Т.В. Золотова, А.С. Марунько

Финансовый университет при Правительстве РФ, Москва, Россия

УДК 311.16
DOI: http://dx.doi.org/10.21686/2500-3925-2024-2-72-82

Методы интеллектуальной обработки данных для исследования влияния окружающей среды на заболеваемость населения в Москве

Цель исследования. *Цель исследования состоит в том, чтобы подтвердить или опровергнуть экологическую детерминированность возникновения социально значимых заболеваний у населения Москве на основе анализа данных по экологическим и здравоохранительным показателям в разрезе муниципальных единиц города.*

Материалы и методы. *В статье проведен анализ российской и зарубежной библиографии по проблеме исследования. На основе собранных и обработанных открытых данных по экологическим показателям и по заболеваемости населения в различных районах Москвы были проведены различные виды анализа для выявления взаимосвязи между этими данными. Для классификации социально значимых заболеваний на основе экологических показателей места проживания были построены модели машинного обучения. Математическую основу методов машинного обучения составляют метод k-ближайших соседей, многослойный перцептрон, градиентный бустинг. Для построения моделей использован программный инструмент Jupyter Notebook, поддерживающий язык программирования Python.*

Результаты. *Корреляционно-регрессионный анализ показал, что между некоторыми выбранными экологическими показателями и возникновением социально значимых заболеваний существует статистически значимая корреляция. Данный результат говорит о возможной взаимосвязи, что является одним из главных выводов данной работы. Разработан веб-интерфейс для автоматизации анализа новых данных с помощью построенных моделей машинного обучения, использованных при проведении регрессионного анализа для построения бинарной логистической модели (предсказание на основе собранных данных людей с социально значимыми заболеваниями) и модели мультиклассовой классификации (предсказание на основе собранных данных, какая именно болезнь может быть выявлена у человека). Проведен анализ используемых моделей машинного обучения, определена наилучшая модель для классификации социально значимых заболеваний.*

Заключение. *В результате проведенного исследования удалось собрать полноценную информацию о различных экологических показателях и наличии или отсутствии различных объектов, оказывающих воздействие на окружающую среду. Эти данные были использованы не только в моделях машинного обучения, но и для формирования объективной оценки экологической обстановки муниципальных единиц города Москвы. Поскольку было реализовано автоматическое обновление рейтинга для динамических данных данный результат может быть использован обычными пользователями, не имеющих достаточных квалификаций в экологии и медицине для самостоятельного анализа экологического состояния районов. Считаем, что такие исследования наверняка приведут к эффективным практическим решениям в данной области.*

Ключевые слова: *экология, окружающая среда, здравоохранение, data mining, корреляционно-регрессионный анализ, машинное обучение, автоматизация.*

Tatiana V. Zolotova, Anna S. Marunko

Financial University under the Government of the Russian Federation, Moscow, Russia

Intelligent Data Processing Methods for Studying the Influence of the Environment on the Morbidity of the Population in Moscow

Purpose of the study. *The purpose of the study is to confirm or refute the environmental determinism of the occurrence of socially significant diseases among the population of Moscow based on the analysis of data on environmental and health indexes in the context of municipal units of the city.*

Materials and methods. *The article analyzes Russian and foreign bibliography on the research problem. Based on collected and processed open data on environmental indexes and population morbidity in various districts of Moscow, various types of analysis were carried out to identify the relationship between these data. To classify socially significant diseases based on environmental indexes of the place of residence, machine learning models were designed. The mathematical basis of machine learning methods is the k-nearest neighbors' method, multilayer perceptron, and gradient boosting. To create the models, the Jupyter Notebook software tool, which supports the Python programming language, was used.*

Results. *Correlation and regression analysis showed that there is a statistically significant correlation between some selected environmental indexes and the occurrence of socially significant diseases. This result indicates a possible relationship, which is one of the main conclusions of this paper. A web interface has been developed to automate the analysis of new data using constructed machine learning models used to conduct regression analysis to create a binary logistic model (prediction based on collected data of people with socially significant diseases) and a multiclass classification models (prediction based on collected data, which it is the disease that can be detected in a person). The machine learning models used were analyzed and the best model for classifying socially significant diseases was determined.*

Conclusion. *As a result of the study, it was possible to collect comprehensive information about various environmental indexes and the presence or absence of various objects that have an impact on the environment. These data were used not only in machine learning*

models, but also to form an objective assessment of the environmental situation of municipal units of Moscow city. Since automatic updating of the rating for dynamic data was implemented, this result can be used by ordinary users who do not have sufficient qualifications in ecology and medicine for independent analysis of the ecological state of areas. We believe that such research will certainly lead to effective practical solutions in this area.

Keywords: ecology, environment, healthcare, data mining, correlation and regression analysis, machine learning, automation.

Введение

В связи с недавними эпидемиологическими потрясениями в виде распространения вируса COVID-19 по всему миру государства и общественность стали концентрироваться на решении проблем здравоохранения: развитие здравоохранительных учреждений, просвещение населения касательно правил личной гигиены, более активное проведение профилактических мероприятий и прививочных кампаний, спонсирование медицинских и исследовательских лабораторий. При этом здоровью населения угрожали и не перестают угрожать и прочие заболевания, эпидемии которых необходимо отслеживать и предотвращать. В том числе это относится к болезням, которые законодательно занесены в России в перечень социально значимых: туберкулёз, гепатиты B и C, злокачественные новообразования, сахарный диабет, психические расстройства и расстройства поведения, ВИЧ, инфекции, передающиеся преимущественно половым путем, и болезни, характеризующиеся повышенным кровяным давлением.

Чтобы обеспечить своевременное предотвращение, выявление и лечение данных заболеваний, стоит обратить внимание на факторы, влияющие как на распространение заболеваний, так и на тяжесть их симптомов и последствий. Одним из таких факторов является состояние окружающей среды, которое, как правило, характеризуется повышенными темпами загрязнения в крупных мегаполисах, как Москва. При этом воздух, вода, почва имеют разные степени и причины загрязнения во всех районах Москвы, от чего варьируется его влияние на здоровье местных жителей. Именно поэтому исследовать и контролировать как состояние экологической обстановки, так и заболеваемость населения социально значимыми и прочими болезнями – это важная и актуальная на данный момент проблема.

При этом нарушение установленных норм по сохранению экологического благосостояния общества – это обычное явление на данный момент. Основные проблемы экологического регулирования – это малая стоимость санкций по сравнению со стоимостью обеспечения более экологичной деятельности либо отдельного человека, либо предприятия [1]. Более того, законодательная база охраны окружающей среды в России недостаточна, чтобы регулировать все необходимые области деятельности человека, и чтобы поощрять следование принципам устойчивого развития (например, эксплуатация возобновляемых ресурсов). Также нормы и правила экологического регулирования необходимо регулярно актуализировать для освещения новых выявленных проблем экологической обстановки и для формализации создаваемых впоследствии мер контроля. Таким образом, правовая база России для защиты окружающей среды от неблагоприятных антропогенных факторов сформирована не до конца и имеет ряд недоработок, из-за которых невозможно перейти к урегулированию более сложных экологических проблем, в том числе проблемы влияния экологической обстановки на здоровье населения.

В настоящее время имеется большое количество научных статей, затрагивающих тему загрязнения окружающей среды [1–5] и влияние этой проблемы на здоровье человека, однако полноценных научных трудов действительно не так много. Ю.П. Гичев – один из немногих российских ученых, который исследует уже на протяжении нескольких десятилетий адаптацию человека в экстремальных условиях среды, последствия вредного влияния загрязнения окружающей среды на здоровье человека и развитие экологически обусловленной патологии. Книга Ю.П. Гичева «Экологическая детерминированность основных заболеваний и сокращения продолжительности жизни» на данный момент является актуальным и полным источником информации по данной теме в России [2]. Стоит также упомянуть книгу «Environmental & Pollution Science» от профессоров Университета Аризоны Ian L. Pepper, Charles P. Gerba, Mark L. Brusseau [5], в которой корректно и точно исследованы факторы и процессы загрязнения окружающей среды. Авторы «Environmental & Pollution Science» являются докторами наук по экологической микробиологии и экологической химии, и их труд стал классикой за рубежом по теме загрязнения окружающей среды, увидев уже три издания.

Методы машинного обучения являются сейчас часто используемым инструментом для исследования окружающей среды. Так в работе «Machine Learning for Ecology and Sustainable Natural Resource Management» [6] авторы Grant R. W. Humphries, Dawn R. Magness и Falk Huettmann – специалисты по анализу данных и машинному обучению и по биологии и экологии – подробно описыва-

ют стандартные методы анализа данных и модели машинного обучения, применимые при исследовании окружающей среды.

Несмотря на то, что в национальных стратегиях развитых стран формирование и укрепление здоровья населения является одним из приоритетных направлений, заболеваемость населения некоторыми видами заболеваний не анализируется с точки зрения взаимосвязи с загрязнением окружающей среды той или иной местности. Исключение составляют такие промышленные регионы, как, например, Красноярский край. Проводимые на данную тему исследования как в России, так и за рубежом концентрируются только на исследовании влияния атмосферного загрязнения на заболеваемость населения респираторными заболеваниями [7, 8]. Таким образом, исследователи в данной области не могут быть уверены в существовании экологической детерминированности социально значимых заболеваний. Однако понимание причин происхождения заболеваний – это ключевой фактор для их своевременной профилактики, выявления и лечения.

Также в Москве внедрена развитая система для осуществления мониторинга экологических показателей по всему городу: качество воздуха и воды, шумовое загрязнение, состояние зеленых насаждений, а также метеоданные. Данные измерения ежедневно агрегируются и отображаются на Портале открытых данных Правительства Москвы. При этом в свободном доступе отсутствуют какие-либо инструменты для анализа и оценки экологической обстановки территории, которые бы автоматически обновлялись и учитывали регулярно дополняемые открытые данные. В свою очередь это позволило бы местным жителям получать на регулярной основе актуальную оценку состояния окружающей среды в месте их проживания и принимать на ее основе решения, например, в отношении профилактических мероприятий. Более того, необходимо учитывать, что в крупных мегаполисах на экологию, в основном, влияют антропогенные факторы, которые могут очень быстро изменять состояние локальной биосферы [9], поэтому отсутствие регулярных оценок может негативно повлиять на благосостояние населения при радикальных и/или быстрых переменах в окружающей среде.

Оценки экологической обстановки на территории Москвы в открытом доступе, в основном, состоят из нерегулярных исследований и научных публикаций, в которых «экорейтинги» или иные виды оценок создаются вручную специалистами-экологами. Например, поиск актуального экорейтинга районов Москвы приводит к анализу, проведенному в 2020 году группой компаний по экологической экспертизе. Результатом данного исследования является экологическая карта Москвы, с помощью которой пользователи могут ознакомиться с оценками экологической обстановки по муниципальным единицам. Однако в данной карте не реализованы автоматические обновления при получении новых данных по критериям, отобранным для формирования оценок (например, новые потенциально опасные объекты или новые данные по загрязнению атмосферного воздуха), а также отсутствует возможность ознакомиться с алгоритмом оценки. Данная карта не является комплексным инструментом для проведения анализа экологической обстановки и может быть использована только для простейшего анализа состояния окружающей среды местными жителями, который не будет основан на актуальных данных. Более того, оценка каждого района была рассчитана как среднее арифметическое всех критериев, в то время как более эффективным методом оценки является применение векторов коэффициентов приоритета для каждого параметра [10].

Автоматизированные решения в открытом доступе для оценки предрасположенности местных жителей к социально значимым или иным заболеваниям также не были найдены. Подобная медико-экологическая аналитика так же, как и оценка экологической обстановки, производится в отдельных исследованиях для специфичных районов России, отраслей промышленности или видов заболеваний [10–12].

Таким образом, на данный момент собирающиеся регулярно открытые данные по экологическим показателям не учитываются в проведенных исследованиях, и пользователи не имеют доступа к актуальным исследованиям и оценкам состояния окружающей среды и рисков заболевания социально значимыми болезнями в их месте проживания. Создание полностью или частично автоматизированных решений для проведения подобных оценок экологической обстановки и предрасположенности населения к заболеваниям в рамках такого мегаполиса, как Москва, позволит определить наличие и силу влияния состояния окружающей среды на заболеваемость населения, а также даст понимание о необходимости профилактических процедур или своевременного медицинского обследования.

Исследование взаимосвязи состояния окружающей среды в муниципальных единицах Москвы и заболеваемости местного населения болезнями из перечня социально значимых

На основе показателей, взятых из открытых данных и свидетельств местных жите-

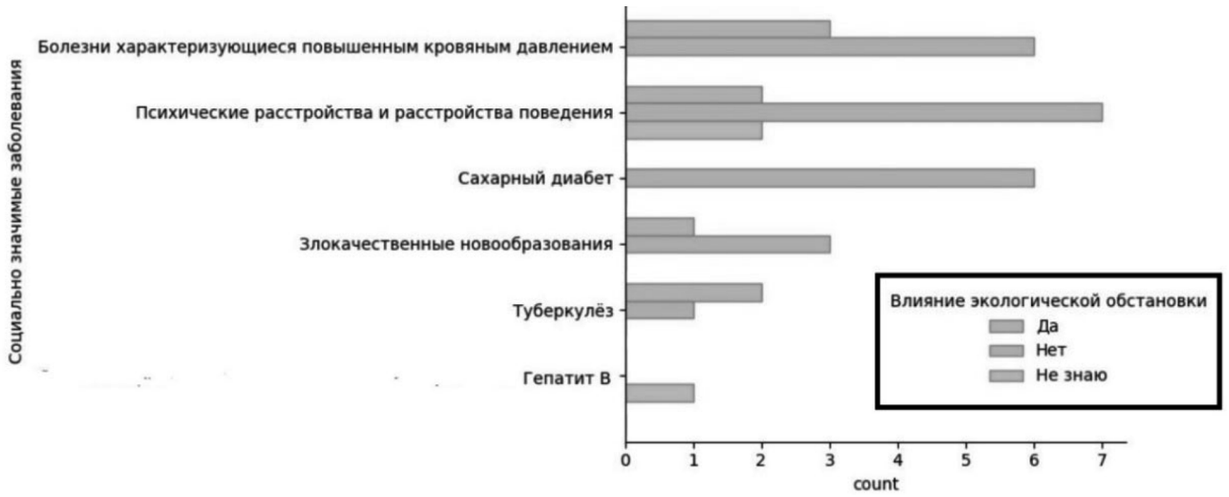

Рис. 1. Диагностированные социально значимые заболевания у населения по результатам опроса
Fig. 1. Diagnosed socially significant diseases in the population based on survey results

лей, можно утверждать о наличии загрязнения окружающей среды в Москве. Вместе с тем имеет место заболеваемость местных жителей социально значимыми болезнями. Возникает вопрос как связаны между собой эти две проблемы, т.е. имеется ли взаимосвязь между заболеваемостью населения и загрязнением окружающей среды. К исследованию данного вопроса мы и переходим.

Перед проведением анализа необходимо было собрать данные по экологическим показателям и по выявленным у населения социально значимым заболеваниям. В источниках открытых данных практически невозможно найти подробную информацию по заболеваемости населения Москвы, тем более в разрезе по районам города, поэтому было принято решение собрать данные путем проведения социологического опроса на онлайн-платформе. С опросом онлайн можно ознакомиться по данной ссылке: https://forms.gle/e1QatSiww6QgP4hz6. Главной целью опроса было определить следующие параметры: пол, возраст, район проживания и наличие социально значимых заболеваний. По результатам опроса удалось опросить респондентов из 71 района Москвы (~50%). Самые распространенные заболевания: болезни, характеризующиеся повышенным кровяным давлением, психические расстройства и сахарный диабет (рис. 1).

При изучении полученных ответов было выявлено, что опрашиваемые не всегда правильно интерпретировали структуру раздела о заболеваниях. Поэтому, в первую очередь в начале предварительной обработки данных, полученных по итогам анкетирования, необходимо было исправить искажения, пустые ответы (N/A), так как в случае отсутствия диагностированного заболевания ответы о влиянии экологической обстановки на возникновение заболевания не являются применимыми.

По каждому экологическому показателю была применена отдельная методология для сбора данных. По Превышению ПДК загрязняющих веществ в районах Москвы отбирались вещества, которые способны вызывать заболевания или интоксикацию организма. например, из более 15 измеряемых показателей загрязнения атмосферного воздуха были выделены 5 таких, которые могли бы подтвердить экологическую детерминированность рассматриваемых заболеваний, а также наиболее полно были отражены в

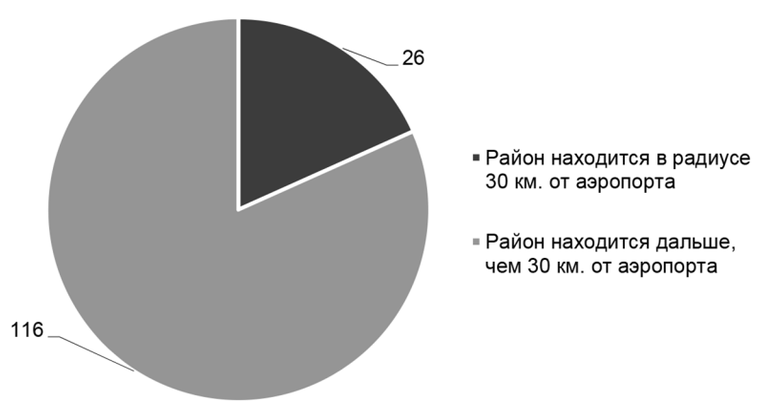

Рис. 2. Наличие аэропорта в радиусе 30 километров от районов
Fig. 2. Availability of an airport within a radius of 30 kilometers from the districts

разрезе всех районов Москвы: оксид углерода, диоксид серы, сероводород, диоксид азота, оксид азота. При этом районы под действием влияния аэропортов, как негативного фактора, который надо учитывать, определялись согласно ФЗ от 01.07.2017 N 135, устанавливающего радиусы приаэродромной территории. Таким образом, считалось, что район находится под влиянием аэропорта, если он расположен в радиусе 30 км от него (рис. 2). Также сбор данных по критериям включал в себя изучение официальных перечней различных зон/объектов: перечень промышленных зон, список ТЭЦ Москвы, список очистных сооружений и т.д.

Таким образом, для проведения оценки состояния окружающей среды по каждому району Москвы было выделено 11 различных параметров, информация по которым собиралась либо с помощью источников открытых данных, либо самостоятельно и вручную.

Для выявления экологической детерминированности социально значимых заболеваний был проведен корреляционно-регрессионный анализ: сначала проведено изучение корреляций между выделенными экологическими параметрами и заболеваемостью, а затем построена регрессионная и другие виды предиктивных классифицирующих моделей для дополнительного исследования возможной взаимосвязи.

Два полученных набора данных – экологические показатели и показатели заболеваемости, собранные в рамках социологического опроса – были объединены в единый набор. При этом по районам проживания опрошенных определялись соответствующие, наиболее актуальные экологические показатели. При изучении полученного набора данных, было обнаружено, что он не является сбалансированным, так как

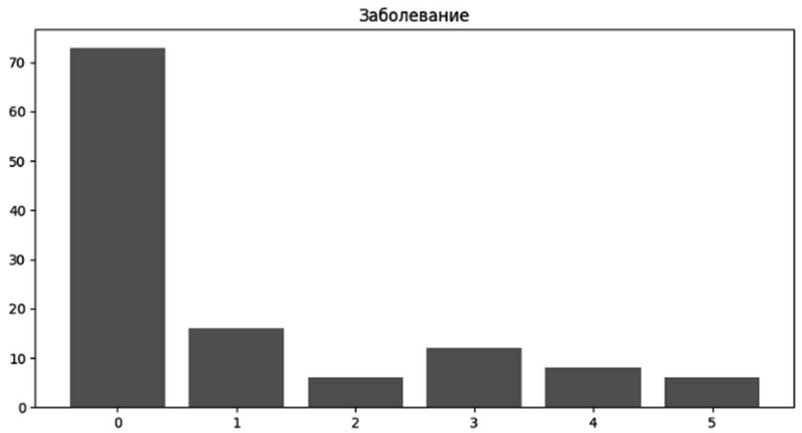

Рис. 3. Начальное распределение классов социально значимых заболеваний

Fig. 3. Initial distribution of classes of socially significant diseases

класс, у которого отсутствуют выявленные социально значимые заболевания (нулевой класс), количественно превышал остальные классы более чем в два раза (рис.3).

Такая несбалансированность могла оказать негативное влияние на результаты анализа. Поэтому, сохраняя первоначальное распределение данных, были добавлены новые искусственные записи в примерно таком же распределении (oversampling) с помощью алгоритма Synthetic Minority Oversampling Technique (далее – SMOTE) [13].

После балансирования классов в выборке был посчи-

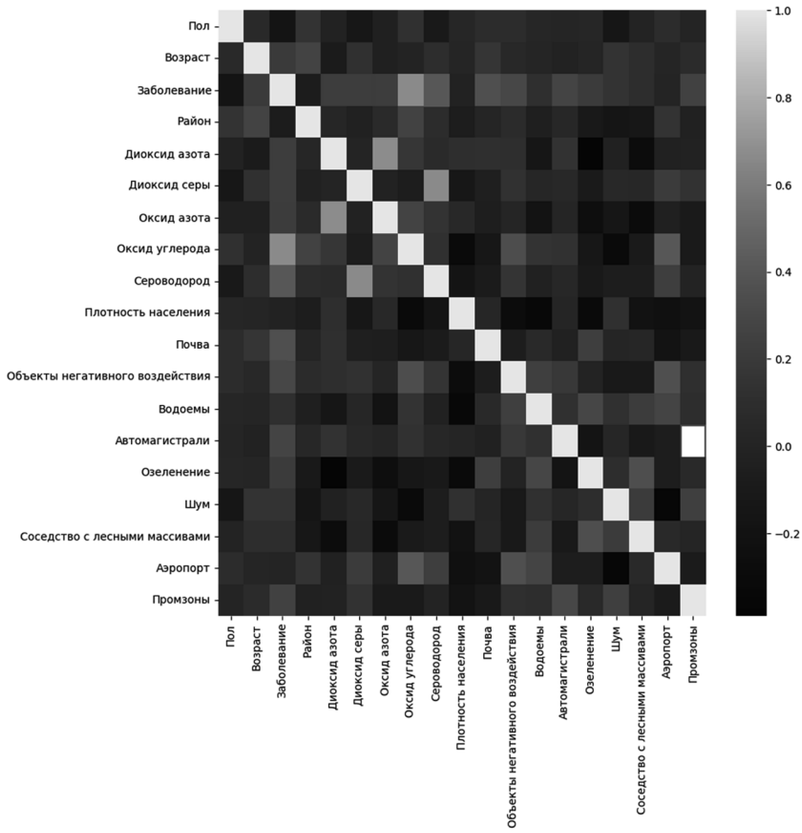

Рис. 4. Тепловая карта коэффициентов корреляции Пирсона

Fig. 4. Heat map of Pearson correlation coefficients

тан коэффициент корреляции Пирсона

$$r_{xy} = \frac{\sum\left(x_i - M_x\right)\left(y_i - M_y\right)}{\sqrt{\sum\left(x_i - M_x\right)^2\left(y_i - M_y\right)^2}}$$

для количественной оценки связи между заболеваемостью и состоянием окружающей среды.

При изучении тепловой карты полученных корреляций (рис. 4) по столбцу/строке «Заболевание» было отмечено присутствие более высоких корреляций между наличием заболеваний и экологическими параметрами: например, загрязняющие вещества в атмосферном воздухе, показатели загрязнения почвы, наличие объектов негативного воздействия и автомагистралей.

Проведен расчет *p*-значения и *t*-критерия Стьюдента (метод статистической проверки гипотез), чтобы убедиться, что коэффициенты корреляций статистически значимы:

$$t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}},\ \alpha = 0{,}1.$$

В рамках расчета *t*-критерия Стьюдента были применены следующие гипотезы: H_0. Корреляция между двумя параметрами равна нулю; H_1. Между двумя параметрами существует статистически значимая корреляция.

По результатам проверки гипотез был сделан вывод о существовании статистически значимой корреляции между выявленными социально значимыми заболеваниями у населения и возрастом, наличием загрязняющих веществ в атмосферном воздухе (в основном, оксида углерода и сероводорода), загрязнением почвы, наличием объектов негативного воздействия, в том числе автомагистралей и промышленных зон.

Поскольку корреляционный анализ выявил существующую взаимосвязь, то далее был праведен регрессионный анализ с применением методов машинного обучения для построения бинарной логистической модели (предсказание на основе собранных данных людей с социально значимыми заболеваниями) и модели мультиклассовой классификации (предсказание на основе собранных данных, какая именно болезнь может быть выявлена у человека).

Для проведения первой части регрессионного анализа столбец «Заболевание» был приведен к бинарному формату, где 1 – наличие социально значимого заболевания у респондента, а 0 – отсутствие заболевания. Затем была построена модель логистической регрессии с использованием метрики ROC-AUC или ROC-кривой, которая позволила оценить, насколько хорошо ранжируются значения из выборки на два класса, а не их абсолютные значения [6, 14]:

$$\frac{TPR}{FPR},\ TPR = \frac{TP}{TP+FN},$$

$$FPR = \frac{FP}{FP+TN}.$$

ROC-кривая отображает соотношение True positive rate (*TPR*) и False positive rate (*FPR*) по результатам классификации.

Вероятность принадлежности объекта выборки к первому классу (наличие заболевания) выражалась через уравнение логистической регрессии:

$$P = \frac{e^{a+bx}}{1+e^{a+bx}}.$$

Таким же образом была решена задача мультиклассовой классификации: чтобы определить наиболее эффективный способ решения такой задачи, были выбраны три модели с разными подходами к классификации [14]:

KNeighborsClassifier (метод *k*-ближайших соседей): оценка сходства объектов на основе их расстояния друг от друга в пространстве признаков;

MLPClassifier (многослойный перцептрон): данная модель будет представлять простейшую полносвязную нейронную сеть, так как для более комплексных нейронных сетей недостаточно данных;

CatBoostClassifier (градиентный бустинг): данная модель уже была использована в рамках первой задачи и будет представлять метод ансамблирования деревьев решений.

Для каждой из этих моделей была использована метрика precision, которая поддерживает мультиклассовую классификацию, а также подходит для несбалансированных датасетов [15]. Несмотря на уже примененный алгоритм SMOTE, набор данных остался отчасти несбалансированным, так как было необходимо сохранить отображение истинного выявленного распределения заболеваемости среди населения. Формула для метрики precision:

$$P = \frac{TP}{TP+FP}.$$

Модели с самыми высокими показателями метрики по результатам обучения могут быть объединены в единый ансамбль для получения более точных предсказаний о предрасположенности населения к социально значимым болезням.

Для проведения обучения с различными комбинациями параметров моделей машинного обучения и выбора оптимального варианта был использован поиск по сетке (Grid Search).

В рамках поиска по сетке для бинарной классификации была выявлена лучшая вариация модели логистической регрессии с параметрами C = 100, solver = saga, максимальное количество итераций = 400 с показателями ROC-AUC на обучающей выборке 0.8492 и на тестовой – 0.8313 (рис. 5). Аналогично было выполнено моделирование второго способа решений бинарной классификации ансамбля деревьев решений с по-

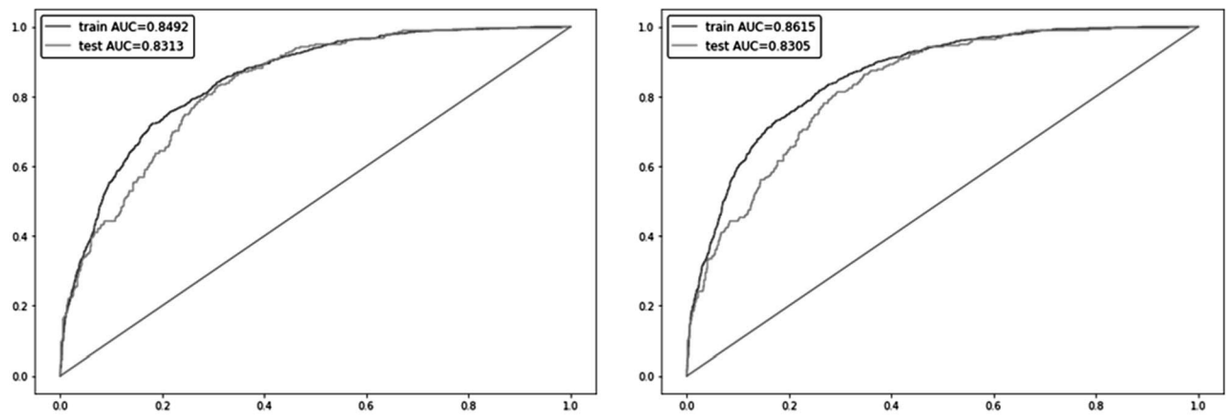

Рис. 5. Показатели метрики ROC-AUC по результатам обучения логистической регрессии (слева) и CatBoostClassifier (справа)

Fig. 5. Indexes of the ROC-AUC metric based on the results of training logistic regression (left) and CatBoostClassifier (right)

мощью CatBoostClassifier (библиотека CatBoost). По итогам поиска по сетке самая результативная модель (с параметрами скорость обучения = 0.1, глубина деревьев = 4, параметр регуляризации листьев деревьев (l_2) = 1, максимальное количество итераций = 400) показала ROC-AUC на обучающей выборке 0.8615 и на тестовой – 0.8305 (рис. 5). Таким образом, можно сделать вывод, что на основе данных об актуальных экологических показателях действительно возможно выявление наличия экологически детерминированных заболеваний, даже при наличии малого количества информации о тенденциях заболеваемости.

Результаты поиска по сетке для мультиклассовой классификации представлены в таблице 1. Качество обучения классификации выборки на отдельные виды социально значимых заболеваний значительно ухудшились по сравнению с моделями для бинарной классификации, так как данная задача более комплексна и решается эффективно моделями машинного обучения при наличии репрезентативного числа объектов в наборе данных.

Самыми точными оказались разновидности модели MLPClassifier, поэтому также был реализован ансамбль из нескольких многослойных перцептронов с различными комбинациями параметров, которые оказались оптимальными по результатам первоначального поиска по сетке. В данном случае был использован «стекинг»: на основе выходных данных пяти различных моделей MLPClassifier была сформирована таблица мета-признаков, по которым была обучена отдельная модель логистической регрессии для формирования конечного предсказания ансамбля.

Как правило, стекинг и иные виды ансамблирования моделей могут повысить точность и эффективность моделирования, однако иногда единичная модель может оказаться точнее, что и произошло в данном случае. Причиной этому мог послужить дисбаланс данных. Таким образом, наиболее точной моделью для классификации социально значимых заболеваний

Таблица 1 (Table 1)

Результаты классификации тестовой выборки по разным моделям
Results of classification of the test sample according to different models

Модель и примененные параметры	Значение метрики Precision на тестовой выборке
MLPClassifier (активация ReLU, скорость обучения 0.1, метод оптимизации SGD)	0.6946
MLPClassifier (активация Tanh, скорость обучения 0.01, метод оптимизации Adam)	0.6913
MLPClassifier (активация Tanh, скорость обучения 0.1, метод оптимизации Adam)	0.6906
CatBoostClassifier (скорость обучения 0.1, глубина 4, число итераций 400, регуляризация L2 2)	0.6835
CatBoostClassifier (скорость обучения 0.1, глубина 4, число итераций 400, регуляризация L2 0.1)	0.6814
KNeighborsClassifier (число соседей 5, веса distance, алгоритм kd_tree)	0.6764
CatBoostClassifier (скорость обучения 0.01, глубина 4, число итераций 400, регуляризация L2 0.1)	0.6722
KNeighborsClassifier (число соседей 5, веса distance, алгоритм auto)	0.6647
KNeighborsClassifier (число соседей 5, веса distance, алгоритм ball_tree)	0.6589

на основе экологических показателей места проживания оказалась простая реализация многослойного перцептрона MLPClassifier (с параметрами: ReLU, 0.1, SGD), что показывает, что есть потенциал для последующего обучения нейронных сетей для решения данной задачи. Однако сперва необходимо развить обучающую и тестовую выборки до более репрезентативного количества записей.

Разработка инструментов для автоматизации анализа и мониторинга экологической обстановки

В работе был реализован «экорейтинг» как инструмент для автоматизированного проведения актуальной оценки экологической обстановки муниципальных единиц Москвы не только специалистами, но и обычными пользователями. Более того, экорейтинг подразумевает под собой определение единой методологии оценки. Иными словами, разнообразные параметры такие, как показатели загрязняющих веществ, плотность населения и другие, были преобразованы в единую и прозрачную числовую оценку.

Для построения экорейтинга был использован способ оценивания по критериям на основе весовых коэффициентов: оценка формировалась из суммы произведения выявленных критериев на их весовые коэффициенты значимости [16]:

$$R_i = \sum_{n=1}^{15} k_{in} w_n,$$

где R_i – рейтинг i-го района, k_{in} – значение для i-го района по n-му критерию, w_n – вес n-го критерия. Веса критериев были определены с помощью метода анализа иерархий, чтобы обеспечить объективность назначаемых критериям значимостям. Матрица попарных сравнений критериев, где по каждой паре критериев было произведено их попарное сравнение значимости на достижение поставленной цели по балльной шкале, была нормирована, а затем было посчитано среднее по каждой строке, что и являлось весами критериев [17, 18]:

$$w_n = \frac{1}{15} \sum_{j=1}^{15} \frac{a_{ij}}{\sum_{i=1}^{15} a_{ij}},$$

где a_{ij} – элемент матрицы, 15 – количество критериев. Так как критерии, перемноженные на веса, суммировались, то, чем выше был оценочный балл – тем более неудовлетворительная экологическая обстановка была в районе. При этом положительные критерии (водоемы, процент озелененных территорий и соседство с лесными массивами) вычитались из рейтинга. Полученные оценки состояния окружающей среды по всем районам Москвы можно разделить на несколько категорий: от 0 до 100 – отлично (около 21% районов); от 100 до 200 – хорошо (около 21% районов); от 200 до 300 – удовлетворительно (около 30% районов); от 300 до 400 – неудовлетворительно (около 18% районов); от 400 и более – плохо (около 10% районов).

Результаты по оценке экологической обстановки районов говорят о преобладании удовлетворительного или более высокого уровня экологического благосостояния территорий Москвы, однако, разумеется, присутствуют районы, требующие повышенного внимания в отношении окружающей среды (например, Зябликово и Новокосино с рейтингами 570.2979 и 587.3477).

Поскольку экорейтинг включал в себя динамичные данные с внешних источников (различные показатели загрязнения по районам с Портала открытых данных Правительства Москвы), то также была реализована автоматическая актуализация экорейтинга при размещении на Портале новых измерений показателей загрязнения с помощью модулей библиотеки Python и официального API-сервиса портала [19].

Для обеспечения свободного доступа к результатам исследования и разработанным инструментам был создан веб-интерфейс. Веб-интерфейс был реализован с помощью сервиса Streamlit, который предоставляет как библиотеку для разработки интерфейса, так и сервер для его размещения в сети Интернет. В первую очередь была создана страница с общими результатами сбора и анализа данных для пояснения целей исследования и предпосылок для построения модели машинного обучения и экорейтинга с помощью таких стандартных объектов библиотеки, как, например, streamlit.expander и streamlit.image. Следующая часть веб-интерфейса позволяет обратиться к построенной модели машинного обучения через элемент streamlit.form с различными видами ввода (текстовое поле, «слайдер» и «select box»): пользователь может ввести свои данные (экологические параметры подтягиваются автоматически по введенному наименованию района из файлов на сервере) и получить примерное распределение предрасположенности человека к социально значимым заболеваниям на основе места его проживания. На сервере происходит обращение к загруженной модели, а также кодирование введенных категориальных признаков и преобразование результатов моделирования в табличный вид для последующего отображения с помощью streamlit.write (рис. 6).

На веб-интерфейсе также размещен экорейтинг и отдельная программа для ежедневной проверки новых измерений по экологическим показателям на Портале открытых данных Правительства Москвы и при их наличии автоматического

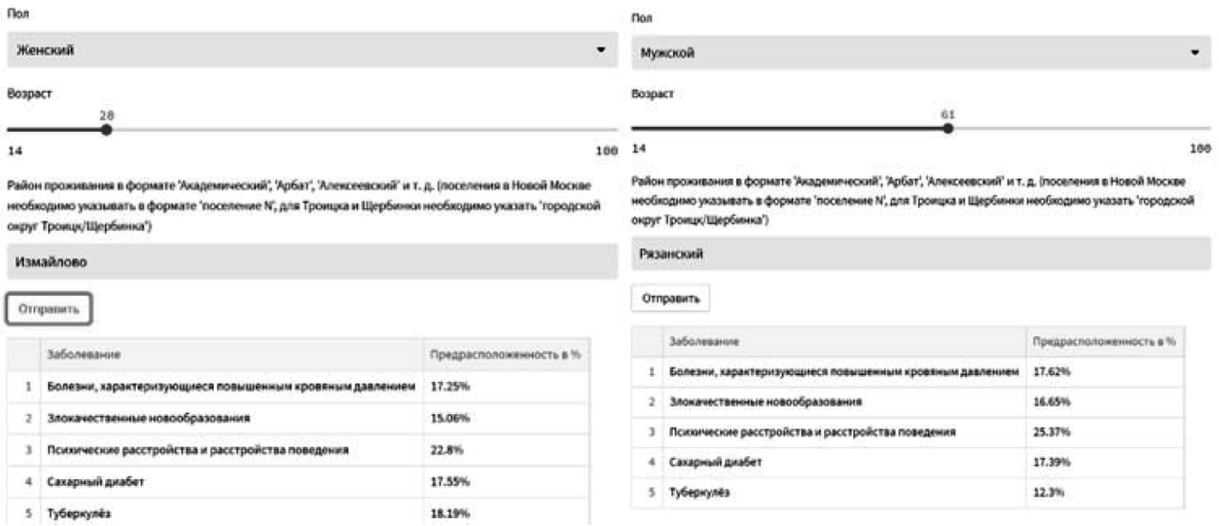

Рис. 6. Примеры работы модели на веб-интерфейсе
Fig. 6. Examples of the model working on the web interface

обновления баллов по муниципальным единицам. Ежедневный запуск реализован с помощью объекта Timer в модуле threading, новые значения записываются в хранящуюся на сервере таблицу экорейтинга. Было решено представить экорейтинг как в виде таблицы, так и с помощью размеченной карты, реализованой с применением python-библиотеки Plotly и возможности ее интеграции с модулем Streamlit [20]. В модуль «scatter_mapbox» был помещен объект DataFrame с координатами районов и представление ранжирования баллов в цветовой схеме для более наглядной визуализации результатов экорейтинга (рис. 7).

Полный код программной реализации веб-интерфейса представлен в репозитории GitHub: https://github.com/AnnBengardt/Ecological-Determinacy-of-Diseases. Сайт размещен на сервере и находится в открытом доступе по ссылке: https://anna-marunko-vkr-ecological-determinacy-of-diseases.streamlit.app/.

Заключение

В рамках проведения исследования после сбора и обработки данных был проведен корреляционно-регрессионный анализ, который показал, что между частью отобранных экологических показателей и наличием социально значимых заболеваний у населения действительно существует статистически значимая корреляция. Данный результат позволяет предположить о существовании взаимосвязи. Поэтому одним из главных выводов данного исследования является то, что при реализации проектов, потенциально оказывающих вредоносное влияние на окружающую среду, критически важно оценивать риски ухудшения здоровья местного населения.

Поскольку статистически значимая корреляция между показателями была подтверждена, то при проведении

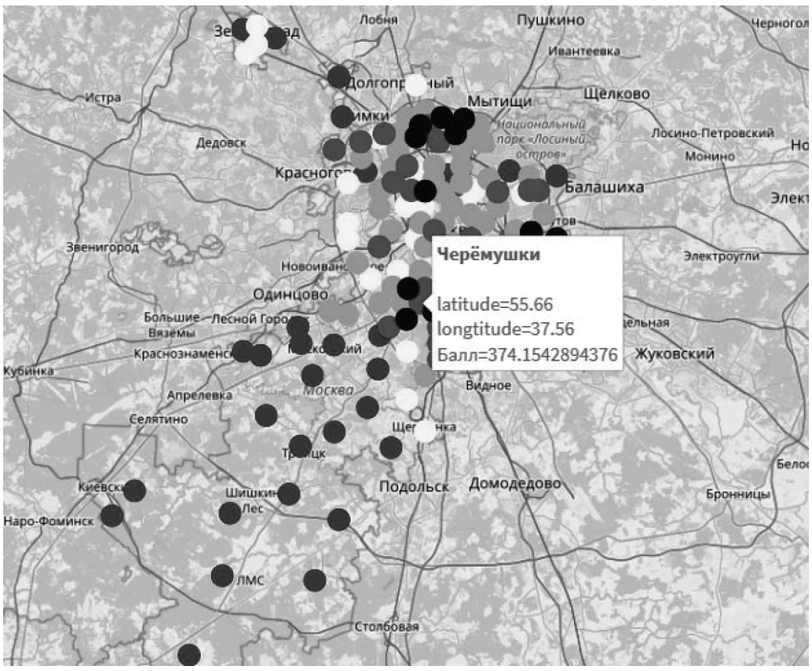

Рис. 7. Визуализация экорейтинга в виде размеченной карты на веб-интерфейсе

Fig. 7. Visualization of eco-rating in the form of a marked map on the web interface

регрессионного анализа были построены модели машинного обучения для предсказания предрасположенности населения к заболеваниям на основе выявленной взаимосвязи. Качество таких предсказаний недостаточно высокое (около 65–70%) для использования в медицинской диагностике, что подтверждает необходимость более глубоких исследований в данной сфере. Выявленную взаимосвязь можно изучать далее на более «совершенных» наборах данных в рамках проектов устойчивого развития и медико-биологических исследований, что приведет к более качественным решениям в этой области.

Литература

1. Гальперин М. В. Общая экология. М.: ИНФРА-М, 2022. 336 с.

2. Гичев Ю. П. Экологическая детерминированность основных заболеваний и сокращения продолжительности жизни. Новосибирск: София, 2021. 130 с.

3. Ефанов А.М., Ляхова О.Л., Мезенцева О.А. Влияние шумового воздействия на здоровье человека // Наука-2020. 2019. № 11. С. 158–162.

4. Лукашевич О.А., Хамдиев И.Ю., Васильев М.В Негативное экологическое влияние аэропортов на окружающую местность // Новые импульсы развития: вопросы научных исследований. 2020. № 7. С. 16–20.

5. Brusseau M.L., Pepper I.L., Gerba C.P., Brusseau M.L. Environmental and Pollution Science. Burlington: Elsevier Inc, 2019. 532 c.

6. Humphries G.R.W., Magness D.R., Huettmann F. Machine Learning for Ecology and Sustainable Natural Resource Management. Cham: Springer Nature Switzerland, 2018. 441 c.

7. Тюрина Т.А. Эволюция взглядов на мир в контексте проблем экологии // Гуманитарные и социальные науки. 2016. № 4. С. 36–40.

8. Семенова Н.П., Ушкарева О.А. Влияние загрязнения атмосферного воздуха на заболеваемость населения Республики Саха (Якутия) // Здоровье населения и среда обитания. 2013. № 10. С. 34–37.

9. Едаменко А.С. Проблемы урбанизированных российских территорий // Концепт. 2018. № 4. С. 1–4.

10. Мун С.А., Зинчук С.Ф. Оценка экологической опасности территорий и онкологической заболеваемости населения Кемеровской области в зависимости от загрязнения атмосферного воздуха // Современные проблемы науки и образования. 2015. № 6. С. 1–11.

11. Мамырбаев А.А. Медико-экологическая оценка здоровья населения в регионах добычи углеводородного сырья. Актобе: НАО ЗКГМУ им. М. Оспанова, 2019. 126 с.

12. Гасангаджиева А.Г., Габибова П.И., Даудова М.Г., Галкина И.В., Гираев К.М., Магомедова З.Я. Медико-экологическая оценка и прогноз социально значимой патологии населения Республики Дагестан // Юг России: экология, развитие. 2019. № 4. С. 147–164.

13. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique // Journal Of Artificial Intelligence Research. 2002. № 16. С. 321–357.

14. Grandini M., Bagli E., Visani G. Metrics for Multi-Class Classification: an Overview [Электрон. ресурс]. Режим доступа: https://arxiv.org/abs/2008.05756. (Дата обращения: 12.04.2023).

15. Bataresh F.A., Yang R. Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering. London: Academic Press, 2020. 266 с.

16. Лохов А.С., Коробов В.Б. Сравнительный анализ применения весовых коэффициентов и коэффициентов значимости в классификационных геоэкологических моделях // Проблемы региональной экологии. 2022. № 4. С. 81–86.

17. Волокобинский М.Ю., Пекарская О.А., Рази Д.А. Принятие решений на основе метода анализа иерархий // Вестник Финансового университета. 2016. № 2. С. 33–42.

18. Dos Santos P.H., Neves S.M., Sant'Anna D.O., Oliveira C.H., Carvalho H.D. The analytic hierarchy process supporting decision making for sustainable development: An overview of applications // Journal of Cleaner Production. 2019. № 7. С. 119–138.

19. Документация API Портала открытых данных города Москвы [Электрон. ресурс]. Режим доступа: https://apidata.mos.ru/Docs. (Дата обращения: 20.04.2023).

20. Документация Streamlit [Электрон. ресурс]. Режим доступа: https://docs.streamlit.io/. (Дата обращения: 24.04.2023).

References

1. Gal'perin M.V. Obshchaya ekologiya = General ecology. Moscow: INFRA-M; 2022. 336 p. (In Russ.)

2. Gichev Yu.P. Ekologicheskaya determinirovannost' osnovnykh zabolevaniy i sokrashcheniya prodolzhitel'nosti zhizni = Environmental determinism of major diseases and reduction of life expectancy. Novosibirsk: Sofia; 2021. 130 p. (In Russ.)

3. Yefanov A. M., Lyakhova O. L., Mezentseva O. A. The influence of noise exposure on human health. Nauka-2020 = Science-2020. 2019; 11: 158-162. (In Russ.)

4. Lukashevich O.A., Khamdiyev I.Yu., Vasil'yev M.V Negative environmental impact of airports on the surrounding area. Novyye impul'sy razvitiya: voprosy nauchnykh issledovaniy = New development impulses: issues of scientific research. 2020; 7: 16-20. (In Russ.)

5. Brusseau M.L., Pepper I.L., Gerba C.P., Brusseau M.L. Environmental and Pollution Science. Burlington: Elsevier Inc; 2019. 532 p.

6. Humphries G.R.W., Magness D.R., Huettmann F. Machine Learning for Ecology and Sustainable Natural Resource Management. Cham: Springer Nature Switzerland; 2018. 441 p.

7. Tyurina T.A. Evolution of worldviews in the context of environmental problems. Gumanitarnyye i sotsial'nyye nauki = Humanitarian and social sciences. 2016; 4: 36–40. (In Russ.)

8. Semenova N.P., Ushkareva O.A. The influence of atmospheric air pollution on the morbidity of the population of the Republic of Sakha (Yakutia). Zdorov'ye naseleniya i sreda obitaniya = Population health and habitat. 2013; 10: 34–37. (In Russ.)

9. Yedamenko A.S. Problems of urbanized Russian territories. Kontsept = Concept. 2018; 4: 1–4. (In Russ.)

10. Mun S.A., Zinchuk S.F. Assessment of the environmental danger of territories and cancer incidence of the population of the Kemerovo region depending on atmospheric air pollution. Sovremennyye problemy nauki i obrazovaniya = Modern problems of science and education. 2015; 6: 1–11. (In Russ.)

11. Mamyrbayev A.A. Mediko-ekologicheskaya otsenka zdorov'ya naseleniya v regionakh dobychi uglevodorodnogo syr'ya = Medical and environmental assessment of population health in hydrocarbon production regions. Aktobe: NAO West Kazakhstan State Medical University named after. M. Ospanova; 2019. 126 p. (In Russ.)

12. Gasangadzhiyeva A.G., Gabibova P.I., Daudova M.G., Galkina I.V., Girayev K.M., Magomedova Z.YA. Medical-ecological assessment and forecast of socially significant pathology of the population of the Republic of Dagestan. Yug Rossii: ekologiya, razvitiye = South of Russia: ecology, development. 2019; 4: 147–164. (In Russ.)

13. Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research = Journal of Artificial Intelligence Research. 2002; 16: 321-357.

14. Grandini M., Bagli E., Visani G. Metrics for Multi-Class Classification: an Overview = Metrics for Multi-Class Classification: an Overview [Internet]. Available from: https://arxiv.org/abs/2008.05756. (cited 12.04.2023).

15. Bataresh F.A., Yang R. Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering. London: Academic Press; 2020. 266 p.

16. Lokhov A.S., Korobov V.B. Comparative analysis of the use of weighting coefficients and significance coefficients in classification geoecological models. Problemy regional'noy ekologii = Problems of regional ecology. 2022; 4: 81–86. (In Russ.)

17. Volokobinskiy M.YU., Pekarskaya O.A., Razi D.A. Decision making based on the hierarchy analysis method. Vestnik Finansovogo universiteta = Bulletin of the Financial University. 2016; 2: 33–42. (In Russ.)

18. Dos Santos PH., Neves S.M., Sant'Anna D.O., Oliveira C. H., Carvalho H. D. The analytic hierarchy process supporting decision making for sustainable development: An overview of applications. Journal of Cleaner Production. 2019; 7: 119-138.

19. Dokumentatsiya API Portala otkrytykh dannykh goroda Moskvy = API documentation of the Open Data Portal of the city of Moscow [Internet]. Available from: https://apidata.mos.ru/Docs. (cited 20.04.2023). (In Russ.)

20. Dokumentatsiya Streamlit = Streamlit documentation [Internet]. Available from: https://docs.streamlit.io/. (cited 24.04.2023).

Сведения об авторах

Татьяна Валерьяновна Золотова
Финансовый университет при Правительстве РФ, Москва, Россия
Эл. почта: tzolotova@fa.ru

Анна Сергеевна Марунько
Финансовый университет при Правительстве РФ, Москва, Россия
Эл. почта: marunko.a@yandex.ru

Information about the authors

Tatiana V. Zolotova
Financial University under the Government of the Russian Federation, Moscow, Russia
E-mail: tzolotova@fa.ru

Anna S. Marunko
Financial University under the Government of the Russian Federation, Moscow, Russia
E-mail: marunko.a@yandex.ru