MACHINE LEARNING BASED SOLUTION FOR PREDICTING VOLUNTARY EMPLOYEE TURNOVER IN ORGANIZATION

Janis Judrups[1], Ronalds Cinks[2], Ilze Birzniece[3], Ilze Andersone[3]
[1]Baltic Computer Academy Ltd, Latvia; [2]University of Latvia, Latvia;
[3]Riga Technical University, Latvia
janis.judrups@bda.lv, ronalds.cinks@lu.lv , ilze.birzniece@rtu.lv, ilze.andersone@rtu.lv

Abstract. The cost associated with employee turnover and the shortage of available workforce in the market creates a situation where employee retention is crucial for the successful operation of an organization. Working remotely, especially in a situation with COVID-19 pandemic, increases the risks of voluntary employee turnover, since it makes the judgment of employee attitudes more difficult for the managers. Voluntary employee turnover (VET) measurements are one of the key indicators for evaluating the effectiveness of personnel management practices in organizations. The paper proposes a solution to decrease the risk of voluntary employee turnover in organizations. The authors propose a machine learning based model to identify the employees prone to voluntary employee turnover based on the employee data gathered and stored by the organization. The model will allow the managers to make a prediction based on data of the risks associated with voluntary employee turnover and to adjust the decision making process based on the information. To create the proposed IT solution for predicting the voluntary employee turnover analysis of models describing it has been performed to identify the most important factors that influence it. 9 factor groups with 67 factors of VET have been identified during the analysis. In the next step, 46 data clusters relevant for the decision making have been identified in specific organization and data from the clusters retrieved for the analysis. Based on the analysis a model for machine learning will be created, developed, and validated for the use in organizations.

Keywords: voluntary employee turnover; machine learning model; decision tree classification; decision support.

Introduction

The cost of employee turnover, and as well the deficit of high performing employees in the market leads to the fact that for organizations to be successful employee retention becomes critical [1]. An increasing amount of organizations worldwide organize activities with the goal of increasing employee retention [2-4]. Measurements of voluntary employee turnover (VET) can work as important indicators for evaluating the effectiveness of different human resource activities [5; 6].

Latvia shows high turnover rates comparing to other European countries. For example, employees who have been at least 10 years with a company within the past 5 years are on average 33% of all employees, but in other European countries this statistic is close to 40-50%. This statistic includes both voluntary and unvoluntary turnover [7].

Latvian companies are experiencing a shortage of qualified employees, furthermore, it is predicted that this shortage will increase [8]. Improvement in the financial situations of companies and their ability to afford qualified employees, the employee shortage will increase the competition for attracting and retaining top talent for rival companies [9].

HR managers and managers need objective data to make good decisions, or else they have to rely on intuition, understanding of human behaviour and experience, which has been shown in many cases to be an inferior approach [10; 11]. This need is further increased due to the current situation of working remotely and in the context of the Covid19 pandemic [12].

A tool that allows managers to better understand the situation with voluntary turnover within the organization would provide the information necessary to manage the human resources more meaningfully and to lower the risks of VET [13; 14].

The aim of this paper is to offer a solution to reduce VET in an organization based on predicting the VET risks of the organization's employees.

The main novelty of the paper is the choice of VET risk prediction factors based on the results of the VET analysis performed in the study. The 9 groups of VET factors identified in the study provide a scientifically based framework for the practical analysis and management of VET risk prediction results. The proposed VET prediction solution architecture with the help of a machine learning algorithm allows to create a tool that can be used in the work process with the possibility to adapt it to the specific data sources and environment of a particular organization.

Concept of VET prediction solution

The idea of the proposed VET predicting solution assumes that by numerically describing the factors characterizing VET, it is possible to find regularities that show the essential characteristics of employees who left voluntarily and who remained. This would help identify the relevant factors characterising VET and to build a profile of an employee at risk of VET. This, in turn, would help identifying individual employees at risk of VET and allow their managers to plan for corrective actions to ensure the employee's willingness to continue working with the organisation.

The following data sources would be used to gather factors characterising VET: (1) internal data of the organization (personal data of the employee, data on work performance, etc.) and (2) data characterizing the industry and business environment (unemployment rate, wage level in the industry, etc.). In the future, various personal data (social networks, etc.) for employee surveys and activities in external systems could serve as additional data sources, if such data are available.

Data used as a basis for factors characterising VET would be divided in (1) static data, which do not change or change rarely (age, education), and (2) dynamic data, which change often (performance results, dynamics of sick leaves, etc,). The use of dynamic data would help the solution to signal regular changes in VET status of an employee.

A machine learning algorithm should be developed and trained on organisation specific data to predict the risk of VET based on the factors characterising VET, which would help managers and HR experts make decisions about corrective actions based on dynamically changing data to reduce the risk of VET.

In practice, the solution would be implemented as a human resource (HR) management system module with three main user roles and their main functions (see Figure 1): (1) HR expert – review of predicted VET results, including factor details and factor dynamics; (2) employee manager – review of predicted results in accordance with the policy set by personnel management specialists, (3) administrator - management of additional data sources and access rights.

The proposed solution can be divided into three main parts - (1) data sources; (2) machine learning algorithm predicting VET; (3) VET prediction system. Expected data sources are HRM or other external systems that contain employee data relevant to the risk of VET. If the external systems are unable to use the API of the solution proposed a custom data integrator is provided. The algorithm service uses the machine learning algorithm developed and trained for the analysis, if data predicting VET and calculates the probability of employee VET.

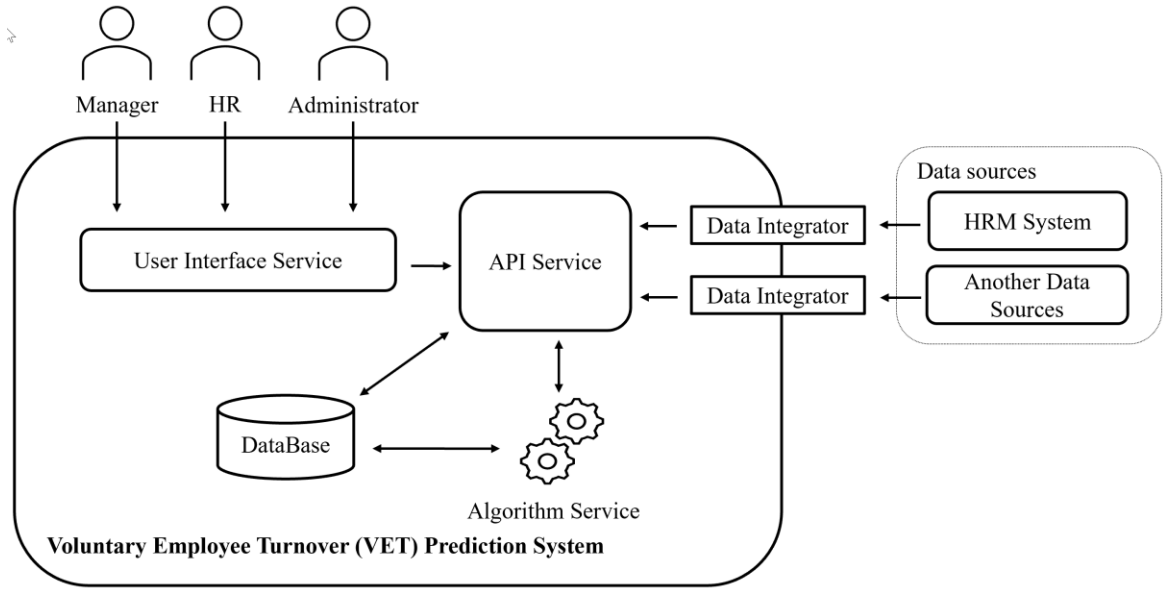

Fig. 1. **Common architecture of VET prediction**

The solution for the prediction of VET requires three main components: (1) data for VET predicting, (2) predicting algorithm, (3) IT system. This paper looks in detail at the aspect of VET data acquisition.

The development and validation of the prediction algorithm in the prototype system are planned for the future.

For the selection of predicting data, it is necessary to understand the aspects and operating models of VET, to identify the factors influencing VET and the data sources describing them. The corresponding data sources will be retrieved from the information systems of the target organization and used to develop a predicting algorithm.

Data Acquisition

Modern approaches in human resource management are relying on three voluntary turnover models – traditional (attitudes X alternatives) models, unfolding model and job embeddedness model.

1. Traditional voluntary turnover models [15] and their derivatives [16] mostly propose two major factors that predict voluntary turnover – (1) job satisfaction (commitment to the organization) and (2) perceived alternatives. When job satisfaction decreases, the chance that an employee will leave the organization increase. This specific aspect has had the most attention in the scientific literature [17]. The factor of perceived alternatives describes the employee's subjective feeling about the chances of getting another job [18].
2. The main assumptions within the unfolding model [18] are that – (1) emotional factors are not the only ones that can push one to terminate the job relationship, (2) an individual can compare the existing job with other opportunities, but es well can skip this step, and (3) a compatibility judgement can be made. The model describes several scenarios that an employee can take to terminate the job relationship. The unfolding model extends the understanding of the decision making behind voluntary turnover, including intuitive and less linear decision-making processes [19].
3. Michell and colleagues developed the job embeddedness concept, which explains why employees choose to stay with an employer, not why they choose to leave [20]. This construct describes the forces that make the employee feel that one cannot leave one's job. The job embeddedness construct has three subfactors: (1) links – to what degree the employee is connected with other employees or activities within the organization, (2) fit – to what extent the job and the community is in congruence with other aspects of the employee's life, and how easy it would be to break these ties, and (3) sacrifice – what would the employee loose in the case of leaving the company. Job embeddedness significantly predicts that employees stay in their current workplaces.

Identified prediction factors of voluntary turnover

Voluntary turnover models identify more than 50 scientifically proven general voluntary turnover predictors, where they could be further dived into more subfactors [21]. Potentially there are many factors that predict voluntary turnover, so it is important to understand which of them have the most impact [22]. Meta-analysis by Rubenstein and colleagues analysed 57 voluntary turnover predictors. They were categorized into 8 factors: (1) individual differences (e.g., age, sex, personality, education), (2) job characteristics (e.g., job safety, pay, role clarity), (3) traditional job attitudes (e.g., job satisfaction, organizational commitment), (4) personal variables (e.g., engagement, coping, stress), (5) organizational context (e.g., size, support, climate), (6) person-environment variables (e.g., fit, job embeddedness, life-work balance, (7) job market situation (e.g. available alternatives), (8) withdrawal cognitions and withdrawal behaviours (e.g., job search, absenteeism, performance, citizenship behaviour) [22].

Simply measuring different variable impact is not the best method to understand the complex relationship between these variables, therefore the theory congruent combinations of variables is important [22]. When focusing just on the predictor factors, it does not explain the process through which the employee goes through before deciding to quit [23]. Understanding the interaction between the different predictors, does not only increase the chance of predicting voluntary turnover, but as well to explain the process [24].

Voluntary turnover models should be evaluated for their ability to explain turnover rates [25]. Sometimes it can be valuable to analyse the available data even if it is not researched before, or even it seems unrelated to the behaviour of interest. In the end, a psychologically inconsistent model could

better predict behaviour than a model congruent with the theory [26]. Multivariate research tends to explain a much larger variance of turnover than separate variables [27; 28].

The use of big data and machine learning is seldom used in psychological research, but the trend is increasing [26]. Research using machine learning for predicting voluntary turnover shows considerably high prediction rates (e.g., Auc = 0.88; [13]).

Within this research both approaches are important – to predict and to explain voluntary turnover. So, it is important to choose both theoretically consistent variables, and the variables that might not have been previously mentioned or research in the literature, but are in line with the existing theory.

1. Demographic variables. Demographic variables have been researched widely in the context of voluntary turnover [22]. For the current research, the following demographic variables will be considered: employees' gender, age, time in the organization, time in the current position, education.
2. Job embeddedness. Job embeddedness theory predicts that family conditions, living location, and professional aspects can significantly impact an employee's decision to terminate the work relationship [20]. For the purposes of the current research, the following variables will be considered: the count and age of employee's children, count of attended training.
3. Job characteristics. Job characteristics have shown their importance in predicting voluntary turnover [22]. For the purposes of the current research, the following variables will be considered: position level in the organization, the count of employees under direct supervision, business trip count in the specified period.
4. Organizational context. Organizational culture, the behaviour of other employees and the employee's place in the organization are important factors, which impact the employee's behaviour [29]. In this research, the following factors will be considered: potential count of steps for development, employee turnover% in the team, turnover% difference between the team and the organization, changes in time.
5. Job attitudes, fairness. Job attitudes are one of the most important voluntary turnover predictors [17]. In this research, this factor will be measured indirectly, through data that are related to job satisfaction, fairness, and organizational commitment [30]: growth opportunities – frequency of position change, last position growth, the time spent in the current position in the specific time period, employees overall income and salary, salary changes and its amount, employees salary against – teams, departments, companies average salary, employees salary change frequency against – salary changes average in the team, department, and company, employees salary change% against – changes in% in the team, department, the company, employees salary against others who have the same position in the company.
6. Person - Organization fit. Employee's fit in the company significantly predicts the decisions for voluntary turnover [22; 31]. To assess employees fit with the organization and in the specific team, in the current research such factors will be considered representing the person-organization fit: age, gender, and time in the organization of the direct supervisor against the employee; colleague age, gender, time in the team, department, organization (against the employee).
7. Alternatives. An employee who feels a lot of alternatives for different job opportunities and has the perception that finding a new place of employment is easy will be more likely to choose the option of voluntary turnover [22; 32]. Although for the current research we do not have data, we argue that it would be important if data would become available to include such data sources into the model.
8. Shock events. The unfolding theory [33] predicts that specific shock events at work or home can increase the chance of the employee to consider quitting one's job. A specific interpretation of the event depends on the individual, but certain events are evaluated similarly by most individuals. For the current research, the following events will be considered: a salary deduction for the employee, direct supervisor leaving the organization, somebody from the team leaving the organization, the birth of a child, returning from parental leave.
9. Withdrawal behaviour. The withdrawal from the organization usually occurs within a longer period and often progresses until results in the voluntary termination of the job relationship. Behaviours that are characterized as withdrawal behaviours are: absenteeism, being late, and increasing the frequency and duration of sickness leaves [21; 34]. For the current research the following variables

will be considered: being late for work; sickness leaves – frequency, average length, changes in time.

Data analysis

The research data were retrieved in an ICT organization with 1400 employees. Data were retrieved from 46 data clusters. Initial data were gathered, combined, and analysed taking into consideration employee working status. Three categories of employees are defined – (1) employees voluntarily left, (2) employees left (without specifying the conditions, including those being fired) and (3) employees still working (see Table 1). Machine learning provides a range of techniques to address the task of distinguishing employees who are more likely to voluntarily leave the organization versus those employees who are loyal. Due to the data nature and properties of different machine learning approaches [35], clustering and classification are chosen as appropriate techniques to consider. For the data preparation and pre-processing step, initial separate data sets were merged into a joint data set. Categorical values are transformed into numerical ones. Data were of high quality and contained no missing values; therefore, no data cleaning was necessary.

Table 1

Summary of research data

Category	Number of employees
Employees voluntarily left	1172
Employees left (without specifying the conditions)	2129
Employees still working	1408

Clustering

To prepare employee data for clustering, additional processing was done, including data normalization and Principal Component Analysis (PCA) to reduce the number of features.

Employee data were clustered using the K-Means algorithm [36] implemented in Scikit-learn [37] and applying various settings. However, neither analysis of each category separately, nor all categories together lead to satisfactory clustering results. Silhouette analysis, which can be used to evaluate the quality of the resulting clusters, gives a weak result of silhouette coefficient between 0.2 and 0.4.

Classification

A classification task was set up to identify the employees prone to voluntary employee turnover. Pre-processing for classification included feature correlation analysis to reduce the number of features. CART algorithm [38] implemented in Scikit-learn was run over two data settings, 2 and 3 classes respectively. The 2-class layout included employees voluntarily left vs. employees still working whereas the 3-class layout separated all 3 categories as separate classes. Both layouts demonstrated considerable accuracy, however, the highest results were achieved for 2-class stetting, reaching average precision of 89% (87% for predicting employees voluntarily left, 90% for employees still working). These initial results clearly show the capabilities of classification to address this problem and motivates further steps to experiment with various data and algorithm settings. The choice of the human-readable classifier is essential in this problem domain as the decision tree allows to examine most influential features [39]. The decision tree reveals that education level, rotation in the enterprise, amount of business trips in more specific combinations are among the critical features when distinguishing between classifications.

Further research directions

Further research includes several steps. First, discussion with stakeholders approving the detected features and particularizing employee subgroups to analyse. Second, development and validation of rigid classification models for detecting employees prone to voluntary employee turnover. Third, design of workflow, which ensures machine learning-based IT solution for monitoring static and dynamic human-factors in the organizations to mitigate the risks of voluntary employee turnover.

Conclusions

1. Based on the results of the VET analysis performed in the study, 46 VET risk prediction factors were selected for the further research. The 9 groups of VET factors identified in the study provide a scientifically based framework for the practical analysis and management of VET risk forecasting results.
2. The paper offers a machine learning based VET prediction solution architecture for the development of a practical decision-making support tool for managers. Such an architecture would allow the developed decision-making tool to be adapted to the specific data sources and environment of the particular organization.
3. To determine whether the chosen factors have influence on the voluntary employee turnover, two machine learning techniques were run on the data sets – K-means clustering and CART decision tree algorithm. While K-means clustering returned poor results, the decision tree achieved average precision of 89%. It was found that employees education level, rotation in the enterprise, amount of business trips in more specific combinations are among the critical features when distinguishing between VET classifications.

Acknowledgements

The research leading to these results has received funding from the project "Competence Centre of Information and Communication Technologies" of the EU Structural funds, contract No. 1.2.1.1/18/A/003 signed between the IT Competence Centre and Central Finance and Contracting Agency, Research No. 1.4 "Development of a modular Personnel Management Solution (PPR) based on smart technologies".

References

[1] Hinkin T. R., Tracey J. B. The cost of turnover: Putting a price on the learning curve. Cornell hotel and restaurant administration quarterly, 41(3), 2000, pp. 14-21.
[2] Fulmer I. S., Gerhart B., Scott K. S. Are the 100 best better? An empirical investigation of the relationship between being a "great place to work" and firm performance. Personnel Psychology, 56(4), 2003, pp. 965-993.
[3] Hom P. W., Roberson L., Ellis A. D. Challenging conventional wisdom about who quits: Revelations from corporate America. Journal of Applied Psychology, 93(1), 2008, 1.
[4] Michele Kacmar K., Andrews M. C., Van Rooy D. L. Sure everyone can be replaced… but at what cost? Turnover as a predictor of unit-level performance. Academy of Management journal, 49(1), 2006, pp. 133-144.
[5] Shaw J. D., Gupta N., Delery J. E. Alternative conceptualizations of the relationship between voluntary turnover and organizational performance. Academy of management journal, 48(1), 2005, pp. 50-68.
[6] Ulrich D., Smallwood N. HR's new ROI: Return on intangibles. Human Resource Management: Published in Cooperation with the School of Business Administration, The University of Michigan and in alliance with the Society of Human Resources Management, 44(2), 2005, pp. 137-142.
[7] OECD. Labour Market Statistics: Employment by job tenure intervals: persons. OECD Employment and Labour Market Statistics. (database); 2016. [online] [23.03.2021]. Available at: http://dx.doi.org/10.1787/data-00295-en.
[8] Ekonomikas ministrija. Informativais zinojums par darba tirgus videja un ilgtermina prognozem, 2016. [online] [23.03.2021]. Available at: https://www.em.gov.lv/lv/nozares_politika/tautsaimniecibas_attistiba/informativais_zinojums_par_darba_tirgus_videja_un_ilgtermina_prognozem/.
[9] Ozolina-Ozola I., Gaile-Sarkane E. Job Change in Latvia: The Role of Labor Market Conditions and Employees' Socio-Demographic Characteristics. Procedia Computer Science, 104, 2017, pp. 197-204.
[10] Meehl P. E. Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. 1954.
[11] Kahneman D. Thinking, fast and slow. Macmillan. 2011.

[12] Rousseau S., Deschacht N. Public Awareness of Nature and the Environment During the COVID-19 Crisis. Environ Resource Econ 76, 2020, pp. 1149–1159.
[13] Rohit P., Ajit P. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms. International Journal of Advanced Research in Artificial Intelligence, 2016.
[14] Rombaut E., Guerry M.A. , The effectiveness of employee retention through an uplift modeling approach, International Journal of Manpower, Vol. 41 No. 8, 2020, pp. 1199-1220.
[15] March J. G., Simon H. A. Organizations. New York, NY: Wiley. 1958.
[16] Mobley W. H. Intermediate linkages in the relationship between job satisfaction and employee turnover. Journal of applied psychology, 62(2), 1977, p. 237.
[17] Steel R. P. Turnover theory at the empirical interface: Problems of fit and function. Academy of Management Review, 27(3), 2002, pp. 346-360.
[18] Lee T. W., Mitchell T. R. An alternative approach: The unfolding model of voluntary employee turnover. Academy of management review, 19(1), 1994, pp. 51-89.
[19] Harman W. S., Lee T. W., Mitchell T. R., Felps P. etc. The psychology of voluntary employee turnover. Current Directions in Psychological Science, 16(1), 2007, pp. 51-54.
[20] Mitchell T. R., Lee T. W. 5. The unfolding model of voluntary turnover and job embeddedness: Foundations for a comprehensive theory of attachment. Research in organizational behavior, 23, 2001, pp. 189-246.
[21] Holtom B. C., Mitchell T. R., Lee T. W. etc. 5 turnover and retention research: a glance at the past, a closer review of the present, and a venture into the future. The Academy of Management Annals, 2(1), 2008, pp. 231-274.
[22] RubensteinA. L., Eberly M. B., Lee T. etc. Looking beyond the trees: A meta-analysis and integration of voluntary turnover research. In Academy of Management Proceedings (Vol. 2015, No. 1, p. 12779). Briarcliff Manor, NY 10510: Academy of Management, 2015.
[23] Mobley W. H. Some unanswered questions in turnover and withdrawal research. Academy of management review, 7(1), 1982, pp. 111-116.
[24] Hom P. W., Griffeth R. W. What is wrong with turnover research? Commentary on Russell's critique. Industrial and Organizational Psychology, 6(2), 2013, pp. 174-181.
[25] Russell C. J. Is it time to voluntarily turn over theories of voluntary turnover?. Industrial and Organizational Psychology, 6(2), 2013, pp. 156-173.
[26] Yarkoni T., Westfall J. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science, 12(6), 2017, pp. 1100–1122.
[27] Price J. L., Mueller C. W. A causal model of turnover for nurses. Academy of management journal, 24(3), 1981, pp. 543-565.
[28] Prestholdt P. H., Lane I. M., Mathews R. C. Nurse turnover as reasoned action: Development of a process model. Journal of Applied Psychology, 72(2), 1987, pp. 221.
[29] Hom P. W., Lee T. W., Shaw J. D. etc. One hundred years of employee turnover theory and research. Journal of Applied Psychology, 102(3), 2017, p. 530.
[30] Pfeffer J. Organizational demography: Implications for management. California management review, 28(1), 1985, pp. 67-81.
[31] Kristof-Brown A. L., Zimmerman R. D., Johnson E. C. Consequences of individuals' fit at work: a meta-analysis of person–job, person–organization, person–group, and person–supervisor fit. Personnel psychology, 58(2), 2005, pp. 281-342.
[32] Griffeth R. W., Hom P. W., Gaertner S. A meta-analysis of antecedents and correlates of employee turnover: Update, moderator tests, and research implications for the next millennium. Journal of management, 26(3), 2000, pp. 463-488.
[33] Lee T. W., Mitchell T. R. An alternative approach: The unfolding model of voluntary employee turnover. Academy of management review, 19(1), 1994, pp. 51-89.
[34] Berry C. M., Lelchook A. M., Clark M. A. A meta-analysis of the interrelationships between employee lateness, absenteeism, and turnover: Implications for models of withdrawal behavior. Journal of Organizational Behavior, 33(5), 2012, 678-699.
[35] Witten I. H., Frank E., Hall M. A. etc. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, 4th ed., 2016.

[36] MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth "Berkeley symposium on mathematical statistics and probability". Vol. 1. No. 14. 1967.
[37] Pedregosa F., Varoquaux, G., Gramfort A. etc. Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011
[38] Breiman L., Friedman J.H., Olshen R.A. etc. Classification And Regression Trees (1st ed.). CRC Press., 1984.
[39] Birzniece I. The Use of Inductive Learning in Information Systems In: Proceedings of 16th "International Conference on Information and Software Technologies", Lithuania, Kaunas, April 22-23, 2010, pp. 95-101.