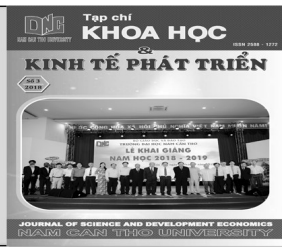
Journal of Science and Development Economics
Nam Can Tho University

Website: jsde.nctu.edu.vn



Water quality prediction using MLP in dynamic environment

Tran Van An[1], Kieu Tien Binh[1], Nguyen Dinh Thuy Huong[2], Ngo Ho Anh Khoi[1]
[1]*Faculty of Information Technology, Nam Can Tho University*
[2]*Vietnam Maritime University*

Received: November 10, 2023
Revised: November 30, 2023
Accepted: December 5, 2023

Keyword: *mssmartyPants, MLP classifier, water quality*

Từ khoá: *chất lượng nước, phân loại MLP*

ABSTRACT

Water quality plays an immensely significant role in safeguarding the health of humans and other species on the planet. Polluted water sources can contain harmful substances such as heavy metals, pesticides, bacteria, and viruses, which can pose threats to human health and other organisms. Water quality is also a crucial factor in protecting the environment and sustaining the ecosystems of freshwater and marine environments. Polluted water sources can impact the lives of aquatic animals and plants, causing harm and endangering biodiversity by disrupting the food chain. Within the scope of this article, we utilize data from "Water Quality" by MssmartyPants using the classical MLP Classifier algorithm. Incorporating artificial intelligence into water quality assessment contributes to increased accuracy and efficiency compared to manual methods, while also providing significant value in academic research.

TÓM TẮT

Chất lượng nước đóng một vai trò vô cùng quan trọng trong việc bảo vệ sức khỏe của con người và các loài khác trên hành tinh. Nguồn nước bị ô nhiễm có thể chứa các chất có hại như kim loại nặng, thuốc trừ sâu, vi khuẩn và vi rút, có thể gây nguy hiểm cho sức khỏe con người và các sinh vật khác. Chất lượng nước cũng là yếu tố quan trọng trong việc bảo vệ môi trường và duy trì hệ sinh thái nước ngọt và môi trường biển. Nguồn nước bị ô nhiễm có thể ảnh hưởng đến đời sống của động vật và thực vật thủy sinh, gây hại và gây nguy hiểm cho đa dạng sinh học do làm gián đoạn chuỗi thức ăn. Trong phạm vi bài viết này, chúng tôi sử dụng dữ liệu từ "Chất lượng nước" của MssmartyPants bằng

thuật toán Phân loại MLP cổ điển. Việc kết hợp trí tuệ nhân tạo vào đánh giá chất lượng nước góp phần tăng độ chính xác và hiệu quả so với phương pháp thủ công, đồng thời mang lại giá trị đáng kể trong nghiên cứu học thuật.

1. INTRODUCTION

The issue of clean water is a global concern and is regarded as one of the major challenges for sustainable development worldwide. According to reports from the World Health Organization (WHO) and UNICEF, approximately 2.2 billion people worldwide still lack access to clean water, and 4.2 billion people lack access to sanitation services (UNESCO, 2016) [1]. Vietnam is among the countries facing the challenge of clean water. Despite having significant water sources like the Red River, Mekong River, and the Gulf of Tonkin, water quality in many regions is encountering difficulties, affecting both human health and the environment. Water sources are polluted due to various factors including waste, livestock waste, chemical-intensive agriculture, industrial activities, and rampant littering by the population - as reported by the Ministry of Natural Resources and Environment, around 60-70% of lakes, river, (Ministry of Natural Resources and Environment, 2023).

UNESCO (2016) [1] discussed how water quality is one of the key challenges that society will face in the 21st century, posing threats to human health, limiting food production, diminishing the ecological functions of ecosystems, and impeding economic growth. In the article "Drinking-water - World Health Organization (WHO)," the author emphasizes the importance of clean and easily accessible water for public health and highlights how improving water supply and sanitation can drive economic growth for countries and significantly contribute to poverty reduction (WHO, 2022) [2]. Lastly, Roy et al. (2019) [3], highlighted how water can be one of the most valuable natural resources after air. In Vietnam, we have several articles addressing water quality issues. Huỳnh Phú et al., (2021) [4] focused on analyzing surface water quality based on the economic development of regions in Bac Lieu Province. Next, Vũ Thị Thanh Hương et al. (2020) [5] predicted water quality in the Bac Hung Hai irrigation system according to socioeconomic development scenarios up to the year 2020. Lastly, Vũ Thị Hồng Nghĩa et al. (2011) [6] presented evaluations of the water quality of the Cau River and proposed environmental management solutions to protect and improve water quality.

Lê Phước Cường et al. (2020) [7] utilized machine learning models such as Linear Regression, Random Forest, Support Vector Machine, K-nearest neighbor, and Cubist to predict groundwater quality near the Cẩm Hà landfill, Hoi An City. Rosly et al. (2015) [8] compared various classification methods such as Naive Bayes (NB), J48 Decision Tree, Sequential Minimal Optimization (SMO), Multi-Layer Perceptron (MLP), and Instance-Based Learning with k-Nearest Neighbors (IBK) for water quality classification of the Kinta River dataset in Perak, Malaysia. Results showed that multi-classification approaches could achieve higher accuracy than individual methods [8]. Nasir et al. (2022) [9] discussed

the use of seven individual classifiers to predict the Water Quality Index (WQI). The stacked model proved successful in predicting water quality, and the CATBOOST method yielded the best prediction results.

Finally, while there are no specific mentions of water quality prediction using the MLP Classifier algorithm, similar algorithms have been used to introduce innovative approaches. This algorithm is likely to yield the best results for water quality diagnosis. The dataset will be sourced from Kangle and the selected dataset named "Water Quality" by MssmartyPants.

2. MATERIALS AND METHODS

Most of the data used for research and learning purposes in this article will come from Kaggle, including the "Water Quality" dataset by Aditya Kadiwal. This dataset contains a fair amount of data errors (missing or lost data in certain rows/columns), which can significantly impact subsequent machine-learning efforts (Kadiwal, 2021) [10]. The "Indian water quality data" by Anbarivan and Anjali Vasudevan contained text-based information, making data normalization challenging, and it's not suitable for classification algorithms, thus not fitting the current purpose of the article of Ramakrishnan (2016) [11]. Lastly, the "Water Quality" dataset by MssmartyPants, comprising 21 columns and over 8000 rows of data, is relatively well-structured and has been extensively used for scientific research purposes, making it the chosen dataset (MsSmartyPants, 2021) [12].

Dataset:

All the listed datasets are related to water quality prediction; however, only one dataset meets the requirements of the research topic, as the others have various issues when used in this context. The dataset must be numerical, have specific classifications, and contain multiple fields to yield objective results. Only the Water Quality dataset from MssmartyPants meets these criteria, which is why it was selected for the internship project. However, this dataset is temporary in nature, as the indices can differ between databases, potentially leading to changes in the dataset's classification in the future. This highlights the fact that this dataset exists in an unstable environment, making it unsuitable for using classical machine learning algorithms in a static environment to increase dataset class sizes. Instead, advanced algorithms need to be employed to handle data in a dynamic environment. Raw data must be processed before being applied to the software's training process. The attributes used for testing, such as aluminum, ammonia, arsenic, barium, cadmium, chloramine, chromium, copper, fluoride, bacteria, viruses, lead, nitrates, nitrites, mercury, perchlorate, radium, selenium, silver, uranium, will be kept unchanged as they are numerical values, requiring no conversion. All conversion processes are entirely conducted within Microsoft Excel software.

After transforming the data from its raw form to a standardized format for software use, the dataset must be saved as a ".csv" file extension. Only one cell is allowed for each data entry, and parameters as well as labels must be separated by commas. According to system regulations, data parameters must come first, followed by the label. After all data standardization processes, a standardized dataset containing 21 features with a total of 8000 data points will be obtained. These indices are sourced from the aforementioned data and have been verified by experts in the field. Each feature will have the following indices:

Prediction label: This parameter carries a decisive value and is particularly important. The label can only take one of two values: "0" for unsafe or "1" for safe. The predictions consist of a total of 7084 unsafe data points and 916 safe data points.

Aluminum: It is found everywhere, in food, water, and cooking utensils made of aluminum. Scientists have found that aluminum has negative effects on health. High doses of aluminum exposure can damage bones and tissues. Bones lose calcium and phosphorus, becoming weaker and causing bone pain (ranges within the dataset from 0 to 5.05).

Ammonia: A colorless gas with a pungent odor, chemical formula NH_3. Ammonia is not highly toxic to humans and animals. However, if it exceeds permissible levels in water, it can transform into cancer-causing agents and other dangerous diseases (ranges within the dataset from 0 to 29.84).

Arsenic: Also known as arsenic, a highly toxic compound, four times more toxic than mercury. Ingesting water with even half a grain of rice's worth of arsenic can kill a healthy person. The World Health Organization (WHO) has reported that for every 10,000 cancer cases, 6 deaths are attributed to water with arsenic levels above the standard of 0.01 mg/l (ranges within the dataset from 0 to 1.05).

Barium: Barium is a solid substance that contributes to pollution in various wastewater treatment systems today. However, most people still have little understanding of the hidden potential dangers and optimal treatment methods for this pollutant in wastewater, tap water, and industrial water (ranges from 0 to 4.94 in the dataset).

Cadmium: Cadmium is one of the three most dangerous metals for the human body, the other two being lead and mercury. Regular consumption of water containing cadmium can increase the risk of diseases such as prostate cancer and lung cancer (ranges from 0 to 0.13 in the dataset).

Chloramine: Not only does chloramine have an unpleasant odor, but it also affects your health. The impact depends on the chlorine residue in the water. According to QCVN 01:2009/ BYT (National Technical Regulation on Drinking Water Quality), the permissible chlorine level in water is 0.3-0.5 mg/l. However, the actual situation currently shows a widespread excess of chlorine (above the standard) in tap water (ranges from 0 to 8.68 in the dataset).

Chromium: Chromium is a heavy metal that can cause cancer if accumulated in the body. According to the World Health Organization, chromium is toxic to the body, and drinking water containing chromium, even in amounts as low as 1-2g, can lead to immediate death (ranges from 0 to 0.9 in the dataset).

Copper: Copper is a fairly common metal found in water. To ensure safety for users, the copper content in water must be less than 2mg/l. The harmful effects of this heavy metal in water include irritation and corrosion of mucous membranes, nerve inhibition, etc (ranges from 0 to 2 in the dataset).

Fluoride: Accumulation of excessive fluoride in the body can affect joints, leading to increased risks of joint pain, immobility, weakened bones, and even bone cancer. Additionally, excessive fluoride in water can increase the risk of thyroid gland diseases (ranges from 0 to 1.5 in the dataset).

Bacteria: The presence of hidden microorganisms in water indicates that the water source is not safe, causing various dangerous symptoms such as diarrhea, vomiting, high fever, etc (ranges from 0 to 1 in the dataset).

Viruses: Similar to bacteria, the presence of viruses in water is extremely dangerous (ranges from 0 to 1 in the dataset).

Lead: The presence of lead in water is due to the corrosion of pipes and industrial wastewater. According to current regulations on clean water and drinking water, the lead content in water must not exceed 0.01 mg/l (ranges from 0 to 0.2 in the dataset).

Nitrates: If water with nitrate is heated, it can form nitrosamines. There are various types of nitrosamines, some of which can increase the risk of cancer (ranges from 0 to 19.83 in the dataset).

Nitrites: Similar to nitrates, water with nitrites heated at high temperatures is extremely dangerous (ranges from 0 to 2.93 in the dataset).

Mercury: Mercury is a metallic element found naturally in air, water, and soil. Even slight exposure to mercury can cause severe health problems, threatening the development of fetuses and the early stages of children's lives. Apart from young children, mercury poisoning can harm the nervous, digestive, and immune systems, affecting the lungs, kidneys, skin, and eyes (ranges from 0 to 0.01 in the dataset).

Perchlorate: Formed from one chlorine atom and four oxygen atoms, perchlorate is a powerful oxidant widely used in rocket fuel, fireworks, and flares. They are also produced as byproducts from chemical manufacturing and herbicides, and therefore, after use, they can contaminate the environment, and seep into water sources, and soil (ranges from 0 to 60.01 in the dataset).

Radium: Health issues related to radium stem from its radioactive properties. Radium primarily appears in groundwater in the form of radioactive isotopes radium-226 and radium-228. As these isotopes decay, they emit alpha particles that can damage human tissue. Alpha radiation is blocked by the skin, so there is no danger when bathing or washing dishes with water containing radium. However, long-term consumption of water with radium can lead to chronic health problems, including an increased risk of bone cancer (ranges from 0 to 7.99 in the dataset).

Selenium: While selenium is recognized as an essential element for human health, it is only needed in very small amounts (the human body requires very little selenium). Prolonged exposure above the permissible limit can lead to selenium poisoning, with symptoms such as depression, anxiety, nervousness, mood swings, nausea, vomiting, garlic-scented breath and sweat, and in some cases, hair loss and brittle nails (ranges from 0 to 0.1 in the dataset).

Silver: When nano silver particles come into contact with the human body, they attack the immune system, destroy cell structures, and gradually weaken health. This prolonged condition can increase the risk of devastating diseases such as Alzheimer's, Parkinson's, and even cancer (ranges from 0 to 0.5 in the dataset).

Uranium: In nature, uranium (VI) forms highly soluble carbonate complexes in alkaline environments. This enhances the mobility and persistence of uranium in soil and groundwater, originating from nuclear waste materials, posing health risks to humans (ranges from 0 to 0.09 in the dataset).

Overall, this database has been relatively well standardized, except the author not providing additional information about certain data fields, requiring further investigation. The experimental dataset consists of two parts: a training dataset containing 5600 data points (70% of the original data) and a test dataset containing 2400 data points (30% of the original data). The positions of these data points will change in each experiment, and each experiment will involve random shuffling both during training and afterward.

The experiment will be conducted using a batch learning model with a batch size of 699. This means the system will perform 699 steps, with each step containing about 129 data points. The model used in the experiment is a batch data model. This model will use the same batch data set by dividing the original data set into smaller batches in 35 steps, meaning each batch contains 129 data points. This number is neither too large nor too small, making it convenient for experimentation.

3. RESULTS AND DISCUSSION

The sliding window approach is a technique used in machine learning to train models by utilizing small data windows. These windows are created by sequentially moving through the dataset, generating a new window at each step. The size of the window can be fixed or variable, and the overlap between consecutive windows can also be controlled. This approach is computationally efficient and can be employed for performing machine learning tasks on data streams. This training model method has several advantages compared to other methods, such as batch learning. Firstly, it enables faster training times as it processes only a small portion of data at each step. Secondly, it can help avoid overload as the model continuously interacts with new data points. Lastly, it is memory-efficient, requiring only a small portion of data to be loaded into memory at a time. Therefore, employing this approach significantly enhances the effectiveness and productivity of ML algorithms (Joseph, 2022) [13].

The model used for conducting scientific experiments has been explicitly discussed in the previous section. Therefore, in this section, the focus is on analyzing and comparing results among various algorithms.

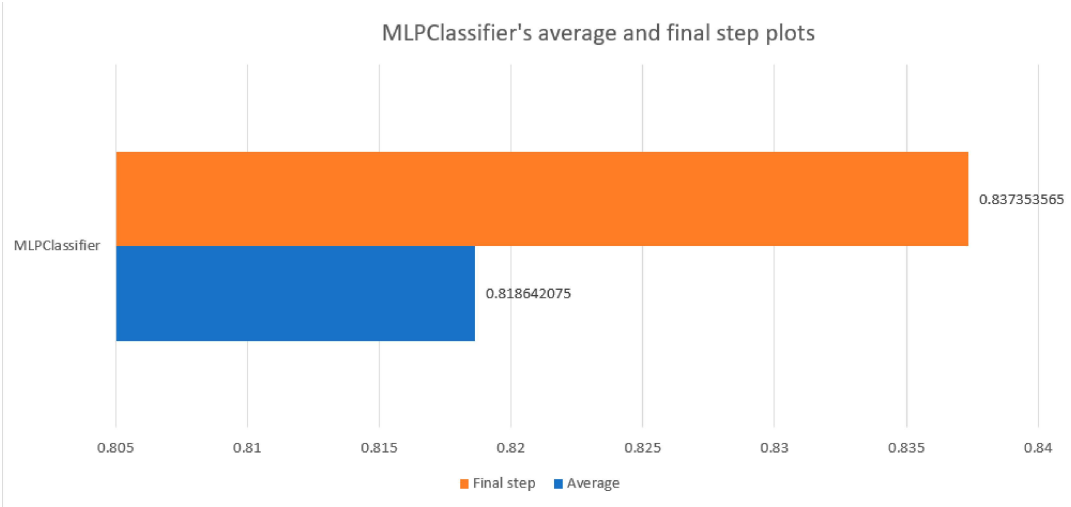


Figure 1. Average of BA and in final step

The MLP Classifier is a machine learning algorithm used for classification tasks. The average curve and final accuracy are typically important metrics for evaluating the performance of a classification model. In your specific case, the average curve accuracy is 0.818, and the final accuracy is 0.837. Accuracy is computed by comparing the number of correct predictions to the total number of samples in the test dataset. The average curve is a graph illustrating how the model's accuracy changes as different parameters or conditions are altered. Through the average curve, you can identify certain parameter thresholds that lead to improved model performance, aiding in optimizing its effectiveness. The final accuracy, on the other hand, represents the ultimate accuracy of the model after optimization and training on the training data. This is the accuracy the model can achieve when used in real-world applications. The difference between the average curve and the final accuracy reflects the model's optimism or realism. If the average curve exhibits higher accuracy than the final accuracy, it could suggest that further fine-tuning of the model could enhance its performance. Conversely, if the final accuracy closely resembles the average curve, this indicates that the model has been well-optimized and can be directly applied to new data. In summary, the average curve and final accuracy of the MLP Classifier provide important information about the classification model's performance on test data. Analyzing and evaluating these metrics helps you gain a better understanding of the model's capabilities and limitations in classifying different samples.

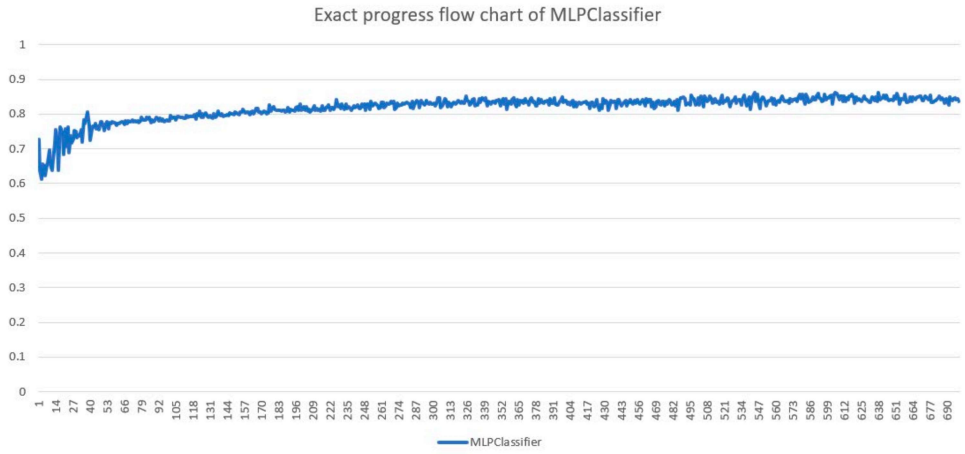


Figure 2. Progress flow chart of MLP Classifier

Increasing Variation Trend: The accuracy progression plot of the MLP Classifier illustrates that the model's accuracy increases over time or the number of iterations. This indicates that the model is learning from the data and improving its accuracy as time goes on. The gradual increase in accuracy suggests that the model is not just initially randomly learning from the data but also capturing more important features as training progresses further. This could imply that the model is enhancing its performance by grasping the data's complexity better. The fluctuation within the range of 0.814 to 0.855 indicates that the model's accuracy is

not stable and varies across iterations. However, this variation falls within a narrow range, which may suggest that the model is nearing the threshold of maximum performance for the current training dataset.

In summary, the accuracy progression plot of the MLP Classifier depicts a model evolving from an initial accuracy of 0.814 to a higher level of 0.855 overtime or iterations, despite minor fluctuations during this process+.

Installation:

MLP Classifier will be the chosen algorithm for the web environment. It includes three forms: the login form, the diagnosis form, and the settings form. To use the software after successful installation, you need to access the portal "http://127.0.0.1:8000" to enter the main page of the system. Here, the login interface will be displayed. In the login interface, there are two ways to log into the system: one is to log in as a developer and the other is to log in as a user.

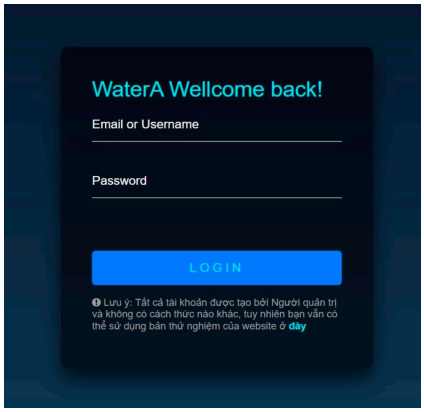


Figure 3. Login screen

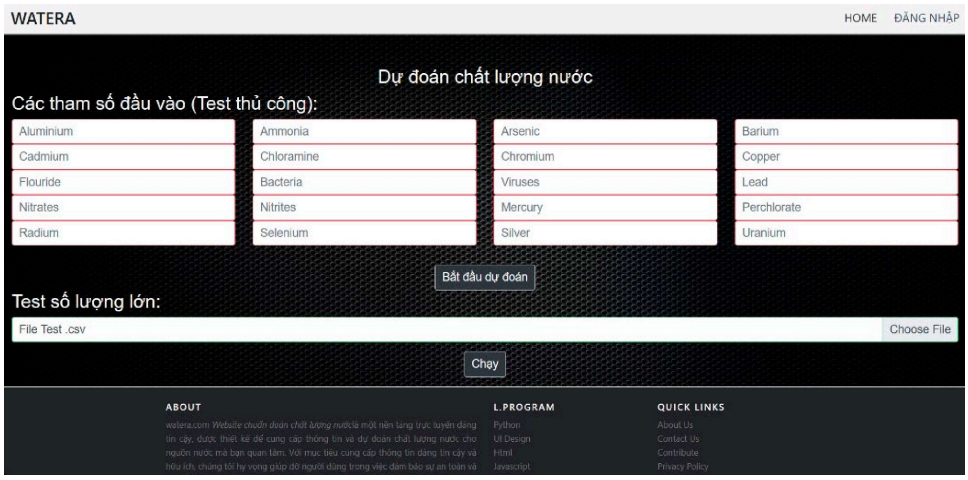


Figure 4. Main interface

As the project is still under development, user account allocation remains limited. Therefore, apart from using the provided accounts, users can also access the website directly to experience its main functionalities. There are two methods for data processing: manual data entry or processing a large amount of data using a ".csv" file. Here is the main interface:

Interface for the list of trained models:

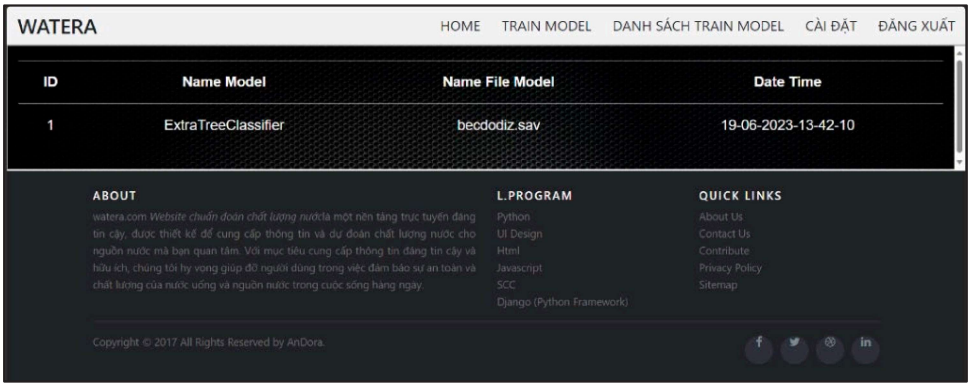


Figure 5. Model installation interface

Configuration settings interface:

System Requirements: Operating System: Windows 10, RAM: Minimum 2GB, Hard Drive Space: Minimum 10GB, and Internet Connection.

Installation process:

Step 1: Install Python and Libraries: Extract the files and open the "SETUP" folder, run the "python-3.9.9-amd64.exe" file to install Python 3.9.9, after Python installation is complete, run the "inLib.bat" file to install the necessary libraries.

Step 2: Remove Excess Data (Optional):

If needed, you can run the "Remove.bat" file to delete unnecessary data files. Only perform this step in the mentioned cases.

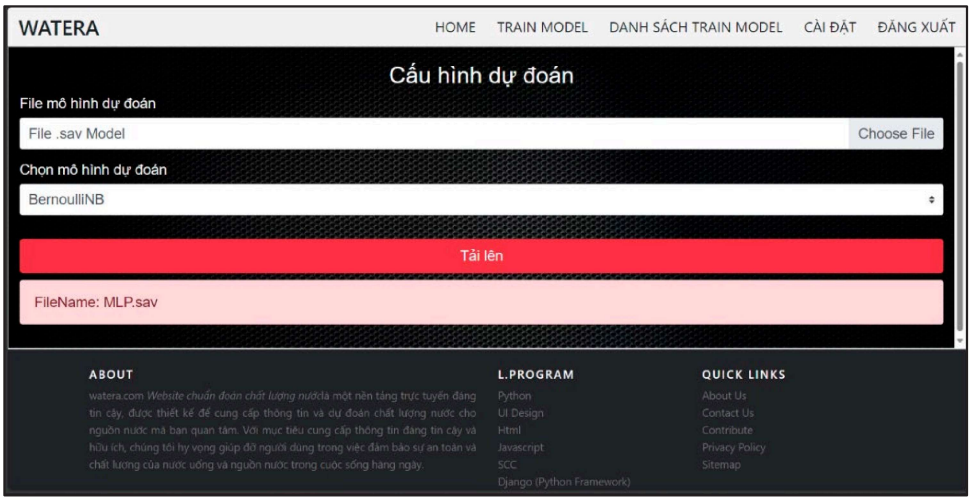


Figure 6. Configuration settings interface

Step 3: Run the Program: To run the program, use the "Runserver.bat" file. This file is configured to execute the command "py manage.py runserver". The program will run on the default port "http://127.0.0.1:8000/".

Using the Software: Ensure your computer is always connected to the internet. Access the address "http://127.0.0.1:8000" to reach the main page of the system. On the main interface page, you will see a template of input fields to predict water quality.

Note that I have used the information you provided to create a general guide. If you encounter any issues or have specific questions about the installation, usage, or operation of the software, you should refer to the documentation provided by the developer or the software's support team.

4. CONCLUSION AND SUGGESTIONS

In conclusion, we have explored three classical algorithms used in the project, delved into the specifics of the "Water Quality" database by MssmartyPants, and explained the need for a specific development direction to effectively utilize the application. This research enables us to comprehend the significance of water quality and the integration of artificial intelligence into diagnostic processes. Water quality remains a pressing concern in our nation's developmental journey, as the lack of clean water sources can impact socioeconomic factors, facilitate disease transmission, and affect human health. Ultimately, prevention is better than cure, and together, we must foster a collective awareness to safeguard clean water sources and mitigate actions that contribute to environmental pollution, especially in water sources.

REFERENCES

[1] UNESCO. (2016). *The global water quality challenge & SDGs.* https://en.unesco.org/waterquality-iiwq/wq-challenge

[2] WHO. (2022). *Drinking-water*. https://www.who.int/news-room/fact-sheets/detail/drinking-water. Retrieved May 13, 2023.

[3] Roy, R. (2019). *An Introduction to Water Quality Analysis.* https://www.researchgate.net/publication/352907194

[4] Huỳnh Phú (2021). *Nghiên cứu phân vùng chất lượng nước mặt theo diễn biến phát triển các vùng kinh tế của tỉnh Bạc Liêu.* https://www.researchgate.net/publication/352250212_Nghien_cuu_phan_vung_chat_luong_nuoc_mat_theo_dien_bien_phat_trien_cac_vung_kinh_te_cua_tinh_Bac_Lieu

[5] Vũ Thị Thanh Hương, Nguyễn Đức Phong, & Nguyễn Xuân Khôi (2020). Nghiên cứu dự báo chất lượng nước trong hệ thống thủy lợi bắc Hưng Hải theo các kịch bản phát triển kinh tế xã hội đến năm 2020. *Tạp Chí Khoa Học và Công Nghệ Thủy Lợi*

[6] Vũ Thị Hồng Nghĩa (2011). *Đánh giá hiện trạng chất lượng nước sông Cầu và đề xuất giải pháp quản lý môi trường nước sông Cầu trên địa bàn Tỉnh Thái Nguyên.* Nghiên cứu quản lý chất lượng nước Sông Cầu trên địa bàn Tỉnh Thái Nguyên (vnu.edu.vn)

[7] Lê, P., & Cường (2020). *Ứng dụng mô hình học máy dự báo chất lượng nước dưới đất: Điển hình tại khu vực thành phố Hội An, Tỉnh Quảng Nam (Application of Machine Learning Models in underground water prediction: A case study in Hoian City, Quangnam Province.* https://media.neliti.com/media/publications/453597-application-of-machine-learning-models-i-ddd67fa8.pdf

[8] Rosly, R., Makhtar, M., Awang, M.K., & Deris, M.M. (2015). *Multi-Classifier models to improve the accuracy of water quality application.*

https://www.researchgate.net/publication/331208114

[9] Nasir, N., Kansal, A., Alshaltone, O., Barneih, F., Sameer, M., Shanableh, A., & Al-Shamma'a, A. (2022). Water quality classification using machine learning algorithms. *Journal of Water Process Engineering*, *48*, 102920. https://doi.org/10.1016/j.jwpe.2022.102920

[10] Kadiwal, A. (2021). *Water Quality*. https://www.kaggle.com/datasets/adityakadiwal/water-potability, Retrieved May 13, 2023.

[11] Ramakrishnan, V. (2016). *India water quality data.* https://www.kaggle.com/datasets/venkatramakrishnan/india-water-quality-data. Retrieved May 13, 2023.

[12] MsSmartyPants. (2021). *Water quality*. https://www.kaggle.com/datasets/mssmartypants/water-quality. Retrieved May 13, 2023.

[13] Joseph (2022). *Sliding window machine learning: What you need to know - reason*. https://reason.town/sliding-window-machine-learning/. Retrieved June 3, 2023.