Wastewater discharges impact on Environment

**Agricultural Wastewater discharge’s impact on sea:**

As per correlation matrix, the correlation was found between agricultural wastewater discharges and total discharges to the sea. Hence, the regression visualisation indicates this correlation. The positive slope suggests that as the discharge of the volume of agricultural wastewater increases, the total discharges to the sea increases respectively

Applied progression formula (Agricultural Wastewater Discharges = 0.1682 × Total Discharges to the Sea (million m³) + 25.2254) indicates that for every additional million cubic meters of agricultural wastewater discharges, it increases total discharge to the sea by an average of 0.1682 (million m³).

The visualization provides a scatter plot of the actual data points (blue) and the predicted regression line (red). The shaded area represents the 95% confidence interval for the regression predictions, providing a range where we expect the actual value to lie with 95% confidence (figure 14).

A graph with red and blue dots

Description automatically generated

Figure 14

The model’s coefficient of determination suggests that approximately 38.82% of the variability in agricultural wastewater discharges can be explained by the total discharges to the sea. The confidence intervals are relatively tight around the regression line, indicating a certain degree of precision in the model’s predictions.

**Agricultural Wastewater discharge’s impact on inland waters:**

Opposite to the impact on the sea, it was found a negative corelation to between agricultural wastewater discharges and total discharges to the inland waters. This implies that an increase in inland discharges is associated with a slight decrease in agricultural wastewater discharges. This relationship could be influenced by various factors, including the management practices for agricultural and industrial wastewater or the geographical distribution of agricultural areas relative to inland water bodies.

Applied progression formula: Agricultural wastewater discharges = −0.01035 × Total discharges to Inland waters (million m³) + 130.1775 Agricultural wastewater discharges = −0.01035 × Total discharges to Inland waters (million m³) + 130.1775 indicates that for every million cubic meters increase in inland discharges, the volume of agricultural wastewater discharges decreases by approximately 0.01035 million cubic meters.

These models quantify the association between the total discharges and agricultural wastewater discharges, with positive association for sea discharges and a slight negative association for inland discharges (figure 15)

A graph with blue dots and red lines

Description automatically generated

Figure 15

**Urban Wastewater discharge’s impact on sea:**

Applied progression model (Urban Wastewater Discharges = 0.03337 × Total Discharges to the Sea (million m³) + 74.2727) suggests that for every million cubic meters increase urban wastewater discharges, the volume of total discharges to sea increases by approximately 0.03337 million cubic meters.

The coefficient of determination, R², for this regression is approximately 0.2462, which indicates that around 24.62% of the variability in urban wastewater discharges can be explained by the total discharge to the sea. It's a modest fit, indicating that while there is some relationship, a significant portion of the variability is not captured by this model, and another factor may be influencing urban wastewater discharges (figure 16).

A graph with a red line and blue dots

Description automatically generated

Figure 16

**Industrial Wastewater discharge’s impact on inland waters:**

The regression analysis demonstrates a positive relationship between the volume of industrial wastewater discharges and the volume of inland water discharges. The model's moderate R² value indicates a substantial association, yet it also suggests that other factors play a role in determining industrial wastewater discharges that the model does not capture.

Applied progression formula: Industrial Wastewater Discharges = 0.2957 × Total Discharges to Inland Waters (million m³) − 79.3546 where coefficient 0.2957 suggests that for each million cubic meters increase in the industrial wastewater discharges, inland water discharges increase by approximately 0.2957 million cubic meters. The R² value is 0.4916, indicating that about 49.16% of the variability in industrial wastewater discharges is explained by the model, showing a moderate fit (figure 17)

Figure 17A graph with a red line and blue dots

Description automatically generated

Random Forest Regression

The Random Forest regression model was used to predict deaths due to unsafe WASH practices, based on various environmental and infrastructural predictors. This evaluation typically includes metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to gauge how well the model predicts continuous numerical outcomes. Here are the key outcomes were calculated:

* Mean Squared Error (MSE): The model achieved an MSE of approximately 0.014. While the MSE is lower than what we observed with the linear regression model, it still represents the average of the squares of the errors—the difference between the observed values and the values predicted by the model.

* R-squared (R²): The R² value is approximately -10.72. Like the linear regression model, a negative R² value indicates that the model does not fit the data well and performs worse than a simple mean-based model.

The visualization (figure 18) of feature importance from the Random Forest model reveals which predictors have the most influence on predicting deaths due to unsafe WASH practices. The importance values help identify which environmental and infrastructural features contribute most to the model's predictions.

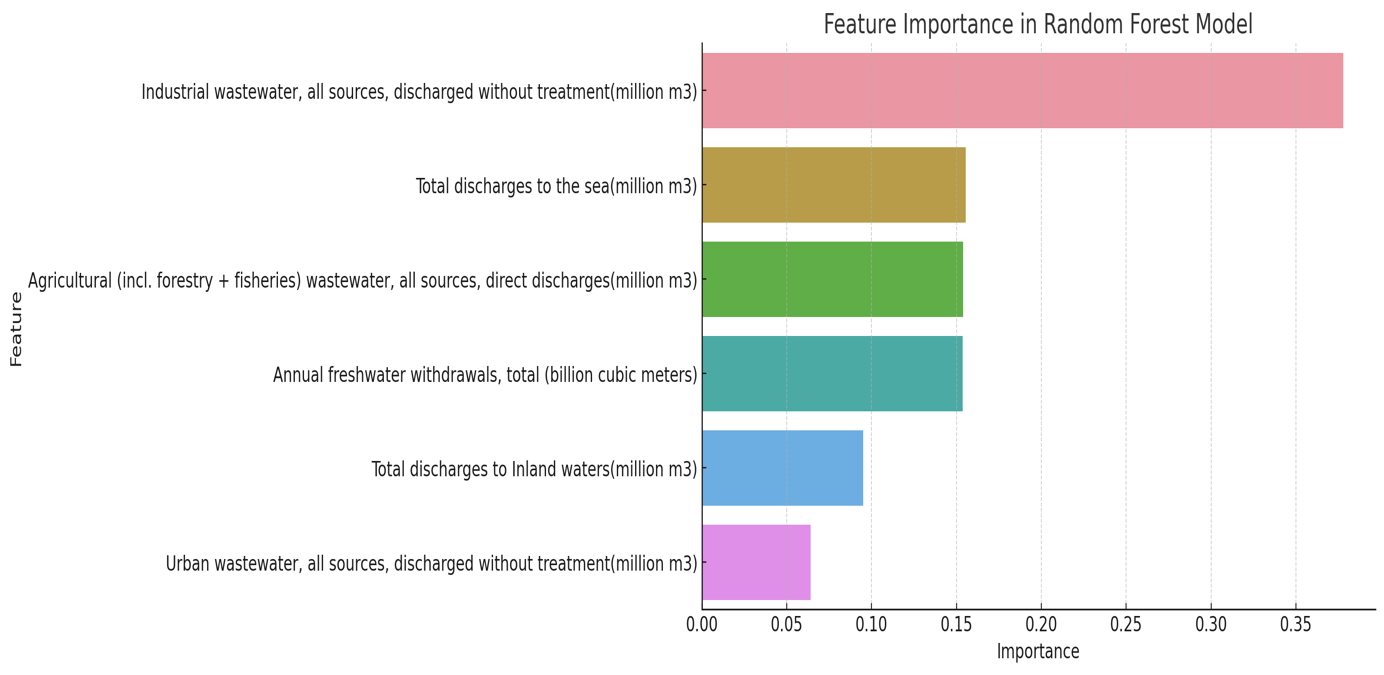


        Figure 18

However, the negative R² value suggests that the Random Forest model, despite its ability to capture non-linear relationships and interactions between predictors, still does not perform well on this dataset. This could be due to underlying complexities in the data that are not captured by the current set of predictors or due to overfitting to the training data. Moreover, the model's poor performance suggests caution in interpreting these importance values as definitive indicators of causal relationships.

Therefore, linear and Random Forest regression analyses struggled to accurately predict deaths due to unsafe WASH practices based on the selected predictors. This highlights the complexity of the relationship between WASH practices and health outcomes, which may be influenced by a wide range of socioeconomic, environmental, and behavioural factors not fully captured in the current dataset (figure 19).

