DOI: 10.25728/mlsd.2023.0795

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ПРИ ИССЛЕДОВАНИИ ЭКОЛОГИЧЕСКОЙ ДЕТЕРМИНИРОВАННОСТИ СОЦИАЛЬНО ЗНАЧИМЫХ ЗАБОЛЕВАНИЙ В Г. МОСКВЕ

Марунько А.С., Золотова Т.В.
Финансовый университет при Правительстве РФ, Москва, Россия
marunko.a@yandex.ru, tzolotova@fa.ru

Аннотация. Поведено исследование взаимосвязи между частью отобранных экологических показателей и наличием социально значимых заболеваний у населения г. Москвы. Статистически значимая корреляция подтвердилась. При проведении регрессионного анализа были построены модели машинного обучения. Разработан веб-интерфейс для автоматизации процессов предсказания заболеваемости населения.

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

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 В и С, злокачественные новообразования, сахарный диабет, психические расстройства и расстройства поведения, ВИЧ, инфекции, передающиеся преимущественно половым путём, и болезни, характеризирующиеся 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.

1. Актуальность и существующие способы решения проблемы влияния загрязнения окружающей среды на здоровье населения и мониторинга экологической обстановки

Целью данного исследования является подтвердить или опровергнуть экологическую детерминированность возникновения социально значимых заболеваний у населения Москвы на основе анализа данных по экологическим и здравоохранительным показателям в разрезе муниципальных единиц города. Будет рассмотрен следующий круг проблем: показатели загрязнения окружающей среды по Москве на основе открытых данных и свидетельств местных жителей, заболеваемость местных жителей социально значимыми болезнями и наличие связи между заболеваемостью и загрязнением 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 на взаимосвязь с загрязнением окружающей среды той или иной местности, если это не промышленный регион, как, например, Красноярский край. Также проводимые на данную тему исследования как в России, так и за рубежом концентрируются только на исследовании влияния атмосферного загрязнения на заболеваемость населения респираторными заболеваниями [1, 2]. Таким образом, можно сделать вывод, что 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 зелёных 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 её 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 [3], 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, проведённому 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. Таким образом, 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 [4].

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 [4, 5, 6].

Таким образом, на данный момент собирающиеся регулярно открытые данные по экологическим показателям не учитываются в проведённых 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.

2. Анализ взаимосвязи экологической обстановки в муниципальных единицах г. Москвы и заболеваемости местного населения болезнями из перечня социально значимых

Перед проведением анализа необходимо было собрать данные по экологическим показателям и по выявленным у населения социально значимым заболеваниям.

Так как в источниках открытых данных практически невозможно найти подробную информацию по заболеваемости населения Москвы, тем более в разрезе по районам города, то было принято решение собрать данные путем проведения социологического опроса на онлайн-платформе. 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 распространённые заболевания: болезни, характеризующиеся повышенным кровяным давлением, психические расстройства и сахарный диабет (рис. 1).

Информация по экологическим показателям, которые можно было бы применить для подтверждения экологической детерминированности социально значимых заболеваний,

предположительно должна была собираться путем поиска и изучения источников открытых данных, в том числе Портала открытых данных Правительства Москвы. Однако во время проведения исследования было обнаружено, что в открытом доступе либо отсутствуют данные по всем выбранным экологическим показателям, необходимым для практической реализации, либо имеющиеся данные являются неполными или неактуальными. Как следствие, сбор данных состоял не только из поиска подходящих наборов данных, но и из ручного создания наборов данных.

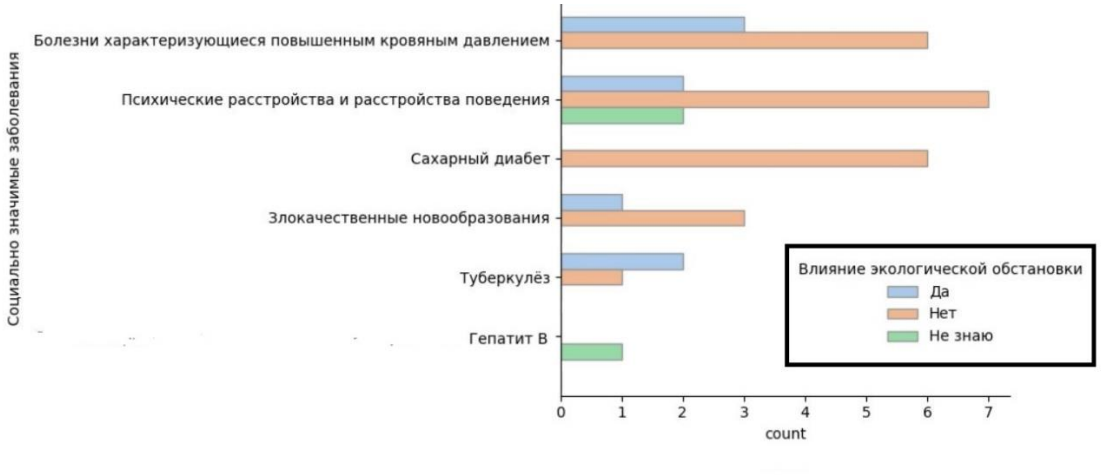

Рис. 1. Диагностированные социально значимые заболевания у населения по результатам опроса

По каждому показателю была отдельная методология для сбора данных. Например, по Превышению ПДК загрязняющих веществ в районах Москвы отбирались вещества, которые способны вызывать заболевания или интоксикацию организма. А районы под действием влияния аэропортов определялись согласно ФЗ от 01.07.2017 N 135, который устанавливает 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 данным критериям включал в себя изучение официальных перечней различных зон/объектов: перечень промышленных зон, список ТЭЦ Москвы, список очистных сооружений и т.д.

Рис. 2. Наличие аэропорта в радиусе 30 километров от районов

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 проведён корреляционно-регрессионный анализ. Два полученных набора данных (экологические показатели и показатели заболеваемости, собранный в рамках социологического опроса) были объединены в единый набор: по районам проживания опрошенных определялись соответствующие, наиболее актуальные экологические показатели. Далее, было обнаружено по изучении полученного набора данных, что он не является сбалансированным, так как класс, у которого отсутствуют выявленные социально значимые заболевания (нулевой класс), количественно превышал остальные классы более чем в два раза (рис. 3). Поскольку это могло оказать негативное влияние на результаты

анализа, но при этом было необходимо сохранить первоначальное распределение данных, то были добавлены новые искусственные записи в примерно таком же распределении (oversampling) с помощью алгоритма Synthetic Minority Oversampling Technique (далее – SMOTE) [7].

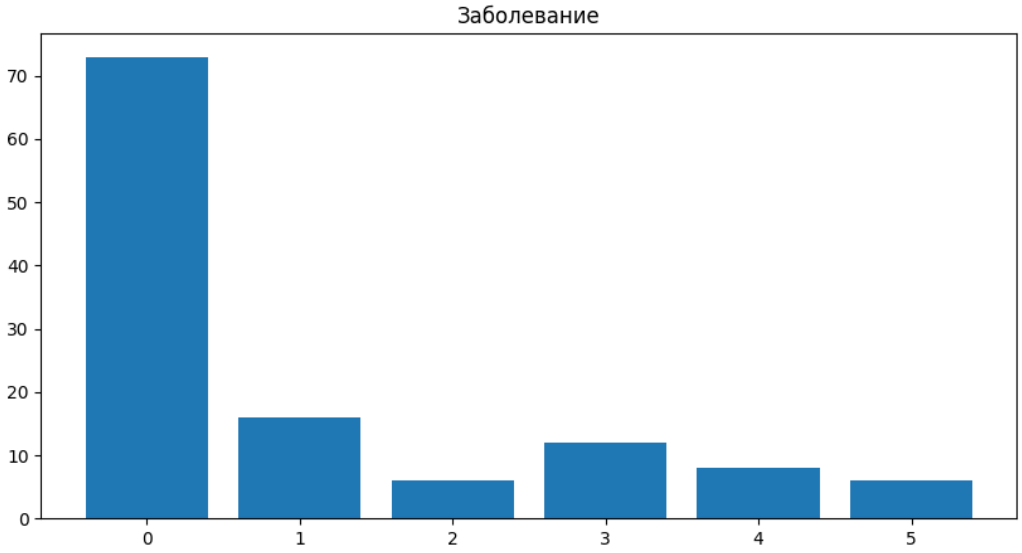

Рис. 3. Изначальное распределение классов социально значимых заболеваний

После балансирования классов в выборке был посчитан коэффициент корреляции Пирсона $r_{xy} = \frac{\sum(x_i - M_x)(y_i - M_y)}{\sqrt{\sum(x_i - M_x)^2 (y_i - My)^2}}$ 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.

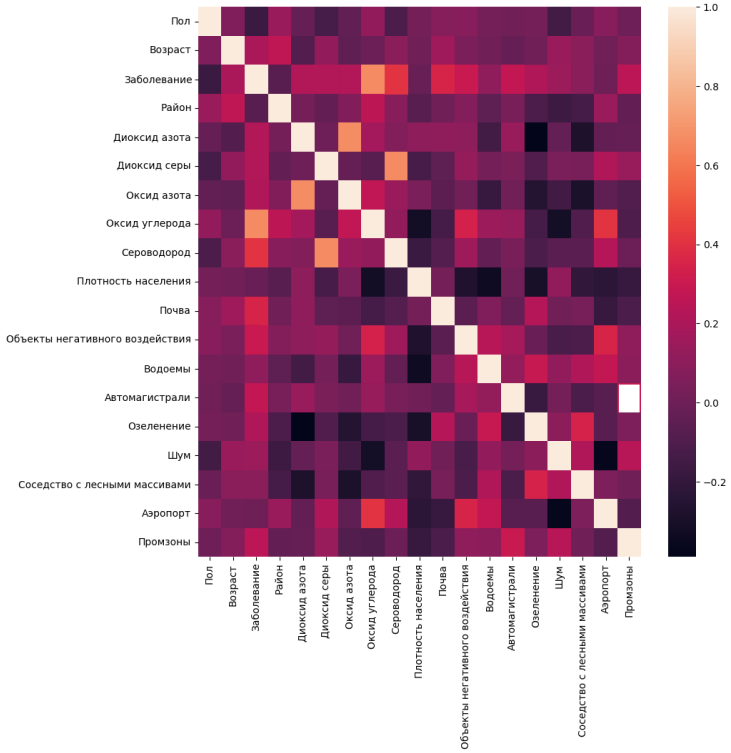

Рис. 4. Тепловая карта коэффициентов корреляции Пирсона

Проведён расчёт *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 расчёта *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 проведён регрессионный анализ с применением 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 приведён 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 [8]: $\frac{TPR}{FPR}$, $TPR = \frac{TP}{TP+FN}$, $FPR = \frac{FP}{FP+TN}$.

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 и оптимальный способ решения такой задачи, были выбраны три модели с разными подходами к классификации [8]:

- 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 [9]. Несмотря на уже применённый 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 сциально значимым болезням.

3. Результаты проведённого анализа и разработки инструментов для автоматизации анализа и мониторинга экологической обстановки

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 помощью CatBoostClassifier (библиотека CatBoost). По итогам поиска по сетке самая результативная модель (с параметрами скорость обучения = 0.1, глубина деревьев = 4, параметр регуляризации листьев деревьев (l2) = 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.

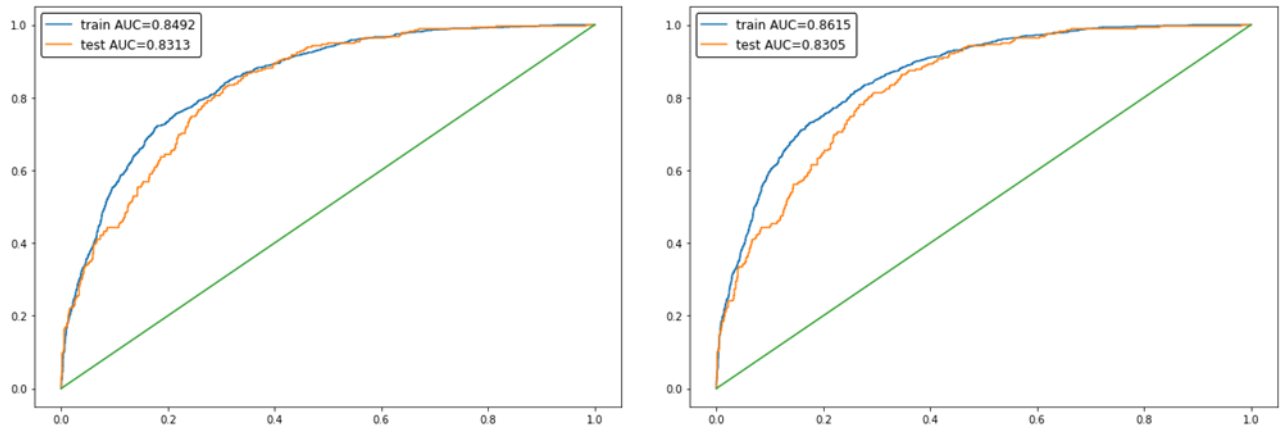

Рис. 5. Показатели метрики ROC-AUC по результатам обучения логистической регрессии (1) и CatBoostClassifier (2)

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.

Таблица 1. Результаты классификации тестовой выборки по трём выбранным моделям

Модель и применённые 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

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 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.

Также был реализован «экорейтинг» как инструмент для автоматизированного проведения актуальной оценки экологической обстановки муниципальных единиц г. 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 [10]: $R_i = \sum_{n=1}^{15} k_{in} w_n$, где R_i – рейтинг *i*-го района, k_{in} – значение для *i*-го района по *n*-ному 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 [11]: $w_n = \frac{1}{15} \sum_{j=1}^{15} \frac{a_{ij}}{\sum_{i=1}^{15} a_{ij}}$, где a_{ij} – элемент матрицы, 15 – количество критериев, w_n – вес *n*-го критерия. 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 озеленённых 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. Также для обеспечения свободного доступа к результатам исследования и разработанным инструментам был создан веб-интерфейс.

4. Заключение

В рамках проведения исследования после сбора и обработки данных был проведён корреляционно-регрессионный анализ, который показал, что между частью отобранных экологических показателей и наличием социально значимых заболеваний у населения действительно существует статистически значимая корреляция. Такой результат позволяет предположить о существовании взаимосвязи, что является одним из главных выводов данного исследования: 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 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, например, 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. *Тюрина Т.А.* Эволюция взглядов на мир в контексте проблем экологии // Гуманитарные и социальные науки. 2016. N 4. – С. 36–40.
2. *Семенова Н.П., Ушкарева О.А.* Влияние загрязнения атмосферного воздуха на заболеваемость населения Республики Саха (Якутия) // Здоровье населения и среда обитания. 2013. N 10. – С. 34–37.
3. *Едаменко А.С.* Проблемы урбанизированных российских территорий // Концепт. 2018. N 4. – С. 1–4.
4. *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. N 6. – С. 1–11.
5. *Мамырбаев А.А.* Медико-экологическая оценка здоровья населения в регионах добычи углеводородного сырья. – Актобе, НАО ЗКГМУ им. М. Оспанова. – 2019. – 126 с.
6. *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. N 4. – С. 147–164.
7. *Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P.* SMOTE: Synthetic Minority Over-sampling Technique // Journal Of Artificial Intelligence Research. – 2002. – N 16. – P. 321-357.
8. *Humphries G.R.W.* Machine Learning for Ecology and Sustainable Natural Resource Management / G.R.W. Humphries, D.R. Magness, F. Huettmann. – Cham : Springer Nature Switzerland, 2018. – 441 p.
9. *Bataresh F.A.* Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering / F.A. Bataresh, R. Yang. – 1st ed. – London : Academic Press, 2020. – 266 p.
10. *Лохов А.С., Коробов В.Б.* Сравнительный анализ применения весовых коэффициентов и коэффициентов значимости в классификационных геоэкологических моделях // Проблемы региональной экологии. 2022. N 4. – С. 81–86.
11. *Волокобинский М.Ю., Пекарская О.А., Рази Д.А.* Принятие решений на основе метода анализа иерархий // Вестник Финансового университета. 2016. N 2. – С. 33–42.