جامعة بيرزيت

BIRZEIT UNIVERSITY

FACULTY OF ENGINEERING AND TECHNOLOGY
DEPARTMENT OF COMPUTER ENGINEERING

Artificial Intelligence

ENCS3340

Project 2 Report
Machine Learning for Classification

Prepared by:

Tariq Odeh -1190699

Yousef Hammad - 1170625

Sec: 1

Instructor: Dr. Adnan Yahya

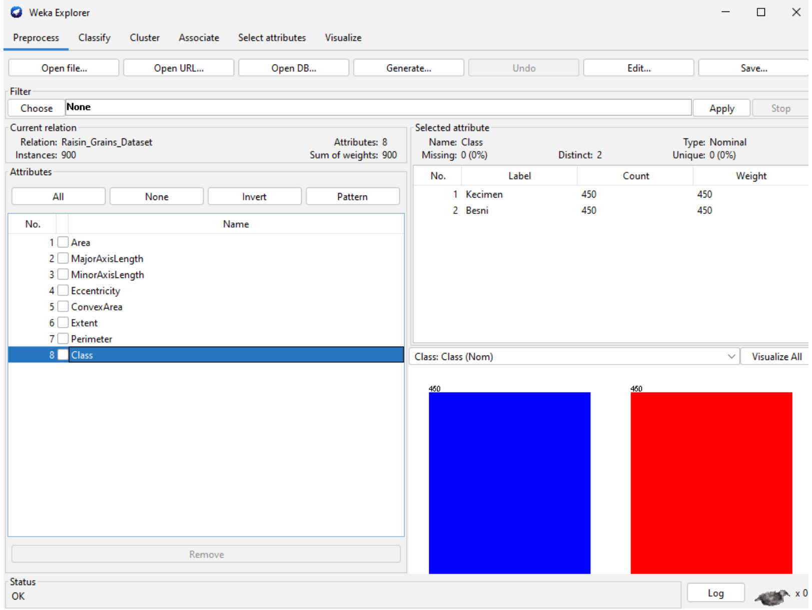
Date: 11th June 2022

Abstract

We'll learn how to assess alternative categorization methods using machine learning tools in this assignment. Additionally, we must compare multiple machine learning algorithms for a classification problem using WEKA, and the dataset we will work on is 2 according to the university id (1170625) with three models Decision.

1. Dataset

From the figure we can see the name of relationship Raisin_Grains _Datasetis, the instances (900) , and there are 8 attributes in the dataset, also in panel below current relation shows the name of attributes (Area, Extent, Perimenter, ...). And in the right panel, the selected attribute statistics are displayed. It exhibits the following: Name, Type, Missing, Unique and Distinct. When select class attributevwe can see that the name is Class, type is normal, missing is zero, unique is zero and the distinct are two distinct values (Kecimen and Besni). For the attribute the nominal type means it accepts no numeric values and the numeric values means it accepts numeric values. For each of the 900 instances, the count of each particular class label is provided in the count column. The output class label for the attribute will be displayed in the histogram.The class label for this dataset is either good or terrible. There are 450 Kecimen examples and 450 Besni examples.



Filter: None

Current relation
Relation: Raisin_Grains_Dataset
Instances: 900
Attributes: 8
Sum of weights: 900

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

No.	Label	Count	Weight
1	Kecimen	450	450
2	Besni	450	450

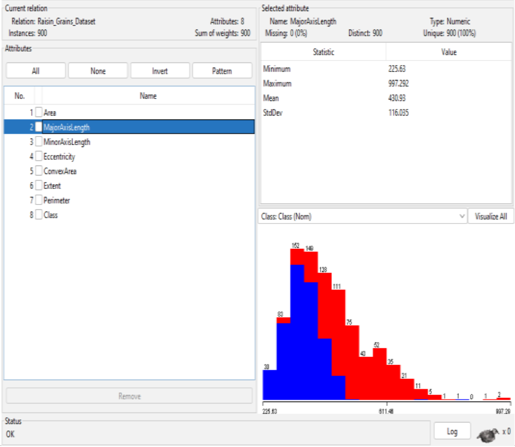
Attributes

No.	Name
1	□ Area
2	□ MajorAxisLength
3	□ MinorAxisLength
4	□ Eccentricity
5	□ ConvexArea
6	□ Extent
7	□ Perimeter
8	□ Class

Class: Class (Nom)

Status
OK

In this figure shown the name of the attribute. The type of the attribute is numeric, thus it's a number. It has 900 different values in 900 different circumstances. It's unique since there are 900 different values that don't match. Minimum monetary value for this parameter is 225. maximum value for this parameter is 997. The mean is calculated by dividing the total number of values by the number of cases. Attribute duration standard deviation and Histogram are used to calculate deviations from the mean.



Current relation
Relation: Raisin_Grains_Dataset
Instances: 900
Attributes: 8
Sum of weights: 900

Selected attribute
Name: MajorAxisLength
Type: Numeric
Missing: 0 (0%)
Distinct: 900
Unique: 900 (100%)

Statistic	Value
Minimum	225.63
Maximum	997.292
Mean	430.93
StdDev	116.035

Attributes

No.	Name
1	□ Area
2	□ MajorAxisLength
3	□ MinorAxisLength
4	□ Eccentricity
5	□ ConvexArea
6	□ Extent
7	□ Perimeter
8	□ Class

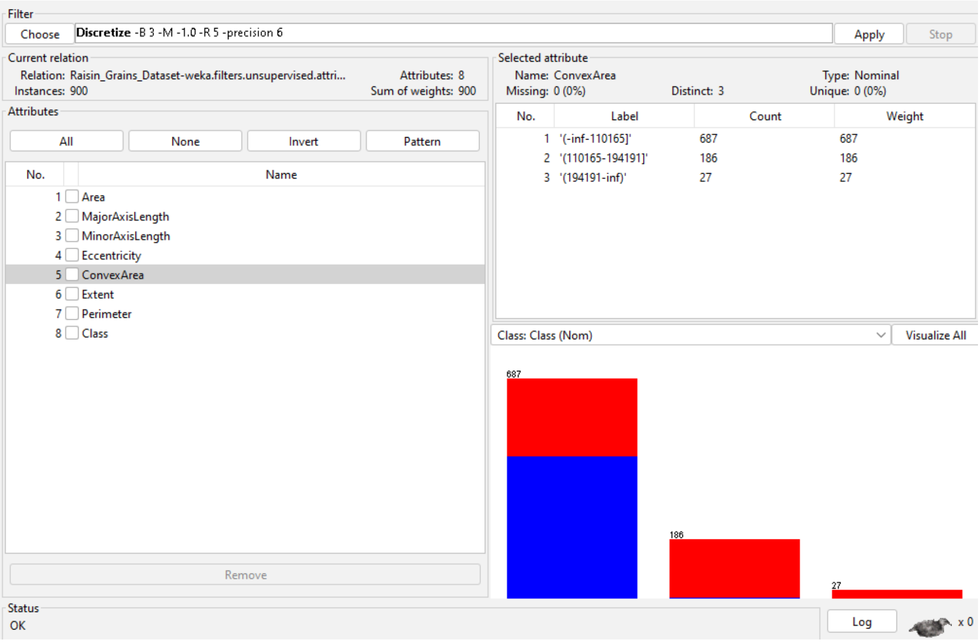
Class: Class (Nom)

Status
OK

2. Decision Tree

Filter

Preprocess for one attribute (ConvexArea) discretization of continuous attributes with 3 pins.



Filter: Discretize -B 3 -M -1.0 -R 5 -precision 6

Current relation
Relation: Raisin_Grains_Dataset-weka.filters.unsupervised.attri... Attributes: 8
Instances: 900 Sum of weights: 900

Selected attribute
Name: ConvexArea Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-110165]'	687	687
2	'(110165-194191]'	186	186
3	'(194191-inf)'	27	27

Attributes

No.	Name
1	Area
2	MajorAxisLength
3	MinorAxisLength
4	Eccentricity
5	ConvexArea
6	Extent
7	Perimeter
8	Class

Class: Class (Nom)

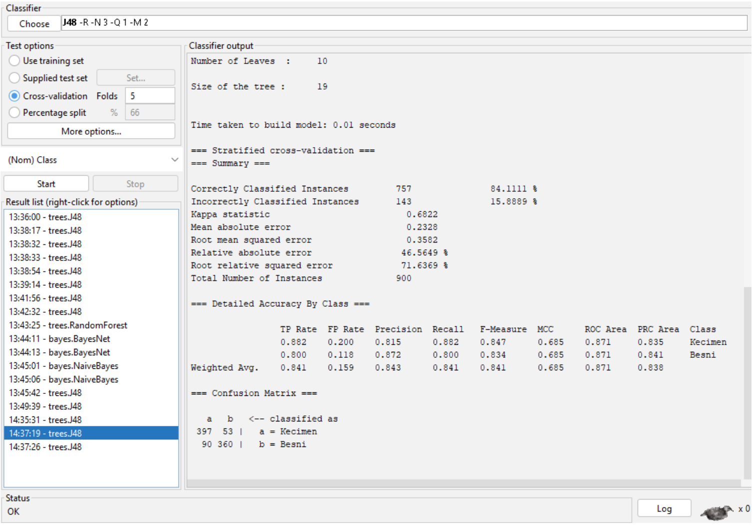
Status
OK

Classifer

in the confusion matrix shown below, it can be seen that the true kecimen have 397 that were classefied correctly and 53 that were classefied as Besni , the true Besni have 360 classefied correctly and 90 that were classefied as keciman.

It can be noticed that the correctly classefied have an accuracy of 84.11% and incorrectly classified have an accuracy of 15.89% .

In the Detailed Accuracy by class it shows the TP rate, FP Rate, Percision, F-Measure,MCC, ROC Area, PRC Area and the class.



Classifier: J48 -R -N 3 -Q 1 -M 2

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 5
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)
13:36:00 - trees.J48
13:38:17 - trees.J48
13:38:32 - trees.J48
13:38:33 - trees.J48
13:38:54 - trees.J48
13:39:14 - trees.J48
13:41:56 - trees.J48
13:42:32 - trees.J48
13:43:25 - trees.RandomForest
13:44:11 - bayes.BayesNet
13:44:13 - bayes.BayesNet
13:45:01 - bayes.NaiveBayes
13:45:06 - bayes.NaiveBayes
13:45:42 - trees.J48
13:49:39 - trees.J48
14:35:31 - trees.J48
14:37:19 - trees.J48
14:37:26 - trees.J48

Classifier output

```
Number of Leaves  :     10

Size of the tree :      19


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         757               84.1111 %
Incorrectly Classified Instances       143               15.8889 %
Kappa statistic                          0.6822
Mean absolute error                      0.2328
Root mean squared error                  0.3582
Relative absolute error                 46.5649 %
Root relative squared error             71.6369 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.882    0.200    0.815      0.882    0.847      0.685    0.871     0.835     Kecimen
                 0.800    0.118    0.872      0.800    0.834      0.685    0.871     0.841     Besni
Weighted Avg.    0.841    0.159    0.843      0.841    0.841      0.685    0.871     0.838

=== Confusion Matrix ===

   a   b   <-- classified as
 397  53 |   a = Kecimen
  90 360 |   b = Besni
```

Status
OK

As an example Tp rate of the Kecimen class will be calculated by dividing the correctly classefied Kecimans with the total true Kecimens $\frac{397}{450} = 0.882$.

FP rate will be calculated by dividing incorrectly classefied Keciman (Kecimans classefied as Besni) on the total Kekiman values $\frac{53}{450} = 0.2$.

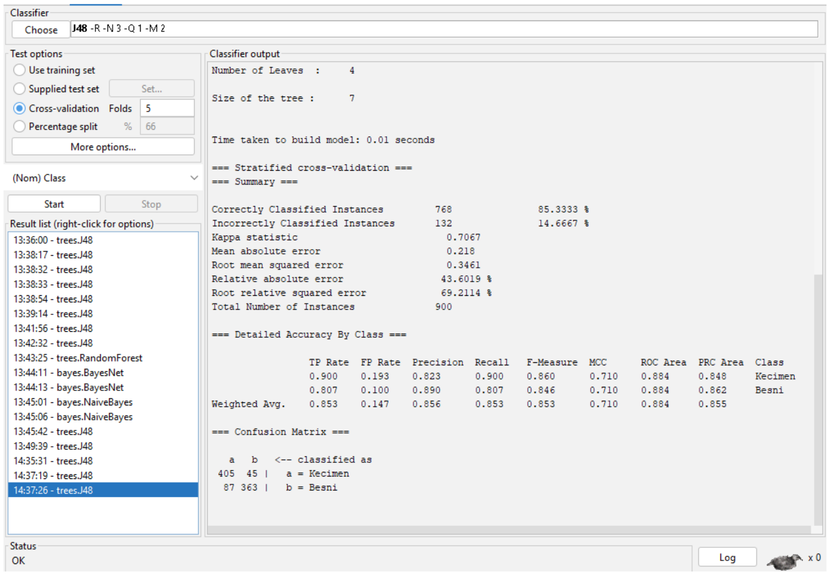
Precision ca be calculated by deviding TP over FP + TP $\frac{0.882}{0.2+0.882} = 0.815$.

Recall should be equal to the TP.

F-Measure can be calculated by $\frac{2* \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2* 0.815 * 0.882}{0.815+ 0.882} = 0.847$

Hyper-parameter

When the hyper parameter of the model is changed to reduced error pruning as shown in figure below it can be noted that the changes in the values displayed on the screen such as correctly classified instances became higher and equal 85.33% and also the incorrectly classified instances changed and it became less and equal 14.66%. In addition, the effects are obvious for the confusion Matrix, TP Rate, FP Rate, Precision, Recall and F- Measure.



Classifier

Choose J48 -R -N 3 -Q 1 -M 2

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 5
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

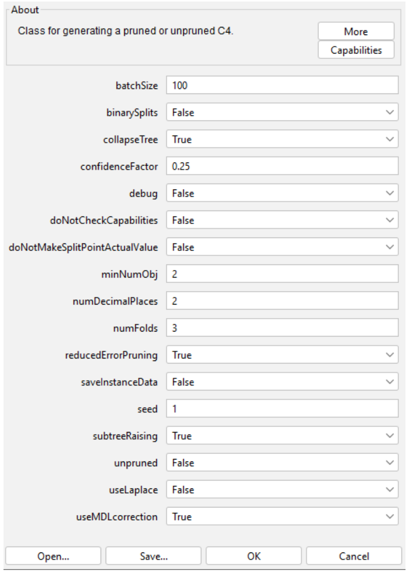
13:36:00 - trees.J48
13:38:17 - trees.J48
13:38:32 - trees.J48
13:38:33 - trees.J48
13:38:54 - trees.J48
13:39:14 - trees.J48
13:41:56 - trees.J48
13:42:32 - trees.J48
13:43:25 - trees.RandomForest
13:44:11 - bayes.BayesNet
13:44:13 - bayes.BayesNet
13:45:01 - bayes.NaiveBayes
13:45:06 - bayes.NaiveBayes
13:45:42 - trees.J48
13:49:39 - trees.J48
14:35:31 - trees.J48
14:37:19 - trees.J48
14:37:26 - trees.J48

Classifier output

```
Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         768               85.3333 %
Incorrectly Classified Instances       132               14.6667 %
Kappa statistic                          0.7067
Mean absolute error                      0.218
Root mean squared error                  0.3461
Relative absolute error                 43.6019 %
Root relative squared error             69.2114 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.900    0.193    0.823      0.900    0.860      0.710    0.884     0.848     Kecimen
                 0.807    0.100    0.890      0.807    0.846      0.710    0.884     0.862     Besni
Weighted Avg.    0.853    0.147    0.856      0.853    0.853      0.710    0.884     0.855

=== Confusion Matrix ===

   a   b   <-- classified as
 405  45 |   a = Kecimen
  87 363 |   b = Besni
```

Status

OK



About

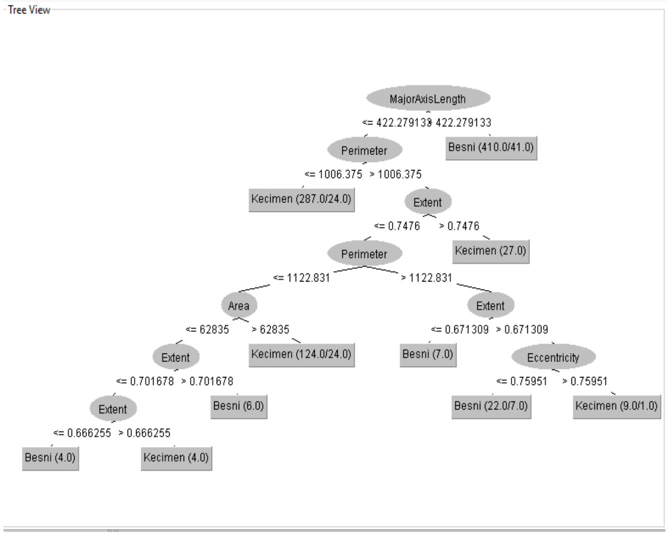
Class for generating a pruned or unpruned C4.

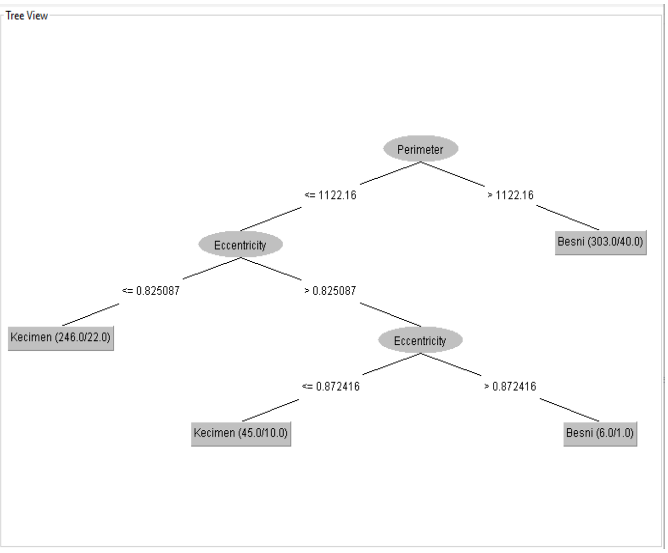
batchSize 100
binarySplits False
collapseTree True
confidenceFactor 0.25
debug False
doNotCheckCapabilities False
doNotMakeSplitPointActualValue False
minNumObj 2
numDecimalPlaces 2
numFolds 3
reducedErrorPruning True
saveInstanceData False
seed 1
subtreeRaising True
unpruned False
useLaplace False
useMDLcorrection True

Visualize the trees

Default Tree

With reduced error pruning set to true

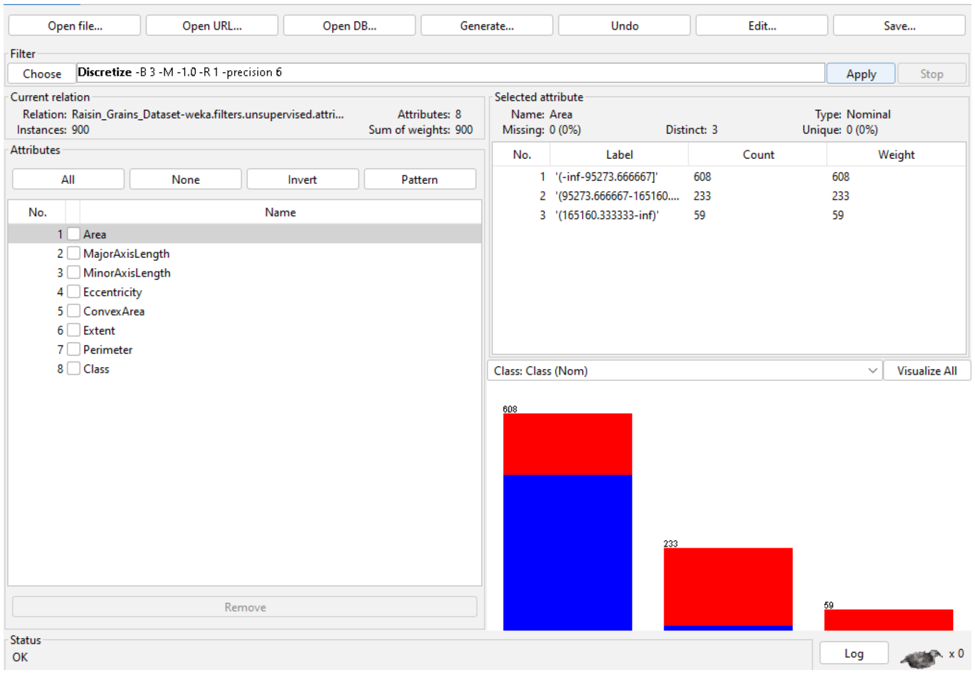




3. Naïve Bayes

Filter

Preprocess for one attribute (Area) discretization of continuous attributes with 3 pins.



Filter: Discretize -B 3 -M -1.0 -R 1 -precision 6

Current relation
Relation: Raisin_Grains_Dataset-weka.filters.unsupervised.attri... Attributes: 8
Instances: 900 Sum of weights: 900

Selected attribute
Name: Area Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-95273.666667]'	608	608
2	'(95273.666667-165160....	233	233
3	'(165160.333333-inf)'	59	59

Attributes

No.	Name
1	□ Area
2	□ MajorAxisLength
3	□ MinorAxisLength
4	□ Eccentricity
5	□ ConvexArea
6	□ Extent
7	□ Perimeter
8	□ Class

Class: Class (Nom)

Status
OK

Classifer

in the confusion matrix shown below, it can be seen that the true kecimen have 426 that were classefied correctly and 24 that were classefied as Besni , the true Besni have 323 classefied correctly and 127 that were classefied as keciman.

It can be noticed that the correctly classefied have an accuracy of 83.22% and incorrectly classified have an accuracy of 16.78% .

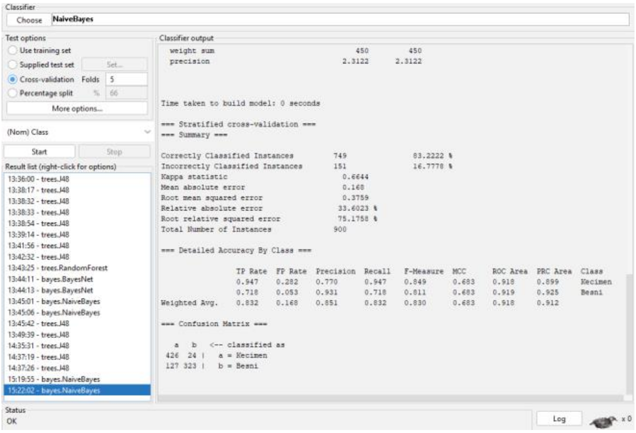
In the Detailed Accuracy by class it shows the TP rate, FP Rate, Percision, F-Measure,MCC, ROC Area, PRC Area and the class.

As an example Tp rate of the Kecimen class will be calculated by dividing the correctly classefied Kecimans with the total true Kecimens $\frac{426}{450} =$ 0.947 .

FP rate will be calculated by dividing incorrectly classefied Keciman (Kecimans classefied as Besni) on the total Kekiman values $\frac{24}{450} = 0.282$.

Precision ca be calculated by deviding TP over FP + TP $\frac{0.947}{0.282+0.947} = 0.77$.

Recall should be equal to the TP.



Classifier: NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 5
○ Percentage split % 66

(Nom) Class

13:36:00 - trees.J48
13:38:17 - trees.J48
13:38:32 - trees.J48
13:38:33 - trees.J48
13:38:54 - trees.J48
13:39:14 - trees.J48
13:41:56 - trees.J48
13:42:32 - trees.J48
13:43:25 - trees.RandomForest
13:44:11 - bayes.BayesNet
13:44:13 - bayes.BayesNet
13:45:01 - bayes.NaiveBayes
13:45:06 - bayes.NaiveBayes
13:45:42 - trees.J48
13:49:39 - trees.J48
14:35:31 - trees.J48
14:37:19 - trees.J48
14:37:26 - trees.J48
15:19:55 - bayes.NaiveBayes
15:22:02 - bayes.NaiveBayes

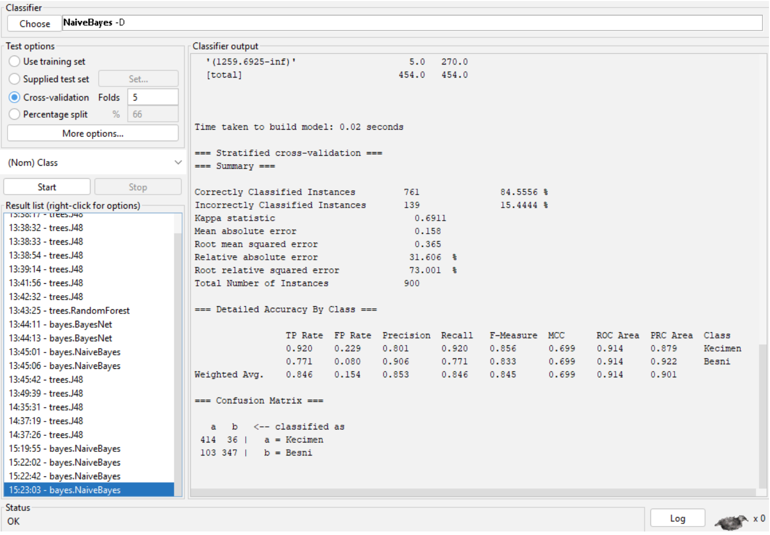
```
  weight sum                             450        450
  precision                           2.3122     2.3122



Time taken to build model: 0 seconds

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

Correctly Classified Instances         749               83.2222 %
Incorrectly Classified Instances       151               16.7778 %
Kappa statistic                          0.6644
Mean absolute error                      0.168
Root mean squared error                  0.3759
Relative absolute error                 33.6023 %
Root relative squared error             75.1758 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.947    0.282    0.770      0.947    0.849      0.683    0.918     0.899     Kecimen
                 0.718    0.053    0.931      0.718    0.811      0.683    0.919     0.925     Besni
Weighted Avg.    0.832    0.168    0.851      0.832    0.830      0.683    0.918     0.912

=== Confusion Matrix ===

   a   b   <-- classified as
 426  24 |   a = Kecimen
 127 323 |   b = Besni
```

Status
OK

F-Measure can be calculated by $\frac{2* \text{Precision} * \text{Recall}}{\text{Precision}+ \text{Recall}} = \frac{2* 0.77 * 0.947}{0.77+ 0.947} = 0.849$

Hyper-parameter

When the hyper parameter of the model is changed to supervised discretization as shown in figure below it can be noted that the changes in the values displayed on the screen such as correctly classified instances became higher and equal 84.56% and also the incorrectly classified instances changed and it became less and equal 15.44%. In addition, the effects are obvious for the confusion Matrix, TP Rate, FP Rate, Precision, Recall and F- Measure.



Classifier: NaiveBayes -D

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 5 ○ Percentage split % 66

(Nom) Class

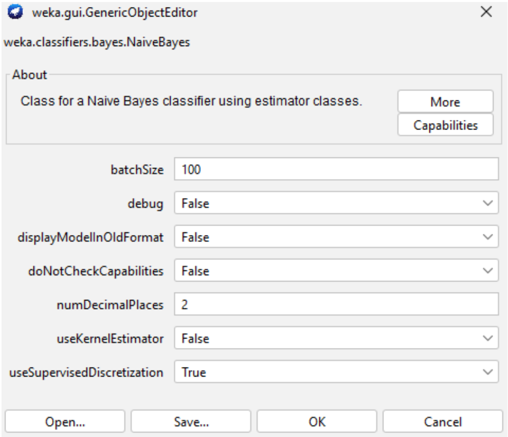
Result list (right-click for options)

```
13:38:32 - trees.J48
13:38:33 - trees.J48
13:38:54 - trees.J48
13:39:14 - trees.J48
13:41:56 - trees.J48
13:42:32 - trees.J48
13:43:25 - trees.RandomForest
13:44:11 - bayes.BayesNet
13:44:13 - bayes.BayesNet
13:45:01 - bayes.NaiveBayes
13:45:06 - bayes.NaiveBayes
13:45:42 - trees.J48
13:49:39 - trees.J48
14:35:31 - trees.J48
14:37:19 - trees.J48
14:37:26 - trees.J48
15:19:55 - bayes.NaiveBayes
15:22:02 - bayes.NaiveBayes
15:22:42 - bayes.NaiveBayes
15:23:03 - bayes.NaiveBayes
```

Classifier output

```
  '(1259.6925-inf)'                    5.0   270.0
  [total]                            454.0   454.0



Time taken to build model: 0.02 seconds

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

Correctly Classified Instances         761               84.5556 %
Incorrectly Classified Instances       139               15.4444 %
Kappa statistic                          0.6911
Mean absolute error                      0.158
Root mean squared error                  0.365
Relative absolute error                 31.606  %
Root relative squared error             73.001  %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.920    0.229    0.801      0.920    0.856      0.699    0.914     0.879     Kecimen
                 0.771    0.080    0.906      0.771    0.833      0.699    0.914     0.922     Besni
Weighted Avg.    0.846    0.154    0.853      0.846    0.845      0.699    0.914     0.901

=== Confusion Matrix ===

   a   b   <-- classified as
 414  36 |   a = Kecimen
 103 347 |   b = Besni
```

Status: OK



weka.gui.GenericObjectEditor

weka.classifiers.bayes.NaiveBayes

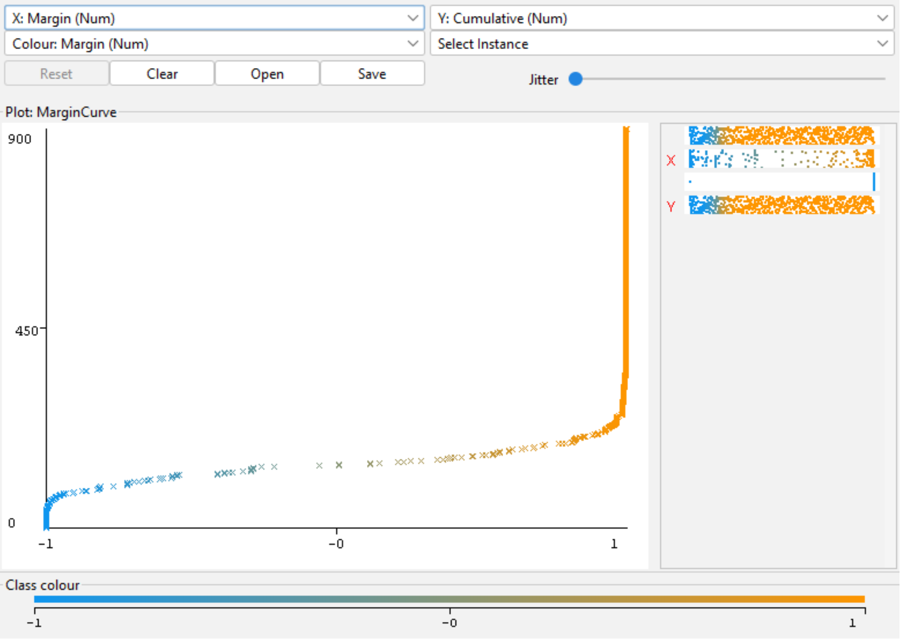
About

Class for a Naive Bayes classifier using estimator classes.

batchSize	100
debug	False
displayModelInOldFormat	False
doNotCheckCapabilities	False
numDecimalPlaces	2
useKernelEstimator	False
useSupervisedDiscretization	True

Visualize the curve

Use Supervised discretization



X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

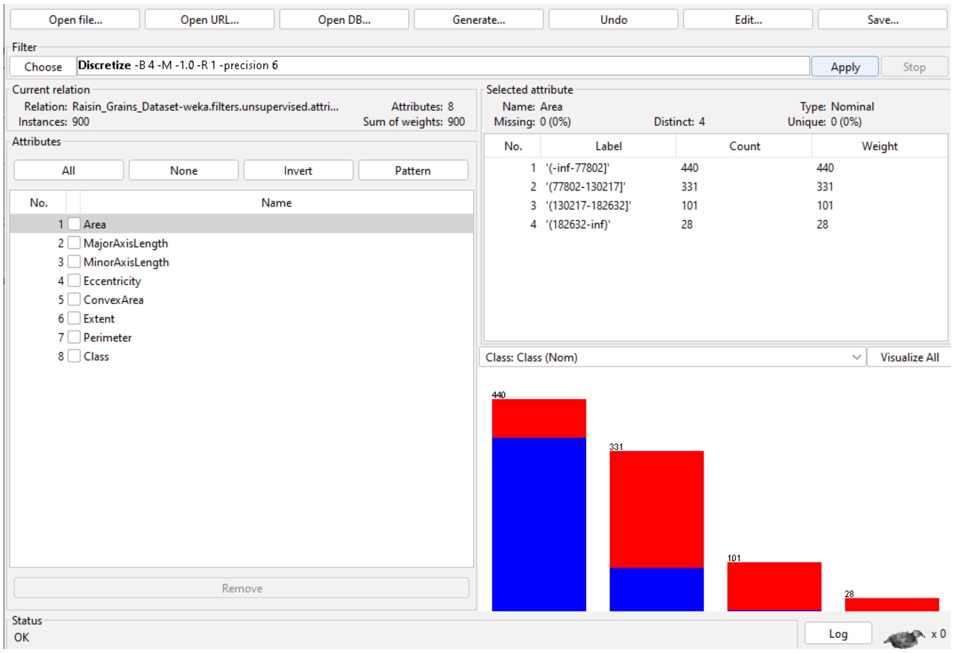
Select Instance

Plot: MarginCurve

4. Lazy IBK

Filter

Preprocess for one attribute (Area) discretization of continuous attributes with 4 pins.



Filter: Discretize -B 4 -M -1.0 -R 1 -precision 6

Current relation
Relation: Raisin_Grains_Dataset-weka.filters.unsupervised.attri... Attributes: 8
Instances: 900 Sum of weights: 900

Attributes

No.		Name
1	□	Area
2	□	MajorAxisLength
3	□	MinorAxisLength
4	□	Eccentricity
5	□	ConvexArea
6	□	Extent
7	□	Perimeter
8	□	Class

Selected attribute
Name: Area Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-77802]'	440	440
2	'(77802-130217]'	331	331
3	'(130217-182632]'	101	101
4	'(182632-inf)'	28	28

Status
OK

Classifer

In the confusion matrix shown below, it can be seen that the true kecimen have 368 that were classefied correctly and 82 that were classefied as Besni , the true Besni have 366 classefied correctly and 84 that were classefied as keciman.

It can be noticed that the correctly classefied have an accuracy of 81.56% and incorrectly classified have an accuracy of 18.44% .

In the Detailed Accuracy by class it shows the TP rate, FP Rate, Percision, F-Measure,MCC, ROC Area, PRC Area and the class.

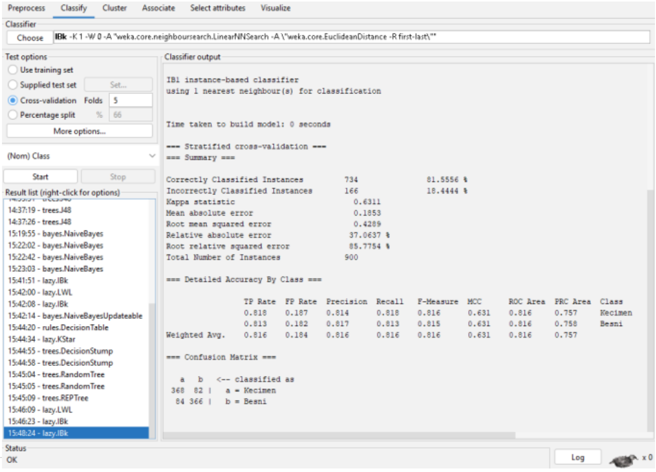
As an example Tp rate of the Kecimen class will be calculated by dividing the correctly classefied Kecimans with the total true Kecimens $\frac{368}{450} = 0.818$.

FP rate will be calculated by dividing incorrectly classefied Keciman (Kecimans classefied as Besni) on the total Kekiman values $\frac{82}{450} = 0.187$.

Precision ca be calculated by deviding TP over FP + TP $\frac{0.818}{0.187+0.818} = 0.814$.

Recall should be equal to the TP.

F-Measure can be calculated by $\frac{2* \text{Precision} * \text{Recall}}{\text{Precision}+ \text{Recall}}$ = $\frac{2* 0.814 * 0.818}{0.814+ 0.818} = 0.816$



Classifier: IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 5
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)
14:37:19 - trees.J48
14:37:26 - trees.J48
15:19:55 - bayes.NaiveBayes
15:22:02 - bayes.NaiveBayes
15:22:42 - bayes.NaiveBayes
15:23:03 - bayes.NaiveBayes
15:41:51 - lazy.IBk
15:42:00 - lazy.LWL
15:42:08 - lazy.IBk
15:42:14 - bayes.NaiveBayesUpdateable
15:44:20 - rules.DecisionTable
15:44:34 - lazy.KStar
15:44:55 - trees.DecisionStump
15:44:58 - trees.DecisionStump
15:45:04 - trees.RandomTree
15:45:05 - trees.RandomTree
15:45:09 - trees.REPTree
15:46:09 - lazy.LWL
15:46:23 - lazy.IBk
15:48:24 - lazy.IBk

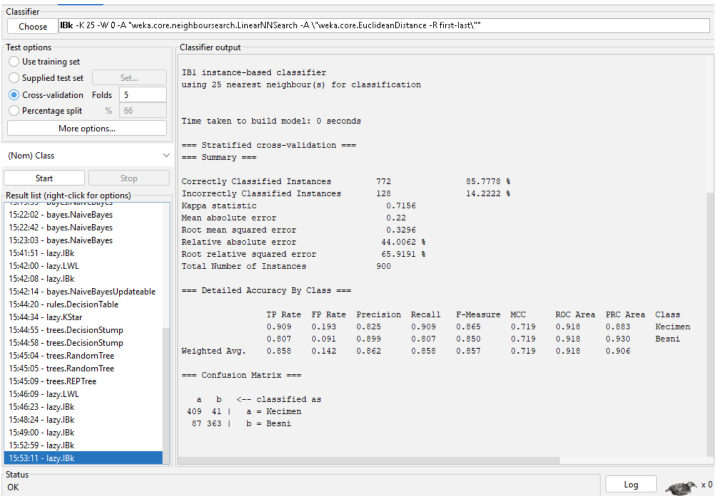
Classifier output

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


Time taken to build model: 0 seconds

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

Correctly Classified Instances         734               81.5556 %
Incorrectly Classified Instances       166               18.4444 %
Kappa statistic                          0.6311
Mean absolute error                      0.1853
Root mean squared error                  0.4289
Relative absolute error                 37.0637 %
Root relative squared error             85.7754 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.818    0.187    0.814      0.818    0.816      0.631    0.816     0.757     Kecimen
                 0.813    0.182    0.817      0.813    0.815      0.631    0.816     0.758     Besni
Weighted Avg.    0.816    0.184    0.816      0.816    0.816      0.631    0.816     0.757

=== Confusion Matrix ===

   a   b   <-- classified as
 368  82 |   a = Kecimen
  84 366 |   b = Besni
```

Status
OK

Hyper-parameter

When the hyper parameter of the model is changed to KNN as shown in figure below it can be noted that the changes in the values displayed on the screen such as correctly classified instances became higher and equal 85.77% and also the incorrectly classified instances changed and it became less and equal 14.22%. In addition, the effects are obvious for the confusion Matrix, TP Rate, FP Rate, Precision, Recall and F-Measure.



Classifier

Choose IBk -K 25 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 5
○ Percentage split % 66

(Nom) Class

Result list (right-click for options)

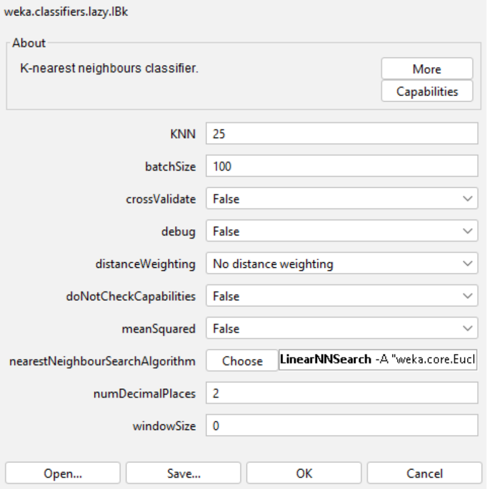
15:22:02 - bayes.NaiveBayes
15:22:42 - bayes.NaiveBayes
15:23:03 - bayes.NaiveBayes
15:41:51 - lazy.IBk
15:42:00 - lazy.LWL
15:42:08 - lazy.IBk
15:42:14 - bayes.NaiveBayesUpdateable
15:44:20 - rules.DecisionTable
15:44:34 - lazy.KStar
15:44:55 - trees.DecisionStump
15:44:58 - trees.DecisionStump
15:45:04 - trees.RandomTree
15:45:05 - trees.RandomTree
15:45:09 - trees.REPTree
15:46:09 - lazy.LWL
15:46:23 - lazy.IBk
15:48:24 - lazy.IBk
15:49:00 - lazy.IBk
15:52:59 - lazy.IBk
15:53:11 - lazy.IBk

Classifier output

```
IB1 instance-based classifier
using 25 nearest neighbour(s) for classification


Time taken to build model: 0 seconds

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

Correctly Classified Instances         772               85.7778 %
Incorrectly Classified Instances       128               14.2222 %
Kappa statistic                          0.7156
Mean absolute error                      0.22
Root mean squared error                  0.3296
Relative absolute error                 44.0062 %
Root relative squared error             65.9191 %
Total Number of Instances              900

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.909    0.193    0.825      0.909    0.865      0.719    0.918     0.883     Kecimen
                 0.807    0.091    0.899      0.807    0.850      0.719    0.918     0.930     Besni
Weighted Avg.    0.858    0.142    0.862      0.858    0.857      0.719    0.918     0.906

=== Confusion Matrix ===

   a   b   <-- classified as
 409  41 |   a = Kecimen
  87 363 |   b = Besni
```

Status

OK



weka.classifiers.lazy.IBk

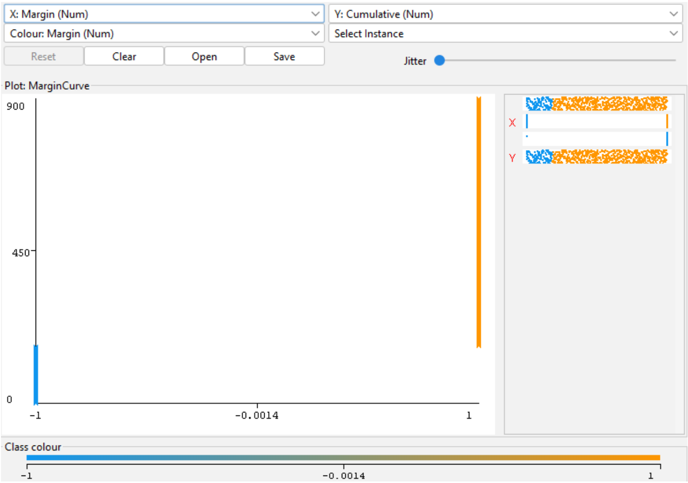
About

K-nearest neighbours classifier.

KNN	25
batchSize	100
crossValidate	False
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	Choose LinearNNSearch -A "weka.core.Eucl
numDecimalPlaces	2
windowSize	0

Visualize the curves

Default



X: Margin (Num)

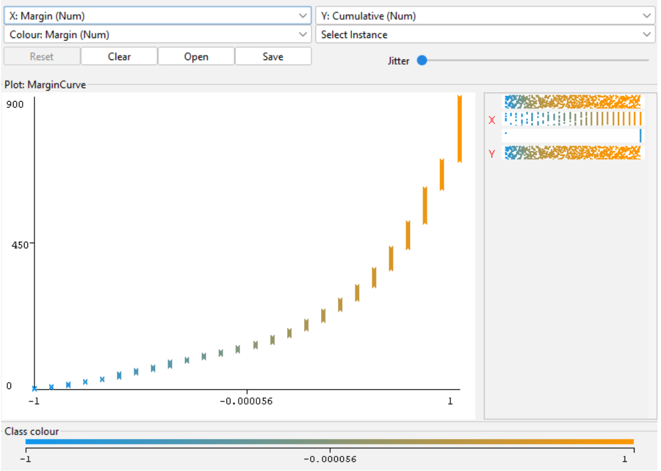
Y: Cumulative (Num)

Colour: Margin (Num)

Select Instance

Plot: MarginCurve

Use KNN



X: Margin (Num)

Y: Cumulative (Num)

Colour: Margin (Num)

Select Instance

Plot: MarginCurve