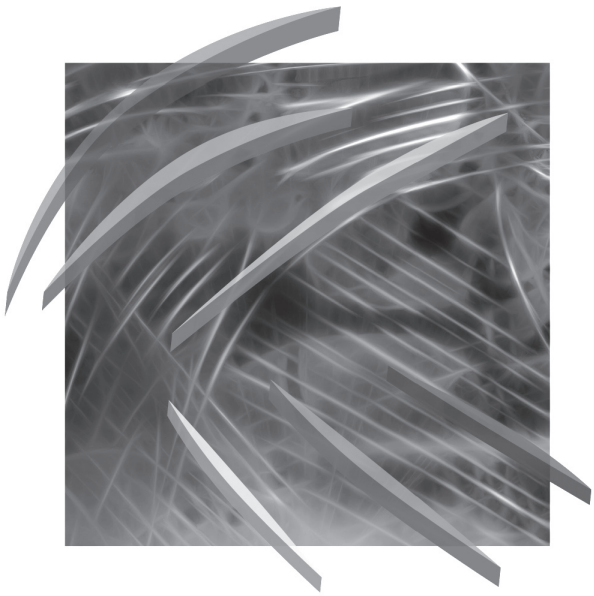
INFORMATYKA EKONOMICZNA
BUSINESS INFORMATICS

21 • 2011



Publishing House of Wrocław University of Economics
Wrocław 2011

Copy-editing: Agnieszka Flasińska, Elżbieta Macauley, Tim Macauley,

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Cibis

Typesetting: Małgorzata Czupryńska

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
and in The Central and Eastern European Online Library www.ceeol.com

Information of submitting and reviewing papers is available on the Publishing House's website www.wydawnictwo.ue.wroc.pl

ISSN 1507-3858 (Business Informatics)
ISSN 1899-3192 (Research Papers of Wrocław University of Economics)

The original version: printed

Printing: Printing House TOTEM
Print run: 200 copies

Contents

Streszczenia

ISSN 1507-3858

Maciej Pondel
Wrocław University of Economics, Wrocław, Poland
maciej.pondel@ue.wroc.pl

A COMPARISON OF DECISION TREE DATA MINING ALGORITHMS IN SAS ENTERPRISE MINER AND MS SQL SERVER DATA MINING

Abstract: Today's business generates tremendous volumes of data. Most of the data are essential to provide operational processes. Though they are useful to support operational business processes they are also used for analysts. Analytical tools support processes on a higher – managing – level. This group of solutions is called Business Intelligence. A part of the business intelligence tools is data mining software. Various software producers develop their own Data Mining tools. Author of this paper has chosen two of them: Microsoft SQL Server 2008 and SAS Enterprise Miner and compared one of the available Data Mining methods.

Key words: data mining, SAS Enterprise Miner, MS SQL Server Data Mining, decision trees.

1. Introduction to chosen tools

We can divide data analysis into two main groups. The first is building reports using transactional databases, data warehousing and OLAP. These technologies provide to the decision makers information about the business. The second group is extracting knowledge from data. This group is called Data Mining. Data Mining methods provide knowledge extracted from data to decision makers. Data mining as a process can provide real benefits to the organization that uses it.

MS SQL Server 2008 is first of all a database server that is equipped with a lot of features allowing to build a whole Business Intelligence system. In this tool we can find:

– Integration Services that is an ETL tool,
– Data Warehousing,
– Analysis Services that is multidimensional OLAP database,
– Data Mining tools,
– Reporting Services – user interface for designing reports and server distributing reports to the users.

SAS is the producer of one of the most prized analytical solutions among all software companies. SAS provides a powerful set of analytical tools that contains, among others, such solutions as:

– Enterprise Guide – data analysis tool for advanced reports building,
– Enterprise Miner – advanced data mining tool,
– Enterprise Data Integration Server – ETL tool,
– Intelligence Storage – data warehousing tool,
– OLAP Server – multidimensional OLAP database.

Other great IT Enterprises have also among their main product tools for Business Intelligence. We can find data mining tools offered by great database providers (as a part of database product):
– Oracle – Oracle Data Mining (ODM),
– IBM – DB2 Intelligent Miner.

There are also other enterprises providing analytical tools supporting data mining. For example:
– SPSS – Modeler,
– Statsoft – Statistica Data Miner.

2. Data mining methods

Data mining involves four main group of tasks [Owoc, Hauke, Pondel 2003; Witten, Frank 2005]:
– Association rules – discovering the relations between attributes describing records. This method is most commonly used to analyze customers behaviour. It allows to determine which products are bought together to make more efficient selling offers.
– Clustering – joining objects into the classes having similar properties. It is commonly used to divide the customers into groups that behave similarly (are likely to buy the same product or to abandon our services).
– Classification – in this method we are looking for relations between attributes describing analyzed objects and the target variable that is the result of some fact or operation. Based on the discovered relations we build a model that will be used to classify new cases (objects). Classification is commonly used in credit scoring (discovering which attributes describing customers influence him to be a reliable customer), customer churn, abuse detection, detection of money laundering and many more.
– Regression – attempting to find a function which models the data that we possess. It allows to predict the values that we do not know for example the future values describing some business process.

Both tools MS SQL Server and SAS Enterprise Miner have algorithms covering the listed tasks (Table 1).

SAS Enterprise Miner contains much more options and algorithms. The idea of working with the tool is also different.

Table 1. Comparison of data mining algorithms

Tasks	SAS Enterprise Guide	MS SQL Server
Association rules	Association, Market Basket	Microsoft Association Rules
Clustering	Clustering SOM/Kohonen Variable clustering	Microsoft Clustering, Microsoft Sequence Clustering Microsoft Neural Network
Classification	Decision trees, Gradient boosting, DM Neural Auto Neural Memory-Based Reasoning Neural network Rule Induction Two Stage	Microsoft Naïve Bayes Microsoft Decision Trees Algorithm Microsoft Logistic Regression Microsoft Neural Network
Regression	Regression DMINE, Least Angle Regression, Linear Regression Logical Regression	Microsoft Time Series Algorithm Microsoft Logistic Regression

Source: based on [Mendrala, Szeliga 2009; MacLennan, Tang, Crivat 2009].

3. Comparison of the idea of use

SAS Enterprise miner is a complex tool focused on the whole data mining process. Such process consists of the following steps (see [Hand 2005; Beck 1997]):

1. Data preparation – collection of data from different sources, data cleaning and transformation. SAS Enterprise Miner contains dedicated features of data preparation with implemented tasks of detecting the most common data defects for data mining.

2. Model building – the model is based on the chosen data mining task and on the chosen algorithm.

3. Model assessment – we need to determine the accuracy of the build model and examine if it is worth applying in the real business. The accuracy is calculated during the process of model validating and testing. SAS Enterprise miner is equipped with a huge number of methods assessing the generated model.

4. Applying the model – the current data should be analyzed by the prepared model. There is a task dedicated to this role in SAS Enterprise miner.

All the tasks should be configured and run in specific sequence that the user can arrange. An example of such a sequence is presented on Figure 1.

MS SQL Server Data mining is delivered to the users in 3 alternative ways. We can build the models using:

– SQL Server Business Intelligence Development Studio,

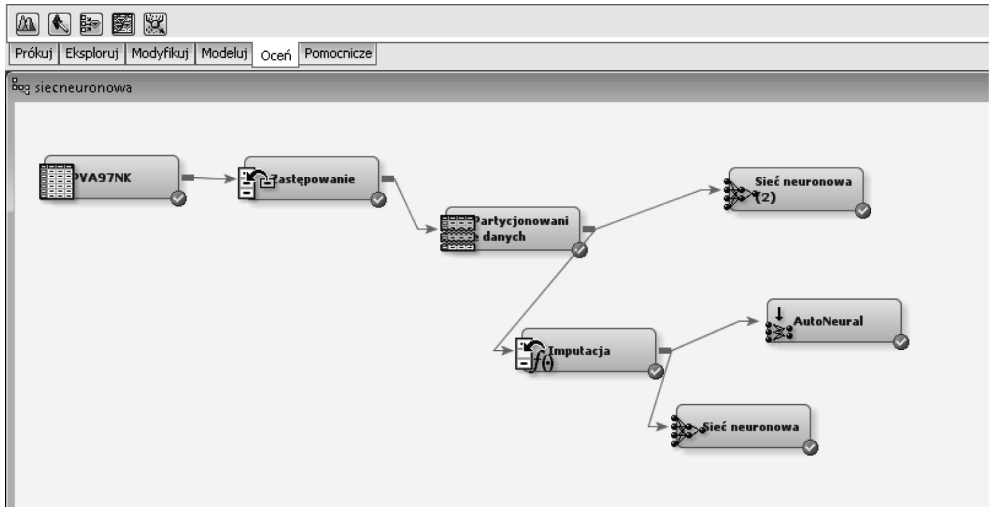


Figure 1. Example of Data Mining process in SAS Enterprise Miner

Source: own elaboration.

– Query Language for performing Data Mining operations – Data Mining Extensions to SQL (DMX),
– SQL Server 2008 Data Mining Add-Ins for Microsoft Office 2007.

SQL Server Business Intelligence Development Studio is a tool dedicated to developers. It is based on programming environment of Microsoft – Visual Studio. Using this tool the user can perform the whole data mining process but not in such a consistent way as in SAS. To prepare data, the user needs to create a project in Microsoft ETL tool called Integration Services. This tool is not dedicated only for data mining which is why the process of data preparation requires from users much more knowledge than in SAS. Using BI Development Studio, the user is also able to:

– build a model on chosen data set with chosen data mining technique and algorithm,
– explore the model with Mining Model Editor,
– asset the model with a set of Accuracy Charts.

BI Development studio is presented on Figure 2.

Next option of building data mining models is creating them using DMX queries. This language gives us whole functionality of building and browsing the model. Obviously it is the most difficult option to manage data mining.

The easiest way of performing data mining is to use MS Excel ADD-In that allows to connect to SQL Server and perform data mining. This is a tool more for analysts than for developers. It combines Excel features such as simplicity with advanced options of MS SQL Server Data mining. We provide data for data mining in Excel and from the Excel user interface we launch data mining operations. Excel

is also the user interface of browsing the results of data mining. In this case we can only build a model and assess it. To prepare the data we need to use standard Excel functions. This process is shown on Figure 3.

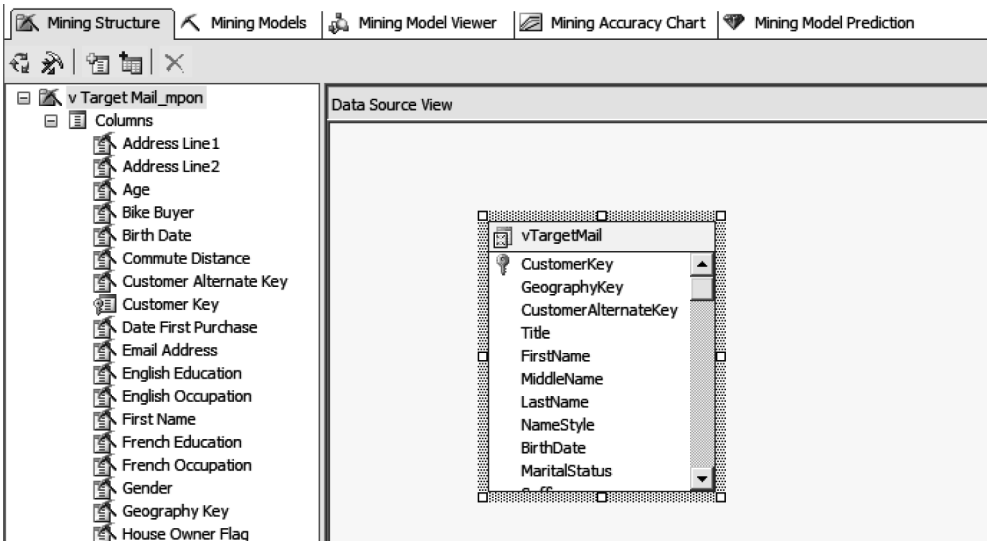


Figure 2. Data Mining with BI Development studio

Source: own elaboration.

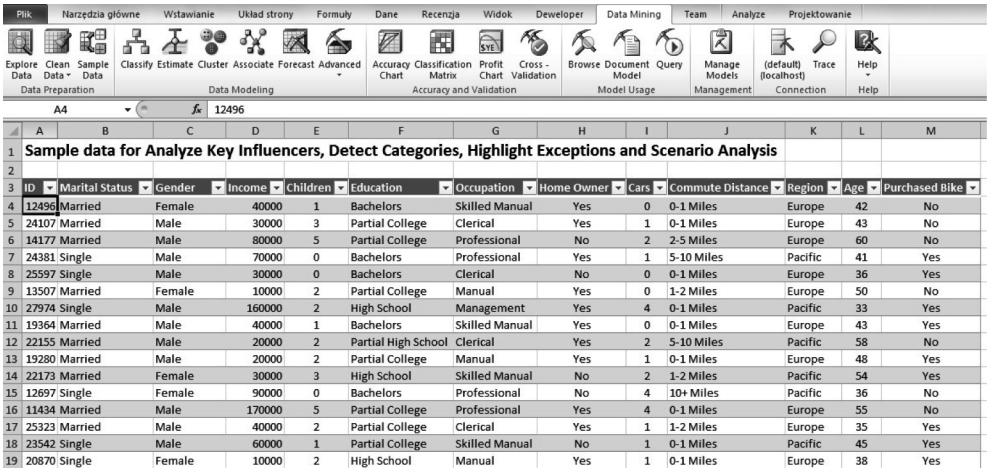


Figure 3. Data mining in MS Excel

Source: own elaboration.

4. Decision trees

Decision tree is the method of data mining belonging to predictive modelling. In predictive modelling, also called supervised prediction or supervised learning, our aim is to identify relationships between the input values and the target. Input values

⊟ Reguła podziału	
Kryterium przedziałowe	ProbF
Kryterium nominalne	ProbChisq
Kryterium porządkowe	Entropia
Poziom istotności	0.2
Braki danych	Użyj w wyszukiwaniu
Używaj jednorazowo	Nie
Maksymalne rozgałęzienie	2
Maksymalna głębia	6
Minimalna wielkość kategoryzacji	5
⊟ Węzeł	
Wielkość liściowa	5
Liczba reguł	5
Liczba reguł zastępczych	0
Wielkość podziału	
⊟ Poszukiwanie podziału	
Wyczerpujące	5000
Próba węzła	20000
⊟ Poddrzewo	
Metoda	Ocena
Liczba liści	1
Miara oceny	Decyzja
Ułamek ocen	0.25
⊟ Walidacja krzyżowa	
Wykonuj walidację krzyżową	Nie
Liczba podzbiorów	10
Liczba powtórzeń	1
Ziarno	12345
⊟ Istotność na podstawie obserwacji	
Istotność na podstawie obserwacji	Nie
Liczba istotności jednozmiennych	5
⊟ Korekta wartości p	
Korekta Bonferroniego	Tak
Czas korekty Kassa	Przed
Informacje wejściowe	Nie
Liczba zmiennych wejściowych	1

Figure 4. Setting parameters in SAS Decision tree

Source: own elaboration.

are attributes describing the case and the target value as the result of classification. Decision trees provide prediction rules to score new cases. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. A tree can be "learned" by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node all has the same value of the target variable, or when splitting no longer adds value to the predictions. [Wikipedia 2011]. Decision trees are very easy to understand even by beginner data miners which is why they were chosen to comparison in SAS and MS SQL Server.

Every data mining method in SAS Enterprise Miner has more options and more parameters to set. In decision trees we have two main ways of building the tree.

1. In an interactive way. This means that user is able to build the tree on his own deciding on which attribute the next leaf will be based.

2. Traditional way. The tree is built by the algorithm from the beginning to the end.

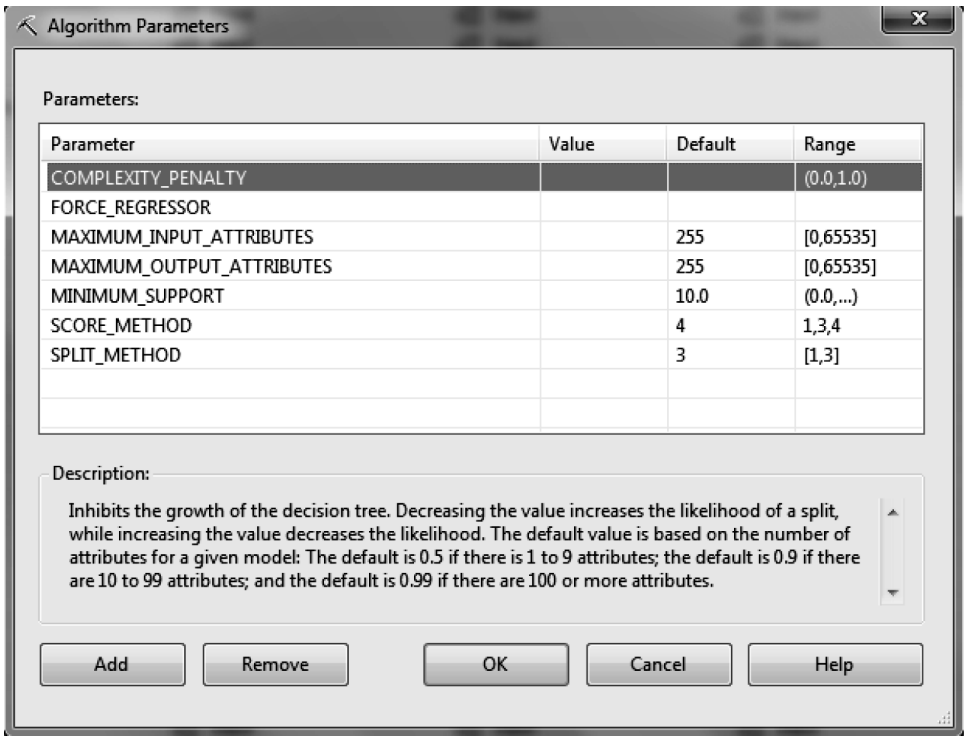


Figure 5. Setting parameters in MS Decision Tree

Source: own elaboration.

In Microsoft Decision Trees the interactive way of building a tree does not exist.

Counting the number of parameters steering the building of the tree:

– SAS Enterprise miner has 28 parameters,
– MS SQL Server has 7 parameters.

Setting those parameters is shown on Figures 4 and 5.

To compare the efficiency of those two algorithms in two different tools we will provide the same database that is prepared for building a model and we will test it on the same test set. To this test we will use the default parameters configuration. The resulting decision trees are shown in Figures 6 and 7.

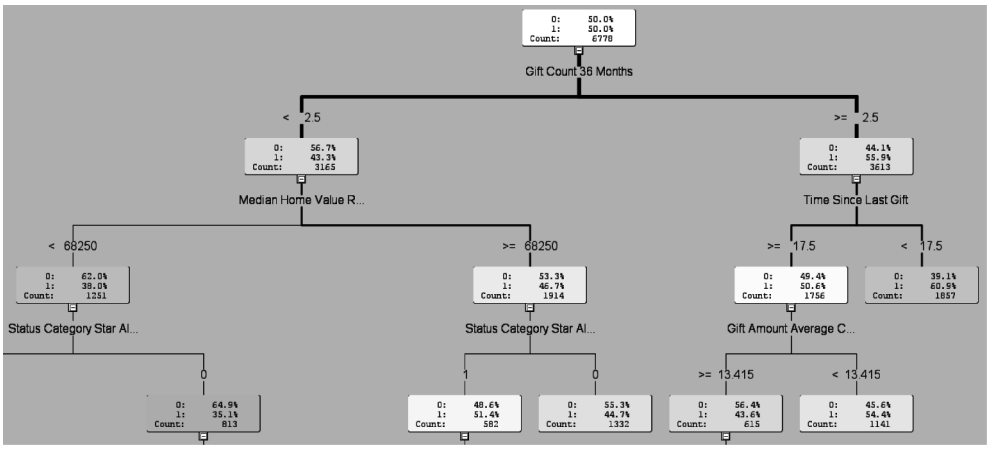


Figure 6. SAS EM Decision Tree

Source: own elaboration.

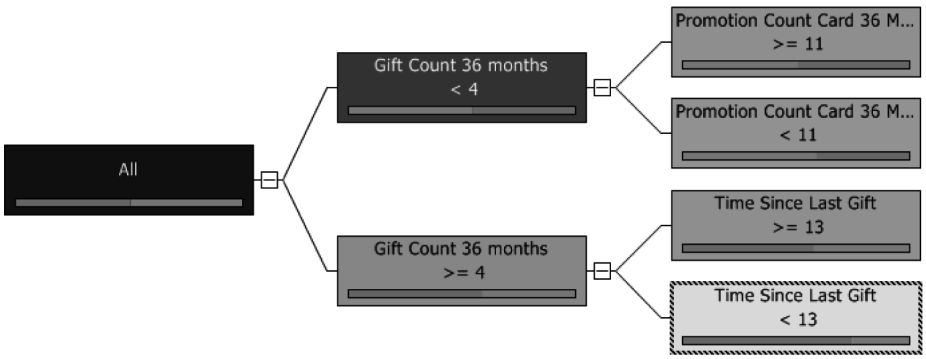


Figure 7. MS Decision Tree

Source: own elaboration.

We can observe the decision tree generated if SAS is much deeper and contains much more leaves. But how can it influence the correctness of the classification of the test set? The answer is shown below in tables that show the classification matrix for these two trees.

Table 2. Classification matrix for SAS decision tree

	0(Actual)	1(Actual)
0	59.**55%**	42.17%
1	40.45%	57.**83%**
Correct	59.**55%**	57.**83%**
Misclassified	40.**45%**	42.**17%**
Total correct	58.**69%**	
Total misclassified	41.**31%**	

Source: own elaboration.

Table 3. Classification matrix for MS decision tree

	0(Actual)	1(Actual)
0	71.**28%**	57.26%
1	28.72%	42.**74%**
Correct	71.**28%**	42.**74%**
Misclassified	28.**72%**	57.**26%**
Total correct	57.**21%**	
Total misclassified	42.**79%**	

Source: own elaboration.

As shown in the tables, although the decision trees look completely different their aggregated correctness is similar. The particular results are much different but the difference of total correctness is about 1.5 percentage points.

5. Summary

SAS Enterprise miner is a much more advanced data mining tool than MS SQL Server 2008R2. It has much more data mining algorithms and they are much more sophisticated. Users are able to tune them with the parameters in more advanced ways. Moreover SAS provides a whole data mining process closer to the methodology. The

biggest advantage of MS SQL Server Data mining is simplicity of use, in particular when they use MS Excel Data Mining Add-In. As the test on the decision tree proved – SAS model was more efficient than MS one, but the difference was quite little. We can state that SAS is the tool for more exacting users and more advanced. MS SQL Data Mining can be a good alternative for beginners or for users already using the Microsoft SQL Server database.

References

Beck A., *Herb Edelstein discusses the usefulness of data mining*, http://www.tgc.com/dsstar/971014/100007.html.

Hand D., What you get is what you want? Some dangers of Black Box Data Mining, [in:] *M2005 Conference Proceedings*, SAS Institute, Cary, NC, 2005.

MacLennan J., Tang Z., Crivat B., *Data Mining with Microsoft SQL Server 2008*, Wiley Publishing, Indianapolis 2009.

Mendrala D., Szeliga M., *Serwer SQL 2008 Usługi biznesowe*, Helion, Gliwice 2009.

Owoc M., Hauke K., Pondel M., Building Data Mining Models in the Oracle 9i Environment, [in:] *Informing Science + IT Education Conference, Pori, Finland, 24-27 June 2003*, Pori 2003.

Wikipedia 2011, http://en.wikipedia.org/wiki/Decision_tree_learning.

Witten I., Frank E., *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers (Elsevier), San Francisco 2005.

PORÓWNANIE ALGORYTMÓW DRZEW DECYZYJNYCH W NARZĘDZIACH SAS ENTERPRISE MINER I MS SQL SERVER DATA MINING

Streszczenie: W dzisiejszym biznesie generowana jest ogromna ilość danych. Większość tych danych jest niezbędna do zapewnienia sprawnego przebiegu procesów operacyjnych. Są one również bardzo ważne z analitycznego punktu widzenia. Narzędzia analityczne wspierają procesy wyższego poziomu – zarządzania. Taka grupa narzędzi nosi nazwę Business Intelligence. Część tej grupy stanowią narzędzia drążenia danych. Wielu producentów oprogramowania ma w swojej ofercie narzędzia drążenia danych. Autor na potrzeby tego artykułu wybrał dwa: Microsoft SQL Server 2008 oraz SAS Enterprise Miner i porównał jedną z dostępnych metod drążenia danych.

Słowa kluczowe: drążenie danych SAS Enterprise Miner, MS SQL Server Data Mining, drzewa decyzyjne.