CS699 A2 Data Mining (Spring 2022)

Project Topic: How do the existing factors influence forest fires?

Team Members: Jiongkui Jin & Sha Hu

Date: 04/05/2022

1. Introduction and Motivation

Forest fires have been one of the most essential factors which lead to property loss and environment damage in North America. A great deal of manpower and material resources have been put into fighting the forest fires, however, the problem still remains severe and brings great harm to the natural environment.

Therefore, prevention and control of forest fires become even more important and a lot of valuable data has been gathered to analyze the potential reasons and affecting existing factors in the formation of forest fires.

By executing data mining and data analysis on this problem, our team aims to find potential reasons for forest fires in North America by analyzing the degree of correlation between existing factors and forest fires. After that, we summarized the findings and elaborated potential implications as well as possible improvements by combining the analysis results with the actual situation.

2. Data Preparation

We chose real data from EarthData which provides access to all DAAC data via a map web-based interface. The dataset we chose provides the results of field measurements and estimates of carbon stocks as well as combustion rates that characterize burned and unburned forest. The districts of the forests are southern boreal near the La Ronge and Weyakwin communities in central Saskatchewan (SK), a typical forest area in Canada.

The dataset was published in 2020 while the measurements were completed in 2016 at 47 stands that burned in the 2015 Saskatchewan wildfires and at 32 unburned stands in comparable adjacent areas.

Stands were characterized through field observations, sampling of the vegetative community and soils as well as basic landscape geophysical traits. Among them, the sampling of the vegetative community includes tree species, abundance, biophysical measurements, stand age, coarse woody debris, history of fires or logging. The sampling of soils includes soil moisture class, unburned and burned soil organic layer depth, samples for bulk density and carbon analyses.

From these results, the pre-fire carbon stocks and carbon combustion values from both the above and below ground pools were estimated using a combination of linear and mixed-effects modeling and were calibrated against carbons stocks from the unburned stands.

Estimates of uncertainties were generated for above and below ground carbon stocks and combustion values using a Monte Carlo framework paired with classic uncertainty propagation techniques.

After selection of this dataset, we executed data preprocessing using Weka and Excel. During this process, we firstly removed the irrelevant attributes using Weka. Secondly, we handled the missing value by excel. Fill in it automatically with the

attribute mean for all samples belonging to the same class.All of the missing values (-9999) were replaced by the average number of the rest values in the same attribute. Finally, we removed the outliers using Weka.

The details of Data processing is in appendix1.

3. Data Mining

Firstly, we determined our data mining goal by analyzing the current forest fires situation in North America, especially Canada. After that, we found our desired dataset in EarthData.

Secondly, we pursued data preprocessing by deleting irrelevant attributes, handling missing values and removing outliers.

Thirdly, we chose five classification algorithms and five attribute selection methods.

Fourthly, for each classification algorithm, we build a classification model from the preprocessed dataset and test the model using 10-fold cross-validation, then collect and keep the performance result.

Fifthly, from each reduced training dataset, we build five classification models (using the five classification algorithms we selected above) and test them on the corresponding reduced test dataset.

Sixthly, we compared the 25 models and selected one model as the best model.

Finally, we compared the best model with the model that was built using the same classification algorithm from the preprocessing dataset.

The data mining tools, classification algorithms and attribute selection are explained detailedly as following:

3.1 Data Mining Tools

We used Weka, Excel , Word and R to pursue data mining. Among them, Weka is used for preprocessing data, building and testing models as well as comparing different models. Excel is used to execute data preparation. Word and R is used for tables.

3.2 Classification Algorithms

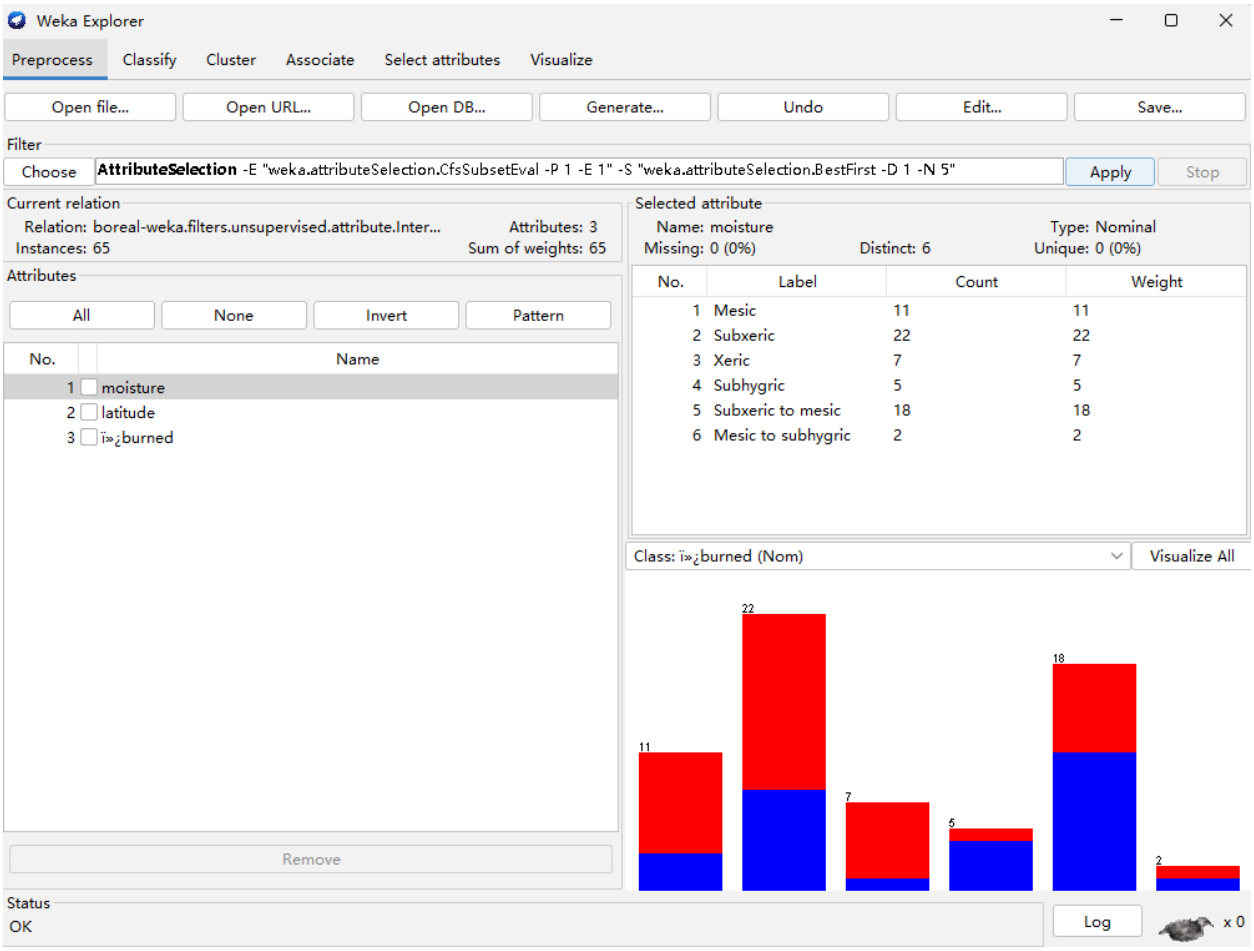
We chose five classification algorithms including Function-Logistic, Bayes-NativeBayes, Lazy-IBK(K=10), meta-Bagging(RandomForest) and tree-J48.

3.3 Attribute Selection

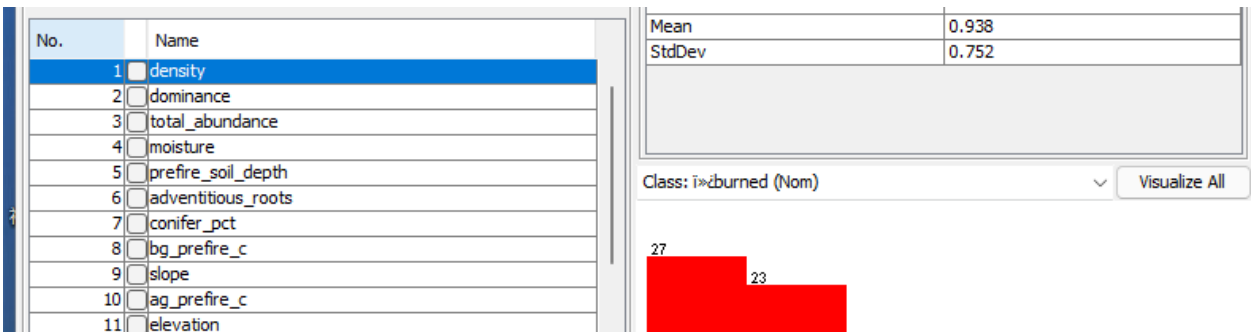
We imported the prepared dataset into Weka and chose Filter-supervised-attribute-AttributeSelection. Among all the attribute selection methods, we chose five and used them to determine our sets of attributes.

The first attribute selection method we used is BestFirst of CfsBubsetEval Evaluator. The other four attribute selection methods we used are Ranker Method in CorrelationAttributeEval, GainRatioAttributeEval, OneRAttributeEval and ClassifierAttributeEval Evaluators.

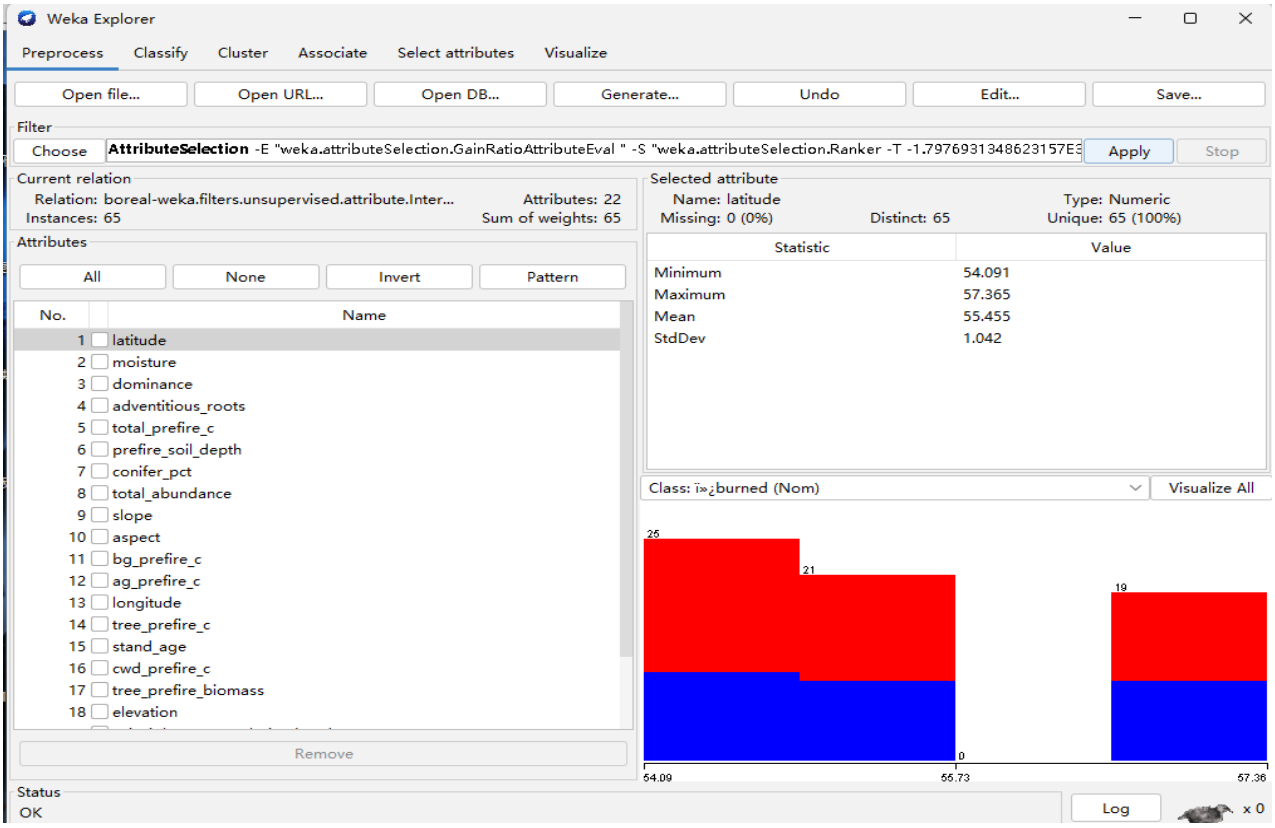
The sets of attributes we selected using those five attribute selection methods are as following:



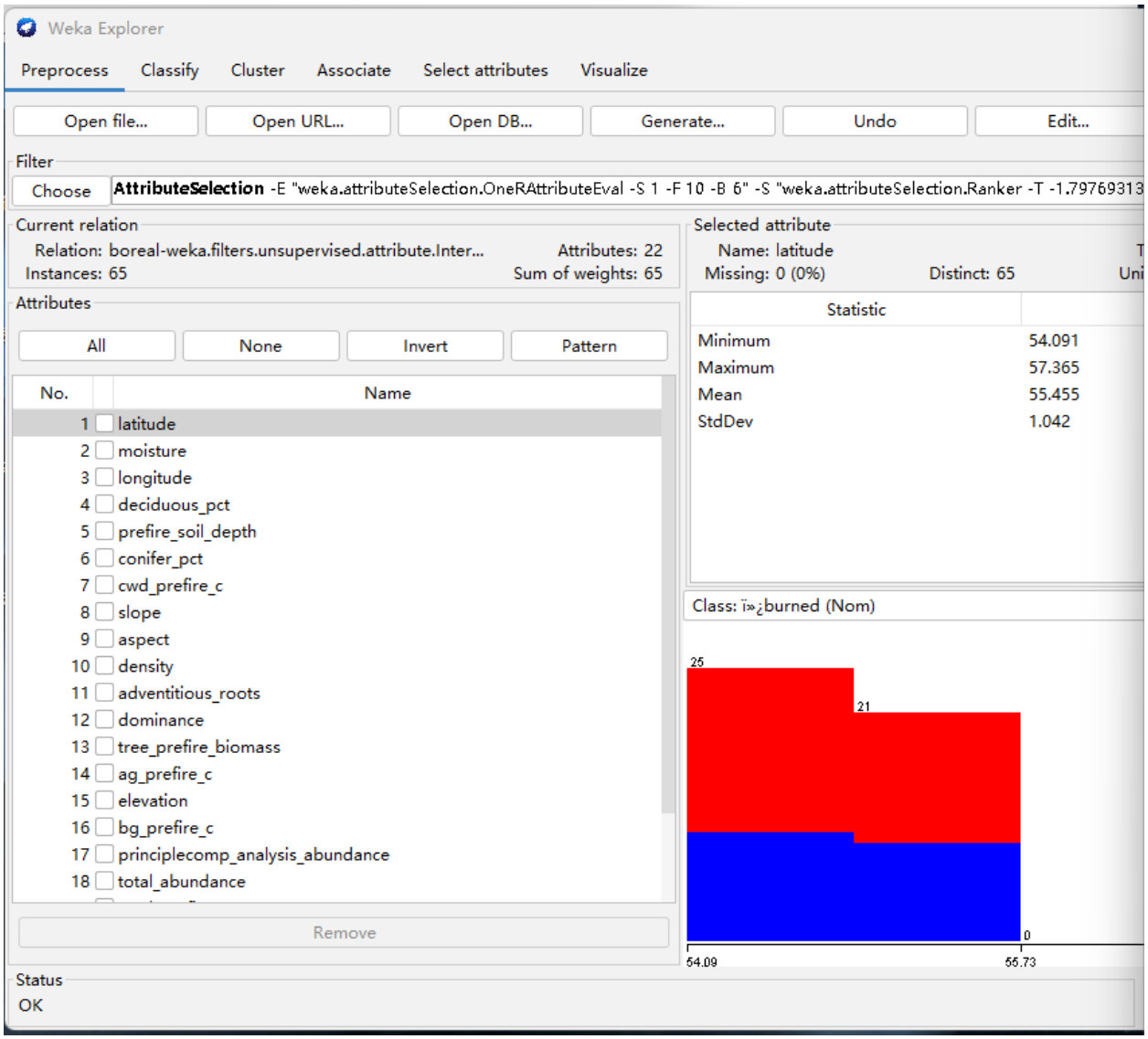
We selected the moisture and latitude attributes except the class attribute using the first attribute selection method. After that, we put the selected attributes into a new data set which is called Dataset.1 in this report.



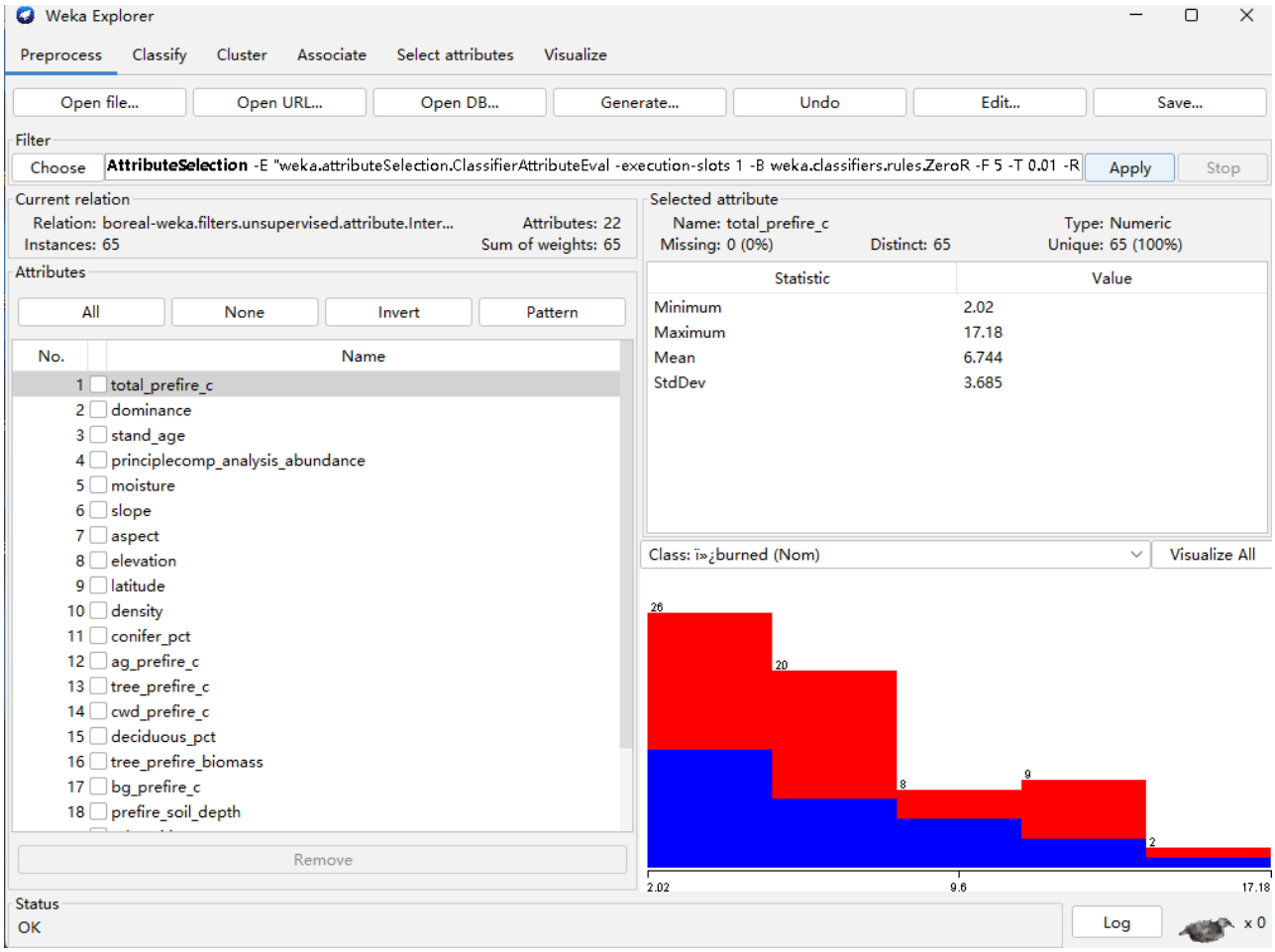
The Dataset.2 in this report includes density, dominance, total_abundance and moisture attributes except the class attribute. This dataset was selected using the Ranker Method of CorrelationAttributeEval Evaluator.



The third attribute selection method is the Ranker Method of GainRatioAttributeEval Evaluator. We selected the first four attributes according to the ranks of them which are latitude, moisture， dominanc and abentitious_roots attributes. After that, we put them into a new data set called Dataset.3



The Dataset.4 contains latitude, moisture, longitude and deciduous_pct attributes except the class attribute. We selected these four attributes using the Ranker Method of OneRAttributeEval Evaluator.



The Dataset.5 contains total_prefire_c, dominance, stand_age and principlecomp_analysis_abundance attributes except the class attribute was selected using the same Ranker Method of ClassifierAttributeEval Evaluator.

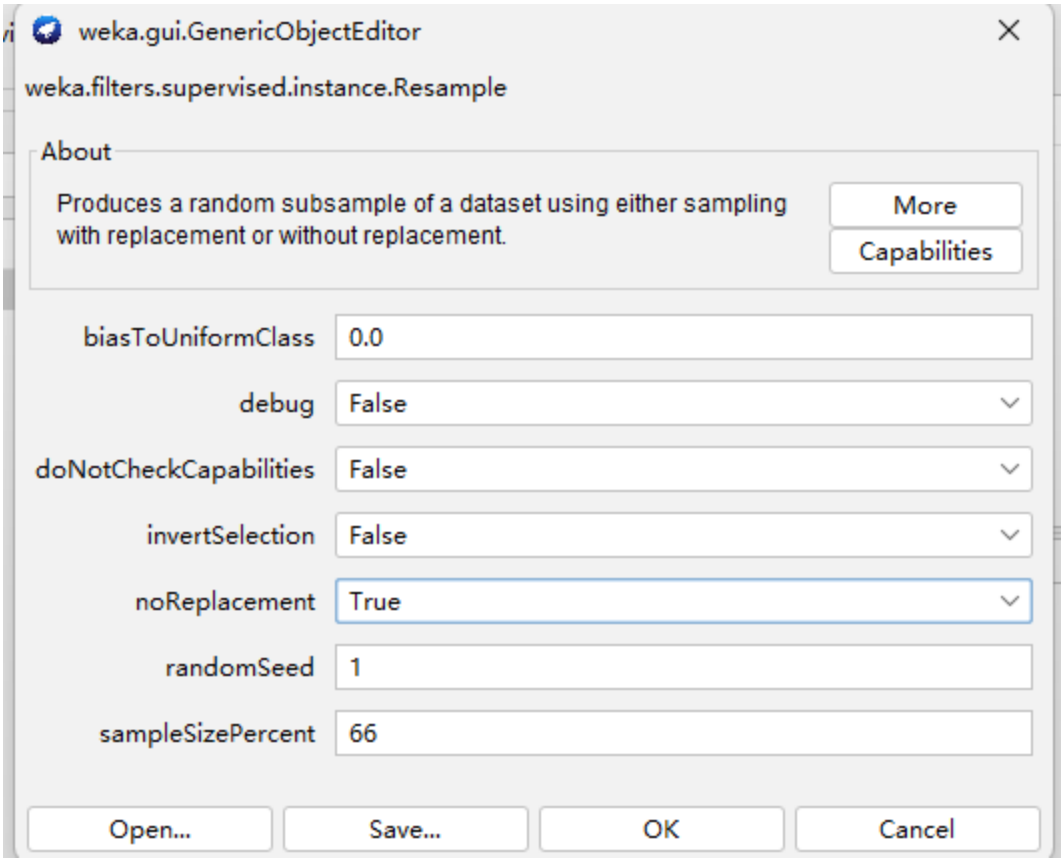
4. Analysis

We executed the data analysis process in four steps which are explained detailedly as following:

4.1

We split the reduced dataset into a training dataset and a test dataset using weka's filters- supervised- instance- resample.

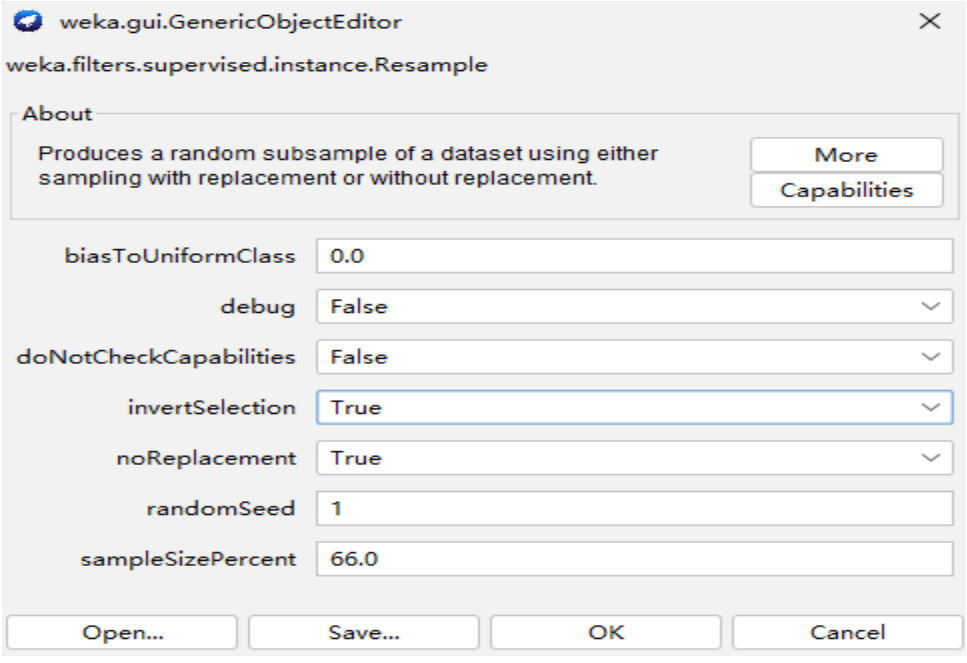
The parameters are set as shown in the figure below.



We applied the filter, and got the training set of the Dataset.1, then we saved the result as Dataset.1-train.

After that, we returned to Dataset.1 and used weka's filters- supervised- instance- resample.

The parameters are set as shown in the figure below.



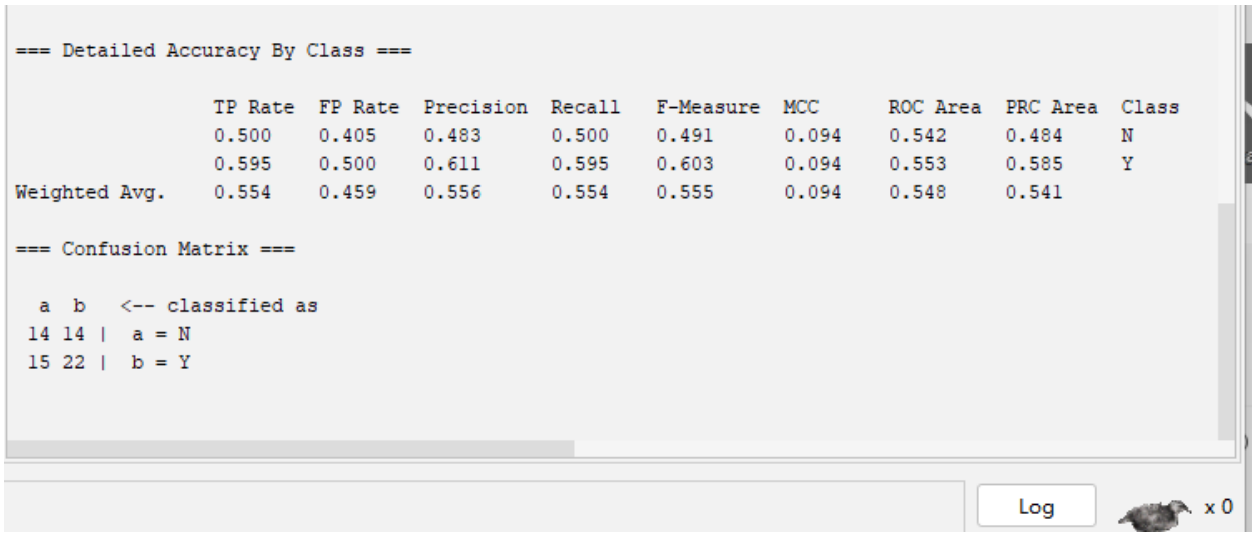
Then we applied the filter, got the test dataset of the Dataset.1 and saved the result as Dataset.1-test.

We pursued the same process for Dataset.2,3,4,5 using the same method and saved the results as Dataset.2-train,Dataset.2-test.Dataset.3-train,Dataset.3-test.Dataset.4-train,Dataset.4-test.Dataset.5-train,Dataset.5-test.

4.2.Build and test the models

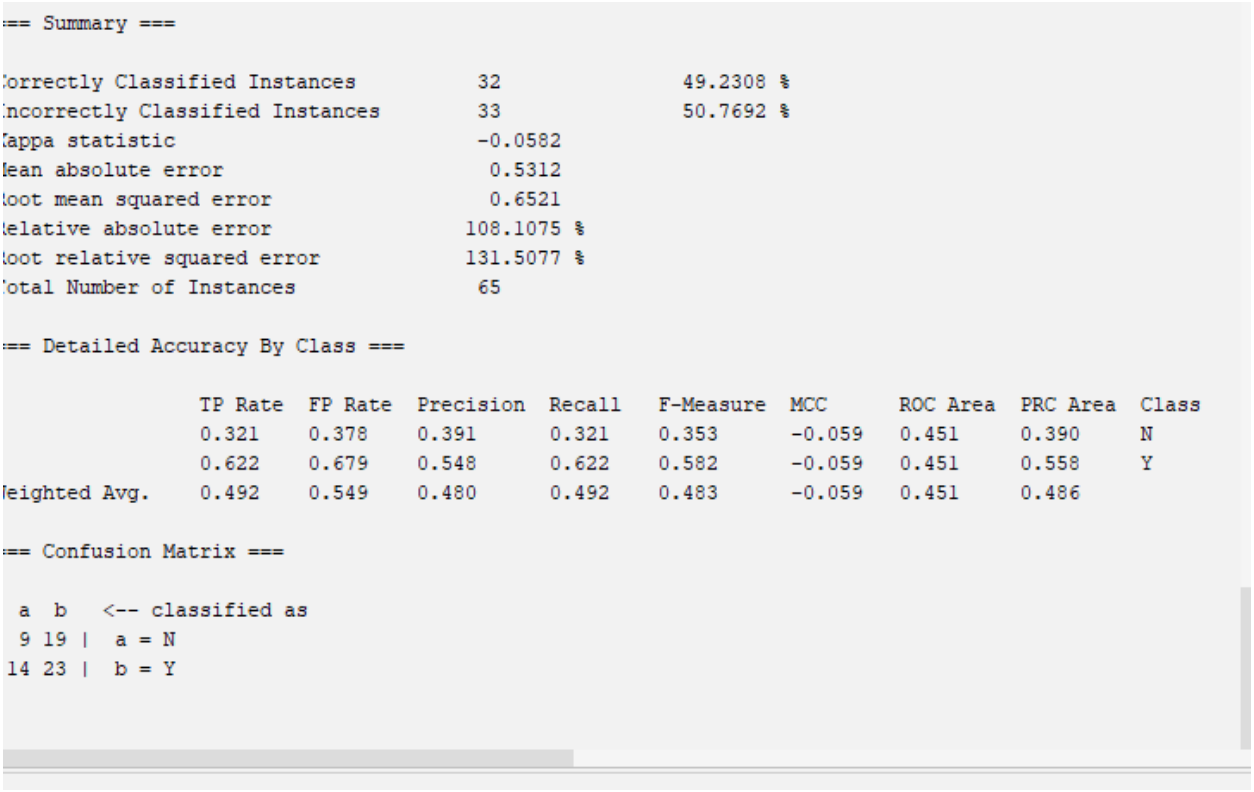
4.2.1 For each classification algorithm, we built a classification model from the preprocessed dataset and tested it using 10-fold cross-validation. The results are as following:

Logistic



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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.405    0.483      0.500    0.491      0.094    0.542     0.484     N
                 0.595    0.500    0.611      0.595    0.603      0.094    0.553     0.585     Y
Weighted Avg.    0.554    0.459    0.556      0.554    0.555      0.094    0.548     0.541

=== Confusion Matrix ===

  a  b   <-- classified as
 14 14 |  a = N
 15 22 |  b = Y
```


NativeBayes



```
=== Summary ===

Correctly Classified Instances          32               49.2308 %
Incorrectly Classified Instances        33               50.7692 %
Kappa statistic                         -0.0582
Mean absolute error                      0.5312
Root mean squared error                  0.6521
Relative absolute error                108.1075 %
Root relative squared error            131.5077 %
Total Number of Instances               65

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.321    0.378    0.391      0.321    0.353      -0.059   0.451     0.390     N
                 0.622    0.679    0.548      0.622    0.582      -0.059   0.451     0.558     Y
Weighted Avg.    0.492    0.549    0.480      0.492    0.483      -0.059   0.451     0.486

=== Confusion Matrix ===

  a  b   <-- classified as
  9 19 |  a = N
 14 23 |  b = Y
```

Bagging(RandomForest)

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.250    0.162    0.538      0.250    0.341      0.109    0.673     0.576     N
               0.838    0.750    0.596      0.838    0.697      0.109    0.673     0.716     Y
Weighted Avg.  0.585    0.497    0.571      0.585    0.544      0.109    0.673     0.656

=== Confusion Matrix ===

  a  b   <-- classified as
  7 21 |  a = N
  6 31 |  b = Y
```

Log x 0

IBK(K=10)

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.214    0.405    0.286      0.214    0.245      -0.202   0.503     0.430     N
               0.595    0.786    0.500      0.595    0.543      -0.202   0.503     0.625     Y
Weighted Avg.  0.431    0.622    0.408      0.431    0.415      -0.202   0.503     0.541

=== Confusion Matrix ===

  a  b   <-- classified as
  6 22 |  a = N
 15 22 |  b = Y
```

Log x 0

J48

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.500    0.432    0.467      0.500    0.483      0.067    0.534     0.465     N
               0.568    0.500    0.600      0.568    0.583      0.067    0.534     0.593     Y
Weighted Avg.  0.538    0.471    0.543      0.538    0.540      0.067    0.534     0.538

=== Confusion Matrix ===

  a  b   <-- classified as
 14 14 |  a = N
 16 21 |  b = Y
```

Log x 0

4.2.2 Build and test the models(25models)

Then we used the five classification algorithms to build 25 classification models from the five dataset of the five sets of attributes after the process of attributes selection. The results are exhibited as graphs of training dataset and test dataset as following:

Dataset.1 Logistic

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.611    0.167    0.733      0.611    0.667      0.459    0.843     0.811     N
                 0.833    0.389    0.741      0.833    0.784      0.459    0.843     0.887     Y
Weighted Avg.    0.738    0.294    0.738      0.738    0.734      0.459    0.843     0.854

=== Confusion Matrix ===

  a  b   <-- classified as
 11  7 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.385    0.444      0.400    0.421      0.016    0.500     0.550     N
                 0.615    0.600    0.571      0.615    0.593      0.016    0.500     0.574     Y
Weighted Avg.    0.522    0.506    0.516      0.522    0.518      0.016    0.500     0.564

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.1 NativeBayes

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.167    0.750      0.667    0.706      0.510    0.844     0.810     N
                 0.833    0.333    0.769      0.833    0.800      0.510    0.844     0.887     Y
Weighted Avg.    0.762    0.262    0.761      0.762    0.760      0.510    0.844     0.854

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = N
  4 20 |  b = Y
```

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.400    0.385    0.444      0.400    0.421      0.016    0.469     0.541     N
                0.615    0.600    0.571      0.615    0.593      0.016    0.469     0.540     Y
Weighted Avg.   0.522    0.506    0.516      0.522    0.518      0.016    0.469     0.540

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.1 IBK(K=10)

Train

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.611    0.167    0.733      0.611    0.667      0.459    0.750     0.652     N
                0.833    0.389    0.741      0.833    0.784      0.459    0.750     0.779     Y
Weighted Avg.   0.738    0.294    0.738      0.738    0.734      0.459    0.750     0.724

=== Confusion Matrix ===

  a  b   <-- classified as
 11  7 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.500    0.462    0.455      0.500    0.476      0.038    0.565     0.467     N
                0.538    0.500    0.583      0.538    0.560      0.038    0.565     0.668     Y
Weighted Avg.   0.522    0.483    0.527      0.522    0.524      0.038    0.565     0.581

=== Confusion Matrix ===

 a b   <-- classified as
 5 5 | a = N
 6 7 | b = Y
```

Log x 0

Dataset.1 Bagging-RandomForest

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.042    0.944      0.944    0.944      0.903    0.995     0.994     N
                 0.958    0.056    0.958      0.958    0.958      0.903    0.995     0.997     Y
Weighted Avg.    0.952    0.050    0.952      0.952    0.952      0.903    0.995     0.996

=== Confusion Matrix ===

  a  b   <-- classified as
 17  1 |  a = N
  1 23 |  b = Y
```

Log x 0

1.evaluator: CfsBubsetEval search Methon:BestFirst

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.231    0.625      0.500    0.556      0.280    0.612     0.585     N
                 0.769    0.500    0.667      0.769    0.714      0.280    0.612     0.696     Y
Weighted Avg.    0.652    0.383    0.649      0.652    0.645      0.280    0.612     0.648

=== Confusion Matrix ===

  a  b   <-- classified as
  5  5 |  a = N
  3 10 |  b = Y
```

Log x 0

Dataset.1 J48

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.167    0.750      0.667    0.706      0.510    0.839     0.745     N
                 0.833    0.333    0.769      0.833    0.800      0.510    0.839     0.845     Y
Weighted Avg.    0.762    0.262    0.761      0.762    0.760      0.510    0.839     0.802

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.400    0.385    0.444      0.400    0.421      0.016    0.481     0.437     N
               0.615    0.600    0.571      0.615    0.593      0.016    0.481     0.549     Y
Weighted Avg.  0.522    0.506    0.516      0.522    0.518      0.016    0.481     0.500

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.2 Logistic

Train

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.722    0.125    0.813      0.722    0.765      0.609    0.876     0.844     N
               0.875    0.278    0.808      0.875    0.840      0.609    0.876     0.916     Y
Weighted Avg.  0.810    0.212    0.810      0.810    0.808      0.609    0.876     0.885

=== Confusion Matrix ===

  a  b   <-- classified as
 13  5 |  a = N
  3 21 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.400    0.462    0.400      0.400    0.400      -0.062   0.438     0.476     N
               0.538    0.600    0.538      0.538    0.538      -0.062   0.431     0.525     Y
Weighted Avg.  0.478    0.540    0.478      0.478    0.478      -0.062   0.434     0.504

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 6 7 | b = Y
```

Log x 0

Dataset.2 NativeBayes

Train

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.611    0.167    0.733      0.611    0.667      0.459    0.789     0.748     N
               0.833    0.389    0.741      0.833    0.784      0.459    0.789     0.848     Y
Weighted Avg.  0.738    0.294    0.738      0.738    0.734      0.459    0.789     0.805

=== Confusion Matrix ===

  a  b   <-- classified as
 11  7 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.300    0.385    0.375      0.300    0.333      -0.088   0.508     0.555     N
               0.615    0.700    0.533      0.615    0.571      -0.088   0.508     0.580     Y
Weighted Avg.  0.478    0.563    0.464      0.478    0.468      -0.088   0.508     0.569

=== Confusion Matrix ===

 a b   <-- classified as
 3 7 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.2 IBK(K=10)

Train

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.889    0.333    0.667      0.889    0.762      0.556    0.830     0.739     N
               0.667    0.111    0.889      0.667    0.762      0.556    0.830     0.839     Y
Weighted Avg.  0.762    0.206    0.794      0.762    0.762      0.556    0.830     0.796

=== Confusion Matrix ===

  a  b   <-- classified as
 16  2 |  a = N
  8 16 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.308    0.556      0.500    0.526      0.195    0.538     0.554     N
                 0.692    0.500    0.643      0.692    0.667      0.195    0.538     0.673     Y
Weighted Avg.    0.609    0.416    0.605      0.609    0.606      0.195    0.538     0.621

=== Confusion Matrix ===

 a b   <-- classified as
 5 5 | a = N
 4 9 | b = Y
```

Log x 0

Dataset.2 Bagging-RandomForest
Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.167    0.810      0.944    0.872      0.770    0.964     0.948     N
                 0.833    0.056    0.952      0.833    0.889      0.770    0.964     0.975     Y
Weighted Avg.    0.881    0.103    0.891      0.881    0.882      0.770    0.964     0.963

=== Confusion Matrix ===

  a  b   <-- classified as
 17  1 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.200    0.385    0.286      0.200    0.235      -0.199   0.438     0.414     N
                 0.615    0.800    0.500      0.615    0.552      -0.199   0.438     0.583     Y
Weighted Avg.    0.435    0.619    0.407      0.435    0.414      -0.199   0.438     0.510

=== Confusion Matrix ===

 a b   <-- classified as
 2 8 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.2 J48

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.167    0.750      0.667    0.706      0.510    0.839     0.745     N
                 0.833    0.333    0.769      0.833    0.800      0.510    0.839     0.845     Y
Weighted Avg.    0.762    0.262    0.761      0.762    0.760      0.510    0.839     0.802

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.385    0.444      0.400    0.421      0.016    0.481     0.437     N
                 0.615    0.600    0.571      0.615    0.593      0.016    0.481     0.549     Y
Weighted Avg.    0.522    0.506    0.516      0.522    0.518      0.016    0.481     0.500

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.3 Logistic

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.611    0.208    0.688      0.611    0.647      0.410    0.850     0.843     N
                 0.792    0.389    0.731      0.792    0.760      0.410    0.850     0.891     Y
Weighted Avg.    0.714    0.312    0.712      0.714    0.712      0.410    0.850     0.870

=== Confusion Matrix ===

  a  b   <-- classified as
 11  7 |  a = N
  5 19 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.400    0.615    0.333      0.400    0.364      -0.214   0.338     0.363     N
                0.385    0.600    0.455      0.385    0.417      -0.214   0.335     0.514     Y
Weighted Avg.   0.391    0.607    0.402      0.391    0.394      -0.214   0.336     0.448

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 8 5 | b = Y
```

Log x 0

Dataset.3 NativeBayes

Train

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.667    0.208    0.706      0.667    0.686      0.462    0.850     0.834     N
                0.792    0.333    0.760      0.792    0.776      0.462    0.850     0.891     Y
Weighted Avg.   0.738    0.280    0.737      0.738    0.737      0.462    0.850     0.867

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = N
  5 19 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.400    0.385    0.444      0.400    0.421      0.016    0.454     0.477     N
                0.615    0.600    0.571      0.615    0.593      0.016    0.454     0.624     Y
Weighted Avg.   0.522    0.506    0.516      0.522    0.518      0.016    0.454     0.560

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.3 IBK(K=10)

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.722    0.125    0.813      0.722    0.765      0.609    0.881     0.815     N
                 0.875    0.278    0.808      0.875    0.840      0.609    0.881     0.881     Y
Weighted Avg.    0.810    0.212    0.810      0.810    0.808      0.609    0.881     0.853

=== Confusion Matrix ===

  a  b   <-- classified as
 13  5 |  a = N
  3 21 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.500    0.385    0.500      0.500    0.500      0.115    0.569     0.456     N
                 0.615    0.500    0.615      0.615    0.615      0.115    0.569     0.675     Y
Weighted Avg.    0.565    0.450    0.565      0.565    0.565      0.115    0.569     0.580

=== Confusion Matrix ===

 a b   <-- classified as
 5 5 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.3 Bagging-RandomForest

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.042    0.944      0.944    0.944      0.903    0.998     0.997     N
                 0.958    0.056    0.958      0.958    0.958      0.903    0.998     0.998     Y
Weighted Avg.    0.952    0.050    0.952      0.952    0.952      0.903    0.998     0.998

=== Confusion Matrix ===

  a  b   <-- classified as
 17  1 |  a = N
  1 23 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.500    0.308    0.556      0.500    0.526      0.195    0.623     0.626     N
                0.692    0.500    0.643      0.692    0.667      0.195    0.623     0.722     Y
Weighted Avg.   0.609    0.416    0.605      0.609    0.606      0.195    0.623     0.681

=== Confusion Matrix ===

 a b   <-- classified as
 5 5 | a = N
 4 9 | b = Y
```

Log x 0

Dataset.3 J48

Train

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.722    0.083    0.867      0.722    0.788      0.660    0.928     0.879     N
                0.917    0.278    0.815      0.917    0.863      0.660    0.928     0.934     Y
Weighted Avg.   0.833    0.194    0.837      0.833    0.831      0.660    0.928     0.910

=== Confusion Matrix ===

  a  b   <-- classified as
 13  5 |  a = N
  2 22 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.300    0.385    0.375      0.300    0.333      -0.088   0.465     0.419     N
                0.615    0.700    0.533      0.615    0.571      -0.088   0.465     0.556     Y
Weighted Avg.   0.478    0.563    0.464      0.478    0.468      -0.088   0.465     0.496

=== Confusion Matrix ===

 a b   <-- classified as
 3 7 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.4 Logistic
Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.778    0.208    0.737      0.778    0.757      0.566    0.889     0.881     N
                 0.792    0.222    0.826      0.792    0.809      0.566    0.889     0.919     Y
Weighted Avg.    0.786    0.216    0.788      0.786    0.786      0.566    0.889     0.903

=== Confusion Matrix ===

  a  b   <-- classified as
 14  4 |  a = N
  5 19 |  b = Y
```

Log x 0

Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.538    0.364      0.400    0.381      -0.137   0.454     0.511     N
                 0.462    0.600    0.500      0.462    0.480      -0.137   0.454     0.556     Y
Weighted Avg.    0.435    0.573    0.441      0.435    0.437      -0.137   0.454     0.536

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 7 6 | b = Y
```

Log x 0

Dataset.4 NativeBayes
Trian

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.722    0.167    0.765      0.722    0.743      0.560    0.833     0.816     N
                 0.833    0.278    0.800      0.833    0.816      0.560    0.833     0.848     Y
Weighted Avg.    0.786    0.230    0.785      0.786    0.785      0.560    0.833     0.834

=== Confusion Matrix ===

  a  b   <-- classified as
 13  5 |  a = N
  4 20 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.400    0.385    0.444      0.400    0.421      0.016    0.492     0.528     N
               0.615    0.600    0.571      0.615    0.593      0.016    0.492     0.577     Y
Weighted Avg.  0.522    0.506    0.516      0.522    0.518      0.016    0.492     0.556

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.4 IBK(K=10)

Train

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.611    0.125    0.786      0.611    0.688      0.510    0.824     0.682     N
               0.875    0.389    0.750      0.875    0.808      0.510    0.824     0.851     Y
Weighted Avg.  0.762    0.276    0.765      0.762    0.756      0.510    0.824     0.779

=== Confusion Matrix ===

  a  b   <-- classified as
 11  7 |  a = N
  3 21 |  b = Y
```

Log x 0

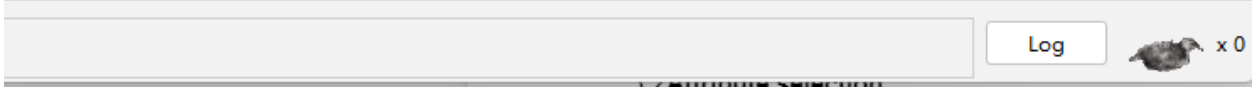
Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.200    0.308    0.333      0.200    0.250      -0.122   0.454     0.422     N
               0.692    0.800    0.529      0.692    0.600      -0.122   0.454     0.584     Y
Weighted Avg.  0.478    0.586    0.444      0.478    0.448      -0.122   0.454     0.514

=== Confusion Matrix ===

 a b   <-- classified as
 2 8 | a = N
 4 9 | b = Y
```

Log x 0

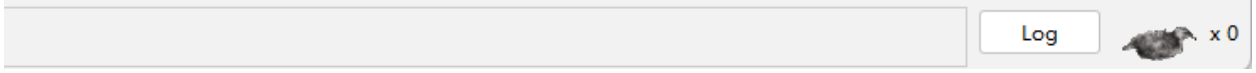
Dataset.4 Bagging-RandomForest
Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     N
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Y
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

  a  b   <-- classified as
 18  0 |  a = N
  0 24 |  b = Y
```



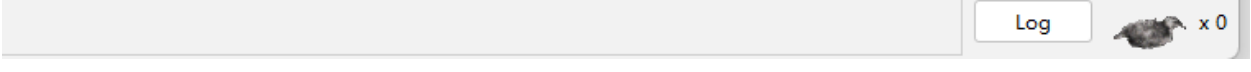
Test

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.400    0.385    0.444      0.400    0.421      0.016    0.523     0.545     N
                 0.615    0.600    0.571      0.615    0.593      0.016    0.523     0.629     Y
Weighted Avg.    0.522    0.506    0.516      0.522    0.518      0.016    0.523     0.592

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 5 8 | b = Y
```



Dataset.4 J48
Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.889    0.167    0.800      0.889    0.842      0.716    0.927     0.874     N
                 0.833    0.111    0.909      0.833    0.870      0.716    0.927     0.928     Y
Weighted Avg.    0.857    0.135    0.862      0.857    0.858      0.716    0.927     0.905

=== Confusion Matrix ===

  a  b   <-- classified as
 16  2 |  a = N
  4 20 |  b = Y
```



Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.400    0.538    0.364      0.400    0.381      -0.137   0.412     0.386     N
               0.462    0.600    0.500      0.462    0.480      -0.137   0.412     0.523     Y
Weighted Avg.  0.435    0.573    0.441      0.435    0.437      -0.137   0.412     0.464

=== Confusion Matrix ===

 a b   <-- classified as
 4 6 | a = N
 7 6 | b = Y
```

Log x 0

Dataset.5 Logistic

Train

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.667    0.208    0.706      0.667    0.686      0.462    0.738     0.688     N
               0.792    0.333    0.760      0.792    0.776      0.462    0.738     0.813     Y
Weighted Avg.  0.738    0.280    0.737      0.738    0.737      0.462    0.738     0.759

=== Confusion Matrix ===

  a  b   <-- classified as
 12  6 |  a = N
  5 19 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.300    0.385    0.375      0.300    0.333      -0.088   0.431     0.439     N
               0.615    0.700    0.533      0.615    0.571      -0.088   0.431     0.559     Y
Weighted Avg.  0.478    0.563    0.464      0.478    0.468      -0.088   0.431     0.506

=== Confusion Matrix ===

 a b   <-- classified as
 3 7 | a = N
 5 8 | b = Y
```

Log x 0

Dataset.5 NativeBayes

Trian

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.444    0.250    0.571      0.444    0.500      0.204    0.701     0.624     N
                0.750    0.556    0.643      0.750    0.692      0.204    0.701     0.794     Y
Weighted Avg.   0.619    0.425    0.612      0.619    0.610      0.204    0.701     0.721

=== Confusion Matrix ===

  a  b   <-- classified as
  8 10 |  a = N
  6 18 |  b = Y
```

Log x 0

Test

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

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.400    0.615    0.333      0.400    0.364      -0.214   0.415     0.394     N
               0.385    0.600    0.455      0.385    0.417      -0.214   0.415     0.619     Y
eighted Avg.   0.391    0.607    0.402      0.391    0.394      -0.214   0.415     0.521

== Confusion Matrix ===

a b   <-- classified as
4 6 | a = N
8 5 | b = Y
```

Log x 0

Dataset.5 IBK(K=10)

Train

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.500    0.208    0.643      0.500    0.563      0.306    0.689     0.601     N
                0.792    0.500    0.679      0.792    0.731      0.306    0.689     0.715     Y
Weighted Avg.   0.667    0.375    0.663      0.667    0.659      0.306    0.689     0.666

=== Confusion Matrix ===

  a  b   <-- classified as
  9  9 |  a = N
  5 19 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.300    0.538    0.300      0.300    0.300      -0.238   0.358     0.368     N
                0.462    0.700    0.462      0.462    0.462      -0.238   0.358     0.535     Y
Weighted Avg.   0.391    0.630    0.391      0.391    0.391      -0.238   0.358     0.463

=== Confusion Matrix ===

 a b   <-- classified as
 3 7 | a = N
 7 6 | b = Y
```

Log x 0

Dataset.5 Bagging-RandomForest

Train

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.889    0.125    0.842      0.889    0.865      0.760    0.968     0.966     N
                0.875    0.111    0.913      0.875    0.894      0.760    0.968     0.975     Y
Weighted Avg.   0.881    0.117    0.883      0.881    0.881      0.760    0.968     0.971

=== Confusion Matrix ===

  a  b   <-- classified as
 16  2 |  a = N
  3 21 |  b = Y
```

Log x 0

Test

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

                TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                0.300    0.538    0.300      0.300    0.300      -0.238   0.323     0.396     N
                0.462    0.700    0.462      0.462    0.462      -0.238   0.323     0.474     Y
Weighted Avg.   0.391    0.630    0.391      0.391    0.391      -0.238   0.323     0.440

=== Confusion Matrix ===

 a b   <-- classified as
 3 7 | a = N
 7 6 | b = Y
```

Log x 0

Dataset.5 J48

Train

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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.429     N
                 1.000    1.000    0.571      1.000    0.727      ?        0.500     0.571     Y
Weighted Avg.    0.571    0.571    ?          0.571    ?          ?        0.500     0.510

=== Confusion Matrix ===

  a  b   <-- classified as
  0 18 |  a = N
  0 24 |  b = Y
```

Log x 0

Test

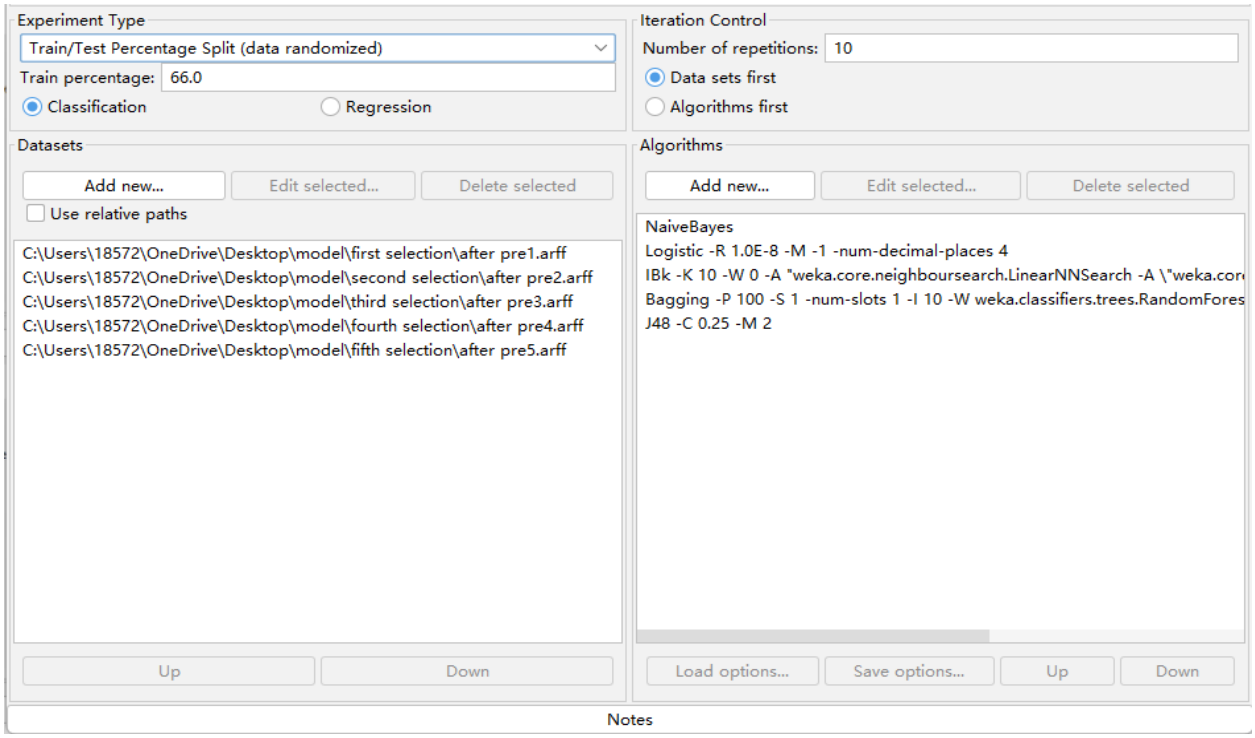
```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.435     N
                 1.000    1.000    0.565      1.000    0.722      ?        0.500     0.565     Y
Weighted Avg.    0.565    0.565    ?          0.565    ?          ?        0.500     0.509

=== Confusion Matrix ===

  a  b   <-- classified as
  0 10 |  a = N
  0 13 |  b = Y
```

Log x 0

4.3 Compare the 25 classification models

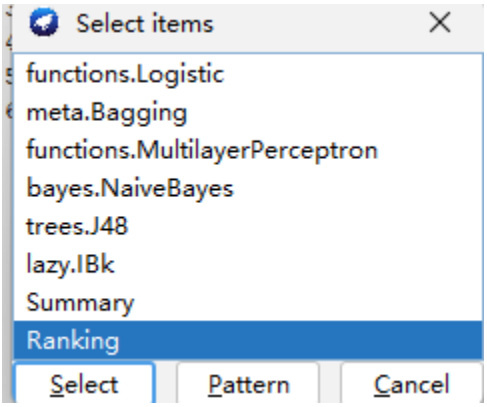
We used the weka-Experimenter to make a comparison of the above 25 models.

Firstly, we uploaded the five Dataset(1,2,3,4,5) and chose the five algorithms we have selected.



Then, we selected “run” and clicked “start” when we finished choosing Analyze.
In Analyze, click “experiment”.
Choose dataset in rows, Scheme in cols.

Select Ranking in Test base.



In Comparison field，choose Percent_correct (The comparison measurement)

Click perform test.

```
>-<   >   < Resultset
  1   2   1 functions.Logistic
  0   1   1 meta.Bagging
  0   0   0 lazy.IBk
  0   0   0 bayes.NaiveBayes
 -1   0   1 trees.J48
```

We also used other measures to compare these 25 models. The comparison results are as follows:

Mean_absolute_error

```
>-<   >   < Resultset
  4   4   0 lazy.IBk
  4   4   0 bayes.NaiveBayes
  1   1   0 trees.J48
 -1   3   4 functions.Logistic
 -8   1   9 meta.Bagging
```

F-measure

```
>-<   >   < Resultset
  2   2   0 meta.Bagging
  0   0   0 trees.J48
  0   0   0 lazy.IBk
 -1   0   1 functions.Logistic
 -1   0   1 bayes.NaiveBayes
```

Area_under_ROC

```
>-<   >   < Resultset
  5   6   1 meta.Bagging
  0   0   0 lazy.IBk
 -1   0   1 trees.J48
 -1   2   3 functions.Logistic
 -3   0   3 bayes.NaiveBayes
```

Matthews_correlation

```
>-<   >   < Resultset
  2   2   0 meta.Bagging
  1   2   1 functions.Logistic
  0   0   0 lazy.IBk
 -1   0   1 trees.J48
 -2   0   2 bayes.NaiveBayes
```

The results of the comparison of the models are given below:

	Percent_correct	Mean_absolute_error	F-measure	Area_under_ROC	Matthews_correlation
winner	Logistic	IBK	Bagging	Bagging	Bagging
Win times	1	4	2	5	2

According to the results. Bagging algorithms win 3 times (the largest number), so we selected the Bagging(RandomForest) algorithm as the best algorithm.

After selecting the best algorithm, we went back to “setup”, Deleted other algorithms except the Bagging algorithm. Then we went back to Analyze and chose Scheme in rows, Dataset in cols.

The comparison of the five datasets using the Bagging algorithm are as follows:

Use Percent_correct to compare

```
Dataset                 (1) 'after_s | (2) 'afte (3) 'afte (4) 'afte (5) 'bore
-------------------------------------------------------------------------------
meta.Bagging            (10)   68.18 |   53.10     65.80     70.50     55.43
-------------------------------------------------------------------------------
                             (v/ /*) |   (0/1/0)   (0/1/0)   (0/1/0)   (0/1/0)
```

Use Mean_absolute_error

```
Dataset                  (1) 'after_ | (2) 'aft (3) 'aft (4) 'aft (5) 'bor
-------------------------------------------------------------------------
meta.Bagging             (10)   0.37 |   0.50 v   0.40 v   0.40     0.48 v
-------------------------------------------------------------------------
                             (v/ /*) |  (1/0/0)  (1/0/0)  (0/1/0)  (1/0/0)
```

Use F-measure

```
Dataset                  (1) 'after_ | (2) 'aft (3) 'aft (4) 'aft (5) 'bor
-------------------------------------------------------------------------
meta.Bagging             (10)   0.61 |   0.41     0.55     0.63     0.43
-------------------------------------------------------------------------
                             (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)  (0/1/0)
```

Area_under_ROC

```
Dataset                  (1) 'after_ | (2) 'aft (3) 'aft (4) 'aft (5) 'bor
-------------------------------------------------------------------------
meta.Bagging             (10)   0.75 |   0.46     0.73     0.75     0.53 *
-------------------------------------------------------------------------
                             (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)  (0/0/1)
```

Matthews_correlation

```
Dataset                  (1) 'after_ | (2) 'aft (3) 'aft (4) 'aft (5) 'bor
-------------------------------------------------------------------------
meta.Bagging             (10)   0.35 |   0.04     0.29     0.40     0.08
-------------------------------------------------------------------------
                             (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)  (0/1/0)
```

According to the results, “v” and “*” infer that the comparison has statistical significance. Therefore, when we use Mean_absolute_error. It has higher statistical significance than others.

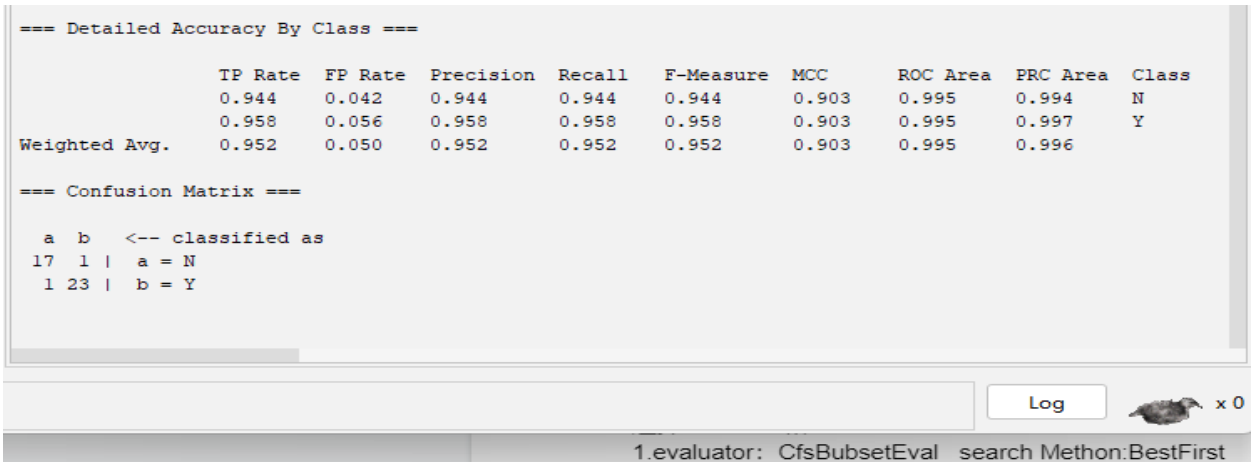
We used Mean_absolute_error as our measure to choose the best model. The lower the Mean_absolute_error is, the better is the model.

Therefore, model Dataset.1 using Bagging(RandomForest) algorithm appears to be the best model.

4.4.

In the final stage of this data analysis process, we compared the performance of our best model with the performance of the model that was built using the same classification algorithm from the dataset with all attributes. The results are as follows:

Best model

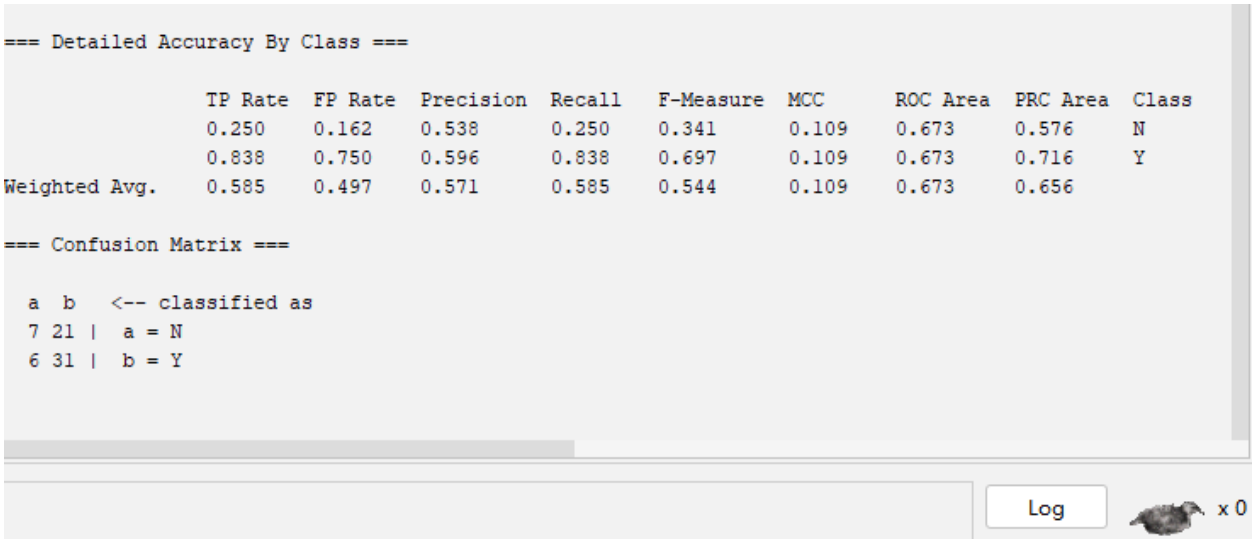


```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.944    0.042    0.944      0.944    0.944      0.903    0.995     0.994     N
                 0.958    0.056    0.958      0.958    0.958      0.903    0.995     0.997     Y
Weighted Avg.    0.952    0.050    0.952      0.952    0.952      0.903    0.995     0.996

=== Confusion Matrix ===

  a  b   <-- classified as
 17  1 |  a = N
  1 23 |  b = Y
```

Log x 0

1.evaluator: CfsBubsetEval search Methon:BestFirst

The model that was built using the same classification algorithm from the dataset with all attributes



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

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.250    0.162    0.538      0.250    0.341      0.109    0.673     0.576     N
                 0.838    0.750    0.596      0.838    0.697      0.109    0.673     0.716     Y
Weighted Avg.    0.585    0.497    0.571      0.585    0.544      0.109    0.673     0.656

=== Confusion Matrix ===

  a  b   <-- classified as
  7 21 |  a = N
  6 31 |  b = Y
```

Log x 0

The comparison results are gathered and showed as following:

	TPR	FPR	Precision	Recall	F-Measure	MCC	ROC	PRC
Best	0.952	0.050	0.952	0.952	0.952	0.903	0.995	0.994
All attributes	0.585	0.497	0.571	0.585	0.544	0.109	0.673	0.656

(All are weighted Avg.)

Obviously, the index has increased significantly which indicates that the model is much better after filtering the attributes.

5. Summary of Findings

According to the best model we choose. We find that the latitude and moisture are the most important factors existing in the formation of forest fires.

Therefore, we made a comparison towards the results of the best model and visualized the results using R. The data visualization result is shown below:

```
> y
                      X.burned
moisture                N  Y
  'Mesic to subhygric'  1  1
  'Subxeric to mesic'  11  7
  Mesic                 3  8
  Subhygric             4  1
  Subxeric              8 14
  Xeric                 1  6
>
```

Obviously, compared to unburned forests, the burned forests are much drier. Therefore, the forest with lower moisture degree has higher potential for the occurrence of forest fires.

Besides the moisture degree, latitude has great influence on the formation of forest fires too. However, because of the narrow fluctuation range, we could not determine the accurate influence which latitude has on the formation of forest fires. We could only lead to the conclusion that the occurrence potential of forest fires differ in different latitudes. One of the possible reasons is that different latitude contributes to the formation of different natural environments.

6. Potential Implications and Improvements

6.1 Practical Applications

Moisture is one of the most essential factors which has a great impact on the formation of forest fires. Therefore, the government should put more manpower and material resources into the prevention of forest fires in the forests with drier environments. Besides that, the government should try to increase the degree of moisture in the forests, especially the ones suffering higher potential of forest fires.

We should also pay more attention to the latitude factor when analyzing the natural environment of the forests. Latitude appears to be one of the most important factors affecting the formation of forest fires. Although the fluctuation range is narrow, different latitudes could lead to obvious differences in the occurrence potential of forest fires.

The government should also attach importance to the forest in obviously higher or lower latitude. This could lead to great changes in the possibility of forest fires as well as the difficulty to prevent, control and fight the forest fires.

6.2 Potential Improvements

This report only discusses the natural factors influencing forest fires. However, human factors could have greatly affected the formation of forest fires too. Therefore, we could consider more comprehensive factors when discussing data mining goals and gathering initial dataset next time. We could also make analyses of the influences the combination of several existing factors could have on the occurrence of our mining goal.

We only selected five classification algorithms to build the models. Next time, we could include more methods in the models building process to get more accurate dataset analysis results.

Reference: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1740

Appendix: Data preprocessing is detailed explained in the supplementary appendix document.