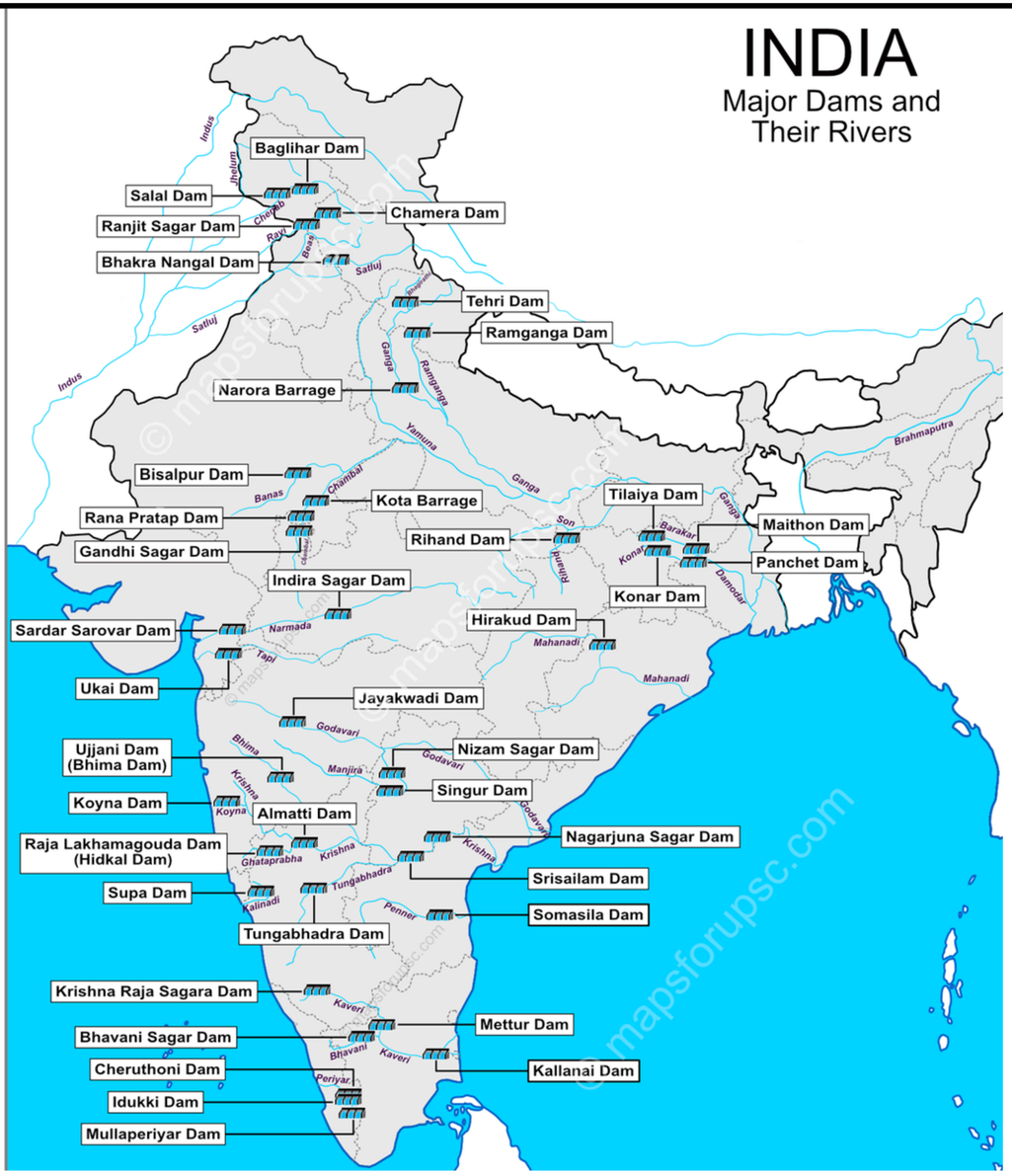
SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

AI-Driven Water Reservoir Analysis and Prediction



Managing water reservoirs efficiently is crucial for agriculture, energy, and sustainability.

Project Goal

Use machine learning to predict reservoir storage levels, efficiency, and risk of depletion based on historical and environmental data.

Motivation

Traditional monitoring is manual and reactive; data science enables proactive water management & forecasting. This system enables **data-driven, consistent, and fast decisions.**

Core Tools Used Python, Pandas, Scikit-learn, TensorFlow, and XGBoost.

Name: GAUTAM SUKHANI
PRN: 22070521059
SEM and SEC: SEM 7, SEC A

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Understanding the Data

The model was trained on Cleaned Water Reservoir—a structured dataset comprising approximately 55,000 records and 10 columns features related to records of multiple reservoirs with hydrological and environmental parameters.

Demographics

- Reservoir_Name, Year, Month, Storage_Capacity, Current_Storage
- Rainfall_mm, Inflow_cusecs

Target Attributes

- Outflow_cusecs, Evaporation_mm
- Storage_Efficiency_% (target variable)

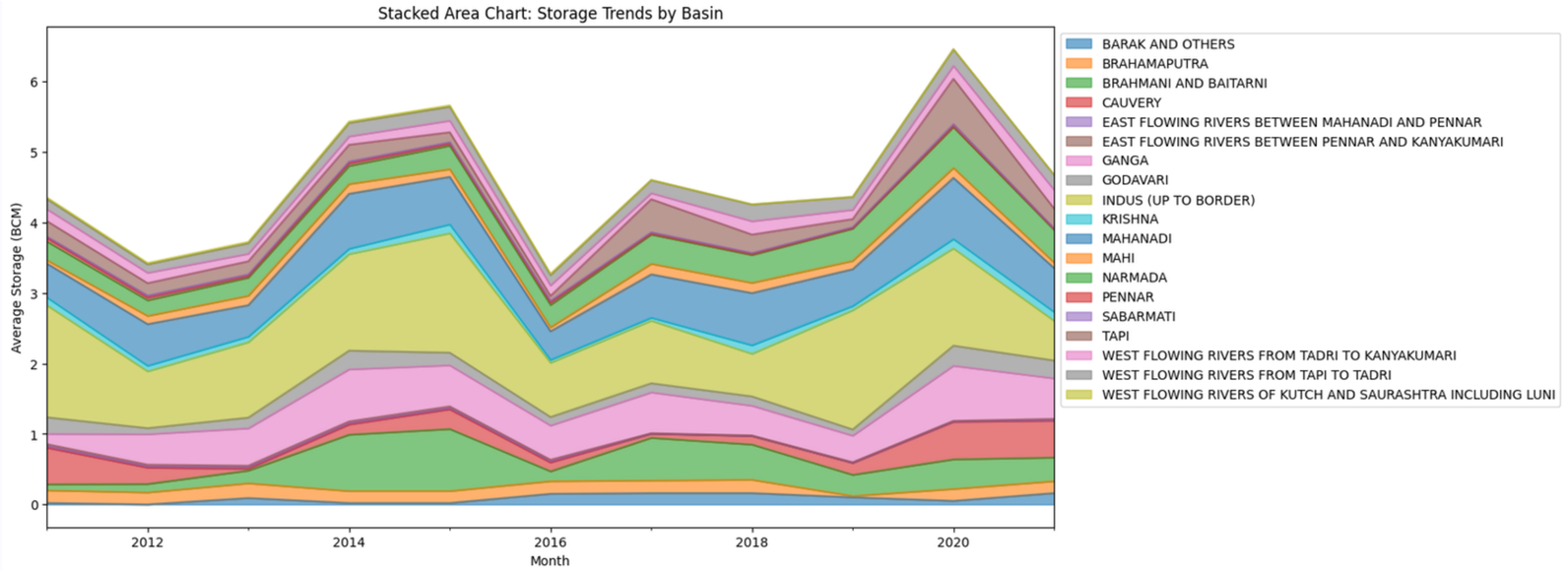
Data Preprocessing Steps

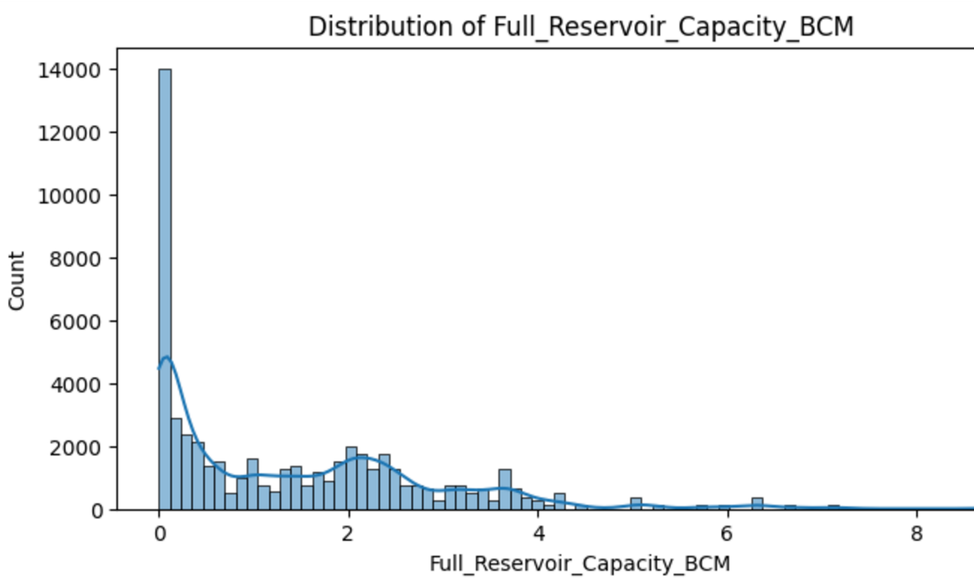
- Missing value treatment and outlier removal
- Feature scaling (MinMaxScaler)
- Encoded categorical fields like Reservoir_Name

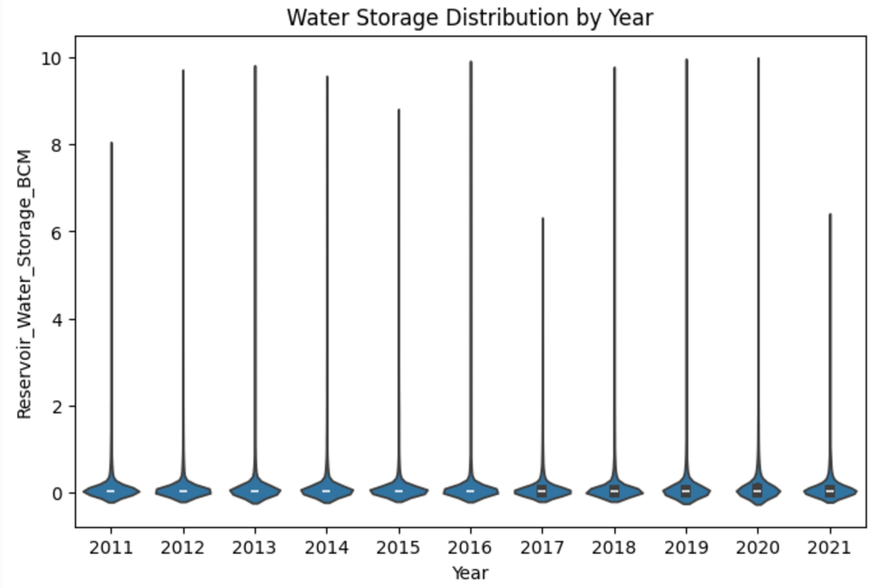
	Year	Month	Full_Reservoir_Capacity_BCM	Reservoir_Water_Level_M	Reservoir_Water_Storage_BCM
count	53625.000000	4719	53625.000000	53625.000000	53625.000000
mean	2015.720000	2016-04-30 15:16:21.818181632	36.026746	196.888303	0.195662
min	2011.000000	2011-05-01 00:00:00	0.000000	0.000000	0.000000
25%	2013.000000	2013-05-01 00:00:00	0.100000	139.190000	0.020000
50%	2016.000000	2016-05-01 00:00:00	1.785000	143.750000	0.020000
75%	2018.000000	2019-05-01 00:00:00	8.840000	151.180000	0.020000
max	2021.000000	2021-05-01 00:00:00	7414.290000	1002.790000	9.720000
std	3.013597	NaN	367.648471	172.001115	0.649604

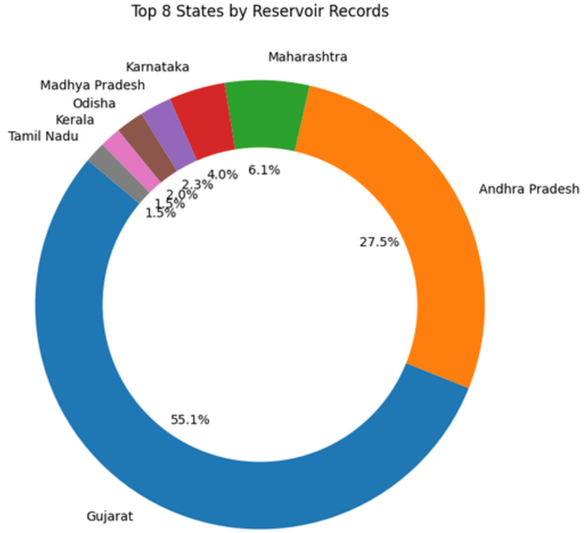
SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

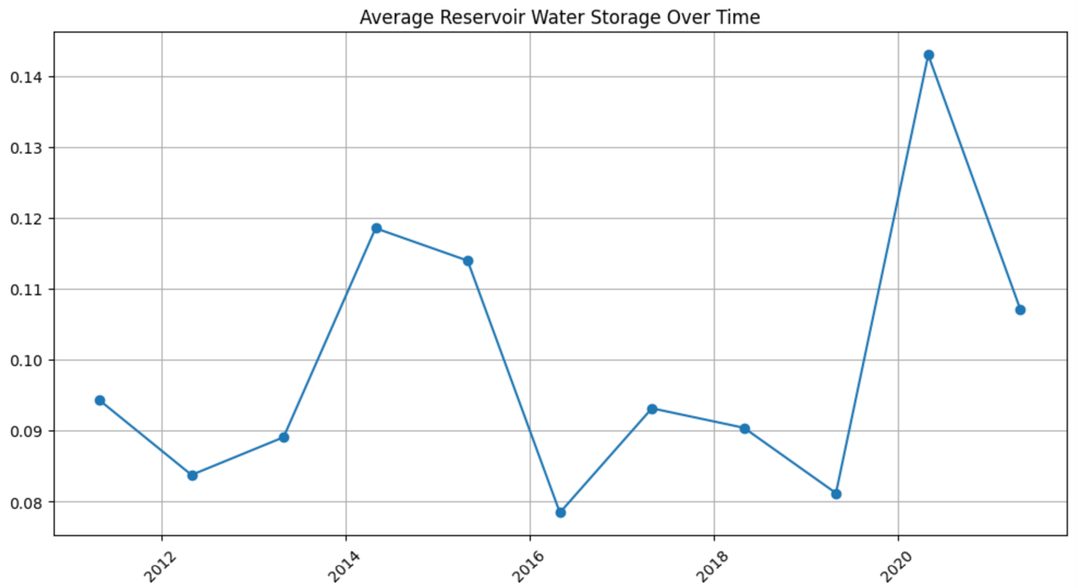
Exploratory Data Analysis











Linear Regression Model: Establishing the Baseline

A Linear Regression model was implemented as the initial baseline to predict the numeric Storage_Efficiency_%

0.74

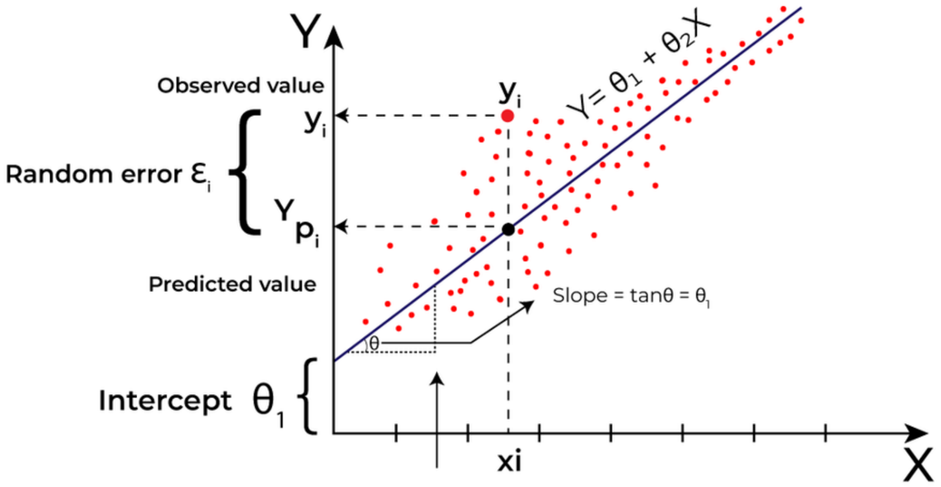
R-Squared

Explained variance

0.32

RMSE

Root Mean Square Error

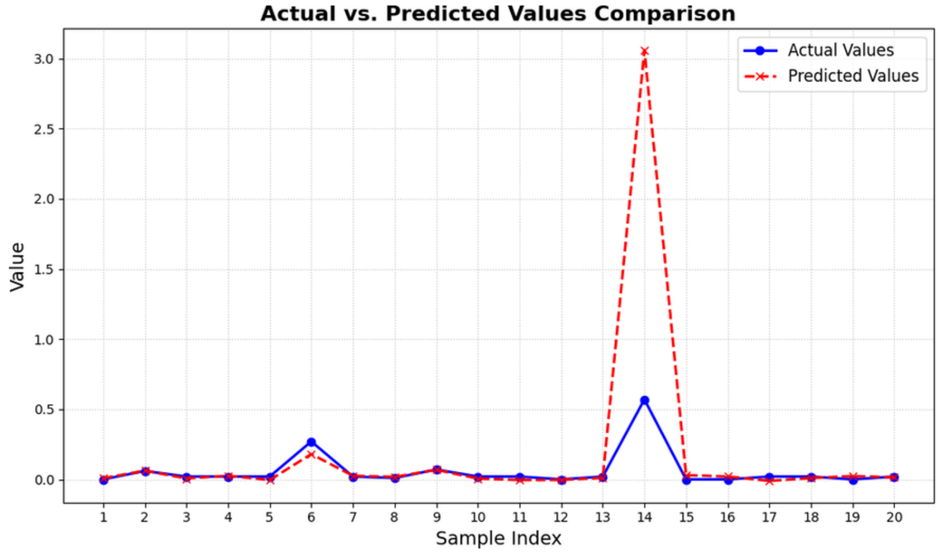


Observations

- The model effectively captured the primary linear trends within the data.
- Provided excellent interpretability, clearly showing the captured general trends well (rainfall → efficiency increase).

Conclusion

While a good baseline, the linear regression is limited in its ability to model the complex, non-linear interactions inherent in risk assessment. Further complexity is required. Struggled with non-linear rainfall-evaporation effects.

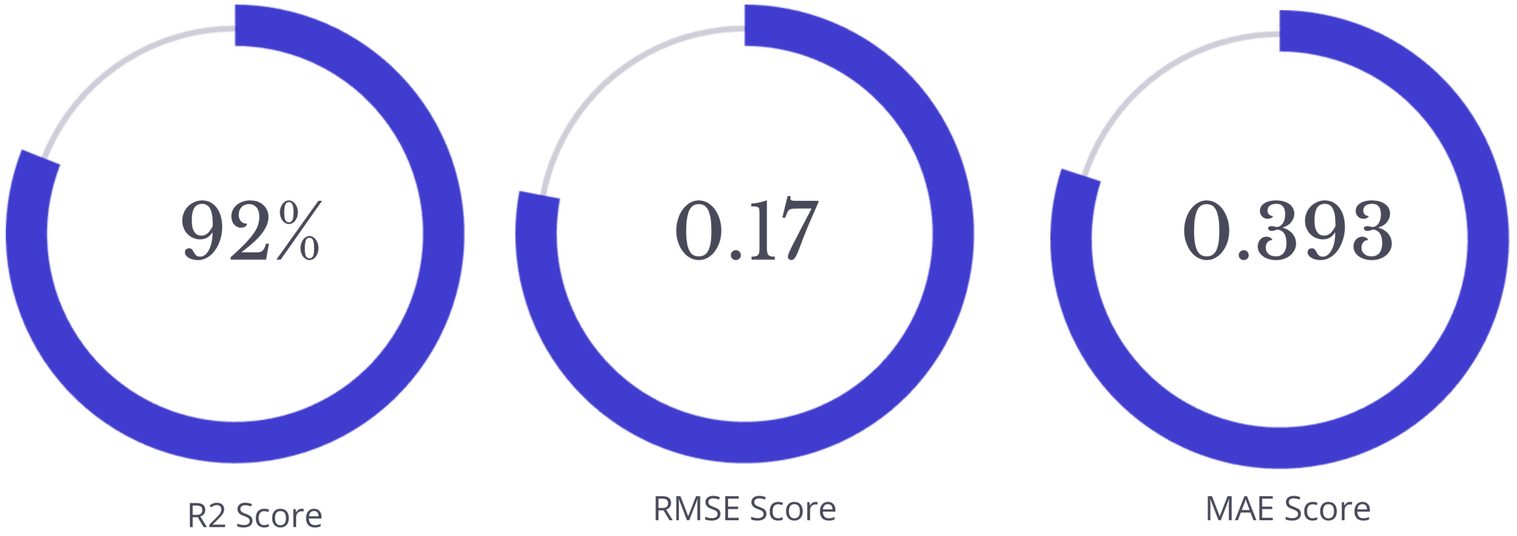


SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

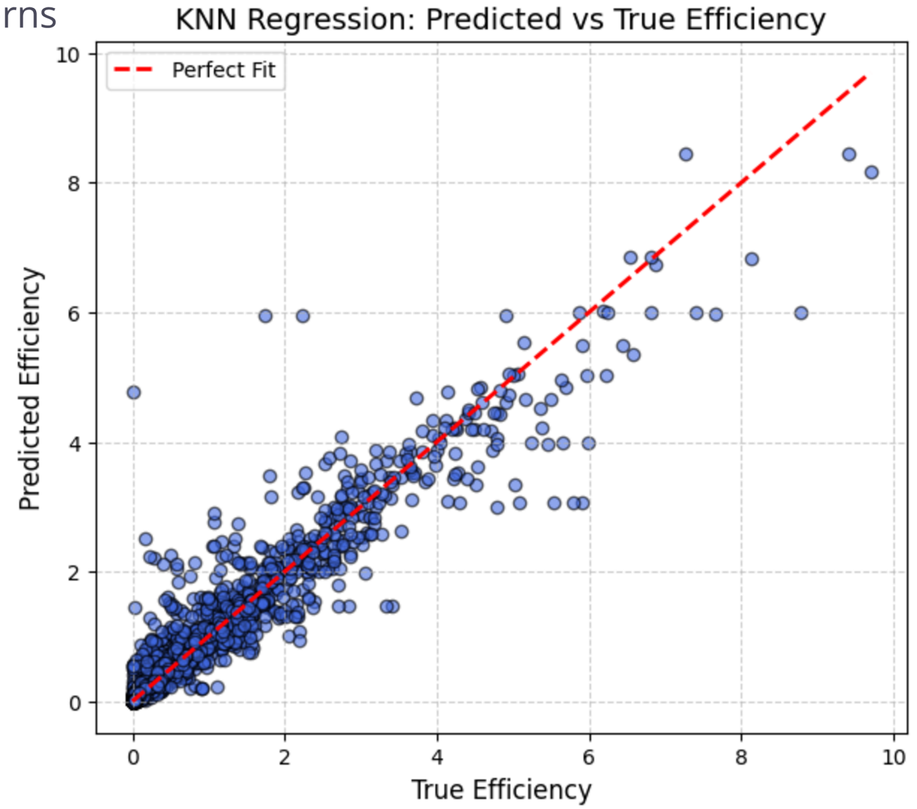
K-Nearest Neighbors (KNN) Classifier

KNN was employed as a classification model to predict efficiency by nearest historical patterns

The optimal value of K = 5 was determined through thorough cross-validation.



R2 Score | RMSE Score | MAE Score



Strengths

- Performs well and efficiently on structured, moderate-sized datasets.
- Performs well where reservoir behavior is regionally similar.

Limitations

- Computationally expensive during inference for very large datasets.
- Performance is highly sensitive to the initial scaling and normalization of features.

Conclusion: KNN provides a solid local classifier, effective for pattern matching, but its reliance on proximity limits global generalization and scalability.

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Random Forest Regressor

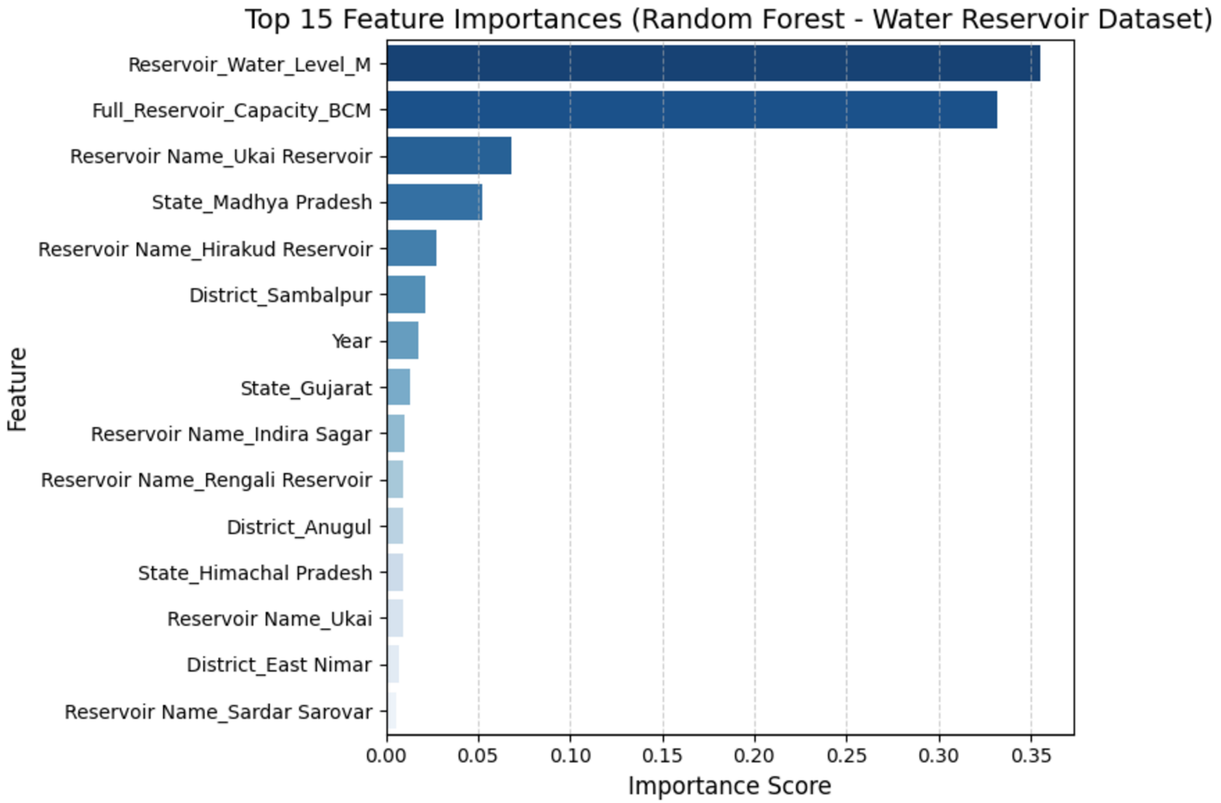
The Random Forest, an ensemble model utilizing multiple decision trees with bagging, was selected to capture complex feature interactions and mitigate overfitting.

R2 Score

98%

RMSE Score

0.15



Key Advantages

- Provides highly interpretable feature importance rankings.
- Demonstrates robustness against noise, outliers, and class imbalance.
- Strong generalization capabilities due to ensemble averaging.

Conclusion: Achieved high classification accuracy with a balanced generalization, making it the best candidate for stable production environments.

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

XGBoost Regressor: Optimized Performance

XGBoost (Extreme Gradient Boosting) was utilized for the continuous prediction of the boosted gradient model optimizing residuals iteratively. The model was rigorously tuned using grid search for hyperparameters like learning rate and max depth.

RMSE (Lower is Better)

R-Squared (Explained Variance)

Predictive Power

Demonstrated excellent predictive performance, surpassing the linear model.

Data Handling

Handles missing values internally, reducing the need for extensive manual imputation.

Limitation

Requires careful monitoring to mitigate the slight risk of overfitting on small or noisy datasets.

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

K-Means Clustering: Applicant Segmentation

K-Means, an unsupervised model, was applied to segment applicants into distinct risk groups, providing underwriters with clear, pre-grouped data.

The optimal number of clusters (**k=3**: High, Moderate, Low) was determined. Silhouette Score = 0.65

Cluster 1

High-efficiency reservoirs — high inflow and rainfall

Cluster 3

Low-efficiency — low inflow, high evaporation

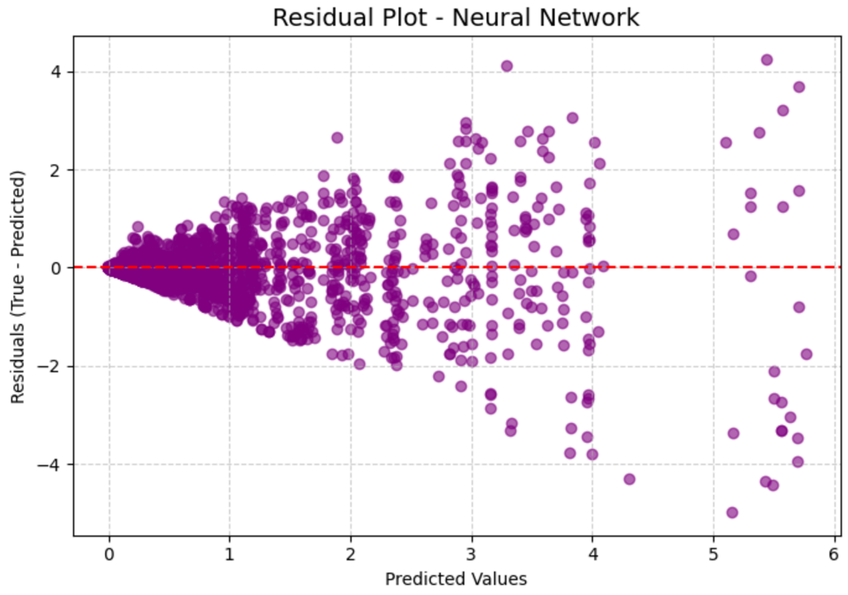
Cluster 2

Moderate efficiency — seasonal fluctuation

Conclusion: The clustering approach is invaluable for policymakers prioritize low-performing reservoirs..

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Neural Network: Exploring Deep Learning Potential



A Deep Learning model, built using TensorFlow/Keras with three hidden layers and ReLU activation,128-64-32 neurons. represents our most complex classification architecture. Optimizer: Adam; Loss: Mean Squared Error.

R2 Score
75%

RMSE Score
0.31

MAE Score
0.09

Core Strengths

- Exceptional ability to capture complex, non-linear feature relationships.
- Highly adaptable and scalable to future, massive data volumes.

Deployment Status

Achieved the highest overall classification accuracy. While requiring more computational resources for training and careful tuning, this architecture is ideal for high-performance production

SYMBIOSIS
INSTITUTE OF TECHNOLOGY, NAGPUR

Performance Comparison and Key Takeaways

Linear Regression	R^2 = 74%	Simple, interpretable	Misses non-linearity
KNN	R^2 = 92%	Easy to implement	Slow for large data
Random Forest	R^2 = 98%	Best performance	Slower training
XGBoost	R^2 = 93%	Balanced performance	Slight overfitting risk
K-Means	Silhouette = 0.65	Segmented Reserviors	Needs manual K selection
Neural Network	R^2 = 72%	Captures complexity	Heavy computation

Model Synthesis

- Random Forest achieved the best predictive accuracy.
- XGBoost provided strong interpretability and balanced performance.
- K-Means added useful grouping insights for management planning.

Efficiency Gains

AI-based models are projected to improve efficiency by **30–40%**.

Future Scope:

- Integrate satellite rainfall data.
- Deploy model as web dashboard.
- Add anomaly detection for real-time alerts.