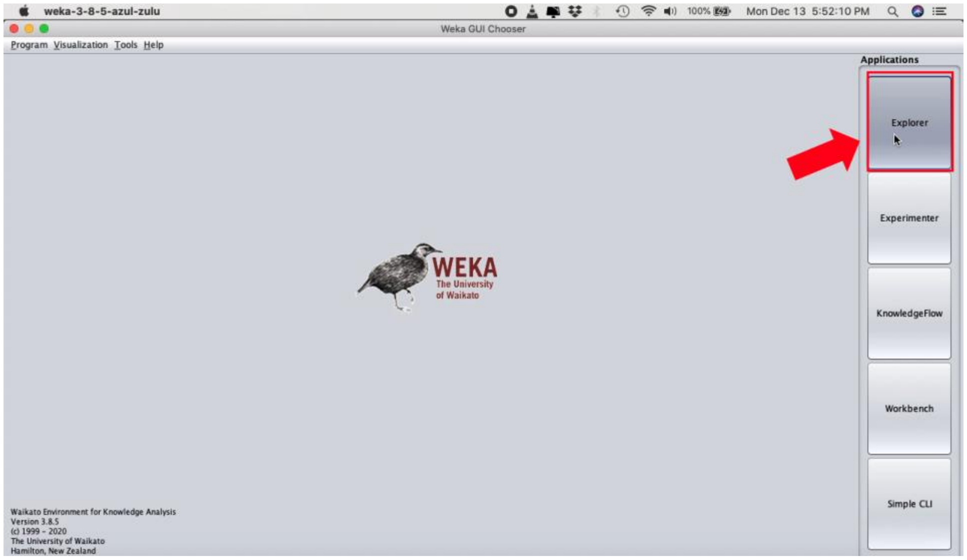
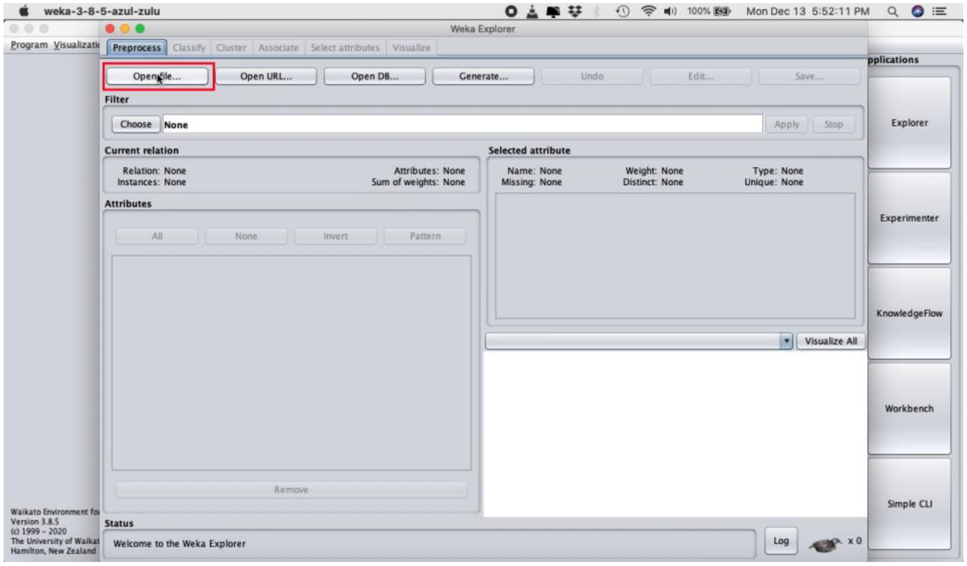
Data Processing using K-Nearest Neighbor (KNN) in WEKA

In this part, Waikato Environment for Knowledge Analysis (WEKA) v.3.8.5 is used to create a KNN model. First, we start by choosing the Explorer menu.



Waikato Environment for Knowledge Analysis
Version 3.8.5
(c) 1999 - 2020
The University of Waikato
Hamilton, New Zealand

Next we will have Weka Explorer window as shown as below. Choose Open file to pick the dataset that we are going to use.



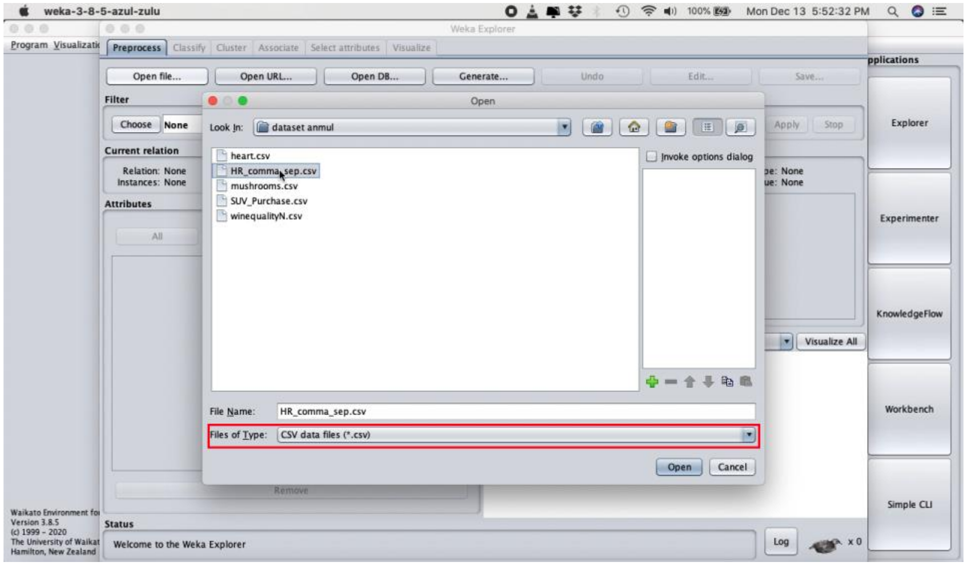
Filter: None

Relation: None
Instances: None
Attributes: None
Sum of weights: None

Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None

Welcome to the Weka Explorer

Change the type of files from ARFF into CSV, which is the format of the dataset we use.



Filter: None

Relation: None
Instances: None

Look In: dataset anmul

- heart.csv
- HR_comma_sep.csv
- mushrooms.csv
- SUV_Purchase.csv
- winequalityN.csv

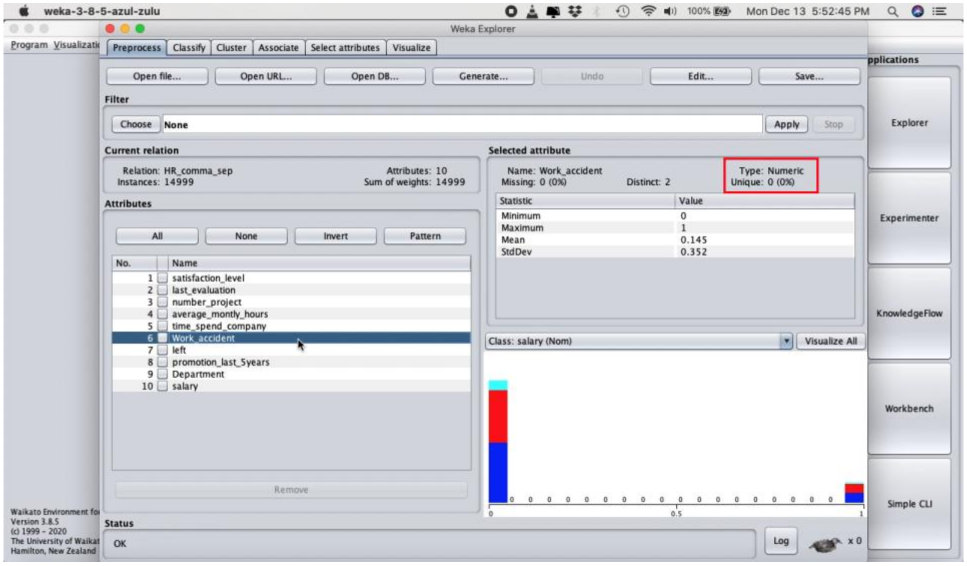
File Name: HR_comma_sep.csv

Files of Type: CSV data files (*.csv)

Waikato Environment fo
Version 3.8.5
(c) 1999 - 2020
The University of Waikat
Hamilton, New Zealand

Status: Welcome to the Weka Explorer

Now the window will show the data with its ten attributes.



Filter: None

Relation: HR_comma_sep
Instances: 14999
Attributes: 10
Sum of weights: 14999

No.	Name
1	satisfaction_level
2	last_evaluation
3	number_project
4	average_montly_hours
5	time_spend_company
6	Work_accident
7	left
8	promotion_last_5years
9	Department
10	salary

Name: Work_accident
Missing: 0 (0%)
Distinct: 2
Type: Numeric
Unique: 0 (0%)

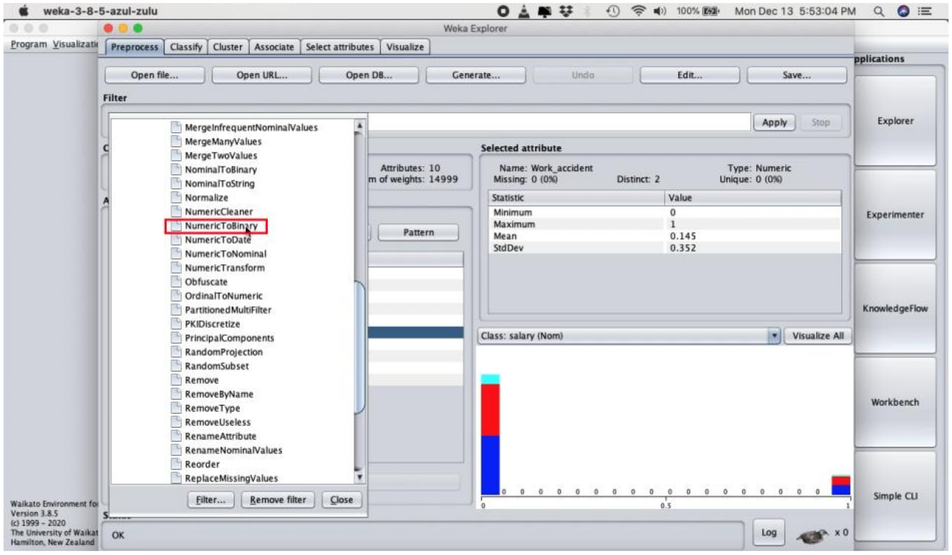
Statistic	Value
Minimum	0
Maximum	1
Mean	0.145
StdDev	0.352

Class: salary (Nom)

Waikato Environment fo
Version 3.8.5
(c) 1999 - 2020
The University of Waikat
Hamilton, New Zealand

Status: OK

The program automatically detects categorical variable as numerical variable since it uses 0-1 scale. To change the attribute of these variables into binary, we use a NumericToBinary filter.



Attributes: 10
Sum of weights: 14999

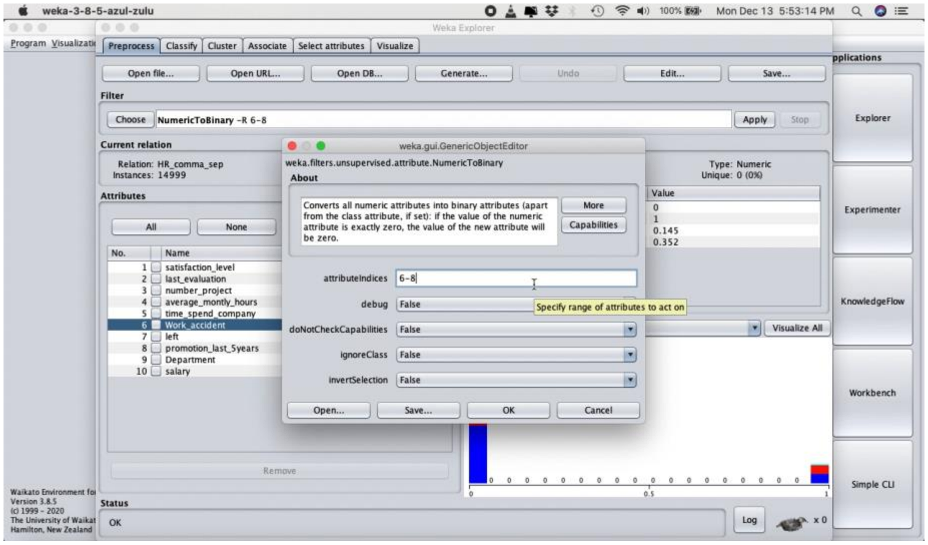
Selected attribute
Name: Work_accident
Type: Numeric
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.145
StdDev	0.352

Class: salary (Nom)

Status
OK

Put the number of the variables we are going to change. In this case *work_accident*, *left*, dan *promotion_last_5years* are marked as the 6th, 7th, and 8th attribute, so we put the number 6-8. Lastly, choose Apply in the Filter tab.



NumericToBinary -R 6-8

Current relation
Relation: HR_comma_sep
Instances: 14999

Attributes

No.	Name
1	satisfaction_level
2	last_evaluation
3	number_project
4	average_montly_hours
5	time_spend_company
6	Work_accident
7	left
8	promotion_last_5years
9	Department
10	salary

weka.filters.unsupervised.attribute.NumericToBinary

Converts all numeric attributes into binary attributes (apart from the class attribute, if set): if the value of the numeric attribute is exactly zero, the value of the new attribute will be zero.

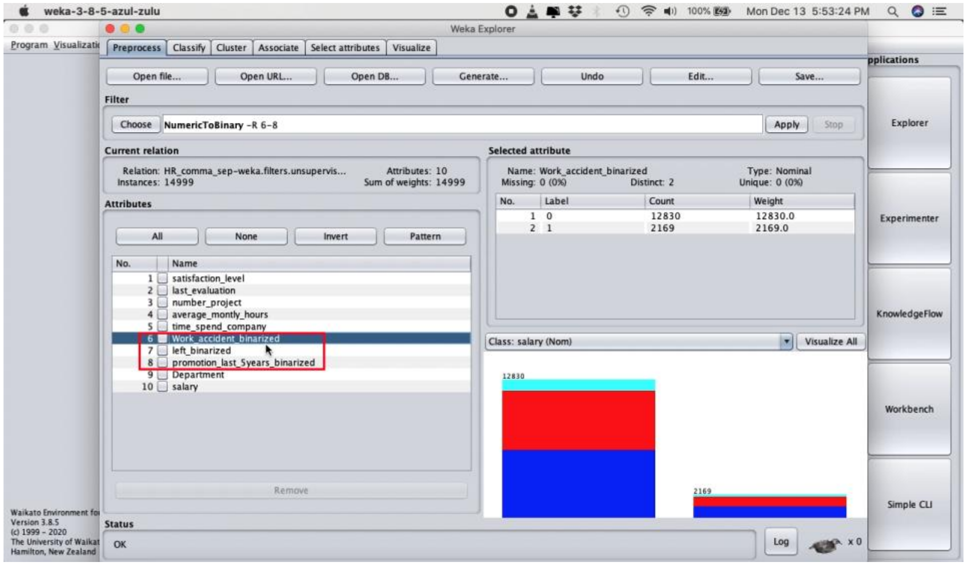
attributeIndices: 6-8
debug: False
doNotCheckCapabilities: False
ignoreClass: False
invertSelection: False

Type: Numeric
Unique: 0 (0%)

Value
0
1
0.145
0.352

Status
OK

The variables *work_accident*, *left*, and *promotion_last_5years* now have binarized labels behind their names.



Filter: NumericToBinary -R 6-8

Current relation
Relation: HR_comma_sep-weka.filters.unsupervis... Attributes: 10
Instances: 14999 Sum of weights: 14999

Selected attribute
Name: Work_accident_binarized Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count	Weight
1	0	12830	12830.0
2	1	2169	2169.0

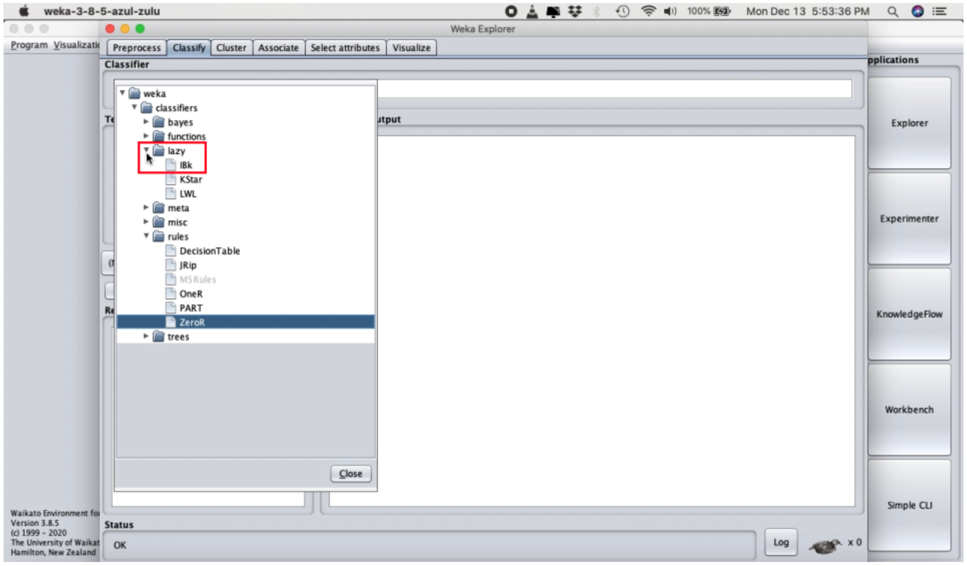
Attributes

No.	Name
1	satisfaction_level
2	last_evaluation
3	number_project
4	average_montly_hours
5	time_spend_company
6	Work_accident_binarized
7	left_binarized
8	promotion_last_5years_binarized
9	Department
10	salary

Class: salary (Nom)

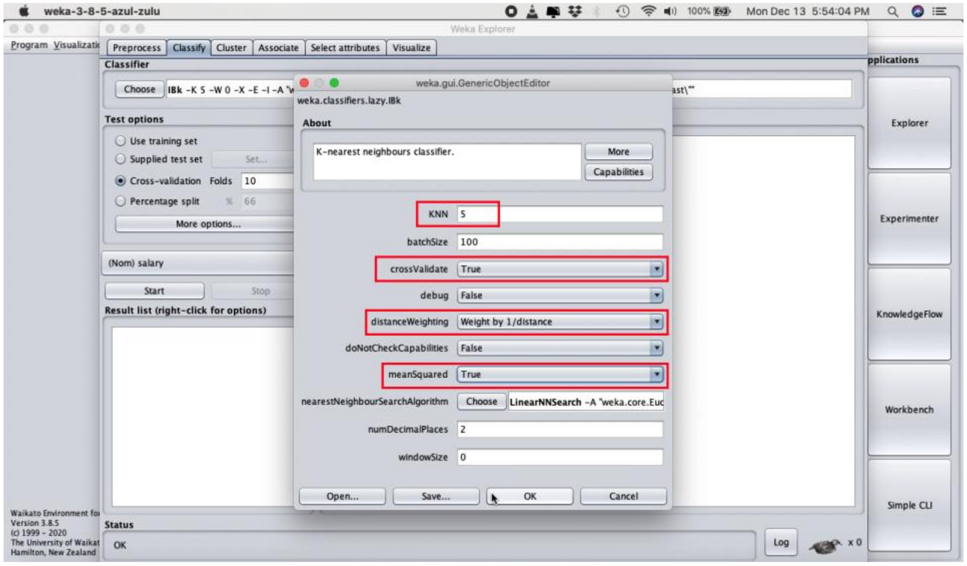
Status
OK

Now we are ready to build a KNN model. In Classify tab, choose Classifier menu and pick IBk (*Instance Based Learner*), which is another name of KNN model in WEKA.



Status
OK

In this study we set the number of *k* to 5, set crossValidate to *True*, and picked weighted voting for distanceWeighting calculation.



Classifier: IBk -K 5 -W 0 -X -E -I -A "w… ast\""

Test options

- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) salary

weka.gui.GenericObjectEditor

weka.classifiers.lazy.IBk

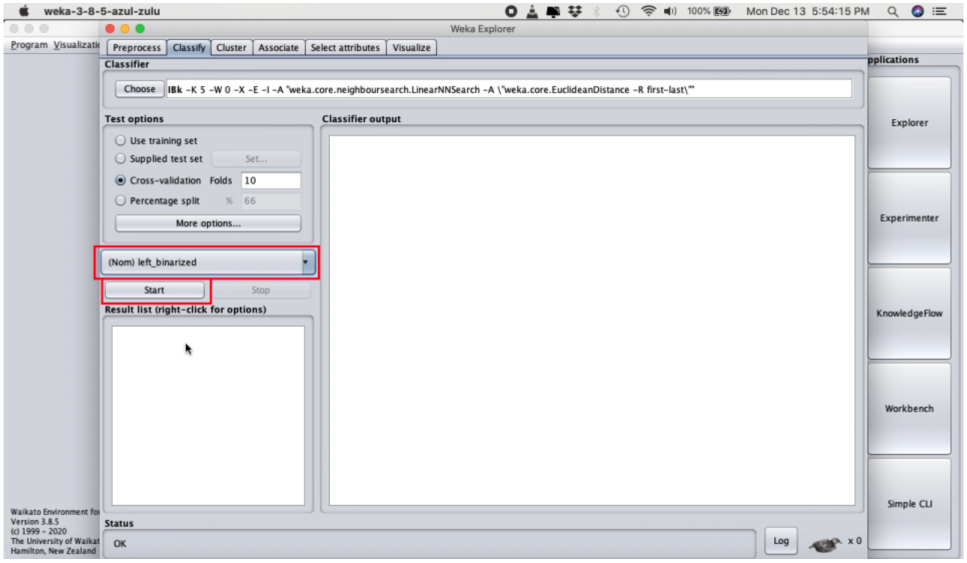
About

K-nearest neighbours classifier.

KNN	5
batchSize	100
crossValidate	True
debug	False
distanceWeighting	Weight by 1/distance
doNotCheckCapabilities	False
meanSquared	True
nearestNeighbourSearchAlgorithm	LinearNNSearch -A "weka.core.Euc
numDecimalPlaces	2
windowSize	0

Status: OK

Next we determine the response variable. In this case the response variable is *left* which determines whether an employee would leave the company or not. Then we can start the model building by choosing Start.



Classifier: IBk -K 5 -W 0 -X -E -I -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

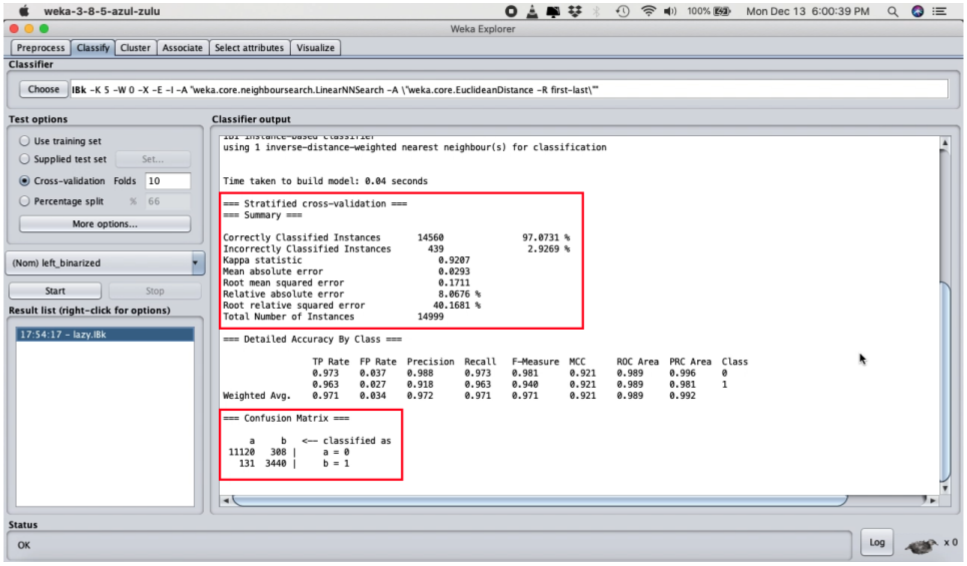
Test options

- ○ Use training set
- ○ Supplied test set
- ● Cross-validation Folds 10
- ○ Percentage split % 66

(Nom) left_binarized

Status: OK

The picture below shows the output given after the model is done.



Classifier: IBk -K 5 -W 0 -X -E -I -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) left_binarized

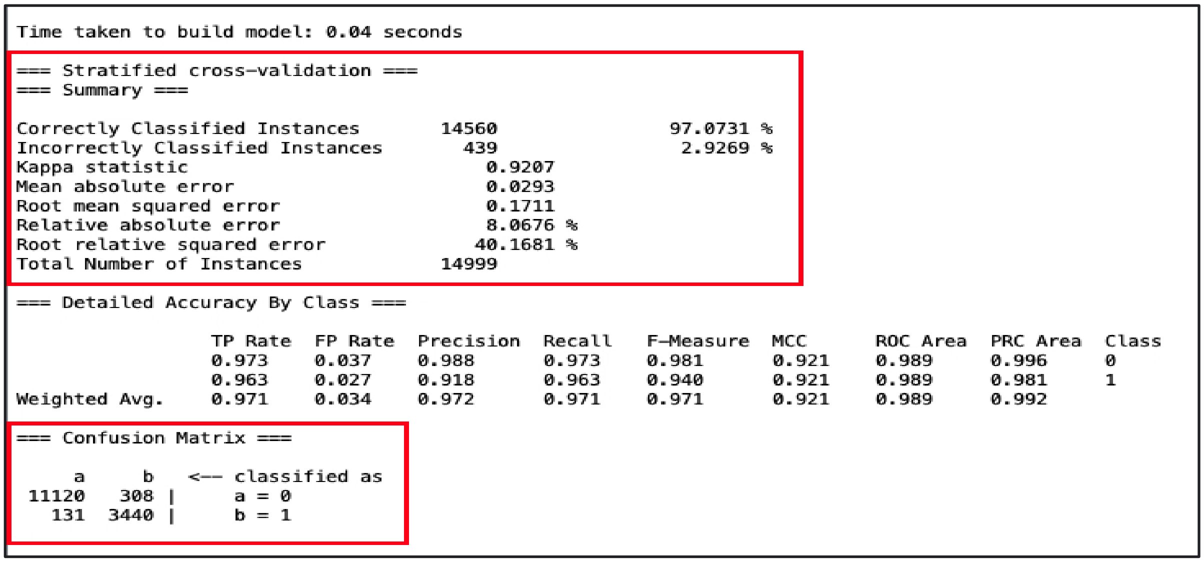
Result list (right-click for options): 17:54:17 - lazy.IBk

Status: OK

```
IB1 instance-based classifier
using 1 inverse-distance-weighted nearest neighbour(s) for classification


Time taken to build model: 0.04 seconds

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

Correctly Classified Instances       14560               97.0731 %
Incorrectly Classified Instances       439                2.9269 %
Kappa statistic                          0.9207
Mean absolute error                      0.0293
Root mean squared error                  0.1711
Relative absolute error                  8.0676 %
Root relative squared error             40.1681 %
Total Number of Instances            14999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.037    0.988      0.973    0.981      0.921    0.989     0.996     0
                 0.963    0.027    0.918      0.963    0.940      0.921    0.989     0.981     1
Weighted Avg.    0.971    0.034    0.972      0.971    0.971      0.921    0.989     0.992

=== Confusion Matrix ===

     a     b   <-- classified as
 11120   308 |     a = 0
   131  3440 |     b = 1
```

Result

From the figure it can be seen that the model successfully predicts employee turnover correctly with 97.07% accuracy. In the confusion matrix, it can be seen that the true positive, false positive, true negative, and false negative values are 3440, 308, 11120, and 131 respectively. The Kappa statistical value is a measure that compares the accuracy of observations with the expected accuracy that is formulated randomly. Landis & Koch (1977) said that the Kappa value in the range of 0.81-1 shows an almost perfect match between expectations and observations, meaning that the model is very good at predicting employee turnover.



```
Time taken to build model: 0.04 seconds

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

Correctly Classified Instances       14560               97.0731 %
Incorrectly Classified Instances       439                2.9269 %
Kappa statistic                          0.9207
Mean absolute error                      0.0293
Root mean squared error                  0.1711
Relative absolute error                  8.0676 %
Root relative squared error             40.1681 %
Total Number of Instances            14999

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.973    0.037    0.988      0.973    0.981      0.921    0.989     0.996     0
                 0.963    0.027    0.918      0.963    0.940      0.921    0.989     0.981     1
Weighted Avg.    0.971    0.034    0.972      0.971    0.971      0.921    0.989     0.992

=== Confusion Matrix ===

     a     b   <-- classified as
 11120   308 |     a = 0
   131  3440 |     b = 1
```