Addressing ESG and DEIB Goals in Water Access Analysis

Course: Computational Data Analytics with Python

Assignment A1: Individual Project Report

Due Date: March 01, 2025

Student Name: Abi Joshua George

Student ID: 46656697

Table of Contents

1. Introduction and Overview of the ESG Issue 3
1.1 Background on Water Access Disparities: 3
1.2 Why Water Access is an ESG Issue: 3
1.3 Role of Data Analysis in Addressing Water Access Issues: 4
2. Methodology and Techniques Used 4
2.1 Data Source: 4
2.2 Preprocessing Steps: 5
2.3 Exploratory Data Analysis (EDA): 5
2.4 Regression Models (For Forecasting Rural Water Access): 5
2.5 Classification Models (To Classify Water Accessibility Levels): 6
2.6 Clustering Models (For Country Clustering and Uncovering Patterns): 6
2.7 Model Evaluation Metrics: 6
3. Key Findings 7
3.1 Regression Model Results: 7
3.2 Classification Model Results: 7
3.3 Clustering Model Results: 8
4. Interpretation in the Context of ESG 9
4.1 Environmental Impact (E): 9
4.2 Social Impact (S): 9
4.3 Governance and Policy (G): 9
4.4 Contribution to ESG Goals: 10
5. Ethical Considerations Related to DEIB 10
6. Social and Environmental Impact 11
7. References 12

1. Introduction and Overview of the ESG Issue

Background on Water Access Disparities:

Access to clean and safe drinking water is a natural human right, but millions of people all over the world are still struggling to achieve a steady supply of water. Most conspicuous is the disparity between urban and rural water supply. Residents in urban areas generally have great infrastructures in place that supply them with clean water regularly, but rural areas often depend on minimal water supply because of the lack of proper infrastructure, geographical limitations, and limited finances.

Based on the World Health Organization (WHO), approximately 2 billion individuals worldwide depend on water sources with fecal contamination, which causes serious health consequences like cholera, diarrhea, and other waterborne illnesses (WHO, 2023). This is common in developing countries where water treatment facilities and infrastructure are limited. In many cases, rural folk are compelled to rely on unimproved water supplies such as surface water, with adverse health and economic effects.

Why Water Access is an ESG Issue:

The ESG framework provides a lens through which companies, governments, and stakeholders measure sustainability and ethical conduct. Water Access is directly connected to all three pillars of ESG:

- **Environmental Impact:** Proper water management for sustainability is needed for climate resilience, conservation, and pollution control. Water scarcity because of climate change threatens both society and ecosystems and hence is an urgent environmental concern.

- **Social Equity:** Access to safe drinking water is a social justice issue as it impacts public health, gender equality, and poverty reduction. Women and children in many rural areas bear the burden of walking long distances to fetch water, with less time for education and economic activity.

- **Governance & Policy:** Policies, investments, and infrastructure must be developed and implemented by governments, corporations, and non-profits to make water more accessible. International initiatives such as the United Nations Sustainable Development Goal (SDG) 6: Clean Water and Sanitation, demand universal and equal access to water by 2030 (United Nations, 2023).

Several organizations have incorporated water stewardship programs as a measure of adopting ESG principles. For example, Coca-Cola and Nestlé are committed to enhance water efficiency and community water projects, while governments globally have introduced policies to build sustainable water infrastructure neglected areas.

Role of Data Analysis in Addressing Water Access Issues:

Analysis of data is important in identifying and mitigating the water access problem. Through statistical techniques and machine learning models, trends, patterns, the most significant risk factors for water disparity can be uncovered. Exploratory data analysis (EDA), regression, classification, and clustering models are applied in this project to:

- Assess regional disparities in rural water access.
- Predict the most significant factors impacting water availability, such as population growth, geographic position, and infrastructure investment.
- Group and classify countries according to levels of water accessibility to identify areas with the highest level of intervention needed.

Through data analysis, we can generate insights that can be utilized in informing decision-making by policymakers, NGOs, as well as corporate investors investing in water sustainability programs. This research informs ESG-compliant strategies through identification of at-risk regions as well as prioritizing investment in these areas for sustainable water systems.

2. Methodology and Techniques Used

Data Source:

The dataset used in this analysis contains information on the levels of rural water access in countries, which was obtained from credible sources like the World Health Organization (WHO), the United Nations (UN), and the World Bank. The dataset has variables for:

- Water access indicators (e.g., Rural At Least Basic Water Access, Rural Limited Access, Rural Unimproved Water Access).
- Demographics (e.g., population, urbanization rate, per capita GDP).
- Geographical characteristics (e.g., country, iso3, WHO region, SDG region).

- Policy and infrastructure characteristics (e.g., UNICEF reporting region and governance type).

Preprocessing Steps:

Before performing the analysis, the dataset underwent data cleaning and preprocessing steps such as:

- Managing missing values using imputation methods wherever required.
- Encoding categorical features (e.g., regions, continents) for machine learning model compatibility.
- Scaling numeric features so that fair comparison is possible among variables.

Exploratory Data Analysis (EDA):

EDA was performed to reveal patterns, relationships, and disparities in water availability across various areas. The main methods are:

- Scatterplots and boxplots for comparing the degree of water access by continents.
- Time-series analysis to study trends in water accessibility across the years.
- Correlation analysis to determine relationships between access to water and socio-economic indicators.

Regression Models (For Forecasting Rural Water Access):

Regression analysis of the model was used in estimating the percentage of rural populations with access to at least basic drinking water. The models tested are:

- Linear Regression, a baseline model for comparison.
- Decision Tree Regressor to model non-linear relationships.
- Random Forest Regressor (Ensemble learning for better accuracy).
- K-Nearest Neighbors (KNN) Regressor to measure performance based on proximity-based learning.

Classification Models (To Classify Water Accessibility Levels):

Classification models were used for grouping nations into high-access and low-access according to water supply. The models used are:

- Logistic Regression to assess linear decision boundaries.
- Decision Tree Classifier, for interpretability and structure-based classification.
- Random Forest Classifier, for better generalization and performance.
- K-Nearest Neighbors (KNN) Classifier, for classification of data points based on similarity.

Clustering Models (For Country Clustering and Uncovering Patterns):

Clustering algorithms were used to find natural groupings within the data in order to target interventions:

- K-Means Clustering, for clustering countries based on similar water access attributes.
- Hierarchical Clustering, for displaying relationships between clusters using dendrograms.
- DBSCAN (Density-Based Clustering), for anomaly detection and unusual pattern discovery.

Model Evaluation Metrics:

To ensure robust model selection in each model type, the following evaluation metrics were used:

Model Type	Metrics Used
Regression	Mean Absolute Error (MAE), Mean Squared Error (MSE), R^2 Score
Classification	Accuracy, Precision, Recall, F1-Score, Confusion Matrix
Clustering	Elbow Method, Silhouette Score

By using a mix of supervised and unsupervised learning techniques, this analysis provides data-driven insights into water access challenges in households. This methodology ensures that the analysis aligns with ESG principles, specifically in understanding social equity (S) and environmental sustainability (E).

3. Key Findings

Regression Model Results:

For predicting rural water access,

- **Best Performing Model:** The Random Forest Regressor demonstrated the lowest error rates, making it the most reliable model for predicting rural water access levels.

- **Urban percentage:** Higher urbanization rates were correlated with higher rural water access, suggesting that urban development indirectly improves rural infrastructure.

- **Continental disparities:** Africa had the lowest predicted rural water access compared to other continents, emphasizing the need for region-specific interventions.

Classification Model Results:

For identifying countries with low vs high water access,

- **Best Performing Model:** Random Forest Classifier achieved the highest accuracy and F1-score, effectively distinguishing between countries with high and low water access levels.

- **Logistic Regression** had more false negatives, meaning it struggled to correctly classify countries with low water access.

- **Tree-based models (Decision Tree, Random Forest**) performed better due to their ability to capture non-linear relationships.

Clustering Model Results:

- **Optimal Clustering Method:** K-Means Clustering produced the best silhouette score, indicating that the identified clusters are well-defined.

- **Distinct Clusters Identified:**
  - **Cluster 1:** High rural water access (>90%), mainly developed nations.
  - **Cluster 2:** Moderate access (60% to 90%), emerging economies with stable water infrastructure.
  - **Cluster 3:** Low access (<60%), countries facing severe water crises, mainly in sub-Saharan Africa and parts of South Asia.

Prediction Results:

Following the in-depth comparison of all models, Random Forest Regressor performed best based on its higher accuracy and lower error metrics. In order to compare its predictive capacity, the model was applied on the dataset to predict rural water access.

Model Performance Metrics,

The model's performance was tested with standard regression performance metrics:

- Mean Absolute Error **(MAE): 0.02**
- Mean Squared Error **(MSE): 0.01**
- R^2 Score: **0.97**

These results inform us that the model makes very accurate predictions, with minimal deviation from actual values.

To validate the accuracy of the model, an Actual vs. Predicted Values plot was generated. The visualization attests that most predictions closely adhere to actual values, with data points forming a near-perfect diagonal trend. This suggests that the Random Forest Regressor is correctly detecting underlying patterns in rural water access data.

Implications of the Results,

The ability of the model to generate high-accuracy predictions for water access levels has significant practical implications for development institutions and policymakers. The model can assist:

- Infrastructure planning – Identifying regions where water accessibility is most needed.
- Resource allocation – Enabling effective allocation of investments and interventions.

While the model itself works very well, future refinement can be done by adding additional socioeconomic and environmental factors to increase the accuracy and reliability of projections.

4. Interpretation in the Context of ESG

Environmental Impact (E):

- Water Scarcity: The key findings highlight environmental concerns in water resource management, particularly for developing nations.
- Infrastructure and Global Warming: The reduced access to water in some of the clusters requires climate-driven water scarcity opportunities and investment in sustainable water infrastructure.
- Governments must use predictive analysis to forecast future water access scenarios and invest accordingly.

Social Impact (S):

- Water Equity: The classification models capture extreme disparities in access to water, revealing a social gap that needs addressing.
- Health and Hygiene: Poor access to water is related to health crises, rendering the topic an issue for policy response.
- The cluster analysis ensures the targeting of areas of high vulnerability to investment in infrastructure and assistance.

Governance and Policy (G):

- Data models provide quantifiable evidence to governments to help them design improved policies for access to water.
- Monitoring and Accountability: Governments and institutions can track progress toward SDG 6 (Clean Water and Sanitation) through predictive analytics.

- International Cooperation: The results encourage global cooperation in efforts aimed at improving the distribution and management of water to underprivileged regions.

Contribution to ESG Goals:

ESG Goal	How this Analysis helps
SDG 6 (Clean Water & Sanitation)	Identifies at-risk regions and enables predictive policymaking.
SDG 10 (Reduced Inequalities)	Highlights disparities in water access, promoting equity-driven solutions.
SDG 13 (Climate Action)	Provides insights into climate-related risks affecting water availability.

5. Ethical Considerations Related to DEIB

Fairness and Inclusivity within Data Analysis,

Ethical considerations is important in ensuring data-driven conclusions do not reinforce inequalities, particularly with respect to global water accessibility. Bias during data collection can lead to misleading assumptions, affecting resource allocation and policy decision (United Nations, 2015). Sustaining fairness within this analysis aligns with Sustainable Development Goal 6 (SDG 6), which seeks global and equitable water accessibility.

Bias in Data and Model Predictions,

One of the primary issues in this analysis is the geographic bias as only fewer data points were available for African and South American nations than for North American and European nations, which will impact model applicability. Sampling bias was also an issue as more reported water access areas were overrepresented, and this could have caused overexaggerated predictions for underserved areas (Mehrabi et al., 2021).

To solve this issue, the dataset was categorized by continent and grouped for regional fairness, and model accuracy classification was subjected to a regional test. Owing to high rates of performance, differential rates of recall measures suggest that some regions were more frequently classified in error and called for fairness audits.

Improvement of Ethical Concerns,

Following are the methods employed in order to improve fairness concerns:

- Balanced Data Representation – Categorizing by continents ensured regional equality.

- Bias Detection in Model Predictions – Classification accuracy differences were explored by region.

- Ethical Visualization – Information was presented in neutral forms to avoid misinterpretations (Holstein et al., 2019).

Future improvement could be with fairness-aware machine learning techniques, such as re-weighting underrepresented data and incorporating socioeconomic factors for a more comprehensive analysis.

Real-World Implications,
These findings can be employed to inform policy and sustainability action. Policymakers can use AI models in order to identify at-risk regions, allocate resources more effectively, and forecast future patterns of water scarcity. Businesses can also integrate these findings into company sustainability programs for promoting equitable water management (United Nations, 2015).

6. Social and Environmental Impact

Water access is a worldwide issue, particularly in rural regions where infrastructure limitations and climate change exacerbate the scarcity gap. Machine learning models are applied in this study to analyze worldwide trends in water access, which provides valuable insights for policymakers, business, and sustainability professionals. By identifying patterns in water accessibility, such models can guide intervention to improve resource allocation and investment in infrastructure (World Health Organization, 2022).

The study observed significant differences in rural water availability, with the poorest regions experiencing the greatest challenges. Machine learning models, particularly Random Forest classifier and regressor, had very good predictive power, and thus they can be considered as possible tools for assessing water accessibility. Data bias and regional gaps, however, remain persistent challenges, emphasizing the importance of responsible AI deployment in decision-making (Mehrabi et al., 2021).

The key findings of the analysis are:

- Predictive modeling can help support early intervention programs by governments and NGOs.

- AI-powered clustering techniques help identify areas requiring urgent water access improvement.

- Sustainability initiatives can include these insights to support ESG programs, supporting SDG 6 (Clean Water and Sanitation) (United Nations, 2015).

Environmental and Social Considerations,

The environmental impact of water availability is two-faceted—climate change degrades natural sources of water, and mismanagement of water causes depletion of resources. The findings of the study emphasize the role of sustainable policies, such as rainwater harvesting, efficient irrigation, and conservation of groundwater. Socially, increased access to water reduces gender inequality, with women in most regions being tasked with water collection (World Bank, 2021).

Conclusion: Selecting the Best Model,

Based on the regression, classification, and clustering analysis of all models, Random Forest was the most effective in predictive tasks (high accuracy, low error rate) and classification (99% accuracy, lowest misclassification rate).

For clustering, K-Means produced the most interpretable country clusters based on water availability, while DBSCAN struggled with forming meaningful clusters. These results show the importance of selecting a model based on the use case and quality of data.

7. References

- World Health Organization (WHO). (2023). *Drinking Water*. Retrieved from https://www.who.int/news-room/fact-sheets/detail/drinking-water
- United Nations. (2023). *Sustainable Development Goal 6: Clean Water and Sanitation*. Retrieved from https://sdgs.un.org/goals/goal6
- United Nations. (2015). Sustainable Development Goals (SDG 6): Ensure availability and sustainable management of water and sanitation for all. Retrieved from https://sdgs.un.org/goals

- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.
- Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? CHI Conference on Human Factors in Computing Systems.
- World Health Organization. (2022). Global progress on water, sanitation, and hygiene: SDG 6 update. Retrieved from https://www.who.int
- World Bank. (2021). Gender and water: Addressing inequities in access and management. Retrieved from https://www.worldbank.org