Scientific article UDC 330.47 DOI: https://doi.org/10.57809/2023.2.4.7.6

# APPLYING MACHINE LEARNING METHODS IN ELECTRONIC DOCUMENT MANAGEMENT SYSTEMS

#### Laylo Abdukhalilova<sup>1</sup>, Oksana Iliashenko<sup>2</sup> 💿 , Dayana Alchinova<sup>2</sup> 💿 🖾

<sup>1</sup> Tashkent State Transport University, Tashkent, Uzbekistan; <sup>2</sup> Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

## <sup>⊠</sup> alchinovadayana@gmail.com

**Abstract.** This article discusses methods and schemes for using machine learning in automating document management business processes. The scenarios described in the article can be useful for companies involved in document workflow automation or related areas (for example, email services) or ECM systems in general, and represent a generalization of the experience of specialists in the use of machine learning methods in document management. Several machine learning models used in business process automation solutions in practice are also considered. The results of the study, based on the analysis of all the points considered in the article, identified the main possible areas of development that arise when using machine learning models in electronic document management systems, which will be useful for data scientists developing such areas of AI, as machine learning.

Keywords: business processes, documentation support for management, machine learning, electronic document management systems, data analytics

**Citation:** Abdukhalilova L., Ilyashenko O., Alchinova D. Applying machine learning methods in electronic document management systems. Technoeconomics. 2023. 2. 4 (7). 61–71. DOI: https://doi.org/10.57809/2023.2.4.7.6

This is an open access article under the CC BY-NC 4.0 license (https://creativecommons. org/licenses/by-nc/4.0/)

Научная статья УДК 330.47 DOI: https://doi.org/10.57809/2023.2.4.7.6

# ПРИМЕНЕНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В СИСТЕМАХ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

Лайло Абдухалилова<sup>1</sup>, Оксана Ильяшенко<sup>2</sup> 💿, Даяна Альчинова<sup>2</sup> 💿 🖾

<sup>1</sup> Ташкентский государственный транспортный университет, Ташкент, Узбекинстан; <sup>2</sup> Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Россия

□ alchinovadayana@gmail.com

Аннотация. В данной статье рассматриваются методы и сценарии применения машинного обучения в автоматизации бизнес-процессов документационного обеспечения управления. Описанные в статье сценарии могут быть полезны компаниям, занимающимся автоматизацией электронного документооборота или смежными областями (например, почтовые сервисы) или в целом ЕСМ-системами, и являются обобщением опыта авторов в области использования методов машинного обучения в системах электронного документооборота. Также будут рассмотрены несколько моделей машинного обучения, используемые в решениях автоматизации бизнес-процессов в практике. В результатах исследования на основе анализа всех рассмотренных пунктов статьи выведены основные возможные зоны развития, возникающих при использовании моделей машинного обучения в системах электронного документооборота, что будет полезно дата-сайентистам, развивающим такого направления искусственного интеллекта, как машинное обучение.

Ключевые слова: бизнес-процессы, документационное обеспечение управления, машинное обучение, системы электронного документооборота, аналитика данных

Для цитирования: Абдухалилова Л., Ильяшенко О., Альчинова Д. Применение методов машинного обучения в системах электронного документооборота // Техноэкономика. 2023. Т. 2, № 4 (7). С. 61–71. DOI: https://doi.org/10.57809/2023.2.4.7.6

Это статья открытого доступа, распространяемая по лицензии CC BY-NC 4.0 (https:// creativecommons.org/licenses/by-nc/4.0/)

### Introduction

Modern business, as one of the priority tasks of strategic planning, announces process automation to maintain the company's sustainability in an increasingly competitive environment. One of the breakthrough technologies in solving automation problems is artificial intelligence (AI) technologies, including machine learning methods. Machine learning is a branch of computer science that allows systems to learn and improve their performance based on previous experience or history.

Machine learning is based on an algorithmic approach that allows systems to learn from data definitions and the resulting knowledge to solve problems at the first moment in time. Any AI algorithm is based on a mathematical model of varying degrees of complexity. A mathematical model is a representation of a system, process or phenomenon using mathematical explanations and results.

The purpose of this study is to study existing methods of using machine learning models in automating business processes related to document management support. The article uses an analytical approach based on an analysis of relevant literature and sources to consider the prospects for using intelligent microservices in a system for increasing document flow (EDMS). Scientific articles, books and electronic resources related to the use of AI and microservices to obtain documentation support for management have been studied. In addition, a practical analysis of the results of research and implementation of intelligent microservices in the EDMS was carried out. The advantages and disadvantages of such a situation were explored, as well as issues of consideration, situation and depending on the quality of the data. The novelty of scientific research lies in the qualitatively new use of machine learning technologies to automate the processes of working with documents in the company's ecosystem by increasing the volume of routine operations performed by enterprise employees. The results of this study can help contracting companies implement EDMS, as well as improve the efficiency and effectiveness of management processes.

## **Materials and Methods**

Tasks of an electronic document management system that can be modernized using machine learning:

# 1. Data storage

The task of storing data in electronic document management systems includes classification and selection of an acceptable storage method depending on the type of data. Machine learning models created for automatic classification will allow you to achieve fast and efficient document search by building connections between system objects. The model will detect classes of documents that have high value and require timely consideration first (Gogoulou, 2019). By combining the model with the notification service, the system will send such documents to users so they can start working quickly. This scheme allows the system to adapt the business process to changing conditions and makes it possible to achieve greater flexibility and stability.

### 2. Data processing

In an electronic document management system, objects have their own attributes to determine how data is stored, for example, data type, field type, section, etc. In cases where new objects appear in the system, a machine learning model that has been trained to recognize new object attributes and building a connection between these attributes and processes that use objects with similar attributes will enable the system to increase the speed of processing large arrays of different types and types of data (Iliashenko, 2020). The presence of such a goal in the company's strategy as system scalability from one process of one company to an entire consortium will contribute to the improvement of such models.

#### 3. Data analysis

Analyzing data and making decisions based on this analysis is the main task of companies with a Data-driven approach. For such companies, the issue of using machine learning in data analysis will be relevant, since the effectiveness of using models is based on data on decisions made in existing successful cases, which will be embedded in the model, and which will serve as a model (teacher) for it (Devlin, 2019). The presence of such data in a company in an acceptable volume will allow the model to predict the consequences of decisions made with a high percentage of forecast quality. To understand the accuracy of the model, data scientists can increase the number of iterations of running these models. As with all decisions made, the consequences will not be immediately visible. Users of the system should remember this and not forget that ultimately, it is not the machine learning model that is responsible for the result, but the users themselves (Mogilko, 2020).

### 4. Generation of system objects

Neural networks, which are deep learning models, allow companies to minimize or completely eliminate the routine tasks of creating and working with objects. In relation to electronic document management systems, machine learning, namely deep learning, will be able to optimize the process of developing content and documents by generating memorized objects and their reproduction based on these objects of a system of the same type (Sambetbayeva, 2022). Users only need to configure the generation of individual sections of documents according to specified rules.

5. Data security

Similar to the point on data processing, the system includes document flow, using machine learning, you can effectively manage documents with confidential information and personal data of the company's employees and its counterparties. Machine learning models built to automatically identify such information will classify documents containing confidential information and automatically determine access rights to them. To understand how machine learning models work with text and can be applied in various company cases, let's consider the stages of computer processing of texts in natural language:

1. Definition of linguistic resources - collection of text data (tagged and/or untagged) for processing and, if necessary, various kinds of dictionaries, for example, thesauruses in a form suitable and convenient for machine processing.

2. Cleaning text data - bringing texts to a general form without losing elements that affect the result of processing.

3. Graphematic analysis, segmentation or tokenization - breaking each text into units that are subject to further processing (tokens).

4. Morphological analysis:

- determination of the part of speech (part-of-speech, POS) and morphological features of the word, such as gender, number, case, declension;

- lemmatization is the process of transforming word tokens into the initial form, which in the case of processing Russian-language text means the representation of nouns in the nominative case and singular, adjectives, like nouns, but in the masculine gender, verbs, gerunds and participles in the indefinite form of the verb;

- stemming is the process of bringing word tokens to a base by cutting off suffixes and endings.

5. Exclusion of stop words, or "noise words" – removal of tokens representing frequently used but meaningful words (prepositions, conjunctions, parts, interjections, introductory words, some verbs, pronouns) (Popova, 2019).

6. Vectorization – consideration of each token in the text of a small numerical value.

The result of the steps is a representation of the source texts, suitable for further processing in order to solve a higher-level problem, for example, a text classification problem. Further processing of the prepared texts is carried out on the basis of two approaches: engineering and based on machine learning.

### **Results and Discussion**

The engineering approach is to use rules compiled by linguists in the form of dictionaries, templates and other linguistic sources. Machine learning approaches use natural language texts as resources to build a mathematical model that can exhibit features in earlier samples of the data presented. Based on the identified features, the algorithm determines feature values for new data. Let's consider additional machine learning models:

Natural language processing (NLP) is an interdisciplinary field of computer science and linguistics. It's primarily about giving computers the ability to support and manipulate human language. It involves processing natural language datasets such as text or speech corpora using either rule-based or probabilistic machine learning approaches. The goal is a way to "understand" computer requirements, including the contextual subtleties of the language within them. The technology can then pinpoint the information and ideas contained in the documents, as well as categorize and organize the documents themselves (Cherkasov, 2018; Anufrieva, 2023).

To solve NLP problems, an engineering approach is being used less and less and increasingly based on machine learning. But the first one should not be neglected, since it is effectively "applicable to narrow subject areas with clear rules for naming significant objects and a small variety of required language constructs" (Ponomarev, 2023).

BERT model

The transfer learning approach represented by Bidirectional Encoder Representations from Transformers is a recent language representation model. BERT is designed to pre-train deep bidirectional representations to extract context-sensitive features from input text. These representations fall under the category of "embeddings," which is an important concept in the field of NLP (Ilyin, 2023).

The term embedding refers to fixed-length vector representations of text that are capable of encoding syntactic and semantic information. BERT embeddings can be successfully used for any NLP tasks such as language inference and name and object recognition.

What makes the BERT model powerful is that high accuracy is achieved without any specific network architecture at the fine-tuning level. The model consists of a sequence of embedding layers, transformers, and a prediction layer (Figure 1). Given the incredibly good performance of BERT embedding in capturing input context, it is expected that the size of the text completing the BERT model, with a conversation history encoding mechanism, can lead to a high-performance conversational machine model understanding (Obukhov, 2020).

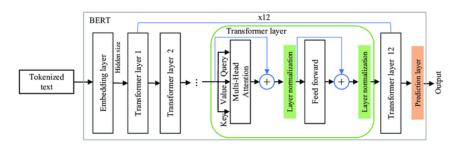


Fig. 1. BERT model architecture and its components

Model ELMo

The ELMo (Embeddings from Language Models) model takes context into account before embedding a word. This is how contextualized word-embeddings appeared (Davletov, 2023). Contextualized embeddings give words different vectors based on their semantics in the context of a sentence. Instead of using fixed word embeddings, ELMo looks at the entire sentence before assigning each word its embedding. It uses a bi-directional long short-term memory model (bi-directional LSTM), trained specifically for the task of creating such embeddings (Castelblanco, 2020).

As we train the model on a large data set, we begin to learn language constructs. ELMo creates contextualized embedding by grouping latent effects (other embedding) in this way (concatenation follows weighted addition) (Figure 2).

Let's look at the main scenarios for using machine learning.

Scenarios for the use of machine learning models largely depend on the implementation processes of management information support and EDMS components (Bolshakova, 2017).

The electronic document management system consists of the following elements:

- a scan of the document included in the system on paper;

- document registration card (DC), displaying information about the document, namely information about the sender, counterparty, subject of the document, date of sending, etc.;

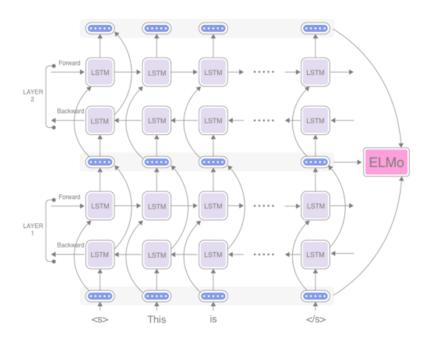


Fig. 2. ELMo model architecture

- information about the system user who works in the document's RC (Registrar) (Chernov, 2016).

Let's look at how machine learning models are used in the medium term and involve document flow.

1. Automatic routing of documents based on topic tagging can be represented in the form of the following script steps (Fig. 3):

1) maintaining a directory of thematic organization, where the responsible department of the employee is responsible for each topic;

2) highlighting semantic tags in the document text;

3) correlation of semantic tags with thematic ones (calculation of the probability of assignment, ordering by probability);

4) identifying the artist by tag.

The results of automatic routing of documents based on thematic tagging are the acceleration of delivery of the document to the department/performer, eliminating the processing of routine correspondence by the manager (Ivanovsky, 2021).

Examples of extracted entities are "Counterparty", "Related Documents and Instructions". *Results:* 

1. Related information on selected entities. For example, searching for correspondence with a selected counterparty.

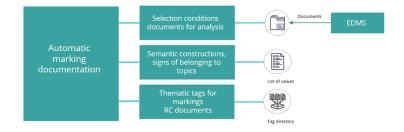


Fig. 3. Diagram of components and connections between them used in automatic routing

2. Automatic formation of links between documents.

In 15% of documents, on average, connections are not made at the registration stage. Scenario stages:

1) registration of an incoming document, attaching files;

2) recognition of file scans;

3) analysis of the text of files, search for data from related documents;

4) search for related documents in the EDMS;

5) establishing connections between documents.

Outcome: the ability to move to other system objects with a logical connection.

3. Automatic creation of an object in the system based on a document included in the system (Fig 4).

Scenario stages:

1) scanning an incoming document - transferring information into the system;

2) creation of the RC document;

3) text processing in the document;

4) automatic filling of document RC attributes.

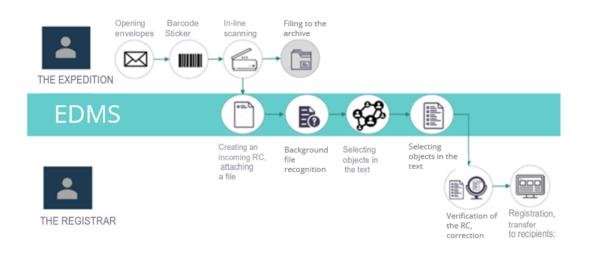


Fig. 4. Scheme of the process of automatically entering an incoming document into the system

# 4. PDF OCR

Optical Character Recognition is the process of processing and recognizing images or videos containing text content, as well as extracting text and layout of the information contained in them. OCR is most used to convert a non-editable PDF document file containing an image of the document into an editable Word document (Iliashenko, 2022).

Modern OCR systems use intelligent character recognition (ICR) technology to read text just like a human would (Figure 5). They use advanced machine learning techniques for human reading skills. A machine learning system called a neural network analyzes text at many levels by repeatedly processing the image. It looks for various image attributes (curves, lines, intersections and loops) and combines the results of different levels of analysis to produce a final result.

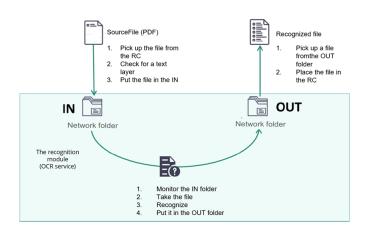


Fig. 5. Architecture of the OCR module in the document management system

# Conclusion

The areas of development that arise when using machine learning models in electronic document management systems are the following:

1. Enriching the base of customers who have a need to solve business cases. Improvement of supervised machine learning models occurs at the expense of existing data, based on which the model can be calibrated, since there are real "prediction" values.

2. Expanding the pool of datasets for conducting experiments. Machine learning models must run through several epochs and iterations. To do this, you initially need a large array of data, on the basis of which new and varied data can be generated.

3. Development of the IT infrastructure of companies, namely IT capacities. During the time spent training the model, new data appears in living systems that can make changes to the model. For such maneuvers, companies must have the necessary capacity.

Integration of machine learning technologies into electronic document management systems will give companies the opportunity to solve problems of effective work with documents, from reorganization in the file structure to provide direct access to information to automated decision-making on this document. Electronic document management systems that integrate artificial intelligence technologies will have a positive impact on the company's financial policy in the future. Every organization in the future will save money by replacing routine operations performed by staff with an ensemble of digital digital services using AI technologies. To effectively integrate machine learning into an EDMS, a company must define the goals it wants to achieve through this merger. In addition, the company must have qualified specialists to select machine learning models for its specific task, because training a model requires resources.

The use of machine learning methods in electronic document management systems can significantly improve the efficiency and accuracy of document processing and analysis processes, which helps improve the productivity and competitiveness of enterprises.

#### REFERENCES

Anufrieva V. 2023. Improvement of the system of working with suppliers at enterprises of the power engineering sector. Technoeconomics, 2, 3 (6), 69–78. DOI: https://doi.org/10.57809/2023.2.3.6.6

Bolshakova E. I., Vorontsov K. V., Efremova N. E., Klyshinsky E. S., Lukashevich N. V., Sapin A. S. 2017. Automatic processing of texts in natural language and data analysis: textbook, 269.

**Castelblanco A.** 2020. Machine learning techniques for identity document verification in uncontrolled environments: A case study. Lecture Notes in Computer Science, 12088, 271-281.

doi:10.1007/978-3-030-49076-8 26

Cherkasov D.Yu., Ivanov V.V. 2018. Machine learning. Science, technology and education, 5 (46).

**Chernov T.S., Ilyin D.A.** 2016. Investigation of methods of image segmentation of text blocks of documents using structural analysis and machine learning algorithms. Bulletin of the Russian Foundation for Basic Research, 4(92), 55-71. doi:10.22204/2410-4639-2016-092-04-55-71

**Davletov A. R.** 2023. Modern methods of machine learning and OCR technology for automation of document processing. Bulletin of Science, 5 (67), 676-698.

Devlin J., Chang M.-W., Lee K., Toutanova K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

**Gogoulou E.** 2019. Using Bidirectional Encoder Representations from Transformers for Conversational Machine Comprehension. KTH Royal Institute Of Technology, 78.

**Iliashenko O.Yu., Iliashenko V.M., Bikkulova Z.U.** 2020. Artificial intelligence as a tool with an observable process of providing medical care. In the collection: System analysis in design and management. collection of scientific papers of the XXIV International Scientific and Educational-Practical Conference, 272-284.

**Iliashenko V.M.** 2022. Machine learning technologies in the oil and gas industry. In the collection: Increasing the managerial, economic, social and innovative-technical potential of enterprises, industries and national economic complexes. Collection of articles of the XIII International Scientific and Practical Conference, 116-119.

**Ilyin I.V., Ilyashenko O.Yu., Devezas T.S., Shchenikov E.M.** 2023. Forecasting electricity consumption using machine learning methods. Ecology, environmental protection, carbon neutrality and development. Collection of proceedings of the Chinese-Russian ASRTU Symposium, 36-38.

**Ivanovsky N. I.** 2021. Application of machine learning algorithms for classification of documentation. Vestnik VNIIDAD, 2, 35-39.

Mogilko D.Y., Ilin I.V., Iliashenko V.M., Svetunkov S.G. 2020. BI capabilities in a digital enterprise business process management system. Lecture Notes in Networks and Systems, 95, 701-708.

**Obukhov A.** 2020. Algorithm of adaptation of electronic document management system based on machine learning technology. Progress in Artificial Intelligence, 9 (4), 287-303. doi:10.1007/s13748-020-00214-2

**Ponomarev O.V., Svetunkov S.G.** 2023. Economics as a science: an explanatory dictionary of general scientific terms and concepts: textbook, 214.

**Popova E. S.** 2019. Using artificial neural networks to solve text classification problems. Proceedings of the 29th international conference on computer graphics and vision "Graphicon", 270–273.

Sambetbayeva M., Kuspanova I., Yerimbetova A., Serikbayeva S., Bauyrzhanova S. 2022. Development of intelligent electronic document management system model based on machine learning methods. Eastern-European Journal of Ente rprise Technologies, 1 (2 (115), 68–76.

How transfer learning came to NLP. URL: https://habr.com/ru/articles/487358/ (accessed: 12.07.23).

Optical character recognition. URL: https://en.wikipedia.org/wiki/Optical\_character\_recognition (accessed: 12.07.23).

What is Optical Character Recognition (OCR): Overview and Use Cases. URL: https://www.labellerr.com/blog/optical-character-recognition-overview/ (accessed: 13.07.23).

What are the advantages of Optical Character Recognition? URL: https://makemeanalyst. com/what-are-the-advantages-of-optical-character-recognition/ (accessed: 14.07.23).

### СПИСОК ИСТОЧНИКОВ

Anufrieva V. 2023. Improvement of the system of working with suppliers at enterprises of the power engineering sector. Technoeconomics, 2, 3 (6), 69–78. DOI: https://doi.org/10.57809/2023.2.3.6.6

Bolshakova E. I., Vorontsov K. V., Efremova N. E., Klyshinsky E. S., Lukashevich N. V., Sapin A. S. 2017. Automatic processing of texts in natural language and data analysis: textbook, 269.

**Castelblanco A.** 2020. Machine learning techniques for identity document verification in uncontrolled environments: A case study. Lecture Notes in Computer Science, 12088, 271-281. doi:10.1007/978-3-030-49076-8\_26

**Cherkasov D.Yu., Ivanov V.V.** 2018. Machine learning. Science, technology and education, 5 (46).

**Чернов Т.С., Ильин Д.А.** 2016. Исследование методов сегментации изображений текстовых блоков документов с помощью алгоритмов структурного анализа и машинного обучения. Вестник Российского фонда фундаментальных исследований, 4(92), 55-71. doi:10.22204/2410-4639-2016-092-04-55-71

Давлетов А. Р. 2023. Современные методы машинного обучения и технология ОСК для автоматизации обработки документов. Вестник науки, 5 (67), 676-698.

Devlin J., Chang M.-W., Lee K., Toutanova K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

**Gogoulou E.** 2019. Using Bidirectional Encoder Representations from Transformers for Conversational Machine Comprehension. KTH Royal Institute Of Technology, 78.

**Iliashenko O.Yu., Iliashenko V.M., Bikkulova Z.U.** 2020. Artificial intelligence as a tool with an observable process of providing medical care. In the collection: System analysis in design and management. collection of scientific papers of the XXIV International Scientific and Educational-Practical Conference, 272-284.

**Iliashenko V.M.** 2022. Machine learning technologies in the oil and gas industry. In the collection: Increasing the managerial, economic, social and innovative-technical potential of enterprises, industries and national economic complexes. Collection of articles of the XIII International Scientific and Practical Conference, 116-119.

**Ilyin I.V., Ilyashenko O.Yu., Devezas T.S., Shchenikov E.M.** 2023. Forecasting electricity consumption using machine learning methods. Ecology, environmental protection, carbon neutrality and development. Collection of proceedings of the Chinese-Russian ASRTU Symposium, 36-38.

**Ивановский Н. И.** 2021. Применение алгоритмов машинного обучения для классификации документации. Вестник ВНИИДАД, 2, 35-39.

Mogilko D.Y., Ilin I.V., Iliashenko V.M., Svetunkov S.G. 2020. BI capabilities in a digital enterprise business process management system. Lecture Notes in Networks and Systems, 95, 701-708.

**Obukhov A.** 2020. Algorithm of adaptation of electronic document management system based on machine learning technology. Progress in Artificial Intelligence, 9 (4), 287-303. doi:10.1007/s13748-020-00214-2

**Ponomarev O.V., Svetunkov S.G.** 2023. Economics as a science: an explanatory dictionary of general scientific terms and concepts: textbook, 214.

**Popova E. S.** 2019. Using artificial neural networks to solve text classification problems. Proceedings of the 29th international conference on computer graphics and vision "Graphicon", 270–273.

Sambetbayeva M., Kuspanova I., Yerimbetova A., Serikbayeva S., Bauyrzhanova S. 2022. Development of intelligent electronic document management system model based on machine learning methods. Eastern-European Journal of Ente rprise Technologies, 1 (2 (115), 68–76.

How transfer learning came to NLP. URL: https://habr.com/ru/articles/487358/ (accessed: 12.07.23).

Optical character recognition. URL: https://en.wikipedia.org/wiki/Optical\_character\_recognition (accessed: 12.07.23).

What is Optical Character Recognition (OCR): Overview and Use Cases. URL: https://www.labellerr.com/blog/optical-character-recognition-overview/ (accessed: 13.07.23).

What are the advantages of Optical Character Recognition? URL: https://makemeanalyst. com/what-are-the-advantages-of-optical-character-recognition/ (accessed: 14.07.23).

# INFORMATION ABOUT AUTHORS / ИНФОРМАЦИЯ ОБ АВТОРАХ

**ABDUKHALILOVA Laylo T.** – professor, Candidate of Economic Sciences. E-mail: laylo.66@mail.ru **АБДУХАЛИЛОВА Лайло Тухтасиновна** – профессор, к.э.н. E-mail: laylo.66@mail.ru

ILIASHENKO Oksana Yu. – Associate Professor, Candidate of Pedagogic Sciences. E-mail: ioy12@yandex.ru ИЛЬЯШЕНКО Оксана Юрьевна – доцент, кандидат педагогических наук. E-mail: ioy12@yandex.ru ORCID: https://orcid.org/0000-0001-9292-7759

ALCHINOVA Dayana Yu.— business analyst. E-mail: alchinovadayana@gmail.com АЛЬЧИНОВА Даяна Юрьевна— бизнес-аналитик. E-mail: alchinovadayana@gmail.com ORCID: https://orcid.org/0009-0005-0963-4506

Статья поступила в редакцию 15.11.2023; одобрена после рецензирования 26.11.2023; принята к публикации 27.11.2023.

The article was submitted 15.11.2023; approved after reviewing 26.11.2023; accepted for publication 27.11.2023.