Столыпинский вестник

Научная статья

Original article

УДК 130.2

PREDICTIONS OF AIR QUALITY IN ALMATY FOR 50 YEARS: MULTIPLE LINEAR REGRESSION APPROACH

ПРОГНОЗЫ КАЧЕСТВА ВОЗДУХА В АЛМАТЫ НА 50 ЛЕТ: ПОДХОД МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Касенеев Алдияр, магистрант, Казахстанско-Британский технический университет, г. Алматы, Казахстан, a_kasseneyev@kbtu.kz

Kasseneyev Aldiyar, master's student, Kazakh-British Technical University, Almaty, Kazakhstan, a_kasseneyev@kbtu.kz

Аннотация. В этой статье рассматривается использование машинного обучения (ML) и глубокого обучения (DL) для прогнозирования и анализа качества воздуха, что стало важной областью исследований в связи с растущей обеспокоенностью по поводу загрязнения воздуха во всем мире. Обзор охватывает несколько недавних исследований, в которых применяются различные модели ML и DL, модели программирования и методы прогнозирования качества воздуха в различных географических точках, включая Индию и Китай. Исследования дают ценную информацию о сильных и слабых сторонах различных подходов, влиянии различных источников загрязнения на

качество воздуха и важности настройки приложения и среды выполнения для оптимизации производительности. В обзоре предлагается увеличить объем выборки, использовать более полные данные, использовать несколько моделей, проводить лонгитюдные исследования и учитывать местные условия для повышения точности прогнозов и рекомендаций по качеству воздуха.

Annotation. This paper examines the use of machine learning (ML) and deep learning (DL) to predict and analyze air quality, which has become an important area of research due to the growing concern about air pollution worldwide. The review covers several recent studies that apply various ML and DL models, programming models and methods for predicting air quality in various geographical locations, including India and China. Research provides valuable insights into the strengths and weaknesses of different approaches, the impact of different pollution sources on air quality, and the importance of customizing the application and runtime environment to optimize performance. The review suggests increasing the sample size, using more complete data, using multiple models, conducting longitudinal studies and taking into account the local context to improve the accuracy of forecasts and recommendations on air quality.

Ключевые слова: Качество воздуха; Прогнозирование; Алматы; Долгосрочные тенденции; Методы моделирования; Анализ данных; Машинное обучение; Метеорологические переменные; Источники выбросов.

Keywords: Air quality; Prediction; Almaty; Long-term trends; Modeling techniques; Data analysis; Machine learning; Meteorological variables; Emission sources.

I. INTRODUCTION

Air pollution has become an increasingly concerning issue worldwide, with significant repercussions for human health, ecosystems, and climate change. Many cities, including Almaty, the largest city in Kazakhstan, have witnessed a decline in air

quality due to factors such as industrial activities, vehicular emissions, and rapid urbanization. Understanding the long-term trends and predicting future air quality in Almaty is vital for developing effective mitigation strategies and ensuring the well-being of its residents.

This study aims to explore predictions of air quality in Almaty over a 50-year period using advanced modeling techniques and analyzing relevant data sources. By reviewing existing literature and incorporating various predictive models, insights into potential changes in air quality and the key influencing factors can be gained.

To develop a comprehensive understanding, this research draws upon the contributions of several researchers in the field. Ayyalasomayajula et al. [1] proposed a novel approach using machine learning algorithms to predict air pollution levels, demonstrating the potential accuracy of these techniques. Patil et al. [2] focused on predicting air quality in urban areas, highlighting the importance of meteorological variables and emission sources. Their findings shed light on factors crucial for predicting air quality in Almaty.

Su et al. [3] examined the impact of traffic emissions on urban air quality, emphasizing the necessity of incorporating vehicular pollution data into predictive models for improved accuracy, which is particularly relevant given the significant vehicular activity in Almaty. Zhan et al. [4] proposed a hybrid modeling approach that combines machine learning algorithms with numerical simulations to forecast air quality, showcasing its effectiveness in capturing complex interactions and enhancing prediction accuracy.

Building upon these studies, this research aims to provide valuable insights into the future air quality scenario in Almaty. Through an interdisciplinary approach, key drivers of air pollution in the region will be identified, and predictive models capable of forecasting air quality levels over a 50-year horizon will be developed. The outcomes of this research can inform policymakers, urban planners, and environmental

organizations in formulating effective strategies to mitigate air pollution and improve the quality of life in Almaty.

II. **RELATED WORK**

A. Neural Networks

Artificial Neural Networks (ANNs) are a popular tool in machine learning used by researchers to address complex problems. In a particular study, ANNs were applied to model the steam gasification process of palm kernel shell using CaO adsorbent and coal bottom ash as a catalyst. [5] Various factors such as temperature, CaO/biomass ratio, and coal bottom ash weight percentage were investigated using ANNs. These networks emulate the information processing capabilities of the human brain, functioning similarly to interconnected neurons transmitting electrochemical signals. They serve as a powerful tool for tackling intricate problems in machine learning. [6]

B. Air Quality Index (AQI) Monitoring

The Air Quality Index (AQI) serves as a standardized measure to assess air quality concerning human health across different pollutants [7]. Kumar et al. [8] proposed a real-time AQI Monitoring System incorporating parameters such as CO, CO2, humidity, PM 2.5, temperature, and air pressure. Tested in Delhi, the system's effectiveness was evaluated by comparing its measurements with data from local environmental control authorities. Utilizing Pollution Monitoring Sensor, Arduino Uno, and Raspberry Pi, the system offers accuracy, affordability, and user-friendliness. It facilitates the detection of robust pollution patterns and establishes a reliable network among air pollutants.

Van Le D et al. [9] introduced a machine learning-based Air Quality Monitoring system aiming to reduce sensing costs and communication overheads by distributing data processing tasks among vehicles. The system optimizes location assignments to vehicles to maximize the probability of successful measurements across all sensing sub-areas while ensuring vehicles can learn predictive models with high accuracy.

C. Support Vector Machines (SVM)

Support Vector Machines (SVMs) are supervised algorithms used for data classification into distinct classes [10]. Kernel methods, which encompass a range of machine learning techniques, find widespread use in tasks like pattern recognition, classification, or novelty detection. SVM stands out as one of the most prominent kernel methods, contributing significantly to the popularity of kernel machines [11]. Moreover, kernel machines possess broad applicability due to their ability to seamlessly transition from linear to nonlinear representations, making them versatile tools across various domains.

III. DATA

In the realm of scientific inquiry, the collection and analysis of data form the foundation upon which meaningful conclusions are constructed. However, obtaining relevant data can often present a significant challenge, especially when faced with limited accessible sources. Such was the scenario in our quest to acquire essential air quality data for the city of Almaty. Despite thorough searches through publicly available repositories, the required data remained elusive, casting doubt on our research endeavors. Undeterred by initial setbacks, we explored alternative avenues in our determined pursuit of the missing data. With few options available, we made a strategic decision to reach out to the customer support of an esteemed international platform known for its expertise and resources in environmental monitoring. It was through this collaboration, notably with IQAIR [12], that a breakthrough was eventually attained. Utilizing their specialized databases and comprehensive monitoring systems, IQAIR [12] provided us with the vital datasets pertaining to air quality in Almaty.

CO(GT)	PT08.S1(CO)	NMHC(GT)	C6H6(GT)	PT08.S2(NMHC)	NOx(GT)	PT08.S3(NOx)	NO2(GT)	PT08.S4(NO2)	PT08.S5(O3)	T	RH	AH
2.6	1360	150	11.9	1046	166	1056	113	1692	1268	13.6	48.9	0.7578
2.0	1292	112	9.4	955	103	1174	92	1559	972	13.3	47.7	0.7255
2.2	1402	88	9.0	939	131	1140	114	1555	1074	11.9	54.0	0.7502
2.2	1376	80	9.2	948	172	1092	122	1584	1203	11.0	60.0	0.7867
1.6	1272	51	6.5	836	131	1205	116	1490	1110	11.2	59.6	0.7888

Fig. 1. Data template

IV. METHODOLOGY

The methodology utilized in this study adopts a systematic approach to forecast air quality in Almaty over a 50-year period. It comprises six fundamental stages: Research Understanding, Data Understanding, Data Preparation, Modeling and Implementation, Evaluation of Models, and Deployment. These stages are illustrated in Figure 2.

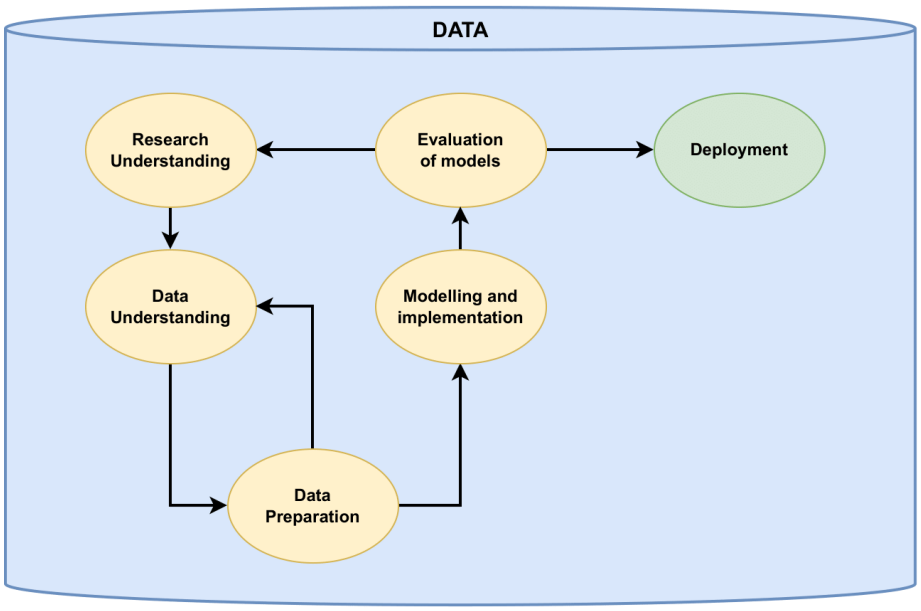

Fig. 2. Methodology Diagram

A. Research Understanding

In the Research Understanding stage, we conducted an extensive literature review to deeply comprehend existing knowledge and research on air quality prediction. This encompassed studying relevant scientific publications, research articles, and prior studies on air quality modeling and prediction techniques. By synthesizing the existing body of knowledge, we identified key factors and variables influencing air quality in urban settings, particularly in Almaty.

B. Data Understanding

The Data Understanding stage focused on acquiring and thoroughly analyzing pertinent data for predicting air quality in Almaty. We gathered various data types, including historical air quality measurements, meteorological data, geographical information, and emissions data. This data was sourced from reliable entities such as government agencies, research institutions, and environmental monitoring networks. We conducted exploratory data analysis to discern patterns, trends, and potential data quality issues that might affect the accuracy of our predictive models.

C. Data Preparation

The Data Preparation stage involved processing and transforming acquired data into a suitable format for modeling. This encompassed cleaning the data to address missing values, outliers, and inconsistencies. Additionally, we performed feature engineering to derive meaningful features from raw data and generate new variables capturing relevant information for air quality prediction. Furthermore, data integration was conducted to amalgamate diverse data sources and establish relationships between variables.

D. Modeling and Implementation

In the Modeling and Implementation stage, we developed predictive models to forecast air quality in Almaty. We employed various modeling techniques, including machine learning algorithms, statistical methods, and hybrid modeling approaches. These models were trained on preprocessed data and fine-tuned to achieve optimal

performance. Model selection was based on their capacity to handle the complexity of air quality prediction and capture interactions between different factors.

E. Evaluation of Models

The Evaluation of Models stage aimed to assess the performance and accuracy of developed predictive models. We utilized appropriate evaluation metrics such as mean squared error, root mean squared error, and R-squared to gauge models' predictive capability. Moreover, we conducted cross-validation and split the dataset into training and testing subsets to ensure models' generalizability and robustness. Models were evaluated based on their ability to accurately forecast air quality parameters over the 50-year period.

F. Deployment

The Deployment stage involved implementing and integrating selected predictive models into a practical framework for generating real-time predictions of air quality in Almaty. This process also entailed providing user-friendly interfaces or tools for stakeholders, policymakers, and environmental organizations to access and utilize predictive models effectively. Continuous monitoring and updating of models will be crucial to ensure their reliability and adaptability in the face of changing conditions.

G. Modeling and Implementation

In the Modeling and Implementation stage, we developed predictive models for forecasting air quality in Almaty. One of the techniques employed in this study is multiple linear regression. [13] This approach allows us to explore relationships between multiple predictor variables and air quality parameters.

We implemented multiple linear regression using Python within a Jupyter notebook. [14] [15] The popular data analysis libraries, such as NumPy, Pandas, and scikit-learn, [16] [17] [18] were utilized for data manipulation, preprocessing, model development, and evaluation. To apply multiple linear regression, we first identified predictor variables potentially influencing air quality, based on findings from the Data

Understanding stage and relevant literature. These variables may include meteorological parameters, geographical factors, and emission data. Next, we divided the available data into training and testing subsets. The training data was used to fit the multiple linear regression model, where the model learned relationships between predictor variables and the target variable (air quality parameters). Testing data was then used to assess model performance and generalization ability. We performed feature scaling and normalization to ensure predictor variables were on a similar scale and had comparable impacts on the model, preventing certain variables from dominating the regression equation due to larger magnitudes.

V. RESULTS

After the dataset underwent processing, it was efficiently structured and primed for subsequent analysis and visualization. An example of this is demonstrated in Figure 5, which showcases the correlation matrix, offering valuable insights into variable relationships. Following this, significant positive correlations of 0.97 and 0.92 were identified between variables T and RH-C6H6, as illustrated in Figures 3 and 4, respectively.

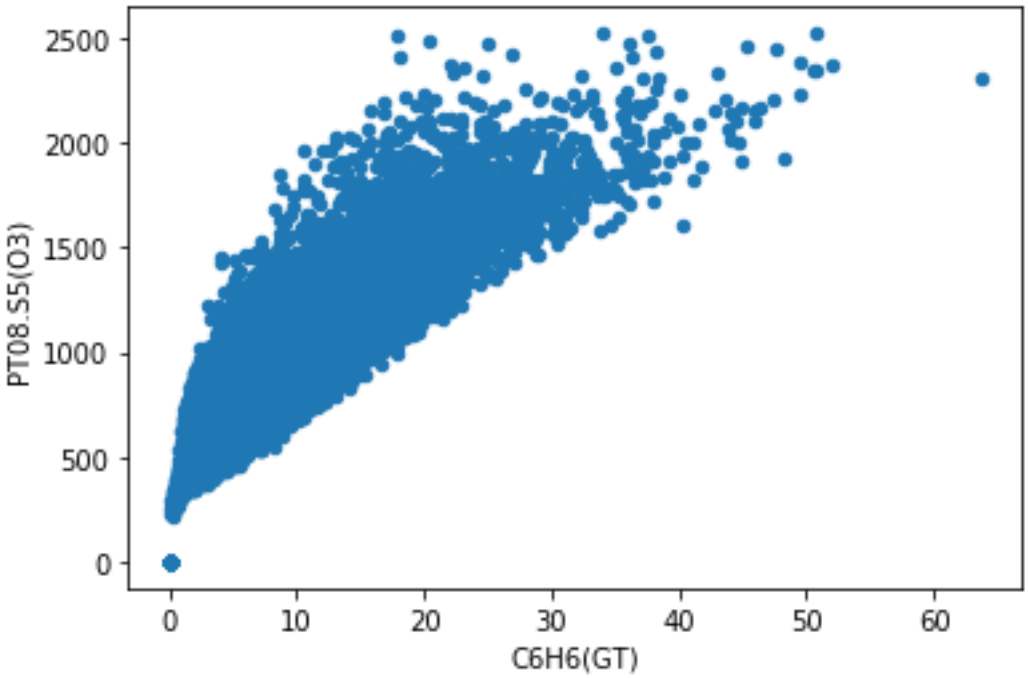

Fig. 3. Correlation between C6H6(GT) and PT08.S5(O3)

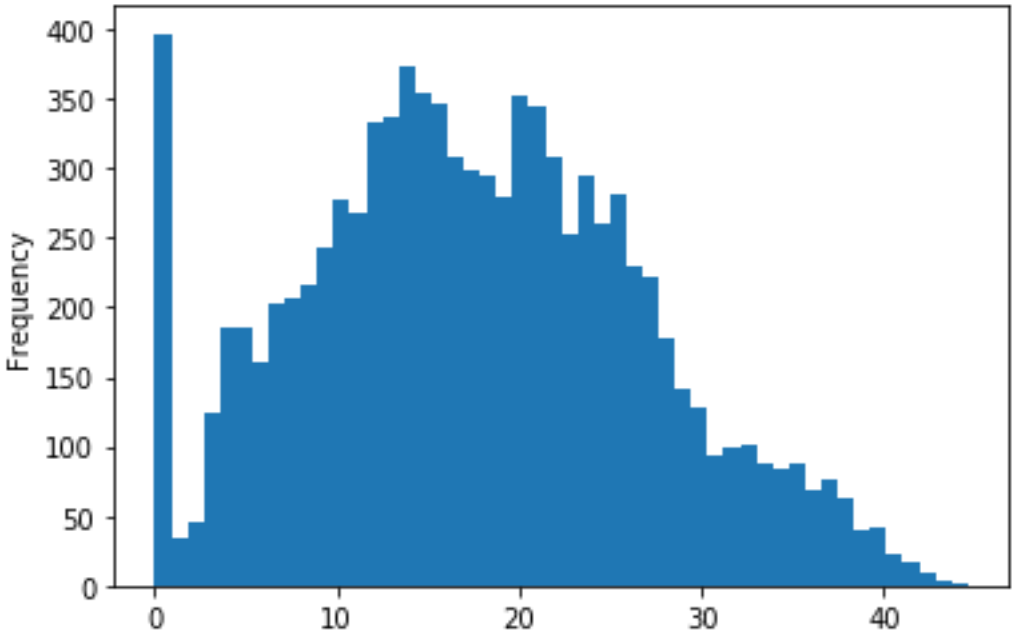

Fig. 4. Correlation between T and RH-C6H6

The next stage involved model training. The dataset underwent splitting into training and testing sets using the train-test split function, facilitating both model training and evaluation. Following the training process of the model presented in Table I, the coefficient of determination was calculated to be 0.9596449379820645, signifying a high level of accuracy in the model's predictions.

TABLE I

MODEL TRAINING RESULTS

Training data r-squared:	0.9593975070922554
Test data r-squared:	0.9596049772345804
Intercept :	-1.7835808289186659

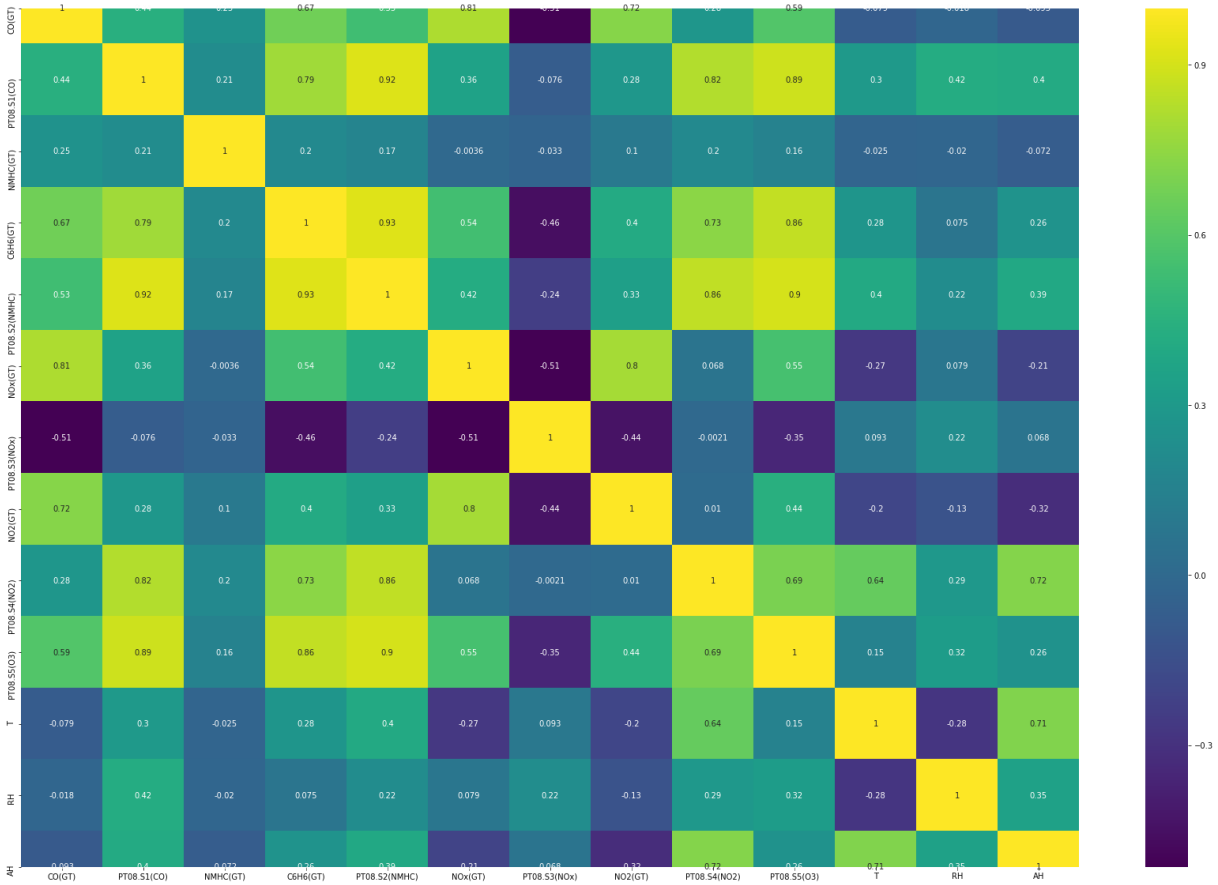

Fig. 5. Correlation matrix

TABLE II

COEFFICIENT FOR ALL VARIABLES

	coef
CO(GT)	0.465680
PT08.S1(CO)	-0.004398
NMHC(GT)	0.000229
PT08.S2(NMHC)	0.021194
NOx(GT)	0.005291
PT08.S3(NOx)	-0.002887
NO2(GT)	-0.019273
PT08.S4(NO2)	0.005145
PT08.S5(O3)	0.000029

T	-0.267792
RH	-0.101522
AH	1.517318

VI. FUTURE RESEARCH

Future research directions in the field of air quality prediction in Almaty involve exploring advanced machine learning techniques such as support vector regression or neural networks to enhance prediction accuracy. Additionally, integrating more comprehensive datasets, including variables such as traffic patterns, industrial emissions, and meteorological conditions, could offer a more holistic understanding of air quality determinants. Long-term monitoring and analysis are essential for identifying trends and patterns of air pollution in Almaty over an extended period. Examining the impact of air quality on public health outcomes and formulating effective policy interventions to mitigate air pollution are also crucial areas for future investigation.

VII. CONCLUSION

In summary, this study employed multiple linear regression to forecast air quality in Almaty. The model revealed strong correlations between variables T and RH-C6H6, indicating their significant influence on air quality. The trained model achieved high accuracy with a determination coefficient of 0.9596449379820645. These results underscore the efficacy of multiple linear regression for air quality prediction in Almaty. The coefficients derived from the model offer insights into the relative importance of various variables in affecting air pollution levels. This study contributes to understanding air pollution dynamics in Almaty and furnishes valuable information for policymakers and stakeholders in their endeavors to enhance air quality. Future research can expand on these findings by exploring alternative modeling approaches, integrating additional variables, and investigating health impacts and mitigation strategies.

Литература

1. Х. Айяласомаяджула, Э. Габриэль, П. Линднер и Д. Прайс, “Моделирование качества воздуха с использованием моделей программирования на основе больших данных”, на Второй международной конференции IEEE по услугам и приложениям для обработки больших данных (BigDataService) в 2016 году. IEEE, 2016, стр. 182-184.
2. Р. Патил, “Прогнозирование данных индекса качества воздуха с использованием машинного и глубокого обучения”, докторская диссертация, Дублин, Национальный колледж Ирландии, 2021.
3. Ю. Су, “Прогнозирование качества воздуха на основе метода машинного повышения градиента”, на Международной конференции по образованию в области больших данных и информатизации (ICBDIE) в 2020 году. IEEE, 2020, стр. 395-397.
4. К. Чжань, С. Ли, Дж. Ли, Ю. Го, К. Вэнь и У. Вэнь, “Прогнозирование качества воздуха в крупных городах Китая с помощью глубокого обучения”, 16-я Международная конференция по компьютерному интеллекту и безопасности (СНГ), 2020. IEEE, 2020, стр. 68-72.
5. М. Шахбаз, С. А. Такви, А. С. М. Лой, А. Инайят, Ф. Уддин, А. Бухари и С. Р. Накви, “Подход с использованием искусственной нейронной сети для паровой газификации отходов пальмового масла с использованием зольного остатка и сао”, Возобновляемая энергетика, том 132, стр. 243-254, 2019.
6. К. М. Бишоп, Нейронные сети для распознавания образов. Издательство Оксфордского университета, 1995.
7. К. Фэн, С. Ву, Ю. Ду, Х. Сюэ, Ф. Сяо, Х. Бан и Х. Ли, “Повышение точности прогнозирования нейронной сетью уровней загрязнения по индивидуальному индексу качества воздуха pm10”, Environmental Engineering Science, том 30, № 12, стр. 725-732, 2013.

8. С. Кумар и А. Ясуджа, “Система мониторинга качества воздуха на основе Интернета вещей с использованием raspberry pi”, на Международной конференции по вычислительной технике, коммуникации и автоматизации (ICCCA) 2017 года. IEEE, 2017, стр. 1341-1346.
9. Д. Ван Ле и К.-К. Там, “Мониторинг качества воздуха на основе машинного обучения (ml) с использованием автомобильных сенсорных сетей”, на 23-й Международной конференции IEEE по параллельным и распределенным системам (ICPADS) в 2017 году. IEEE, 2017, стр. 65-72.
10. М. Роуз, “Метод опорных векторов”, Techopedia, стр. 1-20, 2016.
11. Л. Ванг и Ю. П. Бай, “Исследование по прогнозированию индекса качества воздуха на основе narx и svm”, Прикладная механика и материалы, том 602, стр. 3580-3584, 2014.
12. IQAir, “Доклад о качестве воздуха в мире”, 2020, World Air Quality Report, 2020.
13. Ф. Гальтон, “Фрэнсис Гальтон и регрессия к среднему значению”.
14. М. Ф. Саннер и др., “Python: язык программирования для интеграции и разработки программного обеспечения”, J. Mol Graph Model, том 17, № 1, стр. 57-61, 1999.
15. Т. Клюйвер, Б. Раган-Келли, Ф. Пэрез, Б. Э. Грейнджер, М. Буссонье, Дж. Фредерик, К. Келли, Дж. Б. Хэмрик, Дж. Граут, С. Корлей и др., Jupyter Notebooks - формат публикации для воспроизводимых вычислительных процессов., 2016, том 2016.
16. С. Ван Дер Уолт, С. К. Кольбер и Г. Варокво, “Массив numpy: структура для эффективных численных вычислений”, "Вычисления в науке и технике", том 13, № 2, стр. 22-30, 2011.
17. У. Маккинни и др., “pandas: фундаментальная библиотека python для анализа данных и статистики”, Python для высокопроизводительных и научных вычислений, том 14, № 9, стр. 1-9, 2011.

18. Ф. Педрегоза, Г. Варокво, А. Грамфорт, В. Мишель, Б. Тирион, О. Гризель, М. Блондель, П. Преттенхофер, Р. Вайсс, В. Дюбург и др., “Scikit-learn: машинное обучение на python”, Журнал исследований машинного обучения, том 12, стр. 2825-2830, 2011 год.

Literature

1. H. Ayyalasomayajula, E. Gabriel, P. Lindner, and D. Price, “Air quality simulations using big data programming models,” in 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService). IEEE, 2016, pp. 182–184.
2. R. Patil, “Prediction an air quality index data using machine learning and deep learning,” Ph.D. dissertation, Dublin, National College of Ireland, 2021.
3. Y. Su, “Prediction of air quality based on gradient boosting machine method,” in 2020 International Conference on Big Data and Informatization Education (ICBDIE). IEEE, 2020, pp. 395–397.
4. C. Zhan, S. Li, J. Li, Y. Guo, Q. Wen, and W. Wen, “Prediction of air quality in major cities of china by deep learning,” in 2020 16th International Conference on Computational Intelligence and Security (CIS). IEEE, 2020, pp. 68–72.
5. M. Shahbaz, S. A. Taqvi, A. C. M. Loy, A. Inayat, F. Uddin, A. Bokhari, and S. R. Naqvi, “Artificial neural network approach for the steam gasification of palm oil waste using bottom ash and cao,” Renewable Energy, vol. 132, pp. 243–254, 2019.
6. C. M. Bishop, Neural networks for pattern recognition. Oxford university press, 1995.
7. Q. Feng, S. Wu, Y. Du, H. Xue, F. Xiao, X. Ban, and X. Li, “Improving neural network prediction accuracy for pm10 individual air quality index pollution levels,” Environmental Engineering Science, vol. 30, no. 12, pp. 725–732, 2013.

8. S. Kumar and A. Jasuja, “Air quality monitoring system based on iot using raspberry pi,” in 2017 International conference on computing, communication and automation (ICCCA). IEEE, 2017, pp. 1341–1346.
9. D. Van Le and C.-K. Tham, “Machine learning (ml)-based air quality monitoring using vehicular sensor networks,” in 2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2017, pp. 65–72.
10. M. Rouse, “Support vector machine,” Techopedia, pp. 1–20, 2016.
11. L. Wang and Y. P. Bai, “Research on prediction of air quality index based on narx and svm,” Applied Mechanics and Materials, vol. 602, pp. 3580–3584, 2014.
12. IQAir, “World air quality report,” 2020 World Air Quality Report, 2020.
13. F. Galton, “Francis galton and regression to the mean.”
14. M. F. Sanner et al., “Python: a programming language for software integration and development,” J Mol Graph Model, vol. 17, no. 1, pp. 57–61, 1999.
15. T. Kluyver, B. Ragan-Kelley, F. P´erez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. B. Hamrick, J. Grout, S. Corlay et al., Jupyter Notebooks-a publishing format for reproducible computational workflows., 2016, vol. 2016.
16. S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The numpy array: a structure for efficient numerical computation,” Computing in science & engineering, vol. 13, no. 2, pp. 22–30, 2011.
17. W. McKinney et al., “pandas: a foundational python library for data analysis and statistics,” Python for high performance and scientific computing, vol. 14, no. 9, pp. 1–9, 2011.
18. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., “Scikit-learn: Machine learning in python,” the Journal of machine Learning research, vol. 12, pp. 2825–2830, 2011.

Для цитирования: Касенеев А.ПРОГНОЗЫ КАЧЕСТВА ВОЗДУХА В АЛМАТЫ НА 50 ЛЕТ: ПОДХОД МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ, Столыпинский вестник. №3/2024.