ECONOMIC THEMES (2016) 54(4): 563-586

http://www.economic-themes.com/

CHAID DECISION TREE: METHODOLOGICAL FRAME AND APPLICATION

Marina Milanović

Faculty of Economics, University of Nis, Serbia

✉ *milanovicm.ar67@gmail.com*

Milan Stamenković

Faculty of Economics, University of Kragujevac, Serbia

✉ *m.stamenkovic@kg.ac.rs*

UDC 005.322:316.46
Original scientific paper

Abstract: Technological advancement across human activities has brought about accelerated generation of huge amounts of data. Consequently, researchers are faced with the problem how to determine adequate ways of turning the available data mass into useful knowledge. Data analysis adapted to these changes when data mining was developed as an approach to data analysis from different perspectives which reveals significant hidden regularities. This paper presents conceptual characteristics of decision tree, an important data mining method which is, due to its explorative nature, exceptionally suitable for detection of data structure when analysing various problem situations. The empirical section of the paper demonstrates applicative characteristics of this method using CHAID algorithm in leadership studies: an interdependence of selected personal characteristics and the manager's leadership style has been investigated. The aim of the paper is to develop a classification model for identification of the dominant leadership style. The study was conducted on the sample of 417 managers of privately owned small-sized enterprises in Serbia, using a specially designed questionnaire. The classification model identified the set of six statistically significant personal characteristics as predictors of dominant leadership style.

Received: 12.12.2016.
Accepted: 01.02.2017.

Keywords: data mining, decision tree method, CHAID algorithm, leadership styles, leaders' personal characteristics, classification rules, dependency
JEL classification: C38, D83, M50

1. Introduction

In the modern world, the systems for management of data, information and knowledge are offering new potential for improvement of business and quality of business decisions. The analysis of economic activity in the past decades reveals a strong influence and intense application of IC technology, which has resulted in informatisation of nearly all business processes. These trends have lead to accelerated generation and storage of huge amounts of complex data. On the other end of the spectrum, technological advancement has enabled software support for the application of quantitative methods and created a new challenge: how to turn huge, easily available mass of data into highly valuable information, i.e. knowledge.

In order to adjust the manner of data analysis to the changes caused by the abundance of data and its speeded flow, and to bridge the gap between the availability and use of data, the data mining (DM) was developed as a multidisciplinary approach, suitable for quick analysis of large amount of data and identification of significant patterns hidden deep inside databases.

Bearing in mind that there is a wide spectrum of DM methods which cannot all be presented here, this study focuses on: ▸ the decision tree theoretical framework as a method which is particularly appropriate for the purpose of exploratory knowledge discovery, and ▸ implementation possibilities of CHAID decision tree algorithm in solving problem situations, pertaining to business activities and economy in general, defined in the form of predictive DM tasks. In the empirical part of the paper, the application of CHAID algorithm is demonstrated in the field of leadership, which is, under conditions of intense and turbulent changes, one of the key factors of business success. In this context, the following research question is formulated: Is it possible, and with which probability, to determine the dominant leadership style of a manager using a particular set of (statistically significant) personal characteristics?

Respecting the importance of the impact of personal characteristics on leadership style, verified by numerous authors in the theory of leadership and management, the main objective of this paper is to show that the dominant style of leadership behaviour of managers can be identified through the development of classification model using CHAID algorithm. Accordingly, the basic hypothesis of the paper is as follows: CHAID procedure, as a form of multivariable analysis is a usefull methodological tool which, through data dimensionality reduction, enables identification of personal characteristics (variables) which have the strongest interaction with the dominant leadership style (as the dependent variable).

After the introduction, a review of relevant literature and empirical studies is presented. This is followed by conceptual and methodological determinations of decision tree method, with special emphasis on CHAID method. The theoretical

grounding is followed by a detailed presentation of paticular empirical research based on the implementation of CHAID algorithm, which includes: the definition of the problem situation, the research methodology, settings of the algorithm parameters, the obtained results and their interpretation, as well as quality assessment of the developed model. The final section of the paper briefly summarises the study's achievements and limitations together with suggestions for further work.

2. Literature Review

Great interest for the analysis of *DM* methods and their application in various problem situations, conditioned by tendencies mentioned in the introduction, contributed to the publication of numerous forms of scientific and professional publications that contain different aspects of observation and results of conducted theoretical and empirical research.

Investigating relevant bibliographic units (Kantardzic, 2011; Nisbet et al., 2009; Shmueli et al., 2010; Vercellis, 2009; Witten & Frank, 2005; Han et al., 2012), in which the characteristics of fundamental DM methods are presented (including also a decision tree method), with mainly hypothetical illustrations, it can be noticed the presence of standardised and comprehensive form of presentation of relevant aspects of DM analysis, ranging from the data preparation and model creation to its evaluation and implementation. When it comes to decision tree, these authors provide sufficient information to understand the principle of basic algorithms' functioning and fundamental issues pertaining to the method's successful application in the search of DM solutions to the problems of classification and prediction. In the study entirely dedicated to decision tree for business intelligence and DM as "a simple, but powerful form of multiple variable analysis", author de Ville (2006, p.1), examines most relevant dimensions of the entire group of decision tree methods and algorithms.

Of the many issues, which are elaborated in detail towards achieving, already proclaimed in the introduction, the objective of the study "to increase the utility and decrease the futility of using decision trees" (deVille, 2006, p.viii), in particular the following two stand out: the role of statistics in decision tree development process and the synergetic effects of the integration of decision tree with other DM methods. However, the most comprehensive study of decision trees as data mining method is Rokach and Maimon (2008). In their study, the authors state that decision tree, originally stemming from logic, management and statistics, is one of the most promising and popular approaches in the investigation and identification of useful regularities in large sets of complex data. Covering almost all aspects of this methodological approach, its importance in the realisation of a

wide spectrum of DM tasks, such as classification, regression, estimation, clustering and selection of variables is particularly emphasised.

Numerous studies conducted in different areas confirm the universality of predictive modeling based on DM methodology. In addition, in these studies, frequent applications of all algorithms from decision tree class have been identified. Considering that the modeling in the empirical part of this paper is carried out using the CHAID[1] algorithm, a brief overview of scientific papers in which this algorithm is implemented in solving practical issues and DM tasks in the field of economics and education (in order to discover information relevant for management in education) follows.

Soldić-Aleksić (2009) in her study emphasizes the importance of market segmentation, as one of the most important concepts of modern marketing, and also, on the empirical example, demonstrates the application of logistic regression (for selection of predictor variables) and CHAID analysis for generating a predictive model of market segmentation of visitors of one fast food restaurant. Furthermore, the implementation of CHAID analysis in marketing and its usefulness in (tourist) market segmentation (including its comparison with the results of discriminant analysis) has been confirmed in the study by Díaz-Pérez and Bethencourt-Cejas (2016). Kim et al. (2010) examine and assess consumer preferences, the behaviour of Japanese tourists in Korea as well as their intentions to revisit the country, using CHAID procedure. Öcal et al. (2015) conducted their research in order to develop a predictive model to evaluate financial success / failure of manufacturing companies applying C5.0 and CHAID algorithms. As business activities planning is based on historical data and as material planning is a very important function for manufacturing and service providing companies, Kağnicioğlu and Moğol (2014) create and examine CHAID decision tree for material planning in housekeeping processes in hotel industry. Novotná (2012) focuses on companies' credit rating modeling using data from the sample of 4892 companies from eight Central and East European countries using discriminant analysis, logistic regression and decision tree (CHAID and CART algorithms). As leasing has become a substantive financing source in modern economy, Horvat et al. (2012) focus on the development of classification model for the detection of frauds in leasing agreements using CHAID algorithm. Popescu et al. (2014) examine macro economic development indicators for EU countries, trying to improve the process of economic performance evaluation using hierarchical cluster analysis and CHAID methodology to determine the variables that can be used as best predictors of high *vs* low economic performance countries, including determination of consistent classification rules. Student performance modeling and

[1] *CHAID* is the acronym for *Ch*i Square *A*utomatic *I*nteraction *D*etection.

the comparison of results obtained using different DM algorithms for the needs of faculty/university management is the topic of numerous recently published papers. To illustrate, Baran and Kılıç (2015) use *CHAID* algorithm to examine students' characteristics (such as demographic characteristics, study habits and technology familiarity) and their association with students' achievement.

This literature overview lists numerous (although not all) possibilities for CHAID algorithm application in solving DM tasks converted into concrete problem situations in the field of economics. Therefore, the contribution of this paper manifests in the application of CHAID algorithm in the field of leadership studies in order to identify the (combination of) managers' key personal characteristics which determine the dominant leadership style. Although there are a large number of publications and research projects addressing the issue of leadership (Stojanović Aleksić, 2007), on the one hand, and numerous scientific papers in which different application possibilities of *CHAID* algorithm are demonstrated, on the other hand, the originality of the presented research idea is reflected in the fact that in the online databases and publications, particular studies that combine these two concepts were not found.

3. Decision Tree

Decision tree is one of the most frequently used predictive methods to create DM model in order to solve various DM tasks. Basically, this method is used in classification, as well as in realisation of other predictive *DM* tasks such as prediction, regression and estimation, as well as for data description, vizualisation and dimensionality reduction. It is a method of inductive reasoning and supervised learning. It belongs to the category of non-parametric methods.

3.1. The Basic Concepts of the Hierarchical Structure of Decision Tree

As a form of multivariable analysis, decision tree is a method of data modeling which, depending on the selected criterion and the values of input variables divides (classifies) units of observation of a heterogenous and usually large population into a certain number of smaller, predefined homogenous classes (categories, groups) of the output variable. Input variables are independent (predictor) variables, which may be categorical or numerical, while the output (target) variable is dependent (usually categorical variable). The resulting model of interdependence between the input variables and output variable is graphically represented in the form of a tree, which is reflected in the method's name.

Decision tree structure comprises a hierarchically organised set of data groups, called nodes[2], which are interconnected by tree branches. In the beginning of the hierarchical structure is the root of the tree (i.e. root node). It refers to the dependent variable and includes all the observed data of the training sample that is to be divided into certain classes during the process of model development. The root node has no input branches, but it can have two or more output branches, unless all the elements belong to the same class, in which case the node has no output. Contrary to the root node, all other nodes have exactly one input branch (Rokach & Maimon, 2008).

The node with output branches is termed the internal node. Every internal node corresponds to one input variable and has one input and, depending on the possible values of the output variable, two or more output branches. Outputs, or responses (presented as tree branches) can be of the binary type, or in the form of a choice between several possible values or value intervals. Internal nodes are also called test nodes or decision nodes, because in each of them some condition for the particular variable is examined. The nodes with a single input branch and no output branches are called terminal nodes or tree leaves. Terminal nodes are the possible solutions to the problem.

Thus, every branch, as an element of the tree structure, ends with the decision or terminal node. Nodes and branches of the tree are also the elements of knowledge about the relationship between the target variable and input variables. This knowledge is expressed in the form of *if-then* rules. Basically, decision tree comprises all the observed data. Their sorting into homogenous groups takes place along the path from the root to the terminal node. Each path from the root to the leaf presents a single rule. The purpose of producing a decision tree is to generalise the data structure by determining a set of logical *if-then* rules, which will enable precise classification of not only existing but new data and prediction of the value of dependent variable.

Generally, in the field of operational research and decision theory, the concept of the decision tree is interpreted as a graphical representation of the decision-making problem in conditions of uncertainty, and it represents a hierarchical model of a set of decisions and natural states (events) with their possible outcomes (i.e. effects). On the other hand, in DM field the primary purpose of decision tree is the classification of a certain set of input data. In addition, the performed classification does not result in a single decision. It serves to formulate a set of rules to be applied when solving the observed problem. Conceptually, the decision tree is a hierarchical model consisting of a set of rules for the partition of the heterogenous input data set into groups which are homogenous regarding the dependent variable categories.

---

[2] *Node* is a common, generic term for decision tree elements

3.2. Key Issues in Recursive Partitioning Procedure

Methodologically, decision tree is produced by the procedure of recursive partitioning of data into groups in such a way to maximise homogeneity – purity (i.e. to minimise entropy) regarding the target variable in each of the obtained groups. The learning process takes place by comparing the potential divisions, first, of the entire set of observed data based on each independent variable and the selection of the best partition according to the relevant criteria. Subsequently, the tree is developed by consequtive repetition of these steps in every node on every level of the tree's hierarchical structure. The process ends when homogenous groups have been obtained or when one of the defined criteria for recursion termination has been met. Accordingly, the first level of branches and internal nodes represents the independent variable which has the strongest connection with the dependent variable.

The presented framework for recursive partitioning of data to produce decision tree implies that certain methodological problems should be solved and that model parameters should be specified before or during the application of the given algorithm. These questions pertain to (Vercellis, 2009):

- Rules for heterogenous population splitting into smaller homogenous groups,
- Stopping criteria of recursive process, and
- Pruning of the tree.

There are numerous criteria for quantification of group homogeneity and for the selection of the best split of the data. When classification tree is concerned (if the dependent variable is categorical or if a numeric valuable has been transformed into a categorical), most popular criteria (measures) include: Gini coefficient, entropy measure, information gain and Chi-square criterion. The basic idea behind these criteria is to increase the homogeneity of the newly formed groups (nodes) regarding the categories of the target variable, in comparison to the group (node) which has been splitted.

If dependent variable is a quantitative one, and its values in the formed nodes are numerical, the common criteria for measuring the purity and splitting of the regression tree are the sum of sqared deviation from the node's average and *F*-test statistic. In fact, in the formation of a regression tree, the selection of independent variables and the division of nodes is carried out so as to minimise the variance within the group and maximise the variance between groups (Tufféry, 2011).

Introduction of certain limitations, in the form of criteria for stopping the process of tree growth, makes it possible to avoid the negative effects of the formation of the tree of great complexity. These criteria are a set of rules used during the development of a decision tree to determine whether it is necessary to

create more new branches and nodes, or to convert particular node into a leaf (Vercellis, 2009). They refer to the combination and fulfillment of the following general conditions for the termination of tree branching (Tufféry, 2011; Maimon & Rokach, 2010):

- Tree depth (the number of tree levels) has reached certain limit,
- Number of leaves (and rules, automatically) has reached certain maximum,
- Further splitting of any node will result in one or more new nodes whose number of observations is below the predefined minimal node size,
- The tree quality is adequate, and
- Futher division will not significantly improve the tree quality.

To solve the problem of overfitting in training sample, tree pruning method, which is based on elimination of statistically insignificant branches and redundant information, is used. There are two main ways of tree pruning (Gorunescu, 2011):

- *A priori* determination of tree growth (prepruning), which implies predefined rules for stopping further tree growth, and
- *A posteriori* reduction of the formed tree (postpruning), which means that, after a tree of maximum complexity has been produced, certain branches will be eliminated until the desired tree complexity is achieved.

The quality assessment of the results obtained by decision tree method is the issue closely related to the foregoing methodological aspects. Although the obtained models can be assessed using various criteria (predictive accuracy, speed, robustness, scalability and understandability), the parameter which is most frequently used for evaluation is the model accuracy. The model accuracy is defined as the proportion of correctly classified data using the designed model, contrary to the concept of error which indicates wrongly classified observations. A simple tool for the analysis of model performance based on the representation of the number of correctly and wrongly classified observations for each dependent variable category is the two dimensional square matrix, also called the classification matrix or the confusion matrix. Its elements represent testing results of the predictive model and they are used as the primary source of data for the calculation of model quality indicators.

In terms of model performance evaluation and its predictive capability testing it is very important to examine the issue regarding the way of data division into adequate subsets/subsamples for model training, validation and testing. There are several methods for splitting of the initial set of data in function of performance assesment of the classification model. Given that in the empirical part of this research, a cross-validation method is used, a brief overview of the characteristics

of this method follows. (For more details regarding model quality assesment see, eg. Vercellis, 2009; Witten & Frank, 2005).

In the cross validation method the original set of observed data is first randomly divided into *k* disjunctive partitions of approximately same size, and then, the evaluation process is conducted through *k* iterations, as follows: in each iteration, a single subset/partition is selected for testing while the union of other subsets (*k*-1) is used for model training. Training and testing are carried out the same number of times. In the end of the procedure, overall accuracy is calculated as arithmetic mean of *k* individual accuracies. In practice, larger number of iterations is usually preferred in order to obtain more robust assessment of classification model accuracy (Vercellis, 2009). A typical form of cross-validation (predominantly used in DM softwares) is the tenfold cross-validation, where the initial set is divided into 10 subsets. Generally, cross-validation method is demanding in terms of calculation, but can produce good results even with small number of observations, being based on their maximum exploitation.

Solutions to the previously mentioned methodological issues led to the development of numerous decision tree algorithms. Such multitude of algorithms is certainly caused by the method's popularity in research community and efforts to eliminate, through appropriate modifications and improvements, the shortcomings of existing algorithmic solutions. On the other hand, a practical attractiveness of the decision tree method is associated with a whole range of its positive characteristics, such as: conceptual simplicity, flexibility and ability to work with all types of variables, interpretability, calculation speed, handling missing data and robustness, in terms of extreme values. Despite the positive aspects, an analyst opting for this method must be aware of its drawbacks. Its main shortcoming is sensitivity to the change of input data used to train the model, so that small changes can result in completely different tree configurations.

Since in this study CHAID algorithm is applied, its fundamental determinations are presented below.

3.3. CHAID Algorithm

The CHAID algorithm, proposed by a statistician Kass in the late 1970s, is one of the most popular statistically based methods of supervised learning for decision tree development. Essentially, being one of multivariate dependency methods, CHAID algorithm is used for the detection of association between the categorical dependent variable and multiple independent variables which can be categorical and/or metric (in which case, their coding and transformation into categorical variables must be done previously).

The acronym CHAID denotes automatic and iterative procedure of tree development based on Pearson's Chi-square statistic and corresponding *p-value*.

Actually, as indicated by the algorithm's name, the basic criterion for recursive splitting of heterogenous population into homogenous groups according to the dependent variable categories (including the formation of independent variables categories and selection of statistically significant independent variables) is the Chi-square test statistic. Accordingly, the algorithm minimises variations of dependent variable within groups and maximises it between groups (Soldić-Aleksić, 2009).

Essentially, the CHAID method involves testing of hypotheses about the (in)dependence of two variables in each step of the algorithm's implementation. The logic of testing and formulating the conclusions is identical to the traditional procedure for statistical hypothesis testing, whereby, a software algorithm support enables rapid computation of multiple tests and easy (user-friendly) implementation of heuristic approach in finding the best partition of the observed data set.

Accordingly, the CHAID algorithm enables implementation of the following processes:

- The selection of the relevant independent variables from the set of input variables in such way that, in the resulting hierarchically arranged structure, as the first independent variable for the partition of input data is selected the variable with the lowest *p-value*, and is, therefore, most strongly associated with the dependent variable. In the procedure of hypothesis testing, if the *p-value* is equal to or lower than the predefined level of significance α, then the alternative hypothesis, which suggests a dependency between variables is accepted, which, in the context of tree development, denotes node splitting using a given independent variable. Otherwise, the node is considered to be the terminal node. Tree building ends when *p-values* of all the observed independent variables are higher than a certain split threshold.

- Merging the categories (values) of each independent variable so that a certain number of nodes, with statistically significant difference between them, appear on the tree. In fact, the algorithm identifies pairs of values of independent variables which are least different from the dependent variable, so that the number of categories of predictor variables depends on the Chi-square test results and *p-value*. If the obtained *p-value* is higher than a certain merge threshold, the algorithm merges particular categories with no statistically significant differences. After that, the search for a new merging pair continues until the pairs, for which the *p-value* is smaller than the defined level of significance α, are not identified.

Accordingly, it is possible to identify two key functions of statistical tests in CHAID analysis: ▸ combination of individual values and determination of predictor variable categories, and ▸ the selection of predictor variables according to the statistical significance of their association with the dependent variable (deVille, 2006). When non-binary predictor variables are concerned, the test value increases

along with the number of categories (branches) into which they are split. However, variables with more categories are more probable to be identified as statistically significant in relation to the dependent variable, compared to the independent variables with less categories. to the dependent variable. To eliminate this side effect caused by multiple testing in a single cycle of decision tree formation, Kass suggested using *Bonferroni adjustment* (Kass, 1980).

From the viewpoint of realisation of DM tasks, important feature of this algorithm refers to the fact that its application is extended to cases when the variables are numerical. If it comes to the classification problems, in determining the significance of the relationship and the best split for each tree building level, Chi-square test is used (including the likelihood ratio test if the variable is ordinal). For regression type of problems *F*-test is used as the criterion of numerical variables division (Rokach & Maimon, 2008).

Such applicability to both classification and regression problems is one of the key advantages of this algorithm. It is especially convenient for generation of non-binary trees, which makes this algorithm very popular in market research (Nisbet et al., 2009). On the other hand, one of the key disadvantages of the CHAID method is that it requires large amounts of data, because they are at every tree level split into several groups which may become too small for reliable analysis (Nisbet et al., 2009). In principle, due to the very nature of DM analysis, to apply the CHAID method it is necessary to have large samples. However, as a diagnostic approach, this method can be used with smaller samples as well.

4. Empirical Research

Respecting: (a) the role of leadership as an important factor of business success, (b) the results of traditional studies on the relationship between personal characteristics and style of leadership behaviour of managers, and (c) the fact that the sector of small and medium-sized enterprises[3] is one of the driving forces of economic development (RZS, 2015), especially in transition countries, empirical research in this paper, based on the implementation of CHAID algorithm and development of corresponding classification model is focused on the examination of dependence between the (selected) personal characteristics and dominant leadership style of managers of small private enterprises (including micro-enterprises and entrepreneurs) in Serbia.

[3] According to the average number of employees, as one of the criteria for categorization of enterprises, micro, small and medium enterprises differ, whereby micro enterprises (which also include entrepreneurs, as individuals who are self-employed) employ up to 9 small between 10 and 49, and medium between 50 and 249 persons.

4.1. Defining Problem Situation

It is common knowledge that modern business environment is characterised by complexity, uncertainty, heterogeneity and turbulence. In such circumstances, the need for finding the optimal combination of business activities in the function of achieving the company's objectives, as well as the fulfillment of specific requirements and (often conflicting) goals of different stakeholders, emphasizes the importance of managerial activities related to "the system of decision making, information and communication, and processes of management of human resources and employees' behaviour" (Petković et al., 2010). These activities pertain to leadership function of management. Consequently, leadership as a process of influencing and motivating the followers (individuals or groups, associates and other employees) to engage in achieving certain organisational objectives (Stojanović Aleksić, 2007) is a very important factor for business success. However, it certainly should not be overlooked that it is a two-way process of mutual influence between leaders and their followers.

Bearing in mind the aforementioned context, it is quite understandable the existence of large research interest for the phenomenon of leadership. Following aspects of leadership studies are particularly important for the purpose of this paper:

- Definition and analysis of characteristics of different categories of leadership styles[4], as manners (general patterns) in which leaders behave and establish relationships with their followers, and
- The analysis of influences of different factors on particular leadership style, among which most significant are: personal characteristics of leaders, followers, and the environment.

Based on numerous studies on these issues, it is possible to extract the following general findings, relevant from the point of view of the primary purpose of this paper: universal or the best style of leadership is impossible to define or clearly to identify in business practice. This is due to the complexity of human activities and human behaviour (Gonos & Gallo, 2013). A combination of

[4] In time, different categories of leadership style were identified. They are determined primarily by the following basic criteria: manager's approach to followers' motivation, manner in which decisions are made and whether employees participate in that process, sources of power used to influence the followers and managers' capacity to adjust to different situations (Petković et al., 2010). The theory differentiates between classical and modern leadership styles. According to Hawthorn and Iowa studies, the first science-based leadership studies classical leadership styles are: autocratic (or authoritarian), democratic (or participative) and *laissez-faire* (or liberal). (For details, see Stefanović & Stefanović, 2007; Petković et al., 2010; Stojanović Aleksić, 2007; Gonos & Gallo, 2013).

leadership styles adjusted to the circumstances is the best solution (Dulčić & Vrdoljak-Raguž, 2007).

In addition, research studies exploring the relationships between personal characteristics of leaders (belonging to different groups: personal characteristics of a general nature, economic, socio-psychological, etc.) and leadership styles report different findings ranging from statistically significant relationship in certain situations to extremely weak interdependence. In spite of the fact that some characteristics are relevant in some, but negligible in other situations, the search for common denominators and combination of a large number of characteristics that determine a leader's personality profile, always results in a certain knowledge that contribute to understanding and interpretation of the essence of the phenomenon of leadership.

4.2. Research Methodology

In accordance with the objective, and context of the defined problem situation, for the purpose of data collection, survey research was conducted, using a specially designed questionnaire, whose structure consists of the following parts:

- The first part contains questions about the enterprise in which the respondent is engaged.

- The second part includes questions related directly to the respondents, and are associated with their demographic and socio-psychological characteristics, private life and work experience. In designing of questionnaire and selection of the respondents' characteristics, those that often appear as the subject of consideration in the research studies of a similar nature, were taken into account. However, the biggest influence on the design of this part of the questionnaire, is associated with expert suggestions and proposals, which are obtained through the authors' personal communication with researchers in the field of leadership.

- The third part contains 18 statements related to the certain aspects of leadership behaviour, and served to identify each respondent's dominant leadership style according to the three categories identified in Iowa study: authoritarian, democratic and *laissez-faire*. For designing of this part of the questionnaire, Northouse's questionnaire model (Northouse, 2012) was used with certain adjustments, as a diagnostic model of leadership styles. To measure respondents' (dis)agreement with each of the statements a five-point Likert scale (1 – I strongly disagree, 2 – I disagree, 3 – I neither agree nor disagree, 4 – I agree, and 5 – I strongly agree) was used.

The questionnaires were distributed in June 2016 in two cities in Sumadija county. The respondents were selected among managers and owners of randomly chosen micro and small enterprises from official data bases of three accounting

agencies. The questionnaires were delivered in person or by e-mail with the cover letter which contained instructions, guaranteed anonymity and confidentiality of responses, and emphasised that the results will exclusively be used for scientific research.

By the defined deadline, a total of 417 valid questionnaires were returned to the authors of research[5]. The collected data were coded and analysed using the Statistical Package for the Social Sciences – SPSS, version 21.0.

Once the data was collected and the adequate base formed, the data from the first and second part of the questionnaire were assessed using descriptive statistical analysis tools. Reliability and internal consistency of items (statements) in third part is checked by determining the value of Cronbach's alpha coefficient and its comparison to a critical value of 0.7 (Hair et al., 2010). For the determination of the dominant leadership style of respondents, summarising respondents' answers by groups (subsets) of statements, that are defined by the applied Northouse's model, was performed. As each group of statements corresponds to one of the three leadership style categories, the highest sum indicates the dominant style. In addition, using the method of multivariate analysis – CHAID decision tree, the relation between selected characteristics of the respondents and their dominant leadership style is established and investigated. Finally, the discovered regularities were presented and interpreted, in an appropriate manner.

4.3. Empirical Results

In this section, the results of data analysis and application of CHAID algorithm, which are directly related to the defined hypotheses, will be presented, evaluated and discussed.

4.3.1. Sample Characteristics

Statistically (methodologically) speaking, the questionnaire items are variables in the process of modeling. Table 1 contains variables which were selected from the formed base and shows the structure of analysis sample of 417 respondents.

[5] The usual survey problems occurred once more: a large number of potential respondents refused to fill in the questionnaire. Nevertheless, the response rate was high when compared with the number of distributed questionnaires (although not in comparison to the number of contacts we had previously made). This was to be expected. For this reason, we distributed the questionnaires only after we had discussed the purpose of the research with each potential respondent and obtained his/her consent to participate. Thanks to the understanding and time spent to complete the questionnaire by 417 respondents, the completion of this study is provided.

This table also incorporates results of summed managers' attitudes regarding the 18 statements from the third part of the questionnaire. These statements, presented in a Likert scale, also get the status of variables, for which an acceptable level of internal consistency was confirmed (Cronbach's alpha being 0.71). Following Northouse's procedure, the respondents were classified into one of three categories of the derived variable *dominant leadership style*. Table 1 reveals that majority of the leaders in our study incline towards democratic style (55.2%). Autocratic style was identified in 30.5% of managers, while the smallest group (14.4%) contains managers whose dominant leadership behaviour belongs to *laissez-faire* style.

4.3.2. Application of CHAID analysis and its results

In order to build a decision tree using CHAID algorithm, according to its nature, firstly, a specification of the used variables, was carried out as follows:

- The variable, *dominant leadership style* is defined as a dependent variable. It is a nominal variable (with three values) so model creation can be based on Chi-square splitting criterion.

- Variables that are given in Table 1, and marked with the symbols from X_1 to X_{10} are used as independent variables. These are defined as nominal and ordinal variables. Some of them were originally numerical variables, but they are incorporated into the CHAID procedure, after their transformation into categorical variables was performed.

A summary of key elements of the application of CHAID algorithm is given in Table 2. It consists of two parts, where the first one refers to the specification of the elements and parameters necessary to run the algorithm, while the second presents part of the results of the algorithm's application.

In the context of applied statistical programme for the development of CHAID model (IBM, 2012), apart from the elements and criteria defined and displayed in the first part of Table 2, Pearson's Chi-square test is chosen, and a common level of significance ($\alpha=0.05$) for node splitting and independent variable categories merging is determined, with automatic adjustment of *p-values* (to solve the problem of incorrect rejection of null hypothesis in the context of multiple comparation). The presence of missing values was not detected in any of the variables. During the process of model construction ten-fold cross-validation model is applied. Figure 1 illustrates the resulting CHAID tree.

The results of CHAID procedure (presented in Table 2 and Figure 1) indicate that created model contains, within four levels of the tree depth, a total of 21 nodes, of which 12 are terminal. In addition, from a total of 10 initially specified independent variables, the final model includes 6, while the remaining 4 were not

statistically significant from the point of association with the dominant leadership style. In other words, based on the values of Chi-square criterion and corresponding *p-values*, the best subset of predictors was selected. In this way, by applying CHAID algorithm, a significant reduction of the model' dimensionality is achieved. The selected variables can be used for future research and development of new, parametric models. In the resulting model, apart from the reduction of the number of predictors, the reduction of their categories was also achieved. For example, variable X_5 in its original form has 5 categories (Table 1), and 2 categories in the model (Figure 1).

Table 1. Sample structure according to variables used in *CHAID* analysis

Variable names and symbols	*Values (modalities)*	*Code*	*Structure*		*Type of the variables*	
			f_i	%	MS	IDV/DV
Structure of employees by gender (X_1)	higher % of men	1	256	61.4	NV	IDV
	higher % of women	2	161	38.6		
Gender (X_2)	Male	1	280	67.1	NV	IDV
	Female	2	137	32.9		
Age (X_3)	less than 35 years	1	54	12.9	RV ↓ OV	IDV
	35 – 44	2	126	30.2		
	45 – 54	3	144	34.5		
	55 years and more	4	93	22.3		
Level of education (X_4)	M.Sc. & PhD	1	17	4.1	OV	IDV
	Faculty	2	170	40.8		
	Higher school	3	68	16.3		
	High school	4	162	38.8		
Marital status (X_5)	Single	1	61	14.6	NV	IDV
	Widowed	2	13	3.1		
	Married	3	306	73.4		
	Common law marriage	4	9	2.2		
	Divorced	5	28	6.7		
An only child (X_6)	Yes	1	109	26.1	NV	IDV
	No	2	308	73.9		
Childhood spent in? (X_7)	Village	1	129	30.9	NV	IDV
	City	2	288	69.1		
Popularity in childhood (X_8)	Yes	1	291	69.8	NV	IDV
	No	2	126	30.2		
Role model (X_9)	Yes	1	158	37.9	NV	IDV
	No	2	259	62.1		

Leadership work experience (X_{10})	less than 6 years	1	82	19.7	RV ↓ OV	IDV
	6 – 10	2	104	24.9		
	11 – 20	3	142	34.1		
	21 years and more	4	89	21.3		
Dominant leadership style (Y)	Authoritarian	A	127	30.5	NV	DV
	Democratic	D	230	55.2		
	Laissez-faire	L	60	14.4		

Note: Acronyms and symbols used in the table have the following meanings: MS = Measurement Scales; NV = Nominal Variable; OV = Ordinal Variable; RV = Racio Variable; RV→OV = transformed numerical (racio) variable into categorical variable (ordinal); f_i = frequency; IDV = Independent Variable; DV = Dependent Variable.

Source: Authors' calculations

Table 2. Model summary

Specifications	Growing method	CHAID
	Dependent variable	Y
	Independent variables	X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}
	Validation	cross-validation
	Maximum tree depth	4
	Minimum cases in parent node	40
	Minimum cases in child node	15
Results	Independent variables included	X_5, X_7, X_6, X_{10}, X_2, X_4
	Number of nodes	21
	Number of terminal nodes	12
	Depth	4

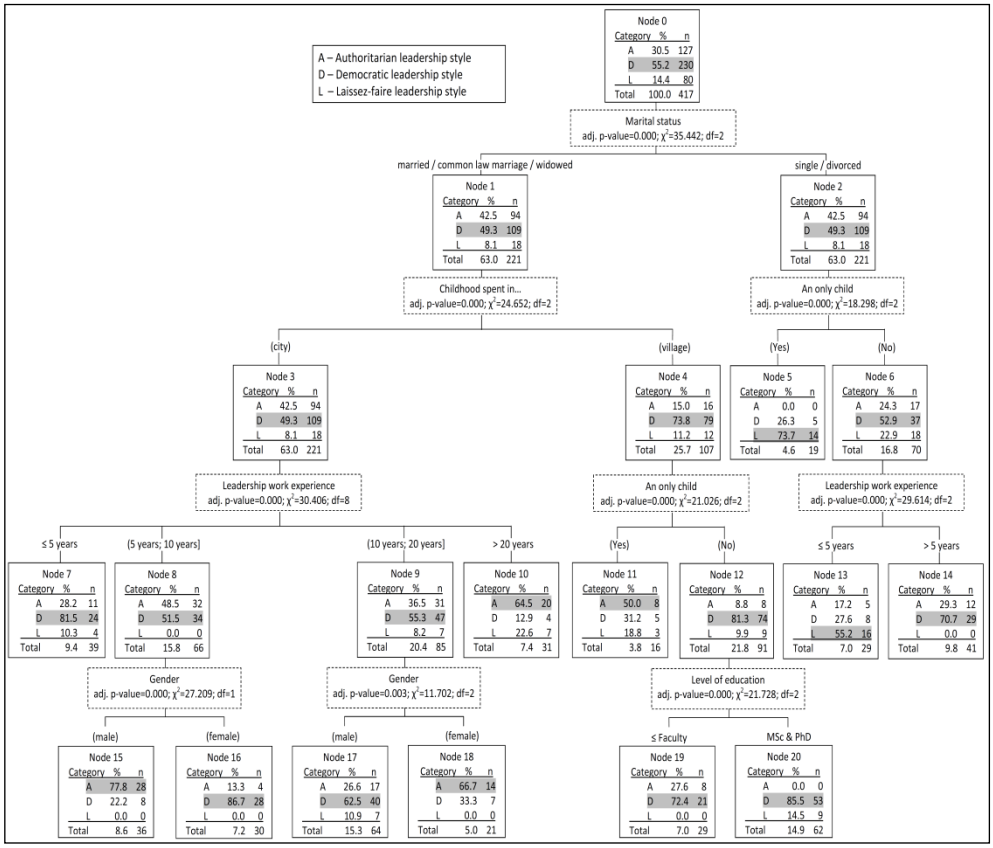
Source: Authors' calculations

4.3.3. Analysis of Modeling Results

As can be seen in Figure 1, the most significant independent variable is *manager's marital status* X_5, which means that it is most strongly associated with the dependent variable and has most power in division of observations into groups. In other words, for the observed data this variable has the biggest potential to differentiate and classify managers into three groups according to the leadership style. (Statistical significance of X_5 was determined, with $\alpha=0.05$ using following values: $\chi^2=35.442$, $df=2$, $p\text{-value}=0.000$.) As the first discriminator, it splits the root node, i.e. the sample of 417 respondents, into two groups containing different categories of X_5 presented as node 1 and node 2.

The majority of respondents (328) belong to node 1 which groups values of X_5 coded as 2/3/4. The rest of respondents belong to node 2 containing X_5 values coded as 1/5. Both groups are dominated by managers with democratic style. As for the percentage distribution of other two categories of leadership style, the first group contains significantly higher proportion of managers with autocratic, and the second with *laissez-faire* style. For X_5, within the first level of the tree, nodes 1 and 2 are parent nodes.

Figure 1. CHAID decision tree



Source: Authors' calculations

Within the second level of the tree, two statistically significant variables are identified: *childhood spent in?* (X_7) and *an only child* (X_6). Variable X_7 is significant for splitting of node 1 (χ^2*=24.652, df=2, p-value=0.000*). According to this its categories following two groups of respondents are obtained: managers who spent their childhood in urban areas belong to node 3, while those from rural areas belong to node 4. The split of node 2 is based on X_6 (χ^2*=18.298, df=2, p-*

value=0.000), producing following two groups: node 5 and node 6, which contain managers who did and did not grow up as an only child, respectively. Regarding this tree level, for X_7 nodes 3 and 4 are parent nodes, and for X_6 node 5 is terminal, while node 6 is parental.

In the third level of the tree depth, variable that is statistically significant for splitting the nodes 3 and 6 is *leadership work experience*, X_{10}. For the managers distributed within node 3, four categories of this variable were formed, and represented with nodes 7, 8, 9 and 10 ($\chi^2=30.406$, *df=6, p-value=0.000*). In addition, for the managers in node 6, a significant reduction of modalities was performed and only two categories of the same variable were formed as nodes 13 and 14 ($\chi^2=29.614$, *df=2, p-value=0.000*). In this tree level, X_6 appears again as a statistically significant predictor for node 4, forming thus nodes 11 and 12 ($\chi^2=21.026$, *df=2, p-value=0.000*). When variable X_{10} is concerned, nodes 8 and 9 are parental, while nodes 7 and 10 are terminal, together with nodes 13 and 14. As for the X_6, node 11 is terminal, while node 12 is parental.

Within the fourth, final level of the tree, two variables are identified as statistically significant: *gender*, X_2, and *level of education*, X_4. Variable X_2 (with two categories) is significant for splitting of nodes 8 and 9 ($\chi^2=27.209$, *df=1, p-value=0.000*, and $\chi^2=11.702$, *df=2, p-value=0.003*, respectively). In this way, nodes 15, 16 and 17, 18 are formed. It is interesting to point out that, in the empirical research conducted by Stojanović Aleksić et al. (2016), based on a sample of 79 randomly selected leaders in business organisations and institutions in Serbia, using the Chi-square test of independence, a statistically significant relationship between leaders' gender and dominant leadership style was also identified. The split of node 12 is based on X_4 ($\chi^2=21.728$, *df=2, p-value=0.000*) resulting in nodes 19 and 20, which denote integrated categories of X_4, coded as 1/2, i.e. 3/4, respectively. All the nodes in the final level of decision tree are terminal.

Generally, of the total number of terminal nodes in the formed tree structure, four (marked as 10, 11, 15 and 18) pertain to autocratic, six (marked as 7, 14, 16, 17, 19 and 20) to democratic, while two (marked as 5 and 13) refer to *laissez-faire* style. In fact, the paths from the root to the terminal nodes generate a set of rules for classification of managers into one of the defined categories of the variable *dominant leadership style.* This clearly indicates that the developed model and knowledge depicted in the decision tree can be formulated as *if-then* rules. For example, the rule in node 16 can be interpreted as follows: if a manager is widowed, married, or in a common-law marriage, from urban environment, has held leadership position from 6 to 10 years and is female then, we can state with 0.867 probability, that dominant leadership style of that particular manager is democratic style. The other derived rules can be interpreted in a similar manner.

4.3.4. Accuracy assessment of the classification model

The process of classification is not completely finished until all its performance have been assessed. Tables 3 and 4 present basic information about the performance of the developed CHAID model in terms of its accuracy and predictive potential.

Table 3. Risk

Estimate	
Re-substitution	Cross-validation
0.297	0.365

Source: Authors' calculations

Table 4. Classification matrix

Dominant leadership style		Predicted			*% of correctly classified*
		A	D	L	
Observed	A	70	52	5	55.1%
	D	24	193	13	83.9%
	L	10	20	30	50.0%
	Overall %	24.9%	63.5%	11.5%	70.3%

Source: Authors' calculations

Table 3 presents prediction risk as a percentage of inaccurately classified observations. To be precise, our findings suggest that, if the characteristics of a manager in terms of the six independent variables are known, the risk that the manager will be inaccurately classified in terms of dominant leadership style (based on entire sample) is 29.7%, while that risk, when a test sample is used in model cross-validation, is 36.5%. Table 4 presents classification matrix containing, by categories of the dependent variable, empirical and modeled values, i.e. actual (observed) and predicted classifications. In accordance with the foregoing, it can be stated that overall accuracy of the model is 70.3%. In other words, the model has accurately classified 293 (main diagonal of the matrix) out of 417 managers in the observed sample. Observed by the categories of the dependent variable, significant differences in classification accuracy can be seen. However, the percentage structure of modeled (predicted) values according to the categories of the dependent variable (24.9% : 63.5% : 11.5%) is not significantly different from that of the original data (30.5% : 55.2% : 14.4%), and it faithfully represents the key relations among them.

In general, the best measure of the model's performance, is not its raw accuracy, but its usefulness and effectiveness in achieving the primary purpose for which it was formed in domain of solving a specific problem.

5. Conclusion

In this paper, the possibilities of simultaneous observation of multiple variables, i.e. (selected) personal characteristics of managers in function of creating a classification model for the identification of their dominant style of leadership behaviour, were examined and evaluated. As a methodological basis for creating a model, CHAID decision tree method is used, whose applicability has been verified in numerous problem areas and situations. In fact, on a concrete empirical example, the domain of application of this algorithm is expanded and demonstrated its usefulness in the new problem context.

The analysis of dependency between selected personal characteristics and leadership styles, conducted on a sample of 417 managers of small privately owned Serbian enterprises, extracted a set of six statistically significant predictors of a dominant leadership style (in the hierarchical structure, the variable *marital status* was in the strongest interaction with the dependent variable – *dominant leadership style*), and as a result of their combination, 12 classification rules were generated. From the aspect of analyeed sample, identified regularities are reliable, because they represent the result of methodologically adequate application of CHAID algorithm.

However, as there are numerous controversies in the theory of leadership and management pertaining to the factor that dominantly determines managers' leadership behaviour, the study's findings should not be generalized *a priori*. The relevance of the selected variables was determined for the observed sample, which does not rule out the possibility that the analysis of a different sample might yield a completely different combination of statistically significant variables. For this reason, it is necessary to continually observe a larger number of samples to achieve and confirm stability of presented findings. Although in this paper, a quantitative approach to the defined problem was emphasised, it is also extremely important that the results of the modeling are discussed from the socio-psychological point of view by experts from the theory of leadership, organisational behaviour, sociology and psychology.

This research has certain limitations which pertain to sample size, due to the dominant share of the small business sector in the structure of the Serbian economy. Apart from usage of a larger sample, future studies should include more input variables in order to improve model performance and help examine the research question from different perspectives. Generally speaking, the obtained model can serve as a basis for further studies regarding the investigation of

relations between economic results, employees' satisfaction and leadership styles, using other classifications of leadership styles or for comparative analysis with large enterprises.

References

Baran, B. & Kılıç, E. (2015). Applying the CHAID algorithm to analyze how achievement is influenced by university students' demographics, study habits, and technology familiarity. *Educational Technology & Society*, 18 (2), 323–335.

de Ville, B. (2006). *Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner*. Cary, NC: SAS Institute Inc.

Díaz-Pérez, M. F. & Bethencourt-Cejas, M. (2016). CHAID algorithm as an appropriate analytical method for tourism market segmentation. *Journal of Destination Marketing & Management*, 5 (3), 275-282, doi: 10.1016/j.jdmm.2016.01.006.

Dulčić, Ž. & Vrdoljak-Raguž, I. (2007). Stilovi vođstva hotelskih menadžera Dubrovačko-neretvanske županije–empirijsko istraživanje. *Ekonomski pregled*, 58 (11), 709-731.

Gonos, J. & Gallo, P. (2013). Model for leadership style evaluation. *Management*, 18(2), 157-168.

Gorunescu, F. (2011). *Data Mining: Concepts, Models and Techniques*. Berlin: Springer.

Hair, J.F.Jr., Black, W., Babin, B. & Anderson, R. (2010). *Multivariate data analysis* (8th ed). Upper Saddle River (New Jersey): Pearson Prentice Hall.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques*, (3rd ed). Amsterdam (etc.): Elsevier Inc.

Horvat, I., Pejić Bach, M. & Merkač Skok, M. (2014), Decision tree approach to discovering fraud in leasing agreements. *Business Systems Research*, 5(2), 61-71, doi: 10.2478/bsrj-2014-0010.

IBM (2012). *IBM SPSS Decision Trees 21*. Retrieved from: http://www.sussex.ac.uk/its/pdfs/SPSS_Decision_Trees_21.pdf Accessed on: September, 2016.

Kağnicioğlu, H. C. & Moğol, M. (2014). Implementation of CHAID algorithm: a hotel case. *International Journal of Research in Business and Social Science*, 3(4), 42-52, doi: 10.20525/ijrbs.v3i4.116.

Kantardzic, M. (2011). *Data mining: concepts, models, methods, and algorithms* (2nd ed). Hoboken, New Jersey: John Wiley & Sons.

Kass, V. G. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical society*, 29 (2), 119-127.

Kim, S.S., Timothy, J. D. & Hwang, J. (2011). Understanding Japanese tourists' shopping preferences using the Decision Tree Analysis method. *Tourism Management*, 32(3), 544-554, doi: 10.1016/j.tourman.2010.04.008.

Maimon, O. & Rokach, L. (Eds.) (2010). *Data mining and knowledge discovery handbook* (2nd ed). New York: Springer.

Nisbet, R., Elder, J. & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Amsterdam (etc.): Elsevier Inc.

Northouse, P.G. (2012). *Introduction to leadership: Concepts and Practice* (2nd ed). Los Angeles (etc.): SAGE Publ.Inc.

Novotná, M. (2012). The use of different approaches for credit rating prediction and their comparison. In: Proceedings of the 6th International Conference on *Managing and Modelling of Financial Risks* (pp. 448-457).

[Availabe at SSRN: https://ssrn.com/abstract=2867. Accessed August, 23,2016.]

Öcal, N., Ercan, K. M. & Kadıoğlu, E. (2015). Predicting financial failure using decision tree algorithms: an empirical test on the manufacturing industry at Borsa Istanbul. *International Journal of Economics and Finance*, 7(7), 189-206, doi:10.5539/ijef.v7n7p189.

Republički zavod za statistiku (RZS), (2015). *Preduzeća u Republici Srbiji prema veličini, u 2014. godini* (Radni dokument, 90). Beograd: Republički zavod za statistiku.

Petković, M., Janićijević N., Bogićević Milikić B. (2010). *Organizacija* (8th ed). Beograd: Ekonomski fakultet Univerziteta u Beogradu.

Popescu, M. E., Andreica, M. & Micu, D. (2014). A method to improve economic performance evaluation using classification tree models. *European Journal of Business and Social Sciences*, 3 (4), 249-256.

Rokach, L. & Maimon, O. (2008). *Data mining with decision trees:theory and applications*. New Jersey (etc.): World Scientific.

Shmueli, G., Patel, N.R. & Bruce, P.C. (2010). *Data mining for business intelligence concepts, techniques and applications in Microsoft Office Excel with Xlminer* (2nd ed). Hoboken, New Jersey: John Wiley & Sons.

Soldić-Aleksić, J. (2009). Prediktivni model segmentacije tržišta: primena modela logističke regresije i CHAID procedure. *Marketing*, 40 (3), 129-138.

Stefanović, N. & Stefanović, Ž. (2007). *Liderstvo i kvalitet*. Kragujevac: Univerzitet u Kragujevcu, Mašinski fakultet.

Stojanović-Aleksić, V. (2007). *Liderstvo i organizacione promene*. Kragujevac: Univerzitet u Kragujevcu, Ekonomski fakultet.

Stojanović Aleksić, V., Stamenković, M. & Milanović, M. (2016). Analiza liderskih stilova u organizacijama u Srbiji: uticaj pola. *Teme*, XL(4), 1383-1397.

Tufféry, S. (2011). *Data mining and statistics for decision making*. Chichester: John Wiley & Sons.

Vercellis, C. (2009). *Business intelligence: data mining and optimization for decision making*. Chichester: John Wiley & Sons.

Witten, H. I. & Frank E. (2005). *Data mining: practical machine learning tools and techniques* (2nd ed). Amsterdam (etc.): Elsevier Inc.

CHAID STABLO ODLUČIVANJA: METODOLOŠKI OKVIR I PRIMENA

Apstrakt: Tehnološki napredak u svim sferama ljudskog delovanja uzrokovao je ubrzano generisanje ogromnih količina podataka. Konsekventno, pred istraživačima se pojavio problem identifikovanja odgovarajućih načina za pretvaranje lako dostupnih velikih količina podataka u korisno znanje. U pravcu prilagođavanja analize podataka nastalim promenama, razvijen je data mining pristup analize podataka iz različitih perspektiva i otkrivanja signifikantnih zakonitosti duboko skrivenih u njima. Polazeći od navedenog, u radu se prezentuju koncepcijska određenja stabla odlučivanja, kao značajnog

data mining metoda, koji je, shodno svojoj eksplorativnoj prirodi, izuzetno pogodan za otkrivanje struktura podataka pri razmatranju različitih problemskih situacija. U empirijskom delu rada demonstrirana su aplikativna svojstva ovog metoda na osnovu CHAID algoritma u proučavanju fenomena liderstva sa aspekta ispitivanja međuzavisnosti odabranih personalnih karakteristika i liderskog stila menadžera. Osnovni cilj rada je razvoj klasifikacionog modela za identifikovanje dominantnog stila liderstva. Istraživanje je sprovedeno na uzorku od 417 menadžera malih srpskih preduzeća u privatnom vlasništvu, korišćenjem posebno dizajniranog upitnika za te svrhe. Klasifikacioni model rezultirao je skupom od šest statistički značajnih personalnih karakteristika kao prediktora dominantnog liderskog stila.

Ključne reči: *data mining*, metod stabla odlučivanja, *CHAID* algoritam, liderski stilovi, personalne karakteristike lidera, klasifikaciona pravila

Authors' biographies

Marina Milanović graduated from and defended her Master's thesis at the Faculty of Economics, University of Kragujevac, Serbia. At the same Faculty, she was engaged as a teaching and research assistant, within the Department of Statistics and Informatics. She is currently working on Ph.D. thesis at the Faculty of Economics, University of Niš, Serbia. She is the author and coauthor of numerous scientific and research papers in various kinds of publication. Fileds of her research interests are: statistical data analysis in economy, data mining, and statistical process control.

Milan Stamenković is a teaching and research assistant at the Department of Statistics and Informatics, Faculty of Economics, University of Kragujevac, Serbia. He is engaged in teaching the subject Fundamentals of Statistics. He is currently working on Ph.D. thesis at the same Faculty. He is the author and co-author of numerous scientific and research papers in various kinds of publication. Fileds of his research interests are: statistical data analysis, application of multivariate statistical methods in economy and demographic analysis.