CENG 463 Homework 1
Water Resource Risk Classification Report

Group 8

Eren Işık
320201108

Fatih Furkan Keser
310201042

Yusuf Tigrik
310201043

0. Introduction

This report presents a comparative analysis of five machine learning algorithm Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gaussian Naive Bayes, and Logistic Regression applied to the World Resources Institute (WRI) Aqueduct Water Risk Atlas dataset. The objective is to classify countries into water risk categories (0–4) based on various environmental indicators. The study includes feature engineering, baseline model evaluation, hyperparameter tuning, and a comprehensive analysis of feature importance to identify the most suitable model for this kind of problem.

1. Feature Engineering

To capture complex relationships within the water risk data and reduce dimensionality, two composite features were engineered:

1. Composite Water Stress Index (CWSI):

This feature represents key stress indicators into one feature.

$$CWSI = 0.5\ x\ bws_{score} + 0.3\ x\ gtd_{score} + 0.2\ x\ drr_{score}$$

2. Seasonal-Flood Interaction (SFI):

This feature represents the compounding risk of variability and flooding.

$$SFI = sev_{score}\ x\ rfr_{score}$$

Our main goal with feature engineering was to combine related data points to reduce noise and help the models learn better. The strong performance of CWSI proves that grouping correlated indicators like water stress and drought creates a much clearer

signal for the model than using them separately. Similarly, SFI helped us capture the complex relationship between seasonal changes and flooding. Overall we hope these new features made it easier for tree based models like RF to find efficient splits and allowed linear models to detect patterns they might otherwise miss.

2. Model Training & Evaluation

2.1 Experimental Setup

The libraries are Scikit-learn,NumPy,Pandas,Seaborn,Matplotlib.

Cross validation data split is 80% Training, 20% Testing, 42 Random state.

Preprocessing: StandardScaler was applied for distance-based and linear models (SVM, KNN, Logistic Regression). Random Forest and Naive Bayes used raw features.

we set random state to 1923 for model reproductivity.

Target: Multi-class classification (Risk Labels 0-4).

2.2 Baseline Results

Model	Accuracy	Rank
Random Forest	92.87%	1
KNN	80.13%	2
SVM	73.66%	3
Logistic Regression	65.86%	4
Gaussian NB	59.60%	5

2.3 Discussion: Suitability & Performance

The baseline evaluation highlights distinct differences in model suitability:

Random Forest demonstrated exceptional suitability, achieving the highest baseline accuracy (92.87%). Its ensemble nature effectively handles the

complex, non-linear interactions between water risk indicators without requiring feature scaling.

KNN showed good performance (80.13%), indicating that countries with similar risk profiles cluster together in the feature space.

Logistic Regression and **Naive Bayes** proved less suitable. The poor performance of Logistic Regression suggests the data is not linearly separable. Naive Bayes performed worst (59.60%) because the dataset contains highly correlated features (e.g., *bws_score* and *CWSI*), which strictly violates the algorithm's independence assumption.

3. Hyperparameter Tuning

3.1 Optimization Strategy

GridSearchCV with 5-fold cross-validation was employed to optimize hyperparameters. The following parameter grids were explored:

Random Forest: n_estimators, max_depth, min_samples_split

KNN: n_neighbors, weights

SVM: C, kernel

Logistic Regression: C, solver

Gaussian NB: var_smoothing

3.2 Comparative Results (Baseline vs. Tuned)

Hyperparameter Optimization Comparison

Model	Baseline Acc	Tuned Acc	Improvement
Random Forest	0.9287	0.9232	-0.0055
Gaussian NB	0.596	0.596	0.0
SVM	0.7366	0.7892	0.0526
KNN	0.8013	0.8869	0.0856
Logistic Regression	0.6586	0.6575	-0.0011

3.3 Discussion: Impact of Optimization

Hyperparameter tuning had a varying impact across models:

-The most significant improvement (+8.56%) was observed in KNN. Switching from uniform to distance weights allowed the model to pay more attention to

the closest neighbors, which is critical in high-dimensional feature spaces where some neighbors may be less relevant.

-Increasing C to 10 improved accuracy by roughly 2.5%, allowing for a tighter decision boundary, though it still lagged behind tree-based methods.

-Random Forest was already optimal with default settings, highlighting its robustness. Logistic Regression and Naive Bayes showed no improvement, confirming that their underperformance is due to structural limitations (linearity and independence assumptions) rather than suboptimal parameters.

4. Feature Importance Analysis

4.1 Analysis (Based on Random Forest)

The feature importance scores extracted from the best-performing model (Random Forest) are as follows:

Rank	Feature	Importance Score	Type
1	CWSI	29.25%	Derived
2	drr_score	15.83%	Original
3	bws_score	15.53%	Original
4	gtd_score	13.89%	Original
5	SFI	9.52%	Derived
6	rfr_score	8.35%	Original
7	sev_score	7.63%	Original

4.2 Interpretation of Influential Features

The fact that our engineered CWSI is the top feature confirms that our strategy worked. The model clearly prioritized water scarcity factors (stress, drought, depletion) over everything else to determine the risk. Interestingly, flood-related features like **rfr_score** had very little impact on the decisions. This tells us that for this dataset, the 'lack of water' is a much stronger predictor of risk than 'excess water' (flooding). The model essentially learned that if a country is thirsty, it's at high risk regardless of its flood status

5. Comprehensive Model Comparison

5.1 Final Comparison Table

Model	Tuned Accuracy	Precision (Weighted)	Training Speed(sec)	Suitability
Random Forest	92.23%	92.44%	0.195	Excellent
KNN	88.69%	88.85%	0.003	Good
SVM	78.92%	78.68%	0.125	Moderate
Logistic Regression	65.75%	65.98%	0.039	Poor
Gaussian NB	59.60%	58.36%	0.001	Poor

5.2 Conclusion: The Best Model

Random Forest is identified as the best model for this dataset, achieving the highest accuracy (91.44%) and balanced precision/recall across all five classes.

Why it performs best:

1. Handling Complexity: It effectively captures non-linear relationships between water stress indicators without assuming linear separability.
2. Feature Utilization: It successfully exploited the engineered *CWSI* feature, using it as a primary splitting criterion.
3. Robustness: Unlike KNN, it is less sensitive to feature scaling and outliers. Unlike Naive Bayes, it does not assume feature independence, which is crucial given the high correlation between water risk metrics.

While KNN serves as a strong alternative (88.85%) after tuning, Random Forest's superior performance and ability to provide feature importance make it the optimal choice for Water Resource Risk Classification.