Data Mining Process

Ljupce Markusheski, Ph. D.[1] **Igor Zdravkoski, Ph D**[2] **Miroslav Andonovski, Ph D**[3]

Faculty of Economy – Prilep, "Marksova" 133, Prilep, Republic of North Macedonia, ljmarkusoski@t.mk.

Faculty of Economy – Prilep, "Marksova" 133, Prilep, Republic of North Macedonia,
igor.zdravkoski@hotmail.com.

Faculty of Economy – Prilep, "Marksova" 133, Prilep, Republic of North Macedonia,
miroslav.andonovski@uklo.edu.mk.

Abstract: Data Mining is a powerful tool for companies to extract the most important information from their data warehouse. These tools allow you to predict future trends and behaviors in order to be able to provide activities based on specific knowledge. Such activities are much more effective, and thus, more economical. This tool allows you to obtain information that would be too time-consuming to acquire in the traditional way. At the same time, this tool allows you to obtain information that would probably be omitted by experts due to their unpredictability.

Data Mining is ready for immediate introduction to business due to three factors that are now well advanced:

- Mass gathering of information by companies
- The enormous computing power of computers
- Ready algorithms

Data Mining got its name from the similarities between searching for valuable information in large databases and searching for a new vein of ore, eg iron in the mountains. Both of these activities require a huge amount of work and a precise search to be able to find a place where real and real value is located. If we provide a database of sufficient size and quality, Data Mining will allow us to gain new business opportunities.

Keywords: Data Minig tools, databases, data warehouse, knowledge.

1. Introduction

Data Mining is a powerful tool for companies to extract the most important information from their data warehouse[1]. These tools allow you to predict future trends and behaviors in order to be able to provide activities based on specific knowledge. Such activities are much more effective, and thus, more economical. This tool allows you to obtain information that would be too time-consuming to acquire in the traditional way. At the same time, this tool allows you to obtain information that would probably be omitted by experts due to their unpredictability.[2]

Data are any facts, numbers, or text that can be processed by a computer. Many organizations accumulate vast and growing amounts of data in a variety of formats and databases. These data may be loosely grouped into three categories: operational or transactional data, such as company sales, costs, inventory, payroll, and accounting; non-operational data, such as industry sales, forecast data, and macro-economic data; and metadata, which is data about the data themselves, such as elements related to a database's design or query protocol.

The patterns, associations, and relationships among all these data can provide information. For example, analysis of retail point-of-sale transaction data can yield information on which products are selling and when. Information can then be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide

[1] Jiawei Han, Micheline Kamber, Jian Pei, DATA MINING, Concepts and Tehniques, Third Edition, Simon Fraser University Elsevier Inc., 2012, page 2.

[2]http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items to combine with promotional efforts for the best sales or profit results.[3]

Most companies have already accumulated large amounts of data. Data Mining can be quickly implemented into existing databases, which will allow you to get answers such as:, Which customers will most likely answer our offer and why?

Evolution began as information began to be collected on computer hard drives. This allowed access to information in real time and data mining according to the needs of an individual user. Data Mining has made it possible to revolutionize this process from retrospective to prospect. It allowed the current data available to predict the future, which allows immediate decision-making.

Data Mining is ready for immediate introduction to business due to three factors that are now well advanced:

- Mass gathering of information by companies
- The enormous computing power of computers
- Ready algorithms

Data Mining got its name from the similarities between searching for valuable information in large databases and searching for a new vein of ore, eg iron in the mountains. Both of these activities require a huge amount of work and a precise search to be able to find a place where real and real value is located. If we provide a database of sufficient size and quality, Data Mining will allow us to gain new business opportunities by providing the following factors:

Automated prediction of future behaviors and trends

Predicting behaviors and trends by traditional methods is very time-consuming, thanks to Data Mining the same information can be drawn straight from the database in a quick way. This is clearly seen in direct marketing. Mailing sent to customers of the past is used to predict the most profitable group of recipients in future mailings.

Automatic discovery of previously unknown models.

The Data Mining tool combing databases can detect previously unknown models, e.g. customer behavior. For example, there may be a correlation between the model of the bought car and the favorite color. In a traditional way, it is difficult to catch such correlations, especially since they first have to be predicted. Data Mining does it for us.[4]

2. Data Mining process

Data mining is a promising and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence (AI) and statistical.

Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace… etc, to increase their business efficiency. Therefore, the needs for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background. As the result, in 1990, a cross-industry standard process for data mining (CRISP-DM) first published after going through a lot of workshops, and contributions from over 300 organizations.[5]

[3] https://www.encyclopedia.com/science-and-technology/computers-and-electrical-engineering/computers-and-computing/data-mining

[4] http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

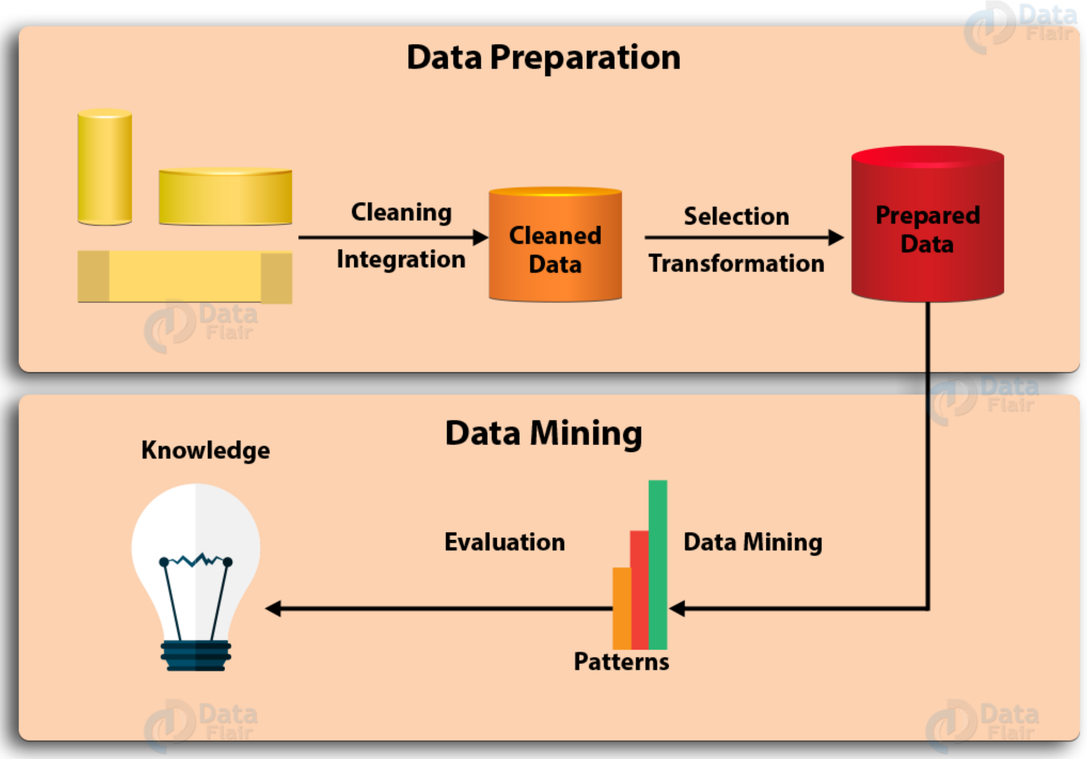
[5] http://www.zentut.com/data-mining/data-mining-processes/

http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

As we can see on diagram 1 Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining.

Data preparation process includes data cleaning, data integration, data selection and data transformation. Whereas the second phase includes data mining, pattern evaluation, and knowledge representation.

Diagram 1: Data Mining Process



Sources: https://data-flair.training/blogs/data-mining-process

a. Data cleaning

In the phase of data mining process, data gets cleaned. As we know data in the real world is noisy, inconsistent and incomplete.

It includes a number of techniques. Such as filling in the missing values, combined compute. The output of the data cleaning process is adequately cleaned data.

b. Data integration

In this phase of Data Mining process data in integrated from different data sources into one. As data lies in different formats in a different location. We can store data in a database, text files, spreadsheets, documents, data cubes, and so on. Although, we can say data integration is so complex, tricky and difficult task. That is because normally data doesn't match the different sources.

c. Data selection

This is the process by which data relevant to the analysis is retrieved from the database. As this process requires large volumes of historical data for analysis. So, usually, the data repository with integrated data contains much more data than actually required. From the available data, data of interest needs to be selected and stored.

d. Data transformation

In this process, we have to transform and consolidate the data into different forms. That must be suitable for mining. Normally this process includes normalization, aggregation, generalization etc.

e. Data mining

In this phase of Data Mining process, we have applied methods to extract patterns from the data. As these methods are complex and intelligent. Also, this mining includes several tasks. Such as classification, prediction, clustering, time series analysis and so on.

f. Pattern evaluation

The pattern evaluation identifies the truly interesting patterns. That is representing knowledge based on different types of interesting measures. A pattern is considered to be interesting if it is potentially useful. Also, easily understandable by humans. Further, it validates some hypothesis. That someone wants to confirm new data with some degree of certainty.

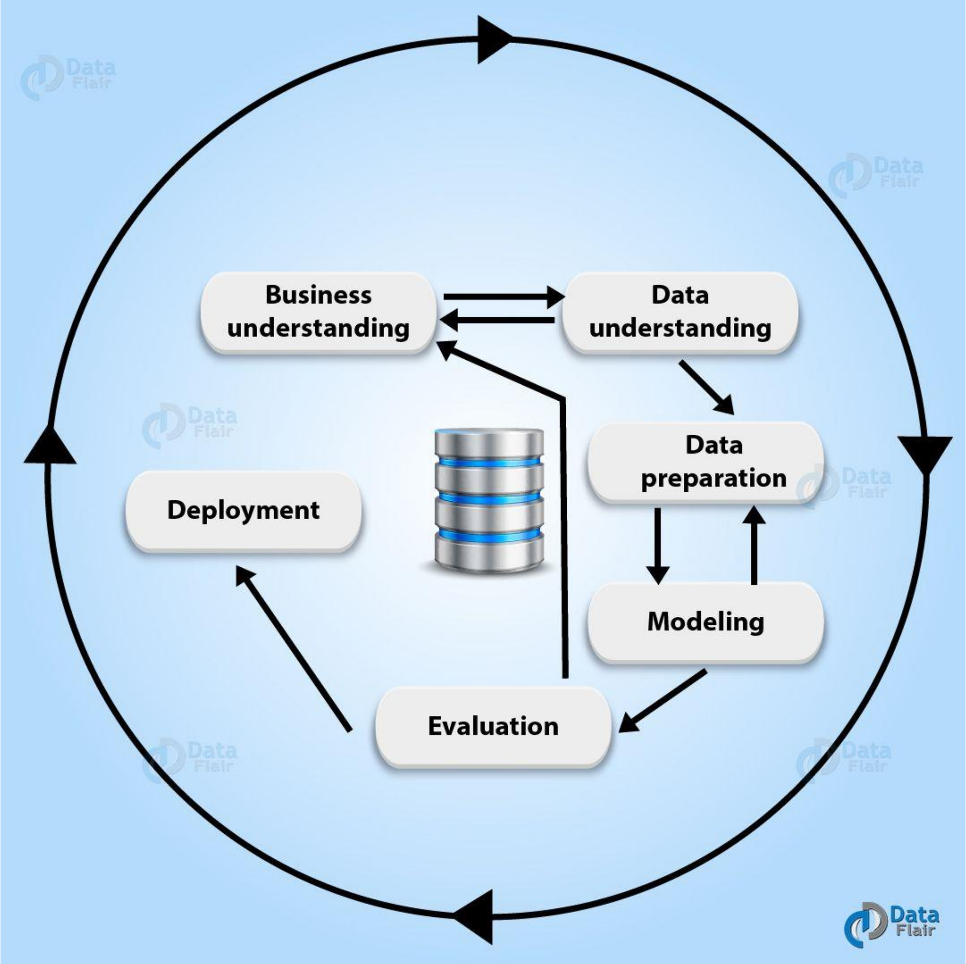
g. Knowledge representation

In the phase of Data Mining process, we have to represent data to the user in an appealing way. Also, that information is mined from the data. To generate output different techniques are need to be applied.[6]

3. Stages of Data Mining Process

As we can see on diagram 2 Data Mining Process has six stages, and it’s a cyclical process.

Diagram 2: six phases of Data Mining Process



Sources: https://data-flair.training/blogs/data-mining-process/

a. Business understanding

[6]https://data-flair.training/blogs/data-mining-process/ http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

- First, we have to understand the requirements. Then have to find what are the business requirements.
- Next, the current situation has to access by finding out the different resources, assumptions. Also, by considering other important factors.
- Then, to achieve the business objectives we need to create data mining.
- Finally, we have to establish a new data mining plan to achieve both business and data mining goals. The plan should be as detailed as possible.

b. Data understanding

- First, this phase starts with the collection of initial data. As in this, we have to collect data from available sources. As we have to collect data to get familiar with the data. Also, in order to make data collection, we need some activities that need to perform. Such as data load and data integration.
- Next, the "gross" or "surface" properties of acquired data need to examine and report.
- Then, we need to explore the data needs by tackling the data mining questions. That can be addressed using querying, reporting, and visualization.
- Finally, have to examine the data quality by answering some important questions. Such as "Is the acquired data complete?", "Is there any missing values in the acquired data?"

c. Data preparation

In this data, the preparation process our 90% time consumed in our project. Also, it's outcome is the final data set. Once we identify the data sources, then we need to select, clean, construct and have to format in the desired form. The data exploration task has to be done at a greater depth. That need to be carry during this phase to notice the patterns. That is based on business understanding.

d. Modelling[7]

- First, we have to select modelling techniques that we need to use for the prepared dataset.
- Next, we have to generate a test scenario to validate the quality and validity of the model.
- Then, by using modelling tools we have to prepare one or more models on the dataset.
- Finally, by involving these models need to assess involving stakeholders. That is to make sure that created models are met business initiatives.

e. Evaluation

- Particularly, in this case, have to evaluate the result in the context of the business goal.
- In this phase, due to new patterns, new business requirements occurs. That patterns have to discover in the model results or from other factors. Gaining business understanding is an iterative process in data mining. The go or no-go decision must make in this step to move to the deployment phase.

f. Deployment

The information, which we gain through data mining process, we need to present it. The information has to represent in such a way that stakeholders can use it whenever they want it. Based on the business requirements, the deployment phase could be creating a report. Also, as complex as a repeatable data mining process across the organization. In this plans for deployment, maintenance, have to create for implementation.

[7] Florin Gorunesku, DATA MINING, Concepts Models, and Techniques, Springer-Verlag Berlin Heidelberg, 2011, page 29.

and also future supports. From the project point, the final report needs to summary the project experiences. And, the review the project to see what need to improved created learned lessons.

As a result, we have studied the Data Mining Process. Along with this have learned stages of data with diagram and cross-industry standard process (CRISP-DM). Furthermore, if you feel any query feel free to ask in a comment section.[8]

4. Techniques of Data Mining

Just as a carpenter uses many tools to build a sturdy house, a good analyst employs more than one technique to transform data into information. Most data miners go beyond the basics of reporting and OLAP (On-Line Analytical Processing, also known as multi-dimensional reporting) to take a multi-method approach that includes a variety of advanced techniques. Some of these are statistical techniques while others are based on artificial intelligence (AI) .

Cluster Analysis

Cluster analysis is a data reduction technique that groups together either variables or cases based on similar data characteristics[9]. This technique is useful for finding customer segments based on characteristics such as demographic and financial information or purchase behavior. For example, suppose a bank wants to find segments of customers based on the types of accounts they open. A cluster analysis may result in several groups of customers. The bank might then look for differences in types of accounts opened and behavior, especially attrition, between the segments. They might then treat the segments differently based on these characteristics.

Linear Regression

Linear regression is a method that fits a straight line through data. If the line is upward sloping, it means that an independent variable such as the size of a sales force has a positive effect on a dependent variable such as revenue. If the line is downward sloping, there is a negative effect. The steeper the slope, the more effect the independent variable has on the dependent variable.

Correlation

Correlation is a measure of the relationship between two variables. For example, a high correlation between purchases of certain products such as cheese and crackers indicates that these products are likely to be purchased together. Correlations may be either positive or negative. A positive correlation indicates that a high level of one variable will be accompanied by a high value of the correlated variable. A negative correlation indicates that a high level of one variable will be accompanied by a low value of the correlated variable.

Positive correlations are useful for finding products that tend to be purchased together. Negative correlations can be useful for diversifying across markets in a company's strategic portfolio. For example, an energy company might have interest in both natural gas and fuel oil since price changes and the degree of substitutability might have an impact on demand for one resource over the other. Correlation analysis can help a company develop a portfolio of markets in order to absorb such environmental changes in individual markets.

Factor Analysis

Factor analysis is a data reduction technique. This technique detects underlying factors, also called "latent variables," and provides models for these factors based on variables in the data. For example, suppose you have a market research survey that asks the importance of nine product attributes. Also suppose that you find three underlying factors. The variables that "load" highly on these factors can offer some insight about what these factors might be. For example, if three attributes such as technical support, customer service, and availability of training courses all load highly on one factor, we might call this factor "service." This technique can be very helpful in finding important underlying characteristics that might not be easily observed but which might be found as manifestations of variables that can be observed.

[8] https://data-flair.training/blogs/data-mining-process/ http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

[9] Charu C. Aggarwal, DATA MINIG, The Textbook, Springer International Publishing Switzerland, 2015, page 16.

Another good application of factor analysis is to group together products based on similarity of buying patterns. Factor analysis can help a business locate opportunities for cross-selling and bundling. For example, factor analysis might indicate four distinct groups of products in a company. With these product groupings, a marketer can now design packages of products or attempt to cross-sell products to customers in each group who may not currently be purchasing other products in the product group.

Decision Trees

Decision trees separate data into sets of rules that are likely to have different effects on a target variable[10]. For example, we might want to find the characteristics of a person likely to respond to a direct mail piece. These characteristics can be translated into a set of rules. Imagine that you are responsible for a direct mail effort designed to sell a new investment service. To maximize your profits, you want to identify household segments that, based on previous promotions, are most likely to respond to a similar promotion. Typically, this is done by looking for combinations of demographic variables that best distinguish those households who responded to the previous promotion from those who did not.

Neural Networks

Neural networks mimic the human brain and can "learn" from examples to find patterns in data or to classify data. The advantage is that it is not necessary to have any specific model in mind when running the analysis. Also, neural networks can find interaction effects (such as effects from the combination of age and gender) which must be explicitly specified in regression. The disadvantage is that it is harder to interpret the resultant model with its layers of weights and arcane transformations. Neural networks are therefore useful in predicting a target variable when the data are highly non-linear with interactions, but they are not very useful when these relationships in the data need to be explained.

Association Models

Association models examine the extent to which values of one field depend on, or are predicted by, values of another field. Association discovery finds rules about items that appear together in an event such as a purchase transaction. The rules have user-stipulated support, confidence, and length. The rules find things that "go together." These models are often referred to as Market Basket Analysis when they are applied to retail industries to study the buying patterns of their customers.[11]

5. Users of data mining

Data mining is at the heart of analytics efforts across a variety of industries and disciplines.

Communications

In an overloaded market where competition is tight, the answers are often within your consumer data. Multimedia and telecommunications companies can use analytic models to make sense of mountains of customers data, helping them predict customer behavior and offer highly targeted and relevant campaigns.

Insurance

With analytic know-how, insurance companies can solve complex problems concerning fraud, compliance, risk management and customer attrition. Companies have used data mining techniques to price products more effectively across business lines and find new ways to offer competitive products to their existing customer base.

Education

[10] Jared Dean, BIG DATA, DATA MINING and MACHINE LEARNING, Value Creation for Business Leaders and Practitioners, John Wiley and Sons Inc., Hoboken, New Jersey, 2014 page 101.

[11]https://www.encyclopedia.com/science-and-technology/computers-and-electrical-engineering/computers-and-computing/data-mining http://lean-management.pl/technologie/wprowadzenie-do-data-mining/

With unified, data-driven views of student progress, educators can predict student performance before they set foot in the classroom – and develop intervention strategies to keep them on course. Data mining helps educators access student data, predict achievement levels and pinpoint students or groups of students in need of extra attention.

Manufacturing

Aligning supply plans with demand forecasts is essential, as is early detection of problems, quality assurance and investment in brand equity. Manufacturers can predict wear of production assets and anticipate maintenance, which can maximize uptime and keep the production line on schedule.

Banking

Automated algorithms help banks understand their customer base as well as the billions of transactions at the heart of the financial system. Data mining helps financial services companies get a better view of market risks, detect fraud faster, manage regulatory compliance obligations and get optimal returns on their marketing investments.

Retail

Large customer databases hold hidden customer insight that can help you improve relationships, optimize marketing campaigns and forecast sales. Through more accurate data models, retail companies can offer more targeted campaigns – and find the offer that makes the biggest impact on the customer.[12]

Car company

The company needs information about the fastest growing car markets where it can sell its cars. For this data is required about past sales and the other companies present with which new company will be required to compete. This requires data mining to be performed to be able to reach to a conclusion after a number of analyses like annual sales data, growth rate, preferred segment, people’s preference for a car, affordability.

Insurance company

The company needs information about the disasters which have occurred in the past and which might occur in the future. Identification of all possible disasters and then assigning the probability to each as per which the final policy will be framed. Now since disasters are unknown a number of studies need to be carried out, and historic data needs to be collected and analyzed. The data will be available from a number of sources which will then be collected, stored, sorted and analyzed.[13]

6. Summary

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is concern about individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals' buying habits and preferences.

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit card accounts on several different databases. The address (or even the name) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered.

Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the usefulness of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained. The result is increased pressure for faster, more powerful data mining queries. These more efficient data mining systems often cost more than their predecessors.

[12] https://www.sas.com/en_us/insights/analytics/data-mining.html

[13] https://planningtank.com/computer-applications/data-mining

We can find many articles on the web that show the advantages use of Big Data technology. He describes many threats resulting from analyzes and tools based on this solution. Which side of the dispute has so right? There is no unconditionally correct answer. Any solution, which revolutionizes the way of thinking to the extent that Big Data changes approach to data brings innumerable benefits. However, at access to such huge amounts of data is not enough to make them bad used. Currently, almost every Internet user in a lesser or to a greater extent is surrounded by the effects of the Big Data i technology It is not necessarily a bad thing. If, thanks to this, the user saves time because suggested to him are searches tailored to his preferences then this is on certainly added value, and user experience resulting from the use of some applications are significantly changed for the better. There is no doubt that these benefits may cost us a leak of data about our activities and movements Internet.

The most justified way out in this situation is definitely not closing for any new technologies, which is undoubtedly the analysis based on Big Data. Education in a given area is always the right approach to such issues. We will learn more about how the technology works, the easier it is predict the consequences resulting from its use, and thus - we become its conscious users. We can have more influence on what data we share about each other. This allows us to enjoy the benefits flowing using this technology, while understanding the risks they pose with being its recipient.

REFERENCES

[1] Jiawei Han, Micheline Kamber, Jian Pei, DATA MINING, Concepts and Tehniques, Third Edition, Simon Fraser University, Elsevier Inc., 2012.

[2] Wil van der Aalst, PROCESS MINING, Data, Science in Action, Second Edition, Springer-Verlag Berlin Heidelberg, 2016.

[3] Charu C. Aggarwal, DATA MINIG, The Textbook, Springer International Publishing Switzerland, 2015.

[4] Dean, BIG DATA, DATA MINING and MACHINE LEARNING, Value Creation for Business Leaders and Practitioners, John Wiley and Sons Inc., Hoboken, New Jersey, 2014.

[5] Florin Gorunesku, DATA MINING, Concepts Models, and Techniques, Springer-Verlag Berlin Heidelberg, 2011.

WEB PAGES

1. http://lean-management.pl
2. https://www.encyclopedia.com
3. https://www.investopedia.com
4. https://www.newgenapps.com
5. https://www.computerworld.pl
6. http://www.zentut.com
7. https://data-flair.training/blogs
8. http://support.sas.com
9. https://www.sas.com
10. https://planningtank.com
11. https://www.ngdata.com
12. http://customerservicezone.com