Development of methods for assessing the performance of teachers using of TUIT-LMS data

F A Alisherov\textsuperscript{1}, S Q Iskandarov\textsuperscript{2,5}, S Sh Bekturdiyev\textsuperscript{3} and O K Khujaev\textsuperscript{4}

\textsuperscript{1}Founder of Algorismus Ltd, Amir Timur, 24a, Tashkent, Uzbekistan
\textsuperscript{2}Phd researcher, Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Amir Timur, 108, Tashkent, 100200, Uzbekistan
\textsuperscript{3}Assistant teacher of the department of information processing and control systems, Tashkent State Technical University named after Islam Karimov, Tashkent, Uzbekistan
\textsuperscript{4}Urgench branch of Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, 110, al-Khwarizmi str., Urgench, 220100, Uzbekistan

\textsuperscript{5}E-mail: sanjariskandarov@gmail.com

Abstract. The use of data mining methods is one of the current trends, which is widely used in finance, health, telecommunications, e-learning and others. Much of the research in the field of education is focused on the assessment of students' performance. But the impact of teachers on the quality of education is also significant. The traditional way to evaluate a teacher’s performance is to conduct an assessment survey that takes into account the student’s point of view. The article solves the problem of classification of teacher's activity using Generalized Linear Model, Deep Learning, Decision Tree, Random Forest methods based on the results of the survey and textual data based on the survey data conducted in the TUIT-LMS system and determines the reliability of the results.

1. Introduction

The development of information and communication technologies (ICTs) has changed the way knowledge is acquired and opened a new path for distance learning. In the modern era, e-education has become widespread. Learning management systems (LMS - Learning management system) allowed teachers and students to easily and clearly assimilate educational materials [1].

Data mining is the process of extracting important knowledge from large amounts of data in data warehouses, databases, or other data carriers.

The gradual introduction of data mining methods into the education system is gaining practical importance. The application of data mining methods in the education system is aimed at teaching and assessing students, storing, archiving and analyzing information. Data mining methods are used in the education system mainly in the following areas [2]:

- psychology and psychometrics;
- machine learning;
- education statistics;
- information visualization and computer modeling.
• give recommendations on the choice of subject and teacher.
• formation of knowledge for management, etc.

Applying Data Mining techniques to data in education is called "Educational Data Mining (EDM)". EDM programs are used to measure and predict student or teacher performance and develop recommendations and suggestions. Data Mining can be used to provide students with a variety of recommendations to help them learn more effectively and efficiently[11].

2. Materials and methods
Educational Data Mining is an area that provides interesting opportunities for researchers and practitioners. This area helps higher education institutions with effective methods to increase student productivity and improve student achievement. Data Mining is an important tool that helps organizations improve decision making and analyze new patterns and relationships between large amounts of data [10,12].

Data mining in the education system consists of the processes of formulating, testing and improving hypotheses, as shown in figure 1 [14].

![Figure 1. Data mining in the education system.](image)

Data mining techniques can be classified as follows:

• statistics and visualization;
• intelligent analysis of web data;
• clustering and classification;
• the rule of associativity;
• intellectual analysis of text information.

It is possible to predict, classify, make recommendations and support decision-making using the processing of textual and numerical data of the TUIT-LMS system using intellectual analysis methods. In particular, the results of surveys conducted among students at TUIT-LMS can be used to assess and analyze the activities of teachers using Data mining methods (data mining).

Data mining techniques are used to analyze large amounts of data. In general, the processes in data mining methods can be described as follows [15,16].
2.1. Decision Tree method
It divides text data into classes based on a given set of data. In this case, the data is formed in the form of a tree. The main tree node starts choosing from a significant parameter, and also divides the data into two classes [7]. As a result, text information is divided into two classes. The process can be illustrated in figure 2 below.

![Figure 2](image)

**Figure 2.** The structure of the summary tree method.

One of the advantages of the decision tree method is that it sorts the most significant value and depends on the algorithm. The ID3 and Best First Tree (BFTree) algorithms of the decision tree method are presented as follows.

**ID3 method.** You can classify text parameters based on a given dataset \( T \) using the following expression:

\[
I(S_1, S_2, \ldots, S_n) = -\sum_{i=1}^{n} \frac{S_i}{S} \log \left( \frac{S_i}{S} \right)
\]

(1)

here \( n \) - is the number of characters in the class. \( S_i \) - after the differences in the informativeness of the parameters are calculated, the entropy for the vertices of the decision tree is calculated

\[
E(A) = \sum_{j=1}^{m} \frac{S_j}{S} I(S_j, \ldots, S_{n_j})
\]

(2)

Where \( m \) is the number of network dependencies of the parameter \( A \). It follows from this that, based on the highest values of the informativeness of the parameter \( A \), the best parameter is selected for classification, which is calculated based on the following expression

\[
Gain(A) = I(S_1, S_2, \ldots, S_n) - E(A)
\]

(3)

2.2. Naive Bayes method
This method is based on probability theory and Bayes theorem. He creates a model in the form of probabilities. Therefore, the division into classes is calculated as a product of probabilities [17]:

\[
\arg \max P(v_j) \prod_{i=1}^{n} P(a_i | v_j)
\]

(4)

2.3. Generalized Linear
The Generalized Linear Model (GLM) allows you to generalize simple linear regression to response variables that have error distribution patterns other than normal distribution, such as a Gaussian distribution. The generalized linear model assumes that each result of the dependent variables is formed
from a known distribution in an exponential family, a large probability distribution that includes the usual, binomial, Poisson and gamma distributions[1].

This model consists of 3 components:

- Exponential family of probability distributions;
- Linear predictive $\mu = X\beta$;
- connection function to $g(E(Y | X) = \mu = g^1(\eta)$.

2.4. Gradient Boosted Trees
This method is formalized in the algorithm “AdaBoost” of Freund and Schapire, originally applied to classification problems. Boosting works by sequentially applying weak learners to repeatedly re-weighted versions of the training data. After each boosting iteration, misclassified examples have their weights increased, and correctly, classified examples their weights decreased. Hence, each successive classifier focuses on examples that have been hard to classify in the previous steps[8].

2.5. Support Vector Machine (SVM)
Support Vector Machine (SVM) is a new learning algorithm developed on the basis of Statistical Learning Theory (STL) established by Vapnik and others. It is based on the VC dimension theory and the principle of structural risk minimization. Based on the limited sample information, we seek the best compromise between the complexity of the model and the learning ability, which greatly overcomes the problems of dimensional disaster and local minima in traditional machine learning, so as to obtain better generalization ability. SVM is used to solve classification problems, finding the hyper plane with the greatest separation space (delimited by its margins) between the categories of the chosen variable. The hyper plane is the equation separating the two classes, while the support vectors are the data points closest to the margin boundaries[9].

2.6. Deep Learning
Deep Learning ("deep learning") is a part of a family of machine learning methods based on artificial neural networks. Deep Learning is used in the following areas: bioinformatics, drug design, medical imaging analysis, material validation and development of programs and recommendations for board games[3] Deep Learning is a machine learning algorithm that uses multiple layers to step by step extract higher-level dependencies from raw data. The latent factor model for content-based recommendations uses Deep Learning to create features. The model uses a hybrid and content-oriented approach and serves to improve recommendations in several tasks.

2.7. Random Forest
(Random Forest or Random Resolution Forest) is a learning method that uses mixed algorithms to achieve the best results for classification, regression, and other problems. This method is commonly used during training to build many decision trees and solve problems such as grades (classification) or mean / mean guess. Random forests are modifications of harvested trees that create a large collection of trees, built to further improve predictive performance[5]. The sequence of this method is as follows.

- Step 1. An arbitrary $T_d$ training sample is extracted from the $T_d$ data.
- Step 2. A tree called $R_d$ forms $T_d$ as a root.
- Step 3. Forming the $N_i$ tree.
- Step 4. If $N_i$ consists of only one instance of the class, then [6] is returned.

Otherwise $N_i$ randomly chooses the possible $p\%$ of the node (fission properties) inside. For more information, when dividing, a function called F. is selected. For node trees $N_i$, $N_{i1}$, ..., $N_{if}$, functional
nodes $F$ are created. Where $F_i$ are possible values $F_{i1} F_{i2}$. Then the contents of $N_i$ are replaced with $T_{Dj}$. $T_{Dj}$ is considered as the set $N_i$ corresponding to $F$.

3. Proposal method

The proposal method. Since the founding of the university, several services have been introduced for students, including the possibility of students choosing subjects within the selected disciplines, the ability to choose teachers for the subjects. There are several ways to organize this. These include the organization of trial lessons for 2 weeks when choosing a teacher. Of course, this is a process that requires additional effort, time and resources. Therefore, it is advisable to conduct surveys of students about their teacher, receive feedback and use the results of their analysis. The organization of the questionnaire in the form of numerical scores simplifies the process of its analysis. But providing students with grades about teachers in the form of text, then analyzing their textual information using data mining techniques, further enhances the quality. The accumulation of such information over the years will be an important factor in creating a quality educational environment. The large size of the data creates difficulties for simple methods of data analysis, therefore, it is advisable to use the knowledge gained through the mining of large amounts of data. The model analyzes the result under which methodology the quality indicator is considered better. The proposed method is depicted in figure 3 below:

![TUIT-LMS Model of knowledge and recommendations development based on intellectual analysis of data](image-url)
4. Results and discussion
Data collection is the first step from various sources; the data is preprocessed and fills the empty columns. For implementation data is loaded or imported into Rapid Miner. Then the data is divided into training and testing data set in 80% and 20% ratio. Finally, the predicted outcome will be produced. The model learns from the data if new data comes the model predict the result. The process starts at data collection and ends at prediction of knowledge. The student evaluation data have 10 attributes in which 10 independent variables (from Q1 to Q10). A total of 1,970 students participated in the survey and surveyed teachers who taught each subject. Given that there are an average of seven subjects, there will be 13,790 survey results. This total accounts for 137,900.

| Variable | Description | Possible Values |
|----------|-------------|-----------------|
| Q1       | Teacher can be interested in his subject. | \{1,2,3,4,5\} |
| Q2       | The teacher has prepared for each lesson. | \{1,2,3,4,5\} |
| Q3       | The teacher has a deep knowledge of science. | \{1,2,3,4,5\} |
| Q4       | The teacher used a variety of pedagogical methods to achieve the learning objectives | \{1,2,3,4,5\} |
| Q5       | The teacher was able to direct the students' activity effectively. | \{1,2,3,4,5\} |
| Q6       | The teacher was able to make time for the students outside of class. | \{1,2,3,4,5\} |
| Q7       | The teacher asked questions and answered them. | \{1,2,3,4,5\} |
| Q8       | Do you understand the teacher's instructions for working with tasks | \{1,2,3,4,5\} |
| Q9       | The teacher provided the students with literature on each topic | \{1,2,3,4,5\} |
| Q10      | Overall level of training. | \{1,2,3,4,5\} |

Rapid Miner was chosen as the testing environment for the above methods. Rapid Miner is an open source software tool for IAD(intellectual analysis data) tasks. It is an integrated system that does not select an operating system and can integrate with R products, and has a user-friendly graphical interface. Data analysis processes are described in graphical form with a breakdown into modules[18]. The test results are shown in table 2

| Methods               | Accuracy | Training time (ms) |
|-----------------------|----------|--------------------|
| Generalized Linear Model | 85.9     | 172                |
| Deep Learning         | 78.8     | 635                |
| Decision Tree         | 81.9     | 252                |
| Random Forest         | 86.5     | 96                 |
| Naive Bayes           | 87.7     | 16                 |
| Gradient Boosted Trees | 84.4     | 508                |
| Support Vector Machine | 76.5     | 18546              |

Based on the above results, the number of incoming objects only affects the accuracy of the result. This can be represented in the diagram as shown in figure 4. The time difference traveled to training the data set in each model is shown in the figure 5.
5. Conclusion
In conclusion, the diagram above shows that in the proposed model, the Generalized Linear Model, Naive Bayes, Gradient Boosted Trees and Random Fores methods are close to each other and show higher reliability, but the time spent on training data set is Generalized Linear Model 1.8 times more and spends 10.75 times more time than Naive Bayes. On the other hand, Gradient Boosted Trees spend 3 times less time. When using the Deep Learning, Trees, Decision Tree and Support Vector Machine method, the result reliability turns out to be less than the above. The worst outcome remains the Support Vector Machine method in terms of accuracy and training data set time on the data being taught. If we want to develop recommendation systems for the education system, it makes sense to use the Generalized Linear Model, Naive Bayes, Gradient Boosted Trees and Random Fores methods. The most reliable and time-consuming of these is the Naive Bayes method.

References
[1] Van Barneveld et al. 2012 Analytics in higher education: Establishing a common language Educause Learning Initiative
[2] Romero C et al. 2008 Data mining algorithms to classify students Proceedings of Educational Data Mining (Montréal, Québec, Canada)
[3] Schmidhuber J 2015 Deep Learning in Neural (Networks: An Overview Neural Networks) 61
[4] Yann LeCun et al. 2015 Deep Learning Nature 521(7553) 436-44
[5] Marvin N W and Andreas Z 2017 Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R Journal of Statistical Software Articles 77(1)
[6] Philipp R, Marvin N and Anne L 2019 Hyperparameters and tuning strategies for random forest WIREs Data Mining and Knowledge Discovery 9(3)
[7] Vijayalakshmi V, Panimalar K and Janarthanan S 2020 Predicting the performance of instructors using Machine-learning algorithms High Technology Letters 26(12)
[8] García E, Romero C, Ventura S and de Castro C 2011 A collaborative educational association rule mining tool The Internet and Higher Education 14(2)
[9] Kamilov M M et al. 2019 The method of applying the algorithm of calculating grades for finding similar diagnostics in medical information systems International Journal of Innovative Technology and Exploring Engineering 8(6)
[10] Villanueva A, Moreno L G and Salinas M J 2018 Data mining techniques applied in educational environments: Literature review Digital Education Review 33
[11] Christopher K, Do Xuan Anh and Nicolas H 2016 Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500 FAU Discussion Papers in Economics 03
[12] Mateus Maia et al. 2020. Convolutional Support Vector Models: Prediction of Coronavirus Disease Using Chest X-rays Information 11(548)
[13] Alfonso U, Antonino F, Massimo L R, Valentina R and Riccardo R. 2019 Data Mining Prediction
Methods Encyclopedia of Bioinformatics and Computational Biology (Elsevier) p 413

[14] Dekker G W, Pechenizkiy M and Vleeshouwers J M 2009 Predicting students drop out: a Case study Proceedings of the 2nd International Conference on Educational Data Mining

[15] Urazmatov T Q, Nurmetova B B and Kuzibayev X S 2020 Analysis of big data processing technologies IOP Conference Series: Materials Science and Engineering 862(4)

[16] Rakhimboev K J, Ismailov M A and Khalmuratov O U 2020 Parametric method for evaluation the state of complex organizational objects IOP Conference Series: Materials Science and Engineering 919(5)

[17] Fanbo Meng, Linying Xu 2018 An improved native bayes classifier for imbalanced text categorization based on k-means and chi-square feature selection Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control

[18] Rapid Miner Inc. Headquarters official web portal Retrieved from https://academy.rapidminer.com/