

WEB-ПРИЛОЖЕНИЕ ДЛЯ ИССЛЕДОВАНИЯ ВЛИЯНИЯ ГИПЕРПАРАМЕТРОВ НА ЭФФЕКТИВНОСТЬ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В РАННЕЙ ДИАГНОСТИКЕ ЗАБОЛЕВАНИЙ

Е.В. Тимошенко, А.Ф. Ражков

Могилевский государственный университет имени А.А. Кулешова

WEB APPLICATION FOR RESEARCH OF THE INFLUENCE OF HYPERPARAMETERS ON THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS IN EARLY DIAGNOSIS OF DISEASES

E.V. Timoschenko, A.F. Razhkov

Mogilev State A. Kuleshov University

Аннотация. В статье представлена программная реализация web-приложения для исследования влияния гиперпараметров на эффективность алгоритмов машинного обучения при решении задач классификации биомедицинских данных и прогнозирования наличия заболеваний как продолжение предыдущих исследований по оптимизации гиперпараметров алгоритмов машинного обучения для задач классификации биомедицинских данных. Web-приложение призвано облегчить процесс настройки моделей, предоставляя удобный инструмент для проведения экспериментов, позволяет загружать набор биомедицинских данных, выбирать алгоритм классификации и задавать значения для соответствующих гиперпараметров. Также оно может быть использовано в качестве инструмента поддержки принятия врачебного решения, обеспечивая более точную диагностику на основе анализа клинических данных пациентов.

Ключевые слова: гиперпараметры, модель машинного обучения, алгоритмы классификации, оценка эффективности, ранняя диагностика заболеваний, прогнозная аналитика, web-приложение.

Для цитирования: Тимошенко, Е.В. Web-приложение для исследования влияния гиперпараметров на эффективность алгоритмов машинного обучения в ранней диагностике заболеваний / Е.В. Тимошенко, А.Ф. Ражков // Проблемы физики, математики и техники. – 2025. – № 1 (62). – С. 113–119. – DOI: https://doi.org/10.54341/20778708_2025_1_62_113. – EDN: JTWROE

Abstract. The article presents a software implementation of a web application for studying the influence of hyperparameters on the effectiveness of machine learning algorithms when solving problems of classification of biomedical data and predicting the presence of diseases as a continuation of previous studies on optimizing hyperparameters of machine learning algorithms for problems of classification of biomedical data. The web application is designed to facilitate the process of setting up models, providing a convenient tool for conducting experiments, making it possible to load a set of biomedical data, select a classification algorithm and set values for the corresponding hyperparameters. It can also be used as a medical decision support tool, providing more accurate diagnosis based on the analysis of the patient clinical data.

Keywords: hyperparameters, machine learning model, classification algorithms, performance assessment, early diagnosis of diseases, predictive analytics, web application.

For citation: Timoschenko, E.V. Web application for research of the influence of hyperparameters on the performance of machine learning algorithms in early diagnosis of diseases / E.V. Timoschenko, A.F. Razhkov // Problems of Physics, Mathematics and Technics. – 2025. – № 1 (62). – P. 113–119. – DOI: https://doi.org/10.54341/20778708_2025_1_62_113 (in Russian). – EDN: JTWROE

Введение

В современной медицине накапливаются огромные объемы данных, содержащих ценную информацию о состоянии здоровья пациентов. При этом существует множество серьезных заболеваний, которые крайне важно идентифицировать на ранних стадиях для своевременного назначения соответствующего лечения и повышения шансов на благоприятный исход. Эффективная обработка и анализ биомедицинских данных играют ключевую роль в выявлении заболеваний на начальных этапах их развития, когда симптомы могут быть не столь очевидными, но

ранняя диагностика и терапия способны значительно улучшить прогноз для пациента.

Методы машинного обучения находят широкое применение для решения задач прогнозирования и классификации в различных сферах, включая медицину. Одним из наиболее эффективных подходов машинного обучения для целей прогнозной аналитики в здравоохранении является классификация, позволяющая на основе набора признаков отнести пациента к одной из заранее определенных категорий, таких как наличие или отсутствие того или иного заболевания.

Однако эффективность алгоритмов машинного обучения во многом зависит от правильного выбора гиперпараметров, определяющих поведение модели. Гиперпараметры представляют собой параметры, значения которых задаются вручную перед обучением модели и не подлежат автоматической настройке в процессе оптимизации, однако они оказывают значительное влияние на способность модели эффективно обобщать данные и демонстрировать высокую точность прогнозирования. Подбор оптимальных значений гиперпараметров часто требует проведения большого количества экспериментов и анализа полученных результатов. Ранее авторами проводились такие исследования, в результате которых были найдены алгоритмы машинного обучения, обладающие наивысшей производительностью, для некоторых наборов биомедицинских данных [1], [2].

В статье представлена программная реализация web-приложения для исследования влияния гиперпараметров на эффективность алгоритмов машинного обучения при решении задач классификации биомедицинских данных и прогнозирования наличия заболеваний. Такое приложение призвано облегчить процесс настройки моделей, предоставляя удобный инструмент для проведения экспериментов. Web-приложение позволяет загружать набор биомедицинских данных, выбирать алгоритм классификации и задавать значения для соответствующих гиперпараметров. Затем производится обучение моделей с различными комбинациями гиперпараметров, и пользователь может оценить их влияние на точность прогнозов, анализируя значения метрик качества классификации, наиболее распространенной из которых является точность - доля правильно классифицированных объектов. Однако в случае несбалансированных наборов данных, когда один класс значительно превосходит по численности другие, точность становится неинформативной и даже вводящей в заблуждение метрикой. Эта проблема особенно актуальна в задачах медицинской диагностики, где ошибки классификации могут иметь серьезные последствия для здоровья пациентов. Например, при прогнозировании редких, но опасных для жизни заболеваний крайне важно минимизировать количество ложноотрицательных результатов, даже если это приведет к некоторому росту ложноположительных.

1 Анализ предметной области и постановка задачи

Согласно данным Всемирной организации здравоохранения, в число наиболее опасных и распространенных заболеваний, приводящих к высокому уровню смертности, входят сердечно-сосудистые заболевания (17,9 миллиона случаев смерти в 2019 году) [3], рак молочной железы

(685 000 случаев смерти в 2020 году) [4], сахарный диабет (1,5 миллиона смертей в 2019 году) [5], хронические болезни почек (1,2 миллиона смертей в 2017 году) [6] и болезни печени (1,3 миллиона смертей в 2017 году) [7]. Ранняя диагностика этих патологий крайне важна для своевременного начала лечения и повышения шансов на благоприятный исход.

Повышение осведомленности населения о серьезных заболеваниях и мотивация людей из групп риска регулярно проходить профилактические медицинские обследования имеют решающее значение для раннего выявления патологий. Ранняя диагностика не только повышает шансы на успешное лечение, но и позволяет существенно сэкономить время и финансовые ресурсы за счет более простых и экономически эффективных схем терапии на начальных стадиях болезни [8]. Одним из перспективных инструментов для достижения этих целей является разработка систем прогнозирования риска и вероятности наличия заболеваний с использованием современных технологий машинного обучения [9].

Предлагаемое web-приложение, реализующее подобную систему диагностического скрининга, не ставит своей целью полностью заменить врача-специалиста при постановке окончательного диагноза. Это было бы неоправданно, учитывая сложность и многофакторность большинства патологических процессов. Однако применение эффективных алгоритмов машинного обучения и глубокого обучения для анализа медицинских данных пациентов позволяет с высокой точностью прогнозировать риски и вероятности развития заболеваний.

Таким образом, web-приложение выступает в роли системы поддержки принятия врачебных решений [10], предоставляя специалистам дополнительный инструмент для более ранней диагностики и стратификации рисков. Для пациентов же оно служит мотивацией для регулярного прохождения обследований, помогая своевременно выявить возможные проблемы со здоровьем на доклинической стадии.

2 Методы и технологии разработки

Для разработки web-приложения был использован стек технологий, включающий Python, Scikit-Learn и Streamlit. Python был выбран в качестве языка программирования, поскольку он предоставляет богатый инструментарий для эффективного решения задач машинного обучения.

Библиотека Scikit-Learn является одним из наиболее популярных и мощных инструментов в экосистеме Python для создания моделей машинного обучения. Она предоставляет широкий спектр алгоритмов для классификации, регрессии, кластеризации и других задач, а также инструменты для предобработки данных, отбора признаков, кросс-валидации и оценки

производительности моделей. В рамках данного проекта Scikit-Learn использовалась для реализации различных алгоритмов классификации. Библиотека позволяет легко настраивать гиперпараметры этих алгоритмов и обучать модели на загруженных пользователем наборах данных.

Фреймворк Streamlit был использован для разработки web-интерфейса в качестве инструмента создания интерактивных web-приложений на Python. Он обеспечивает кэширование вычислений, что позволяет ускорить работу приложения и повысить его отзывчивость.

3 Архитектура web-приложения

Web-приложение было разработано с использованием модульного подхода (рисунок 3.1), разделяя функциональность на отдельные компоненты. Основные модули включают:

Модуль загрузки данных отвечает за загрузку и предобработку наборов данных, предоставляемых пользователем.

Модуль алгоритмов классификации предоставляет широкий спектр алгоритмов классификации, что дает пользователям возможность гибко выбирать и применять оптимальный алгоритм для задачи классификации.

Модуль оценки производительности вычисляет метрики качества для обученных моделей, такие как точность (accuracy), точность (precision), полнота (recall), F1-мера (F1-score), среднеквадратичная ошибка (Mean Squared Error, MSE).

Модуль интерфейса разработан с использованием Streamlit и предоставляет интуитивно понятный пользовательский интерфейс для взаимодействия с приложением.

Web-приложение спроектировано на основе клиент-серверной архитектуры (рисунок 3.2). Приложение используется для построения и развертывания моделей машинного обучения на веб-сервере. Это позволяет пользователям взаимодействовать с моделями машинного обучения в режиме реального времени через web-интерфейс.

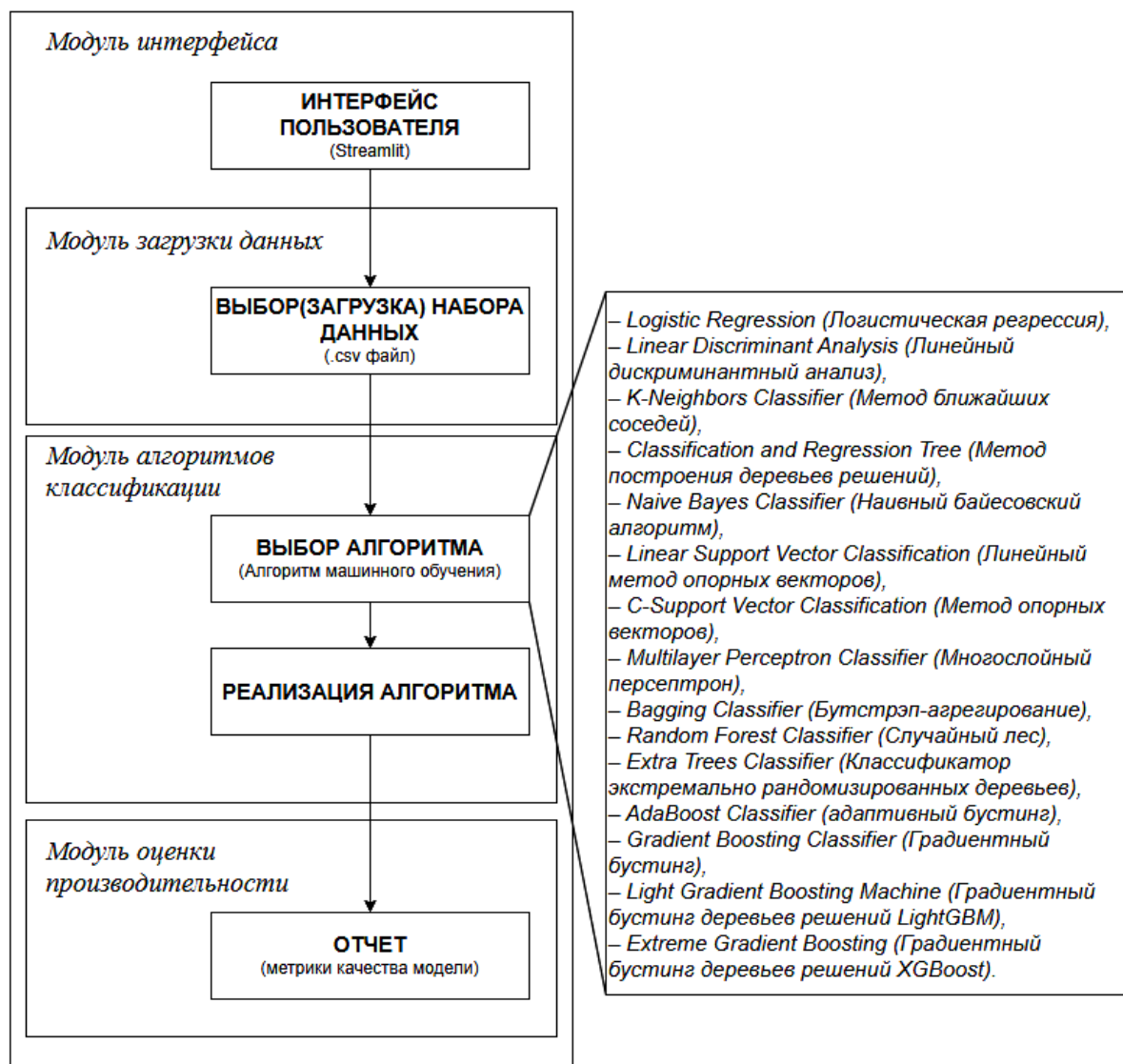


Рисунок 3.1 – Модульный подход

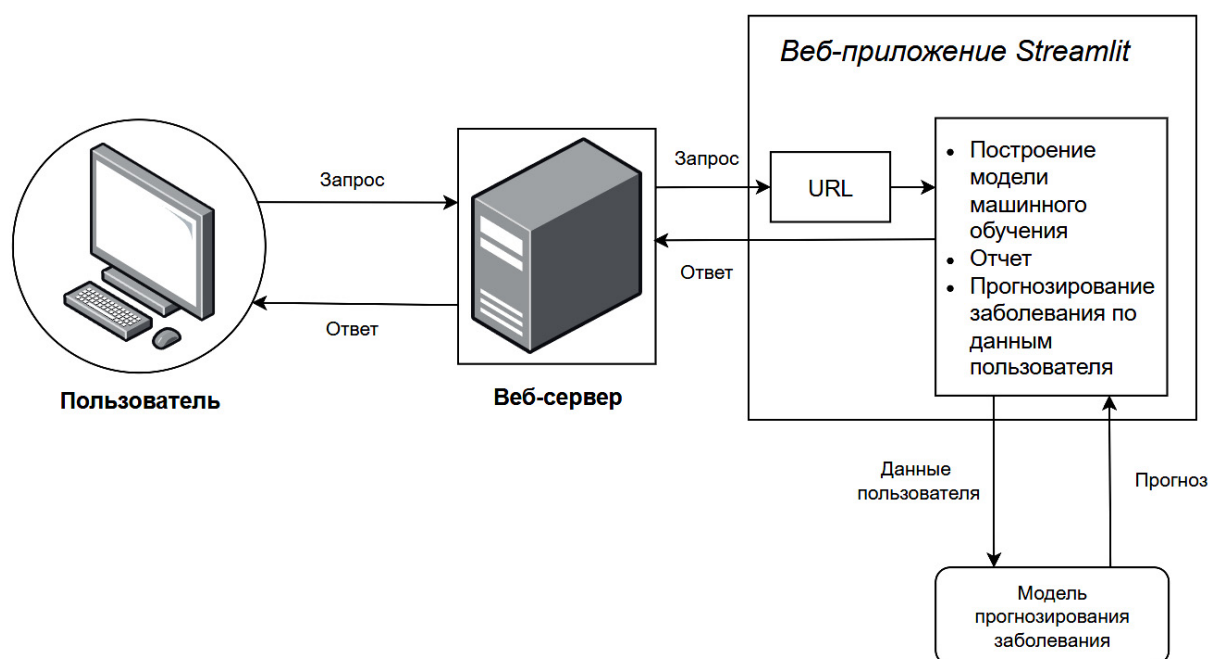


Рисунок 3.2 – Архитектура web-приложения

Такая архитектура обеспечивает модульность и масштабируемость приложения, позволяя легко добавлять новые алгоритмы классификации или метрики оценки производительности в дальнейшем.

Выбранный стек технологий Python, Scikit-Learn и Streamlit обеспечивает гибкость в построении и обучении моделей машинного обучения, а также делает их доступными через интуитивный web-интерфейс, облегчая исследование влияния гиперпараметров на эффективность алгоритмов классификации.

4 Интерфейс web-приложения

Интерфейс разработанного web-приложения соответствует правилам проектирования взаимодействия человека с компьютером [11] для удобства конечного пользователя. Интерфейс является интуитивно понятным, ориентированным преимущественно на пользователей со средним уровнем опыта заполнения форм с данными в интернете. Дизайн приложения также реализован согласно основным правилам проектирования взаимодействия человека с компьютером, поэтому были выбраны оптимальные размер и цвет шрифта, размер и цвет основных функциональных кнопок, размер текстовых полей для ввода данных, выравнивание их содержимого, цвет фона. Используется простой и хорошо организованный вывод информации, где боковая часть интерфейса предназначена для выбора набора данных, алгоритма и его гиперпараметров, а основная часть интерфейса используется для вывода набора данных, метрик выбранной модели машинного обучения и ввода данных пациента

(рисунок 4.1). В каждом поле ввода предусмотрены возможные значения для выбора, а также проверка правильности ввода, чтобы пользователи не смогли отправить неверную информацию.

Для того, чтобы начать работу с приложением, пользователю нужно сперва выполнить следующие действия на боковой панели главной страницы web-приложения: выбрать/загрузить набор данных, выбрать алгоритм и его гиперпараметры (рисунок 4.2).

Разработанное web-приложение, представляющее собой систему прогнозной аналитики на основе моделей машинного обучения, позволяет пользователям загружать собственные наборы данных, где каждая строка соответствует отдельному объекту, описанному набором атрибутов. После загрузки данных осуществляется выбор алгоритма машинного обучения и настройка его гиперпараметров. Обучение моделей машинного обучения происходит на 80% загруженного набора данных, в то время как оставшиеся 20% используются для оценки точности предсказаний. По завершении обучения пользователь получает основные метрики качества модели, такие как точность (accuracy), точность (precision), полнота (recall), F1-мера (F1-score), среднеквадратичная ошибка (Mean Squared Error, MSE). Кроме того, пользователь может ввести атрибуты нового объекта, и приложение предоставит прогноз целевой переменной для этого объекта. Для валидации точности прогнозирования были использованы наборы данных по сердечно-сосудистым заболеваниям, раку молочной железы, диабету, хронической болезни почек, заболеваниям печени.

Приложение совершит обучение выбранной модели и отобразит результаты оценки в виде метрик (рисунки 4.3, 4.4).

Для диагностики пациента необходимо ввести его данные в форму (рисунок 4.5), после чего будет представлена вероятность наличия заболевания (рисунок 4.6).

Набор данных и алгоритмы классификации

Выбрать набор данных:

Сердечно-сосудистые заболе...

Выбрать алгоритм классификации:

(LR) Logistic Regression

Больше информации об алгоритме

Параметры:

penalty

elasticnet

dual

False

tol (0.01-5.0)

1,01

C (0.01-1000.0)

0,01

fit_intercept

False

intercept_scaling (0.01-100.0)

0,01

Машинное обучение в здравоохранении

Прогнозирование сердечно-сосудистых заболеваний

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	Diagnosis
0	63	1	3	145	233	1	0	150	0	2.3000	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5000	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4000	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8000	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6000	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4000	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3000	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5000	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6000	2	0	2	1
10	54	1	0	140	239	0	1	160	0	1.2000	2	0	2	1

Размер набора данных: (303, 14)

Используемый алгоритм классификации: (LR) Logistic Regression

Графики

Метрики модели:

Precision: 0.829
Recall: 0.853
F1-Score: 0.841
Mean Squared Error: 0.18
Accuracy: 81.967 %

Время выполнения: 0.9682 с.

Рисунок 4.1 – Главная страница web-приложения

Набор данных и алгоритмы классификации

Выбрать набор данных:

Сердечно-сосудистые заболе...

Выбрать алгоритм классификации:

(LR) Logistic Regression

Больше информации об алгоритме

Параметры:

penalty

elasticnet

dual

False

tol (0.01-5.0)

1,01

Выбрать алгоритм классификации:

(LR) Logistic Regression

(LDA) Linear Discriminant Analysis

(KNN) K-Neighbors Classifier

(CART) Classification and Regressi...

(NB) Naive Bayes Classifier

(LSVC) Linear Support Vector Clas...

(SVC) C-Support Vector Classificat...

Выбрать набор данных:

Сердечно-сосудистые заболеван...

Рак молочной железы

Диабет

Хроническая болезнь почек

Заболевания печени

Рисунок 4.2 – Выбор алгоритма машинного обучения, набора данных, гиперпараметров

0	63	1	3	145	233	1	0	150	0	2.3000	0	0	1
1	37	1	2	130	250	0	1	187	0	3.5000	0	0	2
2	41	0	1	130	204	0	0	172	0	1.4000	2	0	2
3	56	1	1	120	236	0	1	178	0	0.8000	2	0	2
4	57	0	0	120	354	0	1	163	1	0.6000	2	0	2
5	57	1	0	140	192	0	1	148	0	0.4000	1	0	1
6	56	0	1	140	294	0	0	153	0	1.3000	1	0	2
7	44	1	1	120	263	0	1	173	0	0	2	0	3
8	52	1	2	172	199	1	1	162	0	0.5000	2	0	3
9	57	1	2	150	168	0	1	174	0	1.6000	2	0	2

Рисунок 4.3 – Вывод набора данных

Используемый алгоритм классификации: (LR) Logistic Regression

Графики

Метрики модели:

Precision: 0.829
Recall: 0.853
F1-Score: 0.841
Mean Squared Error: 0.18
Accuracy: 81.967 %

Время выполнения: 0.8519 с.

Рисунок 4.4 – Вывод метрик выбранной модели

Значения пациента

Дополнительная информация

В данном разделе Вы можете ввести свои собственные значения для прогнозирования наличия заболевания. Введите необходимые значения ниже, и Вы получите свой статус на основе этих значений.

1 - Высокий риск

0 - Низкий риск

age (20.0000-86.0000)

20

sex (0.0000-1.0000)

0

cp (1.0000-4.0000)

1

trestbps (76.0000-218.0000)

76

chol (74.0000-616.0000)

74

fbs (0.0000-1.0000)

0

restecg (1.0000-3.0000)

1

Рисунок 4.5 – Ввод данных пациента

Ваш статус:

1 - Вы находитесь в группе высокого риска!

Рисунок 4.6 – Вывод статуса пациента

Заключение

Приложение может найти применение в различных сферах, предоставляя ценные инструменты как для исследователей, так и для практикующих специалистов. Для студентов и исследователей приложение послужит платформой для изучения влияния гиперпараметров на производительность алгоритмов машинного обучения. Оно позволит проводить эксперименты с различными моделями на реальных наборах данных, способствуя глубокому пониманию принципов работы алгоритмов машинного обучения [12]. Для врачей и медицинских работников приложение выступает в качестве системы поддержки принятия врачебных решений. Оно предоставляет дополнительный инструмент для более точной диагностики заболеваний, анализируя клинические данные пациентов [13]. Таким образом, приложение имеет широкий спектр применения, позволяя как развивать научные исследования, так и улучшать практику здравоохранения.

В дальнейшем планируется расширение функциональности web-приложения за счет внедрения гибридных нейросетевых моделей, дополнительных метрик оценки производительности моделей, а также возможности визуализации и интерпретации результатов прогнозирования. Кроме того, будет рассмотрена возможность интеграции web-приложения с медицинскими информационными системами для более эффективного использования в клинической практике.

ЛИТЕРАТУРА

1. Ражков, А.Ф. Оптимизация гиперпараметров алгоритмов машинного обучения для решения задач классификации данных / А.Ф. Ражков, Е.В. Тимощенко // Современное программирование: материалы IV Международной научно-практической конференции (г. Нижневартовск, 08 декабря 2021 года) / отв. ред. Т.Б. Казиахмедов. – Нижневартовск: Нижневартовский государственный университет, 2022. – С. 267–274.
2. Тимощенко, Е.В. Исследование производительности алгоритмов машинного обучения в задачах классификации данных / Е.В. Тимощенко, А.Ф. Ражков // Проблемы физики, математики и техники. – 2023. – № 4 (57). – С. 94–102.
3. *Cardiovascular diseases (CVDs)* [Электронный ресурс] // Всемирная организация здравоохранения. – Режим доступа: <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>. – Дата доступа: 13.03.2024.
4. *Breast cancer* [Электронный ресурс] // Всемирная организация здравоохранения. – Режим доступа: <https://www.who.int/news-room/fact-sheets/detail/breast-cancer>. – Дата доступа: 13.03.2024.

5. *Diabetes* [Электронный ресурс] // Всемирная организация здравоохранения. – Режим доступа: <https://www.who.int/news-room/fact-sheets/detail/diabetes>. – Дата доступа: 13.03.2024.

6. *GBD Chronic Kidney Disease Collaboration*. Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet*. – 2020 Feb 29. – Vol. 395, № 10225. – P. 709–733. – DOI: 10.1016/S0140-6736(20)30045-3.

7. *Mortality from chronic liver disease: Recent trends and impact of the COVID-19 pandemic* / U. Fedeli, Amidei C. Barbiellini, V. Casotto, E. Grande, M. Saia, A. Zanetto, F.P. Russo // *World J Gastroenterol*. – 2023 Jul 14. – № 29 (26). – P. 4166–4173. – DOI: 10.3748/wjg.v29.i26.4166.

8. Lee, S. Early detection of disease and scheduling of screening examinations / S. Lee, H. Huang, M. Zelen // *Stat Methods Med Res*. – 2004 Dec. – № 13 (6). – P. 443–456. – DOI: 10.1191/0962280204sm377ra.

9. *AI in health and medicine* / P. Rajpurkar, E. Chen, O. Banerjee [et al.] // *Nat Med*. – 2028. – № 28. – P. 31–38. – DOI: <https://doi.org/10.1038/s41591-021-01614-0>.

10. Hak, F. Towards effective clinical decision support systems: A systematic review / F. Hak, T. Guimarães, M. Santos // *PLoS One*. – 2022 Aug 15. – № 17 (8). – P. e0272846. – DOI: 10.1371/journal.pone.0272846.

11. *Defining Recommendations to Guide User Interface Design: Multimethod Approach* / C. Diehl, A. Martins, A. Almeida, T. Silva, Ó. Ribeiro, G. Santinha, N. Rocha, A.G. Silva // *JMIR Hum Factors*. – 2022 Sep 30. – № 9 (3). – P. e37894. – DOI: 10.2196/37894.

12. Тимощенко, Е.В. Методы интеллектуального анализа данных в виртуальном практикуме для целей цифровизации образования / Е.В. Тимощенко, А.Ф. Ражков // *Цифровая трансформация*. – 2021. – № 4 (17). – С. 52–62.

13. Malykh, V.L. Decision support systems in medicine / V.L. Malykh // *Program Systems: Theory and Applications*. – 2019. – № 10:2 (41). – P. 155–184. (In Russian). – DOI: 10.25209/2079-3316-2019-10-2-155-184.

Поступила в редакцию 24.06.2024.

Информация об авторах

Тимощенко Елена Валерьевна – к.ф.-м.н., доцент
Ражков Александр Федорович – аспирант