Project 1

Yu SU 55323707

1.

17.2.4

Subset size	Attributes in best subset	Classification accuracy
9	Si Al Ba Mg Na Ca Ri K Type	77.1028
8	Al Ba Mg Na Ca Ri K Type	77.5701
7	Ba Mg Na Ca Ri K Type	78.972
6	Mg Na Ca Ri K Type	78.0374
5	Na Ca Ri K Type	77.1028
4	Ca Ri K Type	73.8318
3	Ri K Type	64.9553
2	K Type	49.0654
1	Type	35.514
0		

17.2.5

No, it is not an unbiased estimate of accuracy on future data. For each attribute selection step, if we only use the data in each fold to make choice is OK, but we actually use(look) the whole test set to make choice, which will cause the overfitting (make the accuracy looks better than it should be)

2.

The right way is before the cross-validation, we should leave one of the folds out as the final test set. And then do the execution of 17.2.4 on the other folds. After we get the best subsets, we could use the left test fold to test the classification accuracy of these subsets. This accuracy is the final unbiased estimate of accuracy on future data.

3.

17.4.8

```
Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 17 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.2948    2 wage-increase-first-year
 0.1893    3 wage-increase-second-year
 0.1624   11 statutory-holidays
 0.1341   14 contribution-to-dental-plan
 0.1164   16 contribution-to-health-plan
 0.1091   12 vacation
 0.0855   13 longterm-disability-assistance
 0.0717    9 shift-differential
 0.0548    7 pension
 0.0484    5 cost-of-living-adjustment
 0.0333   15 bereavement-assistance
 0.0307    4 wage-increase-third-year
 0.024    10 education-allowance
 0.0195    8 standby-pay
 0         6 working-hours
 0         1 duration
```

Four most important attribute:
2 wage-increase-first-year
3 wag-increase-second-year
11 statutory-holidays
14 contribution-to-dental-plan

17.4.9

CfsSubsetEval/BestFirst:

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 114
        Merit of best subset found:    0.363

Attribute Subset Evaluator (supervised, Class (nominal): 17 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 2,3,5,11,12,13,14 : 7
                     wage-increase-first-year
                     wage-increase-second-year
                     cost-of-living-adjustment
                     statutory-holidays
                     vacation
                     longterm-disability-assistance
                     contribution-to-dental-plan
```

J48/BestFirst:

```
=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 138
        Merit of best subset found:    0.842

Attribute Subset Evaluator (supervised, Class (nominal): 17 class):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 1,2,4,6,11,12 : 6
                     duration
                     wage-increase-first-year
                     wage-increase-third-year
                     working-hours
                     statutory-holidays
                     vacation
```

The attributes are selected by both methods:

2 wage-increase-first-year

11 statutory-holidays

12 vacation

Relate to ranker:

The rank of these three attributes is 1,3,6. That means in different algorithm the best attributes are not same. CfsSubsetEval focus on relation and the infogain focus on individual information gain.

17.4.10

Number of copies	accuracy	Selected attribute
0	75.5208%	2,6,7,8
1	75.5208%	2,6,7,8
2	75.5208%	2,6,7,8
3	75.5208%	2,6,7,8
4	75.5208%	2,6,7,8

No matter how many redundant copies in the data, the accuracy is always same, so we could say the NaiveBayes could eliminate the redundant attributes.

17.4.11

Accuracy:

Number of copies	BestFirst/wrapper	BestFirst/CfsSubset	Ranker/infogain
0	73.9583%	75.5208%	74.349%
1	73.9583%	75.5208%	74.0885%
2	73.9583%	75.5208%	73.4375%
3	73.9583%	75.5208%	72.0052%
44	73.9583%	75.5208%	72.0052%

Selected attributes

Number of copies	Bestfirst/wrapper	BestFirst/CfsSubset	Ranker/infogain
0	1,2,3,6,7 : 5	2,6,7,8:4	2,6,8,5,4,1,7,3:8
1	1,2,3,6,7 : 5	2,6,7,8:4	2,6,8,5,4,10,1,7:8
2	1,2,3,6,7 : 5	2,6,7,8:4	2,6,8,5,4,11,10,1 :8
3	1,2,3,6,7 : 5	2,6,7,8:4	2,6,8,5,4,12,11,10:8
4	1,2,3,6,7 : 5	2,6,7,8:4	2,6,8,5,4,12,13,1 :8

First two methods could eliminate the redundant attribute because their result does not change when there are new copy in the data.
While the method of ranker could not eliminate the redundancy. The copy will influence the result, because the ranker method focuses on individual single information gain of each data. There in no relation between each attribute, so it can not eliminate the redundancy.

17.4.12

Optimized:

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

Cross-validated Parameter selection.
Classifier: weka.classifiers.lazy.IBk
Cross-validation Parameter: '-K' ranged from 1.0 to 10.0 with 10.0 steps
Classifier Options: -K 7 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

IB1 instance-based classifier
using 7 nearest neighbour(s) for classification


Time taken to build model: 0.52 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         568               73.9583 %
Incorrectly Classified Instances       200               26.0417 %
Kappa statistic                          0.4009
Mean absolute error                      0.3149
Root mean squared error                  0.4372
Relative absolute error                 69.2814 %
Root relative squared error             91.7214 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.850    0.466    0.773      0.850    0.810      0.406    0.763     0.832     tested_negative
               0.534    0.150    0.656      0.534    0.588      0.406    0.763     0.623     tested_positive
Weighted Avg.  0.740    0.356    0.732      0.740    0.732      0.406    0.763     0.759

=== Confusion Matrix ===

   a   b   <-- classified as
 425  75 |   a = tested_negative
 125 143 |   b = tested_positive
```

Original:

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

Cross-validated Parameter selection.
Classifier: weka.classifiers.lazy.IBk
Classifier Options: -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

IB1 instance-based classifier
using 1 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         539               70.1823 %
Incorrectly Classified Instances       229               29.8177 %
Kappa statistic                          0.3304
Mean absolute error                      0.2988
Root mean squared error                  0.5453
Relative absolute error                 65.7327 %
Root relative squared error            114.3977 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.794    0.470    0.759      0.794    0.776      0.331    0.650     0.732     tested_negative
               0.530    0.206    0.580      0.530    0.554      0.331    0.650     0.469     tested_positive
Weighted Avg.  0.702    0.378    0.696      0.702    0.698      0.331    0.650     0.640

=== Confusion Matrix ===

   a   b   <-- classified as
 397 103 |   a = tested_negative
 126 142 |   b = tested_positive
```

The accuracy is 70.1823% without CV parameter selection.

After CV parameter selection, the accuracy is 73.9583%.

The selected value is K = 7

17.4.13

Optimized:

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

Cross-validated Parameter selection.
Classifier: weka.classifiers.trees.J48
Cross-validation Parameter: '-C' ranged from 0.1 to 0.5 with 5.0 steps
Cross-validation Parameter: '-M' ranged from 1.0 to 10.0 with 10.0 steps
Classifier Options: -C 0.2 -M 10



Time taken to build model: 2.3 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         571               74.349  %
Incorrectly Classified Instances       197               25.651  %
Kappa statistic                          0.433
Mean absolute error                      0.3133
Root mean squared error                  0.4302
Relative absolute error                 68.9416 %
Root relative squared error             90.2662 %
Total Number of Instances              768


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

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.56: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.56
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   insu <= 120: tested_negative (11.0/4.0)
|   |   |   |   |   |   insu > 120: tested_positive (10.0/2.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   insu <= 14
|   |   |   |   preg <= 5: tested_negative (11.0/2.0)
|   |   |   |   preg > 5: tested_positive (10.0/4.0)
|   |   |   insu > 14: tested_positive (14.0/4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 2.3 seconds
```

Original:

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

Cross-validated Parameter selection.
Classifier: weka.classifiers.trees.J48
Classifier Options: -C 0.25 -M 2
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768
```

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

plas <= 127
|   mass <= 26.4: tested_negative (132.0/3.0)
|   mass > 26.4
|   |   age <= 28: tested_negative (180.0/22.0)
|   |   age > 28
|   |   |   plas <= 99: tested_negative (55.0/10.0)
|   |   |   plas > 99
|   |   |   |   pedi <= 0.56: tested_negative (84.0/34.0)
|   |   |   |   pedi > 0.56
|   |   |   |   |   preg <= 6
|   |   |   |   |   |   age <= 30: tested_positive (4.0)
|   |   |   |   |   |   age > 30
|   |   |   |   |   |   |   age <= 34: tested_negative (7.0/1.0)
|   |   |   |   |   |   |   age > 34
|   |   |   |   |   |   |   |   mass <= 33.1: tested_positive (6.0)
|   |   |   |   |   |   |   |   mass > 33.1: tested_negative (4.0/1.0)
|   |   |   |   |   preg > 6: tested_positive (13.0)
plas > 127
|   mass <= 29.9
|   |   plas <= 145: tested_negative (41.0/6.0)
|   |   plas > 145
|   |   |   age <= 25: tested_negative (4.0)
|   |   |   age > 25
|   |   |   |   age <= 61
|   |   |   |   |   mass <= 27.1: tested_positive (12.0/1.0)
|   |   |   |   |   mass > 27.1
|   |   |   |   |   |   pres <= 82
|   |   |   |   |   |   |   pedi <= 0.396: tested_positive (8.0/1.0)
|   |   |   |   |   |   |   pedi > 0.396: tested_negative (3.0)
|   |   |   |   |   |   pres > 82: tested_negative (4.0)
|   |   |   |   age > 61: tested_negative (4.0)
|   mass > 29.9
|   |   plas <= 157
|   |   |   pres <= 61: tested_positive (15.0/1.0)
|   |   |   pres > 61
|   |   |   |   age <= 30: tested_negative (40.0/13.0)
|   |   |   |   age > 30: tested_positive (60.0/17.0)
|   |   plas > 157: tested_positive (92.0/12.0)

Number of Leaves  :     20

Size of the tree :      39
```

without CV parameter selection:
The accuracy is 73.8281%, the tree size is 39

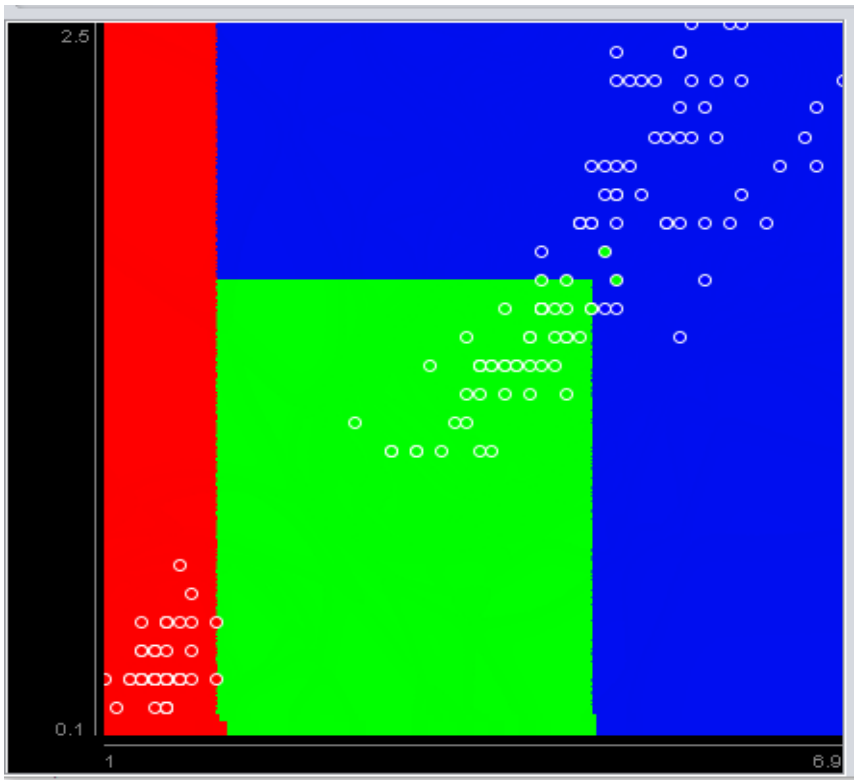
With CV parameter selection:
The accuracy is 74.349%, the tree size is 29.

The selected value is C = 0.2, M = 10.

4.

17.3.8

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

JRIP rules:
===========

(petallength <= 1.9) => class=Iris-setosa (50.0/0.0)
(petalwidth <= 1.6) and (petallength <= 4.9) => class=Iris-versicolor (47.0/0.0)
 => class=Iris-virginica (53.0/3.0)

Number of Rules : 3


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         147               98      %
Incorrectly Classified Instances         3                2      %
Kappa statistic                          0.97
Mean absolute error                      0.0252
Root mean squared error                  0.1122
Relative absolute error                  5.6604 %
Root relative squared error             23.7915 %
Total Number of Instances              150
```



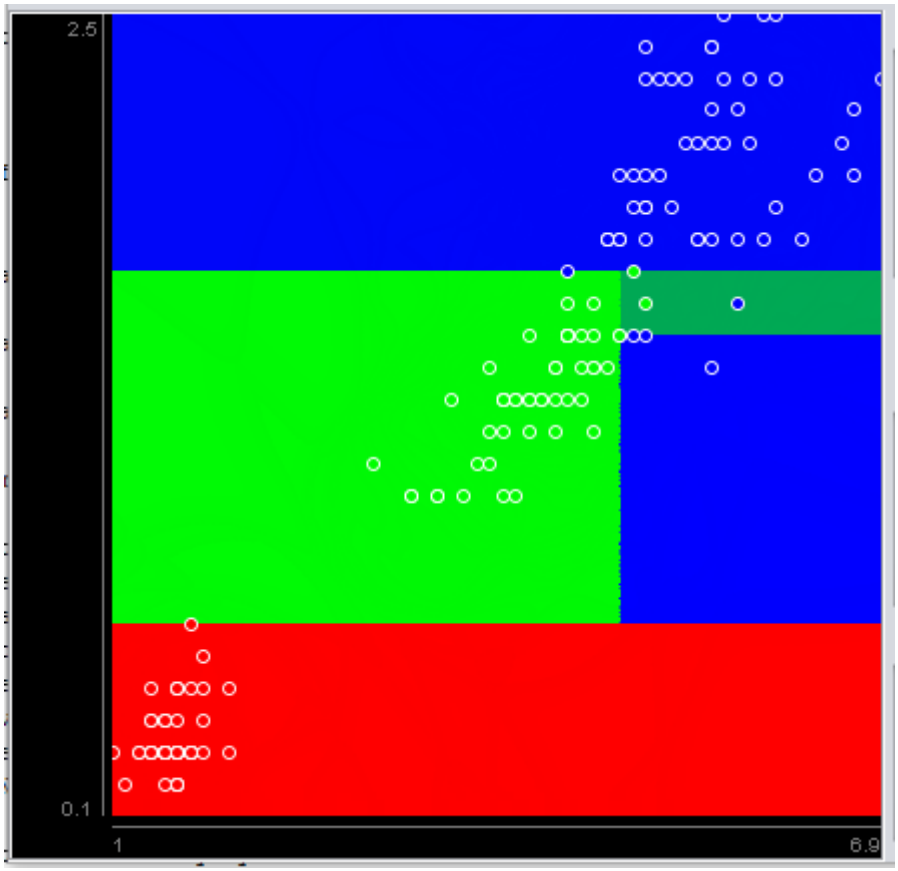
The JRip classify the spot in three types with generate three rules.

17.3.9

petallength <=1.9 =>class=Iris-setosa;
petallength <=4.9 and petallength >1.9 and petalwidth <=1.6 =>class=Iris-versicolor;
petallength >4.9 or petallength <=4.9 and petallength >1.9 and pentallwidth >1.6 => class=Iris-virginica;

17.3.10

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

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
|   petalwidth <= 1.7
|   |   petallength <= 4.9: Iris-versicolor (48.0/1.0)
|   |   petallength > 4.9
|   |   |   petalwidth <= 1.5: Iris-virginica (3.0)
|   |   |   petalwidth > 1.5: Iris-versicolor (3.0/1.0)
|   petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         147               98      %
Incorrectly Classified Instances         3                2      %
Kappa statistic                          0.97
Mean absolute error                      0.0233
Root mean squared error                  0.108
Relative absolute error                  5.2482 %
Root relative squared error             22.9089 %
Total Number of Instances              150
```



J48 generate a decision tree with 5 leaves to classify the spot.

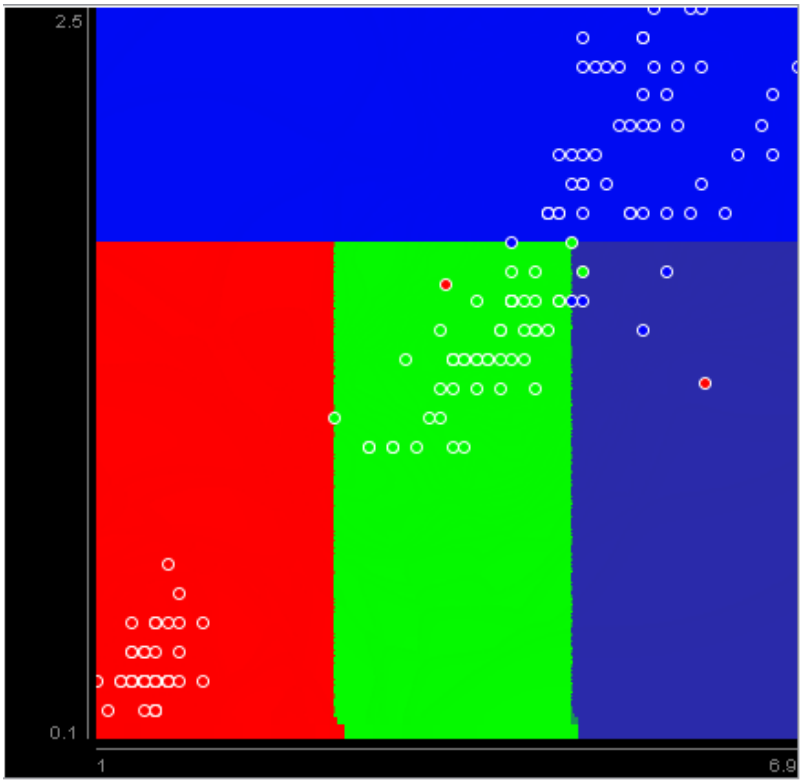
7.3.11

3: >6 and <50
2: >=50 and <=75
1: >75

17.3.12

When we add some noise, the J48 bound hardly changed, but JRip changed much. So we could say:
J48 has better efficiency to count noise than JRip.
J48 could solve the overfitting better

JRip:



J48: