Charles Sturt University

Data mining and visualization for business intelligence

Assessment

Practical Assessment

Submitted by:

Johan Sebastian Ramirez Vallejo

Contents

Figures

Tables

1. Data Exploration

The dataset contains numeric type data, and it was retrieved by WEKA software, it was required to convert the data to nominal type data to help the algorithm to make sense of the data, in this case, the values represent a scale of 0 – 4, which represents categories where the records are distributed in this scale across to the attributes to determine the most common response across the survey. Therefore, Nominal data is used to label variables without providing any quantitative values. So, the records are segmented in each category. This helps to create a profile and formulate hypotheses according to the problems that are being analysed. For example, the hypothesis who tend to divorce to take prevention actions.

It is generally observed about the attribute values are distributed by the result class where “blue” means no result in divorce, and “Red” means results in divorce, demonstrating that couples who strongly agree (0 on the scale) with the attribute tend to **no divorce**, in contrast, who strongly disagree with the highest category (4 in the scales) the attribute tend to **divorce**.

1.1 Attributes Distribution

Each attribute has a numeric domain of 0-4. An attribute value of 0 means the participant strongly agreed with the attribute while 4 means the participant strongly disagreed with the attribute. The class 0 = Blue means- no divorce, and 1 = Red – Divorce. Selection Attributes, a feature in WEKA support to determine the correlation this attribute evaluator was correlation attribute evaluation with standards parameters. The results showed a rank list where the top is the most correlated attributes and the bottom the fewer correlated attributes. It is important to highlight that the attribute evaluator can evaluate nominal and numeric data.

The dataset was modified and created sub-datasets in terms of rank to evaluate the 6 most, least and author-correlated attributes as shown below.

1.1.1 Explanation of the 6 most correlated attributes

- 9. I enjoy travelling with my wife.
- 18. My spouse and I have similar ideas about how marriage should be
- 19. My spouse and I have similar ideas about how roles should be in marriage
- 35. I can insult my spouse during our discussions.
- 36. I can be humiliated when we have discussions.
- 40. We're just starting a discussion before I know what is going on.

1.1.2 Explanation of the 6 Least correlated attributes

- 6. We do not have time at home as partners.
- 43. I mostly stay silent to calm the environment a little bit.
- 45. I would rather stay silent than discussing with my spouse.
- 46. Even if I am right in the discussion, I stay silent to hurt my spouse.
- 48. I feel right in our discussions.
- 52. I would not hesitate to tell my spouse about her/his inadequacy.

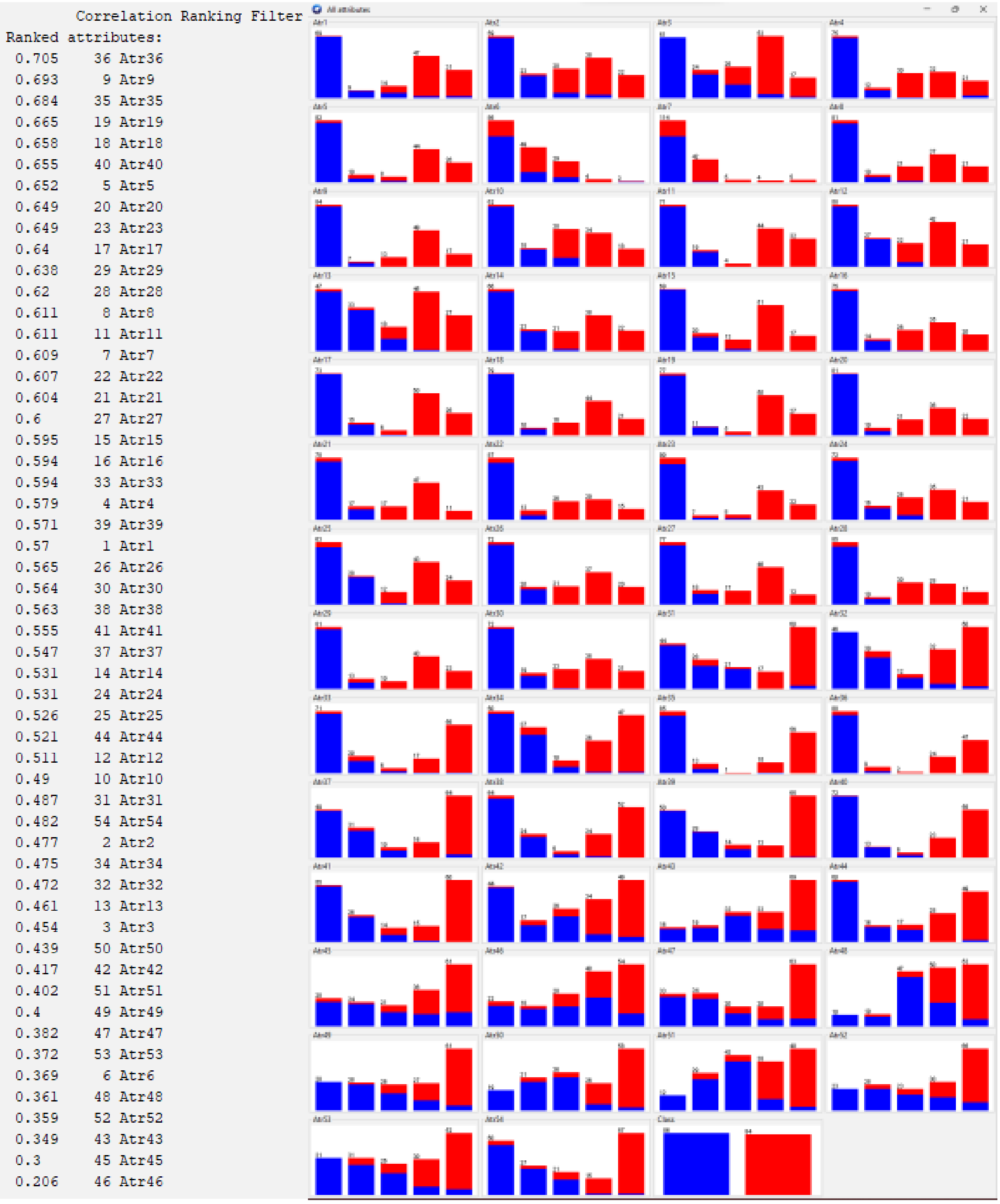
1.1.3 Explanation of the 6 most correlated attributes Author

- 2. I know we can ignore our differences, even if things get hard sometimes.
- 6. We do not have time at home as partners.

- 11. I think that one day in the future when I look back, I see that my spouse and I have been in harmony with each other.
- 18. My spouse and I have similar ideas about how marriage should be
- 26. I know my spouse's basic anxieties.
- 40. We're just starting a discussion before I know what is going on.

1.2 Correlation Rank

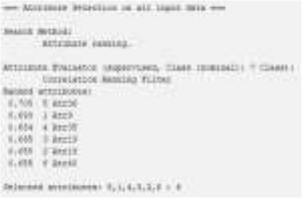
The correlation rank was marked by the select attribute feature in WEKA, where the attributes were listed in a rank and, they were visualized to explore the distribution of each attribute.



```
        Correlation Ranking Filter
Ranked attributes:
 0.705    36 Atr36
 0.693     9 Atr9
 0.684    35 Atr35
 0.665    19 Atr19
 0.658    18 Atr18
 0.655    40 Atr40
 0.652     5 Atr5
 0.649    20 Atr20
 0.649    23 Atr23
 0.64     17 Atr17
 0.638    29 Atr29
 0.62     28 Atr28
 0.611     8 Atr8
 0.611    11 Atr11
 0.609     7 Atr7
 0.607    22 Atr22
 0.604    21 Atr21
 0.6      27 Atr27
 0.595    15 Atr15
 0.594    16 Atr16
 0.594    33 Atr33
 0.579     4 Atr4
 0.571    39 Atr39
 0.57      1 Atr1
 0.565    26 Atr26
 0.564    30 Atr30
 0.563    38 Atr38
 0.555    41 Atr41
 0.547    37 Atr37
 0.531    14 Atr14
 0.531    24 Atr24
 0.526    25 Atr25
 0.521    44 Atr44
 0.511    12 Atr12
 0.49     10 Atr10
 0.487    31 Atr31
 0.482    54 Atr54
 0.477     2 Atr2
 0.475    34 Atr34
 0.472    32 Atr32
 0.461    13 Atr13
 0.454     3 Atr3
 0.439    50 Atr50
 0.417    42 Atr42
 0.402    51 Atr51
 0.4      49 Atr49
 0.382    47 Atr47
 0.372    53 Atr53
 0.369     6 Atr6
 0.361    48 Atr48
 0.359    52 Atr52
 0.349    43 Atr43
 0.3      45 Atr45
 0.206    46 Atr46
```

Figure 1 Correlation Rank and Visualization

1.2.1 Most correlated Attributes



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

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 7 Class):
        Correlation Ranking Filter
Ranked attributes:
 0.706  5 Atr36
 0.[illegible]  1 Atr9
 0.654  4 Atr20
 0.[illegible]  3 Atr19
 0.[illegible]  2 Atr18
 0.[illegible]  6 Atr40

Selected attributes: 5,1,4,3,2,6 : 6
```

Figure 2 Most correlated Attributes

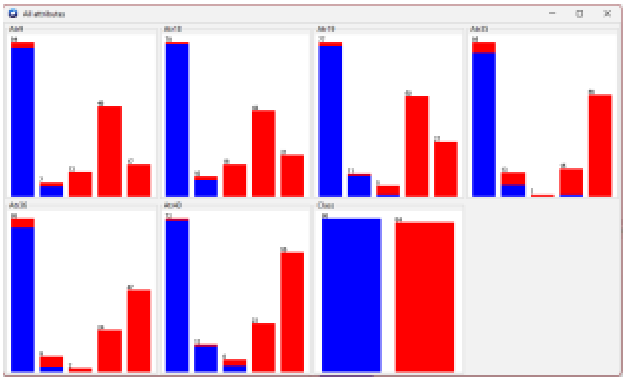
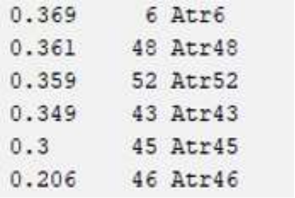


Figure 3. Visualization of most correlation attributes

1.2.2 Least Correlated Attributes



```
0.369      6 Atr6
0.361     48 Atr48
0.359     52 Atr52
0.349     43 Atr43
0.3       45 Atr45
0.206     46 Atr46
```

Figure 4. Least Correlation Attributes

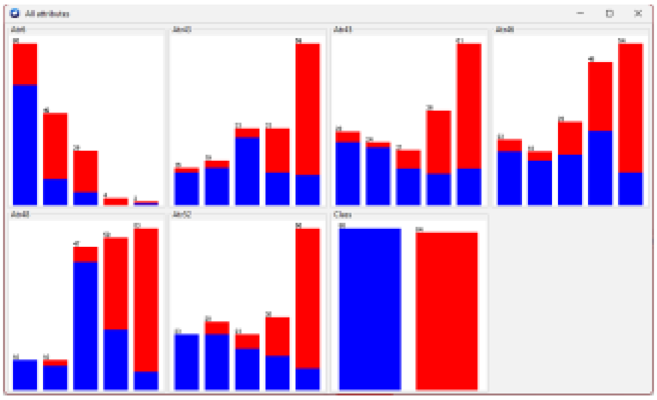
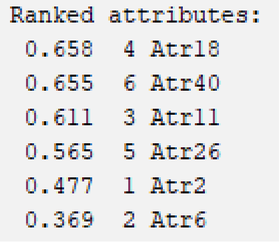


Figure 5. Visualization least correlation attributes

1.2.3 Comparison 6 most correlation attributes and 6 least correlation attributes

Comparing the most correlated attributes against the least correlated attributes, show that people who strongly agree do not get a divorce (The positive skew), and people who are more disagree with the attribute tend to divorce (The negative skew). Then, the most correlated attributes showed that the subset of attributes is highly correlated to the class, so the segmentation of the class is determined by the extreme values in the 0-4 scale. On the other hand, the lower correlation does not show the correlation to both classes which means that people who agree with the attribute do not get a divorce but there is another proportion of people who disagree re also no get a divorce then cannot be conclusive to any hypothesis or predict the results if we see this uncorrelated attribute. For example, attribute 6 in the 6 least correlated attributes shows a positive skew for both classes that indicates that people who answer the question tend to agree and cannot be classified as who may get a divorce or not. Then, there is no group segmentation to determine the correlation to the class.

1.2.4 Author Correlation



```
Ranked attributes:
 0.658   4 Atr18
 0.655   6 Atr40
 0.611   3 Atr11
 0.565   5 Atr26
 0.477   1 Atr2
 0.369   2 Atr6
```

Figure 6. Author Rank attributes

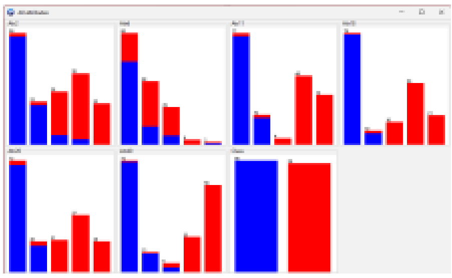
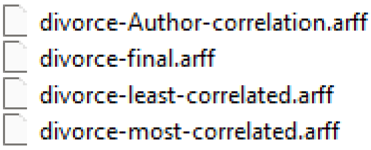


Figure 7. Visualization author attributes

2. Model building

The dataset was explored, and it was created new sub-datasets which represent the most correlated attributes, and least correlated attributes given by the selection attribute WEKA feature, where was used the rank method to determine a list that indicates the less to more rank where 0 low correlate and 1 high correlate. The pre-process module in WEKA was selected and filtered the corresponding dataset and saved it. then it was created models with different techniques such as Artificial Neural Network ANN(MLP), Simple Tree J48, and Logistic Regression. Therefore, it was analysed the correlated attributes chosen by the author.



divorce-Author-correlation.arff
divorce-final.arff
divorce-least-correlated.arff
divorce-most-correlated.arff

Figure 8. Data sets used

2.1 Analysis datasets

It was determined the nominal dataset and sub-datasets, and they were used to create two models using two test options 10-fold cross-validation, and 70:30 train:test. The model measure for comparison was accuracy and kappa which indicate the results of each model.

Classification	Number of features (Dataset)	Accuracy (10fold cross valid.)	Kappa (10fold cross valid.)	Build Time (secs)	Accuracy (70:30 train:test split)	Kappa (70:30 train:test split)	Build Time (secs)
Technique							
ANN(MLP)	All 54 attributes (divorce _nominal. arff)	97.6471 %	0.9529	42.32 sec	96.0784 %	0.9217	40.44 sec
J48		95.2941 %	0.9059	0.01 sec	96.0784 %	0.9217	0 sec
Logistic Regression		95.2941 %	0.9059	0.17 sec	96.0784 %	0.9217	0.04 sec
ANN(MLP)	6 most correlated attributes / (divorce _nominal _most6.arff)	97.6471 %	0.9529	1.43 sec	94.1176 %	0.8828	1.28 sec
J48		95.8824 %	0.9176	0 sec	96.0784 %	0.9217	0 sec
Logistic Regression		97.6471 %	0.9529	0.01 sec	98.0392 %	0.9607	0.01 sec
ANN(MLP)	6 least correlated attributes / (divorce_ Nominal _least6.arff)	90 %	0.8001	1.57 sec	80.3922 %	0.6065	1.48 sec
J48		89.4118 %	0.7881	0 sec	74.5098 %	0.4848	0 sec
Logistic Regression		88.8235 %	0.7765	0.01 sec	84.3137 %	0.6852	0.01 sec
ANN(MLP)	Attributes proposed by the authors (divorce_ Nominal _author6.arff)	98.2353 %	0.9647	1.48 sec	98.0392 %	0.9607	1.54 sec
J48		95.2941 %	0.9058	0 sec	96.0784 %	0.9217	0 sec
Logistic Regression		98.2353 %	0.9647	0.01 sec	96.0784 %	0.9217	0 sec

Table 1 Analysis table divorce data set

2.2 Discussion

Algorithms used

Artificial neuron Network ANN(MLP): This is a technique that was created to simulate how a human brain works in a simple way. These behave like neurons are interconnected to transmit the inputs generated by activation functions from sensory organs to the core. According to the distribution of the dataset, the values are automatically updated until achieving the target outputs values based on the learning rules. The network created after the training process can classify the data given for the test (Yöntem et al., 2019). The model accuracy is based on the number of neurons that best fit the data.

Simple Tree J48: It is a supervised learning approach and predictive decision support tool that create representation from observations to possible effects, characterized by nodes useful for decision-making, and at the end of every node is the outcome of making those decisions (Ngai et al., 2011). Some advantages stand in that requires little data preparation which means can handle missing values. It can handle categorical and numerical data and can represent multi-output problems. the decision tree by its selves do not perform predictively compare to other algorithms, but, using it as a base of a complex algorithm such as random forest is useful (Hanafy & Ming, 2021).

Logistic regression: is a classical technique that removes the variables which have no relations with the claimed model. Therefore, the technique is helpful to predict the expected outcome of a binary dependent variable, and it is defined by a given set of predictor variables in this case a linear combination of risk factors that is relevant to the probability of observing an event. (Pesantez-Narvaez et al., 2019). The main characteristics are based that the maximum likelihood estimates are easily found, and the predictions are given 0 o 1 which is easily interpreted as the event probability event. Some disadvantages – it may lead to overfitting - The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

a. Compare your results for Accuracy (10fold cross valid.) and Accuracy (70:30 train:test split). Discuss the difference and justify the suitability of 10-fold cross-validation.

Cross-validation is a standard evaluation technique, which consists in running in a systematic way a percentage split, for example, the dataset is split into 10 pieces or folds and uses each split as a testing dataset against the remaining 9 pieces. This process gives 10 evaluation results and computes the average; at the 11th time, it is invoked the data set to obtain the final model. Therefore, this process takes more time to build a model, the results indicate that ANN (MLP) algorithm takes more time than the 70:30 train: Test validation. This indicates that building the model takes 42.32 sec, but this takes more time to make the cross-validation against the model, cross-validation does this repetitive time until it reaches the k factor to get the final test. Then, the total time obtained is approximately 7 min to build the model and evaluate the model. Hence, 70:30 train:test split validation is faster in terms that this validation method split the dataset in 70% for training and 30% for testing and builds the final model in 40.44 sec. The final result indicates that the model with cross-validation obtained an accuracy of 97.6471 %, slightly superior to 70:30 train:test which got 96.0784 %. This shows that cross-validation has better performance, but this does not work well for all models, for example, the performance on J48 and logistic regression is less accurate than the 70:30 train:test. The purpose of cross-validation is to evaluate the predictive ability of the model with new data, and it offers the option of evaluating the data with different data sets to obtain more accurate results.

b. From your table, analyses the results of 6 most correlated attributes and attributes proposed by the authors. Can you justify the difference, if any? [5 marks]

Comparing the results obtained between the 6 most correlated attributes selected in the selection method by WEKA, and the 6 attributes selected by the author, it is conclusive that the attributes observed by an expert can improve considerably the accuracy of the model even using different algorithms. The analysis table 1 shows that the accuracy of the algorithm ANN(MLP), J48 and logistic Regression are 97.6471 %, 95.8824 %, and 97.6471 % respectively, they are slightly less than the accuracy of the author choose, are 98.2353 %, 95.2941 %, 98.2353 % of the same algorithms. This makes more effective attributes that were used by the author. Hence, the algorithm ANN(MLP) and logistic regression have better performance on the attributes selected by the author correlation feature, their attributes have a common meaning and failed attempts to repair, love map and negative conflict behaviours (Yöntem et al., 2019).

In the document it is clear that the author used values of significance to choose the attributes given by the correlated-base feature selection method, but, first It is difficult to say if the author use the correlation with numeric or nominal values because although the correlation feature can handle nominal and numeric data (Hall, 1999), which the author do not indicate how was used. On the other hand, the correlation feature ranks the attributes indicating the correlation coefficient and the author indicates that used the values of significance obtained by applying the correlation base feature from WEKA. Then, it is not comparable to the method, but effectively slightly increases the performance of the model. Therefore. It is conclusive to say that the author used their expertise in the subject to determine the attributes combined with an uncertain method

The author attributes show correlation with the class except attribute 6 which does not clearly show the data segmentation or correlation to a specific class. it could be concluded that eliminating attribute 6 and building the model again would expect to obtain the same rate of accuracy

c. From your table, compare the results of the 6 most correlated attributes to All 54 attributes. What do they tell you about ANN and logistic regression? [5 marks]

Methods were applied to all 54 attributes and 6 most correlated attributes. The results obtained with direct application of methods with all 54 attributes showed that the highest success rate was 97.6471 % by ANN using 10-fold cross-validation while the logistic regression method obtained 96.0784 %. In contrast, the 6 most correlated attributes were 97.6471 % for ANN and 98.0392 % for logistic regression which is the highest in this comparison. the most successful result is obtained with the logistic regression model applied together with the correlation and using the test evaluation 70:30 train:test. However, it is recognised that the cross-validation purpose of cross-validation is to evaluate the predictive ability of the model with new data, then, the model built with cross-validation is more reliable. In addition, logistic regression is less prone to overfitting than ANN due to that it is simpler compared to the complex structure created by the ANN technique. This also is shown during the build model time where ANN take 42.36 secs when are 54 attributes and 1.40 secs when there are 6 attributes while logistic regression takes 0.01 secs to build the models in both cases. then again, logistic regression overperformed the ANN technique to predict and classify accurately faster.

d. Given that accuracy of predictions is a priority, for each of the 4 configurations of datasets, identify and justify which model you will choose.

The table above shows different attempts to build a model using the divorce dataset and correlation-based feature to evaluate the techniques against subsets that show differences in the correlation results. The study demonstrates the effectiveness of using the correlation feature to identify the critical attributes that can determine the resulting class, in this case, divorce or not divorce. The analysis is also used to evaluate evaluation where cross-validation remains as effective to the capability to evaluate new data and increase the accuracy of a model. Therefore, the model suggested is the logic regression that used cross-validation on the 6 most effective attributes chosen by the author understanding that those attributes choosing by the experts. The accuracy rate obtained was 98.2353% which is the same obtained by the ANN technique, but this is prone to overfitting for the complex structure, where Logistic regression is less and is simplest. Another reason is that Logistic regression requires less computational resources than the others. then, the best model to predict the outcome is Logistic Regression.

3. References

Hall, M. A. (1999). Correlation-based feature selection for machine learning.

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks, 9*(2), 42.

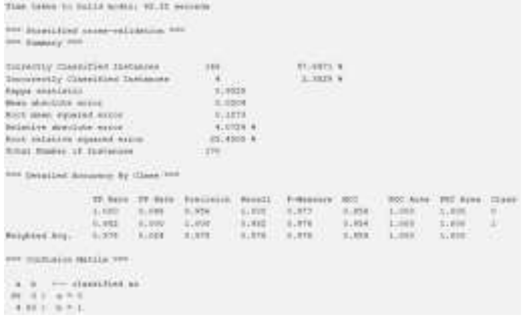
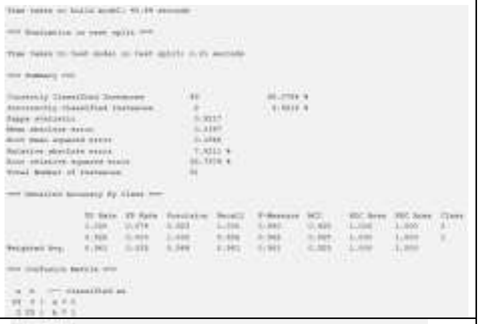
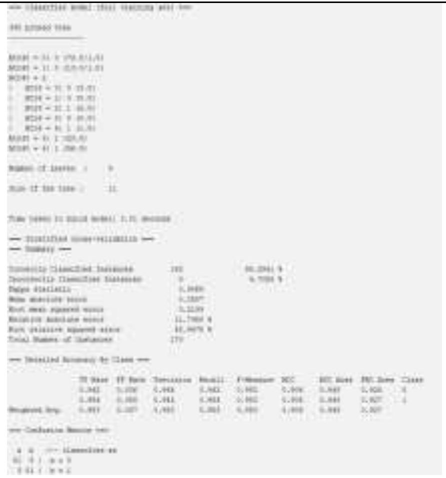
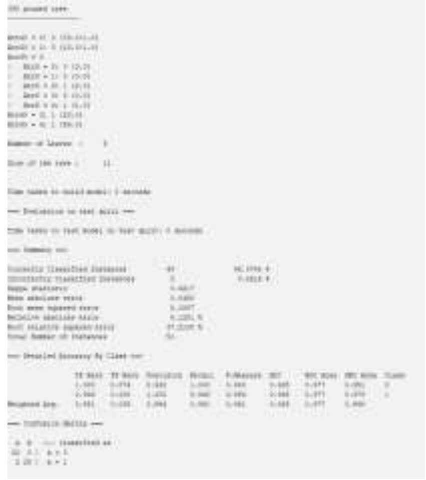
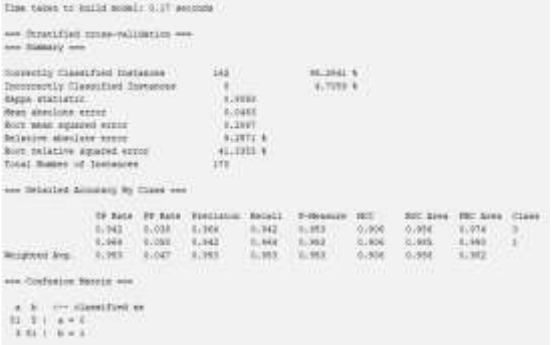
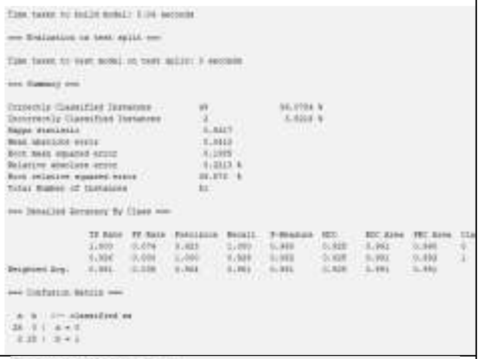
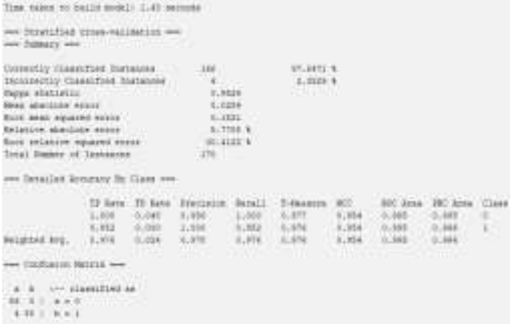
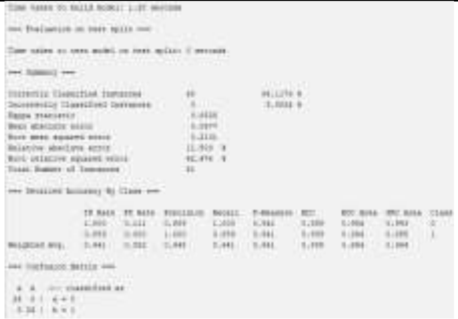
Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision support systems, 50*(3), 559-569.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks, 7*(2), 70.

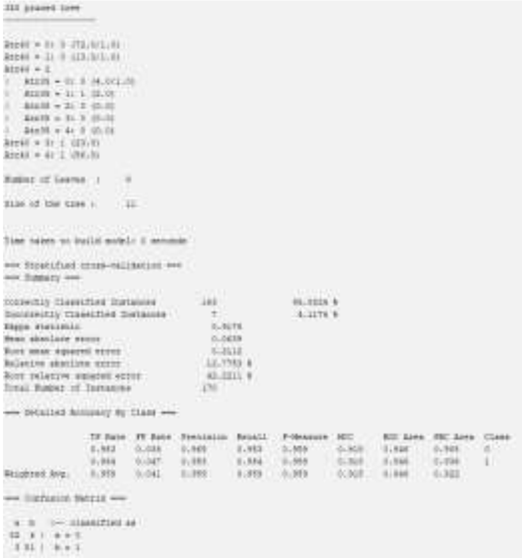
Yöntem, M. K., Kemal, A., Ilhan, T., & KILIÇARSLAN, S. (2019). Divorce prediction using correlation based feature selection and artificial neural networks. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, 9*(1), 259-273.

4. Appendix A Results

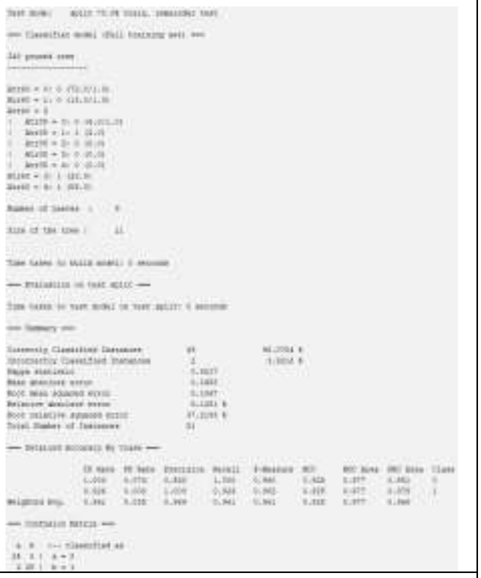
The system used in the study for the application Intel(R) Core (TM) i7-8550U CPU @ 1.80GHz 1.99 GHz processor, 8.00 GB DDR3 RAM.

Classification Technique	Number of features (Dataset)	Model (10fold cross valid.)	Model 70:30 train:test split)
ANN(MLP)	All 54 attributes (di.vorce_ nominal.arff)		
J48			
Logistic Regression			
ANN(MLP)	6 most correlated attributes / (divorce_ nomina l_most6.arff)		

J48

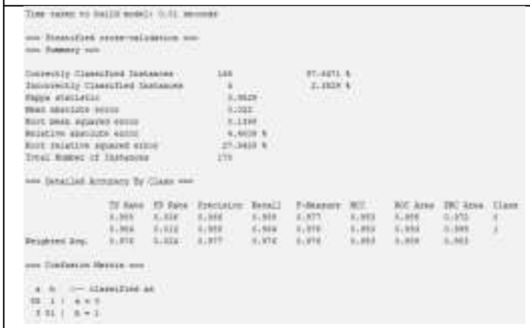


```
J48 pruned tree
------------------

Atr40 = 0: 0 (72.0/1.0)
Atr40 = 1: 0 (13.0/1.0)
Atr40 = 2
|   Atr35 = 0: 0 (4.0/1.0)
|   Atr35 = 1: 1 (2.0)
|   Atr35 = 2: 0 (0.0)
|   Atr35 = 3: 0 (0.0)
|   Atr35 = 4: 0 (0.0)
Atr40 = 3: 1 (23.0)
Atr40 = 4: 1 (56.0)

Number of Leaves  :     8

Size of the tree :      11


Time taken to build model: 0 seconds

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

Correctly Classified Instances         163               95.8824 %
Incorrectly Classified Instances         7                4.1176 %
Kappa statistic                          0.9176
Mean absolute error                      0.0439
Root mean squared error                  0.2112
Relative absolute error                 12.7753 %
Root relative squared error             42.2211 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.953    0.036    0.965      0.953    0.959      0.918    0.946     0.945     0
                 0.964    0.047    0.953      0.964    0.959      0.918    0.946     0.934     1
Weighted Avg.    0.959    0.041    0.959      0.959    0.959      0.918    0.946     0.922

=== Confusion Matrix ===

  a  b   <-- classified as
 82  4 |  a = 0
  3 81 |  b = 1
```

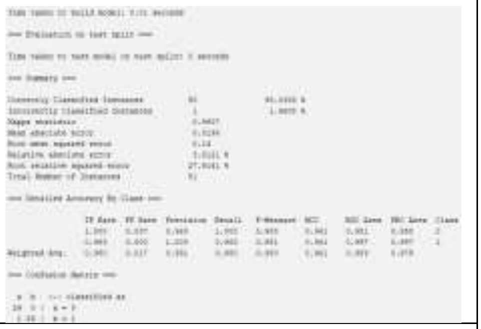


```
Test mode:    split 70.0% train, remainder test

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

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

Atr40 = 0: 0 (72.0/1.0)
Atr40 = 1: 0 (13.0/1.0)
Atr40 = 2
|   Atr35 = 0: 0 (4.0/1.0)
|   Atr35 = 1: 1 (2.0)
|   Atr35 = 2: 0 (0.0)
|   Atr35 = 3: 0 (0.0)
|   Atr35 = 4: 0 (0.0)
Atr40 = 3: 1 (23.0)
Atr40 = 4: 1 (56.0)

Number of Leaves  :     8

Size of the tree :      11


Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9217
Mean absolute error                      0.0403
Root mean squared error                  0.1947
Relative absolute error                  8.1251 %
Root relative squared error             37.2291 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.074    0.923      1.000    0.960      0.923    0.977     0.962     0
                 0.926    0.000    1.000      0.926    0.962      0.923    0.977     0.979     1
Weighted Avg.    0.961    0.035    0.964      0.961    0.961      0.923    0.977     0.966

=== Confusion Matrix ===

  a  b   <-- classified as
 24  0 |  a = 0
  2 25 |  b = 1
```

Logistic Regression



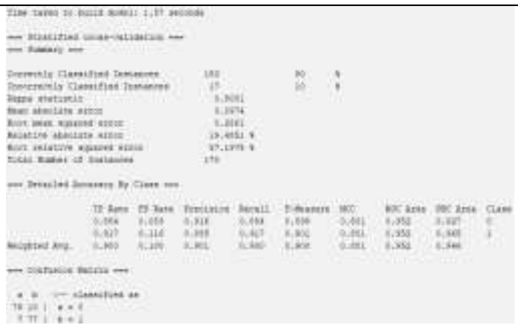
```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         166               97.6471 %
Incorrectly Classified Instances         4                2.3529 %
Kappa statistic                          0.9529
Mean absolute error                      0.022
Root mean squared error                  0.1399
Relative absolute error                  4.4009 %
Root relative squared error             27.9819 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965    0.012    0.988      0.965    0.976      0.953    0.956     0.972     0
                 0.988    0.035    0.965      0.988    0.976      0.953    0.956     0.989     1
Weighted Avg.    0.976    0.024    0.977      0.976    0.976      0.953    0.956     0.983

=== Confusion Matrix ===

  a  b   <-- classified as
 83  3 |  a = 0
  1 83 |  b = 1
```



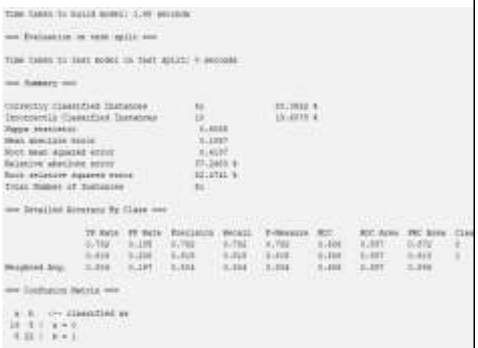
```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          50               98.0392 %
Incorrectly Classified Instances         1                1.9608 %
Kappa statistic                          0.9607
Mean absolute error                      0.0196
Root mean squared error                  0.14
Relative absolute error                  3.9221 %
Root relative squared error             27.9581 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.037    0.960      1.000    0.980      0.961    0.981     0.960     0
                 0.963    0.000    1.000      0.963    0.981      0.961    0.981     0.981     1
Weighted Avg.    0.980    0.017    0.981      0.980    0.980      0.961    0.981     0.978

=== Confusion Matrix ===

  a  b   <-- classified as
 24  0 |  a = 0
  1 26 |  b = 1
```

ANN(MLP)

6 least correlated attributes / (divorce_nominal_least6.arff)

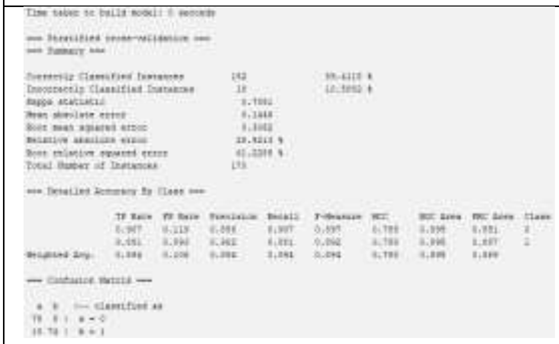


```
Time taken to build model: 1.57 seconds

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

Correctly Classified Instances         153               90      %
Incorrectly Classified Instances        17               10      %
Kappa statistic                          0.8001
Mean absolute error                      0.0974
Root mean squared error                  0.2861
Relative absolute error                 19.4851 %
Root relative squared error             57.1975 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884    0.083    0.916      0.884    0.899      0.801    0.952     0.927     0
                 0.917    0.116    0.885      0.917    0.901      0.801    0.952     0.965     1
Weighted Avg.    0.900    0.100    0.901      0.900    0.900      0.801    0.952     0.946

=== Confusion Matrix ===

  a  b   <-- classified as
 76 10 |  a = 0
  7 77 |  b = 1
```

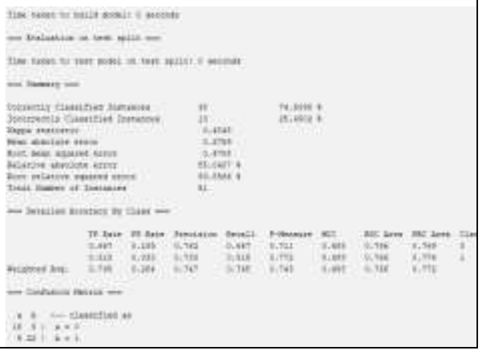


```
Time taken to build model: 1.49 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          41               80.3922 %
Incorrectly Classified Instances        10               19.6078 %
Kappa statistic                          0.6058
Mean absolute error                      0.1937
Root mean squared error                  0.4237
Relative absolute error                 37.2463 %
Root relative squared error             82.2741 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.792    0.185    0.792      0.792    0.792      0.606    0.807     0.672     0
                 0.815    0.208    0.815      0.815    0.815      0.606    0.807     0.819     1
Weighted Avg.    0.804    0.197    0.804      0.804    0.804      0.606    0.807     0.848

=== Confusion Matrix ===

  a  b   <-- classified as
 19  5 |  a = 0
  5 22 |  b = 1
```

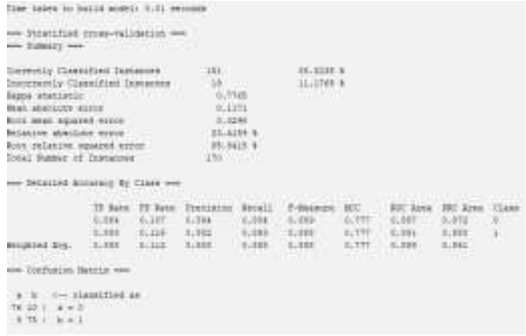
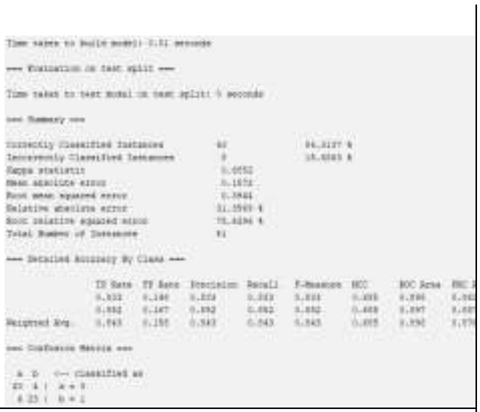
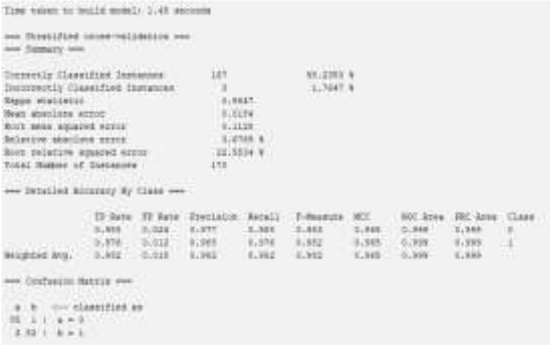
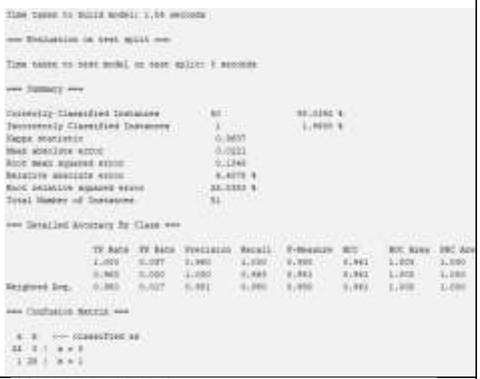
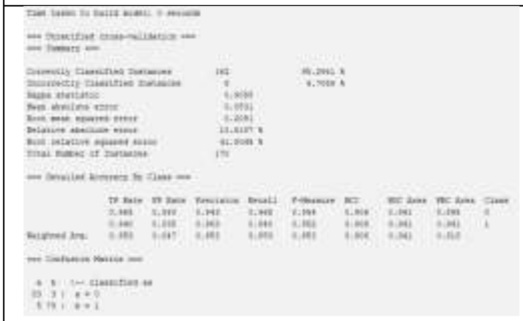
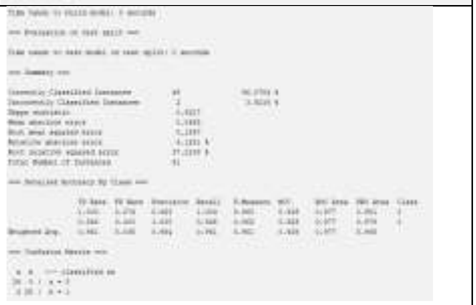
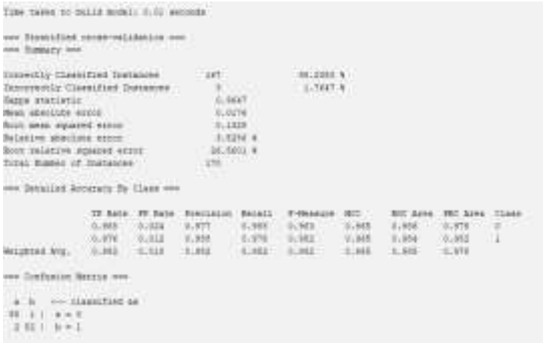
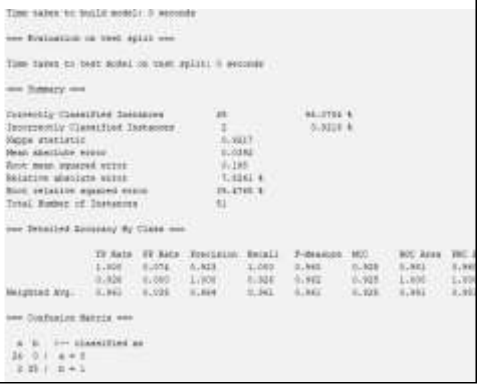
J48



```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         152               89.4118 %
Incorrectly Classified Instances        18               10.5882 %
Kappa statistic                          0.7882
Mean absolute error                      0.1448
Root mean squared error                  0.3062
Relative absolute error                 28.9213 %
Root relative squared error             61.2208 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.907    0.119    0.886      0.907    0.897      0.788    0.898     0.851     0
                 0.881    0.093    0.902      0.881    0.892      0.788    0.898     0.887     1
Weighted Avg.    0.894    0.106    0.894      0.894    0.894      0.788    0.898     0.869

=== Confusion Matrix ===

  a  b   <-- classified as
 78  8 |  a = 0
 10 74 |  b = 1
```



```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          38               74.5098 %
Incorrectly Classified Instances        13               25.4902 %
Kappa statistic                          0.4847
Mean absolute error                      0.2789
Root mean squared error                  0.4793
Relative absolute error                 55.6427 %
Root relative squared error             93.0666 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.667    0.185    0.762      0.667    0.711      0.489    0.706     0.749     0
                 0.815    0.333    0.733      0.815    0.772      0.489    0.766     0.776     1
Weighted Avg.    0.745    0.264    0.747      0.745    0.743      0.489    0.738     0.772

=== Confusion Matrix ===

  a  b   <-- classified as
 16  8 |  a = 0
  5 22 |  b = 1
```

Logistic Regression			
ANN(MLP)	Attributes proposed by the authors (divorce_nominal_author6.arff)		
J48			
Logistic Regression			

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         151               88.8235 %
Incorrectly Classified Instances        19               11.1765 %
Kappa statistic                          0.7765
Mean absolute error                      0.1171
Root mean squared error                  0.3286
Relative absolute error                 23.4159 %
Root relative squared error             65.9415 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.884    0.107    0.894      0.884    0.889      0.777    0.887     0.872     0
                 0.893    0.116    0.882      0.893    0.888      0.777    0.891     [illegible]     1
Weighted Avg.    0.888    0.112    0.888      0.888    0.888      0.777    0.889     [illegible]

=== Confusion Matrix ===

  a  b   <-- classified as
 76 10 |  a = 0
  9 75 |  b = 1
```

```
Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          43               84.3137 %
Incorrectly Classified Instances         8               15.6863 %
Kappa statistic                          0.6852
Mean absolute error                      0.1572
Root mean squared error                  0.3844
Relative absolute error                 31.3569 %
Root relative squared error             75.6296 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 0.833    0.148    0.833      0.833    0.833      0.685    0.896     0.86
                 0.852    0.167    0.852      0.852    0.852      0.685    0.897     0.87
Weighted Avg.    0.843    0.158    0.843      0.843    0.843      0.685    0.896     0.87

=== Confusion Matrix ===

  a  b   <-- classified as
 20  4 |  a = 0
  4 23 |  b = 1
```

```
Time taken to build model: 1.48 seconds

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

Correctly Classified Instances         167               98.2353 %
Incorrectly Classified Instances         3                1.7647 %
Kappa statistic                          0.9647
Mean absolute error                      0.0174
Root mean squared error                  0.1128
Relative absolute error                  3.4788 %
Root relative squared error             22.5534 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.024    0.977      0.988    0.983      0.965    0.998     0.998     0
                 0.976    0.012    0.988      0.976    0.982      0.965    0.999     0.999     1
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.965    0.999     0.999

=== Confusion Matrix ===

  a  b   <-- classified as
 85  1 |  a = 0
  2 82 |  b = 1
```

```
Time taken to build model: 1.56 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          50               98.0392 %
Incorrectly Classified Instances         1                1.9608 %
Kappa statistic                          0.9607
Mean absolute error                      0.0221
Root mean squared error                  0.1346
Relative absolute error                  4.4078 %
Root relative squared error             26.3383 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Are
                 1.000    0.034    0.957      1.000    0.978      0.961    1.000     1.000
                 0.966    0.000    1.000      0.966    0.982      0.961    1.000     1.000
Weighted Avg.    0.980    0.015    0.981      0.980    0.980      0.961    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 22  0 |  a = 0
  1 28 |  b = 1
```

```
Time taken to build model: 0 seconds

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

Correctly Classified Instances         162               95.2941 %
Incorrectly Classified Instances         8                4.7059 %
Kappa statistic                          0.9058
Mean absolute error                      0.0731
Root mean squared error                  0.2081
Relative absolute error                 13.8137 %
Root relative squared error             41.0068 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.965    0.060    0.943      0.965    0.954      0.906    0.961     [illegible]     0
                 0.940    0.035    0.963      0.940    0.952      0.906    0.961     0.961     1
Weighted Avg.    0.953    0.047    0.953      0.953    0.953      0.906    0.961     [illegible]

=== Confusion Matrix ===

  a  b   <-- classified as
 83  3 |  a = 0
  5 79 |  b = 1
```

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9217
Mean absolute error                      [illegible]
Root mean squared error                  [illegible]
Relative absolute error                  [illegible] %
Root relative squared error             [illegible] %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.074    0.923      1.000    0.960      0.923    0.977     [illegible]     0
                 0.926    0.000    1.000      0.926    0.962      0.923    0.977     [illegible]     1
Weighted Avg.    0.961    0.035    0.964      0.961    0.961      0.923    0.977     [illegible]

=== Confusion Matrix ===

  a  b   <-- classified as
 24  0 |  a = 0
  2 25 |  b = 1
```

```
Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         167               98.2353 %
Incorrectly Classified Instances         3                1.7647 %
Kappa statistic                          0.9647
Mean absolute error                      0.0176
Root mean squared error                  0.1329
Relative absolute error                  3.5294 %
Root relative squared error             26.5801 %
Total Number of Instances              170

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.988    0.024    0.977      0.988    0.983      0.965    0.956     0.978     0
                 0.976    0.012    0.988      0.976    0.982      0.965    0.954     0.952     1
Weighted Avg.    0.982    0.018    0.982      0.982    0.982      0.965    0.955     0.970

=== Confusion Matrix ===

  a  b   <-- classified as
 85  1 |  a = 0
  2 82 |  b = 1
```

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances          49               96.0784 %
Incorrectly Classified Instances         2                3.9216 %
Kappa statistic                          0.9217
Mean absolute error                      0.0392
Root mean squared error                  0.198
Relative absolute error                  7.8261 %
Root relative squared error             39.4785 %
Total Number of Instances               51

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC A
                 1.000    0.074    0.923      1.000    0.960      0.925    0.961     0.96
                 0.926    0.000    1.000      0.926    0.962      0.925    1.000     1.00
Weighted Avg.    0.961    0.035    0.964      0.961    0.961      0.925    0.981     0.98

=== Confusion Matrix ===

  a  b   <-- classified as
 24  0 |  a = 0
  2 25 |  b = 1
```