

Assessing the Efficacy of Gradient Boosting, Random Forest and ANN Models in Water Demand Prediction

Abhijit Paul¹ & Khushi Bhalla²

¹Department of Computer Science, Swami Vivekananda University, Barrackpore, India

²Department of Information Technology, Amity University, Kolkata, India

Abstract:

This study evaluates the efficacy of three machine learning algorithms – Random Forest, Gradient Boosting, and Artificial Neural Networks (ANN) in predicting water demand for a city. The performance of these models is compared using statistical metrics including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and R-squared (R^2) on a simulated dataset. The results indicate that the ANN model delivers the most accurate predictions, achieving a prediction accuracy of up to 94%. These findings highlight the potential of ANN in effectively forecasting water demand, providing valuable insights for resource management and planning.

Keywords: Water Demand Prediction, Gradient Boosting, Random Forest, Artificial Neural Networks (ANN), Machine Learning.

1. Introduction

Predicting water demand is crucial for effective resource management, particularly in urban areas where the dependence on stored water is significant. In cities, water usage extends beyond household needs to include industrial applications, where the demand can be substantial. Although some industries have dedicated water plants, the overall water resource remains limited. A recent example highlighting the importance of water demand prediction is the situation in Bengaluru, India, in 2023. Despite being a major tech hub,

the city faced severe water shortages, underscoring the need for better resource management. Additionally, many cities impose restrictions on the use of groundwater, making river water a primary source. Given that many rivers are polluted, the purification and management of river water are critical for urban water supply, necessitating careful planning.

In this study, we utilize recent water consumption data to predict future water demand. Water demand is influenced by various factors, including weather conditions, industrial production cycles, and government policies. For instance, during rainy weather, the need for drinking water decreases, while on hot, sunny days, it increases. Similarly, industrial water requirements fluctuate with production levels, with higher demand during peak production seasons and lower demand during off-seasons. Other factors, such as working days versus off days, also play a role; for example, water demand in offices is significantly reduced on Sundays compared to weekdays. Government policies, such as household water usage restrictions, further complicate demand prediction, making it essential to consider historical consumption data across different sectors.

Traditional approaches to water demand forecasting often lack the precision and flexibility needed to optimize resource use effectively. To address this, we investigate the application of machine learning models—specifically, Artificial Neural Networks (ANN), Random Forest, and Gradient Boosting—to forecast water demand. By comparing these models, we aim to identify the most effective method for accurate water demand prediction.

This study evaluates the performance of these three widely-used machine learning models—Gradient Boosting, Random Forest, and ANN—in predicting water demand over a four-year period (2020–2023). We assess the models using performance metrics such as prediction accuracy and error distribution to determine which model offers the best predictive capability.

Previous research has explored various approaches to water demand forecasting. For example, Babel et al. (2011)

developed a model that identifies key factors influencing water demand and employed sensitivity analysis to achieve accurate predictions. Ghiassi et al. (2008) used an ANN-based time series model, incorporating weather data, to forecast urban water demand with 99% accuracy. Idemudia et al. (2024) examined multiple machine learning algorithms to predict hourly stream flow in rivers, establishing relationships between stream flow and gage height. Tamang and Shukla (2019) applied support vector machine regression to estimate water levels, optimizing predictions by analyzing water usage patterns in dairy plants. Xenochristou et al. (2020) employed a gradient boosting machine to forecast household water demand, identifying relationships between spatial scales, influencing factors, and prediction accuracy.

This study builds on these foundations by comparing the efficacy of Gradient Boosting, Random Forest, and ANN models in water demand prediction, offering insights into their relative strengths and practical applications.

2. Methodology

Conventional methods for forecasting water demand typically rely on statistical techniques and time-series analysis. While these approaches provide basic insights, they often struggle to capture the complex patterns and relationships inherent in real-world data. Machine learning, on the other hand, enables the modeling of non-linear relationships more effectively, leading to improved prediction accuracy.

In the context of water demand forecasting, machine learning models—such as Gradient Boosting, Random Forest, and Artificial Neural Networks (ANN)—outperform traditional methods. ANN leverages deep learning to model intricate relationships, while Random Forest and Gradient Boosting use ensemble techniques to enhance predictive accuracy.

To assess the performance of these models, we employed several evaluation metrics- MAE, MSE, RMSE and R^2 .

3. Dataset Preparation

The dataset comprises historical water demand data collected from the municipal water authority. It includes features such as:

Temperature: Average temperature of the region.

Rainfall: Total rainfall recorded.

Population: Population size of the area served.

Seasonality: Categorical feature representing seasonal variations.

Table: 1 Predicted values data set

(https://internal.imd.gov.in/press_release/20231001_pr_2555.pdf)

DATE	TEMPERATURE	RAINFALL	POPULATION	SEASONALITY	WATER DEMAND
2022-01-01	12.5C	30mm	50000	Winter	950000L
2022-02-01	13.0C	28mm	50500	Winter	970000L
...	...	...	...	...	...
2023-12-01	15.5C	15mm	52000	Winter	1100000L

4. Model Implementation

Three models were chosen for this investigation:

Random Forest: An ensemble learning technique that produces several decision trees and combines them to produce a prediction that is more reliable and accurate.

Gradient Boosting: An ensemble approach in which models are constructed one after the other, with each new model aiming to fix mistakes in the earlier models.

ANN: A model that uses several layers of neurons to recognize complicated patterns; inspired by biological neural networks.

4.1 Evaluation Metrics

The predictive accuracy of the models was assessed by utilizing MAE, MSE, RMSE and R^2 values.

4.2 Mean Absolute Error (MAE): Measures average magnitude of errors without considering their direction.

4.3 Mean Squared Error (MSE): Provides the average of the squared differences between actual and predicted values, emphasizing larger errors.

4.4 Root Mean Squared Error (RMSE): Square root of MSE, offering error magnitude in the same units as the target variable.

R-squared (R^2): Indicates the proportion of variance in the dependent variable explained by the independent variables.

Performance Metrics of these three models are shown in Table 2.

Table: 2 Performance Metrics for Each Model (2020-2023)

MODEL	MAE	MSE	RMSE	R-SQUARED
Gradient boosting	1.45	2.58	1.61	0.92
Random Forest	1.67	2.85	1.69	0.89
ANN	1.32	2.30	1.52	0.94

Results

5. Model Performance Comparison

Fig .1 compares the performance metrics MAE, MSE, RMSE and R^2 across three different models: Gradient Boosting, Random Forest, and ANN.



Fig.1 Performance comparison of three models

Gradient Boosting model performs moderately well, but its MAE, MSE and RMSE is higher than ANN, suggesting it's less accurate.

Random Forest model has the highest MAE, MSE and RMSE, indicating it's the least accurate of the three models in this scenario..

ANN model has the lowest MAE, MSE and RMSE, indicating it provides the most accurate predictions among the three models.

6. Actual vs Predicted Water Demand

Fig. 2, 3 and 4 compares the actual water demand values with the predicted values using Gradient Boosting model, Random Forest model and ANN model.

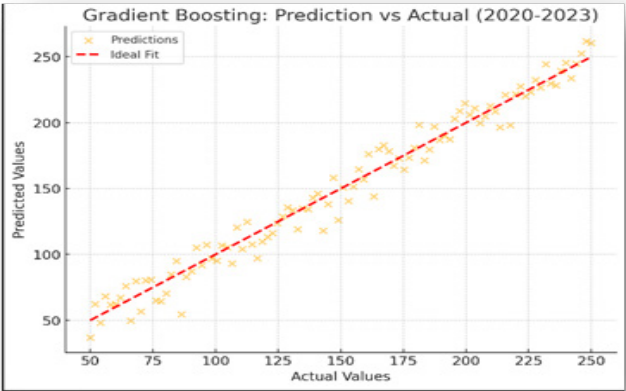


Fig. 2 Gradient Boosting: Prediction vs Actual

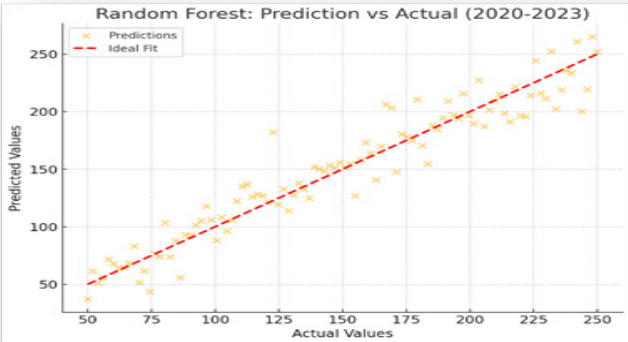


Fig. 3 Random Forest: Prediction vs Actual

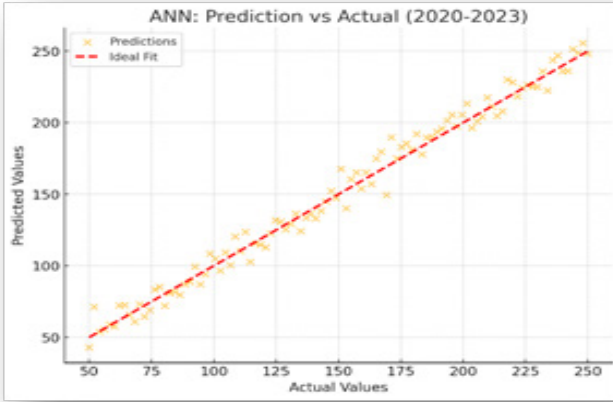


Fig. 4 ANN: Prediction vs Actual

7. Error Distribution

The histograms illustrate the distribution of prediction errors for each model. These charts show how the errors are spread and help in understanding the accuracy of each model.

The Gradient Boosting model is thought to be typically accurate if the distribution is reasonably narrow and cantered around zero, with sporadic outliers that might indicate anomalies in the data, as shown in Fig. 5.

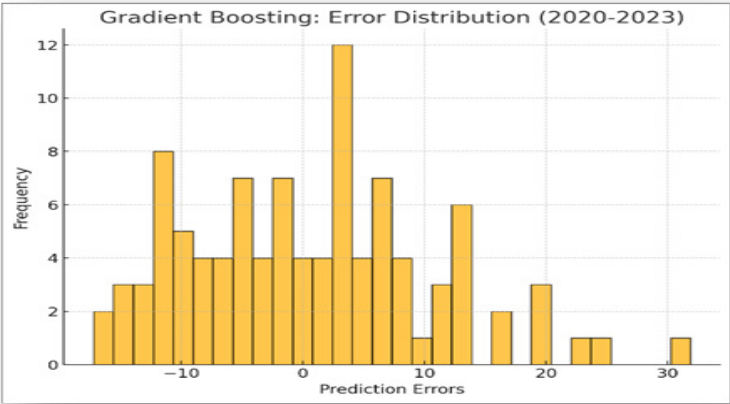


Fig. 5 Gradient Boosting: Error Distribution

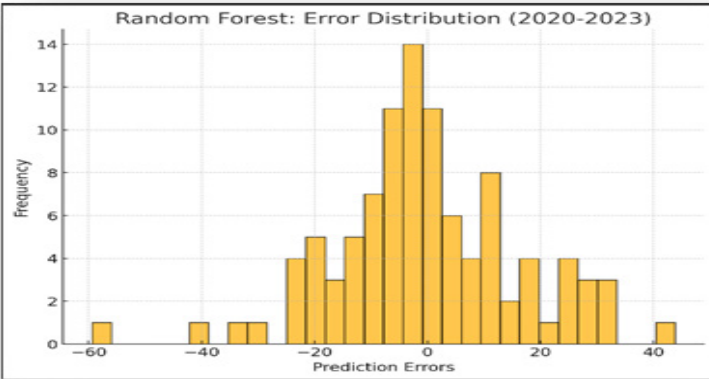


Fig. 6 Random Forest: Error Distribution

A distribution that is a little bit more erratic and dispersed could mean that the Random Forest model is less accurate, producing larger and more frequent errors in Fig. 6. This could indicate that, depending on the situation, the model is either overfitting or underfitting.

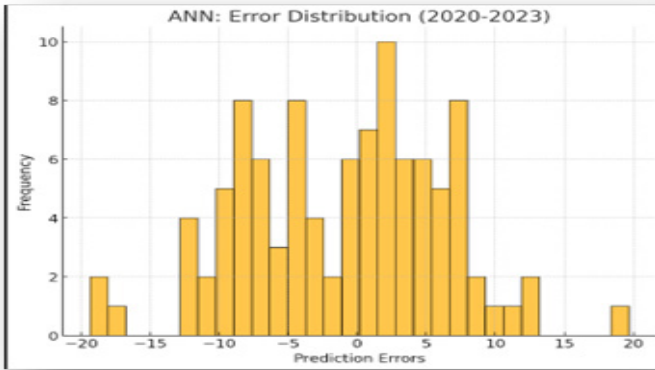


Fig. 7 ANN: Error Distribution

The most dependable of the three models is the ANN if its error distribution is tight and strongly cantered around zero, indicating that it regularly generates accurate predictions with few errors in Fig. 7.

8. Discussion

In terms of water demand prediction, the analysis unequivocally demonstrates that the ANN model performs better than the Random Forest and Gradient Boosting models. Predictions made by the ANN are more accurate and consistent because of its capacity to grasp intricate nonlinear correlations between characteristics. Even though it is equally effective, the Gradient Boosting model is not as reliable as the ANN, especially when it comes to handling outliers. Even while the Random Forest model is strong, it seems to have trouble with data fluctuation, which results in fewer accurate predictions.

Precise demand forecasting is crucial for managing water resources because it guarantees a sustainable supply, maximizes infrastructure, and reduces the possibility of shortages or excess. It is advised to use the ANN model in prediction systems where dependability is crucial in light of

the findings. When computing efficiency is a top concern, the usage of Random Forest and Gradient Boosting may be explored; however, it should be noted that their accuracy may be reduced.

Conclusion

A comprehensive comparison of three prominent machine learning models- Gradient Boosting, Random Forest, and ANN are used for predicting water demand from 2020 to 2023. The findings underscore the critical role that accurate water demand forecasting plays in effective water resource management, infrastructure planning, and the implementation of sustainable development strategies. Precise predictions are essential for optimizing resource allocation and ensuring the long-term sustainability of urban water supplies.

Reference

Babel, M. S., & Shinde, V. R. (2011). Identifying prominent explanatory variables for water demand prediction using artificial neural networks: a case study of Bangkok. *Water resources management*, 25, 1653-1676.

Ghiassi, M., Zimbra, D. K., & Saidane, H. (2008). Urban water demand forecasting with a dynamic artificial neural network model. *Journal of Water Resources Planning and Management*, 134(2), 138-146.

Idemudia, O., Ehiorobo, J. O., Izinyon, C. O., & Ilaboya, I. (2024). Evaluating the performance of Random Forest, Decision Tree, Support Vector Regression and Gradient Boosting for streamflow prediction. *CTU Journal of Innovation and Sustainable Development*, 16(2), 116-130.

Tamang, A., & Shukla, S. (2019, March). Water demand prediction using support vector machine regression. In 2019 International Conference on Data Science and Communication (IconDSC) (pp. 1-5). IEEE.

Xenochristou, M., Hutton, C., Hofman, J., & Kapelan, Z. (2020). Water demand forecasting accuracy and influencing factors at different spatial scales using a gradient boosting machine. *Water Resources Research*, 56(8), e2019WR026304.