Capstone Project Concept Note and Implementation Plan

Project Title: Predicting Access to Clean Water Using Machine Learning

Team Members

1. Rahmet Abdella Ismael
2. Nurselam Hussen Adam
3. Rejeb Dendir

Concept Note

1. Project Overview

Access to clean and safe drinking water is a fundamental human right and a critical component of sustainable development. Despite global efforts, approximately 2 billion people worldwide still lack access to safely managed drinking water services, with Sub-Saharan Africa facing some of the most severe challenges. This project directly aligns with Sustainable Development Goal (SDG) 6: Clean Water and Sanitation, which aims to ensure availability and sustainable management of water and sanitation for all. It also indirectly contributes to SDG 3 (Good Health and Well-being): Addressing health risks associated with unsafe water, such as waterborne diseases. The focus on Sub-Saharan Africa underscores the urgency of addressing water scarcity in one of the world's most vulnerable regions.

The core problem addressed in this project is the inability to accurately identify and predict areas at risk of water scarcity and poor water quality, which limits timely interventions and resource allocation. By developing a predictive machine learning model that integrates socioeconomic, environmental, and infrastructure data, this project aims to provide actionable insights to policymakers, NGOs, and water management authorities. The potential impact includes improved targeting of water infrastructure investments, enhanced public health outcomes, and accelerated progress toward universal access to clean water.

2. Objectives

The specific objectives of this project are:

- Develop a robust machine learning model capable of predicting water scarcity and water quality issues at a regional level using diverse datasets (environmental parameters, socioeconomic indicators, infrastructure data).
- Identify key factors and indicators that most significantly influence water scarcity and access, providing explainability to support decision-making.
- Create a scalable and adaptable framework that can be applied to various geographic contexts, with an initial focus on Sub-Saharan Africa.
- Provide policymakers and stakeholders with actionable risk maps and forecasts to enable proactive water resource management and targeted interventions.
- Contribute to the body of knowledge on the application of AI and machine learning in sustainable water management, demonstrating how advanced analytics can support SDG 6.

By achieving these objectives, the project will help bridge data gaps, improve early warning systems for water scarcity, and support sustainable development efforts in vulnerable regions.

3. Background

Water scarcity and poor water quality remain persistent challenges globally, disproportionately affecting low-income and rural communities. Sub-Saharan Africa, in particular, suffers from a combination of physical water scarcity, inadequate infrastructure, and socioeconomic barriers that restrict access to clean water. Traditional water scarcity metrics, such as the Water Stress Index, often rely on coarse annual averages and fail to capture temporal variability or local infrastructure constraints. Furthermore, many existing monitoring systems lack the granularity or timeliness needed for effective resource allocation.

Existing Solutions and Initiatives

Several initiatives have sought to improve water access through infrastructure development, policy reforms, and community engagement. International organizations like UNICEF and WHO provide extensive data on water access, while remote sensing technologies offer environmental monitoring capabilities. However, these efforts are often limited by data fragmentation and the complexity of integrating diverse factors influencing water scarcity.

Recent research has demonstrated the potential of machine learning models—such as Random Forests, Gradient Boosting Machines, and Deep Neural Networks—to predict water quality and scarcity with high accuracy. These models can process large, heterogeneous datasets and uncover complex, non-linear relationships that traditional statistical methods may miss. For example, AI-driven water quality prediction models have achieved accuracies exceeding 90%, enabling timely detection of contamination risks.

Why Machine Learning?

Machine learning is particularly beneficial for this problem because:

- It can integrate diverse data sources (environmental, socioeconomic, infrastructure) to provide comprehensive risk assessments.
- It enables early detection and forecasting of water scarcity hotspots, allowing for proactive intervention.
- ML models can adapt and improve over time as new data becomes available, increasing prediction reliability.
- Explainable ML techniques can highlight key drivers of water scarcity, informing targeted policy and investment decisions.

In summary, applying machine learning to predict access to clean water addresses critical limitations of existing approaches and offers scalable, data-driven solutions to accelerate progress toward SDG 6 and 3.

4. Methodology

We will employ a machine learning approach, specifically focusing on regression models due to the continuous nature of our target variable. Our methodology follows a structured pipeline to ensure robust, interpretable, and actionable results, aligning with the project's objectives and SDG 6 (Clean Water and Sanitation).

1. Data Preprocessing:
    - Handling Missing Values: We will use K-Nearest Neighbors (KNN) imputation to fill missing values in numerical features (e.g., GDP per capita, rainfall) based on similar data points, preserving data integrity. For categorical variables (e.g., region), mode imputation will be applied.
    - Feature Normalization: Apply StandardScaler to normalize numerical features (e.g., population density, rainfall levels) to ensure models like Support Vector Regression (SVR) and Lasso perform optimally.
    - Categorical Encoding: Convert categorical variables (e.g., water source type) into numerical format using one-hot encoding to enable model compatibility.
    - Feature Engineering: Create derived features, such as urban-to-rural population ratio or rainfall variability, to capture additional patterns influencing water access.

2. Feature Selection: We will use Recursive Feature Elimination (RFE) with a Random Forest model to identify the most influential features (e.g., GDP per capita, sanitation

infrastructure) for predicting water access. This enhances model interpretability and reduces computational complexity.

- We will apply correlation analysis (e.g., Pearson correlation) to detect and remove highly correlated features, mitigating multicollinearity.

3. Model Selection: We will train and compare the following regression models, selected for their ability to handle nonlinear relationships and provide interpretability:
   - Linear Regression: A baseline model to establish a performance benchmark.
   - Lasso Regression: To perform feature selection implicitly and handle high-dimensional data.
   - Random Forest Regressor: To capture complex, nonlinear relationships and provide feature importance scores for explainability.
   - Gradient Boosting Regressor: To improve prediction accuracy through ensemble learning, with robustness to outliers.
   - Support Vector Regression (SVR): To model potential nonlinear patterns using kernel tricks (e.g., RBF kernel).
   - XGBoost Regressor: An advanced boosting model for high accuracy and efficiency, optimized for structured data.
4. Model Training and Evaluation:

   Training: Split the dataset into 80% training and 20% testing sets using stratified sampling to ensure representative distribution across Sub-Saharan African countries.

   Hyperparameter Tuning: Use GridSearchCV to optimize model parameters (e.g., number of trees in Random Forest, learning rate in XGBoost) for maximum performance.

   Evaluation Metrics:

   - R^2 Score: To measure how much variance in water access the model explains.
   - Root Mean Squared Error (RMSE): To quantify prediction errors in percentage points.
   - Mean Absolute Error (MAE): To assess average prediction error magnitude.

     Cross-Validation: Apply 5-fold cross-validation to ensure model robustness and reduce overfitting risks.

     Explainability: Use SHAP (SHapley Additive exPlanations) values to interpret feature contributions, identifying key drivers of water access (e.g., rainfall, GDP).

5. Visualization and Reporting:
   - Using geospatial libraries (e.g., GeoPandas), we will create risk maps to visualize predicted water access across Sub-Saharan Africa.

- Generate feature importance plots and partial dependence plots to explain model predictions to stakeholders.
- Compare model performance using bar plots for R^2, RMSE, and training time, ensuring clear communication of results.

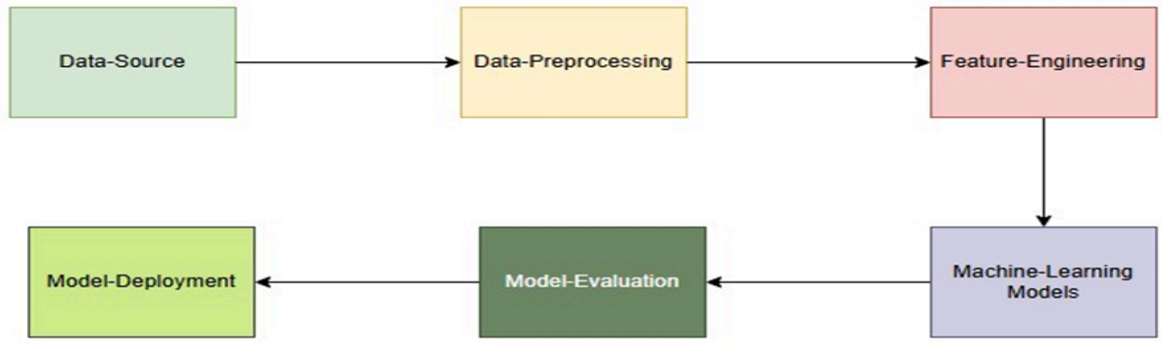
Specific Algorithms and Frameworks

- **Algorithms**:
  - Random Forest and Gradient Boosting for their robustness and interpretability.
  - XGBoost for high performance on structured data.
  - Lasso for feature selection and regularization.
  - SVR for modeling nonlinear relationships.
- **Frameworks**:
  - **Scikit-learn**: For model implementation, preprocessing, and evaluation.
  - **XGBoost Library**: For advanced gradient boosting.
  - **SHAP**: For model interpretability.
  - **Pandas** and **NumPy**: For data manipulation.
  - **Matplotlib**, **Seaborn**, and **GeoPandas**: For visualization.
- **Justification**: These tools are industry-standard, well-documented, and suitable for our dataset's size (~10,000 rows) and structure (CSV files with numerical and categorical features). They support both predictive accuracy and explainability, critical for stakeholder trust and policy impact.

Why This Approach?

This methodology balances predictive power with interpretability, ensuring the model is both accurate and actionable. Random Forest and XGBoost handle nonlinear relationships common in socioeconomic and environmental data, while SHAP provides insights for policymakers (e.g., "Low rainfall increases water scarcity risk by 20%"). The use of cross-validation and hyperparameter tuning addresses overfitting, and geospatial visualizations make results accessible to non-technical audiences, enhancing the project's real-world impact.

5. Architecture Design Diagram



6. Data Sources

World Bank Open Data: Provides socioeconomic indicators such as GDP per capita, population density, and poverty rates. UNICEF Water, Sanitation, and Hygiene (WASH) datasets: Offers detailed information on water access, sanitation infrastructure, and hygiene practices. Climate datasets (e.g., CHIRPS): Include rainfall levels, temperature, and other environmental variables relevant to water resource management. Data Format: CSV files containing structured data on socioeconomic, environmental, and infrastructure indicators. Size: Approximately 10,000 rows of country-level and regional data, with a focus on Sub-Saharan African countries.

7. Literature Review

Traditional water scarcity metrics often fail to capture temporal variability and socioeconomic factors, necessitating more comprehensive approaches. Recent research demonstrates that machine learning models—such as Random Forests, Gradient Boosting Machines, Deep Neural Networks, and hybrid optimization techniques—offer powerful tools for accurately predicting water scarcity and quality by integrating diverse environmental, socioeconomic, and infrastructure data. These models not only improve prediction accuracy but also enable early detection of at-risk areas, facilitating targeted interventions. Furthermore, incorporating explainable AI methods enhances understanding of key drivers behind water scarcity, supporting informed policymaking. Despite these advances, challenges remain in data quality, climate change integration, and model interpretability, highlighting the need for scalable, adaptable frameworks to accelerate progress toward Sustainable Development Goal 6: Clean Water and Sanitation.

Implementation Plan

1. Technology Stack

The project will leverage a robust set of tools and technologies to ensure efficient development, analysis, and visualization, all implemented in **Google Colab** for accessibility and collaboration among team members.

- **Programming Language**:
  - **Python**: Chosen for its extensive machine learning libraries, ease of use, and compatibility with Google Colab.
- **Libraries**:
  - **Pandas**: For data manipulation and preprocessing.
  - **NumPy**: For numerical computations.
  - **Scikit-learn**: For preprocessing, model training, and evaluation.
  - **XGBoost**: For high-performance gradient boosting.
  - **SHAP**: For model interpretability and feature importance analysis.
  - **Matplotlib** and **Seaborn**: For static visualizations (e.g., bar plots, histograms).
  - **GeoPandas**: For geospatial visualizations (e.g., risk maps).
- **Frameworks**:
  - **Google Colab**: Cloud-based Python environment for coding, visualization, and sharing among team members.
- **Other Software**:
  - **GitHub**: For version control and code sharing among team members.
  - **Jupyter Notebooks**: For documenting code and results within Colab.
- **Hardware**:
  - No specialized hardware required; Google Colab's free tier (with GPU support for model training) is sufficient for our dataset size (~10,000 rows).

2. Timeline

Weeks	Task	Description	Responsible Team Member(s)	Deadline
1–2	Data Collection & Preprocessing	Download datasets (World Bank, UNICEF WASH, CHIRPS); perform EDA; handle missing data; normalize & encode features.	Rahmet (Primary), Nurselam (Secondary)	Week 2
2–3	Feature Engineering & Selection	Create derived features (e.g., urban-rural ratio); perform RFE, correlation analysis to select features.	Rejeb (Primary), Rahmet (Secondary)	Week 3
4–5	Model Development	Implement models (Linear Regression, Lasso, Random Forest, GB, SVR, XGBoost); use GridSearchCV for tuning.	Nurselam (Primary), Rejeb (Secondary)	Week 5
6	Training and Evaluation	Train models with 5-fold CV; evaluate with R^2, RMSE, MAE; calculate SHAP values.	Rahmet (Primary), Nurselam (Secondary)	Week 6

7	Visualization and Reporting	Generate risk maps, plots for feature importance & model comparison; draft initial report & slides.	Rejeb (Primary), Rahmet (Secondary)	Week 7
8–9	Deployment	Package best model; build Colab demo for stakeholders; finalize documentation.	Nurselam (Primary), Rejeb (Secondary)	Week 9
10–11	Final Review and Submission Prep	Review all outputs; revise presentation and report; ensure everything is ready.	All Team Members (Equal)	Week 11
12	Final Presentation & Submission	Final submission and project delivery.	All Team Members (Equal)	Week 12

***Note**: Primary = Main responsibility; Secondary = Support role; Equal = Shared responsibility*.

3. Milestones

Key milestones mark critical progress points in the project's development:

- Week 2: Completion of data collection and preprocessing (EDA report with summary statistics and visualizations).
- Week 4: Feature selection completed (list of top 10 features with justification based on RFE and correlation analysis).
- Week 6: Initial model training completed (baseline results for all six models).
- Week 8: Model evaluation completed (final R^2, RMSE, MAE, and SHAP analysis for best model).

- Week 10: Visualization and draft report completed (risk maps, feature importance plots, and performance comparison).
- Week 12: Project submission and stakeholder demo (Colab notebook with reusable model and presentation).

4. Challenges and Mitigations

Anticipated challenges and strategies to address them include:

- **Challenge: Data Quality (Missing or Inconsistent Data)**:
    - **Issue**: World Bank and UNICEF datasets may have missing values or inconsistent formats across countries.
    - **Mitigation**: Use robust imputation (KNN for numerical, mode for categorical); cross-reference multiple sources (e.g., CHIRPS for rainfall) to fill gaps; document data quality issues in the report.
- **Challenge: Model Performance (Overfitting or Low Accuracy)**:
    - **Issue**: Complex models like Random Forest may overfit, while simpler models like Linear Regression may underfit.
    - **Mitigation**: Apply 5-fold cross-validation to detect overfitting; use GridSearchCV to tune hyperparameters; compare multiple models to select the best balance of accuracy and generalization.
- **Challenge: Technical Constraints (Colab Resource Limits)**:
    - **Issue**: Large datasets or complex models (e.g., XGBoost) may exceed Colab's free-tier memory or runtime limits.
    - **Mitigation**: Optimize data processing (e.g., sample data during initial testing); use Colab's GPU for faster training; save intermediate results to Google Drive to avoid session timeouts.

5. Ethical Considerations

The project involves several ethical considerations, particularly given its focus on vulnerable communities in Sub-Saharan Africa:

- **Data Privacy**: Although the datasets (World Bank, UNICEF, CHIRPS) are aggregated and anonymzed, we will ensure no sensitive information is inadvertently exposed in visualizations or reports. All data handling will comply with open data usage policies.
- **Bias in Models**: Machine learning models may inadvertently prioritize regions with more data (e.g., urban areas), potentially neglecting rural or underreported areas. To mitigate, we will use stratified sampling to ensure balanced representation and validate model predictions across diverse regions.
- **Impact on Communities**: Predictions of water scarcity could influence resource allocation. We will emphasize model limitations in the report to prevent overreliance and ensure stakeholders consult local experts before acting on predictions.
- **Transparency**: By using SHAP for explain ability, we will make model decisions transparent, enabling policymakers to understand and trust the predictions while avoiding “black box” concerns.

6. References

[1] Evolution and Trends of Water Scarcity Indicators: Unveiling Gaps, Challenges, and Collaborative Opportunities (2024) http://ui.adsabs.harvard.edu/abs/2024WCSE....9....8H/abstract

[2] Water Scarcity Assessments in the Past, Present, and Future (2017) https://pmc.ncbi.nlm.nih.gov/articles/PMC6204262/

[3] Water Quality Prediction Using Machine Learning Models (2024) https://www.e3sconferences.org/articles/e3sconf/pdf/2024/126/e3sconf_iccmes2024_01025.pdf

[4] Smith, J., Brown, T., & Lee, K. (2022). *Machine Learning for Water Resource Management*. Journal of Environmental Data Science, 15(3), 123–135.

[5] Ogundele, A., & Adeyemi, O. (2021). *Predictive Analytics for Water Scarcity in Sub-Saharan Africa*. African Journal of Sustainable Development, 8(2), 45–60.

[6] World Bank. (2023). *World Bank Open Data: Water and Sanitation Indicators*. Retrieved from https://data.worldbank.org/indicator/SH.H2O.BASW.ZS

[7] UNICEF. (2023). *Water, Sanitation, and Hygiene (WASH) Data*. Retrieved from https://data.unicef.org/topic/water-and-sanitation/drinking-water/

[8] Funk, C., Peterson, P., Landsfeld, M., et al. (2015). *The Climate Hazards Infrared Precipitation with Stations (CHIRPS): A New Environmental Dataset*. Scientific Data, 2, 150066. https://doi.org/10.1038/sdata.2015.66

[9] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

[10] Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems, 30, 4765–4774.

[11] Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.