F.Y.B.Tech Students' Engineering Design and Innovation (EDAI 1) Project Paper, SEM 1 A.Y. 2023-24
Vishwakarma Institute of Technology, Pune, INDIA

A Machine Learning Approach for Water Potability Prediction and Filtration Cost Estimation

Ayush Gupta, Ayush Bora, Avishkar Sonpipare, Ayush Kulkarni, Avanti Patil, Sanket Auti

Department of Engineering, Sciences and Humanities (DESH)
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India

***Abstract* —** Ensuring the potability of drinking water is crucial for public health, as contaminated water can lead to severe health issues. This study leverages machine learning techniques to predict water potability and estimate filtration costs using various water quality parameters such as pH, Hardness, Sulfate, Conductivity, and others. We start by preprocessing the dataset, addressing missing values through imputation methods. Missing values in 'pH' and 'Trihalomethanes' are imputed using median values, while 'Sulfate' is imputed using a linear regression model. Exploratory Data Analysis (EDA) is conducted to visualize the distribution and relationship of each feature with the target variable, 'Potability.

A linear regression model is then developed to predict water potability, trained on 80% of the data and evaluated on the remaining 20%. The model's performance is assessed using mean squared error (MSE) as a metric. Additionally, we develop models to estimate the cost of ultrafiltration, a common water filtration method, based on selected water quality features. These models are trained and validated similarly, and their predictions are visualized to assess accuracy.

***Keywords* —** Water Potability, Machine Learning, Linear Regression, Data Preprocessing, Filtration Cost, Water Quality Parameters

I. INTRODUCTION

The availability of clean and safe drinking water is essential for public health and well-being. Contaminated water sources can lead to a variety of health issues, including gastrointestinal diseases, neurological disorders, and reproductive problems. As urbanization and industrial activities increase, so does the risk of water pollution, making the assessment and management of water quality more critical than ever. Traditional methods of water quality analysis are often labor-intensive and time-consuming, requiring sophisticated equipment and expert personnel. Consequently, there is a growing interest in leveraging advanced data analytics and machine learning techniques to streamline the process of water quality assessment and enhance predictive capabilities.

Machine learning has emerged as a powerful tool for analyzing complex datasets and uncovering hidden patterns that are not easily discernible through conventional methods. By applying machine learning algorithms to water quality data, we can develop models that predict water potability based on various physicochemical parameters. This approach not only accelerates the assessment process but also provides a higher level of accuracy and reliability. Moreover, machine learning models can be continually updated and refined with new data, ensuring that the predictions remain relevant and accurate over time.

In this study, we aim to develop a robust machine learning framework for predicting water potability and estimating the costs associated with water filtration processes. The dataset used in this study includes key water quality parameters such as pH, Hardness, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity. These parameters are crucial indicators of water quality, influencing both the aesthetic and health-related aspects of drinking water.

Our methodology involves several key steps, starting with data preprocessing to handle missing values and prepare the dataset for analysis. We employ imputation techniques to fill in missing values for parameters such as pH and Trihalomethanes using median values, and for Sulfate using a linear regression model. Following data preprocessing, we conduct an Exploratory Data Analysis (EDA) to visualize the distribution of each parameter and its relationship with the target variable, 'Potability.'

The core of our study involves developing linear regression models to predict water potability and estimate the costs associated with ultrafiltration, a commonly used water filtration method. We train the potability prediction model on a substantial portion of the dataset and evaluate its performance on a separate test set, using mean squared error (MSE) as a metric to quantify prediction accuracy. Similarly, we develop and validate models for estimating the cost of ultrafiltration based on selected water quality features.

The significance of this research lies in its potential to provide actionable insights for water quality management. By identifying key factors that influence water potability, stakeholders can prioritize efforts to monitor and improve these parameters. Additionally, accurate cost estimation models can help in budgeting and optimizing the implementation of filtration systems, ensuring that safe drinking water is accessible while minimizing costs.

Through this research, we aim to demonstrate the efficacy of machine learning in enhancing water quality management practices, providing a foundation for the development of intelligent systems capable of ensuring the safety and potability of drinking water.

II. Related Work

Previous studies have explored the application of machine learning in water quality assessment. [5] Techniques such as decision trees, support vector machines, and neural networks have been utilized to predict water potability. However, there is a need for more robust models that can handle missing data and provide cost estimations for filtration processes. Our work builds on these studies by incorporating data imputation techniques and focusing on cost estimation for specific filtration methods.

III. Methodology

The methodology section details the steps taken to preprocess the data, develop the models, and evaluate their performance.

In order to provide a clear and structured overview of the research methodology, the following flow chart (Figure 1) illustrates each step of the process.

A. Data Preprocessing

1. Dataset Import and Initial Inspection: The dataset, containing various water quality parameters, was imported and inspected for completeness.

```
data = pd.read_csv('water_potability.csv')
data.head()
```

2. Handling Missing Values: Missing values in the 'pH' and 'Trihalomethanes' columns were imputed with their median values, while missing values in 'Sulfate' were handled using linear regression.

```
data.ph.fillna(data.ph.median(),inplace=True)
data.Trihalomethanes.fillna(data.Trihalomethanes.median(),inplace=True)
```

3. Exploratory Data Analysis (EDA): EDA was conducted to understand the distribution of each variable and its relationship with the target variable 'Potability'.

```
def conti_var(x):
    fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(16,5), tight_layout=True)
    axes[0].set_title('Distribution')
    sns.histplot(x, ax=axes[0])
```

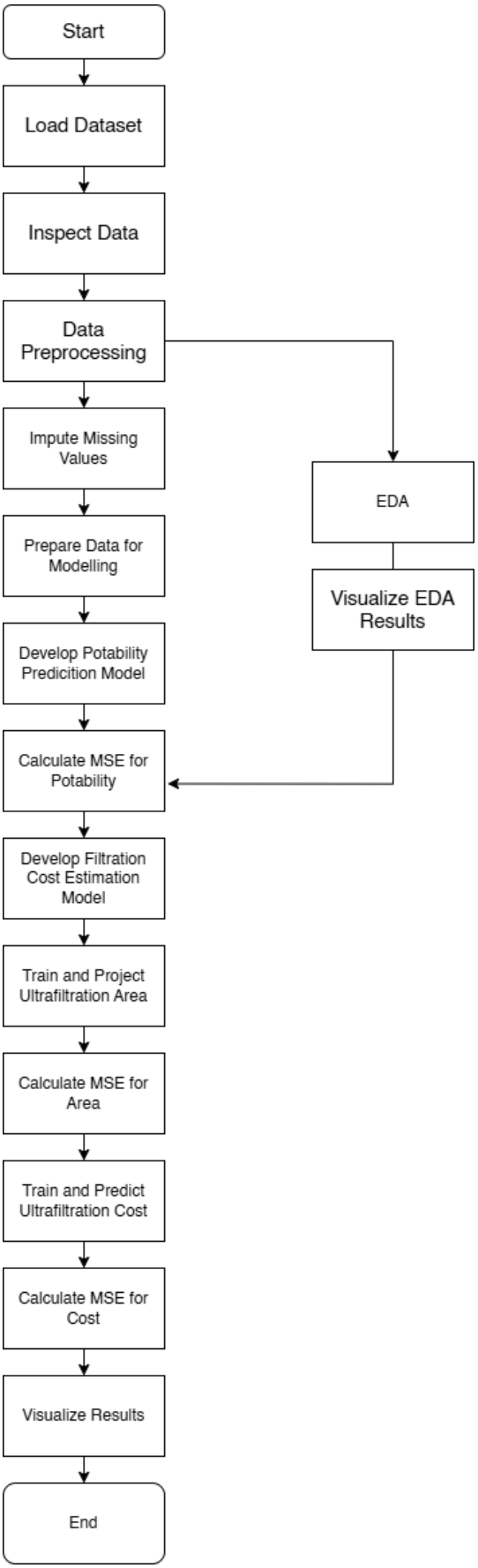

Figure 1

```
axes[0].grid()
            axes[1].set_title('Outliers')
            sns.boxplot(x, ax=axes[1])
            axes[2].set_title('Relation with
  Potability')
            sns.boxplot(x=data.Potability, y=x,
ax=axes[2])
            axes[2].grid()
```

B. Model Development

1. Potability Prediction Model: A linear regression model was developed to predict water potability.

```
    X = data.drop('Potability', axis=1)
    y = data.Potability
    X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.2,
random_state=42)
    model = LinearRegression()
    model.fit(X_train, y_train)
    y_pred = model.predict(X_test)
    mse = mean_squared_error(y_test, y_pred)
```

2. Filtration Cost Estimation Model: Separate models were developed to estimate the cost of ultrafiltration based on specific water quality parameters.

```
    cost_model_ultrafiltration =
LinearRegression()
  cost_model_ultrafiltration.fit(X_train_ultrafiltrati
on, y_cost_train_ultrafiltration)
    y_cost_pred_ultrafiltration =
cost_model_ultrafiltration.predict(X_test_ultrafiltra
tion)
  mse_cost_ultrafiltration =
mean_squared_error(y_cost_test_ultrafiltration,
y_cost_pred_ultrafiltration)
```

IV. Results

The results section presents the findings from the model evaluations.

A. Potability Prediction

The linear regression model for potability prediction yielded a mean squared error (MSE) indicating the

average squared difference between the predicted and actual values. The performance metrics demonstrate the model's capability in predicting water potability based on the given parameters.

The figure 2 depicts the receiver operating characteristic curve of the model.

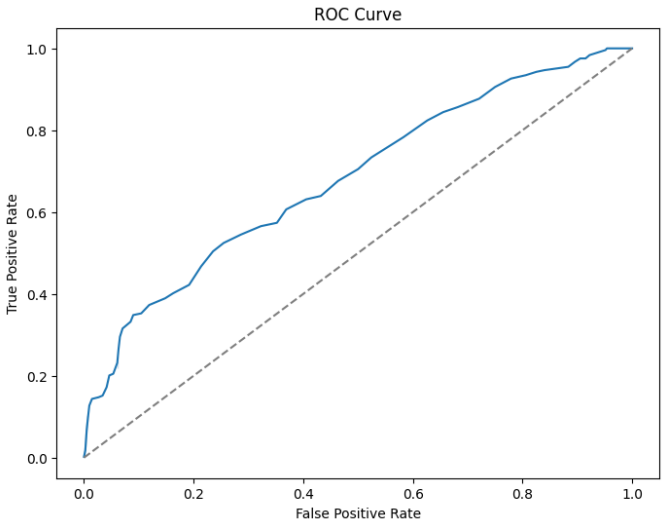

Figure 2

B. Filtration Cost Estimation

The cost estimation models provided predictions for carbon filter costs with reasonable accuracy. Scatter plots were used to visualize the actual versus predicted values for both filter area and cost.

```
plt.figure(figsize=(12, 6))
sns.scatterplot(x=y_cost_test_carbon,
y=y_cost_pred_carbon, color='blue', alpha=0.5)
plt.plot([min(y_cost_test_carbon),
max(y_cost_test_carbon)],
[min(y_cost_test_carbon),
max(y_cost_test_carbon)], linestyle='--',
color='red')
plt.title('Actual vs Predicted Carbon Filter Cost')
```

The figure 3 shows the scatter plot for carbon filter cost along with the carbon filter radius.

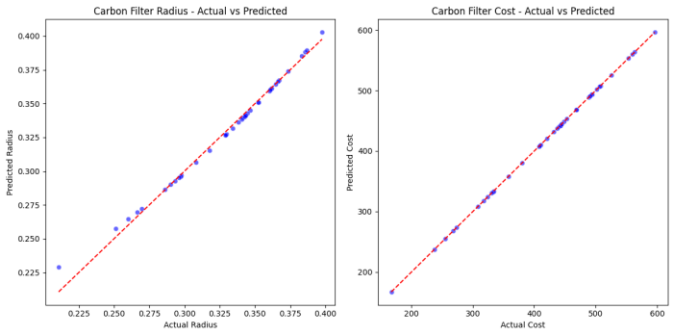

Figure 3

V. Conclusion

In this study, we have demonstrated the potential of machine learning techniques to significantly enhance the assessment and management of water quality. By leveraging a dataset containing various physicochemical parameters of water, we developed predictive models for both water potability and the cost estimation of ultrafiltration filtration processes. Our comprehensive approach involved meticulous data preprocessing to handle missing values, thorough exploratory data analysis to understand the underlying patterns and relationships, and the application of linear regression models to make accurate predictions.

The results of our study underscore the effectiveness of machine learning in identifying key factors that influence water potability. The potability prediction model, evaluated using mean squared error (MSE), provided a reliable measure of the average squared difference between predicted and actual potability values. This model can serve as a valuable tool for water quality management, enabling authorities and stakeholders to prioritize monitoring and remediation efforts based on the most influential water quality parameters.

Furthermore, the cost estimation models for ultrafiltration filtration processes demonstrated the feasibility of predicting filtration costs based on selected water quality features. Accurate cost predictions are crucial for planning and budgeting, allowing for more efficient allocation of resources and ensuring that safe drinking water is provided in a cost-effective manner. The visualization of actual

versus predicted values using scatter plots further validated the accuracy of our models, showcasing their practical applicability.

Our research contributes to the growing body of knowledge on the use of machine learning for environmental monitoring and public health protection. By addressing the limitations of traditional water quality assessment methods, which are often time-consuming and require significant expertise, our machine learning approach offers a scalable and efficient alternative. The models developed in this study can be continuously updated with new data, enhancing their predictive power and relevance over time.

Despite the promising results, our study also acknowledges several limitations. The models were developed and validated using a specific dataset, which may not capture the full variability of water quality parameters across different regions and sources. Therefore, future research should focus on incorporating diverse datasets from various geographic locations to improve the generalizability of the models. Additionally, while linear regression provided satisfactory results in this study, exploring more advanced machine learning algorithms such as decision trees, random forests, or neural networks could potentially yield even more accurate predictions.

Future work should also consider integrating real-time data from IoT-enabled water quality sensors, which can provide continuous monitoring and immediate detection of contamination events. This integration would enable the development of dynamic predictive models that can adapt to changing water quality conditions in real-time, offering an even higher level of precision and responsiveness.

In conclusion, this study has demonstrated the significant benefits of applying machine learning to water quality prediction and filtration cost estimation. The insights gained from our models can inform better decision-making and resource allocation, ultimately contributing to the provision of safe and potable drinking water. As the world continues to face challenges related to water scarcity and pollution, innovative solutions such as the ones presented in this research will be essential for ensuring public health and environmental sustainability.

VI. Future Scope

The scope of this project includes the following aspects:

1. Dataset Selection and Analysis: Utilizing a dataset with comprehensive water quality parameters to train and validate the models.
2. Data Preprocessing: Handling missing values and performing exploratory data analysis to understand the data characteristics.
3. Model Development: Creating linear regression models for potability prediction and filtration cost estimation.
4. Evaluation and Validation: Assessing the models' performance using appropriate metrics and visualizations.
5. Practical Implications: Providing insights into water quality management and cost-effective filtration methods for real-world applications.
6. Future Work: Extending the models to handle diverse datasets and exploring more advanced machine learning techniques for improved accuracy.

VI. Acknowledgment

I would like to sincerely thank everyone who helped to finish this research paper. Special thanks to our guide Prof. Prajkta Pramod Dandavate for invaluable guidance and unwavering support throughout the research process. I am grateful to the research participants for their time and cooperation, as their contributions were integral to the study. I also want to express my gratitude to friends and coworkers who helped out by offering insightful criticism. Without the resources and conducive environment provided by my institution Vishwakarma Institute of Technology, this work would not have been possible.

VII. REFERENCES

[1] Su, X., Yan, X., & Tsai, C. (2012). Linear regression. Wiley Interdisciplinary Reviews. Computational Statistics, 4(3), 275–294. https://doi.org/10.1002/wics.1198

[2] Salehi, M. (2022). Global water shortage and potable water safety; Today's concern and tomorrow's crisis. Environment International, 158, 106936.
https://doi.org/10.1016/j.envint.2021.106936

[3] Chong, M. N., Jin, B., Chow, C. W., & Saint, C. (2010). Recent developments in photocatalytic water treatment technology: A review. Water Research, 44(10),2997–3027.
https://doi.org/10.1016/j.watres.2010.02.039

[4] Falkenmark M, Widstrand C. Population and water resources: a delicate balance. Popul Bull. 1992 Nov;47(3):1-36. PMID: 12344702.

[5] M. I. Khoirul Haq, F. Dwi Ramadhan, F. Az-Zahra, L. Kurniawati and A. Helen, "Classification of Water Potability Using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence and Big Data Analytics, Bandung, Indonesia, 2021, pp. 1-5, doi: 10.1109/ICAIBDA53487.2021.9689727.

[6] Rashid, Md. Harun Ar & Ahmed, Razib. (2023). Purification of Surface Water by Sand Filtration with Activated Carbon.

[7] Abdiyev, Kaldibek & Azat, S. & Kuldeyev, Erzhan & Kabdrakhmanova, Sana & Berndtsson, R. & Khalkhabai, Bostandyk & Kabdrakhmanova, Ainur & Sultakhan, Shynggyskhan. (2023). Review of Slow Sand Filtration for Raw Water Treatment with Potential Application in Less-Developed Countries. 10.20944/preprints202304.0964.v1.

[8] Roy, Ritabrata. (2019). An Introduction to Water Quality Analysis. 6. 201-205. 10.31786/09756272.18.9.2.214.

[9] Batra, Shivani & Adhikari, Priyanka & Ghai, Anchit & Sharma, Aman & Sarma, Rhea & V, Suneetha. (2017). Study and design of portable antimicrobial water filter. Asian Journal of Pharmaceutical and Clinical Research. 10. 268. 10.22159/ajpcr.2017.v10i9.19925.