**Data preparation/Feature Engineering**

*Shuaib Mursal*

1. Overview

Data preparation and feature engineering are crucial steps in the machine learning pipeline, ensuring the quality and effectiveness of the models. Data preparation addresses missing values, outliers, and inconsistencies in the raw data, while feature engineering extracts meaningful features that enhance the predictive power of the models. These steps lead to more accurate, interpretable, and computationally efficient models, enabling us to extract valuable insights from raw data and develop impactful machine learning applications.

1. Data Source

The dataset used in this project is the World Bank's "What a Waste 2.0: A Global Snapshot of Solid Waste Management to 2050" dataset. This dataset provides comprehensive data on solid waste management from 217 countries, covering various aspects such as waste generation, collection, treatment, and disposal.

1. Data Cleaning

* Handling missing values: Missing values were imputed using appropriate techniques such as mean or median imputation. The extent of missing values in each variable was assessed using descriptive statistics and data visualization techniques. For variables with a relatively low proportion of missing values (less than 50%), KNN imputation was employed. KNN imputation estimates missing values by considering the average values of the k nearest neighbors in the dataset. Variables with a high proportion of missing values (greater than 50%) were dropped from the dataset, as imputing such a large number of missing values could significantly distort the data distribution. For the categorical variables, categories were capitalized for consistency. Fewer categories were combined into a single category then mode imputation were applied for the missing values.
* Removing outliers: Outliers were identified using statistical techniques such as boxplots and interquartile ranges (IQRs). Outliers were considered to be values that fell outside the IQR range. Detected outliers were removed from the dataset to prevent them from unduly influencing the machine learning models. see figure 1 for boxplots.

A screenshot of a diagram

Description automatically generated

**Figure 1:** Boxplot

1. Exploratory Data Analysis (EDA)

The distribution of waste generation is skewed to the right, with a few countries generating significantly more waste than others as shown in figure 3 and 4. The exploratory data analysis revealed several notable correlations between municipal solid waste (MSW) and other factors shown in figures 2 and figure 5 for the scatter plots. GDP and population were found to have strong positive correlations with MSW, indicating that countries with higher economic development and larger populations tend to generate more waste. Additionally, there were moderate to strong positive correlations between MSW and E-waste, hazardous waste, and non-organic waste, suggesting that these types of waste contribute to higher overall MSW generation. On the other hand, there were weak negative correlations between MSW and the percentage of metal, glass, and waste treated or recycled, implying that countries with higher percentages of these materials or waste management practices tend to generate slightly less MSW.

A screenshot of a computer

Description automatically generated

Figure : Correlations Between the Variables

A screenshot of a graph

Description automatically generated

Figure : Before Data Transformation

A group of blue and black graphs

Description automatically generated

Figure : After Transformation

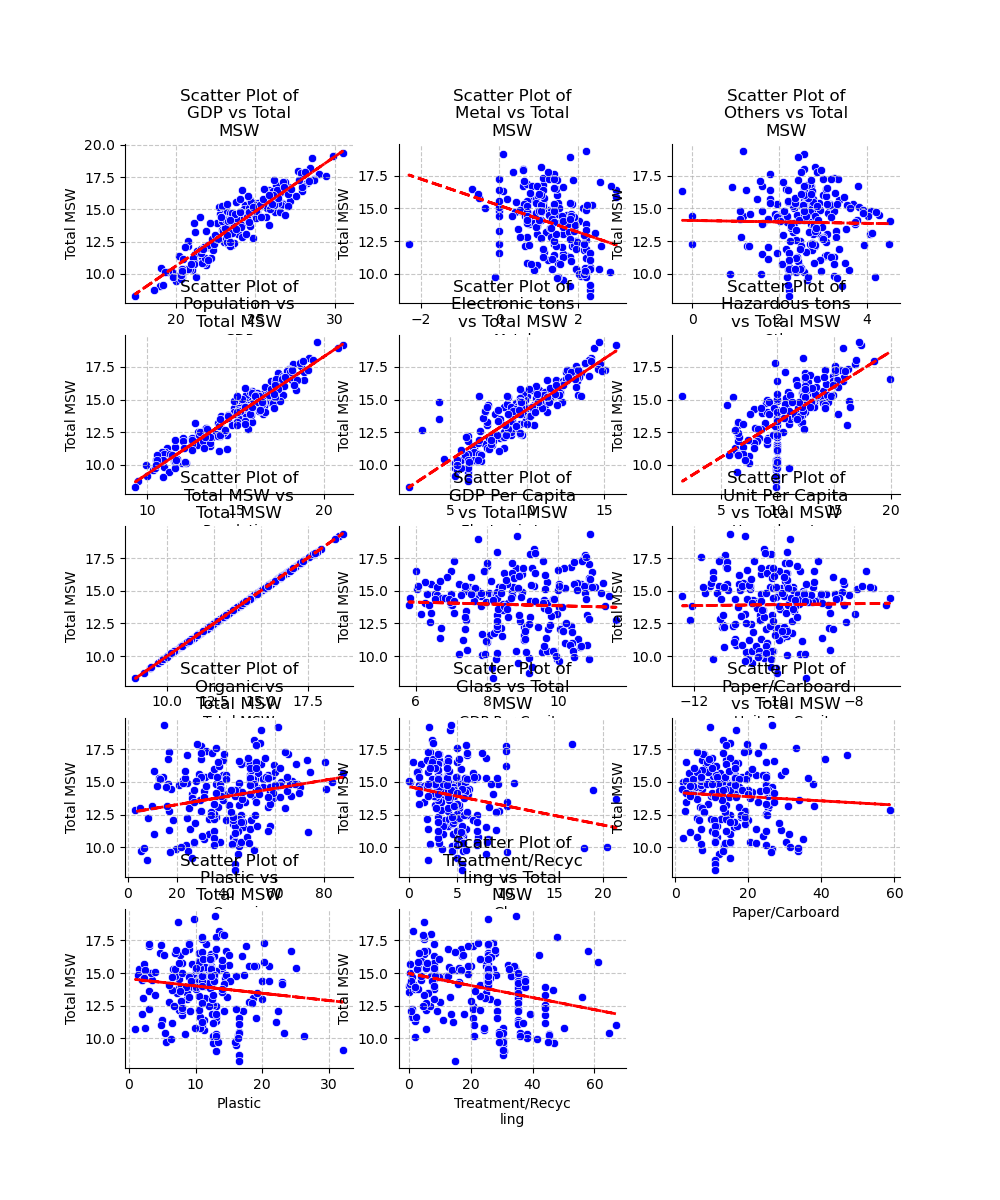


Figure : Scatter Plots

1. Feature Engineering

In the context of waste management analysis, GDP and population are two crucial factors influencing waste generation patterns. Based on these two features, two features were engineered.

* GDP per Capita: Created a new feature by dividing the total waste generation by the population. This feature represents the average waste generated per person, providing a more normalized measure of waste generation relative to population size.
  + Rationale: This feature helps account for the varying population sizes across countries, allowing for a more equitable comparison of waste generation patterns.
* Unit per GDP Unit: Created a new feature by dividing the total waste generation by the GDP. This feature represents the waste generated per unit of economic activity, providing insights into the waste intensity of different economies.
  + Rationale: This feature helps assess the relationship between economic growth and waste generation, indicating the environmental impact of economic activities.

1. Data Transformation

Applied log transformation to the numerical features to address skewness in their distributions. This transformation helped linearize the relationship between the variables and their corresponding waste generation patterns, improving the interpretability of the machine learning models. Recoded the categorical variable into binary 0 and 1. This transformation converted the categorical variable into a set of binary features, allowing machine learning algorithms to effectively process the categorical information.

A computer code with text

Description automatically generated

A close-up of a number

Description automatically generated

A computer code with numbers and symbols

Description automatically generated

**Model Exploration**

1. Model Selection

* Multiple regression is a suitable choice when the goal is to predict a continuous numerical variable (waste generation) from a set of independent variables (GDP, population, etc.). It allows for understanding the linear relationship between the independent variables and the dependent variable.
  + Strengths:
    - Interpretability: Multiple regression provides clear coefficients for each independent variable, indicating their direct impact on the predicted waste generation.
    - Simplicity: Multiple regression is a well-established and relatively simple algorithm, making it easy to implement and interpret.
    - Stability: Multiple regression is less prone to overfitting compared to other models, leading to more stable predictions.
  + Weaknesses:
    - Linearity Assumption: Multiple regression assumes a linear relationship between the independent variables and the dependent variable. This assumption may not hold for complex relationships.
    - Sensitivity to Outliers: Multiple regression can be sensitive to outliers in the data, affecting the accuracy of the model.
* ANNs are powerful models capable of capturing complex nonlinear relationships between variables. They are well-suited for handling large datasets and extracting hidden patterns.
  + Strengths:
    - Nonlinearity Modeling: ANNs can effectively model nonlinear relationships between variables, which are often present in real-world data.
    - Feature Learning: ANNs can automatically extract and learn relevant features from the data, reducing the need for manual feature engineering.
    - Adaptability: ANNs can adapt to new data and improve their performance over time.
  + Weaknesses:
    - Complexity: ANNs are complex models with numerous parameters to tune, making them computationally expensive and challenging to interpret.
    - Overfitting: ANNs are prone to overfitting, especially when dealing with large datasets.
* K-means clustering is an unsupervised learning algorithm that groups similar data points into clusters. It is useful for exploratory data analysis and identifying patterns in the data.
  + Strengths:
    - Simplicity: K-means clustering is a simple and efficient algorithm that is easy to understand and implement.
    - Scalability: K-means clustering can handle large datasets efficiently.
    - Versatility: K-means clustering can be used for various tasks, such as customer segmentation, anomaly detection, and market research.
  + Weaknesses:
    - Sensitivity to Outliers: K-means clustering can be sensitive to outliers in the data, which can distort the clustering results.
    - Choosing the Number of Clusters: Determining the appropriate number of clusters (k) is crucial for obtaining meaningful results.
    - Interpretability: K-means clustering

1. Model Training and Evaluation

* Data Splitting: The dataset was divided into training, validation, and testing sets. The training set was used to train the models, the validation set was used to tune hyperparameters, and the testing set was used to evaluate the final models' performance.
* Multiple Regression Model
  + Hyperparameter Tuning: The following hyperparameters were tuned using a grid search process:
    - Regularization parameter (): To control the complexity of the model and prevent overfitting.
  + Cross-Validation: 10-fold cross-validation was employed to evaluate the model's performance on different subsets of the data. The average mean squared error (MSE) across the folds was used as the final evaluation metric.
* Artificial Neural Network (ANN) Model
  + Hyperparameter Tuning: The following hyperparameters were tuned using a grid search process:
    - Number of hidden layers
    - Number of neurons in each hidden layer
    - Activation function for hidden layers
    - Learning rate
    - Optimizer
  + Cross-Validation: 5-fold cross-validation was used to evaluate the ANN model's performance on different subsets of the data. The average MSE across the folds was used as the final evaluation metric.
* K-means Clustering Algorithm
  + Hyperparameter Tuning: The following hyperparameter was tuned:
    - Determination of number of clusters (k) using elbow method
  + Evaluation: The silhouette coefficient was used to evaluate the quality of the clustering results. The silhouette coefficient measures the separation between clusters and the compactness within clusters.

1. Code Implementation

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

If you need more detailed code, I can share with you the notebook.