and the rest 9 of them are misclassified.
- For class good, out of 37 instances, only 2 has been misclassified as bad and the rest 35 has been correctly classified as good.

- **Analysis :**

- Here in labor data set, Classification accuracy using simple cart classifier is 78.9474% while it is 80.7018 % using decision stump classifier.
- The accuracies of both the classifiers vary by a margin of 2%.
- This similarity in the accuracy because both the decision tree (Fig 4.1) and classification (Fig 4.4) use only the single attribute of wage-increase-first-year and pension for classification.

- **Classification accuracy for different training and test sets :**

- **Using Simple Cart :**

Simple Cart	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20

Number of Instances	46	34	28	23	11
Correctly Classified Instances	38	30	24	22	8
Incorrectly Classified Instances	8	4	4	1	3
Percentage of Correctly Classified Instances	82.6087	88.2353	85.7143	95.6522	72.7273
Percentage of Incorrectly Classified Instances	17.3913	11.7647	14.2857	4.3478	27.2727

- **Using Decision Stump :**

Decision Stump	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	46	34	28	23	11
Correctly Classified Instances	34	28	23	20	9
Incorrectly Classified Instances	12	6	5	3	2
Percentage of Correctly Classified Instances	73.913	82.3529	82.1429	86.9565	81.8182
Percentage of Incorrectly Classified Instances	26.087	17.6471	17.8571	13.0435	18.1818

- **Analysis :**

- From the above tables it can noticed that for 60 vs 40 set the percentage of classification accuracy is higher followed by 40 vs 60.

- Using simple cart we have observed the classification accuracy of 95.6522 % and using decision stump it is 86.9565 % .
- In this we can infer that higher number of training set results in higher classification accuracy also very less number of test data does not yield good result.
- There is no change in decision tree or the number of leaf nodes and the size of the tree even when done with different percentage of training and test sets.

- **By Changing the parameters of the Classifiers :**

- **Using Simple Cart :**

- Debug

Debug value	True	False
Correctly Classified Instances	45	45

Classification Accuracy	78.9474	78.9474
Number of Leaf nodes	2	2
Size of the tree	3	3

- Heuristic values

heuristic value	True	False
Correctly Classified Instances	45	45
Classification Accuracy	78.9474	78.9474
Number of Leaf nodes	2	2
Size of the tree	3	3

- minNumObject

minNumObject value	0.0	1.0	2.0	3.0	4.0	5.0	10.0
Correctly Classified Instances	43	44	45	44	42	43	41
Classification Accuracy	75.4386	77.193	78.9474	77.193	73.6842	75.4386	71.9298
Number of Leaf nodes	2	2	2	2	2	2	2
Size of the tree	3	3	3	3	3	3	3

- numFoldsPruning

numFoldsPruning value	2	4	5	6	8	10
Correctly Classified Instances	44	44	45	44	43	43
Classification Accuracy	77.193	77.193	78.9474	77.193	75.4386	75.4386
Number of Leaf nodes	10	2	2	4	4	6
Size of the tree	19	3	3	7	7	11

- seed

seed value	0	1	2	4	5	10
Correctly Classified Instances	47	45	45	45	46	44
Classification Accuracy	82.4561	78.9474	78.9474	78.9474	80.7018	77.193
Number of Leaf nodes	8	2	2	8	4	7
Size of the tree	15	3	3	15	7	13

- sizePerValue

sizePer value	0.2	0.3	0.5	0.7	0.9	1.0
Correctly Classified Instances	42	43	43	44	44	45
Classification Accuracy	73.6842	75.4386	75.4386	77.193	77.193	78.9474
Number of Leaf nodes	2	2	2	4	2	2
Size of the tree	3	3	3	7	3	3

- useOneSE value

useOneSE value	True	False
Correctly Classified Instances	43	45
Classification Accuracy	75.4386	78.9474
Number of Leaf nodes	2	2
Size of the tree	3	3

- usePrune Value

usePrune value	True	False
Correctly Classified Instances	45	44
Classification Accuracy	78.9474	77.193
Number of Leaf nodes	2	10
Size of the tree	3	19

- **Analysis :**

- Debug and heuristic parameters does not have any effect on accuracy.
- minNumObject parameter forms uneven accuracy curve for different values and has the highest accuracy at value 2.0.
- sizePer has the highest accuracy at value 1.0 and its value cannot be 0.1 as classifier cannot have more fields than instances.
- The parameter numFoldsPruning forms a bell curve for different values and has the highest accuracy at value 5.
- useOneSE has the highest accuracy when its value is set to false and usePrune has the highest accuracy when its value is said to true.
- The parameter seed has the highest accuracy at value 0.

- **Using decision Stump :**

- debug

Debug value	True	False
Correctly Classified Instances	46	46
Classification Accuracy	80.7018	80.7018

- The change in value of parameter debug does not have any effect on the accuracy.

- **Noise :**

- **Adding Missing Values :**

- **Using Simple Cart :**

No.	duration Numeric	wage-increase-first-year Numeric	wage-increase-second-year Numeric
26	3.0	2.0	2.0
27	2.0	4.5	4.5
28		3.0	3.0
29	2.0	5.0	4.0
30	3.0	2.0	
31	3.0	4.5	4.5
32	3.0	3.0	2.0
33	2.0	2.5	2.5
34	2.0		5.0
35	3.0	2.0	2.5
36	2.0	2.0	2.0

- **Classification Accuracy :**

Correct instances	42
Incorrect instances	15
Correct %	73.6846
Incorrect %	26.3158
Number of leaf nodes	2
Size of tree	3

- **After Misclassifying:**
- **Changes in the values of the class attribute after misclassifying :**

Attribute Name	Label	Count
Class	Bad	20
	Good	37

- **Classification Accuracy :**

Correct instances	42
Incorrect instances	15
Correct %	73.6842
Incorrect %	26.3158
Number of leaf nodes	3

Size of tree	5

- **Using Decision Stump (After Adding Missing Values):**

- **Classification Accuracy :**

Correct instances	46
Incorrect instances	11
Correct%	80.7018
Incorrect%	19.2982

- **After Misclassifying :**

- **Classification Accuracy :**

Correct instances	43
Incorrect instances	14
Correct %	75.4386
Incorrect %	24.5614

- **Analysis :**

- For the labor dataset, for simple cart classifier the accuracy reduces from 78.9474 in the original to 73.6846when missing values are introduced, but reduces to 73.6842 when data is misclassified.
- For decision Stump classifier the accuracy of 80.7018 in the original remains same when missing values are introduced, but reduces to 75.4386 when data is misclassified.
- As already discussed for Iris dataset, missing values have very less or no impact on the accuracy of the dataset, while noise reduces the accuracy of the dataset.
- In this case also, there is same result as in Iris. As the number of missing values increase, the accuracy might further decrease.
- For this dataset, the original accuracy might have increased had there been less or missing values in the original dataset.

- **Class Distribution :**

- Class Distribution for this Labor dataset is 0.35 : 0.65.
- Here, the observations are unevenly distributed among the two classes.
- Yes, Class Distribution matters for our experiment. Accuracy is minimized when the distribution is uniform and uneven classes tends to have more accuracy.

- For this Vote data set, the total number of instances is 57 of which 90% will be training set and the rest 10% will be test set for each of the 10 iterations.
- Number of Training set : 51
- Number of Test set : 06

- Data is randomly divided into training and test set when the cross validation is done.
- Class Distribution will be not the same as that of full data set in Training and Test set.
- No, it does not matter. If class distribution across all folds is same, then cross-validated accuracy estimate might be biased and not really estimate generalization accuracy.

Dataset : Diabetes

- **Attributes :**

- Number of attributes : 9

Sl No.	Attribute Name	Minimum Value	Maximum Value	Mean	Standard Deviation
1.	Preq	0	17	3.845	3.37
2.	Plas	0	199	120.895	31.973
3.	Pres	0	122	69.105	19.356
4.	Skin	0	99	20.536	15.952
5.	Insu	0	846	79.799	115.244
6.	Mass	0	67.1	31.993	7.884
7.	Pedi	0.078	2.42	0.472	0.331
8.	Age	21	81	33.241	11.76
9.	Class				

- **Attribute : Class**

Sl No.	Attribute Name	Label	Count
1.	Class	Tested_negative	500
		Tested_positive	268

- **Decision Tree Algorithms :**

- **Using Simple Cart :**

- Test Mode : 10 fold cross validation

```
CART Decision Tree

plas < 127.5: tested_negative(391.0/94.0)
plas >= 127.5
|  mass < 29.95: tested_negative(52.0/24.0)
|  mass >= 29.95: tested_positive(150.0/57.0)
```

Fig 5.1 : Decision Tree for Diabetes (Using Simple Cart)

- Number of leaf nodes : 3
- Size of the tree : 5

- **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 577
- Incorrectly Classified Instances : 191
- Percentage of Correctly Classified Instances : 75.1302
- Percentage of Incorrectly Classified Instances : 24.8698

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.868     0.466      0.776     0.868     0.82       0.727    tested_negative
                 0.534     0.132      0.684     0.534     0.6        0.727    tested_positive
Weighted Avg.    0.751     0.35       0.744     0.751     0.743      0.727
```

Fig 5.2 : Detailed accuracy by class for Diabetes (Using Simple Cart)

- **Confusion Matrix :**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 434  66 |   a = tested_negative
 125 143 |   b = tested_positive
```

Fig 5.3 : Confusion Matrix for Diabetes (Using Simple Cart)

- **Observation :**

- For class tested_negative, out of 500 instances, 434 of them are correctly classified and 66 of them are misclassified as tested_positive.
- For class tested_positive, out of 268 instances, 125 has been misclassified as tested_negative and the rest 143 has been correctly classified as tested_positive.

- **Using Decision Stump :**

- **Test Mode : 10 fold cross validation**

- **Classifications:**

```
Classifications

plas <= 127.5 : tested_negative
plas > 127.5 : tested_positive
plas is missing : tested_negative
```

Fig 5.4 : Classifications for Diabetes (Using decision stump)

- **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 552
- Incorrectly Classified Instances : 216
- Percentage of Correctly Classified Instances : 71.875
- Percentage of Incorrectly Classified Instances : 28.125

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.796     0.425      0.777     0.796     0.787      0.684    tested_negative
                 0.575     0.204      0.602     0.575     0.588      0.684    tested_positive
Weighted Avg.    0.719     0.348      0.716     0.719     0.717      0.684
```

Fig 5.5 : Detailed accuracy by class for Diabetes (Using decision stump)

- **Confusion Matrix :**

```
=== Confusion Matrix ===

   a   b   <-- classified as
 398 102 |   a = tested_negative
 114 154 |   b = tested_positive
```

Fig 5.6 : Confusion matrix for Diabetes (Using decision Stump)

- **Analysis :**

- For class tested_negative, out of 500 instances, 398 of them are correctly classified and 102 of them are misclassified as tested_positive.
- For class tested_positive, out of 268 instances, 114 has been misclassified as tested_negative and the rest 154 has been correctly classified as tested_positive.
- In this case, Classification accuracy using simple cart classifier is 75.1302% and it is 71.875% using decision stump classifier.
- Again it can be noticed from the decision tree (Fig 2.1) and classification (Fig 2.4) that decision stump uses only a single attribute of plass for classification whereas simple cart uses two of plass and mass.
- Hence in this case, the performance of simple cart classifier is better when compared with decision stump classifier.

- **Classification accuracy for different training and test sets :**

- **Using Simple Cart :**

Simple Cart	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	614	461	384	307	154
Correctly Classified Instances	398	339	251	228	119
Incorrectly Classified Instances	216	122	133	79	35
Percentage of Correctly Classified Instances	64.8208	73.5358	65.3646	74.2671	77.2727
Percentage of Incorrectly Classified Instances	35.1792	26.4642	34.6354	25.7329	22.7273

- **Using Decision Stump :**

Decision Stump	20 vs 80	40 vs 60	50 vs 50	60 vs 40	80 vs 20
Number of Instances	614	461	384	307	154
Correctly Classified Instances	403	306	282	227	114
Incorrectly Classified Instances	211	155	102	80	40
Percentage of Correctly Classified Instances	65.6352	66.3774	73.4375	73.9414	74.026
Percentage of Incorrectly Classified Instances	34.3648	33.6226	26.5625	26.0586	25.974

- **Analysis :**

- From the above tables it can noticed that for 80 vs 20 set the percentage of classification accuracy is higher.
- Using simple cart we have observed the classification accuracy of 77.2727 % and using decision stump it is 74.026 %.

- In this we can infer that higher number of training set results in higher classification accuracy and vice versa

➢ **By Changing the parameters of the Classifiers :**

➢ **Using Simple Cart :**

- Debug

Debug value	True	False
Correctly Classified Instances	577	577
Classification Accuracy	75.1302	75.1302
Number of Leaf nodes	3	3
Size of the tree	5	5

- Heuristic value

heuristic value	True	False
Correctly Classified Instances	577	577
Classification Accuracy	75.1302	75.1302
Number of Leaf nodes	3	3
Size of the tree	5	5

- minNumObject

minNumObject value	0.0	1.0	2.0	3.0	4.0	5.0	10.0
Correctly Classified Instances	577	577	577	565	568	572	578
Classification Accuracy	75.1302	75.1302	75.1302	73.5677	73.7583	74.4792	75.2604
Number of Leaf nodes	3	3	3	21	13	13	3
Size of the tree	5	5	5	41	25	25	5

- numFoldsPruning value

numFoldsPruning value	2	4	5	6	8	10
Correctly Classified Instances	576	569	577	571	574	574
Classification Accuracy	75	74.0885	75.1302	74.349	74.7396	74.7396
Number of Leaf nodes	2	6	3	24	17	6
Size of the tree	3	11	5	47	33	11

- seed

seed value	0	1	2	4	5	10
Correctly Classified Instances	571	577	572	578	573	568
Classification Accuracy	74.349	75.1302	74.4792	75.2604	74.6094	73.7583
Number of Leaf nodes	3	3	13	20	3	6
Size of the tree	5	5	25	39	5	11

- sizePer Value

sizePer value	0.1	0.3	0.5	0.7	0.9	1
Correctly Classified Instances	557	551	577	577	574	568
Classification Accuracy	72.526	71.7448	75.1302	75.1302	74.7396	73.7583
Number of Leaf nodes	3	2	3	6	6	6
Size of the tree	5	3	5	11	11	11

- useOneSE value

useOneSE value	True	False
Correctly Classified Instances	582	568
Classification Accuracy	75.7813	73.7583
Number of Leaf nodes	3	6
Size of the tree	5	11

- usePrune value

usePrune value	True	False
Correctly Classified Instances	568	551
Classification Accuracy	73.7583	71.7448
Number of Leaf nodes	6	80
Size of the tree	11	159

➢ **Analysis :**

- Debug and heuristic parameters does not have any effect on accuracy.
- The parameters minNumObject forms an exponential curve and it has the highest accuracy at value 10.
- The parameter numFoldsPruning, seed value and sizePer forms an uneven curve for different values and has the highest accuracy at values 5, 4, 0.9 respectively.
- useOneSE and usePrune has the highest accuracy when its value is set to true.

➢ **Using Decision Stump :**

- debug :

Debug value	True	False
Correctly Classified Instances	552	552
Classification Accuracy	71.875	71.875

- debug parameter does not have any effect on accuracy.

➢ **Noise :**

➢ **Adding Missing Values :**

➢ **Using Simple Cart :**

Relation: pima_diabetes

No.	preg Numeric	plas Numeric	pres Numeric	skin Numeric	insu Numeric	mass Numeric	pedi Numeric	age Numeric	class Nominal
1	6.0	148.0	72.0	35.0	0.0	33.6	0.627	50.0	test...
2	1.0	85.0	66.0	29.0	0.0		0.351	31.0	test...
3	8.0	183.0	64.0	0.0	0.0	23.3	0.672	32.0	test...
4	1.0	89.0	66.0	23.0	94.0	28.1	0.167	21.0	test...
5	0.0		40.0	35.0	168.0		2.288	33.0	test...
6	5.0	116.0	74.0	0.0	0.0	25.6	0.201	30.0	test...
7	3.0	78.0	50.0	32.0	88.0	31.0	0.248	26.0	test...
8	10.0	115.0	0.0	0.0	0.0	35.3	0.134	29.0	test...
9	2.0		70.0	45.0	543.0	30.5	0.158	53.0	test...
10	8.0	125.0	96.0	0.0	0.0	0.0	0.232	54.0	test...

➢ Number of leaf nodes : 10
➢ Size of the tree : 19

➢ **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 572
- Incorrectly Classified Instances : 196
- Percentage of Correctly Classified Instances : 74.4792
- Percentage of Incorrectly Classified Instances : 25.5208

➢ **After Misclassifying Examples :**

Attribute Name	Minimum	Maximum	Mean	Standard Deviation
Plas	0	200	120.88	32.474
Mass	0	67.1	32.026	7.889

Number of leaf nodes :3

Size of the tree :5

➢ **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 566
- Incorrectly Classified Instances : 202
- Percentage of Correctly Classified Instances : 73.6979
- Percentage of Incorrectly Classified Instances : 26.3021

- **Using Decision Stump :**
- **After Adding Missing Values :**

- **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 551
- Incorrectly Classified Instances : 217
- Percentage of Correctly Classified Instances : 71.7448
- Percentage of Incorrectly Classified Instances : 28.2552

- **After Misclassifying Examples :**

- **Classification Accuracy :**

- Total Number of Instances : 768
- Correctly Classified Instances : 549
- Incorrectly Classified Instances : 219
- Percentage of Correctly Classified Instances : 71.4844
- Percentage of Incorrectly Classified Instances : 28.5156

- **Analysis :**

- For this dataset, for simple cart classifier the accuracy reduces from 75.1302 in the original to 74.4792 when missing values are introduced, but reduces to 74.7396 when data is misclassified.
- For decision Stump classifier the accuracy of 71.875 in the original reduces to 71.7448 when missing values are introduced, but reduces to 71.4844 when data is misclassified.
- Missing values have very less or no impact on the accuracy of the dataset, while noise reduces the accuracy of the dataset.
- The accuracy may decrease is the number of missing values increase.

- **Class Distribution :**

- Class Distribution for this Diabetes dataset is 0.65 : 0.35.
- Here, the observations are unevenly distributed among the two classes.

- Yes, Class Distribution matters for our experiment. Accuracy is minimized when the distribution is uniform and uneven classes tends to have more accuracy.

- For this Vote data set, the total number of instances is 768 of which 90% will be training set and the rest 10% will be test set for each of the 10 iterations.
- Number of Training set : 691
- Number of Test set : 77

- Data is randomly divided into training and test set when the cross validation is done.
- Class Distribution will be not the same as that of full data set in Training and Test set.
- No, it does not matter. If class distribution across all folds is same, then cross-validated accuracy estimate might be biased and not really estimate generalization accuracy.

➢ **Conclusion :**

After conducting experiments and having done the analysis on the four different data sets the following things can be inferred:

- After conducting the experiments on four different data sets, using two different classifiers i.e simple cart and decision stump it can be inferred that simple cart is better classifier than decision stump as it considers more than one attribute in decision tree while decision stump uses only one. In case of any data set with more than 2 classes, performance of decision stump is less.

- From the analysis of the experiment done it can be concluded that the numeric parameters gives higher accuracy for the default value and decreases as it goes away from the default value. For simple cart accuracy is high when useOneSE is set true and usePrune is set false. Debug and heuristic values do not affect the accuracy.

- The accuracy of the classifier mostly depends on the complexity and parameters of the algorithm. When there are very less number of training set the classifier tends to give very less accuracy. This also depends on class distribution.

- For a data set, when noise is added the accuracy reduces. The presence of missing values in the data set has very less or no impact on the accuracy. Sometimes, when there are too many missing values the accuracy might decrease. When there are misclassification in the data set the accuracy tends to go down by a bigger margin.

- Class Distribution also has its impact on accuracy. Accuracy is minimized when the distribution is uniform and uneven classes tends to have more accuracy. Forcing the training set to have same class distribution as the original data might not fetch correct results.