Water Usage and Conservation Analysis

College Name: Symbiosis Institute Of Technology

Student Names:	Krutika Gadge	(22070521178)
	Poornima Mendhekar	(22070521162)
	Mansee Dakhole	(22070521164)

Problem Statement

Water scarcity is an escalating global issue exacerbated by population growth, climate change, and inefficient usage. Addressing this requires advanced tools and insights to optimize resource management and promote sustainable practices. This study aims to provide data-driven solutions to this pressing problem.

Brief Overview:

This project focuses on analyzing global water consumption patterns to predict future trends and provide actionable insights for optimizing water usage and improving conservation efforts.

Key Objectives:

- Analyze historical water usage data to understand consumption patterns.
- Predict future water usage using machine learning (ML) models.
- Propose actionable strategies to promote water conservation and efficiency.

Dataset Overview

Dataset Description:

- Source: Global Water Usage Statistics dataset from Kaggle.
- Size:179 rows and 4 columns.

Key Features:

- Country: The name of the country.
- Yearly Water Used (m³): Total water consumption per year.
- Daily Water Used Per Capita (liters): Average daily water usage per person.
- Population: Total population of the country.

Data Preprocessing:

- Removed commas from numeric columns for consistency.
- Converted columns to appropriate numeric data types.
- Engineered new features such as "Yearly Water Per Capita" for deeper analysis.

Visual Analysis:

- Histogram: Displays the distribution of yearly water usage across countries.
- Boxplot: Highlights outliers in daily water usage per capita, aiding in identifying extreme usage patterns.

Methodology

Approach:

Steps Taken:

- **Data Preprocessing:** Handled missing values and cleaned the dataset for consistency.
- **Exploration:** Analyzed trends, distributions, and correlations using visual tools.
- **Feature Engineering:** Created derived metrics like "Yearly Water Per Capita."
- **Encoding:** Encoded categorical variables for machine learning compatibility.
- **Model Building**: Implemented Linear Regression for baseline predictions.
- Used **Random Forest** to enhance predictive performance.
- **Evaluation:** Assessed models using Mean Squared Error (MSE) and R^2 metrics.

Algorithms Used:

Machine Learning Algorithms:

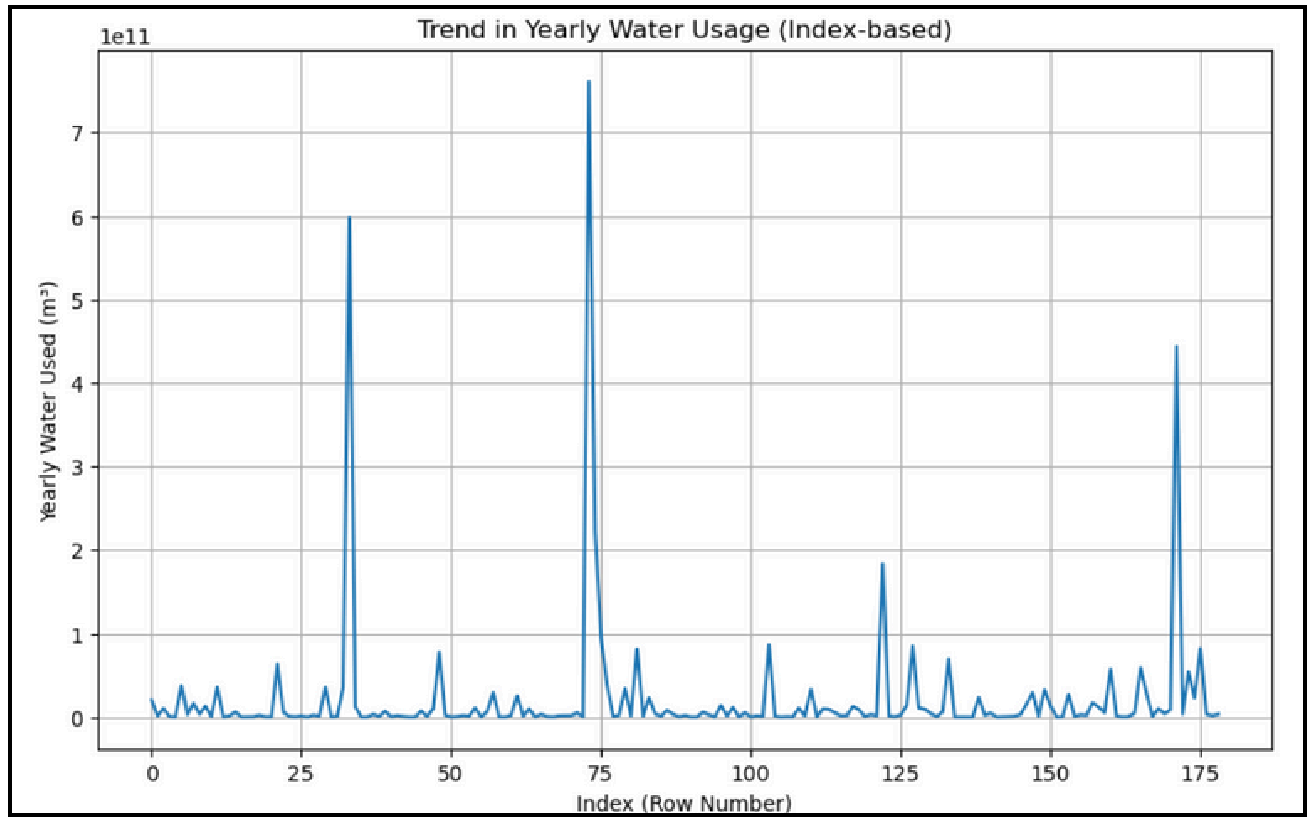
1. **Linear Regression**:A simple, interpretable model to establish a baseline.
2. **Random Forest :**Captures complex relationships and handles non-linearity effectively.

Reasons for Selection:

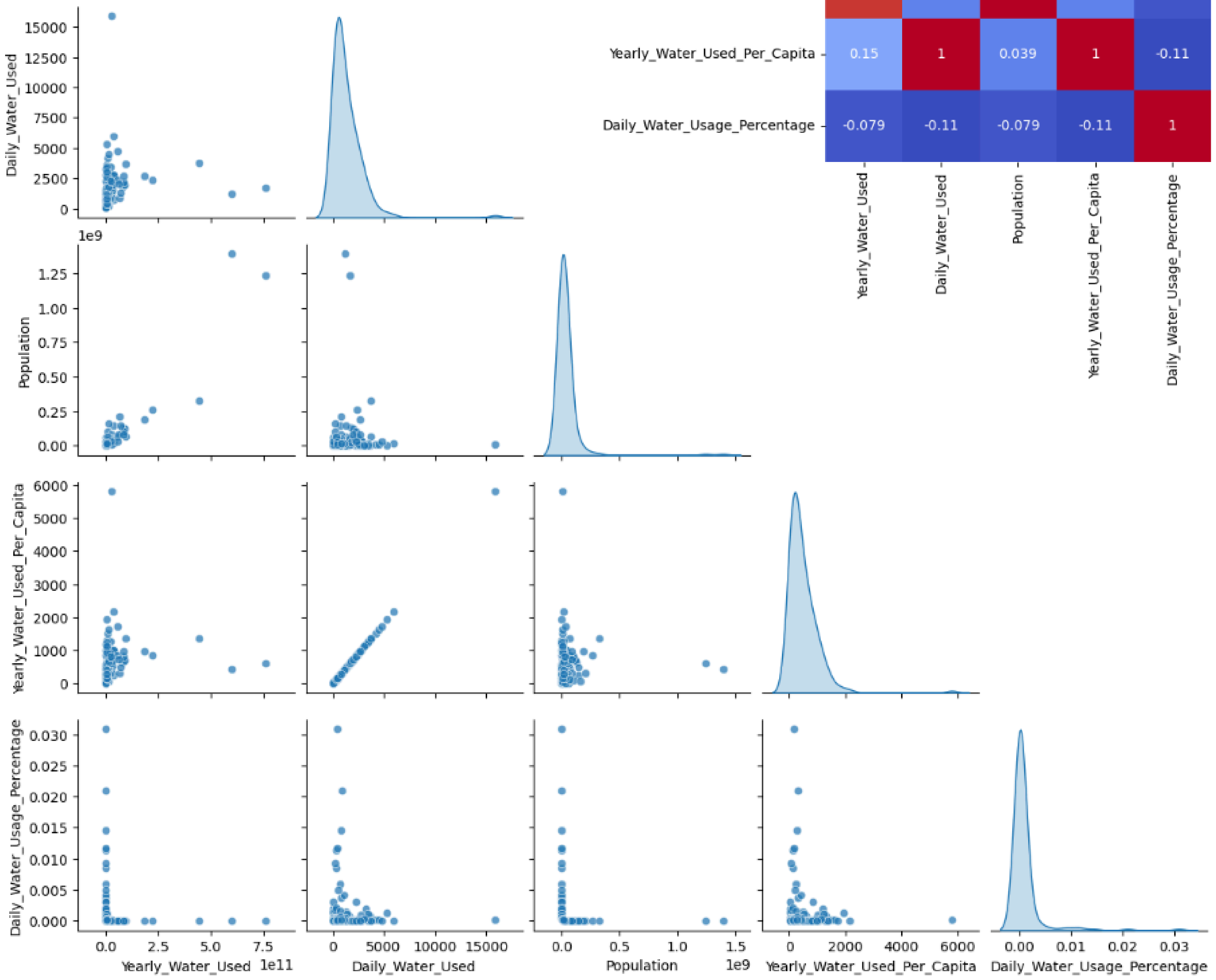
- Linear Regression provides quick insights and is easy to interpret.
- Random Forest improves accuracy by considering multiple decision trees and addressing overfitting.

Data visualization:

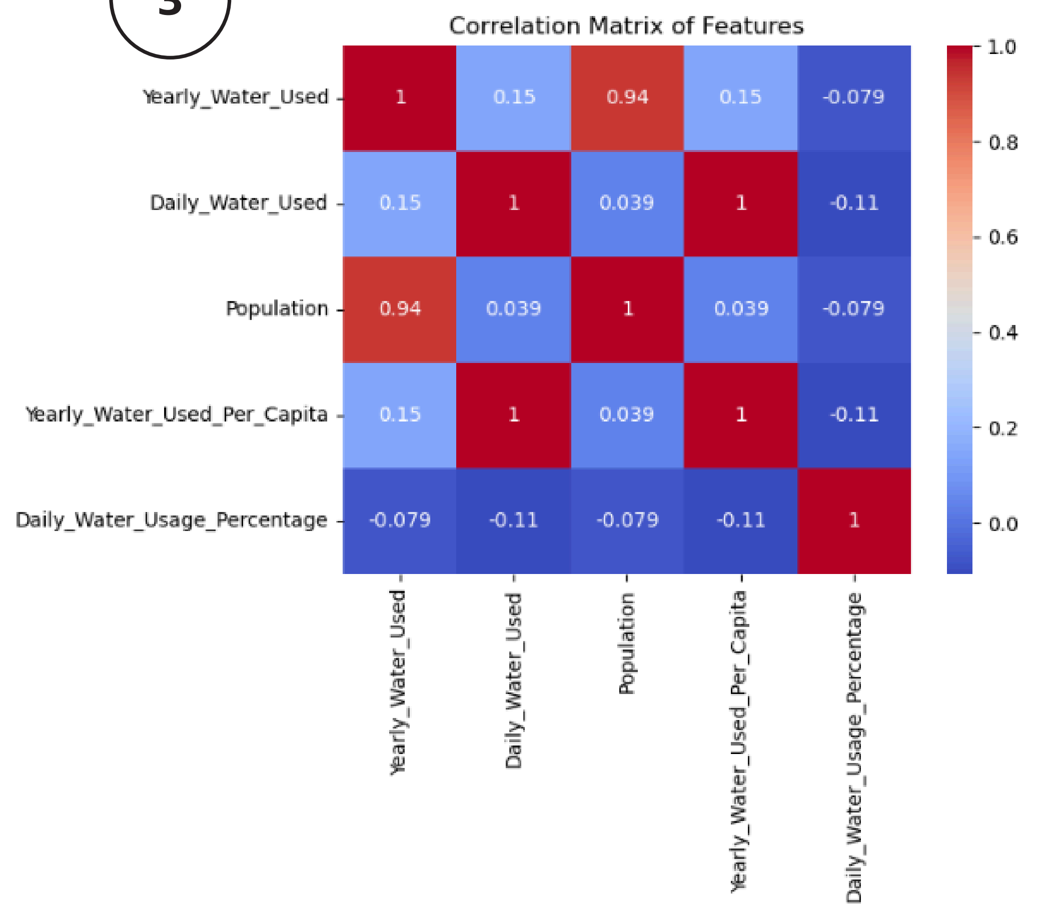
1



2



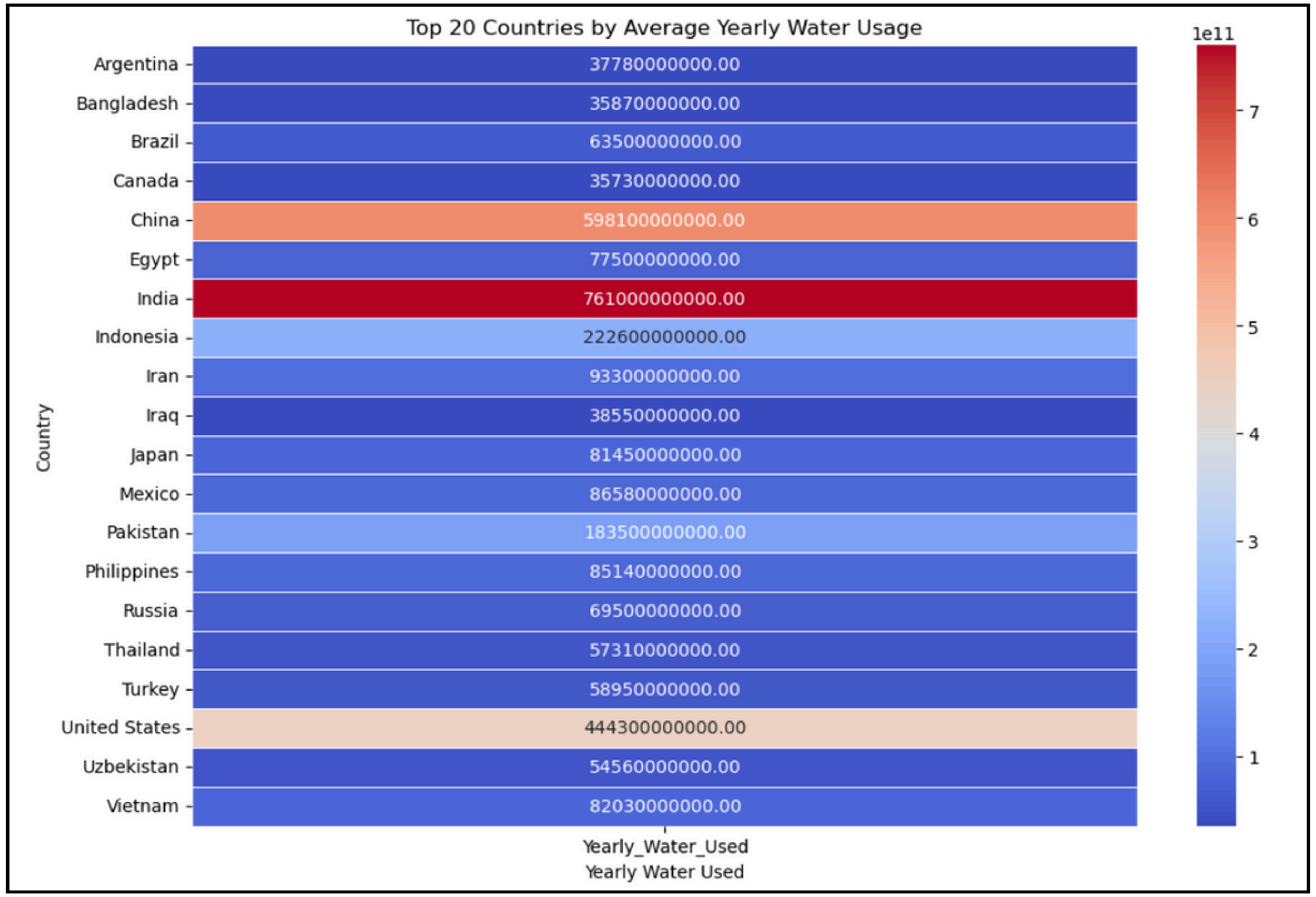
3



Correlation Matrix of Features

	Yearly_Water_Used	Daily_Water_Used	Population	Yearly_Water_Used_Per_Capita	Daily_Water_Usage_Percentage
Yearly_Water_Used	1	0.15	0.94	0.15	-0.079
Daily_Water_Used	0.15	1	0.039	1	-0.11
Population	0.94	0.039	1	0.039	-0.079
Yearly_Water_Used_Per_Capita	0.15	1	0.039	1	-0.11
Daily_Water_Usage_Percentage	-0.079	-0.11	-0.079	-0.11	1

4

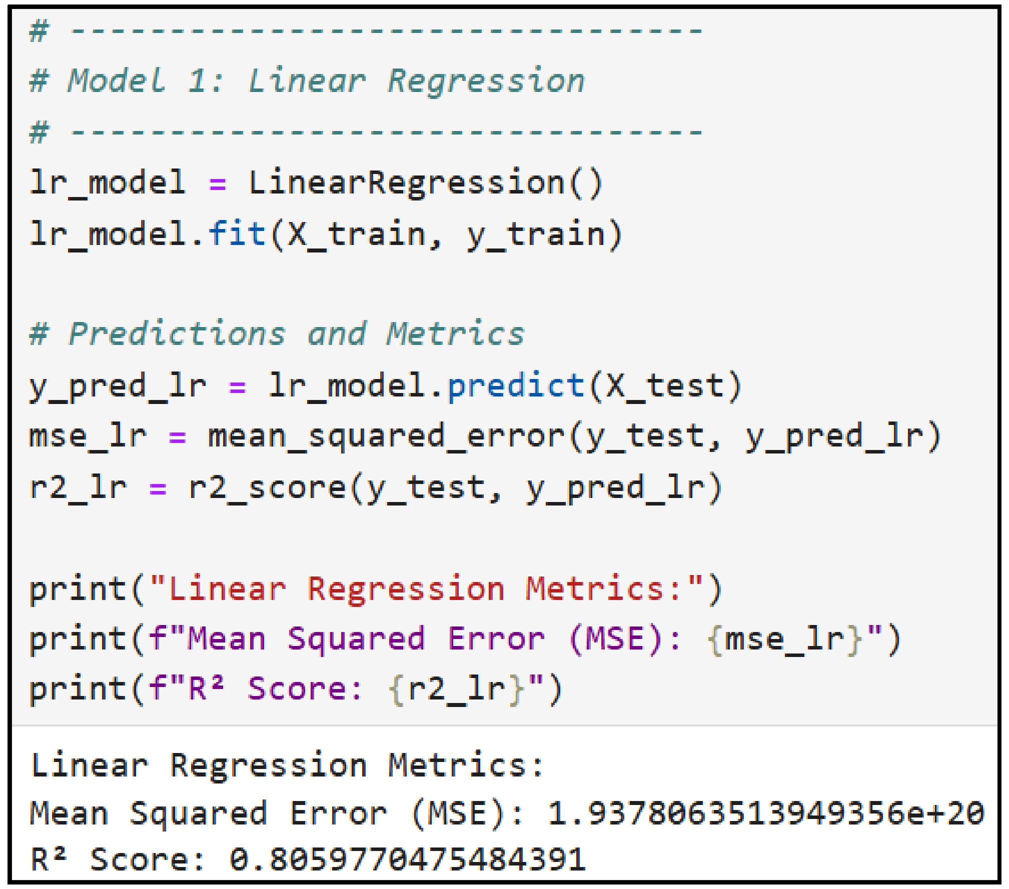


Top 20 Countries by Average Yearly Water Usage

Country	Yearly_Water_Used
Argentina	37780000000.00
Bangladesh	35870000000.00
Brazil	63500000000.00
Canada	35730000000.00
China	598100000000.00
Egypt	77500000000.00
India	761000000000.00
Indonesia	222600000000.00
Iran	93300000000.00
Iraq	38550000000.00
Japan	81450000000.00
Mexico	86580000000.00
Pakistan	183500000000.00
Philippines	85140000000.00
Russia	69500000000.00
Thailand	57310000000.00
Turkey	58950000000.00
United States	444300000000.00
Uzbekistan	54560000000.00
Vietnam	82030000000.00

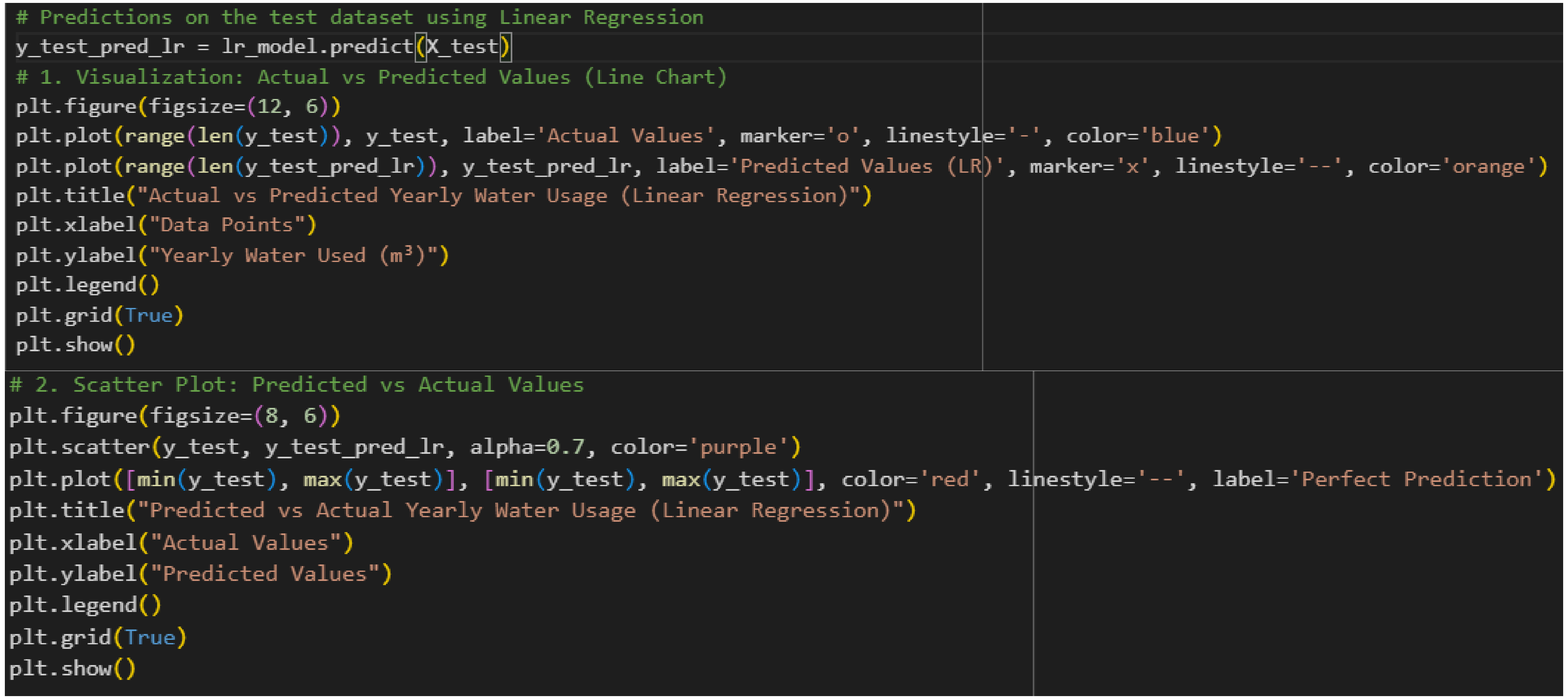
Yearly Water Used

MODEL BUILDING



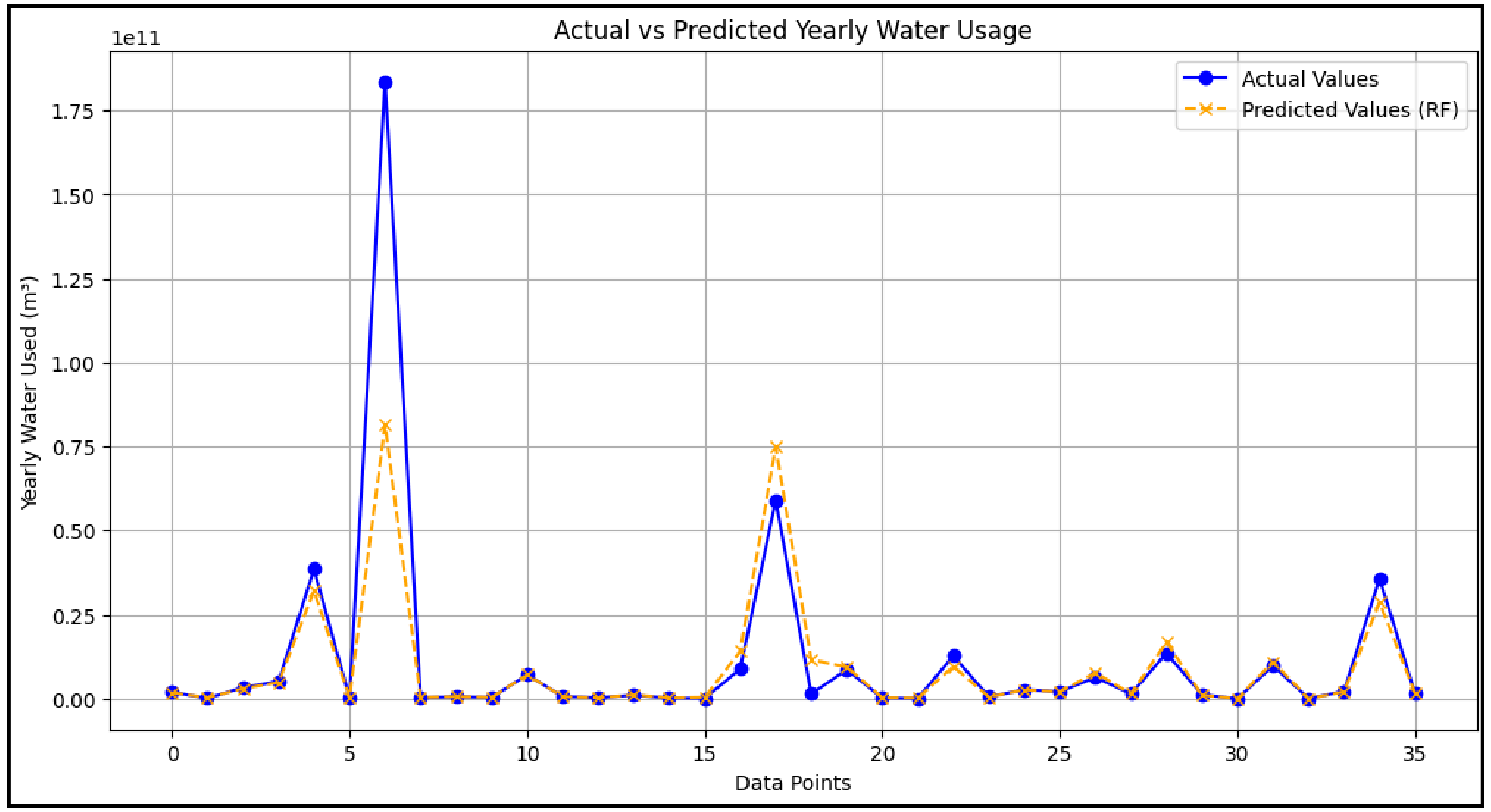
```
# --------------------------------
# Model 1: Linear Regression
# --------------------------------
lr_model = LinearRegression()
lr_model.fit(X_train, y_train)

# Predictions and Metrics
y_pred_lr = lr_model.predict(X_test)
mse_lr = mean_squared_error(y_test, y_pred_lr)
r2_lr = r2_score(y_test, y_pred_lr)

print("Linear Regression Metrics:")
print(f"Mean Squared Error (MSE): {mse_lr}")
print(f"R² Score: {r2_lr}")
```

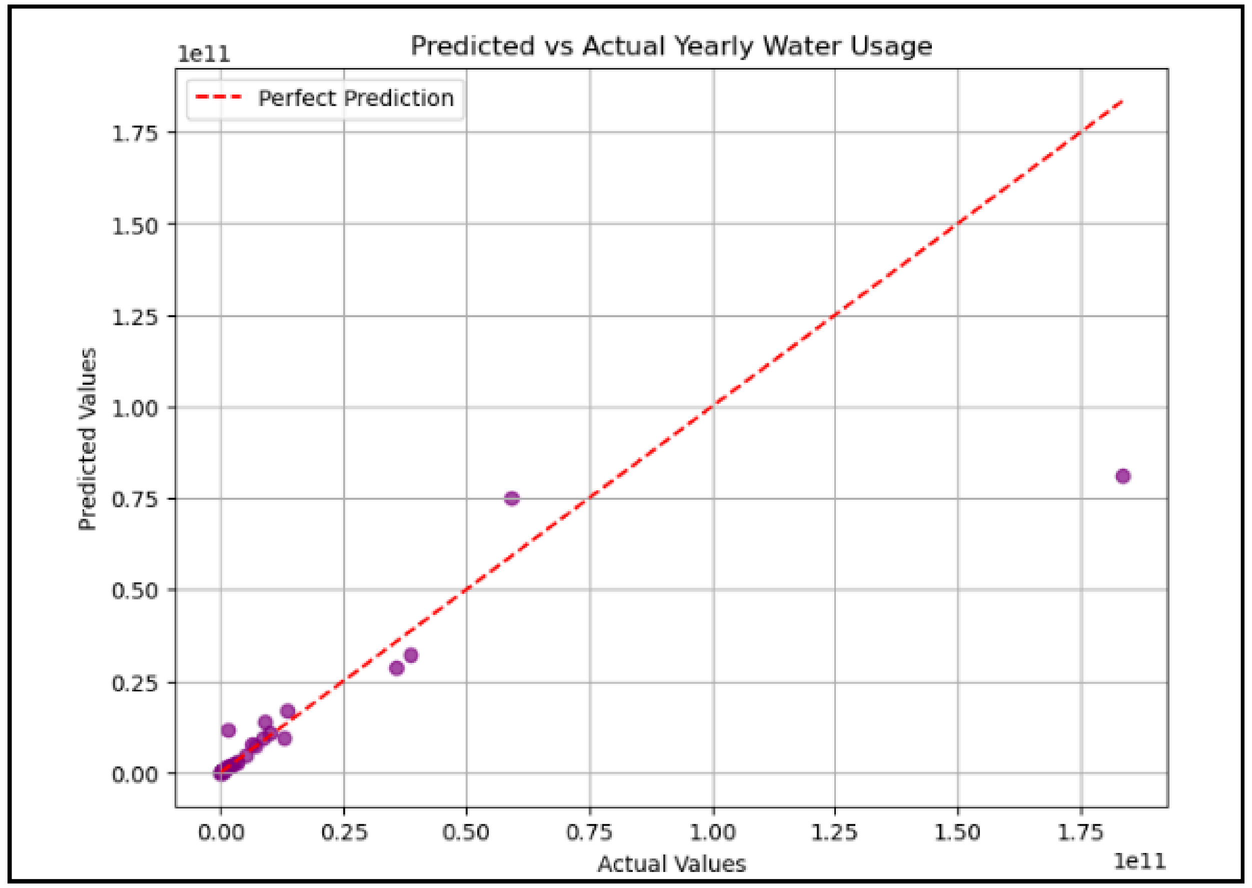
```
Linear Regression Metrics:
Mean Squared Error (MSE): 1.9378063513949356e+20
R² Score: 0.8059770475484391
```



```
# Predictions on the test dataset using Linear Regression
y_test_pred_lr = lr_model.predict(X_test)
# 1. Visualization: Actual vs Predicted Values (Line Chart)
plt.figure(figsize=(12, 6))
plt.plot(range(len(y_test)), y_test, label='Actual Values', marker='o', linestyle='-', color='blue')
plt.plot(range(len(y_test_pred_lr)), y_test_pred_lr, label='Predicted Values (LR)', marker='x', linestyle='--', color='orange')
plt.title("Actual vs Predicted Yearly Water Usage (Linear Regression)")
plt.xlabel("Data Points")
plt.ylabel("Yearly Water Used (m³)")
plt.legend()
plt.grid(True)
plt.show()
```

```
# 2. Scatter Plot: Predicted vs Actual Values
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_test_pred_lr, alpha=0.7, color='purple')
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color='red', linestyle='--', label='Perfect Prediction')
plt.title("Predicted vs Actual Yearly Water Usage (Linear Regression)")
plt.xlabel("Actual Values")
plt.ylabel("Predicted Values")
plt.legend()
plt.grid(True)
plt.show()
```





Github Link:

https://github.com/Krutikagadge/Water-Usage-and-Conservation-Analysis

Conclusion

Summary:

- Historical patterns reveal significant disparities in water usage across countries.
- Machine learning models, particularly Random Forest, achieved a prediction accuracy of [insert R^2 score].
- Recommendations include optimizing water distribution and reducing wastage.

Future Work:

- Incorporate additional variables like rainfall patterns, industrial water usage, and climate data.
- Experiment with advanced time-series models like Long Short-Term Memory (LSTM) for better predictions.

References

1. Kaggle Dataset: Dataset Link
2. Documentation: Scikit-learn.
3. Github Link: https://github.com/Krutikagadge/Water-Usage-and-Conservation-Analysis

Thank You