FORECASTING MUNICIPAL SOLID WASTE GENERATION IN MALAYSIA

NUR HIDAYAH BINTI AHMAD SHAFII

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2025

FORECASTING MUNICIPAL SOLID WASTE GENERATION IN MALAYSIA

NUR HIDAYAH BINTI AHMAD SHAFII

RESEARCH REPORT SUBMITTED TO THE FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY UNIVERSITI MALAYA, IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF DATA SCIENCE

FACULTY OF COMPUTER SCIENCE & INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
KUALA LUMPUR

2025

*(Please delete this part) Depending on the medium of thesis/dissertation, pick either the English or Bahasa Malaysia version and delete the other.

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: (I.C/Passport No:)

Matric No:

Name of Degree:

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Field of Study:

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the Universiti Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date: 24 Jan 2025

Name: PROF. TS. DR. RAFIDAH MD NOOR

Designation: DEPARTMENT OF COMPUTER SYSTEM & TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND
INFORMATION TECHNOLOGY
UNIVERSITI MALAYA
50603 KUALA LUMPUR.

UNIVERSITI MALAYA

PERAKUAN KEASLIAN PENULISAN

Nama: (No. K.P/Pasport:)

No. Matrik:

Nama Ijazah:

Tajuk Kertas Projek/Laporan Penyelidikan/Disertasi/Tesis ("Hasil Kerja ini"):

Bidang Penyelidikan:

Saya dengan sesungguhnya dan sebenarnya mengaku bahawa:

(1) Saya adalah satu-satunya pengarang/penulis Hasil Kerja ini;
(2) Hasil Kerja ini adalah asli;
(3) Apa-apa penggunaan mana-mana hasil kerja yang mengandungi hakcipta telah dilakukan secara urusan yang wajar dan bagi maksud yang dibenarkan dan apa-apa petikan, ekstrak, rujukan atau pengeluaran semula daripada atau kepada mana-mana hasil kerja yang mengandungi hakcipta telah dinyatakan dengan sejelasnya dan secukupnya dan satu pengiktirafan tajuk hasil kerja tersebut dan pengarang/penulisnya telah dilakukan di dalam Hasil Kerja ini;
(4) Saya tidak mempunyai apa-apa pengetahuan sebenar atau patut semunasabahnya tahu bahawa penghasilan Hasil Kerja ini melanggar suatu hakcipta hasil kerja yang lain;
(5) Saya dengan ini menyerahkan kesemua dan tiap-tiap hak yang terkandung di dalam hakcipta Hasil Kerja ini kepada Universiti Malaya ("UM") yang seterusnya mula dari sekarang adalah tuan punya kepada hakcipta di dalam Hasil Kerja ini dan apa-apa pengeluaran semula atau penggunaan dalam apa jua bentuk atau dengan apa juga cara sekalipun adalah dilarang tanpa terlebih dahulu mendapat kebenaran bertulis dari UM;
(6) Saya sedar sepenuhnya sekiranya dalam masa penghasilan Hasil Kerja ini saya telah melanggar suatu hakcipta hasil kerja yang lain sama ada dengan niat atau sebaliknya, saya boleh dikenakan tindakan undang-undang atau apa-apa tindakan lain sebagaimana yang diputuskan oleh UM.

Tandatangan Calon Tarikh:

Diperbuat dan sesungguhnya diakui di hadapan,

Tandatangan Saksi Tarikh:

Nama:

Jawatan:

FORECASTING MUNICIPAL SOLID WASTE GENERATION IN MALAYSIA

ABSTRACT

Accurate prediction of municipal solid waste (MSW) generation is crucial for Malaysia due to its heavy reliance on landfills for waste disposal. The present study compares three prediction algorithms (Multiple Linear Regression (MLR), Random Forest (RF), and Artificial Neural Networks (ANN)) to forecast MSW generation in Malaysian states that operates under Act 672. As previous studies provided limited comparisons of regression models and influential variables for waste generation, this research addresses these gaps by conducting a comprehensive analysis based on socioeconomic and demographic factors using the same dataset. Among the untuned models, the results show that MLR achieved the highest R^2 of 0.82 followed by RF (R^2 = 0.70) and ANN (R^2 = 0.58). After the hyperparameter optimization and cross-validation, the optimised RF model outperformed the others with improved R^2 of 0.85 and lowest RMSE of 7354.93. Its ability to handle complex relationships further established RF as the most accurate and reliable predictive model. The proposed model identified GDP per capita as the primary factor influencing solid waste generation, followed by crude death rate, fertility rate, and labor force rate. The predictive model is capable of forecasting the yearly average MSW generation for each state over the next five years.

Keywords: Municipal solid waste; Machine learning; Waste prediction; Socioeconomic and demographic factor

RAMALAN PENJANAAN SISA PEPEJAL PERBANDARAN DI MALAYSIA

ABSTRAK

Di Malaysia, ramalan penjanaan sisa pepejal perbandaran (MSW) penting kerana kebergantungan pada kaedah pengurusan sisa menerusi tapak pelupusan sampah amat tinggi. Kajian ini membandingkan tiga model ramalan (Regresi Berganda (MLR), Hutan Rawak (RF), dan Rangkaian Neural Tiruan (ANN)) untuk meramalkan penjanaan MSW di negeri-negeri Malaysia yang beroperasi di bawah Akta 672. Oleh sebab kajian perbandingan bagi model regresi dan pembolehubah yang berpengaruh untuk penjanaan sisa adalah terhad, penyelidikan ini menangani jurang ini dengan menjalankan analisis komprehensif berdasarkan faktor sosioekonomi dan demografi menggunakan set data yang sama. Keputusan kajjian menunjukkan bahawa MLR mencapai R^2 tertinggi sebanyak 0.82 diikuti oleh RF ($R^2 = 0.70$) dan ANN ($R^2 = 0.58$). Selepas pengoptimuman hiperparameter dan pengesahan silang, model RF yang ditala menunjukkan peningkatan R^2 sebanyak 0.85 dan RMSE terendah sebanyak 7354.93 berbanding model lain. Keupayaannya untuk mengendalikan perhubungan yang rumit seterusnya mengukuhkan RF sebagai model ramalan yang paling tepat dan boleh dipercayai. Model yang dicadangkan mengenal pasti KDNK per kapita sebagai faktor utama yang mempengaruhi penjanaan sisa pepejal, diikuti dengan kadar kematian kasar, kadar kesuburan dan kadar tenaga buruh. Model ramalan dalam kajian ini mampu meramalkan purata tahunan penjanaan MSW bagi setiap negeri dalam tempoh lima tahun akan datang.

Kata kunci: Sisa pepejal perbandaran; Model pembelajaran mesin; Ramalan sisa; Faktor sosioekonomi dan demografi

ACKNOWLEDGEMENTS

All praise and gratitude to Allah S.W.T who has blessed me with the opportunity to pursue my studies and granted me countless blessings throughout my life. His guidance and mercy have been the foundation of my strength and perseverance in completing this journey.

I extend my heartfelt thanks to my supervisor, Prof. Ts. Dr. Rafidah binti Md Noor for her guidance and invaluable insights throughout this thesis. Her support has been instrumental to my success.

My ultimate gratitude is dedicated to my beloved parents, Masrokah Ramelan and Ahmad Shafii Karim, for their warmth, patience, understanding and unwavering support. To my sister, Lailatul Fitriyah, whose opinions I deeply value and who often believes in me more than I believe in myself—thank you for your trust and encouragement. A heartfelt thanks to my brother, Galang Nukman Aiman, for his constant support and presence.

Not to be forgotten are my friends, especially Premi, Ummul Farah, and Anies Zaidi, whose invaluable feedback and constant encouragement they bring to my life.

A special thanks to the data science community, whose open resources and shared knowledge have made this research possible. Your contributions are truly invaluable.

Lastly, this project has been an incredible experience, and I sincerely thank everyone who directly or indirectly contributed to this remarkable journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

ANN	:	Artificial Neural Network
DLTSF	:	Deep Learning Time Series Forecasting
ETSX	:	Error-Trend-Seasonality with external variables
GDP	:	Gross Domestic Product
LSTM	:	Long Short-Term Memory
MAE	:	Mean Absolute Error
MLR	:	Multiple Linear Regression
MSW	:	Municipal Solid Waste
PSO	:	Particle Swarm Optimization
RF	:	Random Forest
RMSE	:	Root Mean Squared Error
SDGs	:	Sustainable Development Goals
SWCorp	:	Solid Waste Management and Public Cleansing Corporation

LIST OF APPENDICES

CHAPTER 1: INTRODUCTION

1.1 Background

Municipal waste management (MSW) is a critical issue especially in developing countries. There are several types of waste management practices are used in Malaysia which are recycling, composting, incineration, inert landfill, sanitary landfill, and other disposal sites (Wahidah & Ghafar, 2017). However, landfill disposal remains the most common method due to its cost-effectiveness and simplicity (Lim et al., 2016). Malaysia experienced challenges to manage its increasing waste generation as most of the waste will ultimately be deposited in landfills. With the current trends, landfill capacities are rapidly approaching their limits. Ineffective waste management not only causes the spread of foodborne diseases but also contributes 30% of global greenhouse gas emissions, consumes up to 70% of freshwater withdrawals, and generates 10-90% of air pollutants (Gatto, 2024). It is not only through methane gas production in landfills that is 21 times more potent than carbon dioxide in terms of global warming potential (Mahasan, 2023) but also through the wasted resources and energy used in food production, processing, and transportation (Liegeard & Manning, 2020).

The SDGs are a comprehensive collection of 17 interconnected objectives established by the United Nations (Nik Mahdi et al., 2023). The SDG Dashboards and Trends shown in Figure 1.1 is designed to track global issues and promote sustainable development in Malaysia. Each goal is overseen by a dedicated ministry responsible for planning, implementation, monitoring, and reporting on its performance. While Malaysia has successfully achieved SDG 1: No Poverty, progress on SDG 7 (Affordable and Clean Energy), SDG 12 (Responsible Consumption and Production), SDG 13 (Climate Action), SDG 16 (Peace, Justice, and Strong Institutions), and SDG 17 (Partnerships for the Goals) has stagnated with major challenges remaining to achieve the goal. The effort to manage waste generation is aligned with SDG 12: Responsible Consumption and Production.

Moreover, it also supports SDG 13: Climate Action in mitigating climate-related risks by reducing methane emissions and improving resource efficiency.

(United Nations, 2024)

Figure 1.1 SDG Dashboards and Trends

Several studies have been conducted to investigate the application of machine learning with socioeconomic and demographics factors in MSW prediction. When applied to MSW prediction, machine learning helps to uncover complex patterns and relationships within data that are often overlooked by traditional predictive models. While machine learning models have strong potential for MSW predictions, the performance depends heavily on high-quality input data and appropriate model selection. Furthermore, the complexity of advanced models requires significant computational resources and expertise for optimisation to address challenges such as data noise, irrelevant variables, and regional variability (Hoy et al., 2022; Niu et al., 2021). Despite these challenges, the potential of machine learning to improve the accuracy and reliability of MSW prediction continues to be extensively explored.

Jereme et al. (2016) highlight that despite Malaysia's high per capita waste generation, waste treatment in the country remains limited, with a lack of innovative strategies compared to developed nations. Addressing these challenges and making progress towards the SDGs will require multi-stakeholder collaboration, innovative solutions, and a shift towards more sustainable consumption patterns. This could include improved food waste tracking and management systems, enhanced public education campaigns, and the development of circular economic approaches in the food sector. By focusing on these areas, Malaysia can work towards meeting its SDG commitments while addressing the pressing issue of food waste.

1.2 Problem Statement

In Malaysia, the daily generation of MSW rates exceeds 39,000 tons daily with over 30% being food waste. Households have been identified as the largest contributors to food waste in Malaysia (Jereme et al., 2016; Ng et al., 2023). MSW management is a complex challenge associated with various factors such as socioeconomic and demographic factors. Araiza-Aguilar et al. (2020), Yusoff et al. (2018), Azadi & Karimi-Jashni (2016) and Ghinea et al. (2016) emphasize the role of demographic factors such as urbanization and population as key drivers of MSW generation. However, these studies do not consider socioeconomic variables like GDP as influential factors in MSW generation.

In contrast, Dissanayaka & Vasanthapriyan (2019) analyzed factors such as GDP growth rate, crude birth rate, and total population among other variables and found that socioeconomic variables influence MSW generation in Sri Lanka. Interestingly, their study revealed a negative correlation between population and waste generation in Sri Lanka. This suggests that in some cases densely populated areas may have better waste management systems or more conservative consumption patterns. Despite this, densely populated areas often face more significant challenges in managing food waste due to the increased volume generated (Adelodun & Choi, 2020; Bharadi et al., 2022) although the financial and environmental costs of wasting food are well well acknowledged (Phooi et al., 2022).

While demographic factors are well-established predictors of MSW generation, their influence varies based on regional and economic contexts. Similarly, socioeconomic factors also shape waste generation patterns, though their impacts are less consistently documented. These observations highlight the need for further research to integrate socioeconomic and environmental factors into predictive models for waste generation.

The recent advancements in predictive modeling using machine learning gained popularity in waste management literature due to its ability to discover hidden patterns in forecasting waste generation. However, the application of these techniques in waste management remains underexplored in many regions, including Malaysia. While research by Nasir et al. (2023) focused on time series forecasting of solid waste in specific states, and Hoy et al. (2022) evaluated a Bayesian-optimized neural network with ensemble learning, these examples are exceptions rather than the norm. Earlier efforts, such as those by Yusoff et al. (2018) using neural networks and Zulkipli et al. (2018) utilizing integrated system dynamics provide valuable insights but are insufficient to fully uncover the potential of modern machine learning models in Malaysia.

In the previous studies related to predictive models in waste management, research in Malaysia has primarily focused on time-series approaches or advanced machine learning algorithms especially ANN. This focus has left a gap in exploring the relative performance of simpler models like MLR and RF despite their potential. Dissanayaka & Vasanthapriyan (2019) highlight the potential of ANN, MLR and RF in forecasting waste trends. Araiza-Aguilar et al. (2020)demonstrated that MLR achieved high predictive accuracy. Their model reached an R^2 of 0.975 with a 7.7% mean absolute error when modeling MSW generation using variables like population density and migration rate in Mexico. In contrast, A. Kumar et al. (2018) reported moderate accuracy for MLR with R^2 values of 0.782 for biodegradable and 0.676 for non-biodegradable MSW prediction using socioeconomic variables. These findings show variability in model performance across different contexts and highlight the need to explore a range of predictive models for comparison within the Malaysian context.

Therefore, this research aims to evaluate the predictive performance of three machine learning models namely MLR, RF and ANN. Given the scale of MSW generation in Malaysia and its unique demographic and socioeconomic characteristics, it is important to identify the most effective method for predicting waste generation. The findings will contribute to data-driven waste management strategies and support Malaysia's commitment to achieve the SDGs, particularly those related to sustainable cities and responsible consumption.

1.3 Research Questions

The research questions for this study

i. What are the key variables that influence the amount of waste generation?
ii. How accurate and reliable can Multiple Linear Regression, Random Forest, and Artificial Neural Networks predict solid waste generation?

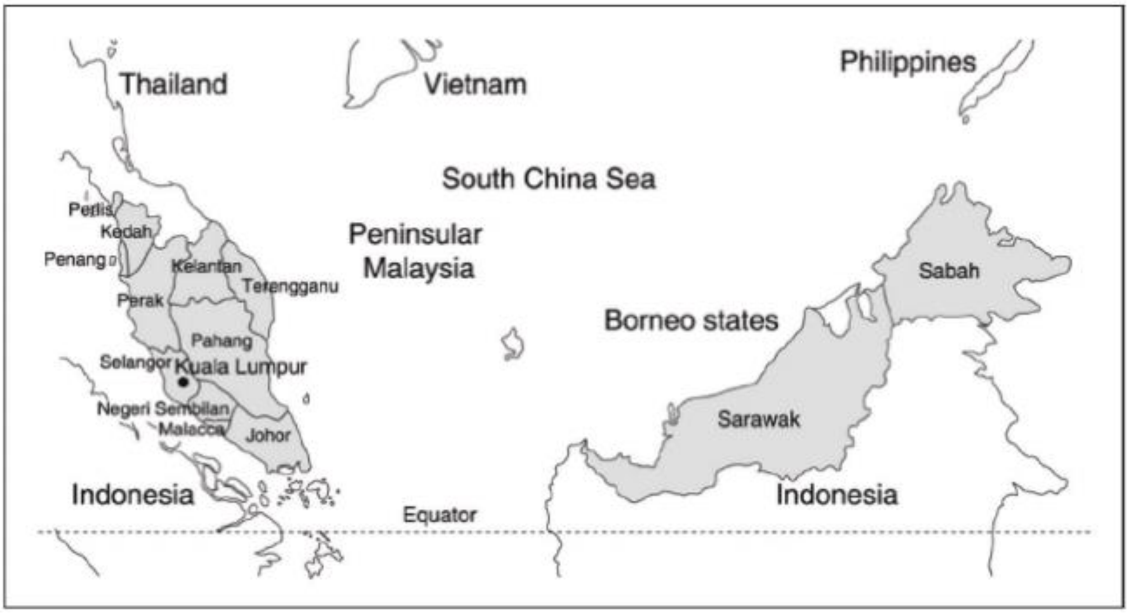
1.4 Research Objectives

The objectives of this study

i. To identify and analyze influential variables that affect the amount of waste generation.
ii. To evaluate the predictive performance of Multiple Linear Regression, Random Forest, and Artificial Neural Networks models for solid waste generation.

1.5 Research Scope

Malaysia is one of the leading economies in the Southeast Asian countries. It has a diverse multicultural society shaped by British colonial policies. As illustrated in Figure 1.2 Malaysia comprises three federal territories of W.P Kuala Lumpur, W.P Labuan, and W.P Putrajaya, and 13 states of Johor, Kedah, Kelantan, Malacca, Negeri Sembilan, Pahang, Penang, Perak, Perlis, Selangor, Sabah, Sarawak, and Terengganu. The country is geographically divided into two main regions by the South China Sea: Peninsular Malaysia and East Malaysia on the island of Borneo (Daud, 2021).



(Daud, 2021)

Figure 1.2 Maps of Malaysia

This study focuses on states and federal territory that operate under the Solid Waste and Public Cleansing Management Act 2007 (Act 672). Act 672 was introduced to standardize and regulate solid waste management and public cleansing services. This law is administered by SWCorp Malaysia under the Ministry of Housing and Local Government. It provides a legal framework for waste management with the goal to increase efficiency, reduce environmental harm, and promote sustainable waste management practices. Under Act 672, the management of waste is privatized and handled by concessionaires like Alam Flora, SWM Environment, and E-Idaman to ensure compliance with federal standards. However, not all states in Malaysia have adopted Act 672 due to political, administrative, and jurisdictional reasons. Some states chose to retain autonomy over waste management due to concerns over privatization, the cost implications of federal intervention, and the belief that their existing waste management systems are effective. Only six states (Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perlis) and two federal territories (Wilayah Persekutuan Kuala Lumpur, Putrajaya) that operate under Act 672. However, there is no data on Putrajaya's waste as the city does not have its own landfill and its solid waste is disposed of at Tanjung 12, Selangor.

The dataset used in this research covers solid waste and recyclable waste generation from 2017 to 2021. These data were acquired from the SWCorp's archive website and the yearly statistics published by the Ministry of Housing and Local Government (KPKT). The socioeconomics and demographics factors were sourced from Department of Statistics Malaysia (DOSM) webpage. While solid waste generation data are reported monthly, socio-economic and demographic data are available only on an annual basis. To address these limitations, we assumed that monthly changes in socio-economic and demographic factors are minimal and treated them as constant throughout each year. All variables were then combined into a unified dataset for analysis.

In order to predict solid waste generation in Malaysia, the study focused on the predictive performance evaluation of three machine learning models MLR, RF and ANN. The analysis will use a dataset that includes variables such as GDP, GDP per capita, population size, urban and elderly population, fertility rates, household numbers, labor force participation, employment ratio, crude birth rate and crude death rates. By analyzing the relationships between these socio-economic and demographic factors against solid waste generation, the study seeks to identify the most effective method for forecasting future waste trends.

1.6 Significance of the study

This study holds significant relevance to several key stakeholders, amongst them are businesses, marketers and academic researchers. The study provides critical insights into waste generation trends to tackle one of Malaysia's most pressing environmental challenges.

Accurate forecasting of waste generation will help policymakers and waste management authorities such as SWCorp and municipal councils to develop more effective strategies to reduce waste. The predictive models developed in this research offer valuable tools to optimize waste collection schedules, enhance recycling efforts, and manage landfill capacities. Additionally, the findings support resource allocation decisions and enable these authorities to plan effectively for future waste management challenges.

The study supports Malaysia's efforts to achieve SDG 12: Responsible Consumption and Production by providing insights that reduce waste generation and promote efficient resource use. It supports efforts to minimize environmental impacts through improved recycling practices and sustainable consumption. It also contributes to SDG 13: Climate Action by addressing waste-related greenhouse gas emissions from landfills through better waste forecasting and management. These efforts collectively help Malaysia transition toward more sustainable and climate-resilient waste management systems.

In addition, this research benefits businesses, particularly those involved in food production and retail. Understanding the relationship between waste generation and population demographics can inform more sustainable practices and reduce economic losses to accommodate market changes. Marketers can use these insights to align campaigns with consumer behaviors and promote environmentally friendly products.

Additionally, the study further contributes to academia by offering a comprehensive review of previous literature regarding the relationship between socio-economic and demographic variables with waste generation. It demonstrates the application of machine learning models in environmental management. This provides a strong basis for further research into waste generation and management, particularly in using predictive analytics for sustainable environmental practices in Malaysia.

Lastly, the findings of the study offer valuable benefits to Malaysian society by addressing the pressing issue of solid waste. With landfills reaching capacity, he insights provided aim to raise community awareness about the environmental impacts of waste, encourage better household waste management practices, and emphasize the application of machine learning for predicting MSW generation. These efforts collectively support the goal of building a more sustainable and environmentally conscious society.

CHAPTER 2: LITERATURE REVIEW

2.1 Municipal Solid Waste

The Solid Waste and Public Cleansing Management Act 2007 defines MSW management as the regulation of waste generation, storage, collection, transfer, and disposal of unwanted materials or surplus substances no longer in use. However, it excludes scheduled wastes under the Environmental Quality Act 1974, sewage under the Water Services Industry Act 2006, and radioactive waste under the Atomic Energy Licensing Act 1984 (Ng et al., 2023; Syifaa et al., 2023). The Ministry of Local Government Development (KPKT) oversees solid waste management with operational support provided by the SWCorp and policy oversight by the National Solid Waste Management Department (NSWMD).

The composition of MSW in Malaysia is predominantly influenced by household waste, followed by commercial, institutional waste and industrial waste. Based on the news in Bernama entitled "Experts: Food Wastage Expected to Last till End-Syawal", it reported that Malaysia experiences a significant increase in food waste up to 44.5 per cent compared to other wastes during festive seasons. The data from the SWCorp supported the increasing pattern, as it indicated that the total solid waste disposed of during Ramadan rose by 21% from 2019 to 2022 (Mahasan, 2023).

The significant waste production in Malaysia results largely from population growth and urbanization (Ng et al., 2023). Factors such as rapid economic development, increased urban migration, and evolving lifestyle patterns have caused the waste generation rate to rise by 3% to 4% annually (Zulkipli et al., 2018). The COVID-19 pandemic significantly increased waste generation in Malaysia, particularly household and clinical waste (Cheng et al., 2022). Household waste, including plastic packaging from e-commerce and food deliveries, surged during lockdowns, while clinical waste, such as masks and PPE, rose by 27%, reaching 35.41 tons/day.

(Ministry of Housing and Local Government, 2023)

Figure 2.1 Solid Waste Management Facility in Malaysia

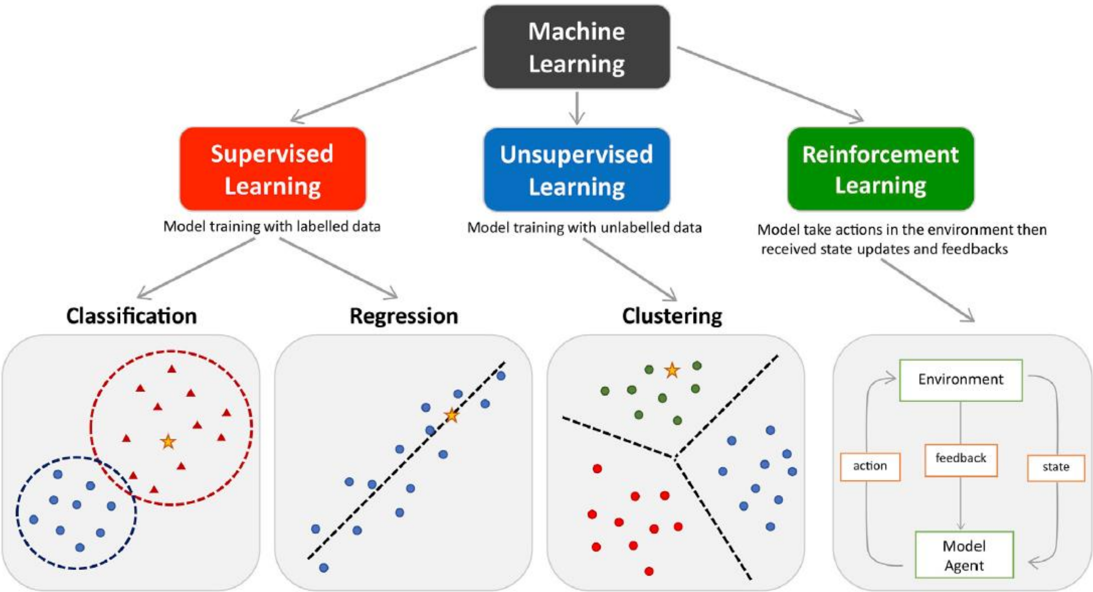
In Malaysia, waste disposal is managed through various types of facilities. Disposal sites (landfills), transfer stations, thermal treatment plants, and waste-to-energy (WTE) facilities as shown in Figure 2.1 serve as the primary options for states adopting the Solid Waste and Public Cleansing Management Act 2007 (Act 672). Landfills remain the primary method with 141 facilities nationwide of MSW disposal in Malaysia due to their low cost and simplicity compared to more advanced methods like incineration, which requires specialized technical expertise and higher operational expenses (Syifaa et al., 2023). This includes 22 sanitary landfills equipped with engineering solutions to minimize environmental risks, 114 non-sanitary landfills, and 5 inert landfills designed for non-reactive and non-decomposable waste. Malaysia also has five transfer stations that facilitate waste management by transferring waste from smaller collection vehicles to larger transport vehicles for more efficient long-distance transport. Additionally, the country operates four thermal treatment plants that use technologies to reduce the volume of waste sent to landfills and one WTE facility that converts waste into usable energy (Ministry of Housing and Local Government, 2023; Ng et al., 2023). Other facilities supporting waste management in Malaysia include Material Recovery Facilities for recycling, Communal/Commercial Composting Facilities for organic waste, Biogas Facilities for energy production and Refuse-Derived Fuel Facilities for creating fuel from waste (SWCorp Malaysia, n.d.)

Despite the infrastructure, challenges persist. Approximately 65% of MSW disposed of in landfills consists of recyclable materials (Ng et al., 2023). Open dumping is commonly used as it accommodates the high organic content of Malaysian waste (Syifaa et al., 2023). The high moisture content of Malaysian solid waste that influenced by the country’s climate and population lifestyle, reduces the efficiency of incineration as a disposal method (Ng et al., 2023; Zulkipli et al., 2018). Additionally, Malaysia faces growing challenges with landfill management as many sites are reaching or exceeding capacity limits. Building new landfills is becoming increasingly difficult due to land scarcity, rising land prices, and increased demand driven by population (Syifaa et al., 2023). These factors highlight the urgent need for more sustainable and efficient waste management solutions.

2.2 Machine Learning

Machine learning is a branch of artificial intelligence that uses data and algorithms to replicate human learning and improve accuracy over time. The foundational concept of machine learning from experience was first proposed by Alan Turing in the mid-20th century (Peng et al., 2021) It plays an important tool in data science to make predictions and uncover critical insights.

As illustrated in Figure 1, machine learning consists of three main types: supervised, unsupervised and reinforcement learning. These types differ in their outcomes, validation methods, and refinement processes. Supervised learning uses labeled datasets to predict future outcomes. It is divided into classification models that categorize input data into categories and regression models that predict continuous outcomes (MathWorks, n.d.; Sagi, 2024). In contrast, unsupervised learning uses unlabeled datasets to uncover hidden patterns or intrinsic structures input data. The most common unsupervised machine learning is the clustering model. Lastly, reinforcement learning uses trial and error to adjust actions, explore options, receive feedback, and improve performance without depending on labeled data. It requires significant computational resources (Peng et al., 2021; Sagi, 2024).



(Peng et al., 2021)

Figure 2.2 Main Types of Machine Learning

Selecting the appropriate machine learning algorithm depends on the problem, data type, and available resources. There is no single best method, as trial and error plays a key role. However, supervised learning is suitable for predictions and unsupervised learning is ideal to uncover data patterns. It's also common to combine different machine learning techniques to address complex problems.(MathWorks, n.d.; Sagi, 2024).This study employed supervised machine learning.

Forecasting uses economic concepts, mathematics, statistics, and econometric analysis. Accurate exchange rate predictions are crucial for investment and business, while forecasts of solid waste generation are vital for ensuring a healthier environment (Nasir et al., 2023). Individuals and authorities depend on forecasts to make decisions that shape the economy's future direction. Accurate MSW forecasts are crucial for planning waste collection, ensure sustainable solutions, and address potential challenges (Fokker et al., 2023; Nasir et al., 2023)

Models are evaluated to identify the most accurate forecasts. Error measures such as the coefficient of determination (R^2), RMSE and MAE are commonly used to differentiate between good and poor models (Fokker et al., 2023; Nasir et al., 2023). R^2 shows the proportion of variation in the dependent variable explained by the model, ranging from 0 to 1. For example, an R^2 of 0.90 means the model accounts for 90% of the variation, with 10% due to other factors. RMSE measures the average error magnitude in the same units as the target variable. Lower RMSE values indicate better performance, but an RMSE of zero may suggest overfitting. MAE reflects the average absolute error, also in the target's units. For instance, an MAE of 3 means predictions are on average off by 3 units. There is no single metric that can provide a complete evaluation. A combination of metrics will offer better insights. High R^2 and low RMSE values usually indicate a well-fitting model (Wohlwend, 2023). Table 2 compares the evaluation metrics commonly used in regression analysis. The equation is described below

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{1}$$

$$\text{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}} \tag{2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3}$$

Where $\hat{y}_i, \dots, \hat{y}_n$ represent observed values and $y_i, \dots, y_n$ are the predicted values and n is the number of solid waste sample sizes.

Table 2.1 Comparison of Evaluation Metrics

Metric	Range	Units	Benefits	Limitations
R^2	$0 > \pm 1$	Dimensionless	- Measures the proportion of variance explained by the model. - Easy to interpret.	- Only measures explained variance. - Sensitive to irrelevant features and non-linear data. - Unsuitable for cross-dataset comparisons.
RMSE	0 to ∞	Same as target variable	- Penalizes large errors, making it useful when large errors are critical.	- Sensitive to outliers. - Hard to interpret as an average error due to squaring of differences.
MAE	0 to ∞	Same as target variable	- Represents average error magnitude. - Less sensitive to outliers. - Simple to compute.	- Does not penalize large errors disproportionately. - Ignores the direction of errors. - Limited compatibility with gradient-based optimization algorithms.

Training, testing, and cross-validation are essential techniques for building reliable machine learning models. A training set trains the algorithm and a testing set evaluates its accuracy by comparing predicted values with true values. Weight adjustments refine the model during cross-validation to prevent overfitting or underfitting. Researchers commonly split datasets into 70–80% for training and 20–30% for testing to ensure unbiased evaluation. After training the dataset, hyperparameter tuning is performed to optimize model parameters. Cross-validation is then used to evaluate the model's performance and ensure its ability to generalize to unseen data. The most common cross-validation method is K-Fold Cross-Validation, where the dataset is divided into K equal parts (folds). The model is trained on K−1 folds and validated on the remaining fold, cycling through all folds to produce an average performance score. Commonly, K is set to 5 or 10 parts. In each iteration, 4 parts are used for training and 1 for validation. The best parameters are chosen based on validation results and used to train the full training dataset (Jainvidip, 2024).

2.2.1 Multiple Linear Regression

Multiple Linear Regression analysis is a fundamental statistical technique used to determine the relationship between a dependent variable and one or more independent variables (Dissanayaka & Vasanthapriyan, 2019). The MLR equation can be written as:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k \tag{4}$$

Where Y is the dependent variable, b_0 represents the intercepts when all predictors are zero, X is the independent variables and the coefficient of $b_0, b_1, +b_2 + \cdots + b_k$ measures the average change in the dependent variable for a one unit change in the independent variable.

MLR analysis provides reliable predictions for linear relationships. It assesses the statistical significance of variables and determines their impact. It handles multiple variables effectively and is suitable for complex datasets. Unlike advanced models such as ANN, MLR offers a clearer interpretation of the relationships between variables, making it useful for model comparison (Araiza-Aguilar et al., 2020). For example, Araiza-Aguilar et al. (2020) used MLR to forecast MSW generation based on social and demographic variables. Population emerged as the most influential predictor, with an adjusted R^2 of 0.975. Similarly, A. Kumar & Samadder (2017) developed MLR model to predict household solid waste generation rates in Dhanbad, India that focused on biodegradable and non-biodegradable waste. Using socioeconomic factors such as household size, income, education, occupation, and kitchen fuel type as predictors, the models achieved R^2 values of 0.782 for biodegradable waste and 0.676 for non-biodegradable waste.

Despite its strength, there is a certain limitation of MLR model. The non-linear relationships and sensitivity to outliers can significantly affect its accuracy (Swaak, 2021). For example, Ghinea et al. (2016) applied MLR to explore socio-economic factors such as number of residents, population aged 15–59, urban life expectancy that influenced waste generation. They demonstrated that the S-curve time series model is the most suitable for MSW prediction for total waste and individual waste fractions. Similarly, Azadi & Karimi-Jashni (2016) compared MLR with ANN to predict mean seasonal MSW generation rates in Fars Province, Iran. MLR effectively identified key predictors such as population and waste collection frequency but showed weaker predictive accuracy with a lower R^2 of 0.70 and higher RMSE of 95.13 due to multicollinearity and reliance on linear assumptions. Both studies showed that while MLR remains useful for its simplicity and interpretability, it often struggles to capture non-linear relationships inherent in waste generation dynamics.

2.2.2 Random Forest

The Random Forest algorithm (RF) is a machine learning method that uses an ensemble of decision trees. Each tree in the forest makes a class prediction, and the final prediction is determined by a majority vote (Dissanayaka & Vasanthapriyan, 2019). It is a popular choice due to its consistence prediction accuracy with minimal tuning and its effectiveness in handling non-linear parameters. While it has several tunable hyperparameters, the default settings often produce good results (Boehmke & Greenwell, 2020).

Grid search is a common method for hyperparameter tuning(Bhat, 2023). It specifies a range of values for each hyperparameter and trains the model on all possible combinations of these values. According to Boehmke & Greenwell (2020) and Bhat (2023), the main hyperparameters to consider include:

i) Number of trees (n_estimator)

The number of trees in the forest determines the stability of error rates and prediction performance. A good starting point is 10 times the number of features. More trees improve prediction stability and variable importance estimates but increase computation time linearly.

ii) Number of features at each split (max_features)

Controls split-variable randomization to balance low tree correlation with predictive strength. Higher values of number of features at each split work better with noisy data, as they are more likely to select strong predictors. It is recommended to test five evenly spaced values at first.

iii) Node size (min_samples_leaf and min_samples_split)

Node size is an important parameter that influences the performance of a RF model. It defines the minimum number of data points required in a tree node. Larger values of min_samples_split prevent overfitting but may reduce flexibility. Larger values of min_sammples_leaf simplify trees and improve generalization but may miss complex patterns. Choosing an optimal node size helps the model handle noise and outliers effectively while making accurate predictions

iv) Maximum tree depth (max_depth)

The maximum depth of individual trees controls how many levels each tree can grow. Deeper trees can capture more complex patterns but are more prone to overfitting.

RF's effectiveness has been demonstrated as another predictive model for MSW. Dissanayaka & Vasanthapriyan (2019) compared RF, ANN and MLR models to predict

MSW generation using socioeconomic factors in India. RF achieved (R^2 = 0.9608) outperformed MLR (R^2 = 0.6973) but performed slightly below ANN (R^2 = 0.9923). While RF demonstrated strong predictive performance due to its ability to capture complex interactions, it required proper parameter tuning for optimal results. In another study, A. Kumar et al. (2018) analyzed RF, ANN and Support Vector Machine (SVM) to predict waste rates based on factors of income, education, occupation, and type of house in Dhanbad, India. In this case, RF achieved the weaker accuracy of R^2 =0.66 compared to ANN (R^2 =0.75) and SVM (R^2 = 0.74).

2.2.3 Artificial Neural Network

An artificial neural network (ANN) is a computational model inspired by the structure of the human brain where it is made up of many network of interconnected neurons. Each neuron receives inputs as weighted sums of outputs from connected neurons (S. Kumar et al., 2020). ANN typically consists of three layers: an input layer, one or more hidden layers, and an output layer. The input layer processes single input values, which are transmitted through the hidden layers using synaptic weights. These connections link every neuron in one layer to all neurons in the subsequent layer. Finally, the output layer generates numerical results (Ali & Ahmad, 2019).

Training of ANNs requires adjustments to the weights of input, intermediate, and output connections. Hoy et al. (2022) emphasized that excessive neurons in a model could results in overfitting, while too few neurons result in underfitting. The ideal configuration depends on the size of input vectors and the classification of input-output relationships. Factors such as the size of training data, algorithm selection, transfer functions, network structure, and data representation significantly influence ANN performance. Studies have demonstrated that ANN models improve their ability to solve problems with sufficient training (Niu et al., 2021). Regarding the comparison activation function shown in Table

2.2, ReLU is widely used due to its simplicity and computational efficiency while Sigmoid and Tanh are less preferred due to gradient issues.

Table 2.2 Comparison of Common Activation Functions in Neural Networks

Activation Function	Advantages	Disadvantages	Applications
Sigmoid	- Non-linear - Differentiable - Outputs between 0 and 1, interpreted as probabilities	- Saturates easily - Suffers from the vanishing gradient problem - Computationally expensive	Binary classification, probabilistic outputs
Tanh (Hyperbolic)	- Non-linear - Differentiable - Zero-centered, aiding faster convergence compared to Sigmoid	- Saturates easily - Computationally heavier than ReLU	Suitable for hidden layers in deep networks
ReLU	- Non-linear - Computationally efficient - Faster convergence - Avoids saturation issues	- Not differentiable at zero - Dying ReLU problem - Outputs not zero-centered	General-purpose, standard for deep learning architectures
Leaky ReLU	- Solves Dying ReLU problem by allowing small gradients for negative inputs	Outputs not zero-centered	Used in deep networks to avoid dead neurons
SELU	- Self-normalizing - Maintains stability in deep networks - Eliminates batch normalization	- Limited to specific network architectures - Computationally expensive	Self-normalizing architectures requiring stability
Softmax	- Converts outputs to probabilities - Differentiable - Interpretable for classification tasks	- Sensitive to large input values, requiring numerical adjustments for stability	Output layer for multi-class classification

ANN models have been widely applied for MSW prediction due to their ability to model non-linear and dynamic relationships. Several studies demonstrated ANN's effectiveness in MSW prediction but also highlighted challenges related to overfitting, data quality, and computational demands.

Ali & Ahmad (2019) applied an ANN time series model structured to forecast monthly MSW generation in Kolkata, India. After testing various configurations, the study identified the 1-19-1 ANN structure as the most suitable model. It achieved R^2 of 0.9267 and predicted waste generation to increase from 4,500 tons per day in 2017 to 5,205 tons per day by 2030. On the other hand, Yusoff et al. (2018) experimented with different ANN architectures to predict MSW in Malaysia based on population growth. The optimal architecture, with two hidden layers (10 and 5 nodes) achieved 98.8% accuracy and forecasted a 29.03% increase in waste generation by 2031. Both studies emphasized that ANN performance relies heavily on high-quality input data, including seasonal and demographic factors, and recommended comparisons with alternative models for future research.

Despite its strength, ANN faces several challenges. Niu et al. (2021) evaluated traditional ANN model to forecast MSW in Suzhou, China. The ANN model achieved an R^2 of 0.94 on the training dataset but dropped to 0.74 on the testing dataset. This finding indication of overfitting and limited generalization. The study compared ANN with Long Short-Term Memory (LSTM) networks and ARIMA models. Although ARIMA struggled to capture non-linear trends, Long Short-Term Memory (LSTM) model performed better with R^2 of 0.92 on testing data. Similarly, S. Kumar et al. (2020) applied an ANN model with a time-series autoregressive technique to predict MSW generation in Greater Noida, India. The optimal 1-20-1 architecture achieved high accuracy (R^2 = 0.9411) but faced challenges due to seasonal variations, and data noise. In China, Abbasi & El Hanandeh (2016) compared ANN with SVM, Adaptive Neuro-Fuzzy Inference Systems (ANFIS), and k-Nearest Neighbors (kNN) to predict MSW generation in Logan

City, Australia. The ANN designed with a single hidden layer of eight neuron achieved the lowest accuracy (R^2 = 0.46). The study highlighted that ANN struggled with overfitting and sensitivity to irrelevant data, which reduced its accuracy. Sodanil & Chatthong (2014) developed time-series ANN model to forecast monthly MSW generation in Bangkok. The optimal ANN structure was identified as 3-35-1 achieved a prediction trend accuracy R^2 value of 0.629. These studies consistently highlight the need for refining ANN structures, incorporating external variables, and integrating optimization techniques to improve generalization.

Some studies explored optimization techniques to enhance ANN performance. Hoy et al. (2022) optimized a Bayesian-optimized ANN model to forecast MSW generation in Malaysia. The optimized ANN outperformed default ANN configurations. By focusing on eight waste types (e.g., food, garden, paper, and plastic), the study forecasted MSW generation to reach 42,873 tons per day by 2030, with food waste comprising 44%. Similarly, Elshaboury et al. (2021) integrated particle swarm optimization (PSO) with a feedforward ANN, creating an ANN–PSO model to predict MSW in Polish cities. The model used economic, demographic, and social factors such as population, revenue per capita, and employment-to-population ratio to forecast waste generation. This hybrid ANN–PSO model outperformed traditional ANN models (R^2 = 0.96 vs 0.68). Despite its improved accuracy, the ANN–PSO model depended heavily on city-specific data and required adjustments to adapt to different regional conditions. Both studies suggested further research into optimization algorithms and dimensionality-reduction techniques to enhance performance.

2.2.4 Summary

In summary, ANN models are frequently used in MSW prediction due to its ability to handle non-linear relationships and complex datasets. The model delivers higher accuracy compared to MLR and RF. While ANN achieves strong performance in MSW prediction, it relies heavily on computational resources and high-quality data, requiring careful configuration to address these limitations.

RF often outperforms MLR and handles noisy data, high-dimensional datasets, and non-linear interactions effectively. However, its accuracy depends on proper hyperparameter optimization, and it may perform poorly with smaller datasets. Despite these challenges, RF remains a valuable tool for MSW prediction when accuracy and interpretability are priorities. MLR is less accurate than RF and ANN in most cases due to its inability to model non-linear relationships. However, its simplicity and interpretability make it suitable in certain contexts, and its relevance in waste generation prediction should not be dismissed entirely.

To address the lack of comprehensive comparisons between these models, this research will evaluate ANN, RF, and MLR using the same dataset to identify the most effective model for predicting MSW generation in Malaysia. This study will evaluate the performance metrics of these models in a structured manner and provide valuable insights into their application in MSW management.

2.3 Previous Research on Influential Variables of MSW Generation

The prediction of MSW generation has seen the application of various statistical and machine learning techniques. Socioeconomic and demographic indicators, such as population size, employment ratio, urbanization, GDP, fertility rate, and education levels, are widely recognized as critical predictors (Dissanayaka & Vasanthapriyan, 2019; Elshaboury et al., 2021; Ghinea et al., 2016; Intharathirat et al., 2015). According to Zulkipli et al. (2018), the solid waste generation in Malaysia is influenced by several external factors, with population being a primary driver. The growing population, rapid urbanization, and shifting lifestyles in Malaysia have resulted in a projected solid waste generation of over 15 million tonnes by 2025. While population growth is a key factor driving this increase, reducing the population is not a practical solution. Instead, implementing systematic and innovative waste management strategies is essential to address this challenge effectively. Additionally, Yusoff et al. (2018) supports the finding that waste generation increases proportionally with population growth. It predicts a 29.03% increase in solid waste generation by 2031 compared to 2012. It highlights the need for advanced prediction tools like ANNs to address the growing waste challenges effectively.

While traditional indicators like population and household size are proven to be significant predictors, other variables such as crude death rates and fertility rates have shown varying results across studies. To address these gaps, this research aims to identify external indicators that influence waste generation by focusing on a range of socioeconomic and demographic variables. Specifically, the variables considered in this research include GDP, population estimates, urban indicators, population aged 60+, fertility rate, number of households, labor force rate, employment-population ratio, crude birth rate, and crude death rate. By analyzing these external factors, the study seeks to determine their impact on waste generation and contribute to a deeper understanding of the relationship between these indicators and solid waste management. Table 2.3 summarizes machine learning techniques, findings and respective independent variables based on similar studies addressing solid waste generation.

Table 2.3 Summary of Influential Variables in MSW Generation

Reference	ML Techniques	Independent Variables	Findings
Fokker et al. (2023)	Seasonal Naïve Benchmark, ETSX ensemble models, and Quantile Regression with external variables.	Hourly fill rates, weather variables, and event data	The ETSX model accurately predicted MSW generation in 74% of cases. Poor weather conditions, such as precipitation, wind gusts, and thunderstorms, lead to less waste disposal.
Abdella Ahmed et al. (2022)	LSTM, DLTSF	Socio-economic zones (poor, social, privileged) and waste types (plastic, glass, paper, carton, organic)	Solid waste generation per person was highest in the privileged zone (0.86 kg/day) compared to social (0.65 kg) and poor (0.42 kg) zones. The DLTSF model forecasted MSW types with a mean RMSE of 0.03371.
Hoy et al. (2022)	Bayesian-Optimised ANN, Ensemble Learning	Eight waste composition variables. Socioeconomic indicators (population, fertility rate, life expectancy, working hours, GDP, human capital index, CO2 emissions, energy, and electricity consumption).	Bayesian-optimized neural networks reduced overfitting and forecast uncertainty (3.64–27.7%) compared to default models (11.1–44,400%). Malaysia's MSW in 2030 is projected at 42,873 t/d, with 44% as food waste. Each waste type showed correlations with socioeconomic indicators.
Elshaboury et al. (2021)	ANN coupled with PSO algorithm	Population, employment ratio, revenue per capita, business types, and REGON entities per 10,000 people.	The ANN–PSO model achieved higher accuracy (R = 0.96) compared to traditional ANN models.

Table 2.3 Summary of Influential Variables in MSW Generation (continued)

Reference	ML Techniques	Independent Variables	Findings
Araiza-Aguilar et al. (2020)	MLR	Population density, migration rate, socioeconomic factors	MLR achieved an adjusted R^2 of 0.975 with 7.7% mean absolute percentage error for predicting waste generation.
Dissanayaka & Vasanthapriyan (2019)	MLR, RF, ANN	Socio-economic indicators (total population, GDP growth rate, crude birth rate)	ANN achieved the highest accuracy ($R^2 = 0.9923$), followed by Random Forest ($R^2 = 0.9608$) and MLR ($R^2 = 0.6973$). Crude birth rate and GDP growth rate showed strong positive correlations with MSW generation.
Yusoff et al. (2018)	ANN using two hidden layers (10 and 5 nodes, respectively)	Population data	The ANN model with two hidden layers achieved $R^2 = 0.988$. Solid waste generation is predicted to increase by 29.03% from 2012 to 2031.
Zulkipli et al. (2018)	An integrated dynamical solid waste management model	Population, waste generation rate, and socio-economic factors (income per service center, household size, and income per household).	Malaysia's solid waste generation is expected to exceed 15 million tonnes by 2025, driven mainly by population growth. Reducing the population alone is not enough to reduce total solid waste.
A. Kumar et al. (2018)	ANN, MLR, SVM	Socioeconomic parameters (e.g., income, education, occupation, housing type)	Higher socioeconomic groups had the highest plastic waste generation (51 g/c/d), while lower groups generated the least (8 g/c/d). Informal recyclers played a major role in recycling and revenue generation.

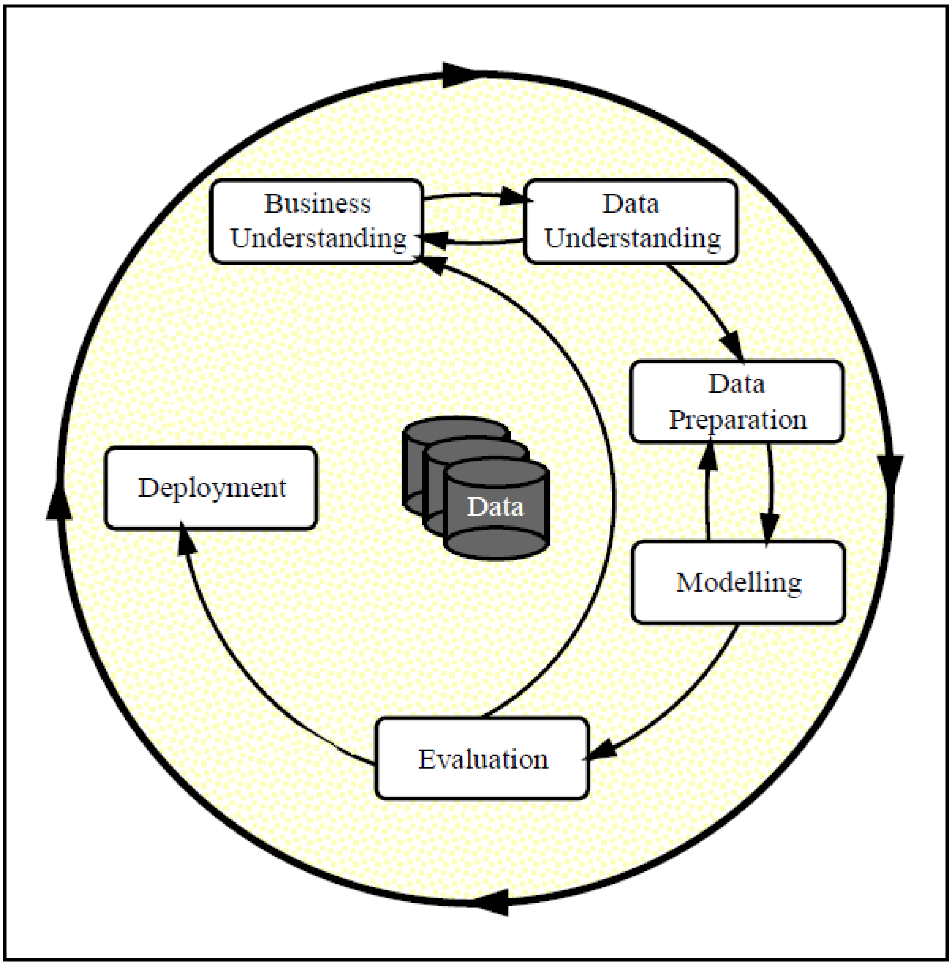
Table 2.3 Summary of Influential Variables in MSW Generation (continued)

Reference	ML Techniques	Independent Variables	Findings
A. Kumar & Samadder (2017)	MLR	Household size, income, education, fuel usage	Biodegradable waste and non-biodegradable waste prediction showed R^2 values of 0.782 and 0.676, respectively.
Johnson et al. (2017)	Gradient Boosted Regression Tree (GBRT)	Demographic and socioeconomic data, weather variables	The gradient boosting regression model achieved $R^2 > 0.88$ and accurately captured waste generation trends affected by holidays, weather, and seasons.
Azadi & Karimi-Jashni (2016)	ANN, MLR	Population, collection frequency, maximum seasonal temperature, altitude	ANN outperformed MLR in predicting seasonal MSW generation rates.
Ghinea et al. (2016)	MLR, Waste Prognostic Tool, Time Series Analysis	Socio-economic indicators (number of residents, population aged 15–59, urban life expectancy)	Regression analysis identified the population aged 15–59 and total MSW as key factors influencing waste generation. The S-Curve model best predicts total MSW.
Intharathirat et al. (2015)	Grey Models: GM(1,1), GM(1,n), GMC(1,n)	Consumption expenditure, household size, employment proportion, population density, and urbanization.	The GMC(1,5) model achieved top accuracy (MAPE: 1.16%) and predicts a 1.40% annual increase in MSW, from 43,435 tonnes/day in 2013 to 55,177 tonnes/day in 2030. Population density, urbanization, and employment drive the increase. Demographic factors have a higher impact than socio-economic factors.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This chapter describes various stages used to conduct this study. This study is divided into six phases based on the Cross Industry Standard Process for Data Mining (CRISP-DM) data life cycle as shown Figure 3.1. Each stage is important to turn raw data into meaningful insights.



(Wirth & Hipp, 2000)

Figure 3.1 CRISP-DM Data Lifecycle Process

The study begins with the business understanding phase to identify project objectives and interpret them into specific data mining goals. Next, the data understanding phase involves data collection and initial exploration to ensure data relevance to uncover patterns and relationships. This stage is an important phase as it determines the feasibility and reliability of the final output from the expectation that can be achieved using the data.

The third stage is data preparation. In this stage, raw data is transformed into a suitable modeling format. Next, the modeling phase will apply various algorithms to generate predictive or descriptive insights. A model assessment might need to be done in this stage to evaluate the existing data for result improvement. In the evaluation phase, it ensures that the developed models align with business objectives and that their performance is satisfactory. Testing and training data percentage will be used in this stage to ensure that the model from the modelling phase is reliable and accurate. Finally, the deployment phase implements the results and transforms the findings into actionable knowledge for stakeholders. In the deployment phase, the process can range from straightforward tasks like report writing to more complex tasks such as establishing a repeatable data mining workflow (Wirth & Hipp, 2000).

3.2 Business Understanding

There are two primary objectives in the research. Generally, this research is conducted to recommend the most accurate and suitable predictive model based on the demographics and socioeconomic factors that influence solid waste generation. Figure 3.2 illustrates the structured approach used in this project that outlines the steps required to achieve the research objective. The process begins with understanding the requirements of the research, followed by a review of existing research and methodologies to gather insights into best practices, previous findings, and relevant data sources. The literature review summary identifies the factors that might influence solid waste generation. This step help to identify the factors that influence solid waste generation which contributes to the first research objective. Next, relevant data is collected from three main categories and prepared for the modelling phase. Since the data comes from different sources, it is essential to standardize and merge it into a uniform format. During the modelling phase, the cleaned data is divided into training and testing sets, and parameters are defined for each model. The models are evaluated based on their performance. If the models fail to meet the research objectives, the process returns to the data preparation stage to refine the approach and improve the outcomes. Once the models achieve the required performance, the results are visualized to communicate findings effectively.

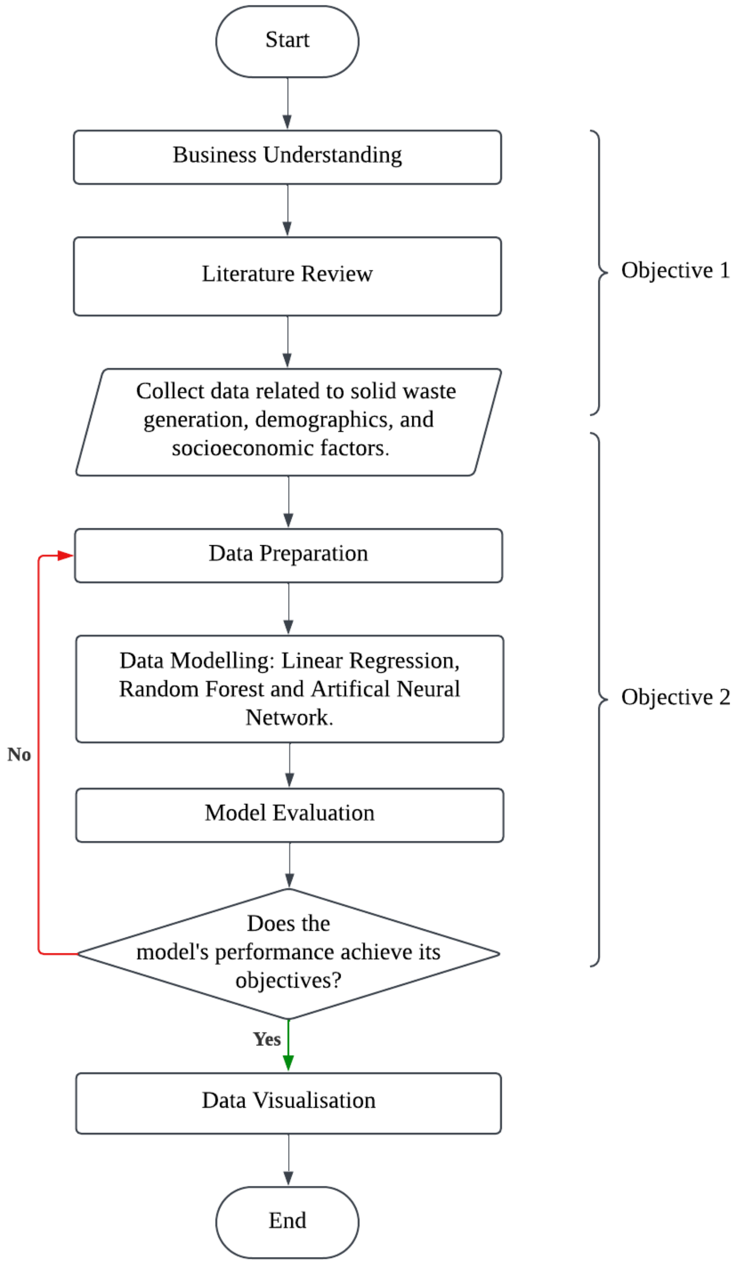


Figure 3.2 Research Workflow

3.3 Data Understanding

In the data understanding phase of the CRISP-DM, the data collection process plays an important role. Data collection for this study was primarily conducted through online sources. The challenge in collecting waste management data in Malaysia is its limited public availability. In addition, there are no centralized platforms where such data can be purchased. Typically, requesting data via email to the relevant department resulted only in annual summaries rather than detailed monthly or daily figures.

Despite these challenges, we successfully acquired data on solid waste and recyclable waste generation from 2017 to 2021. These datasets were sourced from the Solid Waste Corporation’s archive website and the yearly statistics published by the Ministry of Housing and Local Government (KPKT). The dataset included six states and a federal territory that operate under the Solid Waste and Public Cleansing Management Act 2007 (Act 672): Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perlis and Wilayah Persekutuan Kuala Lumpur.

Table 3.1 showed the definition of the parameters in the dataset. The socioeconomic factors (GDP, Number of Households, Labour Force Rate, Employment-Population Ratio) and demographic factors (Population Estimate, Urban Indicator, Population Aged 60+, Fertility Rate, Crude Birth Rate, Crude Death Rate) were obtained from Department of Statistics Malaysia (DOSM) webpage. Overall, the dataset consists of 504 rows and 16 columns.

Table 3.1 Data Description

Parameter	Description
Year	The year for which the data is collected.
Month	The month for which the data is collected.
State	Name of the state in Malaysia.
Facility	Type of solid waste management facility.
Solid Waste Entered to Disposal Site	The collection of solid waste deposited at a designated disposal facility in states that adopted the Solid Waste and Public Cleansing Management Act 2007 (Act 672) measured in tonne.
Recyclable Household Waste Collection	The collection of recyclable waste collected from households in states that adopted the Solid Waste and Public Cleansing Management Act 2007 (Act 672) by concession companies once a week measured in tonne.
GDP	Total value of goods and services produced after deducting the cost of goods and services used in production, but before deducting the consumption of fixed capital.
Population Estimate	The estimated total number of people living measured in thousands.
Urban Indicators	The estimated total number of people living aged 15 to 59 years old measured in thousands.
Population Aged 60+	The estimated total number of people aged more that 60 measured in thousands.
Fertility Rate	The average number of children that are born to a woman over her lifetime based on current age-specific fertility rates.
Number of Household	The total number of households.
Labour force rate	Rate of the working-age population that is either employed or actively looking for work.
Employment-Population Ratio	Rate of the number of employed people to the working-age population.
Crude Birth Rate	Number of live births per 1000 population
Crude Death Rate	Number of deaths per 1000 population.

3.4 Data Preparation

After collecting the data, we examined the data type, missing values and unreliable data using the sanity check. The initial overview revealed that the columns ‘Month’, ‘State’, ‘Facility’, and ‘Solid Waste Entered to Disposal Site (Tonne)’ were categorized as object data types. We identified the unique values within these columns to ensure appropriate handling. Upon checking, the dataset contained no duplicate entries. However, the missing data summary indicated that 5.75% and 4.76% of values were missing in the 'Solid Waste Entered to Disposal Site (Tonne)' and 'Recyclable Waste Collection (Tonne)' columns, respectively.

3.4.1 Data Reformatting

From the sanity check output, several data preprocessing steps were taken to ensure consistency and prepare the data for analysis. The 'Solid Waste Entered to Disposal Site (Tonne)' column was converted to a float data type for numerical analysis. The 'Month' column was mapped from Malay to English month names. The 'Date' column was converted to DateTime format for time-based analysis. Lastly, the 'State' column was standardized by capitalizing only the first letter of each word for consistency and converted to categorical type.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 504 entries, 0 to 503
Data columns (total 18 columns):
 #   Column                                        Non-Null Count  Dtype
---  ------                                        --------------  -----
 0   Year                                          504 non-null    category
 1   Month                                         504 non-null    category
 2   State                                         504 non-null    category
 3   Facility                                      504 non-null    category
 4   Solid Waste Entered to Disposal Site (Tonne)  474 non-null    float64
 5   Recycleable Waste Collection (Tonne)          480 non-null    float64
 6   GDP                                           504 non-null    float64
 7   Population                                    504 non-null    float64
 8   Urban Population                              504 non-null    float64
 9   Elderly Population                            504 non-null    float64
 10  Fertility Rate                                504 non-null    float64
 11  Number of Household                           504 non-null    int64
 12  Labour Force Rate                             504 non-null    float64
 13  Employment Ratio                              504 non-null    float64
 14  Crude Birth Rate                              504 non-null    float64
 15  Crude Death Rate                              504 non-null    float64
 16  Date                                          504 non-null    datetime64[ns]
 17  GDP per capita                                504 non-null    float64
dtypes: category(4), datetime64[ns](1), float64(12), int64(1)
memory usage: 58.3 KB
```

Figure 3.3 Dataset Overview After Reformatting

Figure 3.3 illustrated the data overview after the data reformatting. The 'Gross Domestic Product (GDP)','Population Estimate ('000)', 'Urban Indicators ('000) (population aged 15 - 59 years)', and 'Population Aged 60+ ('000)' were converted from thousands to their actual unit values for accuracy analysis. Furthermore, we created new features named GDP per capita for more granular economic insights. Lastly, several columns were renamed for clarity. For example, ‘Population Estimate ('000)’ column was renamed to ‘Population’ column.

3.4.2 Outlier and Missing Values Treatment

In this section, boxplots were used to detect skewness and outliers in the dataset. Boxplots are an effective tool for visualization of the distribution of numerical data and identification of potential outlier (Agarwal, 2019).

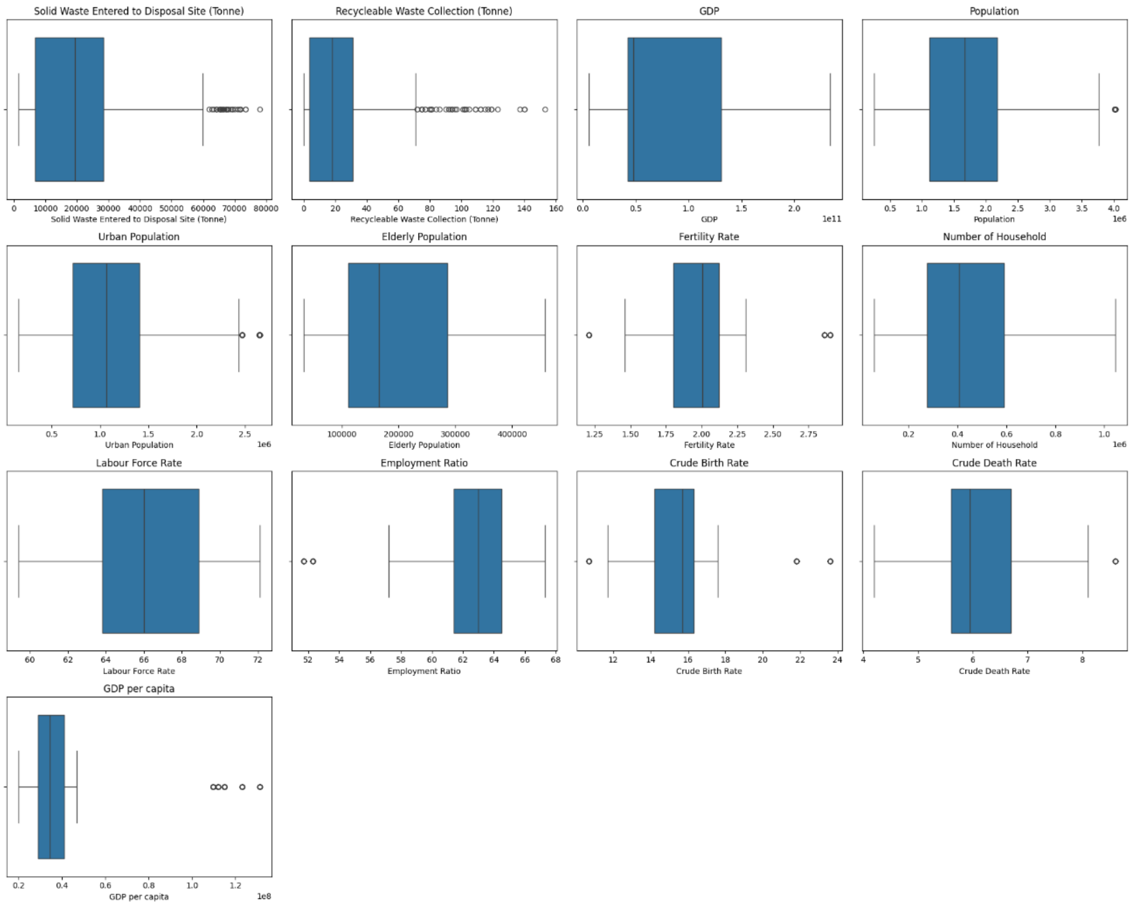


Figure 3.4 Boxplot Visualization of Skewness and Outliers

Figure 3.4 showed that the outlier values in the boxplots are not excessively extreme. Since the data originated from a reliable source, these outliers may represent exceptional cases rather than errors. For example, high GDP outliers could indicate states with large industries. Also, high fertility rates or crude birth rates might reflect cultural or policy-driven trends. Hence, we opted to retain the data to ensure integrity and transparency.

The boxplot of the Solid Waste Entered to Disposal Site (Tonne) and Recycleable Waste Collection (Tonne) revealed right-skewed distribution. In such cases, median imputation is suitable for numerical data with a heavily skewed distribution and outliers (Firdose, 2023). Moreover, missing values imputation with mean or median by relevant grouping will show more refined result (Huey, 2021). For example, missing values imputation within each state will ensure the overall trends remain consistent as each state has its own unique characteristics.

To address missing values in the 'Solid Waste Entered to Disposal Site (Tonne)' and 'Recyclable Waste Collection (Tonne)' columns, we will apply median imputation by state and month grouping. This approach ensures consistency with the unique temporal and regional patterns in the dataset. If missing values persist after group-based imputation, using the overall median as a fallback is both appropriate and practical.

3.4.3 Irrelevant Column

As mentioned in Chapter 1, any state or territory other than Johor, Kedah, Melaka, Negeri Sembilan, Pahang, Perlis, and Wilayah Persekutuan Kuala Lumpur must be excluded from the dataset. Furthermore, Recyclable Waste Collection' represented the same source as general waste generation. Hence, it was excluded from the heatmap correlation and data modeling phases to avoid redundancy.

If the explanatory variables do not meet the independence criteria, multicollinearity may occur due to high correlations between two or more variables. This issue increases the model's standard error and adds uncertainty to the estimated coefficients (A. Kumar & Samadder, 2017). In the study, we used correlation heatmap and variation inflation factor (VIF) to identify the multicollinearity issues. Finally, we trained models to predict 'Solid Waste Entered to Disposal Site' based on other features to avoid data leakage. When the most suitable method is identified, its predictions will be used as an input to model 'Recyclable Waste Collection'.

Table 3.2 Statistical Parameters for MSW Prediction

Statistical Parameters	Solid Waste Entered to Disposal Site (Tonne)	Fertility Rate	Employment Ratio	Crude Death Rate	GDP per capita
count	480	480	480	480	480
mean	24648.770	1.981	62.426	6.202	4.34E+07
min	1925.280	1.210	59.900	4.200	2.01E+07
25%	9661.395	1.795	64.325	5.600	3.0E+07
50%	20278.455	1.990	66.300	6.000	3.47E+07
75%	30398.424	2.120	68.900	6.700	4.12E+07
max	77915.860	2.310	72.100	8.600	1.31E+08
std	19491.200	0.245	2.897	0.994	2.94E+07

After removing the irrelevant column, the cleaned dataset will be saved for future analysis. Table 3.2 Statistical Parameters for MSW Prediction. Table 3.2 depicts the waste generation does not follow a normal distribution. The large difference between the mean and median suggests a right-skewed distribution. It suggests that certain states or urbanized regions disproportionately contribute to waste generation. Further analysis could focus to examine unique characteristics of the states to better understand the drivers of waste production.

The next step is to perform exploratory data analysis (EDA) to uncover patterns, trends, and relationships within the data. As part of the analysis, a correlation heatmap will be used to evaluate the relationship between the variables. This helps identify strongly correlated variables, which can guide feature selection and highlight redundant or irrelevant variables. Based on this analysis, some variables will be dropped to improve model performance and reduce complexity. The findings and insights from the EDA will be presented in detail in the next chapter.

3.5 Data Modelling

In the modelling phase, supervised predictive models were designed based on insights from the literature review. The data was shuffled prior to splitting to avoid bias. If the data is ordered, models trained on unshuffled data may learn patterns that don't generalize well to unseen data. Shuffled data also ensures that each fold in the dataset represents the entire data distribution which will result in more reliable evaluation metrics (Dutta, 2024).

The dataset was divided into training and testing sets, with 80% (384 samples) used for training the model and 20% (96 samples) reserved for testing and validation. This percentage split was chosen based on findings from previous research, which demonstrated that an 80/20 division provides an effective balance between model training and evaluation. A larger training set was chosen to enable the models to effectively learn patterns from the historical fill-rate percentage (Fokker et al., 2023) and ensure good training performance (Hoy et al., 2022). The validation set was used to select optimal model hyperparameters, while the test set was used to evaluate the final model's performance by comparing its predictions with unseen fill-rate percentages.

Three predictive models were developed: MLR, RF and ANNs. These models were trained on the training dataset to predict the target variable. For the ANN model, the Sequential API from Keras was utilized, where layers are added in a linear sequence. The network architecture consists of two hidden layers and an output layer. The first hidden layer has 64 neurons with a ReLU activation function and weights initialized using a normal distribution. The second hidden layer has 32 neurons, also using the ReLU activation function and normal weight initialization. The output layer consists of a single neuron with no activation function.

3.6 Data Evaluation

In order to assess the performance of the developed models, three popular evaluation metrics were used to ensure the reliability and accuracy of performance. The metrics are MAE, RMSE and R^2. For better performance, a predictive model should have high R^2 and low RMSE (A. Kumar & Samadder, 2017; Wohlwend, 2023).

Once the models were evaluated, cross-validation of the R^2 score was performed to assess the model's performance. By splitting the training data into multiple folds, cross-validation reduces the risk of overfitting and ensures that the evaluation metrics are reliable and not overly influenced by a single training-test split.

Then, hyperparameter tuning was performed for the selected model to enhance its performance. The feature importance for tree-based models and the residual distribution analysis will also be conducted to better understand the model's behaviours. The final model will be used for predictions and further analysis to enable future forecasting.

3.7 Deployment

In the deployment phase, the trained model was saved using the Joblib library to avoid the need for retraining. This allows the pre-trained model to be easily loaded for predictions or integration into a production environment. Joblib is ideal for machine learning models as it efficiently stores data as byte strings (Sharma, 2024). It ensures reliability during file operations and supports saving multiple model versions for easy comparison.

3.8 Software/Tools Used

In this study, Python 3 was selected as the programming language for its extensive library support and ease of error debugging. Libraries such as Pandas, NumPy, and Matplotlib were used for data analysis, mathematical operations, and visualization, respectively.

Google Colab was chosen as the primary platform to run Python code efficiently. Various machine learning algorithms were applied to develop the prediction model. Lastly, Microsoft Office tools which include Word, Power BI, and PowerPoint, were used for documentation, analytics, and creating presentation slides.

CHAPTER 4: RESULT AND DISCUSSION

4.1 Introduction

This chapter presents the findings of the study through detailed exploration of the dataset and the results of the predictive models. The chapter begins with an analysis of the dataset using exploratory data analysis (EDA) to uncover patterns, relationships, and trends. Following this, the results from the data modeling phase and the evaluation of model performance are discussed. These insights provide a comprehensive understanding of the data and the effectiveness of the predictive models to achieve the research objectives.

4.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) provides foundational insights that guide the entire machine learning process. Firstly, we evaluate the correlation heatmap to understand the relationship between the target variable which is solid generation and other variables in the dataset. Table 4.1 presents the correlation coefficient scale used in this project. Variables with coefficients between 0.0 and ±0.2 are considered weakly related to the target variable and may be dropped to reduce noise. Variables between ±0.8 and ±1.0 show strong relationships and are further examined to address potential redundancy and avoid multicollinearity.

Table 4.1 Correlation Coefficient Scale

Range of Correlation Coefficient	Strength of the Relationship
0.0 to ± 0.2	Little
± 0.2 to ± 0.4	Weak
± 0.4 to ± 0.6	Moderate
± 0.6 to ± 0.8	Strong
± 0.8 to ± 1.0	Very Strong

(Pennsylvania State University, 2024)

Based on the heatmap in Figure 4.1, several key relationships between the features in the dataset were identified. Variables such as Population, Urban Population, Elderly Population, Number of Households, and Employment Ratio dropped due to their little correlation with the target variable.

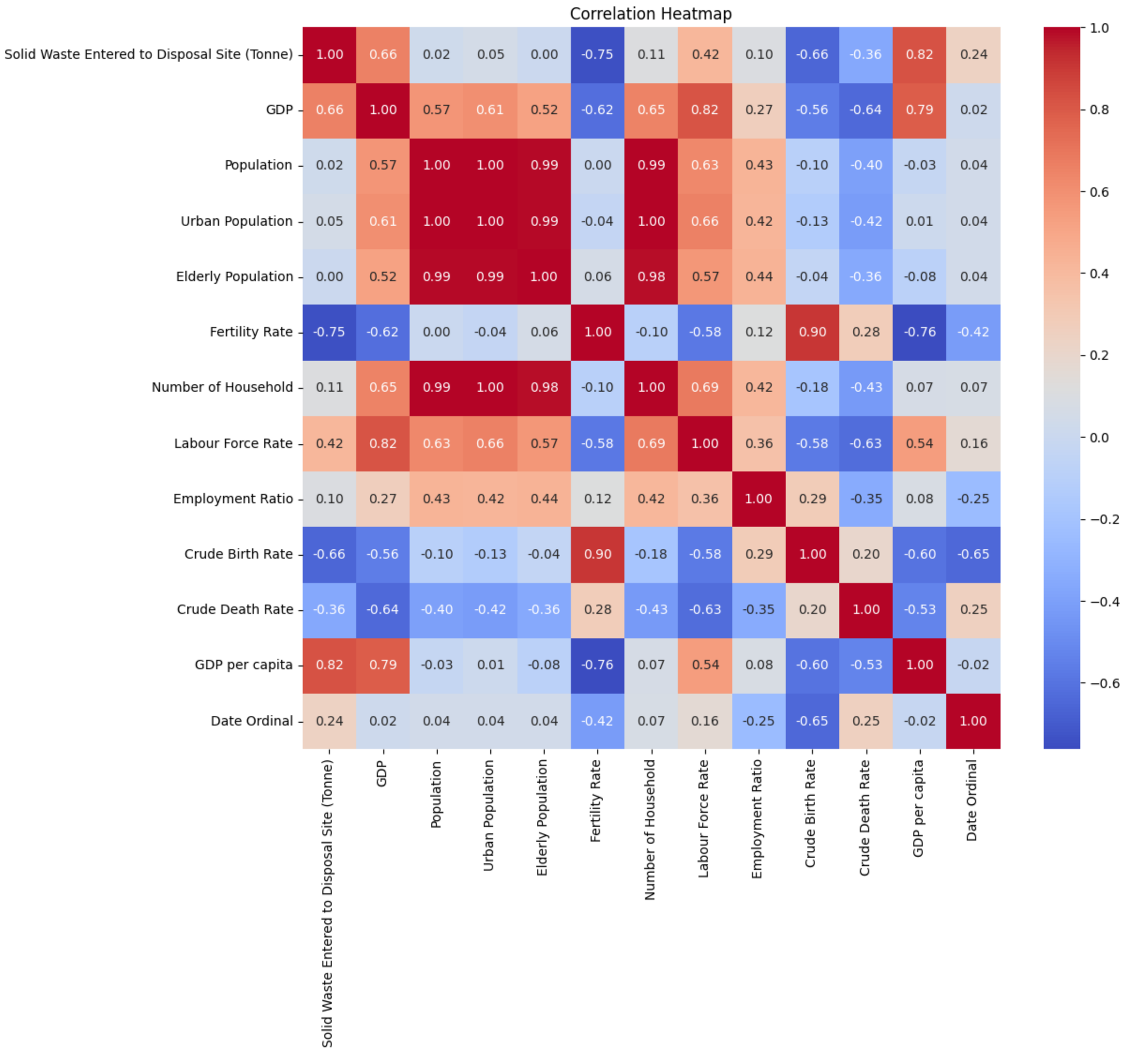


Figure 4.1 Correlation Heatmap

There were strong correlations observed between certain variables. For example, Population was highly correlated with Urban Population (1.0), Elderly Population (0.99), and Number of Households (0.99). Next, Fertility Rate showed a strong correlation with Crude Birth Rate (0.90). Since Fertility Rate (-0.75) has a stronger correlation with the target variable than Crude Birth Rate (-0.66), Fertility Rate was retained. Lastly, GDP was highly correlated with GDP per Capita (0.79) and Labour Force Rate (0.82). Given that GDP per Capita (0.82) has a stronger correlation with the target variable than GDP (0.66), GDP per Capita was retained.

In contrast to studies by Elshaboury et al. (2021), Araiza-Aguilar et al. (2020) andYusoff et al. (2018) which highlight a strong correlation between Population and waste generation, these studies do not consider GDP as an influential factor. Similarly, Hoy et al. (2022), found that different types of MSW compositions were correlated with various socioeconomic indicators. In this study, Population and Employment Ratio were strongly correlated with waste generation. Unlike other studies, all variables in the research were positively correlated with waste generation. This may be due to the inclusion of specific MSW types, which can influence the socioeconomic indicators relate to waste generation.

Among other variables, Dissanayaka & Vasanthapriyan (2019) found that Crude Birth Rate, followed by GDP growth rate and Total Population, significantly influenced waste sgeneration in Sri Lanka. Interestingly, their study reported that Total Population was negatively correlated with waste generation.

In this study, several variables were dropped due to high correlations or lower relevance to the target variable. Urban Population, Population, Number of Households, Elderly Population, Employment Ratio, GDP, and Crude Birth Rate were removed. These retained variables provide better predictive power and align with findings from other studies.

Next, we assessed the degree of multicollinearity using the Variance Inflation Factor (VIF). According to Ghinea et al. (2016; Hoy et al., 2022) and Hoy et al. (2022), a VIF of 1 indicates no correlation among variables, a VIF between 1 and 5 suggests moderate correlation, and a VIF greater than 5 indicates high correlation. As shown in Table 1.2, all features demonstrated moderate correlation with waste generation. Therefore, we can conclude that the features are acceptable for inclusion in the predictive model.

Table 4.2 Variance Inflation Factor Matrix

Feature	VIF
const	2791.09
Fertility Rate	4.99
GDP per capita	4.23
Labour Force Rate	2.47
Crude Death Rate	2.41
Date Ordinal	1.97

To further analyze these relationships, this section is divided into three parts:

4.2.1 Univariate Analysis

(i) Total Solid Waste Generation by State Analysis

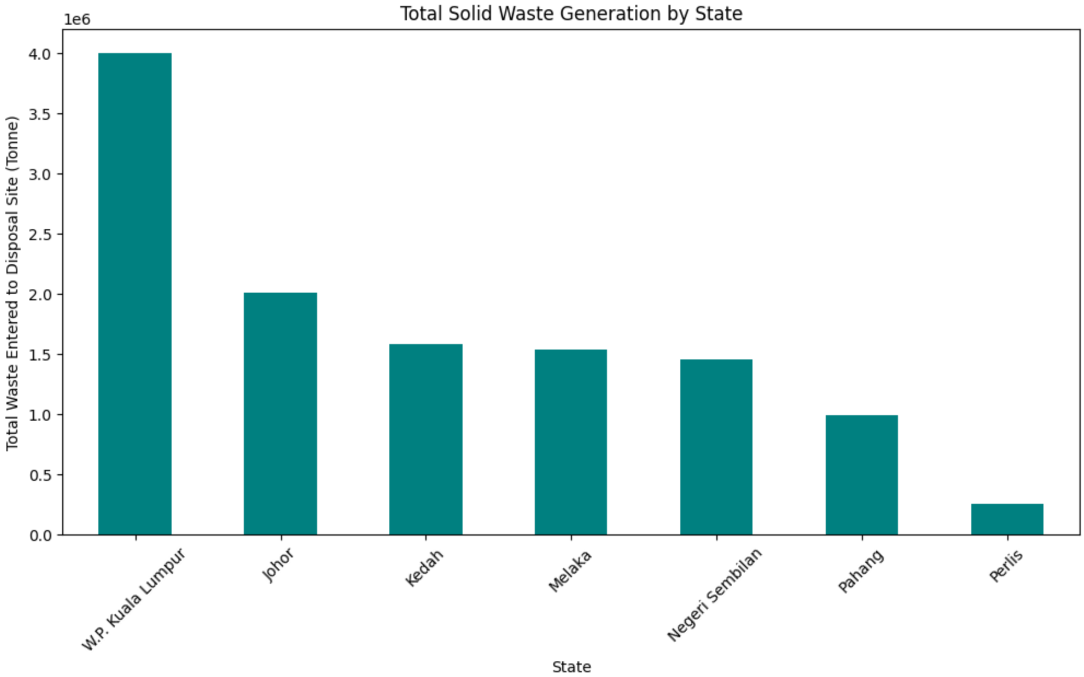


Figure 4.2 Total Solid Waste Generation by State

The bar chart in Figure 4.2 illustrates the total solid waste generation by state in Malaysia. W.P. Kuala Lumpur produces the highest amount of waste, exceeding 4 million tonnes, followed by Johor at slightly over 2 million tonnes. On the other hand, Pahang produces approximately half of Johor's waste, while Perlis contributes significantly less waste than all other states. The chart highlights a clear disparity in waste generation, with W.P. Kuala Lumpur leading by a wide margin.

(ii) State-wise Analysis of Key Socioeconomic Indicators

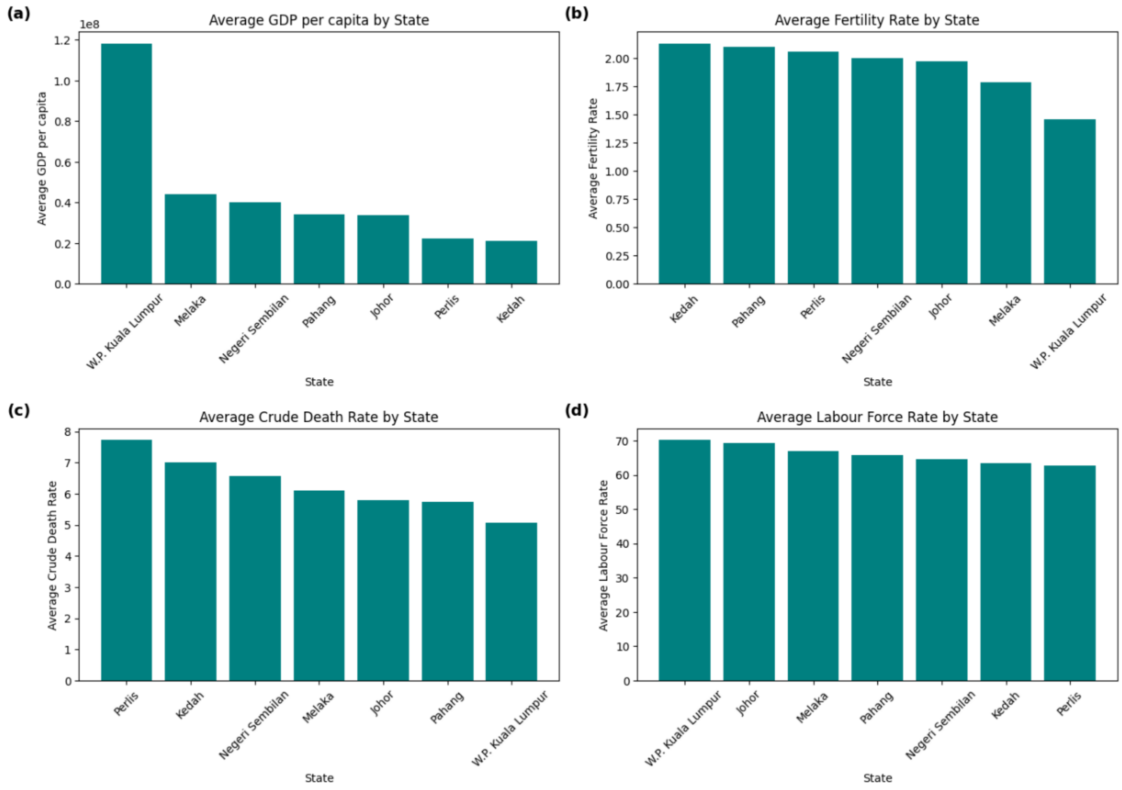


Figure 4.3 State-wise Analysis of Key Socioeconomic Indicators

Figure 4.3 highlights clear disparities in economic performance, fertility rates, mortality rates, and labour force participation across states. Figure 4.3 (a) shows that W.P. Kuala Lumpur stands out with a significantly higher average GDP per capita than all other states with values of approximately 120 million. It reflects higher economic activity in W.P Kuala Lumpur compared to other states. In contrast, Perlis and Kedah recorded the lowest GDP per capita among the states, with values of approximately 20 million each. Phooi et al. (2022) emphasized that rural households donated less food waste than urban areas and that high living standards lead to low environmental consciousness, even when people understand the financial and environmental costs of wasting food.

Based on Figure 4.3 (b), the fertility rate is highest in Kedah, Pahang, and Perlis, all showing an average close to 2.0. In contrast, W.P Kuala Lumpur recorded an average fertility rate at 1.5 which is the lowest compared to other states. Next, Figure 4.3 (c) illustrated that Perlis and Kedah have the highest crude death rates among the states, at approximately 7.8 and 7.0 respectively. It indicates relatively higher mortality levels in these regions. Meanwhile, W.P. Kuala Lumpur records the lowest crude death rate at around 5.0. These figures highlight significant variations in mortality trends across the states. According to RECYCLING Magazine (2019), states with higher crude death rates signals an increased strain on healthcare systems and waste management. In terms of labour force rate, there is no significant trend in labour force participation among the states as presented in Figure 4.3 (d).

Overall, the labour force participation rate across all states is consistently above 60%. Nevertheless, W.P. Kuala Lumpur and Johor have the highest labour force participation rates, with values of approximately 70% each. It reflects strong workforce engagement and possibly better employment opportunities, industrial activities, and access to economic centers in these states.

(iii) Recyclable Waste Collection Distribution by State

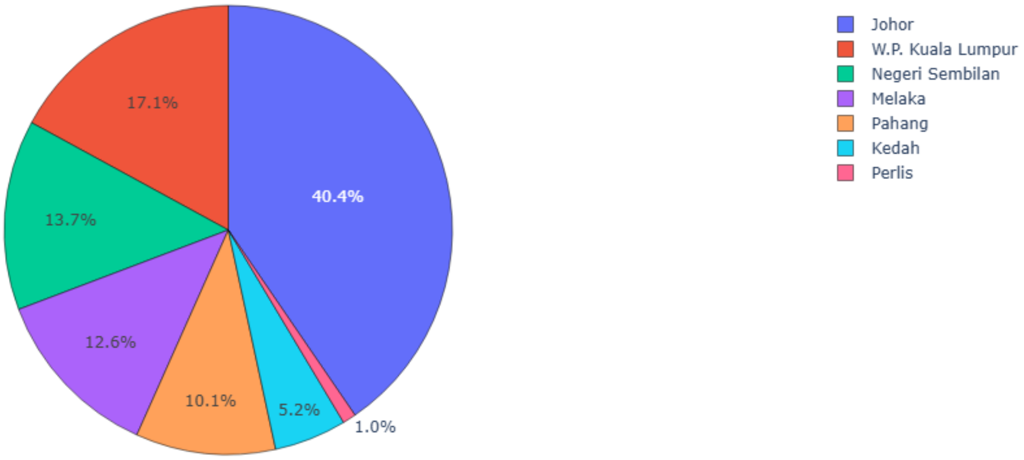


Figure 4.4 Recyclable Waste Collection Distribution by State

The distribution of recyclable waste collection by state in Figure 4.4 shows that Johor holds the largest contribution at 40.4% compared to other states. W.P. Kuala Lumpur follows with 17.1%, while Negeri Sembilan (13.7%), Melaka (12.6%), and Pahang (10.1%) contribute mid-range proportions. Kedah and Perlis have the smallest total recyclable waste collection, at 5.2% and 1.0%, respectively. The distribution shows a clear dominance by Johor, which accounts for nearly half of the total recyclable waste collection. The higher recyclable waste collection in Johor and W.P. Kuala Lumpur may indicate a combination of higher waste generation and relatively effective waste management practices. Effective waste management is particularly important in urbanized and economically active regions like Johor and W.P. Kuala Lumpur, where higher consumption often leads to greater recyclable waste output. Jamaludin et al. (2022) proposed two models for household food waste: linear (ending in landfills) and circular (reuse/recycling). They found that only low-income, less-educated groups chose the linear model, likely due to unawareness of the food waste problem's severity. Low-income groups had the highest food waste rates.

These findings align with A. Kumar et al. (2018) who observed that socioeconomic groups play a critical role in recycling behavior and plastic waste management. In developing countries, the composition of plastic waste varies from 5% to 8% of total MSW with much of it discarded after single use. The study found that middle socioeconomic regions have the most active in recycling and recovering plastic waste, achieving a 93% recycling rate. This was primarily due to their tendency to sell recyclable plastic waste to informal waste buyers (IWBs) for revenue generation. In contrast, low socioeconomic regions demonstrated the lowest recycling and recovery rates (44%) due to a lack of awareness about recycling. Despite this, the households still engaged in plastic recovery for income. High socioeconomic regions had a recycling rate of only 67% although it generated the highest waste. This lower rate was attributed to their perception that selling recyclable plastic waste yielded minimal financial benefit. These findings suggest that financial incentives and awareness significantly influence recycling practices across socioeconomic groups.

4.2.2 Bivariate Analysis

The bubble chart visualizes the relationship between solid waste generation against influencing factors. The color gradients indicate the state and the bubble sizes represent the total contribution of each state to overall waste.

(i) Analysis of GDP per Capita and Solid Waste Generation by State

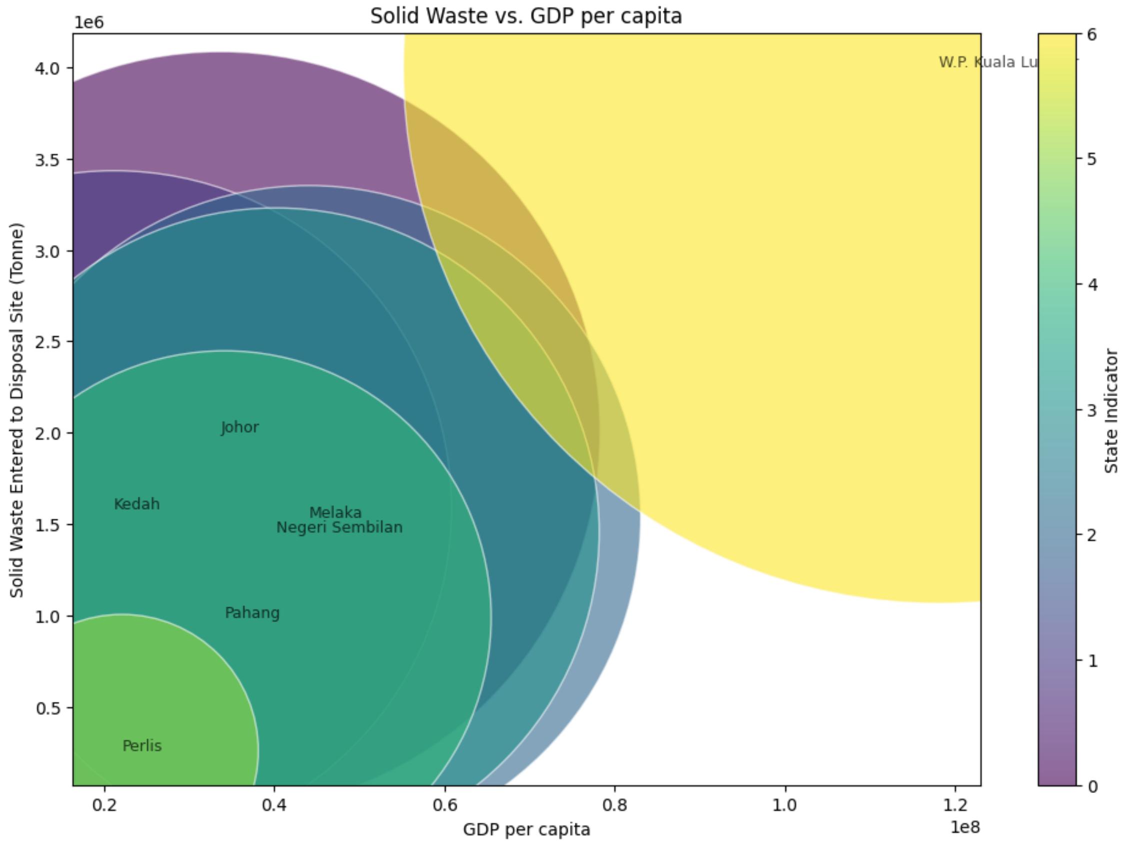


Figure 4.5 Solid Waste against GDP per Capita

The bubble chart in Figure 4.5 illustrates a clear positive correlation between GDP per capita and solid waste generation among Malaysian states. W.P. Kuala Lumpur leads with the highest GDP per capita and produces the largest amount of solid waste, exceeding 4 million tonnes. Kedah contributes approximately 4 times more waste than Perlis, even though their GDP per capita values are relatively close (approximately RM 30 million for Kedah and RM 20 million for Perlis). This disparity could be influenced by other factors, such as population density, urbanization, or industrial activity.

Despite Johor having a GDP per capita roughly one-third that of W.P. Kuala Lumpur and not the second-largest GDP per capita contributor, it ranks as the second-highest waste contributor that generate with approximately 2 million tonnes of waste. This substantial waste generation surpasses Melaka, which has the second-highest GDP per capita. It may be attributed to Johor's higher labour force participation rate (Figure 4.3 (d)), as it ranks as the second highest among all states in this regard. The active workforce in Johor likely drives increased economic activities and consumption patterns that influence its significant waste output.

These findings align with previous studies, such a s Abdella Ahmed et al. (2022) and A. Kumar et al. (2018) which associate higher socioeconomic status with greater waste generation due to increased consumption of packaged goods. Wealthier households tend to generate higher volumes of carton, glass, and plastic waste due to high purchase rates in affluent regions. However, A. Kumar & Samadder (2017) emphasized that lower socioeconomic region tends to produce more non-biodegradable wastes as it is easily available at low cost for domestic use.

(ii) Analysis of Fertility Rates and Solid Waste Generation by State

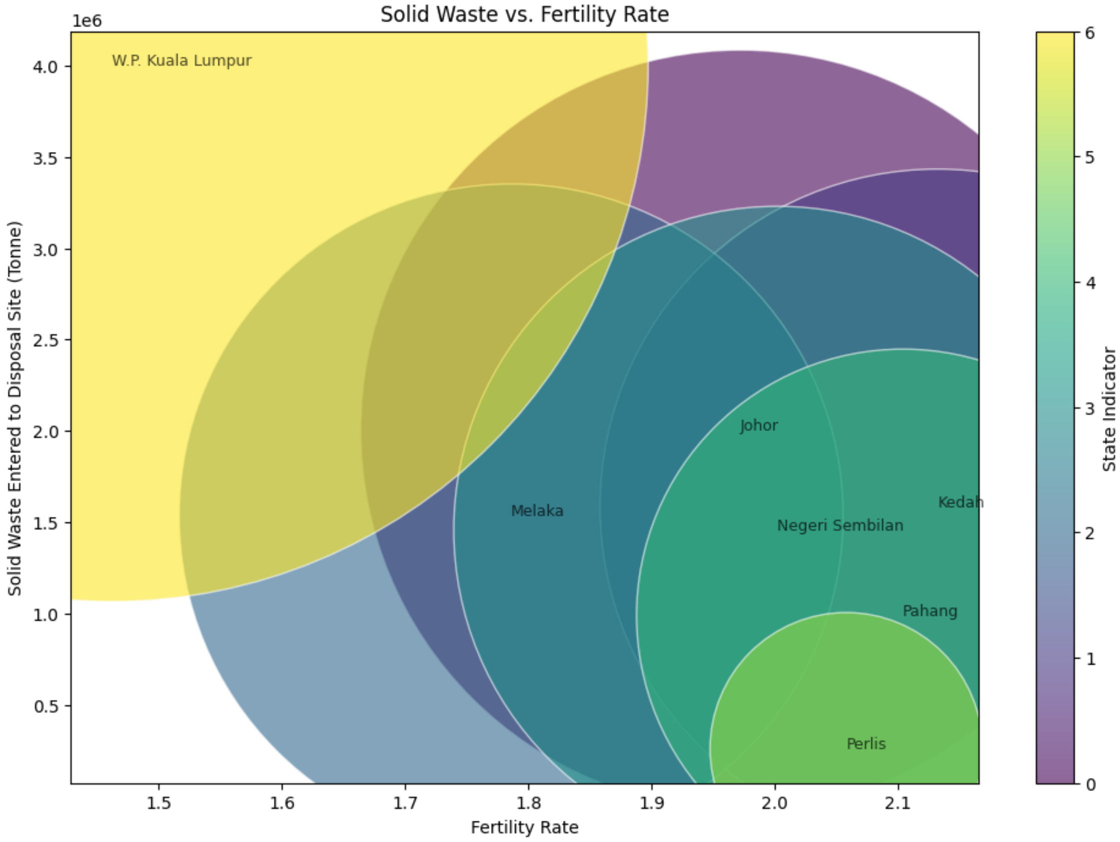


Figure 4.6 Solid Waste against Fertility Rate

Figure 4.6 highlights a negative correlation between fertility rates and solid waste generation. W.P. Kuala Lumpur, despite having the lowest fertility rate of approximately 1.5, produces the highest amount of solid waste, exceeding 4 million tonnes. Conversely, Perlis, with the highest fertility rate of approximately 2.1, contributes the least to solid waste generation at around 500,000 tonnes. This pattern suggests that states with lower fertility rates often experience higher levels of urbanization and economic activity. These factors contribute to significantly higher waste generation. In contrast, states with higher fertility rates and less urbanization tend to produce less waste.

These findings align with Hoy et al. (2022), who identified a strong negative correlation between fertility rate and specific types of (MSW, such as food, garden waste, paper, plastic, glass, and textiles. The study suggests that fertility rate has a stronger influence on certain waste types compared to total population. This indicates that the age structure of a population—reflected in fertility rates—plays a role to determine both the type and quantity of waste generated. For example, higher fertility rates often associated with households with children, may lead to increased generation of paper, plastic, and metal waste due to consumption patterns specific to family-oriented needs.

This correlation is further supported by Novriadhy et al. (2021)who focused on the role of Household Age Structure (HAS) in waste generation at the household level in Palembang City. Their findings show that households with higher proportions of children under five or women of childbearing age tend to produce less total waste. This is due to distinct consumption patterns in high-fertility households that generate specific waste types like diapers and pharmaceutical waste. In contrast, urban households in low-fertility regions generate more waste from packaged goods and luxury items which contribute to higher levels of non-biodegradable waste.

In summary, the negative correlation between fertility rates and solid waste generation highlights the influence of demographic structures on waste patterns. Higher fertility rates in less urbanized areas result in lower waste generation with distinct waste types, and vice versa.

(iii)Analysis of Crude Death Rate and Solid Waste Generation by State

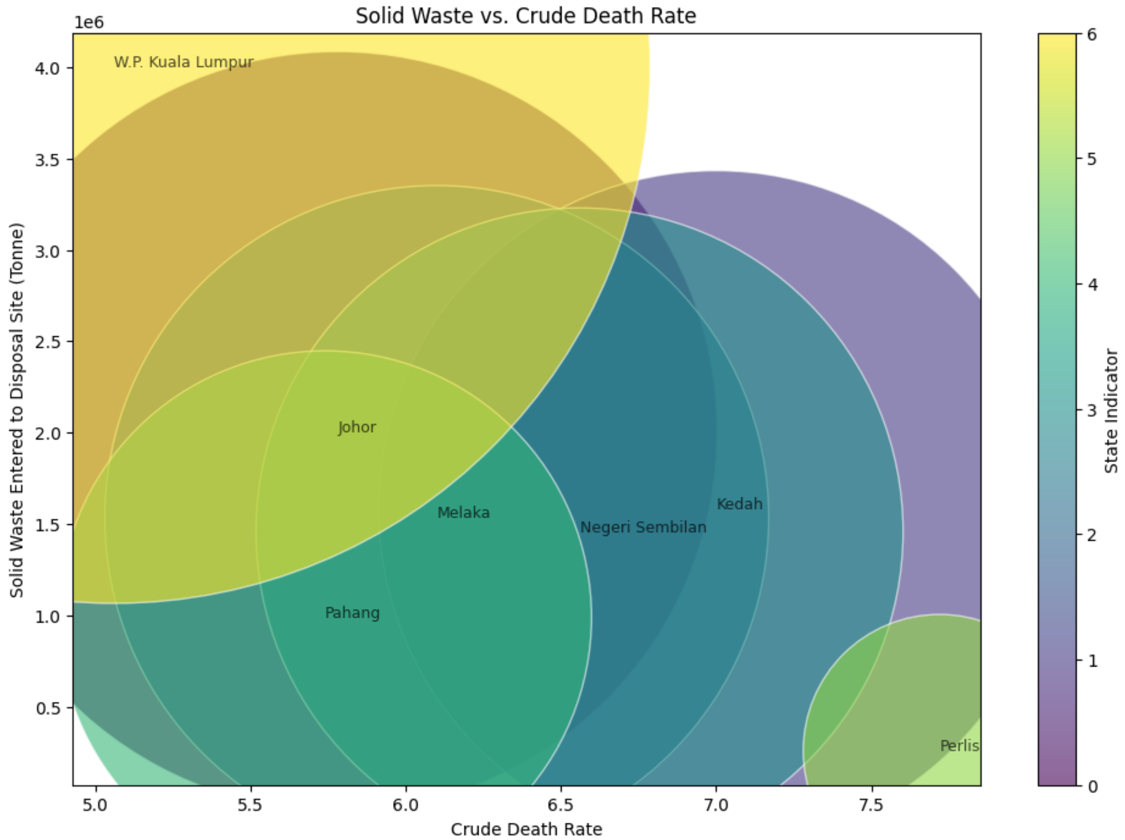


Figure 4.7 Solid Waste against Crude Death Rate

W.P. Kuala Lumpur generates the highest solid waste at over 4 million tonnes, despite having the lowest crude death rate of approximately 5.0 as illustrated in Figure 4.7. Meanwhile, Perlis with the highest crude death rate of 7.5, contributes only 500,000 tonnes. Crude death rates often reflect regions with older populations. These areas typically experience lower economic activity which result in reduced consumption and waste generation. States with lower crude death rates, such as W.P. Kuala Lumpur and Johor, produce more waste due to higher consumption levels, while states like Perlis and Kedah, with higher crude death rates generate less waste.

Amiruddin et al. (2023) found that young couples and single-person households who cook at home often generate food waste after cooking. Younger people tend to waste more food Jamaludin et al. (2022) due to less experience managing household food

Radzymińska et al. (2016) and lower involvement in food waste prevention Abd Razak et al. (2018). According to RECYCLING Magazine (2019), growing elderly require more healthcare products. The ongoing use and disposal of medical items increased the volume of waste, particularly non-biodegradable or hard-to-recycle materials. Furthermore, their unfamiliarity with modern waste management practices can create inefficiencies in disposal processes. These findings support the analysis that regions with lower crude death rates are indicative of younger populations and higher economic activity. These factors contribute to increased consumption levels, resulting in greater waste production.

(iv) Analysis of Labour Force Rate and Solid Waste Generation by State

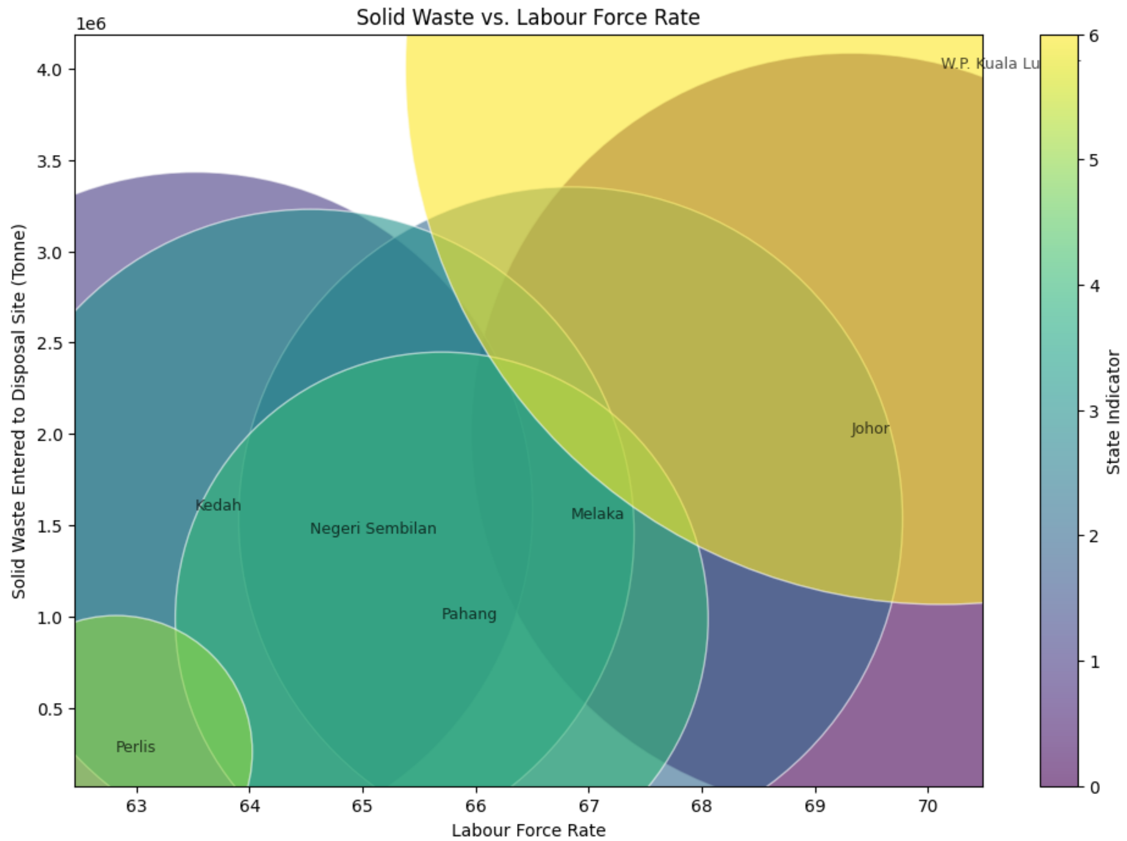


Figure 4.8 Solid Waste against Labour Force Rate

The bubble chart in Figure 4.8 shows a positive relationship between labour force participation and solid waste generation. Higher labour force participation typically indicates increased economic activity. For example, W.P. Kuala Lumpur has the highest labour force rate of 70% and the largest amount of solid waste generation that exceeds 4 million tonnes. It indicates a strong correlation between economic activity and waste output. Furthermore, Perlis has the lowest labour force rate of approximately 64% and smallest amount of solid waste generation around 0.5 million tonnes. It reflect its smaller population and lower economic activity compared other states. This increased economic activity leads to greater waste accumulation. As larger populations with higher labour force participation typically generate more waste through household, industrial, and commercial activities.

Supporting this, Elshaboury et al. (2021) found that labour force participation rates significantly influence the population's purchasing power which is a critical factor in waste disposal trends. Areas with higher labour force participation often experience greater economic activity. It led to increased consumption of goods and services and higher waste output. Conversely, regions with lower labour force participation may generate less waste due to reduced purchasing power and lower resource utilization. A. Kumar et al. (2018)provided additional insights into waste generation patterns where populations with lower labour force participation generate more inert waste due to reliance on solid fuels such as coal, firewood, and briquettes for daily living. This reliance on solid fuels contributes significantly to overall waste generation, particularly in areas where cleaner or alternative fuels are not readily accessible.

These findings underline the complex relationship between labour force participation, economic factors, and waste management. While higher labour force participation often leads to increased waste generation due to economic activity, the type and nature of waste also depend on factors such as employment type, income levels, and access to resources.

(v) Insights on Total Solid Waste Generation

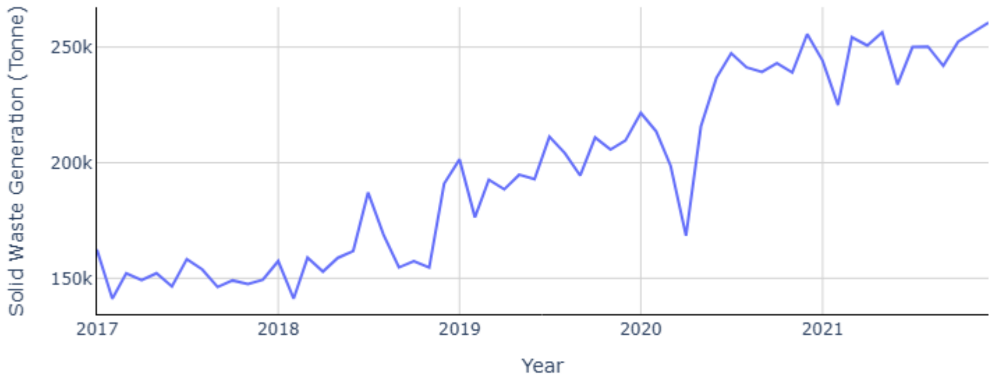


Figure 4.9 Total Solid Waste Generation Over Time

The line chart in Figure 4.9 depicts the trend of total solid waste generation in Malaysia from 2017 to 2021. Overall, the data reveals a steady increase in waste generation over time, with significant fluctuations in certain periods. From 2017 to late 2018, solid waste generation remained relatively stable. In 2019, total waste generation experienced a sharp rise, surpassing 200,000 tonnes. It reflects a significant growth during this period. In 2020, a sharp decline in total solid waste generation is observed. This drop likely coincided with the COVID-19 pandemic, which may have affected economic activity and waste production (Nasir et al., 2021; Norkhadijah et al., 2023). By late 2020, the total solid waste generation stabilized at around 250,000 tonnes and maintained its consistency through 2021.

Densely populated areas face more significant challenges in managing food waste due to the increased volume generated (Bharadi et al., 2022). In the study conducted by Ismail & Azeman (2021) that focus on Kuantan, Pahang, a region in Malaysia with a relatively larger population, greater number of landfills and a higher volume of waste compared to other states. The increased waste generation is attributed to higher consumption rates and the convenience culture prevalent in urban settings, where food is often over-purchased and discarded easily. This is supported by a study by Amiruddin et al. (2023), which found that higher population density in urban areas leads to greater pressure on waste management systems. As a result, there are inadequate disposal practices and the overuse of landfills.

(vi) Cumulative Analysis of Solid and Recyclable Waste Generation Over Time

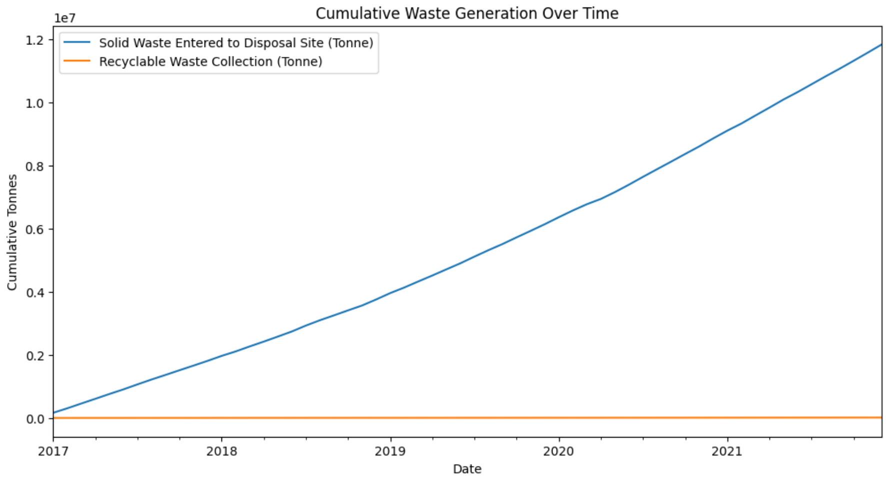


Figure 4.10 Disparities in Waste Disposal and Recycling Efforts Over Time

The chart in Figure 4.10 illustrates the cumulative waste generation over time for both solid waste and recyclable waste in Malaysia from 2017 to 2021. The solid waste entered into disposal sites shows a consistent and steep upward trend, reaching approximately 12 million tonnes by the end of 2021. The growing accumulation of solid waste over the years is likely driven by improper management of waste streams. According to Ishak (2024), ASEAN countries received 17% of global plastic waste between 2017 and 2021. In 2021, Malaysia turns into global dumping ground as it more than imported 500,000 tonnes of plastic waste and exported only 11,000 tonnes. This increase on cumulative waste generation also resulted from China's 2018 ban on importing most plastics and other materials, which redirected waste shipments to Southeast Asia.

In contrast, the recyclable waste collection remains relatively stagnant and contributes only a small fraction of the total waste generated. The findings align with Ishak (2024) and Nasir et al. (2021) where Malaysia lacks adequate recycling facilities and public understanding about proper waste segregation. Most waste (89%) ends up in landfills, reducing recycling opportunities. Challenges such as illegal dumping, poor practices by waste operators, and unclear enforcement further hinder progress. Although efforts to promote a circular economy exist, delays and lack of accountability slow improvements.

This disparity suggests a limited emphasis on recycling efforts compared to the overwhelming volume of waste being disposed of. Malaysia needs better infrastructure, public education, and stricter regulations with incentives for sustainable waste management to increase recycling rates.

4.2.3 Multivariate Analysis

(i) Statewise Trends in Recyclable Waste Collection Over Time

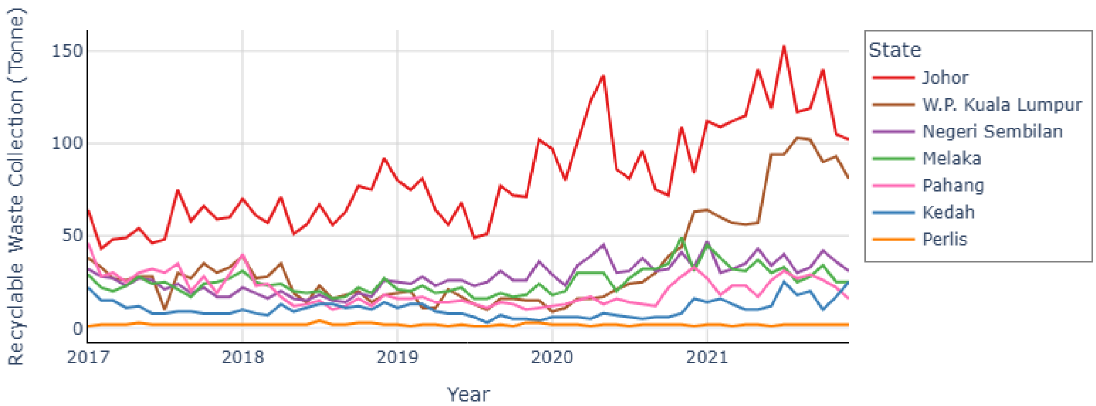


Figure 4.11 Recyclable Waste Collection by State Over Time

Figure 4.11 show the trend of recyclable waste collection by state in Malaysia from 2017 to 2021. Johor consistently leads in recyclable waste collection with a peak reaching over 150 tonnes around 2020. W.P. Kuala Lumpur follows with a gradual increase in recycling rates with a peak reach at approximately 100 tonnes in 2021. Overall, Negeri Sembilan, Melaka, Pahang, and Kedah show slight increases in recyclable waste collection particularly after late 2020. The trend also highlights significant disparities among states, with Johor and W.P. Kuala Lumpur demonstrated stronger recycling efforts compared to other states, which exhibit only modest progress. Meanwhile, Perlis remains the lowest contributor, with collection levels consistently low throughout the years. This might due to its smaller population and limited waste management resources.

The COVID-19 pandemic in early 2020 brought significant changes in consumption patterns and waste management behaviors. According to SWCorp director, there was increased awareness about sustainability during the pandemic where more individuals likely participated in recycling initiatives (Bernama, 2020). This shift may have contributed to the slight upward trends seen in several states during and after 2020.

The sudden spike in recyclable waste collection observed in 2021 for Johor and W.P. Kuala Lumpur is likely linked to the launch of the Malaysian Recycling Alliance (MAREA) in January 2021. MAREA focuses on urban recycling efforts that aim to enhance the recycling value chain and improve post-consumer packaging recovery. This spike reflects the success of campaigns promoting waste separation at the source. It also highlights advancements in recycling infrastructure and increased public awareness driven by MAREA's multi-stakeholder approach(Invest KL, 2022)

In conclusion, this analysis highlight the challenges faced by less urbanized states to increase recycling efforts. A focused approach that combined infrastructure development, public awareness campaigns, and tailored policies for rural and less-developed regions, is essential. This will ensure that all states can contribute meaningfully to Malaysia's sustainability goals.

(ii) Statewise Trends in Solid Waste Generation Over Time

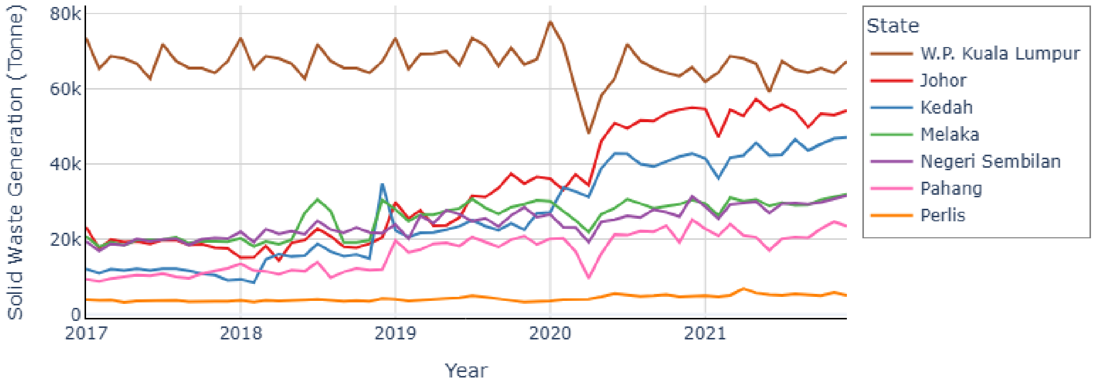


Figure 4.12 Solid Waste Generation by State Over Time

The line chart in Figure 4.12 shows the trend of solid waste generation by state in Malaysia from 2017 to 2021. Generally, all states except W.P. Kuala Lumpur exhibit a gradual increase in waste generation after 2020. W.P. Kuala Lumpur consistently generates the highest amount of solid waste, maintaining values between 60,000 and 80,000 tonnes over the years with a sharp decrease in 2020 which likely due to the COVID-19 pandemic. Johor follows as the second-largest contributor, with waste generation ranging from 40,000 to 60,000 tonnes, showing a steady upward trend. Meanwhile, the smallest states which is Perlis consistently generates the least waste that remained below 10,000 tonnes throughout the period. Most food waste happens at the consumer level due to culture of excess and convenience (Bharadi et al., 2022). The availability of abundant food choices and the convenience of discarding food without significant financial consequences contribute to this problem (Azeman et al., 2021).

Norkhadijah et al. (2023) explained that the disparity in solid waste generation among states is influenced by population size and GDP per capita. For instance, Johor and W.P. Kuala Lumpur were the top waste producers in 2019 since they contributed 72.3% of Malaysia’s total GDP per capita. This is largely attributable to their higher populations, with Johor having 3.76 million residents and Kuala Lumpur 7.78 million in 2019.

On the other hand, the waste generation trends in Melaka and Negeri Sembilan are almost identical throughout the years. This is likely due to their similar population sizes, with Negeri Sembilan at 1.13 million and Melaka at 0.93 million in 2019 as mentioned by Nasir et al. (2021). Also, Negeri Sembilan has faced challenges in managing household waste and reported running out of suitable locations to dispose of waste in landfills which may have contributed to increased waste. Moreover, there is a noticeable spike in Melaka's waste generation around the middle of 2018. This is consistent with Nasir et al. (2021) where the spike is linked to the start of Ramadan in Malaysia. During this period, more waste is generated as restaurants and eateries become crowded with people breaking their fast or having supper.

The volume of solid waste plummeted for most states in early 2020. This coincides with Malaysia's total lockdown in March 2020 due to the COVID-19 pandemic. These measures reduced commercial, industrial, and even household activities in many areas, leading to a sharp decline in waste generation. Additionally, the increase in recycling rates during this period likely contributed to the reduction in waste volumes. Nasir et al. (2023) described the sudden drop in solid waste generation as a 'random shock' (short-term memory) due to reduction of solid waste production as people stayed at home during the lockdown.

Starting in April 2020, waste volumes began to rise again across all states. It is likely due to the implementation of the Conditional Movement Control Order (CMCO), which allowed limited economic activities to resume. Norkhadijah et al. (2023) emphasized that the relaxation of restrictions that allow eateries, hawkers, and markets to operate until 10pm, led to higher waste generation observed particularly after June 2020. Moreover, the rise in plastic waste, particularly from personal protective equipment (PPE), along with the growing volume of packaging waste driven by increased online shopping further shaped waste generation patterns during this period.

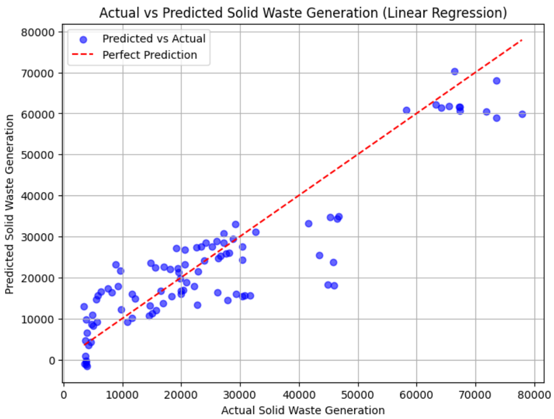
For Johor and W.P. Kuala Lumpur, waste generation after 2021 remains constant. This might be attributed to the urban recycling efforts initiated by the Malaysian Recycling Alliance (MAREA). As discussed in the recyclable waste collection analysis, these effort could have helped stabilize waste production. Table 4.3 presents a summary comparison of waste generation and recycling trends across different states and categories.

Table 4.3 Summary Comparison of Waste Generation and Recycling Trends

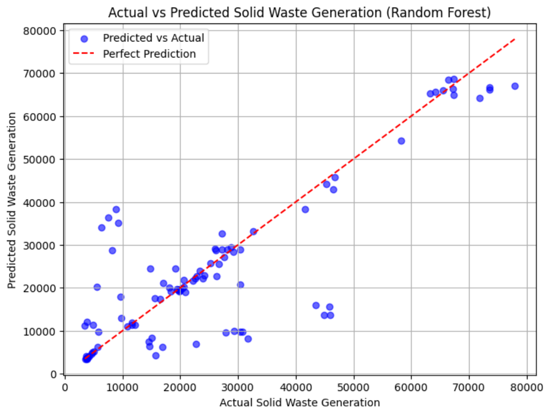
Category	States	Solid Waste Generation	Recyclable Waste Collection
High Waste, High Recycling	W.P. Kuala Lumpur, Johor	Kuala Lumpur: Highest waste, sharp decrease in 2020 due to COVID-19. Johor: Second-highest contributor with steady upward trend.	Johor: Leads in recycling with the highest peak in 2020. Kuala Lumpur: Gradual increase with highest peak in 2021.
Moderate Waste, Moderate Recycling	Kedah, Melaka, Negeri Sembilan, Pahang	Moderate waste generation with Melaka and Negeri Sembilan trends almost identical likely due to their similar population sizes.	Slight increases particularly after late 2020.
Low Waste, Low Recycling	Perak	Lowest throughout the years.	Lowest throughout the years.
COVID-19 Impact Group	All States	Significant drop in 2020 as people stayed home, rebound post-2020 with a drastic increase in plastic waste after June 2020.	Slight improvement due to increased sustainability awareness
Malaysian Recycling Alliance (MAREA) Group	W.P. Kuala Lumpur, Johor	Stabilized waste production post-2021 due to urban recycling efforts.	Reflects success of targeted urban recycling campaigns.

4.3 Data Evaluation

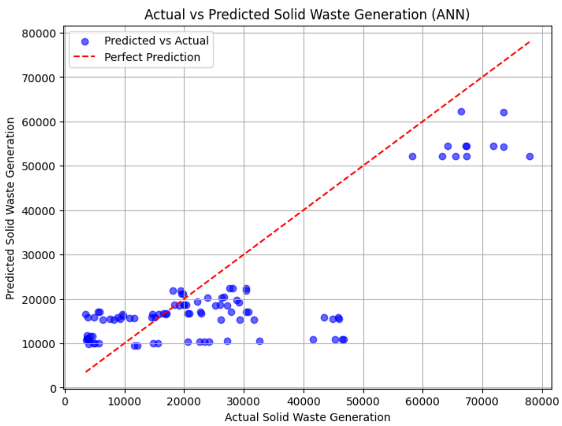
In this section, the results of all predictive models are evaluated using statistical metrics to determine their accuracy and reliability in predicting solid waste output. Key predictors such as GDP per capita, crude death rate, fertility rate, and labour force participation rate were identified as determinants in solid waste output. These variables capture the socioeconomic and demographic factors that contribute to waste production across different regions. The categorical variable state was one-hot encoded to ensure the model can better capture the unique waste generation patterns specific to each state.



(a) Multiple Linear Regression



(b)Random Forest



(c) ANN

Figure 4.13 Graph of Actual vs Predicted

When the models were developed using the MLR algorithm as shown in Figure 4.13 (a), the scatterplot shows a relatively close clustering of points along the perfect prediction line for mid-range values of solid waste generation. According to Azadi & Karimi-Jashni (2016), it indicates a strong correlation between the two variables if the data points form a straight line. If the slope of this line is close to one, it suggests that the variables have a proportional and balanced relationship. The differences in the performance of the developed models can be attributed to the use of various prediction algorithms. However, some deviation is observed for higher actual values. Random Forest algorithm in Figure 4.13 (b) also shows a larger spread of points away from the line particularly for higher values of solid waste generation despite some clustering near the line. The ANN scatterplot in Figure 4.13 (c) reveals a less consistent alignment with the perfect prediction line. Many points deviate significantly.

4.3.1 Model Performance Comparison

In the study, three performance metrics are evaluated. Table 4.4 reveals clear trends in the evaluation of various modeling techniques across different metrics.

Table 4.4 Evaluation Performance on the Test Set

Model	MAE	RMSE	R^2
Multiple Linear Regression	6240.22	8351.79	0.82
Random Forest	6225.69	10804.01	0.70
Artificial Neural Network	9662.38	12760.36	0.58

Since the R^2 score is the most widely used performance metric in the reviewed literature on waste generation, it will be the primary measure for comparison. According to A. Kumar & Samadder (2017), the R^2 value typically below 0.50 unless the models include numerous independent variables. With respect to R^2, the results indicated that MLR and RF have a good fit the test data with MLR as the best prediction model. MLR achieves the highest R^2 score of 0.82 which outperforms RF (R^2 = 0.70) and ANN (R^2 =

0.58). The scores for each model particularly the ANN, are lower compared to those reported in previous literature (Azadi & Karimi-Jashni, 2016; Dissanayaka & Vasanthapriyan, 2019; A. Kumar et al., 2018; Swaak, 2021). This can be explained by relatively small datasets and the external features included in the study. ANN typically require large amounts of data to learn complex patterns effectively, and model cannot generalize well to unseen inputs with limited data (Abbasi & El Hanandeh, 2016; S. Kumar et al., 2020; Nasir et al., 2023).

Based on MAE score, RF model performed most accurate with a score of 6225.69. It showed that RF has better average prediction accuracy than the other models. Despite Random Forest sensitivity to outliers, it performs well on this dataset and provides the most reliable results. Furthermore, ANN model showed the worst performance according to the MAE with score of 9662.38. With only two hidden layers (64 and 32 neurons, respectively) and normal weight initialization, the model might have been too shallow or lacked sufficient capacity to fully exploit the features in the dataset. Additionally, the absence of advanced techniques such as dropout for regularization or batch normalization could have limited its ability to generalize effectively.

With respect to RMSE, ANN demonstrated higher RMSE compared to MLR and RF. It indicates a greater sensitivity to larger errors. ANN shows the least suitable model due to its lowest R2 and the highest error metrics. In the study by Dissanayaka & Vasanthapriyan (2019) and A. Kumar et al. (2018), RF consistently outperformed MLR because it was tuned through hyperparameter optimisation and cross-validation. These techniques allow RF to better handle complex, non-linear relationships and improve its predictive accuracy by selecting the most optimal model parameters. Given these findings, further cross-validation between RF and MLR will help determine the more accurate and reliable model.

Table 4.5 shows the five folds cross-validation result of MLR and RF. The mean R^2 scores for MLR and RF are 0.84 and 0.78, respectively. Since standard MLR does not require hyperparameter tuning, we will proceed to optimize the hyperparameters of the RF model and re-evaluate its R^2 scores to explore potential performance improvements. RF was selected for hyperparameter tuning due to its potential for improvement through adjustments in tree depth, the number of trees, and other key parameters.

Table 4.5 Multiple Linear Regression and Random Forest Cross-Validation Result

MLR Cross-Validation R^2 Scores	0.87	0.86	0.84	0.78	0.84
RF Cross-Validation R^2 Scores	0.83	0.77	0.83	0.69	0.79

From the hyperparameter tuning using GridSearchCV, the optimized Random Forest model was identified with the following parameters: a maximum depth of 5, the square root of the total number of features considered at each split, minimum samples split of 5, 128 decision trees in the ensemble, and a random seed value of 42 for reproducibility. This optimized model achieved an R^2 of 0.85, MAE of 5239.97, and RMSE of 7537.56.

After hyperparameter optimization, the RF model was re-evaluated on both validation and test datasets to confirm its performance gains. The results show that the optimized RF model outperforms the MLR model in terms of R^2 and error metrics as described in Table 4.6.

Table 4.6 Optimised Random Forest Cross-Validation Result

Cross-Validation R^2 Scores	0.89	0.86	0.87	0.80	0.84
Cross-Validation RMSE Scores	7227.05	7187.17	6457.82	8888.83	7013.78

The cross-validation results of the optimized RF model in Table 4.6 showed that all R^2 scores are relatively high, with an average of 0.85. Although one-fold has an R^2 of 0.80, it remains acceptable as it captures approximately 80% of the variance and the performance is consistent across all folds. The model's average of R^2 of 0.85 and RMSE of 7354.93 indicate improved accuracy and reliability compared to the untuned version (R^2 = 0.70, RMSE = 10804.01) and MLR (R^2 = 0.82, RMSE = 8351.79). These findings align with the study by Dissanayaka & Vasanthapriyan (2019) where hyperparameter optimization of RF model allowed the model to balance bias and variance. Similarly, A. Kumar & Samadder (2017) emphasizes that without cross-validation the model is prone to biased predictions and overfitting as it may not perform consistently on unseen data. The use of 5-fold cross-validation balanced training and testing helped validate the model's performance and optimized its settings which resulted in reliable and accurate predictions of waste generation rates (A. Kumar et al., 2018). We can conclude that the optimised RF model is better suited for predicting solid waste generation due to balanced accuracy and generalization compared to MLR.

4.3.2 Feature Importance

Since the optimised RF model outperforms the other models, permutation importance scores for the selected features are calculated using this model. The inclusion of the state variable which was one-hot encoded to represent unique regional patterns further improves the model's predictions.

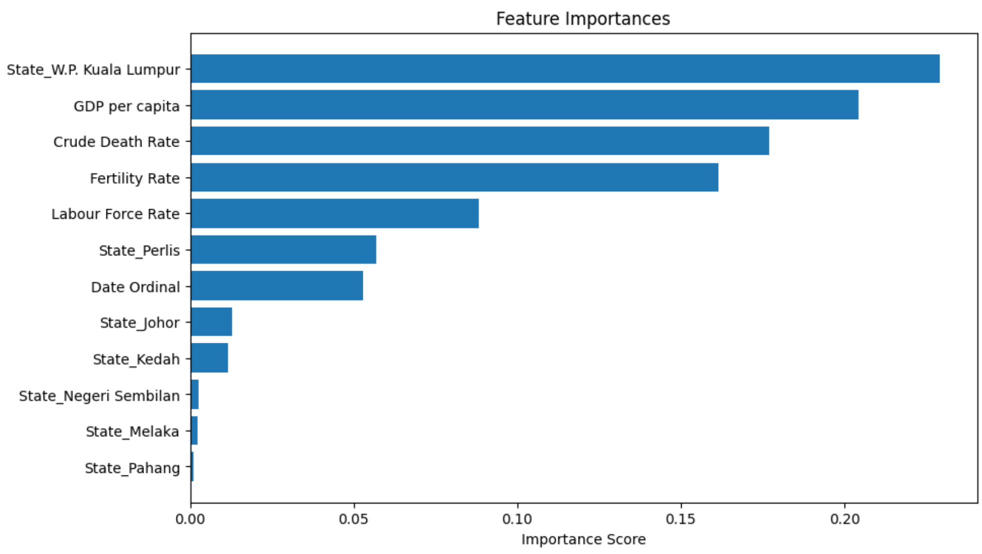
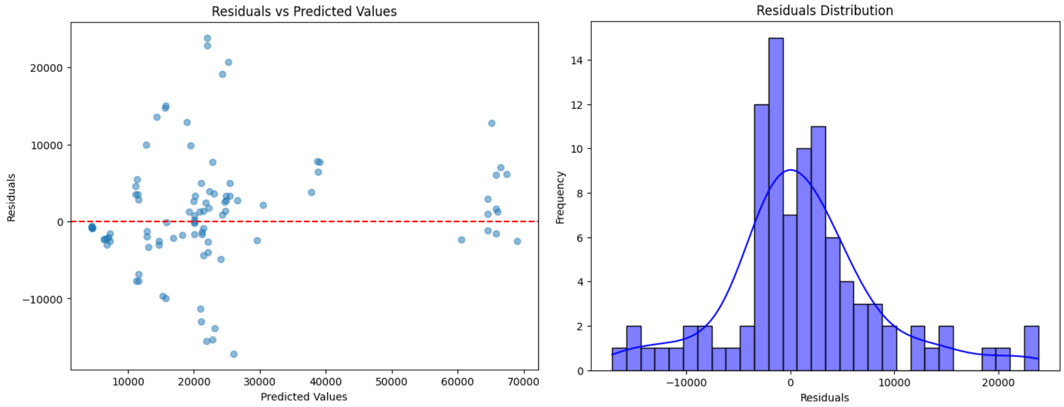


Figure 4.14 Graph of Feature Importance

The feature importance analysis in Figure 4.14 highlights the critical role of socioeconomic and demographic factors in solid waste generation. The high importance of 'State_W.P. Kuala Lumpur' and GDP per capita in the feature analysis indicated the influence of urbanization and economic activities on waste generation. In contrast, states like Pahang, Melaka, and Negeri Sembilan have lower importance scores indicated their comparatively smaller contributions to waste generation. The findings are similar to Abdella Ahmed et al. (2022), Dissanayaka & Vasanthapriyan (2019), Azadi & Karimi-Jashni (2016) and Ghinea et al. (2016) where urban centers generate higher per capita waste due to greater consumption and diverse waste streams. Over the last 20 years, the rapid pace of urbanization and population growth has caused household solid waste in Malaysia to double (Ng et al., 2023). Higher-income urban populations tend to produce more recyclable waste such as plastics, paper, and glass due to their consumption habits. Conversely, lower-income areas generate less waste but may still contribute significantly to organic and mixed waste streams (Abdella Ahmed et al., 2022; Cheng et al., 2022).

4.3.3 Residual Analysis



(a) Residuals vs. Predicted (b) Residuals Distribution

Figure 4.15 Residual Analysis of the Model Prediction

Figure 4.15 (a) evaluates the performance of the optimized RF model. Most residuals are scattered around the horizontal line at 0. This indicates that the model's predictions are relatively accurate for the majority of the data points and model does not exhibit significant bias. There is an indication of heteroscedasticity where the spread of residuals increases as the predicted values increase. This suggests that the model may struggle to predict higher values with consistent accuracy. There are a few extreme residuals where errors exceed ±20,000. These outliers indicate instances where the model underperformed, possibly due to unusual patterns or variability in the data that the model could not fully capture.

The histogram in Figure 4.15 (b) provides insights into the error distribution of the optimised RF model. The residuals are primarily centered around 0, with the highest frequency observed near the mean. This indicates that the model's predictions are accurate for most data points with minimal bias. However, there is a slight skew to the positive values reveals a tendency for the model to slightly underestimate some predictions.

4.3.4 Analysis of Actual vs Predicted MSW Generation

In summary, while the optimised model demonstrates strong predictive performance, its accuracy could be further improved by addressing variability in extreme predictions and reducing the spread of residuals for higher predicted values.

Table 4.7 Performance Metrics on Training and Test Datasets

Optimised Random Forest	R^2	RMSE
Training	0.88	6623.75
Test	0.85	7537.56

The table presents the R^2 and RMSE values for the optimised RF model, evaluated on both the training and test datasets. On the training dataset, the model achieves an R^2 of 0.88, indicating that it explains 88% of the variability in the target variable. The corresponding RMSE value of 6623.75 suggests a low level of prediction error. This indicates the model's ability to fit the training data accurately.

On the test dataset, the model achieves an R^2 of 0.85, indicating that it captures 85% of the variability in the target variable on unseen data. The RMSE of 7537.56 on the test dataset shows slightly higher error compared to the training dataset but remains within an acceptable range. The test RMSE is slightly higher than the training RMSE, which is expected. However, the difference is not large which suggests minimal overfitting. This makes the model reliable for predicting solid waste generation.

(i) Analysis of Yearly Mean Predicted Solid Waste by State

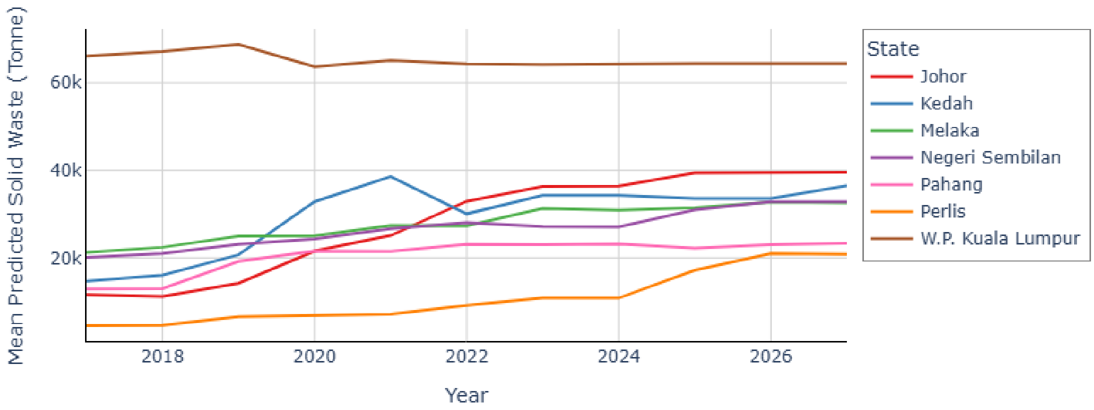


Figure 4.16 Yearly Mean Predicted Solid Waste by State

The graph in Figure 4.16 shows the mean predicted differences in waste generation trends across regions. from 2018 to 2026. W.P. Kuala Lumpur consistently leads with the highest predicted waste generation with values above 60,000 tonnes throughout the period. This reflects the state’s high degree of urbanization and economic activity contribute significantly to solid waste production. Johor follows as the second-largest contributor with predicted waste generation steadily increasing from approximately 40,000 tonnes in 2018 to nearly 45,000 tonnes by 2026.

In contrast, states such as Kedah and Melaka exhibit moderate growth trends. By 2026, their waste generation is projected to converge around 30,000 tonnes. Smaller states like Negeri Sembilan, Pahang, and Perlis show significantly lower levels of waste generation.

Perlis, the smallest state, remained below 20,000 tonnes throughout the observed period. While most states show steady growth in waste generation, trends begin to stabilize around 2024–2026.

These trends emphasize the huge differences between urbanized states, such as W.P. Kuala Lumpur and Johor, and less urbanized regions like Perlis and Pahang. The findings highlight the need to develop waste management strategies that address the growing waste volumes in highly urbanized states while ensuring smaller states are prepared for future increases. It allows authorities to strategically plan landfill capacities across Malaysia to promote sustainable environmental management.

(ii) Analysis of Actual vs Predicted Total Solid Waste Generation Over Time

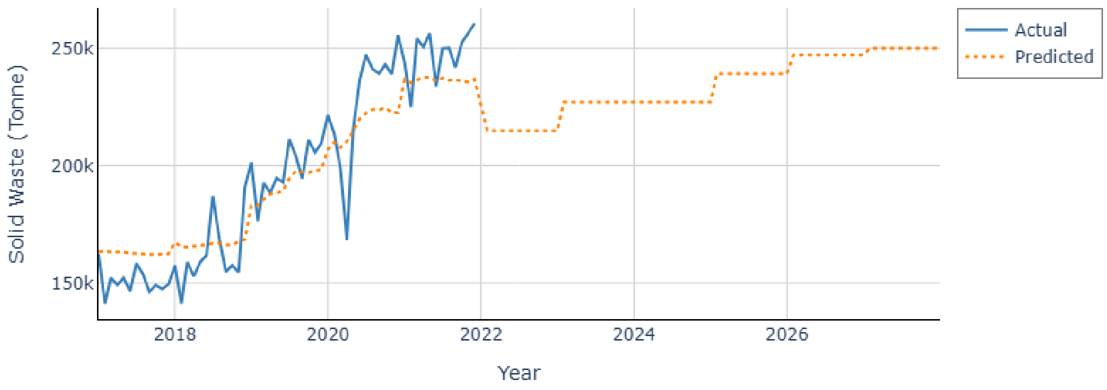


Figure 4.17 Actual vs Predicted Total Solid Waste Generation Over Time

Figure 4.17 compares the actual solid waste generation with the predicted solid waste generation over time. The actual data fluctuates due to its monthly recording. In contrast, the predicted values are derived from yearly aggregated features (e.g., GDP per capita, fertility rate, labor force rate), which remain constant throughout each year. Therefore, the predicted values is smoother in trend.

Since the independent variables (e.g., GDP, fertility rate) are annual values, it is assumed that these values remain unchanged across months within the same year. This assumption simplifies the prediction process but does not capture the finer granularity of monthly variations.

This limitation highlights the impact of using annual data for monthly predictions. The absence of high-resolution features reduces the model's accuracy in predicting monthly or seasonal variations. Incorporating monthly or seasonal feature data could significantly improve the model’s ability to predict monthly solid waste generation and align better with actual data.

(iii) Analysis of Yearly Mean Actual vs Predicted Solid Waste Generation

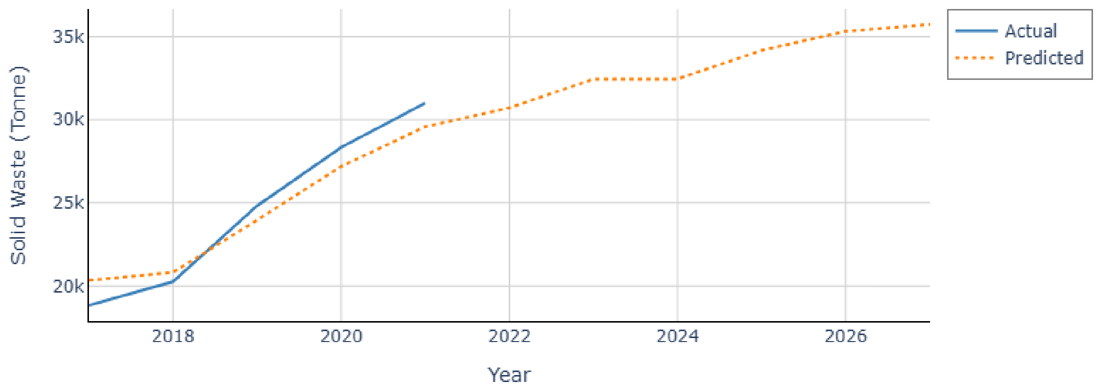


Figure 4.18 Yearly Mean Actual vs Predicted Solid Waste Generation

The graph in Figure 4.18 compares the yearly mean solid waste of actual with predicted values. From 2017 to 2021, the predicted values closely align with the actual yearly mean. This suggests that the model effectively captures the overall upward trend in solid waste generation during this period.

Beyond 2021, the model forecasts a continued but slower increase in waste generation, with the trend stabilizing at approximately 35,000 tonnes by 2026. This suggests that the model assumes waste generation will reach a saturation point, possibly influenced by the socioeconomic or demographic variables included in the predictions. While the model accurately reflects long-term trends, it fails to capture short-term variations or anomalies, such as the drop in 2020. This limitation arises from the use of yearly aggregated features, which lack the granularity needed to account for monthly or seasonal changes.

Table 5.1 provides a comparison of actual and predicted average solid waste generation from 2017 to 2021 by state, along with predicted values for future years (2022–2027). The regional disparities highlight the importance of designing waste management strategies to the specific needs of each state.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

5.1.1 Identification and Analysis of Influential Variables Affecting Waste Generation

In conclusion, the research identified several influential variables in predicting MSW generation in Malaysia. The GDP per capita, crude death rate, fertility rate, and labor force participation were determined as significant contributors to waste generation trends based on their high correlation coefficients and feature importance. The high importance of some states and GDP per capita in the feature analysis emphasize the impact of urbanization and economic activity on waste generation. This aligns with previous studies that highlight the interconnected relationship between urbanization and GDP per capita, where urbanization leads to higher GDP. Consequently, higher waste generation due to diverse waste streams.

5.1.2 Evaluation of Predictive Performance of Multiple Linear Regression, Random Forest, and Artificial Neural Networks for Solid Waste Generation

Three forecasting techniques of MLR, RF and ANNs for MSW management were evaluated in this study. Specific metrics such as RMSE, MAE and R-squared are used to evaluate the forecasting efficacy. The extensive evaluations, comparisons, and findings related to the predictive model are discussed in Section 4.3. Generally, MLR and RF have demonstrated good result but ANN underperformed.

Among the models tested and hyperparameter tuning of RF, the optimized RF model emerged as the best predictive model. With parameters including a maximum depth of 5, the square root of the total number of features considered at each split, minimum samples split of 5, 128 decision trees in the ensemble, and a random seed of 42, the model achieved the highest R^2 (0.85) and the lowest error metrics (MAE: 5239.97, RMSE: 7537.56).

These results confirm the suitability of the optimized RF model for forecasting solid waste generation. Therefore, we can conclude that simpler ML models can achieve good results as long as they meet the problem's requirements, with parameter tuning being relatively straightforward. In contrast, advanced ML models typically require larger datasets and more intensive optimization efforts to achieve improved accuracy and effectiveness.

Using the prediction algorithm, future solid waste generation from 2022 to 2027 was forecasted for each state. Based on a comparison with historical actual values, the proposed predictive models demonstrate strong performance and promising results for forecasting future waste generation. The findings also conclude that the disproportionately high solid waste generation in W.P. Kuala Lumpur and Johor are worrying. It also highlight the challenges of waste management in Malaysia, particularly the limited success of recycling initiatives compared to the overwhelming volume of solid waste. It is alarming that current landfills may soon be unable to accommodate future waste. Without intervention, this overload could lead to severe environmental pollution and hinder economic growth.

5.2 Recommendations

Several actions are recommended to mitigate this growing issue. Poor public awareness stems from a lack of knowledge about food waste and its environmental consequences often stems from insufficient education at a young age. Misunderstandings about date labels such as "use-by" and "best before" contribute significantly to food waste. Therefore, there should be an increase in public composting awareness and promotion of proper waste separation and disposal practices at the household level.

Next, consumer preferences for fresh and aesthetically perfect produce can lead retailers to discard imperfect but edible items. Regulatory authorities and industry associations should work to reduce the strict cosmetic standards for the fresh produce in retailers. Additionally, the government should offer tax incentives to businesses that donate surplus food to charities. This would make it economically viable for companies to participate in food recovery efforts and support sustainable practices.

Furthermore, the lack of precise inventory tracking results in perishable goods remaining in storage beyond their optimal selling period. Consequently, large quantities of food are disposed of without any opportunity for recovery or redistribution. Retailers should adopt advanced technology to predict food demand and manage raw ingredients as well as to monitor dynamic pricing

Overall, these insights emphasize the need for predictive models' integration into waste management planning to address rising waste challenges. This research could benefit key stakeholders such as the Ministry of Housing and Local Government (KPKT), Solid Waste Management, Public Cleansing Corporation (SWCorp) and local municipal council. For example, the Ministry of Housing and Local Government (KPKT) could use these findings to develop targeted policies for reducing waste generation and improving recycling rates. Similarly, SWCorp and local municipal councils can use the predictive models to allocate resources more efficiently, plan waste management infrastructure, and address areas with disproportionately high waste generation. Such efforts would align with Malaysia's commitment to achieve its SDGs and promote sustainable urban development.

5.3 Limitations

This study has several limitations that should be addressed in future research. First, the use of annual socioeconomic and demographic data limited the model's ability to capture monthly or seasonal variations in waste generation. Since waste production can be influenced by seasonal trends, the incorporation of more granular data could significantly improve the predictive accuracy and responsiveness of the model.

Next, the size of the dataset was relatively small. This limited dataset size may reduce the model's ability to generalize across different regions or capture rare patterns in waste generation. Increasing the size of the dataset by collecting additional data points could improve the accuracy and reliability of the predictions. Future research should use longer historical records or expand the dataset to include all Malaysian states and federal territories, even those without direct landfill operations, as their waste generation dynamics also contribute to the national context. A larger dataset would also enable the use of more advanced machine learning models that perform better with higher data volumes, such as ANNs.

While the ANN model demonstrated lower accuracy compared to other models, this does not exclude its potential for improvement. Further exploration of its architecture and performing hyperparameter optimization could improve its ability to handle complex patterns and address outliers. The higher errors for regions or periods with unusual waste generation patterns in this study reveals that ANN or alternative deep learning approaches might be better suited for handling these anomalies in future work.

Lastly, this study did not include variables that could have a significant impact on waste generation, such as policy changes, technological advancements in waste management, economic disruptions, and public awareness campaigns. These factors play a critical role in shaping waste management trends and their exclusion limits the comprehensiveness of the model. Including such dynamic variables in future research could provide a more holistic understanding of the factors influencing waste generation.

REFERENCES

Abbasi, M., & El Hanandeh, A. (2016). Forecasting municipal solid waste generation using artificial intelligence modelling approaches. *Waste Management*, *56*, 13–22. https://doi.org/10.1016/j.wasman.2016.05.018

Abd Razak, S., Abd Ghafar, S. W., Mohd Padzil, N. A., Kamaruddin, A., Mat Zin, N., Saim, M., & MS Suhaimi, A. H. (2018). Household food wastage prevention in Malaysia: An Issue Processes Model perspective. *Economic and Technology Management Review*, *13*, 51–62.

Abdella Ahmed, A. K., Ibraheem, A. M., & Abd-Ellah, M. K. (2022). Forecasting of municipal solid waste multi-classification by using time-series deep learning depending on the living standard. *Results in Engineering*, *16*. https://doi.org/10.1016/j.rineng.2022.100655

Adelodun, B., & Choi, K. S. (2020). Impact of food wastage on water resources and GHG emissions in Korea: A trend-based prediction modeling study. *Journal of Cleaner Production*, *271*. https://doi.org/10.1016/j.jclepro.2020.122562

Agarwal, V. (2019, November 30). *Outlier detection with boxplots*. Medium. https://medium.com/@agarwal.vishal819/outlier-detection-with-boxplots-1b6757fafa21

Ali, S. A., & Ahmad, A. (2019). Forecasting MSW generation using artificial neural network time series model: a study from metropolitan city. *SN Applied Sciences*, *1*(11). https://doi.org/10.1007/s42452-019-1382-7

Amiruddin, N. N. H. N., Rozamri, N. A., Baharudin, F., & Mohamad, I. N. (2023). A Study on household food waste management and composting practice. *IOP*

Conference Series: Earth and Environmental Science, *1205*(1). https://doi.org/10.1088/1755-1315/1205/1/012018

Araiza-Aguilar, J. A., Rojas-Valencia, M. N., & Aguilar-Vera, R. A. (2020). Forecast generation model of municipal solid waste using multiple linear regression. *Global Journal of Environmental Science and Management*, *6*(1), 1–14. https://doi.org/10.22034/gjesm.2020.01.01

Azadi, S., & Karimi-Jashni, A. (2016). Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province, Iran. *Waste Management*, *48*, 14–23. https://doi.org/10.1016/j.wasman.2015.09.034

Azeman, A. S., Ismail, S. N., Mohamad, M. A., Mohamad, N. H., & Mohd, N. S. (2021). Consumer food waste intention in Klang Valley: A review and analysis. *International Journal of Academic Research in Business and Social Sciences*, *11*(16). https://doi.org/10.6007/ijarbss/v11-i16/11222

Bernama. (2020, August 2). *SWCorp aims 40 per cent recycling rate by 2025.* BERNAMA. https://www.bernama.com/en/news.php?id=1866343

Bharadi, V., Jadhav, P., Nanche, O., & Munj, O. (2022). Food waste management using machine learning. *International Journal of Creative Research Thoughts (IJCRT)*, *10*(4). www.ijcrt.org

Bhat, S. (2023, January 11). *A comprehensive guide to random forest regression*. Medium. https://medium.com/@bhatshrinath41/a-comprehensive-guide-to-random-forest-regression-43da559342bf

Boehmke, D., & Greenwell, B. (2020, February 1). *Hands-on machine learning with R*. https://bradleyboehmke.github.io/HOML/random-forest.html

Cheng, K. M., Tan, J. Y., Wong, S. Y., Koo, A. C., & Amir Sharji, E. (2022). *A review of future household waste management for sustainable environment in Malaysian cities*. https://doi.org/10.20944/preprints202205.0074.v1

Daud, S. (2021). The COVID-19 pandemic crisis in Malaysia and the social protection program. *Journal of Developing Societies*, *37*(4), 480–501. https://doi.org/10.1177/0169796X211041154

Dissanayaka, D. M. S. H., & Vasanthapriyan, S. (2019). Forecast municipal solid waste generation in Sri Lanka. *2019 International Conference on Advancements in Computing (ICAC*, 210–215.

Dutta, S. (2024, July 17). *What is shuffling the data? A guide for students*. Medium. https://medium.com/@sanjay_dutta/what-is-shuffling-the-data-a-guide-for-students-0f874572baf6

Elshaboury, N., Abdelkader, E. M., Alfalah, G., & Al-Sakkaf, A. (2021). Predictive analysis of municipal solid waste generation using an optimized neural network model. *Processes*, *9*(11). https://doi.org/10.3390/pr9112045

Firdose, T. (2023, May 29). *Filling missing values with mean and median*. Medium. https://tahera-firdose.medium.com/filling-missing-values-with-mean-and-median-76635d55c1bc

Fokker, E., Koch, T., & Dugundji, E. R. (2023). Short-term time series forecasting for multi-site municipal solid waste management. *Procedia Computer Science*, *220*, 170–179. https://doi.org/10.1016/j.procs.2023.03.024

Gatto, A. (2024). *Towards sustainable global food and biomass systems: Interactions between food loss and waste reductions, dietary shifts, and transitioning to a circular bio-based economy* [Doctoral dissertation]. Wageningen University.

Ghinea, C., Drăgoi, E. N., Comăniţă, E. D., Gavrilescu, M., Câmpean, T., Curteanu, S., & Gavrilescu, M. (2016). Forecasting municipal solid waste generation using prognostic tools and regression analysis. *Journal of Environmental Management*, *182*, 80–93. https://doi.org/10.1016/j.jenvman.2016.07.026

Hoy, Z. X., Woon, K. S., Chin, W. C., Hashim, H., & Fan, Y. Van. (2022). Forecasting heterogeneous municipal solid waste generation via Bayesian-optimised neural network with ensemble learning for improved generalisation. *Computers and Chemical Engineering*, *166*. https://doi.org/10.1016/j.compchemeng.2022.107946

Huey, F. T. (2021, July). *Substituting missing data with the group average: Why it's good to be cautious*. Towards Data Science. https://towardsdatascience.com/substituting-missing-data-with-the-group-average-why-its-good-to-be-cautious-d64bead7a029

Intharathirat, R., Abdul Salam, P., Kumar, S., & Untong, A. (2015). Forecasting of municipal solid waste quantity in a developing country using multivariate grey models. *Waste Management (New York, N.Y.)*, *39*. https://doi.org/10.1016/j.wasman.2015.01.026

Ishak, M. Y. (2024, June 7). *Malaysia battles to avoid waste chokehold*. Eco-Business . https://www.eco-business.com/opinion/malaysia-battles-to-avoid-waste-chokehold/

Ismail, S. N., & Azeman, A. S. (2021). Assessing consumers' behavior towards food waste in Pahang, Malaysia. *Jurnal Intelek*, *16*(2), 48–59. https://doi.org/10.24191/ji.v16i2.399

Jainvidip. (2024, July 2). *Understanding train, test, and validation data in machine learning*. Medium. https://medium.com/@jainvidip/understanding-train-test-and-validation-data-in-machine-learning-f8276165619c

Jamaludin, H., Elmaky, H. S. E., & Sulaiman, S. (2022). The future of food waste: Application of circular economy. *Energy Nexus*, *7*. https://doi.org/10.1016/j.nexus.2022.100098

Jereme, I. A., Siwar, C., Begum, R. A., & Abdul Talib, B. (2016). Addressing the problems of food waste generation in Malaysia. *International Journal of ADVANCED AND APPLIED SCIENCES*, *3*(8), 68–77. https://doi.org/10.21833/ijaas.2016.08.012

Johnson, N. E., Laniuk, O., Cazap, D., Liu, L., Starobin, D., Dobler, G., & Ghandehari, M. (2017). Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. *Waste Management*, *62*, 3–11. https://doi.org/10.1016/j.wasman.2017.01.037

Ministry of Housing and Local Government. (2023). *KPKT Statistics 2023*.

Kumar, A., & Samadder, S. R. (2017). An empirical model for prediction of household solid waste generation rate – A case study of Dhanbad, India. *Waste Management*, *68*, 3–15. https://doi.org/10.1016/j.wasman.2017.07.034

Kumar, A., Samadder, S. R., Kumar, N., & Singh, C. (2018). Estimation of the generation rate of different types of plastic wastes and possible revenue recovery from informal recycling. *Waste Management*, *79*, 781–790. https://doi.org/10.1016/j.wasman.2018.08.045

Kumar, S., Gaur, A., Kamal, N., Pathak, M., Shrinivas, K., & Singh, P. (2020). Artificial neural network-based optimum scheduling and management of forecasting municipal solid waste generation: A case study of Greater Noida in Uttar Pradesh (India). *Journal of Physics: Conference Series*, *1478*(1). https://doi.org/10.1088/1742-6596/1478/1/012033

Invest KL. (2022, June 22). *Leading the zero-waste movement in Greater Kuala Lumpur*. Invest KL. https://www.investkl.gov.my/insights/spotlight-greater-kl/leading-the-zero-waste-movement-in-greater-kuala-lumpur

Liegeard, J., & Manning, L. (2020). Use of intelligent applications to reduce household food waste. In *Critical Reviews in Food Science and Nutrition* (Vol. 60, Issue 6, pp. 1048–1061). Taylor and Francis Inc. https://doi.org/10.1080/10408398.2018.1556580

Lim, W. J., Chin, N. L. , Yusof, A. Y. , Yahya, A. , & Tee, T. P. (2016). Food waste handling in Malaysia and comparison with other Asian countries. *International Food Research Journal*. http://www.ifrj.upm.edu.my

Mahasan, N. S. (2023). *Experts: Food wastage expected to last till end-Syawal*. Bernama. https://www.bernama.com/en/bfokus/news.php?id=2181359#

MathWorks. (n.d.). *What is machine learning?* MATLAB & Simulink. Retrieved December 8, 2024, from https://www.mathworks.com/discovery/machine-learning.html

Nasir, N., Shariff, S. S. R., Januri, S. S., Zulkipli, F., & Md Yasin, Z. A. M. (2023). Time series forecasting of solid waste generation in selected states in Malaysia.

International Journal of Advanced and Applied Sciences, *10*(4), 76–87. https://doi.org/10.21833/ijaas.2023.04.009

Nasir, N., Zulkipli, F., Faizal, M., Ghadafy, M., & Azman, N. H. (2021). Forecasting solid waste generation in Negeri Sembilan and Melaka. *Journal of Quality Measurement and Analysis JQMA*, *1*, 61–77. http://www.ukm.my/jqma

Ng, K. S., Yeoh, L., Iacovidou, E., Wan Ab Karim Ghani, W. A., & Yamaguchi, A. (2023). *Towards sustainable municipal solid waste management in Malaysia*. https://eng.ox.ac.uk/synergors

Nik Mahdi, N. A., Fernando, Y., & Abdalla, Y. A. (2023). Understanding the sustainable development goals concept: Malaysia report and trend. *Journal of Governance and Integrity*, *5*(3), 317–327. https://doi.org/10.15282/jgi.5.3.2022.8938

Niu, D., Wu, F., Dai, S., He, S., & Wu, B. (2021). Detection of long-term effect in forecasting municipal solid waste using a long short-term memory neural network. *Journal of Cleaner Production*, *290*. https://doi.org/10.1016/j.jclepro.2020.125187

Norkhadijah, S., Ismail, S., Azwa, N., Tamrin, M., Abidin, E. Z., Rasdi, I., Shamsuddin, A. S., Udin, N. M., Alam, S., & Darul Ehsan, S. (2023). Assessing the impact of COVID-19 on solid waste generation and environmental health footprint: A case study. In *Malaysian Journal of Medicine and Health Sciences* (Vol. 19, Issue SUPP10).

Novriadhy, D., Komalasari, O., Hatta, H., & Oktarina, R. (2021). The effect of household's age structure on waste generation in Palembang City. *IOP Conference Series: Earth and Environmental Science*, *810*(1). https://doi.org/10.1088/1755-1315/810/1/012024

Peng, J., Jury, E. C., Dönnes, P., & Ciurtin, C. (2021). Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: Applications and challenges. In *Frontiers in Pharmacology* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fphar.2021.720694

Phooi, C. L., Azman, E. A., Ismail, R., Arif Shah, J., & Koay, E. S. R. (2022). Food waste behaviour and awareness of Malaysians. *Scientifica*, *2022*. https://doi.org/10.1155/2022/6729248

Radzymińska, M., Jakubowska, D., & Staniewska, K. (2016). Consumer attitude and behaviour towards food waste. *Journal of Agribusiness and Rural Development*, *10*(1). https://doi.org/10.17306/JARD.2016.20

RECYCLING Magazine. (2019, July 3). *Waste management and the elderly: A neglected problem?* RECYCLING Magazine. https://www.recycling-magazine.com/2019/07/03/waste-management-and-the-elderly-a-neglected-problem/

Sagi, O. (2024, March 24). *Three types of machine learning you should know*. Pecan AI. https://www.pecan.ai/blog/3-types-of-machine-learning/

Sharma, P. (2024, December 3). *Joblib: How to save and load machine learning models in Python*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2023/02/how-to-save-and-load-machine-learning-models-in-python-using-joblib-library/

Sodanil, M., & Chatthong, P. (2014). Artificial Neural Network-based time series analysis forecasting for the amount of solid waste in Bangkok. *Ninth International Conference on Digital Information Management (ICDIM 2014)*, 257.

Pennsylvania State University. (2024, August 4). *STAT 200: Elementary statistics*. Pennsylvania State University. https://online.stat.psu.edu/stat200/book/export/html/239

Swaak, B. (2021). *Machine learning in waste management: A model comparison for the prediction of biodegradable waste generation*. Tilburg University School of Humanities and Digital Sciences.

SWCorp Malaysia. (n.d.). *Kemudahan pengurusan sisa pepejal*. SWCorp Malaysia. Retrieved December 11, 2024, from https://www.swcorp.gov.my/tapak-pelupusan/

Syifaa, N., Shakil, M., Zahida, A., Azhar, M., & Othman, N. (2023). Solid waste management in Malaysia: An overview. In *Information Management and Business Review* (Vol. 15, Issue 1).

United Nations. (2024). *Sustainable Development Report*. https://dashboards.sdgindex.org/profiles/malaysia

Wahidah, S., & Ghafar, A. (2017). *Food waste in Malaysia: Trends, current practices and key challenges*. http://ap.fftc.agnet.org/ap_db.php?id=774&print=1

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining*, 29–40.

Wohlwend, B. (2023, July 14). *Regression Model Evaluation Metrics: R-Squared, Adjusted R-Squared, MSE, RMSE, and MAE* . Medium. https://medium.com/@brandon93.w/regression-model-evaluation-metrics-r-squared-adjusted-r-squared-mse-rmse-and-mae-24dcc0e4cbd3

Yusoff, S. H., Din, U. N. K. A., Mansor, H., Midi, N. S., & Zaini, S. A. (2018). Neural network prediction for efficient waste management in Malaysia. *Indonesian Journal of Electrical Engineering and Computer Science*, *12*(2), 738–747. https://doi.org/10.11591/ijeecs.v12.i2.pp738-747

Zulkipli, F., Nopiah, Z., Basri, N., Cheng, J., Zulkepli, J., & Khlaid, K. (2018). Integrated dynamical model for Malaysian solid waste management using system dynamics. *International Journal of Engineering & Technology (UAE)*, *7*(3), 131–135. www.sciencepubco.com/index.php/IJET

APPENDIX A: ACTUAL VS PREDICTED MSW BY STATE

Table 5.1 Average Solid Waste Generation (Tonnes) by State

States	Year	Average Solid Waste Generation (Tonnes)	
		Actual	Predicted
Johor	2017	9596.836	11599.392
	2018	9183.351	11205.262
	2019	15058.687	14188.005
	2020	23046.639	21551.183
	2021	26701.939	25112.468
	2022	NA	32949.386
	2023	NA	36349.920
	2024	NA	36444.229
	2025	NA	39427.810
	2026	NA	39592.261
	2027	NA	39592.261
Kedah	2017	11425.127	14693.536
	2018	16340.531	16025.041
	2019	23090.143	20693.891
	2020	37820.491	32861.034
	2021	43438.349	38586.035
	2022	NA	30059.111
	2023	NA	34304.746
	2024	NA	34333.764
	2025	NA	33532.600
	2026	NA	33592.265
	2027	NA	36485.255
Melaka	2017	19513.613	21226.322
	2018	22501.287	22415.193
	2019	27965.873	24996.498
	2020	28028.753	25069.018
	2021	29839.488	27438.467
	2022	NA	27323.891
	2023	NA	31290.543
	2024	NA	30916.054
	2025	NA	31434.139
	2026	NA	32786.320
	2027	NA	32617.219

Table 5.1 continued

States	Year	Average Solid Waste Generation (Tonnes)	
		Actual	Predicted
Negeri Sembilan	2017	19347.630	20103.938
	2018	22059.339	20963.820
	2019	25300.337	23101.757
	2020	25544.146	24319.799
	2021	29216.321	26639.655
	2022	NA	28068.269
	2023	NA	27122.623
	2024	NA	27104.509
	2025	NA	30961.343
	2026	NA	32866.305
	2027	NA	32818.596
Pahang	2017	10319.815	12858.601
	2018	11811.331	12995.150
	2019	18893.747	19224.502
	2020	19805.173	21570.289
	2021	21536.132	21474.961
	2022	NA	23118.196
	2023	NA	23026.096
	2024	NA	23156.320
	2025	NA	22236.428
	2026	NA	23032.603
	2027	NA	23347.009
Perlis	2017	3652.868	4528.550
	2018	3777.119	4588.253
	2019	4048.003	6536.903
	2020	4670.422	6930.706
	2021	5393.670	7144.981
	2022	NA	9138.999
	2023	NA	10841.763
	2024	NA	10821.626
	2025	NA	17199.297
	2026	NA	20997.478
	2027	NA	20846.162

Table 5.1 continued

States	Year	Average Solid Waste Generation (Tonnes)	
		Actual	Predicted
W.P. Kuala Lumpur	2017	67205.032	66085.585
	2018	67205.032	67124.115
	2019	69137.483	68775.708
	2020	64709.505	63657.520
	2021	65179.255	65098.392
	2022	NA	64297.644
	2023	NA	64150.410
	2024	NA	64357.828
	2025	NA	64357.828
	2026	NA	64357.828
	2027	NA	64357.828