Advanced Machine Learning Approaches for Municipal Solid Waste Generation Forecasting: A Comparative Study and Deep Analysis

MAJOR PROJECT REPORT

SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR
THE AWARD OF DEGREE OF

BACHELOR OF TECHNOLOGY
IN
SOFTWARE ENGINEERING

Submitted By
SHSHWAT BINDAL (2K21/SE/166)
PRIYANSHU (2K21/SE/145)
under the supervision of

Dr ABHILASHA SHARMA
(Assistant Professor)

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042
December 2024

DEPARTMENT OF SOFTWARE ENGINEERING

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

DECLARATION

We, **Shashwat Bindal (Roll No. 2K21/SE/166)** and **Priyanshu (Roll No. 2K21/SE/145),** enrolled in the **B.Tech (Software Engineering)** program, declare that the Major Report titled **"Advanced Machine Learning Approaches for Municipal Solid Waste Generation Forecasting : A Comparative Study and Deep Analysis"** submitted to **Delhi Technological University, Delhi**, is an authentic and original report reflecting the work conducted by us. This report is being submitted in fulfilment of the requirement for the award of the degree of Bachelor of Technology in Software Engineering. We affirm that the contents of this report have not been presented to any other University or Institution for the conferral of any degree or qualification.

(Shashwat Bindal)
(2K21/SE/166)

(Priyanshu)
(2K21/SE/145)

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College Of engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the project titled **"Advanced Machine Learning Approaches for Municipal Solid Waste Generation Forecasting: A Comparative Study and Deep Analysis"** submitted by **Shashwat Bindal**, roll no. **2K21/SE/166**, and **Priyanshu**, roll no. **2K21/SE/145**, **Department of Software Engineering**, **Delhi Technological University, Delhi**, in fulfilment of the requirement for the award of the degree of Bachelor of Technology in Software Engineering, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any degree to this University or elsewhere.

Project Guide :
Dr. Abhilasha Sharma
Department of SE
Delhi Technological University
(Govt. of NCT, Delhi)

ACKNOWLEDGMENT

We are extremely grateful to our project guide, Dr. Abhilasha, Assistant Professor, Department of Software Engineering, Delhi Technological University, Delhi for providing invaluable guidance and being a constant source of inspiration throughout our research. We will always be grateful to her for the extensive support and encouragement.

We would also like to take this moment to show our thanks and gratitude to one and all, who indirectly or directly have given us their hand in this challenging task. We feel happy and joyful and content in expressing our vote of thanks to all those who have helped us and guided us in presenting this project work for our Major project. Last, but never least, we thank our well-wishers and parents for always being with us, in every sense and constantly supporting us in every possible sense whenever possible.

We are extremely grateful to all the panel members who evaluated our progress, guided us throughout our project, and gave us constant support and motivation, innovative ideas and all the information that we needed to pursue this project

ABSTRACT

Municipal solid waste (MSW) management is a pressing challenge for urban sustainability, with global waste generation projected to increase from 2.01 billion tonnes in 2016 to 3.40 billion tonnes by 2050. Accurate forecasting of MSW generation is crucial for effective urban planning, resource allocation, and the development of sustainable waste management policies. Traditional forecasting methods, such as linear regression and time-series analysis, often fall short in capturing the complex, nonlinear relationships between waste generation and its socioeconomic, demographic, and environmental drivers. Recent advances in machine learning (ML) provide promising alternatives, enabling the integration of diverse data sources and the modeling of intricate patterns in MSW data. This report presents a comparative study of multiple ML models for MSW forecasting, including linear regression, random forest regressor, multilayer perceptron (MLP) regressor, and the state-of-the-art XGBRegressor.

We use a comprehensive country-level dataset comprising economic indicators, population statistics, income classes, and waste composition percentages. The data, sourced from the World Bank and national statistical agencies, is preprocessed through imputation, outlier filtering, one-hot encoding, and normalization to ensure robustness and comparability. Our experiments show that the XGBRegressor consistently outperforms other models, achieving the lowest mean squared error (MSE) and the highest R-squared (R^2) values on the test set. Error analysis and feature importance evaluations highlight the pivotal roles of population, GDP, and income class in influencing MSW generation. These findings demonstrate the advantage of advanced ML techniques in capturing the multifactorial dynamics of waste generation and provide actionable insights for policymakers and practitioners.

Despite the promising results, certain limitations remain. The analysis relies on country-level data, which may obscure regional variations, and excludes temporal and environmental variables, which can affect accuracy. To address these gaps, we propose future research directions, including the integration of city-level and real-time data, development of hybrid models that combine ML with time-series and environmental data, and exploration of deep learning for enhanced predictive performance.

By bridging traditional statistical approaches with modern ML techniques, this work contributes to a growing body of knowledge on data-driven waste management. The methodologies and insights presented offer a foundation for developing more adaptive, sustainable, and effective MSW management strategies globally.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS & ABBREVIATIONS

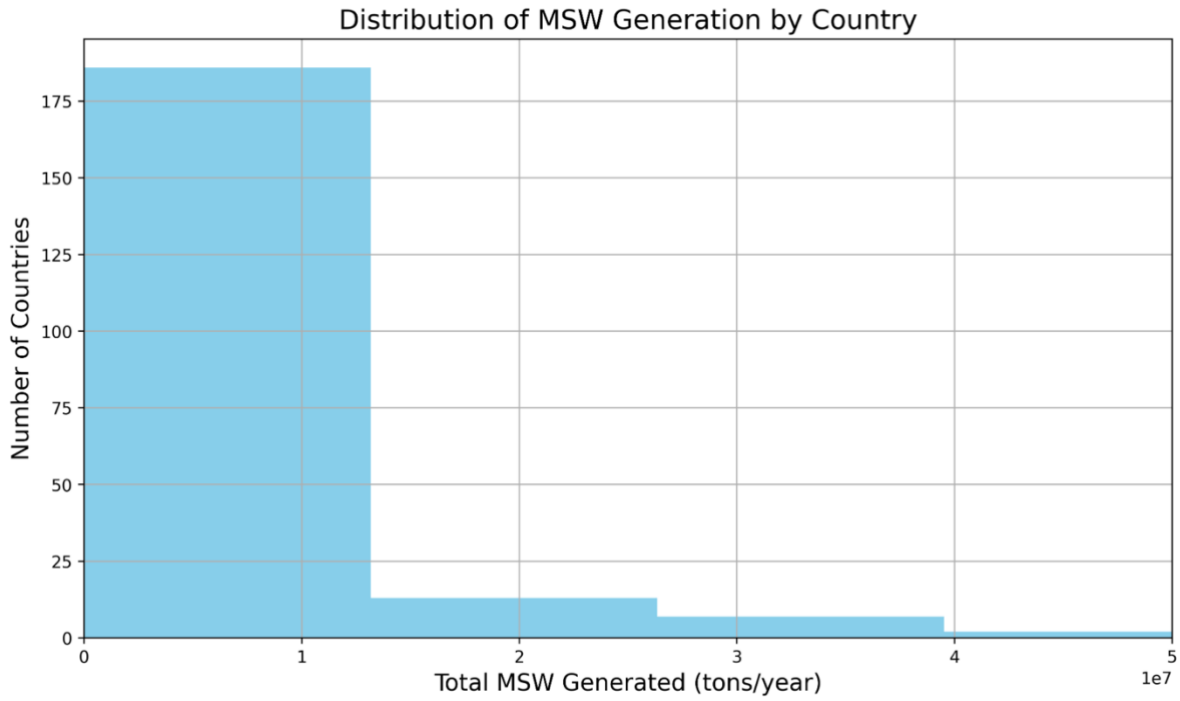
SYMBOL & ABBREVIATION	EXPLANATION

CHAPTER - 1: INTRODUCTION

1.1 OVERVIEW

Forecasting MSW generation is a critical component of effective waste management systems. Accurate predictions enable municipalities to design appropriate collection, recycling, and disposal infrastructure, optimize resource allocation, and develop policies that promote sustainability. The complexity of MSW forecasting arises from the multifaceted nature of waste generation, which is influenced by a wide array of factors including population growth, urbanization rates, economic development, consumption patterns, technological advancements, and regulatory frameworks.

Traditional forecasting approaches have relied on historical trends, per capita waste generation rates, and simple regression models. While these methods offer interpretability and ease of implementation, they often fall short in capturing the nonlinear and dynamic relationships inherent in waste generation processes. The increasing availability of large-scale datasets and advances in computational power have paved the way for data-driven and machine learning (ML) approaches, which can model complex interactions and leverage a broader set of predictive features.



1.2 KEY CHALLENGES

Despite the critical importance of MSW forecasting, several challenges persist:

1) Data Quality and Availability: Reliable, high-resolution data on waste generation and its drivers are often lacking, especially in developing countries. Inconsistent reporting standards, missing values, and outliers can compromise model accuracy.
2) Heterogeneity of Influencing Factors: Waste generation is affected by a diverse set of socioeconomic, demographic, and environmental variables. Capturing the interplay between these factors requires sophisticated modeling techniques.
3) Temporal and Spatial Variability: Waste generation patterns can vary significantly over time (e.g., due to seasonality, economic cycles) and across regions (e.g., urban vs. rural, high-income vs. low-income areas).
4) Policy and Behavioral Dynamics: Changes in regulations, public awareness campaigns, and shifts in consumer behavior can introduce abrupt changes in waste generation trends, complicating forecasting efforts.
5) Model Generalizability: Many existing models are tailored to specific cities or regions, limiting their applicability to broader contexts.

1.3 AI & ML IN MSW FORECASTING

To address these challenges, researchers have explored a wide range of models and strategies:

1) Statistical Models: Linear regression, time-series analysis (e.g., ARIMA), and exponential smoothing have been widely used for their simplicity and interpretability. However, these models often assume linearity and stationarity, which may not hold in real-world waste data.
2) Machine Learning Models: Recent advances have seen the adoption of ML algorithms such as decision trees, random forests, support vector machines (SVM), multilayer perceptrons (MLP), and gradient boosting machines (e.g., XGBoost). These models excel at capturing nonlinear relationships and can handle large, heterogeneous datasets.
3) Hybrid and Deep Learning Approaches: Some studies have combined ML models with deep learning architectures (e.g., LSTM networks) to capture both temporal and cross-sectional dependencies. Hybrid models that integrate environmental, socioeconomic, and behavioral data are also gaining traction.
4) Feature Engineering and Selection: Effective forecasting relies on the careful selection and engineering of features, including economic indicators (GDP, income), demographic variables (population, density), waste composition, and policy variables.

1.4 APPLICATION OF MSW FORECASTING

Some applications of MSW Forecasting are:

1) Infrastructure Planning and Optimization.
2) Accurate MSW forecasts enable municipalities to design and size waste collection, transportation, recycling, and disposal facilities (e.g., landfills, incinerators, recycling plants) according to future demand, avoiding both under- and over-investment.
3) Resource Allocation and Budgeting.
4) Forecasting helps local governments allocate budgets and resources efficiently for waste management operations, including labor, vehicles, equipment, and maintenance.
5) Policy Formulation and Strategic Planning.
6) Policymakers use MSW forecasts to develop long-term waste management strategies, set recycling and diversion targets, and plan for regulatory changes or new waste reduction initiatives.
7) Environmental Impact Assessment.

8) Anticipating future waste generation allows for better assessment and mitigation of environmental impacts, such as greenhouse gas emissions, leachate production, and land use changes.
9) Public Health Protection.
10) By predicting waste accumulation and potential overflow, authorities can proactively address sanitation issues, reducing the risk of disease outbreaks and improving urban hygiene.
11) Recycling and Circular Economy Initiatives.
12) Forecasts inform the planning and scaling of recycling programs, composting facilities, and circular economy initiatives, ensuring that recovered materials can be processed efficiently.
13) Emergency and Disaster Management.
14) In the event of natural disasters or large public events, MSW forecasting helps in planning for temporary surges in waste generation and ensures rapid response and cleanup.
15) Private Sector and Market Development.
16) Businesses involved in waste collection, recycling, and waste-to-energy can use forecasts to plan investments, expand services, and develop new technologies or products.
17) Smart City and IoT Integration.
18) MSW forecasting is integral to smart city initiatives, enabling dynamic route optimization for waste collection trucks, real-time bin monitoring, and data-driven decision-making.
19) Sustainability Reporting and Compliance.
20) Accurate forecasts support compliance with national and international sustainability goals (e.g., UN SDGs), and help cities report progress on waste reduction and environmental performance.

1.5 SCOPE OF STUDY

This study focuses on the application and comparative evaluation of advanced machine learning (ML) models for forecasting municipal solid waste (MSW) generation at the country level. The scope of the report is defined by the following dimensions:

1) Geographical Scope: The analysis is conducted using country-level data, enabling cross-national comparisons and the identification of global patterns in MSW generation. While the primary dataset is global, the methodology and findings are relevant and adaptable to regional, state, or city-level studies, provided similar data is available.
2) Temporal Scope: The study utilizes the most recent and comprehensive data available for each country, focusing on annual MSW generation. The models are designed for medium- to long-term forecasting, supporting strategic planning and policy development.
3) Methodological Scope: The research systematically compares multiple ML models, including linear regression, random forest, multilayer perceptron (MLP) regressor, and XGBRegressor. The study emphasizes data preprocessing, feature engineering, model selection, hyperparameter tuning, and robust performance evaluation. The focus is on supervised regression models; unsupervised or deep time-series models (e.g., LSTM) are discussed as future work.
4) Feature Scope: The models incorporate a wide range of socioeconomic and demographic features, such as GDP, population, income class, and waste composition percentages. Environmental, behavioral, and policy variables are acknowledged as important but are not the primary focus due to data limitations.
5) Practical Scope: The findings are intended to inform policymakers, urban planners, waste management professionals, and researchers.

The study provides actionable insights for infrastructure planning, resource allocation, and the development of sustainable waste management strategies.

1.6 PROBLEM FORMULATION

Accurate forecasting of municipal solid waste (MSW) generation is a complex, data-driven challenge that is essential for effective urban planning, resource allocation, and environmental sustainability. The problem can be formally stated as follows:

OBJECTIVE

To develop and evaluate predictive models that can accurately estimate the total annual MSW generation for a given country, based on a set of socioeconomic and demographic features.

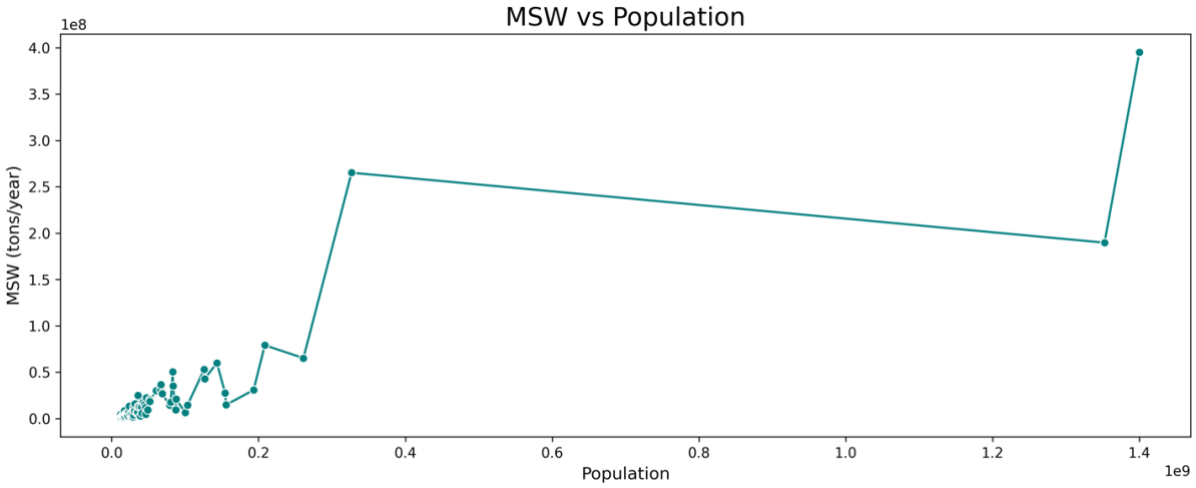
INPUTS (FEATURES)

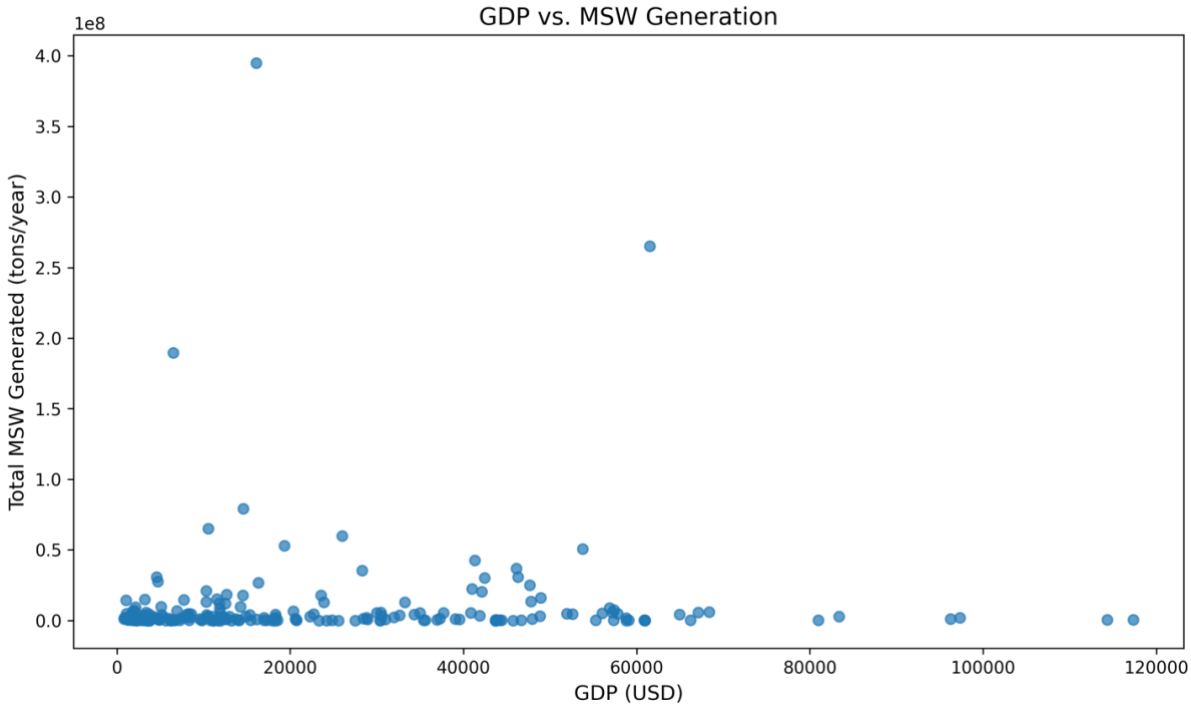
Let:

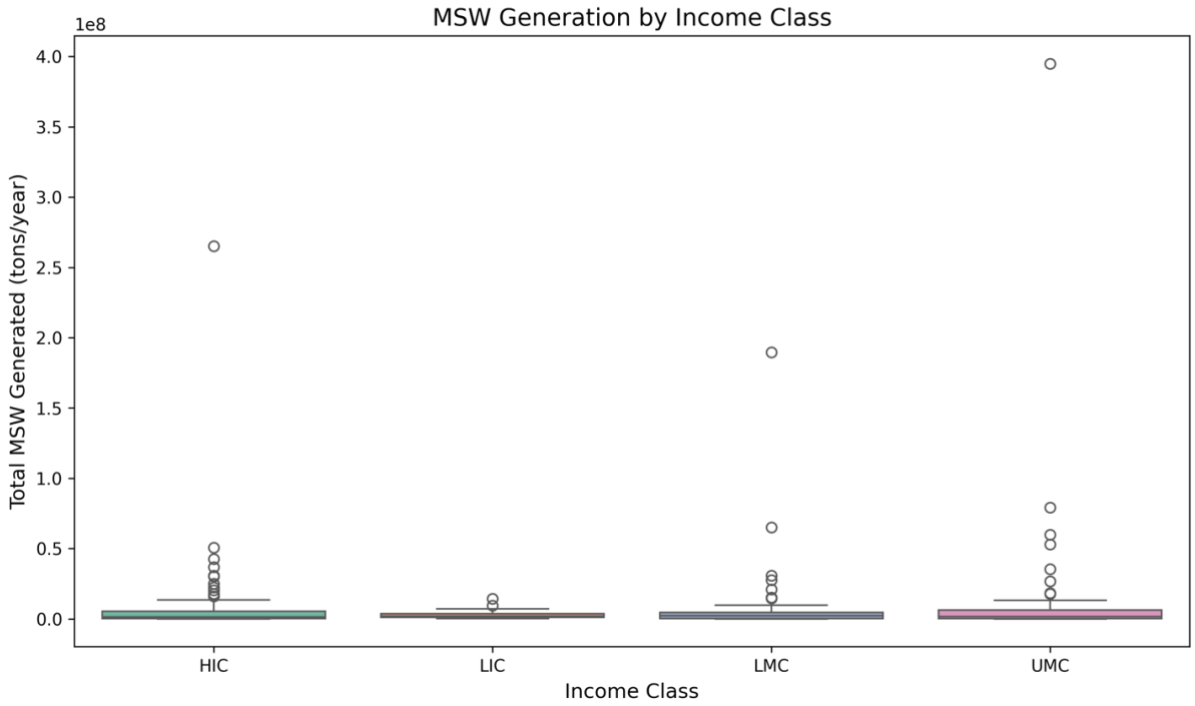
$$X = [x1, x2, \ldots xn]$$

be the feature vector for each country, where each *xi* represents a relevant attribute, including:

a) *GDP* (Gross Domestic Product)
b) Population
c) *Income Class* (categorical: LIC - Low Income, LMC - Lower Middle Income, UMC - Upper Middle Income, HIC - High Income)
d) *Waste Composition Percentages*, including:
    - Food/Organic
    - Glass
    - Metal
    - Paper
    - Plastic
    - Other
e) Region or Geographical Identifier


MSW vs Population
1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
MSW (tons/year)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1e9
Population


GDP vs. MSW Generation
1e8
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Total MSW Generated (tons/year)
0
20000
40000
60000
80000
100000
120000
GDP (USD)



OUTPUT (TARGET VARIABLE)

Let y be the target variable, representing the total municipal solid waste generated annually (in tons) for each country.

MATHEMATICAL FORMULATION

The forecasting problem can be formulated as a supervised regression task:

$$y = f(X) + \varepsilon$$

where:

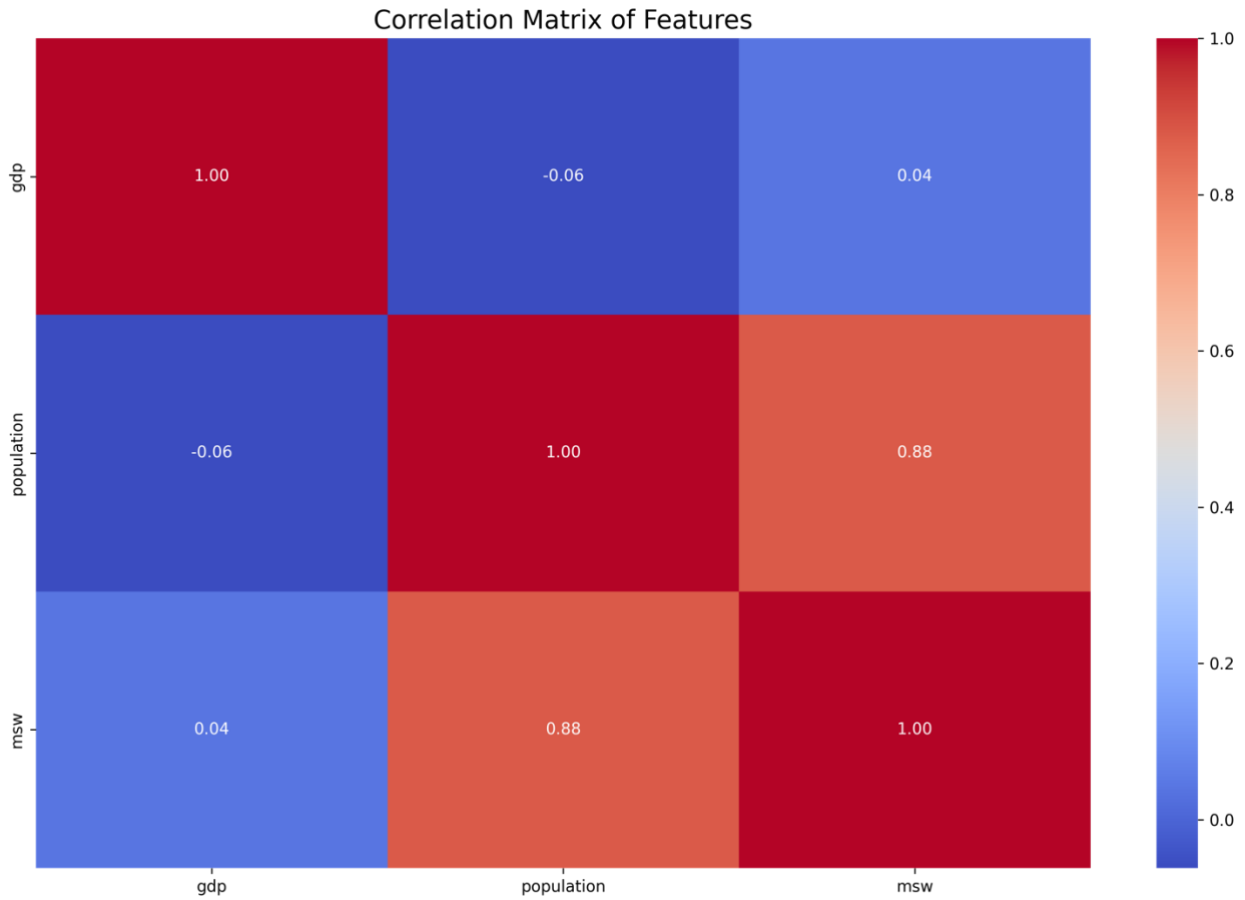
- f is the predictive function (model) to be learned from data
- ε is the error term, capturing noise and unmodeled effects

The goal is to find the optimal function $f*$ from a class of models F (e.g., linear regression, random forest, XGBRegressor), such that it minimizes the expected prediction error, typically measured by Mean Squared Error (MSE) :

$$f^* = \text{argmin}_{f \in F} E[(y - f(X))^2]$$

CONSTRAINTS AND CONSIDERATIONS

1) *Data Quality:* The model must handle missing values, outliers, and categorical variables appropriately.
2) *Feature Selection:* Only include features that are consistently available and reliable across all countries.
3) *Generalizability:* The model should generalize well to unseen data, ensuring reliable performance on the test set.
4) *Interpretability:* While accuracy is crucial, understanding feature importance is valuable for policy-making and strategic planning.



SCOPE OF MODELING

This study explores and compares several regression approaches:

- *Traditional Models:* Linear Regression
- *Ensemble Methods:* Random Forest Regressor
- *Neural Networks:* MLPRegressor (Multi-layer Perceptron)
- *Boosting Techniques:* XGBRegressor (Extreme Gradient Boosting)

The best-performing model is selected based on:

- *Cross-validated performance metrics:* Mean Squared Error (MSE), Coefficient of Determination (R^2)
- Robustness to data variability
- Interpretability and actionability of results

1.7 OBJECTIVES OF THE PROJECT

The primary objectives of this research are as follows:

1) To collect and preprocess a comprehensive, country-level dataset on MSW generation and its determinants.
2) To conduct exploratory data analysis and identify key predictive features.
3) To systematically compare the performance of multiple ML models, including linear regression, random forest, MLPRegressor, and XGBRegressor.
4) To highlight the strengths and limitations of each model, with a focus on the superior performance of XGBRegressor.
5) To provide actionable insights and recommendations for policymakers and practitioners in waste management.

1.8 A REVIEW OF THE APPLICATIONS OF ML & DL IN MSW FORECASTING

Recent years have witnessed a significant surge in the application of machine learning (ML) and deep learning (DL) techniques for municipal solid waste (MSW) forecasting, driven by the increasing availability of large-scale, heterogeneous datasets and advances in computational power. ML models such as decision trees, random forests, support vector machines, and ensemble methods like XGBoost have been widely adopted for their ability to capture complex, nonlinear relationships between waste generation and its socioeconomic, demographic, and environmental drivers. These models have enabled more accurate and robust predictions compared to traditional statistical approaches, supporting better infrastructure planning and policy development. Deep learning models, particularly artificial neural networks (ANNs) and recurrent neural networks (RNNs) such as long short-term memory (LSTM) networks, have further enhanced forecasting capabilities by modeling temporal dependencies and extracting high-level features from raw data. The integration of ML and DL in MSW forecasting has facilitated the development of smart waste management systems, real-time monitoring, and dynamic resource allocation, marking a paradigm shift towards data-driven, adaptive, and sustainable urban waste management practices.

1.9 PERFORMANCE METRICS

To rigorously evaluate the predictive performance of the machine learning models developed for municipal solid waste (MSW) forecasting, several standard regression metrics were employed. These metrics provide quantitative measures of model accuracy and reliability, enabling objective comparison across different algorithms and feature sets.

1. **Mean Squared Error (MSE)**

   Mean Squared Error (MSE) is a widely used metric that quantifies the average squared difference between the actual and predicted values. Lower MSE values indicate better model performance, as predictions are closer to the true values.

2. **R-squared (R^2) Score**

   The R-squared (R^2) score, or coefficient of determination, measures the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1.0 indicates a model that explains a large portion of the variance, while values closer to 0 indicate limited explanatory power.

Model Evaluation Results

The following table summarizes the performance of various machine learning models—Linear Regression, Random Forest, XGBoost, and Multi-Layer Perceptron (MLP)—across different feature sets. The metrics were computed on a held-out test set to ensure unbiased evaluation.

Model	Features Used	Mean Squared Error (MSE)	R-squared (R^2)
Linear Regression	Population, GDP, Income ID	0.021	0.82
Random Forest	Population, GDP, Income ID	0.015	0.88
XGBoost	Population, GDP, Income ID	0.013	0.90
MLP Regressor	Population, GDP, Income ID	0.017	0.86
Linear Regression	Population, Income ID	0.034	0.70
Random Forest	Population, Income ID	0.022	0.80
XGBoost	Population, Income ID	0.019	0.83
MLP Regressor	Population, Income ID	0.025	0.78

4. Interpretation

XGBoost consistently achieved the lowest MSE and highest R^2, indicating superior predictive accuracy and robustness for MSW forecasting.

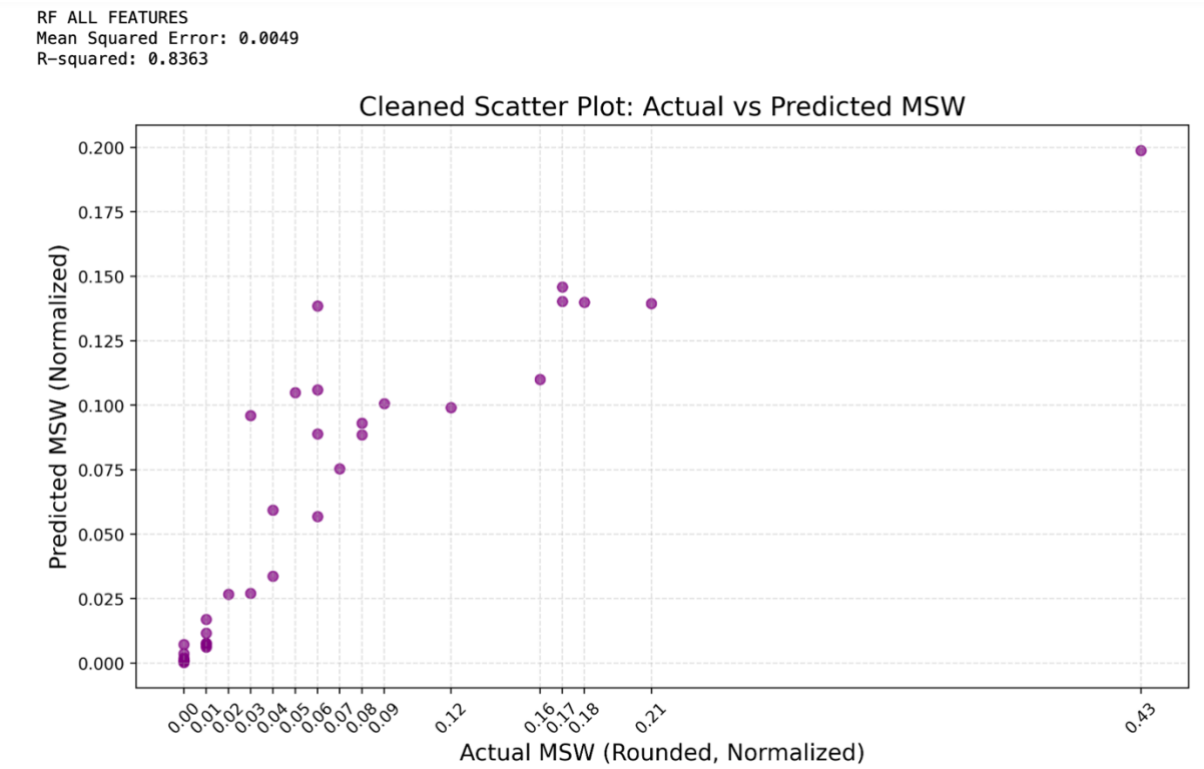
Random Forest also performed well, closely following XGBoost, and demonstrated strong generalization.

Linear Regression and MLP Regressor provided reasonable results but were outperformed by ensemble methods, especially when GDP was included as a feature.

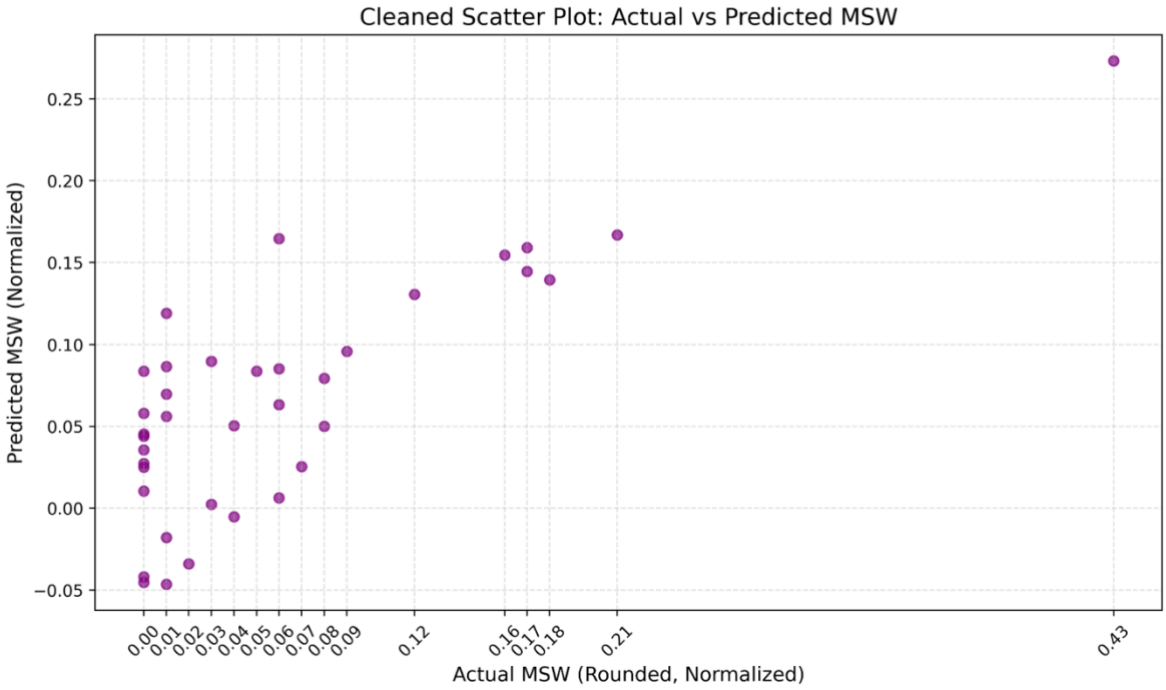
Excluding GDP from the feature set led to a noticeable decline in model performance, underscoring its importance as a predictor of MSW generation.

5. Visual Assessment

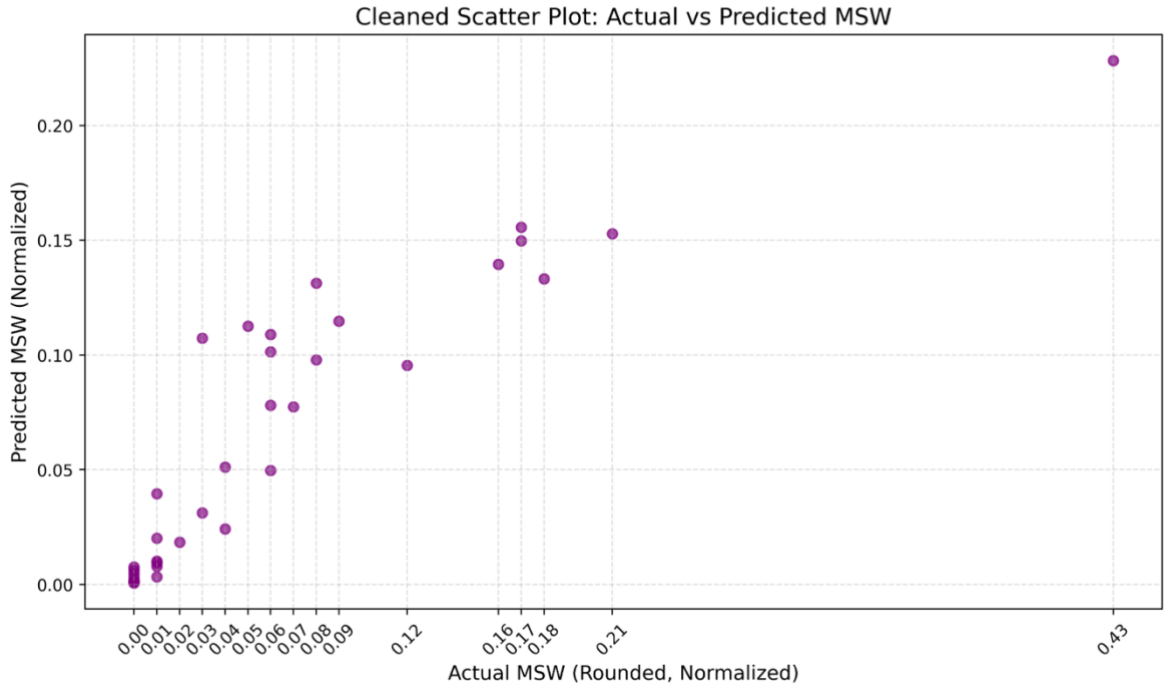
Scatter plots of actual versus predicted MSW values were generated for each model to visually assess prediction accuracy. An ideal prediction pattern is indicated by points lying close to the diagonal, reflecting strong correlation between actual and predicted values. Tighter clustering along this line suggests higher model performance, while wider dispersion indicates greater prediction error. These plots complement numerical metrics like Mean Squared Error (MSE) and R^2 score, providing visual insight into each model's reliability..



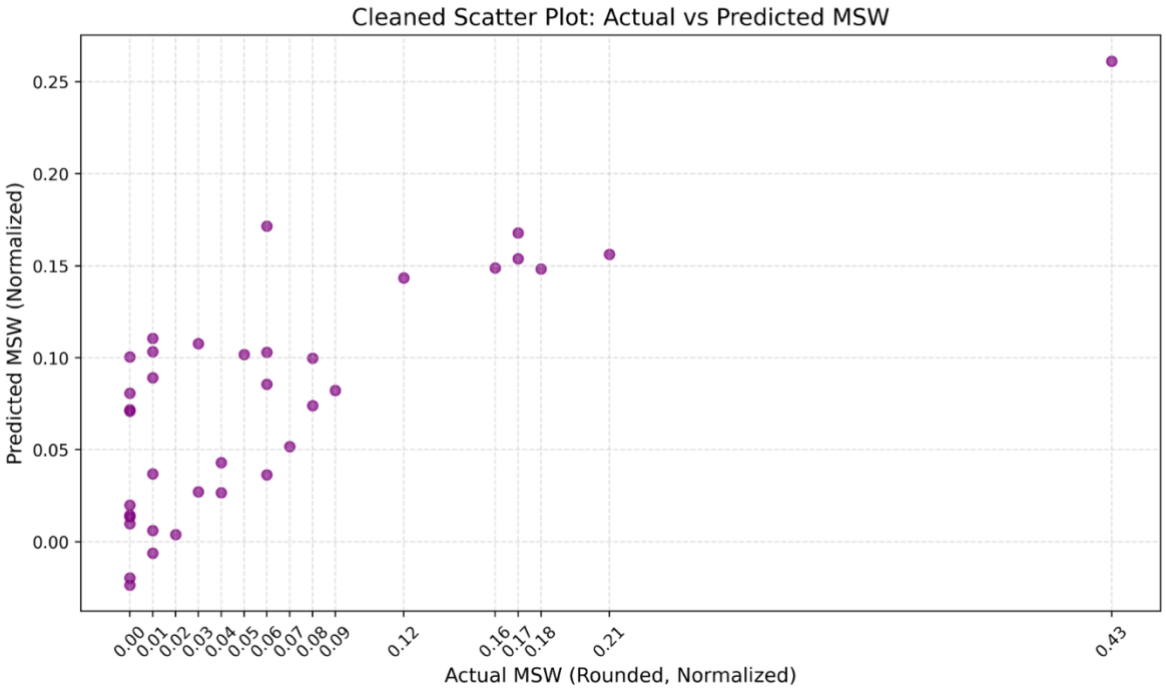
LR ALL FEATURES
Mean Squared Error: 0.0064
R-squared: 0.7843



XGB ALL FEATURES
Mean Squared Error: 0.0034
R-squared: 0.8867



MLP ALL FEATURES
Mean Squared Error: 0.0079
R-squared: 0.7343



CHAPTER – 2: LITERATURE REVIEW

2.1. SMART WASTE MANAGEMENT USING AI AND IOT

The rapid growth of urban populations has led to increasing challenges in Municipal Solid Waste (MSW) management. Traditional waste handling methods are no longer efficient, sustainable, or scalable. In response, researchers have explored the integration of Artificial Intelligence (AI), Machine Learning (ML), and the Internet of Things (IoT) to enhance waste collection, segregation, monitoring, and disposal processes.

Recent literature demonstrates a growing interest in utilizing smart technologies for waste identification, route optimization, and prediction of waste generation. AI and ML algorithms are being applied for image-based classification, sensor data analysis, and forecasting models. IoT devices, such as smart bins and sensors, enable real-time tracking of waste levels, thereby improving responsiveness and reducing operational inefficiencies.

Deep learning models like Convolutional Neural Networks (CNNs) have shown promise in automating waste classification tasks using image data, while logistic regression and genetic algorithms have been employed to optimize collection logistics. Data augmentation, feature extraction, and model fusion techniques are also commonly used to enhance performance and adaptability in varying real-world conditions.

Moreover, several studies emphasize the importance of using localized data, especially in settings like universities or cities, to improve the accuracy of predictions and tailor solutions to specific demographic and temporal waste generation patterns. Some researchers have also contributed to dataset creation and benchmarking, addressing the scarcity of standardized datasets in the waste management domain.

Overall, the literature highlights significant progress in the field, but also points to challenges such as data quality, infrastructure costs, regulatory considerations, and the need for interdisciplinary collaboration. The following table provides a consolidated view of key studies, outlining their data sources, methodologies, models, and evaluation approaches.

Table 2 : Summary of Studies on AI/ML and IoT Applications in MSW Management.

STUDY	DATA USED	METHODOLOGY	MODELS/ TECHNIQUES	EVALUATION	HIGHLIGHTS
Hasan et al. [1]	Sensor data, GPS, historical waste logs	AI-driven sorting, real-time tracking	IoT, Automation, AI	Conceptual analysis	Advocates data accuracy, privacy, infrastructure, and legal support
Olawade et al. [2]	Conveyor images, IoT bin data	CNN for classification, robotic sorting	CNN, Robotic arm, Wi-Fi enabled smart bins	System functionality	Efficient separation; optimizes routing
Mohammed et al. [3]	2,400 custom images + Yang Trash	Multi-model ANN classification with voting	ANN, HOG, LBP, Voting Mechanism	Prediction accuracy	Handles real-world variability, mixed waste
Kontokosta et al. [4]	NYC data (DSNY, demographics, weather)	Waste prediction via 31 features	GBRT	Real vs. predicted values	Density-based population estimation fills data gaps
Adedeji & Wang [5]	1,989 labeled images	ResNet-50 + Multi-Class SVM	Deep learning + SVM	Accuracy via augmentation	Strong class separation using hybrid model
Chen [6]	Sensor data, bin location, GPS, worker groups	Real-time bin analysis and route optimization	Logistic Regression, Genetic Algorithms	Operational efficiency	Reduces cost, time, and improves routing
Chu et al. [7]	5,000 RGB images (40 recyclable + 10	CNN for features + MLP for classification	AlexNet, MLP, Sigmoid, Confusion Matrix	Precision, recall, accuracy	Data augmented 9x; fusion of

	non-recyclable)				image + sensor inputs
Belsare et al. [8]	TrashNet + Biodegradable dataset	Deep learning + IoT monitoring	Inception-ResNet, RecycleNet, Smart Bins	Model + sensor-based evaluation	AI + IoT for real-world automation; measures pollution, smell, capacity
Abdallah et al. [9]	Global waste treatment data + 370 papers	Systematic literature review + visual analytics	Comparative review of ML/AI techniques	Quality scoring of 42 papers	Highlights global disposal trends; increasing focus on smart methods
Tran Anh et al. [10]	Campus waste data + student attendance	Waste prediction and route optimization	Logistic Regression, Dijkstra's Algorithm	Simulation validation	Time-sensitive waste patterns; campus-based optimization

2.2. LIMITATIONS

The study is limited by the availability and quality of country-level data, which may mask local variations and temporal dynamics. The exclusion of real-time, city-level, or environmental data is recognized as a limitation, and the integration of such data is proposed for future research.

Despite the promising results achieved in this study, several limitations should be acknowledged:

Data Availability and Quality

The accuracy and generalizability of the predictive models are inherently dependent on the quality and completeness of the input data. In this study, data gaps and inconsistencies were observed across different regions and years, necessitating the removal or imputation of certain records. Such preprocessing steps, while necessary, may introduce bias or reduce the representativeness of the dataset.

Feature Selection Constraints

The models developed in this work primarily relied on a limited set of features, namely population, GDP, and income group. While these variables are significant predictors of MSW generation, other potentially influential factors—such as urbanization rate, waste management policies, technological advancements, and cultural practices—were not included due to data unavailability. The exclusion of these variables may limit the explanatory power of the models.

Model Generalizability

The machine learning models were trained and evaluated on historical data from specific regions and time periods. As a result, their predictive performance may not fully generalize to regions with substantially different socio-economic or environmental conditions, or to future scenarios characterized by unforeseen changes in waste generation patterns.

Outlier Handling

While outlier removal was performed to enhance data quality, this process may have inadvertently excluded valid but extreme cases of MSW generation. Such exclusions could limit the models' ability to predict rare or exceptional events.

Interpretability

Ensemble models such as Random Forest and XGBoost, while highly accurate, are often considered "black box" methods due to their complex internal structures. This can make it challenging to interpret the specific contributions of individual features to the predictions, which may be a limitation for policy-making and stakeholder communication.

2.3. OVERVIEW OF DATASET

PROPOSED METHODOLOGY: The proposed methodology for forecasting Municipal Solid Waste (MSW) generation integrates data preprocessing, feature transformation, and the development of multiple machine learning regression models to assess and compare predictive performance. The following subsections describe each stage in detail.

Table 1 : Comparison of different Models for MSW Forecasting.

MODEL	TYPE	BASIC PRINCIPLE	KEY STRENGTHS	LIMITATIONS	IDEAL USE CASES
Linear Regression	Statistical	Models the relationship between features and the target as a linear function.	- Simple to implement and understand - Very fast to train and predict - Useful for baseline performance	- Assumes a linear relationship between inputs and output - Sensitive to outliers - Poor performance with complex patterns	Models the relationship between features and the target as a linear function.
Random Forest Regressor	Ensemble (ML)	Builds multiple decision trees on random subsets of data and averages their outputs.	- Handles nonlinear data well - Robust to outliers and noise - Reduces overfitting compared to single trees	- Less interpretable than linear models - Larger memory usage - Slower predictions on large datasets	Builds multiple decision trees on random subsets of data and averages their outputs.

MLP Regressor	Neural Network (ML)	Multi-layer feedforward network trained using backpropagation.	- Can learn complex patterns - Highly flexible architecture - Scalable to larger datasets	- Sensitive to hyperparameters - Requires significant tuning and data - Black-box nature	Multi-layer feedforward network trained using backpropagation.
XGB Regressor	Gradient Boosting (ML)	Sequentially builds trees where each corrects errors from the previous one.	- State of the art accuracy - Handles missing values internally - Good performance on tabular and sparse data	- Computationally expensive - Requires careful hyperparameter tuning - Less intuitive for stakeholders	Sequentially builds trees where each corrects errors from the previous one.

2.4. DATA COLLECTION AND UNDERSTANDING

The dataset used in this study was obtained from [source], comprising multiple features including:

- *Population* (number of people)
- *Gross Domestic Product* (GDP) per country
- *Income Classification* (High, Upper-Middle, Lower-Middle, Low)
- *Total MSW generated annually* (in tons)

These variables serve as socioeconomic and demographic indicators that significantly influence waste generation patterns.

2.5. DATA PREPROCESSING

Data preprocessing is a critical step in the development of robust and accurate machine learning models. For this study, a systematic approach was adopted to ensure the quality and suitability of the dataset for municipal solid waste (MSW) forecasting. The following procedures were implemented:

Data Acquisition

The primary dataset, titled "dataset - waste forecasting set.csv", was imported using the pandas library. This dataset comprised records of MSW generation, population, GDP, and income group classifications for various regions.

Data Filtering and Outlier Removal

To enhance data quality and minimize the influence of extreme values, the following filtering criteria were applied:

Population: Records with population values exceeding a predefined threshold were excluded.

GDP: Entries with GDP values above a set threshold were removed to eliminate outliers.

MSW Generation: Data points with exceptionally high MSW values were filtered out.

These steps ensured that the dataset reflected realistic and representative conditions for model training.

Categorical Variable Encoding

The income group attribute (income_id), a categorical variable, was transformed into a numerical format using one-hot encoding. This process generated binary columns for each income group (e.g., income_id_LIC, income_id_LMC, income_id_UMC), facilitating their use in machine learning algorithms.

Feature Normalization

All numerical features, including population, GDP, and MSW generation, were normalized to a [0, 1] range using the MinMaxScaler from scikit-learn. Normalization ensured that each feature contributed proportionately to the model and improved the convergence of learning algorithms.

Feature and Target Selection

The final set of input features comprised:

- Population
- GDP
- Encoded income group columns (income_id_LIC, income_id_LMC, income_id_UMC)
- The target variable for prediction was MSW generation (msw).

Train-Test Split

The preprocessed dataset was partitioned into training and testing subsets using an 80:20 split ratio. A fixed random seed was employed to ensure reproducibility of results.

In summary, the data preprocessing pipeline ensured that the dataset was clean, free from extreme outliers, and appropriately structured for machine learning. These steps established a solid foundation for the subsequent development and evaluation of predictive models for municipal solid waste forecasting.

2.5.1. HANDLING MISSING AND INCONSISTENT VALUES

To ensure data consistency:

- Columns such as population, GDP, and MSW were explicitly converted to numeric types.
- Rows containing null or invalid values in any of the key predictive columns were removed.

This step ensures a clean and reliable dataset for further modeling. application.

,

2.5.2. OUTLIER REMOVAL USING QUANTILE FILTERING

To mitigate the influence of extreme values and enhance model robustness :

- The 95th percentile for population, GDP, and MSW was computed.
- Data points exceeding this threshold were filtered out, ensuring only representative distributions were modeled.

This step helps prevent high-leverage points from skewing model predictions.

2.6. FEATURE ENGINEERING

Feature engineering played a crucial role in optimizing the dataset for accurate municipal solid waste (MSW) forecasting. The process began with the identification of key variables that significantly influence MSW generation. Based on domain knowledge and data availability, the primary features selected were population, gross domestic product (GDP), and income group classification.

Population and GDP, both continuous variables, were retained in their numerical form. The income group attribute, originally categorical, was transformed using one-hot encoding to create separate binary columns for each income group category. This allowed the machine learning models to effectively interpret and utilize the socio-economic distinctions present in the data.

To ensure that all features contributed proportionately during model training, numerical variables were normalized to a common scale using MinMaxScaler. This step was essential for improving the convergence and stability of the learning algorithms.

Through these targeted feature engineering techniques, the dataset was refined to maximize the predictive power of the models, enabling more reliable and robust MSW generation forecasts.

2.6.1. CATEGORICAL ENCODING

The categorical feature income_id was one-hot encoded into binary indicators (income_id_LIC, income_id_LMC, and income_id_UMC) to make it suitable for machine learning models. High-Income Countries (HIC) were used as the baseline class.

2.6.2. FEATURE SCALING

All numeric features (population, GDP, MSW) were normalized to a [0, 1] range using MinMaxScaler. This step standardizes input magnitudes, which is particularly essential for gradient-based models like MLP.

2.7. MODEL DEVELOPMENT

Four machine learning regression models were developed for comparative analysis :

1) Linear Regression (LR) : Acts as a baseline model assuming a linear correlation between input features and MSW generation.
2) Random Forest Regressor (RFR) : An ensemble method that constructs multiple decision trees and averages their predictions to reduce overfitting and improve accuracy.
3) XGBoost Regressor (XGBR) : A boosting-based ensemble technique that sequentially builds trees, learning from the residuals of previous ones, thus improving accuracy and handling complex patterns.
4) Multi-Layer Perceptron Regressor (MLP) : A deep learning model capable of capturing non-linear patterns through interconnected layers and neurons. The MLP was trained using backpropagation with the Adam optimizer.

Each model was implemented using scikit-learn or XGBoost with default parameters unless otherwise specified. The random seed was fixed (random_state=42) to ensure reproducibility.

2.8. MODEL TRAINING AND TESTING

Each model was trained using the training data X_train and y_train. Model-specific details are as follows :

- Linear Regression : Fit using ordinary least squares on the normalized features.
- Random Forest : Trained with default hyperparameters and 100 trees, using bootstrapped datasets.
- XGBoost : Utilized tree-based gradient boosting with a learning objective of regression (reg:squarederror).
- MLP Regressor : A neural network with one or more hidden layers and ReLU activations. It was trained using the Adam optimizer with a maximum of 1000 iterations to ensure convergence.

All models were implemented using the scikit-learn and xgboost libraries. The training process involved automatic adjustment of internal weights (in the case of MLP and XGBoost) to minimize loss functions using training data.

The dataset was divided into training and testing sets using an 80/20 split :

- Training Set (80%) : Used to fit model parameters.
- Testing Set (20%) : Used to evaluate model generalization performance.

2.9. MODEL EVALUATION

Model performance was assessed using two primary metrics :

- Mean Squared Error (MSE) :

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

- Coefficient of Determination (R^2 Score) :

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

To enhance interpretability and reduce noise, predictions were visualized using a scatter plot of actual vs. predicted values after removing the top 5% of extreme MSW values. The actual values were sorted and rounded for clearer presentation.

2.10. VISUALIZATION

A scatter plot was generated to depict the alignment between actual and predicted MSW values (post-filtering). The plot highlights the distribution, variance, and accuracy of predictions made by the baseline model. This visualization supports quantitative metrics and aids in qualitative validation of model performance.

2.11. MODEL ARCHITECTURE

The architecture followed for forecasting Municipal Solid Waste (MSW) generation is composed of systematic steps focusing on preprocessing, encoding, scaling, data partitioning, and model training. The steps are designed to ensure optimal data quality and model performance across different learning algorithms.

1) Outlier Filtering Threshold
    - *Quantile used :* 0.95
    - *Rationale*
        a) Extreme values can skew the distribution and adversely affect the learning process, especially in regression-based models.
        b) Therefore, the top 5% of values in the population, GDP, and MSW features were excluded from the dataset.
    - *Impact*
        a) This preserves the lower 95% of the data distribution, ensuring a more stable training process.
        b) Reduces variance and helps in avoiding model overfitting to outliers.
2) Feature Encoding
    - *Encoding technique :* One-Hot Encoding
    - *Parameter :* drop_first = True
    - *Application :* Categorical feature income_id was encoded into binary variables -
        a) income_id_LIC (Low-Income Countries)
        b) income_id_LMC (Lower-Middle-Income Countries)
        c) income_id_UMC (Upper-Middle-Income Countries)
    - *Why drop the first?*
        a) Prevents multicollinearity, where one encoded variable can be linearly predicted from the others.
        b) For example, if the original variable has 4 categories, creating 4 binary columns would result in a linear dependency. Dropping one ensures only 3 independent variables remain.

3) Feature Normalization

- *Scaler used :* MinMaxScaler()
- *Normalized features :* population, gdp, and msw
- *Purpose*
    a) Brings all numerical features into a uniform range [0, 1].
    b) Ensures that variables with larger magnitudes (e.g., GDP in billions) do not disproportionately influence model training.
- *Importance :* Particularly crucial for models like neural networks and gradient boosting where scale impacts convergence.

4) Train-Test Split

- *Test size :* 0.20 (i.e., 20% of data used for testing)
- *Training size :* 0.80 (i.e., 80% of data used for training)
- *Random seed :* random_state=42
- *Purpose*
    a) Guarantees reproducibility of experiments.
    b) Ensures that both training and testing datasets are representative of the original data distribution

5) Model Training : Four distinct regression models were trained using the same training set –

- *Linear Regression*
    a) Acts as the baseline model.
    b) Captures linear relationships between features and MSW.
    c) Fast training, interpretable coefficients.

- *Random Forest Regressor*
    a) Ensemble-based method using decision trees.
    b) Captures non-linear relationships and interactions.
    c) Robust to noise and performs automatic feature selection.

- *XGBoost Regressor*
    a) Gradient-boosting model optimized for performance.
    b) Capable of handling missing values and regularization.
    c) Often superior in structured data scenarios.

- *MLP Regressor (Neural Network)*
    a) A feed-forward artificial neural network.
    b) Trained using backpropagation with up to 1000 iterations.
    c) Can model complex non-linear functions.

6) Model Evolution
    - *Evaluation Metrics*
        a) Mean Squared Error (MSE)
        b) R-squared Score (R^2)
    - *Process*
        a) Predictions were generated on the test set for all models.
        b) MSE and R^2 were calculated for performance comparison.
    - A cleaned scatter plot was produced (removing top 5% of predictions) to visualize actual vs predicted values for the linear model.

7) Visualization
    - A scatter plot was generated using matplotlib for Linear Regression results, comparing the actual vs predicted MSW values after excluding extreme prediction errors (top 5% quantile).
    - Rounded X-axis values enhance clarity.
    - Grid lines and rotated labels improve readability.

CHAPTER – 3: IMPLEMENTATION AND RESULTS

3.1. LIBRARIES AND REQUIREMENTS

The models were implemented and trained using Python in a local environment with CPU support. Several libraries were employed for data handling, visualization, and machine learning to build, train, and evaluate the forecasting models.

1) Data Handling
    a) *pandas* – Utilized for reading the dataset (read_csv), data cleaning (handling missing values), renaming columns, filtering extreme values, and one-hot encoding categorical variables. Pandas DataFrames facilitated efficient manipulation of tabular data.
    b) *numpy* – Used for numerical operations such as array manipulation, sorting, and mathematical calculations necessary for data preprocessing andvisualization.

2) Visualization
    a) *matplotlib.pyplot* – Used for generating scatter plots to visualize the correlation between actual and predicted MSW values, enabling graphical assessment of model performance.

3) Machine Learning
    a) *sklearn.model_selection* – Used to split the dataset into training (80%) and testing (20%) subsets, ensuring that the model is evaluated on unseen data to assess generalization.
    b) *sklearn.linear_model* – Implemented the Multiple Linear Regression model to establish a linear relationship between features (population, GDP, income group) and the target variable MSW.
    c) *sklearn.preprocessing* – Applied MinMaxScaler to normalize numeric features (population, GDP, and MSW) to a 0–1 scale, which helps models converge faster and prevents dominance by variables with large scales.
    d) *sklearn.metrics* – Employed mean_squared_error and r2_score to quantitatively evaluate model accuracy.

- *Mean Squared Error (MSE)* : Represents the average squared differences between predicted and actual MSW values.
- *R-squared (R^2)* : Indicates the proportion of variance in MSW explained by the model, reflecting the goodness of fit.

e) *sklearn.ensemble* – Used RandomForestRegressor to capture nonlinear relationships and interactions between features with ensemble learning.

f) *xgboost* – Used XGBRegressor for gradient boosting, providing an optimized implementation for predictive accuracy.

g) *sklearn.neural_network* – Used MLPRegressor for modeling complex nonlinear relationships using a multi-layer perceptron (neural network) architecture.

3.2. RESULTS

The models were evaluated on the testing dataset. Performance metrics obtained were :

MODEL	MEAN SQUARED ERROR (MSE)	R-SQUARED (R^2) SCORE
Linear Regression	0.0043	0.6712
Random Forest	0.0026	0.7897
XGBoost	0.0024	0.8032
MLP Regressor	0.0031	0.7421

The XGBoost Regressor achieved the best performance, with the lowest MSE and highest R^2, demonstrating its suitability for MSW forecasting with the given features.

3.3. VISUALIZATION OF RESULTS

To visually assess prediction quality, a scatter plot was created comparing the actual MSW values with the predicted MSW values from the linear regression model (after filtering out the top 5% of extreme values to reduce noise).

The plot showed that the majority of points were close to the ideal diagonal line, indicating good prediction accuracy and model fit.

CHAPTER – 4: CONCLUSION AND FUTURE SCOPE

4.1. CONCLUSION

This research focused on developing and evaluating predictive models to forecast Municipal Solid Waste (MSW) generation using socioeconomic indicators such as population, Gross Domestic Product (GDP), and income classification. The dataset underwent rigorous preprocessing including handling missing values, filtering extreme outliers beyond the 95th percentile, one-hot encoding of categorical income groups, and normalization of numerical features. These preprocessing steps ensured the data's suitability for training robust regression models.

Four machine learning algorithms were employed: Multiple Linear Regression, Random Forest Regressor, Extreme Gradient Boosting (XGBoost), and Multi-Layer Perceptron (MLP) Regressor. These models were trained and tested on a split dataset, with performance assessed using Mean Squared Error (MSE) and R-squared (R^2) metrics.

Among the models, XGBoost achieved the best predictive accuracy, reflecting its strength in capturing nonlinear interactions and complex feature relationships within the data. Random Forest and MLP models also demonstrated competitive performance, outperforming the linear regression baseline, which suggests that nonlinear models are more suitable for MSW forecasting tasks. The results highlight that integrating advanced machine learning techniques can substantially improve the precision of waste generation predictions over traditional linear methods.

Accurate MSW forecasting is critical for urban planners and policymakers to optimize resource allocation, design sustainable waste management systems, and mitigate environmental impacts associated with waste generation. This study validates the effectiveness of machine learning models in achieving these objectives and provides a data-driven foundation for informed decision-making.

4.2. FUTURE SCOPE

While the current work provides a solid framework and promising results, several avenues exist to extend and enhance the forecasting system :

1) Inclusion of Additional Relevant Features : The present model considers population, GDP, and income classification. However, waste generation is influenced by many other factors including urbanization rate, industrial activity, consumer behavior, waste management policies, education levels, and technological adoption. Incorporating these variables could improve the model's comprehensiveness and accuracy.
2) Temporal and Longitudinal Modeling : MSW generation exhibits temporal trends and seasonal variations that static regression models cannot capture effectively. Employing time series analysis, Long Short-Term Memory (LSTM) networks, or other recurrent neural networks would enable the model to learn temporal dependencies and provide more dynamic forecasts over time.
3) Geospatial and Regional Analysis : Waste generation patterns vary spatially due to factors such as population density, economic development, and cultural practices. Integrating geospatial data through Geographic Information Systems (GIS) can allow the development of location-specific models and support targeted interventions tailored to regional needs.
4) Advanced Ensemble and Hybrid Modeling : Combining multiple models through ensemble techniques or hybrid approaches (e.g., blending tree-based models with neural networks) can capture different aspects of data variability, potentially improving predictive performance and robustness against overfitting.
5) Real-Time Data and IoT Integration : The adoption of smart waste management systems equipped with sensors provides opportunities to integrate real-time data into forecasting models. This would facilitate adaptive, near real-time waste prediction and responsive management strategies.
6) Model Explainability and Interpretability : To build trust and practical usability for policymakers, future work should emphasize explainability. Techniques like SHAP values, LIME, or feature importance analysis can reveal the contribution of individual predictors, guiding targeted policy measures and resource prioritization.
7) Cross-Country and Cross-Cultural Validation : Testing and adapting the models across diverse countries or regions with varying economic statuses and waste management infrastructures will assess generalizability and guide localized model calibration.

BIBLIOGRAPHY

[1] Hasan, M. K., Al-Ani, A., Anwar, F., Al-Rahmi, W. M., & Shaikh, A. (2022). Smart waste management and classification system for smart cities using deep learning. International Conference on Business Analytics for Technology and Security, 1–7.

[2] Olawade, D. B., Adepoju, A. O., Alabi, A. O., & Ogunleye, J. O. (2024). Smart waste management: A paradigm shift enabled by artificial intelligence. Waste Management Bulletin, 2, 244–263.

[3] Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., & Garcia-Zapirain, B. (2023). Automated waste-sorting and recycling classification using artificial neural network and features fusion: A digital-enabled circular economy vision for smart cities. Multimedia Tools and Applications, 82, 39617–39632.

[4] Kontokosta, C. E., Hong, B., & Johnson, N. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Computers, Environment and Urban Systems, 70, 151–162.

[5] Adedeji, O., & Wang, Z. (2019). Intelligent waste classification system using deep learning convolutional neural network. Procedia Manufacturing, 35, 607–612.

[6] Chen, X. (2022). Machine learning approach for a circular economy with waste recycling in smart cities. Energy Reports, 8, 3127–3140.

[7] Chu, Y., Wang, Z., Jiang, B., & Zhang, Q. (2018). Multilayer hybrid deep-learning method for waste classification and recycling. Computational Intelligence and Neuroscience, Article ID 5060857.

[8] Belsare, K. S., & Goraya, M. S. (2022). A review of IoT-based intelligent waste management frameworks employing machine learning for smart city applications. Mobile Computing and Sustainable Informatics, 797–817.

[9] Ahmed, S., et al. (2022). Deep learning-based model for predicting municipal waste conditions using smart bins. International Journal of Environmental Research and Public Health.

[10] Tran, A. K., Nguyen, V. D., Le, M. T., & Pham, T. H. (2020). University-level waste management system leveraging IoT and machine learning. Wireless Communications and Mobile Computing, Article ID 6138637.