INTRODUCTION TO KNOWLEDGE DISCOVERY IN DATA PROCESS IN CONTEXT OF INDUSTRY 4.0

Michal Fridrich[1]

Abstract

This article aims to describe the KDD process, or Knowledge discovery in data, and to present some of the available software tools for this process in connection with Industry 4.0. Methods used to interpret the results in this article include research of professional sources, analysis and synthesis of acquired knowledge, and inductive and deductive approaches. In various branches of the economy, whether it is business economics or macroeconomics, we encounter an ever-increasing amount of generated data. This data can be very useful, as information can be drawn from it, which can then be used to optimize processes that lead to strengthening competitiveness in today's very turbulent market. The more data is created, the more it can cause complications in their identification. Errors also often occur during data generation, due to which the data is not uniform. The results of this work will include a description of the eight important steps of the KDD process.

Keywords

KDD Process, Data Mining, Industry 4.0, Data Processing

I. Introduction

The current trend of industrial development, when there is increasing digitization and automation in industry and thus changes in the labor market, is referred to as the fourth industrial revolution or Industry 4.0. The concept of data mining has been known for half a century, but only now is it gaining a completely new dimension. This is mainly due to the applicability of data mining in various industries, which is also related to the emergence of algorithms for machine learning. Machines are gaining the ability to take over activities previously performed by humans. The first mentions of Industry 4.0 appeared in 2011, the whole concept was then presented in 2013 in Hannover. Ideas about the fourth industrial revolution are no longer "only" the domain of academics in laboratories, but with the development of related technologies and the entry into the era of "big data," they find their application across many different fields (Oliff & Liu, 2017).

Many theories criticize this development of the industry. There are two main reasons for this criticism. The first reason is that the machines that will be part of the so-called Internet of Things will be controlled over the Internet, which means that they will become vulnerable to hacker attacks from outside. This creates the need to develop new security platforms, which are connected to more sophisticated data encryption not only in data warehouses but also, for example when machines communicate with each other. When such an attack is carried out, paralysis of the entire production line or logistics could occur, which is unacceptable at a time when most automotive companies deliver material "just in time".

The second reason is the economic point of view. On the one hand, production will become very efficient, but on the other hand, the people who operated the machines up to the present time will lose their jobs. According to some theorists, this could be solved by not having the machines fully automated, but still partially controlled by humans (Zezulka et al., 2016). However, development in this area cannot be stopped. The volumes of data that companies work with are growing almost exponentially, which is also due to increasing computing power. This is also why it is important to somehow process the data into a readable form so that companies do not get lost in it, simply put.

[1]Michal Fridrich, PRIGO University, V. Nezvala 801/1, 73601 Havířov, Czech Republic. E-mail: michal.fridrich@prigo.cz.

Among other things, data mining tools are used for this. However, data mining is no longer the only step in working with data. If we look at the whole issue comprehensively, we often call the whole process KDD or knowledge discovery in data.

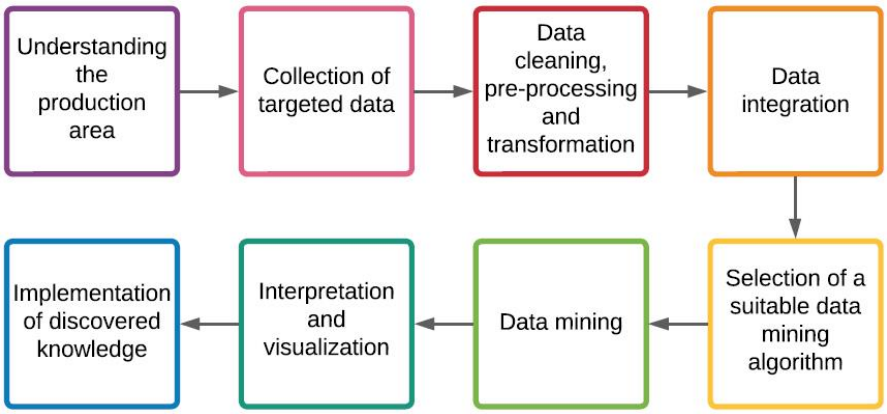
II. Knowledge Discovery in Data

Information and knowledge provide a competitive advantage and improve the global market position for both individuals and organizations. Thanks to them, it is possible to solve ever more complex situations and learn new skills. In most industries, manufacturing is highly competitive, and financial margin is the factor that forms the line between success and failure. To meet the tough challenges, the company must achieve the lowest possible costs in manufacturing while maintaining a highly-skilled, flexible, and efficient workforce. In higher-wage economies, this can generally be achieved primarily through the efficient use of knowledge. However, knowledge can take many forms (Harding et al., 2009). One of the challenges is identifying the knowledge and information that is crucial for the company. In modern times, the volume of data collected in production environments from sensors, barcodes, or camera systems is growing at an unprecedented rate. This data is related to, for example, product design, machines, processes, material, inventory, maintenance, planning, control, assembly, or logistics and may contain valuable dependencies, trends, and patterns. Manual analysis of a huge amount of data with different attributes in production databases is very impractical for obtaining useful information. An intelligent and automated data analysis methodology is needed to efficiently extract information from data.

Knowledge Discovery In Databases (KDD-or searching for knowledge in databases) is a process that involves the application of specific algorithms for extracting models (patterns) from data. KDD includes theories, algorithms, and methods from the intersection of several scientific disciplines, including data mining, database technology, machine learning, statistics, artificial intelligence, knowledge systems, and data visualization (S. Mitra et al., 2002). The overall KDD process is presented in Figure 1 and most often includes the following steps, some of which will be more detailed in the following subsections:

1) Understanding the production area;
2) Collection of targeted data;
3) Data cleaning, pre-processing, and transformation;
4) Data integration;
5) Selection of a suitable data mining algorithm;
6) Data mining;
7) Interpretation and visualization;
8) Implementation of discovered knowledge.

Figure 6 KDD process



Source: Own study according to (Harding et al., 2009)

Understanding the production area

This step is the complete foundation for starting the KDD process. The management of a company that decides to collect data and subsequently process it must know very well all the activities taking place in the company and its surroundings.

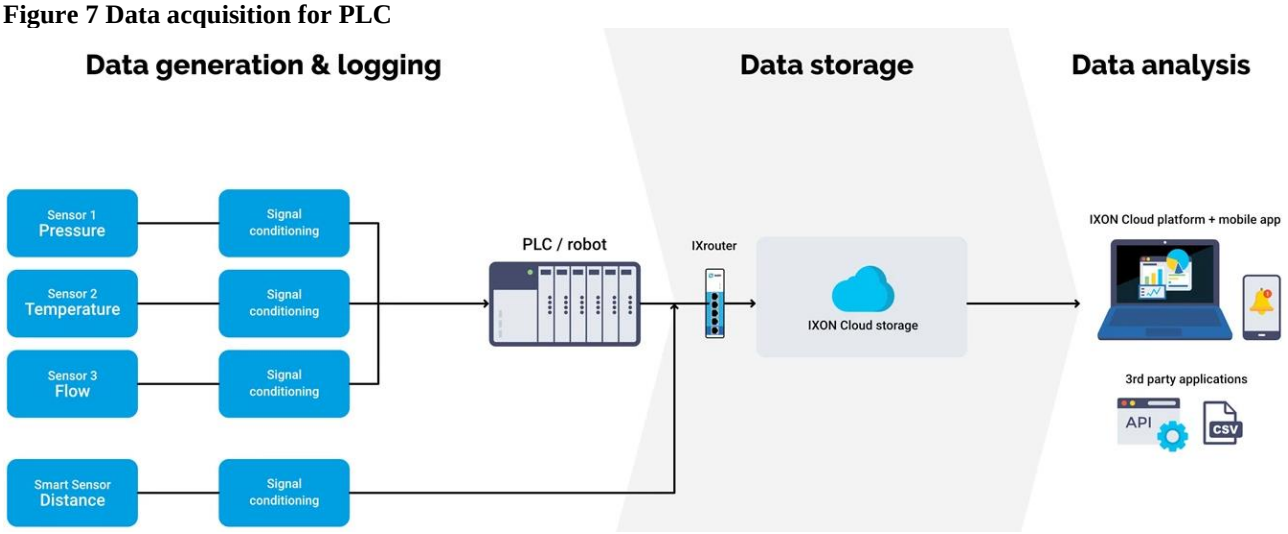
Collection of targeted data

Data can generally be divided into quantitative (can be counted) and qualitative (contains records that cannot be quantified). General methods of data collection include interviews, questionnaires, observations, holding group discussions with interest groups (focus groups), and collecting data from company documents (Ainsworth, 2021).

Data acquisition, or DAQ (Data acquisition), is the process of digitizing data from industrial facilities or environments (such as buildings) for display, analysis, and storage in computers, according to IXON Cloud, which develops software for remote control of IIOT devices, PLC, servers or cloud. Most systems for collecting data from sensors cannot do without a device that converts the analog signal to digital (IXON, 2021).

PLC is an abbreviation for Programmable Logic Controller, a computer used in industry to control motors, robots, and sensors (transducers) - for example, a production line with motors for driving belts, arms, and movements of levers or hydraulic devices. PLC data logging is the process of collecting data from machines and connected devices (sensors) for future analysis. An example of this process can be the measurement of humidity in a factory furnace as a digital value using a smart sensor. If the data is to be evaluated correctly, it is time and date-stamped. Data logging or data collection software/systems for PLC can be used to collect different types of information about machines or buildings such as vibration, light intensity, temperature, electric current, voltage, sound frequency, or pressure measurement. A diagram of one of the options for collecting, storing, and analyzing data can be seen in Figure 2 (IXON, 2021).

Figure 7 Data acquisition for PLC



Source: (IXON, 2021)

One of the software tools, thanks to which it is possible to collect data from the company's sensors, is a tool called WhiteMON, which will be described in more detail in the following paragraphs.

WhiteMON – Software for collecting and processing data from sensors

WhiteMON is a professional environment monitoring solution offered as a service. The WhiteMON system ensures data collection from sensors, storage of measured values, and their clear display in the form of online graphs and tables. Export in HTML, XML and CSV format is supported. WhiteMON can alert you to the critical status of sensors via SMS, e-mail, ICQ, or other channels. Thanks to the web interface, it is possible to see current and historical data at any time and from anywhere without the need for installation on workstations. Multiple users with different passwords and access levels can work with the system at the same time. Users can customize the system, set their own welcome screen, create bulletin boards with the most monitored sensors, etc. WhiteMON can be easily connected to other applications, if necessary, it is possible to immediately work with measured data or events in other company systems (Whitesoft, 2019). The software environment is shown in Figure 3.

WhiteMON offers the following usage options:

- Deduction of values from sensors in the selected interval;
- Online graphic outputs: live / 2 days / 30 days / 100 days;
- Exports of "from-to" historical data;
- One-time or automatic periodic reports in HTML/PDF;
- Notification of sensor status via SMS, e-mail, ICQ;
- Program control of other devices (e.g. ventilation when the temperature is exceeded);
- Data archiving 100 years back;
- Nonstop access to the system via a web browser via Internet;
- Data security – password login, different levels of access;
- Bulletin boards and maps with selected sensors;
- Possible integration with other systems (quality control, ERP, web applications);
- Environment in English.

Figure 8 Software WhiteMON

Source: (Whitesoft, 2019)

What the software can track:

- States and physical quantities (temperature, pressure, humidity, distance, movement, presence of gases);
- Network elements;
- Windows servers;
- Linux servers;
- Mail servers (Exchange, IMAP, POP3, SMTP);
- Web servers (Apache, IIS, Google Analytics, HTTP);
- SQL database (MySQL, MS SQL, MySQL);
- Virtual servers (VMware, Hyper-V, Citrix, Virtuozzo);
- SNMP protocol support;
- The possibility of programming your sensors (EXE, SSH script, VBS, Powershell).

If the company would prefer to use its server to run this software, it is possible to purchase its own license and run the service on its own server. It would be ideal to use software for data collection, convert it into a SQL database, and then use data mining tools in the cloud. This process would run concurrently with and directly interface with the enterprise's visual interactive environment.

Data cleaning, pre-processing, and transformation

Data preparation is a key phase of the KDD process. During this phase, it is important not to make mistakes that would lead to the creation of a defective set of prepared data, since the success of the following KDD steps depends largely on this preparation phase. In addition, data preparation is one of the most time-consuming and difficult phases of the KDD process (Oliff & Liu, 2017). Data quality can be affected by user input errors, input inconsistencies, missing values, typos, or incorrect data generation (Calabrese, 2019).

Automatic data preparation mechanisms offer significant time and resource savings in the KDD process. These resources could then be put to good use in later, less automatable stages of the process, such as the interpretation of results.

Data preparation involves performing operations on the source data in such a way that it is ready for the application of data mining algorithms. Although it is difficult to create a precise list of data preparation steps, most authors emphasize the following procedure (Lara et al., 2014):

Data Collection and Integration: The goal of this task is to collect data from different sources, and help represent, code, and integrate data from different tables to homogenize information.

Data cleaning: In this step, data conflicts are removed, skewed values are reconciled, and problems with noise or missing values are solved. Removing duplicate records is very important.

Data Transformation: Data transformation is the process of converting data from one format or structure to another format or structure.

Data reduction: The goal of this task is to select relevant data for subsequent data mining.

A large number of software tools are available for data preparation. The most important thing is to have the goals of their usefulness set when starting the data preparation. For ordinary data preparation, for example, MS Excel is sufficient.

MS Excel as a tool for data preparation

The tool, which is familiar to most users, contains many useful functions used to prepare data without users knowing about their existence. Among the most used functions for data preparation in Excel is (Trumpexcel, 2014):

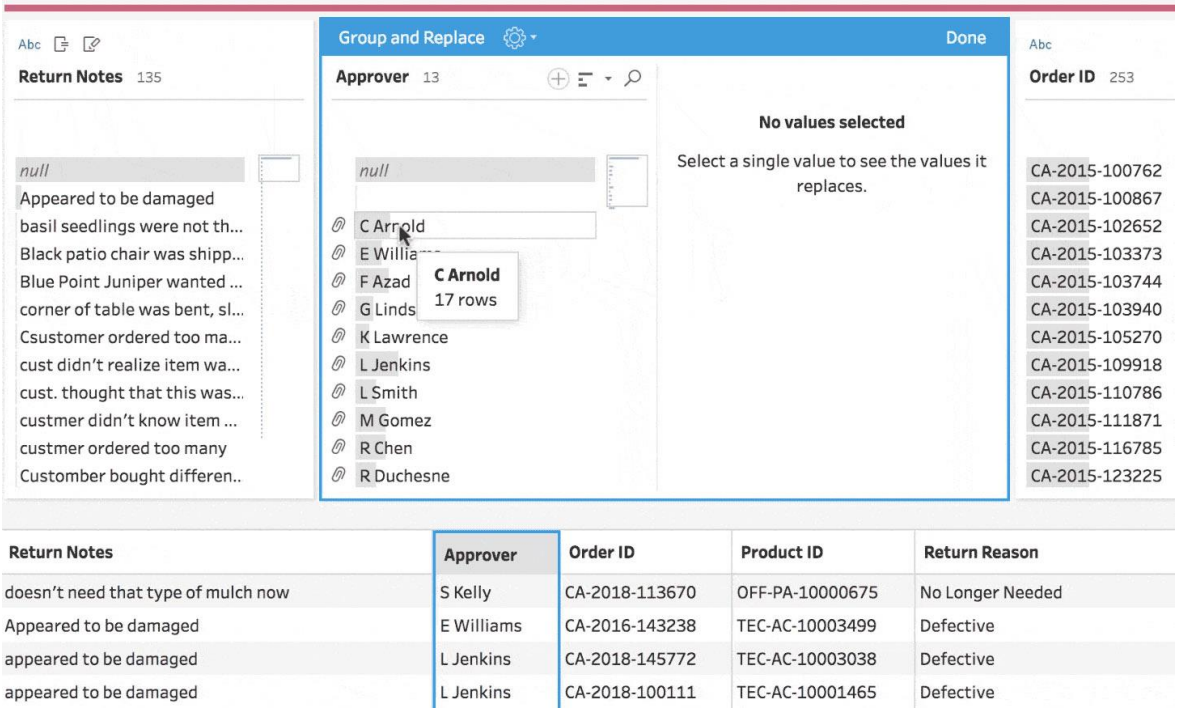
- Erasing excess spaces in cells;
- Selection and treatment of all empty cells;
- Convert numbers that are stored as text to numbers;
- Removal of duplicate records;
- Error highlighting;
- Mass change of text from lowercase or uppercase letters to the correct size;
- Spell check;
- Remove all formatting;
- Using the "find and replace" function.

To automate the preparation of data in large files, it is possible to use the macro functions and create custom scripts suitable for the given document. However, there are also specialized software applications designed directly for data preparation. These applications include, for example, Tableau Prep.

Tableau Prep software application for data preparation

By providing a visual and direct way to combine, modify, clean, and integrate data, Tableau Prep enables analysts and business users to start analyzing data faster. This application consists of two products: Tableau Prep Builder for creating data flows and Tableau Prep Conductor for planning, monitoring, and managing flows across the organization. Three coordinated views allow you to view row-level data, profiles of each column, and the overall data preparation process. Changes made are visible in real-time even for millions of records. The order of the individual steps can be changed by the user without risking permanent changes in the data. This makes it possible to experiment at a high level. Intelligent functions such as fuzzy clustering allow mass changes to be made repeatedly with a single click. The program environment can be seen in Figure 4 (Tableau, 2021).

Figure 9 Tableau Prep application user interface

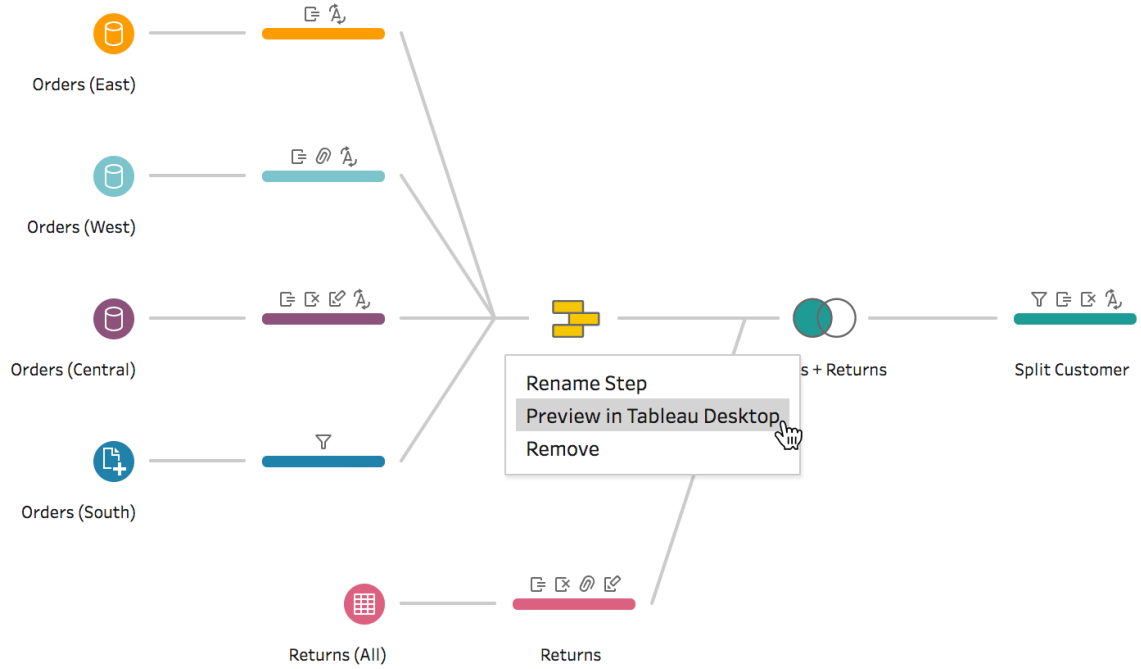


Return Notes	Approver	Order ID	Product ID	Return Reason
doesn't need that type of mulch now	S Kelly	CA-2018-113670	OFF-PA-10000675	No Longer Needed
Appeared to be damaged	E Williams	CA-2016-143238	TEC-AC-10003499	Defective
appeared to be damaged	L Jenkins	CA-2018-145772	TEC-AC-10003038	Defective
appeared to be damaged	L Jenkins	CA-2018-100111	TEC-AC-10001465	Defective

Source: (Tableau, 2021)

The application offers the possibility to connect to data on the local disk, in the network, and the cloud. The advantage is a high degree of connectivity with third-party applications, which include, for example, Excel, Oracle Database, AWS, or SAP. A big advantage is also the possibility of displaying the entire data structure in the form of a diagram (can be seen in Figure 5) and sharing work online with other users.

Figure 10 Display of the data structure in the form of a diagram



Source: (Tableau, 2021)

All data with which the company decides to continue working should meet the following conditions (Wang & Strong, 1996):

- **Validity** – a factor that determines to what extent the data agree with the subject of business;
- **Accuracy** – the data should correspond to their real values;
- **Completeness** – a measure determining whether all the necessary data are known;
- **Consistency** – data should be consistent across all datasets;
- **Uniformity** – data should be in the same units within datasets;
- **Accessibility** – it is important that only users who have the appropriate rights can easily access the data.

After collecting, reducing, transforming, cleaning, and integrating data, data mining follows.

Data mining

Data mining uses well-researched statistical principles to discover patterns in data. Data mining algorithms can be used to predict trends, identify patterns, create rules and recommendations, analyze the sequence of events in complex data sets, and gain new insights (Han & Kamber, 2012). Data mining uses mathematical analysis to infer patterns and trends that exist in data. Usually, these patterns cannot be detected by traditional data exploration because the relationships are too complex or because the amount of data is too large. These patterns and trends can be collected and defined as data mining models. Building a data mining model is part of a larger process that includes everything from asking questions about the data, to building the model itself, to deploying the model in a working environment.

An algorithm in data mining (or machine learning) is a set of heuristic methods and calculations that subsequently create a model from the data. If a model is to be created, it is first necessary to use a given algorithm that analyzes the data and looks for specific types of patterns or trends in it. The algorithm uses the results of data analysis over many iterations to find the optimal parameters for building a data mining model. These parameters are then used across the dataset to extract actionable patterns and detailed statistics. The mining model that the algorithm creates from the data can take various forms, including, for example (Minewiskan, 2021):

- Set of clusters (clustering);
- Decision tree (decision tree);
- Predictive mathematical model (predictive analysis).

Clustering in data mining

A cluster is a group of objects that belong to the same class. This means that data with similar properties are grouped into one cluster and other objects are grouped into another cluster, as can be seen in Figure 6. These groupings are useful for data exploration, identifying anomalies in the data, and making predictions. Clustering algorithms identify relationships in a data set that usually cannot be logically inferred by random observation. Cluster analysis is widely used in many fields, including data analysis, market research, pattern recognition, and image processing. It helps marketers differentiate and characterize different groups of customers in their client base based on buying patterns (Javapoint, 2021).

Figure 11 Grouping data into clusters

Source: Own study according to (Javapoint, 2021)

For the correct functionality of the clustering algorithm, the following two conditions must be met (Minewiskan, 2021):

One key column: Each model must contain one numeric or text column that uniquely identifies each record. Composite keys are not allowed.

Input columns: Each model must contain at least one input column that contains the values used to create the clusters. It is possible to have as many input columns as needed, but depending on the number of values in each column, adding more columns may increase the time required to train the model.

From clustering software applications for commercial use, the author recommends, for example, BayesiaLab, ClustanGraphics3, or IBM SPSS Modeler. If it were necessary to use free open-source applications, then you can choose, for example, between Autoclass C or MDL Clustering software (KDnuggets, 2021).

Decision tree

A decision tree is another data mining algorithm. It is a structure that includes a root node, branches, and leaves. Each internal node indicates a test for an attribute (for example, yes/no), each branch indicates the result of that test, and each leaf node has a class label. The highest node in the tree is the root node. Thanks to this algorithm, it is possible to make decisions in processes more easily. An example can be a decision tree about the introduction of new product development with possible impacts on the process under different scenarios. A condition for a successful decision tree is a previously known output. If the tree contains high-quality input data, then machine learning is very fast and efficient (Tutorialspoint, 2021).

Predictive analysis

Predictive analytics is the use of data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond descriptive statistics and reports of what happened and provide the best assessment of what will happen in the future. The result is more efficient decision-making and the acquisition of new insights that lead to continuous improvement (Kumar & L., 2018).

Predictive algorithms use known results to develop (or train) a model that can be used to predict values for different or new data. Modeling results in predictions that represent the probability of a target variable (for example, yield) based on the estimated significance of a set of input variables. This differs from descriptive models, which help understand what happened, or diagnostic models, which help understand key relationships and determine why something happened. Predictive analytics is used in many different fields, including engineering, healthcare, and banking (SAS, 2021).

Data mining software applications

There are more and more applications on the market that allow you to perform the entire data mining process in a sophisticated manner. Among the most used are, for example, Xplenty, Rapid Miner, Orange Weka, or KNIME (Softwaretestighelp, 2021). The choice of application depends primarily on the industry and the purpose of using the acquired knowledge. Modern data mining software applications can work with large volumes of data that can be connected to data warehouses. A big advantage is that some of them already know data mining algorithms and, thanks to machine learning and artificial intelligence, not only create the necessary models but also interpret and visualize the data.

Interpretation and visualization

This step is the penultimate step in the entire KDD process. The interpretation and visualization of data serve to easily orientate yourself in the results of previous analyses. Data can be interpreted in different ways, which most often include graphs, tables, or figures. The basic prerequisite for effective data display is the simplicity, readability, and good comprehensibility of the given representation (Midway, 2020).

For the presentation of data using graphs, it is necessary to choose a meaningful graph type for the end user. Among the most common types of graphs are (IBM Docs, 2021):

- **Bar graphs** are used to compare discontinuous data or show trends over time;
- **Line charts** are used to show trends over time and compare multiple data series;
- **Pie charts** are used to show mutual ratios;
- **Horizontal bar graphs** are useful for showing trends over time and plotting multiple data series;
- **Area graphs** are useful for highlighting the magnitude of change over time. Stacked area charts are also used to show the relationship of parts to a whole;
- **Scatter plots** are useful for clearly presenting quantitative data;
- **Combo charts** display multiple data series using a combination of columns, areas, and lines within a single chart. This makes them useful for highlighting relationships between different data series;
- **Pareto charts** facilitate process improvement by identifying the root causes of a given event. It numbers the order of the categories from the most numerous to the least numerous. These charts are often used for quality control data to identify and reduce the root causes of problems.

Other, less commonly used chart types are bubble charts, quadrant charts, bullet charts, gauge charts, micrographs, radar charts, and range indicator charts (IBM Docs, 2021).

The next and final step in the KDD process is the implementation of the discovered knowledge.

Implementation of discovered knowledge

After completing the entire KDD process, it is necessary to implement the discovered knowledge into the company process. Discovered knowledge should be used to fulfill predetermined goals and management should determine whether and to what extent the analysis was successful. If there are too many errors in the discovered knowledge, it is possible that at the beginning of the KDD process data were selected with parameters that do not suit the subject of the given analysis. In this case, the whole process needs to be repeated with new and different input data.

III. Conclusion

This article aimed to describe the KDD process, or Knowledge discovery in data, based on a search of professional literary sources. Due to the ever-increasing amount of data that businesses generate,

there is a need to use processes and tools, thanks to which the data can be appropriately used to obtain important information leading to maintaining competitiveness in today's turbulent economic environment. The article presented all the steps of the KDD process, which include Understanding the production area; Collection of targeted data; Data cleaning, pre-processing, and transformation; Data integration; Selection of a suitable data mining algorithm; data mining; Interpretation and visualization; Implementation of discovered knowledge. In some of the sub-steps of the entire process, software tools were also presented, thanks to which these steps can be performed. In the future, the author of this article would like to devote himself to the entire KDD process and, above all, to data mining. This is an area that, in connection with Industry 4.0, will be relevant for many years to come.

References

Ainsworth, Q. (2021). *Data Collection Methods*. Retrieved January 14, 2021, from https://www.jotform.com/data-collection-methods/.

Calabrese, B. (2019). Data Cleaning. In S. Ranganathan, M. Gribskov, K. Nakai, & C. Schönbach (Eds.), *Encyclopedia of Bioinformatics and Computational Biology* (pp. 472–476). Academic Press.

Han, J., & Kamber, M. (2012). *Data mining: Concepts and techniques* (3rd ed). Elsevier.

Harding, J. A., Shahbaz, M., Srinivas, & Kusiak, A. (2009). Data Mining in Manufacturing: A Review. *Journal of Manufacturing Science and Engineering*, *128*(4), 969–976.

IBM Docs. (2021). *IBM Docs*. Retrieved May 15, 2021, from https://prod.ibmdocs-production-dal-6099123ce774e592a519d7c33db8265e-0000.us-south.containers.appdomain.cloud/docs/cs/cognos-analytics/11.1.0?topic=charts-chart-types.

IXON. (2021). *PLC Data Loggers & Acquisition Software + Alternative*. Retrieved May 10, 2022, from https://www.ixon.cloud/knowledge-hub/a-better-alternative-to-data-loggers-and-acquisition-software-for-plc-s.

Javapoint. (2021). *Data Mining Cluster Analysis—Javatpoint*. Retrieved August 16, 2021, from https://www.javatpoint.com/data-mining-cluster-analysis.

KDnuggets. (2021). Clustering and Segmentation Software. *KDnuggets*. Retrieved May 24, 2021, from https://www.kdnuggets.com/software-for-data-mining-analytics-data-science-and-knowledge-discover/clustering-and-segmentation-software/.

Kumar, V., & L., M. (2018). Predictive Analytics: A Review of Trends and Techniques. *International Journal of Computer Applications*, *182*, 31–37.

Lara, J. A., Lizcano, D., Martínez, M. A., & Pazos, J. (2014). Data preparation for KDD through automatic reasoning based on description logic. *Information Systems*, *44*, 54–72.

Midway, S. R. (2020). Principles of Effective Data Visualization. *Patterns*, *1*(9), 100141.

Minewiskan. (2021). *Data Mining Algorithms (Analysis Services—Data Mining)*. Retrieved July 5, 2022, from https://docs.microsoft.com/en-us/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining.

Oliff, H., & Liu, Y. (2017). Towards Industry 4.0 Utilizing Data-Mining Techniques: A Case Study on Quality Improvement. *Manufacturing Systems 4.0 – Proceedings of the 50th CIRP Conference on Manufacturing Systems*, *63*, 167–172.

S. Mitra, S. K. Pal, & P. Mitra. (2002). Data mining in soft computing framework: A survey. *IEEE Transactions on Neural Networks*, *13*(1), 3–14.

SAS. (2021). *Predictive modeling techniques: What they are and how to use them*. Retrieved September 15, 2021, from https://www.sas.com/ko_kr/insights/analytics/predictive-modeling-techniques.html.

Softwaretestighelp. (2021). *Top 15 Best Free Data Mining Tools: The Most Comprehensive List*. Retrieved March 26, 2021, from https://www.softwaretestinghelp.com/data-mining-tools/.

Tableau. (2021). *Tableau Prep*. Retrieved March 28, 2021, from https://www.tableau.com/products/prep.

Trumpexcel. (2014). *10 Super Neat Ways to Clean Data in Excel Spreadsheets*. Retrieved August 8, 2021, from https://trumpexcel.com/clean-data-in-excel/.

Tutorialspoint. (2021). *Data Mining—Decision Tree Induction—Tutorialspoint*. Retrieved January 10, 2021, from https://www.tutorialspoint.com/data_mining/dm_dti.htm.

Wang, R. Y., & Strong, D. M. (1996). Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, *12*(4), 5–33. JSTOR.

Whitesoft. (2019). *WhiteMON | Whitesoft s.r.o.* Retrieved June 3, 2021, from http://www.whitesoft.cz/reseni/telekomunikace/whitemon.

Zezulka, F., Marcon, P., Vesely, I., & Sajdl, O. (2016). Industry 4.0 – An Introduction in the phenomenon. *IFAC-PapersOnLine*, *49*(25), 8–12.