УДК 004.048 Научная статья

https://doi.org/10.32603/2071-8985-2023-16-10-70-76

Исследование генеалогических деревьев кошек бенгальской породы методами машинного обучения с целью выявления наследственных заболеваний

Н. А. Фомченкова, Я. А. Бекенева✉

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина), Санкт-Петербург, Россия

✉yabekeneva@etu.ru

Аннотация. Методами интеллектуального анализа исследовалась вероятность передачи болезни по наследству с помощью базы данных бенгальских кошек, которые имеют высокую предрасположенность к гипертрофической кардиомиопатии. Для выполнения поставленной задачи была выбрана графовая система управления базами данных Neo4j. Программа, реализующая сбор информации из веб-версии базы данных кошек и дальнейшую обработку полученных данных, реализована на языке Python. Применялись различные подходы для определения статуса заболевания HCM особи по ее родословной. В анализе использовались такие методы, как метод случайного леса, логистическая регрессия и многослойные перцептрон. Эксперименты показали, что самый эффективный подход к решению данной задачи – предсказание связей, а самая эффективная модель среди рассматриваемых моделей-кандидатов – метод случайного леса. На практике были рассмотрены способы решения задач машинного обучения с использованием данных структуры графов.

Ключевые слова: родословная, бенгальская кошка, HCM, граф, Data Mining, база данных

Для цитирования: Фомченкова Н. А., Бекенева Я. А. Исследование генеалогических деревьев кошек бенгальской породы методами машинного обучения с целью выявления наследственных заболеваний // Изв. СПбГЭТУ «ЛЭТИ». 2023. Т. 16, № 10. С. 70–76. doi: 10.32603/2071-8985-2023-16-10-70-76.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Original article

The Study of Genealogical Trees of Bengal Cats Using Machine Learning Methods to Identify Hereditary Diseases

N. A. Fomchenkova, Ya. A. Bekeneva✉

Saint Petersburg Electrotechnical University, Saint Petersburg, Russia

✉yabekeneva@etu.ru

Abstract. The aim of the work is to investigate, using methods of intelligent analysis, the probability of heritable disease transmission, using a database of Bengal cats, which have a high predisposition to hypertrophic cardiomyopathy. The Neo4j graph database management system was chosen to accomplish the task. The program realizing the collection of information from the web version of the cat database and further processing of the obtained data was implemented in Python. Various approaches have been used to determine an individual's HCM disease status from their pedigree. The analysis used such methods as: random forest method, logistic regression and multilayer perceptron. Experiments have shown that the most effective approach to solving this problem is the prediction of relationships, and the most effective model among the considered candidate models is the random forest method. In practice, methods for solving machine learning problems using graph structure data were considered.

Keywords: pedigree, Bengal cat, HCM, graph, Data Mining, database

For citation: Fomchenkova N. A., Bekeneva Ya. A. The Study of Genealogical Trees of Bengal Cats Using Machine Learning Methods to Identify Hereditary Diseases // LETI Transactions on Electrical Engineering & Computer Science. 2023. Vol. 16, no. 10. P. 70–76. doi: 10.32603/2071-8985-2023-16-10-70-76.

Conflict of interest. The authors declare no conflicts of interest.

Введение. История развития человечества неразрывно связана с приручением диких животных. Оно положило начало процессу одомашнивания (доместикация) – изменению диких животных, при котором они на протяжении поколений подвергаются искусственному отбору и изолируются от представителей своей дикой формы. Человек отбирал особей с интересующими его признаками (например, устойчивость к климату, выносливость, плодовитость, качество и количество производимых продуктов) и скрещивал их с другими подходящими кандидатами. Это были первые шаги в истории селекции. Человек еще ничего не знал о генетике, рецессивных и доминантных генах, но шаг за шагом своими действиями «изменял» животных.

С открытием и развитием генетики процесс выведения новых пород с заданными характеристиками получил новый импульс. Люди начали понимать, как правильно осуществить селекцию определенного признака. Проблема же заключалась в том, что в результате множества факторов (например, мутаций в результате близкородственного скрещивания) у животных начали проявляться не только требуемые признаки, но и заболевания.

Хорошо известно, что у большинства породистых кошек и собак имеются заболевания и расстройства, к которым их порода предрасположена. У кошек выявлено более 230 наследственных нарушений и генетической предрасположенности к заболеваниям [1]. На данный момент для многих заболеваний были разработаны тесты, схемы лечения и профилактики заболеваний, но все они работают с уже родившейся особью.

В наши дни изучение генеалогических деревьев играет важную роль в генетике, поскольку оно позволяет определить, какие особенности предков – например, предрасположенности к болезням, склонность к обучаемости определенным навыкам или ее отсутствие, физические характеристики – будут передаваться потомкам. Для изучения генеалогических деревьев можно применять различные методы интеллектуального анализа данных. Они позволят определить предрасположенность потомства к тем или иным заболеваниям и проявлению признаков.

Данная статья посвящена изучению способов применения методов интеллектуального анализа данных к генеалогическим деревьям бенгальских кошек с целью предсказать вероятность появления гипертрофической кардиомиопатии у будущего потомства.

Постановка задачи. Цель публикации заключается в исследовании с помощью методов интеллектуального анализа вероятности передачи болезни по наследству при помощи базы данных бенгальских кошек, которые имеют высокую предрасположенность к гипертрофической кардиомиопатии.

Первые представители данного вида были получены соединением дикой азиатской леопардовой кошки (Prionailurus bengalensis) с домашними кошками, особенно с пятнистой египетской мау. Скрещивание с домашними кошками было необходимо в первую очередь по двум причинам: для получения у потомков более дружественного характера и частого бесплодия самцов раннего поколения. Далее происходило скрещивание особей, полученных на первом этапе. Представителем породы считается бенгал, отстающий от дикого предка более чем на 4 поколения [2]. Данная порода кошек была выведена относительно недавно. Самое раннее упоминание о помеси азиатской леопардовой кошки и домашней кошки относится к 1889 г., но как отдельная порода бенгалы выделились гораздо позже. Создателем современной бенгальской породы считается Джин Милл из Калифорнии. Первые свои эксперименты Милл начала еще в 1960-х гг., а в 1975 г. она получила группу бенгальских кошек, которые уже были выведены для использования в генетическом тестировании в Университете Лома Линда.

Бенгальские кошки, как и представители других искусственно выведенных пород, подвержены большому количеству генетических заболеваний. Одно из таких заболеваний – это гипертрофическая кардиомиопатия (Hypertrophic Cardiomyopathy, HCM), наиболее распространенное заболевание

сердца у кошек, при котором происходит утолщение стенок сердца и уменьшение за счет этого объема левого желудочка, откуда кровь при сокращении сердца поступает в общий кровоток [3]. Также утолщение сердечной мышцы ведет к необходимости увеличения усилий, которые организму приходится прикладывать для выкачивания крови из сердца. Все это приводит к снижению эффективности работы сердца и его быстрому износу, но раннее выявление болезни позволяет увеличить шансы выживания особи. Симптомы заболевания значительно различаются от кошки к кошке, но сердечный шум служит самым распространенным признаком. Данное заболевание считается в большинстве случаев наследуемым, но точно сказать, проявится ли оно у потомка или будет рецессивным, нельзя. Поэтому так важно научиться анализировать генеалогическое дерево особи перед его дальнейшим скрещиванием с помощью различных методов Data mining.

В наши дни IT-технологии развиваются с огромной скоростью, поэтому все чаще во время поиска новых решений для различных задач во всех областях обращаются именно к ним. Например, в медицине активно используют различные виды машинного обучения для решения задачи диагностики заболеваний по результатам проведенных анализов, анамнезу и информации о заболеваниях родственников. Также существуют методы, которые позволяют прогнозировать, с какими заболеваниями пациент может столкнуться в будущем.

Одна из таких программ на вход получает данные о пациенте (данные из его медицинской карты, данные исследований, генеалогическое древо вместе с информацией о заболеваниях предков). Далее программа формирует список возможных заболеваний для этого пациента. Помимо наследственных заболеваний также учитываются те, которые часто встречаются в регионе, где пациент проживает. С помощью сверточных нейронных сетей (convolutional neural network, CNN) программа определяет, насколько вероятно возникновение того или иного заболевания. Если вероятность хотя бы одного заболевания превышает заданное программе пороговое значение, то пациент получает рекомендацию обратиться к специалисту. В дальнейшем врач проверяет результат работы и либо подтверждает его, либо опровергает. Результат проверки врача также вносится в результат работы модели, поэтому данная задача относится к классу задач «обучение с подкреплением», что позволит улучшать качество прогнозирования с увеличением данных для обучения. Разработчики данной программы также предложили использовать CNN-MDRP (multimodal disease risk prediction) для структурированных и неструктурированных данных для повышения точности определения заболеваний [4].

Другое исследование было посвящено соединению уже существующей широко распространенной модели Менделя (Mendelian models), которая оценивает вероятность наличия генетических мутаций у особи, используя родословные и истории болезни членов семьи, и градиентного бустинга. Использование одного метода без другого снижает точность результата. Например, если наблюдаемая семейная история включает в себя неправильно зарегистрированную информацию о заболеваниях, нейронные сети позволят корректно обработать такие данные, что в свою очередь даст преимущество гибридному методу над моделью, которая реализует исключительно правильные законы Менделя без использования нейронных сетей. Аналогично точность будет снижаться и в случае применения только градиентного бустинга, поскольку уже существующие модели (как, например, модель Менделя) включают в себя сложные комплексы закономерностей и математических расчетов, которые позволяют получить искомый результат [5].

Третье исследование также было посвящено попытке предсказать вероятность возникновения у испытуемого рака. Поскольку семейный анамнез служит основным фактором риска для многих видов рака, это заболевание хорошо подходило для исследования. Как и во втором исследовании, разработчики применили традиционный метод анализа с использованием менделевских моделей, и нейронные сети. Разработчики адаптируют полносвязную и сверточную нейронные сети для работы с родословными [6]. Данный метод особенно ценен в ситуациях, когда наблюдаемая семейная история включает неверно зарегистрированные диагнозы рака.

Все три исследования, приведенные ранее, проводились на генеалогических деревьях людей.

Несмотря на то, что существует множество тестов на различные заболевания, планы лечений и профилактики, все это служит методами поиска и лечения болезней у уже родившихся особей. Поскольку разведение породистых животных или животных, пригодных для выполнения конкретных задач (спасатели, поводыри, охранники и т. д.), процесс долгий, финансово затратный и

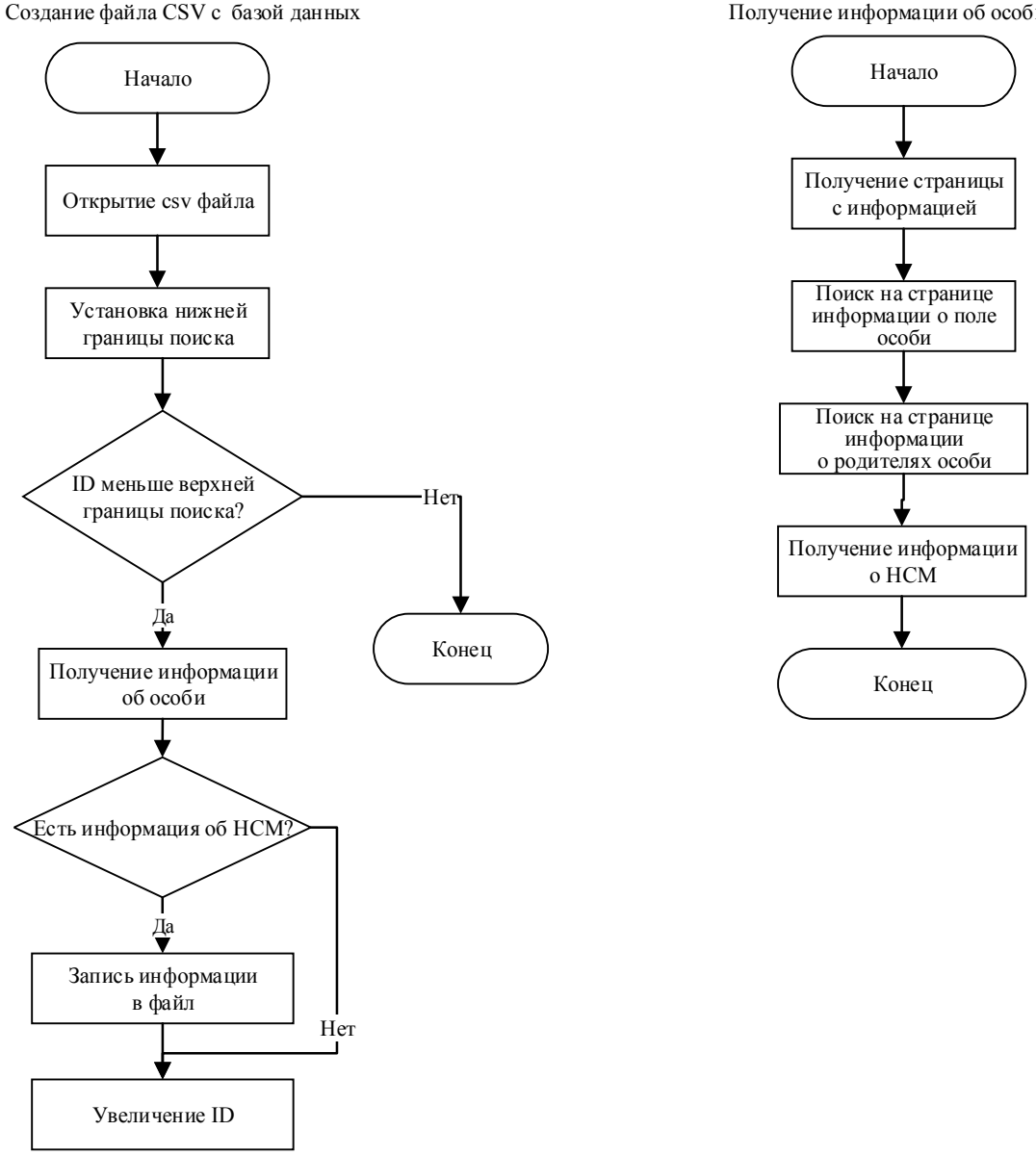
трудоемкий, заводчики крайне заинтересованы в поиске способа определять анамнез и характеристики еще не рожденного потомства. Разумеется, уже сейчас перед скрещиванием пары заводчики производят анализ предков и характеристик будущих родителей, но делается это «на глаз» и основываясь на интуиции. Для повышения надежности подобных оценок возможно использование различных математических методов или методов интеллектуального анализа данных.

Предварительная обработка текста. Для исследования родословных бенгальских кошек была выбрана база данных по адресу https://www.pawpeds.com/cms/index.php. Данная база содержит как информацию о генеалогическом дереве особи, так и ее «карту здоровья». Данная база – не эталон для работы, поскольку многие карты здоровья не заполнены, а часть ветвей дерева обрывается, потому что владельцы не вносили информацию о своих питомцах в эту базу, но при выборе базы данных для анализа были встречены несколько трудностей:

– многие базы разрешают поиск только по индивидуальному номеру особи, который знает только владелец;

– ряд баз доступны только в платном режиме.

При добавлении новой кошки в рассматриваемую базу ей выдается ID. Значение самого первого ID равно 100 000. Для каждой последующей новой добавленной кошки значение ID увеличивается на 1. Таким образом, можно утверждать, что для обхода всех особей в базе достаточно пе-



Алгоритм извлечения базы данных с сайта
Algorithm for extracting a database from a website

ребрать все ID, начиная с первого (равного 100 000) и до последнего выданного, что эквивалентно последней существующей странице.

У каждой особи в базе есть две обязательные страницы. Первая – генеалогическое древо кошки, представленное в виде таблицы с четырьмя столбцами. Тут можно увидеть информацию о четырех поколениях предков особи. Первый столбец в таблице – родители, второй – бабушки и дедушки, и т. д. Вторая – страница с информацией о здоровье кошки, где можно прочитать о проведенных исследованиях и обнаруженных заболеваниях.

Для извлечения данных был написан скрипт на языке Python, который позволяет совершить «обход» (см. рисунок) всех существующих страниц кошек на сайте. Для подготовки данных к анализу использовался файл в формате .csv.

Поскольку в исследовании изучается только заболевание HCM, то любая информация о других заболеваниях будет игнорироваться. Также на этапе извлечения данных не происходит никакой фильтрации значений HCM. Бывают ситуации, когда у особи присутствует два анализа на HCM, тогда программа записывает в файл значение последнего анализа.

После извлечения информации полученная база была дополнительно обработана. Из нее были удалены все кошки, диагноз которых был получен из недостоверного источника, имеющие сомнительный статус CMV, либо чей статус вовсе не был получен или был помечен как «другие диагнозы».

Применение метода классификации к графу. Для начала было решено попробовать применить пайплайн «напрямую», т. е. без каких-либо изменений графа. Рабочий проект на Python был подключен к базе данных Neo4j, в которой хранится граф. Далее была создана проекция графа в памяти программы (рабочего проекта), непосредственно с этой проекцией и будет осуществляться вся дальнейшая работа.

Поскольку свойство hcm – целевое, а такие свойства не могут участвовать в обучении, то из информативных свойств у узлов есть только пол. Для увеличения количества свойств для обучения было решено добавить разные методы предобработки. Для данной задачи подходили только алгоритмы вложения узлов (FastRP, GraphSAGE, Node2Vec, HashGNN), так как они должны отражать топологию графа, отношения между узлами и другую полезную информацию.

Следующим шагом шел процесс обучения моделей-кандидатов и выбор лучшей с применением различных метрик. На этом этапе были применены метрики двух типов: глобальная точность и F1 относительно больных кошек. В качестве кандидатов использовались все три возможных в Neo4j метода машинного обучения: логистическая регрессия, случайный лес и многослойный перцептрон. Некоторые параметры, например penalty в логистической регрессии, были установлены как диапазон, т. е. оптимальное значение должно быть подобрано в процессе обучения.

Применение метрики глобальной точности дало результат свыше 90 %, но необходимо проверить модель на контрольной выборке. Это можно сделать как с помощью Python, применив обученную модель в режиме stream и сравнив полученные значения с полем hcm каждого узла, так и с помощью Neo4j. Было решено воспользоваться вторым вариантом. Для этого каждому узлу контрольной выборки было добавлено новое свойство с классом, полученным с помощью обученной модели. Затем это свойство сравнивалось с целевым свойством с помощью соответствующего запроса.

Видно, что количество верно определенных особей с диагнозом для контрольной выборки равно 0. Это говорит о том, что модель не обучилась, а высокий показатель глобальной метрики – результат того, что модель определяет всю выборку как «здоровую». У этого результата две причины. Во-первых, потому что в изначальном дампе было очень много здоровых кошек относительно больных. Во-вторых, два признака – это очень мало для обучения, к тому же при таком подходе никак не учитывается болезнь самих родителей.

Было решено обучить модель с упором на класс больных кошек, т. е. использовать метрику F1 относительно класса больных особей. Это позволило обучать и тестировать модель так, чтобы она была способна верно определять наличие заболевания. К сожалению, ввиду тех же самых факторов (плохое соотношение больных и здоровых, недостаточное количество признаков, отсутствие информации о болезни родителей) результаты оказались неудовлетворительными, метрика была близка к нулю.

Поскольку применение метода классификации не дало желаемого результата, было предложено искать решение через предсказание связей (link prediction). Для этого к графу, описанному ранее, были добавлены два дополнительных узла, относящиеся к классу hcm, у которых есть одно

свойство status со значениями 1 и 0. К узлу hcm_0 тянулись связи от всех здоровых кошек, а к узлу hcm_1 – от всех больных.

На этапе добавления свойств был использован алгоритм FastRP. Для получения признаков использовался метод cosine. Он вычисляет значение для пары узлов, основываясь на их свойствах.

После добавления моделей-кандидатов они были обучены с использованием метрик aucpr и out_of_bag_error. После перекрестной и итоговой проверок выяснилось, что лучшим оказался метод случайного леса, который показал около 95 % точности.

Однако модель также необходимо проверить на контрольной выборке. Для этого добавим все предсказанные связи между узлами в базу данных и выведем результат для hcm_1. По результатам видно, что модель справилась с задачей определения диагноза гораздо лучше, чем при использовании метода классификации узлов. Многие особи, соединенные с узлом hcm_1, действительно имеют положительный диагноз.

Если смотреть на результат с точки зрения классификации, то на основе контрольной выборки можно высчитать основные метрики. Учитывая, что TP = 14, так как именно столько было верно отнесено к классу больных особей, а TN = = 970, FP = 5, FN = 11 соответственно, то метрики будут равны Acc ≈ 98.3 %, Recall = 56 %, Precision ≈ ≈ 73.7 и F_1 ≈ 63.6 %.

Обсуждение результатов. Результаты проведенного исследования показывают, что лучшим методом для выбранного набора данных является способ предсказания связей. Модель обучилась искать зависимости в графе для определения статуса hcm кошки. Метрика F_1 составила 63.6 %. Лучшей моделью среди моделей кандидатов был признан метод случайного леса.

Выводы и заключение. В ходе выполнения работы были исследованы особенности анализа данных, полученных из родословных бенгальских кошек. Полученные данные в дальнейшем использовались для анализа средствами графовой базы данных Neo4j и языка запросов Cypher. Были использованы различные подходы для определения статуса заболевания HCM особи по ее родословной. В анализе использовались следующие методы: метод случайного леса, логистическая регрессия и многослойные перцептрон. Эксперименты показали, что самым эффективным подходом решения данной задачи является предсказание связей, а самой эффективной моделью среди рассматриваемых моделей-кандидатов – метод случайного леса.

На практике были рассмотрены способы решения задач машинного обучения с использованием данных графовой структуры – логистическая регрессия, случайный лес и многослойный перцептрон. Дальнейшие исследования могут быть направлены на тестирование других методов, повышение качества подготовки данных и повышения точности результатов. Результаты исследования могут использоваться селекционерами для автоматизированного анализа планируемых комбинаций и выявления вероятности получения больных потомков. Полученные результаты могут быть использованы не только для рассмотренного заболевания, но и для ряда других наследственных заболеваний.

Список литературы

1. Feline Hereditary Diseases. URL: https://www.vin.com/apputil/content/defaultadv1.aspx?id=5709928&pid=11372 (дата обращения 23.10.2023).

2. Бенгальская кошка. URL: https://ru.wikipedia.org/wiki/Бенгальская_кошка (домашняя) (дата обращения 23.10.2023).

3. Kittleson M. D., Meurs K. M., Harris S. P. The genetic basis of hypertrophic cardiomyopathy in cats and humans // J. of Veterinary Cardiology, 2015, Vol. 17. P. S53–S73.

4. Disease prediction by machine learning over big data from healthcare communities / M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang // IEEE Access. 2017. Vol. 5. P. 8869–8879.

5. Prediction of hereditary cancers using neural networks / Z. Guan, G. Parmigiani, D. Braun, L. Trippa // The Ann. of Appl. Statistics. 2022. Vol. 16, № 1. P. 495–520.

6. Parmigiani, G., Braun, D. Extending models via gradient boosting: An application to Mendelian models / T. Huang, G. Idos, C. Hong, S. B. Gruber // The Ann. of Appl. Statistics. 2021. Vol. 15, № 3. P. 1126–1146.

Информация об авторах

Фомченкова Наталия Анатольевна – студент гр. 8308 СПбГЭТУ «ЛЭТИ»
E-mail: netta.mer18@gmail.com

Бекенева Яна Андреевна – канд. техн. наук, доцент СПбГЭТУ «ЛЭТИ».
E-mail: yabekeneva@etu.ru
http://orcid.org/0000-0002-7110-6000

References

1. Feline Hereditary Diseases. URL: https://www.vin.com/apputil/content/defaultadv1.aspx?id=5709928&pid=11372 (data obraschenija 23.10.2023).

2. Bengal'skaja koshka URL: https://ru.wikipedia.org/wiki/Bengal'skaja_koshka (domashnjaja) (data obraschenija 23.10.2023). (In Russ.).

3. Kittleson M. D., Meurs K. M., Harris S. P. The genetic basis of hypertrophic cardiomyopathy in cats and humans // J. of Veterinary Cardiology, 2015, Vol. 17. P. S53–S73.

4. Disease prediction by machine learning over big data from healthcare communities / M. Chen, Y. Hao, K. Hwang, L. Wang, L. Wang // IEEE Access. 2017. Vol. 5. P. 8869–8879.

5. Prediction of hereditary cancers using neural networks / Z. Guan, G. Parmigiani, D. Braun, L. Trippa // The Ann. of Appl. Statistics. 2022. Vol. 16, № 1. P. 495–520.

6. Parmigiani, G., Braun, D. Extending models via gradient boosting: An application to Mendelian models / T. Huang, G. Idos, C. Hong, S. B. Gruber // The Ann. of Appl. Statistics. 2021. Vol. 15, № 3. P. 1126–1146.

Information about the authors

Natalia A. Fomchenkova – student gr. 8308 of Saint Petersburg Electrotechnical University.
E-mail: netta.mer18@gmail.com

Yana A. Bekeneva – Cand. Sci. (Eng.), Assistant Professor of Saint Petersburg Electrotechnical University.
E-mail: yabekeneva@etu.ru
http://orcid.org/0000-0002-7110-6000

Статья поступила в редакцию 06.07.2023; принята к публикации после рецензирования 24.10.2023; опубликована онлайн 19.12.2023.

Submitted 06.07.2023; accepted 24.10.2023; published online 19.12.2023.