WASTE SENSOR DATA ANALYSIS

Overview

Introduction to the Dataset

•The dataset contains waste sensor readings collected over time.

•Each record includes sensor ID, timestamp, and properties of detected waste.

Purpose

•To analyze waste characteristics and predict waste types.

•To identify patterns and trends in waste sensor data.

Dataset Details

Dataset Overview

•**Rows:** 20,000 **Columns:** 7

Key Features

•**sensor_id:** Unique identifier for each sensor.

•**timestamp:** Date and time of each sensor reading.

•**waste_type:** Type of waste detected (recyclable, organic, non-recyclable).

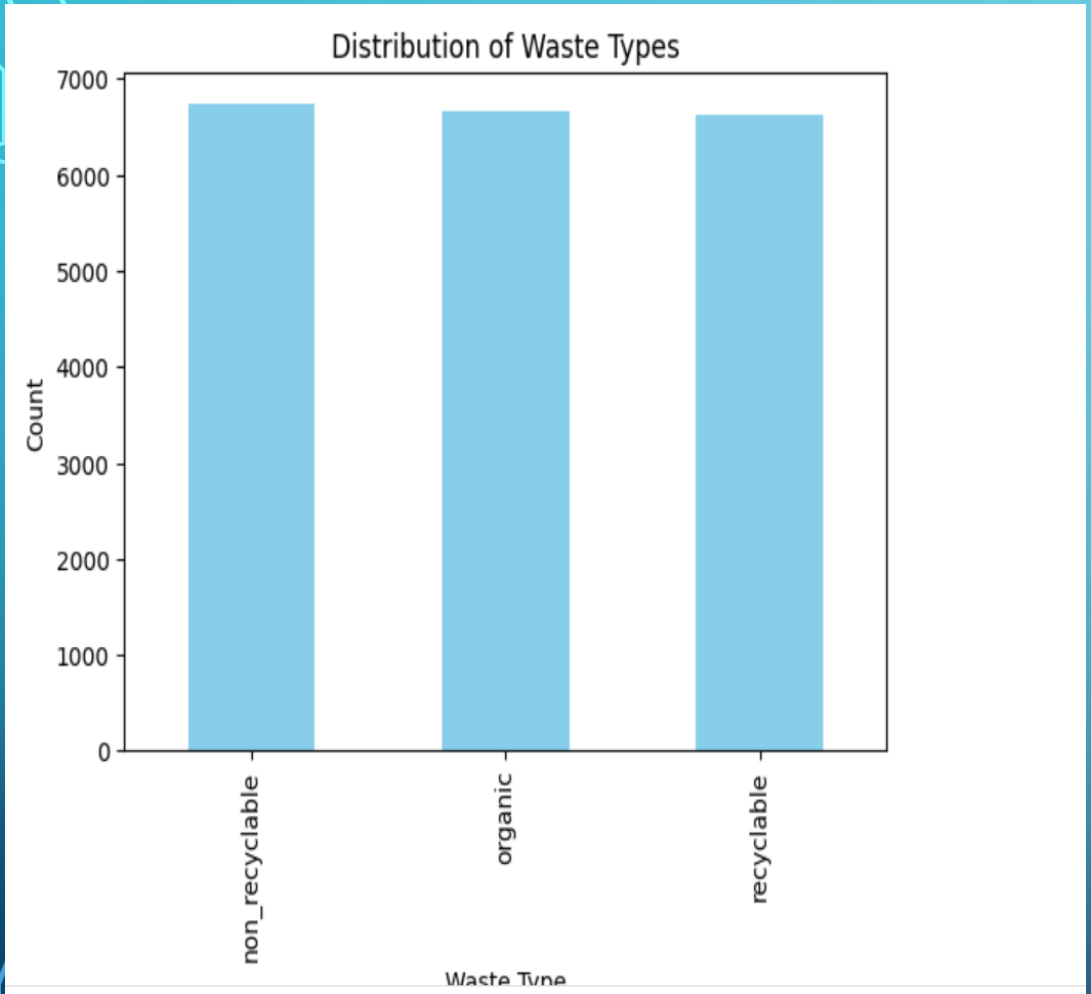
•**inductive_property:** Numerical measurement of the inductive property.

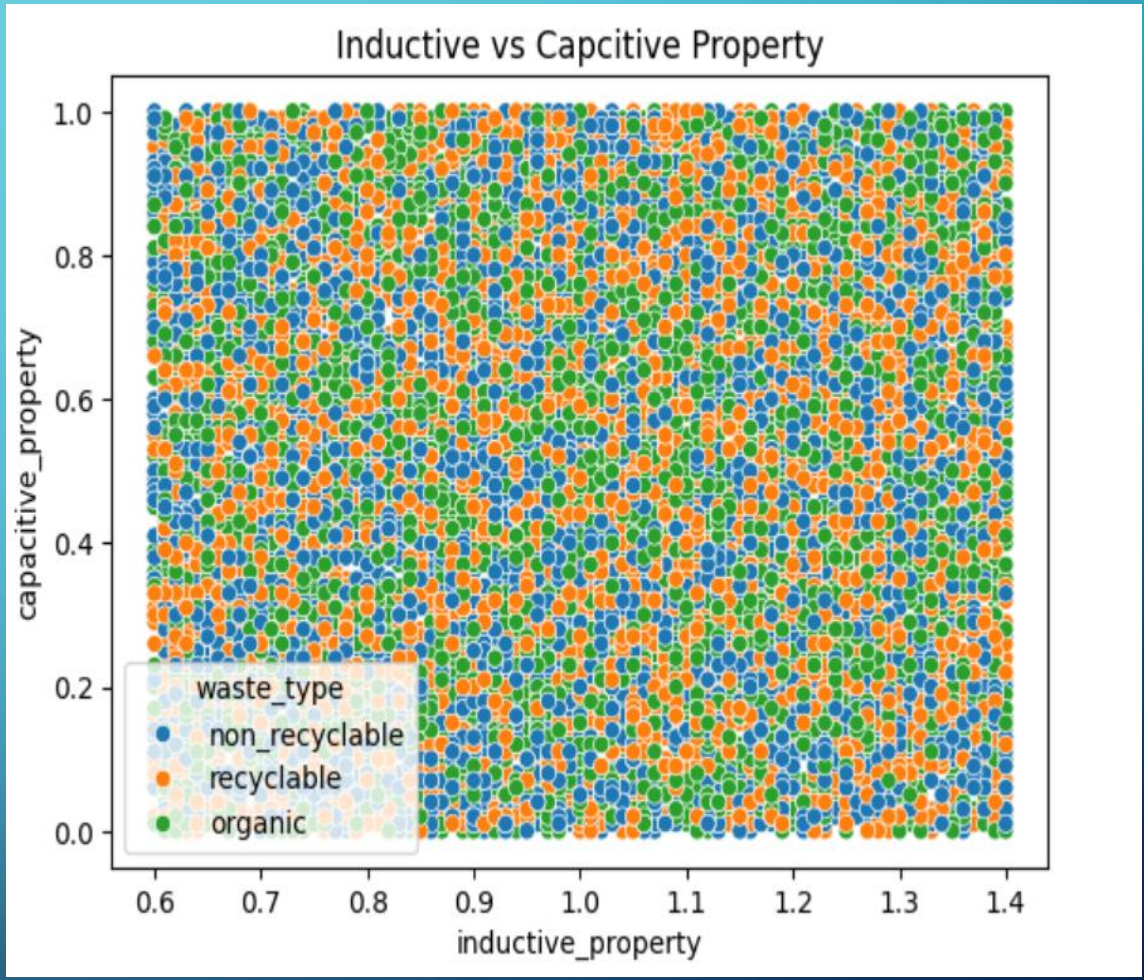
•**capacitive_property:** Numerical measurement of the capacitive property.

•**moisture_property:** Numerical measurement of moisture content.

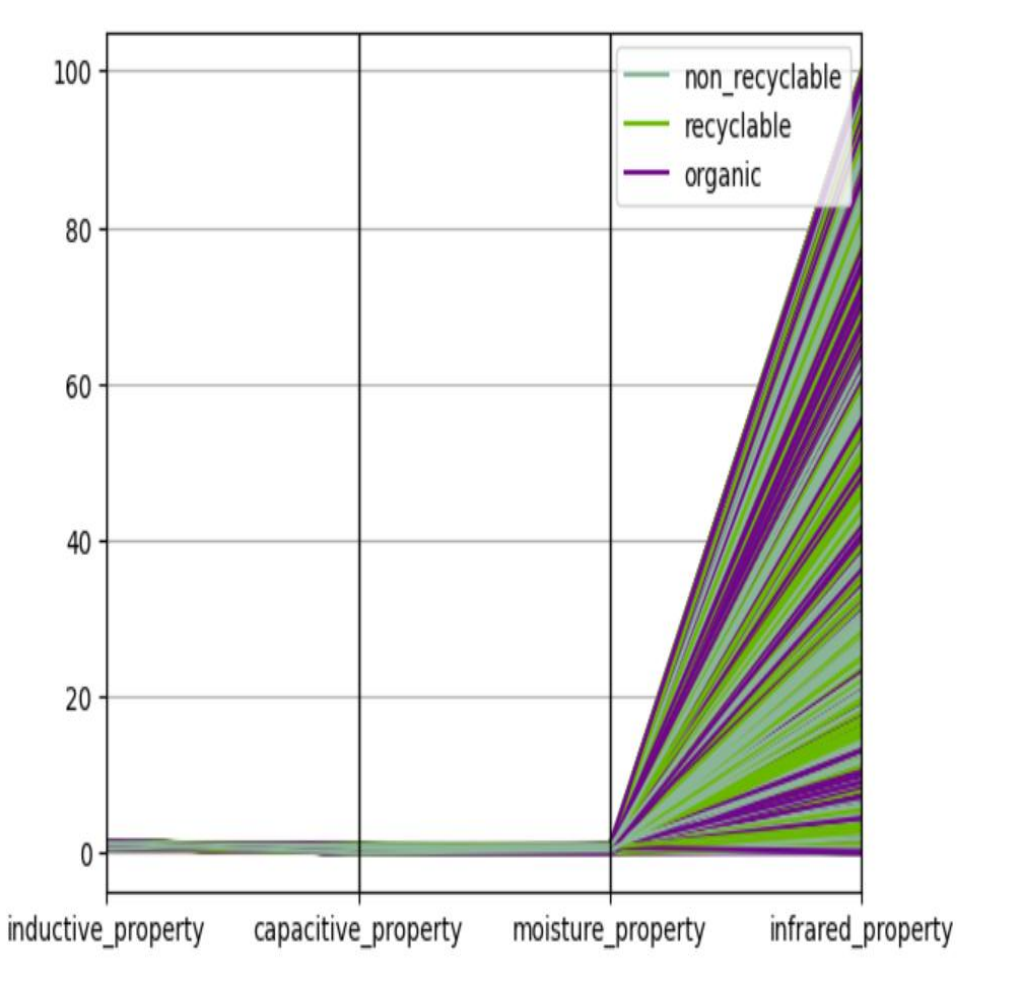
•**infrared_property:** Numerical measurement of the infrared property.

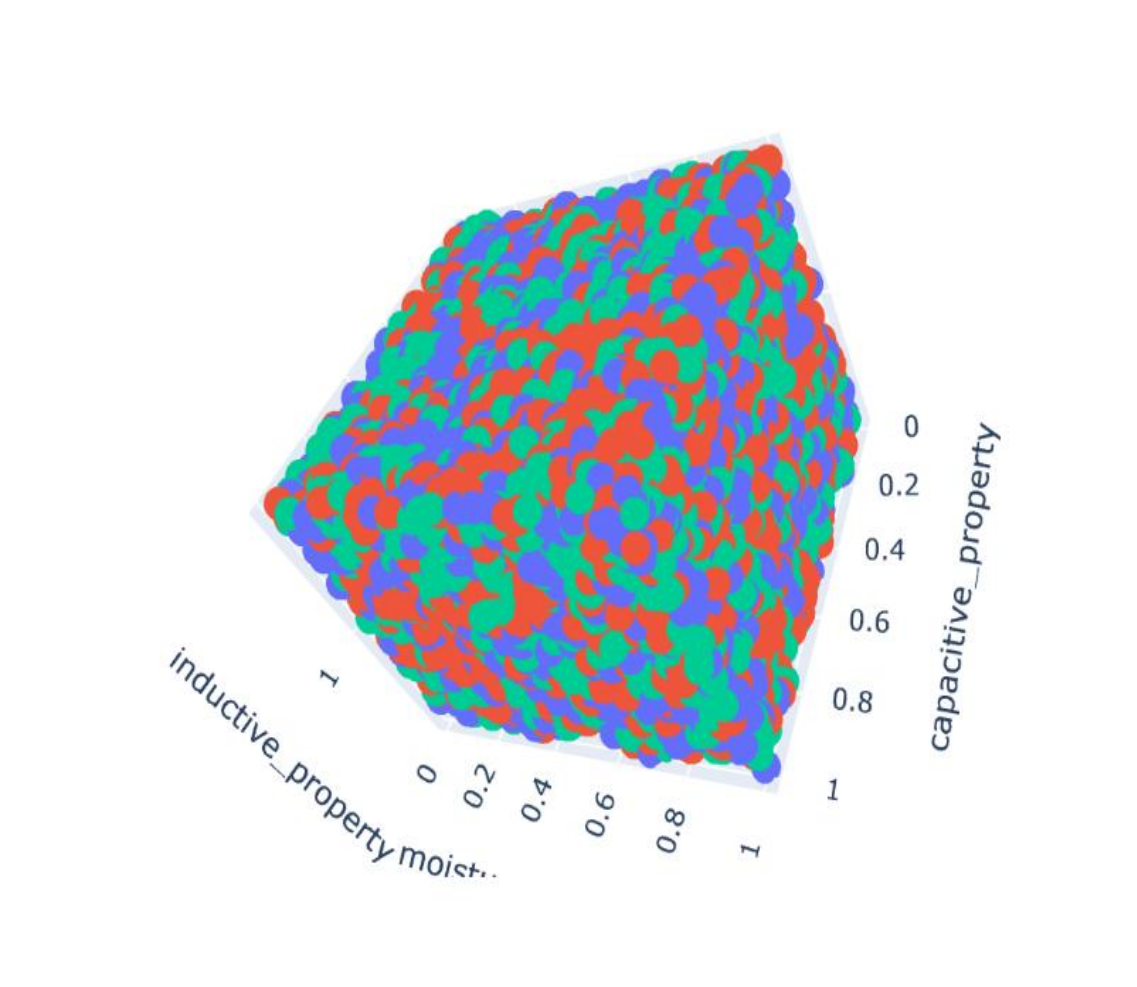
Data Visualizations





Advanced Visualizations





Feature Engineering

New Features Created

- **inductive_capacitive**:
  - Product of inductive_property and capacitive_property.
  - Represents a combined measure of these properties.
- **moisture_infrared**:
  - Product of moisture_property and infrared_property.
  - Captures the interaction between these two properties.

Purpose of Feature Engineering

- Enhance the dataset with meaningful interactions.
- Provide additional insights for machine learning models.
- Improve model performance by capturing complex relationships.

Outlier Handling

Purpose

- Identify and reduce the influence of extreme values.
- Ensure robustness in analysis and modeling.
- Maintain meaningful data distributions.

Method Used: IQR (Interquartile Range)

- **Steps:**
  - Calculate the **Q1 (25th percentile)** and **Q3 (75th percentile)** for each numerical feature.
  - Compute the **IQR**:
    IQR=Q3−Q1IQR = Q3 - Q1IQR=Q3−Q1
  - Define lower and upper bounds:
    Lower Bound=Q1−1.5×IQR\text{Lower Bound} = Q1 - 1.5 \times
    IQRLower Bound=Q1−1.5×IQR
    Upper Bound=Q3+1.5×IQR\text{Upper Bound} = Q3 + 1.5 \times
    IQRUpper Bound=Q3+1.5×IQR
  - Clip values outside the bounds to the nearest bound.

Dimensionality Reduction

PCA (Principal Component Analysis) Implementation

- **Steps:**
  - Applied PCA to reduce the number of features while retaining 95% of the variance in the dataset.
  - Fit the data to extract principal components.

Purpose

- Reduce feature redundancy and complexity.
- Improve model performance and interpretability.
- Mitigate overfitting by focusing on essential data structures.

Model Preparation

Train-Test Split

- **Process:**
  - Split the dataset into training (80%) and testing (20%) subsets.
  - Stratified splitting to maintain class distribution balance.
- **Sizes:**
  - Training Set: 21,411 rows.
  - Testing Set: 5,353 rows.

Feature Scaling

- **Technique:** StandardScaler
  - Scaled numerical features to have a mean of 0 and standard deviation of 1.
- **Purpose:**
  - Normalize feature values for consistency.
  - Improve algorithm performance, especially for distance-based and gradient descent methods.

Classification Models

Algorithms Used

1.Random Forest

- Ensemble method combining multiple decision trees.
- Key Parameters:
  - n_estimators: 50
  - min_samples_split: 5
  - class_weight: balanced

- **XGBoost**
- Gradient boosting algorithm known for high performance.
- Key Parameters:
  - objective: binary:logistic
  - eval_metric: logloss
- **Other Models Evaluated**
- Logistic Regression
- Support Vector Machine (SVM)
- K-Nearest Neighbors (KNN)
- Decision Tree

Model Evaluation

Metrics Used

1. **Accuracy:**
    1. Measures overall correctness of predictions.
    2. Example: Random Forest achieved 85% accuracy.
2. **Classification Report:**
    1. Precision, Recall, and F1-Score for each class.
    2. Helps evaluate model performance on imbalanced data.
3. **Confusion Matrix:**
    1. Visualizes correct and incorrect predictions.
    2. Identifies common misclassification patterns.

Feature Importance

Insights from Random Forest

- **Top Features:**
  - Capacitive Property
  - Inductive Property
  - Moisture-Infrared Interaction
- **Significance:**
  - Capacitive and inductive properties were the most influential in classifying waste types.
  - Engineered features contributed significantly to model performance.

Comparison of Models

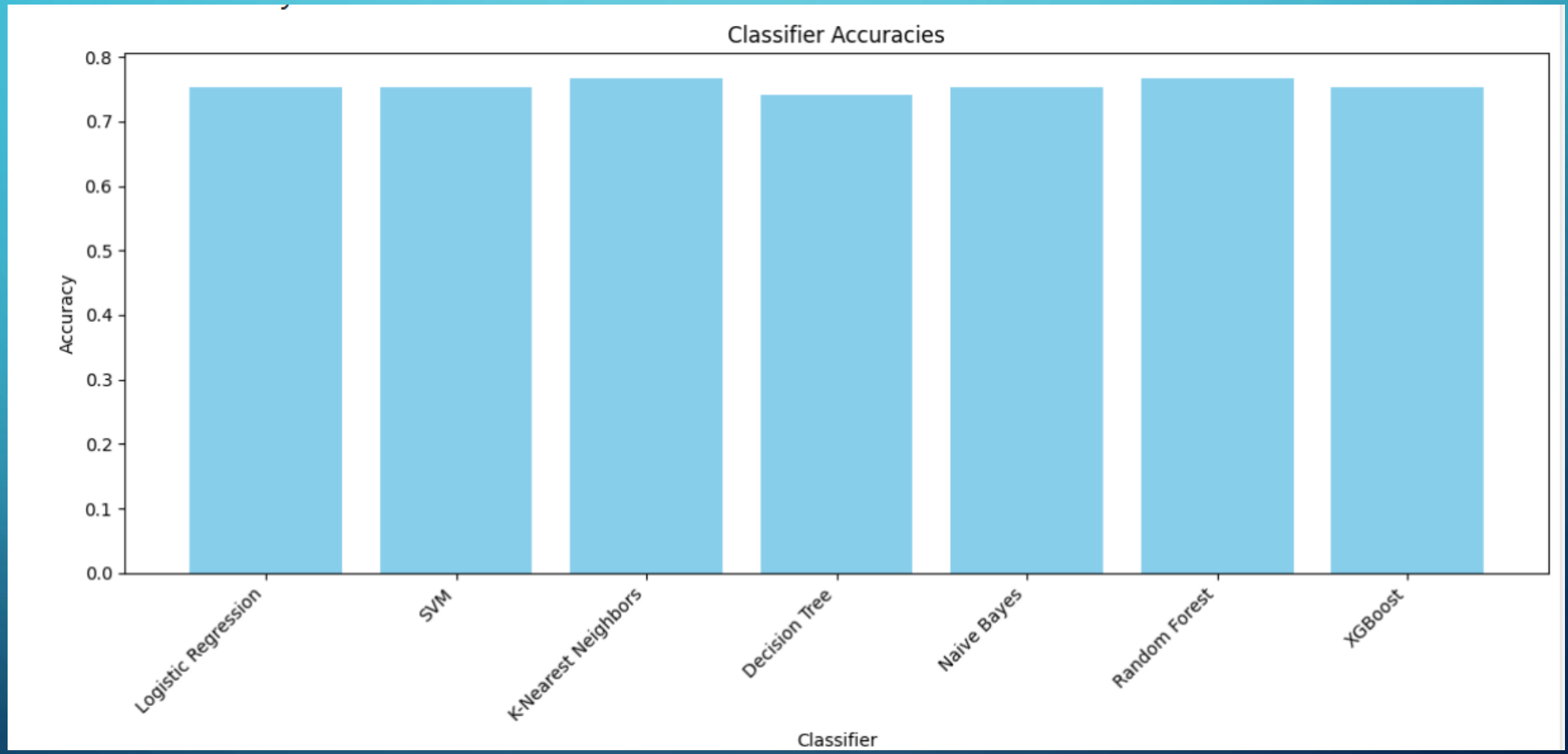
Bar Chart of Classifier Accuracies

- Models Evaluated:
  - Random Forest, XGBoost, Logistic Regression, SVM, etc.
- **Performance:**
  - Random Forest and XGBoost showed the highest accuracy.
  - Example:
    - Random Forest: 85%
    - XGBoost: 83%

Purpose of Comparison

- Identify the best-performing model.
- Visualize model effectiveness for informed decision-making.

Comparison of Models



Conclusion

- The dataset was effectively analyzed to extract meaningful insights using advanced techniques like PCA and feature engineering.
- Random Forest and XGBoost emerged as the best-performing models for waste classification.
- Key properties like capacitive, inductive, and engineered interactions were significant for accurate predictions.

Thank You