University of Novi Sad
Technical faculty "Mihajlo Pupin" Zrenjanin

Proceedings of the 14th International Conference on Applied Internet and Information Technologies AIIT 2024

8 November, 2024, Zrenjanin, Serbia

AIIT

International Conference

University of Novi Sad
Technical faculty
"Mihajlo Pupin"
Zrenjanin
Republic of Serbia

XIV INTERNATIONAL CONFERENCE ON APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2024
PROCEEDINGS

November 8, 2024
Zrenjanin
Serbia

Proceedings publisher and organizer of the conference:
University of Novi Sad, Technical faculty “Mihajlo Pupin”, Zrenjanin, Republic of Serbia

For publisher:
Milan Nikolic, PhD
Dean of Technical faculty “Mihajlo Pupin”,
Zrenjanin, Republic of Serbia

Proceedings editors:
Ivana Berkovic, PhD
Kostandina Veljanovska, PhD

Conference Chairmans:
Ivana Berkovic, Technical faculty “Mihajlo Pupin”, University of Novi Sad, Zrenjanin, Serbia
Kostandina Veljanovska, Faculty of Information and Communication Technologies, University “St. Kliment Ohridski”, Bitola, North Macedonia

Technical preparation of the proceedings:
Zoltan Kazi, PhD
Sinisa Mihajlovic, MSc
Aleksandra Stojkov Loncarski, MSc
Milica Mazalica, MSc

Cover design:
Visnja Ognjenovic, PhD

e-Proceedings
ISBN 978-86-7672-379-9

CIP - Каталогизација у публикацији
Библиотеке Матице српске, Нови Сад

004(082)(0.034.4)

INTERNATIONAL Conference on Applied Internet and Information Technologies (14 ; 2024 ; Zrenjanin)
Proceedings [Elektronski izvor] / XIV International Conference on Applied Internet and Information Technologies AIIT 2024, Zrenjanin, November 8th, 2024 ; [organizer] Technical Faculty "Mihajlo Pupin", Zrenjanin ; [editors Ivana Berkovic, Kostandina Veljanovska]. - Zrenjanin : Tehnički fakultet "Mihajlo Pupin", 2024. - 1 elektronski optički disk (CD-ROM) : tekst, ilustr. ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Naslov sa nasl. ekrana. - Elektronska publikacija u formatu pdf opsega 504 str. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-7672-379-9

a) Информационе технологије -- Зборници

COBISS.SR-ID 158823945

TABLE OF CONTENTS

KNN Algorithm Implementation in Real-World Problem of Water Quality Classification

Kostandina Veljanovska[1], Aneta Trajkovska[1], Nikolce Veljanovski[2]

[1] *University "St. Kliment Ohridski", Bitola, Faculty of Information and Communication Technologies, Department of Intelligent Systems, Bitola, Republic of N. Macedonia*
[2] *NIUL "St. Kliment Ohridski", Bitola, Pece Maticevski 39, 7000 Bitola, Republic of N. Macedonia*

kostandina.veljanovska@uklo.edu.mk; aneta.trajkovska@uklo.edu.mk; nveljan@yahoo.com

Abstract:
Technological evolution is increasingly focused on developing intelligent machine learning solutions that improve the efficiency, functionality and adaptability of everyday systems and processes. Effective utilization of algorithms fundamentally enables systems to learn, adapt to environmental conditions and make decisions based on the knowledge, characteristics that define intelligent agents. This capability includes classifying various processes and refining data through user-defined characteristics to produce accurate results. This paper focuses on determining water quality. Water bodies are essential part of our environment and our aim is to contribute to maintain healthy environment for humans, animals and plants. To maintain water quality on high level is very important in order to protect human and all live creatures' health, to avoid the costs related to medical care, costs of productivity loss, and even loss of life. We have implemented the K-Nearest Neighbors (KNN) algorithm to learn from the dataset and develop a classification process based on various parameters. The algorithm quantifies water quality levels based on the concentration of constituent substances. The KNN algorithm is utilized to train the model with the provided data and its performance is evaluated using predefined coefficient values. Results from the implementation of the model yields an accuracy percentage, demonstrating its effectiveness in determining water quality.

Keywords:
Machine learning, K-Nearest Neighbors algorithm, artificial intelligence, intelligent systems, water quality prediction, water quality classification

1. Introduction

Artificial intelligence includes a wide range of technologies, including machine learning, data mining, augmented reality, robotics, logic programming, speech recognition, image processing, ambient intelligence, data visualization and predictive analytics [1]. These methods are applied through various approaches and analyses tailored to specific needs and problems. The focus of machine learning is to automatically learn to recognize complex patterns and make intelligent decisions based on data. The process of learning is shown in Figure 1.

Figure 1: Machine learning progress

Machine learning algorithms can be broadly categorized into three main types: supervised learning, unsupervised learning and reinforcement learning [2]. Each type serves distinct purposes and is applied in various contexts based on the nature of the data and the specific objectives of the analysis [3].

Clean water in every water body is essential for human health, ecosystems, the environment, and the overall wellbeing of our community. Detection and first of all preventing water pollution is essential in order to safeguard water resources, to ensure a sustainable future, and to protect the environment and also endangered population.

In order to protecting world's water resources and environment requires from us to put collective effort. Humans need to foster a sense of responsibility within the community, and as a result help to preserve clean and healthy waterways and aquatic environments for future generations.

Water pollution as a significant worldwide problem can be caused by a variety of sources including industrial waste, inappropriate disposal of various household or institutional chemicals and agricultural runoff. Water bodies which are polluted with dangerous compounds provide significant risks to aquatic ecosystems, human health, and the economy [4].

Within this article, we aim to provide a comprehensive overview of the implementation of KNN algorithm via practical case for determining the water quality based on the concentration of constituent substances in different locations.

The paper's structure incorporates theoretical background of KNN algorithm, the next chapter is discussion about the importance of water quality. In the chapter four, water quality prediction model is presented. Limitations and alternative ways to improve the KNN algorithm in our case are discussed in the next chapter and at the end we conclude with findings during this research.

2. KNN Algorithm

There are numerous examples of machine learning algorithms implementation in real-world problems solving and also in water quality determination and prediction [5]. As a powerful tool for solving everyday problems, machine learning was used in this research in order to predict level of water pollution. KNN is a widely utilized supervised learning algorithm known for its simplicity and effectiveness in classification tasks [6]. The core principle is to classify a data point based on the classifications of its nearest neighbors within the feature space. KNN operates on the theory that similar instances tend to be located close to one another in the multidimensional space defined by the input features. When a new data point is introduced, KNN assesses its proximity to the existing labeled instances in the dataset [7]. By evaluating the labels of the K nearest neighbors—where K is a user-defined parameter, the algorithm assigns a classification to the new data point based on the majority label among its neighbors [8]. The algorithm's performance hinges on several factors, including the choice of distance metric (such as Euclidean or Manhattan distance), the selection of K and the feature scaling of the input data [9]. Its interpretability and ease of implementation make it a valuable choice for exploration data analysis and situations where a straightforward classification model is desirable.

The KNN algorithm has demonstrated its utility in a diverse array of real-world applications, ranging from recommendation systems to medical diagnostics. Its ease of implementation, coupled with its capacity to handle complex data relationships, makes KNN a valuable tool in various fields [10]. It is used in healthcare (identifying potential diseases or conditions based on the similarity of new patient data to previously diagnosed cases), agriculture (determining the quality of the soil), finance, social media, detecting the air quality [11], etc. Many popular companies utilize KNN in their systems such as Amazon, Spotify, Google, Netflix, LinkedIn, Uber and others.

3. Water Quality

Water pollution appears when different foreign materials contaminate river, lake, sea, ocean or another water body. Often the materials are chemicals, organic, inorganic, biological, radiological, microorganisms or many other substances dissolvable in water which can degrade water quality making the water toxic for people and for environment.

The water, together with the air and the soil are the main elements of our environment. In attempt to address environmental pollution, environmental education is a key factor. [12] Environmental and water pollution education has its main objective to provide learners with knowledge, values and skills that promote the protection and conservation of the environment.

The best thing you can do to prevent water pollution is to educate yourself and your environment about the world's water supply, how you can detect clear water body, understand water pollution and find best means to protect the water resources. Water pollution and water quality education is a very important and effective tool to promote public awareness. We would like to contribute towards this goal by one real world implementation of machine learning algorithm in water quality prediction.

Depending on the aim for water body usage, pollution should be addressed very carefully. Water is a very significant source for all living organisms since all the living organisms need water so it is life itself. Water keeps all of them alive but polluted water is proved to be a very harmful substance. When humans and animals drink polluted water it has serious effects on their health. When their skin is in contact with polluted water, consequences are numerous. Some types of water pollution could be dangerous even for plants. Therefore, it is essential to find models that will react on water pollution quickly and in real-time. Machine learning algorithms as artificial intelligence algorithms implemented in real-time have the capability to gather data form on-site sensors and detect the level of pollution.

4. Water Quality Classification Model

This research focuses on classifying water quality based on specific measurement parameters across various locations. The classification framework delineates water into three distinct quality levels: high, medium and low. This assessment incorporates both the geographical context and the concentrations of various substances present in the water, employing weighting factors to evaluate their impact on overall water quality.

4.1. Methodology for Analyzing Water Quality Classification

Data collection: the dataset includes measurements of various water quality parameters: nitrate, ammonium, total phosphor, phosphate and chlorophyll. There are other parameters that has to be taken into account, but in order to develop this model and to provide quick results, we determine the scope of our research only to these parameters. Our future research will encompass more parameters and we plan to include parameters joined effect prediction on water pollution taking into account few measurements points next to each another.

Data preprocessing phase: in this phase we have ensured data integrity by data cleaning. It was important to understand well the data and do normalizations to ensure that all variables contributed equally to the classification process. Also adding weighting factors based on the relative importance of determining water quality, based on the following metrics Table 1:

Table 1:
Classes of substances for determining water quality

Parameter	Class High Quality	Medium Quality	Class low Quality	Weighting Factor
Nitrate	0-4,5	4,5-8,5	>8,5	5
Ammonium	0-0,6	0,6-1,6	>1,6	5
TP	0-0,15	0,15-0,7	>0,7	2
Phosphate	0-0,075	0,075-0,35	>0,35	4
Chlorophyll	0-60	60-175	>175	3

Classification framework: water quality was classified into three categories: high, medium, and low, as per the values in Table 1.

Water classification was done based on the Table 1 and taking into account that the parameters that have weighting factor 4 and 5 are crucial. We take water quality value according to those parameters which are prevalent. If one or more of the parameters weighting 4 and 5 have certain value, we take water quality to be of that value.

Algorithm implementation: the KNN algorithm was implemented for the classification process, with utilization of Python programing language, Figure 2.

	monitoringSiteIdentifier	observedPropertyDeterminandLabel	phenomenonTimeSamplingDate	resultObservedValue	ClassQuality
0	EESJA9706000	Ammonium	20140527	0.0700	med
1	IT13PI16	Ammonium	20140527	0.0319	high
2	IT13PE04	Ammonium	20140527	0.0253	high
3	IT13AL13	Ammonium	20140527	0.0183	high
4	IT13GU01	Ammonium	20140527	0.0133	high
5	IT13GU03	Ammonium	20140527	0.0010	high

Figure 2: Analyzing the presence of Ammonium in the water and determining the class quality

Training and Testing Split: dataset was divided into training and testing sets to evaluate the algorithm's performance, demonstrated in Table 2 below.

Table 2:
Testing values of the classifier k in the KNN algorithm

Number of data loaded (Database)	Training data	Testing data
23137	18509	4628
1500	1199	300
2500	1999	500
9500	7555	1889

K value determination: various values of k classifier were tested to identify the optimal number of neighbors that yielded the highest classification accuracy, shown in Table 3.

Table 3:
Testing values of the classifier k in the KNN algorithm

Parameter	Accuracies	Macro Avg (%)	Weighted Avg (%)
K=1	0.81	0.82	0.82
K=5	0.79	0.79	0.79
K=20	0.74	0.77	0.76

Model evaluation: according to the results obtained, it can be concluded that although we have a change in input data, the KNN algorithm itself has the highest accuracy when the value of the parameter k=1, Figure 3.

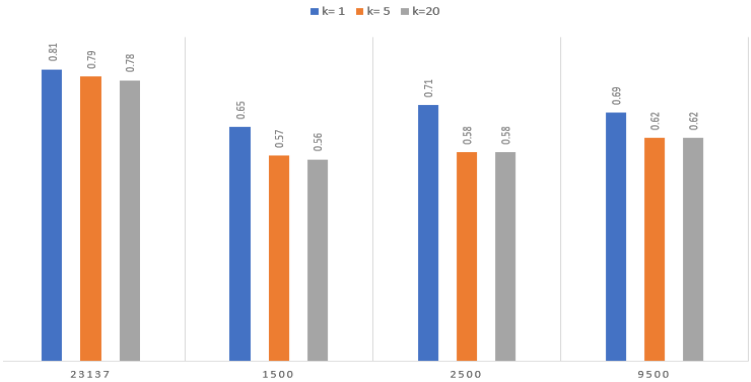

Figure 3: Results obtained from testing the algorithm by loading different numbers of data from the database

4.2. Water Classification Based on Date Period and Location

Representation of the results from classification by the algorithm for date 01.07.2020, results shown that the water has medium quality, more details in Table 4.

Table 4:
Results from the analysis of the values entered on 20200107 for location **SE647907-123283**

Location ID	Nitrate Class	Ammonium Class	Phosphate Class	TP Class	Water Quality (WF)
SE647907-123283	~5.81 med	~1.09 med	~0.54 low	~0.77 low	med

5. Limitations and Alternatives to improve the KNN algorithm

Despite its strengths, KNN also has limitations such as computational inefficiency in large datasets, memory intensive and dimensionality issues. The choice of the parameter K is crucial, too small a value can led to overfitting, while too large a value can result in underfitting. Determining the optimal K often requires cross-validation. As the dataset grows, the K-Nearest Neighbors (KNN) algorithm's can lead to longer processing times and increased resource consumption [13]. There are many alternatives that can be implemented to improve the performance of the KNN algorithm: dimensionality reduction, weighted KNN, using efficient data structures, distance metric variations, hybrid approaches and incremental learning [14].

6. Conclusions

Water pollution and water quality education together with scientific effort to determine water quality for each water body is a very important and effective tool to promote public awareness. Our aim is to contribute towards clean, healthy and safety environment by one real world implementation of machine learning algorithm in water quality prediction.

The comprehensive analysis presented in this paper demonstrates the outcomes of the KNN algorithm. KNN algorithm was structured around functions and the implementation of relevant datasets which underwent preprocessing to facilitate the initial learning phase. This step was crucial for enabling the algorithm to perform accurate classifications based on predefined rules embedded in the code. The experimental process revealed how the algorithm's performance was influenced by variations in the data input, demonstrating its adaptability to different datasets. These datasets served as testing resources to refine the classification of water quality into three distinct categories: low, medium, and high. The results showed how the ratio between the weight factors of individual compounds impacted the overall classification, ensuring accurate distribution across the water quality classes. This approach highlights the significance of the influence of each compound on the classification process and its contribution to a more precise evaluation of water quality.

7. Acknowledgment

The research work undertaken to study the problem of water pollution and main parameters of water quality, as well as the way the water pollution can be estimated and predicted was performed as a long-time joined effort with our colleagues from Kristianstad University, Department of Education and the Environment. We would like to thank Hristina Bodin, PhD. for their contribution in determining water substances that are important in water quality classification and water quality prediction. They also provided necessary data for the algorithm training. This paper is one small part of the above mentioned research that highlights the power of machine learning in problem of water pollution solution.

References:

[1] N. Kühl, M. Schemmer, M. Goutier et al. Artificial intelligence and machine learning. Electron Markets 32, 2235–2244 (2022). https://doi.org/10.1007/s12525-022-00598-0

[2] B. Mahesh. "Machine Learning Algorithms - A Review", IJSR (2020), Vol.9, Issue 1, doi:10.21275/ART20203995

[3] A. Singh, N. Thakur and A. Sharma. "A review of supervised machine learning algorithms", IEEE, 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 1310-1315
[4] Yadav, Shyma. (2024). Water Pollution: The Problems and Solutions. Science Insights. 44. 1245-1251. 10.15354/si.24.re905.
[5] Ahmed, A. N. et al., Machine learning methods for better water quality prediction, Journal of Hydrology, Volume 578, 2019, 124084, ISSN 0022-1694.
[6] A. Moldagulova and R. B. Sulaiman. "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), Amman, Jordan, 2017, pp. 665-671, doi: 10.1109/ICITECH.2017.8079924.
[7] M. Steinbach, P.N. Tan. "kNN: k-Nearest Neighbors", (2009), eBook ISBN: 9780429138423
[8] G. Batista, D. F. Silva. "How k-Nearest Neighbor Parameters Affect its Performance", Argentine symposium on artificial intelligence, doi: b83da505b60e32469152b986cbd6199842403b11
[9] K. Veljanovska. "Machine Learning Algorithms Comparison", International Journal of Engineering and Technical Research (2017), Vol.7, Issue11, ISSN: 2321-0869 (O) 2454-4698 (P)
[10] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
[11] P. Wang, Y. Zhang and W. Jiang, "Application of K-Nearest Neighbor (KNN) Algorithm for Human Action Recognition," IEEE (2021), pp. 492-496, doi: 10.1109/IMCEC51613.2021.9482165.
[12] Kılıç, Z., Water Pollution: Causes, Negative Effects and Prevention Methods, Istanbul Sabahattin Zaim University Journal of the Institute of Science and Technology (2021) 3 (1): 129-132.
[13] Halder, R.K., Uddin, M.N., Uddin, M.A. *et al.* Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *J Big Data* 11, 113 (2024). https://doi.org/10.1186/s40537-024-00973-y
[14] B. Sun, J. Du and T. Gao. "Study on the Improvement of K-Nearest-Neighbor Algorithm," *2009 International Conference on Artificial Intelligence and Computational Intelligence*, Shanghai, China, 2009, pp. 390-393, doi: 10.1109/AICI.2009.312.