CENG 463 - HW1 Report

Water Resource Risk Classification

1. Introduction

The objective of this assignment was to classify countries into Water Resource Risk Categories (0-4) using hydrological indicators from the World Resources Institute. The process involved feature engineering, model training, hyperparameter optimization, and feature importance analysis.

2. Feature Engineering Discussion

Two derived features were created to enhance model performance:

1. Composite Water Stress Index (CWSI): A weighted combination of baseline water stress, groundwater depletion, and drought risk. This consolidates multiple stress factors into a single metric.
2. Seasonal-Flood Interaction (SFI): Interaction between seasonal variability and river flood risk.

Evaluation: As detailed in the Feature Importance section, CWSI proved to be the single most predictive feature in the dataset, validating the effectiveness of this engineering step.

3. Model Training & Hyperparameter Optimization

Five models were trained and tuned using GridSearchCV (5-fold CV). The data was scaled for SVM, KNN, and Logistic Regression. Below is the comparison of Baseline vs. Tuned accuracy:

Model	Baseline Acc	Tuned Acc	Improvement
Random Forest	0.9144	0.9144	0.0000
Gaussian NB	0.6092	0.6092	0.0000
SVM	0.7563	0.7816	+0.0253
KNN	0.8112	0.8836	+0.0724
Logistic Regression	0.6872	0.6872	0.0000

Discussion:
- **KNN** yielded the highest improvement (+7.2%). Tuning the 'weights' parameter to 'distance' likely helped by giving more importance to closer neighbors, refining the decision boundaries.
- **SVM** improved marginally (+2.5%) with kernel and C-parameter tuning.
- **Random Forest** showed no improvement. This indicates the default parameters were already highly effective or the grid search space was not wide enough to find a superior configuration. However, it remained the highest performing model overall.

4. Feature Importance Analysis

Using the best performing model (Random Forest), we analyzed the feature importance scores. The top features were:

Feature Name	Importance Score	Type
CWSI	0.2999	Derived (Task 1)
drr_score	0.1572	Original
bws_score	0.1510	Original
gtd_score	0.1411	Original
SFI	0.0941	Derived (Task 1)

Interpretation: The derived feature **CWSI** is the most influential predictor (approx. 30% importance). This confirms that combining water stress, groundwater depletion, and drought risk creates a stronger signal than these features provide individually. The second derived feature, SFI, contributed moderately.

5. Final Conclusion & Model Selection

Best Model: Random Forest (91.44% Accuracy)

Random Forest is the selected model for this classification task. It significantly outperformed the others (KNN: 88%, SVM: 78%, LR: 68%, GNB: 61%). The Random Forest's ensemble nature allows it to capture complex, non-linear relationships between hydrological indicators better than linear models like Logistic Regression. Additionally, it effectively utilized the derived feature CWSI to maximize predictive accuracy.