Using classifiers for mail promotions. Part I. Building response predictor

Lab 2.1

Lab consists of two parts: classification and business analysis

- Part I. Data mining: build the classifier and use it for the prediction of potential responders
- Part II. Business analytics: how to design the most profitable campaign

Plan

Part I. Data Mining. Classification with WEKA.

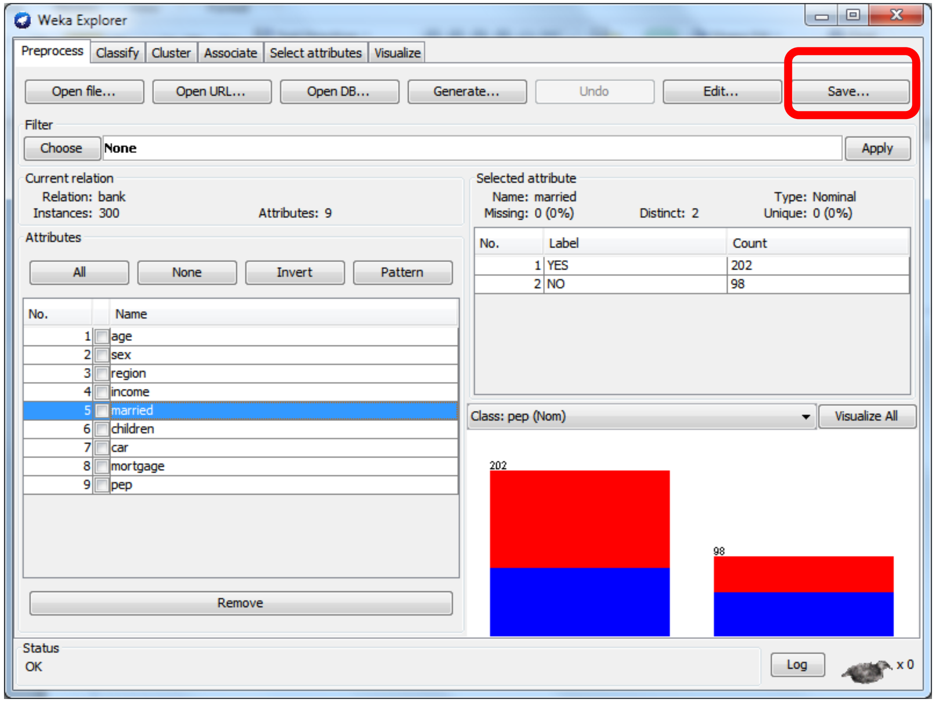
1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Part II. Business analysis

1. Generate Lift chart(s)
2. Cost-benefit analysis
3. Recommendations

Dataset

- Load bank_data.csv into WEKA explorer
- Save file as bank1.arff



Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Dataset: explore available attributes in text editor

- @relation bank-data
- @attribute id {ID12101,ID12102,ID12103,ID12104,ID12105,ID12106,ID12107,ID12108,ID12109,ID12110,ID12111,ID12112,ID12113,ID12114,ID12115,ID12116,ID12117,ID12118,ID12119,ID12120,ID12121,ID12122,ID12123,ID12124,ID12125,ID12126,ID12127,ID12128,ID12129,ID12130,ID12131,ID12132,ID12133,ID12134,ID12135,ID12136,ID12137,ID12138,ID12139,ID12140,ID12141,ID12142,ID12143,ID12144,ID12145,ID12146,ID12147,ID12148,ID12149,ID12150,ID12151,ID12152,ID12153,ID12154,ID12155,ID12156,ID12157,ID12158,ID12159,ID12160,ID12161,ID121
  …
- @attribute age numeric
- @attribute sex {FEMALE,MALE}
- @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
- @attribute income numeric
- @attribute married {NO,YES}
- @attribute children numeric
- @attribute car {NO,YES}
- @attribute save_act {NO,YES}
- @attribute current_act {NO,YES}
- @attribute mortgage {NO,YES}
- @attribute pep {YES,NO}

Class attribute: bought Personal Equity Plan after the last mailing

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Dataset: working with attributes

- @relation bank-data
- @attribute id {ID12101,ID12102,ID12103,ID12104,ID12105,ID12106,ID12107,ID12108,ID12109,ID12110,ID12111,ID12112,ID12113,ID12114,ID12115,ID12116,ID12117,ID12118,ID12119,ID12120,ID12121,ID12122,ID12123,ID12124,ID12125,ID12126,ID12127,ID12128,ID12129,ID12130,ID12131,ID12132,ID12133,ID12134,ID12135,ID12136,ID12137,ID12138,ID12139,ID12140,ID12141,ID12142,ID12143,ID12144,ID12145,ID12146,ID12147,ID12148,ID12149,ID12150,ID12151,ID12152,ID12153,ID12154,ID12155,ID12156,ID12157,ID12158,ID12159,ID12160,ID12161,ID12162,ID12163,ID12164,ID12165,ID12166,ID12167,ID12168,ID12169,ID12170,ID12171,ID12172,ID12173,ID12174,ID12175…
- @attribute age numeric
- @attribute sex {FEMALE,MALE}
- @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
- @attribute income numeric
- @attribute married {NO,YES}
- @attribute children numeric
- @attribute car {NO,YES}
- @attribute save_act {NO,YES}
- @attribute current_act {NO,YES}
- @attribute mortgage {NO,YES}
- @attribute pep {YES,NO}

Non-predictive attribute: remove it and save file

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Dataset: working with attributes

- @relation bank-data

- @attribute age numeric
- @attribute sex {FEMALE,MALE}
- @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
- @attribute income numeric
- @attribute married {NO,YES}
- @attribute children numeric
- @attribute car {NO,YES}
- @attribute save_act {NO,YES}
- @attribute current_act {NO,YES}
- @attribute mortgage {NO,YES}
- @attribute pep {YES,NO}

Numeric attributes age and income: discretize into groups

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Discretize numeric attributes

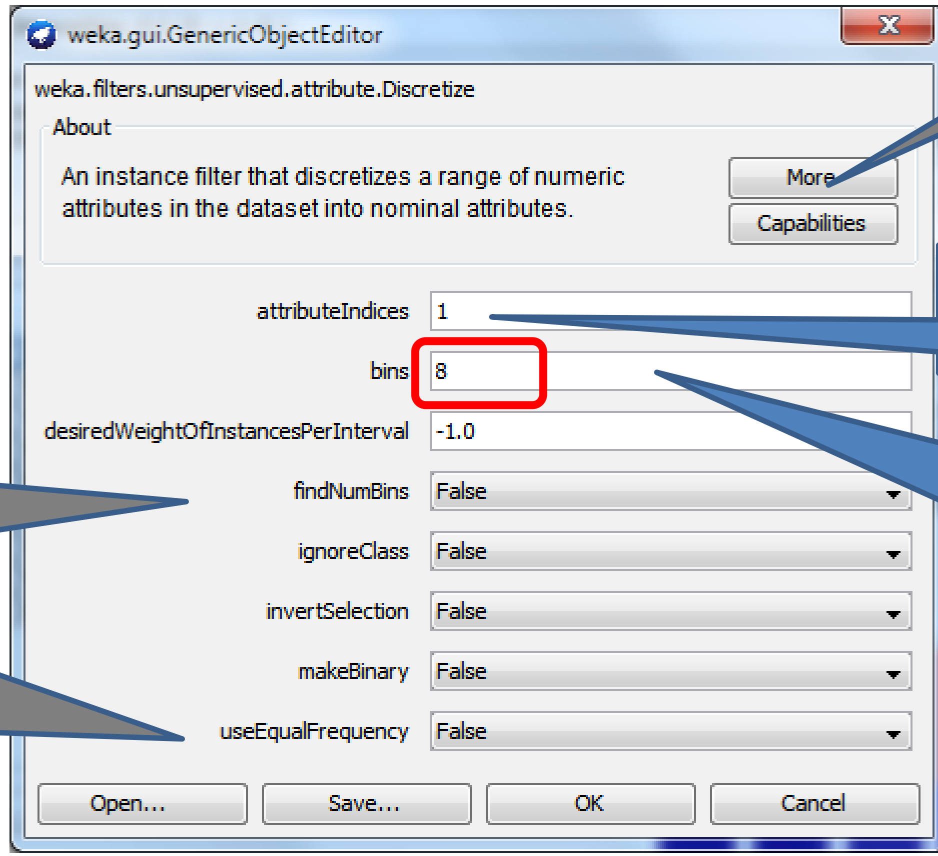
- Simple discretization techniques: distribute numeric values into a predefined number of bins
  - Equal intervals: the bins are defined as equal-size numeric intervals
  - Equal frequency: the bins are defined such as to contain equal number of instances in each interval
- In WEKA: Filter: Choose -> Filters-> Unsupervised -> Attribute-> Discretize.
- Left-click to open parameters window

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Discretize numeric attributes

- Age



Explains parameters

Index of the attribute to apply filter on: 1

Number of bins: based on min-max values and common sense

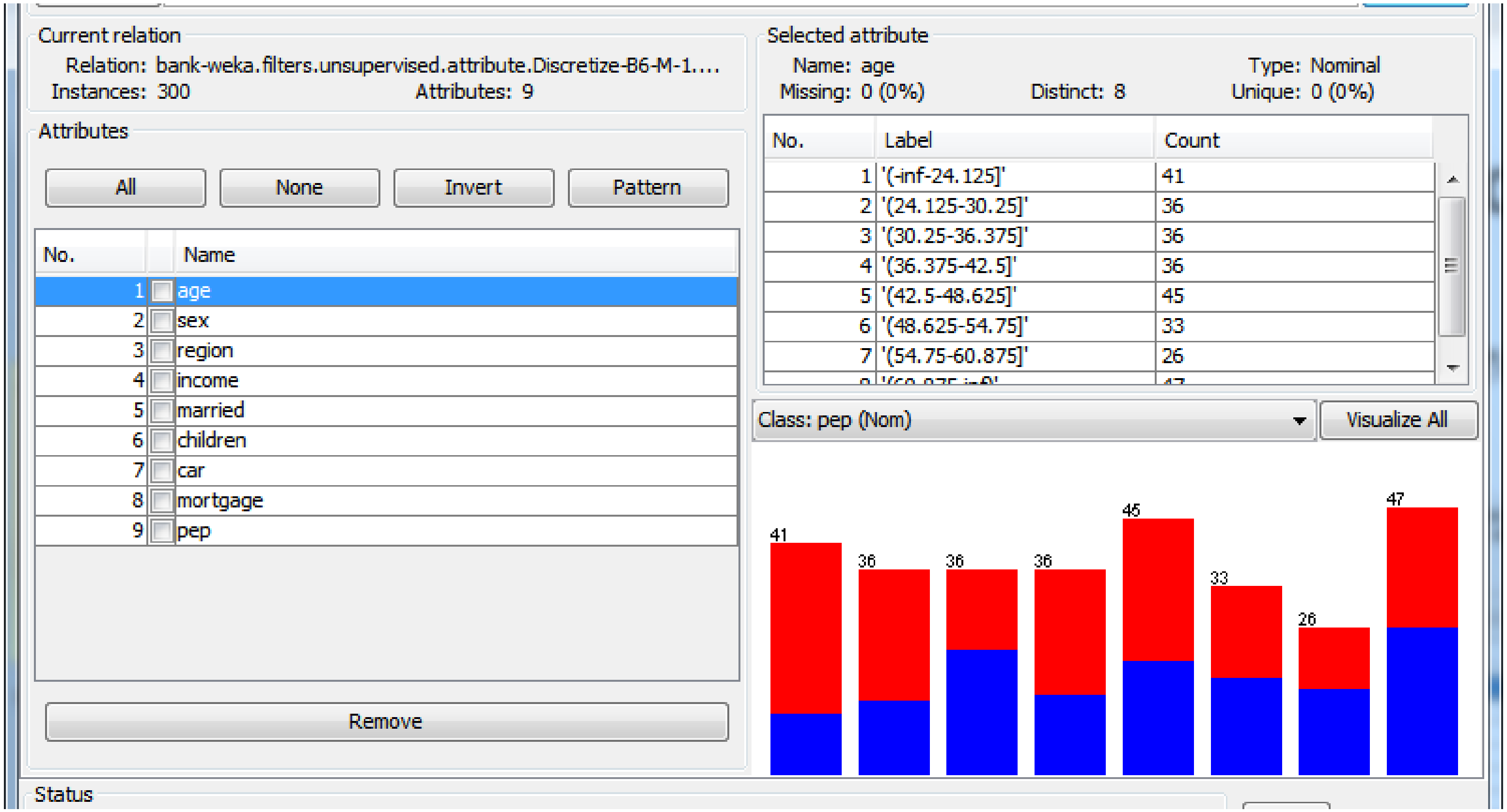
Finds optimal number of bins by data mining techniques

If true - equal frequency binning, if false – equal interval binning

The number of bins is found experimentally, by observing the distribution of the class label in different bins. To play with different settings, use the Undo button

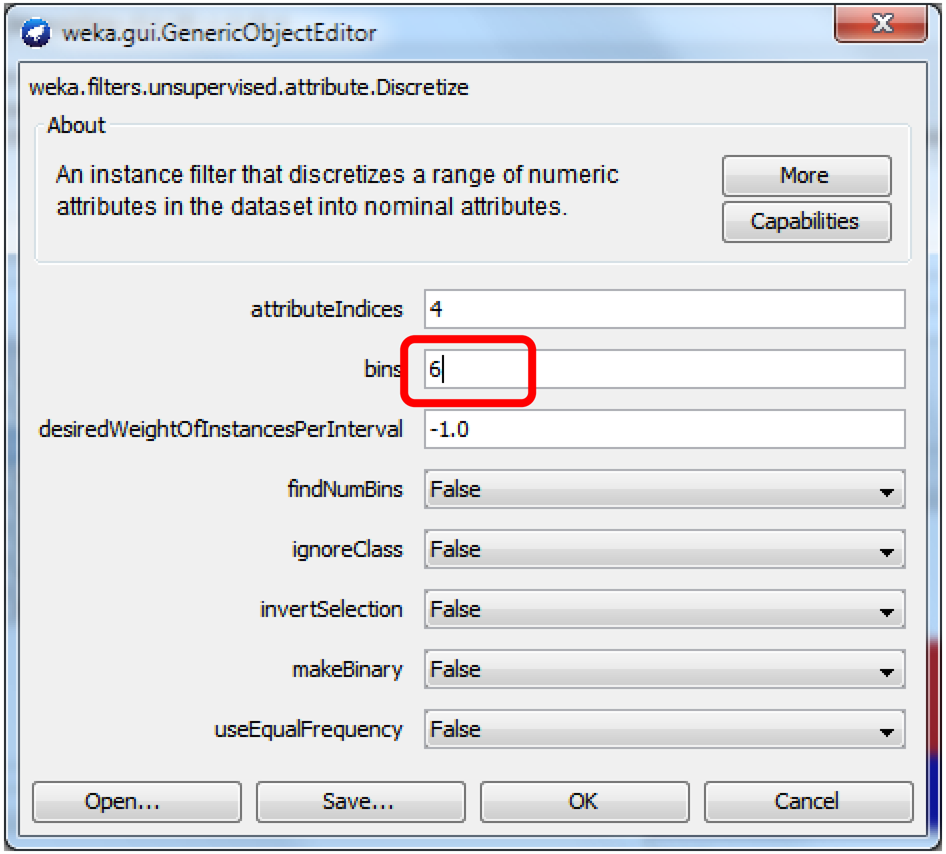
Discretize numeric attributes

- Age after discretization



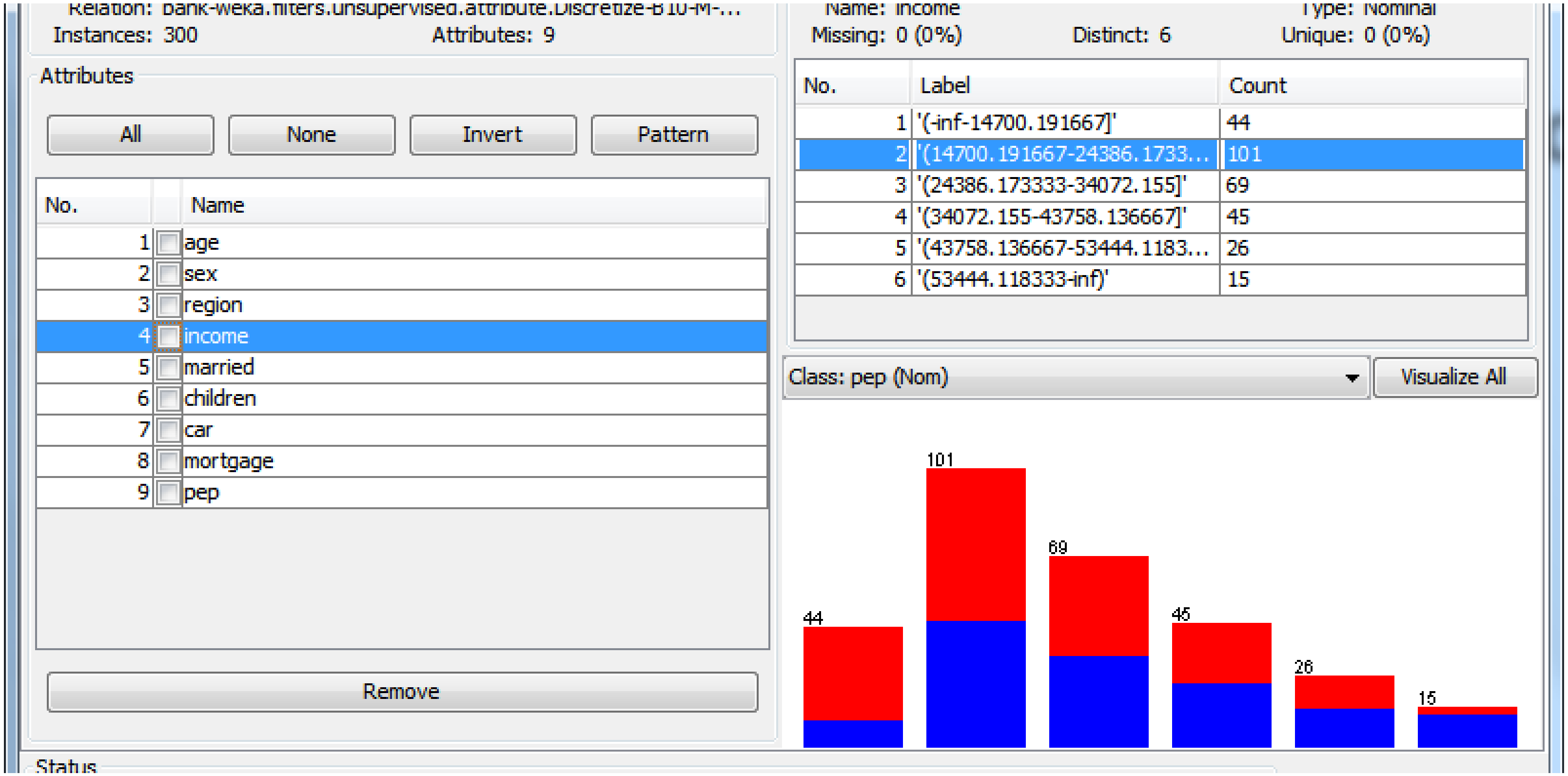
Discretize numeric attributes

- Income



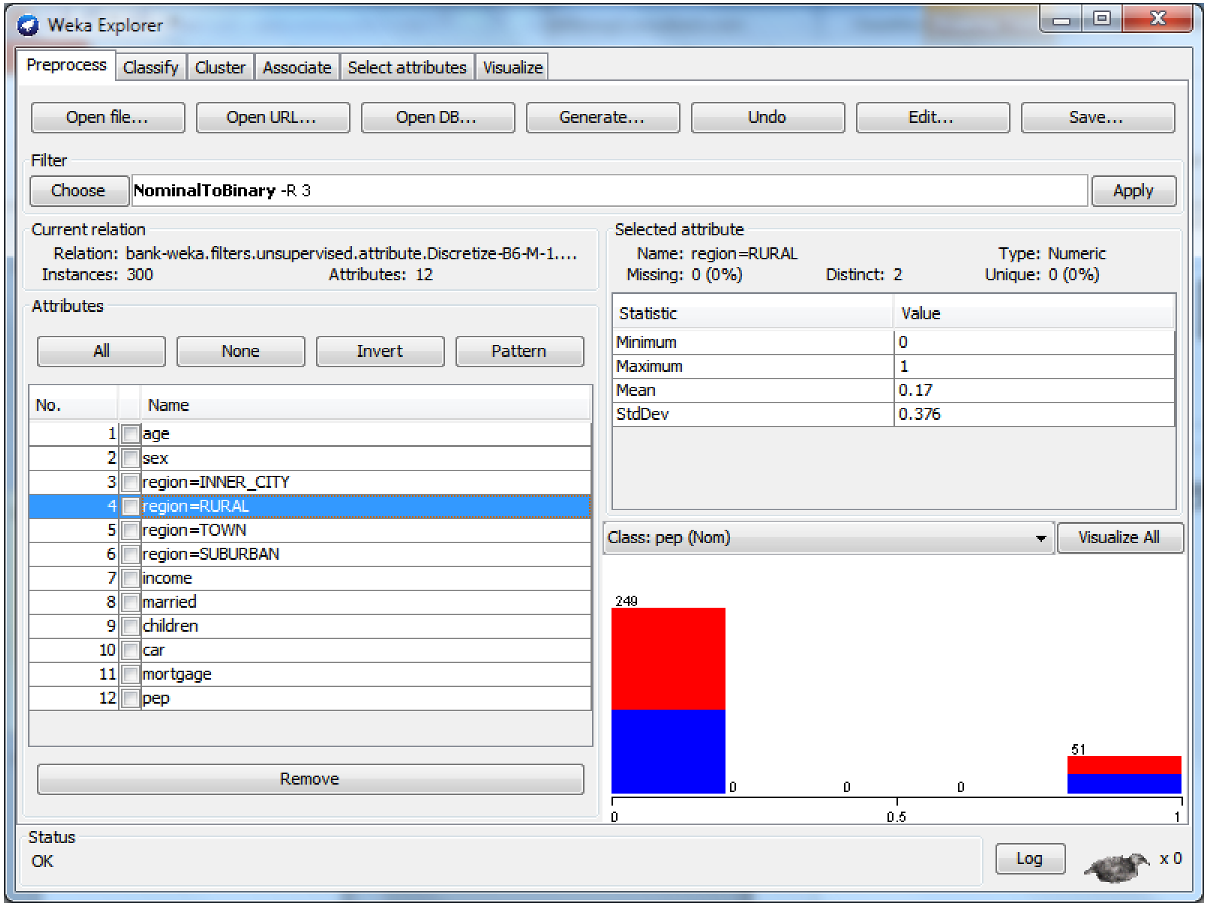
Discretize numeric attributes

- Income after discretization



Optional. [Binaryze multi-valued attributes]

- Undo



Dataset: working with attributes - children

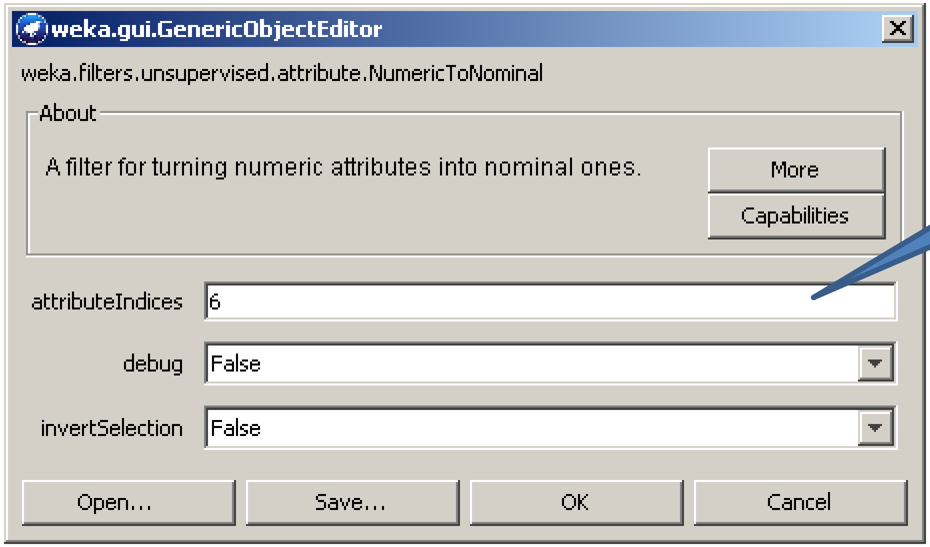
- @relation bank-data

- @attribute age numeric
- @attribute sex {FEMALE,MALE}
- @attribute region {INNER_CITY,TOWN,RURAL,SUBURBAN}
- @attribute income numeric
- @attribute married {NO,YES}
- @attribute children numeric
- @attribute car {NO,YES}
- @attribute save_act {NO,YES}
- @attribute current_act {NO,YES}
- @attribute mortgage {NO,YES}
- @attribute pep {YES,NO}

Not multi-valued: convert to nominal

Convert numeric to nominal

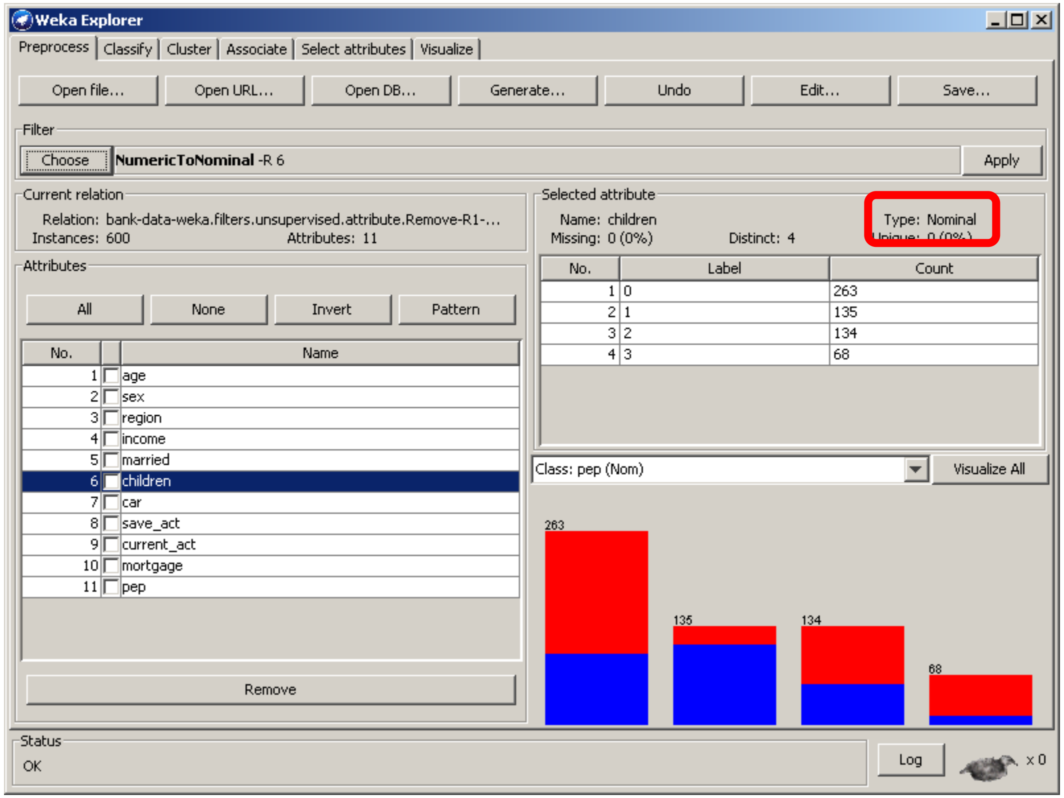
- Filters->Unsupervised->attribute -> NumericToNominal



Index of the attribute to apply filter on: 6

Convert numeric to nominal

- Children after nominalizations: 4 groups

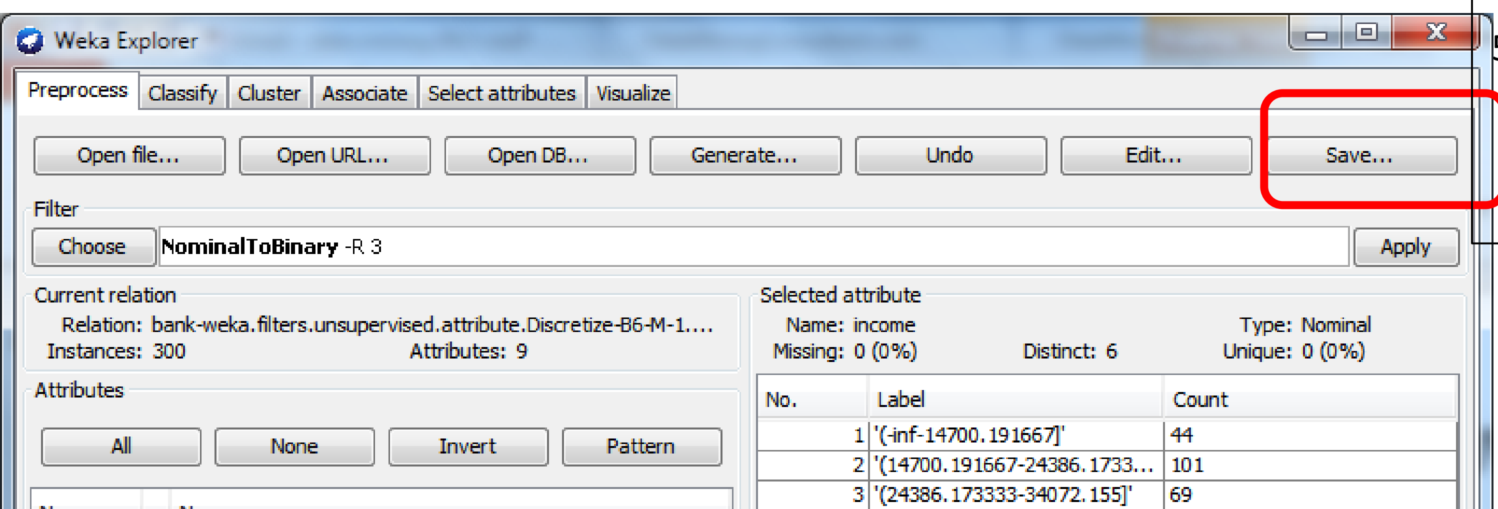


Save the resulting dataset as 'bank2.arff'

- This is the input to our classifiers

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis



Classification

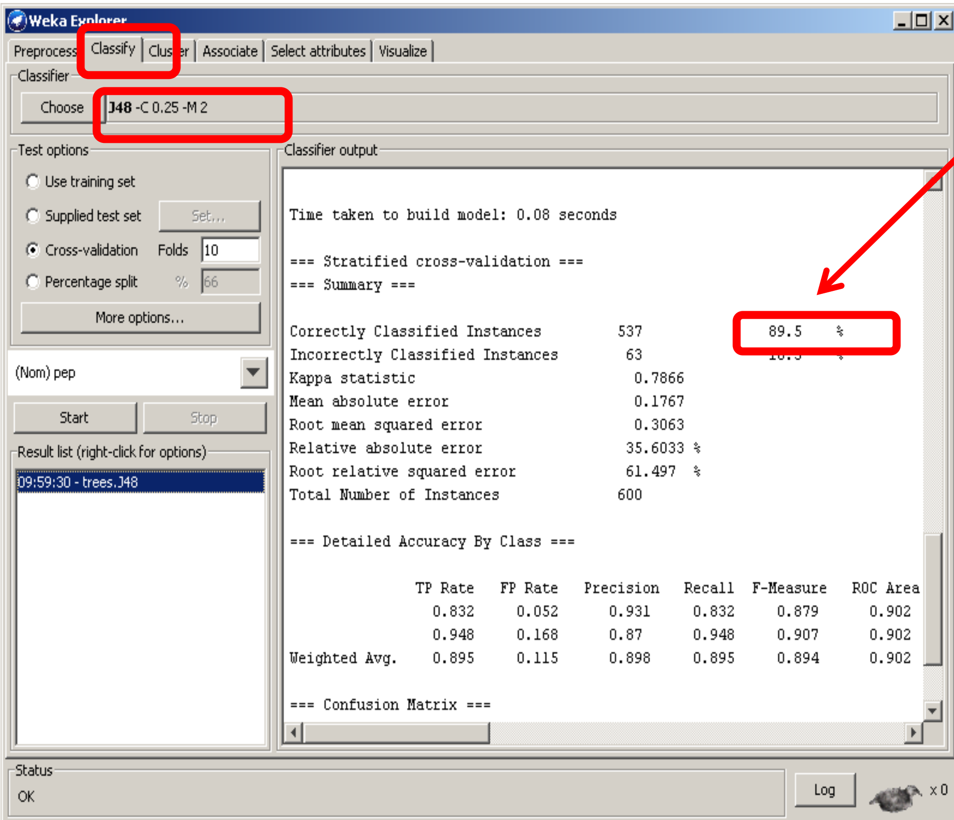
- Our goal: the most accurate classifier

Algorithm	Dataset	Accuracy

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Classification. Trees: J48



Accuracy:
89.5%

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Report

Algorithm	Dataset	Accuracy
J48	bank2.arff	89.5

Attribute selection. Decision tree: J48

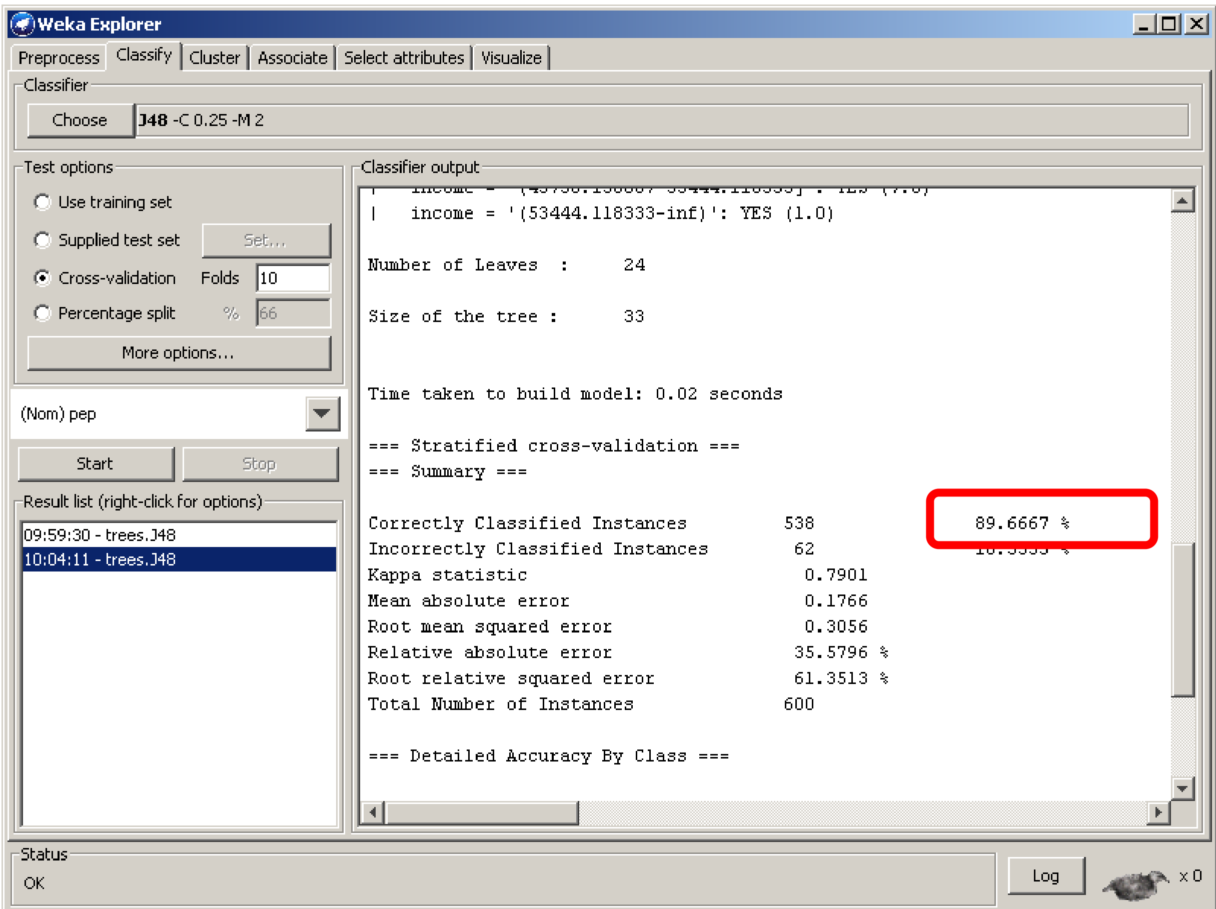
- The most important attributes (used in the tree for splitting nodes): children, married, mortgage, save_act, income

- Let's remove the rest of the attributes (but leave the class attribute!), save file as 'bank3.arff' and try J48 again

```
children = 0
|   married = NO
|   |   mortgage = NO: YES (48.0/3.0)
|   |   mortgage = YES
|   |   |   save_act = NO: YES (12.0)
|   |   |   save_act = YES: NO (23.0)
|   married = YES
|   |   save_act = NO
|   |   |   mortgage = NO: NO (36.0/5.0)
|   |   |   mortgage = YES: YES
(25.0/3.0)
|   |   save_act = YES: NO (119.0/12.0)
children = 1
|   income = '(-inf-14700.191667]': NO
(21.0/3.0)
|   income = '(14700.191667-
24386.173333]': YES (45.0/3.0)
|   income = '(24386.173333-
34072.155]': YES (33.0/2.0)
|   income = '(34072.155-
```

Type I classifiers.
Decision tree: J48 on reduced dataset

- Even better accuracy. Record



Report

Algorithm	Dataset	Accuracy
J48	bank2.arff	89.5
J48	bank3.arff	89.7

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Decision trees: Id3 and Simple cart

Algorithm	Dataset	Accuracy, %
J48	bank2.arff	89.5
J48	bank3.arff	89.7
Id3	bank2.arff	77.0
Id3	bank3.arff	86.0
SimpleCart	bank2.arff	86.8
SimpleCart	bank3.arff	89.5

The best accuracy for decision trees: J48 and on bank3.arff

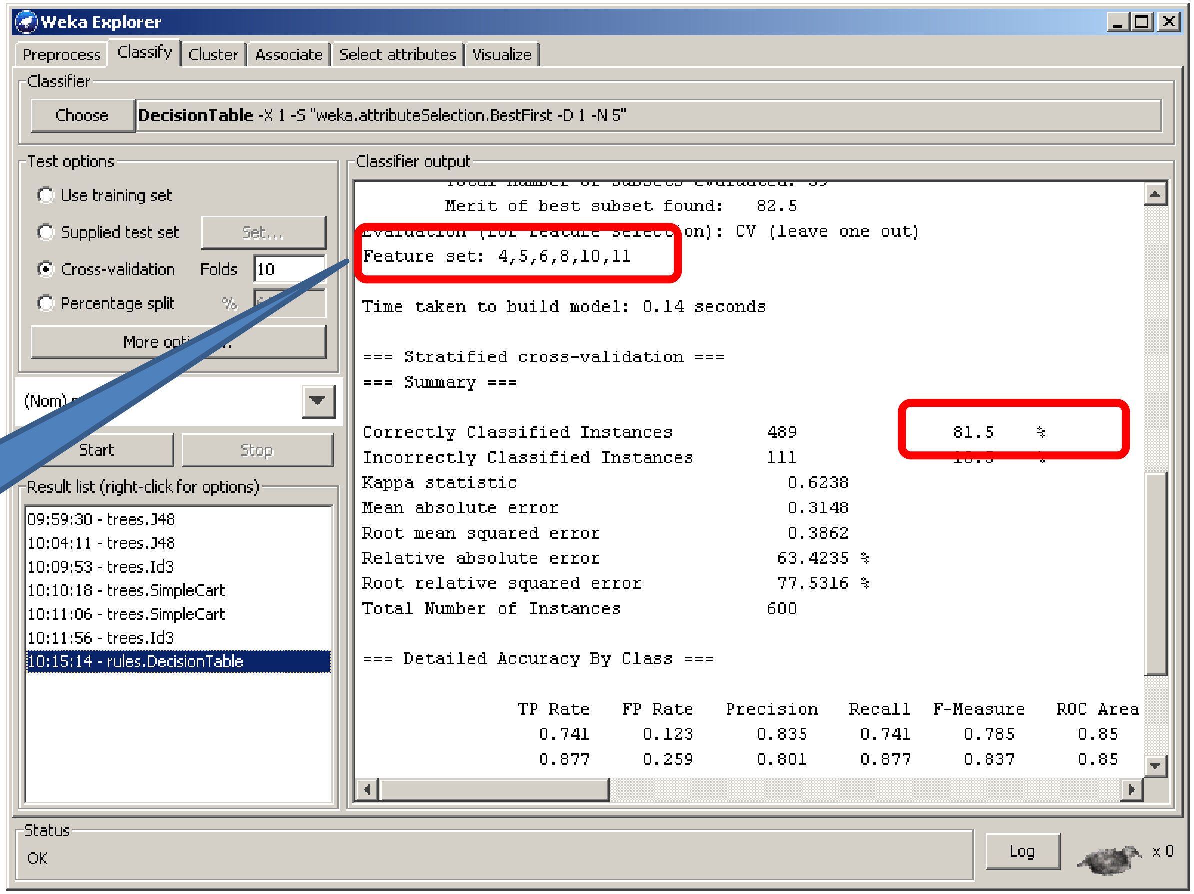
Report so far

Algorithm	Dataset	Accuracy, %
J48	bank3.arff	89.7

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

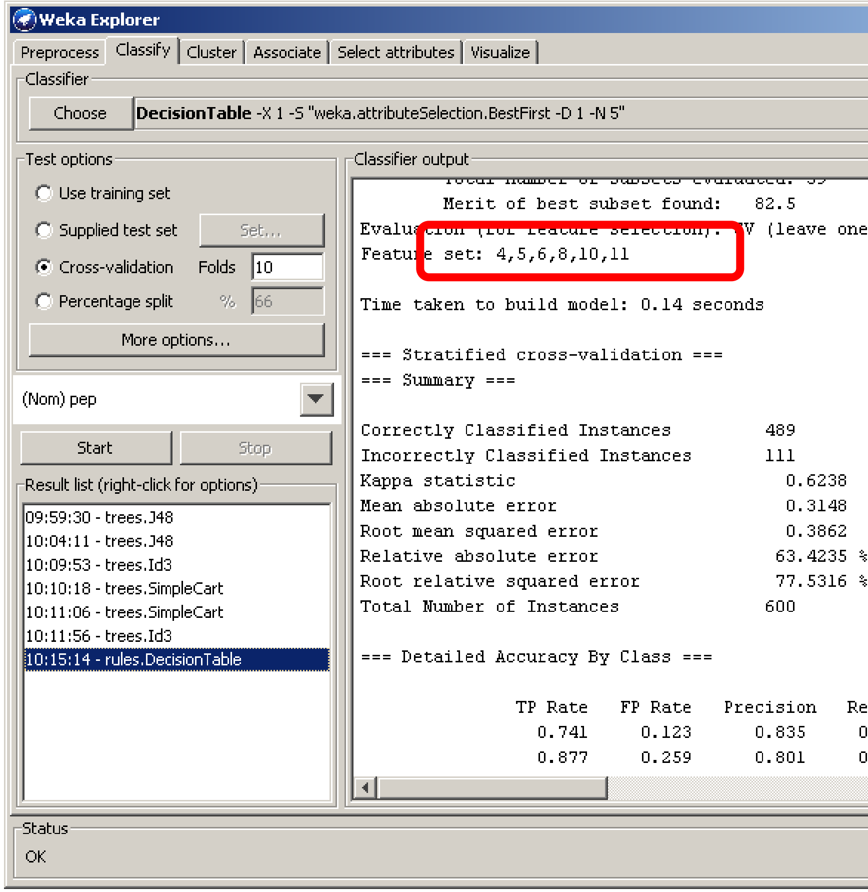
Type 2 classifiers - Rules: DecisionTable on the full dataset bank2.arff



The most important attributes for classification

Attribute selection: DecisionTable on the full dataset

- The most important attributes:
  - 4- income
  - 5- married
  - 6- children
  - 8- save_act
  - 10 - mortgage
- Let's remove the rest
- Save file as bank4.arff
- Re-build decision tree J48: accuracy 89.7 – very high!
- We will use bank4.arff as our input for the rest of the lab



The rest of the Rule learners on bank4.arff

Algorithm	Dataset	Accuracy, %
J48	bank3.arff	89.7
J48	bank4.arff	89.7
JRip	bank4.arff	87.8
Part	bank4.arff	88.3
Prism	bank4.arff	67.3
Ridor	bank4.arff	84.7

Rules

The best result for rule learners

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Report so far

Algorithm	Dataset	Accuracy, %
J48	bank4.arff	89.7
Part	bank4.arff	88.3

Type III classifiers: Naïve Bayes

- For ‘bank2.arff’ (full dataset): 70.5% accurate
- For ‘bank3.arff’ (J48 reduction): 72.5% accurate
- For ‘bank4.arff’ (DecisionTable reduction): 72.5% accurate

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Report

Algorithm	Dataset	Accuracy, %
J48	bank4.arff	89.7
Part	bank4.arff	88.3
NaiveBayes	bank4.arff	72.5

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Generating validation dataset

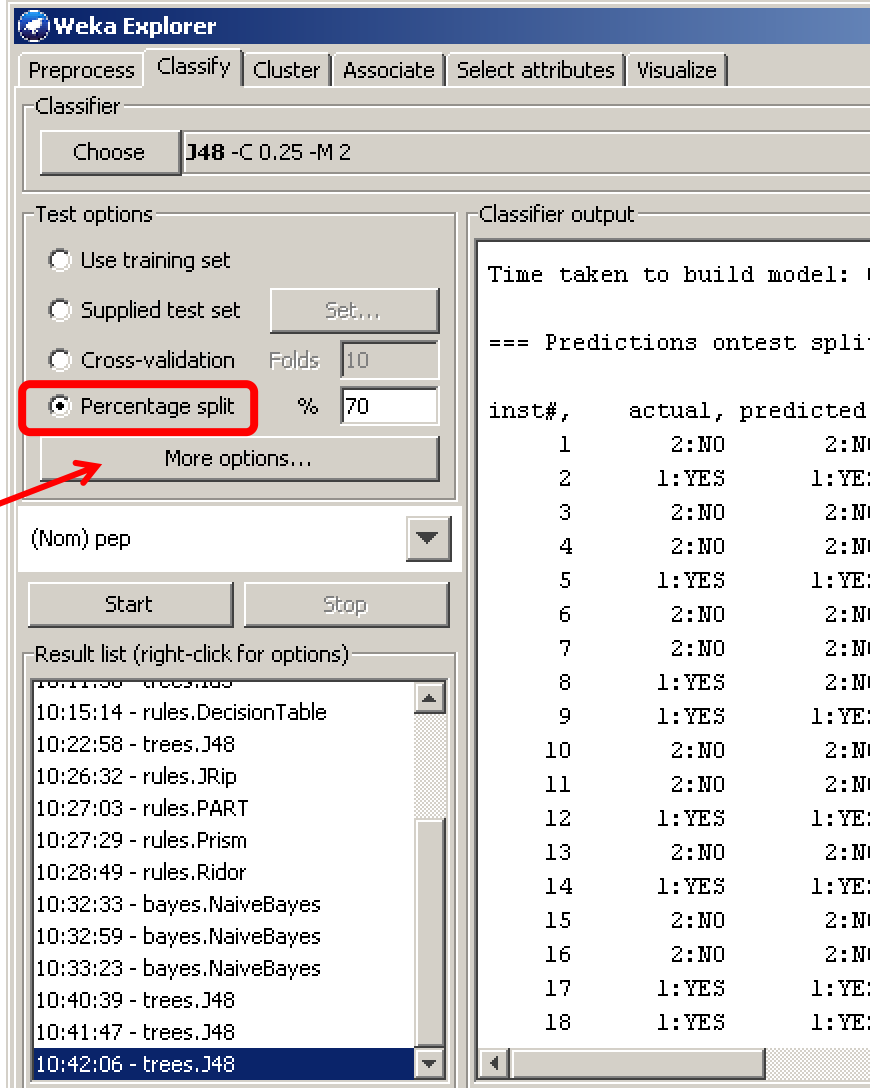
- We will use 70% of the data for training the classifier, and 30% for the validation
- The validation dataset contains actual responses, but we will try to predict them with our best classifier, to see how good is the prediction

Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

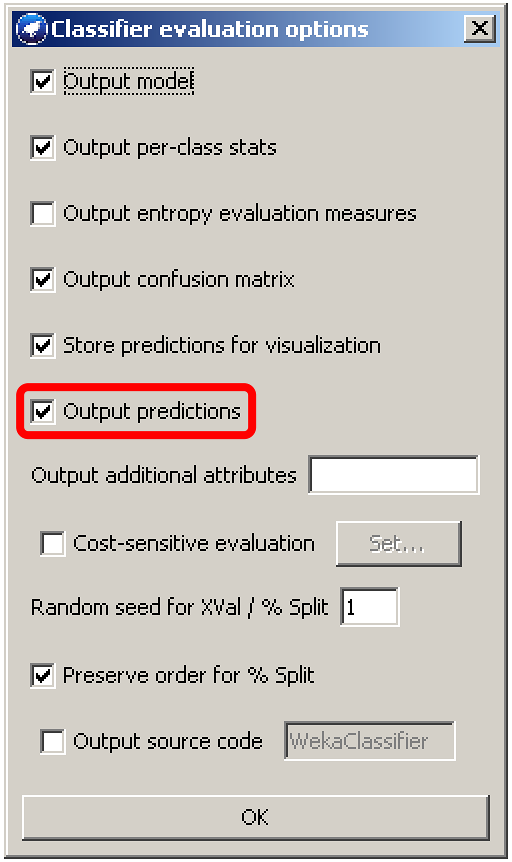
Generating output for business analysis

- Re-open bank4.arff
- Choose one of our best classifiers: J48
- Test options: Percentage split
- Press More Options button



Generating output for business analysis

- Check: Output predictions
- Run J48 Decison tree classifier



Predict class in the validation dataset

- Run J48 using training and validation datasets. Note that the accuracy has decreased.

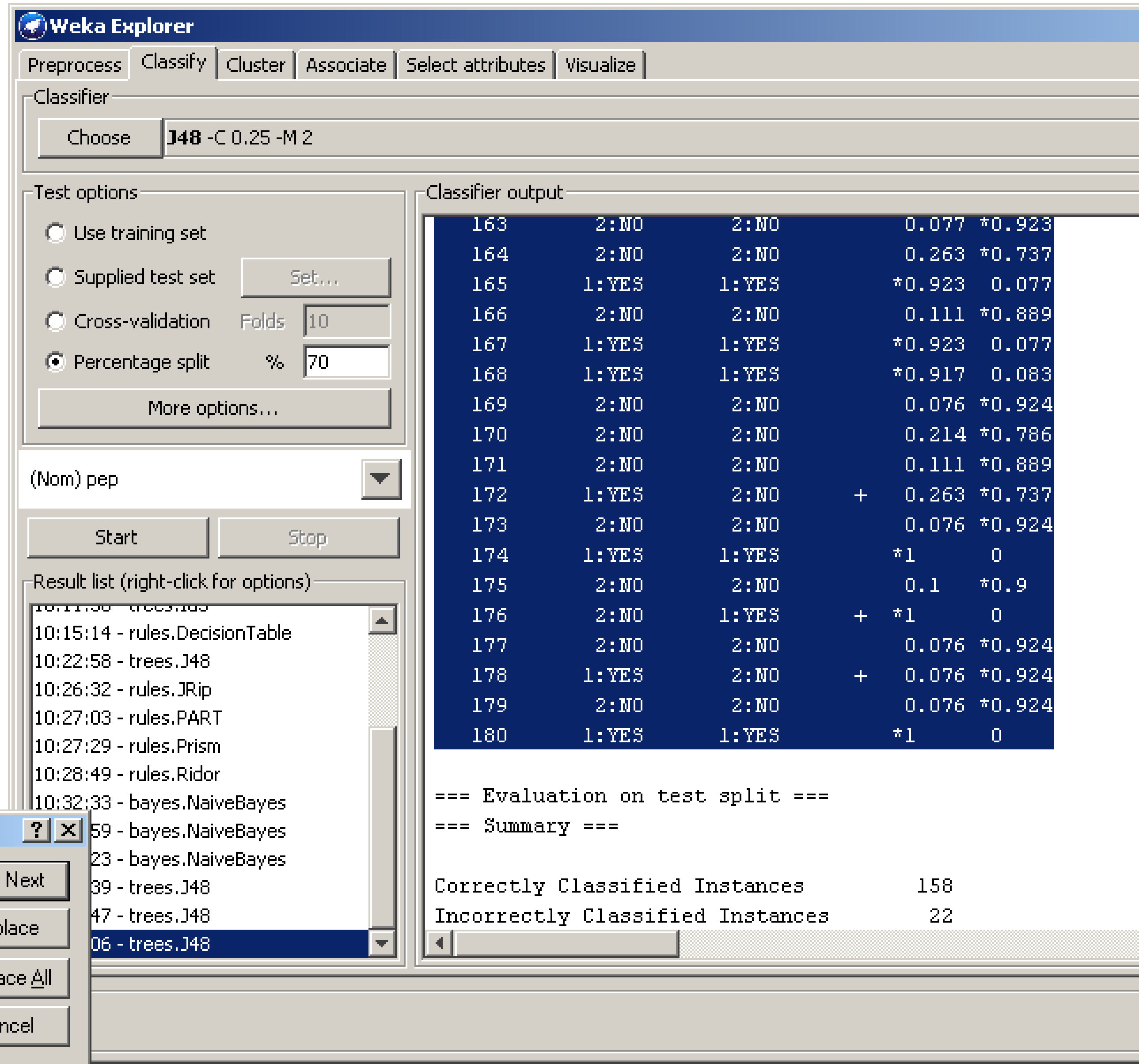
Part I. Data Mining.

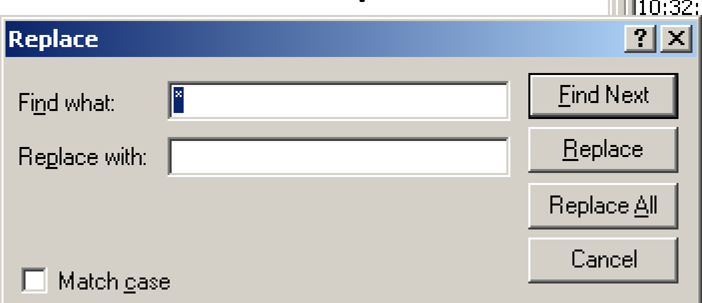
1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Transfer prediction into a text file

- Copy predictions and paste into a text file
- Save file as bank_predicted.txt

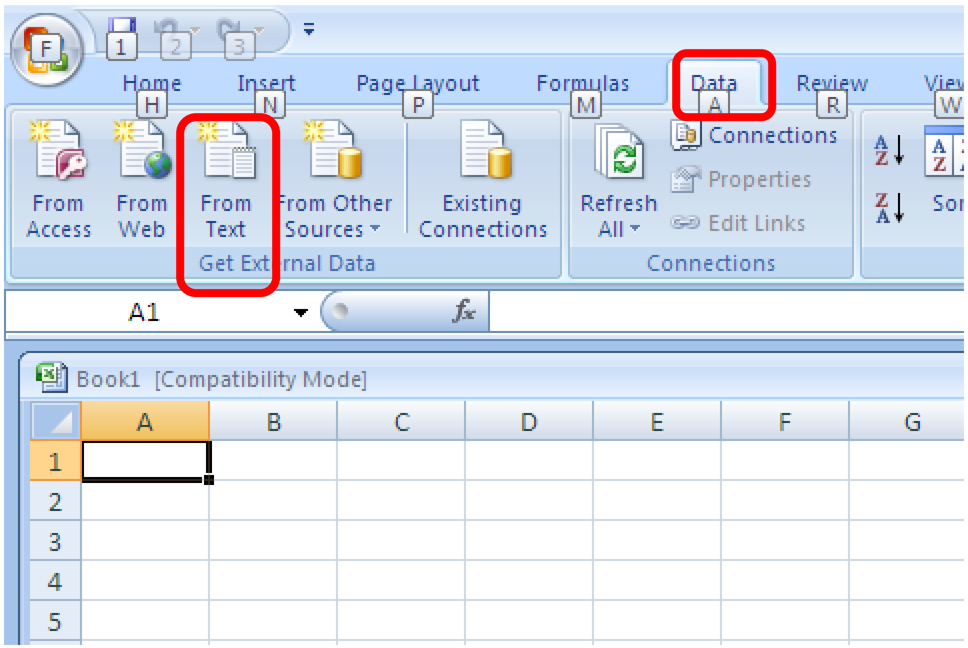
Do find *,+ and replace them with a space





Import predictions into Electronic tables program: example - Excel

- Import data from bank_predicted.txt

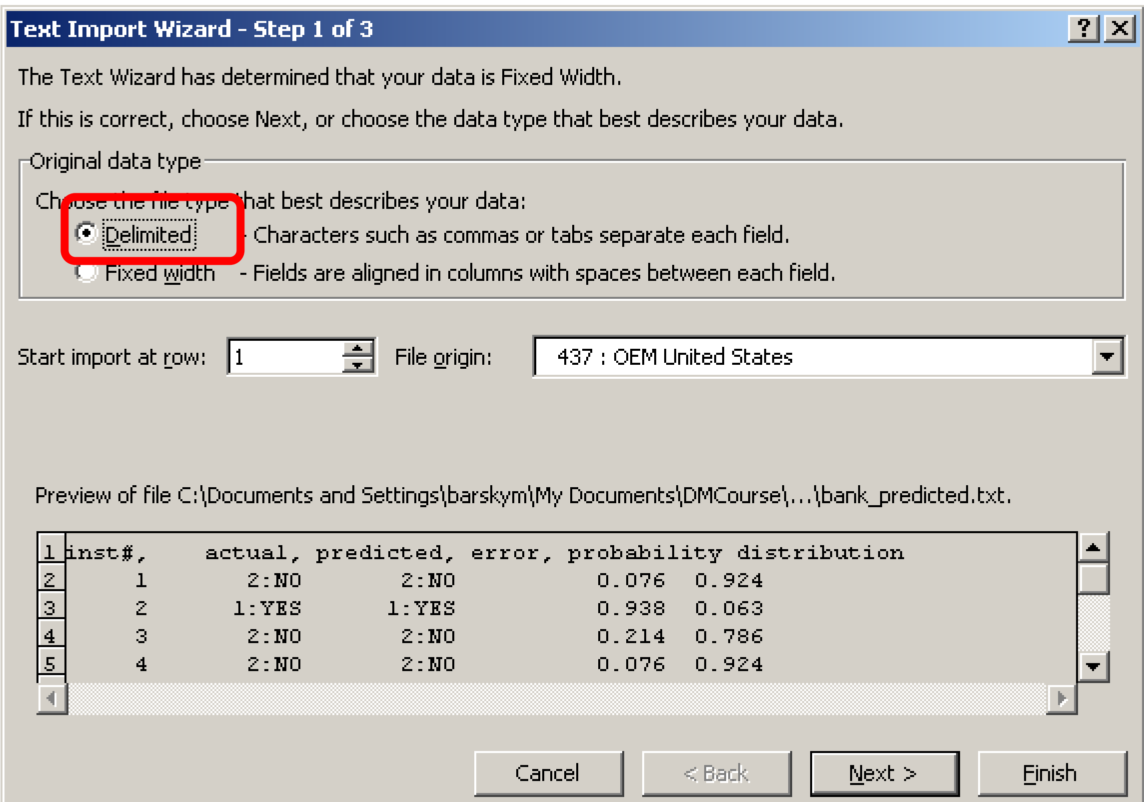


Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Import predictions into Electronic tables program: example - Excel

- Import data from bank_predicted.txt



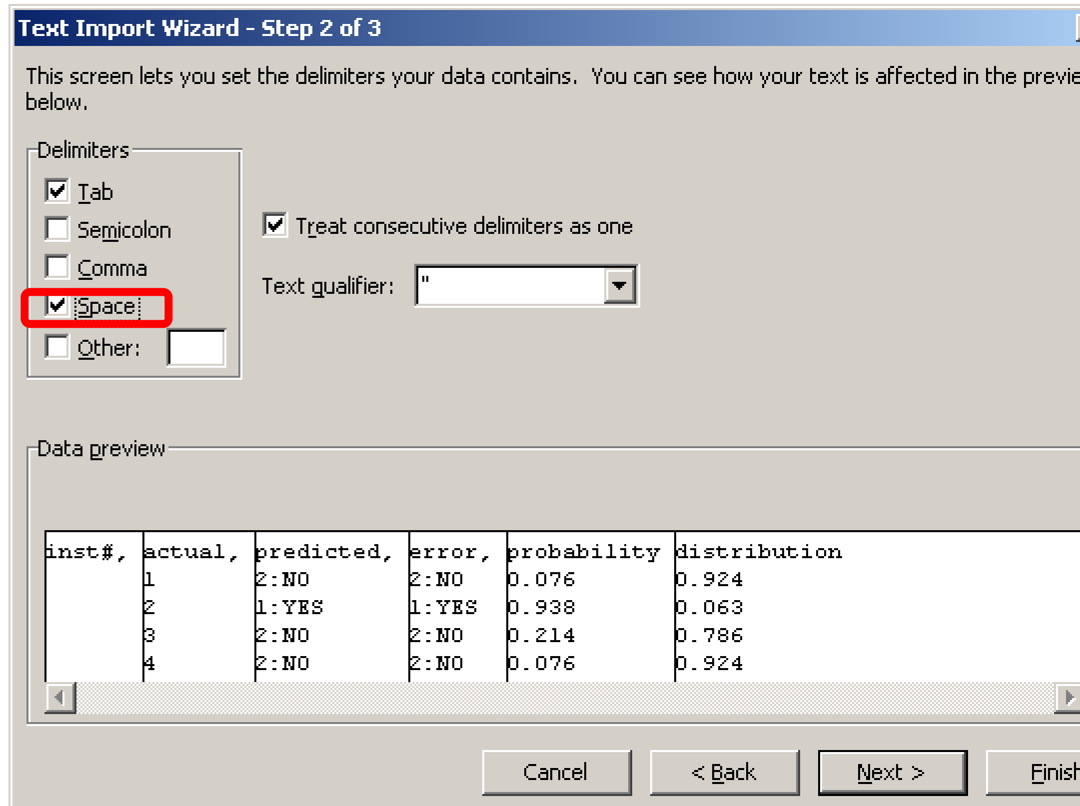
Part I. Data Mining.

1. Prepare data
2. Build several classifiers. Choose the most accurate one.
3. Divide dataset into training and validation datasets
4. Predict class in the validation dataset
5. Prepare output for business analysis

Import predictions into Electronic tables program: example - Excel

- Import data from bank_predicted.txt

- Save file as bank_results.xls (sample file is attached)



Close WEKA

- The data mining part is complete

Part I. Data Mining.

Part II. Business analysis

1. Generate Lift chart(s)
2. Cost-benefit analysis
3. Recommendations