UDC 681.3 : 004.6

# PROBLEM OF DATA ANALYSIS AND FORECASTING USING DECISION TREES METHOD

*T.I. Lytvynenko*

Розглянуто застосування методу дерев рішень до проблеми аналізу даних та прогнозування. Обробка даних базується на реальних спостереженнях рівня продажу груп товарів протягом 2006-2009 рр. R (мова програмування та програмне середовище) застосовано як інструмент для статистичних обчислень. Робота містить порівняння з загальновідомими підходами та можливі шляхи покращення точності отриманих висновків.
Ключові слова: інтелектуальний аналіз даних, прогнозування, прийняття рішень, дерева рішень, мова R.

Рассмотрено применение метода деревья решений к проблеме анализа данных и прогнозирования. Обработка данных базируется на реальных наблюдениях уровня продаж групп товаров на протяжении 2006-2009 гг. R (язык программирования и программная среда) используется в качестве инструмента для статистических вычислений. Работа содержит сравнение с общеизвестными подходами и возможные пути повышения точности полученных результатов.
Ключевые слова: интеллектуальный анализ данных, прогнозирование, принятие решений, деревья решений, язык R.

This study describes an application of the decision tree approach to the problem of data analysis and forecasting. Data processing bases on the real observations that represent sales level in the period between 2006 and 2009. R (programming language and software environment) is used as a tool for statistical computing. Paper includes comparison of the method with well-known approaches and solutions in order to improve accuracy of the gained consequences.
Key words: data mining, forecasting, decision making, decision trees, R language.

## Introduction

This paper is the logical continuation of the publication [1] and represents detailed explanation and practical implementation of the method mentioned in [1] – decision trees. As is stated in [2], all predictive techniques may be divided into two principal groups: statistical and structural methods correspondingly. Decision trees belong to the classical structural approaches. According to [3], a decision tree describes the process graphically and simplifies a major goal of the analysis – to determine the best decisions. In [1] we analyzed approaches to solve the task of sales forecasting. The actual task is to estimate the efficiency of marketing campaigns with the goal to advertise and deliver (or push) the product to the market. These activities are the means to increase the sales. So, if we could predict the sales of some product in advance and then compare it with the actual data, we can fix the profitability of marketing campaigns conducted.

## Problem formalization

As is stated in [1], the predictive methods make forecast based on the statistical relationships between input columns in a dataset. The major idea is to interpret current data in a proper way in order to obtain the objective laws. Input data is represented by a dataset. Three basic categories are considered to be processed: product name, sales and period under consideration. The predictive approaches are applied to a dataset to obtain a probable prognosis for the future period. A decision tree is a drawing [4], consisting of lines and boxes, that shows the different choices which are available to people before they make a decision, and the possible results of these choices. Decision trees [5] are produced by algorithms that identify various ways of splitting a data set into branch-like segments. These segments form an inverted decision tree that originates with a root node at the top of the tree. Graphically a decision tree looks like a flowchart-like structure in which each internal node represents a "test" on an attribute [6]. Commonly a decision tree consists of three types of nodes: decision nodes, chance nodes and end nodes. Classification and regression trees were originally introduces and investigated by Breiman [7] in 1984. As asserted in [8], the main idea behind tree methods is to recursively partition the data into smaller and smaller strata in order to improve the fit as best as possible. Tree models where the target variable can take a finite set of values are called classification trees. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [9]. Classification trees [10] are designed for dependent variables that take a finite number of unordered values, with prediction error measured in terms of misclassification cost. Regression trees are for dependent variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between the observed and predicted values [1]. In this study input data is represented by the table including such categories as product name (or number, to facilitate), month (and year) and sales level (the number of sold units). One of the goals of this research is to investigate seasonal dependencies, that is why primary attention is paid to months, though year is also significant. The period under consideration is from January, 2006, to August, 2015. Statistical dataset may include products, which were not sold at all during the whole period. It is worth noting, that results, based on the real observations, may differ from the ones, obtained by the usage of pseudorandom number generators. The presence of the "zero sales level products", mentioned above, is one of the most substantial factors, which presupposes this phenomenon.

## Existing approaches review

The problem of data analysis and prediction is deeply investigated; nevertheless, a lot of questions are still open. Numerous methods are listed in [2], there they are divided into two principal groups: statistical and structural methods correspondingly.

The first group includes predictive models, which leverages statistics to predict outcomes [1, 11] – ultimate model is represented by analytical formula. The second group consists of step-by-step methods, where terminate algorithm looks like a definite set of stages. Widely used up-to-date statistical approaches are represented by linear regression, exponential smoothing, ARIMA (autoregressive integrated moving average), GARCH ( generalized autoregressive conditional heteroskedasticity). Concerning commercial packages, primary attention in this field is paid to ARIMA, that implemented in such software solutions as SAS, SPSS, Mathematica, Matlab, Microsoft SQL Server (Time Series Algorithm).

The second group includes neural networks, Markov chains method and CART [2]. Generally, their advantages and disadvantages are summarized in pic. 1.

| Class of methods | + | - |
|---|---|---|
| Statistical | • Simplicity of implementation | • Difficult to define the dependency between variables<br>• Difficult to define unknown parameters |
| Structural | • Flexibility<br>• Parallel computing<br>• Combination of categorical and numerical | • Strict dataset prerequisites<br>• Difficult learning algorithm choice<br>• Stopping criteria |

Picture 1. Comparison of statistic al and structural methods

In [1] CART approach was chosen for the further investigations. In addition to R programming, IBM SPSS Statistics was applied to the input data. This software package (originally Statistical Package for the Social Sciences by SPSS Inc.) is adapted to health sciences and marketing, so, according to the character of input data, is appropriate to predictive data mining techniques. SPSS is beneficial for managerial decision-making process due to its business-oriented construction.

## Decision tree algorithm

In CART (classification and regression tree) data are handled in their raw form; no binning is required or recommended [12]. CART splitting rules [12] are represented by the following construction: an instance goes left if CONDITION, otherwise goes right.

General algorithm includes the following steps:

1) Start at the root node;
2) To each element, apply a condition split;
3) If a stopping criterion is reached, exit. Otherwise, apply step 2 to each child node in turn.

CART is chosen due to its obvious advantages [1]. The model is having a single meaning; the process is well-defined and stepwise. This approach is clear for both understanding and implementation. Decision trees are simple to understand and interpret. In [13] it is noticed, that decision trees require relatively little effort from users for data preparation. Moreover, it is worth noting, this approach is also able to cope with different types of data. According to our problem formalization, observations combine categorical (months) and numerical (sales level) data. As is stated in [12], CART belongs to the most substantial data mining algorithms. Tree models where the target variable can take a finite set of values are called classification trees [8]. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees [9]. According to [14], the basic algorithm for decision tree is the greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. It takes a subset of data as input and evaluate all possible splits. The best split decision is chosen to partition the data in two subsets and the method is called recursively. The best split decision is traditionally the split with the highest information gain. He algorithm stops when the stop conditions are met [14]. As is in [15] stated, in machine learning and information theory, information gain is a synonym for Kullback-Leibler divergence, though in the context of decision trees, the term is sometimes used synonymously with mutual

information, which is the expectation value of the Kullback-Leibler divergence of a conditional probability distribution. Detailed information concerning information gain`s computation is provided by [14]. The number of stopping conditions is listed in [14]:

- All the samples belong to the same class, i.e. have the same label since the sample is already "pure";
- Stop if most of the points are already of the same class. This is a generalization of the first approach, with some error threshold;
- There are no remaining attributes on which the samples may be further partitioned;
- There are no samples for the branch test attribute.

Building of the both types is covered by the usage of programming language and environment R. In this paper open-source user interface RStudio is applied to the input data.

## Classification algorithm in R

R language provides a great range of packages for data processing. Two of them – ***rpart*** and ***party*** are the most appropriate for the given problem. ***Rpart*** (recursive partitioning and regression trees) is used for classification by decision trees and generation of regression trees [14]. According to [14], the ***rpart*** programs build classification or regression models of a very general structure using a two stage procedure; the resulting models can be represented as binary trees. The tree is built by the following process: first the single variable is found which best splits the data into two groups ('best' is defined in [14]). The data is separated, and then this process is applied separately to each sub-group, and so on recursively until the subgroups either reach a minimum size (5 for this data) or until no improvement can be made. The resultant model is, with certainty, too complex, and the question arises as it does with all stepwise procedures of when to stop. The second stage of the procedure consists of using cross-validation to trim back the full tree. To grow a tree, we use [14] command ***rpart (formula, data=, method=, control=)***.

## Practical implementation and analysis of the outcome

Firstly, the data concerning just the one position to be processed. It is represented by the following dataset (fragmented in the pic. 2). ***ProductN*** is the number of a sales position. ***Period*** is the month under consideration. In this particular case dataset covers the period between the years 2006 and 2009. ***Sales*** is the number of units, sold during the corresponding month.

Test1.R × data ×

Filter

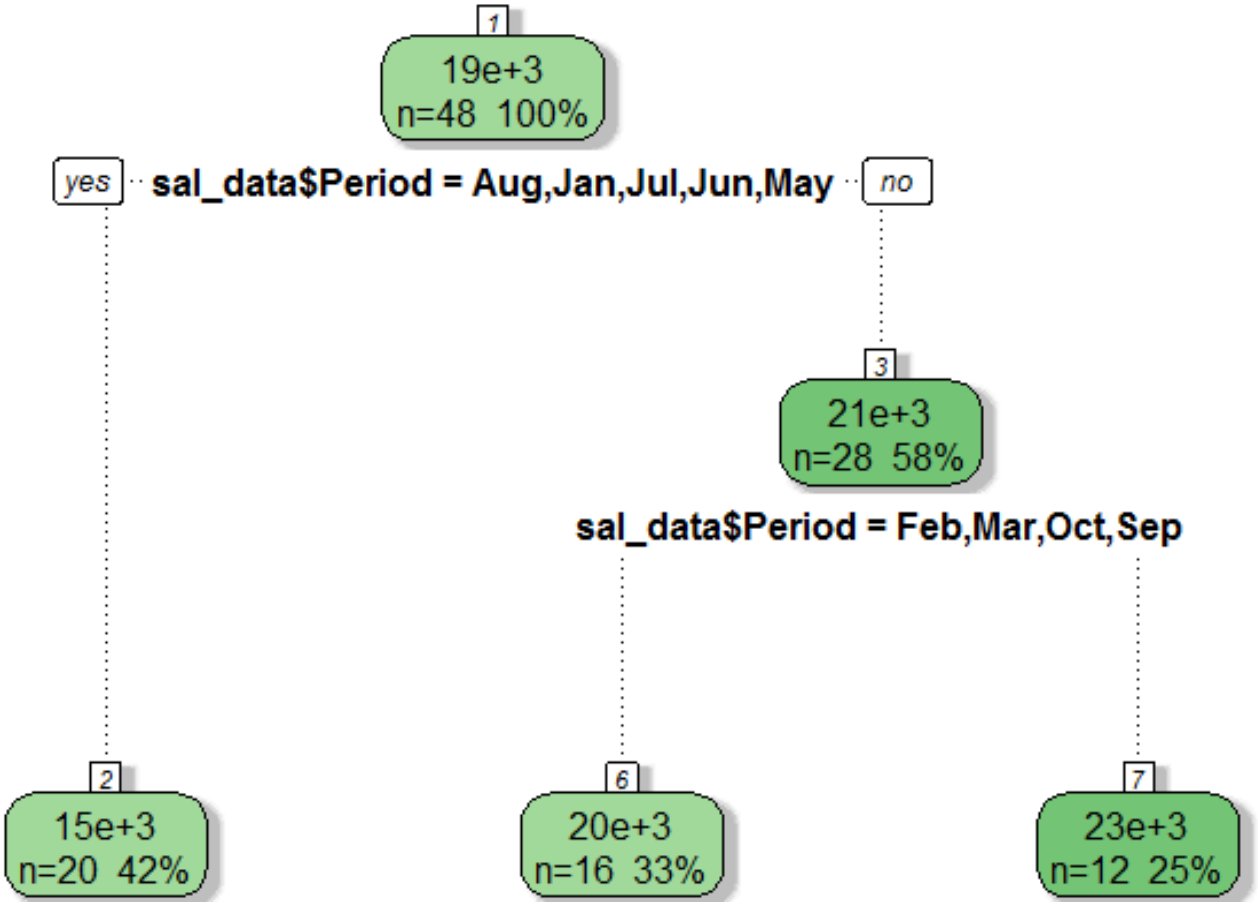
| | ProductN | Period | Sales |
|---|---|---|---|
| 1 | 1 | Jan | 14285 |
| 2 | 1 | Feb | 17282 |
| 3 | 1 | Mar | 18883 |
| 4 | 1 | Apr | 18771 |
| 5 | 1 | May | 12197 |
| 6 | 1 | Jun | 10694 |
| 7 | 1 | Jul | 12783 |
| 8 | 1 | Aug | 15054 |
| 9 | 1 | Sep | 13904 |
| 10 | 1 | Oct | 21427 |
| 11 | 1 | Nov | 19139 |
| 12 | 1 | Dec | 15164 |
| 13 | 1 | Jan | 18768 |
| 14 | 1 | Feb | 18452 |
| 15 | 1 | Mar | 14037 |
| 16 | 1 | Apr | 19859 |
| 17 | 1 | May | 16573 |
| 18 | 1 | Jun | 9810 |
| 19 | 1 | Jul | 13459 |
| 20 | 1 | Aug | 14840 |
| 21 | 1 | Sep | 16512 |
| 22 | 1 | Oct | 18884 |

Picture 2. Input dataset example

The following R libraries are used:

- *library(rpart)*
- *library(rpart.plot)*
- *library(RColorBrewer)*
- *library(rattle)*

Output is the following (pic. 3):




Picture 3. Outcome binary tree

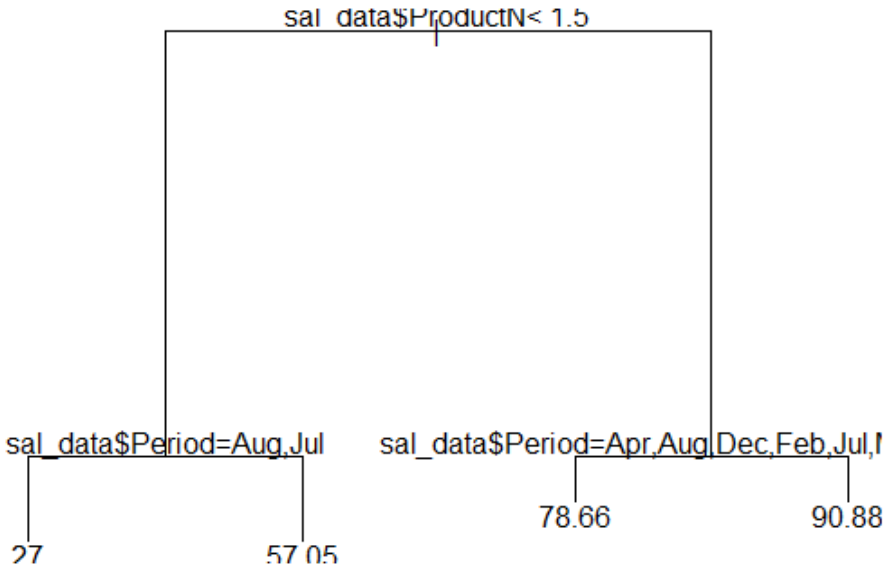
Classification procedure creates the binary tree of possible outcomes, the figures in the nodes represent the number of elements in a dataset, which associate with with alternative, and the percentage of the corresponding observations. According to this tree, the structure of ***ProductN*** sales is (pic. 4):

| Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | |

Picture 4. Seasonal table of the sales level

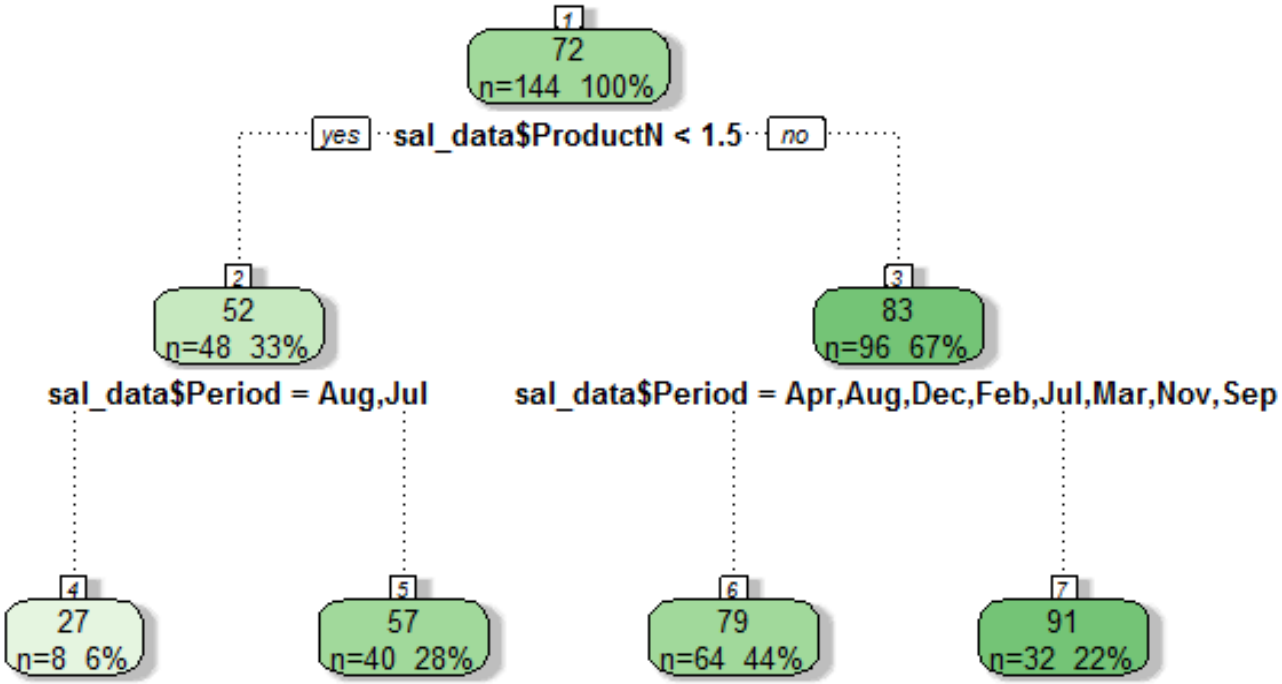
Tree states, that April, November and December belong to the node with the highest sales level (approximately 23000 units per month), though during 2006 their sales made up 18771, 19139 and 15164 respectively. That is why it is

worth noting, that decision trees method takes into account not only figures, but also tendencies. Then, application of the approach to the several product positions is leading to the following outcomes (pic. 5):




Picture 5. Binary tree outcome for a multi-dimensional case

Pic. 6 illustrates the general structure of a decision tree and Picture R is a result of user-oriented transformation.




Picture 6. Outcome decision tree

It is evident from the binary tree, that computing accuracy is substantially worse than in the previous case. Discrepancy between price levels of different products presupposes this phenomenon. Experiments demonstrate that the same situation takes place in case of far greater number of products. Accuracy falls exponentially with the rise of dimension.

That is why general results of the experimental modelling are the following. Major advantages of the method are observed: it simplifies decision-making process because outcome is represented by the binary tree – clear and easy to interpret. On the contrary, simplicity influences the quality, because not all data types may be represented efficiently as a binary structure. CART uses classification and regression principles as a basis (as many software solutions, particularly Microsoft SQL Server Time Series). According to the experimental data, the common drawback of these approaches is their vulnerability in case of existence of abnormal external factors (e. g. World Economic Crisis 2008 for the the data used above). Binary structure is unable to differ mathematically abnormal periods (which may be detected through statistical tests) from practically abnormal ones (caused by external factors). In addition to this, decision tree approach is much more efficient in case of "one-product-position" computations because two branches split may be not structural enough to classify several products, those price varies substantially. Analyzing just one product position prices, decision tree approach achieves the accuracy of a traditional regression method, being simultaneously easier to use and interpret (especially for non-professionals). In future paper I am intended to apply SPSS software package to the same datasets in order to compare results and accuracy.

## Consequences

This work provides the description, explanation and result of the decision tree method and its application to the real observations, using R programming language. It is compared with traditional approaches in order to define common features and drawbacks. Effectiveness of the method checked experimentally. For all methods, both traditional and prospective, computing accuracy is highly dependable on complexity, structure and quality of an input dataset. The most significant disadvantage is vulnerability when it comes to external factors influence. It may be resolved by combining two or more methods and using of machine learning techniques. These alternatives will be presented and practically implemented in the future papers.


1. *Lytvynenko T.I., Panchenko T.V., Redko V.D.* Pre-print: Sales Forecasting using Data Mining Methods. Bulletin of Taras Shevchenko National University of Kyiv, Series Physics & Mathematics.
2. *Чугуева И.А.* Модель прогрозирования временних рядов по выборке максимального подобия. Дис. … канд. техн. наук. – М.; 2012. – 153 с.
3. *TreePlan Software*. Introduction to Decision Trees. Available from: <http://treeplan.com/ chapters/introduction-to-decision-trees.pdf>.
4. *Cambridge* Dictionaries Online: English Dictionary. Available from: <http://dictionary. cambridge.org/dictionary/english/decision-tree>.
5. *SAS Institute*. Decision Trees – What Are They? Available from: <http://support.sas.com/ publishing/pubcat/chaps/57587.pdf>.
6. *Pandey A.K., Goyal N.K.* Early Software Reliability Prediction: A Fuzzy Logic Approach. – New Dehli: Springer India, 2013. – 153 p.
7. *Breiman L., Friedman J., Stone C.J., Olshen R.A.* Classification and Regression Trees / – Chapman and Hall/CRC, 1984. – 368 p.
8. *Gordon L*. Using Classification and Regression Trees (CART) in SAS Enterprise Miner For Applications in Public Health // SAS Global Forum 2013. Paper 089-2013. – 2013. – 8 p.
9. *Jopp F., Reuter H., Breckling B. (Eds.).* Modelling Complex Ecological Dynamics: An Introduction into Ecological Modelling for Students, Teachers & Scientists. – Springer-Verlag Berlin Heidelberg, 2011. – 397 p.
10. *Loh W.-Y*. Classification And Regression Trees // WIREs Data Mining and Knowledge Discovery. – 2011. – Vol.1. – P. 14–23.
11. *Geisser S*. Predictive Inference: An Introduction. Monographs on Statistics and Applied Probability. – NY: Chapman & Hall, 1993. – 265 p.
12. *Wu X., Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, at al.* Top 10 Algorithms In Data Mining // Knowledge and Information Systems. – 2008. – Vol. 14, Iss.1. – P. 1–37.
13. *Deshpande B*. 4 key advantages of using decision trees for predictive analytics. Available from: <http://www. simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics>.
14. *Data* Mining Algorithms in R. Classification. Decision Trees. Available from: <https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/Decision_Trees>.
15. *TU* Darmstadt. Decision Trees. Available from < http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf>.


## References


1. LYTVYNENKO, T., PANCHENKO, T., REDKO,V. (pre-print) Sales Forecasting using Data Mining Methods. Bulletin of Taras Shevchenko National University of Kyiv, Series Physics & Mathematics.
2. CHUGUEVA, I. (2012) Compositional Methods for Software Systems Specification and Verification (PhD Thesis). Kyiv. 177 p.
3. TreePlan Software. Introduction to Decision Trees. Available from: <http://treeplan.com/ chapters/introduction-to-decision-trees.pdf>.
4. Cambridge Dictionaries Online: English Dictionary. Available from: <http://dictionary. cambridge.org/dictionary/english/decision-tree>.
5. SAS Institute. Decision Trees – What Are They? Available from: <http://support.sas.com/ publishing/pubcat/chaps/57587.pdf>.
6. PANDEY, A.K. and GOYAL, N.K. (2013) Early Software Reliability Prediction: A Fuzzy Logic Approach. New Dehli: Springer India. 153 p.
7. BREIMAN, L., FRIEDMAN, J., STONE, C.J. and OLSHEN, R.A. (1984) Classification and Regression Trees. Chapman and Hall/CRC. 368 p.
8. GORDON, L. (2013) Using Classification and Regression Trees (CART) in SAS Enterprise Miner For Applications in Public Health. SAS Global Forum 2013. Paper 089-2013. 8 p.
9. JOPP F., REUTER, H. and BRECKLING, B., Eds. (2011) Modelling Complex Ecological Dynamics: An Introduction into Ecological Modelling for Students, Teachers & Scientists. Springer-Verlag Berlin Heidelberg. 397 p.
10. LOH W.-Y. (2011) Classification and regression trees. WIREs Data Mining and Knowledge Discovery, Vol.1. pp. 14-23.
11. GEISSER, S. (1993) Predictive inference: an introduction. Monographs on Statistics and Applied Probability. NY, Chapman & Hall. 265 p.

12. WU, X., KUMAR, V., QUINLAN, J.R., GHOSH, J., et al. (2008) Top 10 algorithms in data mining. Knowledge and Information Systems, Vol.14, issue 1. pp. 1-37.
13. DESHPANDE, B. 4 key advantages of using decision trees for predictive analytics. Available from: <http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics>.
14. Data Mining Algorithms in R. Classification. Decision Trees. Available from: <https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Classification/Decision_Trees>.
15. TU Darmstadt. Decision Trees. Available from < http://www.ke.tu-darmstadt.de/lehre/archiv/ws0809/mldm/dt.pdf>.

## *Information about author:*

*Lytvynenko Tetiana,*
fourth year student at Taras Shevchenko National University of Kiyv,
Faculty of Cybernetics, Department of Applied Statistics.
1 Ukrainian paper.
http://orcid.org/0000-0003-1662-3379.

## *Affiliation:*

Taras Shevchenko National University of Kiyv,
Faculty of Cybernetics, Department of Applied Statistics,
Academician Glushkov Avenue, 4D, 03680, Kyiv, Ukraine.
Phone: (093) 391 1509.
E-mail: tetiana.tet@gmail.com.