

Применение Python для интеллектуального анализа данных на нефтеперерабатывающем заводе

И.А. Бабкин, Е.И. Алаторцев, К.В. Гусев

МИРЭА – Российский технологический университет, Москва

Аннотация: В статье изложены основные принципы применения технологий искусственного интеллекта и машинного обучения на нефтеперерабатывающих заводах. Рассмотрены преимущественные характеристики языка программирования Python при организации сбора данных с технологических установок, обработке больших данных, построении и валидации моделей прогнозирования. Проанализированы основные библиотеки Python по работе с данными на всех этапах жизненного цикла моделей машинного обучения (Pandas, Matplotlib, Scikit-learn и другие). Особое внимание уделено задаче оптимизации энергопотребления и повышения эффективности работы нефтеперерабатывающих предприятий за счет автоматизации прогнозирования работы их оборудования.

Ключевые слова: машинное обучение, искусственный интеллект, интеллектуальный анализ данных, Python, Scikit-learn, прогнозирование, энергопотребление, нефтепереработка, нефтегазовая отрасль, нефтеперерабатывающий завод.

Введение

Современные российские нефтеперерабатывающие заводы (НПЗ) представляют собой сложные технологические комплексы, оснащенные множеством датчиков контроля работы технологических установок, которые накапливают большие объемы данных, требующие своевременной обработки в ходе постоянного мониторинга деятельности. Данные на НПЗ могут собираться в зависимости от технологического процесса по определенным параметрам, по которым проводится анализ их влияния на процесс переработки нефти, что позволяет не только поддерживать стабильную работу соответствующего технологического оборудования, но и осуществлять прогнозирование потребления технологическими установками энергоресурсов в последующих отчетных периодах [1]. Так как спрос на энергию постоянно увеличивается, принципиальную важность приобретает задача ее сбережения, что способствует снижению расхода энергоносителей на НПЗ [2].

За последние несколько десятилетий средства анализа данных значительно эволюционировали. Если раньше стандартными инструментами изучения накопленных данных были простые статистические расчеты и визуальное построение показателей, то в настоящее время, благодаря научно-техническому прогрессу и последовавшему увеличению вычислительных мощностей, нашли свое промышленное применение разработанные еще в XX веке инструменты, такие, как регрессионный анализ, метод главных компонент, корреляционный анализ и другие. В связи со стремительным развитием компьютерных технологий стало возможным применение более продвинутых методов, включая стохастическое моделирование и машинное обучение. Однако повсеместное ускоряющееся увеличение объема данных наложило новые требования к их оперативной обработке, что повлекло за собой бурный рост в популярности алгоритмов искусственного интеллекта, позволяющих находить в промышленных данных скрытые зависимости и строить на их основе прогностические модели [3].

Интеллектуальная автоматизация управления данными на НПЗ

В настоящее время в области практического применения технологий искусственного интеллекта существует широкий набор методов анализа больших объемов данных. В нефтяной промышленности сбор данных осуществляется с помощью множества датчиков, которыми снабжены технологические установки НПЗ, собирающих в режиме реального времени данные параметров технологического процесса [4]. Исходя из большого объема данных возникает необходимость автоматизации процессов их сбора, хранения и представления. Современные инструменты интеллектуального анализа данных позволяют автоматизировать как процессы обучения и проверки моделей прогнозирования, так и предшествующих им предварительной обработки данных, отбора наиболее важных признаков, а также подбора подходящих алгоритмов машинного обучения [5, 6].

Приведенный подход способствует тому, что значительно совершенствуется весь цикл работы с данными, делая его более эффективным и масштабируемым [7]. Сразу после сбора данных с технологических установок появляется возможность в короткие сроки практически в режиме реального времени получить промежуточные прогнозы по их работе в перспективе, что способствует оперативному принятию управленческих решений по регулированию технологических процессов [8 – 10].

Для реализации процесса машинного обучения, представленного на рис. 1, в промышленных средах, таких как НПЗ, на разных этапах работы с данными могут использоваться ряд специализированных инструментов. За получение сведений с датчиков установок на этапе сбора данных отвечают системы диспетчерского управления, распределенные системы управления и другие решения «промышленного интернета вещей». В дальнейшем эти данные передаются в корпоративные базы данных, озера данных и хранилища данных. Для обработки и подготовки данных используют инструменты извлечения, преобразования и загрузки данных, которые получают данные из множества различных источников, очищают, структурируют и загружают в аналитические системы для последующего использования их в бизнес-задачах.

Основная часть процесса создания моделей машинного обучения в такой системе может выполняться на специализированных платформах, таких как RapidMiner, Loginom, MATLAB и другие. В случае, если у предприятия есть возможность воспользоваться облачными сервисами, допустимо применить решения от крупнейших поставщиков облачных услуг: Amazon Web Services, Microsoft Azure и Google Cloud. В условиях наложенных на Российскую Федерацию ограничений отечественные разработчики успешно используют отечественные решения от Яндекс, Cloud.ru и других.

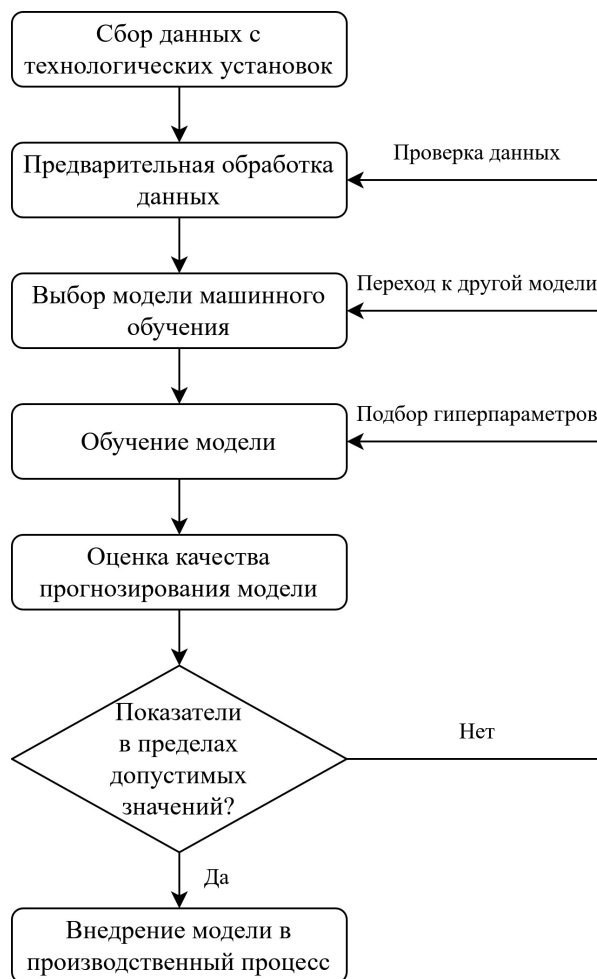


Рис. 1. – Упрощенная модель процесса машинного обучения

Несмотря на наличие множества аналитических платформ, в том числе облачных, среди всех инструментов в области подготовки и эксплуатации моделей машинного обучения закрепился общепринятый стандарт реализации этих процессов – язык программирования Python и библиотеки к нему. Связано это с его ключевыми преимуществами:

- гибкость при реализации своих прогностических моделей, позволяющая очень быстро и эффективно создавать прототипы, тестировать их, объединять, выгружать и интегрировать в другие программные системы на Python;
- открытость разработки, побуждающая свободно распространять свои наработки в области машинного обучения, проверять работоспособность и

точность моделей других пользователей, организовывать соревнования (например, платформа Kaggle) по созданию наиболее эффективных решений на основе искусственного интеллекта в различных областях, что в совокупности создает отличные условия для образования сообщества по машинному обучению, объединяющее в себе экспертов в этой области, энтузиастов, ученых и представителей бизнеса;

- богатая коллекция библиотек, модулей и фреймворков, разработанных специально для научных вычислений, анализа данных и машинного обучения.

Гибкость создаваемых в Python программных решений позволяет проектировать информационные системы, включающие в себя все этапы рассмотренного ранее цикла работы с данными для машинного обучения. Это значительно снижает стоимость и сложность создания решений на основе искусственного интеллекта, так как весь процесс разработки проходит в рамках единой среды программирования, где все функциональные компоненты способны общаться и взаимодействовать друг с другом без необходимости внедрения дополнительных средств коммуникации между модулями (например, брокеров сообщений).

Помимо этого, с помощью Python в рамках единой информационной системы можно поддерживать взаимодействие с рядом разнообразных внешних источников данных, такими как базы данных различных типов, облачными платформами, системами управления и мониторинга, что делает такую систему адаптируемой к существующей инфраструктуре НПЗ, а также масштабируемой в рамках одного предприятия или целого нефтегазохимического комплекса. Универсальность Python при различных сценариях работы с данными закономерно делает его одним из лучших инструментов в области анализа данных и машинного обучения.

Возможности Python для анализа и обработки данных с НПЗ

Рассмотрим конкретные инструменты для работы с данными, доступные в языке программирования Python. В нем представлен широкий спектр библиотек, предназначенных для манипулирования данными на различных этапах их обработки и использования. Среди них есть как инструменты базовой обработки и визуализации информации, так и реализации сложных нейросетевых моделей. Рассмотрим наиболее распространенные в разработке библиотеки Python, сгруппированные по их функциональному назначению:

1. Обработка и подготовка данных:

- NumPy: работа с многомерными массивами и матрицами, эффективное выполнение как стандартных математических операций над ними, так и более сложных вычислений из линейной алгебры, а также специфических операций как преобразования Фурье и полиномиальная алгебра;
- SciPy: более сложные вычисления из областей обработки сигналов, статистических распределений, кластерного анализа и прочего;
- Pandas: основной инструмент работы с табличными данными, позволяющий загружать данные из различных источников, преобразовывать, очищать, обрабатывать и выгружать в необходимый формат;
- Dask: работа с большими данными за счет распределения вычислений, эффективное масштабирование рабочих процессов.

На рис. 2 показан пример представления данных, полученных в ходе исследовательской работы авторов статьи с технологическими установками одного из российских НПЗ, которые после обработки средствами библиотеки Pandas были приведены в табличный вид.

	Дата	Загрузка установки	Расход газа	Коэффициент калорийности газа	Общее потребление топлива	Удельное потребление топлива	Средневзвешенное КПД
0	2023-08-01	87.714851	1.113411	1.650503	1.837689	20.950712	86.319963
1	2023-08-02	86.679225	1.082873	1.676076	1.814979	20.939026	85.949068
2	2023-08-03	85.638053	1.073532	1.674609	1.797747	20.992381	85.939716
3	2023-08-04	85.409293	1.073124	1.679038	1.801816	21.096252	85.935860
4	2023-08-05	85.239241	1.065527	1.653139	1.761464	20.664948	85.927978
...	...	...	...	...	...	...	...
392	2024-08-27	112.285293	1.093418	1.614206	1.765002	15.718906	85.350231
393	2024-08-28	112.626795	1.052864	1.614237	1.699573	15.090307	85.467721
394	2024-08-29	113.220751	1.051649	1.630563	1.714780	15.145455	85.594222
395	2024-08-30	113.520297	1.062982	1.630557	1.733253	15.268225	85.587601
396	2024-08-31	113.606863	1.092680	1.596691	1.744672	15.357097	85.385074

Рис. 2. – Представление данных с помощью библиотеки Pandas

2. Визуализация данных:

- Matplotlib: основная библиотека 2D-визуализации данных для широкого спектра задач с возможностью создания интерактивных графиков;
- Seaborn: построенный на основе Matplotlib более высокоуровневый инструмент визуализации в Python, позволяющий быстро создавать красивые и информативные графики, ограничиваясь при этом в их настройке;
- Plotly: поддержка широкого спектра графиков различной сложности построения при создании динамических информационных панелей для исследования данных.

На рис. 3 представлен пример использования библиотеки Matplotlib для визуализации процесса подбора гиперпараметра (сдвига прямой) для различных типов линейной регрессии с регуляризацией, обученных на тех же данных с технологической установки НПЗ, что и в примере на рис. 2, накопленных за один год ее работы.

3. Искусственный интеллект и методы машинного обучения:

- Scikit-learn: самая распространенная библиотека по созданию, обучению и оценке моделей машинного обучения, имеющая в наличии реализации алгоритмов регрессии, классификации, кластеризации;

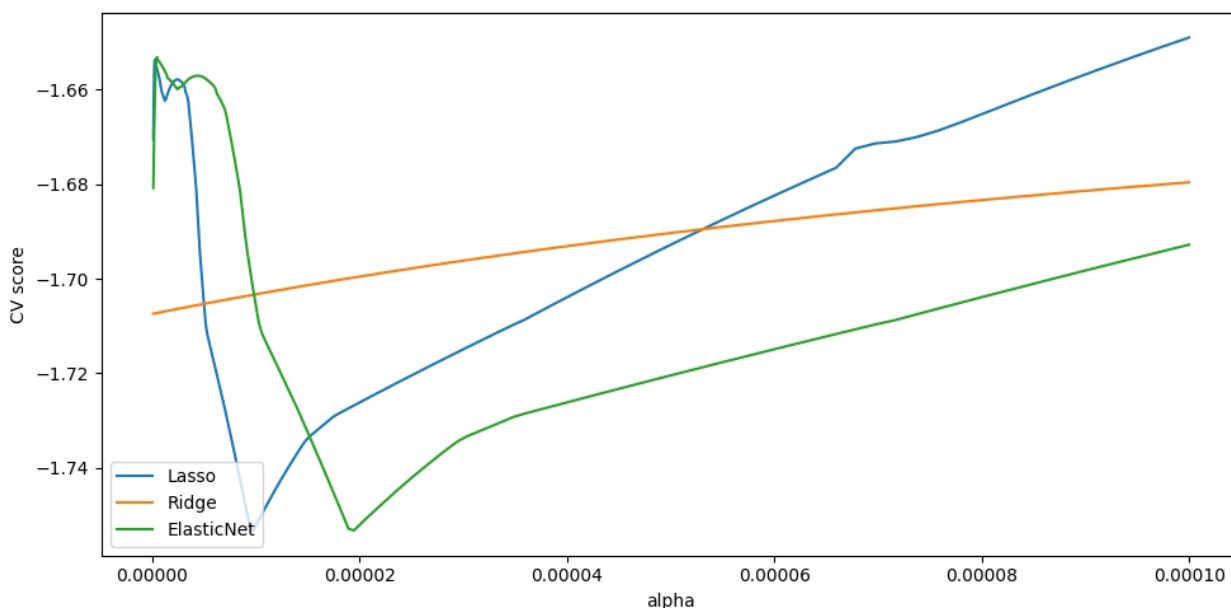


Рис. 3. – Визуализация процесса подбора гиперпараметра с помощью библиотеки Matplotlib

– PyTorch: мощный инструмент для обработки естественного языка, компьютерного зрения и реализации моделей на основе обучения с подкреплением;

– TensorFlow и Keras: создание, обучение, развертывание и оценка нейронных сетей разных размерностей, способных, благодаря Keras, гибко масштабироваться и распределять вычисления;

– Statsmodels: статистическое моделирование с поддержкой проверки гипотез и нахождения доверительных интервалов, анализ временных рядов, диагностическая статистика;

– а также большое число модулей по отдельным инструментам и нейросетевым моделям.

Благодаря широкому спектру специализированных библиотек, Python представляет собой мощный инструмент интеллектуального анализа данных на НПЗ, что отмечается в источниках [11 – 13]. Он предлагает надежную среду для выбора, обучения и оценки моделей машинного обучения,

предназначенных для прогнозирования потребления топлива на предприятии.

В библиотеке Scikit-learn представлены практически все необходимые инженеру инструменты для проведения анализа данных. С помощью нее можно эффективно обрабатывать данные с пропущенными, испорченными или неверно типизированными значениями, полученные с технологических установок. В дальнейшем они могут быть нормализованы или масштабированы (в зависимости от требований модели) и разделены на обучающие, тестовые и валидационные подмножества. При этом весь процесс предварительной обработки данных можно автоматизировать средствами Scikit-learn с помощью построения «пайплайнов».

Scikit-learn позволяет работать с большинством распространенных регрессионных алгоритмов машинного обучения [14]. При этом выбор лучшей модели среди всех доступных можно осуществить за счет методов кросс-валидации [15, 16], реализованных в данной библиотеке. Настройка гиперпараметров модели [17] под изменяющиеся условия эксплуатации технологической установки делает процесс прогнозирования потребления топлива более гибким к непредвиденным рабочим состояниям оборудования.

Если предполагается внедрение в производственные процессы нейросетевых моделей, то используются такие библиотеки, как PyTorch и TensorFlow. Они позволяют создавать как простые модели с полной взаимосвязью нейронов, так и более сложные архитектуры с обратной связью и элементами памяти. Полный потенциал этих библиотек можно реализовать, используя интегрированную в них поддержку вычислений на графическом процессоре, которая значительно ускоряет работу алгоритмов машинного обучения и процесс обучения модели.

Приведенное разнообразие возможностей Python делает его универсальной и мощной платформой для разработки программных решений

в области искусственного интеллекта. Это особенно важно для гибкой интеграции в существующие информационные системы нефтехимических предприятий и масштабирования решений в промышленных масштабах, что делает язык особенно востребованным в задачах анализа данных и цифровизации производств.

Заключение

Сложные промышленные объекты, такие как НПЗ, в условиях «четвертой промышленной революции» («Индустрия 4.0») накапливают огромные массивы необработанных данных. Анализ этих данных мог бы не только контролировать работу оборудования на предприятии, но и выстраивать модели прогнозирования, способствующие экономии энергоресурсов. С развитием вычислительных технологий на смену простым статистическим методам пришли машинное обучение и искусственный интеллект, позволяющие находить ранее неизвестные закономерности в данных и автоматизировать весь цикл работы с ними. На всех этапах – от сбора до получения прогнозных расчетов – все чаще применяется язык программирования Python, благодаря своей гибкости, открытости и широкому набору библиотек. Он позволяет строить решения под конкретные задачи: от подготовки и визуализации данных (NumPy, Pandas, Matplotlib) до создания и внедрения моделей искусственного интеллекта (Scikit-learn, PyTorch, TensorFlow), что делает его одним из наиболее эффективных и универсальных инструментов для промышленной аналитики и цифровизации НПЗ.

Литература

1. Степанова А.И., Хальясмаа А.И., Матренин П.В. Краткосрочное прогнозирование нагрузки предприятия нефтегазовой промышленности с

использованием технологических факторов и аддитивного объяснения Шепли // Известия высших учебных заведений. Проблемы энергетики. 2024. Т. 26. № 4. С. 75-88.

2. Глаголева О.Ф., Капустин В.М. Повышение эффективности процессов подготовки и переработки нефти (обзор) // Нефтехимия. 2020. Т. 60. № 6. С. 745-754.

3. Ефимушкин В.А., Ледовских Т.В., Чукарин А.В. Искусственный интеллект в цифровой трансформации промышленности // Цифровая трансформация. Связь будущего: Материалы XXVIII Международного Форума МАС' 2024, Москва, 26 апреля 2024 года. Москва: Государственный университет просвещения, 2024. С. 119-131.

4. Бабкин И.А., Алаторцев Е.И. Методы машинного обучения для прогнозирования энергопотребления на предприятии нефтехимического комплекса // Информационные технологии, системный анализ и управление (ИТСАУ-2024) : Сборник трудов XXII Всероссийской научной конференции молодых ученых, аспирантов и студентов. В 2-х томах, Таганрог, 05–07 декабря 2024 года. Таганрог: ДиректСайнс (ИП Шкуркин Д.В.), 2024. С. 317-320.

5. Baratchi M., Wang C., Limmer S., van Rijn J.N., Hoos H., Bäck T., Olhofer M. Automated machine learning: past, present and future // Artificial intelligence review. 2024. Vol. 57. № 5. p. 122.

6. Мосева М.С. О задаче разработки автоматического конвейера запуска моделей машинного обучения // Инженерный вестник Дона. 2022. № 4. URL: ivdon.ru/magazine/archive/n4y2022/7605.

7. Chauhan K., Jani S., Thakkar D., Dave R., Bhatia J., Tanwar S., Obaidat M.S. Automated Machine Learning: The New Wave of Machine Learning // 2nd International Conference on Innovative Mechanisms for Industry Applications,

ICIMIA 2020 – Conference Proceedings : 2, Bangalore, 05–07 March 2020. Bangalore, 2020. pp. 205-212.

8. Забайкин Ю.В., Барыкин М.В., Попович А.Э., Заернюк В.М. Анализ экономической эффективности использования методов машинного обучения и искусственного интеллекта для автоматизации управленческих решений в промышленности. Понизовье: National Reserarch, 2024. 208 с.

9. Забайкин Ю.В. Барыкин М.В. Экономические и управленческие аспекты использования предиктивной аналитики и машинного обучения для оптимизации производственных процессов в промышленности // Экономика: вчера, сегодня, завтра. 2024. Т. 14. № 7-1. С. 606-617.

10. Большаков М.А., Ходаковский В.А. Подход к повышению качества моделей машинного обучения в задачах мониторинга сложных систем на основе применения метрических пространств // Инженерный вестник Дона. 2024. № 11. URL: ivdon.ru/magazine/archive/n11y2024/9630.

11. Al Jlibawi A.H.H., Othman M.L., Ishak A., Moh Noor B.S., Sajitt A.H.M.S. Optimization of Distribution Control System in Oil Refinery by Applying Hybrid Machine Learning Techniques // IEEE Access. 2022. Vol. 10. pp. 3890-3903.

12. Ferreira J., Pedemonte M., Torres A.I. Development of a machine learning-based soft sensor for an oil refinery's distillation column // Computers & Chemical Engineering. 2022. Vol. 161. p. 107756.

13. Al-Jamimi H.A., BinMakhashen G.M., Saleh T.A. Multiobjectives optimization in petroleum refinery catalytic desulfurization using Machine learning approach // Fuel. 2022. Vol. 322. p. 124088.

14. Salama M. Optimization of Regression Models Using Machine Learning: A Comprehensive Study with Scikit-learn // International Uni-Scientific Research Journal. 2024. Vol. 5. № 16. pp. 119-129.

15. Yates L.A., Aandahl Z., Richards S.A., Brook B.W. Cross validation for model selection: A review with examples from ecology // *Ecological Monographs*. 2023. Vol. 93. № 1. p. e1557.

16. Jung Y. Multiple predicting K-fold cross-validation for model selection // *Journal of Nonparametric Statistics*. 2017. Vol. 30. № 1. pp. 197-215.

17. Hossain M.R., Timmer D. Machine learning model optimization with hyper parameter tuning approach // *Global Journal of Computer Science and Technology: D. Neural & Artificial Intelligence*. 2021. Vol. 21. № 2. pp. 7-13.

References

1. Stepanova A.I., Khal'yasmaa A.I., Matrenin P.V. Izvestiya vysshihkh uchebnykh zavedeniy. Problemy energetiki. 2024. Vol. 26. № 4. pp. 75-88.

2. Glagoleva O.F., Kapustin V.M. *Neftekhimiya*. 2020. Vol. 60. № 11. pp. 1207-1215.

3. Efimushkin V.A., Ledovskikh T.V., Chukarin A.V. Tsifrovaya transformatsiya. Svyaz' budushchego : Materialy XXVIII Mezhdunarodnogo Forumy MAS' 2024, Moskva, 26 aprelya 2024 goda (Digital Transformation. Connecting the Future : Proceedings of the XXVIII International Forum ICA' 2024, Moscow, April 26, 2024). Moskva: Gosudarstvennyy universitet prosveshcheniya, 2024. pp. 119-131.

4. Babkin I.A., Alatortsev E.I. Informatsionnye tekhnologii, sistemnyy analiz i upravlenie (ITSAU-2024): Sbornik trudov XKhII Vserossiyskoy nauchnoy konferentsii molodykh uchenykh, aspirantov i studentov. V 2-kh tomakh, Taganrog, 05–07 dekabrya 2024 goda (Information Technologies, System Analysis and Management (ITSAU-2024) : Proceedings of the XXII All-Russian Scientific Conference of Young Scientists, Postgraduates and Students. In 2 volumes, Taganrog, December 05-07, 2024). Taganrog: DirektSayns (IP Shkurkin D.V.), 2024. pp. 317-320.

5. Baratchi M., Wang C., Limmer S., van Rijn J.N., Hoos H., Bäck T., Olhofer M. Artificial intelligence review. 2024. Vol. 57. № 5. p. 122.
 6. Moseva M.S. Inzhenernyj vestnik Dona. 2022. № 4. URL: ivdon.ru/magazine/archive/n4y2022/7605.
 7. Chauhan K., Jani S., Thakkar D., Dave R., Bhatia J., Tanwar S., Obaidat M.S. 2nd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2020 – Conference Proceedings : 2, Bangalore, 05–07 March 2020. Bangalore, 2020. pp. 205-212.
 8. Zabaykin Yu.V., Barykin M.V., Popovich A.E., Zaernyuk V.M. Analiz ekonomicheskoy effektivnosti ispol'zovaniya metodov mashinnogo obucheniya i iskusstvennogo intellekta dlya avtomatizatsii upravlencheskikh resheniy v promyshlennosti [Analysis of economic efficiency of using machine learning and artificial intelligence methods to automate management decisions in industry]. Ponizov'e: National Reserarch, 2024. 208 p.
 9. Zabaykin Yu.V. Barykin M.V. Ekonomika: vchera, segodnya, zavtra. 2024. Vol. 14. № 7-1. pp. 606-617.
 10. Bol'shakov M.A., Khodakovskiy V.A. Inzhenernyj vestnik Dona. 2024. № 11. URL: ivdon.ru/magazine/archive/n11y2024/9630.
 11. Al Jlibawi A.H.H., Othman M.L., Ishak A., Moh Noor B.S., Sajitt A.H.M.S. IEEE Access. 2022. Vol. 10. pp. 3890-3903.
 12. Ferreira J., Pedemonte M., Torres A.I. Computers & Chemical Engineering. 2022. Vol. 161. p. 107756.
 13. Al-Jamimi H.A., BinMakhashen G.M., Saleh T.A. Fuel. 2022. Vol. 322. p. 124088.
 14. Salama M. International Uni-Scientific Research Journal. 2024. Vol. 5. № 16. pp. 119-129.
 15. Yates L.A., Aandahl Z., Richards S.A., Brook B.W. Ecological Monographs. 2023. Vol. 93. № 1. p. e1557.
-



16. Jung Y. Journal of Nonparametric Statistics. 2017. Vol. 30. № 1. pp. 197-215.

17. Hossain M.R., Timmer D. Global Journal of Computer Science and Technology: D. Neural & Artificial Intelligence. 2021. Vol. 21. № 2. pp. 7-13.

Дата поступления: 5.05.2025

Дата публикации: 25.06.2025