Waste Management and Recycling in India – Predictive Modeling Report

1. Introduction

Urban sustainability is one of the most pressing challenges of the 21st century. Rapid urbanization in India has led to a significant increase in waste generation, putting pressure on municipal waste management systems.
The **goal of this project** is to build a predictive model to estimate the **Recycling Rate (%)** of waste in Indian cities based on multiple environmental, infrastructural, and socio-economic factors.
By accurately predicting recycling rates, cities can:

- Optimize landfill usage
- Improve waste segregation policies
- Allocate resources efficiently
- Promote public awareness campaigns

2. Methodology

2.1 Data Preprocessing

- **Missing values**:
    - Numeric columns filled with median values.
    - Categorical columns filled with the mode.
- **Latitude/Longitude extraction**:
  Parsed from `Landfill Location (Lat, Long)` column.
- **Data type conversions**:
  Year converted to categorical for better trend modeling.

2.2 Feature Engineering

To capture more meaningful patterns, several new features were created:

1. **Waste_per_Density** = Waste Generated / (Population Density + 1)
2. **Landfill_Utilization** = Waste Generated / (Landfill Capacity + 1)
3. **Efficiency_Awareness** = Municipal Efficiency Score × Awareness Campaigns Count
4. **Cost_per_Ton_Generated** = Cost of Waste Management / Waste Generated
5. **Recycling_Landfill_Ratio** = Recycling Rate / Landfill Capacity
6. **Landfill_Distance_to_Center** = Euclidean distance from the landfill to India's geographic center (lat=22.5937, lon=78.9629)

2.3 Encoding Categorical Variables

- **Target Encoding**: Applied to `Waste Type` and `Disposal Method` to retain relationship with the target variable.
- **One-Hot Encoding**: Applied to `City/District` and `Landfill Name`.

2.4 Model Selection

- **Model Chosen**: Random Forest Regressor
  - Handles both numeric and categorical features after encoding.
  - Resistant to multicollinearity and outliers.
  - Can capture non-linear relationships.

3. Results

3.1 Model Performance

Metric	Value
Train R^2	0.996
Test R^2	0.974
RMSE	2.684

3.2 Visualizations

- **Feature Importance**: Landfill Capacity, Waste Generated, and Efficiency_Awareness were among the most influential predictors.
- **Recycling Trends**: Higher recycling rates were observed in cities with greater awareness campaigns and better municipal efficiency.
- **Geospatial Insights**: Cities closer to central processing facilities tended to have better recycling efficiency.

4. Discussion

4.1 Challenges

- Large number of city dummy variables increased dimensionality — mitigated via encoding.
- Small risk of overfitting due to high R^2 scores — monitored via train/test split and will verify with cross-validation.

4.2 Limitations

- Dataset is simulated; real-world variations may differ.
- Some features like landfill age or weather patterns could further improve predictions.

4.3 Real-world Implications

- Helps municipalities forecast recycling needs and capacity.
- Can optimize landfill management and reduce environmental footprint.

4.4 Future Scope

- Incorporate real-time IoT waste bin data.
- Add socio-economic factors like income levels and education rates.
- Use geospatial clustering to group cities with similar waste patterns.

5. Conclusion

The predictive model for **Recycling Rate (%)** using waste management data from Indian cities achieved **97.4% accuracy (R^2)** on unseen test data.
Through careful preprocessing, feature engineering, and model selection, we have built a tool that can significantly aid municipal decision-making for sustainable urban development.