Prediction of VLDL Cholesterol Value with Machine Learning Techniques

İlhan UYSAL[1] and Cafer ÇALIŞKAN[2]

[1] Burdur Mehmet Akif Ersoy University, Bucak Emin Gülmez Tech. Sciences Vocational School, Computer Tech. Depart., BURDUR, iuysal@mehmetakif.edu.tr, ORCID: 0000-0002-6091-9110

[2] Antalya Bilim University, Faculty of Engineering, Computer Eng. Depart., ANTALYA, cafer.caliskan@antalya.edu.tr, ORCID: 0000-0002-9619-9207

Abstract. Cholesterol is an oil-like substance that is found in the membranes of animal cells and also carried in blood plasma, which has some vital functions in the human body, especially in the endocrine and digestive systems. Very Low-Density Lipoprotein (VLDL) is a lipid that is not gained with nutrients, instead produced by the body itself. However, it is considered to be in the bad cholesterol group since this type of cholesterol threatens cardiovascular health. As a result, it is normally expected to be at the lowest levels in the human body. In this study, It is applied some machine learning techniques to estimate VLDL Cholesterol value by some attributes such as age, sex, creatinine, aspartat transaminaz (AST), alanine transaminaz (ALT), free t4, glucose, and triglyceride. In this, the techniques include the Generalized Linear Model (GLM), Decision Tree (DT), and Gradient Boosted Trees (GBT). It is computed that GLM has the root-mean-squared-error value 0.655 and the correlation value 1.0 so consequently returns the best results compared to others.

Keywords: Biochemistry and hormone tests, VLDL Cholesterol, Machine Learning

1 Introduction

In recent years, there has been a huge increase in the amount of biological data obtained along with the great breakthroughs in technology. Consequently, traditional approaches in data analysis have become inadequate due to the huge amount of data. Therefore, some new approaches take place to process this data more efficiently. The most frequent one is to apply machine learning techniques for facilitating the management of the data. In this sense, machine learning algorithms can potentially reveal more outcome possibilities compared to the traditional methods. For instance, it was much difficult to detect the symptoms of a disease before, however machine learning algorithms can capture them in a more comprehensive way provided that the relative accurate data is present.

Moreover, some machine learning algorithms have great potential to produce results with higher accuracy in the early stages of some health problems.

Machine learning techniques are generally black-box. In other words, it is not possible to name the solution exactly. People normally have to rely on the decisions reached by black-box systems. However, this creates a contradiction. Artificial intelligence techniques that can explain the results or inferences it reaches are called Explainable Artificial Intelligence. The inference processes and results of such systems can be understood by humans. The sensor-based devices that are incorporated with the Internet of Things (IoT) their integration with mobile technologies are being referred to as the Internet of Healthcare Things (IoHT). This study aims to analyze biological data with some machine learning techniques and to assist physicians with medical diagnoses. Also, the solutions in this study have the potential of XAI and IoHT.

This study is mainly structured as follows: The first section includes preliminaries, then the second section discusses some machine learning techniques in the estimation of VLDL Cholesterol. The next section includes the research findings and finally, the last section discusses the results together with the contribution of the machine learning techniques in performance and precision.

2 Preliminaries

Machine Learning (ML) is the use of algorithms that helps with modeling computer systems to learn from data. Nowadays, it is possible to do many jobs by using machine learning. As a field, machine learning has many benefits such as determination of relationships within large amounts of data, an easier processing of image-based data, helping experts with difficult decisions, and more rapid processing of large amounts of data which is impossible to be done by the human brain at short notice [1].

There are various software available for applying machine learning methods. Rapidminer is one of these software that competes with many paid software in terms of usage rates and preference in order. It can also meet the needs of both beginner and expert level users. It also supports many file extensions such as csv, dat, and log. It provides more than 400 algorithms. Database systems can be with ease run with the Rapidminer. Users don’t need to adjust individual steps or parameters because experimental designs are automatically optimized through the meta operators. The software is written in Java language and compatible with languages such as Python, Weka, or R [2]. In this study, the

latest version of Rapidminer Studio (version 9.6) has been used for applying machine learning algorithms such as Generalized Linear Model (GLM), Decision Tree, and Gradient Boosted Trees. In what follows the basics of these methods are explained.

2.1 Generalized Linear Model

The generalized linear model was first discussed in 1972 by Nelder and Wedderburn. This model, which is used mostly in social sciences and medical applications, also plays an important role in the analysis of life data. This algorithm is an extension of traditional linear models as well as fits generalized linear models to the data by maximizing the log-likelihood. It is a combination of both linear and non-linear regression models that takes into account the non-normally distributed dependent variable. It is a powerful alternative to data conversion [3]. In some empirical studies, the generalized linear model is used as an alternative to data transformation in cases where the dependent variable does not conform to normal distribution [4].

A generalized linear model is made up of a linear predictor as in Equation 3.4 and two functions, namely a link and a variance function.

$$n_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} \qquad (1)$$

The link function describes how the mean of the distribution depends on the linear predictor.

$$qE(Y_i) = \mu_i \qquad (2)$$

$$g(\mu_i) = \eta_i \qquad (3)$$

A variance function describes how the variance depends on the mean where the dispersion parameter φ is a constant [5].

$$var(Y_i) = \varphi V(\mu) \qquad (4)$$

2.2 Decision Trees

Decision trees are one of the most popular supervised machine learning algorithms used for both classification and regression tasks [6]. Decision trees are algorithms that show more understandably the information set that the classifier has and show them in a way tree by sorting in a certain arrange the class options and the situations depending on the probabilities [7]. They can be run on categorical and numerical data. In decision trees, it is used some special terms similar to those of a real tree, such as branches and leaves. It has

decision nodes and leaf nodes in a decision tree [8]. Decision nodes are attributes used to make decisions in the dataset, classify or predict while leaf nodes keep decisions. The node at the topmost of the tree is called the root node. To reach a decision, a certain path is followed from the root of the tree to the leaf nodes [9].

With the algorithms used in the decision tree, it is aimed to extract a decision tree from a given data set by minimizing the generalization error [10]. The algorithm selection can be made according to the type of target variables. Iterative Dichotomiser 3 (ID3), Successor of ID3 (C4.5), Classification and Regression Tree (CART), and Chi-square Automatic Interaction Detector (CHAID) are the algorithms frequently used for decision trees [11].

One of the simplest decision tree algorithms that only work with categorical attributes is ID3 by J. R. Quinlan. ID3 generates a decision tree from a dataset that is represented by a table in general. It uses a top-down, greedy search through the given columns when constructing a decision tree, where it selects the attribute that is best for the classification of a given set. To decide what attribute is the best selection in terms of constructing a decision tree, ID3 uses Entropy and Information Gain [12].

Entropy is the measure of uncertainty or randomness in a given data. Intuitively, it indicates the foreseeability of a particular event. If a result of an event has a likelihood of 100%, the value of entropy is zero and if a result is 50%, the value of entropy reaches the maximum value as it projects perfect randomness. It is with the lowest possible probability to determine the outcome, so as a result, the entropy is likely to get the highest possible value. To build a decision tree, it requires to calculate of two types of entropy. The first one is the entropy E(S) using the frequency table of one attribute. To give a current state S and a probability P(x) of an event x of that state S:

$$E(S) = \sum_{x \in X} -\mathrm{P(x)}\ log2\ P(x) \qquad (5)$$

Let A be a selected attribute, S a current state with this attribute A and P(x) a probability of an event x of the same attribute A, then E(S, A) denotes the entropy that is using the frequency table of these two attributes S and A [13].

$$E(S, A) = \sum_{x \in X} [\mathrm{P(x)} \cdot E(S)] \qquad (6)$$

While Equation 4 relates to the Entropy of an attribute A, Equation 3 is the Entropy of the entire set.

Information gain (IG) is a criterion showing how effective it is a given feature is in classification and takes a value between 0 and 1. It is denoted by IG(S, A) for a state S is the important change in entropy after deciding on a certain attribute A. The relative change in entropy to the independent variables can be measured as below in equation 5:

$$IG(S,A) = \mathrm{E(S)} - E(S,A) \qquad (7)$$

Constructing a decision tree is all about selecting each attribute (A) to calculate Information Gain and finding such an attribute that returns the highest IG. The next decision node for the tree will be this attribute [14].

The C4.5 algorithm can be considered as an improved version of the ID3 algorithm. The main difference of this algorithm from the ID3 algorithm is that it uses normalization. While in the ID3 algorithm, entropy or information gain is calculated and decision points are determined, in the C4.5 algorithm, entropy values are kept as a ratio [15].

CART, which is capable of working with both numerical and categorical data, uses the Gini algorithm, regression trees, and random forest algorithms as branching criteria and produces binary trees. In the CART algorithm, it is performed the partitioning process by applying a certain criterion in a node. For this, first, it is taken into account the values with all the qualities, and after all matches, two splits are obtained [16].

The CHAID algorithm was created by Kaas in 1980 when there was no statistically significant difference to calculate the best division by combining the possible category pair of the prediction variable in pairs that fit the target variable. The Chi-square test is used instead of entropy or Gini metrics used to select the most suitable sections. To calculate the best division, the prediction variables are combined until there is no statistically significant difference in a pair that fits the target variable. The main difference between CHAID and other methods, while ID3, C4.5, and CRT derive binary trees, CHAID derives multiple trees [17].

2.3 Gradient Boosted Trees

Gradient Boosted Trees is a machine learning method that is used for regression and classification problems. This creates a model of decision trees, typically combined with weak prediction models [18]. The purpose of any supervised learning algorithm is to identify and minimize a loss function [19].

$$MSE = a_0 + \sum \left(y_i - y_i^p\right)^2 \qquad (8)$$

It is shown y_i *is the* i^{th} target value, y_i^p *is the* i^{th} prediction, and L=(y_i, y_i^p) is the loss function in Equation 8.

.

3 The Research Findings and Discussion

In this study, the dataset is obtained from the Burdur Provincial Health Directorate. It contains the blood test results of the patients at the internal-medicine-polyclinic of Burdur State Hospital from 2017 to 2018. It has 67 different laboratory analyses that consist of biochemistry and hormone test results of exactly 20004 patients. In this dataset, there are exactly 6883 male patients and 13121 female patients [20].

The comparison of the correlation coefficient of models used is given in Figure 1. Although the best performing model is GLM, other models also found very close to 1 (0,999 and 0,996) and a positive relationship.

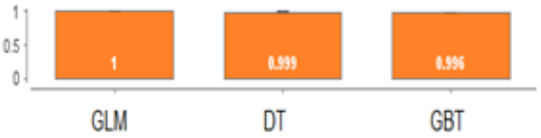

Fig. 1. The comparison of correlation coefficient of models

Root Mean Squared Error (RMSE) measures the error of a model in predicting quantitative data. The comparison of the applied models in terms of RMSE is given in Figure 2. According to our computation, GLM has the best performance and the value is 0.655.

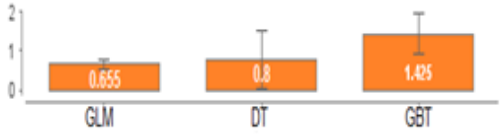

Fig. 2. The comparison of root mean squared error values of models

In comparing performances, another criterion used is the absolute error value, which shows the average vertical distance between each real value and the line that best fits the data. According to our computation, the decision tree method has the best performance in terms of the absolute error value. The decision tree algorithm has an absolute error value equal to 0.109 as is shown in Figure 3.

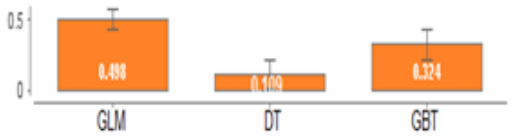

Fig. 3. The comparison of absolute error values of models

The Mean Squared Error (MSE) indicates how close a regression curve is to a series of points. MSE measures the performance of the estimator in the machine learning model and is always positive, and predictors with an MSE value close to zero are assumed to perform better. The comparison of squared error values of the applied models is given in Figure 4. While the MSE value of the GLM algorithm is 0.441, the decision tree algorithm has the MSE value equal to 1.01 and the gradient boosted has it as 2.245.

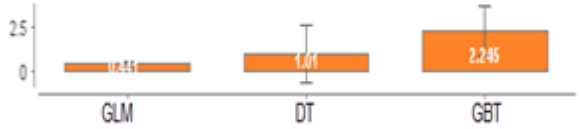

Fig. 4. The comparison of squared error values of models

Below, Figure 5 lists the weights of the attributes in computing the correlation. According to our computation, the most effective attribute on the estimation for VLDL Cholesterol value is triglyceride. Apart from that, other effective ones are Total Cholesterol, HDL Cholesterol, Glucose, LDL Cholesterol, ALT, Free T4, Insulin, Potassium, and Chlorine.

Attribute	Weight
Triglyceride	1.000
Cholesterol	0.365
HDL Cholesterol	0.338
Glucose	0.220
LDL Cholesterol	0.130
ALT	0.129
Free T4	0.095
Insuline	0.069
Potassium	0.069
Chlorine	0.068

Fig. 5. The weights by correlation

3.1 Generalized Linear Model

In this study, it is observed that attributes are supporting the prediction and attributes contradicting the prediction. In Figure 6, the list of supporting and contradicting attributes that are statistically significant in predicting the VLDL Cholesterol values is given with their significance levels. According to this list, the significant supporting attributes are Triglyceride, Lipase, Creatinine, and Free PSA in the reference range by the GLM. The attributes CEA, Chlorine, and Globulin are the contradicting factors in predicting the values.

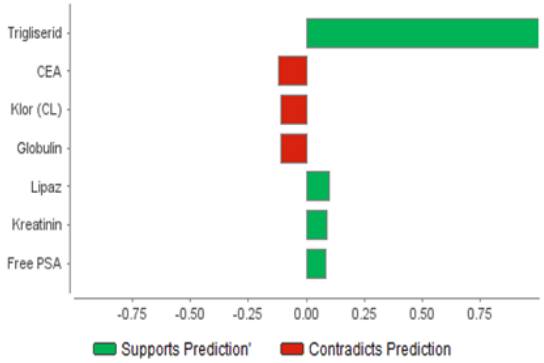

Fig. 6. Important factors for normal values of VLDL Cholesterol by the Generalized Linear Model

Predictions chart below shows the predictions (by using the generalized linear model) versus the actual values for the 40% validation cases. Each plot in the graph represents a specific prediction. As the plots get closer to the orange line, the model gets better [21]. See Figure 7.

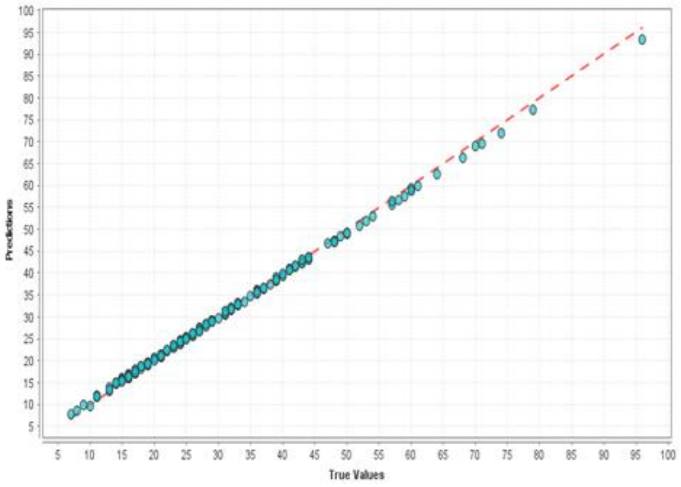

Fig. 7. Predictions Chart of VLDL Cholesterol by the Generalized Linear Model

3.2 Decision Tree

With the assumption that VLDL Cholesterol values are within the reference range, applying the Decision Tree method determines the significant supporting attributes as Triglyceride, Lipase, Creatinine, and Free PSA. The attributes CEA, Chlorine, and Globulin are the contradicting factors in prediction. See Figure 8.

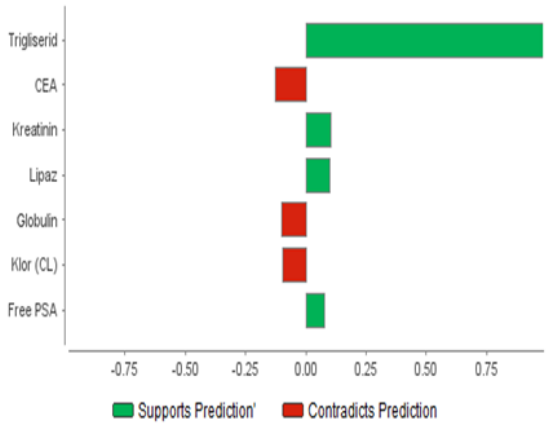

Fig. 8. Important factors for normal values of VLDL Cholesterol by the Decision Tree

The predictions chart obtained for VLDL Cholesterol values by applying the decision tree algorithm is given in Figure 9 below.

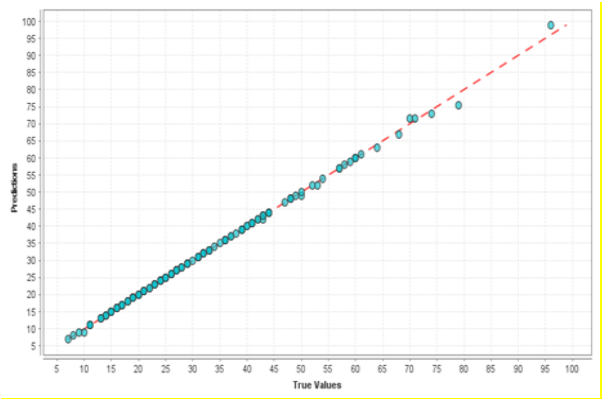

Fig. 9. Predictions Chart of VLDL Cholesterol by the Decision Tree

Optimal Parameters show the model's performance for different parameter settings. Error rates for maximal depth in normal values of VLDL Cholesterol by applying the decision tree algorithm are given in Figure 10. Accordingly, when the maximal depth is at least 10, the error rate drops to 0.6 percent.

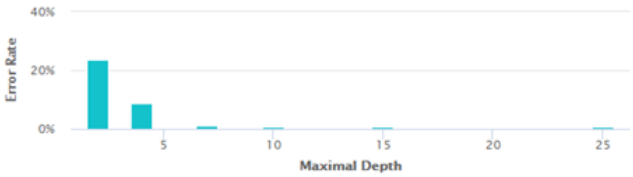

Fig. 10. Error rates for maximal depth

3.3 Gradient Boosted Trees

With the assumption that VLDL Cholesterol values are within the reference range, applying the Gradient Boosted Trees method determines the significant supporting attributes as Triglyceride, Lipase, Creatinine, and Free PSA. The attributes CEA, Chlorine, and Globulin are the contradicting factors in prediction. See Figure 11.

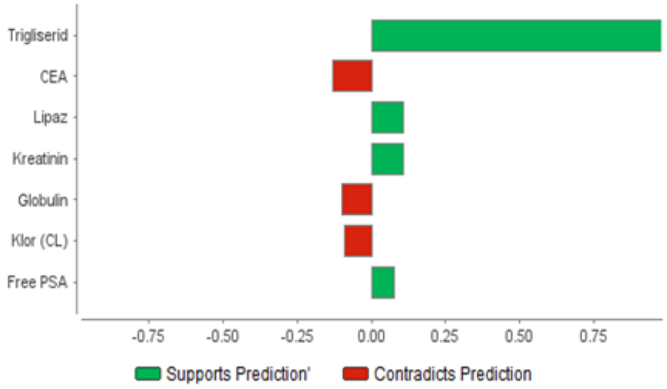

Fig. 11. Important factors for normal values of VLDL Cholesterol by the Gradient Boosted Trees

The predictions chart obtained for VLDL Cholesterol values by applying the gradient boosted trees algorithm is given in Figure 12 below.

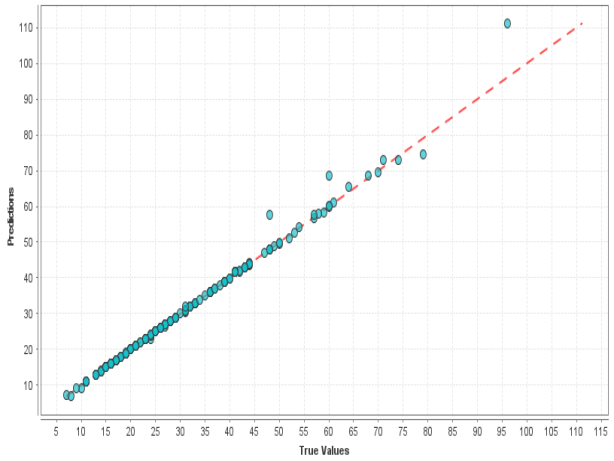

Fig. 12. Predictions Chart of VLDL Cholesterol by the Gradient Boosted Trees

Error rates for parameters in normal values of VLDL Cholesterol by applying the gradient boosted trees algorithm are given in Figure 13. Accordingly, the number of trees has been found as 90, maximal depth 7, learning rate 0.100, and error rate 1.6 percent.

Fig. 13. Error rates for parameters in the Gradient Boosted Trees

4 Conclusion

Doctors request blood tests according to the complaints of patients. For some particular diseases or health problems, certain test results are requested by doctors to find out about the underlying health issues. In this manner, the test results are crucial for them. Although the interconnection between the test results and health issues is sometimes complicated, some studies aim to help doctors to resolve such connections.

In this study, it is observed that age and gender are important factors in the prediction of all test results, as a result, they are also significant for VLDL Cholesterol. In addition, other important factors in predicting the VLDL Cholesterol are Triglycerides, Lipase, Creatinine, and Free PSA. This outcome is verified with various methods such as GLM, Decision Tree, and GBT algorithms in this study. They all have high success rates which are 95 percent or higher. For details see table 1 below.

Table 1. Comparison of Algorithms Performances

ALGORITHM	CORRELATION	RMSE	ABSOLUTE ERROR	SQUARED ERROR
GLM	1	0.655	0.498	0.441
DT	0.999	0.8	0.109	1.01
GBT	0.996	1.455	0.324	2.245

As the solutions reached in this study have the potential of XAI and IoHT, this topic will be a source of inspiration for future studies. Thanks to XAI, patients will know how the solutions given to them are realized, and thanks to IHoT, they will be able to communicate with medical facilities without interruption. In this way, there will be real-time information exchange between patients and medical facilities.

References

1. Turgut S. (2017). Cancer diagnosis using machine learning methods, MSc Thesis, Institute of Graduate Studies in Science and Engineering Department of Computer Engineering, Istanbul, Turkey, 19-20.
2. Uysal I., Bilen M., Ulukus S. (2018). Analysis of Classification Algorithms with Rapidminer, Gece Kitapligi Publishing House, Ankara, 517-521.
3. https://docs.rapidminer.com. Access date: 08.05.2021.
4. Karadag O., Aktas S. (2018). "Generalized Estimating Equations for Genetic Association Studies of Multi-Correlated Longitudinal Family Data, Gazi University Journal of Science, 33 (1): 273-280.
5. Shrestha N. (2019). Estimating the Probability of Earthquake Occurrence and Return Period Using Generalized Linear Models, Journal of Geoscience and Environment Protection, (7): 11-24.
6. Halima E., Zahra B., Hassan A. (2018). A comparative study of algorithms constructing decision trees, Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications - LOPAL '18, 1-5.
7. Danandeh Mehr, A , Bagheri, F , Safari, M. (2020). Electrical Energy Demand Prediction: A Comparison Between Genetic Programming and Decision Tree, Gazi University Journal of Science, 33 (1): 62-72.
8. Quinlan, J. R. (1986). Induction of decision trees, Machine Learning, 1: 81–106.
9. Bilgin M., Editor:Yılmaz A. (2018). Machine Learning, PapatyaBilim Publishing, 90-107.
10. Balaban M.E., Kartal E. (2018). Data Mining and Machine Learning, 2d Edition, Caglayan BookStore, 90-91.
11. Chien-Liang L., Ching-Lung F. (2019). Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan, Journal of Asian Architecture and Building Engineering, 18:6, 539-553.
12. Sudrajat R., Irianingsih I., Krisnawan D. (2017). Analysis of data mining classification by comparison of C4.5 and ID algorithms, IOP Conference Series: Materials Science and Engineering, 166.
13. https://www.thelearningmachine.ai/, Access date: 08.05.2021
14. Wang Y., Li Y., Song Y., Rong X., Zhang S. (2017). Improvement of ID3 Algorithm Based on Simplified Information Entropy and Coordination Degree. Algorithms 10(4): 123-124.

15. Ture M., Tokatli F., Kurt I. (2008). Using KaplanMeirer Analysis Together With Decision Tree Methods (CART, CHAID, QUEST, C4.5, and ID3) In Determining Recurrence-Free Survival of Breast Cancer Patients, Expert Systems With Applications, 3-4.
16. Yang Y., Ajoy V. (2015). Nina F.T., Suzanne S.F., Manufacturability Indices for High-Concentration Monoclonal Antibody Formulations, Computer Aided Chemical Engineering, 37:2147-2152.
17. Ozkan Y., Erol C.S. (2017). Bioinformatics DNA Microarray Data Mining, 2d Edition, PapatyaBilim University Publishing, 227-282.
18. Tapodhir A., Saurav C., Suman N., Abdul M. C., Sanjeev K. (2019). Symptoms to Disease Mapping and Doctor Recommendation System, International Journal of Engineering and Advanced Technology (IJEAT), 9(1): 2249 – 8958.
19. Natekin A., Knoll A. (2013). Gradient Boosting Machines, a tutorial, Frontiers in Neurorobotics, 7: 20-21.
20. Uysal İ., Çalışkan C. (2019). Some Machine Learning Techniques For Medical Diagnosis, Antalya Bilim University Institute of Science Master Thesis, 11.
21. https://docs.rapidminer.com/8.1/studio/operators/validation/visual/lift_chart.html. Access date: 11.05.2021.