DOI: 10.5937/finans2603088V
Primljen/Received: 08.04.2026./ April 8, 2026
Prihvaćen/Accepted: 25.05.2026./ May 25, 2026

UDK 658:004.8
Originalni naučni rad

ARTIFICIAL INTELLIGENCE IN THE FUNCTION OF SOLVING MISSING BUSINESS DATA

Vladimir VASIĆ[1]
Milica BUGARČIĆ[2]
** Odgovorni autor E-mail: vladimir.vasic@bba.edu.rs*

Abstract

In the era of modern economy, in the environment of increasing digitization of business, the market competition between business entities is more and more demanding. In order to withstand the extremely high pace of business imposed by competition, business entities are forced to turn to serious data analytics. In order for data analytics to be able to provide quality information, in addition to secondary data available to everyone, business entities turn to the production of primary data (which falls under the domain of trade secrets). Primary data can be collected by surveying or storing any digital data, which is created in the internal electronic system of a business entity. Also, during the survey, the interviewed person may not have the desire to answer all the questions; as with the collection of digital data from the internal electronic system, occasional downtime and interruptions may occur. So, whether we like it or not, incomplete data is all around us. Essentially, researchers sooner or later (usually sooner) encounter the problem of incomplete data

Keywords*: artificial intelligence, machine learning, data mining, missing data, classification and regression tree algorithm, business data*

JEL: *C45, C40, F14, G21*

[1] Vladimir Vasić, full professor, Belgrade Banking Academy - Faculty of Banking, Insurance and Finance, Zmaj Jovina 12, Belgrade, 011-2621-730, vladimir.vasic@bba.edu.rs, ORCID ID (https://orcid.org/0000-0001-9694-8071)
[2] Milica Bugarčić, associate professor, Belgrade Banking Academy - Faculty of Banking, Insurance and Finance, Zmaj Jovina 12, Belgrade, 011-2621-730, milica.bugarcic@bba.edu.rs, ORCID ID (https://orcid.org/0000-0002-9327-9965)

INTRODUCTION AND LITERATURE REVIEW

Introduction to Machine Learning and Artificial Intelligence

In modern business operations, there is an ever-increasing opportunity to collect and store a large amount of business data. Until recently, the collected data was stored through databases, however this format has been surpassed in terms of storing huge amounts of data. For this reason, data is increasingly being stored in the format of data warehouses as well as data lakes (*Inmon*, 2005, *Sawadogo & Darmont*, 2021)). Such a huge amount of data, it was no longer possible (within real time) to process using statistical tools (*Manyika et al.*, 2011). From that need, the development of quantitative sciences gave rise to a new discipline, which represents a combination of informatics and statistics, as well as new technologies, namely machine learning and artificial intelligence (*Dhar*, 2013; *Jordan & Mitchell*, 2015).

By using a new discipline, we are able to process large amounts of data in real time. This new discipline differs fundamentally (in a philosophical sense) from statistics. When it comes to the statistical approach in modelling business projects, business (qualitative) experts first defined (set) working hypotheses, which were then, through statistical analysis, either accepted or rejected. In artificial intelligence and machine learning, the order of steps is reversed, for the valid reason that we have too much data in big data warehouses (*Breiman*, 2001; *Bzdok et al.*, 2018). For this reason, in order to see if there is useful business information in those data warehouses, without any pre-set working hypotheses, the analysis of the given data begins. Of course, robust tools are used, which have the ability to process large amounts of data in a short time. Only after data processing, all established connections between variables are analyzed by business experts, and decisions are made, which connections are useful and usable, and which are not.

Machine learning and artificial intelligence algorithms

When it comes to supervised learning models, in many cases, decision tree tools are the most successful. A lot of different decision tree algorithms have been developed over time, the most famous of which are: the *classification and regression tree* algorithm, the *chi-square automatic interaction detector* algorithm, then the *exhaustive chi-square automatic interaction detector* algorithm, also the *analytical server chi-square automatic interaction detector* algorithm, then the *random tree* algorithm, as well as the *C5.0* algorithm.

Of all the listed most commonly used algorithms of artificial intelligence and machine learning, the *classification and regression tree* algorithm is the only one used for the problem of incomplete data. This algorithm is implemented in data mining software, when it comes to incomplete data. Also, this is the only, for now, algorithm in the field of artificial intelligence and machine learning, which is slowly being introduced, also in the method of multiple insertion, when it comes to incomplete data. Therefore, it is only a question of the day when other algorithms of artificial intelligence and machine learning will start to be used in the problem of incomplete data.

Literature review

The idea of using the *classification and regression tree* algorithm for imputation of missing values has been proposed by several authors in several different ways. For an overview of the use of the aforementioned algorithm for imputation purposes, you can find at *Saar-Tsechansky and Provost* (2007). In fact, most imputation methods based on decision trees are based on prediction (*Barcena and Tusell*, 2000). Also, some authors (*Harrell*, 2001; *Reiter*, 2005) considered the use of a decision tree in solving the problem of data incompleteness, in the procedure of multiple insertion. Also, *Parker* (2010) discusses the use of decision trees in multiple insertion procedures in various supervised and unsupervised learning models.

ARTIFICIAL INTELLIGENCE ALGORITHM - CLASSIFICATION AND REGRESSION TREE

Therefore, when it comes to the analysis of incomplete data, for the time being, the only artificial intelligence and machine learning tool used is the classification and regression tree algorithm. For this reason, this chapter will study the basics of this algorithm, using a very simple business example. Of course, this does not reduce the generality of the solution method, because with more complex data sets, the classification tree or the regression tree itself becomes quite complicated, but the rules by which decisions will be made, during the growth of the tree, will remain the same.

As part of the example, the data of one of the leading foreign banks in the Republic of Serbia will be used, namely from the client relationship management department. Of course, the sample size will be very small, in order to be able to show the decision rules of the classification and regression tree algorithm. Therefore, the project of interest is determining which bank product the client would opt for, based on its three characteristics.

When a given project is formulated in the form of a supervised learning model, the dependent variable would be, the *product*. The *product* has three categories, or modalities, and they are: *allowed minus*, *cash loan* and *savings*. Of the independent variables, the *education* variable stands out, consisting of three modalities: *secondary*, *colledge* and *university*. The next independent variable in the model would be the *bank-earning* in the previous year. Its modalities would be: the first category up to 12 thousand dinars, the second category would be from 12 thousand dinars to 30 thousand dinars; and the last category would be over 30 thousand dinars. It can be seen that in this business example interrupt variables figure, for the reason that it is much simpler and shorter to show the decision rules, with the *classification and regression tree*. Of course, the principle is the same with continuous variables, but due to the fact that continuous variables have many more different values (compared to an interrupt variable), the tables themselves, in which the statistics responsible for making a decision would be calculated, would be too long (while the principle, as with interrupt variables, is exactly the same).

Table 1. Client characteristics and product chosen

client	education	bank_earning	housing_type	product
1	secondary	12001-30000	at the parents' house	allowed minus
2	college	30001+	own property	cash loan
3	university	30001+	at the parents' house	allowed minus
4	secondary	12001-30000	own property	cash loan
5	college	12001-30000	renter	savings
6	university	0-12000	own property	allowed minus
7	secondary	0-12000	at the parents' house	cash loan
8	college	12001-30000	renter	savings
9	university	12001-30000	own property	cash loan
10	university	30001+	own property	allowed minus
11	college	12001-30000	at the parents' house	cash loan
12	secondary	12001-30000	at the parents' house	allowed minus

Source: Author's analysis.

Table 1 shows a sample of 12 clients. There, the client's characteristics are shown, as well as the bank's product, for which the given client has decided. Therefore, the business project would read: Based on the given three characteristics of the client, it is possible to determine, with great reliability, which bank product the given client would choose. (Of course, in a real business project of a given bank's client relationship management department, many more independent variables are used).

Otherwise, the *classification and regression tree* algorithm were proposed by *Breiman et al.* (1984). The classification tree created by the mentioned algorithm is of strictly binary type, which means that each node of the tree (which has the potential to be further divided) is divided into two newly formed nodes. Also, the *classification and regression tree* algorithm divide the analyzed sample of business data into subsamples with similar values (modalities) of the dependent variable. In this way, the tree actually grows, because the algorithm of the *classification and regression tree*, for each node of the tree, conducts an exhaustive search in relation to all independent variables, and in relation to all their different values, searching for the optimal division, in relation to some specific criterion. When it comes to the algorithm under analysis, it uses the criterion defined by *Kennedy et al.* (1995), which is defined by the expression

$$\Phi(s|t) = 2P_L P_R \cdot Q(s|t) = 2P_L P_R \cdot \sum_{j=1}^{modalities} |P(j|t_L) - P(j|t_R)| \qquad (1)$$

where s denotes a possible division, t denotes a node that has the potential for further division, and *modalities* represent all possible different values of the dependent variable. Then, t_L denotes the left newly formed node originating from the node t, t_R denotes the right newly formed node originating from the node t. P_L denotes the number of observations in the newly formed hub t_L in relation to the total number of observations in the sample. P_R denotes the number of observations in the newly

formed hub t_R in relation to the total number of observations in the sample. $P(j|t_L)$ denotes the number of the *j*th *modality* of the dependent variable in the left newly formed node t_L. And finally, $P(j|t_R)$ denotes the number of the *j*th modality of the dependent variable in the right newly formed node t_R.

Table 2. Possible divisions based on independent variables and their values

possible division	independent variable	t_L	t_R
1	education	secondary	college & university
2	education	college	secondary & university
3	education	university	secondary & college
4	bank_earning	0-12000	12001-30000 & 30001+
5	bank_earning	12001-30000	0-12000 & 30001+
6	bank_earning	30001+	0-12000 & 12001-30000
7	housing_type	at the parents' house	renter & own property
8	housing_type	renter	at the parents' house & own property
9	housing_type	own property	at the parents' house & renter

Source: Author's analysis.

So, when the algorithm goes through all possible partitions, it will choose the one that maximizes the function $\Phi(s|t)$. Within this business example, for the beginning, all 12 clients that are presented in table 1 are classified first in the initial one, i.e. zero hub. As already noted, the *classification and regression tree* algorithm is limited to binary partitioning only, so all possible partitions, with respect to all independent variables and all their possible values, are given in Table 2.

By looking at table 2, we come to the conclusion that when it comes to the zero hub, there are a maximum of nine different possible divisions. On the basis of which possible division the zero hub will be further divided, depends on the previously defined function $\Phi(s|t)$. By using the implemented business model, the properties of the function in relation to which the division of the zero node is performed can be examined. For example, under what conditions does the function $\Phi(s|t)$ take large values. Due to the fact that the function $\Phi(s|t)$ consists of two factors $2P_LP_R$ and $Q(s|t)$, we come to the conclusion that when its factors take large values, then the function $\Phi(s|t)$ itself will take large values.

Table 3. Value of the function $\Phi(s|t)$ and its factors for all possible divisions

possible division	P_L	P_R	$2P_LP_R$	modalities	$P(j\|t_L)$	$P(j\|t_R)$	$Q(s\|t)$	$\Phi(s\|t)$
1	0.333	0.667	0.4442	allowed minus	0.5	0.375	0.5	0.2221
				cash loan	0.5	0.375		
				savings	0	0.25		
2	0.333	0.667	0.4442	allowed minus	0	0.625	1.25	0.5553
				cash loan	0.5	0.375		
				savings	0.5	0		
3	0.333	0.667	0.4442	allowed minus	0.75	0.25	1	0.4442
				cash loan	0.25	0.5		
				savings	0	0.25		
4	0.167	0.833	0.2782	allowed minus	0.5	0.4	0.4	0.1113
				cash loan	0.5	0.4		
				savings	0	0.2		
5	0.583	0.417	0.4862	allowed minus	0.429	0.4	0.116	0.0564
				cash loan	0.429	0.4		
				savings	0.142	0.2		
6	0.25	0.75	0.3750	allowed minus	0.667	0.333	0.668	0.2505
				cash loan	0.333	0.445		
				savings	0	0.222		
7	0.417	0.583	0.4862	allowed minus	0.6	0.286	0.628	0.3053
				cash loan	0.4	0.428		
				savings	0	0.286		
8	0.167	0.833	0.2782	allowed minus	0	0.5	2	0.5564
				cash loan	0	0.5		
				savings	1	0		
9	0.714	0.286	0.4084	allowed minus	0.4	0.428	0.628	0.2565
				cash loan	0.6	0.286		
				savings	0	0.286		

Source: The result of the author's analysis.

For this reason, the factor $Q(s|t)$ is considered first. Given a factor that is actually equal to the expression $\sum_{j=1}^{modalities}|P(j|t_L) - P(j|t_R)|$ will take large values if there is a maximum difference between the probabilities $P(j|t_L)$ and $P(j|t_R)$ in relation to all modalities of the dependent variable. In other words, the given factor will take the maximum value, if the newly formed hubs have as different a proportion of the modality of the dependent variable as possible. This maximum is therefore achieved if the newly formed hubs are as homogeneous as possible. Otherwise, the theoretical maximum for the factor $Q(s|t)$ is two.

The second factor $2P_LP_R$ maximizes its value, when the probabilities P_L and P_R reach large values simultaneously, and this happens in the situation when the number of observations is evenly divided into the newly created left and right nodes. Therefore, the function $\Phi(s \mid t)$ favors balanced divisions, that is, that newly formed hubs have an equal number of observations. Therefore, the function $\Phi(s \mid t)$ prefers the division into newly formed hubs that are homogeneous in terms of the modality of the dependent variable, and which are, as much as possible, equal in terms of the number of observations. Otherwise, the theoretical maximum for the factor $2P_LP_R$ is $2 \cdot 0.5 \cdot 0.5 = 0.5$, so the theoretical maximum of the function $\Phi(s \mid t)$ in this example would be $2P_LP_R \cdot Q(s|t) = 0.5 \cdot 2 = 1.0$.

Table 3 shows the values of the function $\Phi(s \mid t)$ for all possible divisions of the zero node. It is established that the highest value of the given function is 0.5564 and that is for a possible division of eight. Table 2 shows that in a possible division of eight, the independent variable housing type divided the zero node into two newly formed nodes, one of which are renters, and the other with parents, or own real estate. Therefore, in one newly created hub, clients have the characteristic that they are renters according to the type of housing. There are two such clients, and both chose the bank's product - savings. Due to the fact that in a given newly formed node, the dependent variable always has the same modality - saving, the given node cannot be further divided, it is therefore the final node.

Table 4. Possible divisions based on independent variables and their values

possible division	independent variable	t_L	t_R
1	education	secondary	college & university
2	education	college	secondary & university
3	education	university	secondary & college
4	bank_earning	0-12000	12001-30000 & 30001+
5	bank_earning	12001-30000	0-12000 & 30001+
6	bank_earning	30001+	0-12000 &12001-30000
7	housing_type	at the parents' house	renter & own property
8	housing_type	own property	at the parents' house & renter

Source: Author's analysis.

The other newly created node included the remaining observations of the zero node, of which there are ten, and which take the remaining two modalities of the dependent variable (allowed minus and cash loan) and therefore have the potential for further division. Therefore, the remaining ten observations can theoretically be divided in eight ways, which is shown in table 4. Which of these possible divisions will happen depends on the value of the function $\Phi(s \mid t)$ for these divisions. The values of the given function are given in table 5, the analysis of which shows that the maximum of the given function is at division number 3. In table 4, division number three, the independent variable education is given, which is divided into university (on the one hand) and secondary & colledge (on the other hand). Looking now at table 1, the remaining ten elements are divided into two newly formed hubs, one of which has

four elements and the other six elements. Due to the fact that both newly formed hubs are not homogeneous in terms of the modality of the dependent variable, they have the potential for further division.

Table 5. Value of the function $\Phi(s|t)$ and its factors for all possible divisions

division	P_L	P_R	$2P_LP_R$	modalities	$P(j\|t_L)$	$P(j\|t_R)$	$Q(s\|t)$	$\Phi(s\|t)$
1	0.4	0.6	0.4800	allowed minus	0.5	0.5	0	0.0000
				cash loan	0.5	0.5		
				savings	0	0		
2	0.2	0.8	0.3200	allowed minus	0	0.625	1.25	0.4000
				cash loan	1	0.375		
				savings	0	0		
3	0.4	0.6	0.4800	allowed minus	0.75	0.333	0.834	0.4003
				cash loan	0.25	0.667		
				savings	0	0		
4	0.2	0.8	0.3200	allowed minus	0.5	0.5	0	0.0000
				cash loan	0.5	0.5		
				savings	0	0		
5	0.5	0.5	0.5000	allowed minus	0.4	0.6	0.4	0.2000
				cash loan	0.6	0.4		
				savings	0	0		
6	0.3	0.7	0.4200	allowed minus	0.667	0.429	0.476	0.1999
				cash loan	0.333	0.571		
				savings	0	0		
7	0.5	0.5	0.5000	allowed minus	0.6	0.4	0.4	0.2000
				cash loan	0.4	0.6		
				savings	0	0		
8	0.5	0.5	0.5000	allowed minus	0.4	0.6	0.4	0.2000
				cash loan	0.6	0.4		
				savings	0	0		

Source: The result of the author's analysis.

Of course, the calculation of the value of the function $\Phi(s \mid t)$ at new hubs will no longer continue, because the division is performed according to the described procedure. The complete division of all nodes in this example, which formed the *decision tree*, is given in Figure 1. In the given figure, it is noted that the value of the branching level of the tree is five. Then, that all independent variables (education, bank earnings, and housing type) participated in the further branching of the tree, as well as that the number of final (or end) hubs is eight.

In the end, it can be said that the final hubs participate in determining which bank product a potential client would choose. Of course, it should be pointed out that this example elaborated only the essence of decision-making with the *classification and*

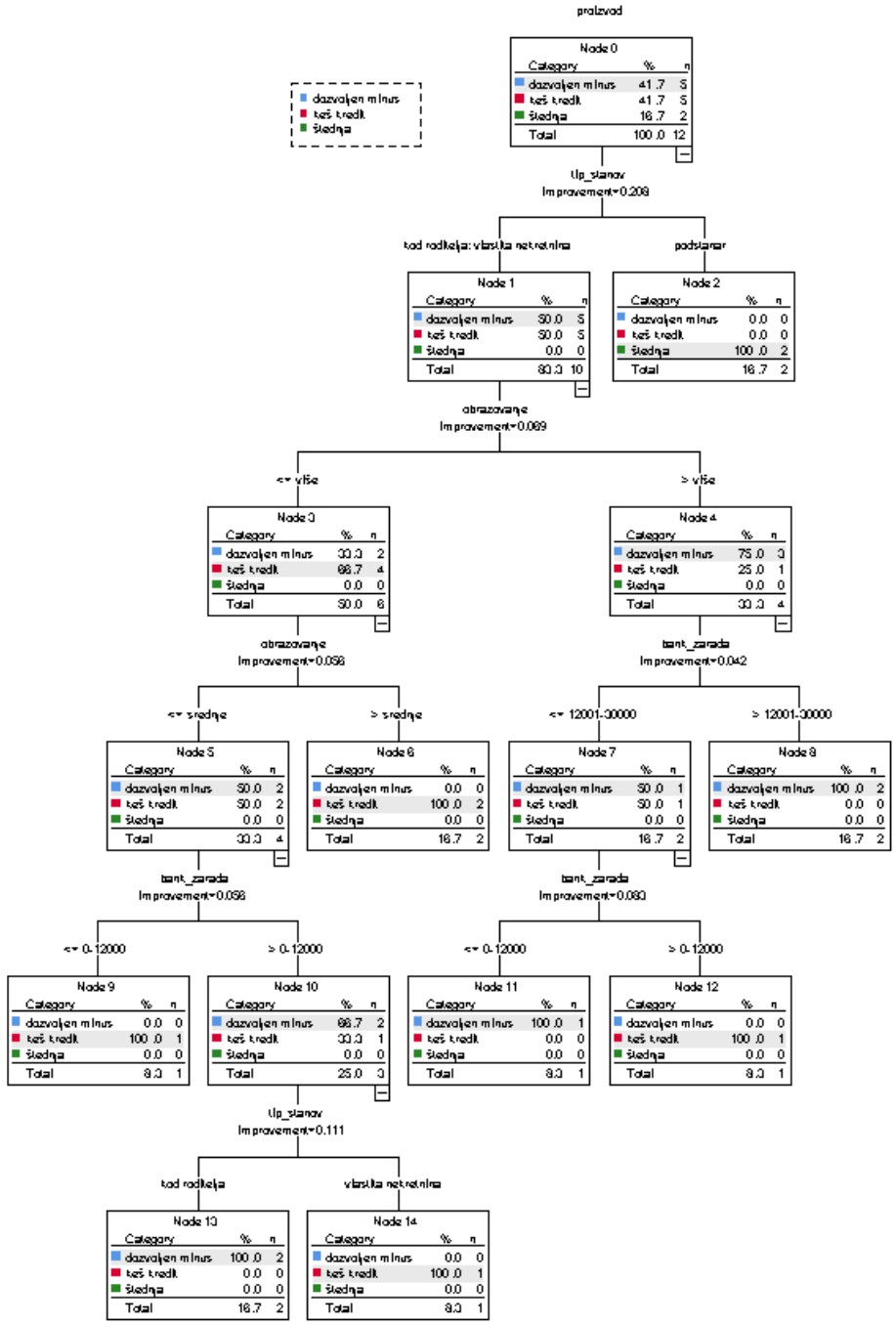
regression tree algorithm, but also that in addition to them, there are other elements that are needed when analyzing large databases.

PRACTICAL APPLICATION OF ARTIFICIAL INTELLIGENCE ALGORITHMS AT INSERTING MISSING DATA

While the sample from the previous business example contained only 12 clients (for the reason that the tables "Possible divisions based on independent variables and their values" and "Value of the function $\Phi(s \mid t)$ and its factors for all possible divisions" could be displayed completely), the business project that is now analyzed is 106 respondents, and is presented in detail in the research "How much does it cost to have a business in Serbia?" from 2003 (*Vasić et al.*, 2003). The same research was conducted for the first time in 2001. Among the many tasks, which were before the given research, there was also the determination of the key factors of the economic environment by businessmen. In that first research, based on six questions, two factors were singled out: the legal environment (*ambience*) and the indirect influence of the state (*influence*). The questions that were asked to the businessmen who participated in the research were: stability of the legal environment (*stability*), simplicity and comprehensibility of the legal environment (*simplicity*), legal possibilities/limitations (*legal*), administrative procedures (*administrative*), working hours of administration and service institutions (*time*), and behavior of state administration officials (*behavior*).

Please note that in this business project there is the presence of incomplete data, which will be solved by using the previously detailed algorithm, and by using *IBM SPSS Modeler* data mining software. As we have already indicated, this algorithm represents a popular class of artificial intelligence and machine learning algorithms. The *classification and regression tree* algorithm performs a search by independent variables and their values, in order to divide the analyzed data into two subsamples. The splitting process is then continued over the subsamples, so that a series of splits creates a binary decision tree. As already mentioned, the dependent variable can be of an intermittent type (in which case the decision tree is a classification tree) or a continuous type (in which case the decision tree is a regression tree).

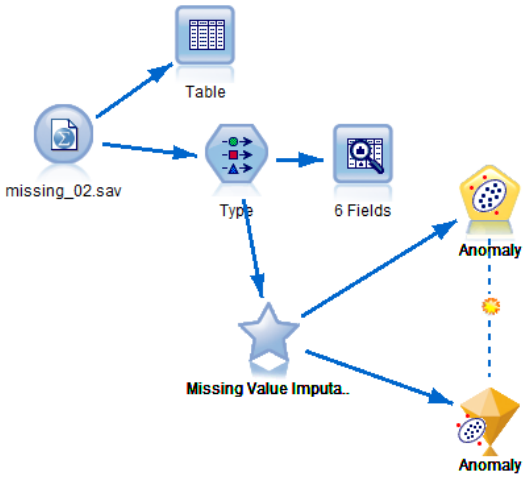
Figure 1. Classification tree view



Source: The result of the author's analysis conducted using IBM SPSS Statistics 26 software.

By the way, the *classification and regression tree* algorithm have properties that are very desirable, when it comes to the procedure of inserting missing data in incomplete variables. First, the algorithm is robust in relation to extreme values, then it is not affected by multicollinearity, as well as asymmetry of variable distributions. Also, the algorithm is flexible, when it comes to the existence of interaction between variables, as well as the existence of non-linear connection between them. Also, the application of the *classification and regression tree* algorithm itself is quite automated, so that only moderate adjustments are needed by the researcher, who plans to use the algorithm for solving the problem of variable incompleteness (*Burgette and Reiter*, 2010).

Figure 2. Display of stream and super-node in solving the problem of incomplete data

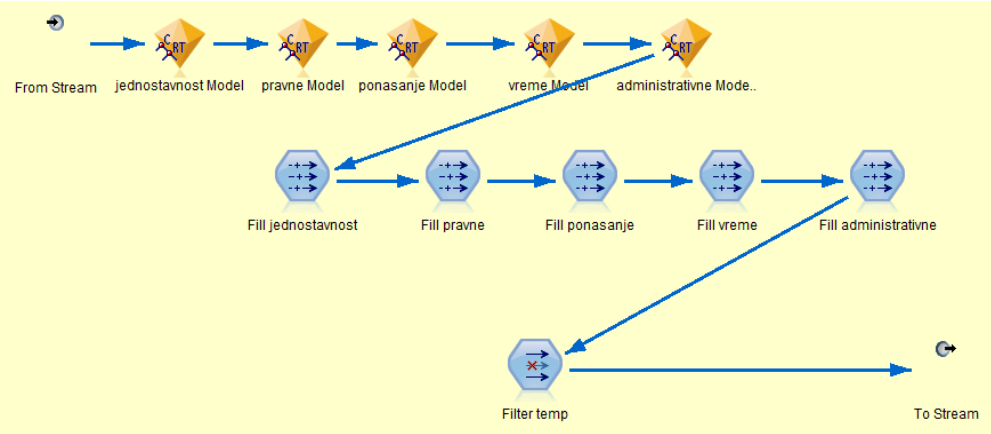


Source: The result of the author's analysis conducted using IBM SPSS Modeler 18.0 software.

In the application considered here, the procedure for inserting incomplete data using the *classification and regression tree* algorithm is shown in Figure 2. When it comes to data mining software (*IBM SPSS Modeler*), the problem of incomplete data is solved by creating a super-node *imputation of missing values* (*IBM*, 2017; *IBM* 2019).

A given super-node (shown in the data mining flow (Figure 2)) is in the shape of a five-pointed star, actually representing a complex analysis. The content of the complex analysis can be viewed by reviewing the content of the super-node, given in Figure 3. Before that, let's recall the content of the data file, which is considered in this business project. Six variables were analyzed, of which the following five had missing values: stability of the legal environment (variable name: *simplicity*), legal possibilities/constraints (variable name: *legal*), administrative procedures (variable name: *behavior*), working hours of administration and service institutions (variable name: *time*), and behavior of state administration officials (variable name: *administrative*).

Figure 3. Content of the super-node in solving the problem of incomplete data



Source: The result of the author's analysis conducted using IBM SPSS Modeler 18.0 software.

In Figure 3, it can be seen that the super-node *imputation of missing values* solved the problem of missing values for each variable by using the *classification and regression tree* algorithm. Therefore, for each variable that had the problem of incomplete data (and there were five of them), imputation of missing values was performed using the *classification and regression tree* algorithm, separately, i.e. individually; and then return to the main stream of data mining.

Finally, note that the same idea of the considered algorithm is used in the procedure of matching average projections (*van Buuren*, 2018), where average projections are calculated using a regression model (instead of a classification tree). Otherwise, the given procedure represents one of the methods for single insertion of data, when it comes to continuous incomplete variables.

CONCLUSION

In the paper, two things are gradually presented for the first time, based on real business data within two separate domestic business projects. The first, a detailed mathematical statistical procedure for solving the division (branching) of the decision tree using the *classification and regression* algorithm (which is given in Tables 1-5, and Figure 1), and the second, the implementation of the algorithm itself within the data mining software, i.e. artificial intelligence and machine learning (Figures 2 and 3).

The paper shows that by applying artificial intelligence, the problem of missing data is solved in a correct and complex (and easy to use) way. In current practice, researchers did not solve the problem of missing data at all (or they did, but in an inappropriate way, such as deleting observations with missing values, or imputing the value of the arithmetic mean in place of the missing data) (*Raghunathan*, 2016). Now, with the application of artificial intelligence tools, this serious problem is overcome with minimal effort, so researchers have the ability to dispose of correctly collected input data, which can lead them to unbiased information and business conclusions.

LITERATURE

Barcena, M. J. and Tusell, F. (2000). Tree-based algorithms for missing data imputation. In Bethlehem, J. G. and Van der Heijden, P. G. M., editors, *COMPSTAT 2000: Proceedings in Computational Statistics*, pages 193-198. Heidelberg, Germany: Physica-Verlag.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3), 199–231. https://doi.org/10.1214/ss/1009213726

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Boca Raton, FL: Chapman & Hall/CRC Press.

Burgette, L. F. and Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172(9):1070-1076.

Bzdok, D., Altman, N., & Krzywinski, M. (2018). Points of Significance: Statistics versus machine learning. *Nature Methods*, 15(4), 233–234. https://doi.org/10.1038/nmeth.4642

Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-69.

Harrell, F. E. (2001). *Regression Modeling Strategies*. New York: Springer-Verlag.

IBM (2017). *IBM SPSS Statistics Algorithms*. Armonk, NY: IBM Corporation.

IBM (2019). *IBM SPSS Statistics 26 Command Syntax Reference*. Armonk, NY: IBM Corporation.

Inmon, W. H. (2005). Building the Data Warehouse. John Wiley & Sons.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Kennedy, R. L., Lee, Y., van Roy, B., Reed, C. D. and Lippman, R. P. (1995). *Solving Data Mining Problems through Pattern Recognition*. Upper Saddle River, NJ: Pearson Education.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Parker, R. (2010). *Missing Data Problems in Machine Learning*. Saarbrucken, Germany: VDM Verlag Dr. Muller.

Raghunathan, T. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press / Taylor & Francis Group.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics*, 21(3):7{30.

Saar-Tsechansky, M. and Provost, F. (2007). Handling missing values when applying classi_cation models. *Journal of Machine Learning Research*, 8:1625{1657.

Sawadogo, P., & Darmont, J. (2021). On data lake architectures and metadata management. *Information Systems*, *98*, Article 101730. doi.org

van Buuren, S. (2018). *Flexible Imputation of Missing Data* (Second Edition). Boca Raton, FL: CRC Press - Taylor & Francis Group.

Vasić, V., P. Bjelić, K. Ognjenović, K. Vukojčić, D. Gajić, Ž. Ivanji, A. Nikolić, J. Momčilović, N. Đunisijević, R. Pupovoc, i I. Ičin. (2003). *Coast of Doing Business in Serbia? II*. Belgrade: G17 Institute. Editor Goran Petković.

CONFLICT OF INTEREST

The authors declare that there are no financial, professional or personal relationships that could lead to bias in the results or interpretation of this research.

VEŠTAČKA INTELIGENCIJA U FUNKCIJI REŠAVANJA NEDOSTAJUĆIH POSLOVNIH PODATAKA

Vladimir VASIĆ
Milica BUGARČIĆ

Apstrakt

U eri savremenog privređivanja, u ambijentu sve veće digitalizacije poslovanja, tržišna utakmica između poslovnih subjekata je sve više i više zahtevnija. Poslovni subjekti da bi izdržali, izuzetno visok tempo u privređivanju koji nameće konkurencija, prinuđeni su da se okrenu ozbiljnoj analitici podataka. Da bi analitika podataka bila u mogućnosti da obezbedi kvalitetne informacije, osim svima dostupnih sekundarnih podatka, poslovni subjekti se okreću produkciji primarnih podataka (koji spadaju u domen poslovne tajne). Primarni podaci mogu se sakupljati anketiranjem ili skladištenjem svakog digitalnog podatka, koji nastaje u internom elektronskom sistemu privrednog subjekta. Takođe, u toku anketiranja, ispitivano lice možda neće imati želju da odgovori na sva pitanja; kao i što kod prikupljanja digitalnih podataka iz internog elektronskog sistema, mogu nastati povremeni zastoji i prekidi. Dakle, želeli mi to, ili ne, nekompletni podaci su svuda oko nas. Suštinski, istraživači se pre ili kasnije (a obično pre) susretnu sa problemom nekompletnih podataka.

***Ključne reči**: veštačka inteligencija, mašinsko učenje, rudarenje podataka, nedostajući podaci, algoritam stabla klasifikacije i regresije, poslovni podaci*