DATAFICATION AS A NECESSARY STEP IN THE PROCESSING OF BIG DATA IN DECISION-MAKING TASKS OF BUSINESS

Martin Misut[1], **Pavol Jurik**[2]

Abstract: The digital transformation of business in the light of opportunities and focusing on the challenges posed by the introduction of Big Data in enterprises allows for a more accurate reflection of the internal and external environmental stimuli. Intuition ceases to be present in the decision-making process, and decision-making becomes strictly data-based. Thus, the precondition for data-based decision-making is relevant data in digital form, resulting from data processing. Datafication is the process by which subjects, objects and procedures are transformed into digital data. Only after data collection can other natural steps occur to acquire knowledge to improve the company's results if we move in the industry's functioning context. The task of finding a set of attributes (selecting attributes from a set of available attributes) so that a suitable alternative can be determined in its decision-making is analogous to the task of classification. Decision trees are suitable for solving such a task. We verified the proposed method in the case of logistics tasks. The analysis subject was tasks from logistics and 80 well-described quantitative methods used in logistics to solve them. The result of the analysis is a matrix (table), in which the rows contain the values of individual attributes defining a specific logistic task. The columns contain the values of the given attribute for different tasks. We used Incremental Wrapper Subset Selection IWSS package Weka 3.8.4 to select attributes. The resulting classification model is suitable for use in DSS. The analysis of logistics tasks and the subsequent design of a classification model made it possible to reveal the contours of the relationship between the characteristics of a logistics problem explicitly expressed through a set of attributes and the classes of methods used to solve them.

UDC Classification: 004.62, **DOI:** https://doi.org/10.12955/pns.v2.156

Keywords: datafication, logistics, classification, DSS, big data.

Introduction

Over the last twenty years, there has been an economic transformation characterized by a rapid shift from the traditional industrial model of production to a new scenario defined by developing a digital or information society (Musik & Bogner, 2019; Eriksson, 2019). This development leads to fundamental changes in businesses. The digital transformation of business in the light of opportunities and focusing on the challenges posed by the introduction of Big Data in enterprises allows a more accurate reflection of the internal and external environmental stimuli. The critical factor is information and knowledge, which we can obtain by analyzing big data (Jeble et al., 2018).

Intuition ceases to be present in the decision-making process, and decision-making becomes strictly data-based (Kościelniak & Puto, 2015). In the traditional decision-making model, internal decisions are based on data generated by transaction processing systems, such as ERP systems, and are supported by decision support systems. Continued development has led to the creation of supply-side and demand-side systems (SRM and CRM), which helped integrate its internal operations with external operations represented by suppliers and customers.

All of these systems use structured data stored in relational databases. However, the situation changes due to a data avalanche's existence and the effort to use its hidden information. With the advent of big data, the information requirements of executives began to change (Merendino et al., 2018). In addition to the traditional data sets described above, there are big data from various sources in structured, semi-structured or unstructured form. It is possible in various ways to use the value hidden in this data for strategic, tactical and operational decisions in companies (Günther et al., 2017). However, the first and fundamental premise is data describing the properties (attributes) of reality, which is the subject of subsequent analysis and extraction. This fact is closely related to the concept of datafication.

This paper presents a method with the application of datafication principles in the creation of classification trees. The classification trees created in this way can then be used to support managerial decisions. The method was verified for logistics problems. The rest of the paper is organized as follows: The section *Datafication and Models of Data Used in Decision Making* briefly describes the relationship between decision-making and the necessary data and the characteristics of datafication and its role in defining the necessary database for decision support. The section after, entitled *Method*, describes the approach to selecting attributes when creating a classification model using the Weka3.8.4 software tool.

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Applied Informatics, martin.misut@euba.sk, ORCID: 0000-0002-5545-2624

[2] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Applied Informatics, pavol.jurik@euba.sk

A detailed description of the proposed method for logistics problems can be found in section *A Logistics Problems Case*. Finally, the *Conclusion* section presents the main findings.

Datafication and Models of Data Use in Decision Making

The quality of decisions made in the current operations of businesses is affected by the efficient use of available big data and small data sets. Modern information and communication technologies form a complex of hardware tools, software tools and organizational solutions working with data that affect how the business is managed. This complexity of the control complex also affects how data is used; therefore, it is not an easy task (Merendino et al., 2018). As Kościelniak and Puto (2015) suggest, data-based decision support should be used to address these challenges.

Jia et al. (2015) perceive data-driven decision making as a continuous process that consists of several steps, including collecting data, converting data into information and, ultimately knowledge, making decisions based on knowledge, monitoring the implementation of decisions, and providing feedback for each process. Several big data decision support models have been described in literature (Athamena & Houhamdi, 2018; Jeble et al., 2018; Jia et al., 2015; Kościelniak & Puto, 2015; Travica, 2017). Most authors cite the determination of selection criteria and the data source as the first step in the decision-making process (Athamena & Houhamdi, 2018; Jeble et al., 2018; Jia et al., 2015; Kościelniak & Puto, 2015). Compared to Travica (2017), the decision cycle begins with the Recognize Big Data Need in the model proposed in it. The existence of several big data decision support models, or rather the absence of a more widely accepted standard model, is based on the fact that big data is still new and requires changes in the organizational culture. Therefore, aligning the potential of big data with the needs of practical decision-making is still a significant task for many companies. In such a case, we must agree with Yousuf and Zainal (2020) that the existence of data suitable for obtaining the necessary information as a basis for a decision is a necessary precondition.

If the essential data do not exist in the given structure, it is necessary to describe the existing natural phenomenon by recording their properties, key for the decision. The properties are recorded in digital form in the so-called attributes, while the values of individual attributes then give the identity of a particular phenomenon. Based on the similarity of values, it is possible to classify phenomena or use them for more complex operations in making decisions. This fact is also used by decision support systems (DSS).

According to Athamena and Houhamdi (2018), decision support systems (DSS) transform input data into valuable information and then convert that information into knowledge to improve the decision-making process and are a vital asset for analysts. In comparison, Power (2002) emphasizes the interactivity of DSS and considers DSS to be an interactive computer information system that helps decision-makers use data. Naturally, advanced methods of artificial intelligence have also found use in DSS. DSS can find the best solution to unstructured or semi-structured problems by applying intelligent models and techniques (Daas et al., 2013). Thus, the precondition for data-based decision-making is relevant data in digital form, resulting from data collection.

Datafication can currently be considered a technological trend that captures many aspects of our lives through data, from which we obtain information that represents a new form of value for us (Fernández-Rovira, et al., 2021). The term datafication was introduced in 2013 by Kenneth Cukier and Victor Mayer-Schöenberger (Mayer-Schönberger & Cukier, 2013). Although datafication has been implicitly existing here for a long time, its mass expansion and conscious use have occurred mainly due to the influence of data and computational possibilities of predictive analysis. Southerton (2020) defined datafication in the Big Data Encyclopedia as follows:

"Datafication refers to the process by which subjects, objects, and practices are transformed into digital data. Associated with the rise of digital technologies, digitization, and big data, many scholars argue datafication is intensifying as more dimensions of social life play out in digital spaces. Datafication renders a diverse range of information as machine-readable, quantifiable data for the purpose of aggregation and analysis. Datafication is also used as a term to describe a logic that sees things in the world as sources of data to be "mined" for correlations or sold, and from which insights can be gained about human behavior and social issues. This term is often employed by scholars seeking to critique such logics and processes. "

We agree with Eriksson (2019) that in the current data-driven environment, a company's successful existence is conditional on total control over the storage, manipulation and extraction of data and related information in the causal context of datafication. Naturally, datafication played a role in launching the big data avalanche. Even though the size of individual data records may be small, the frequency of recording, the number of recorded attributes and the number of data sources may result in significant data accumulation and complexity (Jones, 2019).

However, datafication cannot be confused with digitization, as digitization only means the conversion to digital form (Mejias & Couldry, 2019). The most frequently cited example for understanding the difference between the two concepts is book processing. If the book is scanned page by page and these page images are saved - we are talking about digitizing the content. It is not easy to extract data from a book digitized in this way, as it is encoded in images. After datafication, i.e. using OCR software, the book's content is made available for search or other processing. In this form, the information encoded in the book's text is already available and can be used directly for subsequent management/analysis.

The purpose of data collection and use is the fact that they represent some significant features (though perhaps not all) of reality. Therefore, the initial phase of data collection means that it is necessary to select the aspect of reality that the data will represent (Jones, 2019). Although the cost of recording and storing data is already low, and we expect it to decrease further in the future and we agree with Jones (2019) that this does not mean that all possible data will necessarily be recorded in the future. Nevertheless, it is often necessary to decide which data to record in both the short and the long term. This choice may be influenced by various factors, such as storage capacity constraints, existing technology, and the assessment of the degree of appropriate data details based on expectations regarding their future use. What is recorded will, therefore, generally be a subset of all possible data. Estimating the future use of data helps decide what should be recorded. Therefore, we focused on determining the necessary set of attributes to capture the key properties of a phenomenon so that these attributes can be used for full-fledged decision-making through DSS.

Method

The task of finding a set of attributes (selecting attributes from a set of available attributes) so that a suitable alternative in decision making can be determined on its basis is analogous to the task of classification, i.e. based on which attributes it is possible to classify a given instance into the appropriate class unambiguously. Singhal and Jena (2013) defined classification as finding a set of models that describe and distinguish data classes and concepts to use the model to predict the class whose label is unknown. Song and Ying (2015) recommend decision trees to solve such a problem.

Decision trees can be considered an alternative approach to logistic regression and discriminant analysis if the dependent variable is categorical. The decision tree is a representation of the decision procedure for classifying cases into appropriate classes (Li et al., 2019). It is a graph structure in a tree containing a root node, non-leafy and leaf nodes. Nodes represent a class or testing sign. The edges represent the testing sign values. If the output variable is categorical, each leaf node represents one of the categories of the output variable classes. Then we talk about classification trees.

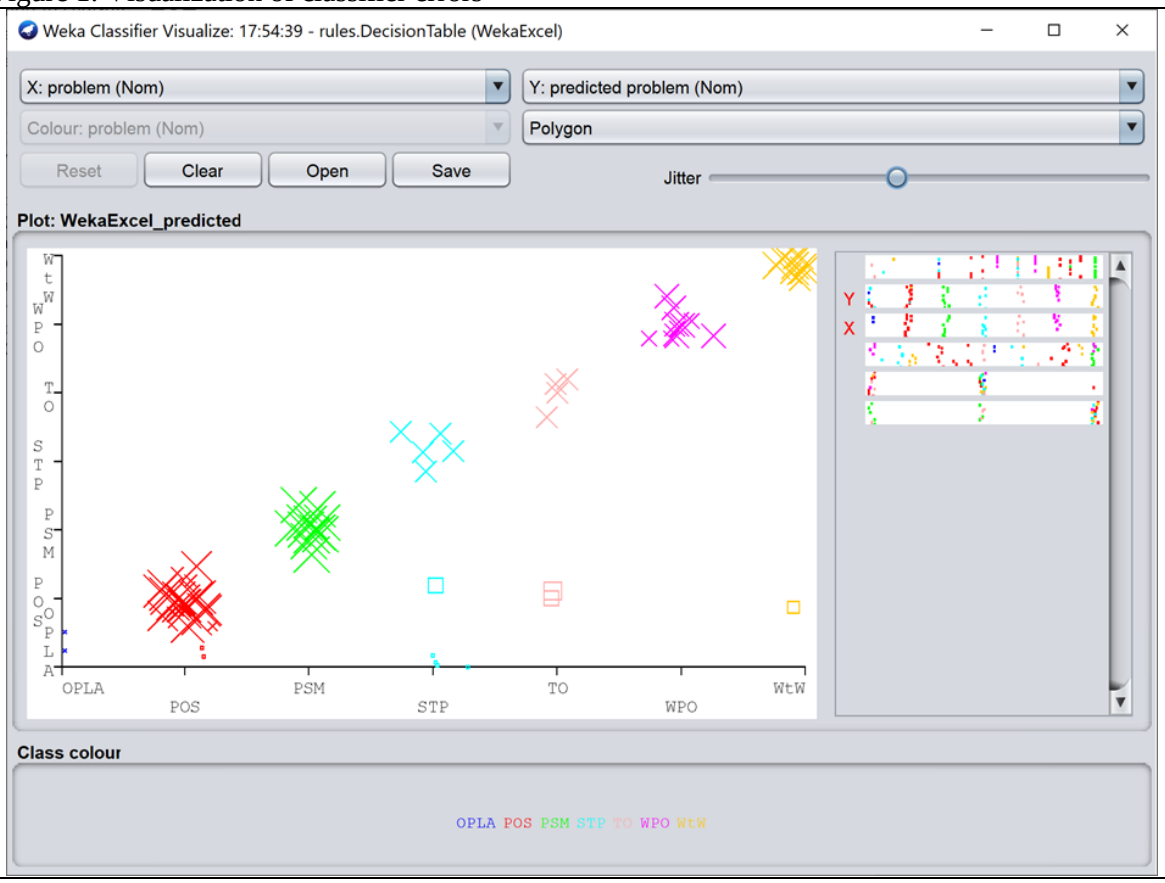
The decision tree is a classifier with a tree structure. Internal nodes are called decision nodes. These specify a test performed on the instance attribute, with each possible test result represented by one branch. The tree leaf indicates the value of the target property of the examples (class instance). The decision tree to classify the example begins at the root of the tree and passes through the individual nodes to the leaf that provides the instance classification. Decision tree induction is a typical inductive approach to knowledge mining through classification (Ahmim et al., 2019).

An essential criterion in the decision tree algorithm is selecting the sign (attribute) to be tested in each tree's decision node (Li et al., 2019). The aim is to select the attribute that best classifies the examples. A statistical property provides an excellent quantitative measure of an attribute's suitability called information gain, which indicates the extent to which an attribute divides training examples into its target classification. This measurement is performed by selecting from the candidate attributes at each step of the tree growth (Myles et al., 2004). The finished tree structure can be rewritten into a set of decision (classification, production) rules. Each classification rule contains a description of one path from the root node to a leaf node. These classification rules can then be used in a decision support system.

Statistical software packages (as SPSS, SAS, ...) can be used to create classification trees. We have decided to use the popular and, for this purpose, suitable tool Weka in version 3.8.4. Weka (Waikato Environment for knowledge analysis) is a suitable tool for performing many data mining tasks. According to Srivastava (2014), these include pre-processing data, selecting attributes, classification, clustering, and improving knowledge discovery using various meta classifiers. There are four steps involved in Weka for classification (Singhal & Jena, 2013), which we followed:

- Preparing the data
- Choose classify and apply an algorithm
- Generate trees
- The result or output.

Figure 1: Visualization of classifier errors



Source: Authors

A Logistics Problems Case

We verified the proposed method as the experimental case for logistic problems. The analysis subject was tasks from logistics and 80 well-described quantitative methods used in logistics to solve them (Brezina, 2003). The analysis of tasks focused not only on the description of the parameters of these tasks but also on the properties, conditions of use and the necessary input data of the methods used to solve a given task. An initial set of attributes was designed for each task, describing its characteristics and conditions of use of algorithms or methods and the type, structure, and variability of the data they work with.

The result of the analysis is a matrix (table), in which the rows contain the values of individual attributes defining a specific logistic task. The columns contain the values of the given attribute for different tasks. As logistic tasks are diverse, it is impossible to determine the value of all attributes for each of them, or the given attribute is irrelevant for identifying the task. In this case, the matrix's corresponding element is equal to NULL in the sense of an empty value. The values of all attributes are coded as categorical to allow classification. Initially, the attributes contained three numeric types of attributes, but we recoded them into categorical ones.

We used the Incremental Wrapper Subset Selection (IWSS) package to process the attributes. This attribute selector is a part of the Weka software. The author of this package, Bermejo (2020), describes how it works as follows:

> *"It first creates a ranking of attributes based on the selected metric. Then it runs an Incremental Wrapper Subset Selection over the ranking (linear complexity) by selecting attributes (using the WrapperSubsetEval class) which improve the performance for a given minimum number of folds out of the folds of the wrapper cross-validation. It contains the theta option, which permits tuning an early stopping (sublinear complexity). It contains the replaceSelection option, which tests at each step of the incremental search swapping a selected attribute by the current candidate. This reduces the mean number of selected attributes without decreasing performance. However, it increases the linear complexity to quadratic."*

As a result, it turned out that of the original number of proposed thirty-seven attributes, the attributes of *Type of problem*, *Type of solution*, *Model*, and *Goal* reduce the entropy rate the most. These attributes, therefore, became a test sign when creating the decision tree. Leaf nodes represent methods or algorithms suitable for solving a Problem specified by attribute values. The summary values of the classification are shown in Table 1, and the visualization of the classification errors for the *Type of problem* is shown in Figure 1.

Table 1: Classification summary

Name of criterium	value
Correctly Classified Instances	70 (87.5%)
Incorrectly Classified Instances	10 (12.5%)
Kappa statistic	0.844
K&B Relative Info Score	48.8845%
Complexity improvement (Sf)	99.0348 bits
Mean absolute error	0.1606
Coverage of cases (0.95 level)	100%
Mean rel. region size (0.95 level)	98.75
Total Number of Instances	80

Source: Authors

As can be seen from the results, this classification model's reliability is not 100%, mainly due to the data's sample size. Nevertheless, this model is suitable for use in DSS, as its outputs will only serve as recommendations. This model's construction highlighted the mapping of the relationship between the task (problem) and its solution method. Given the results achieved, it was not possible to reliably prove the existence of a causal relationship, but the relationship between the structure of the problem and its solution method was sufficiently visible. This relationship is indicated by the fact that there are inherently related methods of solution for similar problems (similar values of attributes).

Conclusion

This paper aimed to point out the importance of datafication as a fundamental precondition for capturing phenomena and facts that exist. Only after data collection can other natural steps take place to acquire knowledge to improve the company's results if we speak in the industry's functioning context. At present, many real-life phenomena are dataficated, and datafication products are already used through big data in many companies. Since there are still problems with data use, which were not the subject of purposeful datafication but arose as a natural consequence of life around us, there are few incentives for purposeful datafication. An example of purposeful data processing can be seen in the issue of logistics tasks discussed here. The analysis of logistics tasks and the subsequent design of a classification model made it possible to reveal the contours of the relationship between the characteristics of a logistics problem explicitly expressed through a set of attributes and the classes of methods used to solve them. Naturally, this direction of research will still need to be given due attention in the future.

Acknowledgements

This work was supported by a project VEGA No. 1/0373/18 entitled "Big data analytics as a tool for increasing the competitiveness of enterprises and supporting informed decisions" by the Ministry of Education, Science, Research and Sport of the Slovak Republic.

References

Ahmim, A., Maglaras, L., Ferrag, M. A., Derdour, M., & Janicke, H. (2019, 29-31 May 2019). *A Novel Hierarchical Intrusion Detection System Based on Decision Tree and Rules-Based Models.* Paper presented at the 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS).

Athamena, B., & Houhamdi, Z. (2018). Model for decision-making process with big data. *Journal of Theoretical and Applied Information Technology, 96*, 5951-5961.

Bermejo, P. (2020). IWSS: Incremental Wrapper Subset Selection (Version 1.0.0): WEKA.

Brezina, I. (2003). *Kvantitatívne metódy v logistike* [Quantitative methods in logistics]. Bratislava, Slovak Republic: Vydavateľstvo EKONÓM.

Daas, D., Hurkmans, T., Overbeek, S., & Bouwman, H. (2013). Developing a decision support system for business model design. *Electron. Mark., 23*(3), 251– 265.

Eriksson, Y. (2019). Digitalization of society: what challenges will users meet? HBiD - Human Behaviour in Design, Proceedings of the 2nd SIG conference, 125-127. DOI: 10.18726/2019_2

Fernández-Rovira, C., Álvarez Valdés, J., Molleví, G., & Nicolas-Sans, R. (2021). The digital transformation of business. Towards the datafication of the relationship with customers. *Technological Forecasting and Social Change, 162*. doi:10.1016/j.techfore.2020.120339

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems, 26*(3), 191-209.

Jeble, S., Kumari, S., & Patil, Y. (2018). Role of Big Data in Decision Making. *Operations and Supply Chain Management: An International Journal, 11*, 36. doi:10.31387/oscm0300198

Jia, L., Hall, D., & Song, J. (2015). *The Conceptualization of Data-driven Decision Making Capability*. Paper presented at the Twenty-first Americas Conference on Information Systems, , Puerto Rico.

Jones, M. (2019). What we talk about when we talk about (big) data. *Journal of Strategic Information Systems, 28*(1), 3-16. doi:10.1016/j.jsis.2018.10.005

Kościelniak, H., & Puto, A. (2015). BIG DATA in Decision Making Processes of Enterprises. *Procedia Computer Science, 65*, 1052-1058. doi:10.1016/j.procs.2015.09.053

Li, M., Xu, H., & Deng, Y. (2019). Evidential Decision Tree Based on Belief Entropy. *Entropy, 21*(9), 897.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*: Houghton Mifflin Harcourt.

Mejias, U. A., & Couldry, N. (2019). Datafication. *Internet Policy Review, 8*(4). doi:10.14763/2019.4.1428

Merendino, A., Dibb, S., Meadows, M., Quinn, L., Wilson, D., Simkin, L., & Canhoto, A. (2018). Big Data, Big Decisions: The Impact of Big Data on Board Level Decision-Making. *Journal of Business Research, 93*, 67-78. doi:10.1016/j.jbusres.2018.08.029

Musik, C., & Bogner, A. (2019). Book title: Digitalization & society. *Österreichische Zeitschrift für Soziologie, 44*(1), 1-14. doi:10.1007/s11614-019-00344-5

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society, 18*(6), 275-285.

Power, D. J. (2002). *Decision support systems: concepts and resources for managers*: Greenwood Publishing Group.

Singhal, S., & Jena, M. (2013). A study on WEKA tool for data pre-processing, classification and clustering. *International Journal of Innovative technology and exploring engineering (IJItee), 2*(6), 250-253.

Song, Y.-Y., & Ying, L. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry, 27*(2), 130.

Southerton, C. (2020). Datafication. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1-4). Cham: Springer International Publishing.

Srivastava, S. (2014). Weka: a tool for data pre-processing, classification, ensemble, clustering and association rule mining. *International Journal of Computer Applications, 88*(10), 26 - 29.

Travica, B. (2017). *Big Data Aspects and Decision Making*. Paper presented at the Sixth European Academic Research Conference on Global Business, Economics, Finance and Social Sciences, 1-3, July 2017, Italy

Yousuf, H., & Zainal, A. (2020). Quantitative Approach in Enhancing Decision Making Through Big Data as An Advanced Technology. *Advances in Science Technology and Engineering Systems Journal, 5*, 109-116. doi:10.25046/aj050515