ISSN 2518-170X (Online)
ISSN 2224-5278 (Print)

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ
АКАДЕМИЯСЫ» РҚБ

«ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ» РҚБ

ХАБАРЛАРЫ

ИЗВЕСТИЯ

РОО «НАЦИОНАЛЬНОЙ
АКАДЕМИИ НАУК РЕСПУБЛИКИ
КАЗАХСТАН»

N E W S

OF THE NATIONAL ACADEMY
OF SCIENCES OF THE REPUBLIC
OF KAZAKHSTAN

SERIES

OF GEOLOGY AND TECHNICAL SCIENCES

3 (471)

MAY – JUNE 2025

THE JOURNAL WAS FOUNDED IN 1940

PUBLISHED 6 TIMES A YEAR

ALMATY, NAS RK

NAS RK is pleased to announce that News of NAS RK. Series of geology and technical sciences scientific journal has been accepted for indexing in the Emerging Sources Citation Index, a new edition of Web of Science. Content in this index is under consideration by Clarivate Analytics to be accepted in the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The quality and depth of content Web of Science offers to researchers, authors, publishers, and institutions sets it apart from other research databases. The inclusion of News of NAS RK. Series of geology and technical sciences in the Emerging Sources Citation Index demonstrates our dedication to providing the most relevant and influential content of geology and engineering sciences to our community.

Қазақстан Республикасы Ұлттық ғылым академиясы «ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы» ғылыми журналының Web of Science-тің жаңаланған нұсқасы Emerging Sources Citation Index-те индекстелуге қабылданғанын хабарлайды. Бұл индекстелу барысында Clarivate Analytics компаниясы журналды одан әрі the Science Citation Index Expanded, the Social Sciences Citation Index және the Arts & Humanities Citation Index-ке қабылдау мәселесін қарастыруда. Webof Science зерттеушілер, авторлар, баспашылар мен мекемелерге контент тереңдігі мен сапасын ұсынады. ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы Emerging Sources Citation Index-ке енуі біздің қоғамдастық үшін ең өзекті және беделді геология және техникалық ғылымдар бойынша контентке адалдығымызды білдіреді.

НАН РК сообщает, что научный журнал «Известия НАН РК. Серия геологии и технических наук» был принят для индексирования в Emerging Sources Citation Index, обновленной версии Web of Science. Содержание в этом индексировании находится в стадии рассмотрения компанией Clarivate Analytics для дальнейшего принятия журнала в the Science Citation Index Expanded, the Social Sciences Citation Index и the Arts & Humanities Citation Index. Web of Science предлагает качество и глубину контента для исследователей, авторов, издателей и учреждений. Включение Известия НАН РК. Серия геологии и технических наук в Emerging Sources Citation Index демонстрирует нашу приверженность к наиболее актуальному и влиятельному контенту по геологии и техническим наукам для нашего сообщества.

News of the National Academy of Sciences of the Republic of Kazakhstan. Series of geology and technology sciences.
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty).
The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. KZ39VPY00025420,** issued 29.07.2020.
Thematic scope: *geology, hydrogeology, geography, mining and chemical technologies of oil, gas and metals*
Periodicity: 6 times a year.
http://www.geolog-technical.kz/index.php/en/

«ҚР ҰҒА» РҚБ Хабарлары. Геология және техникалық ғылымдар сериясы».
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.).
Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 29.07.2020 ж. берілген **№ KZ39VPY00025420** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *Геология, гидрогеология, география, тау-кен ісі, мұнай, газ және металдардың химиялық технологиялары*
Мерзімділігі: жылына 6 рет.
http://www.geolog-technical.kz/index.php/en/

«Известия РОО «НАН РК». Серия геологии и технических наук».
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Собственник: Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **№ KZ39VPY00025420,** выданное 29.07.2020 г.
Тематическая направленность: *геология, гидрогеология, география, горное дело и химические технологии нефти, газа и металлов*
Периодичность: 6 раз в год.
http://www.geolog-technical.kz/index.php/en/

CONTENTS

Geology

NEWS of the National Academy of Sciences of the Republic of Kazakhstan
SERIES OF GEOLOGY AND TECHNICAL SCIENCES
ISSN 2224–5278
Volume 3. Number 471 (2025), 268–280

https://doi.org/10.32014/2025.2518-170X.525

UDC 681.518.5

© Y.A. Tynchenko[1,2], E.V. Khudyakova[3], V.V. Kukartsev[1,2,3], M.N. Stepancevich[3], A.A. Stupina [1,4,5], 2025.
[1]Siberian Federal University, Krasnoyarsk, Russia;
[2]Bauman Moscow State Technical University, Moscow, Russia;
[3]Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow, Russia;
[4]Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia;
[5]Siberian Fire and Rescue Academy of State Fire Service of the Ministry of Emergency Situations of Russia, Zheleznogorsk, Russia.
E-mail: vlad_saa_2000@mail.ru

FORECASTING THE CONTENT OF RARE EARTH ELEMENTS BASED ON GEOCHEMICAL DATA USING ENSEMBLE LEARNING METHODS

Yadviga A. Tynchenko — Siberian Federal University, Krasnoyarsk; Bauman Moscow State Technical University, Moscow, Russia, e-mail: 051301@mail.ru, https://orcid.org/0000-0001-8963-7830;
Elena V. Khudyakova — Doctor of Economics, Professor, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow, Russia, e-mail: evhudyakova@rgau-msha.ru, https://orcid.org/0000-0003-2156-7236;
Vladislav V. Kukartsev — Candidate of Technical Sciences, Associate Professor, Siberian Federal University, Krasnoyarsk, Russia; Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow, Russia; Bauman Moscow State Technical University, Moscow, Russia, e-mail: vlad_saa_2000@mail.ru, https://orcid.org/0000-0001-6382-1736;
Marina N. Stepancevich — Candidate of Economic Sciences, Associate Professor, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow, Russia, e-mail: stepancevich@rgau-msha.ru, https://orcid.org/0000-0002-7125-3027;
Alena A. Stupina — Dr., professor, Siberian Federal University, Krasnoyarsk, Russia; Reshetnev Siberian State University of Science and Technology, Krasnoyarsk, Russia; Siberian Fire and Rescue Academy of State Fire Service of the Ministry of Emergency Situations of Russia, Zheleznogorsk, Russia, e-mail: h677hm@gmail.com, https://orcid.org/0000-0002-5564-9267.

Abstract. This study addresses the challenge of predicting the content of rare earth elements — cerium (Ce), lanthanum (La), and neodymium (Nd)— in geological samples by applying ensemble machine learning methods to geochemical data. Given the growing industrial demand for these elements, efficient

prediction models can significantly enhance exploration strategies and reduce the need for costly field sampling. A geochemical dataset containing concentrations of major and trace elements was analyzed using a combination of algorithms, including XGBoost, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. The ensemble model was optimized through a combination of genetic algorithms and grid search, with XGBoost selected as the primary algorithm due to its high sensitivity to complex nonlinear feature-target dependencies. For classification tasks, the median concentrations of Ce, La, and Nd were used as thresholds to define binary output variables. The model achieved notable predictive accuracy, with a ROC-AUC of 0.83 for La, indicating reliable discrimination between classes. Feature importance analysis identified sulfur (S), aluminum (Al), and calcium (Ca) as key predictors, which aligns with prior studies on their geochemical associations with rare earth elements. While the ensemble model demonstrated strong performance, some misclassifications revealed the need for additional data processing and feature refinement. The study confirms that ensemble learning offers a powerful approach for geochemical modeling and mineral prospecting. The proposed methodology not only enhances prediction accuracy but also provides insight into elemental interactions that underpin the spatial distribution of rare earth metals in geological environments.

Keywords: rare earth elements, geochemical data, machine learning, ensemble models, XGBoost, content prediction, hyperparameter optimization

Acknowledgements: *The research was supported by RSF (project No. 25-26-00377), http://rscf.ru/project/25-26-00377/ .*

© Я.А. Тынченко[1,2], Е.В. Худякова[3], В.В. Кукарцев[1,2,3], М.Н. Степанцевич[3], А.А. Ступина[1,4,5], 2025.

[1]Сібір федералды университеті, Красноярск, Ресей;

[2]Н.Э. Бауман атындағы Мәскеу мемлекеттік техникалық университеті, Мәскеу, Ресей;

[3]К.А. Тимирязев атындағы Ресей мемлекеттік аграрлық университеті – Мәскеу ауыл-шаруашылық академиясы, Мәскеу, Ресей;

[4]Решетнев атындағы Сібір мемлекеттік ғылым және технологиялар университеті, Красноярск, Ресей;

[5]Ресей Төтенше жағдайлар министрлігінің мемлекеттік өртке қарсы қызметінің Сібір өрт сөндіру және құтқару академиясы, Железногорск, Ресей.

E-mail: vlad_saa_2000@mail.ru

ГЕОХИМИЯЛЫҚ ДЕРЕКТЕР НЕГІЗІНДЕ АНСАМБЛЬДІК ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНА ОТЫРЫП, СИРЕК КЕЗДЕСЕТІН ЖЕР ЭЛЕМЕНТТЕРІНІҢ ҚҰРАМЫН БОЛЖАУ

Ядвига А. Тынченко — Сібір федералды университеті, Красноярск, Ресей; Н.Э. Бауман атындағы Мәскеу мемлекеттік техникалық университеті, Мәскеу, Ресей, e-mail: 051301@mail.ru, https://orcid.org/0000-0001-8963-7830;
Елена В. Худякова — экономика ғылымдарының докторы, профессор, К.А. Тимирязев атындағы Ресей мемлекеттік аграрлық университеті – Мәскеу ауыл-шаруашылық академиясы, Мәскеу, Ресей, e-mail: evhudyakova@rgau-msha.ru, https://orcid.org/0000-0003-2156-7236;
Владислав В. Кукарцев — техника ғылымдарының кандидаты, доцент, Сібір федералды университеті, Красноярск, Ресей; К.А. Тимирязев атындағы Ресей мемлекеттік аграрлық университеті – Мәскеу ауыл-шаруашылық академия, Мәскеу, Ресей; Н.Э. Бауман атындағы Мәскеу мемлекеттік техникалық университеті, Мәскеу, Ресей, e-mail: vlad_saa_2000@mail.ru, https://orcid.org/0000-0001-6382-1736;
Марина Н. Степанцевич — экономика ғылымдарының кандидаты, доцент, К.А. Тимирязев атындағы Ресей мемлекеттік аграрлық университеті – Мәскеу ауыл-шаруашылық академиясы, Мәскеу, Ресей, e-mail: stepancevich@rgau-msha.ru, https://orcid.org/0000-0002-7125-3027;
Алена А. Ступина — доктор, профессор, Сібір федералды университеті, Красноярск, Ресей; Решетнев атындағы Сібір мемлекеттік ғылым және технологиялар Университеті, Красноярск, Ресей; Ресей Төтенше жағдайлар министрлігінің мемлекеттік өртке қарсы қызметінің Сібір өрт сөндіру және құтқару академиясы, Железногорск, Ресей, e-mail: h677hm@gmail.com, https://orcid.org/0000-0002-5564-9267.

Аннотация. Бұл зерттеу нақты зерттеулер нәтижесінде алынған геохимиялық деректерге машиналық оқытудың ансамбльдік әдістерін қолдану арқылы геологиялық үлгілерде сирек кездесетін элементтердің – церий (Ce), лантан (La) және неодим (nd) мазмұнын болжау міндетін қарастырады. Осы элементтерге өсіп келе жатқан өнеркәсіптік сұранысты ескере отырып, тиімді болжау модельдері барлау стратегияларын айтарлықтай жақсарта алады және қымбат далалық іріктеу қажеттілігін азайтады. Негізгі және микроэлементтердің концентрациясы бар геохимиялық мәліметтер жиынтығы алгоритмдердің, соның ішінде XGBoost, k-жақын көршілер (KNN), тірек векторлық машиналар (SVM) және логистикалық регрессияның көмегімен талданды. Ансамбль моделі генетикалық алгоритмдер мен торды іздеу комбинациясы арқылы оңтайландырылды, XGBoost негізгі алгоритм ретінде таңдалды, өйткені оның күрделі сызықтық емес тәуелділіктерге сезімталдығы жоғары болды. Жіктеу есептері үшін ce, La және Nd медианалық концентрациялары екілік Шығыс айнымалыларын анықтау үшін шекті мәндер ретінде пайдаланылды. Модель La үшін ROC-AUC 0,83 көмегімен болжаудың айтарлықтай дәлдігіне қол жеткізді, бұл сыныптар арасындағы сенімді айырмашылықты көрсетеді. Белгілердің маңыздылығын талдау күкіртті (S), алюминийді (Al) және кальцийді (CA) негізгі болжаушылар ретінде анықтады, бұл олардың сирек жер элементтерімен геохимиялық байланыстары туралы алдыңғы зерттеулерге сәйкес келеді. Ансамбль моделі жоғары өнімділікті көрсеткенімен, кейбір қате жіктеулер қосымша деректерді өңдеу және белгілерді нақтылау қажеттілігін анықтады. Зерттеу ансамбльдік оқыту геохимиялық модельдеу мен пайдалы қазбаларды барлау үшін күшті тәсілді ұсынатынын растайды. Ұсынылған әдістеме болжаудың дәлдігін арттырып қана қоймайды, сонымен қатар геологиялық ортада сирек

жер металдарының кеңістікте таралуына негіз болатын қарапайым өзара әрекеттесулер туралы түсінік береді.

Түйін сөздер: сирек кездесетін жер элементтері, геохимиялық деректер, машина оқыту, ансамбльдік модельдер, XGBoost, құрамды болжау, гиперпараметрлерді оңтайландыру.

© Я.А. Тынченко[1,2], Е.В. Худякова[3], В.В. Кукарцев[1,2,3], М.Н. Степанцевич[3], А.А. Ступина[1,4,5], 2025.
[1]Сибирский федеральный университет, Красноярск, Россия;
[2]Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия;
[3]Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, Россия;
[4]Сибирский государственный университет науки и технологий им. М.Ф. Решетнёва, Красноярск, Россия;
[5]Сибирская пожарно-спасательная академия ГПС МЧС России, Железногорск, Россия.
E-mail: vlad_saa_2000@mail.ru.

ПРОГНОЗИРОВАНИЕ СОДЕРЖАНИЯ РЕДКОЗЕМЕЛЬНЫХ ЭЛЕМЕНТОВ НА ОСНОВЕ ГЕОХИМИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ АНСАМБЛЕВОГО ОБУЧЕНИЯ

Ядвига А. Тынченко — Сибирский федеральный университет, Красноярск, Россия, e-mail: 051301@mail.ru, https://orcid.org/0000-0001-8963-7830;
Елена В. Худякова — доктор экономических наук, профессор, Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, Россия, e-mail: evhudyakova@rgau-msha.ru, https://orcid.org/0000-0003-2156-7236;
Владислав В. Кукарцев — кандидат технических наук, доцент, Сибирский федеральный университет, Красноярск, Россия, Российский государственный аграрный университет – МСХА имени К.А. Тимирязева, Москва, Россия, МГТУ им. Н.Э. Баумана, Москва, Россия, e-mail: vlad_saa_2000@mail.ru, https://orcid.org/0000-0001-6382-1736;
Марина Н. Степанцевич — кандидат экономических наук, доцент, РГАУ-МСХА имени К.А. Тимирязева, Москва, Россия, e-mail: stepancevich@rgau-msha.ru, https://orcid.org/0000-0002-7125-3027;
Алена А. Ступина — д.т.н., профессор, Сибирский федеральный университет, Красноярск, Россия; Сибирский государственный университет науки и технологий им. М.Ф. Решетнёва, Красноярск, Россия; Сибирская пожарно-спасательная академия ГПС МЧС России, Железногорск, Россия, e-mail: h677hm@gmail.com, https://orcid.org/0000-0002-5564-9267.

Аннотация. В этом исследовании рассматривается задача прогнозирования содержания редкоземельных элементов — церия (Ce), лантана (La) и неодима (Nd) — в геологических образцах путем применения ансамблевых методов машинного обучения к геохимическим данным, полученных в ходе реальных исследований. Учитывая растущий промышленный спрос на эти элементы, эффективные модели прогнозирования могут значительно улучшить

стратегии разведки и снизить потребность в дорогостоящем полевом отборе проб. Геохимический набор данных, содержащий концентрации основных и микроэлементов, был проанализирован с использованием комбинации алгоритмов, включая XGBoost, K-ближайших соседей (KNN), опорных векторных машин (SVM) и логистической регрессии. Ансамблевая модель была оптимизирована с помощью комбинации генетических алгоритмов и поиска по сетке, причем XGBoost был выбран в качестве основного алгоритма из-за его высокой чувствительности к сложным нелинейным зависимостям признак-цель. Для задач классификации медианные концентрации Ce, La и Nd использовались в качестве пороговых значений для определения бинарных выходных переменных. Модель достигла заметной точности прогнозирования с ROC-AUC 0,83 для La, что указывает на надежное различение между классами. Анализ важности признаков выявил серу (S), алюминий (Al) и кальций (Ca) в качестве ключевых предикторов, что согласуется с предыдущими исследованиями их геохимических ассоциаций с редкоземельными элементами. Хотя ансамблевая модель продемонстрировала высокую производительность, некоторые ошибочные классификации выявили необходимость дополнительной обработки данных и уточнения признаков. Исследование подтверждает, что ансамблевое обучение предлагает мощный подход для геохимического моделирования и разведки полезных ископаемых. Предложенная методология не только повышает точность прогнозирования, но и дает представление об элементарных взаимодействиях, которые лежат в основе пространственного распределения редкоземельных металлов в геологических средах.

Ключевые слова: редкоземельные элементы, геохимические данные, машинное обучение, ансамблевые модели, XGBoost, прогнозирование содержания, оптимизация гиперпараметров.

Introduction. In recent decades, rare earth elements (REE) have gained special significance due to their wide application in high-tech industries such as electronics manufacturing, renewable energy, and transportation (Evsyukov, et al., 2024; Kurashkin, et al., 2024; Panfilov, et al., 2024). Metals like cerium (Ce), lanthanum (La), and neodymium (Nd) play a key role in the production of magnets, batteries, and other critically important components. Predicting the content of these elements in rocks is an important step in their exploration and extraction (Epikhin, et al., 2024).

Traditional methods of geochemical data analysis, based on the chemical analysis of samples, often require significant time and financial resources (Degtyareva, et al., 2024). In recent years, with the development of machine learning methods, it has become possible to automate the process of analyzing and predicting metal content based on the geochemical characteristics of rocks (Gerasidi, et al., 2024; Kaung, et al., 2024; Bosikov, et al., 2023).

The aim of this study is to develop an efficient model for predicting the content

of rare earth elements based on geochemical data using ensembles of machine learning models. The study explored various approaches, including algorithms such as XGBoost, Support Vector Machines, and K-Nearest Neighbors. The optimization of the ensemble of models allowed for achieving high prediction accuracy for elements such as cerium (Ce) and lanthanum (La), as confirmed by high ROC-AUC metric values (Fedorova, et al., 2024; Evsyukov, et al., 2024).

The optimization of algorithm hyperparameters plays a key role in achieving high prediction accuracy. In several studies, grid search (GridSearchCV) and genetic algorithms were used for the automatic tuning of model parameters, which improved the ROC-AUC metrics to values above 0.80 (Smirnov, et al. 2023). The most significant factors affecting the content of REE are the concentrations of sulfur (S), aluminum (Al), and calcium (Ca) in rocks.

Other studies have shown that analyzing the mineral composition of rocks using data clustering methods plays an important role in predicting the content of rare earth elements. In particular, the use of deep learning algorithms, such as neural networks, has led to a significant improvement in prediction accuracy. Moreover, recent work has examined the influence of tectonic conditions on the distribution of rare earth elements, which expands the possibilities for prediction.

Several researchers have also proposed combined approaches that integrate traditional geochemical analysis methods with machine learning, achieving better results. The application of geostatistical methods in conjunction with machine learning algorithms allows for more accurate modeling of the spatial distribution of rare earth elements.

Materials and methods. For the study, a geochemical dataset of rock samples collected as part of a project investigating the mineral potential of rare earth elements was used (Migalin et al., 2024). The data were obtained from an open source and contain information about the chemical composition of samples collected from various geological regions. The dataset includes the concentrations of major elements such as aluminum (Al), calcium (Ca), iron (Fe), magnesium (Mg), silicon (Si), as well as the concentrations of rare earth metals such as cerium (Ce), lanthanum (La), and neodymium (Nd) (Panfilova, et al., 2024).

To gain a deeper understanding of the data structure, a statistical analysis of the main characteristics was conducted. Table 1 presents the mean values, medians, and standard deviations for the key elements in the dataset.

Table 1. Averages, medians, and standard deviations for key elements in the dataset

Element	Mean	Median	Std. Deviation
Al (Aluminum)	8.23	8.10	1.75
Ca (Calcium)	4.56	4.40	1.12
Fe (Iron)	5.34	5.10	1.68
Mg (Magnesium)	3.12	3.00	0.94
Si (Silicon)	24.85	24.70	4.58
Ce (Cerium)	0.45	0.42	0.14

La (Lanthanum)	0.22	0.21	0.08
Nd (Neodymium)	0.17	0.16	0.06

For the analysis of target variable concentrations, threshold values were calculated to determine the presence or absence of rare earth elements. In particular, for each element, the median was used as the threshold value for binary classification, where a value above the median indicates the presence of the element in the sample, and below it — its absence (Filina, et al., 2023).

Figure 1 shows histograms of the value distributions for elements such as Ce, La, and Nd. These charts demonstrate that the value distribution for rare earth elements exhibits some asymmetry, which indicates possible differences in element content across different samples.

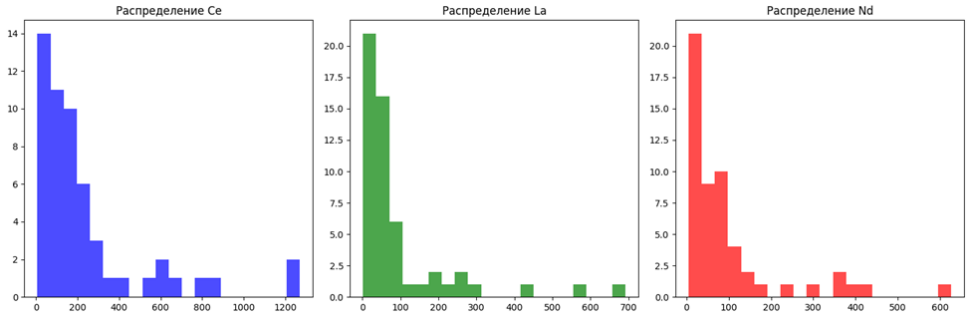

Fig. 1. Histograms of target variables (Ce, La, Nd)

For predicting the content of rare earth elements, an ensemble of models was applied, consisting of the following algorithms:

• XGBoost: Gradient boosting on decision trees. This algorithm effectively identified nonlinear relationships between features and target variables. Hyperparameter optimization for XGBoost included tuning parameters such as n_estimators, max_depth, learning_rate, and subsample.

• K-Nearest Neighbors (KNN): This algorithm was used for classification based on object proximity. Optimal hyperparameter values for the number of neighbors and distance metrics were determined using grid search.

• Support Vector Machines (SVM): The use of SVM allowed for the construction of a nonlinear decision boundary for predicting metal content. A Gaussian kernel function was employed to enhance efficiency.

• Logistic Regression: As a basic linear classification method, logistic regression was added to the ensemble to account for linear dependencies in the data.

To achieve maximum prediction efficiency, an ensemble optimization approach was employed, which included the use of a genetic algorithm and grid search (GridSearchCV). The primary focus was on tuning the hyperparameters of the key ensemble models, such as XGBoost, K-Nearest Neighbors (KNN), and SVM.

XGBoost gradient boosting on decision trees was chosen as the main model of the ensemble due to its ability to detect complex nonlinear dependencies in the data (Kuzkin, et al., 2024; Rozhkova, et al., 2024). For hyperparameter optimization, the grid search (GridSearchCV) method was used, varying parameters such as:

- n_estimators (number of trees),
- max_depth (maximum tree depth),
- learning_rate (learning rate),
- subsample (fraction of the data sample used for building each tree) (Stepanova et al., 2024).

The optimization process was conducted using 3-fold cross-validation, which helped to avoid overfitting and achieved ROC-AUC values exceeding 0.79 for target variables such as Ce and Nd (Olentsova et al., 2024). The best hyperparameters identified during the search were: n_estimators = 100, max_depth = 3, learning_rate = 0.01, and subsample = 0.8.

For the K-Nearest Neighbors (KNN) algorithm, a grid search was also used to tune the following parameters:

- n_neighbors (number of nearest neighbors),
- p (distance metric).

As a result of the optimization, it was determined that the most effective value for the number of neighbors is 9 when using the Manhattan distance metric. This configuration achieved ROC-AUC values of around 0.70.

The SVM algorithm was employed to construct a nonlinear decision boundary using a Gaussian kernel. During the grid search, the following parameters were varied:

- C (regularization parameter),
- gamma (kernel parameter).

The optimal parameter values improved the prediction accuracy for rare earth elements such as Ce and La, reaching ROC-AUC scores of approximately 0.70.

Logistic regression was added to the ensemble as a basic linear method to capture simple linear dependencies between features. In this case, grid search was also used to tune the regularization parameter C. The model demonstrated stable results, complementing the ensemble for predicting metal content.

To evaluate the performance of the ensemble, the following metrics were used:

- ROC-AUC: The primary metric for assessing the model's ability to distinguish between classes (presence and absence of the element);
- F1-score: Used to balance precision and recall;
- Accuracy: The percentage of correct model predictions.

Results and discussion. The results of predicting the content of rare earth elements demonstrated the high efficiency of the proposed ensemble model, which includes the algorithms XGBoost, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression (Borodulin et al., 2024). These

algorithms enabled the modeling of relationships between geochemical features and the presence of elements such as cerium (Ce), lanthanum (La), and neodymium (Nd) in rock samples.

An example prediction for the content of La (lanthanum) showed promising results. Based on the confusion matrix, the model correctly predicted 6 instances for class 0 (absence of La) and 6 instances for class 1 (presence of La). However, the model also made several errors: 4 cases where the presence of La was incorrectly predicted, and 1 case where the model failed to predict the presence of La when it was actually present. These errors may be attributed to a high correlation between features or missing data in some samples.

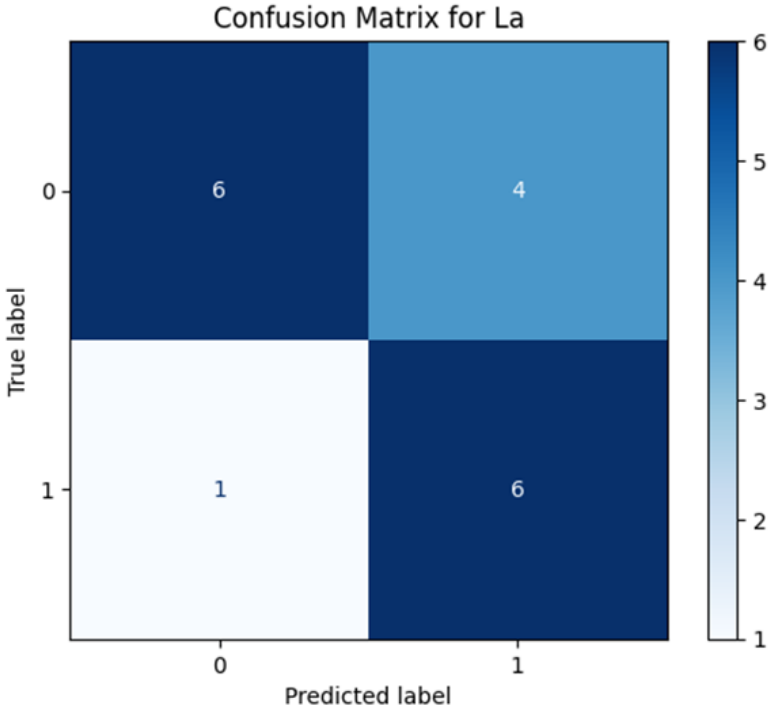

Fig. 2. Confusion matrix

The ROC curve for La (Figure 3) showed an AUC value of 0.83, indicating a strong ability of the model to distinguish between classes. This metric confirms that the model can accurately predict the presence or absence of lanthanum in the samples. An AUC value above 0.80 demonstrates the reliability of the model in this prediction task, despite a certain number of errors in the confusion matrix.

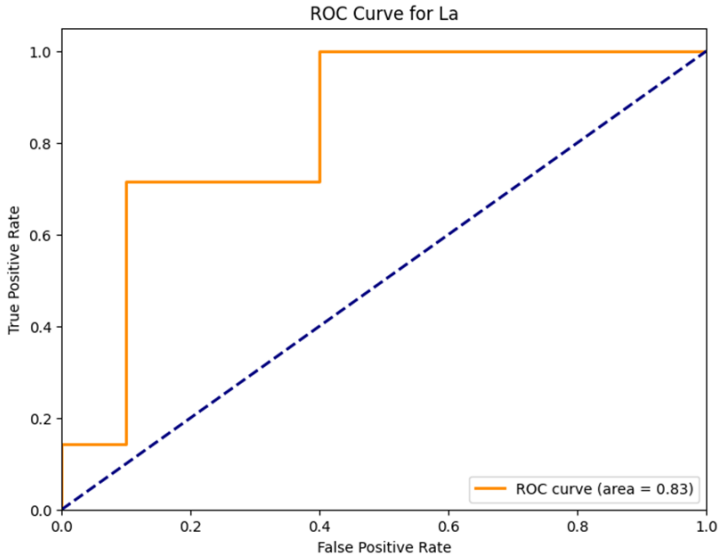

Fig. 3. Curve ROC

The feature importance graph for La (Figure 4) demonstrated that the key variables influencing the predictions are:

- S_pct_ICP-OES (sulfur content),
- Al_pct_ICP-OES (aluminum content),
- Ca_pct_ICP-OES (calcium content).

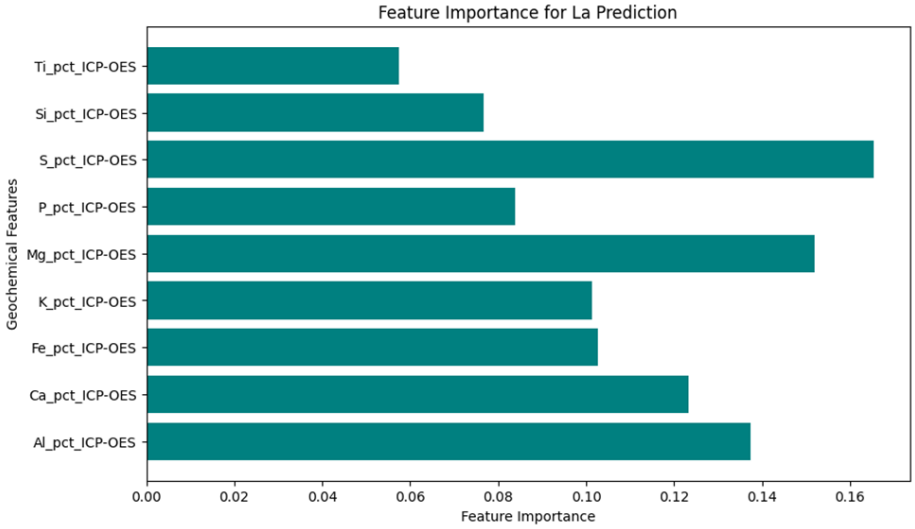

Fig. 4. Feature importance

These features have the greatest impact on the model's predictions. The high importance of sulfur may indicate its interconnection with rare earth elements in geological structures, while aluminum and calcium also play a significant role in predicting the La content. These results are consistent with previous studies showing that rare earth elements often occur together with these chemical elements in rocks.

Despite the high performance of the model, the confusion matrix reveals several misclassifications. In particular, the model erroneously predicted the presence of La 4 times when it was not present, suggesting a possible overestimation of the influence of some features. These errors may be related to a high correlation between features or the complexity of class separation in certain samples. A more detailed data analysis could help improve the model, for example, by using more advanced data processing algorithms such as outlier detection methods or regularization.

Despite the high performance of the model, the confusion matrix reveals several misclassifications. In particular, the model erroneously predicted the presence of La 4 times when it was not present, suggesting a possible overestimation of the influence of some features. These errors may be related to a high correlation between features or the complexity of class separation in certain samples. A more detailed data analysis could help improve the model, for example, by using more advanced data processing algorithms such as outlier detection methods or regularization.

Despite the high performance of the model, the confusion matrix reveals several misclassifications. In particular, the model erroneously predicted the presence of La 4 times when it was not present, suggesting a possible overestimation of the influence of some features. These errors may be related to a high correlation between features or the complexity of class separation in certain samples. A more detailed data analysis could help improve the model, for example, by using more advanced data processing algorithms such as outlier detection methods or regularization.

Conclusion. During the study, an ensemble machine learning model was developed and optimized to predict the content of rare earth elements (Ce, La, and Nd) based on geochemical data. The application of algorithms such as XGBoost, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression enabled the identification of key nonlinear relationships between features, which contributed to improved prediction accuracy.

Hyperparameter optimization using a genetic algorithm and grid search led to significant improvements in the model's performance. The best results were achieved for the element La, with a ROC-AUC of 0.83, indicating high prediction accuracy. Good results were also obtained for Ce and Nd, although their accuracy was slightly lower.

Analysis of feature importance revealed that the key variables with the greatest influence on predictions are the contents of sulfur (S), aluminum (Al), and calcium (Ca). These elements proved to be significant factors in predicting the content of rare earth metals, underscoring their role in geochemical processes.

Despite the successful results, the model made several errors, which indicates the need for further improvements, for example, through the implementation of outlier

detection methods and the use of additional data. It is also advisable to consider employing more complex models, such as deep neural networks, to capture the intricate multidimensional dependencies among features.

Overall, the proposed methodology demonstrated high effectiveness in addressing the tasks of predicting rare earth element content and may be useful in other geochemical analysis and mineral exploration applications.

References

Borodulin A.S. et al. Analyzing Data by Applying Neural Networks to Identify Patterns in the Data. Computational Methods in Systems and Software: Proceedings of the Conference (Chengdu, China, 23–26 October 2024). – 2024. – P. 99–108. (in Eng.)

Bosikov I.I., Klyuev R.V., Revazov V.Ch., Martyushev N.V. Analysis and evaluation of prospects for highquality quartz resources in the North Caucasus [Analiz i otsenka perspektiv resursov vysokokachestvennogo kvartsa na Severnom Kavkaze]. Mining Science and Technology (Russia). – 2023. – №8(4). – P. 278–289. https://doi.org/10.17073/2500-0632-2023-10-165. (in Russian)

Degtyareva K., Khramkov V., Malozyomov B. Use of Computer Simulation Tools to Simulate Processes at the Foundry. 23rd International Symposium INFOTEH-JAHORINA (INFOTEH) (East Sarajevo, March 20–22, 2024). – 2024. – P. 1–5. IEEE. (in Eng.)

Epikhin A.I., Stupina A.A., Panfilov I.A., Bukhtoyarov V.V., Shepeta N.A. Determinants for assessing the energy efficiency of a coal mining enterprise. News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences. – 2024. – 2024(4). – P. 61–71. (in Eng.)

Evsyukov D., Degtyareva K., Boyko A., Kukartseva O. Forecasting the environmental safety of water resources using neural networks. E3S Web of Conferences. – 2024. – 583. – P. 02012. EDP Sciences. (in Eng.)

Evsyukov D., Glinscaya A., Volneikina E., Kukartseva S. Analyzing the influence of parameters on water quality using logistic regression. BIO Web of Conferences. – 2024. – 130. – P. 03001. EDP Sciences. (in Eng.)

Fedorova N., Panfilov I., Moiseeva E., Degtyareva K., Markovskaya E. Analytical methods and tools for business process optimization. BIO Web of Conferences. – 2024. – 113. – P. 05009. EDP Sciences. (in Eng.)

Filina O.A., Martyushev N.V., Malozyomov B.V. Increasing the Efficiency of Diagnostics in the Brush-Commutator Assembly of a Direct Current Electric Motor // Energies. – 2023. – 17(1). – P. 17. (in Eng.)

Gerasidi V.V., Dubrovin R.G., Kukartseva O.I., Panfilov I.A. Analysis of technical operation of Caterpillar Engineering Corporation engines in industry. News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences. – 2024. – 5. – P. 23–35. (in Eng.)

Jackson R., Lin S. The impact of tectonic settings on rare earth element distribution. Tectonophysics. – 2019. – 732. – Pp. 110–126. https://doi.org/10.1016/j.tecto.2019.04.012 (in Eng.)

Kaung P.A., Isakov A.E., Panfilov I.A., Stupina A.A. Principles for forming environmentally safe and economically effective sustainable development of geo resources [Printsipy formirovaniya ekologicheski bezopasnogo i ekonomicheski effektivnogo ustoychivogo osvoyeniya georesursov]. MIAB. Mining Inf. Anal. Bull. – 2024. – 7(1). – P. 159–175. (in Russian)

Kurashkin S., Kravtsov K., Boyko A., Volneikina E. Integration of IT solutions for resource management in environmentally sustainable agriculture. BIO Web of Conferences. – 2024. – 145. – P. 06008. EDP Sciences. (in Eng.)

Kuzkin A.Y., Zadkov D.A., Skeeba V.Y. Viscoplastic properties of chromium-nickel steel in short-term creep under constant stress. Part 1 [Vyazkoplasticheskiye svoystva khromonikelevoy stali pri kratkovremennoy polzuchesti pod postoyannym napryazheniyem. Chast' 1]. CIS Iron and Steel Review. – 2024. – 27. – P. 71–77. (in Russian)

Migalin K.V., Boychuk I.P., Pavlov P.P., Panfilova T.A. Study of ways to influence the combustion process in an ejector pulsating air-breathing engine in order to improve its specific performance. International Journal of Hydrogen Energy. – 2024. – 96. – P. 343–350. (in Eng.)

Olentsova J. et al. Predictive modelling of post-monsoon groundwater quality in Telangana using machine learning techniques. BIO Web of Conferences. – 2024. – 116. – P. 03021.

Panfilov I., Fedorova N., Moiseeva E., Divaeva A., Degtyareva K. Increasing competitiveness of enterprises by optimizing business processes as a factor of sustainable development of industrial region. E3S Web of Conferences. – 2024. – 531. – P. 05019. EDP Sciences. (in Eng.)

Panfilova T.A., Mikhalev A.S., Xiaogang W. Treatment of wastewater from mining industrial enterprises from phenols [Ochistka stochnykh vod gornodobyvayushchikh promyshlennykh predpriyatiy ot fenolov]. MIAB. Mining Inf. Anal. Bull. – 2024. – 7(1). – P. 72–82. (in Russian)

Rozhkova A. et al. Application of statistical data analysis algorithms and determination of the most significant diagnostic factors. BIO Web of Conferences. – 2024. – 116. – P. 03015. (in Eng.)

Smirnov E., Dvoryatkina S., Martyushev N., Shcherbatykh S. Software Package to Support Students' Research Activities in the Hybrid Intellectual Environment of Mathematics Teaching. Mathematics. – 2023. – 11(4). – 952. https://doi.org/10.3390/math11040952 (in Eng.)

Stepanova E. et al. Development of an automated information system for a logging company. BIO Web of Conferences. – 2024. – 116. – P. 03014. (in Eng.)

Publication Ethics and Publication Malpractice
in the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://www.geolog-technical.kz/index.php/en/
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 15.06.2025.
Формат 70x90$^{1}/_{16}$. Бумага офсетная. Печать – ризограф.
14,5 п.л. Заказ 3.

РОО «Национальная академия наук РК»
050010, Алматы, ул. Шевченко, 28, т. 272-13-19