Improving the Wholesales Trough Using the Data Mining Techniques

Goran Vitanov , Igor Stojanovic, Zoran Zdravev

University Goce Delcev – Shtip

goran.vitanov@gmail.com, {igor.stojanovik, zoran.zdravev}@ugd.edu.mk

Abstract. This paper describes the practical use of data mining techniques in wholesales though concrete steps of developing the distribution network. We are choosing a number of techniques for solving the problem through analyzing the specifics of the market situation, defining the business problem, and setting up its model. The cluster analysis enables us to group the data according to similarities of the market segmentation, product categories, regions and groups by turnover, while with the decision trees we are separating the big data collection into consecutive groups. Using these techniques we gain knowledge and have the opportunity to predict the future trends with high probability and on this basis to make a more precise and trustworthy business decision.

Keywords: Data Mining, Cluster, Decision trees, Wholesales.

1 Introduction

The data mining techniques are used in order to foresee a potential future result in other words to predict future possibilities and trends. There are many different automatic analysis techniques of large amount of data with multiple variables, for example: clustering, decision trees, analysis of basket sales, regression models, neural networks, genetic algorithms, hypothesis testing, and many more. In the predicting models the data is gathered, the statistic model is formulated, the prediction is completed, and the model is checked when additional information is available. The analysis combines the knowledge from business and the statistic analytical techniques in order for the user to find information from the existing data. The resulting information enables the companies to better understand the buyers, the sellers, the distributers, etc. Furthermore, this enables the user reach strategic decisions, for example where they can invest, whether to increase or decrease a portfolio of products, or whether to move onto new markets. These analyses not only show what needs to be done, but when and where.

S. Markovski, M. Gusev (Editors): ICT Innovations 2012, Web Proceedings, ISSN 1857-7288

2 Description of the market condition

Increased competition, increased product variety, constant changing of the market segmentation, the complexity of covering certain territories through engaging new commercialists, price increases, the buyers becoming more sophisticated, and the need for using distribution centers are only a portion of the complex dynamics, which is present in the field of distribution. The constant changes in the market environment from the aspect of retail segmentation, the demographic movements, and the increased choices of the buyers, are causing the distribution companies to face further challenges. To be able to survive in such environment the need for analytical, directed, and predicting approach rises. This would result in better understanding the challenges with aim to get more organized, to increase the market coverage, and to increase the profit. Today, in order to reply to the existing challenges there are multiple tools for data mining and prediction, which through the years become more sophisticated and their usage further simplified. Unlike the past, nowadays with the development of the hardware and software these technologies are increasingly becoming an integral part of every large company. This enables the company to bring a newer component in the decision system. The management today can identify the variables that have the most influence over the performance of the business and on this basis to yield better decisions.

An example company that is working on distribution of products from multiple manufacturers in Macedonian market is analyzed. The sale of certain categories of products is changed significantly and the need for more detailed analysis, from the aspect of the changes in market segmentation, is imposed. Based on the database results it has been realized that the number of buyers is changing, which brings the requirement the data to be analyzed in detail. A research is required to understand what is the reason why there parameters are changing, with the goal to increase the profit. The current trend shows that this research should be continued.

3 Information System

The company has six distributed locations with applications connected to central database. The distributed applications are covering the commercial business as (invoicing, orders, purchases, account receivable and account payable, warehouse stocks and etc.). In central location is information system for accounting and management.

Two years ago is developed data warehouse system to support reporting, controlling and planning. Separating and regrouping the information allows seeing a trend, exceptions, patterns, and relationships. This enables to analyze multidimensional data from multiple perspectives using consolidation, drill-down, slicing and dicing. The ETL (extraction, transportation, transformation and loading) solution is developed in order to load the data warehouse.

The operating system is Microsoft Server 2008 and the database system is Microsoft SQL Server 2008. For data mining the company is using Microsoft SQL Server Analysis Services, Data Mining Client for Excel and SPSS.

4 Setting up the model in order to solve the problem

Given the complexity of the problem more analysis is needed in order to get a clear picture of the market condition. The initiating meeting of a data mining project will often involve the data miner, the business expert and the data analyst. It is critical that all three come together to initiate the project. Their different understandings of the problem to be tackled all need to come together to deliver a common pathway for the data mining project. In particular, the data miner needs to understand the business perspective and understand what data is available that relates to the problem and how to get that data, and identify what data processing is required prior to modeling. To be more exact, one should answer the following questions:

- What would the market segmentation be based on the product category?
- What is the connection between a manufacturer and a market segment?
- What is the trend between a regional and a market segment?

Answering these questions lowers the risk when making the decision and it affects the future profitability of the company. In order to complete this, the following techniques of data mining are used as a part of the analysis: Association discovery, Cluster analysis and Decision trees.

4.1 Association discover

In the case of a given collective data out of which every entry contains certain number of elements, the goal of the association would be to find the connection among the elements in the set. The association rule in the data mining can be defined as: let I={i1, i2, ..., in} be a set of n binary attributes which are called: items. Let D={t1, t2, ..., tm} be a set of transactions called data base. Every transaction in D has a unique transaction ID and contains subset of items in I. The rule is established with the implication of the form: X=>Y where X, Y .. I and X .. Y = 0. The set of items X and Y are called predecessor (left side) and successor (right side) of the rule.

Using the association rule we arrive to the following results - Table 1:

Table 1. Number of found items using the association rule.

Support	Size	Itemset
2181	1	Category = TT
1861	1	Grupa Promet >10000
1778	1	Region = South
1704	1	Region = North
1662	1	Category = LKA
1661	1	Region = Исток
1495	1	Category = NKA
1372	1	Grupa Promet = 5001-10000
1369	1	Region = West

1362	1	Grupa Promet = 2001-5000
1023	1	Grupa Promet = 1000-2000
894	1	Grupa Promet = 1000<
780	1	Category = Wholesaler
602	1	Category Producer = КОНЗЕРВИРАН ЗЕЛЕНЧУК
587	1	Category Producer = КОНЗЕРВИ
581	2	Region = South, Category = TT
580	2	Region = Исток, Category = TT
579	1	Category Producer = МЛЕЧНИ ПРОИЗВОДИ
522	2	Region = North, Category = TT
498	2	Region = West, Category = TT
494	1	Category Producer = СУВОМЕСНАТО
492	2	Category = LKA, Region = South
485	1	Category Producer = ГРИЦКИ
483	2	Region = West, Grupa Promet >10000

The system recognizes 2181 buyers in the category TT, 1861 accounts in the "Group of operations" > 10000, 1778 sales in the region "south" etc. This data with further application of the association rule yields the following results:

Table 2. : Results from the association rule usage.

Probability	Importance	Rule
54%	0.22	Group turnover = 1000<, Region = East -> Category = TT
51%	0.19	Group turnover = 1000<, Region = West -> Category = TT
51%	0.19	Group turnover = 1000-2000, Region = West -> Category = TT
46%	0.17	Group turnover = 1000< -> Category = TT
44%	0.13	Group turnover = 1000-2000, Region = East -> Category = TT
43%	0.11	Group turnover = 1000-2000, Region = South -> Category = TT
42%	0.12	Group turnover = 1000-2000 -> Category = TT
42%	0.1	Group turnover = 1000<, Region = North -> Category = TT
41%	0.09	Group turnover = 1000<, Region = South -> Category = TT
40%	0.08	Group turnover = 2001-5000, Region = West -> Category = TT

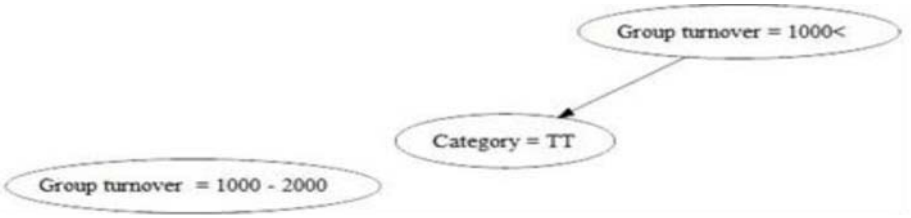


Fig. 1. : Connection strength between categories.

Based on these results it could be concluded that the operations under 1000 denars in all of the regions is connected with the TT channel of sale. The same rule can be

applied for the operations from 1000 to 2000 denars. This means that we have not deduced a separate rule for sales in certain product category, but that the size of the invoice is important. Furthermore, the TT channel during procurement prefers to be more careful of the account size rather than the product portfolio. This is also described in Figure 1.

4.2 Cluster analysis

The goal of the cluster analysis or clustering is to combine a set of objects in a group called "cluster", that way the objects in one cluster have the most similarities among themselves, rather than among other objects in a different cluster. The cluster itself is not a separate algorithm but it is a group of algorithms that enable grouping of objects according to their common attributes. Other terms used to describe clusters are: automatic classification and numerical taxonomy. In this paper the tool "Detect Categories", which is included in "Microsoft Excel", was used to detect clustering of the objects. After using this technique the following results are acquired:

Table 3. Categories created after using "Detect Categories" tool.

Category Name	Row Count
ТТ-Сувомеснато_Запад	2451
Југ-ЛКА	2600
Исток-Средства за садови	2012
НКА-Конзервиран зеленчук	2237

Category	Column	Value	Relative Importance
ТТ-Сувомеснато_Запад	Kategorija	ТТ	41
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	СУВОМЕСНАТО	35
ТТ-Сувомеснато_Запад	Region	Запад	23
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ПИЈАЛОЦИ	10
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	КЕЧАП	10
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ЛЕБ	7
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	ФЛИПС	7
ТТ-Сувомеснато_Запад	Region	Север	6
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	АЛКОХОЛНИ ПИЈАЛОЦИ	2
ТТ-Сувомеснато_Запад	Kategorija Proizvoditel	МАСТИКИ	1
Југ-ЛКА	Region	Југ	77
Југ-ЛКА	Kategorija	LKA	74
Југ-ЛКА	Kategorija Proizvoditel	ДЕЗОДЕРАНСИ	18
Југ-ЛКА	Kategorija Proizvoditel	ГРИЦКИ	5
Југ-ЛКА	Kategorija Proizvoditel	ЧИПС	2
Југ-ЛКА	Kategorija Proizvoditel	МЛЕЧНИ ПРОИЗВОДИ	2

Југ-ЛКА	Kategorija Proizvoditel	БОЈА ЗА КОСА	1
Југ-ЛКА	Kategorija Proizvoditel	ДЕТЕРГЕНТИ	1
Југ-ЛКА	Kategorija Proizvoditel	КОНЗЕРВИ	1
Исток-Средства за садови	Region	Исток	100
Исток-Средства за садови	Kategorija Proizvoditel	СРЕДСТВА ЗА САДОВИ	13
Исток-Средства за садови	Kategorija	ТТ	2
Исток-Средства за садови	Kategorija Proizvoditel	ТОАЛЕТНА ХАРТИЈА	1
НКА-Конзервиран зеленчук	Kategorija	NKA	100
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	КОНЗЕР. ЗЕЛЕНЧУК	13
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	ЧИПС	2
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	ДЕТЕРГЕНТИ	2
НКА-Конзервиран зеленчук	Kategorija	Wholesaler	1
НКА-Конзервиран зеленчук	Kategorija Proizvoditel	КОНЗЕРВИ	1
НКА-Конзервиран зеленчук	Region	Север	1

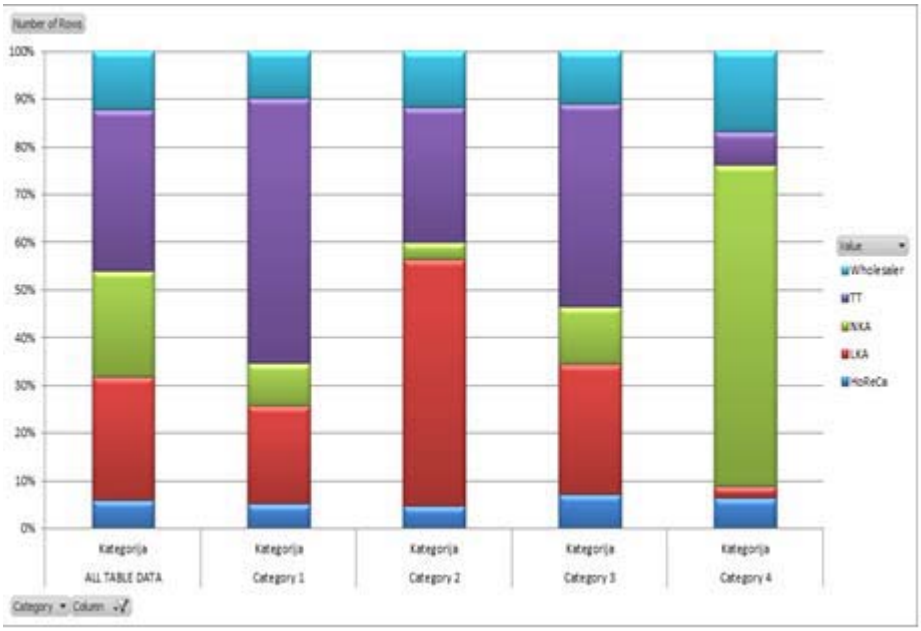


Fig. 2. : Created Clusters

Four categories are formed, which give us a precise picture of the market. Hence the TT channel is important for sales of cold cuts, especially in the western region. The western region is also important for sales of ketchup and beverages. That means if

changes of the distribution network are planned it is important to consider the sales of these groups of products in this region. The southern region is a region in which the LKA buyers are more important. From this we conclude that the investments need to be bigger in this channel from the viewpoint of the marketing activities and the strengthening of the sales force. The east region is important for sale of soaps and detergent and therefore there is a need to support this type of products. The HKA cluster and sales of canned foods gives us direction for increasing the activities of these buyers connected with the sales of canned products. This is an expected trend because NKA is mainly concentrated in the larger towns and the buyers are mainly employed people. By applying the technique Microsoft Cluster on the same data, 9 clusters are created, which are shown on Figure 2 . From the shown data it can be seen that cluster 1 represents the southern region that no groups of products stand out, that in the sales channels TT LKA, is dominant, that the participation of NKA is the smallest cluster 2 represents the north region of NKA sales without important participation of any product, etc.

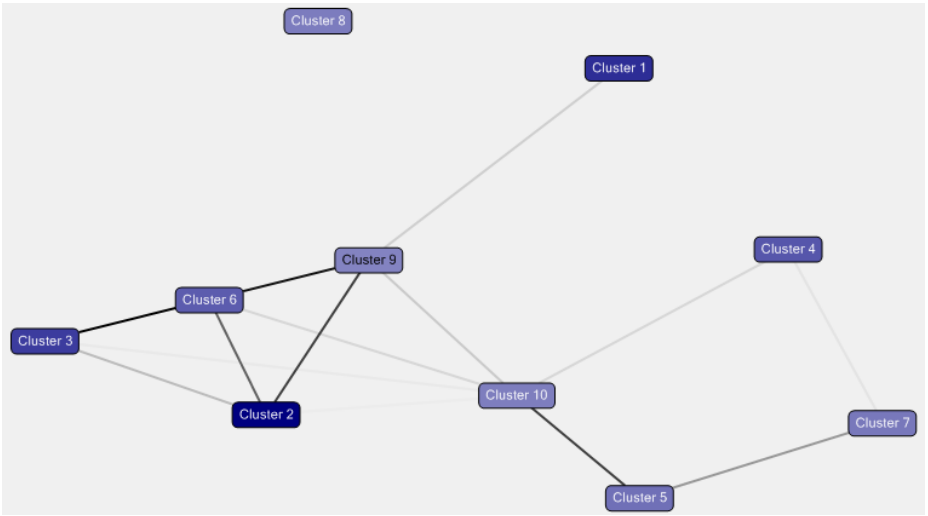
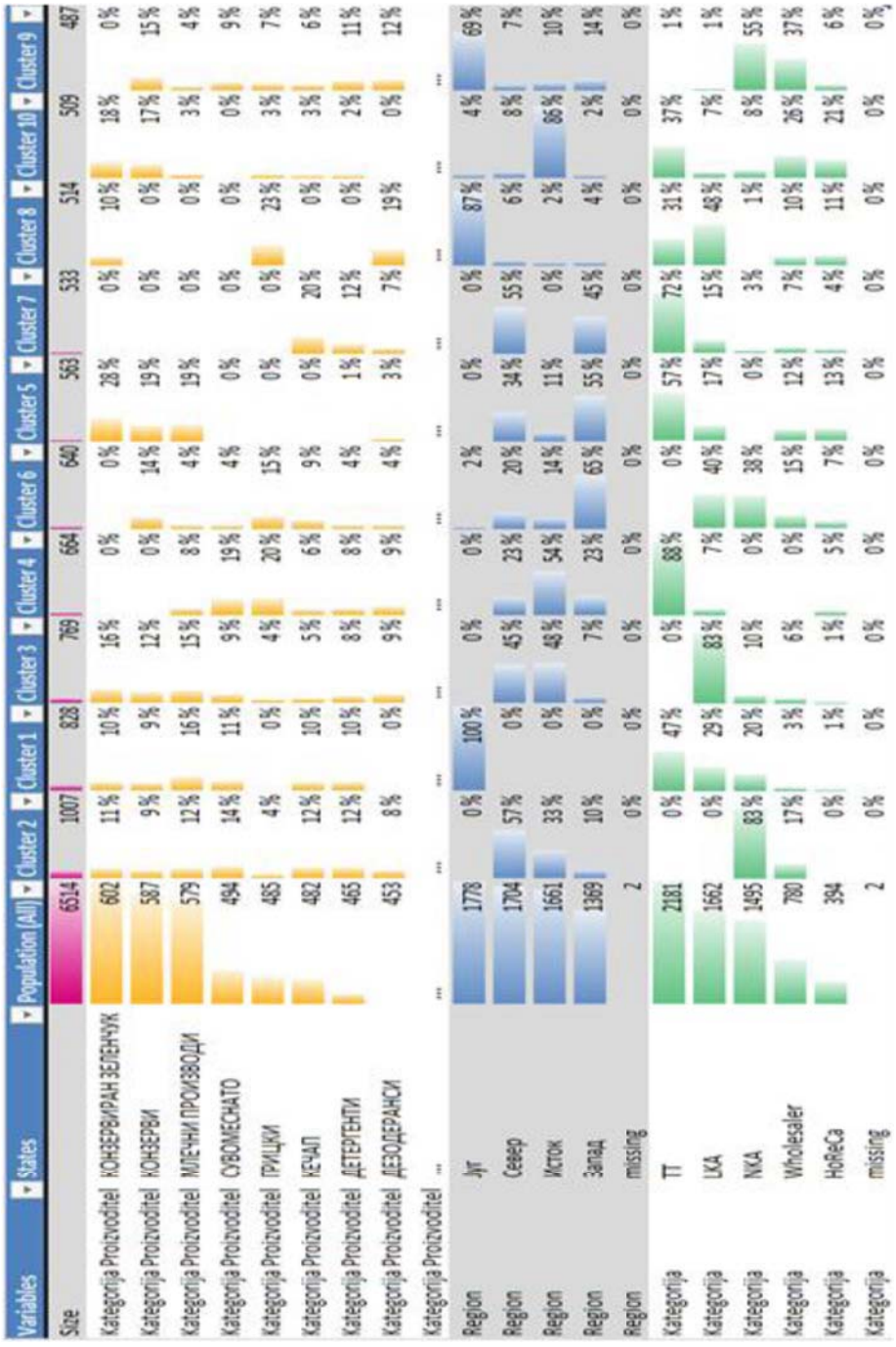


Fig. 3. : Connections among the created clusters



Variables	States	Population (All)	Cluster 2	Cluster 1	Cluster 3	Cluster 4	Cluster 6	Cluster 5	Cluster 7	Cluster 8	Cluster 10	Cluster 9
Size		6514	1007	828	769	664	640	563	533	514	509	487
Kategorija Proizvoditel	КОНЗЕРВИРАН ЗЕЛЕНЧУК	602	11 %	10 %	16 %	0 %	0 %	28 %	0 %	10 %	18 %	0 %
Kategorija Proizvoditel	КОНЗЕРВИ	587	9 %	9 %	12 %	0 %	14 %	19 %	0 %	0 %	17 %	15 %
Kategorija Proizvoditel	МЛЕЧНИ ПРОИЗВОДИ	579	12 %	16 %	15 %	8 %	4 %	19 %	0 %	0 %	3 %	4 %
Kategorija Proizvoditel	СУВОМЕСНАТО	494	14 %	11 %	9 %	19 %	4 %	0 %	0 %	0 %	0 %	9 %
Kategorija Proizvoditel	ГРИЦКИ	485	4 %	0 %	4 %	20 %	15 %	0 %	0 %	23 %	3 %	7 %
Kategorija Proizvoditel	ЧЕЧАП	482	12 %	10 %	5 %	6 %	9 %	0 %	20 %	0 %	3 %	6 %
Kategorija Proizvoditel	ДЕТЕРГЕНТИ	465	12 %	10 %	8 %	8 %	4 %	1 %	12 %	0 %	2 %	11 %
Kategorija Proizvoditel	ДЕЗОДЕРАНСИ	453	8 %	0 %	9 %	9 %	4 %	3 %	7 %	19 %	0 %	12 %
Kategorija Proizvoditel	...	...	...	...	...	...	...	...	...	...	...	...
Region	Jyг	1778	0 %	100 %	0 %	0 %	2 %	0 %	0 %	87 %	4 %	69 %
Region	Север	1704	57 %	0 %	45 %	23 %	20 %	34 %	55 %	6 %	8 %	7 %
Region	Исток	1661	33 %	0 %	48 %	54 %	14 %	11 %	0 %	2 %	86 %	10 %
Region	Запад	1369	10 %	0 %	7 %	23 %	65 %	55 %	45 %	4 %	2 %	14 %
Region	missing	2	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Kategorija	TT	2181	0 %	47 %	0 %	88 %	0 %	57 %	72 %	31 %	37 %	1 %
Kategorija	LKA	1662	0 %	29 %	83 %	7 %	40 %	17 %	15 %	48 %	7 %	1 %
Kategorija	NKA	1495	83 %	20 %	10 %	0 %	38 %	0 %	3 %	1 %	8 %	55 %
Kategorija	Wholesaler	780	17 %	3 %	6 %	0 %	15 %	12 %	7 %	10 %	26 %	37 %
Kategorija	HoReCa	394	0 %	1 %	1 %	5 %	7 %	13 %	4 %	11 %	21 %	6 %
Kategorija	missing	2	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Fig. 4. : Clusters created with Microsoft Cluster Technique

4.3 Decision trees

The decision tree is a powerful and popular technique that can be used for classification and prediction. The attractiveness of the methods based on decision tree is mostly based on the fact that they are rules. These rules can be represented with SQL expressions in order for information from a specific category to be obtained. Because the decision trees combine the research of information and modeling, they are a powerful first step in the process of modeling even if the final model is obtained with some other technique. The decision tree is a structure that can be used to divide a large collection of data into a number of smaller sets of data by using simple decision rules. With each new division, the members of the obtained sets become more similar with each other. The model of the decision tree consists of sets and rules for division of a heterogeneous population into a number of smaller homogenous group with respect to some given variable. In our case, the SPSS tool is used and analysis is done on a category of products as a dependent variable, the sales channel (category) as an independent variable and value as a dependent variable that is influenced by everything.

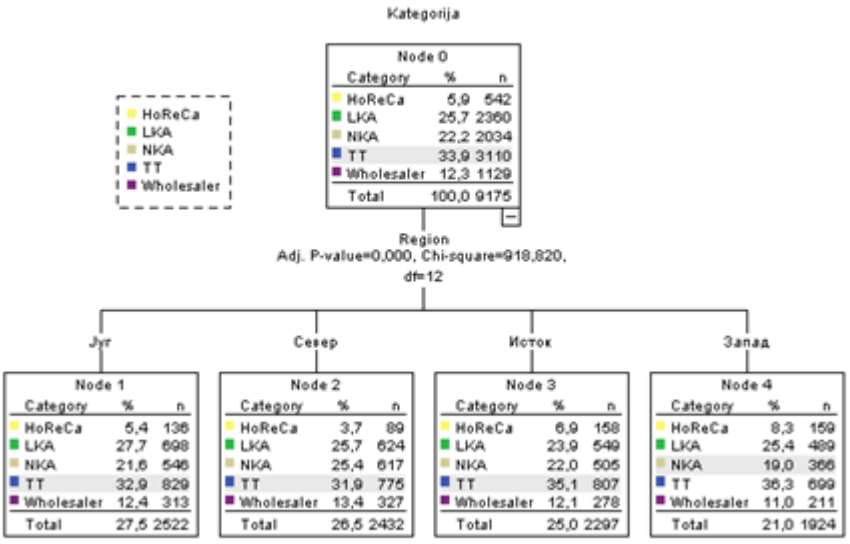


Fig. 5. : Decision tree in SPSS

Through the decision tree it can be seen that there is a regular division of the sales by region. The category buyers are differently arranged by region, where in the north and south region there are changes that benefit NKA and LKA. Based on the derived trend, it can be assumed that the same situation will be present in the other region in the future. This conclusion is based on the rapid development of LKA buyers, who are both in the segment of organized trade and NKA.

5 Conclusion

The giant competition, which in some respect is caused by the globalization, forces every company to start using methods for data mining with the goal to become competitive and to respond correctly to the market challenges. The modern management in no means can let decision making without prior analysis of the result obtained with data mining. To enable this, it is important to create a system that can respond to the current flows. The system will contain hardware, software and hiring computer scientists that will build and constantly improve it. It is important to surpass the classical systems of information in the form of printed reports of the type cards or search of buyers, top buyers, top products and so on. The aim is to implement a system which with the use of the techniques of data mining will enable companies to obtain knowledge and opportunities of prediction of the flows in the future with good probability.

6 Reference

1. Q. M. Mark Whitehorn and M. Keith Burns, "Microsoft Corporation," 07 2008. [Online]. Available: http://technet.microsoft.com/en-us/library/cc719165(v=sql.100).aspx. [Accessed 27 03 2012].
2. G. S. L. Michael J.A. Berry, Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Wiley Publishing, Inc., Indianapolis, Indiana, 2004.
3. "The Transition of Data into Wisdom," 20 03 2012. [Online]. Available: http://www.information-management.com/news/2784-1.html.
4. D. C. A. M. Gene Bellinger, "Data, Information, Knowledge, and Wisdom," [Online]. Available: http://www.systems-thinking.org/dikw/dikw.htm. [Accessed 16 03 2012].
5. Dijcks, Jean Pierre, "Oracle: Big Data for the Enterprise," Oracle, (2012).
6. OLTP vs. OLAP," [Online]. Available: http://datawarehouse4u.info/OLTP-vs-OLAP.html. [Accessed 12 03 201].
7. Brian Larson, Delivering business intelegence with Microsoft SQL Server 2008, Copyright © 2009 by The McGraw-Hill Companies.
8. Javier Torrenteras and Carlos Martinez , SSAS Cube Exploration: Digging Through the Details with Drillthrough..
9. Ian H. Witten, Eibe Frank, Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition, Copyright © 2011 Elsevier Inc.
10. Art Tennick, Practical DMX Queries for Microsoft® SQL Server® Analysis Services 2008, Copyright © 2011 by The McGraw-Hill Companies.