Модель классификации трафика с целью обнаружения роботизированной активности

Д.Э. Шукралиева, А.С. Филатова

Астраханский государственный университет имени В. Н. Татищева, Астрахань

Аннотация: Рассматривается проблема парсинга как растущей угрозы для информационной безопасности. Анализируются методы автоматизированного сбора данных и их последствия, уделено внимание правовым аспектам использования парсинга и юридической ответственности за его осуществление. В качестве решения, основанного на создании и внедрении программного обеспечения с использованием алгоритмов машинного обучения, предлагается метод противодействия с сохранением визуальной целостности страницы – подмена данных для парсеров.

Ключевые слова: парсинг, автоматизированные атаки, защита данных, обнаружение ботов, классификация трафика, машинное обучение, анализ атак, подмена данных, веб-безопасность.

Введение

Парсинг, или скрейпинг, представляет собой один из наиболее распространенных видов автоматизированных атак на веб-ресурсы, целью которых является сбор данных о товарах и ценах с сайтов, особенно в сфере электронной торговли. С каждым годом парсинг становится все более актуальной угрозой для веб-ресурсов. Автоматические боты, способные извлекать информацию с сайтов, создают не только проблемы с доступностью информации, но и ставят под угрозу коммерческие интересы владельцев ресурсов. Несмотря на то, что парсинг не всегда нарушает законы, его последствия могут быть разрушительными.

Парсинг осуществляется через автоматизированные запросы, которые получают доступ к контенту веб-ресурсов. Сервер, согласно правилам протокола HTTP, обязан отвечать на эти запросы, что открывает двери для сбора данных [1]. В отличие от поисковых ботов, которые действуют по согласованию с владельцами сайта и следуют правилам, установленным в файле robots.txt, злонамеренные боты игнорируют эти ограничения и могут деформировать конкурентную среду.

Стратегии парсинга:

Боты извлекают данные, ориентируясь на HTML-код страницы и союзы данных.

При предоставлении web-ресурсами открытых API, которые могут быть использованы для легального извлечения информации, могут быть эксплуатированы злоумышленниками для массового парсинга [2].

Последствия для веб-ресурсов:

Массовые запросы создают излишнюю нагрузку, что может приводить к снижению производительности сайта, тем самым происходит дополнительная нагрузка на сервер.

Утечка пользовательской информации становится реальной проблемой, особенно если злоумышленники получают доступ к чувствительной информации, такой как персональные данные пользователей, интеллектуальная собственность, финансово-значимый контент. В результате утечки конфиденциальная информация и персональные данные могут оказаться на сторонних ресурсах или в базах данных мошенников. Подобные утечки снижают доверие пользователей к добросовестным владельцам web-ресурсов.

Парсинг не всегда является нарушением закона, однако использование полученной информации с нарушением авторских прав, когда контент появится на других ресурсах, может повлечь за собой юридическую ответственность [3].

Рассмотрим 2 судебных дела 2024 года, связанных с нарушением авторских прав с применением парсинга, подчеркивающих важность правовых норм в этой сфере. Первый иск, поданный в федеральный суд Сан-Франциско, утверждает, что OpenAI незаконно копировал текст книг без разрешения правообладателей и, следовательно, не предложил им соответствующее вознаграждение. Второй иск утверждает, что ChatGPT и

DALL·E собирают личные данные пользователей из интернета, тем самым нарушая законы о конфиденциальности.

В качестве мер противодействия парсингу, например, Twitter принял меры по ограничению доступа к своим данным. Что указывает на стремление компаний защитить свои интересы. Такие компании, как Twitter и Reddit, считают данные своей конкурентной стратегией и не желают, чтобы их данные использовались безвозмездно. Это может привести к тому, что компании начнут принимать более строгие меры для ограничения доступа к своим данным, а также искать новые способы монетизации пользовательских данных для обучения моделей ИИ.

В связи с этим становится актуальным вопрос защиты общедоступной информации от несанкционированного копирования. Для эффективного противодействия вредоносным парсерам необходимо разрабатывать современные правовые и технические решения.

Анализ нормативно-правовых актов показал, что парсинг общедоступной информации допустим с точки зрения законодательства [4]. И само определение парсинга в принципе отсутствует. Но практика показывает, что данный процесс очень часто предшествует совершению преступления.

Для защиты от парсинга владельцы web-ресурсов используют следующие методы [5]:

Системы защиты от ботов, например, Active Bot Protection, которые способны блокировать нежелательных ботов на ранних стадиях;

Мониторинг аномального трафика - постоянный анализ трафика с целью обнаружения необычной активности и предотвращения парсинга;

Обфускация страницы, снижающая читабельность исходного кода;

CAPTCHA.

1. Разработка модели классификации сетевого трафика

Для наиболее качественного детектирования действий вредоносных парсеров была предложена классификация трафика [6], представленная на рис.1.

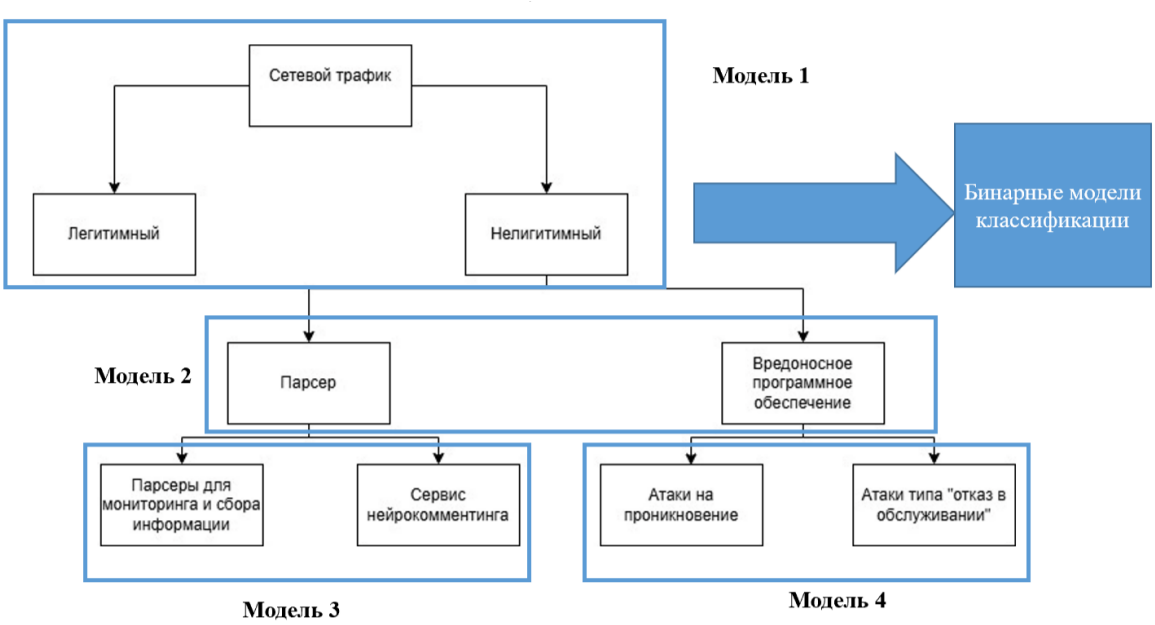

Рис. 1. – Классификация трафика

Для выявления несанкционированного роботизированного процесса парсинга предлагается использовать алгоритмы машинного обучения [7]. Для этого был собран набор данных, который содержит информацию о вредоносном и легитимном трафике во время парсинга, указанный в Таблице №1.

Таблица №1

Описание обучающего множества

Название файла	Тип графика	Количество записей
traffic_dataset.csv	benign_dataset (легитимные образцы)	21873
	parsers_monitoring (образцы парсеров для мониторинга и сбора информации)	8696
	neuro-commenting service (образцы сервисов	7001

	нейрокомментинга)	
	penetration attacks (образцы атаки на проникновение)	6882
	ddos (образцы атаки типа «отказ в обслуживании»)	6993

Анализ литературы показал, что в качестве признаков целесообразно использовать определенные параметры [8]. Они приведенные в Таблице №2.

Таблица №2

Описание признаков и их типов данных

Признак	Описание	Тип данных
len	Количество запросов в сессии	Дискретный тип данных, целочисленное значение (int64)
len_pages	Количество запросов в сессии в страницах (URL заканчивается на .html, .html, php, .asp, .aspx, .jsp	Дискретный тип данных, целочисленное значение (int64)
len_static_request	Количество запросов в сессии в статических страницах	Дискретный тип данных, целочисленное значение (int64)
len_sec	Время сессии в секундах	Непрерывный тип данных, число с плавающей точкой (float64)
len_unique_uri	Количество запросов в сессии, содержащих уникальный URL	Дискретный тип данных, целочисленное значение (int64)
headers_cnt	Количество заголовков в сессии	Дискретный тип данных, целочисленное значение (int64)
has_coockie	Есть ли заголовок Cookie	Бинарный тип данных, логическое значение (bool)
has_referer	Есть ли заголовок Referer	Бинарный тип данных, логическое значение (bool)
mean_time_page	Среднее время на страницу в сессии	Непрерывный тип данных, число с

		плавающей точкой (float64)
mean_time_request	Среднее время на запрос в сессии	Непрерывный тип данных, число с плавающей точкой (float64)
mean_headers	Среднее количество заголовков в сессии	Непрерывный тип данных, число с плавающей точкой (float64)

Были проведены эксперименты, в ходе которых сравнивались разные алгоритмы для построения моделей классификации трафика, представленные в Таблице №3. По полученным результатам были вынесены лучшие решения для моделей – случайный лес, деревья принятия решений [9].

Таблица №3

Сравнительный анализ результатов моделей обучающего множества

Классификаторы машинного обучения	Точность	Полнота	F1-мера	Модель классификации	Параметры
Байесовский классификатор	0,82	0,81	0,82		
Метод опорных векторов	0,87	0,72	0,79		
Деревья принятия решений	0,92	0,98	0,89	Модель 1	X = 0.36131 Entropy = 1.0 Samples = 200 Values = [500/100]
Случайный лес	0,95	0,97	0,96	Модель 2,3,4	bootstrap_features=True max_features < 1.0
Методы бустинга	0,73	0,84	0,82		
CNN	0,65	0,62	0,64		

На основании полученных результатов были построены матрицы ошибок, представленные на рис.5.

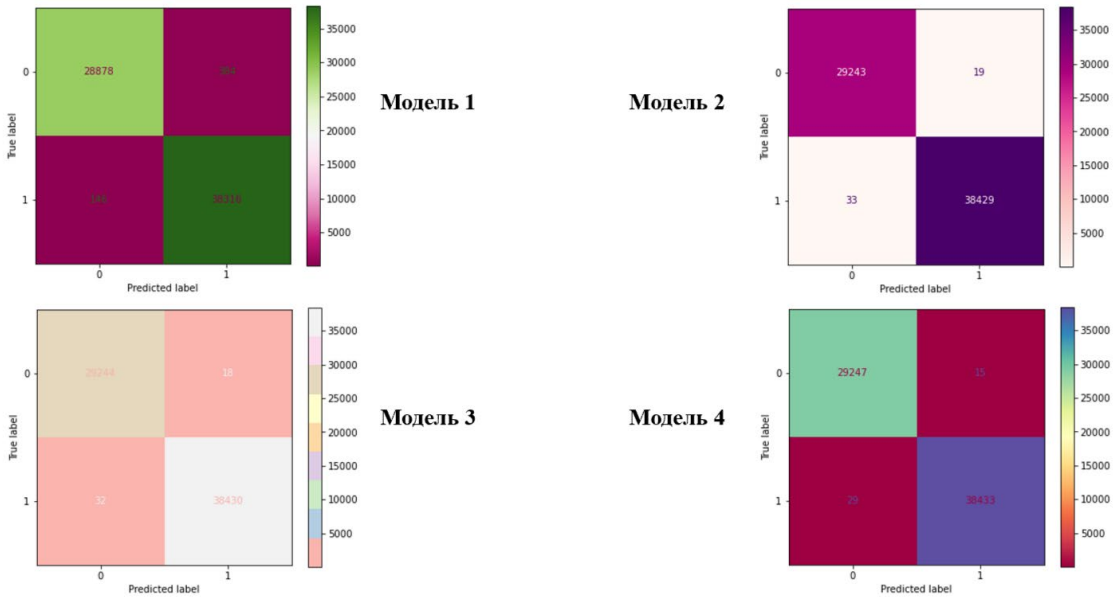

Рис. 5. – Матрица сопряженности, построенные для каждой модели

Доля правильно распознанных образцов трафика Моделью 1, обученной с использованием алгоритма машинного обучения «случайный лес» составила более 89%.

Статистические данные, отражающие процесс обучения моделей 2-4 представлены на Рис.6.

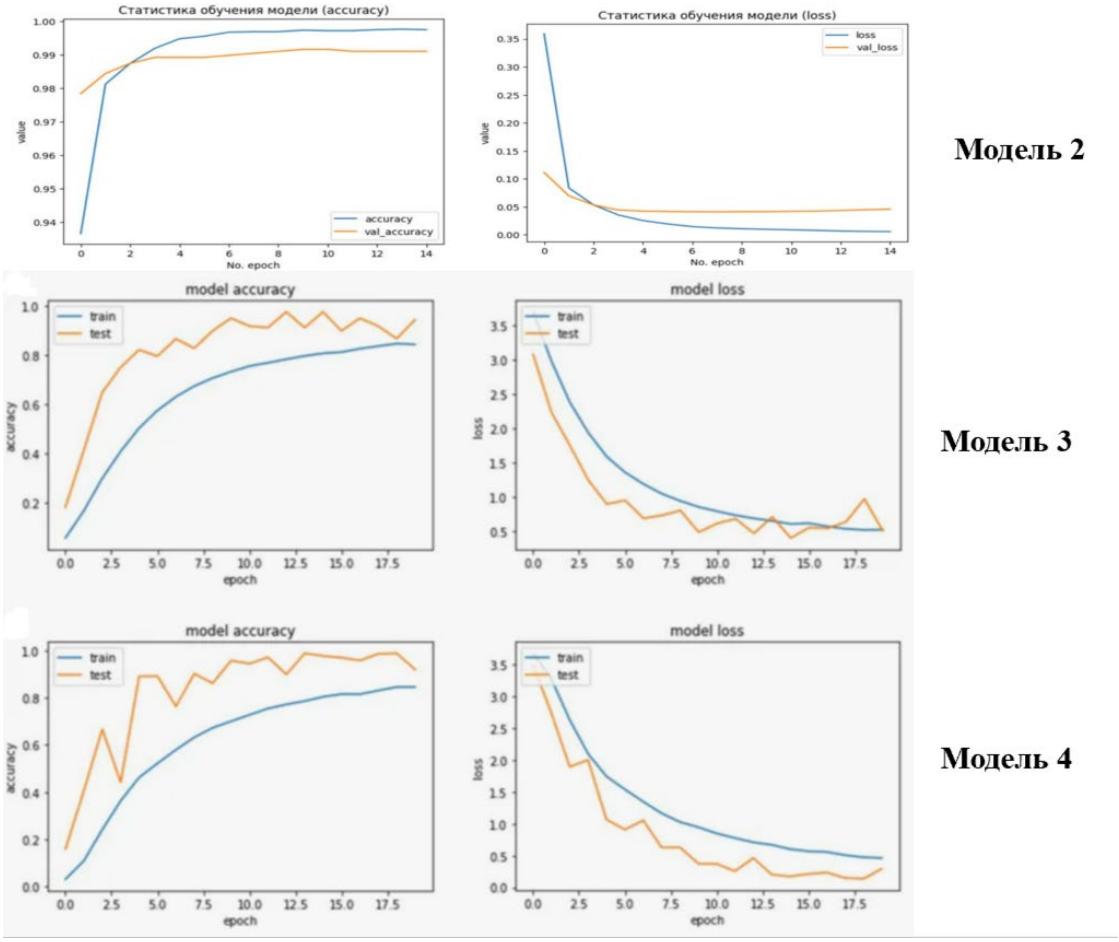

Рис. 6. – Результат обучения моделей с использованием алгоритма «случайный лес»

На основании результатов классификации были разработаны ряд стратегий поведения оператора, представленных в Таблице № 3.

Таблица № 3

Стратегии реагирования на типы трафика.

ТРАФИК	ПОВЕДЕНИЕ ИНФРАСТРУКТУРЫ
Легитимный	Ресурсы работают в обычном режиме
Атаки на проникновение	Центр реагирования на инциденты
Атаки типа «отказ в обслуживании»	Центр реагирования на инциденты, вызов системного администратора
Парсеры для мониторинга и сбора информации	Навязывание ложной информации
Сервисы нейрокомментинга	Проведение анализа стиля сообщения

Отдельно рассмотрим стратегию реагирования «навязывание ложной информации».

2. Навязывание заведомо ложной информации

В различных областях информационной безопасности активно используется методы, основанные на навязывании злоумышленнику заведомо ложной информации. В качестве примера можно привести использование муляжей видеокамер или объектов техники и инфраструктуры.

В приложении «Авито» используется похожая стратегия. В объявлении не указывается достоверный номер телефона продавца или покупателя, но при вызове осуществляется переадресация на правильный номер.

Предлагается расширить данный подход и предложить следующий алгоритм противодействия несанкционированному автоматизированному сбору информации [10]:

– Распознать, что автоматизированный сбор совершают вредоносные парсеры;

– Сформировать «ложные» данные по запросу;

– Передать ложные данные по запросу.

При навязывании ложных данных парсеру внешний вид страницы не должен отличаться от оригинального представления.

При сценарии, когда злоумышленник перебирает web-страницы, адреса которых у него уже есть, где ключевой информацией для него является подробная характеристика об объектах: людях, товарах и т. д. Когда действия злоумышленника будут распознаны, в структуре html-страницы будет создано еще одно поле, в которое будет помещено ложное значение. При этом релевантное значение будет скрыто, а на его месте будет фигурировать ложное поле с ложным значением, что наглядно демонстрирует рис.7.

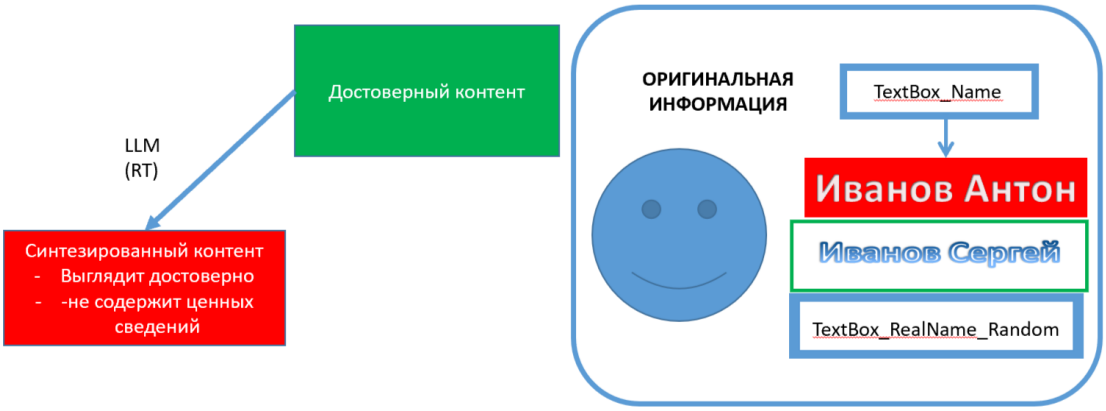

Рис. 7. – Схема навязывания злоумышленнику заведомо ложной информации

Ложное значение, формируется с помощью больших языковых моделей.

3. Внедрение разработанных моделей и подходов к навязыванию заведомо ложной информации в АГУ им. В.Н. Татищева

Web-порталы высших учебных заведений содержит большое число персональных данных, которые могут использоваться для свершений фининговых атак на сотрудников: ФИО руководителей, адреса, телефоны, регалии и.т.п.

При рассмотрении одного из интернет-порталов ФГБОУ ВО «АГУ им. В.Н. Татищева» было обнаружено, что код страниц сайта не содержит в себе никаких признаков защиты от парсинга.

Для организации защиты от несанкционированного сбора информации были внедрены следующие правила работы ресурса:

проводился постоянный мониторинг анализ поступающих запросов на предмет наличия в них признаков парсинга;

в случае обнаружения несанкционированных роботизированных действий код страницы динамически модернизируются с целью навязывания искусственно сгенерированного контента или противодействие роботизированной записи.

В сентябре 2024 года для информационного ресурса «Интернет-портал АГУ» было успешно разработано и внедрено программное обеспечение, которое позволило мониторить и классифицировать трафик.

За время эксплуатации ПО было зафиксировано 4 случая активности вредоносных парсеров. Эти автоматизированные инструменты несанкционированно взаимодействовали с ресурсом, что создало потенциальную угрозу для целостности данных и конфиденциальности пользователей. После подтверждения оператора о действиях вредоносных парсеров внедренным программным обеспечением, были сгенерированы заведомо ложные персональные данные о сотрудниках и студентах АГУ. Всего было сгенерировано 1541 записей.

Своевременное выявление данных угроз позволило повысить защищенность данных ресурсов и, возможно, предотвратить дальнейшие атаки.

Заключение

Парсинг представляет серьезную угрозу для онлайн-бизнеса, и его последствия могут быть разрушительными. Эффективные стратегии защиты помогут минимизировать риски и защитить интересы бизнеса. В условиях постоянной эволюции технологий важно оставаться бдительными и адаптироваться к новым вызовам в области кибербезопасности.

В рамках данной статьи были предложены модели классификации трафика и стратегии реагирования на атаки парсеров. Данные модели легли в основу разработанного программного обеспечения, которое за 2024 год работы в АГУ зафиксировало 4 атаки связанных с несанкционированным роботизированным копированием информации. Несанкционированным парсером было навязано 1541 заведомо ложных записей о сотрудниках и студентах ВУЗа.

Литература

1. Pollard, B., 2019. HTTP/2 in Action. Manning Publications. pp: 3-5. URL: dl.ebooksworld.ir/motoman/Manning.HTTP2.in.Action.EBooksWorld.ir.pdf

2. Демина Р. Ю., Ажмухамедов И. М. Защита web-контента от нелегитимного роботизированного копирования // Вестник ГГНТУ. Технические науки. 2022. Т. 18, № 1. С. 11-17.

3. Демина Р.Ю., Шукралиева Д.Э. Защита веб-контента от несанкционированного автоматизированного сбора информации // Проблемы проектирования, применения и безопасности информационных систем в условиях цифровой экономики. Материалы XXII Международной научно-практической конференции. Ростов-на-Дону: Ростовский государственный экономический университет "РИНХ", 2022. С. 70-75.

4. Flach, P.A., 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, pp: 13-18. URL: cs.put.poznan.pl/tpawlak/files/ZMIO/W02.pdf

5. Chio, C. and D. Freeman, 2018. Machine Learning and Security: Protecting Systems with Data and Algorithms, O'Reilly Media, pp: 32-53. URL: virtualmmx.ddns.net/gbooks/MachineLearningandSecurity.pdf

6. Albon, C., 2018. Machine Learning with Python Cookbook. O'Reilly Media. pp: 35-37.
URL: georgejeswin.github.io/ebooks/bookpy/Machine%20Learning%20with%20Python%20Cookbook_%20Practical%20Solutions%20from%20Preprocessing%20to%20Deep%20Learning%20(%20PDFDrive.com%20).pdf

7. Raschka, S., L. Yuxi and V. Mirjalili, 2022. Machine Learning with PyTorch and Scikit-Learn. Birmingham: Packt Publishing. pp: 2-7. URL: dl.ebooksworld.ir/books/Machine.Learning.with.PyTorch.and.Scikit-Learn.Sebastian.Raschka.Packt.9781801819312.EBooksWorld.ir.pdf

8. Демина Р.Ю., Шукралиева Д.Э. Правовая база защиты web-ресурсов от вредоносного парсинга // Сборник научных трудов II Международной научно-практической конференции: в 6 томах. Казань: Познание, 2023. С. 102-107.

9. Носко В.И. Система автоматизированного построения графа социальной сети // Инженерный вестник Дона, 2012, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015

10. Харламов А.А.; Ермоленко Т.В.; Дорохина Г.В. Сравнительный анализ организации систем синтаксических парсеров // Инженерный вестник Дона, 2013, №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015

References

1. Pollard, B., 2019. HTTP/2 in Action. Manning Publications. pp: 3-5. URL: dl.ebooksworld.ir/motoman/Manning.HTTP2.in.Action.EBooksWorld.ir.pdf

2. Demina R. YU., Azhmukhamedov I. M. Vestnik GGNTU. Tekhnicheskiye nauki. 2022. T. 18, № 1. p. 11-17.

3. Demina R.YU., Shukraliyeva D.E. Problemy proyektirovaniya, primeneniya i bezopasnosti informatsionnykh sistem v usloviyakh tsifrovoy ekonomiki. Materialy XXII Mezhdunarodnoy nauchno-prakticheskoy konferentsii. Rostov-na-Donu: Rostovskiy gosudarstvennyy ekonomicheskiy universitet "RINKH", 2022. pp. 70-75.

4. Flach, P.A., 2012. Machine Learning: The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press, pp: 13-18. URL: cs.put.poznan.pl/tpawlak/files/ZMIO/W02.pdf

5. Chio, C. and D. Freeman, 2018. Machine Learning and Security: Protecting Systems with Data and Algorithms, O'Reilly Media, pp: 32-53. URL: virtualmmx.ddns.net/gbooks/MachineLearningandSecurity.pdf

6. Albon, C., 2018. Machine Learning with Python Cookbook. O'Reilly Media. pp: 35-37. URL:

georgejeswin.github.io/ebooks/bookpy/Machine%20Learning%20with%20Python%20Cookbook_%20Practical%20Solutions%20from%20Preprocessing%20to%20Deep%20Learning%20(%20PDFDrive.com%20).pdf

7. Raschka, S., L. Yuxi and V. Mirjalili, 2022. Machine Learning with PyTorch and Scikit-Learn. Birmingham: Packt Publishing. pp: 2-7. URL: dl.ebooksworld.ir/books/Machine.Learning.with.PyTorch.and.ScikitLearn.Sebastian.Raschka.Packt.9781801819312.EBooksWorld.ir.pdf

8. Demina R.YU., Shukraliyeva D.E. Sbornik nauchnykh trudov II Mezhdunarodnoy nauchno-prakticheskoy konferentsii: v 6 tomakh. Kazan': Poznaniye, 2023. pp. 102-107.

9. Nosko V.I. Inzhenernyj vestnik Dona, 2012, No. 4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015

10. Kharlamov A.A.; Ermolenko T.V.; Dorokhina G.V. Inzhenernyj vestnik Dona, 2013, No. 4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2015.

Дата поступления: 24.05.2025

Дата публикации: 25.07.2025