Organizational development through Business Intelligence and Data Mining

Denis-Cătălin ARGHIR, Ioana-Gilia DUŞA, Miruna ONUŢĂ
The Bucharest University of Economic Studies, Romania
arghir.denis@gmail.com, ioanagd94@gmail.com, miruna.onuta@gmail.com

The article presents the concept of Business Intelligence and their influence on decision making. Examining Business Intelligence systems was accomplished by theoretically comparing of four systems: Microsoft Power Bi, IBM Cognos, Oracle BI, and SAS, focusing on "functionality", "performance", "usage" and "cost" criteria. Functionality testing was done through the Power BI system using a HORECA industry dataset, namely a café retailer. On this dataset has been applied data mining concepts as cluster analysis, KNN classification analysis, and association study, to determine the frequently encountered templates, to categorize buyers into various key categories, and to help the business thrive.
***Keyword**: Business Intelligence, Power BI, Data Mining, Apriori Algorithm, Cluster Analysis, KNN Analysis.*

1 Introduction

The roots of the business intelligence (BI) concept dates back to the nineteenth century when the term BI was originally invented by Richard Millar Devens in the paper "Encyclopaedia of Commercial and Business Anecdotes" published in 1865.
According to Devens [1], the concept was used to describe how Sir Henry Furnese's banker was able to make a profit by receiving information about the banking environment and how he acted before his competitors.
The capacity to collect information and to react on the basis of these denotes the ability showed by Furnese and which underlies the concept of Business Intelligence today.
The next century's development expanded and refined business by first introducing the term "Business Intelligence" in 1989 by Howard Dresner from Gartner Group, which defined this concept, according to [2], as "a method of improving decision-making through the use of fact-based support systems".
Although it is related to enterprise applications, Business Intelligence is not a product or a system, it is a concept that shelters architectures, applications, and databases. Its purpose is to access user's data from an organization as easy as possible by interactive, in real-time access of databases, and also manipulation and analysis of them.
By analysing historical data, BI performs a valuable insight into business activities and business situations, and managers are actually assisted in making decisions through essentials information, including those behavioural and of forecasting.
In the current sense, this term denotes a set of concepts and methods used to improve the quality of business decision-making process and represents a platform for presenting information in a correct, useful and capable way to support the daily activities and decisions of every person in management positions in order to choose the most efficient alternatives.
Business Intelligence is made up of a series of applications and technologies that help gather, store, query, report, and analyse large volumes of data, and provide access to necessary data in company decision-making processes by obtaining analyses and reports.
With today's BI solutions, managers can analyse data directly without needing help from IT staff and without waiting for complex reports to run. This democratization of access to information helps users make informed decisions based

on concrete facts - not on suspicions and instincts.

2. Fields of application of BI

The effects of using a BI system are stunning, because it produces the needed information, at the time it is necessary, providing one of the prerequisites for business success. BI is the art of knowing and harnessing information, gaining competitive advantages.

BI can provide answers to the core issues of an organization, helping it making good decisions to resolve it. Finding answers is based on analysing and comparing historical data, both created within the organization, and data from external sources.

The providing benefits of the BI system, unconcerned of the field of activity, are varied, for example: a producer can quickly find out the need for materials, raw materials or stock based on past sales; a sales manager can create more profitable sales plans following the evolution of the previous period; a distributor can find out the most profitable distribution channels; a service provider has the possibility to anticipate and identify loyalty programs.

3. The architecture of a BI system

The architectural model of a BI system can include the following components:

- Data source - can be extracted from various sources or systems, such as Customer Relationship Management (CRM), Enterprise Resource Planning (ERP), relational databases, Excel spreadsheet files, Comma Separated Values (CSV) or text files (TXT); Once the data has been extracted from external sources and has been transformed according to operational needs, the data is loaded to achieve the final goal, this process is called *Extract Transform Load* (ETL);
- Data processing - once the data has been loaded to the final target, either Data Warehouse or Data Mart can be processed, can be added a series of new data, can be maintained records in a form of data logging;
- Analysis and data presentation - there are various analysis tools, analyses can be made using applications such as:
  - OLAP – useful for dynamic data analysis, fast access to a large amount of data, synchronization of data sources from multiple databases, historical analysis based on time series;
  - DATA MINING – useful for analysing large data sets in order to identify models and relationships for establishing future tendencies - clustering, association, classification;
  - DASHBOARD - useful for a quick view of performance indicators relevant to a business process.

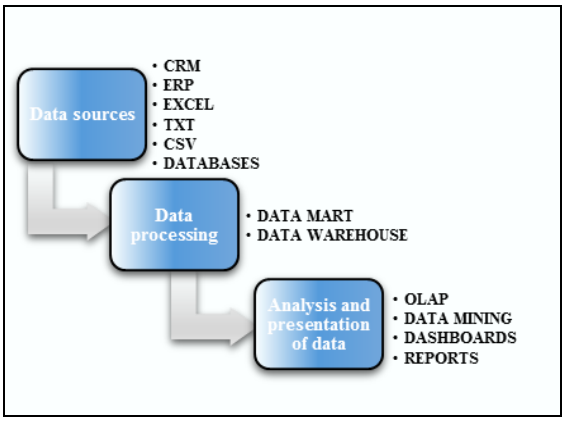


Fig. 1. The architecture diagram of a BI system

In a survey conducted in 2018 [3] on a sample of 600 companies from various industries interested in business intelligence software, has been achieved the top of the most wanted functionalities, which is as follows:

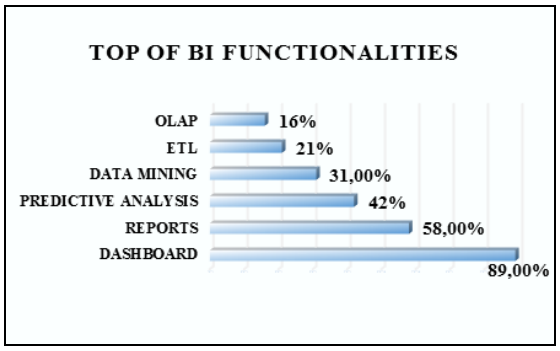


Fig. 2. Top of the most wanted BI features

4. Business Intelligence tools

By using BI tools can be obtained useful information that allows the user to understand at a glance the current state of some relevant business indicators. The tools are varied and start from simple *spreadsheets*, *Online Analytical Processing* (OLAP) - dynamic reporting solutions that allow users to interact with complex structures as time series, data trend; *Dashboard systems*; *Data Mining* – the process of extracting patterns from a large volume of data by combining statistical and artificial intelligence methods with those from database management; *Decision Engineering* – provides the framework that unites a number of good practices for organizing decision-making process; *Data Warehouse* - a data repository designed to facilitate an organization's decision-making process; *Process mining* – extracting knowledge from events recorded by the informatics system; *Exploratory Data Analysis* (EDA) – the exploration of a data set, without a strong dependence on assumptions or models, the objective being to identify patterns in an exploratory manner; *Business performance management* – is a set of managerial and analytical processes that allow the performance management of an organization to achieve one or more purposes.

Among the most popular tools can be listed Microsoft BI & Excel, Oracle BI, IBM Cognos, SAS, Qlik, Tableau, SAP Business Objects, and in terms of accessing business intelligence solutions, the most used by companies are, according to [4]:

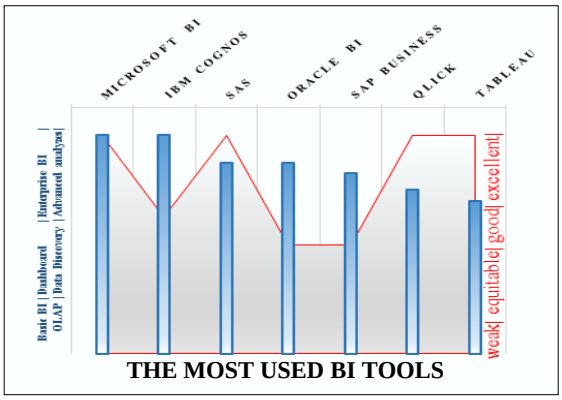


Fig. 3. Most popular BI tools

5. The purpose of using Business Intelligence

Business Intelligence has become a strategic tool to help a company to lead, optimize, discover and innovate to change the landscape of its organization. Business Intelligence systems are useful to modern businesses because they have the ability to provide a continuous flow of information and the capabilities of such a system that implements the BI concept allow employees:

- to align day-to-day operations with overall goals and strategies;
- to identify and understand the relationships between business processes and their impact on performance;
- to access relevant information for analytical responsibilities specific for the analysis;
- to analyse data from documents and to develop them very easily;
- to monitor vital business indicators, such as the current financial reports, the effectiveness, and profitability of sales departments or other relevant measurement indicators.

Business Intelligence represents the ability of an organization to think, plan, predict, solve problems, abstract thinking, understand, innovate, and learn in ways that enhance organizational knowledge, inform in the decision-making process, to

allow effective actions and to help establish and achieve business goals. The role of business intelligence is to create an informational environment in which operational data collected from transactional systems and external sources can be analysed to discover strategic business dimensions.

This information should help organisations to respond to business key issues, make predictions, and act on real-time data to improve the quality and speed of the decision-making process.

Expected benefits do not always justify investing in business intelligence technology.

This means that the development of business intelligence capabilities can only provide information-based decisions, but they cannot implement them.

An analysis of the impact of business intelligence should not focus on the impact at a given point in time, but it should be longitudinal to determine how and why it varies over time. Similarly, to enterprise resource planning systems, there are case studies that have examined long-term success or failure of business intelligence.

On a more theoretical side, in order to substantiate business intelligence research within the research information systems, rigorous preparation based on theory and impact analysis is required.

Given that the rapid evolution of technologies and managerial methods is a significant challenge for theoreticians, most of the previous business intelligence research has not had a theoretical foundation.

In this analysis, the organizational theory of information processing is used to analyse how new business intelligence technologies can favourably enhance information processing capabilities.

Business intelligence clearly reduces decision-making risk and directs operational and marketing activities to generate real value and which can be capitalized, with minimal resources.

Business intelligence projects are not meant to teach managers how to make the right decisions, but instead, help them make decisions based on facts and figures rather than assumptions.

6. The elements that turn BI into a viable business solution

By using a Business Intelligence solution, business people have access to current and quality information, highlighted in a visual and effective way.

Many organizations implement business intelligence systems, but their long-term impact on the quality of the decision-making process and of the performance consequently varies greatly.

An analysis of the factors that influence the continued use of these systems is required and is usually focused on the need for information processing in the continued use of business information and the factors that influence these needs.

Business complexity means, today, that a company needs to perform regularly complex analysis with vast amounts of data.

Many businesses are now implementing business intelligence systems to get timely information about organisational processes and company environments that combine information about past circumstances, present events and projected in future actions to answer to questions that solve various problems.

Business intelligence solutions have seen unprecedented growth over the past decade, and companies that offer them have seen spectacular growth despite all the vicissitudes of the economic environment.

Functionalities offered to users have become increasingly varied covering a wide range of needs ranging from simple tabular reports or graphical reports to the ability to track the organization's main performance indicators in a synthetic and concise manner.

To the extent that Romanian companies want to survive the pressure of European

competition, business intelligence solutions can provide them the necessary means to do this. The only remaining problems are those related to the wish of companies and those time-related because the market already has solutions for any budget.

The benefits of a business intelligence system are obvious - the analysts are optimistic, showing that in the coming years, millions of people will use day-to-day visual analysis tools and BI. The market is already saturated with the range of analytical applications available, which can carry out all sorts of analysis to support decision-making process at all levels.

Other benefits to be taken into account are:

- Reducing downtime spent with periodic reporting activities (collection of reports, consolidations, and various adjustments);
- Reducing time spent with repetitive activities;
- Reduce the role of the IT department in generating reports in favour of the end-user;
- Reduce the time needed to make a decision.

Given that the decision will be even better documented due to the quality of the information provided, we will finally be able to talk about an organization prepared to face any changes in the market, no matter how abrupt they are.

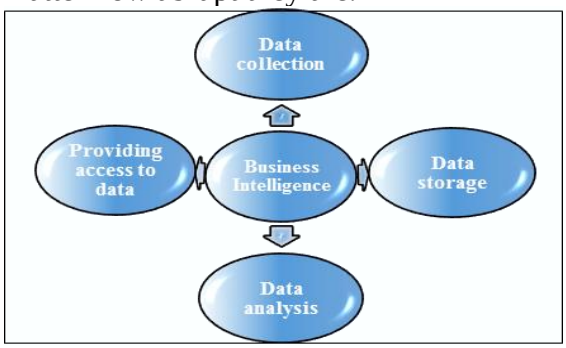


Fig. 4. Benefits provided by BI solutions

7. The Influence of Business Intelligence on Decision Making

Information from business intelligence needs to be integrated into the business process of an organization. This can be achieved by building a decision-making system. The results obtained from this business information are used by operational managers in the form of recommended actions. A prerequisite condition for efficient use is the availability of high-quality data, including good data management, covering:

- identifying users' needs;
- unification of data;
- clearing data;
- improving data quality control.

Most business intelligence initiatives have therefore focused on developing a high-quality business intelligence asset that is used instead of classical reporting systems.

The value of Business Intelligence derives from the ability to extract specific data and to adapt it from a variety of heterogeneous sources.

Managing an enterprise requires efficient data management in order to monitor activities and to evaluate the performance of different business processes.

Nowadays, there are a number of changes in the world of analysis, even for the long term, where BI is beginning to struggle to adapt to new trends. The information has become a profit centre, and processes are now customer-oriented, so business people to have a say about the mode in which analyses are predicted.

With increasing expectations, BI systems are looking to constantly improve their capabilities, because the need for faster data processing has increased, especially that the most data are not in the internal system, companies using information from outside the BI environment of enterprises.

8. Advantages of using Business Intelligence systems

At this moment, the business environment has favoured the spreading of BI applications. There are industries that budget big spending on technology purchases and BI are more or less obvious initiatives that lead to improved

profitability rates. Practically, these applications help to make wise decisions.
BI applications allow the ability to summarize and aggregate by specific and detailed categories at the same time, specific to a particular analysis or process, presenting the exact information and excluding the extra elements. Thus, a decision-maker can monitor the performance variables of a business. Example: sales per region, per product, per quarter, or product return rate for various reasons, customer behaviour analysis based on specific preference analysis.

- Possibility to retrieve data from various computer systems and to carry out a detailed analysis of them for decision support;
- Identifying and adjusting defective processes, even modifying the logic of performing certain activities to meet the company's efficiency requirements;
- Development of modules specifically designed for every single requirement;
- Permanently communication with all the centres for accurate, up-to-date and easy to follow reporting;
- The ability to develop Data Warehouse solutions that support the most demanding reporting requirements;
- Real-time updating of transactional data on which to build a decision support system;
- A business BI system is simple, visual and easy to understand, giving companies the freedom to answer questions immediately as they occur;
- The possibility of creating interactive views in just a few seconds even when working with very large volumes of information.

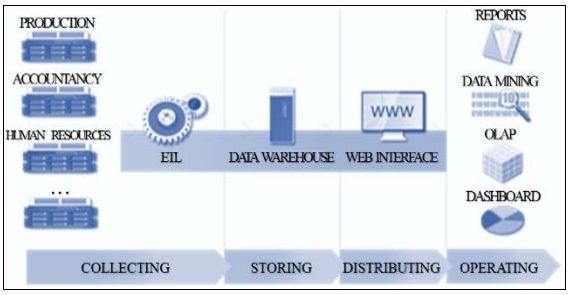


Fig. 5. Advantages of using BI systems [5]

9. Comparison between BI tools

We chose for comparison four BI tools, surprising different aspects (functionalities, performance, usage, costs) to find out which is the most suitable tool for organisational development:

Microsoft Power BI
FUNCTIONALITIES
-*exposure*: there are multiple versions: Power BI Desktop – on-premise version, Power BI service **app.powerbi.com** - Software as a Service (SaaS) online version and Power Bi Mobile - Android, iOS and Windows mobile device version; -*interactivity*: creates real-time reports, creates data analysis models, assures quality, reliability, and scalability of the reporting system; -*modelling*: dynamic and conducive environment for build comprehensive and relevant real-time reports with the possibility of automatically updating them; -*presentation*: graphics - pie, column, line, matrix, Excel reports, dashboards; -*data source*: Excel spreadsheets, Access/ SQL databases, XML files, flat text and CSV files;
PERFORMANCE
-*data integration*: creating relations between tables of type “primary key” - “foreign key”; pivot operations; -*data processing*: data can be processed from Excel spreadsheets or from a local database;
USAGE
-*interface*: depending on the type of version you are working with, data can be viewed as reports, custom dashboards, cubes - either in the form of Mobile or Desktop;
COSTS
-moderate costs;

IBM Cognos

FUNCTIONALITIES
-*exposure*: various tools: Report Studio (interactive and complex report developer tool), Query Studio (simple query and report creation tool), Analysis Studio (multi-dimensional analysis tool that provides drag & drop functionalities for exploration, analysis and comparing large data in a very short time); -*interactivity*: allows filtering, sorting, performing additional calculations, transforming values into graphs and charts; -*modelling*: provides a supportive environment for planning, budgeting, making forecasts and reliable plans in a short time; -*presentation*: export into images, Excel, PPT, PDF; -*data source*: Excel spreadsheets, XML files, flat text and CSV files;
PERFORMANCE
-*data integration*: individual queries can be joined using SQL commands; -*data processing*: data is delivered from cubes to memory for improved performance;
USAGE
-*interface*: provides a web-based architecture with advanced creative capabilities; the dashboard allows desktop customization and access to content;
COSTS
-lower costs than other traditional BI products;

SAS
FUNCTIONALITIES
-*exposure*: various visualization capabilities implemented in the product suite, interface that allows interaction with charts; -*interactivity*: users interact with a dashboard - Dashboard Builder;
-*modelling*: interaction with predefined algorithms and pre-built models for modelling datasets; -*presentation*: export as Excel files, TSV (Tab-Separated Values), CSV etc.; -*data source*: SAS datasets, Microsoft Excel spreadsheets, flat text or CSV files;
PERFORMANCE
-*data integration*: administration of server-based libraries; -*data processing*: is achieved either on the local unit or on the server;
USAGE
-*interface*: interactive, web-based reporting interface, allows the creation of basic queries and reports;
COSTS
-high costs;

ORACLE BI
FUNCTIONALITIES
-*exposure*: multiple view capabilities-basic charts, intuitive reports, diagrams; -*interactivity*: allows filtering, creating interactive parameters; -*modelling*: predictive and reporting functions for dedicated financial planning; -*presentation*: export to PDF, RTF, XML, HTML, Excel and other formats; -*data source*: supported file types are database files, XML flow, HTTP, Web services, Oracle BI analysis, OLAP cubes, LDAP server, XML files, Excel spreadsheets;
PERFORMANCE
-*data integration*: it is created using SQL commands based on table and column queries; they are specified as well as relations between them; -*data processing*: in-memory processing;
USAGE
-*interface*: web-based interface, easy to use for creating reports, with Template Builder functionality;
COSTS
-high prices for large configurations;

By comparing the four tools, we've decided to look in-depth on the features of "Power BI"-a collection of software services, applications and connectors that work together to transform unrelated data sources into a coherent and interactive perspective.

It is a simple and fast system able to create complex analyses based on various data sources, whether simple flat files like text or CSV, Excel spreadsheets or local databases.

Robust and high-quality for organizations, ready for extensive modelling and real-time analysis as well as personalized development, Power BI enables users to easily connect to data sources, visualize and discover what is important for the common good of the business. It can also

serve as an engine for analysis and decision-making for group, divisions, or entire corporations projects. More and more companies from Romania use this tool to monitor business status using dashboards that process data in real time. The primary reason for choosing this tool is that it offers the ability to work both in cloud and on-premise, that it can easily build robust and reusable models using available data, ensuring consistency between reporting and analysis within the organization, and with the Power BI web version and it can distribute various reports in just a few seconds across the organization's departments.
Also, an important feature that has convinced companies in Romania to adopt this tool is the existence of the Mobile version, so managers, the end-users, can have a view of the data anywhere and at any time. They can view custom reports and dashboard, find important information in a due time, and act immediately to reassess situations.

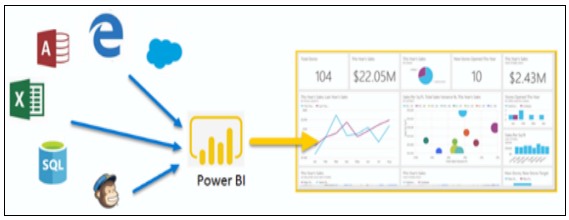


Fig. 6. Power BI Workflow

10. Case Study

Concluding that Power BI is a very useful tool for Business Intelligence analyses, we've decided to test some of the available functionalities on a specific dataset.
The analysed data is a test one and was extracted from „UCI Machine Learning Repository" (*archive.ics.uci.edu [6]*), a website that includes a structured database in various fields across the globe, with varying sizes indicators and instances, classified on different tasks for which can be used, including data mining analysis such as clustering, classification, regression, etc. We chose to study HORECA industry data, from a retailer, namely a coffee shop selling "Delicacies" products.

Why did we choose this data?

We took the data to apply business intelligence analyses, but also of data mining algorithms such as cluster analysis, KNN classification analysis and association study to extract from the dataset the most relevant information that could be used in the business environment, such as grouping data in smaller clusters in order to be able to analyse them as closely as possible, to predict the affiliation of certain classes of a test set, starting from the training set, and to track consumer behaviour on consumption of a particular category of products, in our analysis - "Delicacies", by applying different filters to see how a product sale may be influenced by another product or a mix of products.
After identifying and determining the purpose of the analysis, we imported the data we wanted in comma-separated values (CSV) format into the Power BI tool, assigning data types for each variable based on the data content.

-**Association study**: The submitted data for analysis are "Delicacies" products marketed by a café from the HORECA chain (**Ho**tels - **Re**staurants - **Ca**fes/ Coffee Shops).
For the association study, we used the "Apriori" algorithm. According to [7], by association study, it is desirable to determine consumers' consumption behaviours to find interesting and frequently encountered templates that could help the business to earn more.

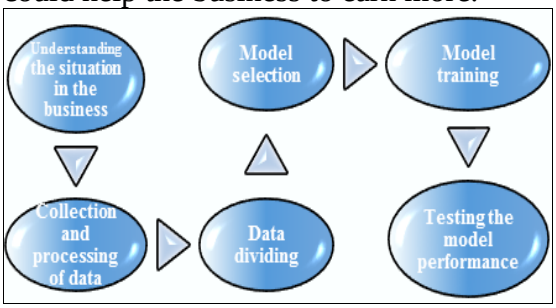


Fig. 7. Steps for implementing analysis

After understanding the need of applying the study and collecting the datasets, we switched to processing them, so, using the BI tool -we concatenated (Merge) the product type with the main ingredient; -we grouped the data set by the order_id; -we transformed the new table, containing the products grouped by the order_id, in a customized list column, with the sequence: ***(Table.Column ([Data], "Product")***; -the list we distributed on the columns (Split) after the delimiter ";" in order to be able to apply the association rules analysis.
To exemplify on a demonstration dataset, we chose 4 purchases made by coffee shop customers:

Fig. 8. Test data for the association study

By distributing the products purchased by each of our 4 customers in table form, we obtained the frequency of purchased products:

Table 1. Frequency of purchased products

ORDER	Tartă Coacaze	Apă Plată	Choux Cafea	Cafea Espresso	Desert Caise
1	1	1	1	1	0
2	0	1	0	0	0
3	0	0	0	0	0
4	1	1	0	0	1
frequency	**50%**	**75%**	**25%**	**25%**	**25%**

ORDER	Tartă Vișine	Prajitură Lămâie	Tartă Lămâie	Prajitură Mascarpone	Prajitură Căpșuni
1	0	0	0	0	0
2	0	1	1	0	0
3	0	1	1	0	0
4	1	0	0	1	1
frequency	**25%**	**50%**	**25%**	**25%**	**25%**

To determine the percentage of how often items appear together in a total of transactions, we calculated the *support* level, one of the features of the association rule:

$$\mathbf{support(X)=\frac{count(X)}{N}}$$

Where:

- *X=the frequency of occurrence of articles (products);*
- *N= total number of transactions/ orders;*

Using the previous formula on the same dataset, it appears that in the case of the combination of "Tartă-Coacăze" (Blueberries Tart) and "Apă-Plată" (Non-carbonated Mineral Water) there is a frequency of 2 appearances in the four orders analysed, resulting a support of 0.5, meaning that 50% of purchases contained the 2 products.

$$\text{support(Apă-Plată \& Tartă-Coacăze)}=$$
$$=\frac{2}{4}=0,5 \quad (1)$$

In order to determine the ratio between the number of customers who buy items that appear as a rule and the number of buyers of the items that appear in the antecedent, will be calculated with another feature of the association rule - *confidence*:

$$\mathbf{confidence(X\rightarrow Y)=\frac{support(X,Y)}{support(X)}}$$
$$\text{support(Apă-Plată)}=\frac{3}{4}=0,75 \quad (2)$$
$$\text{support(Tartă-Coacăze)}=\frac{2}{4}=0,5 \quad (3)$$
$$\text{confidence(Apă-Plată}\rightarrow\text{Tartă-Coacăze)}=$$
$$=\frac{(1)}{(2)}=\frac{0,5}{0,75}=0,(6) \quad (4)$$
$$\text{confidence(Tartă-Coacăze}\rightarrow\text{Apă-Plată)}=$$
$$=\frac{(1)}{(3)}=\frac{0,5}{0,5}=1 \quad (5)$$

We can say, according to (4), that in 60% of the cases, the buyer who bought "Apă-Plată" also bought "Tartă-Coacăze", and according to (5), 100% of the cases "Tartă-Coacăze" were bought together with "Apă-Plată".
To see to what extent the association rule is useful, we will calculate the degree of improvement - *lift*:

$$\mathbf{lift(X\rightarrow Y)=\frac{confidence(X\rightarrow Y)}{support(Y)}}$$
$$\text{lift(Apă-Plată}\rightarrow\text{Tartă-Coacăze)}=$$
$$=\frac{0,(6)}{0,5}=1,33 \quad (6)$$
$$\text{lift(Tartă-Coacăze}\rightarrow\text{Apă-Plată)}=$$
$$=\frac{1}{0,75}=1,33 \quad (7)$$

We can say that both rules are relevant, since they exceed the unit value.
To apply the association algorithm across the dataset, we used the R scripting

functionality within Power BI Desktop as follows:

Having the set called "dataset" as an input, we transform it as transactions by applying the "Apriori" function to find association rules, providing the minimum support and trust levels as parameters.

The first 100 results we recorded in a variable called "output", which is the result of the query.

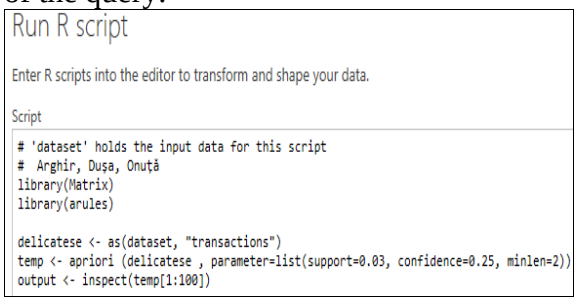
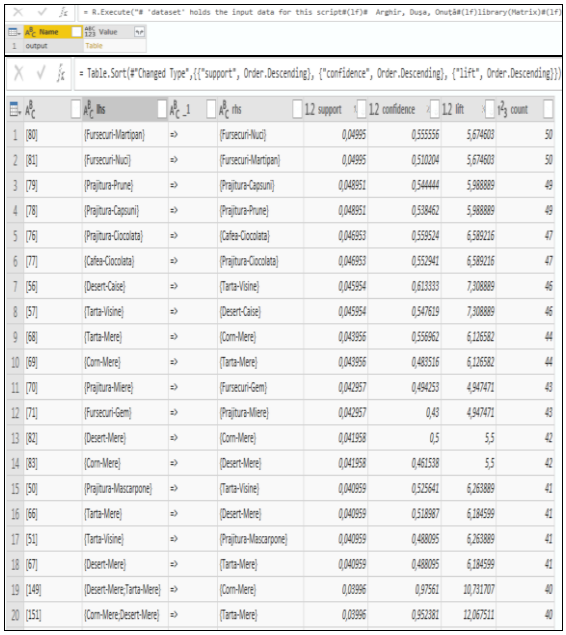


Fig. 9. The R script for obtaining the associations

Thus, the result of the query is generated in Power BI in tabular format and it can be observed the newly generated columns: "lhs" (the primary product purchased), "rhs" (the associated purchase product) and columns "support", "confidence", "lift" (degree of improvement) and "count" (number of appearances of combinations).



	lhs		rhs	support	confidence	lift	count
1	{80}	{Fursecuri-Martipan}	⇒	{Fursecuri-Nuci}	0,04995	0,555556	5,674603	50
2	{81}	{Fursecuri-Nuci}	⇒	{Fursecuri-Martipan}	0,04995	0,510204	5,674603	50
3	{79}	{Prajitura-Prune}	⇒	{Prajitura-Capsuni}	0,048951	0,544444	5,988889	49
4	{78}	{Prajitura-Capsuni}	⇒	{Prajitura-Prune}	0,048951	0,538462	5,988889	49
5	{76}	{Prajitura-Ciocolata}	⇒	{Cafea-Ciocolata}	0,046953	0,559524	6,589216	47
6	{77}	{Cafea-Ciocolata}	⇒	{Prajitura-Ciocolata}	0,046953	0,552941	6,589216	47
7	{56}	{Desert-Caise}	⇒	{Tarta-Visine}	0,045954	0,613333	7,308889	46
8	{57}	{Tarta-Visine}	⇒	{Desert-Caise}	0,045954	0,547619	7,308889	46
9	{68}	{Tarta-Mere}	⇒	{Com-Mere}	0,043956	0,556962	6,126582	44
10	{69}	{Com-Mere}	⇒	{Tarta-Mere}	0,043956	0,483516	6,126582	44
11	{70}	{Prajitura-Miere}	⇒	{Fursecuri-Gem}	0,042957	0,494253	4,947471	43
12	{71}	{Fursecuri-Gem}	⇒	{Prajitura-Miere}	0,042957	0,43	4,947471	43
13	{82}	{Desert-Mere}	⇒	{Com-Mere}	0,041958	0,5	5,5	42
14	{83}	{Com-Mere}	⇒	{Desert-Mere}	0,041958	0,461538	5,5	42
15	{50}	{Prajitura-Mascarpone}	⇒	{Tarta-Visine}	0,040959	0,525641	6,263889	41
16	{66}	{Tarta-Mere}	⇒	{Desert-Mere}	0,040959	0,518987	6,184599	41
17	{51}	{Tarta-Visine}	⇒	{Prajitura-Mascarpone}	0,040959	0,488095	6,263889	41
18	{67}	{Desert-Mere}	⇒	{Tarta-Mere}	0,040959	0,488095	6,184599	41
19	{149}	{Desert-Mere,Tarta-Mere}	⇒	{Com-Mere}	0,03996	0,97561	10,731707	40
20	{151}	{Com-Mere,Desert-Mere}	⇒	{Tarta-Mere}	0,03996	0,952381	12,067511	40

Fig. 10. Output – the result of Apriori algorithm application

With "Forced-Directed Graph 2.0.2" visualization mode in Power BI, together with the "Slicer" filter, we represented in an interactive way the relationships between the main and the related nodes, the link thickness representing the value of the support, with the possibility of selection from the drop-down list the main or associated products.

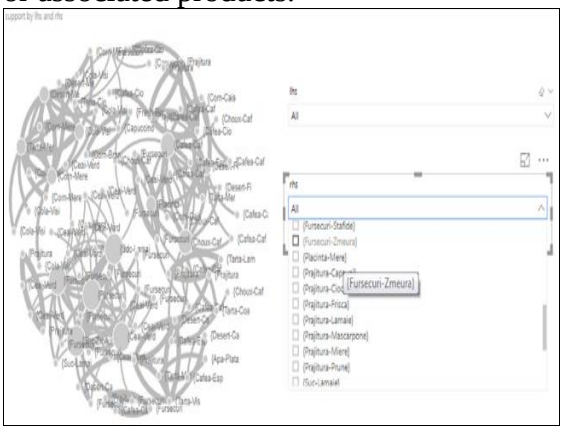


Fig. 11. Representation of associations as a graph (Forced-Directed Graph)

For example, we have selected combinations of main product associated with "Fursecuri-Zmeură" (Raspberry Cookies) and resulted in the following graph, where can be viewed at a glance which are the most preferred customer combinations of products (the most thicker lines):

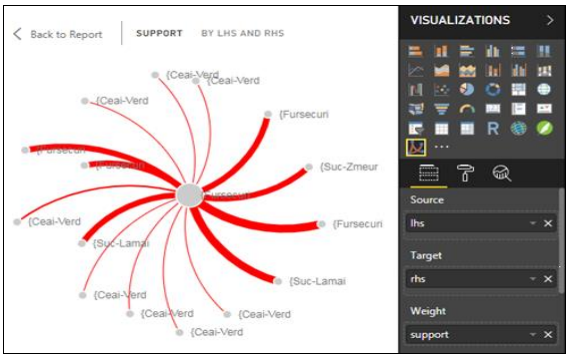


Fig. 12. Extracting all combinations of purchases that have associated the product „Fursecuri-Zmeură"

The strongest link, in the sense of the highest support, is in the case of the combination "Suc-Lămâie" (Lemon-Juice) associated with "Fursecuri-Zmeură" with a support level of 0,03, meaning that 3% of

the purchases (31 purchases) contained the 2 products.

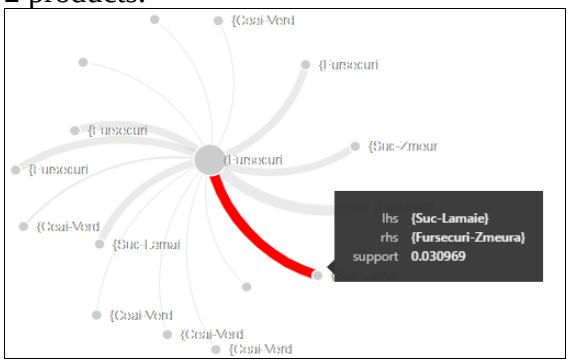


Fig. 13. Detailing the relationship with the highest support level 0,03% „Suc-Lămâie" associated with „Fursecuri-Zmeură"

Another representation can be made using the "R script visual" view:

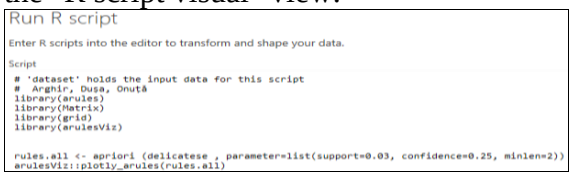


Fig. 14. R script for obtaining a scatterplot representation

Thus, the three dimensions - "support", "confidence" and "lift" can be viewed in a single graph that can be detailed with a simple click:

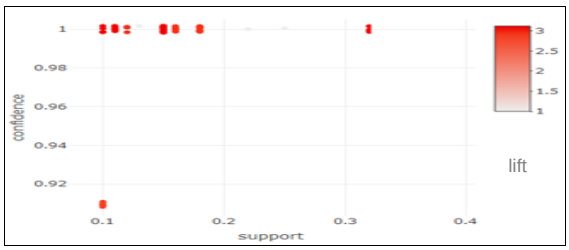


Fig. 15. Scatter-plot representation for product associations (support, confidence and lift)

Following the association study, the management of the unit can create promotional packages that could have positive results in increasing sales of "Delicacies" products.

It follows the use of another data mining analysis that is applicable to business intelligence:

-Cluster Analysis: According to [8] the purpose of clustering occurs from the need to fit, group or classify certain entities or objects in the form of categories or classes, of which delimitation must be very clear. All classification techniques are known as the theory of form recognition.

Cluster analysis is a classification technique characterized by the fact that affecting the forms or objects in clusters or groups is made progressive and without aprioristic knowing the number of classes, depending on verification of two fundamental criteria:

- the objects or forms classified in each class needs to be as similar in terms of certain characteristics;
- objects classified in a class needs to differentiate as much as possible from objects classified in any of the other classes.

The K-means algorithm is a method of dividing a dataset into a specified cluster number introduced by the user (k). This cluster analysis method aims to divide "n" observations in "k" classes where each observation belongs to the class with the closest average.

More specifically, the algorithm assigns "k" centres of the classes (centroids) in "n" points.

The steps of the k-Means algorithm:

- firstly, we chose the initial "k" group centres;
- the "k" groups are created by assigning each instance of that group to whose centre it is closest;
- we recalculate the centres in relation to the new composition of the groups;
- we repeat the algorithm with the second step if the centres have moved;

Table 2. Test data for cluster analysis

ID	PRODUCT	Fats	Carbs	Proteins	Calories	Sugars
0	Căpşuni	1.83	23.88	1.85	110	18.44
1	Lămâie	6.84	23.20	2.02	170	15.38
2	Frişcă	8.15	23.37	2.40	171	14.48
3	Mascar-	7.54	19.46	1.85	151	12.07

	pone					
4	Ciocolată	4.68	21.26	1.62	129	12.50
5	Miere	2.84	28.75	2.25	140	15.16
6	Ciocolată neagră	0.43	23.38	2.10	100	16.40

For exemplification, we chose a demonstrative dataset with 6 records, representing 6 products marketed by the café, with several nutritional features (fats, carbohydrates, proteins, calories and sugars) to promote them in 2 packages, generically called "*dietetic*" and "*caloric*", so we want to create two clusters with these features.

We have selected from the dataset the product with id "2", as having the highest values, and the product with id "6" as having the smallest values around which the two clusters will be created.

Table 3. The chosen values as initial centroids for k=2 clusters

Cluster	Product_ID	Average (centroid)
1	6	(0.43, 23.38, 2.10, 100, 16.40)
2	2	(8.15, 23.37, 2.40, 171, 14.48)

Then, we calculate the distance of each element from the two clusters, using the Euclidean distance:

$$d(p,q)=\sqrt{(p1-q1)^2+(p2-q2)^2+\cdots+\sqrt{+(pi-qi)^2+\cdots+(pn-qn)^2}}$$

$$d(p,q)=\sqrt{(0,43-1,83)^2+(23,38-23,88)^2+\sqrt{+(2,10-1,85)^2+(100-110)^2}\sqrt{(16,40-18,44)^2}}=10,31 \quad (7)$$

$$d(p,q)=\sqrt{(8,15-1,83)^2+(23,37-23,88)^2+\sqrt{+(2,40-1,85)^2+(171-110)^2}\sqrt{+(14,48-18,44)^2}}=61,46 \quad (8)$$

By comparing the distances to the two clusters, we decide that the product with id "0" is closer to Cluster 1. At this point, we need to calculate the new average (media) value of Cluster 1 (centroid):

Media:

$$(0,43+1,83)/2=1,13$$
$$(23,38+23,88)/2=23,63$$
$$(2,10+1,85)/2=1,98$$
$$(100+110)/2=105$$
$$(16,40+18,44)/2=17,42$$

Applying formulas (7), (8) and (9) for the 7 products, are obtained the following solutions:

Table 4. Categorizing products into the 2 clusters

Cluster	Product_ID	Average (centroid)
1	6	(0.43, 23.38, 2.10, 100, 16.40)
	6,0	(1.13, 23.63, 1.98, 105.00, 17.42)
	6,0,4	(2.91, 22.45, 1.80, 117.00, 14.96)
2	2	(8.15, 23.37, 2.40, 171, 14.48)
	2,1	(7.50, 23.29, 2.21, 170.50, 14.93)
	2,1,3	(7.52, 21.37, 2.03, 160.75, 13.50)
	2,1,3,5	(5.18, 25.06, 2.14, 150.38, 14.33)

Thus, we obtain the categorization of the 7 products in the two clusters: Cluster 1 *(products 6, 0, 4)* and Cluster 2 *(products 2, 1, 3, 5)*.

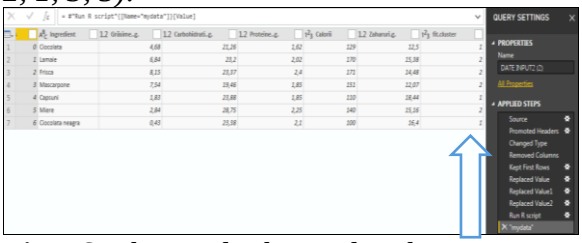


Fig. 16. The result obtained with Power BI on the test set

Returning to the whole dataset, applying the K-Means algorithm with Power BI's R scripting functionality, we have grouped the 50 products in 3 classes, which we can use to promote the products of the café in 3 packages: "*dietetic*", "*medium*" and "*caloric*".

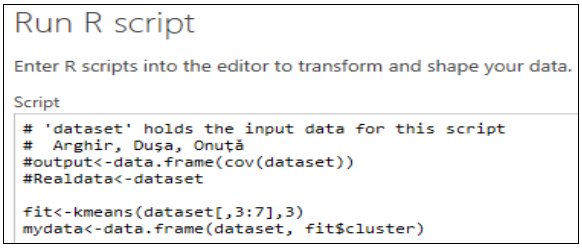
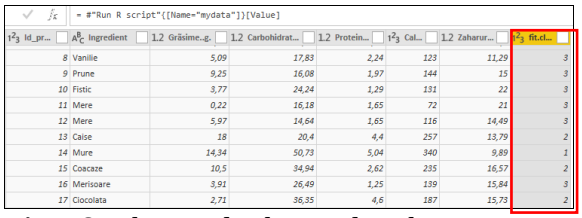


Fig. 17. R script to obtain categorisation of products ranges in 3 clusters

We obtained the categorisation of 50 products in 3 clusters:



Id_pr...	Ingredient	Grăsime...g.	Carbohidrat...	Protein...	Cal...	Zaharur...	fit.cl...
8	Vanilie	5,09	17,83	2,24	123	11,29	3
9	Prune	9,25	16,08	1,97	144	15	3
10	Fistic	3,77	24,24	1,29	131	22	3
11	Mere	0,22	16,18	1,65	72	21	3
12	Mere	5,97	14,64	1,65	116	14,49	3
13	Caise	18	20,4	4,4	257	13,79	2
14	Mure	14,34	50,73	5,04	340	9,89	1
15	Coacaze	10,5	34,94	2,62	235	16,57	2
16	Merisoare	3,91	26,49	1,25	139	15,84	3
17	Ciocolata	2,71	36,35	4,6	187	15,73	2

Fig. 18. The result obtained with Power BI on the whole dataset

We used a Treemap representation to highlight the three newly formed clusters, grouping being done by the "cluster" attribute, the details being represented by "product names" and the values being taken from the attributes "calories", "fats", "proteins" and "sugars".

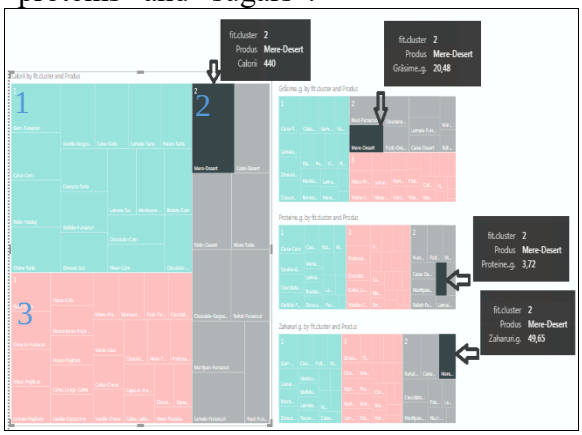


Fig. 19 Treemap representation of obtained clusters

Cluster 3 contains the products with the lowest caloric value, and can call them "dietetic", Cluster 1 is the next one, which can be called "medium" from the caloric point of view, and Cluster 2 contains the most "caloric" products.

For easier visualization of grouping in clusters, we created a "Clustered Bar Chart" for four of the product features and two "Slicer Drop Down List" to compare the average values of "calories", "fats", "proteins" and "sugars" from two clusters at the same time. For not interacting Slicers in the wrong way with all objects, we have blocked -Edit interactions: None- for charts that should remain unchanged:

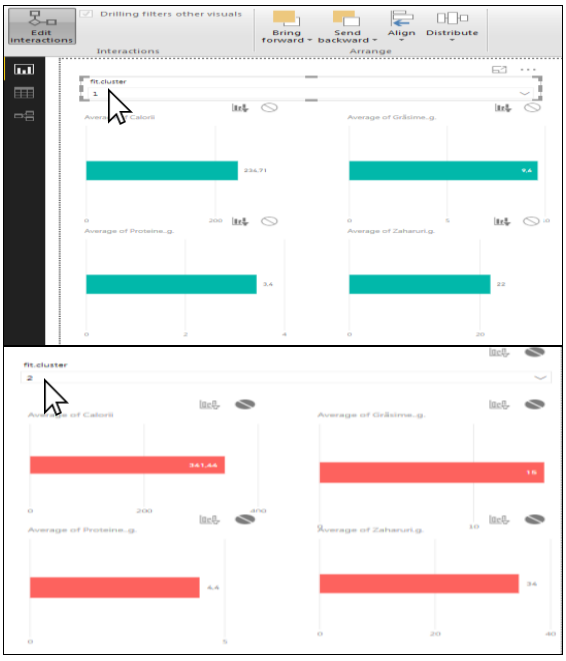


Fig. 20. Comparing the average nutrient values of products in Cluster 1 versus Cluster 2

-The classifier based on K nearest neighbours: to classify the instances, we choose the KNN (K-nearest-neighbours) algorithm, being, according to [9], a simple classification method based on placing all instances in an n-dimensional space.

As mentioned above, we have 3 groups in which the products are divided, namely packages "*dietetic*", "*medium*" and "*caloric*". The coffee shop wants to introduce 16 new products by assigning one of the three attributes according to nutritional value, taking into account the current menu that is already listed with nutrition attributes (34 products).

Thus, we want to identify the proximity of newly introduced products to existing categories.

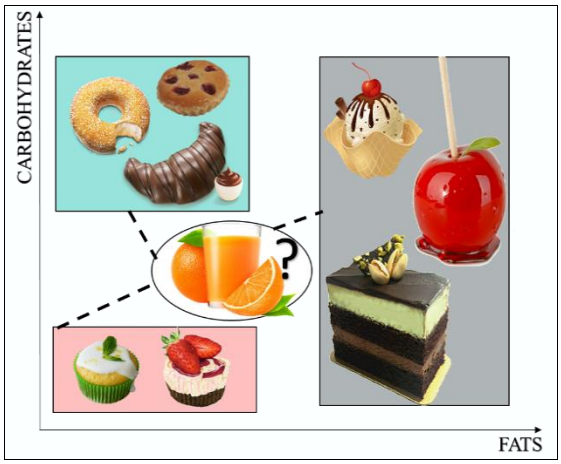


Fig. 21. Graphic representation of the distance between products

To calculate, for example, the category in which the "Fresh-Portocale" *(Orange Juice)* will be placed in the existing range, we will have to calculate the distance between the new introduced product and each products existing in the café menu, using the Euclidean metric, following the reasoning presented at (7), (8); the calculation formula for the n-dimensional case is:

$$d(p,q)=\sqrt{(p1-q1)^2+(p2-q2)^2+\cdots+(pi-qi)^2+\cdots+(pn-qn)^2}$$

By comparing the results obtained, we will choose the smallest value, and we can deduct that the newly introduced product belongs to the category in which the product is at the shortest distance.

To apply the KNN algorithm, we return to the existing dataset, grouped into the three categories "1", "2", "3". We will pick the first 34 records to form an algorithm training set, and the remaining 16 records will be used as a test set to predict belonging to one of the groups. To make the group clearer, we will change the column from numeric type into text, replacing the values: "1" = "medium", "2" = "caloric", "3" = "dietetic".

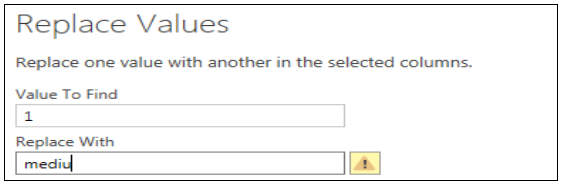


Fig.22. Replacing numerical values with text values

We will normalize the data to bring them to the same scale. For this we will use the R scripting tool, defining a function called "function (x)", which we will apply to the numerical columns in the known dataset, called "dataset":

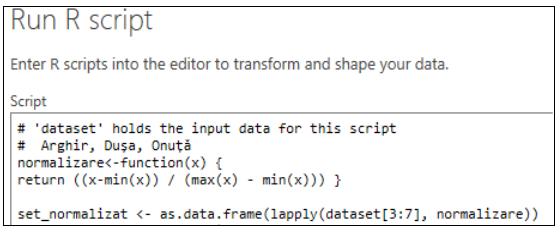


Fig. 23. R Script for dataset normalization

The "lapply" R function has two input variables, *the initial dataset*, and *the normalization function*, building a new normalized set.

The data will be divided into two - in a training set (the first 34 records) and a test set (the next 16 records) on which we will apply the prediction model. We will set prediction labels for the result as being column 8 of the dataset.

```
set_antrenare<-set_normalizat[1:34, ]
set_test<-set_normalizat[35:50, ]
set_antrenare_eticheta <- dataset[1:34, 8]
set_test_eticheta <- dataset[35:50, 8]
```

Fig. 24. R script for obtaining the training set and the test set

Once the data has been prepared, we will apply the KNN algorithm, for which is required "class" library. The R function "KNN" has as input *the train set*, *the test set*, *the prediction labels* (cl) and *the number k* - the closest neighbour for which we calculate the square root, more precisely k will take the value $\sqrt{34}\approx 6$.

```
library("class")
set_test_predictie <- knn(train=set_antrenare, test=set_test,
                          cl=set antrenare_eticheta, k=6)
output<-set_test
output$result<-set_test_predictie
```

Fig. 25. Applying the KNN algorithm based on the training set

We will have four Power BI outputs, namely: -the initial normalized dataset, -a sequence from the normalized set (the first

34 records representing the training data), - another sequence from the normalized set (the last 16 records representing the test set).

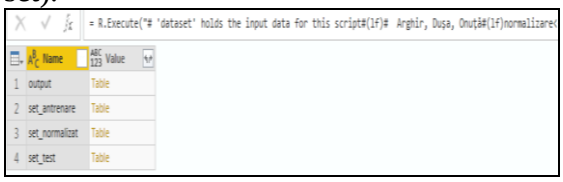
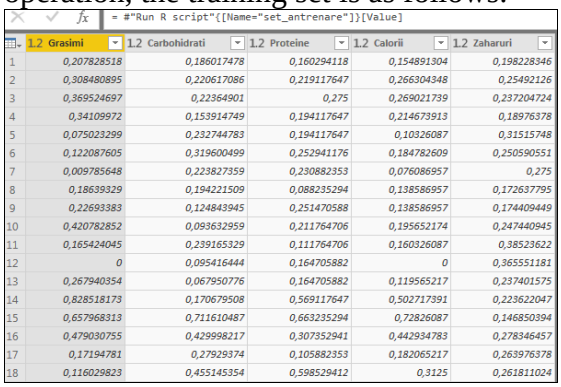


Fig. 26. Power BI Output - normalized set and prediction

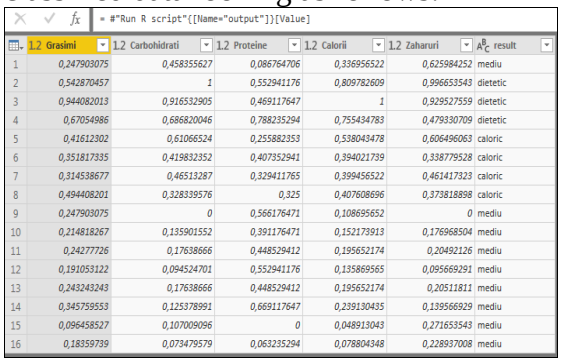
Following the data normalization operation, the training set is as follows:



	Grasimi	Carbohidrati	Proteine	Calorii	Zaharuri
1	0,207828518	0,186017478	0,160294118	0,154891304	0,198228346
2	0,308480895	0,220617086	0,219117647	0,266304348	0,25492126
3	0,369524697	0,22364901	0,275	0,269021739	0,237204724
4	0,34109972	0,153914749	0,194117647	0,214673913	0,18976378
5	0,075023299	0,232744783	0,194117647	0,10326087	0,31515748
6	0,120087605	0,319600499	0,252941176	0,184782609	0,250590551
7	0,009785648	0,223827359	0,230882353	0,076086957	0,275
8	0,18639329	0,194221509	0,088235294	0,138586957	0,172637795
9	0,22693383	0,124843945	0,251470588	0,138586957	0,174409449
10	0,420782852	0,093632959	0,211764706	0,195652174	0,247440945
11	0,165424045	0,239165329	0,111764706	0,160326087	0,38523622
12	0	0,095416444	0,164705882	0	0,365551181
13	0,267940354	0,067950776	0,164705882	0,119565217	0,237401575
14	0,828518173	0,170679508	0,569117647	0,502717391	0,223622047
15	0,657968313	0,711610487	0,663235294	0,72826087	0,146850394
16	0,479030755	0,429998217	0,307352941	0,442934783	0,278346457
17	0,17194781	0,27929374	0,105882353	0,182065217	0,263976378
18	0,116029823	0,455145354	0,598529412	0,3125	0,261811024

Fig. 27. Normalized dataset

Applying the KNN algorithm, we obtain the prediction of belonging of the 16 new products to the three defined groups; the classified data looking as follows:



	Grasimi	Carbohidrati	Proteine	Calorii	Zaharuri	result
1	0,247903075	0,458355627	0,086764706	0,336956522	0,625984252	mediu
2	0,542870457	1	0,552941176	0,809782609	0,996653543	dietetic
3	0,944082013	0,916532905	0,469117647	1	0,929527559	dietetic
4	0,67054986	0,686820046	0,788235294	0,755434783	0,479330709	dietetic
5	0,41612302	0,61066524	0,255882353	0,538043478	0,606496063	caloric
6	0,351817335	0,419832352	0,407352941	0,394021739	0,338779528	caloric
7	0,314538677	0,46513287	0,329411765	0,399456522	0,461417323	caloric
8	0,494408201	0,328339576	0,325	0,407608696	0,373818898	caloric
9	0,247903075	0	0,566176471	0,108695652	0	mediu
10	0,214818267	0,135901552	0,391176471	0,152173913	0,176968504	mediu
11	0,24277726	0,17638666	0,448529412	0,195652174	0,20492126	mediu
12	0,191053122	0,094524701	0,552941176	0,135869565	0,095669291	mediu
13	0,243243243	0,17638666	0,448529412	0,195652174	0,20511811	mediu
14	0,345759553	0,125378991	0,669117647	0,239130435	0,139566929	mediu
15	0,096458527	0,107009096	0	0,048913043	0,271653543	mediu
16	0,18359739	0,073479579	0,063235294	0,078804348	0,228937008	mediu

Fig. 28. The obtaining prediction

In order to verify the accuracy of the prediction model, we will make a matrix representation - confusion matrix. Thus, the first position shows the number of true-positive instances, in which case a "caloric" product is classified correctly as "caloric".

On the main diagonal continues with correct classified instances in which "dietetic" delicacies are predicted as "dietetic" and those "medium" caloric are classified as "medium"; only 1 product was predicted as belonging to the "medium" category, but it belongs to "calorific" category.

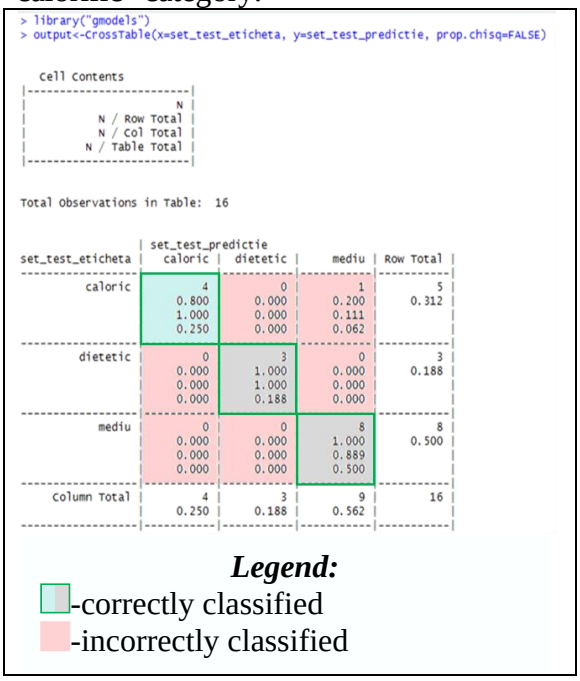


Fig. 29. The accuracy of the prediction model with the KNN algorithm

The quality of a classifier from the perspective of the correct identification of a class is measured using the information in the confusion matrix that contains:

- True Positive (TP) - a product belonging to the class was recognized as belonging to the class;
- True Negative (TN) - a product that does not belong to the class was not recognized as belonging to the class;
- False Positive (FP) - a product that belongs to the class was not recognized as belonging to the class,
- False Negative (FN) - a product that does not belong to the class was recognized to belongs to the class;

Table 5. Confusion matrix

	belongs		does not belong	
recognized	**TP**		**FN**	
	4	4	(0+1)	1
not recognized	**FP**		**TN**	
	(0+0)	0	(3+8)	11

Based on these values, a number of other measures can be calculated:

- **True Positive Rate/** Sensitivity/ Recall/ represent the number of delicacies correctly identified as positive out of total true positives:

$$TPR = \frac{TP}{(TP+FN)} \rightarrow TPR = \frac{4}{4+1} = 80\%$$

- **False Positive Rate** represent the number of delicacies erroneously identified as positive out of total true negatives:

$$FPR = \frac{FP}{(FP+TN)} \rightarrow FPR = \frac{0}{0+11} = 0\%$$

- **False Negative Rate** represent the number of delicacies erroneously identified as negative out of total true positives;

$$FNR = \frac{FN}{(FN+TP)} \rightarrow FNR = \frac{1}{1+4} = 20\%$$

- **True Negative Rate**/ Specificity represent the number of delicacies identified as negative out of total true negatives:

$$TNR = \frac{TN}{(TN+FP)} \rightarrow TNR = \frac{11}{11+0} = 100\%$$

- **Overall Accuracy** is the ratio between the correctly classified delicacies (the values on the main diagonal) and the total number of test values:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \rightarrow ACC = \frac{4+3+8}{16} = \frac{15}{16} = 91,75\%$$

- **Error Rate** is the ratio between the number of incorrectly classified delicacies and the total number of test values:

$$ERR = \frac{FP+FN}{TP+FP+TN+FN} \rightarrow ERR = \frac{1}{16} = 0,062$$

The value of global accuracy that is close to 100% and the error rate close to 0 shows that the prediction model is a very good one.

The result of the prediction compared to the actual data, we represented as a table, with help of two Slicer filters (that do not interact one with other); so we could put two filtering conditions on junction data, in the first Slicer (the real category) by selecting "caloric" and in the second (the predicted category) by selecting "medium" we get the only solution that was not correctly predicted by the model classification, namely "Corn-Ciocolată" (Chocolate Croissant), which was predicted "medium", in reality being in the "caloric" category.

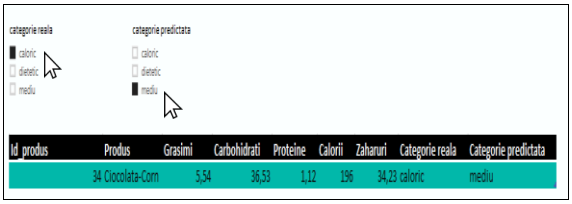


Fig. 30. Selecting an erroneous prediction using Power BI

12. Conclusions and future works

The purpose of applying data mining concepts was to extract the most relevant information and to predict on its basis the belonging of data to certain classes.

Following the application of the "Apriori" algorithm, in were we calculated the support level and the confidence level, we can conclude that in 100% of the cases, customers who bought a specific product - "Tartă-Coacăze" (Blueberries Tart), mandatory bought a second product namely "Apă-Plată" (Non-carbonated Mineral Water).

Instead, only 3% of buyers preferred the combination of "Suc-Lămâie" (Lemon Juice) and "Fursecuri-Zmeură" (Raspberry Cookies). Through cluster analysis, in which we wanted to categorize the same entities in representative categories, were obtained three consumer packages that could be addressed to a variety of buyers: "dietetic", "medium", "caloric".

These packages were obtained by analysing the components: "product name", "calories", "fats", "proteins" and "sugars".

Through the classifier based on the K-nearest-neighbours we wanted to determine in which of the three categories we could introduce a new product. This is very useful for the majority of businesses, as it manages to determine various classifications based on aprioristic knowledge.

In conclusion, we can say that the field of business intelligence is constantly developing, and by transforming raw and unstructured data, can be obtained very useful information.

This process can be accomplished through various data mining algorithms that can take the form of interactive reports that can be presented in a friendly manner, helping the businessman consolidate his future decisions.

In the near future, we want to collect a dataset of major economic interest from Romania, to be subjected to several analyses to offer a wide range of options both in the area of income and profitability and financial analysis for identifying problems and improving these vital indicators.

References

[1] Miller Devens Richard, „Cyclopaedia of Commercial and Business Anecdotes...” Nabu Press, 2011, ISBN: 978-1248003671;
[2] D. J. Power „A Brief History of Decision Support Systems, version 4.0". DSSResources.COM, 2007.
[3] https://selecthub.com/business-intelligence/key-types-business-intelligence-tools/, „What are the Different Types of Business Intelligence Tools?”, Accessed February 28, 2019.
[4] https://www.passionned.com/wp/wp-content/uploads/passionned-parabola-bi-analytics-2019.png?x18199, “References about the most used BI tools”, Accessed February 29, 2019.
[5] http://www.pentalog.ro/html/pentalog/corporate/images/corporate/schema_avantaje_bi, References about BI advantages, Accesed March 2, 2019.
[6] https://archive.ics.uci.edu/ml/index.php, “Seturi de date pentru prelucrări de data mining şi business intelligence”, Accessed March 3, 2019.
[7] http://ip.ase.ro, References about the application of data mining algorithms, Accessed March 7, 2019.
[8] http://ip.ase.ro/Ad11.pdf, References about cluster analysis, Accessed March 9, 2019.
[9] https://www.academia.edu/4424156/, „O îmbunătăţire a performanţelor algoritmului KNN în sistemele de recomandare pe web”, Accessed March 10, 2019.
[10] https://en.wikipedia.org/wiki/IBM_Cognos_Analytics, „IBM Cognos Analytics”, Accessed March 3, 2019.
[11] https://support.sas.com/documentation/onlinedoc/portal/index.html, Accessed March 3, 2019.
[12] https://comparisons.financesonline.com/microsoft-power-bi-vs-sas-business-intelligence, Accessed March 3, 2019.
[13] https://comparisons.financesonline.com/ibm-cognos-vs-oracle-bi, March 3, 2019.
[14] https://www.trustradius.com/compare-products/ibm-cognos-vs-oracle-business-analytics, Accessed March 4, 2019.
[15] https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68, Accessed March 9, 2019.

Denis-Cătălin ARGHIR has graduated the Faculty of Cybernetics, Statistics and Economics Informatics in 2017. He holds a bachelor degree in Economics Informatics and now he is following the master's programme of Database-Support for Business and Psycho-Pedagogical 2nd Module. He followed a series of practice programs at a number of top companies as Microsoft, IBM, High Tech System Software, Cegedim RX, EY Romania and governmental institutions within the Romanian Government to develop his work with databases, business intelligence, internet of things, .NET, Java, and Robotic Process Automation that represent his area of interests.

Ioana-Gilia DUȘA is a student in the final year of the master programme Database-Support for Business at the Faculty of Cybernetics, Statistics and Informatics Economics. She graduated the same faculty, the field of Informatics Economics in 2017 and she has been working as an ETL Developer for almost 3 years. Her interest domains related to computer science are: Data Warehouse, Business Intelligence, informatic systems, and databases.

Miruna ONUȚĂ graduated from The Faculty of Economic Cybernetics, Statistics and Economic Informatics (bachelor domain: Economic Informatics) and currently she is a student in the final year of the master programme: Database-Support for Business at the same university. Since 2018 she works as a business intelligence analyst and she is interested in: BI, databases, data warehouse, cloud computing platforms and services, ETL process.