**Revised Version:**

**Title:** Assessing surface water quality in Bristol, UK, using machine learning techniques in environmental analytics for sustainability and environmental impact assessment.

**Introduction:**

The quality of surface water in urban environments is a critical concern for public health, ecosystem integrity, and sustainable urban development (Vörösmarty et al., 2020). In Bristol, UK, like many growing cities, the management of water resources faces increasing challenges due to urbanization, climate change, and evolving industrial practices (Bristol City Council, 2020). This research proposal aims to address these challenges by leveraging machine learning techniques, specifically Multiple Linear Regression, to assess and predict surface water quality in Bristol using the 2023 dataset from Bristol Open Data. By analyzing this data, the study seeks to provide a comprehensive understanding of the factors influencing water quality and to develop a predictive model that can inform environmental impact assessments and guide sustainable water management practices (Rahmani et al., 2021).

The primary aim of this research is to assess and predict surface water quality in Bristol using machine learning techniques to support environmental impact assessment and sustainable water management. To achieve this aim, the study has several key objectives:

Objectives:

1. Analyze the current state of surface water quality in Bristol using the 2023 dataset from Bristol Open Data.
2. Identify and quantify the key factors influencing surface water quality through statistical analysis and machine learning techniques.
3. Develop and validate a Multiple Linear Regression model to predict specific water quality parameters based on other measured variables within the dataset.
4. Assess the environmental impact of current water quality levels by comparing them with established standards set by the UK Environment Agency.
5. Generate data-driven insights and recommendations for improving water management practices and reducing pollution in Bristol's surface waters based on the 2023 dataset analysis.

By meeting these objectives, this study aspires to contribute valuable insights to the field of urban water management and demonstrate the potential of data analytics in addressing environmental challenges.

**Problem Statement:**

Bristol faces significant challenges in maintaining and improving the quality of its surface water resources. The complex interplay of urban development, industrial activities, and changing climate patterns impacts water quality, potentially affecting public health, ecosystem integrity, and the city's overall environmental sustainability. There is a need for data-driven approaches to understand, predict, and manage surface water quality in this urban environment using publicly available data. This study aims to address this need by applying machine learning techniques to analyze and predict water quality using the 2023 surface water quality dataset from Bristol Open Data.

**Research Scope and Context:**

This research is situated within the broader context of urban water management and environmental data science, with a specific focus on the city of Bristol, UK. As cities grapple with sustainable development challenges, there is an increasing recognition of the need for evidence-based approaches to environmental management (Goonetilleke et al., 2019). The application of machine learning techniques to environmental data offers promising avenues for gaining deeper insights into complex ecological systems and informing policy decisions (Rahmani et al., 2021).

The scope of this study is specifically focused on:

1. Data Source: The research exclusively utilizes the 2023 surface water quality dataset from the Bristol Open Data portal. This focus on a single, recent dataset allows for an in-depth analysis of current water quality conditions in Bristol.
2. Geographical Area: The study is confined to the city of Bristol, UK, providing a detailed examination of water quality issues in this specific urban environment.
3. Temporal Focus: The analysis is limited to the year 2023, offering a snapshot of current water quality conditions rather than a historical trend analysis.
4. Water Quality Parameters: The study will concentrate on key water quality indicators available in the dataset, such as E. coli levels, dissolved oxygen, pH, temperature, and specific chemical pollutants.
5. Analytical Approach: The research will employ statistical analysis and machine learning techniques, specifically Multiple Linear Regression, to analyze the relationships between various water quality parameters.
6. Environmental Impact: The study will assess the current state of water quality in Bristol by comparing observed levels with established environmental standards set by the UK Environment Agency.
7. Predictive Modeling: A Multiple Linear Regression model will be developed to predict specific water quality parameters based on other measured variables within the dataset.
8. Management Insights: Based on the analysis, the study aims to generate data-driven insights and recommendations for improving water management practices in Bristol.

By confining the study to this specific scope, the research aims to demonstrate the potential of open data and machine learning in urban water management, even with limited temporal data. The findings and methodologies developed could serve as a foundation for similar analyses in other urban areas, contributing to the broader field of environmental data science and sustainable urban water management.

**Literature Review**

The study of urban water quality has gained significant attention in recent years, with researchers exploring various approaches to assess, predict, and manage water resources in urban environments. This research builds upon existing literature in the fields of water quality assessment, machine learning applications in environmental science, and urban water management.

Water Quality Assessment in Urban Environments: Urban areas face unique challenges in maintaining water quality due to high population density, industrial activities, and altered hydrological cycles. Grimm et al. (2008) provided a comprehensive overview of urban water quality issues, highlighting the complex interactions between human activities and aquatic ecosystems. More recently, Goonetilleke et al. (2019) examined the potential of stormwater reuse in urban areas, emphasizing the need for innovative approaches to water management in cities.

Machine Learning in Environmental Science: The application of machine learning techniques in environmental science has shown promising results in recent years. Sit et al. (2020) reviewed the progress and future directions of machine learning in water sciences, noting its potential to improve water quality prediction and management. Specifically relevant to this study, Rahmani et al. (2021) demonstrated the exceptional performance of deep learning models in predicting stream temperatures, highlighting the value of integrating diverse data sources in environmental modeling.

Water Quality Studies in Bristol: While broader studies provide context, local research is crucial for understanding Bristol's specific water quality challenges. Langston et al. (2010) examined water quality trends in the Severn Estuary, which is directly relevant to Bristol's water systems. Additionally, annual water quality reports from Bristol Water offer valuable insights into local water management practices and challenges (Bristol Water, 2022).

Machine Learning in Water Quality Prediction: The use of machine learning for water quality prediction has been explored in various contexts. Zhang et al. (2021) provided a comprehensive review of data-driven water quality model development, discussing the advantages and limitations of different machine learning approaches. For this study, the Multiple Linear Regression approach is informed by the work of Pedregosa et al. (2011), who developed the scikit-learn library widely used for implementing machine learning models in Python.

**Data Sources**

This study exclusively utilizes the 2023 Surface Water Quality dataset available from the Bristol Open Data portal (<https://opendata.bristol.gov.uk/>). This comprehensive dataset, provided by Bristol City Council as part of their open governance initiative, covers various water bodies within Bristol's city limits throughout the year 2023. It includes a range of water quality indicators, encompassing physical parameters (such as temperature), chemical parameters (including pH and dissolved oxygen), and biological parameters (like E. coli counts), along with other relevant environmental variables.

The dataset is freely accessible to the public, ensuring transparency and allowing for reproducibility of our research. While this data source provides a detailed snapshot of Bristol's water quality in 2023, it's important to note its limitations. The dataset lacks historical context for trend analysis, and its quality is dependent on the collection and reporting practices of Bristol City Council. Despite these constraints, focusing on this single, rich dataset allows us to demonstrate the potential of open data resources in environmental research and urban water management, particularly in the context of a specific urban area over a defined period.

**Methodology:**

1. Data Collection and Preprocessing:

a) Data Source:

* 2023 surface water quality dataset from Bristol Open Data portal.

b) Data Cleaning:

* + Remove duplicate entries
  + Handle missing values: Impute numerical columns with median, categorical with mode
  + Remove extreme outliers using the Interquartile Range (IQR) method c) Feature Engineering:
  + Create derived features if applicable (e.g., ratios between parameters)
  + Normalize numerical features using Min-Max scaling

1. Exploratory Data Analysis (EDA):
2.  Temporal Analysis of Key Water Quality Indicators: This technique involves creating time series plots for critical water quality parameters such as E. Coli (ec\_cfu), Total Coliforms (tc\_cfu), Temperature (temp), pH, and Dissolved Oxygen (do). As demonstrated by Ren et al. (2020) in their study on urban river water quality, temporal analysis can reveal seasonal patterns, long-term trends, and potential anomalies in water quality data. This approach would help you understand how Bristol's water quality varies over time and identify any recurring patterns or sudden changes that might require further investigation.
3.  Spatial Distribution of Water Quality Parameters: Creating maps or spatial heat maps showing the distribution of key water quality parameters across different sampling sites in Bristol can provide valuable insights. This technique, similar to the approach used by Pu et al. (2021) in their UAV-based water quality assessment, allows for the visualization of spatial patterns in water quality. It can help identify hotspots of pollution or areas with consistently good water quality, potentially linking these patterns to local land use or urban features.
4.  Correlation Analysis of Water Quality Variables: Performing a correlation analysis among all numerical variables in your dataset can reveal important relationships between different water quality parameters. This technique, as employed by Li et al. (2021) in their systematic framework for water quality monitoring, involves creating a correlation matrix and visualizing it using a heatmap. This analysis can help identify which parameters are closely related and might inform the selection of variables for your Multiple Linear Regression model.
5.  Distribution Analysis and Outlier Detection: Creating histograms, box plots, and Q-Q plots for key water quality parameters can help you understand their distributions and identify potential outliers. This approach, similar to the data preprocessing steps described by Wang et al. (2023) in their review of AI applications in environmental monitoring, is crucial for understanding the characteristics of your data and identifying any unusual observations that might require further investigation or data cleaning.

 Compliance Analysis with Environmental Standards: This involves comparing observed levels of key water quality parameters with UK environmental standards. You could create binary variables indicating whether each observation meets the relevant standard, then visualize compliance rates over time or across different sites. This type of analysis, reflecting the environmental impact assessment framework discussed by Oral et al. (2020), can provide insights into the overall health of Bristol's water bodies and identify areas or periods of concern.

Machine Learning Modeling:

a) Feature Selection:

* + Use correlation analysis and domain knowledge to select relevant features

b) Data Splitting:

* + Split the data into training (80%) and testing (20%) sets

c) Model Development:

* + Implement Multiple Linear Regression using scikit-learn
  + Target variable: Choose a key water quality indicator (e.g., E.coli or dissolved oxygen)
  + Features: Include relevant environmental and water quality parameters

d) Model Training and Validation:

* + Train the model on the training dataset
  + Implement 5-fold cross-validation to ensure model stability

e) Model Evaluation:

* + Assess model performance using R-squared (R²) and Root Mean Square Error (RMSE) on the test set
  + Compare with a baseline model (mean prediction) to quantify improvement

1. Environmental Impact Assessment: a) Compare observed water quality levels with environmental standards set by the UK Environment Agency b) Identify locations where water quality falls below acceptable standards c) Analyze the relationship between environmental factors and water quality violations
2. Interpretation and Recommendations: a) Interpret the coefficients of the Multiple Linear Regression model to understand the relative importance of different factors in predicting water quality b) Develop data-driven recommendations for improving water management practices and reducing pollution in Bristol's surface waters

**Analytical Approach**

The analytical approach for this study focuses on three key areas:

1. Exploratory Data Analysis: Statistical summaries and visualizations of water quality parameters will be conducted to identify patterns, trends, and potential outliers in the 2023 Bristol surface water quality dataset.
2. Predictive Modeling: Multiple Linear Regression will be employed to develop a model predicting key water quality indicators. The model will be trained on 80% of the data and validated using 5-fold cross-validation. Performance evaluation will utilize R-squared (R²) and Root Mean Square Error (RMSE) metrics.
3. Environmental Impact Assessment: Predicted and observed water quality levels will be compared with UK Environment Agency standards. This comparison aims to identify areas of concern and inform recommendations for water management practices.

Limitations:

This study acknowledges several limitations:

1. Data Constraints: The analysis is limited to the 2023 dataset from Bristol Open Data, which may have gaps in spatial or temporal coverage.
2. Model Simplicity: Multiple Linear Regression assumes linear relationships between variables, potentially overlooking complex, non-linear dynamics in water quality processes.
3. Temporal Scope: The focus on a single year's data prevents the analysis of long-term trends or seasonal variations in water quality.
4. External Validity: Findings may be specific to Bristol and not directly generalizable to other urban areas with different environmental conditions.
5. Unmeasured Variables: Important factors influencing water quality may not be included in the available dataset, possibly leading to omitted variable bias.
6. Sampling Bias: The location and frequency of sampling points may not provide a complete representation of water quality across all of Bristol's water bodies.

Ethical Considerations

This research adheres to strict ethical guidelines throughout the data lifecycle, from collection to destruction:

1.Data Collection and Generation:

* Only publicly available data from Bristol Open Data will be used, ensuring no infringement on personal privacy.
* We will document the source and date of all data collected to maintain transparency.

2. Data Storage and Security:

* All data will be stored on secure, password-protected servers with restricted access.
* Regular backups will be performed to prevent data loss.
* Compliance with GDPR and local data protection regulations will be ensured.

3. Data Anonymization:

* Specific sampling locations will be generalized to broader areas to prevent identification.
* Any potentially identifying information will be removed or aggregated.

4. Data Usage:

* Data will be used solely for the purposes outlined in this research proposal.
* Analysis methods will be documented in detail to ensure reproducibility.

5. Stakeholder Communication:

* We will maintain transparent communication with relevant stakeholders, including local authorities and environmental agencies.
* Research objectives and potential implications will be clearly communicated.

6. Reporting and Publication:

* Findings will be reported objectively, with clear statements of limitations and potential biases.
* Results will be contextualized within established environmental standards to prevent misinterpretation.

7. Data Destruction:

* Upon completion of the research, raw data will be securely deleted following institutional guidelines.
* Aggregated, anonymized data may be retained for future research, subject to ongoing ethical review.

8. Ongoing Ethical Review:

* The project will undergo regular ethical reviews to ensure continued compliance with ethical standards as the research progresses.

By adhering to these ethical considerations throughout the data lifecycle, we aim to conduct our research with the highest standards of integrity, transparency, and responsibility.

**Associated Risks:**

1. Data Quality: Poor data quality or significant gaps in the Bristol Open Data could lead to unreliable results and misleading conclusions about water quality trends.
2. Model Limitations: The Multiple Linear Regression model may oversimplify complex water quality dynamics, potentially missing important non-linear relationships.
3. Misinterpretation of Results: Findings could be misinterpreted or taken out of context by stakeholders or the public, leading to unwarranted concerns or actions.
4. Stakeholder Sensitivity: Results indicating poor water quality in certain areas could lead to political sensitivity or public concern, requiring careful communication of findings.
5. Scope Limitations: The focus on a single year's data may not capture the full complexity of water quality issues in Bristol, potentially leading to incomplete or short-sighted recommendations.

**Project Timeline (2 Months)**

Week 1-2:

* Finalize research proposal and methodology
* Data collection from Bristol Open Data portal
* Initial data cleaning and preprocessing

Week 3-4:

* Complete data preprocessing
* Conduct exploratory data analysis (EDA)
* Begin feature selection and engineering

Week 5-6:

* Develop and train Multiple Linear Regression model
* Perform model validation and evaluation
* Start environmental impact assessment

Week 7:

* Complete environmental impact assessment
* Analyze results and draw conclusions
* Begin drafting the research report

Week 8:

* Finalize research report
* Prepare executive summary and presentation materials
* Review and refine all project deliverables

Final Week:

* Final proofreading and formatting
* Submit research report and supporting materials

**References:**

Bristol City Council. (2020). Bristol One City Plan 2020. <https://www.bristolonecity.com/wp-content/uploads/2020/01/One-City-Plan_2020.pdf>

Bristol Open Data. (2023). Bristol Surface Water Quality Data. <https://opendata.bristol.gov.uk/explore/?q=water+quality&sort=modified>

Bristol Water. (2022). Annual water quality report 2021. <https://www.bristolwater.co.uk/about-us/water-quality/water-quality-reports>

Environment Agency. (2021). The state of the environment: water quality. <https://www.gov.uk/government/publications/state-of-the-environment/the-state-of-the-environment-water-quality>

Goonetilleke, A., Liu, A., Managi, S., Wilson, C., Gardner, T., Bandala, E. R., Walker, L., Holden, J., Wibowo, M. A., Suhatmini, B., Bonotto, D. M., & Zhang, Q. (2019). Stormwater reuse, a viable option: Fact or fiction? Economic Analysis and Policy, 61, 124-132. <https://doi.org/10.1016/j.eap.2018.08.001>

Grimm, N. B., Faeth, S. H., Golubiewski, N. E., Redman, C. L., Wu, J., Bai, X., & Briggs, J. M. (2008). Global Change and the Ecology of Cities. Science (American Association for the Advancement of Science), 319(5864), 756–760. <https://doi.org/10.1126/science.1150195>

Langston, W. J., Pope, N. D., Jonas, P. J. C., Nikitic, C., Field, M. D. R., Dowell, B., Shillabeer, N., Swarbrick, R. H., & Brown, A. R. (2010). Contaminants in fine sediments and their consequences for biota of the Severn Estuary. Marine Pollution Bulletin, 61(1-3), 68-82. <https://doi.org/10.1016/j.marpolbul.2009.12.014>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Rahmani, F., Lawson, K., Ouyang, W., Appling, A., Oliver, S., & Shen, C. (2021). Exploring the exceptional performance of a deep learning stream temperature model and the value of streamflow data. Environmental Research Letters, 16(2), 024025. <https://doi.org/10.1088/1748-9326/abd501>

Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020). A comprehensive review of deep learning applications in hydrology and water resources. Water Science and Technology, 82(12), 2635-2670. <https://doi.org/10.2166/wst.2020.369>

Date Time

Site ID

Total Coliforms (CFU)

E Coli (CFU)

Total Coliforms (MPN)

E. Coli (MPN)

Presumptive Enterococchi (CFU)

Faecal Streptococchi (CFU)

Temperature

pH

Conductivity

Dissolved Oxygen

Dissolved Oxygen (%)

Phosphate as P

Salinity

Clostridium Perfringens (CFU)

Ammonium as NH4

Turbidity

Ammonium as Nitrate

Nitrite

Nitrate

Faecal Coliforms

Salmonella

BOD

COD

Suspended Solids

ID

Site Name

River

Total Coliform

Environmental Standards for Water:

**Total Coliforms:**

EC Bathing Water Directive Microbiological Standards

|  |  |  |
| --- | --- | --- |
| Organism | Mandatory Standard | Guideline Standard |
| Total coliforms | 10,000 per 100 ml | 500 per 100 ml |

Link: <https://publications.parliament.uk/pa/cm199798/cmselect/cmenvtra/266ii/et0216.htm>

**E.Coli:**

Classification Thresholds (percentile)

Coastal Bathing Waters

Excellent EC: ≤250 cfu/100ml ; IE: ≤100 cfu/100ml (95th percentile)

Good EC: ≤500 cfu/100ml ; IE: ≤200 cfu/100ml (95th percentile)

Sufficient EC: ≤500 cfu/100ml ; IE: ≤185 cfu/100ml (90th percentile)

Poor means that the values are worse than the sufficient

Inland Bathing Waters

Excellent EC: ≤500 cfu/100ml ; IE: ≤200 cfu/100ml (95th percentile)

Good EC: ≤1000 cfu/100ml ; IE: ≤400 cfu/100ml (95th percentile)

Sufficient EC: ≤900 cfu/100ml ; IE: ≤330 cfu/100ml (90th percentile)

Poor means that the values are worse than the sufficient

Link: <https://environment.data.gov.uk/bwq/profiles/help-understanding-data.html>

Temperature:

Temperature in rivers can range from 0 - 25°C

Link:<https://buckinghamshire.moderngov.co.uk/documents/s32273/WaterqualitydatainterpretationfornontechnicalcustomersFeb201.pdf>

pH

the UK drinking water quality regulations include pH as an indicator parameter and specify a minimum pH of 6.5 and a maximum pH of 9.0.

<https://www.dwi.gov.uk/the-physical-and-chemical-properties-of-water/#:~:text=The%20UK%20drinking%20water%20quality,a%20maximum%20pH%20of%209.0>.

Conductivity:

• Conductivity varies in streams and rivers from 300 to 1000 µS/cm. The concentration is affected primarily by the geology of the area – the bedrock through which the water flows.

<https://buckinghamshire.moderngov.co.uk/documents/s32273/WaterqualitydatainterpretationfornontechnicalcustomersFeb201.pdf>

Turbidity:

Turbidity values will vary greatly from near zero in dry weather to ≥ 500 NTU during heavy rain. Values of 25 NTU are common as background levels.

<https://buckinghamshire.moderngov.co.uk/documents/s32273/WaterqualitydatainterpretationfornontechnicalcustomersFeb201.pdf>

Dissolved Oxygen:

The DO concentration in unpolluted waters is around 10 mg/l or between 90-110 %. The critical level of DO in water varies greatly between species. • Concentrations below 5 mg/l or around 50% may adversely affect the functioning and survival of biological communities • Concentrations below 2 mg/l or around 20% may cause fish to die.

Phosphate:

Government guidance recommends that rivers should not exceed annual mean phosphate concentrations of 0.1mg per litre.

<https://assets.publishing.service.gov.uk/media/5a7b1f23ed915d3ed90624fc/defra-stats-observatory-indicators-da3-120224.pdf>

**Thesis Outline with Proposal Integration**

**I. Introduction**

1. **Background Information**
   * Use the section "Introduction" from your proposal, including details about the critical concern for public health, ecosystem integrity, and sustainable urban development in Bristol.
   * Reference the motivations for using machine learning techniques in this context.
2. **Research Problem or Question**
   * Incorporate your "Problem Statement," explaining the challenges Bristol faces in maintaining surface water quality and the need for data-driven approaches.
3. **Objectives of the Study**
   * Use the "Objectives" section, detailing the goals of analyzing current water quality, identifying influencing factors, developing predictive models, and generating insights for water management.
4. **Significance and Motivation of the Research**
   * Expand on the significance of your research, discussing its potential contributions to public health, environmental sustainability, and urban water management practices.

**II. Literature Review**

1. **Overview of Relevant Literature**
   * Summarize the "Literature Review" from your proposal, highlighting key studies on urban water quality, machine learning applications in environmental science, and specific research on Bristol.
2. **Key Theories or Concepts**
   * Discuss the theoretical frameworks and key concepts that underpin your research, such as the interplay between urban development and water quality.
3. **Gaps or Controversies in the Literature**
   * Identify any gaps in existing research that your study aims to address, particularly the limited application of machine learning in urban water management.
4. **Theoretical Framework**
   * Detail the conceptual framework you will use, including the key variables and their relationships as mentioned in your proposal.

**III. Methodology**

1. **Research Design**
   * Describe the overall research design, including your focus on using the 2023 dataset from Bristol Open Data for a detailed analysis of water quality in Bristol.
2. **Data Collection Methods**
   * Use the "Data Collection and Preprocessing" section, detailing the data sources, cleaning processes, and any feature engineering you will perform.
3. **Sampling Techniques**
   * Explain any sampling techniques used to ensure representative data, such as handling missing values and removing outliers.
4. **Data Analysis Procedures**
   * Detail the analytical methods, including exploratory data analysis, correlation analysis, machine learning modelling (Multiple Linear Regression), and environmental impact assessment.

**IV. Results and Analysis**

1. **Presentation of Findings**
   * Present your findings from the exploratory data analysis, correlation analysis, and machine learning models.
2. **Data Analysis and Interpretation**
   * Interpret the results, discussing the significance of key factors influencing water quality and the performance of your predictive models.
3. **Discussion of Results in Relation to Research Questions/Hypotheses**
   * Relate your findings to the research questions posed in your proposal, providing answers and insights based on your analysis.

**V. Discussion**

1. **Interpretation of Results**
   * Provide a detailed interpretation of your results, explaining the implications of your findings for water quality management in Bristol.
2. **Comparison with Existing Literature**
   * Compare your results with findings from the literature review, highlighting any similarities, differences, or new insights.
3. **Implications and Significance of the Findings**
   * Discuss the broader implications of your research for public health, environmental sustainability, and urban water management.
4. **Limitations of the Study**
   * Acknowledge any limitations, such as data constraints, model assumptions, and the scope of your analysis.

**VI. Conclusion**

1. **Summary of Key Findings**
   * Summarize the key findings of your study, emphasizing the most important insights and their implications.
2. **Contributions to the Field**
   * Highlight how your research contributes to the field of environmental science and machine learning, particularly in the context of urban water management.
3. **Recommendations for Future Research**
   * Provide recommendations for future research, suggesting areas where further studies could build on your work.

**VII. References**

* List all the references cited in your thesis, ensuring proper formatting and citation.

**VIII. Appendices (if applicable)**

* Include any additional materials, such as detailed data tables, code, or supplementary analyses, in the appendices.