DOI 10.52209/1609-1825_2023_3_428 UDC 004.048

Using Machine Learning Models to Evaluate the Quality of Text Documents

[1]*KAIBASSOVA Dinara, PhD, Acting Associate Professor, dindgin@mail.ru,
[2]YUCHSHENKO Olessya, PhD, Cand. of Tech. Sci., Associate Professor, Head of EPD, olessyayuchenko@hotmail.com,
[1]MUKHAMETZHANOVA Bigul, PhD, Senior Lecturer, grek79@mail.ru,
[1]SAIMANOVA Zagira, PhD, Senior Lecturer, zagira@mail.ru,
[1]NURTAY Margulan, Master, Lecturer, solano.lifan2@bk.ru,
[1]NPJSC «Abylkas Saginov Karaganda Technical University», Kazakhstan, Karaganda, N. Nazarbayev Avenue, 56,
[2]Voronezh Institute of High Technologies, Russia, Voronezh, Lenin Street, 73a,
*corresponding author.

Abstract. This results of a study on the use of the machine learning method in the processing of text documents to assess the quality of students' written work are presented. Reviewing students' written work, such as exam essays or lab reports, can be quite a time-consuming process, although reviewing such as term papers has become a common time-consuming practice. One of the common problems that arise when evaluating text documents can be the low quality of the source data, which does not comply with the document formatting rules. During the experiment, data was obtained, consisting of several works for testing. The quality assessment indicators used in this study are the compliance of the explanatory note with the structural elements of the document. The best fit model for scoring has been validated against several machine learning algorithms and techniques such as k-nearest neighbors (kNN), support vector regression (SVR), and random forest. The result of the comparison shows that, having advantages in the learning process, kNN provided the highest prediction accuracy. The learning results of all algorithms are clearly presented in tables and graphic illustration.

Keywords: data analysis, text mining, kNN, SVR, Random Forest, quality assessment, educational content.

Introduction

The scientific analysis of education is based on a systematic approach to the educational process. The educational process is a complex, dynamically developing system of social activity. A properly designed information model of business processes can improve the efficiency and quality of their implementation [1]. The multicomponent model of the information educational environment of the university is understood as a system integration of all common requirements, components, information resources and technologies obtained in the course of research that affect the specifics and effectiveness of informatization of educational, control and measurement, extracurricular, research and organizational and managerial activities of institutions of higher education. In addition, it is relevant to develop formal models of knowledge representation that ensure the processing of scientific and educational information at the semantic level in the knowledge management system of the university.

The architecture of the learning achievement analysis environment can be represented as three levels.

The first level – the level of user interaction includes a content management system and additional services. The Content Management System (content management system) provides high-quality tools for creating a portal interface. The interface is designed to organize access to the portal functions.

The second level – the level of knowledge management – includes the business logic of the subject area, which is a set of described rules, principles, dependencies of the objects of the subject area.

The third level is the level of knowledge preservation, represented by a knowledge base based on ontology [2].

In this study, the most appropriate model is selected from several proposed machine learning algorithms for determining the correspondence of text documents to structural elements for assessing the quality of student written works. To do this, the results were tested based on several algorithms and Machine Learning methods, such as k-nearest neighbors (kNN), support vector regression (SVR) and random forest. Checking the text documentation of student work is a laborious process, and also requires a lot of labor. The overall assessment of the work consists of several indicators. One of the indicators is the design of the work in accordance with the requirements of text documents. Automation of this task significantly increases the efficiency of evaluation, as well as labor cost savings. In addition, checking the

work requires higher concentration for a longer period of time, which often leads to errors.

Textual information analysis technologies are changing rapidly under the influence of machine learning. Solving the problem of assessing the quality of text documents of student papers allows you to compare the opinion of a teacher with an alternative opinion of a specialist in the form of a machine learning model trained in a certain set of previous documents, previously evaluated by experts.

Materials and methods

Algorithms of machine learning and neural networks can solve the most complex problems of natural language processing, for example, [3] in the paper, the authors evaluate the quality of responses of students of higher educational institutions using a special algorithm based on the calculation of keywords from a certain data set, the length of the text and other parameters. The disadvantage of this approach is that it is based on statistical data, ignoring the semantic component of the assessment. In the following work [4], a method was proposed for sentimental analysis of students' opinions using machine learning algorithms, such as a reference vector machine, a polynomial naive Bayesian classifier and a random forest. In addition, the selected machine learning methods [5] were used by the authors of the work to develop a model for solving problems such as language learning, assessing the student's reading and writing skills.

The results of checking text documents affect the assessment that the student receives for the work performed (Laboratory report, term paper, course project, etc.). This forecasting problem is the main task of private education and attracts more attention in the field of artificial intelligence and the production of educational data [6].

So, let's consider the concept of forecasting models when evaluating the work to provide the task. The presented model makes it possible to assess the presence in the structure of the work of sections defined by such features as introduction, bibliography, appendix, designations and abbreviations, etc., for a total of 10 features. Let's denote the data sets of machine learning models as follows:

$$D = \{d_1, d_2, ..., d_n\},$$

$$Y = \{y_1(d_1), y_2(d_2), ..., y_2(d_n)\},$$

where D is a set of characters;

Y_i is the evaluation of the d_i – the sign for a given job, n is the number of characters.

Then the evaluation method used in the proposed work has the following form:

$$Y = w_1x_1 + w_2x_2 + ... + w_nx_n, \quad (1)$$

where w is the weight;

x is the input matrix by all attributes, x_1, x_2, ..., x_n are the values of each instance.

As already mentioned, machine learning algorithms such as k-nearest neighbors (kNN), support vector regression (SVR) and random forest were chosen to evaluate students' written papers. Several factors formed the basis for the choice of kNN, SVR and Random Forest algorithms. Firstly, he noted that the amount of data sets available to the authors was limited, and, accordingly, the approach based on hidden semantic analysis was not relevant. An algorithm like KNN learns from a data set only when predicting, which allows you to easily add new data in the future. In the case of SVR, its choice is that it is based on efficiency in large spaces and optimal use of computing power. Random Forest was chosen because of its resistance to retraining with an increase in the number of labels or the amount of data. In addition, it is better to note that they all have the same quality as the ability to learn in small datasets. A regression model is a function that is a comparison between input variables and output variables. The regression problem allows you to choose a function: choose the curve of the function that best matches the known data and predicts the unknown data better. Regression methods are often used to predict continuous grades of students in certain courses [7-8].

Thus, support vector regression is a type of SVM that uses the space between data points as an error and predicts the most likely next point in the dataset. SVR gives us the flexibility to determine how much error is allowed in the sample, and finds the appropriate line (or hyperlight in higher dimensions) that matches the data. The objective function of the SVR is to reduce the coefficients – more precisely, the L2-norm of the coefficient vector – and not the quadratic error. Instead, the error term is handled in constraints, where we set the absolute error to be less than or equal to the specified value, called the maximum epsilon error. We can adjust the Epsilon to get the desired accuracy of our model. Our new objective function (2) and our constraints (3) are as follows:

$$MIN\frac{1}{2}\| w \|^2. \quad (2)$$

Limitations:

$$|y_i - w_ix_i| \le \varepsilon. \quad (3)$$

Random Forest is a method that combines the predictions of several machine learning algorithms to create more accurate predictions than any single model. The algorithm generates a set of assumptions using «decision tree» methods, and then outputs the average value from all of them. Compared to other machine learning methods, the theoretical part of the Random Forest algorithm is simple. All he needs is the formula (4) of the final classifier $a(x)$, which looks like this:

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} b_i(x), \quad (4)$$

where is the N – number of trees;

b – resolute tree;

x – a sample created based on the data provided.

So, the Random Forest method is a universal machine learning algorithm with a teacher. It can be used in various tasks, but it is mainly used in classification and regression tasks. The ability to efficiently process data with multiple labels and classes;

Sensitivity to any monotonic transformation of feature values;

Continuous and discrete labels are equally well handled;

There are methods for assessing the significance of individual features;

Internal evaluation of the generalizing ability of the model;

High parallelization and scaling;

Random forests are very flexible and have very high accuracy.

To assess the quality of all three models, indicators such as Mean Absolute Error, Mean Squared Error, mean Absolute percentage Error and Accuracy were used. The formulas for calculating these indicators (5)-(9) are as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \overline{y_i}|, \quad (5)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2, \quad (6)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y_i})^2}, \quad (7)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \overline{y_i}|}{y_i}, \quad (8)$$

$$Accuracy = (1 - MAPE) * 100\%. \quad (9)$$

Research results

The experimental work was carried out with 285 documents, which were divided in the ratio of 80% – training and 20% – test sample. To do this, the train_test_split function of the Python scikit-learn library was used. The fragment of the received data has the following form, shown in Figure 1:

The application of the KNN algorithm was carried out with the choice of the value of k in the range from 1 to 20. The RMS error (RMSE) was chosen to determine the accuracy of the trained model. The relationship between the values of K and RMSE is further shown on the error curve in Figure 2.

The graphs in which the results of the forecasts in the test sample are compared with the actual values are shown in Figure 3.

This curve is commonly known as the elbow method. The results of the RMS calculation showed that the optimal value is k = 5. Using the optimal parameter selection tool for machine learning of the GridSearchCV model led to the same conclusion. When using the Random forest algorithm, the number of trees was equal to 100, the maximum depth of each tree was limited to 3 values. The Support Vector

file	introduction_exists	references_exists	application_exists	designations_exists	pictures_count	tables_count	foreword	usage_scope	norm_links	terms
doc2.pdf	0	1	1	0	0	0	1200	50	64	675
doc172.pdf	0	1	1	1	1	0	4812	0	0	4537
doc120.pdf	0	0	1	1	3	0	2913	140	1189	2598
doc11.pdf	0	1	0	1	4	0	1934	74	63	1458
doc163.pdf	0	1	1	1	3	0	1850	402	164	1578
doc18.pdf	0	0	0	1	0	0	636	34	27	0
doc101.pdf	0	0	1	1	6	0	4935	463	1933	4662
doc83.pdf	0	0	1	1	9	0	4928	222	4467	4784
doc44.pdf	0	0	1	0	0	0	651	0	0	0
doc177.pdf	0	1	1	1	1	0	6796	303	5908	6482
doc15.pdf	0	0	1	1	11	0	3004	186	1864	3129
doc21.pdf	0	0	0	0	7	0	1988	67	422	1604
doc131.pdf	0	1	1	0	3	3	782	29	486	575
doc106.pdf	0	0	0	1	0	0	1629	77	248	1212
doc4.pdf	0	1	1	0	0	0	1466	133	1126	1201
doc5.pdf	0	0	0	0	0	1	838	102	435	0

Figure 1 – A fragment of the received data

Regressor method used a regularization parameter with a value of 1.0 and epsilon 0.2 (formulas 2-3).

To assess the quality of work of all three models, such indicators as Mean Squared Error, mean Absolute Error, Mean Absolute percentage Error and Accuracy (according to formulas 5-9) were used. Their

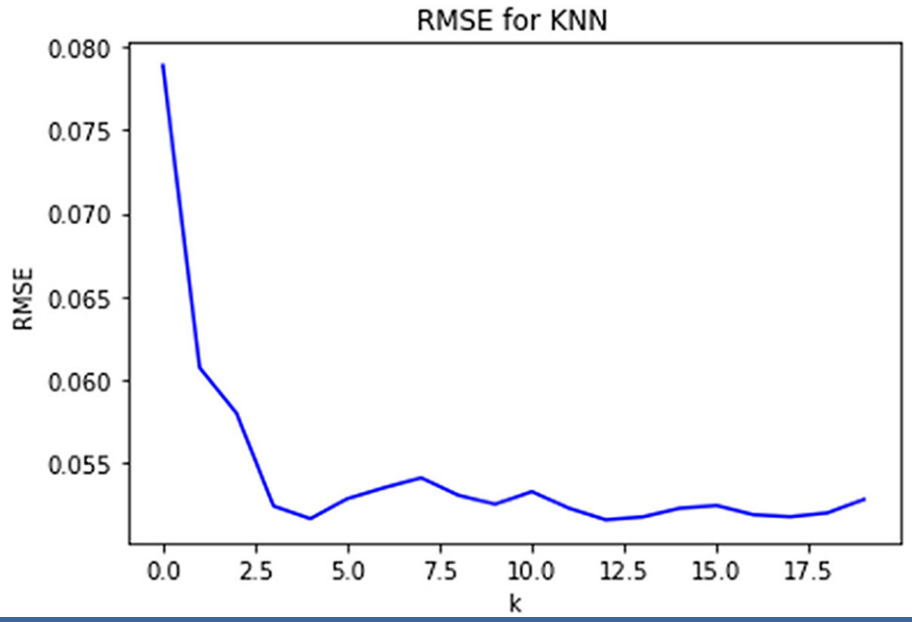
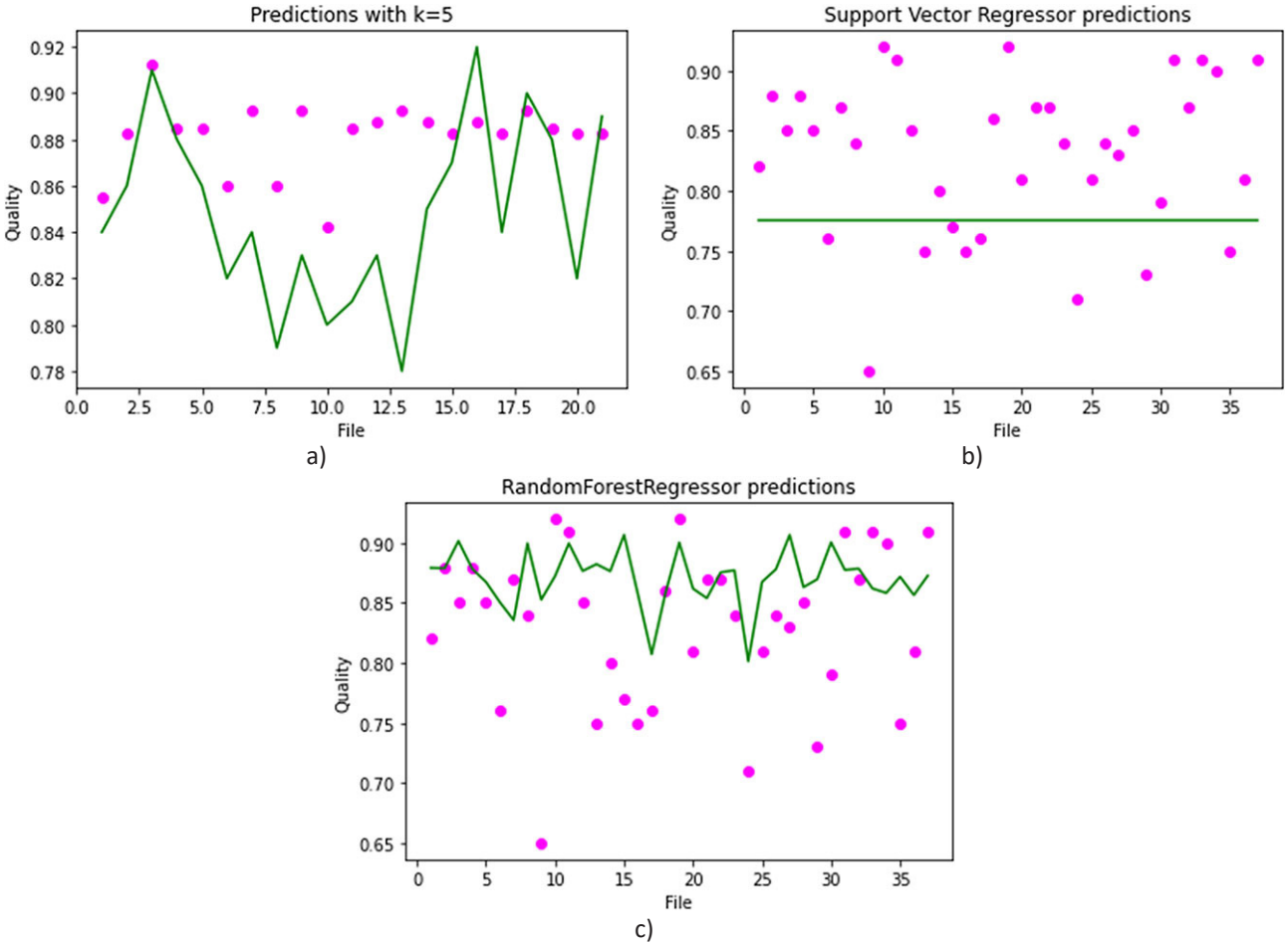


Figure 2 – Forecast error curve



a) kNN algorithm, b) SVR algorithm, c) Random Forest algorithm

Figure 3 – Real and predictable data comparison graphs when using algorithms

values are given in Table.

The results of the experimental work showed that each of the selected algorithms showed significant results in predicting the value of evaluating the quality of text documents. However, it should be noted that the choice of training parameters for algorithms such as Random Forest and SVR was carried out only by default, and not as a result of comparative analysis. Nevertheless, we managed to achieve quite good results.

Values of quality indicators by algorithms

Metric	Algorithms		
	kNN	SVR	Random Forest
MAE	0.038	0.073	0.056
MSE	0.002	0.007	0.005
RMSE	0.047	0.084	0.073
MAPE	4.571	8.653	7.294
Accuracy	95.43	91.35	92.71

Conclusion

Thus, in conclusion, the following conclusions can be drawn:

- the possibility of using machine learning algorithms and data analysis to assess the quality of academic work in terms of reducing labor costs. In addition, such use provides educational institutions with transparency and objectivity of the academic process;

- a step in assessing the quality of text documents, the kNN algorithm showed the best results, accurately indicating the minimum error values.

In the future, it is planned to conduct a study of assessing the quality of the content of documents using semantic analysis. It is planned that a text corpus obtained during the processing of a data set will be used for this, with subsequent expansion of its volume.

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19677319).

REFERENCES

1. Kaibassova D., Nurtay M. (2022). The comparative analysis of Machine Learning models for quality assessment of textual academic works // 2022 International Conference on Smart Information Systems and Technologies (SIST-2022).
2. Yavorskiy V., Kaibassova D., Klyuyeva Y. (2022). Issues of developing measures to analyze storage medium for educational achievements of students // 2022 IEEE 7th International Energy Conference (ENERGYCON 2022).
3. Bharadia, Sharad & Sinha, Prince & Kaul, Ayush. (2018). Answer Evaluation Using Machine Learning.
4. D. Dsouza, Deepika, D. Nayak, E. Machado, Adesh N.D. (2019). «Sentimental Analysis of Student Feedback using Machine Learning Techniques», International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-1S4, June 2019, pp. 986-991.
5. Maqsood S., Shahid A., Tanvir Afzal M., Roman M., Khan Z., Nawaz Z., Aziz MH. (2022). Assessing English language sentences readability using machine learning models. // PeerJ Computer Science 8:e818.
6. Kaibassova D., Sagatbekova D., Sagatbekova A. (2021). «Applying cluster analysis of educational content for identifying similar documents», Bulletin Abai KazNPU, the series of «Physical and Mathematical sciences» 76, 4 (Dec. 2021), 162-167. https://doi.org/10.51889/2021-4.1728-7901.22
7. Polyzou, A., and Karypis, G. (2016). Grade prediction with models specific to students and courses. Int. J. Data Sci. Anal. 2, 159-171. https://doi.org/10.1007/s41060-016-0024-z
8. Hu, Q., Polyzou, A., Karypis, G., and Rangwala, H. (2017). «Enriching coursespecific regression models with content features for grade prediction», in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (Tokyo: IEEE), 504-513.

Мәтіндік құжаттардың сапасын бағалау үшін Machine Learning модельдерін қолдану

[1]***КАЙБАСОВА Динара Женисбековна,*** *PhD, доцент м.а., dindgin@mail.ru,*
[2]***ЮЩЕНКО Олеся Александровна,*** *PhD, т.ғ.к., доцент, РББ бастығы, olessyayuchenko@hotmail.com,*
[1]***МУХАМЕТЖАНОВА Бигуль Олжабаевна,*** *PhD, аға оқытушы, grek79@mail.ru,*
[1]***САЙМАНОВА Загира Бекетаевна,*** *PhD, аға оқытушы, zagira@mail.ru,*
[1]***НҰРТАЙ Марғұлан,*** *магистр, оқытушы, solano.lifan2@bk.ru,*
[1]*«Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КеАҚ, Қазақстан, Қарағанды, Н. Назарбаев даңғылы, 56,*
[2]*Воронеж жоғары технологиялар институты, Ресей, Воронеж, Ленин көшесі, 73а,*
**автор-корреспондент.*

Аңдатпа. *Студенттердің жазба жұмыстарының сапасын бағалау үшін мәтіндік құжаттарды өңдеуде Machine Learning әдісін пайдалану бойынша зерттеу нәтижелері келтірілген. Студенттердің жазба жұмыстары, мысалы, емтихандар бойынша эссе сұрақтары немесе зертханалық жұмыстардың есептері, курстық жобалар/жұмыстар сияқты тексеру әдеттегі тәжірибеге айналғанымен, айтарлықтай уақытты қажет ететін үдеріс болуы мүмкін. Мәтіндік құжаттарды бағалау кезінде туындайтын жалпы мәселелердің бірі*

құжатты пішімдеу ережелеріне сәйкес келмейтін бастапқы деректер сапасының төмендігі болуы мүмкін. Эксперимент барысында тестілеу үшін бірнеше жұмыстардан тұратын деректер алынды. Осы зерттеуде қолданылатын сапаны бағалау көрсеткіштері түсіндірме жазба құжаттарының құрылымдық элементтеріне сәйкестігі болып табылады. Бағалау үшін қолайлы ең жақсы үлгіні k-ең жақын көршілер (kNN), тірек векторларының регрессиясы (SVR) және кездейсоқ орман сияқты бірнеше Machine Learning алгоритмдері мен әдістері негізінде нәтижелері тексерілді. Салыстыру нәтижесі оқу процесінде артықшылықтарға ие бола отырып, kNN ең жоғары болжау дәлдігін бергенін көрсетеді. Барлық алгоритмдердің оқу нәтижелері кестелерде және графикалық иллюстрацияларда анық көрсетілген.

Кілт сөздер: *деректерді талдау, мәтіндік деректерді өңдеу, kNN, SVR, Random Forest, сапаны бағалау, білім мазмұны.*

Использование моделей машинного обучения для оценки качества текстовых документов

[1]***КАЙБАСОВА Динара Женисбековна,*** *PhD, и.о. доцента, dindgin@mail.ru,*
[2]***ЮЩЕНКО Олеся Александровна,*** *PhD, к.т.н., доцент, начальник РИО, olessyayuchenko@hotmail.com,*
[1]***МУХАМЕТЖАНОВА Бигуль Олжабаевна,*** *PhD, старший преподаватель, grek79@mail.ru,*
[1]***САЙМАНОВА Загира Бекетаевна,*** *PhD, старший преподаватель, zagira@mail.ru,*
[1]***НҰРТАЙ Марғұлан,*** *магистр, преподаватель, solano.lifan2@bk.ru,*
[1]*НАО «Карагандинский технический университет имени Абылкаса Сагинова», Казахстан, Караганда, пр. Н. Назарбаева, 56,*
[2]*Воронежский институт высоких технологий, Россия, Воронеж, ул. Ленина, 73а,*
**автор-корреспондент.*

Аннотация. *Представлены результаты исследования по использованию метода машинного обучения при обработке текстовых документов для оценки качества письменной работы студентов. Проверка письменных работ студентов, таких как эссе на экзаменах или отчеты по лабораторным работам, могут быть довольно трудоемким процессом, хотя проверка, такая как курсовые проекты/работы, стала обычной времязатратной практикой. Одной из распространенных проблем, возникающих при оценке текстовых документов, может быть низкое качество исходных данных, не соответствующее правилам форматирования документа. В ходе эксперимента были получены данные, состоящие из нескольких работ для тестирования. Показателями оценки качества, используемые в этом исследовании, является соответствие пояснительной записки структурным элементам документа. Наилучшая модель, подходящая для оценки, была проверена на основе нескольких алгоритмов и методов машинного обучения, таких как k-ближайших соседей (kNN), регрессия опорных векторов (SVR) и случайный лес. Результат сравнения показывает, что, имея преимущества в процессе обучения, kNN обеспечил наивысшую точность прогнозирования. Результаты обучения всех алгоритмов наглядно представлены в таблицах и графических иллюстрациях.*

Ключевые слова: *анализ данных, обработка текстовых данных, kNN, SVR, Random Forest, оценка качества, образовательный контент.*

REFERENCES

1. Kaibassova D., Nurtay M. (2022). The comparative analysis of Machine Learning models for quality assessment of textual academic works // 2022 International Conference on Smart Information Systems and Technologies (SIST-2022).
2. Yavorskiy V., Kaibassova D., Klyuyeva Y. (2022). Issues of developing measures to analyze storage medium for educational achievements of students // 2022 IEEE 7th International Energy Conference (ENERGYCON 2022).
3. Bharadia, Sharad & Sinha, Prince & Kaul, Ayush. (2018). Answer Evaluation Using Machine Learning.
4. D. Dsouza, Deepika, D. Nayak, E. Machado, Adesh N.D. (2019). «Sentimental Analysis of Student Feedback using Machine Learning Techniques», International Journal of Recent Technology and Engineering (IJRTE), Volume-8, Issue-1S4, June 2019, pp. 986-991.
5. Maqsood S., Shahid A., Tanvir Afzal M., Roman M., Khan Z., Nawaz Z., Aziz MH. (2022). Assessing English language sentences readability using machine learning models. // PeerJ Computer Science 8:e818.
6. Kaibassova D., Sagatbekova D., Sagatbekova A. (2021). «Applying cluster analysis of educational content for identifying similar documents», Bulletin Abai KazNPU, the series of «Physical and Mathematical sciences» 76, 4 (Dec. 2021), 162-167. https://doi.org/10.51889/2021-4.1728-7901.22
7. Polyzou, A., and Karypis, G. (2016). Grade prediction with models specific to students and courses. Int. J. Data Sci. Anal. 2, 159-171. https://doi.org/10.1007/s41060-016-0024-z
8. Hu, Q., Polyzou, A., Karypis, G., and Rangwala, H. (2017). «Enriching coursespecific regression models with content features for grade prediction», in 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA) (Tokyo: IEEE), 504-513.