Feature Selection

Graphics and Notes "Data Mining and Predictive Analytics" 2015 Elsevier : Vijay Kotu and Bala Deshpande

Terri Hoare – May 2023

Data Mining
Feature Selection

- Feature selection in predictive analytics refers to the process of identifying the few most important variables or attributes that are essential in a model for an accurate prediction

- Feature selection optimises the performance of the data mining algorithm by deceasing the number of dimensions of the problem (curse of dimensionality). The speed of an algorithm is typically dependant on the number of attributes (dimensions)

- Feature selection simplifies the problem and makes it easier for the analyst to interpret the outcome of the modelling by removing redundant attributes

- Highly correlated attributes which do not add more information are removed. In multiple regression type models, two or more correlated attributes tend to result in unstable or counter intuitive results (multicollinearity)

- Feature selection is needed to remove independent variables (attributes) that may be strongly correlated to each other and ensures that only independent variables (attributes) that may be strongly correlated to the dependent attribute are kept

Data Mining
Feature Selection cont.

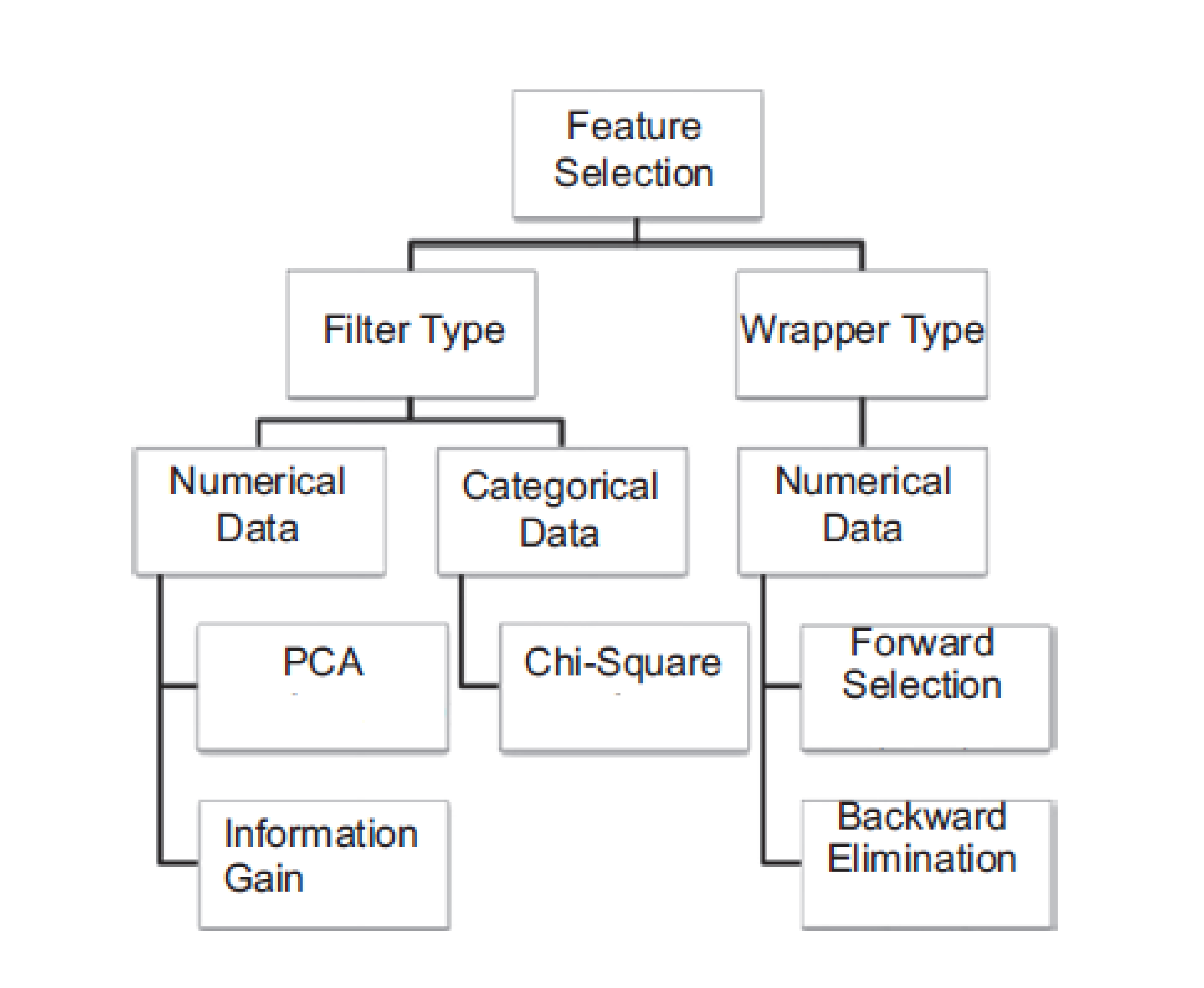
- **Dimension reduction methods** combine or merge actual attributes in order to reduce the number of attributes of a raw data set. **Principal Component Analysis** is very important in data mining

- **Feature selection methods** work more like filters that eliminate some attributes. There are two types **filter** and **wrapper.**

  - ➢ **Filter** approaches work by selecting only those attributes that rank among the top in meeting certain stated criteria. It is typically applied before the modelling process, typically when the number of features or attributes is really large or when computational expense is a criterion

  - ➢ **Wrapper** approaches work by iteratively selecting via a feedback loop only those attributes that improve the performance of an algorithm

Data Mining
Feature Selection : Taxonomy


Feature Selection

Filter Type

Wrapper Type

Numerical Data

Categorical Data

Numerical Data

PCA

Information Gain

Chi-Square

Forward Selection

Backward Elimination

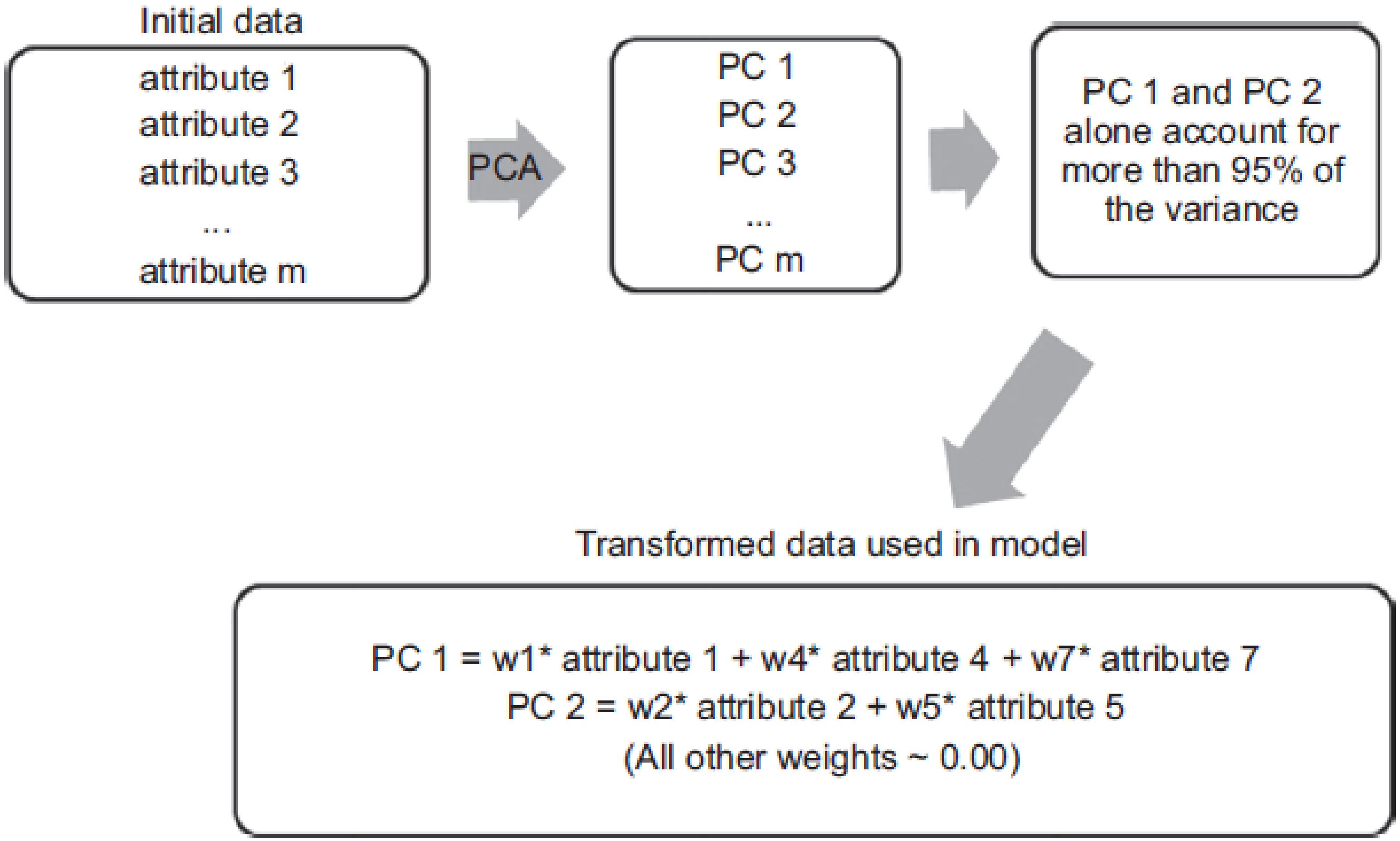
Data Mining Feature Selection
Filter Based : Principal Component Analysis

Description PCA (Principal Component Analysis) is a dimension reduction method. It combines the most important attributes into a fewer number of transformed attributes. It is one of the most widely used techniques in data mining.

When a data set has a large number of variables, many of which are correlated, the result can be unmanageable. PCA provides a method to simplify the data set down to only those important variables that summarise all the information in the data set. Thus it captures the attributes that contain the greatest amount of variability in the data set. It does this by transforming the existing variables into a set of "principal components" or **new variables** that have the following properties:-

- they are uncorrelated with each other
- they cumulatively contain / explain a large amount of variance within the data
- they can be related back to the original variables via weightage factors (the original variables with very low weightage factors in their principal components are effectively removed from the data set)

Data Mining Feature Selection
Filter Based : PCA cont.



Initial data

attribute 1
attribute 2
attribute 3
...
attribute m

PCA

PC 1
PC 2
PC 3
...
PC m

PC 1 and PC 2 alone account for more than 95% of the variance

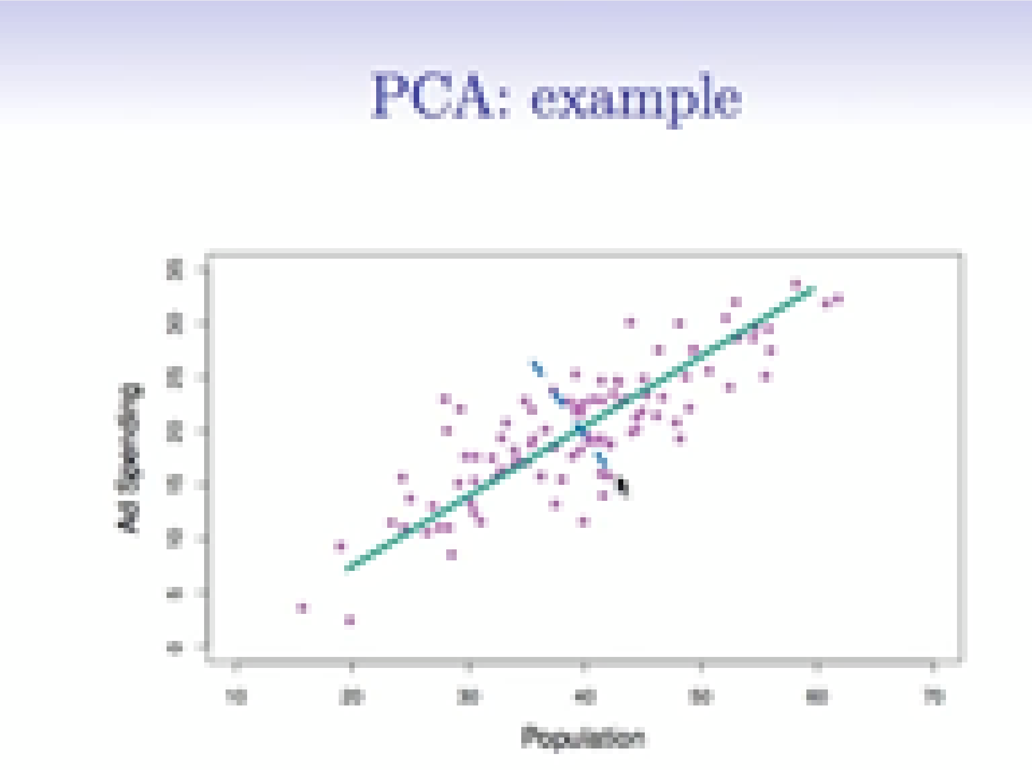
Transformed data used in model

PC 1 = w1* attribute 1 + w4* attribute 4 + w7* attribute 7
PC 2 = w2* attribute 2 + w5* attribute 5
(All other weights ~ 0.00)

A conceptual framework illustrating the effectiveness of using PCA for feature selection. The final data set includes only PC1 and PC2.

Data Mining Feature Selection
Filter Based : PCA : How it Works

Taken from **Stanford Statistical Learning Online**



PCA: example

PCA-1 is the solid green line (maximum variability) and PCA-2 the dotted blue line. **Loading** vectors are depicted from the centre point (intersection) and in the case of PCA-1 north east. All points that is (population, ad sampling) can be projected onto PCA-1 green line and PCA-2 blue line by how far along the lines the point lies (**Scores**). Values left of centre negative. Values right of centre positive.

Data Mining Feature Selection
Filter Based : PCA : How it Works cont.

Example from Stanford Statistical Learning Online

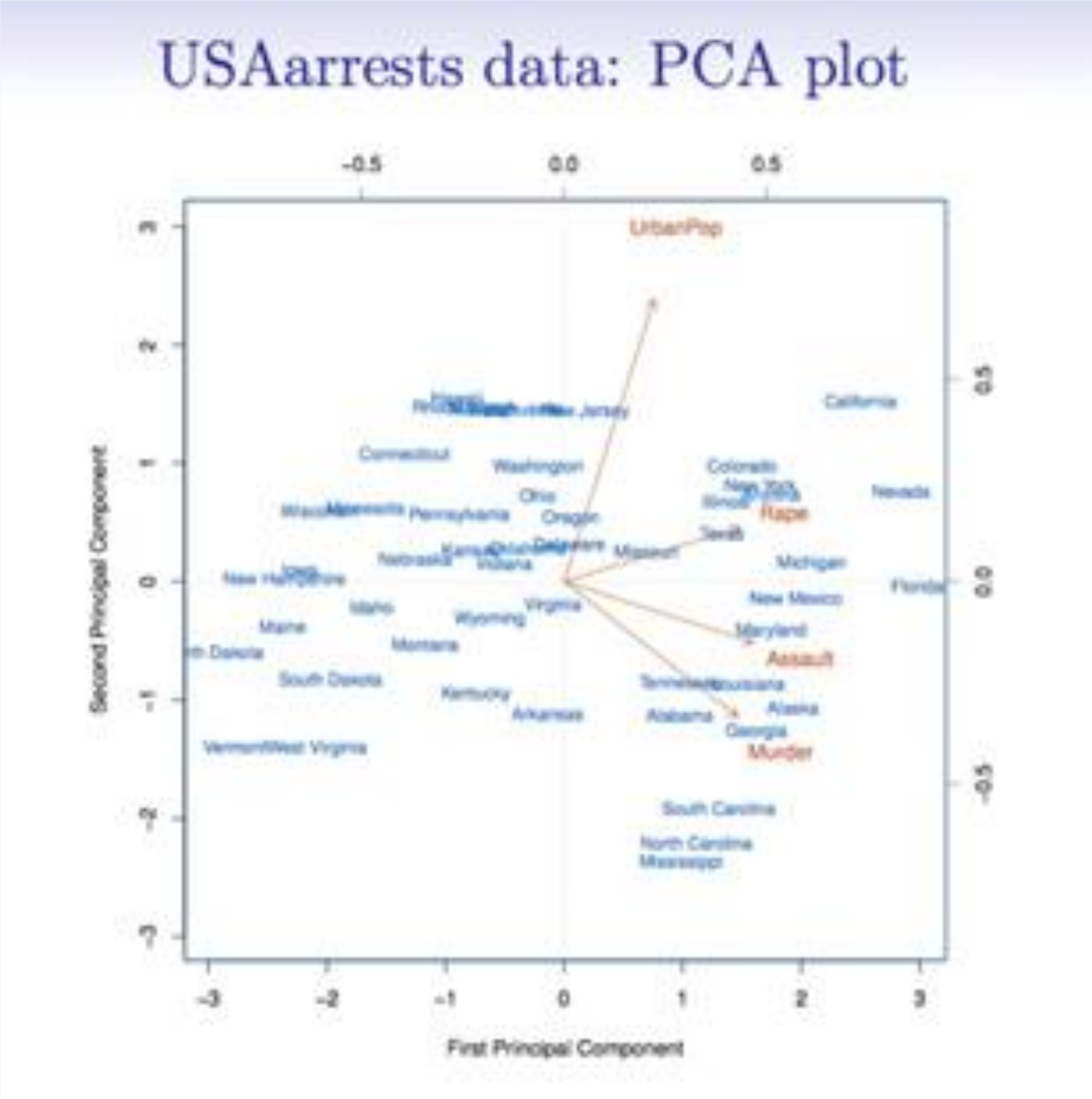
US Arrests data. For each of the fifty States in the United States, the data set contains the number of arrests per 100 000 residents for each of three crimes : Assault, Murder, and Rape. Also recorded is UrbanPop (the percent of the population in each state living in urban areas).

On the following slide, the blue State Names represent the **scores** for the first two principal components.

The orange arrows indicate the first two principal component **loading vectors** (with axes on the top and right) for each of the four attributes (Assault, Murder, Rape, UrbanPop) for example for Rape the first principal component is 0.54 and the second principal component is 0.17.

The Bi-Plot plots the principal component scores for each example of State in the data set together with the principal component loadings. The loadings give an interpretation of the States data for example the large positive rating on Rape, Assault, and Murder show that the States plotted right are high crime states whereas the states plotted left are low crime states.

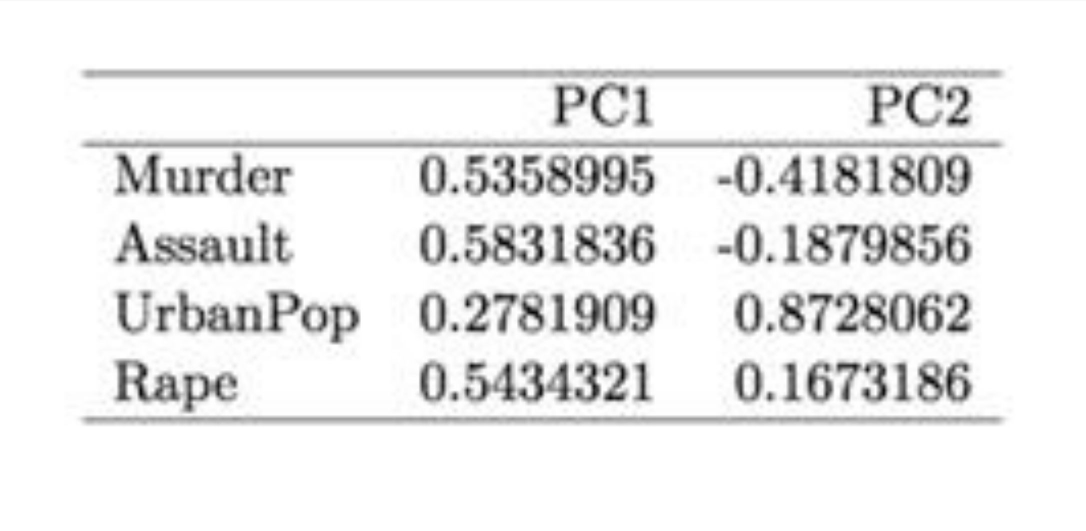
Data Mining Feature Selection
Filter Based : PCA : How it Works cont.



USAarrests data: PCA plot

Data Mining Feature Selection
Filter Based : PCA : How it Works cont.

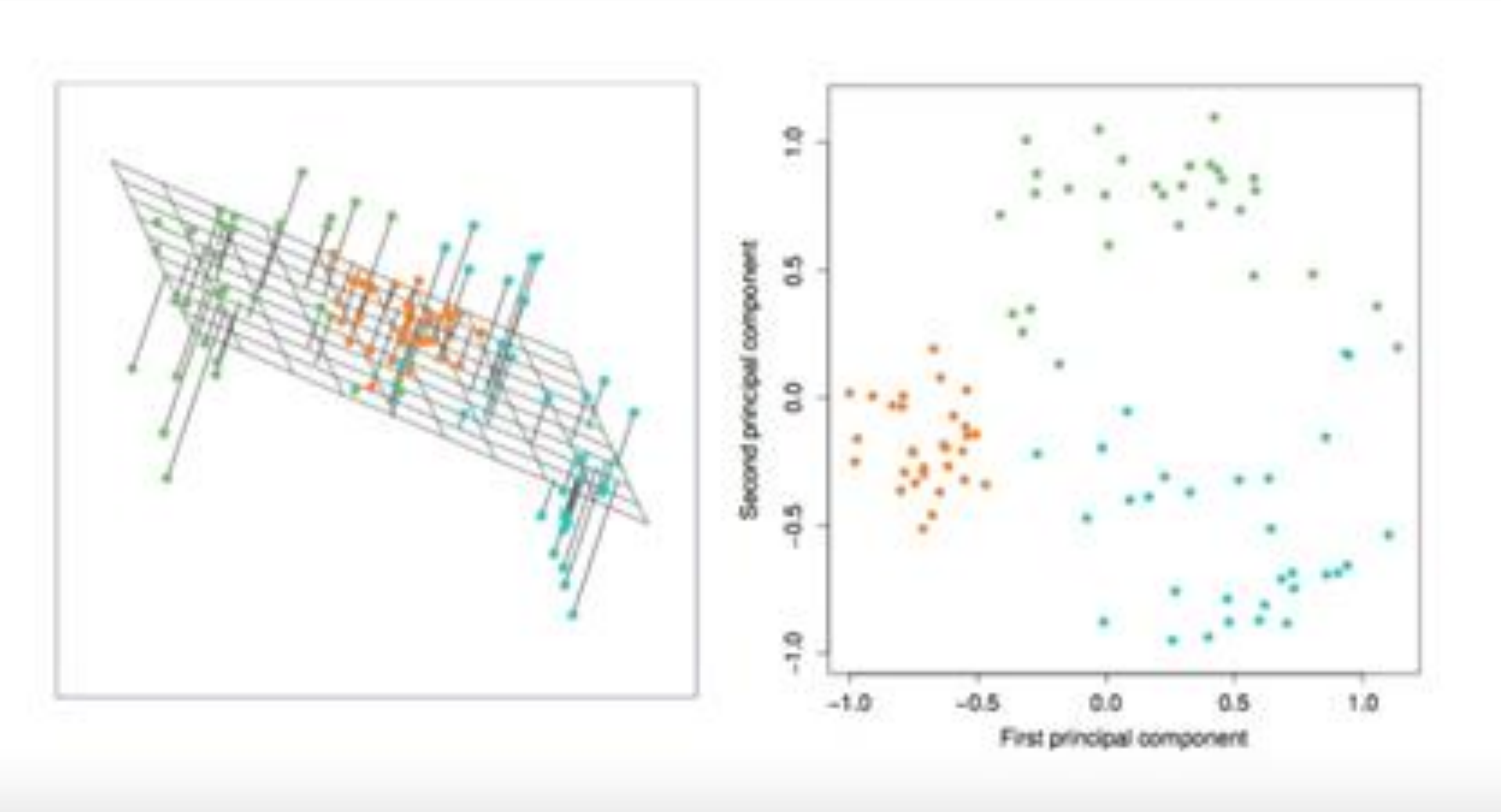
Statistical Learning Online : Table of loadings



	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

Data Mining Feature Selection
Filter Based : PCA : How it Works cont.

Statistical Learning Online : PCA Another Interpretation : Hyperplane



Data Mining Feature Selection
Filter Based : PCA : How it Works cont.

Statistical Learning Online : PCA Another Interpretation : Hyperplane

- The first principal component loading vector has a very special property: it defines the line in p dimensional space that is closest to the n observations (using average squared Euclidean distance as a measure of closeness)

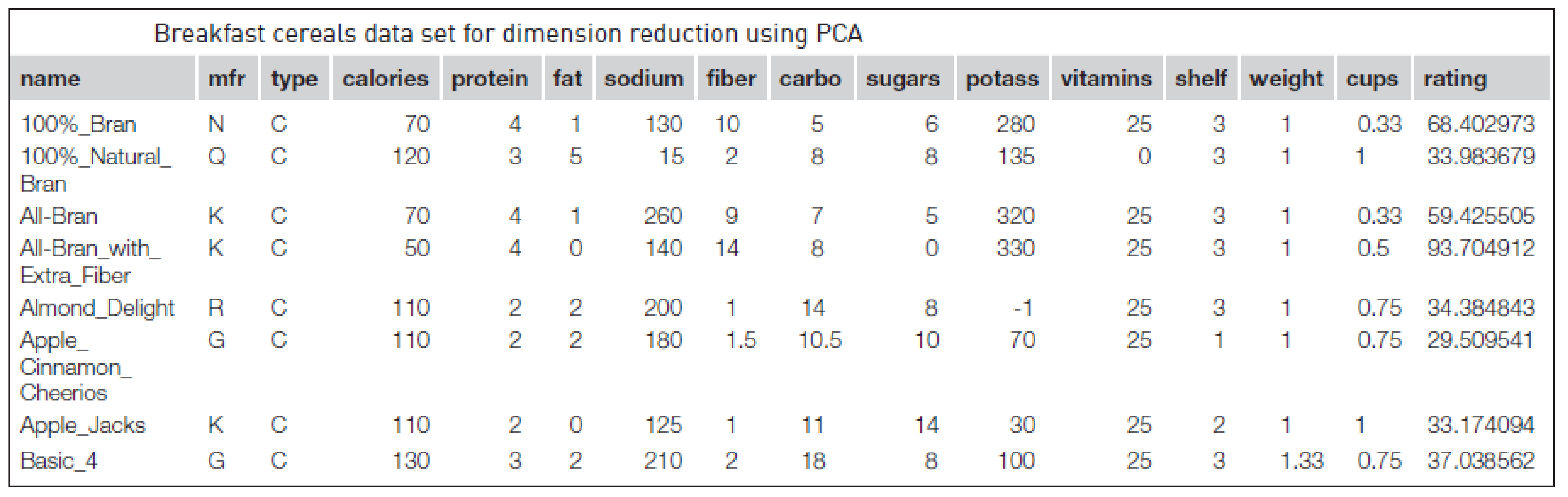
- The notion of principal components as the dimensions that are closest to the n observations extends beyond just the first principal component

- For instance, the first two principal components of a dataset span the plane that is closest to the n observations, in terms of average squared Euclidean distance

Data Mining Feature Selection
Filter Based : PCA : How to Implement

12_Feature_12.2_PCA_cereals.rmp   12_Feature_12.2_cereals-PCA.xlsx

Data set includes information on ratings and nutritional information on 77 breakfast cereals. There are 16 attributes of which 13 are numerical. The objective is to reduce this set of 13 numerical predictors to a much smaller list using PCA. First run is on un-normalised data, the second on normalised data.



Breakfast cereals data set for dimension reduction using PCA

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.402973
100%_Natural_ Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.983679
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.425505
All-Bran_with_ Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.704912
Almond_Delight	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.384843
Apple_ Cinnamon_ Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.509541
Apple_Jacks	K	C	110	2	0	125	1	11	14	30	25	2	1	1	33.174094
Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.038562

Data Mining Feature Selection
Filter Based : PCA : How to Implement cont.

12_Feature_12.2_PCA_cereals.rmp 12_Feature_12.2_cereals-PCA.xlsx

- **Step 1 – Data Preparation**
  Remove non-numeric attributes **Select Attributes**
  Read in excel data **Read Excel**

  Step 2 – PCA Operator
  Principal Components Analysis operator
  Default V**ariance Threshold** of 0.95 or 95% to select only those attributes that collectively account for or explain 95% of the total variance in the data

  Step 3 – Execution and Interpretation
  Three PCA related tabs:-

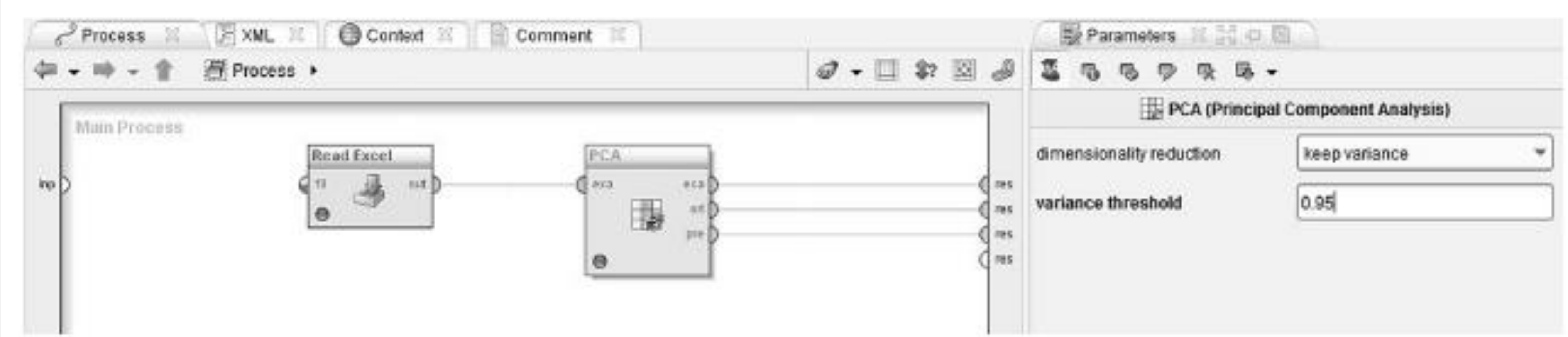
  1. **Eigenvalues** - Information on the contribution each PC individually and cumulatively

  2. **Eigenvectors** - Sort the eigenvectors (weighting factors) for each PC and choose the two to three highest (absolute) valued weighting factors for PC’s 1 to 3 to choose the attributes

  3. **Cumulative Variance Plot**

Data Mining Feature Selection
Filter Based : PCA : How to Implement cont.

Step 1&2 – Process



PCA (Principal Component Analysis)

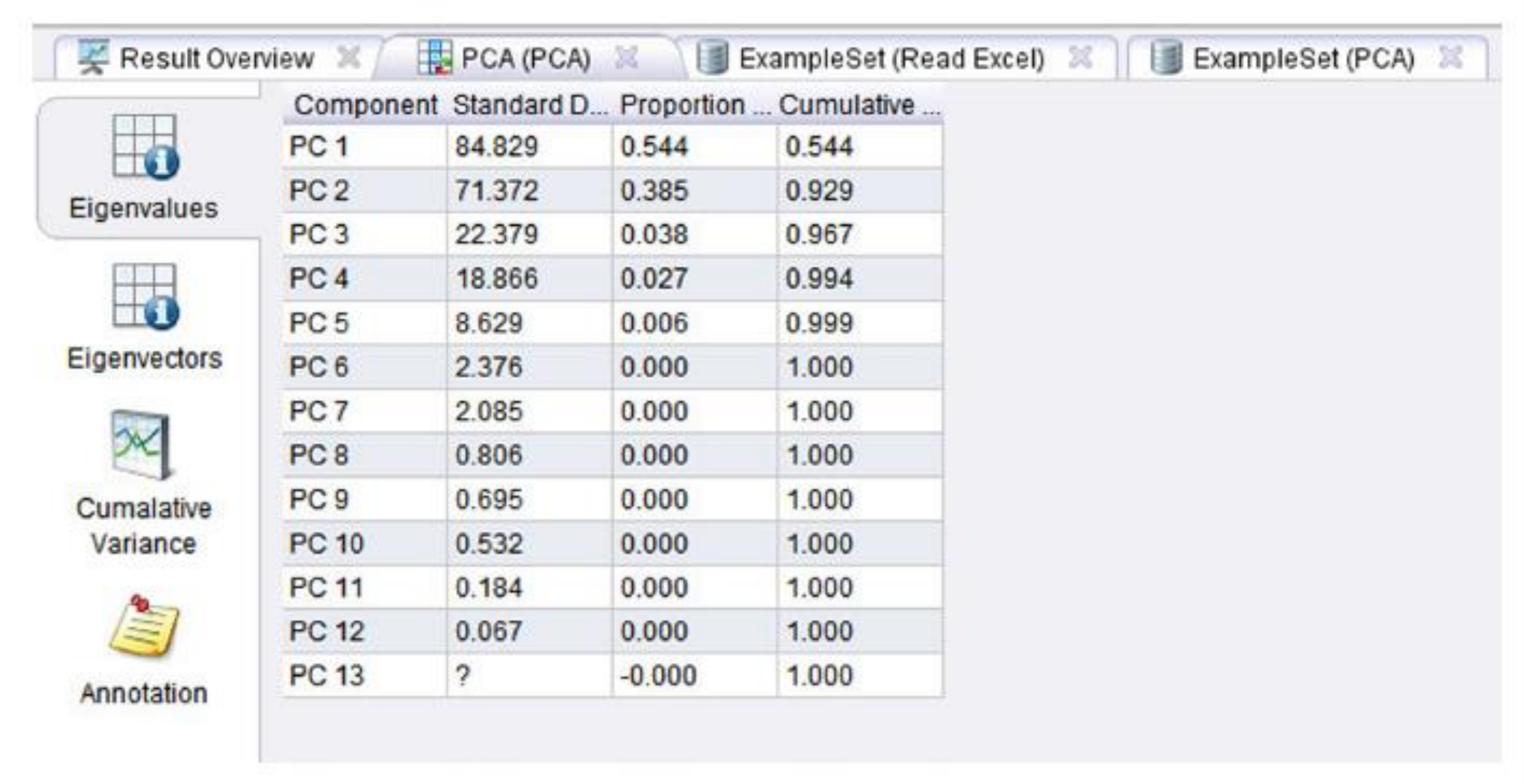
dimensionality reduction: keep variance

variance threshold: 0.95

Configuring the PCA operator.

Data Mining Feature Selection
Filter Based : PCA : How to Implement cont.

Step 3 – Execution results – Eigenvalues Tab

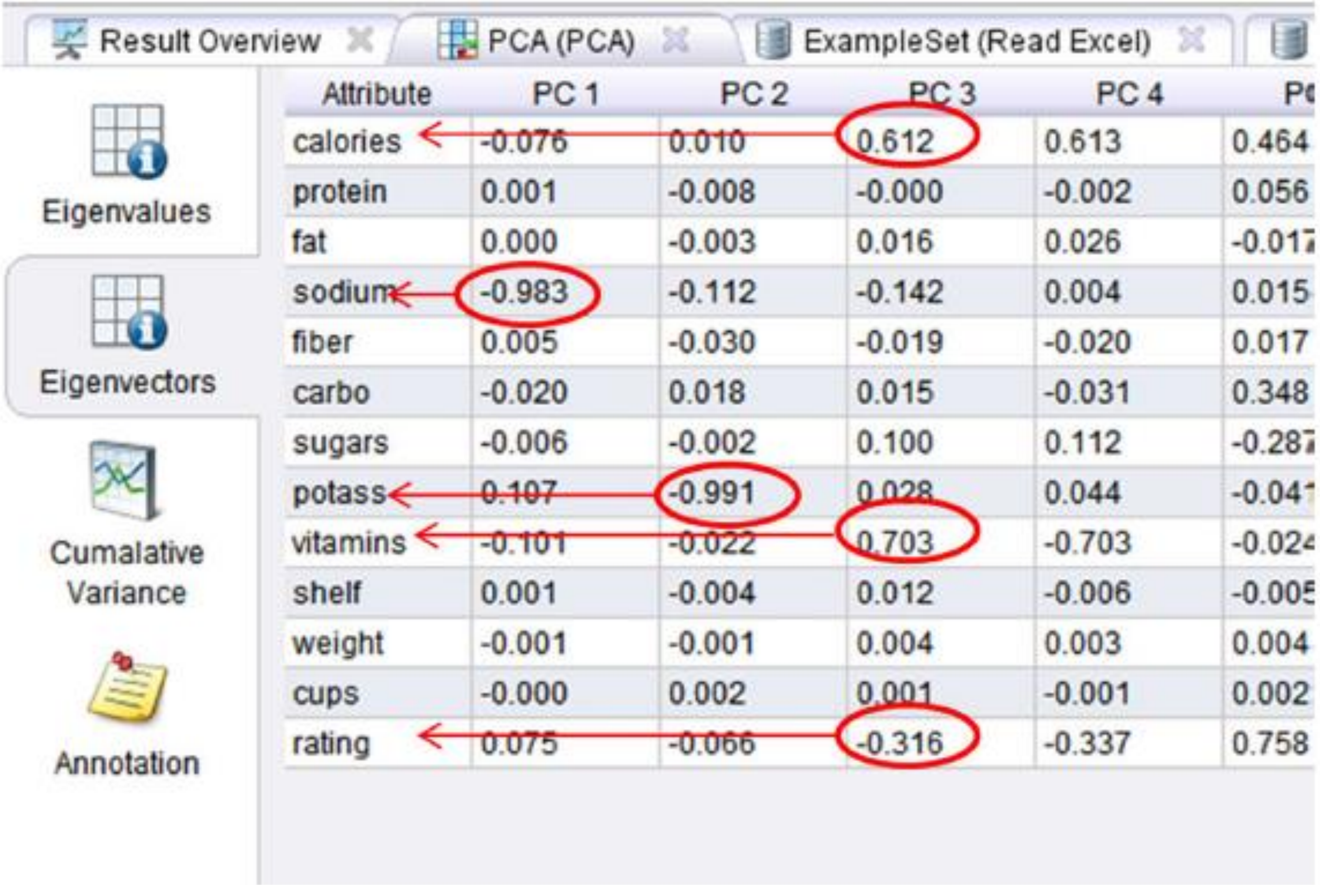


Component	Standard D...	Proportion ...	Cumulative ...
PC 1	84.829	0.544	0.544
PC 2	71.372	0.385	0.929
PC 3	22.379	0.038	0.967
PC 4	18.866	0.027	0.994
PC 5	8.629	0.006	0.999
PC 6	2.376	0.000	1.000
PC 7	2.085	0.000	1.000
PC 8	0.806	0.000	1.000
PC 9	0.695	0.000	1.000
PC 10	0.532	0.000	1.000
PC 11	0.184	0.000	1.000
PC 12	0.067	0.000	1.000
PC 13	?	-0.000	1.000

Output from PCA.

Data Mining Feature Selection
Filter Based : PCA : How to Implement cont.

Step 3 – Execution results – Eigenvectors Tab



Attribute	PC 1	PC 2	PC 3	PC 4	P
calories	-0.076	0.010	0.612	0.613	0.464
protein	0.001	-0.008	-0.000	-0.002	0.056
fat	0.000	-0.003	0.016	0.026	-0.017
sodium	-0.983	-0.112	-0.142	0.004	0.015
fiber	0.005	-0.030	-0.019	-0.020	0.017
carbo	-0.020	0.018	0.015	-0.031	0.348
sugars	-0.006	-0.002	0.100	0.112	-0.287
potass	0.107	-0.991	0.028	0.044	-0.041
vitamins	-0.101	-0.022	0.703	-0.703	-0.024
shelf	0.001	-0.004	0.012	-0.006	-0.005
weight	-0.001	-0.001	0.004	0.003	0.004
cups	-0.000	0.002	0.001	-0.001	0.002
rating	0.075	-0.066	-0.316	-0.337	0.758

Selecting the reduced set of attributes using the Eigenvectors tab from the PCA operator.

Data Mining Feature Selection
Filter Based : PCA : Discussion Points

- For the lecture example PCA reduces the number of attributes for 13 to 5, a more than 50% reduction in the number of data points that any model would realistically need to consider! Huge performance improvements possible on larger data sets!

- A very effective and widely used tool particularly when all attributes are numeric

- Risks to consider :-

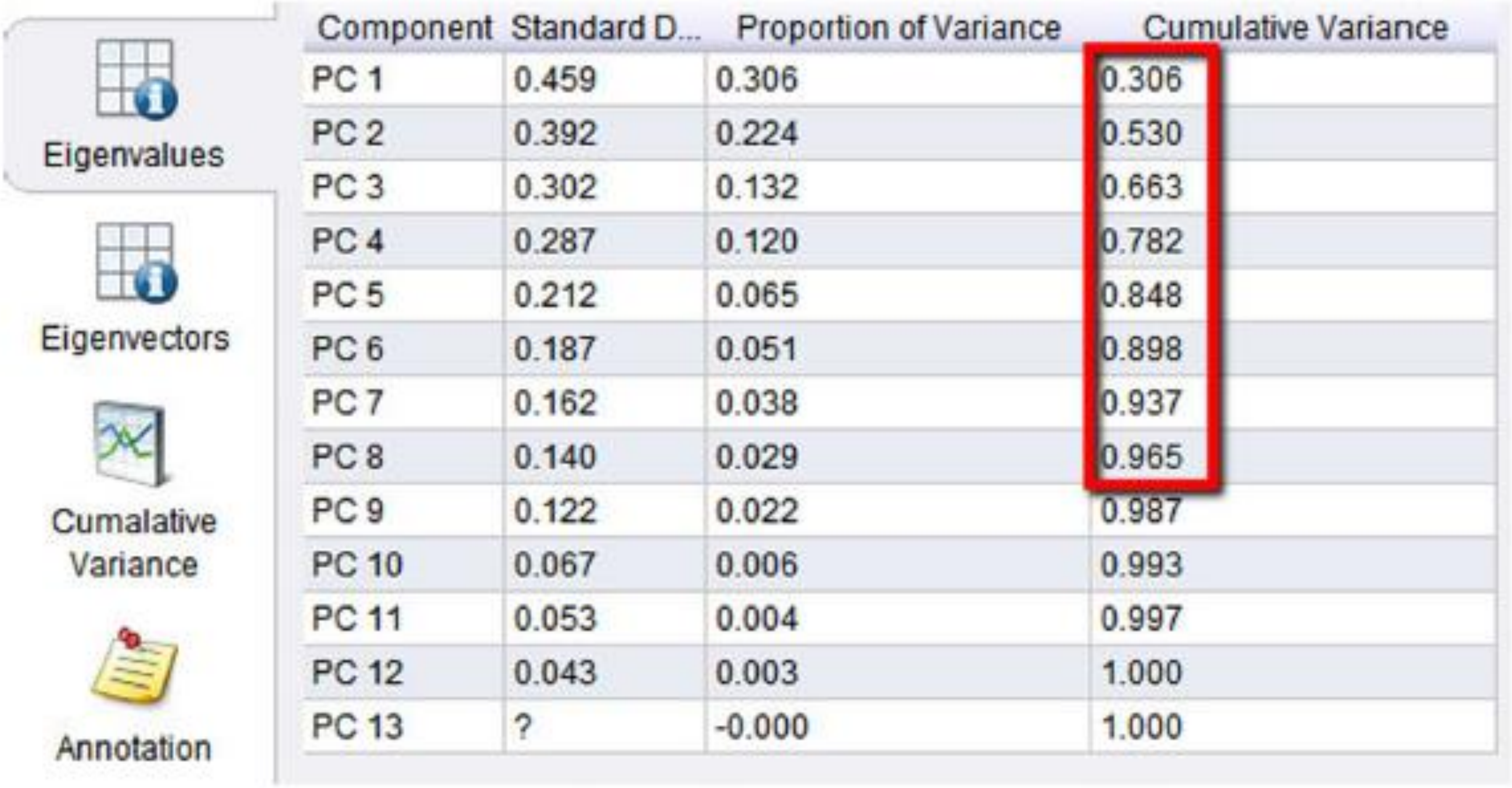
    - Results must be evaluated in the context of the data, If data is very noisy, the PCA might end up suggesting that the noisiest are the most significant as they account for most of the variation for example the crowd in a study of which musical instruments are influencing the harmonics at a rock concert!

Data Mining Feature Selection
Filter Based : PCA : Discussion Points cont.

- Risks to consider cont.:-
  - Adding uncorrelated data does not always help if it is noise. Similarly coincidental correlations could be involved for example number of hours worked in a garment factory and pork prices within a certain period of time. Care should be taken that results make business sense

  - PCA is very sensitive to scaling effects in the data. In the example potassium ranges from -1 to 330 and sodium ranges from 1 to 320. Comparatively most other factors range in the single or low double digits. Therefore sodium and potassium dominated because they contributed to the maximum variance in the data. To minimize the effects of scaling the data should be normalized with the **Normalize** operator. With this data transformation all attributes are reduced to a range between 0 and 1. For this example this results in 8 PCs being needed to account for 95% variance in the data but a more correct result! Refer next slide…

Data Mining Feature Selection
Filter Based : PCA : Discussion Points cont.



Component	Standard D...	Proportion of Variance	Cumulative Variance
PC 1	0.459	0.306	0.306
PC 2	0.392	0.224	0.530
PC 3	0.302	0.132	0.663
PC 4	0.287	0.120	0.782
PC 5	0.212	0.065	0.848
PC 6	0.187	0.051	0.898
PC 7	0.162	0.038	0.937
PC 8	0.140	0.029	0.965
PC 9	0.122	0.022	0.987
PC 10	0.067	0.006	0.993
PC 11	0.053	0.004	0.997
PC 12	0.043	0.003	1.000
PC 13	?	-0.000	1.000

Interpreting RapidMiner output for Principal Component Analysis.

Clustering Models
Filter Based : PCA Summary

- **Description (Model N/A)**
  PCA is in reality a dimension reduction method. It combines the most important attributes into a fewer number of transformed attributes

- **Input**
  Numerical attributes

- **Output**
  Numerical attributes (reduced set). Does not really require a label

- **Pros**
  Efficient way to extract predictors that are uncorrelated to each other. Helps to identify attributes with highest variance

- **Cons**
  Very sensitive to scaling effects, requires normalisation of attribute values before application. Focus on variance sometimes results in selecting noisy attributes

- **Use Cases**
  Most numerical valued data sets that require dimension reduction

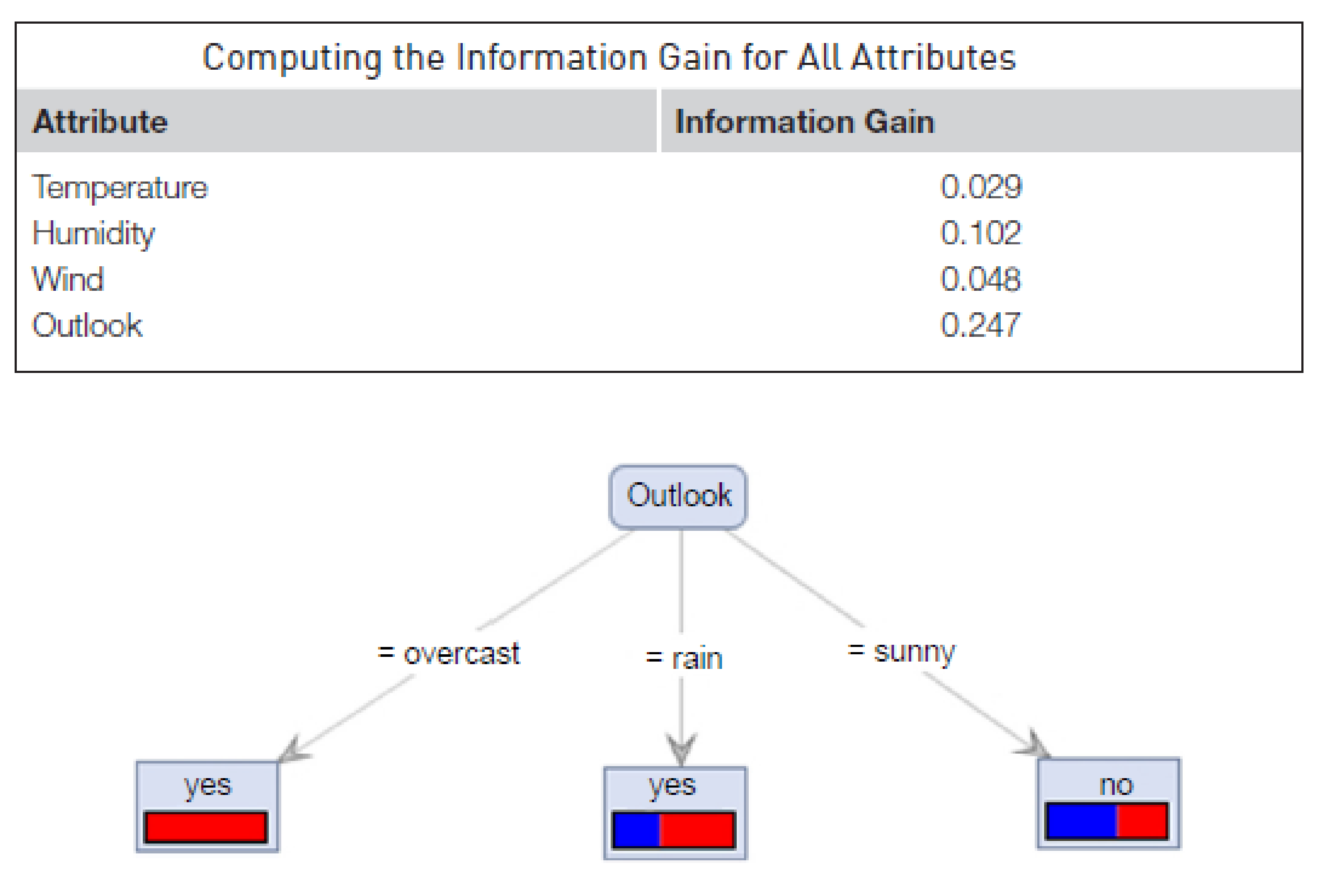
Data Mining Feature Selection
Filter Based : Info Gain

Description Selecting attributes based on relevance to the target or label

The key to feature selection is to include attributes that have a strong correlation with the predicted or dependent variable. With these techniques we can rank attributes based on the amount of information gain and then select only those that meet or exceed some threshold

Data Mining Feature Selection
Filter Based : Info Gain : How it Works

As used for building decision trees. Selection of the attribute on which to split the tree



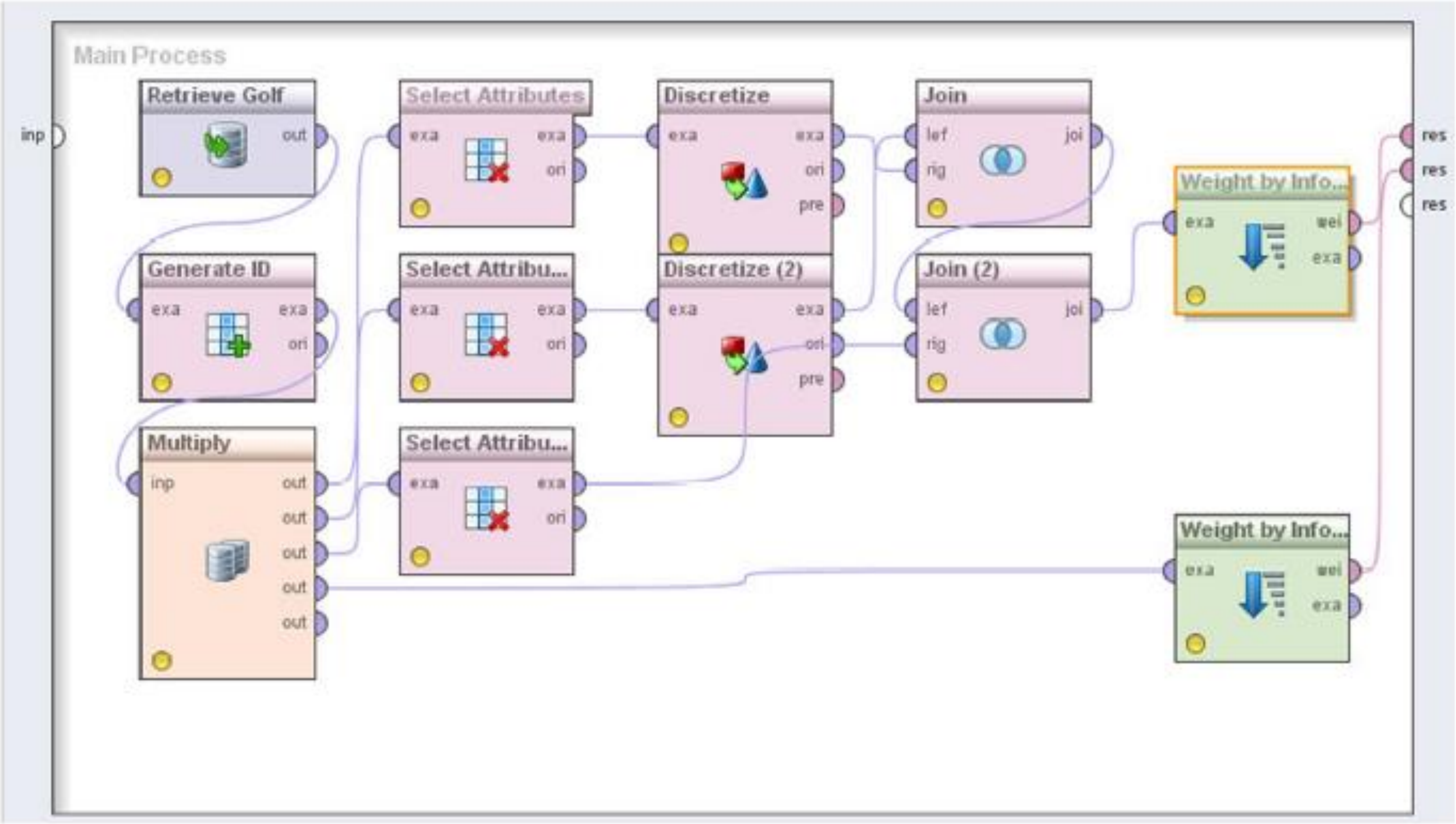
Computing the Information Gain for All Attributes

Attribute	Information Gain
Temperature	0.029
Humidity	0.102
Wind	0.048
Outlook	0.247

Data Mining Feature Selection
Filter Based : Info Gain : How to Implement

12_Feature_12.3_infoGain.rmp

Implemented in RapidMiner by the **Weight by Information Gain operator**

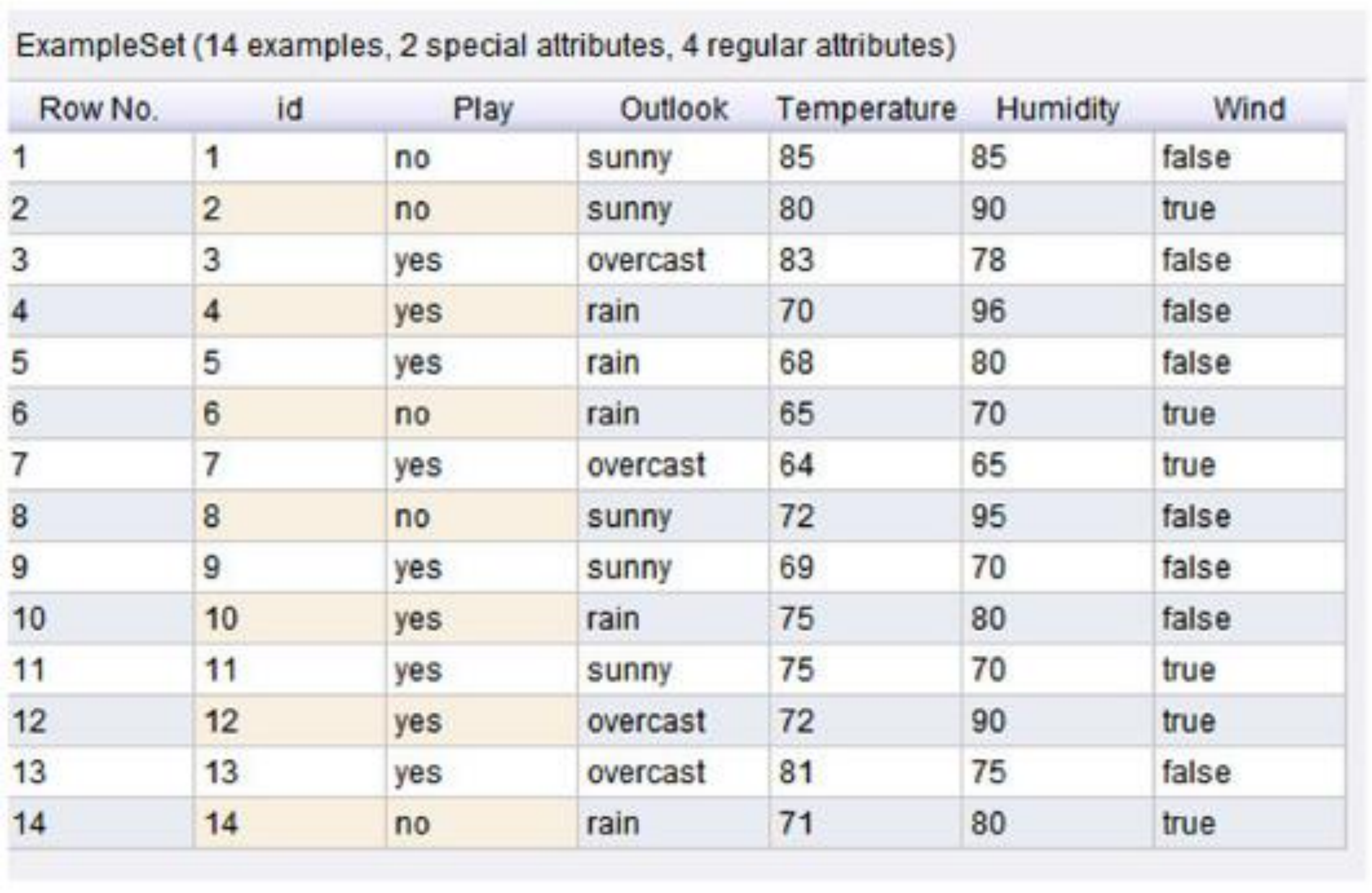


Process to discretize the numeric Golf data set before running information gain–based feature selection.

Data Mining Feature Selection
Filter Based : Info Gain : How to Implement cont.

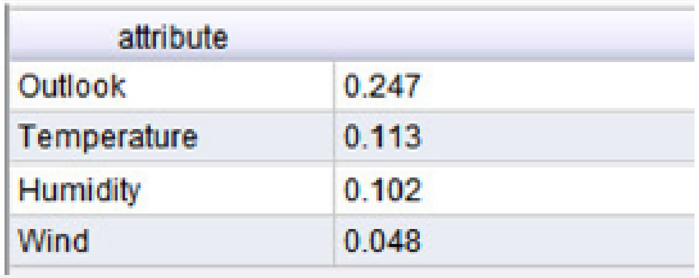
12_Feature_12.3_infoGain.rmp

Results of applying Information Gain in RapidMiner



ExampleSet (14 examples, 2 special attributes, 4 regular attributes)

Row No.	id	Play	Outlook	Temperature	Humidity	Wind
1	1	no	sunny	85	85	false
2	2	no	sunny	80	90	true
3	3	yes	overcast	83	78	false
4	4	yes	rain	70	96	false
5	5	yes	rain	68	80	false
6	6	no	rain	65	70	true
7	7	yes	overcast	64	65	true
8	8	no	sunny	72	95	false
9	9	yes	sunny	69	70	false
10	10	yes	rain	75	80	false
11	11	yes	sunny	75	70	true
12	12	yes	overcast	72	90	true
13	13	yes	overcast	81	75	false
14	14	no	rain	71	80	true

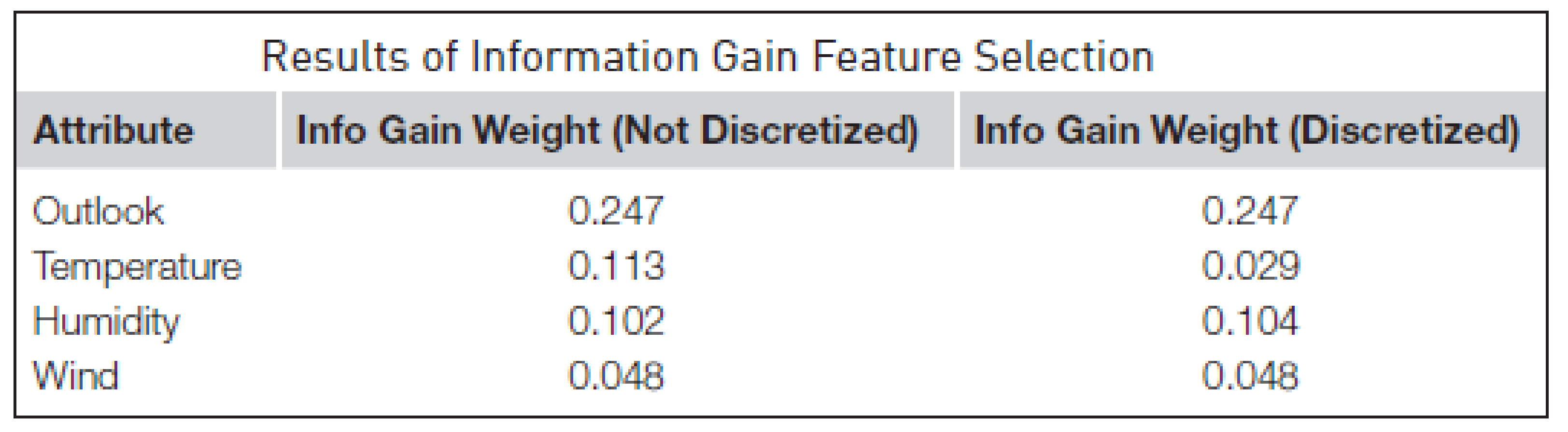


attribute	
Outlook	0.247
Temperature	0.113
Humidity	0.102
Wind	0.048

Results of information gain–based feature selection.

Data Mining Feature Selection
Filter Based : Info Gain : Discussion Points

- Not scale sensitive
- Can work with both numerical and non-numerical data sets
- Different results obtained for discretised and non-discretised data. Attributes Temperature and Humidity can be discretised. For example discretising the numeric value for Humidity into three bands (high, medium, low).



Results of Information Gain Feature Selection

Attribute	Info Gain Weight (Not Discretized)	Info Gain Weight (Discretized)
Outlook	0.247	0.247
Temperature	0.113	0.029
Humidity	0.102	0.104
Wind	0.048	0.048

Clustering Models
Filter Based : Info Gain : Summary

- **Model**
  Similar to decision tree model

- **Input**
  No restrictions on variable type for predictors

- **Output**
  Data sets require a label. Can only be applied on data sets with nominal label

- **Pros**
  Same as decision trees. Intuitive to explain to nontechnical business users. Normalizing predictors is not necessary

- **Cons**
  Same as decision trees. Tends to over-fit the data. Small changes in input can yield substantially different trees. Selecting the right parameters can be challenging

- **Use Cases**
  Applications for feature selection where target variable is categorical or numeric

Data Mining Feature Selection
Filter Based : Chi-Square

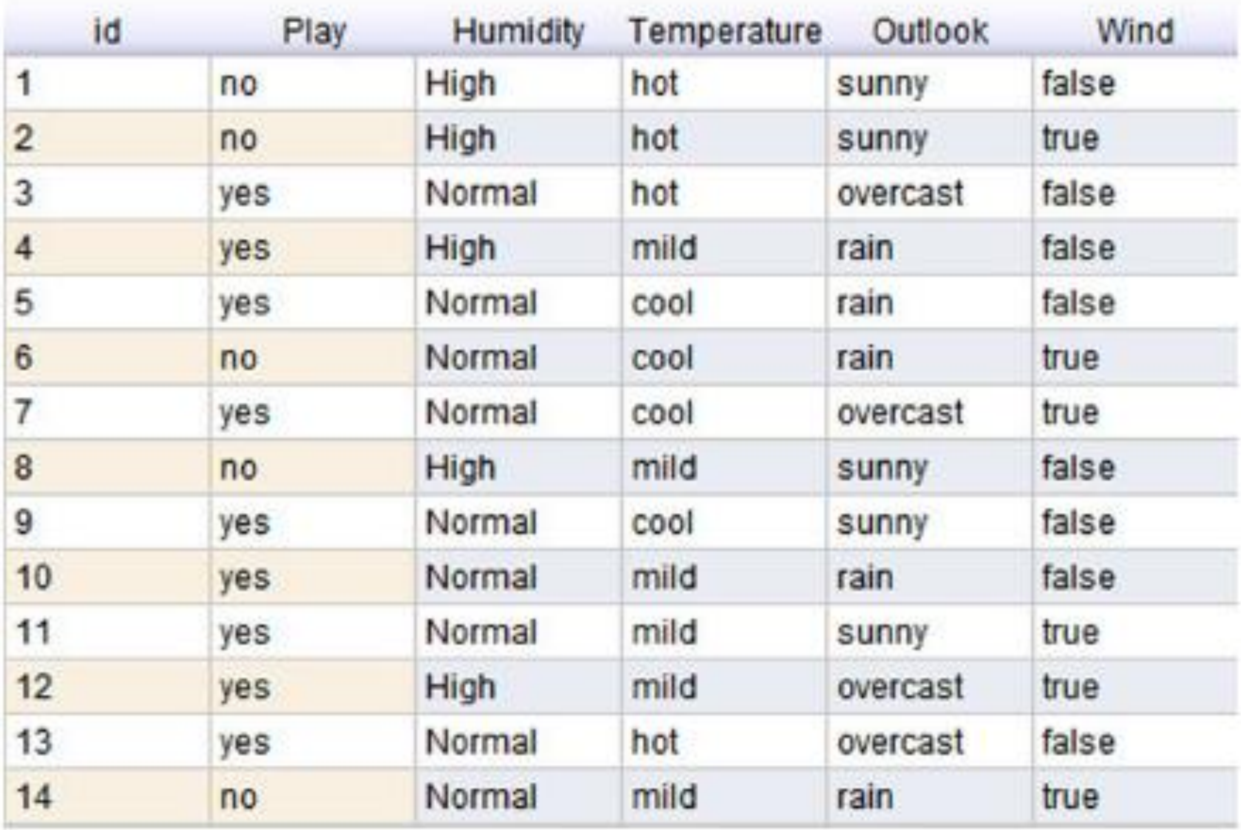
Description Selecting attributes based on relevance to the target or label

Chi-square-based filtering can be used as a means to distinguish between high influence attributes and low or no influence categorical (or nominal) attributes. For example are men or women the primary decision makers when it comes to purchasing big-ticket items?

The chi-square test checks if the frequencies of occurrences across any pair of attributes, such as Outlook = overcast and Play = yes are correlated

Data Mining Feature Selection
Filter Based : Chi-Squared : How it Works

Golf data set all attributes converted to nominal.

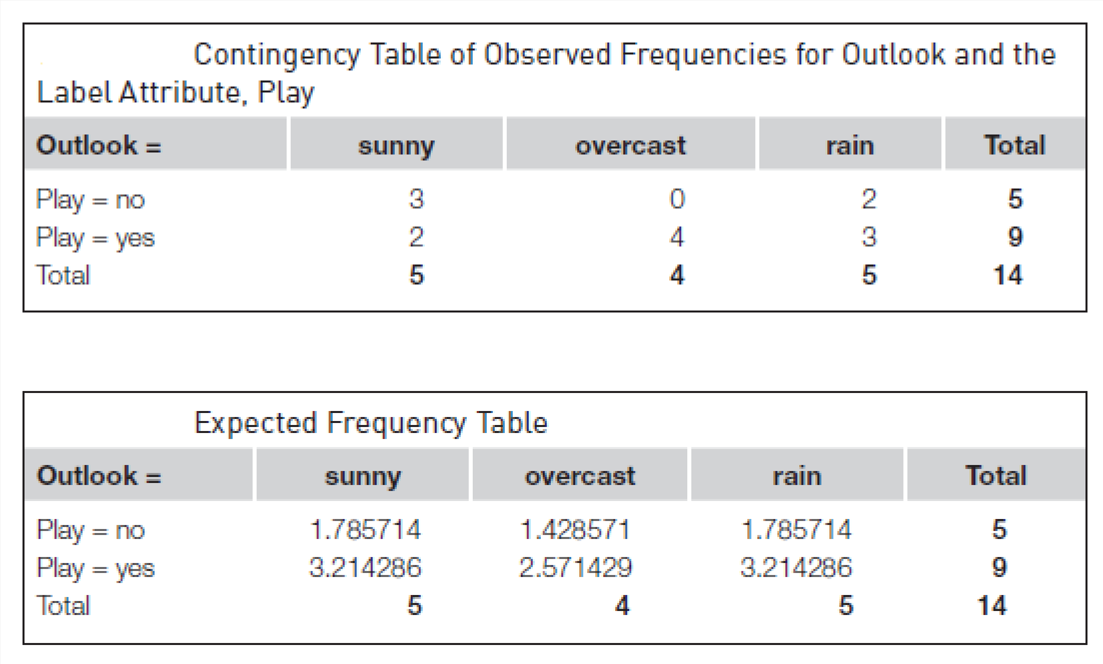


id	Play	Humidity	Temperature	Outlook	Wind
1	no	High	hot	sunny	false
2	no	High	hot	sunny	true
3	yes	Normal	hot	overcast	false
4	yes	High	mild	rain	false
5	yes	Normal	cool	rain	false
6	no	Normal	cool	rain	true
7	yes	Normal	cool	overcast	true
8	no	High	mild	sunny	false
9	yes	Normal	cool	sunny	false
10	yes	Normal	mild	rain	false
11	yes	Normal	mild	sunny	true
12	yes	High	mild	overcast	true
13	yes	Normal	hot	overcast	false
14	no	Normal	mild	rain	true

Converting the golf example set into nominal values for chi-square feature selection.

Data Mining Feature Selection
Filter Based : Chi-Squared : How it Works cont.

Expected frequency given by P(A) * P(B) * N where N is the sum of all occurrences in the data set. For example Expected frequency for the event [Play=No and Outlook=Sunny] = (5/14 * 5/14 * 14) = 1.785



Contingency Table of Observed Frequencies for Outlook and the Label Attribute, Play

Outlook =	sunny	overcast	rain	Total
Play = no	3	0	2	**5**
Play = yes	2	4	3	**9**
Total	**5**	**4**	**5**	**14**

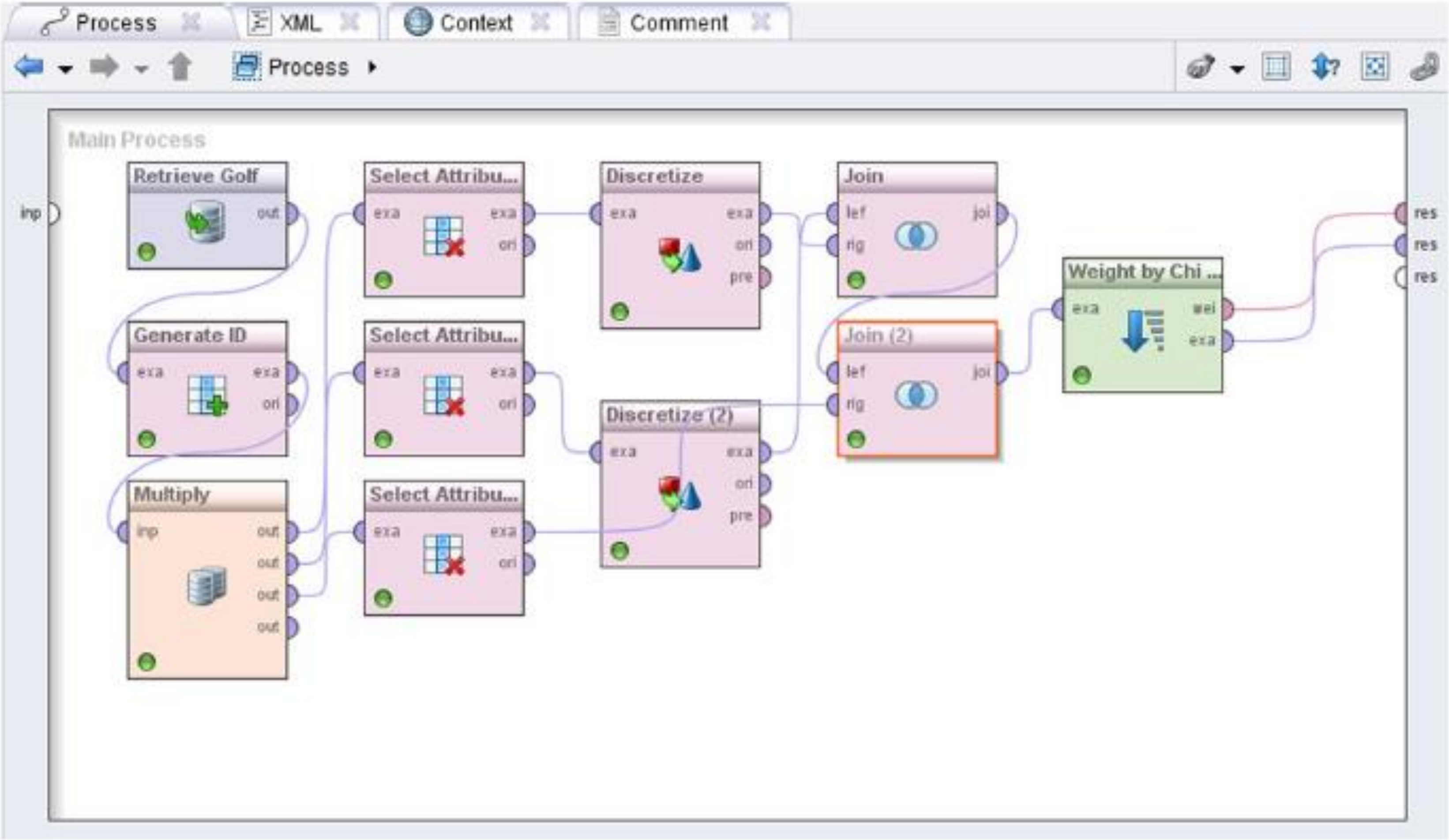
Expected Frequency Table

Outlook =	sunny	overcast	rain	Total
Play = no	1.785714	1.428571	1.785714	**5**
Play = yes	3.214286	2.571429	3.214286	**9**
Total	**5**	**4**	**5**	**14**

The chi-square test statistic is computed by summing the difference between the observed frequency and the expected frequency for each attribute. The test of independence between any two parameters is done by checking if the observed chi-square is less than a critical value dependent on a user selected confidence level. For feature selection the chi-square values are used to rank the attributes.

Data Mining Feature Selection
Filter Based : Chi-Squared : How to Implement

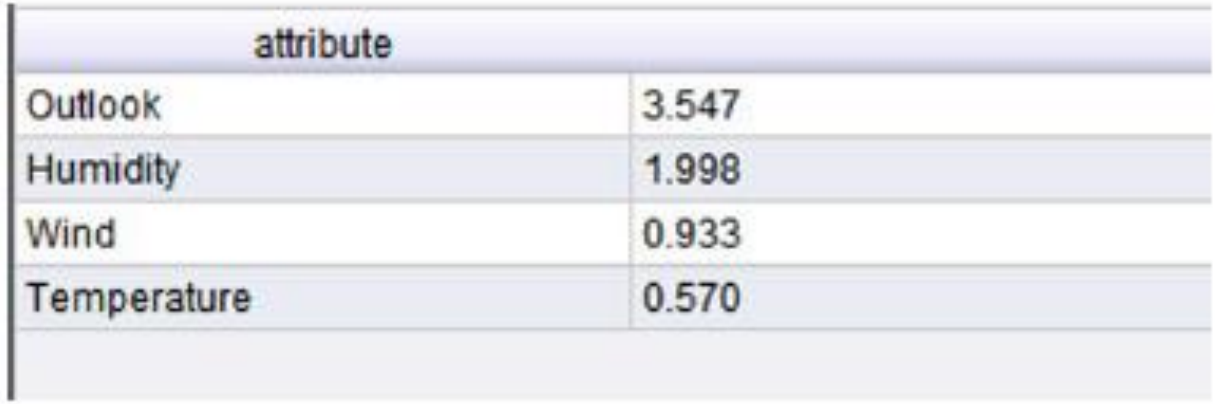
12_Feature_12.4_ChiSq.rmp



Process to rank attributes of the Golf data set by the chi-square statistic.

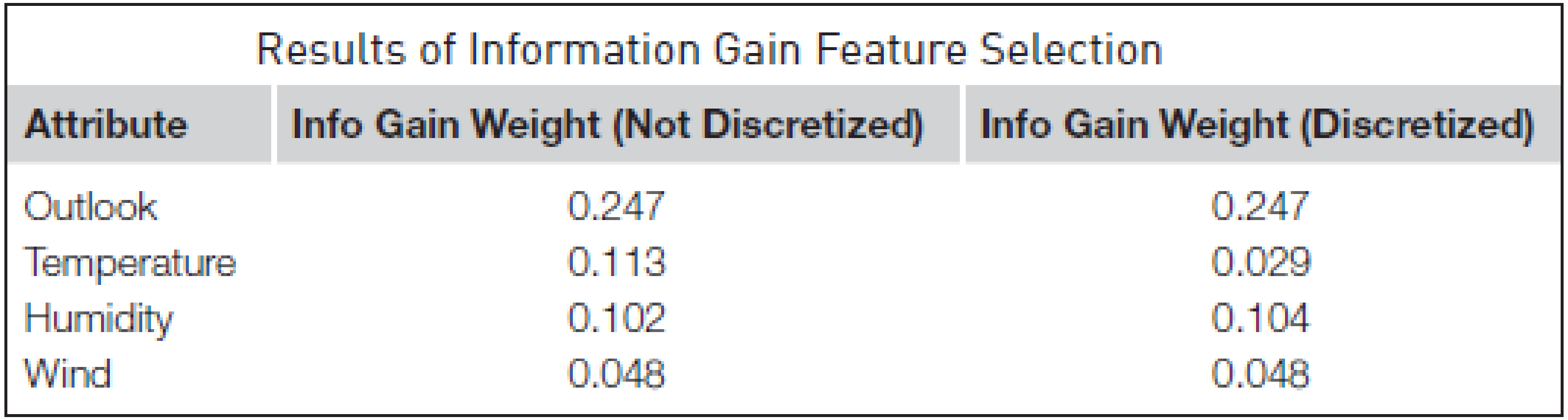
Data Mining Feature Selection
Filter Based : Chi-Squared : How to Implement cont.

Result ranking is the same as for Information Gain Weight (Discretized)



attribute	
Outlook	3.547
Humidity	1.998
Wind	0.933
Temperature	0.570

Results of the attribute weighting by the chi-square method.



Results of Information Gain Feature Selection

Attribute	Info Gain Weight (Not Discretized)	Info Gain Weight (Discretized)
Outlook	0.247	0.247
Temperature	0.113	0.029
Humidity	0.102	0.104
Wind	0.048	0.048

Clustering Models
Filter Based : Chi-Square : Summary

- **Model**
  Uses the chi-square test of independence to relate predictors to label

- **Input**
  Categorical (polynomial attributes)

- **Output**
  Data sets require a label. Can only be applied on data sets with a nominal label

- **Pros**
  Very robust. A fast and efficient scheme to identify which categorical variables to select for a predictive model

- **Cons**
  Sometimes difficult to interpret

- **Use Cases**
  Applications for feature selection where all variables are categorical

Data Mining Feature Selection
Wrapper Based : Forward Selection

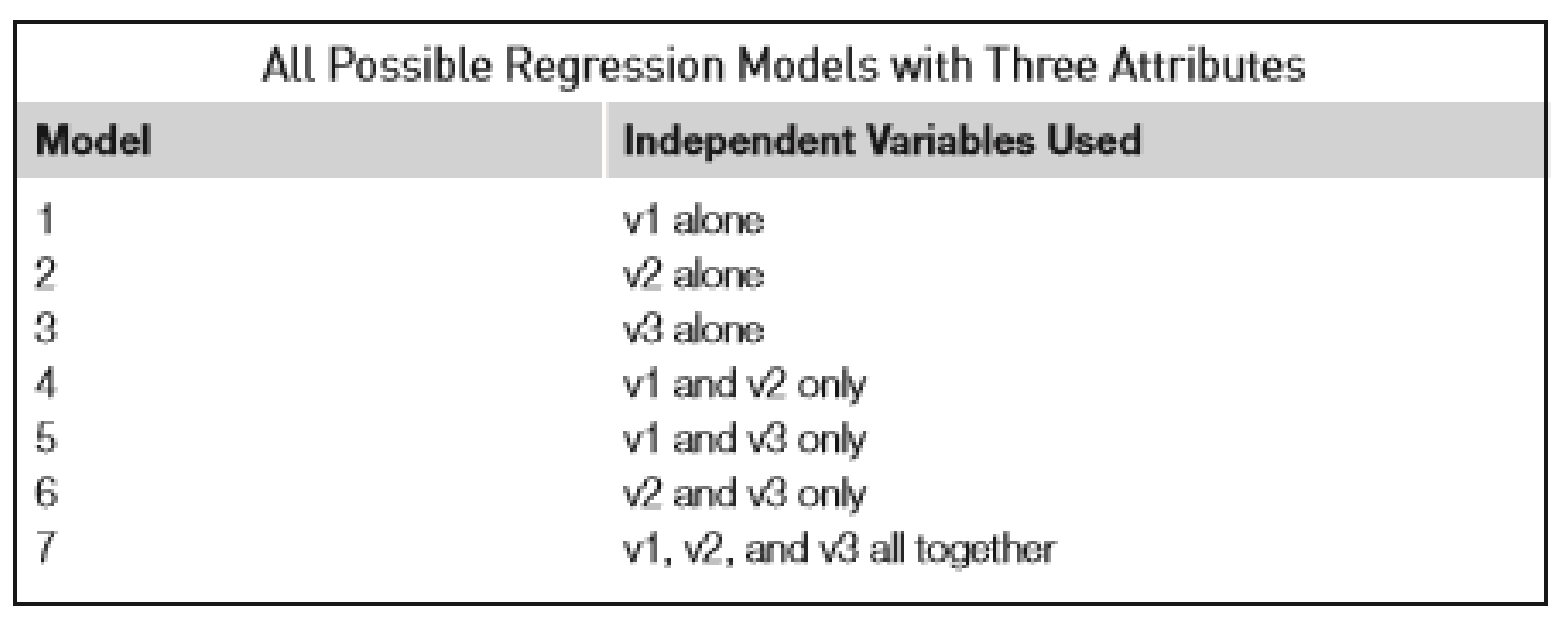
Description Selecting attributes based on relevance to the target or label

In the case of building a regression model, in general if a dataset contains k different attributes, then conducting all possible regression searches implies that we build $2^k - 1$ separate regression models and pick the one that has the best performance. Cleary this is impractical.

The Wrapper approach iteratively chooses features to add or to remove from the current attribute pool based on whether the newly added or removed attribute improves the accuracy.

Forward selection starts with one variable and builds a baseline model, then adds a second variable and builds a new model to compare with the baseline. If the performance is better it proceeds in the same manner with a third variable. If the second model did not improve on the baseline, then a new model is built with the first and third variables to compare against the baseline. The process continues similarly until a required level of model performance is achieved.

Data Mining Feature Selection
Wrapper Based : Forward Selection



All Possible Regression Models with Three Attributes

Model	Independent Variables Used
1	v1 alone
2	v2 alone
3	v3 alone
4	v1 and v2 only
5	v1 and v3 only
6	v2 and v3 only
7	v1, v2, and v3 all together

Wrapper Based : Forward selection : Summary

- **Model**
  Works in conjunction with modelling methods such as regression

- **Input**
  All attributes should be numeric

- **Output**
  The label may be numeric or binominal

- **Pros**
  Multicollinearity problems can be avoided. Speeds up the training phase of the modelling process

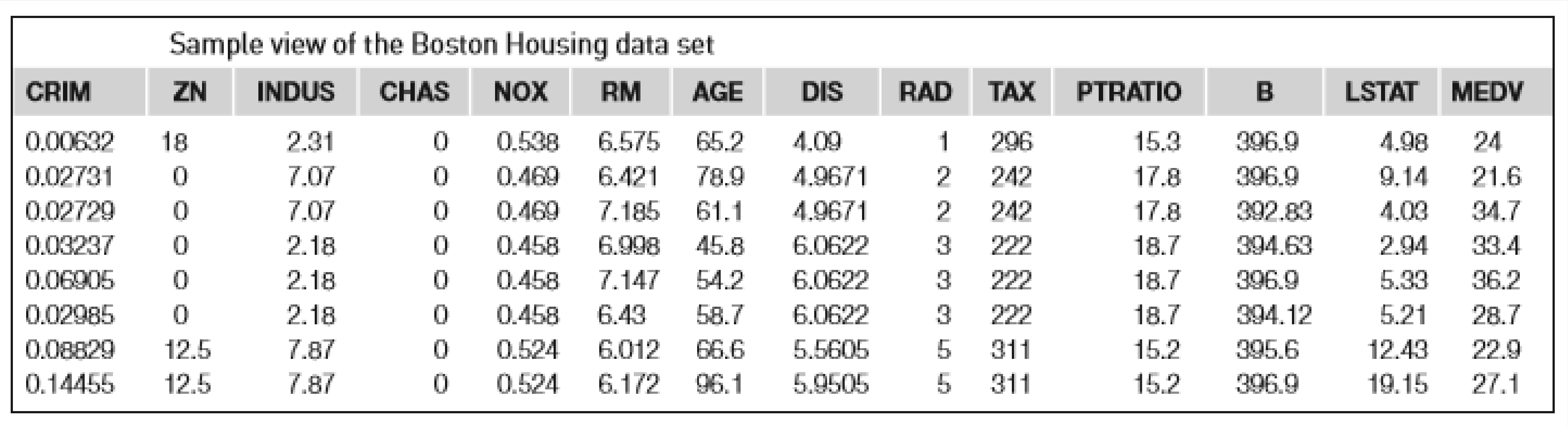
- **Cons**
  Once a variable is added to the set, it is never removed in subsequent iterations even if its influence on the target diminishes

- **Use Cases**
  Data sets with a large number of input variables where feature selection is required

Data Mining Feature Selection
Wrapper Based : Backward Elimination

Description Selecting attributes based on relevance to the target or label

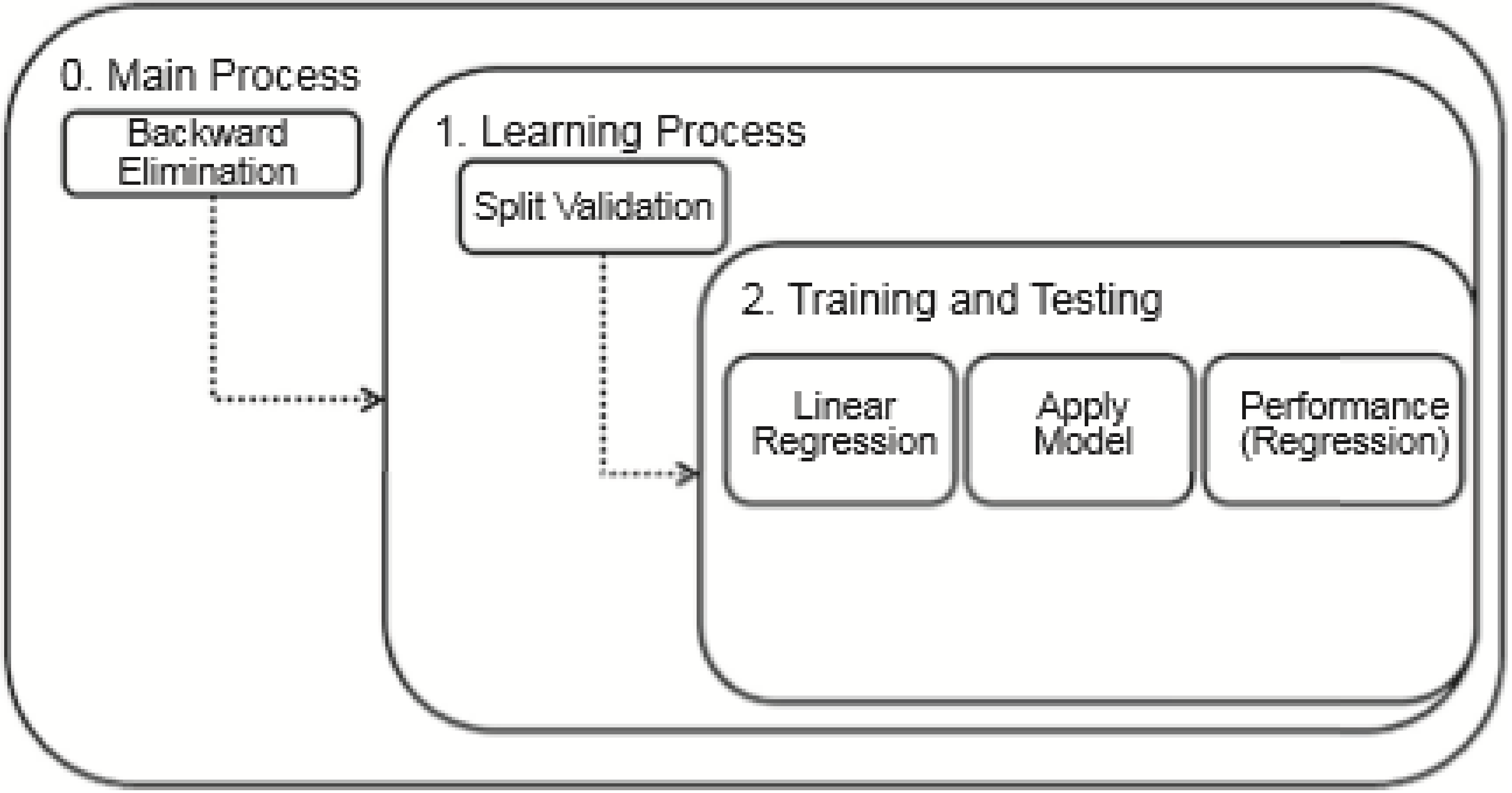
The goal is to build a high quality multiple regression model that includes as few attributes as possible without compromising the predictive ability of the model



Sample view of the Boston Housing data set

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1

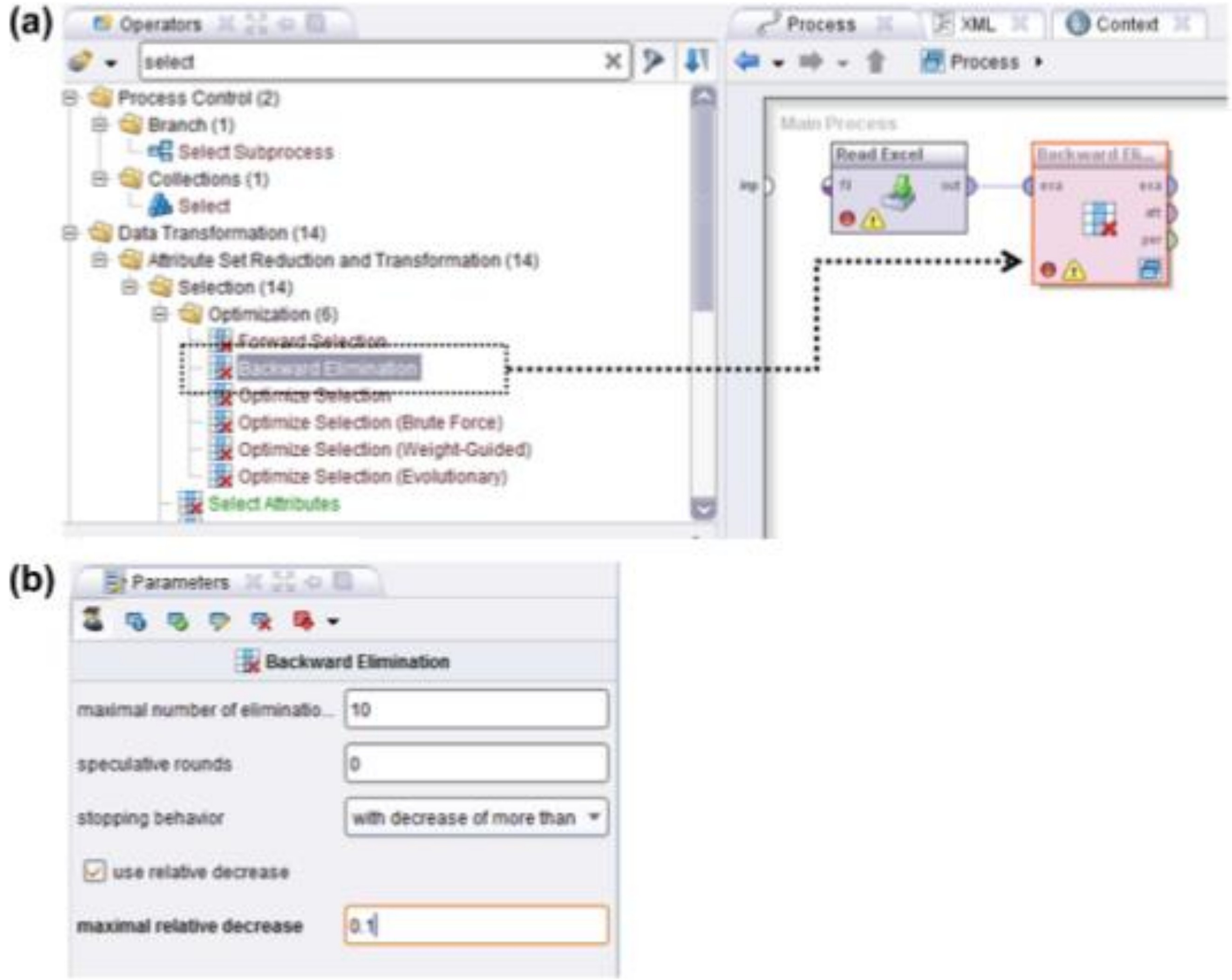
Wrapper Based : Backward Elimination : Process



Wrapper function logic used by RapidMiner.

Wrapper Based : Backward Elimination : Process

12_Feature_12.5_backElim.rmp



maximal number of eliminatio... 10

speculative rounds 0

stopping behavior with decrease of more than

☑ use relative decrease

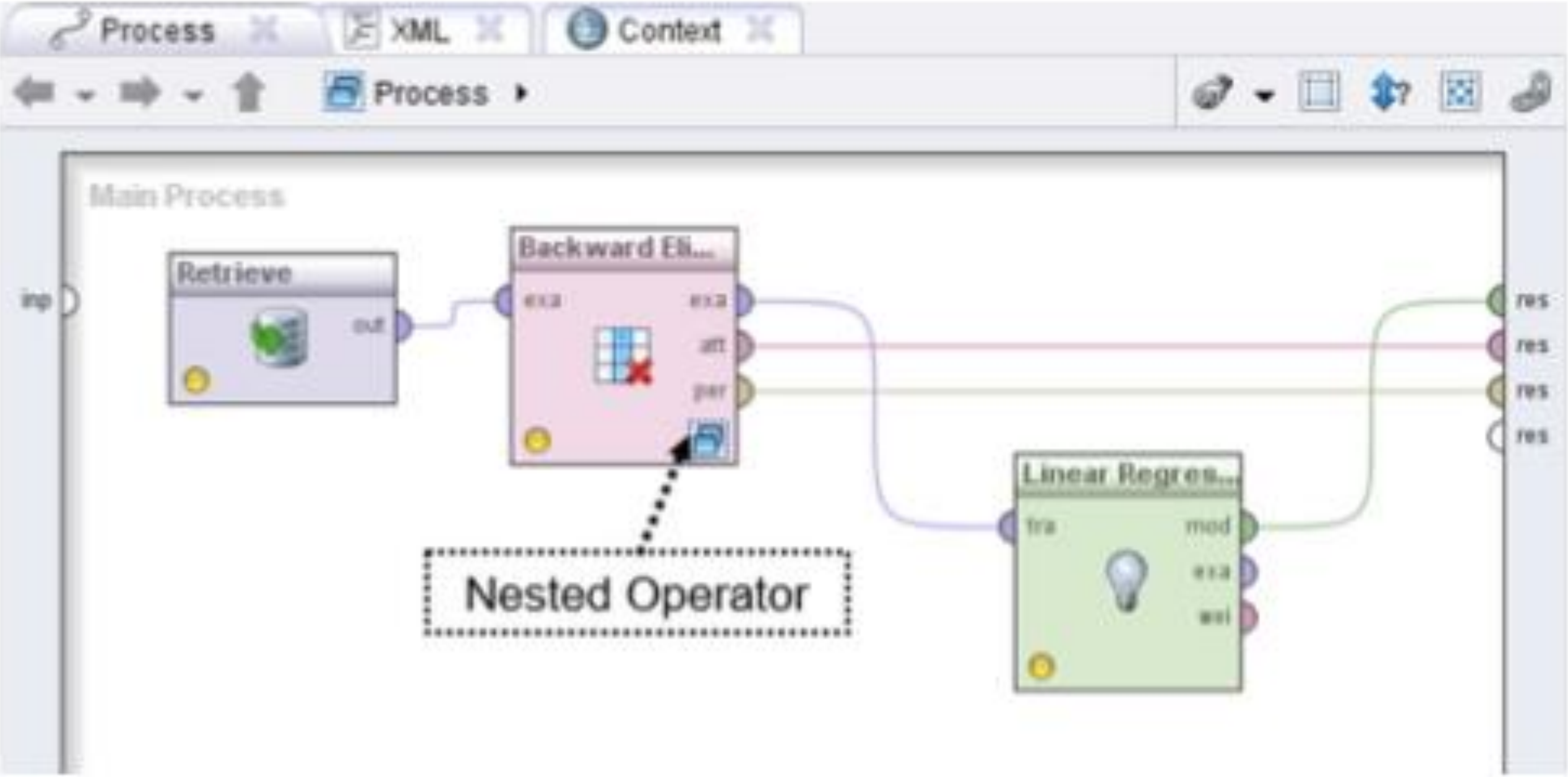
maximal relative decrease 0.1

Configuring the Backward Elimination operator. a) Selecting the Backward Elimination nested operator and b) configuring the parameters.

Wrapper Based : Backward Elimination : Process

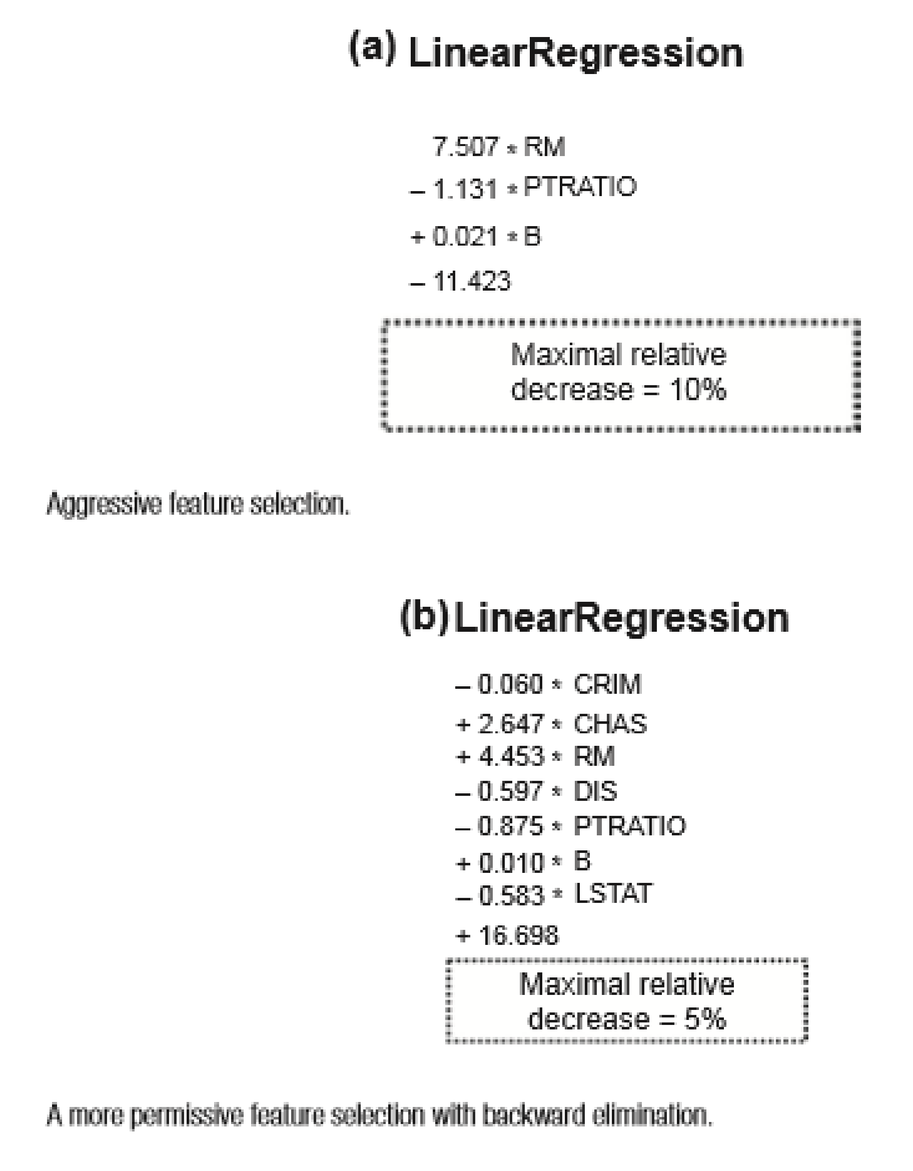
Now let us look at the configuration of the Backward Elimination operator. Here we can specify several parameters to enable feature selection. The most important one is the “stopping behaviour.” Our choices are “with decrease,” “with decrease of more than,” and “with significant decrease.” The first choice is very parsimonious—a decrease from one iteration to the next will stop the process. But if we pick the second choice, we have to now indicate a “maximal relative decrease.” In this example, we have indicated a 10% decrease. Finally, the third choice is very stringent and requires achieving some desired statistical significance by allowing you to specify an alpha level. But we have not said by how much the performance parameter should decrease yet! This is specified “deep inside” the nesting: all the way at the Performance operator that was selected in the Testing window of the Split Validation operator. In this example, the performance criterion was “squared correlation.”

Wrapper Based : Backward Elimination : Process



Final setup of the backward elimination wrapper process.

Wrapper Based : Backward Elimination : Results



(a) LinearRegression

7.507 * RM
– 1.131 * PTRATIO
+ 0.021 * B
– 11.423

Maximal relative decrease = 10%

Aggressive feature selection.

(b) LinearRegression

– 0.060 * CRIM
+ 2.647 * CHAS
+ 4.453 * RM
– 0.597 * DIS
– 0.875 * PTRATIO
+ 0.010 * B
– 0.583 * LSTAT
+ 16.698

Maximal relative decrease = 5%

A more permissive feature selection with backward elimination.

Wrapper Based : Backward Elimination : Results

Comparing the two regression equations we can see that nine attributes have been eliminated. Perhaps the 10% decrease was too aggressive. As it happens, the R^2 for the final model with only three attributes was only 0.678. If we were to change the stopping criterion to a 5% decrease, we will end up with an R^2 of 0.812 and now have 8 of the 13 original attributes. You can also see that the regression coefficients for the two models are different as well. The final judgment on what is the right criterion and its level can only be made with experience with the data set and of course, good domain knowledge. Each iteration using a regression model either removes or introduces a variable, which improves model performance. The iterations stop when a pre-set stopping criterion or no change in performance criterion (such as adjusted R^2 error) is reached. The inherent advantage of wrapper-type methods are that multicollinearity issues are automatically handled. However, you get no prior knowledge about the actual relationship between the variables.

Wrapper Based : Backward Elimination : Summary

- **Model**
  Works in conjunction with modelling methods such as regression

- **Input**
  All attributes should be numeric

- **Output**
  The label may be numeric or binominal

- **Pros**
  Multicollinearity problems can be avoided. Speeds up the training phase of the modelling process

- **Cons**
  Need to begin with a full model which can sometimes be computationally intensive

- **Use Cases**
  Data sets with a large number of input variables where feature selection is required

Feature Selection Discussion Points

Feature selection or dimension reduction is a very important paradigm in data mining. A central hypothesis among all the feature selection methods is that good feature selection results in attributes or features that are highly correlated with the class, yet uncorrelated with each other. Dimension reduction is best understood with real practice. The same technique can yield quite different results based on the selection of analysis parameters. This is where data visualization can play an important role. Sometimes, examining a correlation plot between the various attributes, like in a scatterplot matrix, can provide valuable clues about which attributes are likely redundant and which ones can be strong predictors of the label variable. While there is usually no substitute for domain knowledge, sometimes data is simply too large or mechanisms are unknown. This is where feature selection can prove very useful.