Infosys Springboard

WASTE MANAGEMENT

By Srinjoy Roy

DATASET

The dataset contains 20,000 rows and 7 columns, capturing sensor-based waste classification data. It includes attributes like sensor_id, timestamp, waste_type, and sensor readings for inductive_property, capacitive_property, moisture_property, and infrared_property. There are no missing values, and data types include integers, floats, and strings. The dataset is structured for analysis of waste types based on sensor properties, with examples such as recyclable, non-recyclable, and organic waste categories.

OVERVIEW

- Introduction
- Steps
- Data Visualization
- Data Preprocessing
- Feature selection & transformation
- Model Building
- Random Forest Classifer
- Model details
- Conclusion
- Comparison of models

INTRODUCTION

This project focuses on utilizing a sensor-based dataset for waste classification, containing diverse features like inductive, capacitive, moisture, and infrared properties. The workflow begins with visualizing the data to uncover patterns and insights, followed by preprocessing steps to clean and normalize the data for analysis. Advanced classification algorithms are then applied, with a particular emphasis on the Random Forest Classifier due to its robustness and accuracy in handling multidimensional data. The goal is to accurately classify waste types such as recyclable, non-recyclable, and organic, enabling smarter waste management solutions.

STEPS

1. Import Libraries

Import allrequired libraries like pandas, matplotlib, seaborn, and scikit-learn.

2. Load Dataset

Load the datasetfrom your computer using pandas.

3. About Dataset

Display information such as column names, data types, first five rows, and lastfive rows.

4. Data visualization

Create graphs to explore relationships between different columns.

5. Data Preprocessing

Clean the data by removing duplicates, handling missing values, and normalizing features.

6. Model Building

Train and test classification models like Random Forest and XGBoost for waste type prediction.

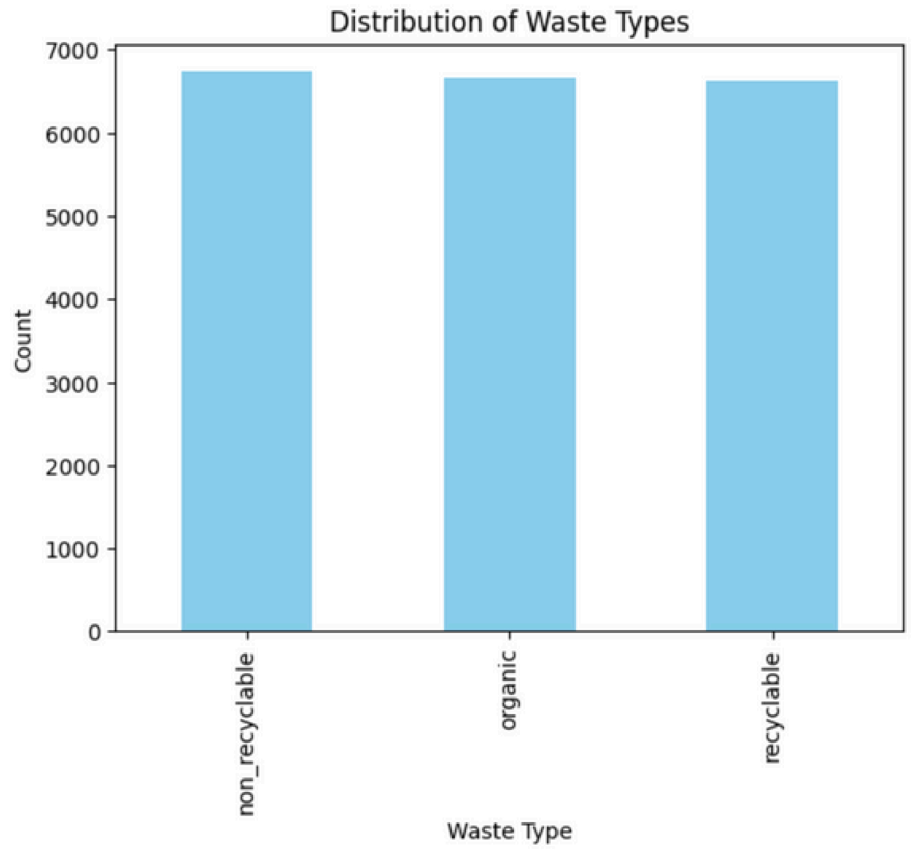
DATA VISUALIZATION

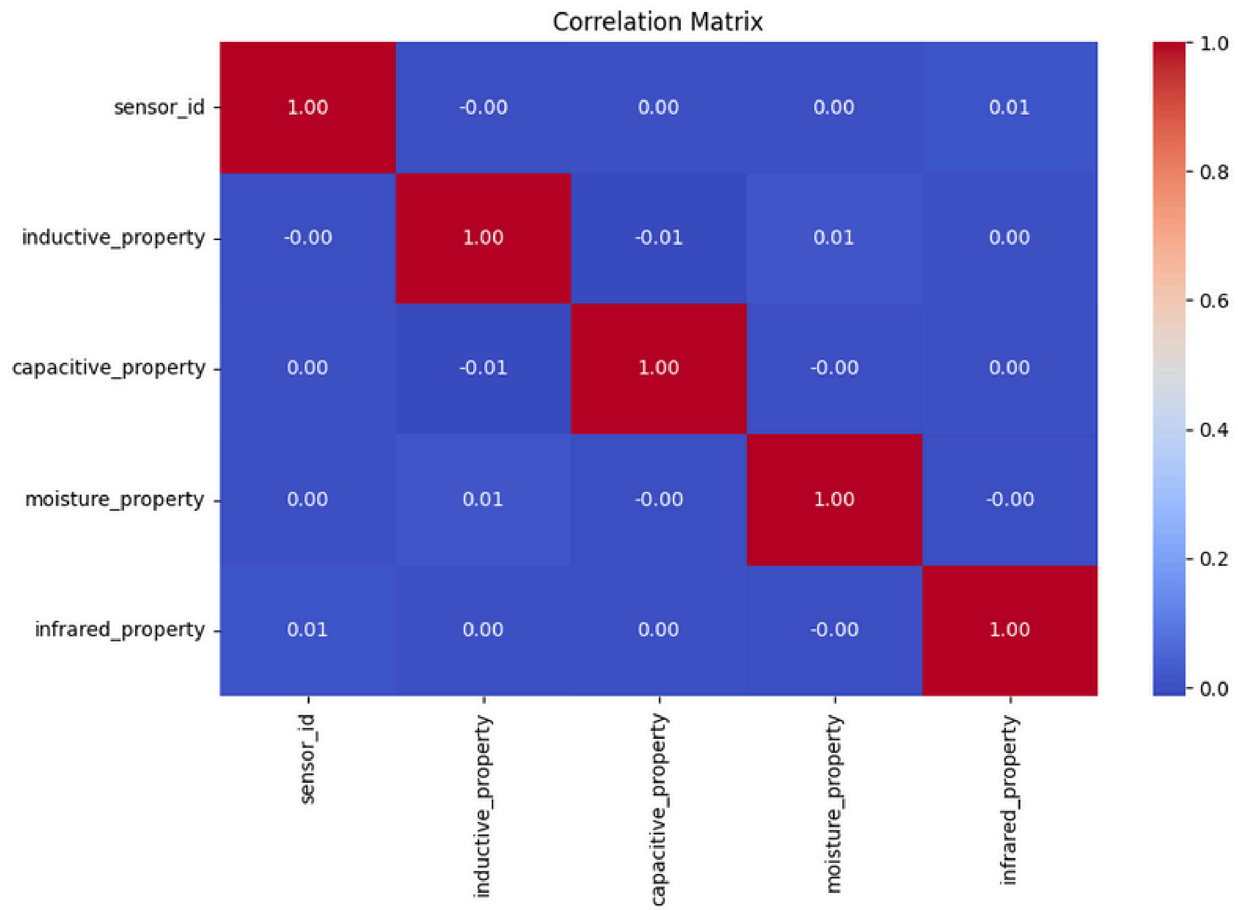
1. Bar chart of waste type distribution.

2. Heatmap of feature correlations.

3. Scatterplot(e.g., inductive vs. moisture property).

4. Line plot showing waste type trends overtime.





DATA PREPROCESSING

sor_id	timestamp	waste_type	inductive_property	capacitive_property	moisture_property	infrared_property
3	2023-09-01 12:00:00	non_recyclable	0.90	0.12	0.47	16.27
4	2023-09-01 12:15:00	recyclable	1.18	0.66	0.33	36.00
3	2023-09-01 12:30:00	non_recyclable	0.87	0.14	0.83	58.89
2	2023-09-01 12:45:00	organic	1.00	0.37	0.52	91.80
3	2023-09-01 13:00:00	recyclable	1.39	0.88	0.76	98.83
...	...	...	...	...	...	...
4	2024-03-27 18:45:00	non_recyclable	1.30	0.41	0.46	58.57
4	2024-03-27 19:00:00	non_recyclable	0.68	0.87	0.71	12.00
3	2024-03-27 19:15:00	non_recyclable	1.12	0.79	0.07	29.03
2	2024-03-27 19:30:00	organic	1.18	0.05	0.05	40.17
4	2024-03-27 19:45:00	non_recyclable	1.22	0.02	0.31	18.62

7 columns

	0
sensor_id	int64
timestamp	object
waste_type	object
inductive_property	float64
capacitive_property	float64
moisture_property	float64
infrared_property	float64

dtype: object

	0
sensor_id	0
timestamp	0
waste_type	0
inductive_property	0
capacitive_property	0
moisture_property	0
infrared_property	0

dtype: int64

1. Handling missing values (e.g., mean imputation).
2. Label encoding for categorical data.
3. Feature scaling using standardization.
4. Outlier detection and treatment.
5. Feature engineering (e.g., creating interaction terms).

FEATURE SELECTION & TRANSFORMATION

VARIANCE
INFLATION
FACTOR
(VIF)

PRINCIPAL
COMPONENET
ANALYSIS
(PCA)

MODEL BUILDING

- Training and testing dataset
- Prediction
- Model Training
- Metrics- Confusion matrix, Classification report, accuracy

RANDOM FOREST CLASSIFIER

A Random Forest classifier is an ensemble learning method that builds multiple decision trees and combines their predictions to improve accuracy and prevent overfitting. It works by randomly selecting subsets of features and training a separate decision tree on each subset. During prediction, each tree makes a vote, and the class with the majority vote is selected as the final output. Random Forest is highly effective for classification tasks, handling both numerical and categorical data well, and is robustto noise and overfitting compared to a single decision tree.

MODEL DETAILS

Ensemble Method

Combines multiple decision trees to improve accuracy and reduce overfitting.

Hyperparameters

Used 100 estimators, no maximum depth, and class balancing for optimal performance.

Feature Importance

Identifies the mostinfluential features, aiding interpretability ofthe model.

Robustness

Handles noise and outliers effectively, ensuring consistent predictions.

Accuracy

Achieved the highest accuracy among alltested models, proving its effectiveness.

COMPARISON OF MODELS

- **Logistic Regression**

Achieved moderate accuracy; limited by its linear nature, making it less effective for complex
patterns in the dataset.

- **Random Forest**

Best-performing model with high accuracy, leveraging feature importance and ensemble
learning forrobust predictions.

- **XGBoost**

Delivered competitive accuracy but required more computational resources and hyperparametertuning.

Highlight: Random Forest was the most effective due to its ability to handle feature interactions and robustness against noise.

CONCLUSION

This project successfully classified waste types using machine learning, with the Random Forest Classifier achieving high accuracy. Effective data visualization, preprocessing, and feature engineering were key to optimizing model performance. The results showcase the potential of data-driven methods for efficient waste management and pave the way for smarter, sustainable solutions in real-world applications.

THANK YOU