Human Resource Intelligent Recommendation Method based on Improved Decision Tree Algorithm

XIAOLONG JIANG
Faculty of Humanities and Social Sciences,
Macao Polytechnic University,
Macao, 23200,
CHINA

Abstract: - Due to the current complexity of human resource recommendation methods and the large number of recommended errors, it is difficult to meet the actual requirements of human resource recommendation. Therefore, this paper proposes an intelligent human resource recommendation method based on an improved decision tree algorithm. Firstly, based on the improved decision tree algorithm, classify resources to accelerate learning speed and avoid errors during learning. Secondly, collect human resources data and store the preprocessed human resources data in the human resources data warehouse. Finally, successful implementation of intelligent human resource recommendations. The experimental results are as follows: After using the human resource intelligent recommendation method based on the improved decision tree algorithm, the number of successful hires was higher than the other two methods. Compared with the method proposed in this paper, method 1 reduced 1098 positions, and method 2 reduced 1008 positions. The effectiveness of the human resource intelligent recommendation method based on the improved decision tree algorithm has been demonstrated.

Key-Words: - Decision tree algorithm, Human resource management, Intelligent technology, Human resources database, ID3 algorithm, C4.5 Decision Tree Algorithm, Comparison experiment.

Received: March 12, 2024. Revised: October 11, 2024. Accepted: November 15, 2024. Published: December 23, 2024.

1 Introduction

This document proposes a decision-making-based approach for handling human resource recommendations to improve the overall quality and execution efficiency of human resource recommendations. Better collect data from the main sources of human resources data, categorize the data by source category, and use it as a reference database for human resources. Currently, we are addressing the shortcomings of preliminary human resource data through expansion, improvement, transformation, and data collection; Merge the data from the data table into the improved indicator table C4.5 ID3. This indomitable attitude has led to a trend of taking action, repeatedly citing different sources of proposals, producing the same thematic results, and meeting and supporting human resource recommendations. The conclusion drawn by experts is that adopting an improved decision-making method based on comprehensive human resource recommendations is more successful than the other two methods. Combined with the method used in this document, Method 1 reduced 1098 projects, and Method 2 reduced 1008 projects. A comprehensive approach to strengthening human resource consulting through more effective decision-making, [1].

Many experts and scholars have spent a lot of time and effort developing human resource recommendation methods, proposing many effective human resource recommendation methods. Initially, this was a personnel recommendation method based on an expert system. This approach first involves establishing a knowledge base of human resource recommendations, which includes a large number of suggested rules. According to the suggested guidelines, suggestions regarding human resources have been put forward. However, this approach is not effective as the results of human resource recommendations are closely related to the suggested rules, resulting in low trust in human resource recommendations. Then, a human resource recommendation algorithm based on K-means clustering emerged, which classifies the raw data and recommends appropriate positions based on analysis. However, the limitations of this method are also evident, often leading to incorrect results in human resource recommendations and the inability to obtain high-quality human resource recommendations, [2].

This project aims to address the existing issues in talent recommendation systems and develop a new decision tree-based talent recommendation method to improve recommendation quality, reduce recommendation errors, and provide a decision-making basis for enterprises. On this basis, this study will also conduct comparative experiments on different HR recommendation methods to test the effectiveness and superiority of the algorithm, [3].

2 Human Resource Recommendation based on Decision Tree Algorithm

2.1 Overall Recommendation Process Design

The entire process of human resource recommendation is to collect human resource data, combine it with human resource attributes, preprocess the collected human resource data, and store the preprocessed human resource information in the human resource database. Utilize human resource data from the database, use a decision tree algorithm to execute human resource recommendations, compile a suggestion list based on the received suggestions, and ultimately determine human resource recommendations. Figure 1 shows the overall process of human resource recommendations, [4].

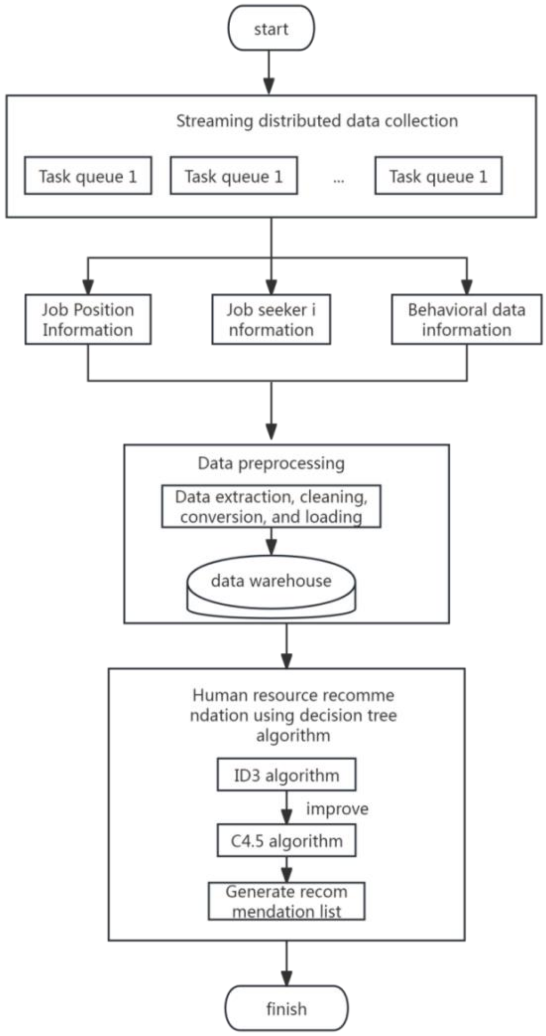

Fig. 1: Overall process of human resource recommendation

2.2 Streaming Distributed Data Collection

How to extract useful talent information from massive data is the key to designing and implementing personalized talent recommendations.

A method of collecting personnel information through distributed clusters. This project will fully leverage the high availability parallelism of clusters and complete real-time and high-quality collection of personnel data through divided task queues, thereby effectively improving the effectiveness of data preprocessing and algorithm generation.

By dividing the queue of the same task in each subsystem, both the speed and bandwidth of data collection can be effectively improved. At the same time, data collection tasks can be extended by configuring subsystems, effectively improving the scalability of human resource data collection. The main server utilizes the efficient data flow storage capability in the memory model to achieve flow processing based on the memory model, thereby collecting distributed human resource data in real time. Figure 2 shows the process of collecting and allocating human resource data through the process, [5].

The specific collection process includes:

(1) Initialize the data collection server cluster group, define the primary server, and assign task roles to each subsystem, [6].

(2) According to partition rules, separate each server, randomly assign different HR channel source categories to that partition, and create task queues related to partitions of the same source category.

(3) The main server processes and separates the task queue human resource data collected by each server in a streaming manner, and classifies the human resource data on a large scale based on the quantitative standards of micro batch processing intervals. Relative micro-package tasks can be processed as temporary packages.

(4) By executing multiple micro-batch processing tasks simultaneously on the main server, a relative flow of human resource data can be achieved. By utilizing memory-based processing, human resource data streams are classified

according to source categories, unified human resource data categories, and classified human resource data are retained to form a set of human resource data, laying the foundation for future human resource data preprocessing, [7].

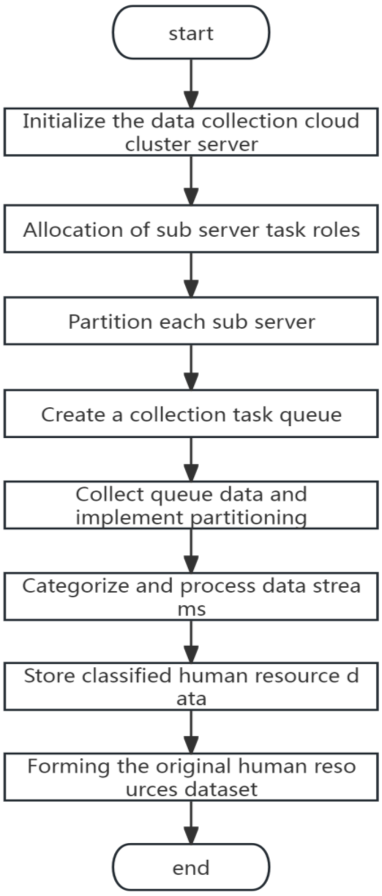

Fig. 2: Flow-based distributed human resource data collection process

2.3 Data Preprocessing

The raw human resources dataset obtained through data flow collection includes user information, performance information, behavioral data, etc. User information refers to the job search information registered by job seekers on the platform; Job information refers to the job requirements and basic working conditions posted by recruitment departments or individuals on specific platforms; Behavioral data refers to work-related operations performed by users when applied to a specific platform, such as applications, bookmarks, and viewing. Due to the different sources of collected raw data, there are many issues such as data duplication, loss, inconsistency, and inaccuracy. If directly used for subsequent recommendation algorithms, it will reduce the accuracy of recommendation results. Therefore, the human resources baseline dataset should be preprocessed to improve the accuracy of human resources recommendations. In order to address the shortcomings of the original human resources dataset mentioned above, a data preprocessing program has been developed, which mainly consists of three parts: data retrieval, data cleaning and transformation, and data uploading. Especially, extracting the required data from the collected raw human resources data according to the needs of the data warehouse is a part of data extraction; The main task of the data cleaning and transformation component is to filter and remove noncompliant data from the retrieved data, such as incomplete data, incorrect data, and duplicate data. According to specific rules, inconsistent data and data granularity are transformed to meet the requirements of the target data storage. The data download task is to transfer the extracted, cleaned, and transformed data to the target data repository. The entire preprocessing process is shown in Figure 3, [8].

The preprocessing of the above data can effectively process the defect data collected in the original human resources dataset, generate actual human resources data, and establish a data repository as input for subsequent recommendation algorithms to implement personalized human resources recommendations, [9].

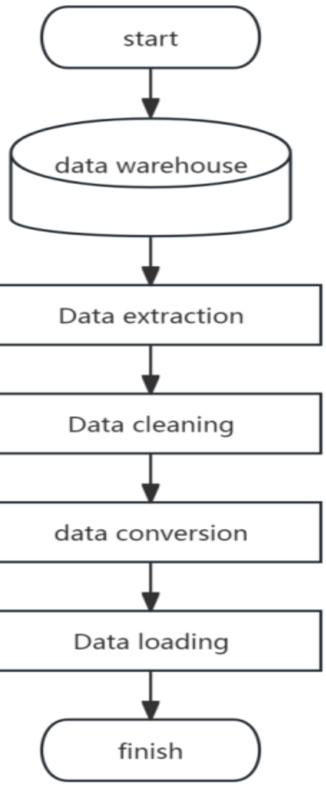

Fig. 3: Data preprocessing process

2.4 Human Resource Recommendation using Decision Tree Algorithm

2.4.1 ID3 Algorithm

The basic idea of the ID3 algorithm is to use the optimal category characteristics as the partitioning criterion and repeat it until a decision tree that meets the requirements is generated. The consistency of data allocation is the foundation of attribute purity. Less consistent characteristics will be purer. In addition, when the purity of attributes is high, fewer branches can be generated, indicating that this feature is beneficial for classification. In the database, information entropy is used to measure whether the attributes between the test samples are pure. As the experimental data decreases, the purity of the sample also increases accordingly. If the result is 0, it means that all examples have the same target attribute. Assuming that the preprocessed data is D, which is a set of multiple attributes, the entropy D can be expressed as:

$$\inf\ o(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{1}$$

In the formula： p_i The likelihood of the i-th category occurring in the entire training data set D can be expressed by the ratio of the sampled data of that category to the total number of samples; M refers to target attributes with various numerical values, [10].

Let attribute A have v different values in training dataset D, and divide the training dataset D into v sub datasets based on the v class values. Therefore, the expected information of attribute A on training dataset D can be expressed as:

$$\inf\ o_A(D) = \sum_{j=1}^{\upsilon} \frac{|D_j|}{|D|} \inf\ o(D_j) \tag{2}$$

In the formula： $\inf\ o_A(D)$ Indicates the impurity of the training dataset D after partitioning it through attribute A, [11].

Before and after segmenting training sample group D, the differences in these impurities can be expressed through information gain, and the specific formula is as follows:

$$gain(A) = \inf\ o(D) - \inf\ oA(D) \tag{3}$$

In the formula： $gain(A)$ The larger the size, $\inf\ o_A(D)$The smaller the value, the higher the purity of the training dataset D divided by attribute A, indicating that attribute A is more advantageous for classification.

When the number of different values for a certain attribute in the learning dataset is large, the information coefficient of that attribute is also high, which will result in evaluating it every time information augmentation is used to select these attributes. This will lead to certain flaws, such as the classification values implemented using these attributes not being significant when they have a large number of different values (such as numbers or names). In addition, when the classification type is larger, it can also lead to too many branches in the decision tree. If the differential continuity value is delayed, the ID3 algorithm cannot differentiate these attributes when considering hierarchical structure and age. Therefore, in order to effectively improve these ID3 defects, algorithm based ID3 algorithms can be improved, and C4.5 decision tree algorithms can be provided, [12].

2.4.2 C4.5 Decision Tree Algorithm

The C4.5 decision tree algorithm is different from the ID3 algorithm because it uses the information amplification coefficient as a measure to select classification attributes. Information separation can be defined as:

$$split\_\inf\ o_A(D) = -\sum_{j=1}^{\upsilon} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|} \tag{4}$$

The information separation in the formula can be understood as the entropy of various values j related to attribute A in the training dataset D.

The information gain rate can be defined as:

$$gain\_ratio(A) = \frac{gain(A)}{split\_ratio(A)} \tag{5}$$

The specific process of the C4.5 decision tree algorithm is as follows:

Algorithm: C4.5 decision tree (training dataset D, candidate attribute set D.Attributelist)

Input: Training dataset D, candidate D's attribute set.

Export: Decision Tree

1) Construct the root node n;

2) If the number of types in training dataset D is 1, then n is a worksheet node, and n is identified by this unique type, returning n;

3) If d. attributelist has no attributes, or if the number of samples d is less than the specified value, then n is the node of the worksheet, and n is the type that most commonly appears when returning n;

4) Calculate the information gain for each attribute in D. Attributeless, where M is the attribute corresponding to the maximum information gain;

5) The property of testing n is m;

6) If the test attribute is continuous, please find the separation threshold for that attribute;

7) Give each node N a brand new node L. If the subset of sample D for node L is 0, separate node L to create a new worksheet node and label it as the most common type D. In addition, using D 'as input, configure the algorithm on node L and continue separation;

8) Convert each branch of the decision tree into an "if it is" rule, store it in a two-dimensional array, and display each path from root to leaf through a two-dimensional array of each row;

9) Node classification error executed and cut. Before the specified end condition is met, that is, the data within the node belongs to the same category, [13].

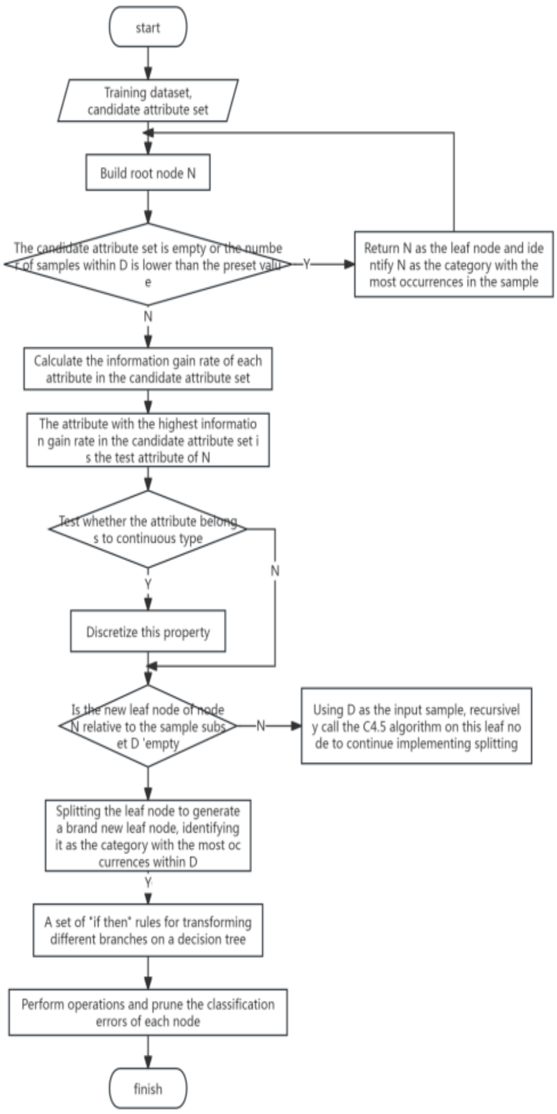

Fig. 4: C4.5 Decision Tree Algorithm Process

The process of C4.5 decision tree algorithm is shown in Figure 4.

By using the C4.5 decision tree algorithm mentioned above and recommending the required human resources as a prerequisite for termination, we will obtain the corresponding human resource recommendations. A list of suggestions for completing human resources recommendations has been prepared based on the survey results, [14].

2.4.3 Talent Resource Recommendation Steps of Decision Tree Algorithm

Step 1: Collect a talent resource recommendation dataset and perform preprocessing.

Step 2: Divide the talent resource recommendation dataset into a training sample set and a testing sample set.

Step 3: Establish the root node N of the tree. If the type is 1, the root node is the leaf node. If there are no attributes in the attribute list, the root node is the leaf node.

Step 4: Calculate the information gain of each attribute in the attribute list and save the attribute with the highest information gain.

Step 5: Set the attribute of the root node to the attribute with the highest information gain

Step 6: Give each node a brand new leaf node.

Step 7: Transform the branches of the decision tree into a recommendation rule.

Step 8: Prune the erroneous nodes to obtain a recommendation rule decision tree.

Step 9: Obtain human resource recommendation results based on the established recommendation rule decision tree.

3 Experimental Testing and Analysis

To test the effectiveness of the improved decision tree-based recommendation HR method, experimental tests were conducted to compare its management effectiveness with traditional methods 1 and 2. In order to verify the correctness of the proposed method, this chapter will use it for testing and verify whether it meets established experimental standards, and obtain good recommended results. Before conducting the experiment, preparation work must be carried out to ensure the accuracy of the experiment.

3.1 Test Preparation

On this basis, the algorithm was improved and optimized using MATLAB as a simulation environment. Table 1 shows the parameters of the experimental environment.

Table 1. Test Environment

name	Model and parameters
CPU	Intel® Pentium®Dual-Core E3600
operating system	Windows 2005 EE SP3
Memory	Kingston 4G
Application software	WWeka V5.7.1
Hard disk	Seagate STAT 450G

At present, the main algorithms on the testing environment market are Knime, Rapidminer, and Weka, which are based on Weka and can be easily improved through Weka. Therefore, in this experimental study, Weka was chosen as the working platform and Java was chosen as the development environment. During the research process, detailed adjustments were made to the experimental environment parameters. Table 2 shows the detailed parameters of the WEKA test.

Table 2. Test Parameters

parameter	numerical value
Memory	Maxheap=1024M
Classification algorithm	C4.9.ConfidenceFactor=0.75,|minNumObj=3
file format	CSV
Test Options	Cross-validation, 10 folder

3.2 Experimental Results and Analysis

Based on these experimental conditions, three recommended human resource methods were compared. 5000 human resources data were selected as the sample for the experiment. The vacancy rates of each method were verified and compared using the three suggested human resources methods. The experimental results are shown in Table 3, [15].

Table 3. Test Results

Sample data/piece	1 number of personnel who can recommend methods for improving decision-feeding algorithms in human resources intelligence	Number of personnel in Method 1/person	Number of personnel in Method 2
1000	562	328	526
2000	1305	639	1274
3000	1952	1037	1540
4000	2306	1208	1864
5000	3047	1961	2019

From the experimental results in Table 3, it can be seen that compared with the other two methods recommended by HR, this algorithm achieves a higher probability of effectiveness. Under this method, the number of employees is corresponding. This algorithm will decrease in practical applications as the number of samples increases. Compared with Experiment 1, this algorithm has 1098 and 1008 experimental results. The experimental results show that the artificial intelligence recommendation algorithm based on the improved decision tree algorithm proposed in this paper is feasible.

4 Summary

This article conducts in-depth research on the decision tree method and its application in human capital intelligent recommendation methods and combines corresponding data preprocessing with distributed data collection processes to establish a human capital information warehouse. The data stored in this warehouse becomes the input training dataset for decision tree calculation, and the algorithm is used to obtain a list of recommendation results, thereby achieving human resource recommendations. In practice, the proposed method has advantages in both recommendation quality and effectiveness and can be used on various human resources recruitment platforms to provide different users with the required human resources information.

Acknowledgement:
This work was supported by Project of Logistics Management and Engineering Teaching Steering Committee of the Ministry of Education (No. JZW2023228).

References:

[1] Singer G, Cohen I. An Objective-Based Entropy Approach for Interpretable Decision Tree Models in Support of Human Resource Management: The Case of Absenteeism at Work. *Entropy*, 2020, 22(8): 821-831. DOI: 10.3390/e22080821.

[2] Mao LZW. Analysis of entrepreneurship education in colleges and based on improved decision tree algorithm and fuzzy mathematics. *Journal of intelligent & fuzzy systems: Applications in Engineering and*

Technology, 2021, 40(2): 2095-2107. DOI: 10.3233/JIFS-189210.

[3] Kang IG, Kim N, Loh WY, Bichelmeyer BA. A Machine-Learning Classification Tree Model of Perceived Organizational Performance in U.S. Federal Government Health Agencies. *Sustainability*, 2021, 13(18): 79-86. DOI: 10.3390/su131810329.

[4] Gao Y, Huang C, Hu M, Feng J, Yang X. Research on Book Personalized Recommendation Method Based on Collaborative Filtering Algorithm. *IOP Publishing*, 2019, 252(5):1-6. DOI:10.1088/1755-1315/252/5/052099.

[5] Serek A, Orynbekova K, Talasbek A, Kariboz D, Saimassay G, Bogdanchikov A. Recommendation System for Human Resource Management by the Use of Apache Spark Cluster. Kaskelen, Kazakhstan, *2023 17th International Conference on Electronics Computer and Computation (ICECCO)*, 2023: 1-4. DOI: 10.1109/ICECCO58239.2023.10147129.

[6] Wang X. College Student Employment Management Recommendation System Based on Decision Tree Algorithm. *2022 International Conference on Education, Network and Information Technology (ICENIT)*, Liverpool, United Kingdom, 2022: 169-173. DOI: 10.1109/ICENIT57306.2022.00044.

[7] Kolankar P, Patel R, Dangi N, Sharma S, Jain S. Exploiting the Most Similar Cases Using Decision Tree to Render Recommendation. *Data Science and Analytics*, 2020, 1230: 290-304. DOI: 10.1007/978-981-15-5830-6_25.

[8] Wagner HNR, Koeke H, Daehne S, Niemann S, Huehne C, Khakimova R. Decision Tree-based Machine Learning to Optimize the Laminate Stacking of Composite Cylinders for Maximum Buckling Load and Minimum Imperfection Sensitivity. *Composite Structures*, 2019, 220(JUL.): 45-63. DOI: 10.1016/j.compstruct.2019.02.103.

[9] Rosewelt, Antony L, Renjit, Arokia J. A content recommendation system for effective e-learning using embedded feature selection and fuzzy DT based CNN. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 2020, 39(1):795-808. DOI: 10.3233/JIFS-191721.

[10] Shulman E, Wolf L. Meta Decision Trees for Explainable Recommendation Systems. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, 2019, 7: 365-371. DOI: 10.1145/3375627.3375876.

[11] Adnan M, Sarno R, Sungkono KR. Sentiment Analysis of Restaurant Review with Classification Approach in the Decision Tree-J48 Algorithm. *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE, Semarang, Indonesia, 2019: 121-126. DOI: 10.1109/ISEMANTIC.2019.8884282.

[12] Qomariyah NN, Heriyanni E, Fajar AN, Kazakov D. Comparative Analysis of Decision Tree Algorithm for Learning Ordinal Data Expressed as Pairwise Comparisons. *2020 8th International Conference on Information and Communication Technology (ICoICT)*. Yogyakarta, Indonesia, 2020: 1-4. DOI:10.1109/ICoICT49345.2020.9166341.

[13] Golshanrad P, Rahmani H, Karimian B, Karimkhani F, Weiss G. MEGA: Predicting the best classifier combination using meta-learning and a genetic algorithm. *Intelligent Data Analysis*, 2021, 25(6): 1547-1563. DOI: 10.3233/IDA-205494.

[14] Mannapov I. The Improvement of Decision Tree Construction Algorithm Based On Quantum Heuristic Algorithms. *Lobachevskii Journal of Mathematics*, 2023, 44: 724-732. DOI: 10.1134/S1995080223020269.

[15] Kherif O, Benmahamed Y, Teguar M, Boubakeur A, Ghoneim SSM. Accuracy Improvement of Power Transformer Faults Diagnostic Using KNN Classifier with Decision Tree Principle. *IEEE Access*, 2021, 9: 81693-81701. DOI: 10.1109/ACCESS.2021.3086135.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)
The author contributed to the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself
This work was supported by Project of Logistics Management and Engineering Teaching Steering Committee of the Ministry of Education (No. JZW2023228).

Conflict of Interest
The author has no conflicts of interest to declare.