ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ / ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI: https://doi.org/10.60797/COMP.2025.5.3

РАЗРАБОТКА АНАЛИТИЧЕСКОЙ СИСТЕМЫ ОЦЕНКИ ВОЗНИКНОВЕНИЯ РИСКОВ ЗДОРОВЬЮ НАСЕЛЕНИЯ НА БАЗЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Научная статья

Липатова А.В.[1,] *, Потапченко Т.Д.[2]
[1] ORCID : 0009-0006-9729-0068;
[1] МТС Диджитал, Москва, Российская Федерация
[2] Московский технический университет связи и информатики, Москва, Российская Федерация

* Корреспондирующий автор (myagkova_anastasia[at]mail.ru)

Аннотация

В работе рассмотрена специфика создания и описание возможностей использования аналитической системы оценки возникновения рисков здоровью населения на базе алгоритмов машинного обучения. Проведен анализ литературных источников и исследований авторов по тематике оценки антропогенных экологических факторов и проблем их негативного влияния на здоровье людей. Выполнена формализация использованного набора данных для анализа уровня загрязненного воздуха, описана его структура, обозначены входные признаки, приведены результаты разведывательного, корреляционного анализов, построены модели машинного и глубокого обучение, проведено исследование их работы и оценка значений метрик, характеризующих точность их работы. Выполнен анализ результатов, выявлены наиболее эффективные модели и обозначены пути дальнейшего совершенствования работы.

Ключевые слова: машинное обучение, интеллектуальный анализ данных.

DEVELOPMENT OF AN ANALYTICAL SYSTEM FOR ASSESSING THE EMERGENCE OF PUBLIC HEALTH RISKS ON THE BASIS OF MACHINE LEARNING ALGORITHMS

Research article

Lipatova A.V.[1,] *, Potapchenko T.D.[2]
[1] ORCID : 0009-0006-9729-0068;
[1] MTS Digital, Moscow, Russian Federation
[2] Moscow Technical University of Communications and Informatics, Moscow, Russian Federation

* Corresponding author (myagkova_anastasia[at]mail.ru)

Abstract

The work examines the specifics of creating and describing the possibilities of using an analytical system for assessing the emergence of risks to public health on the basis of machine learning algorithms. Literature sources and authors' studies on the evaluation of anthropogenic environmental factors and the problems of their negative impact on human health have been analysed. The formalization of the used data set for the analysis of the level of polluted air has been performed, its structure has been described, input attributes have been indicated, the results of exploratory and correlation analyses have been given, machine and deep learning models have been built, their operation has been studied and the values of metrics characterising the accuracy of their operation have been estimated. The results are analysed, the most effective models are identified and ways of further improvement are outlined.

Keywords: machine learning, data mining.

Введение

Проблема автоматизации процессов анализа разнородных и больших объемов экологических данных приобретает все большую актуальность и востребованность, что во многом связано с необходимостью внедрения эффективных мер превентивного препятствия развитию человеческих заболеваний, вызванных различными антропогенными факторами [1], [2]. В контексте существующих риск-ориентированных подходов по оценке последствий возникновения техногенных факторов и их влияния на здоровье населения дополнительную целесообразность приобретают современные подходы к анализу данных, основанные на применении технологий искусственного интеллекта (ИИ), алгоритмов или моделей машинного (МО) и глубокого (ГО) обучения, в том числе искусственных нейронных сетей (ИНС) [3]. Подобные концепции имеют преимущества над существующими статистическими и математическими подходами благодаря поддержке процедур формирования обобщающей способности моделей, унификации предиктивных алгоритмов и возможностям интерпретации результатов в наглядном виде [4], [5].

Анализ литературных источников и проблематики

В настоящее время проблеме анализа данных в области оценки рисков здоровья населению в контексте экологического загрязнения посвящено значительное число научных трудов, рассмотрим популярные практики и подходы, нацеленные на автоматизацию процессов решения интеллектуальных задач.

В статье [6] авторами рассматриваются ИИ и методы МО для предсказания загрязнения воздуха и последствий для здоровья, включая хронические респираторные заболевания. Исследователи подчеркивают высокую точность гибридных моделей, комбинирующих различные алгоритмы для прогнозирования загрязняющих веществ. Они

оценивают модели по метрикам точности, таким как RMSE и MAE, отмечая их эффективность в раннем оповещении о рисках для здоровья, однако дисбаланс входных выборок данных вносит существенные коррективы в полноту.

Исследование [7] основано на применении модели случайного леса для анализа влияния антропогенных выбросов и метеорологических факторов на долгосрочные изменения уровня загрязнения воздуха в восточном Китае. Результаты проведенного анализа данных показали, что значительное снижение загрязнения связано с уменьшением различных антропогенных выбросов в атмосферу ряда локаций. Модель МО позволила выявить тренды сезонных колебаний в загрязнениях воздуха, что позволило провести более точную оценку рисков здоровья населения при различном уровне концентрации вредных веществ, при этом точность модели оказалась достаточно высокой, более 86%.

Авторы [8] оценивают несколько разных моделей ГО, включая LSTM и Bi-LSTM, для прогнозирования уровней загрязненности воздуха примесями PM2.5 и CO. Исследователи установили, что модель Stacked LSTM показала более высокий уровень точности для PM2.5, а Encoder-Decoder LSTM – для значений CO. Результаты использования моделей могут быть использованы для информирования о краткосрочных (горизонт прогнозирования составил от 1 до 3 дней) значениях рисков для здоровья. Также в рамках данной работы рассмотрено применение алгоритмов классификации (SVM) для выявления уровня корреляции между степенью загрязненности воздуха и заболеваниями, связанными с дыхательной и сердечно-сосудистой системами.

В другом исследовании [9] на базе применения МО авторы создали две MS2Quant модели для прогнозирования эффективности ионизации и модель MS2Tox для оценки токсичности продуктов аквакультур. Созданные авторами модели применимы для определения потенциально опасных химических веществ в воде на основе анализа данных по спектрам масс и позволяют повысить уровень быстродействия идентификации и классификации загрязнителей в сточных водах, которые потенциально оказывают влияние на оценку риска здоровью.

В работе [10] авторами исследованы методы ансамблевого обучения для оценки качества грунтовых вод в районе бассейна Гуанчжун. В частности, применены модели LightGBM в комбинации с анализом неопределенности и SHAP подходом для прогноза параметров качества загрязненной воды. Модели позволяют учесть влияние антропогенных и природных факторов, что помогает выявить ключевые риски для здоровья, однако их точность сильно зависит от значений входных гиперпараметров.

Таким образом следует отметить, что в настоящее время в научной среде много внимания уделяется специфике применения методов МО и ГО для задач автоматизации анализа экологически значимых для здоровья населения данных, в связи с чем данная тематика является актуальным и востребованным направлением.

Цель работы заключается в разработке аналитической системы оценки возникновения рисков здоровью населения на базе алгоритмов машинного обучения.

Разработка концепции системы

Рассматриваемая нами задача сводиться к многоклассовой классификации. Задача многоклассовой классификации в машинном обучении – это задача предсказания, где модель должна определить, к какому из нескольких возможных классов принадлежит наблюдаемый объект. Математическая постановка этой задачи в рамках оценки рисков нанесения вреда здоровью населения может быть выражена следующим образом. Наш входной набор данных может быть представлен как $X=\{x_1, x_2, ...,x_n\}$, где каждый объект x_i является вектором признаков из пространства $\Re^d$. Каждому объекту x_i сопоставляется метка класса $y_i \in \{1,2,...,K\}$, где K – количество классов (6 классов в рамках нашей задачи).

Требуется построить функцию $f:R^d \rightarrow \{1,2,...,K\}$, которая для любого входного объекта x будет предсказывать метку класса y (риск здоровью населения). Модель МО строится с использованием обучающей выборки $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$, которая может быть сформирована из информативных входных признаков и ее задача — найти аппроксимацию функции f на основе этих данных. Если $P(y=k|x)$ – вероятность принадлежности объекта x классу k, то функция $f(x)$ предсказывает класс с максимальной апостериорной вероятностью:

$$f(x) = \arg\max_{k \in \{1,2,...,K\}} P(y = k \mid x) \quad (1)$$

Для обучения модели используется функция потерь, которая измеряет расхождение между предсказанными классами и реальными метками классов.

Одной из часто используемых функций потерь является кросс-энтропия:

$$L(y, y') = -\sum_{k=1}^{K} y_k \log\left(y'_k\right) \quad (2)$$

где y_k – бинарный индикатор (0 или 1), указывающий, относится ли объект к классу k, $y'=P(y=k|x)$ – вероятность того, что объект принадлежит классу k, предсказанная моделью. Модель оптимизируется путем минимизации функции потерь L с использованием методов оптимизации, таких как градиентный спуск. Итоговое предсказание выполняется как выбор класса с максимальной вероятностью на основе обученной модели. Общий пайплайн работы системы для проведения исследований приведен на рис.1.

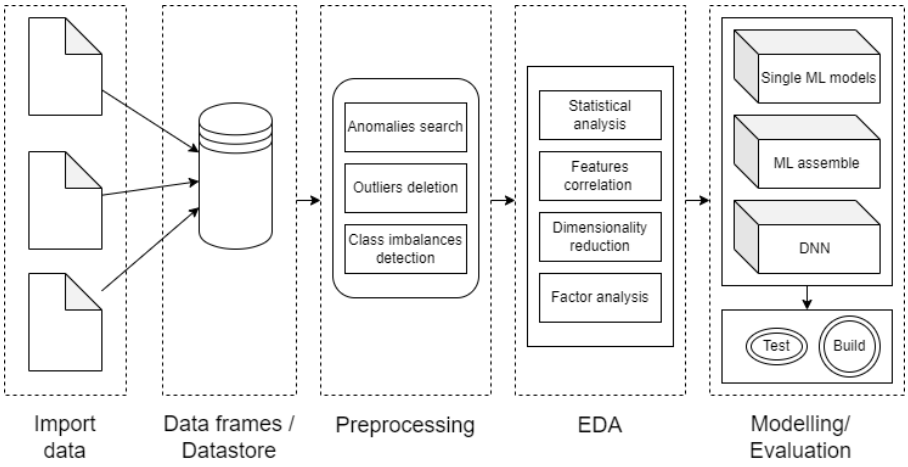

Рисунок 1 - Общий пайплайн работы системы
DOI: https://doi.org/10.60797/COMP.2025.5.3.1

Импортированные наборы данных из датасета посредством функций библиотеки Pandas сохраняются в виде объектов dataframe, после чего осуществляется процедура препроцессинга (предварительной обработки данных посредством поиска аномалий, выбросов и устранения дисбаланса выходного класса), затем осуществляется набор процедур разведывательного анализа данных (статистический, корреляционный и факторный анализ, а также опциональное снижение размерности данных в случае большого числа входных признаков), после чего в рамках этапа моделирования создаются отдельные модели машинного обучения, формируется ансамбль моделей и реализуются глубокие полносвязные модели ИНС на базе разделенных на обучающие и тестовые выборки данных, подвергнутых процедуре кроссвалидации.

Дисбаланс классов реализован на базе подхода, основанного взвешивания классов, т.е. путем расчета значений весов в виде обратной величины частоты класса в выборке (фактически возрастает уровень штрафа модели для менее распространенных классов в датасете). Генерация синтетических данных не предусмотрена, что положительно сказывается на достоверности классификации. Сформированные модели оцениваются по выбранным метрикам оценки качества (точности) их работы, тестируются и их итоговые объекты сериализуются в отдельные файлы для последующей загрузки для использования на новых данных с целью оценки рисков вреда населению.

Описание датасета

В процессе осуществления процедуры поиска доступных наборов данных для проведения исследований влияния различных антропогенных факторов на здоровье населения выявлено, что в свободном доступе отсутствуют комплексные датасеты, отражающие разные аспекты экологической загрязненности. В большей степени на платформах анализа данных и в открытых репозиториях превалируют наборы данных по загрязнению воздушных масс различных регионов мира, в том числе в странах Индии. В качестве базового набора данных возьмем Air Pollution Dataset from India and Nepal (APD) [11]. Представляет собой составной набор данных, который содержит изображения, собранные в Индии и Непале, описывающие и характеризующие уровень рисков вреда людям от уровня загрязнения воздуха различными вредными веществами в различных условиях, а также текстовые наборы данных с детализацией описания данных по значимым признакам. Региональная специфика датасета заключается в учете визуальных изображений разных регионов, что дополняет общую информацию, представленную в табличном виде в формате csv, конкретизируя особенности распределения загрязнений в разных локациях Индии и Непала.

Особенность набора данных заключается в том, что изображения сделаны с разными уровнями загрязненности и могут быть использованы для анализа с применением методов компьютерного зрения и МО. Размер выборки составляет около 12 000 записей, распределение целевых классов приведено в примерно равных пропорциях (от 13 до 21%). Также данные агрегированы на базе сбора информации из 2 разных государств с отличными друг от друга экологическими и социально-экономическими условиями (Индия и Непал), что позволяет проводить сравнительный анализ данных по локациям.

Следует отметить, что в этом наборе данных предусмотрен потенциал анализа не только данных, полученных с измерительных средств состава воздуха, но и анализ визуальных признаков загрязнения (что позволяет сформировать большее признаковое пространство и учесть сложноформализуемые факторы), что может быть полезно для более комплексной оценки. Т.е. данные могут использоваться совместно с метеорологической информацией и измерениями здоровья населения для комплексной оценки рисков. Структурно датасет разделен на два каталога: Combined_Dataset и Country_wise_Dataset. Датасет включает информацию о городе Биратнагар Непала и о городах Индии: Дели, Нагаленд, Бангалор, Большая Нойда, Фаридабад, Мумбаи, Тамил Наду. Входные признаки датасета хранятся в файле формата csv и содержат информацию о расположении локации, имени файла (изображении), дате (год, месяц, день, час), а также показателях загрязненности воздуха (PM2.5, PM10, O3, CO, SO2, NO2) и целевом классе AQI_Class. В качестве целевого признака предусмотрено 6 разных классов загрязнения воздуха, которые представлены в наборе данных:

1. Хорошее (Good), соответствует числовому диапазону (0-50), в этом случае качество воздуха считается удовлетворительным, а загрязнение воздуха представляет небольшой или нулевой риск населению.

2. Умеренное (Moderate), соответствует числовому диапазону (51-100), для данного класса качество воздуха приемлемое, однако для некоторых загрязняющих веществ может быть умеренная проблема со здоровьем для очень небольшого числа людей, которые необычно чувствительны к загрязнению воздуха, т.е. риски населению в целом минимальны.

3. Нездорово для чувствительных групп (Unhealthy_for_Sensitive_Groups), соответствует числовому диапазону (101-150), в этом случае люди, относящиеся к чувствительным группам, могут испытывать последствия для здоровья, но маловероятно, что население в целом будет испытывать высокий риск развития хронических заболеваний, можно интерпретировать класс как низкий уровень риска.

4. Нездоровый (Unhealthy), соответствует числовым значениям в диапазоне (151-200), для данного класса выходного признака более половины представителей общественности может испытывать проблемы со здоровьем, обострением заболеваний, а у представителей уязвимых групп могут возникнуть серьезные проблемы со здоровьем. Средний уровень риска.

5. Очень нездоровый (Very_Unhealthy), соответствует числовому диапазону (201-300), в данном случае риск необратимых пагубных последствий для здоровья населения высок для всех групп.

6. Опасный/тяжелый (Severe), соответствует числовым значениям в диапазоне (301-500), характерно для критических и чрезвычайных ситуаций, в том числе аварий, высокая вероятность необратимого вреда здоровью населения, уровень риска критический.

Разведывательный анализ данных

Первоначально осуществлен импорт библиотек для обработки данных, создания структур (коллекций numpy, pandas) с целью обеспечения необходимых манипуляций с входными признаками, визуализации данных (matplotlib, seaborn), а также ряда пакетов библиотеки sklearn для выполнения процедур преобразования категориальных данных (строковых или текстовых меток) в числовые значения, нормализации данных, подключения метрик оценки моделей и объектов для их непосредственного создания (например, DecisionTreeClassifier). Проведем корреляционный анализ признаков, результат приведен на рис.2.

Attributes	Location	Filename	Year	Month	Day	Hour	AQI	PM2.5	PM10	O3	CO	SO2	NO2
Location	1	0.976	-0.241	0.205	0.213	0.185	0.737	0.751	0.654	-0.008	-0.026	0.422	0.606
Filename	0.976	1	-0.224	0.222	0.295	0.144	0.699	0.740	0.656	-0.033	0.009	0.451	0.610
Year	-0.241	-0.224	1	-0.976	0.317	0.208	-0.105	-0.034	-0.084	0.249	0.065	0.075	0.104
Month	0.205	0.222	-0.976	1	-0.340	-0.263	0.045	-0.007	0.069	-0.260	-0.008	-0.057	-0.111
Day	0.213	0.295	0.317	-0.340	1	0.142	0.046	0.278	0.124	0.092	0.162	0.371	0.267
Hour	0.185	0.144	0.208	-0.263	0.142	1	0.256	0.139	0.096	0.582	-0.390	-0.191	0.147
AQI	0.737	0.699	-0.105	0.045	0.046	0.256	1	0.806	0.664	0.054	-0.216	0.224	0.487
PM2.5	0.751	0.740	-0.034	-0.007	0.278	0.139	0.806	1	0.813	0.035	-0.062	0.281	0.709
PM10	0.654	0.656	-0.084	0.069	0.124	0.096	0.664	0.813	1	0.137	-0.059	0.169	0.571
O3	-0.008	-0.033	0.249	-0.260	0.092	0.582	0.054	0.035	0.137	1	-0.349	-0.314	0.106
CO	-0.026	0.009	0.065	-0.008	0.162	-0.390	-0.216	-0.062	-0.059	-0.349	1	0.398	-0.007
SO2	0.422	0.451	0.075	-0.057	0.371	-0.191	0.224	0.281	0.169	-0.314	0.398	1	0.332
NO2	0.606	0.610	0.104	-0.111	0.267	0.147	0.487	0.709	0.571	0.106	-0.007	0.332	1

Рисунок 2 - Таблица оценки корреляции между входными признаками набора данных
DOI: https://doi.org/10.60797/COMP.2025.5.3.2

Как можно отметить, кроме неинформационного признака имени файла изображения, высокие значения корреляции свойственны для признаков, характеризующих загрязненность воздуха вредными примесями (PM2.5 и PM10), что обусловлено характером их оценки и близостью подходов к измерению оборудования. В рамках проведения предварительного анализа, очистки и предобработки данных в контексте рассматриваемой задачи принято решение удалить из объекта dataframe не информативный признак Filename.

В контексте проведения исследований данных получаем статистическое описание по входным признакам с помощью функции Pandas describe(), которая выводит количество, среднее значение, стандартное отклонение и диапазон данных, результат приведен на рис.3.

```
# Describe data
dataframe.describe()
```

	Year	Month	Day	AQI	PM2.5	PM10	O3	CO	SO2	NO2
count	10281.000000	10281.000000	10281.000000	10281.000000	10281.000000	10281.000000	10031.000000	9807.000000	9034.000000	9920.000000
mean	2022.948254	2.683980	12.116526	167.517848	142.942723	145.403790	39.393336	101.412380	13.305071	37.899775
std	0.221524	1.764717	8.277427	102.798851	130.398412	103.952927	33.371867	116.346153	9.876013	39.627358
min	2022.000000	2.000000	1.000000	15.000000	4.000000	7.000000	1.000000	0.000000	2.000000	0.670000
25%	2023.000000	2.000000	3.000000	97.000000	35.000000	64.000000	12.000000	4.000000	4.400000	7.000000
50%	2023.000000	2.000000	13.000000	152.000000	70.080000	113.000000	31.000000	52.000000	10.000000	20.000000
75%	2023.000000	3.000000	20.000000	230.000000	257.000000	198.000000	59.660000	174.000000	20.000000	64.000000
max	2023.000000	10.000000	28.000000	450.000000	500.000000	480.000000	225.000000	410.000000	57.000000	169.000000

Рисунок 3 - Результат оценки статистических показателей
DOI: https://doi.org/10.60797/COMP.2025.5.3.3

В контексте подготовки данных выполнена процедура преобразования - кодирование меток посредством класса LabelEncoder, посредством которой категориальные данные преобразуются в числовые значения, чтобы сделать их совместимыми с математическими операциями и моделями.

На базе проведенных манипуляций установлено, что признаки месяц и год обладают высокой корреляцией и некоторой противоречивостью, в связи с чем они исключены из итогового набора данных. В процессе реализации процедуры анализа пропусков посредством вызова метода isnull() установлено наличие более 2000 пропусков в признаках O3, CO, SO2, NO2, в связи с чем было выполнено заполнение пропущенных значений путем расчета и подстановки средних значений посредством вызова функции mean().

Разработка и исследование моделей

Для разработки моделей МО использован язык программирования Python, библиотеки sklearn, matplotlib, seaborn, keras, tensorflow [12], [13], на базе чего сформированы отдельные модули Jupiter Notebooks, в каждом из которых реализованы процессы импорта программных зависимостей (библиотек), входных данных (тренировочной и тестовой выборок), созданы (обучены и протестированы) советующие модели, проведена оценка их эффективности на базе описанных выше метрик, а также выполнена сериализация моделей в файлы объектов pickle. В качестве моделей МО реализованы: дерево решений (DT), SVM, случайный лес (RF), XGBoost, глубокие ИНС (сверточная – CNN и рекурентная – LSTM).

Выполнено разделение выборки данных на обучающее и тестовое подмножество в пропорции 75% для обучения и 25% для тестирования моделей МО. С целью выделения значений метрик в отдельные логи принято решение реализовать их сохранения в соответствующие переменные. В связи с необходимостью проведения сводного анализа результатов оценок моделей на предмет точности решения задачи классификации использована визуализация в форме матрицы ошибок средствами seaborn, результаты построения таких матриц для всех созданных моделей МН приведены на рис.4. Для удобства выходные классы рисков преобразованы в числовой диапазон от 0 до 5 по порядку.

Рисунок 4 - Матрицы ошибок моделей дерева принятия решений (*а*), SVM (*б*), случайного леса (*в*), XGBoost (*г*), рекуррентной (*д*) и сверточной (*е*) ИНС
DOI: https://doi.org/10.60797/COMP.2025.5.3.4

Сводные результаты отражают высокую точность моделей, при этом наиболее эффективной моделью с точки зрения точности классификации является XGBoost. С целью дополнительного анализа моделей разработана визуализация для сопоставления точности классификации всех моделей в мультиклассовой форме. Усредненные зависимости по ROC кривым моделей на диаграмме в виде обособленной визуализации приведены на рис.5. Как можно отметить, характер кривых изменчив в разных диапазонах, наиболее близкими к идеальным значениям (более сглаженным и приближенным к 1) являются оценки моделей ансамблей (случайный лес и XGBoost).

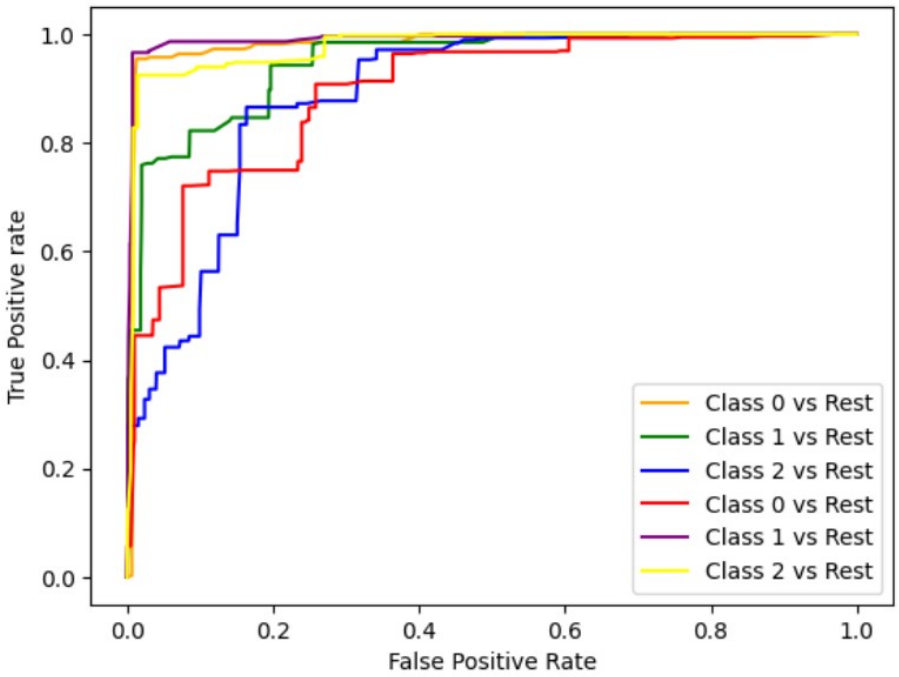

Рисунок 5 - Усредненные зависимости по ROC кривым моделей
DOI: https://doi.org/10.60797/COMP.2025.5.3.5

Для детального исследования характера обучения моделей ИНС сформированы графические зависимости с оценками Accuracy и Loss (рис.6 и рис.7).

Как можно заметить, в целом модели ИНС достигают высоких значений точности, при этом модель LSTM быстрее достигает значений точности около 0,98 (до 15й эпохи) в сравнении с CNN (после 25й эпохи), после чего рост фактически замедляется, периодически наблюдаются незначительные колебания, что свидетельствует о рисках переобучения, однако подобранные значения регуляризации препятствуют данному негативному явлению. Первоначальные значения ошибок у модели также являются более высокими для модели CNN, при этом скорость обучения сверточной модели существенно быстрее чем у LSTM.

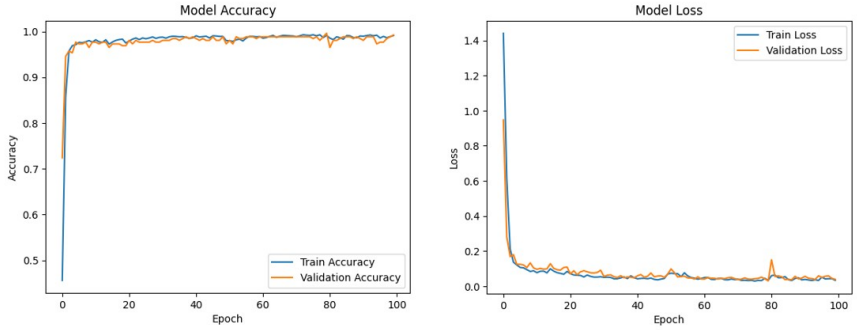

Рисунок 6 - Зависимости значений Accuracy и Loss от эпох обучения рекуррентной ИНС
DOI: https://doi.org/10.60797/COMP.2025.5.3.6

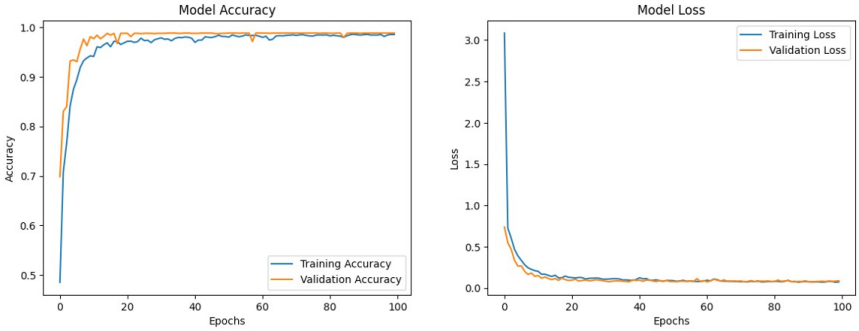

Рисунок 7 - Зависимости значений Accuracy и Loss от эпох обучения сверточной ИНС
DOI: https://doi.org/10.60797/COMP.2025.5.3.7

Результаты сравнительного анализа метрик созданных моделей МО приведены на рис.8.

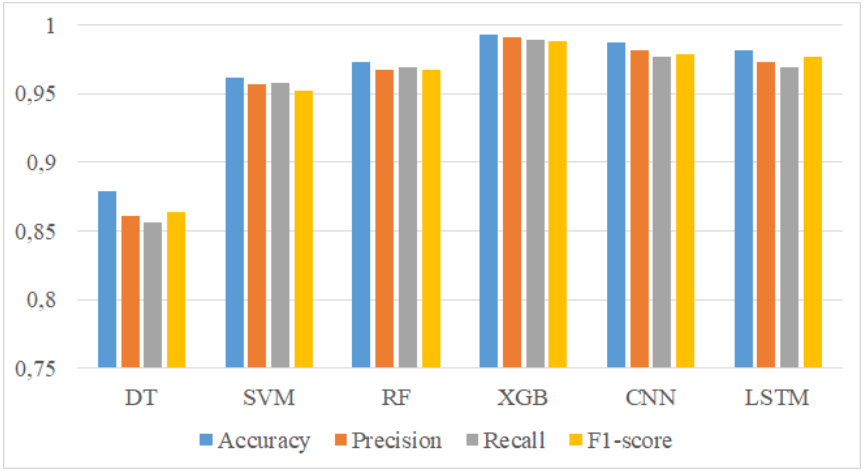

Рисунок 8 - Гистограмма сравнения метрик моделей МО
DOI: https://doi.org/10.60797/COMP.2025.5.3.8

Для дополнительного анализа результатов использования моделей сформирована гистограмма оценки значимости признаков датасета, приведенная на рис.9. Оценка значимости признаков (feature importance) позволяет определить, какие из них наиболее влияют на предсказания модели. Это помогает улучшить качество модели, исключив неинформативные или избыточные признаки, и понять, какие факторы наиболее важны для прогнозирования целевой переменной.

Как можно заметить наибольший уровень значимости характерен для признаков AQI, PM2.5 и PM10, что позволяет сделать вывод о необходимости формирования на них акцента при построении моделей и дальнейшей их оптимизации.

Таким образом, следует отметить, что наиболее быстрой и одновременно наименее точной является модель дерева решений, модель опорных векторов является наиболее точной одиночной моделью, однако ее временные затраты в 5-6 раз выше, чем у дерева решений и в 2 раза выше, чем у моделей на базе ансамблей.

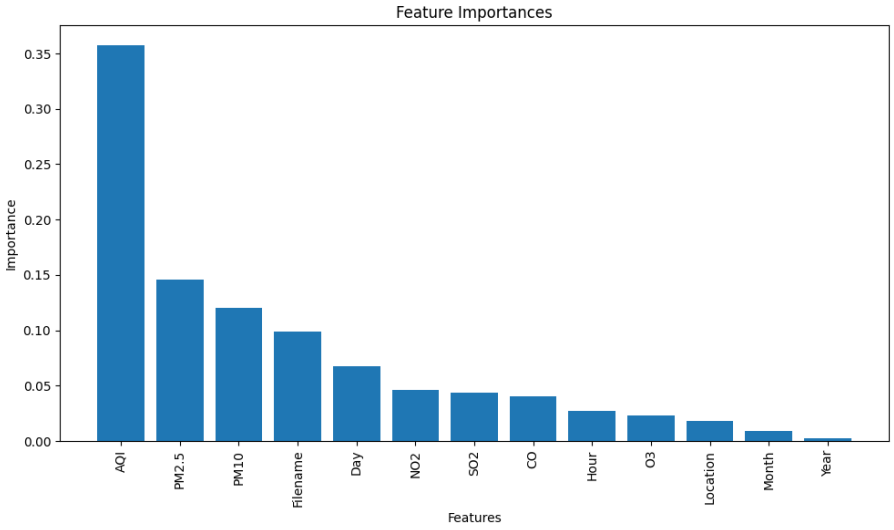

Рисунок 9 - Гистограмма оценки значимости признаков датасета
DOI: https://doi.org/10.60797/COMP.2025.5.3.9

Модели ИНС являются точными, однако требуют значительно больше ресурсов для обучения по причине сложности их структуры и наличия вложенных (скрытых) слоев и большого числа нейронных связей, при этом более ресурсоемкой является LSTM модель. При этом наиболее эффективными с точки зрения соотношения точности и быстродействия являются модели на базе ансамблей, среди которых положительным образом выделяется модель XGBoost.

Заключение

В результате проведенных исследований установлена практическая целесообразность применения разных моделей МО и ГО для решения задачи классификации по оценке рисков вреда здоровью населения от загрязнений воздуха. В целом точность сформированных моделей является достаточно высокой и составляет более 90%, однако скорость их обучения и использования на тестовых данных является разной, с точки зрения наилучшего соотношения по точности и производительности следует отметить модели ансамблей (Random Forest и XGBoost).

В настоящий момент система ограничена рядом аспектов, в частности на данный момент процессы обучения и настройки моделей выполняются только в последовательном режиме, отсутствует поддержка распределённой архитектуры CUDA и входные данные могут вводиться посредством текстового файла (без интерактивного интерфейса пользователя). Следует отметить необходимость подбора моделей и значений их гиперпараметров под конкретные наборы данных, одни из перспективных путей в данном направлении является применение алгоритмов оптимизации, в том числе grid search подхода, что может быть рассмотрено в последующих исследованиях в данной области.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Аскаров К.А. Оценка риска здоровью населения, проживающего в регионе расположения Жезказганского медеплавильного завода ТОО «KAZAKHMYS SMELTING» / К.А. Аскаров, К.К. Алимбетов, Г.Т. Байтуякова [и др.] // Вестник Казахского Национального медицинского университета. — 2022. — №1. — С. 521–529.

2. Имашева Б.С. Оценка риска здоровью населения, проживающего в регионе расположения объектов Павлодарского алюминиевого завода АО «Алюминий Казахстана» / Б.С. Имашева, К.А. Аскаров, М.С. Имашев [и др.] // Наука и Здравоохранение. — 2022. — № 3 (24). — С. 66–77.

3. Камдина Л.В. Развитие методического инструментария оценки влияния антропогенных факторов промышленного производства на качество жизни населения: автореф. дис. ... канд. экон. наук / Камдина Людмила Владимировна. — 2019. — 23 с.

4. Виноградова Е.А. Способы экологического мониторинга воздуха с применением технологических решений для их обработки и анализа, включая машинное обучение, искусственный интеллект и аналитику больших данных / Е.А. Виноградова, М.М. Дмитриев, А.С. Дудрявец [и др.] // Современные технологии: проблемы и тенденции развития. Монография. — Петрозаводск: Новая Наука, 2021. — С. 174–189.

5. Железный С.В. Использование методов машинного обучения для прогнозирования загрязненности атмосферного воздуха / С.В. Железный, А.И. Ситников, А.А. Толстых // Вестник Воронежского института МВД России. — 2019. — № 3. — С. 73–81.

6. Subramaniam S. Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review / S. Subramaniam, N. Raju, A. Ganesan [et al.] // Sustainability. — 2022. — № 14. — 9951.

7. Qian Z. Machine Learning Explains Long-Term Trend and Health Risk of Air Pollution during 2015–2022 in a Coastal City in Eastern China / Z. Qian, Q. Meng, K. Chen [et al.] // Toxics. — 2023. — № 11. — 481.

8. Tello-Leal E. Evaluation of Deep Learning Models for Predicting the Concentration of Air Pollutants in Urban Environments / E. Tello-Leal, U.M. Ramirez-Alcocer, B.A. Macías-Hernández [et al.] // Sustainability. — 2024. — № 16. — 7062.

9. Wu X. A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China / X. Wu, Q. Zhang, F. Wen [et al.] // Water. — 2022. — № 14. — 3408.

10. Sepman H. Machine Learning Tools Can Pinpoint High-Risk Water Pollutants / H. Sepman, P. Peets, L. Jonsson [et al.] // Proceedings. — 2023. — № 92. — 68.

11. Air Pollution Image Dataset from India and Nepal. — URL: https://www.kaggle.com/datasets/adarshrouniyar/air-pollution-image-dataset-from-india-and-nepal (accessed: 11.10.2024).

12. Algore M. Python For Data Analysis: The Ultimate and Definitive Manual to Learn Data Science and Coding With Python. Master The basics of Machine Learning, to Clean Code and Improve Artificial Intelligence / M. Algore. — 2021. — 87 p.

13. Antao T.R. High Performance Python for Data Analytics (MEAP) / T.R. Antao. — New York: Manning Publications, 2020. — 98 p.

Список литературы на английском языке / References in English

1. Askarov K.A. Ocenka riska zdorov'ju naselenija, prozhivajushhego v regione raspolozhenija Zhezkazganskogo medeplavil'nogo zavoda TOO «KAZAKHMYS SMELTING» [Health risk assessment of the population living in the region of Zhezkazgan copper smelter location 'KAZAKHMYS SMELTING' LLP] / K.A. Askarov, K.K. Alimbetov, G.T. Bajtujakova [et al.] // Vestnik Kazahskogo Nacional'nogo medicinskogo universiteta [Bulletin of the Kazakh National Medical University]. — 2022. — №1. — P. 521–529. [in Russian]

2. Imasheva B.S. Ocenka riska zdorov'ju naselenija, prozhivajushhego v regione raspolozhenija ob'ektov Pavlodarskogo aljuminievogo zavoda AO «Aljuminij Kazahstana» [Health risk assessment of the population living in the region where the facilities of Pavlodar aluminium smelter of JSC ‘Aluminium of Kazakhstan’ are located] / B.S. Imasheva, K.A. Askarov, M.S. Imashev [et al.] // Nauka i Zdravoohranenie [Science and Public Health]. — 2022. — № 3 (24). — P. 66–77. [in Russian]

3. Kamdina L.V. Razvitie metodicheskogo instrumentarija ocenki vlijanija antropogennyh faktorov promyshlennogo proizvodstva na kachestvo zhizni naselenija [Development of methodological tools for assessing the impact of anthropogenic factors of industrial production on the quality of life of the population]: abst. dis. ... PhD in Economics / Kamdina Ljudmila Vladimirovna. — 2019. — 23 p. [in Russian]

4. Vinogradova E.A. Sposoby jekologicheskogo monitoringa vozduha s primeneniem tehnologicheskih reshenij dlja ih obrabotki i analiza, vkljuchaja mashinnoe obuchenie, iskusstvennyj intellekt i analitiku bol'shih dannyh [Methods of ecological air monitoring with application of technological solutions for their processing and analysis, including machine learning, artificial intelligence and big data analytics] / E.A. Vinogradova, M.M. Dmitriev, A.S. Dudrjavec [et al.] // Sovremennye tehnologii: problemy i tendencii razvitija. Monografija [Modern Technologies: Problems and Trends of Development. Monograph]. — Petrozavodsk: New Science, 2021. — P. 174–189. [in Russian]

5. Zheleznyj S.V. Ispol'zovanie metodov mashinnogo obuchenija dlja prognozirovanija zagrjaznennosti atmosfernogo vozduha [Use of machine learning methods for forecasting atmospheric air pollution] / S.V. Zheleznyj, A.I. Sitnikov, A.A. Tolstyh // Vestnik Voronezhskogo instituta MVD Rossii [Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia]. — 2019. — № 3. — P. 73–81. [in Russian]

6. Subramaniam S. Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review / S. Subramaniam, N. Raju, A. Ganesan [et al.] // Sustainability. — 2022. — № 14. — 9951.

7. Qian Z. Machine Learning Explains Long-Term Trend and Health Risk of Air Pollution during 2015–2022 in a Coastal City in Eastern China / Z. Qian, Q. Meng, K. Chen [et al.] // Toxics. — 2023. — № 11. — 481.

8. Tello-Leal E. Evaluation of Deep Learning Models for Predicting the Concentration of Air Pollutants in Urban Environments / E. Tello-Leal, U.M. Ramirez-Alcocer, B.A. Macías-Hernández [et al.] // Sustainability. — 2024. — № 16. — 7062.

9. Wu X. A Water Quality Prediction Model Based on Multi-Task Deep Learning: A Case Study of the Yellow River, China / X. Wu, Q. Zhang, F. Wen [et al.] // Water. — 2022. — № 14. — 3408.

10. Sepman H. Machine Learning Tools Can Pinpoint High-Risk Water Pollutants / H. Sepman, P. Peets, L. Jonsson [et al.] // Proceedings. — 2023. — № 92. — 68.

11. Air Pollution Image Dataset from India and Nepal. — URL: https://www.kaggle.com/datasets/adarshrouniyar/air-pollution-image-dataset-from-india-and-nepal (accessed: 11.10.2024).

12. Algore M. Python For Data Analysis: The Ultimate and Definitive Manual to Learn Data Science and Coding With Python. Master The basics of Machine Learning, to Clean Code and Improve Artificial Intelligence / M. Algore. — 2021. — 87 p.

13. Antao T.R. High Performance Python for Data Analytics (MEAP) / T.R. Antao. — New York: Manning Publications, 2020. — 98 p.