CSC4008 Assignment9
118010045

1. Compare the results at different parameter settings (different k)

1) 5-fold cross-validation

Python: **71.8%**

```
Please input k for the k-fold cross-validation: 5
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    5-fold cross_validation

=== Summary===

Correctly Classified Instances           718      71.800 %
Incorrectly Classified Instances         282      28.200 %
Total Number of Instances               1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.756       0.881       0.337       0.814       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 617   83  |   a = good
 199  101  |   b = bad
```

Weka: **75.1%**

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

Correctly Classified Instances         751               75.1    %
Incorrectly Classified Instances       249               24.9    %
Kappa statistic                          0.3756
Mean absolute error                      0.2924
Root mean squared error                  0.4184
Relative absolute error                 69.5786 %
Root relative squared error             91.2977 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.860    0.503    0.799      0.860    0.829      0.379    0.790     0.893     good
                 0.497    0.140    0.603      0.497    0.545      0.379    0.790     0.579     bad
Weighted Avg.    0.751    0.394    0.741      0.751    0.743      0.379    0.790     0.799

=== Confusion Matrix ===

   a   b   <-- classified as
 602  98 |   a = good
 151 149 |   b = bad
```

2) 10-fold cross-validation

Python: **71.3%**

```
Please input k for the k-fold cross-validation: 10
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    10-fold cross_validation

=== Summary===

Correctly Classified Instances           713      71.300 %
Incorrectly Classified Instances         287      28.700 %
Total Number of Instances               1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.754       0.876       0.333       0.810       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 613   87  |   a = good
 200  100  |   b = bad
```

Weka: **75.4%**

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

Correctly Classified Instances         754               75.4    %
Incorrectly Classified Instances       246               24.6    %
Kappa statistic                          0.3813
Mean absolute error                      0.2936
Root mean squared error                  0.4201
Relative absolute error                 69.8801 %
Root relative squared error             91.6718 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.864    0.503    0.800      0.864    0.831      0.385    0.787     0.891     good
                 0.497    0.136    0.611      0.497    0.548      0.385    0.787     0.577     bad
Weighted Avg.    0.754    0.393    0.743      0.754    0.746      0.385    0.787     0.797

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 151 149 |   b = bad
```

3) 15-fold cross-validation

Python: **70.7%**

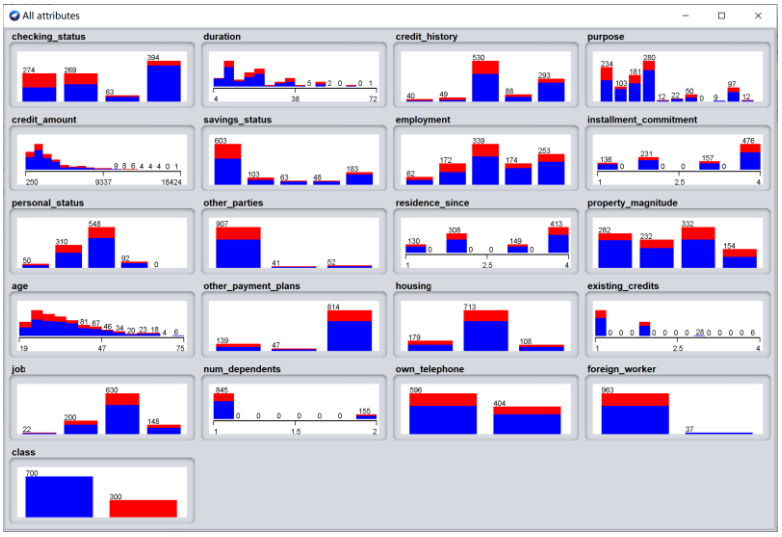
```
Please input k for the k-fold cross-validation: 15
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    15-fold cross_validation

=== Summary===

Correctly Classified Instances           707      70.700 %
Incorrectly Classified Instances         293      29.300 %
Total Number of Instances               1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.754       0.877       0.338       0.811       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 606   85  |   a = good
 198  101  |   b = bad
```

Weka: **74.9%**

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

Correctly Classified Instances         749               74.9    %
Incorrectly Classified Instances       251               25.1    %
Kappa statistic                          0.3668
Mean absolute error                      0.2946
Root mean squared error                  0.4211
Relative absolute error                 70.1027 %
Root relative squared error             91.8973 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.863    0.517    0.796      0.863    0.828      0.371    0.785     0.891     good
                 0.483    0.137    0.602      0.483    0.536      0.371    0.785     0.574     bad
Weighted Avg.    0.749    0.403    0.738      0.749    0.740      0.371    0.785     0.796

=== Confusion Matrix ===

   a   b   <-- classified as
 604  96 |   a = good
 155 145 |   b = bad
```

2. Compare the results at different parameter settings (different selections of attributes)

According to the observation of all the attributes, the 'employment', 'credit_amount', 'age', 'existing_credits', 'personal_status', 'property_magnitude', 'other_payment_plans', 'own_telephone' and 'housing' do not have decisive and powerful influence to the classifier.



1) To maximum the accuracy of Weka, 5 attributes including 'personal_status', 'property_magnitude', 'other_payment_plans', 'own_telephone' and 'housing' are deleted.
The accuracy of Weka can be elevated from 75.4% to 76.2%
Python: **71.4%**

```
Please input k for the k-fold cross-validation: 10
Relation:     german_credit
Instances:    1000
Attributes:   16
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              other_parties
              residence_since
              age
              existing_credits
              job
              num_dependents
              foreign_worker
              class
Test mode:    10-fold cross_validation

=== Summary===

Correctly Classified Instances:         714      71.400 %
Incorrectly Classified Instances:       286      28.600 %
Total Number of Instances:             1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.749       0.890       0.303       0.813       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 623   77  |   a = good
 209   91  |   b = bad
```

Weka: **76.2%**

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

Correctly Classified Instances         762               76.2    %
Incorrectly Classified Instances       238               23.8    %
Kappa statistic                          0.4014
Mean absolute error                      0.3004
Root mean squared error                  0.4178
Relative absolute error                 71.4933 %
Root relative squared error             91.1748 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.870     0.490     0.806      0.870    0.837      0.405    0.784     0.886     good
               0.510     0.130     0.627      0.510    0.563      0.405    0.784     0.574     bad
Weighted Avg.  0.762     0.382     0.752      0.762    0.754      0.405    0.784     0.793

=== Confusion Matrix ===

   a   b   <-- classified as
 609  91 |   a = good
 147 153 |   b = bad
```

2) To maximum the accuracy of Python program, 6 attributes including 'employment', 'credit_amount', 'age', 'existing_credits', 'property_magnitude' and 'other_payment_plans' are deleted.

The accuracy of Python program can be elevated from 71.3% to 73.2%

Python: **73.2%**

```
Please input k for the k-fold cross-validation: 10
Relation:     german_credit
Instances:    1000
Attributes:   15
              checking_status
              duration
              credit_history
              purpose
              savings_status
              installment_commitment
              personal_status
              other_parties
              residence_since
              housing
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    10-fold cross_validation

=== Summary===

Correctly Classified Instances:          732     73.200 %
Incorrectly Classified Instances:        268     26.800 %
Total Number of Instances:              1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.776       0.869       0.413       0.819       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 608   92  |   a = good
 176  124  |   b = bad
```

Weka: **75.2%**

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

Correctly Classified Instances         752               75.2    %
Incorrectly Classified Instances       248               24.8    %
Kappa statistic                          0.3788
Mean absolute error                      0.311
Root mean squared error                  0.4116
Relative absolute error                 74.0113 %
Root relative squared error             89.8265 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.860    0.500    0.801      0.860    0.829      0.382    0.781     0.884     good
                 0.500    0.140    0.605      0.500    0.547      0.382    0.781     0.597     bad
Weighted Avg.    0.752    0.392    0.742      0.752    0.745      0.382    0.781     0.798

=== Confusion Matrix ===

   a   b   <-- classified as
 602  98 |   a = good
 150 150 |   b = bad
```

3. Conclusion (The second question of Assignment9)

During all the attempts of different k values, the 10-fold cross-validation is the best for both Weka and the Python program. Delete the attributes 'property_magnitude' and 'other_payment_plans', both accuracy of Weka and the Python program can be elevated.

4. Implement the “Bagging” strategy (The third question of Assignment9)

The “Bagging” strategy lifts the accuracy from 71.3% to **72.0%** (71.9%-72.1%).

```
Please input k for the k-fold cross-validation: 10
Please input b for the iteration times: 10
Relation:     german_credit
Instances:    1000
Attributes:   21
              checking_status
              duration
              credit_history
              purpose
              credit_amount
              savings_status
              employment
              installment_commitment
              personal_status
              other_parties
              residence_since
              property_magnitude
              age
              other_payment_plans
              housing
              existing_credits
              job
              num_dependents
              own_telephone
              foreign_worker
              class
Test mode:    10-fold cross_validation

=== Summary===

Correctly Classified Instances:          721      72.100 %
Incorrectly Classified Instances:        279      27.900 %
Total Number of Instances:              1000

=== Detail Accuracy By Class ===

Precision   Recall      Specificity F-Measure   Class
0.762       0.874       0.363       0.814       good

=== Confusion Matrix ===

   a    b    <-- classifed as
 612   88  |   a = good
 191  109  |   b = bad
```