DMR - module Data Mining Recipes de Statistica

DMR - module Data Mining Recipes de Statistica

Analyse statistique traditionnelle

- Départ = obtenir réponse questions avec données à recueillir, expériences, sondages,.. *(Small Data)*
- Estimation paramètres modèles postulés
- Ajustement : moindres carrés, modèles parcimonieux
- Méthodes: régression, contrôle qualité, planification d'expériences (DOE), analyse variance (ANOVA), PCA, GLM, GLMZ, séries chronologiques, PLS, …
- Validation : Tests d'hypothèses

VERSUS

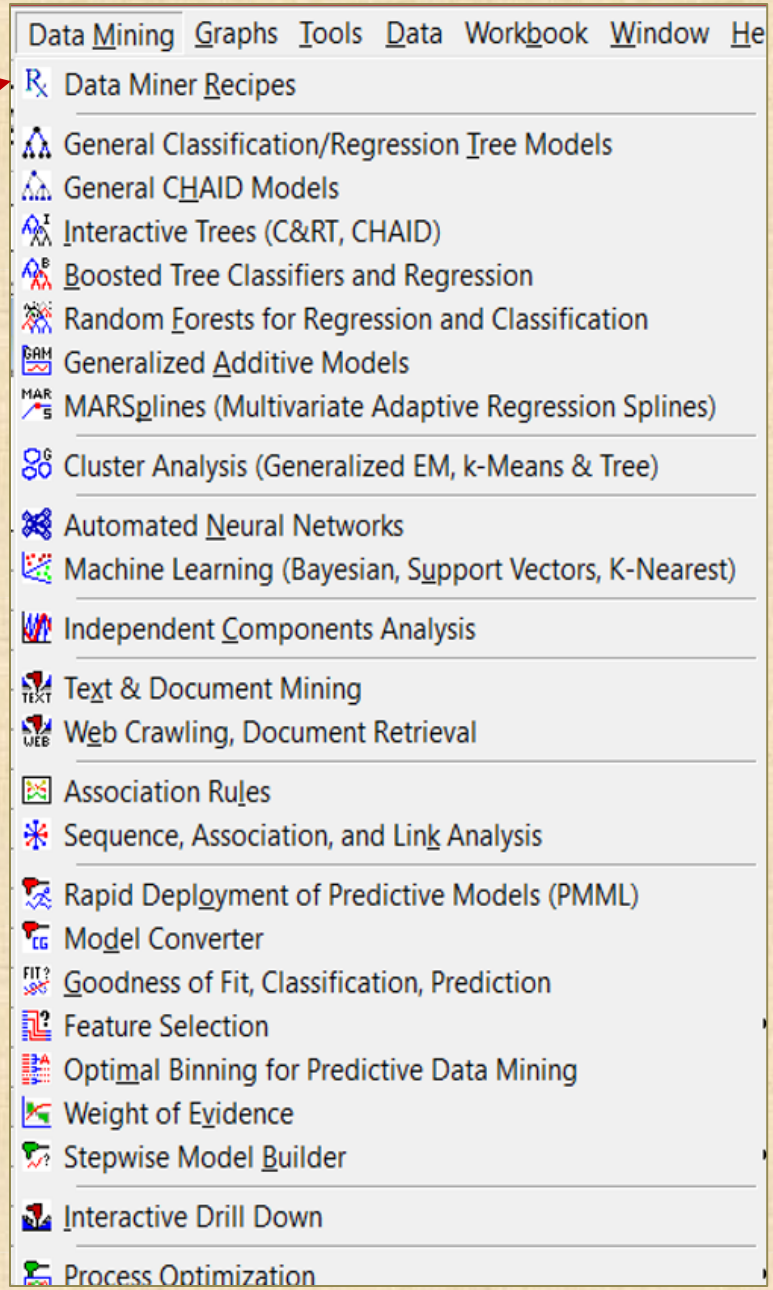
Data Mining - Apprentissage Statistique – Machine Learning

- ✓ But: découverte connaissances, détection patterns,…
- ✓ Au départ: - départ = *données nombreuses existantes (Big Data)*
- ✓ - pas d'hypothèses à priori
- ✓ Algorithmes: régression MARS, réseaux neurones, CRT, forêts
- ✓ aléatoires, machines à support vectoriels (SVM), réseaux
- ✓ bayésiens, text mining … liste 24 analyses (page suivante)
- ✓ Validation : séparation des données, critères performance, ...

Modules du Data Mining de Statistica
Procédure Data Miner Recipes

Data Mining 24 procédures

- C&RT (arbres classification)
- MARS (regression splines)
- Support Vectors Machine (SVM)
- Cluster Analysis (k-Means)
- ANN (réseaux neurones)
- Machine Learning
- Text & Document Mining
- Rapid Deployment
- Model Converter
- Goodness of Fit, Classification, prediction
- Stepwise Model Builder
- Process Optimization



DMR – module Data Mining Recipes de Statistica

STATISTICA Data Miner Recipes (DMR) provide a systematic method for building advanced analytic models to relate one or more target (dependent) quantities to a number of input (independent) predictor variables. The target variables can be continuous or categorical. Continuous target variables usually associated with regression tasks, categorical variables are used in classification problems.
DMR is capable of building predictive models for regression and classification problems.

But

- Simplifier interface usager
- Simplifier processus Data Mining / ML
- Rapidité
- Intégrer avec la plateforme Statistica
- Faciliter évaluation nouvelles données
- Déploiement solution entreprise

ÉTAPES

1. Data preparation
2. Data analysis
3. Data redundancy
4. Dimension reduction
5. Model building
6. Model evaluation deployment

Statistica Data Miner: ÉTAPES

1. Select the desired option on the Data Mining tab (ribbon bar)
   or the Data Mining menu (classic menus)
   Select a data source - Select the variables for the analyses
2. Display the Node Browser and select the desired analyses
   or data management operation
3. Run (update) the data miner project
4. Customize analyses, edit results, save results
5. Deploy solution (models) for new data
6. Prepare project for final customer deployment (in the "field")

Simplifier et automatiser le Data Mining

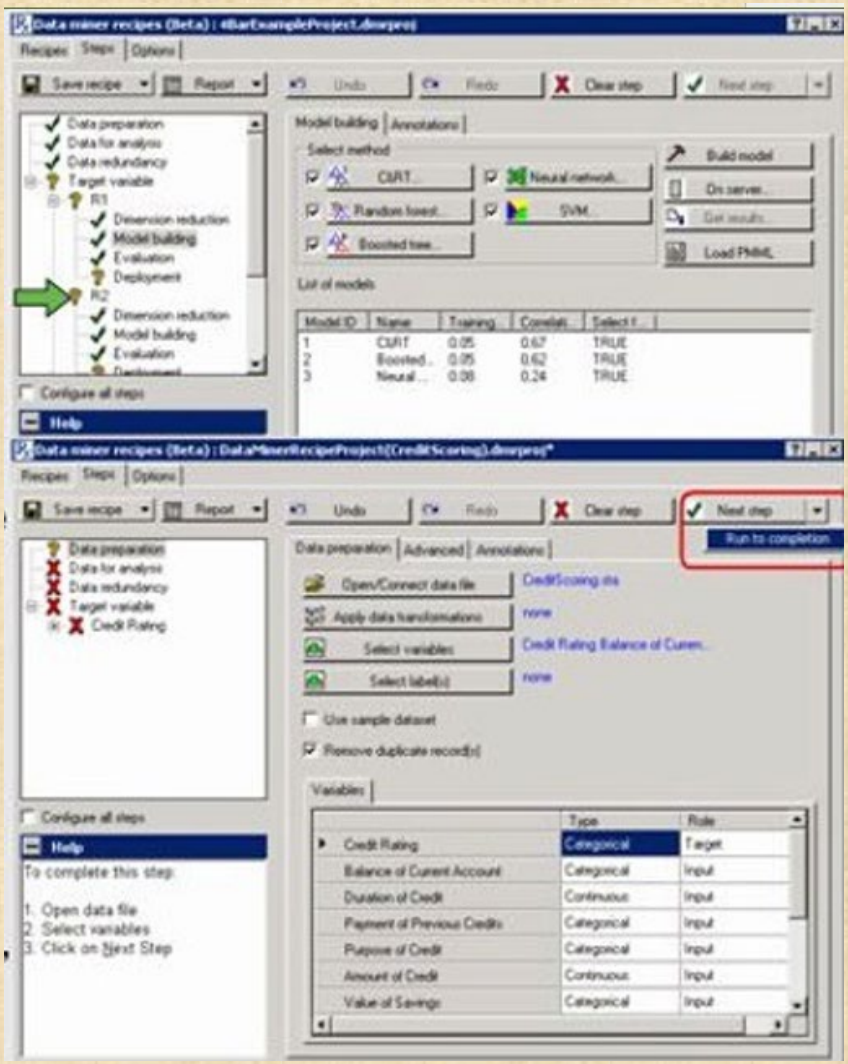
Métaphore ***recipes*** **(recettes)**

- Il n'y a qu'un **nombre limité** de recettes en 2 grandes catégories d'analyse:

- **SUPERVISÉE**
  - ✓ modèles **prédictifs** pour une variable de réponse continue … **REGRESSION**
  - ✓ modèles **prédictifs** pour une variable de réponse catégorique … **CLASSIFICATION**

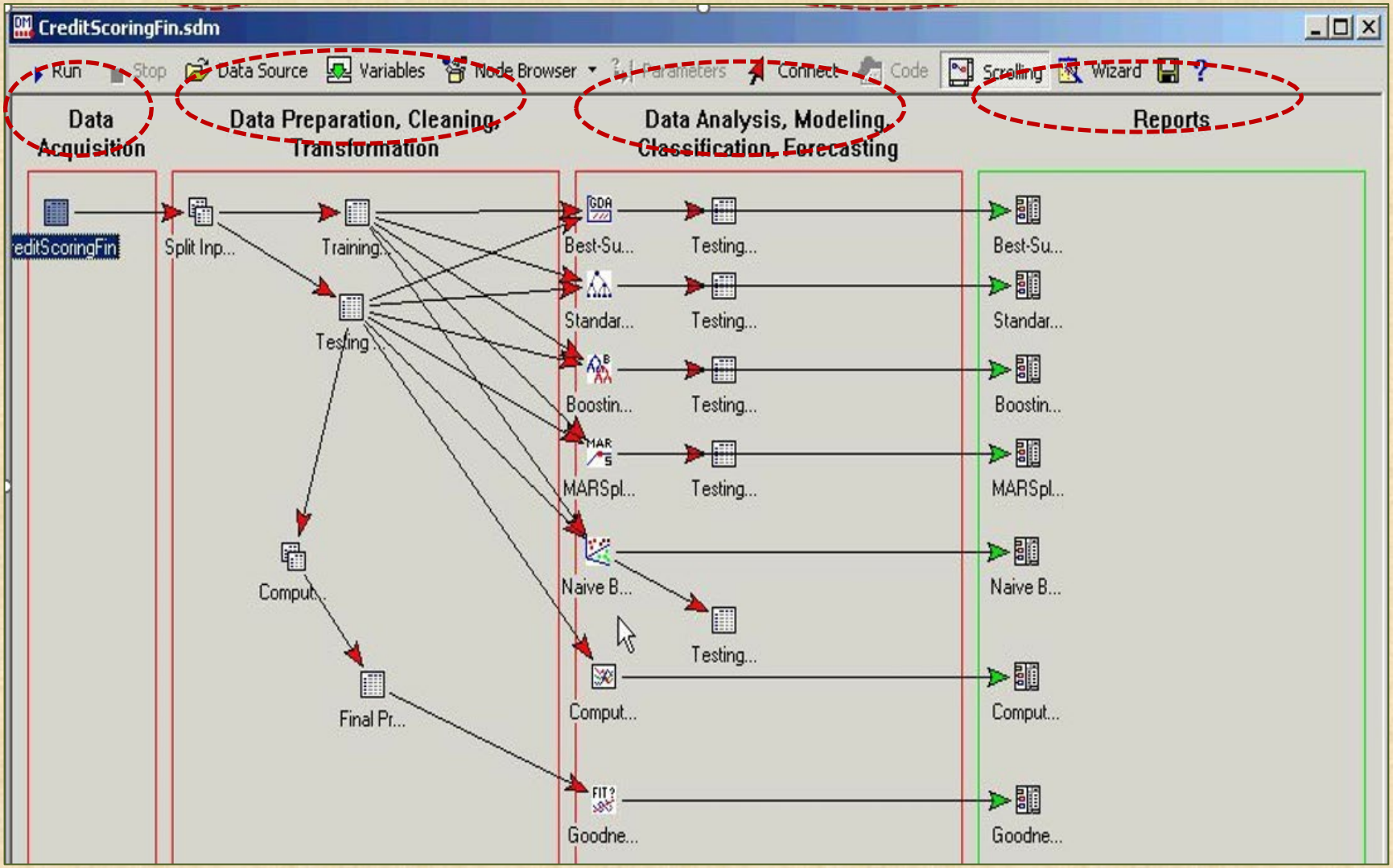
- **NON SUPERVISÉE**
  clustering: k-Means , SVM,…

- Simplifier l'utilisation de **méthodes avancées** pour résoudre des problèmes



Espace de travail - Statistica Workspace .sdm

ne sera pas utilisé dans cette présentation et le cours MTH8302



Data Miner Recipe DMR – quelques règles utiles -1

référence : Statistica, Wikipedia

Large numbers of categories and partitioning into training and testing data.

The effectiveness of DMR can deteriorate when one or more inputs are categorical in nature and **contain many categories**. For example, in applications with credit-scoring data that contain many categorical predictors [i.e., home-town (ZIP-codes), types of previous loans issued, etc.] it has been found that **careful data pre-processing may greatly enhance the effectiveness of the DMR methodology**. Otherwise, the predictive model may end up with too many input variables, which can deteriorate the performance through the curse of dimensionality

curse of dimensionality fléau de la dimension voir https://fr.wikipedia.org/wiki/Fl%C3%A9au_de_la_dimension

Le **fléau de la dimension** ou **malédiction de la dimension** (*curse of dimensionality*) est un terme inventé par Richard Bellman en 1961 pour désigner divers phénomènes qui ont lieu lorsque l'on cherche à analyser ou organiser des données dans des espaces de **grande dimension** alors qu'ils n'ont pas lieu dans des espaces de dimension moindre.

The inclusion of an input in a predictive modeling algorithm, such as neural networks for example, adds another dimension to the space in which the data cases reside. The more inputs a neural network has, the more data points we need to train the network effectively so it can capture the underlying structure in the data (i.e., to model the relationship between the input-target variables). Thus, with the addition of every input, the number of data points needed to train the network will grow rapidly.

issue of too many categorical variables we recommend that you use the various methods available in STATISTICA Data Miner for **carefully pre-processing your data, e.g., to combine categories** while maximizing the ”possible relationships” to the target(s) of interest.

In general, whenever possible the data analyst should **perform an initial "common-sense" screening** of (large numbers of) inputs, and not include "obviously unnecessary" inputs. In addition, one may even try to identify those inputs that carry a small amount of information regarding the prediction problem and eliminate them from the analysis. Although this may lead to some loss of information in the data, it may, on the contrary, considerably reduce the cure of dimensionality, which can significantly increase the performance of the neural network.

DMR employs two methods for combating the curse of dimensionality. In the **data redundancy stage (**for more information on DMR stages DMR identifies those **variable pairs that contain similar information (i.e., correlated variables), and eliminates one of them**, thereby reducing the dimensionality of the problem. Further dimensionality reduction is applied in the dimension reduction step, where DMR can use a tree-based algorithm (as well as the much faster single-pass predictor screening algorithm) for identifying and eliminating those variables that contain little or no relevant information about the target variables.

Data Miner Recipe DMR – quelques règles utiles - 2

référence : Statistica, Wikipedia

Measuring the Wrong Inputs

So, what if you could not find any valid and accurate predictive models using STATISTICA DMR (e.g., in the model building step, the correlations between observed and predicted target values for the hold-out or validation sample are less than .4 or so)?
You should continue to try and retry to find ”good models” using different settings in the model building step.
If you continue to fail in finding ”good models,” the inevitable conclusion should be either:
case 1 There is a large amount of noise on the target variables, which masks the real signal **(i.e., small signal to noise ratio).**
case 2 There is no strong relation between the input and target variables.

In case 1, you should refine the data set by including more accurate measurements.
In case 2, you should ask the question: Why are you measuring these inputs in the first place?
If you cannot find accurate models for the given inputs and targets even though you have applied the most general and advanced neural networks methods for building predictive models, then the inputs are not relevant for the targets. If the target variable of interest is related to important outcomes for your business, e.g., product quality, credit risk, etc., then it follows that none of the input parameters (i.e., inputs) currently measured and stored to describe the process are relevant for the quality of the outputs from the process.

This outcome may occur and is in itself an interesting result. If the inputs (process parameters) are not relevant to outcomes (targets), then why measure (or buy information on) them in the first place? Not measuring the right inputs is a serious problem since it means that you are not ”looking at the right things” in order to improve quality.

When the model building step fails to produce accurate models, ask yourself whether the current measurement system used to describe and ”track” and predict your processes is useful. Of course, there is always the possibility that the predictive data mining algorithms fail to detect the extremely complex relationships that connect the inputs to the targets.

If you fail to find a good predictive model after letting DMR search through a substantial amount of potential models and methods of high complexity, there is a fair chance that the measurement and data collection system should be reconsidered.

Data Miner Recipe DMR – quelques règles utiles - 3

référence : Statistica, Wikipedia

Data Preparation: Missing Data, Outliers, Overall Data Quality

Another issue that may foil the successful implementation of STATISTICA DMR is that the data available for modeling are ”buggy” and not ”reliable.”

There are simple and clear-cut recommendations to remedy poor-quality data.

Again, if the quality of the data is so poor that it cannot be used for model building, then it is likely that it also cannot be used for reporting purposes or to satisfy requirements for regulatory compliance (e.g., FDA 21 CFR Part 11 requirements).

Generally in data analysis, data preparation is the most important activity used to ensure success.

Missing data, outliers. The recipe for model building tools acknowledges that missing data and ”bad measurements” (outliers) typically occur in data and provides options for dealing with these data issues.

Specifically, cases with missing data and outliers can be replaced by means (for the respective continuous inputs), or they can be removed from the data for model building.

Both of these methods work generally well as long as no more than 10% to 15% of the original data are ”modified” in this manner.

If more than 15% of the data cases available for model building have missing or bad data values (outliers), then you should explore the differences between the cases (observations) that are missing and those that are not.

The STATISTICA analytic platform has a large number of general and graphical options to visualize and review whether the missing data or outliers are different (with respect to the target variables of interest or other inputs) from those cases (observations) that have complete and ”good” measurements for all variables. For example, if you find that when a particular input has missing data, the quality (target values) of the respective cases is generally better than that of cases where the input is recorded (not-missing), then this finding clearly deserves further scrutiny.

Data Miner Recipe DMR – quelques règles utiles - 4

référence : Statistica, Wikipedia

Analyzing Very Large Data Sets and Sampling

In some applications, very large data sets are available for model building.
In general, this is a good situation (better than too little data) in that a lot of raw data are available for STATISTICA DMR to use in building accurate and useful predictive models. However, the larger the data set the more processing time is required to build (train) the models. This is especially true of support vector machines and tree algorithms, which can grow in size with the size of the data set.

Why sub-sampling of data is useful. The fact that data for model building in DMR need to be stored in the computer's memory does not, however, pose a problem at all. **In general, it is good practice to sub-sample the cases from a large data set available for modeling and leave multiple hold-out samples to validate the predictive model.**

At first, this issue may not be clear or intuitive. The accuracy and representative nature of statistical estimates from samples depend only on the (reasonable) sample size, not the population from which the sample was drawn.
For example, if you take a sample of 1,000 cases from a large data set and compute a mean for a given variable, you usually get excellent precision (i.e., the confidence bounds for the mean will be very small) regardless of how large the complete data set is.

In other words, the accuracy of your models will usually be just as good when you take a reasonable subset of the (large) data available for modeling. **In addition, doing so then makes the remaining cases available for validation.**

After model building, if the accuracy of the final models is found to be very good in the data that were not used for building models, then you can generally be more confident that the models have good predictive validity.

How many cases (observations) should you have for DMR modeling? That question is difficult to answer.
In general, the number of cases (observations) should be a multiple of the number of inputs that are used for model building (e.g., 10 to 100 times as many cases as variables is sometimes used as a rule of thumb).
The number of inputs that are available for building predictive models may itself be very large (see below) in which case using even fewer cases will still yield accurate models.
This rule of thumb can be used as a general guideline **as long as the overall file size (submitted to, for example, neural networks) is within reasonable limits (e.g., less than 100,000 data points overall).**
Note that STATISTICA DMR is not limited to any particular file size.
Even very large data sets can be analyzed; however, long processing times may be expected in this case.

Data Miner Recipe DMR – quelques règles utiles - 5

référence : Statistica, Wikipedia

Large numbers of inputs.

Another issue is how to deal with large numbers of inputs (variables).
In general, the DMR redundancy and dimension reduction steps are very effective for extracting the relevant diagnostic inputs from a large number of candidate inputs.

In extreme cases (e.g., when there are **thousands of inputs available**, or a few categorical inputs with thousands of categories), **it is advisable to pre-process** the list of inputs prior to applying DMR and identify a subset of inputs that are likely useful (diagnostic) for model building. The STATISTICA Data Miner documentation explains various useful approaches to solve this issue
(ref : feature extraction and feature selection in the topic
Using STATISTICA Data Miner with Extremely Large Data Sets).

Predictive Models in DMR

STATISTICA Data Miner Recipes uses various predictive models to relate the target variables to the inputs. The use of more advanced analytic model makes DMR model building an effective tool that you can use in one stage for making collective predictions.

Among these models, DMR uses Support Vector Machines (SVM), Classification and Regression Trees C&RT, Random Forests, Boosting Trees, and Neural Networks.
For an overview of these tools, please consult the STATISTICA Data Miner Manual.

Exemple DMR : prédire risque crédit

Data: 1000 dossiers de crédit ---------- 5 premières observations

	1 ID	2 Y_Credit Rating	3 Train-80-Test20	4 Balance of Current Account	5 Duration of Credit	6 Payment of Previous Credits	7 Purpose of Credit	8 Amount of Credit	9 Further running credits	10 Value of Savings	11 Employed by Current Employer for
1	1	bad	Test	no running account	36	no problems with current credits	retraining	$3 003,00	no further running cre	no savings	5-8 years
2	2	good	Test	no balance	48	hesistant	retraining	$17 085,60	at other banks	>1400	1-5 years
3	3	bad	Train	>$300	36	no previous credits	used car	$15 363,60	no further running cre	no savings	unemployed
4	4	good	Train	no running account	24	paid back	new car	$8 986,60	no further running cre	no savings	> 8 years
5	5	good	Train	>$300	24	no previous credits	retraining	$1 761,20	no further running cre	no savings	5-8 years

13 Marital Status	14 Gender	15 Living in Current Household for	16 Most Valuable Assets	17 Age	18 Type of Apartment	19 Number of previous credits at this bank	20 Occupation
single	male	< 1 year	life insurance	22	rented	2- 4	skilled employee
single	male	1-5 years	life insurance	46	rented	one	executive/self-employed
divorced/living apart/married	female	1-5 years	life insurance	24	rented	2- 4	executive/self-employed
divorced/living apart/married	female	>8 years	ownership of house or land	42	owned	2- 4	executive/self-employed
single	male	< 1 year	no assets	23	rented	one	skilled employee

variables: var1 var2 … var20

var1 = ID = numéro de ligne

Y var2 = variable de réponse

var3 = Train 80% test 20%

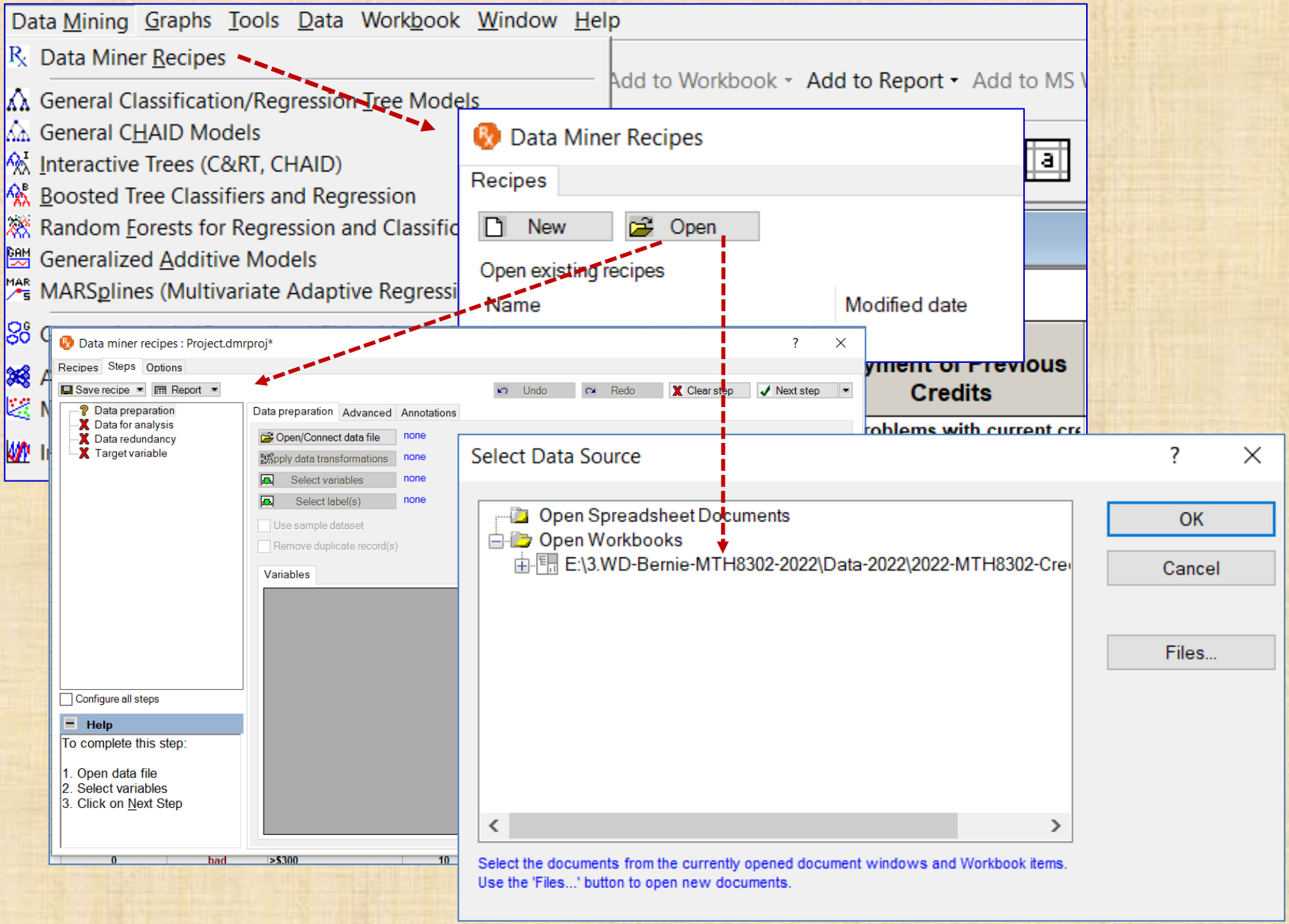
17 Variables X

var4 … var10 : informations crédit candidat

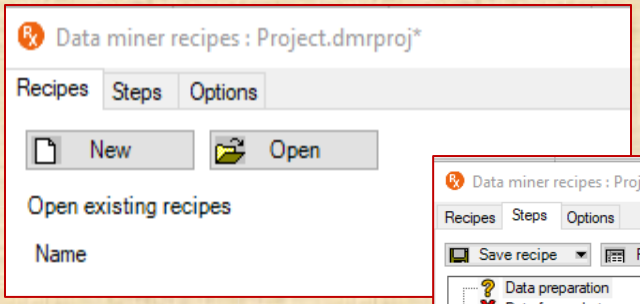
var11 … var20 : informations personnelles emprunteur

	1 Name	2 Long Name (label or formula)
1	ID	=v0;
2	Y_Credit Rating	
3	Train-80-Test20	=rnd(1)<=0,80;
4	Balance of Current Account	
5	Duration of Credit	'Duration of Credit in Months
6	Payment of Previous Credits	
7	Purpose of Credit	
8	Amount of Credit	'Amount of Credit in Dollars
9	Further running credits	
10	Value of Savings	'Value of Savings in Dollars
11	Employed by Current Employer for	
12	Installment in % of Available Income	
13	Marital Status	
14	Gender	
15	Living in Current Household for	
16	Most Valuable Assets	
17	Age	
18	Type of Apartment	
19	Number of previous credits at this bank	
20	Occupation	

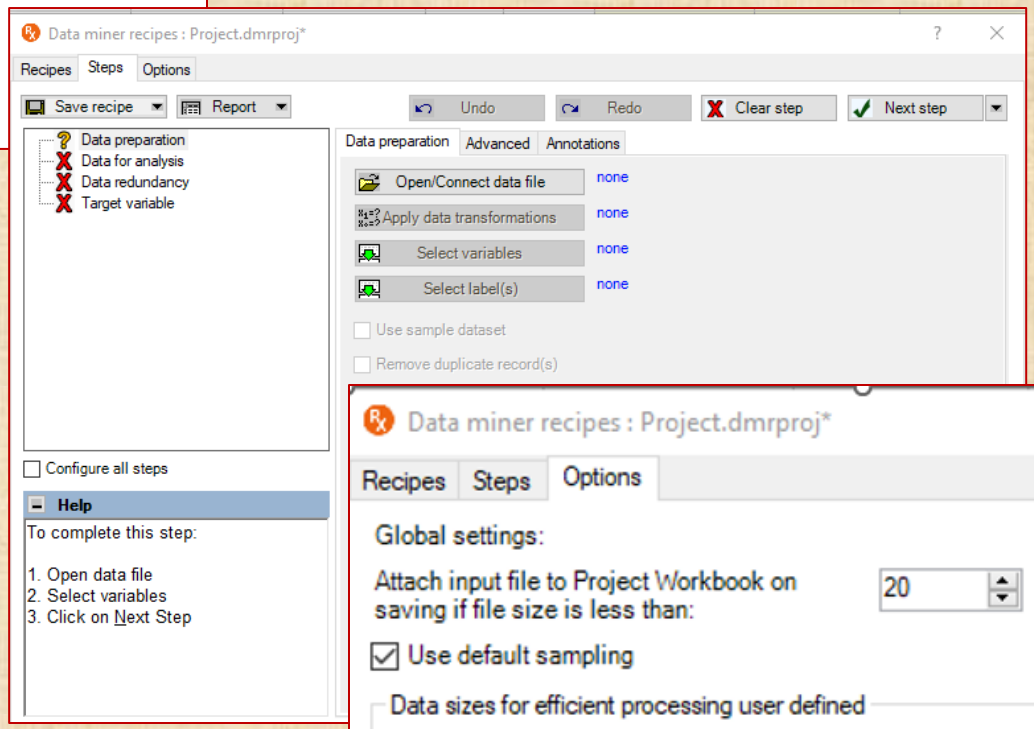
DMR – module Data Mining Recipes - mise en œuvre

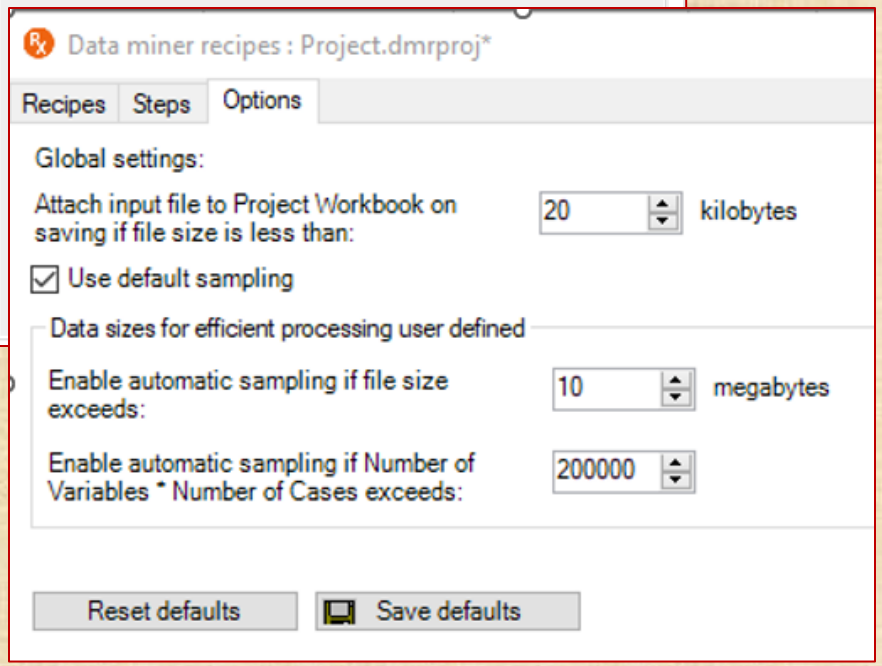


DMR – module Data Mining Recipes - mise en œuvre

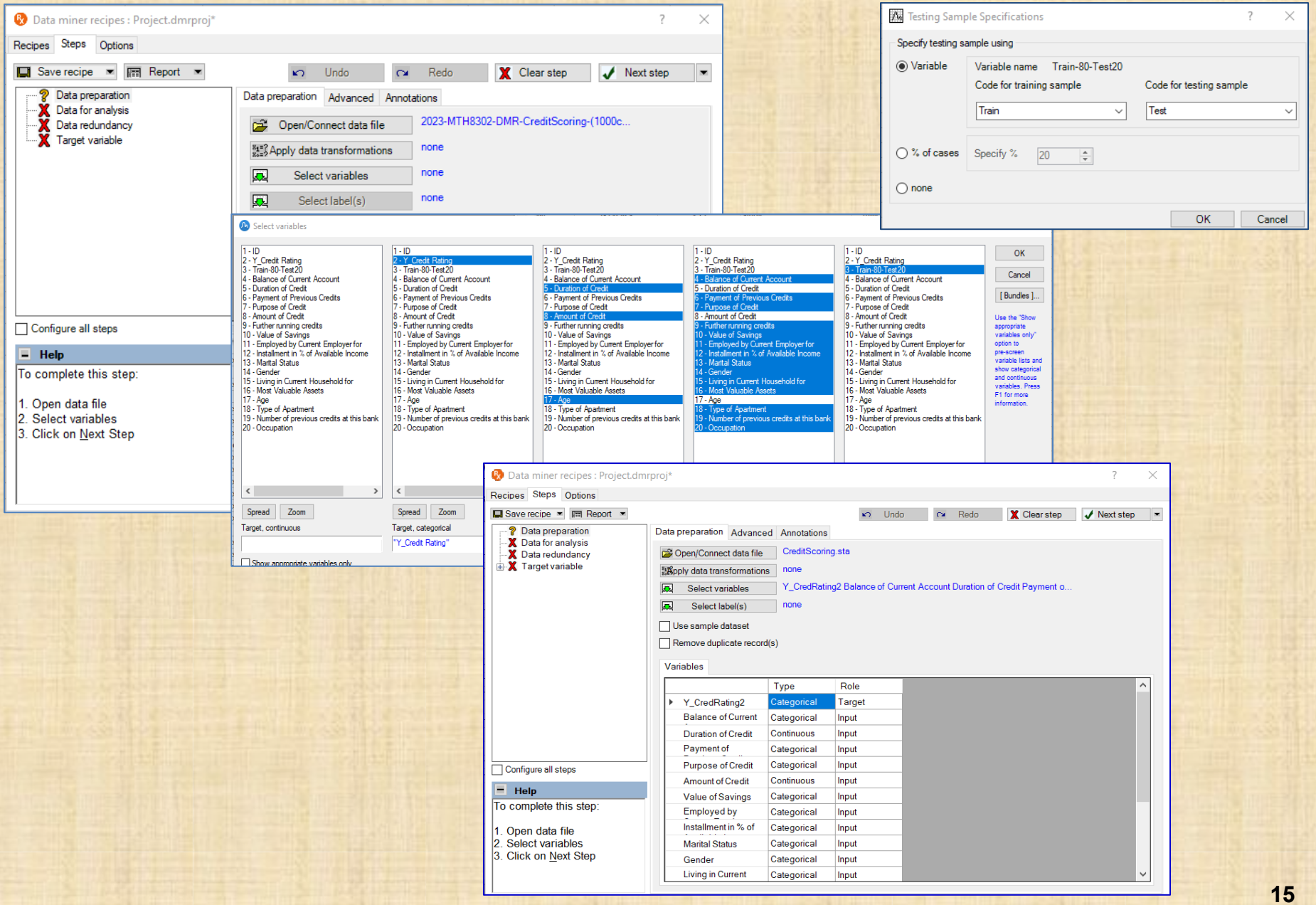


Project.dmrproj fichier pour stocké projet de DMR



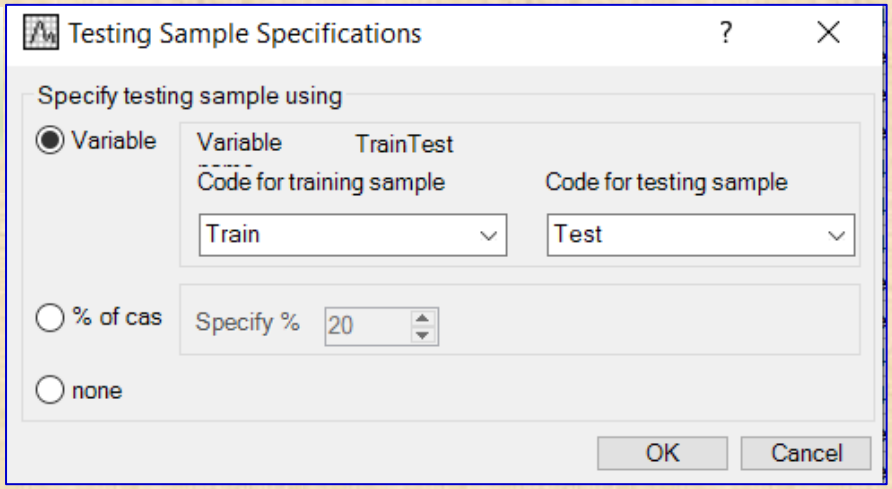


DMR – module Data Mining Recipes - mise en œuvre

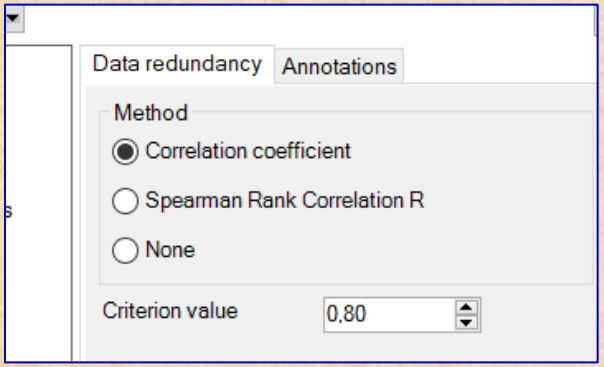


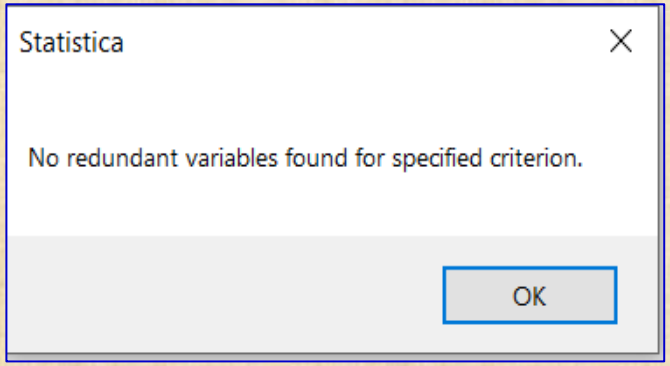
DMR – module Data Mining Recipes - mise en œuvre

Données pour l'analyse

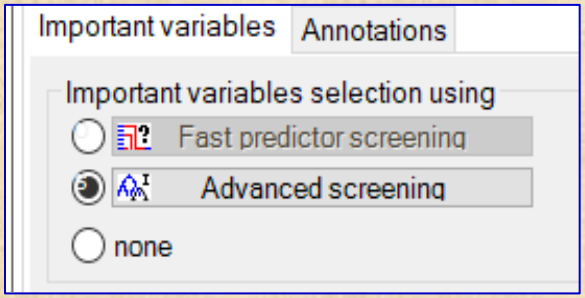


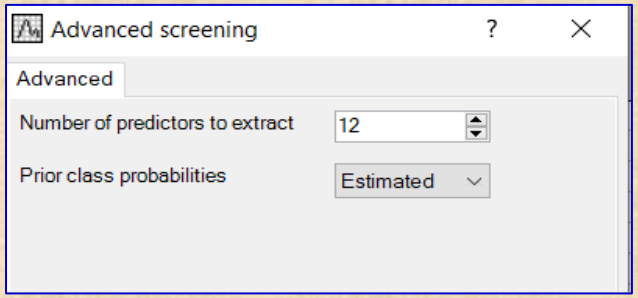
Données redondantes





Variables importantes



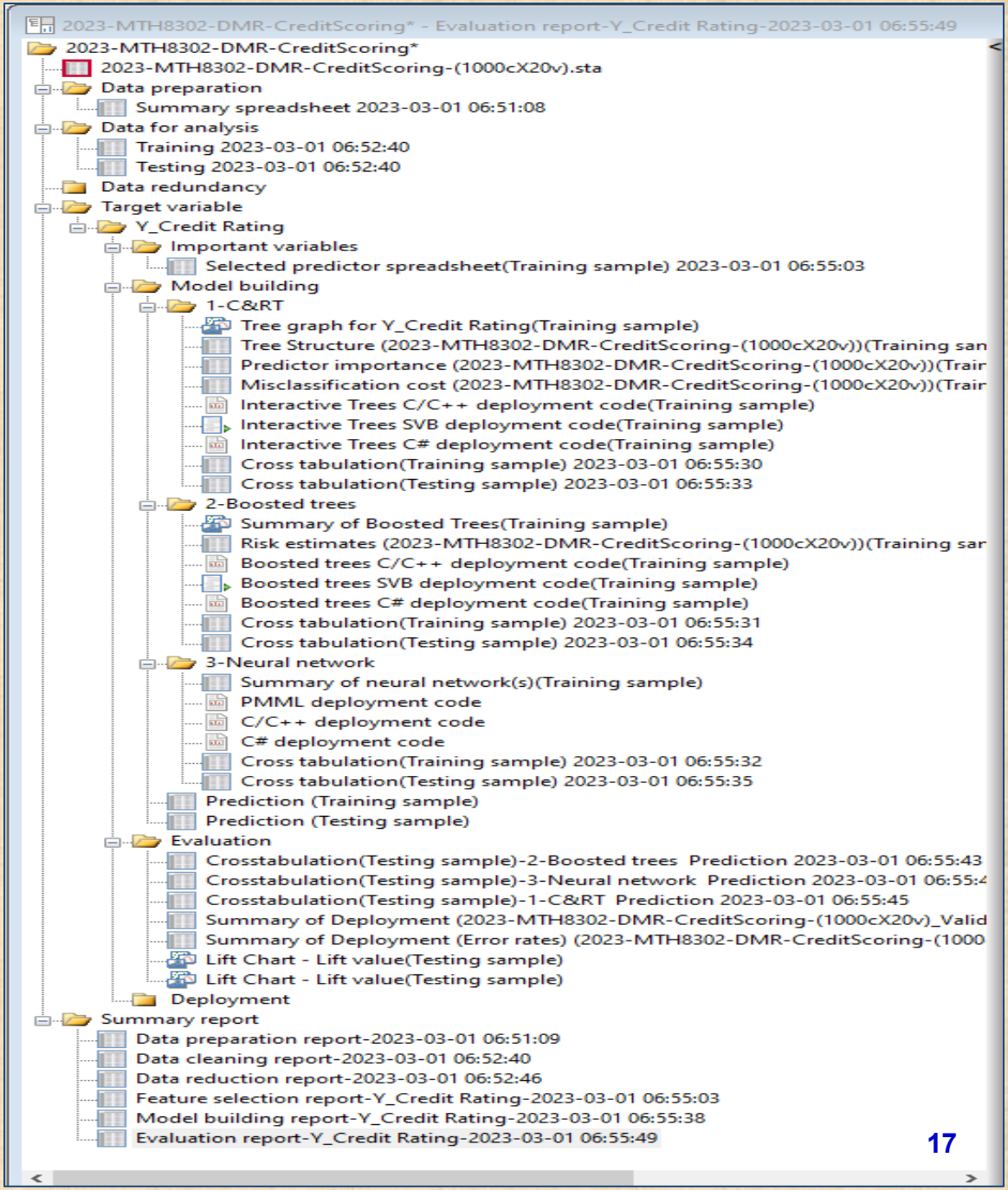


Exemple DMR :

prédire risque crédit

Résumé analyses

2023-MTH8302-DMR-CreditScoring.stw

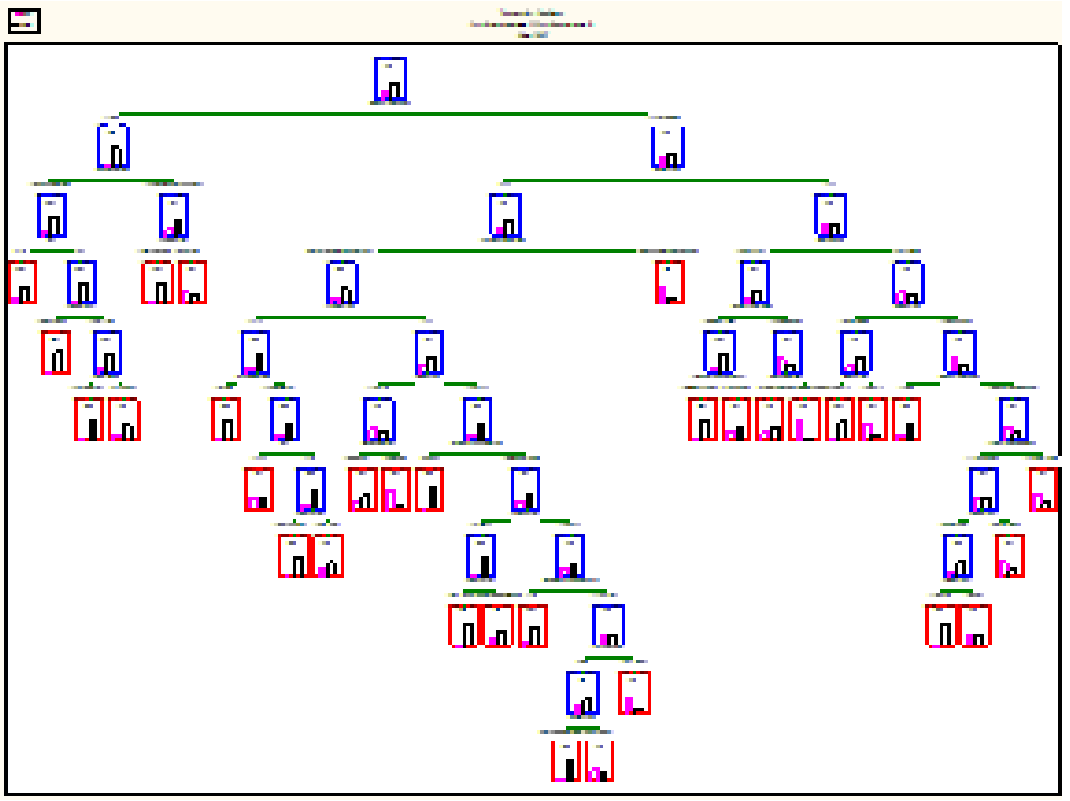


Exemple DMR : prédire risque crédit - Quelques résultats / graphiques

CRT

Predictor importance (2023
Response: Y_Credit Rating
Model: C&RT
Important variables selecti

	Variable Rank	Importance
Purpose of Credit	100	1,000000
Balance of Current Account	96	0,957519
Amount of Credit	88	0,877871
Duration of Credit	85	0,846685
Payment of Previous Credits	78	0,780257
Age	75	0,753659
Value of Savings	64	0,637098
Installment in % of Available Income	60	0,600407
Further running credits	49	0,486066
Type of Apartment	42	0,419032
Employed by Current Employer for	40	0,402100
Most Valuable Assets	36	0,363008
Living in Current Household for	31	0,308631
Occupation	28	0,280503
Marital Status	22	0,218014
Number of previous credits at this bank	19	0,193050
Gender	13	0,133005



Summary Frequency Table (Prediction)
Table: Y_Credit Rating(2) x Model-1-Prediction(2)

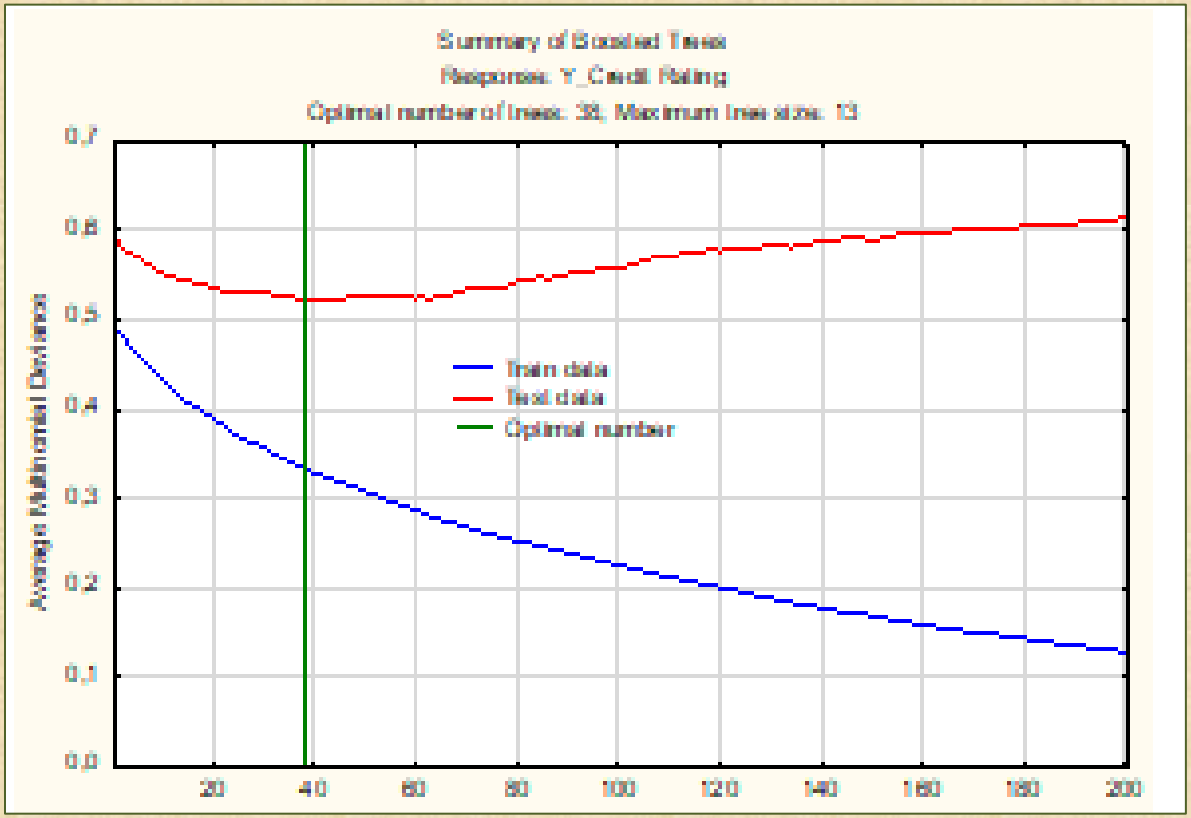
	Y_Credit Rating	Model-1-Prediction bad	Model-1-Prediction good	Row Totals
Count	bad	200	28	22
Column Percent		59,70%	5,97%	
Row Percent		87,72%	12,28%	
Total Percent		24,88%	3,48%	28,36
Count	good	135	441	57
Column Percent		40,30%	94,03%	
Row Percent		23,44%	76,56%	
Total Percent		16,79%	54,85%	71,64
Count	All Grps	335	469	80
Total Percent		41,67%	58,33%	

Summary Frequency Table (Prediction)
Table: Y_Credit Rating(2) x Model-1-Prediction(2)

	Y_Credit Rating	Model-1-Prediction bad	Model-1-Prediction good	Row Totals
Count	bad	48	24	7
Column Percent		52,17%	23,08%	
Row Percent		66,67%	33,33%	
Total Percent		24,49%	12,24%	36,73
Count	good	44	80	12
Column Percent		47,83%	76,92%	
Row Percent		35,48%	64,52%	
Total Percent		22,45%	40,82%	63,27
Count	All Grps	92	104	19
Total Percent		46,94%	53,06%	

Exemple DMR : prédire risque crédit - Quelques résultats / graphiques

Boosted trees



	Summary Frequency Table (Prediction) Table: Y_Credit Rating(2) x Model-2-Prediction(2)			
	Y_Credit Rating	Model-2-Prediction bad	Model-2-Prediction good	Row Totals
Count	bad	118	110	22
Column Percent		84,29%	16,57%	
Row Percent		51,75%	48,25%	
Total Percent		14,69%	13,68%	28,36%
Count	good	22	554	57
Column Percent		15,71%	83,43%	
Row Percent		3,82%	96,18%	
Total Percent		2,74%	68,91%	71,64%
Count	All Grps	140	664	80
Total Percent		17,41%	82,59%	

	Summary Frequency Table (Prediction) Table: Y_Credit Rating(2) x Model-2-Prediction(2)			
	Y_Credit Rating	Model-2-Prediction bad	Model-2-Prediction good	Row Totals
Count	bad	25	47	7
Column Percent		71,43%	28,19%	
Row Percent		34,72%	65,28%	
Total Percent		12,76%	23,98%	36,73%
Count	good	10	114	12
Column Percent		28,57%	70,81%	
Row Percent		8,06%	91,94%	
Total Percent		5,10%	58,16%	63,27%
Count	All Grps	35	161	19
Total Percent		17,86%	82,14%	

Exemple DMR : prédire risque crédit - Quelques résultats / graphiques

Réseau de neurones

Summary of active networks (2023-MTH8302-DMR-CreditScoring-(1000cX20v))
Subset of Summary_of_active_networks_(2023-MTH8302-DMR-CreditScoring-(1000
Variables: *
Include cases: 1

1 Index	2 Net. name	3 Training perf.	4 Test perf.	5 Validation perf.	6 Training algorithm	7 Error function	8 Hidden activation	9 Output activation
3	MLP 64-8-2	78,41615	77,50000		BFGS 4	SOS	Tanh	Logistic

Summary Frequency Table (Prediction)
Table: Y_Credit Rating(2) x Model-3-Prediction(2)

	Y_Credit Rating	Model-3-Prediction bad	Model-3-Prediction good	Row Totals
Count	bad	105	123	22
Column Percent		88,98%	19,01%	
Row Percent		46,05%	53,95%	
Total Percent		13,06%	15,30%	28,36
Count	good	52	524	57
Column Percent		33,12%	80,99%	
Row Percent		9,03%	90,97%	
Total Percent		6,47%	65,17%	71,64
Count	All Grps	157	647	80
Total Percent		19,53%	80,47%	

Summary Frequency Table (Prediction)
Table: Y_Credit Rating(2) x Model-3-Prediction(2)

	Y_Credit Rating	Model-3-Prediction bad	Model-3-Prediction good	Row Totals
Count	bad	29	43	7
Column Percent		70,73%	27,74%	
Row Percent		40,28%	59,72%	
Total Percent		14,80%	21,94%	36,73
Count	good	12	112	12
Column Percent		29,27%	72,26%	
Row Percent		9,68%	90,32%	
Total Percent		6,12%	57,14%	63,27
Count	All Grps	41	155	19
Total Percent		20,92%	79,08%	

Exemple DMR : prédire risque crédit - Quelques résultats / graphiques

évaluation

Summary of Deployment (Error rates) (2023-MTH			
	2-Boosted trees	3-Neural	1-C&RT
Misclassification error rate	0,290816	0,280612	0,346939