УДК 004.9
doi: 10.21685/2227-8486-2023-3-12

ОСОБЕННОСТИ РЕАЛИЗАЦИИ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОСЛЕОПЕРАЦИОННЫХ ОСЛОЖНЕНИЙ С ИСПОЛЬЗОВАНИЕМ БИБЛИОТЕКИ STREAMLIT

О. Ю. Кузнецова[1], Р. Н. Кузнецов[2], А. В. Кузьмин[3]

[1, 2, 3] Пензенский государственный университет, Пенза, Россия
[1] ellekasandra@yandex.ru, [2] nahab007@rambler.ru, [3] a.v.kuzmin@pnzgu.ru

Аннотация. *Актуальность и цели.* MLOps (Machine Learning Operations) является актуальной и важной темой в сфере машинного обучения. Он объединяет практики и процессы, необходимые для эффективной разработки, развертывания и управления моделями машинного обучения. *Материалы и методы.* Для прогнозирования осложнений после хирургического вмешательства разрабатывался пользовательский web-интерфейс с использованием Streamlit. В данной работе применялся конвейер машинного обучения с помощью библиотеки Scikit-learn и создавалось web-приложение с использованием платформы Streamlit с открытым исходным кодом. Это web-приложение обладает простым пользовательским интерфейсом, позволяющим создавать прогнозы осложнений после операции у пациентов. *Результаты*. Был реализован пользовательский интерфейс с использованием библиотеки Streamlit для модели машинного обучения. *Выводы*. В результате были рассмотрены особенности реализации модели машинного обучения с использованием библиотеки Streamlit, разработки пользовательского интерфейса. В качестве примера использован набор данных прогнозирования послеоперационных осложнений.

Ключевые слова: машинное обучение, прогнозирование, послеоперационные осложнения, логистическая регрессия, *k*-ближайших соседей, дерево решений, метод опорных векторов, многослойный персептрон, случайный лес, Streamlit

Для цитирования: Кузнецова О. Ю., Кузнецов Р. Н., Кузьмин А. В. Особенности реализации модели машинного обучения для прогнозирования послеоперационных осложнений с использованием библиотеки Streamlit // Модели, системы, сети в экономике, технике, природе и обществе. 2023. № 3. С. 167–176. doi: 10.21685/2227-8486-2023-3-12

INVESTIGATION MACHINE LEARNING MODEL USING STREAMLIT

O.Yu. Kuznetsova[1], R.N. Kuznetsov[2], A.V. Kuzmin[3]

[1, 2, 3] Penza State University, Penza, Russia
[1] ellekasandra@yandex.ru, [2] nahab007@rambler.ru, [3] a.v.kuzmin@pnzgu.ru

Abstract. *Background*. MLOps (Machine Learning Operations) is a relevant and important topic in the field of machine learning. It brings together the practices and processes needed to effectively develop, deploy, and manage machine learning models. *Materials and methods.* To predict complications after surgery, a Web-based user interface using Streamlit was developed. In this paper, the machine learning pipeline was applied using the Scikit-learn library and a Web application was created using the Streamlit platform, which is

open source. This web application has a simple interface for users that allows you to create forecasts of postoperative complications in patients. *Results*. The user interface was implemented using the Streamlit library for the machine learning model. *Conclusions*. As a result, the features of implementing a machine learning model using the Streamlit library and developing a user interface were considered. A data set for predicting postoperative complications was used as an example.

Keywords: machine learning, forecasting, prediction of postoperative complications, logistic regression, *k*-nearest neighbors, decision tree, support vector machine, multilayer perceptron, random forest, Streamlit

For citation: Kuznetsova O.Yu., Kuznetsov R.N., Kuzmin A.V. Investigation machine learning model using Streamlit. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve = Models, systems, networks in economics, technology, nature and society*. 2023;(3):167–176. (In Russ.). doi: 10.21685/2227-8486-2023-3-12

Введение

Основная актуальность MLOps (Machine Learning Operations) связана с растущим внедрением машинного обучения во многие отрасли, такие как финансы, здравоохранение, розничная торговля, производство и многое другое. Однако просто разработать модель машинного обучения недостаточно. Она также должна быть эффективно внедрена и поддерживаться [1–3].

Практика MLOps привносит модели ML в процесс производства. Это объединяет приложения ML с принципами DevOps (development and operations), где развертывание и обслуживание моделей ML могут быть автоматизированы в производственной среде [4, 5].

Набор инструментов MLOps упрощает управление жизненным циклом ML, обеспечивает надежность и быструю доставку. В этой работе были исследованы широко используемые на практике инструменты:

1. Kubeflow – это проект, созданный компанией Google. Позволяет создавать и масштабировать модели ML. Kubeflow обеспечивает управление моделями машинного обучения поверх Kubernetes, поддерживая этапы разработки, развертывания и мониторинга в течение всего жизненного цикла приложения машинного обучения через автоматизированные рабочие процессы [6]. Но в данном проекте отсутствуют функции управления версиями данных и управления версиями конвейера, что усложняет разработку.

2. MLflow – это не облачная платформа с открытым исходным кодом для управления сквозным жизненным циклом ML, выполняющая основные функции: отслеживание, проектирование, моделирование. MLflow обеспечивает возможность слежения за процессом ML, предоставляя пользователям функциональность для наблюдения за экспериментами, сравнения параметров и конечных результатов моделей ML через ведение записей и запросов всех вводимых и выводимых данных. Этот проект может служить инструментом для инкапсуляции кода ML [7]. Нет возможности развертывания облачной инфраструктуры.

3. DVC (Data Version Control) представляет собой инструментарий, объединяющий управление версиями данных (DVC), непрерывное машинное обучение (CML) и поддерживающие сервисы для управления ML-моделями, наборами данных и экспериментами. Учет различных версий данных становится критически важным в процессе работы, особенно когда объем данных трудно обрабатываем. Кроме того, DVC обеспечивает эффективный совместный

обмен знаниями между командами [8]. В данном инструменте отсутствует возможность настройки гиперпараметров.

4. Streamlit является инструментом на Python, который упрощает и ускоряет создание веб-приложений. Его простота использования не требует сложных настроек. Streamlit обеспечивает возможность итеративного кодирования, при этом результаты можно наблюдать в реальном времени в процессе разработки. При помощи встроенного сервера пользователи могут немедленно развертывать свои приложения в вебе, прослеживая их функционирование через облачные сервисы Streamlit [10].

Целью работы является реализация процесса обучения конвейера машинного обучения с использованием библиотеки Scikit-learn и создания web-приложения с помощью платформы с открытым исходным кодом Streamlit по прогнозированию осложнений после операции на примере желчнокаменной болезни.

Материалы и методы

В исследовании использована анонимизированная выборка пациентов, страдающих желчнокаменной болезнью. Общее число пациентов в выборке составляет 109 человек, из которых 63 пациента не имеют осложнений, а 46 пациентов имеют осложнения. В рамках исследования были собраны данные о различных показателях здоровья у этих пациентов с желчнокаменной болезнью. Эта информация была использована для проведения анализа и изучения связей между этими параметрами и их возможным влиянием на наличие осложнений у пациентов с желчнокаменной болезнью [9]. Приложение машинного обучения было разработано в три этапа:

1. Построение модели машинного обучения (использовался случайный лес, прогнозирование послеоперационных осложнений).

2. Разработка web-интерфейса (библиотека Streamlit).

3. Развертывание web-приложения на платформе Streamlit.

Обычно процесс MLOps начинается с формулировки бизнес-задачи и исследования требований от специалистов в соответствующей области (рис. 1).

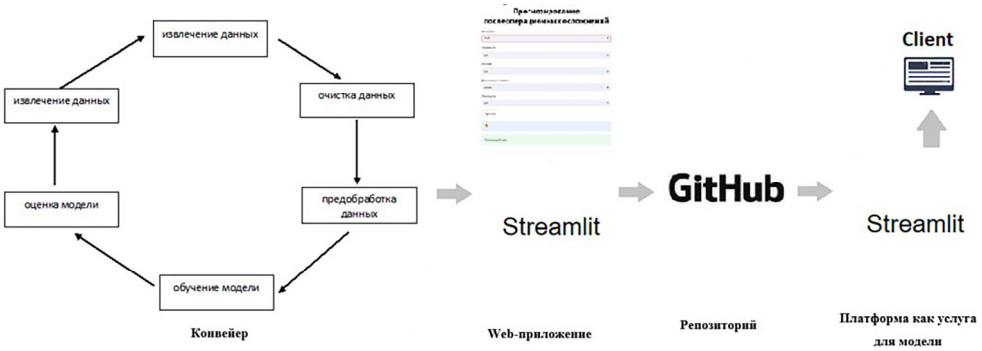

Рис. 1. Архитектура приложения для модели машинного обучения

На основе этих требований дизайнеры определяют тип моделей для разработки, учитывают необходимые функции и выбирают способ сбора данных и их доступность. Соответственно, на каждом из этих этапов для достижения

поставленных целей выполняются операции, такие как обработка данных и настройка модели ML с использованием библиотеки Scikit-learn. Следующей операцией является создание web-приложения для настроенной модели ML, в ней применяется библиотека Streamlit. Так как Streamlit также предоставляет платформу как услугу (PaaS), которая позволяет разрабатывать и развертывать web-приложения без необходимости управления инфраструктурой, web-приложение развертывалось именно на этой платформе. Для развертывания на платформе необходимые исходники кода ML были сохранены в репозитории GitHub. Наконец, пользователю необходима только ссылка на созданное web-приложение для осуществления тестирования пациентов.

Результаты

Первый этап включает серию шагов по предварительной обработке, отбору информативных показателей из входных данных, обучение модели машинного обучения для прогнозирования осложнений. Данные шаги были реализованы в пользовательском интерфейсе с использованием Streamlit [10–14]. Результаты представлены на рис. 2. Это библиотека для быстрой разработки пользовательского интерфейса (UI) для анализа данных и машинного обучения. Она позволяет создавать интерактивные приложения с минимальным количеством кода.

	Hemoglobin	Erythrocytes	Color_index	Leukocytes	Neutrophils	Neutrophils_segmented	Lymph
0	146	4.5	0.97	7.6	9	65	
1	141	4.5	0.94	8.7	2	70	
2	162	4.8	1	8.1	6	80	
3	162.1	4.8	1	9	12	78	
4	142.8	4.6	0.93	12.5	6	48	

Установите флажок на боковой панели, чтобы просмотреть набор данных.

Описание статистических данных

	Hemoglobin	Erythrocytes	Color_index	Leukocytes	Neutrophils	Neutrophils_segmented	Lymph
count	109	109	109	109	109	109	
mean	137.7743	4.3972	0.9341	8.5862	8.7339	62.6881	2
std	18.0734	0.4364	0.0428	4.0164	7.5273	8.676	1
min	69	2.5	0.82	3	1	39	
25%	128	4.3	0.91	5.4	3	58	
50%	140	4.5	0.93	7.5	6	64	
75%	149	4.6	0.97	10.3	11	69	
max	173	5.1	1	20.5	43	80	

Рис. 2. Описательная статистика данных

В данной работе использовался статистический тест X^2 [15] для отбора показателей. Применялась библиотека Scikit-learn [16, 17], которая предоставляет класс SelectKBest, его можно использовать с набором различных

статистических тестов для выбора информативных показателей [16]. Результат работы данного алгоритма представлен на рис. 3.

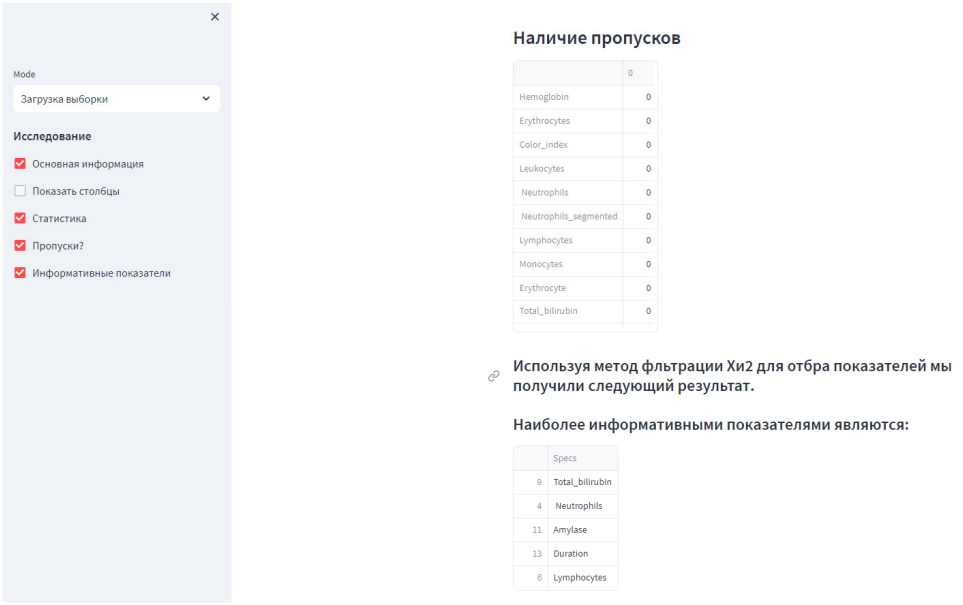

Рис. 3. Отбор информативных показателей

Последним шагом по этапу построение модели машинного обучения является выбор модели. Шесть различных моделей классификации можно протестировать на сайте для получения наилучшей точности прогнозирования послеоперационных осложнений (рис. 4).

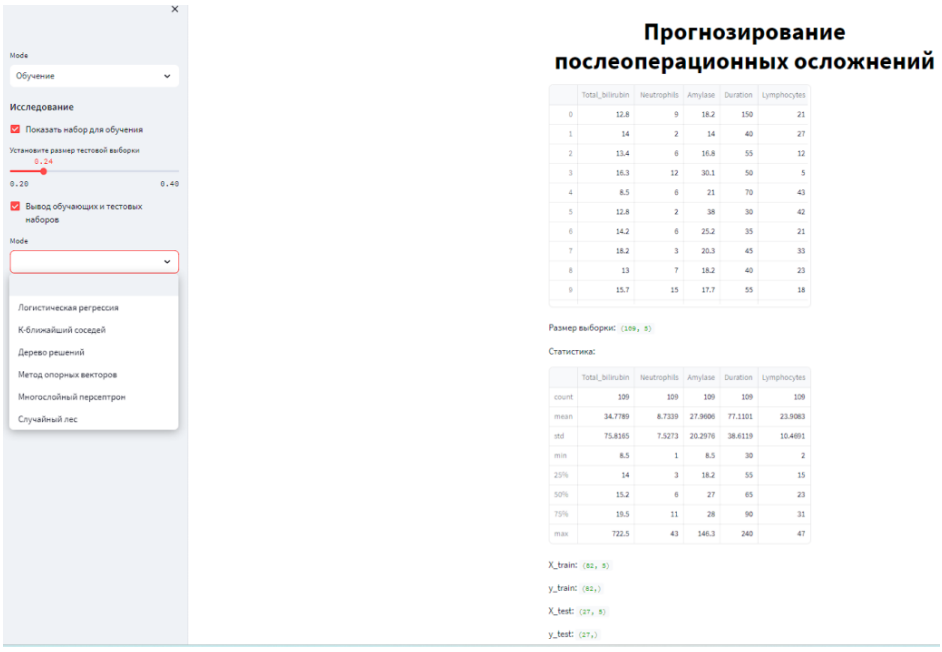

Рис. 4. Выбор модели машинного обучения

На рис. 4 показано, как можно настроить параметры для обучения, это размер тестовой выборки и сам метод обучения. После выбора метода происходит обучение и тестирование модели, также можно вывести матрицу неточности для оценки качества классификации.

На заключительном этапе пользователь может протестировать пациента на настроенной модели.

На рис. 5 показано, как проходит прогнозирование послеоперационных осложнений на новых данных. В зависимости от выбранного метода машинного обучения пользователь может протестировать пациента на разных моделях.

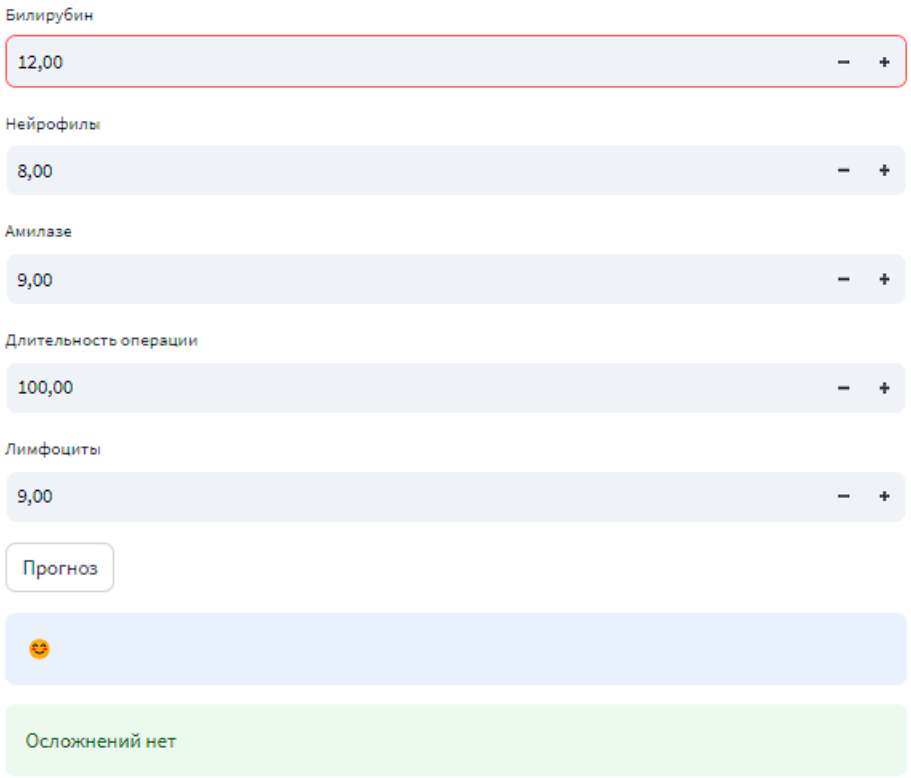

Рис. 5. Прогнозирование осложнений

Обсуждение

Результаты тестирования обученных моделей приведены в табл. 1.

Таблица 1

Оценка точности моделей

Модель	Точность
Логистическая регрессия	66,70
K-ближайших соседей	62,50
Дерево решений	66,70
Жесткий классификатор голосования	75,00
Многослойный персептрон	70,80
Случайный лес	83,30

Сравнение реализации различных классификаторов показало, что наиболее точный классификатор – случайный лес, это позволило повысить точность классификации до 83,30 %. Для реализации аналогичных моделей машинного обучения, представленных в [9], использовались модели с методом обучения жесткого классификатора голосования, модель машинного обучения на основе метода случайного леса дала более точный прогноз, чем жесткий классификатор голосования. Различия в точности модели связаны с различиями в реализации данных методов в библиотеке Scikit-learn.

Выводы

Был проведен обзор платформ альтернативных технологий для MLOps. Развертывание приложений на данных платформах связано с высокими затратами, а также некоторые платформы не предоставляют услуги пользователям из России.

В связи с этим развертывание web-приложений на платформе Streamlit становится актуальным по нескольким причинам. Во-первых, Streamlit предоставляет простой и удобный способ создания пользовательского интерфейса для web-приложений на основе Python. Во-вторых, Streamlit является платформой с открытым исходным кодом, что позволяет разработчикам адаптировать и расширять функциональность с помощью разработчиков сообщества. В-третьих, Streamlit обладает хорошей производительностью и масштабируемостью. Он может легко обрабатывать запросы от одного пользователя или от нескольких пользователей одновременно.

Реализован процесс обучения конвейера машинного обучения с использованием библиотеки Scikit-learn и создания web-приложения с помощью платформы с открытым исходным кодом Streamlit по прогнозированию осложнений после операции на примере желчнокаменной болезни, который позволил повысить точность классификации до 83,30 %.

Список литературы

1. Lathkar M. High-Performance Web Apps with FastAPI: The Asynchronous Web Framework Based on Modern Python. Nanded, Maharashtra, India, 2023. P. 1–309.
2. Singh P. Deploy Machine Learning Models to Production. 2021. P. 1–150.
3. Lane K. The Design of Web APIs. NY : Shelter Island, 2019. P. 1–355.
4. Kreuzberger D., Kühl N., Hirschl S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture // IEEE Access. 2023. January. doi: 10.1109/ACCESS.2023.3262138
5. Clin J. A Comparative Study of Machine Learning Algorithms in Predicting Severe Complications after Bariatric Surgery // The Future of Artificial Intelligence in Clinical Medicine. 2019. Vol. 668, iss. 8 (5). P. 1–27.
6. Документация платформы Kubernetes. URL: https://www.kubeflow.org/docs (дата обращения: 12.05.2023).
7. Документация платформы MLflow. URL: https://mlflow.org/docs/latest/index.html (дата обращения: 12.05.2023).
8. Документация платформы DVC. URL: https://dvc.org (дата обращения: 12.05.2023).
9. Кузнецов Р. Н., Кузнецова О. Ю. Анализ ансамблевых методов классификации для прогнозирования послеоперационных осложнений у больных желчно-

каменной болезнью // Проблемы информатики в образовании, управлении, экономике и технике : сб. ст. XXII Междунар. науч.-техн. конф. (г. Пенза, 9–10 декабря 2022 г.). Пенза : Приволжский Дом знаний, 2022. С. 83–86.
10. Raghavendra S. Beginner's Guide to Streamlit with Python: Build Web-Based Data and Machine Learning Applications. Dharwad, Karnataka, India, 2023. P. 1–203.
11. Khorasani M. Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework. 2022. P. 1–480.
12. Shukla S., Maheshwari A., Johri P., Comparative Analysis of Ml Algorithms & Stream Lit Web Application // 3rd International Conference on Advances in Computing, Communication Control and Networking. Greater Noida, India, 2021. P. 175–180.
13. Parker A., Heflin A., Jones L. C. Analyzing University of Virginia Health publications using open data, Python, and Streamlit // Journal of the Medical Library Association. 2021. Vol. 109 (4). P. 688–689. doi: 10.5195/jmla.2021.1360. PMID: 34858105; PMCID: PMC8608219
14. Gopisetti L. D., Kummera S. K. L., Pattamsetti S. R. [et al.]. Multiple Disease Prediction System using Machine Learning and Streamlit // 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). Tirunelveli, India, 2023. P. 923–931.
15. Горбаченко В. И., Кузнецов Р. Н., Кузнецова О. Ю. Отбор информативных признаков для прогнозирования послеоперационных осложнений при желчнокаменной болезни // Проблемы информатики в образовании, управлении, экономике и технике : сб. науч. ст. по материалам XVI Междунар. науч.-техн. конф. Пенза : Приволжский Дом знаний, 2016. С. 91–97.
16. Документация библиотеки Scikit-learn. URL: https://scikit-learn.org/stable/index.html (дата обращения: 12.05.2023).
17. Документация библиотеки Streamlit. URL: https://docs.streamlit.io/library/get-started/main-concepts (дата обращения: 12.05.2023).

References

1. Lathkar M. *High-Performance Web Apps with FastAPI: The Asynchronous Web Framework Based on Modern Python*. Nanded, Maharashtra, India, 2023:1–309.
2. Singh P. *Deploy Machine Learning Models to Production*. 2021:1–150.
3. Lane K. *The Design of Web APIs*. NY: Shelter Island, 2019:1–355.
4. Kreuzberger D., Kühl N., Hirschl S. Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*. 2023;January. doi: 10.1109/ACCESS.2023.3262138
5. Clin J. A Comparative Study of Machine Learning Algorithms in Predicting Severe Complications after Bariatric Surgery. *The Future of Artificial Intelligence in Clinical Medicine*. 2019;668(8):1–27.
6. *Dokumentatsiya platformy Kubernetes = Kubernetes platform documentation*. (In Russ.). Available at: https://www.kubeflow.org/docs (accessed 12.05.2023).
7. *Dokumentatsiya platformy MLflow = MLflow platform documentation*. (In Russ.). Available at: https://mlflow.org/docs/latest/index.html (accessed 12.05.2023).
8. *Dokumentatsiya platformy DVC = DVC platform documentation*. (In Russ.). Available at: https://dvc.org (accessed 12.05.2023).
9. Kuznetsov R.N., Kuznetsova O.Yu. Analysis of ensemble classification methods for predicting postoperative complications in patients with gallstone disease. *Problemy informatiki v obrazovanii, upravlenii, ekonomike i tekhnike: sb. st. XXII Mezhdunar. nauch.-tekhn. konf. (g. Penza, 9–10 dekabrya 2022 g.) = Problems of informatics in education, management, economics and technology : collection of articles XXII*

International scientific and technical. conf. (Penza, December 9–10, 2022). Penza: Privolzhskiy Dom znaniy, 2022:83–86. (In Russ.)
10. Raghavendra S. *Beginner's Guide to Streamlit with Python: Build Web-Based Data and Machine Learning Applications*. Dharwad, Karnataka, India, 2023:1–203.
11. Khorasani M. *Web Application Development with Streamlit: Develop and Deploy Secure and Scalable Web Applications to the Cloud Using a Pure Python Framework*. 2022:1–480.
12. Shukla S., Maheshwari A., Johri P., Comparative Analysis of Ml Algorithms & Stream Lit Web Application. *3rd International Conference on Advances in Computing, Communication Control and Networking*. Greater Noida, India, 2021:175–180.
13. Parker A., Heflin A., Jones L. C. Analyzing University of Virginia Health publications using open data, Python, and Streamlit. *Journal of the Medical Library Association*. 2021;109(4):688–689. doi: 10.5195/jmla.2021.1360. PMID: 34858105; PMCID: PMC8608219
14. Gopisetti L.D., Kummera S.K.L., Pattamsetti S.R. et al. Multiple Disease Prediction System using Machine Learning and Streamlit. *5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*. Tirunelveli, India, 2023:923–931.
15. Gorbachenko V.I., Kuznetsov R.N., Kuznetsova O.Yu. Selection of informative signs for predicting postoperative complications in cholelithiasis. *Problemy informatiki v obrazovanii, upravlenii, ekonomike i tekhnike: sb. nauch. st. po materialam XVI Mezhdunar. nauch.-tekhn. konf. = Problems of informatics in education, management, economics and technology : collection of scientific articles based on the materials of the XVI international scientific-technical conf.* Penza: Privolzhskiy Dom znaniy, 2016:91–97. (In Russ.)
16. *Dokumentatsiya biblioteki Scikit-learn = Documentation of the library Scikit-learn*. (In Russ.). Available at: https://scikit-learn.org/stable/index.html (accessed 12.05.2023).
17. *Dokumentatsiya biblioteki Streamlit = Documentation of the Streamlit library*. (In Russ.). Available at: https://docs.streamlit.io/library/get-started/main-concepts (accessed 12.05.2023).

Информация об авторах / Information about the authors

Ольга Юрьевна Кузнецова
кандидат технических наук,
доцент кафедры информационно-вычислительных систем,
Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40)
E-mail: ellekasandra@yandex.ru

Olga Yu. Kuznetsova
Candidate of technical sciences,
associate professor of the sub-department of information and computing systems,
Penza State University
(40 Krasnaya street, Penza, Russia)

Роман Николаевич Кузнецов
начальник управления информатизации,
Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40)
E-mail: nahab007@rambler.ru

Roman N. Kuznetsov
Head of the department of informatization,
Penza State University
(40 Krasnaya street, Penza, Russia)

Андрей Викторович Кузьмин
доктор технических наук, профессор,
профессор кафедры информационно-вычислительных систем,
Пензенский государственный университет
(Россия, г. Пенза, ул. Красная, 40)
E-mail: a.v.kuzmin@pnzgu.ru

Andrey V. Kuzmin
Doctor of technical sciences, professor,
professor of the sub-department of information and computing systems,
Penza State University
(40 Krasnaya street, Penza, Russia)

Авторы заявляют об отсутствии конфликта интересов /
The authors declare no conflicts of interests.

Поступила в редакцию/Received 03.07.2023
Поступила после рецензирования/Revised 11.08.2023
Принята к публикации/Accepted 05.09.2023