1. Feature Engineering

To enhance the predictive power of our models, we derived two new features based on hydrological indicators:

- **Composite Water Stress Index (CWSI):** Designed to represent the overall water pressure by aggregating weighted stress factors.
    - *Formula:* CWSI = 0.5 * bws_score + 0.3 * gtd_score + 0.2 * drr_score
- **Seasonal-Flood Interaction (SFI):** Captures the combined risk of high seasonal variability and river flood probability.
    - *Formula:* SFI = sev_score * rfr_score

2. Data Preprocessing & Baseline Models

Before training, we removed the gid_0 column as it is an identifier with no predictive value. Since distance-based algorithms (SVM, KNN) were used, we applied **StandardScaler** to normalize all features to a mean of 0 and variance of 1.

Baseline Performance Results

We trained five models with default parameters. The accuracy scores on the test set are as follows:

Model	**Baseline Accuracy**
Random Forest	**0.9188**
KNN	0.8101
SVM	0.7563
Logistic Regression	0.6872
Gaussian Naive Bayes	0.6092

Evaluation:

In the initial run, Random Forest demonstrated the highest accuracy. This suggests that the dataset contains non-linear patterns and complex feature interactions that tree-based models handle significantly better than linear models like Logistic Regression.

3. Hyperparameter Optimization

We utilized GridSearchCV with 5-fold cross-validation to optimize hyperparameters and ensure robustness.

Tuned Results & Best Parameters

Model	Tuned Test Accuracy	Improvement	Best Parameters
Random Forest	**0.9177**	*Stable*	max_depth: 20, n_estimators: 200
KNN	**0.8869**	**+7.6%**	n_neighbors: 9, weights: distance
SVM	**0.7816**	+2.5%	C: 10, kernel: rbf
Logistic Regression	0.6872	-	C: 1, solver: lbfgs
Gaussian Naive Bayes	0.6092	-	var_smoothing: 1e-09

Discussion on Tuning:

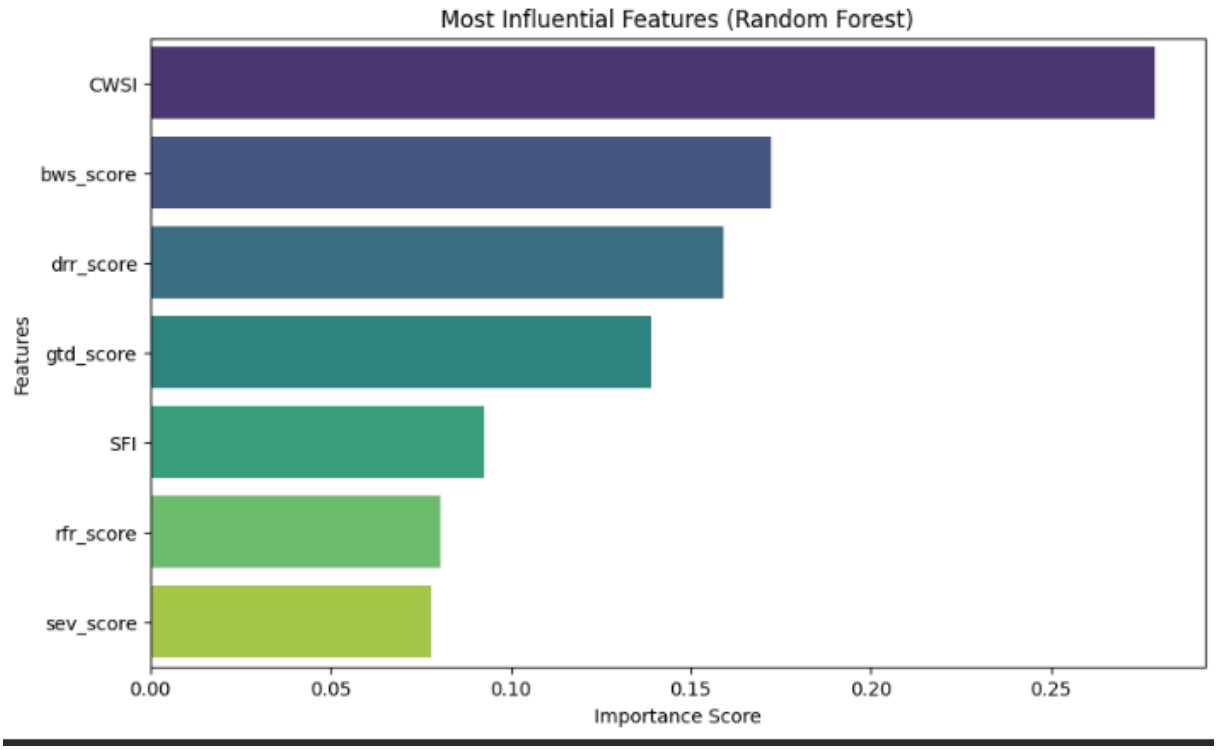
- **KNN Improvement:** Tuning the number of neighbors (n_neighbors=9) and using distance-based weighting provided a significant accuracy boost (+7.6%), making KNN a strong competitor.
- **SVM:** The selection of the rbf kernel over linear confirms that the decision boundaries between risk categories are non-linear.
- **Random Forest:** While the accuracy remained stable, tuning parameters like max_depth helped prevent potential overfitting.

4. Feature Importance Analysis

Using the optimized Random Forest model, we analyzed which features contributed most to the classification decisions.

Top Influential Features:

1. **CWSI (0.279)** – *Most Important*
2. bws_score (0.172)
3. drr_score (0.159)
4. gtd_score (0.139)
5. SFI (0.093)



Interpretation:

The best-performing Random Forest model was utilized to analyze feature importance scores. This analysis provided critical insights into which hydrological indicators drive the water risk classification.

Most Influential Features:

The analysis yielded a significant finding: our derived feature, CWSI (Composite Water Stress Index), emerged as the single most influential predictor, ranking 1st among all features. The dominance of CWSI confirms that creating a composite index—aggregating Baseline Water Stress, Groundwater Depletion, and Drought Risk—provides a much stronger and more holistic signal to the model than treating these features purely individually. Furthermore, the Seasonal-Flood Interaction (SFI) feature ranked 5th, validating its inclusion as it effectively captures the interplay between variability and flood risk.

Least Influential Feature:

Conversely, the analysis identified sev_score (Seasonal Variability) as the least influential feature. This low ranking is particularly insightful when contrasted with the high performance of the derived SFI feature (which is a product of sev_score and rfr_score). It suggests that seasonal variability alone is not a strong differentiator for water risk categories; however, when combined with river flood risk (as in SFI), it becomes highly significant. This further validates the feature engineering process, demonstrating that the model prioritizes the interaction of environmental factors over their standalone values.

5. Final Conclusion

Based on our comprehensive analysis:

1. **Best Model: Random Forest** is the superior model for this dataset with ~92% accuracy. Its ensemble nature allows it to capture complex, non-linear hydrological relationships effectively without heavy assumptions.
2. **Key Takeaway:** The significant performance jump in KNN after tuning and the dominance of the CWSI feature highlight the importance of both **hyperparameter optimization** and **domain-specific feature engineering** in machine learning pipelines.