



UNIVERSITAS  
**MERCU BUANA**

**KOMPARASI METODE REGRESI LINEAR DAN REGRESI RANDOM  
FOREST TERHADAP VOLUME PENGANGKUTAN SAMPAH**



*TUGAS AKHIR*

UNIVERSITAS  
Eka Pramudianzah  
41518010159  
MERCU BUANA

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2022**



**KOMPARASI METODE REGRESI LINEAR DAN REGRESI RANDOM  
FOREST TERHADAP VOLUME PENGANGKUTAN SAMPAH**

*Tugas Akhir*

Diajukan Untuk Melengkapi Salah Satu Syarat  
Memperoleh Gelar Sarjana Komputer

UNIVERSITAS  
MERCU BUANA

Oleh:

Eka Pramudianzah  
41518010159

PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS MERCU BUANA  
JAKARTA  
2022

## LEMBAR PERNYATAAN ORISINALITAS

Yang bertanda tangan dibawah ini:

NIM : 41518010159

Nama : Eka Pramudianzah

Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi Random Forest Terhadap Volume Pengangkutan Sampah

Menyatakan bahwa Laporan Tugas Akhir saya adalah hasil karya sendiri dan bukan plagiat. Apabila ternyata ditemukan didalam laporan Tugas Akhir saya terdapat unsur plagiat, maka saya siap untuk mendapatkan sanksi akademik yang terkait dengan hal tersebut.

Jakarta, 27 Juni 2022



Eka Pramudianzah



UNIVERSITAS  
MERCU BUANA

## SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa : Eka Pramudianzah  
NIM : 41518010159  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi Random Forest Terhadap Volume Pengangkutan Sampah

Dengan ini memberikan izin dan menyetujui untuk memberikan kepada Universitas Mercu Buana **Hak Bebas Royalti Noneksklusif** (*None-exclusive Royalty Free Right*) atas karya ilmiah saya yang berjudul diatas serta perangkat yang ada (jika diperlukan).

Dengan Hak Bebas Royalti/Noneksklusif ini Universitas Mercu Buana berhak menyimpan, mengalihmedia/formatkan, mengelola dalam bentuk pangkalan data (*database*), merawat dan mempublikasikan tugas akhir saya.

Selain itu, demi pengembangan ilmu pengetahuan di lingkungan Universitas Mercu Buana, saya memberikan izin kepada Peneliti di Lab Riset Fakultas Ilmu Komputer, Universitas Mercu Buana untuk menggunakan dan mengembangkan hasil riset yang ada dalam tugas akhir untuk kepentingan riset dan publikasi selama tetap mencantumkan nama saya sebagai penulis/pencipta dan sebagai pemilik Hak Cipta.

Demikian pernyataan ini saya buat dengan sebenarnya.

Jakarta, 27 Juni 2022



Eka Pramudianzah

## SURAT PERNYATAAN LUARAN TUGAS AKHIR

Sebagai mahasiswa Universitas Mercu Buana, saya yang bertanda tangan di bawah ini :

Nama Mahasiswa : Eka Pramudianzah  
NIM : 41518010159  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi Random Forest Terhadap Volume Pengangkutan Sampah

Menyatakan bahwa :

1. Luaran Tugas Akhir saya adalah sebagai berikut :

No	Luaran	Jenis	Status
1	Publikasi Ilmiah	Jurnal Nasional Tidak Terakreditasi	Diajukan ✓
		Jurnal Nasional Terakreditasi	
		Jurnal International Tidak Bereputasi	Diterima
		Jurnal International Bereputasi ✓	
Disubmit/dipublikasikan di :	Nama Jurnal	: International Conference on Engineering and Information Technology for Sustainable Industry 2022 (ICONETSI)	
	ISSN	: 978-1-4503-8771-2	
	Link Jurnal	: <a href="https://iconetsi.sgu.ac.id/2022/">https://iconetsi.sgu.ac.id/2022/</a>	
	Link File Jurnal Jika Sudah di Publish	: -	

2. Bersedia untuk menyelesaikan seluruh proses publikasi artikel mulai dari submit, revisi artikel sampai dengan dinyatakan dapat diterbitkan pada jurnal yang dituju.
3. Diminta untuk melampirkan scan KTP dan Surat Pernyataan (Lihat Lampiran Dokumen HKI), untuk kepentingan pendaftaran HKI apabila diperlukan

Demikian pernyataan ini saya buat dengan sebenarnya.

Mengetahui  
Dosen Pembimbing TA

Rahmat Budiarto, Dr. Prof

Jakarta, 27 Juni 2022



Eka Pramudianzah




## LEMBAR PERSETUJUAN PENGUJI

### LEMBAR PERSETUJUAN PENGUJI

NIM : 41518010159  
Nama : Eka Pramudianzah  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi  
Random Forest Terhadap Volume Pengangkutan  
Sampah

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 27 Juli 2022



(Yaya Sudarya Triana, Ph.D)

UNIVERSITAS  
MERCU BUANA

## LEMBAR PERSETUJUAN PENGUJI

NIM : 4151801059  
Nama : Eka Pramudianzah  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi  
Random Forest Terhadap Volume Pengangkutan  
Sampah

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 27 Juli 2022



(Saruni Dwiasnati, ST, MM, M.Kom)

UNIVERSITAS  
MERCU BUANA



## LEMBAR PERSETUJUAN PENGUJI

### LEMBAR PERSETUJUAN PENGUJI

NIM : 41518010159  
Nama : Eka Pramudianzah  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi  
Random Forest Terhadap Volume Pengangkutan  
Sampah

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 27 Juli 2022



(Dr. Harwikarya, MT)

UNIVERSITAS  
MERCU BUANA



## LEMBAR PENGESAHAN

NIM : 41518010159  
Nama : Eka Pramudianzah  
Judul Tugas Akhir : Komparasi Metode Regresi Linear dan Regresi Random Forest Terhadap Volume Pengangkutan Sampah

Tugas Akhir ini telah diperiksa dan disidangkan sebagai salah satu persyaratan untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Mercu Buana.

Jakarta, 27 Juli 2022

Menyetujui,



(Rahmat Budiarto, Dr. Prof)  
Dosen Pembimbing

Mengetahui,



(Wawan Gunawan, S.Kom, MT)  
Koord. Tugas Akhir Teknik Informatika



Ir. Emil R. Kaburuan, Ph.D., IPM)  
Ka. Prodi Teknik Informatika

## KATA PENGANTAR

Puji syukur kita panjatkan kehadirat Tuhan Yang Maha Esa yang senantiasa melimpahkan rahmat dan hidayah-Nya sehingga laporan tugas akhir ini dapat diselesaikan dengan sebaik-baiknya. Laporan tugas akhir ini ditulis sebagai salah satu syarat untuk menyelesaikan program Pendidikan Strata Satu Teknik Informatika di Universitas Mercu Buana yang berjudul “**Komparasi Metode Regresi Linear dan Regresi Random Forest terhadap Volume Pengangkutan Sampah**”.

Penulis menyadari bahwa tanpa bantuan, dukungan serta bimbingan dari beberapa pihak, laporan tugas akhir ini tidak dapat diselesaikan dengan sangat baik. Oleh karena itu, penulis mengucapkan terima kasih kepada:

1. Bapak Amir Hamzah dan Ibu Nur Cahyani Santoso yang selalu memberikan dukungan serta doa dalam kelangsungan pembuatan laporan tugas akhir.
2. Bapak Emil Robet Kaburuan, ST, MA, Ph.D., sebagai Ketua Program Studi Teknik Informatika di Universitas Mercu Buana.
3. Bapak Rahmat Budiarto, Dr. Prof., sebagai Dosen pembimbing dalam penelitian tugas akhir.
4. Ibu Saruni Dwiasnati, ST.MM., M.Kom., sebagai Dosen Pembimbing Akademik.
5. Seluruh Dosen Program Studi Teknik Informatika Universitas Mercu Buana.
6. Semua teman-teman yang selalu memberi semangat kepada penulis serta semua pihak yang telah membantu selama proses pengerjaan laporan tugas akhir.

Akhir kata, penulis berharap bahwa laporan tugas akhir ini dapat dijadikan acuan Pemerintah Provinsi DKI Jakarta dalam pengambilan keputusan tentang permasalahan pengangkutan sampah di Sungai Situ Waduk yang akan terjadi pada tahun 2022

Jakarta, 27 Juni 2022

Penulis

xi

## DAFTAR ISI

HALAMAN SAMPUL.....	i
HALAMAN JUDUL .....	i
LEMBAR PERNYATAAN ORISINALITAS .....	ii
SURAT PERNYATAAN PERSETUJUAN PUBLIKASI TUGAS AKHIR... iii	
SURAT PERNYATAAN LUARAN TUGAS AKHIR.....	iv
LEMBAR PERSETUJUAN PENGUJI .....	v
LEMBAR PENGESAHAN .....	viii
ABSTRAK .....	ix
ABSTRACT.....	x
KATA PENGANTAR.....	xi
DAFTAR ISI.....	xii
NASKAH JURNAL .....	1
KERTAS KERJA.....	8
BAB 1. LITERATUR REVIEW.....	9
BAB 2. ANALISIS DAN PERANCANGAN.....	21
BAB 3. SOURCE CODE.....	25
BAB 4. DATASET.....	51
BAB 5. TAHAPAN EKSPERIMEN.....	54
BAB 6. HASIL SEMUA EKSPERIMEN DAN KESIMPULAN.....	65
DAFTAR PUSTAKA .....	122
LAMPIRAN DOKUMEN HAKI.....	125
LAMPIRAN KORESPONDENSI .....	128

# Analysis of Waste Transportation Volume in Jakarta Province using Linear Regression and Random Forest Regression

Eka Pramudianzah<sup>†</sup>

Informatics Dept., Faculty of Computer Science

Mercu Buana University

Jakarta, Indonesia

41518010159@student.mercubuana.ac.id

Yaya Sudarya Triana<sup>†</sup>

Information System Dept., Faculty of Computer Science

Mercu Buana University

Jakarta, Indonesia

yaya.sudarya@mercubuana.ac.id

Rahmat Budiarto<sup>†</sup>

Informatics Dept., Faculty of Computer Science

Mercu Buana University

Jakarta, Indonesia

rahmat.budiarto@mercubuana.ac.id

## ABSTRACT

The accumulation of waste volume in the river waters of DKI Jakarta is still a significant problem that cannot be solved optimally because the population continues to increase every year, so the tonnage of waste also increases as well as some residents of DKI Jakarta still throw garbage into the river. In predicting the level of waste volume, the DKI Jakarta Provincial Environment Agency must make decisions, so it is necessary to carry out a prediction stage regarding the increase in waste in the future. For this reason, this research performs a prediction stage by utilizing two machine learning algorithms: Linear Regression and Random Forest Regression. The experiment used historical data on waste volume transportation from January to June 2021. The experimental results showed that the Random Forest Regression had the lowest error values of 0.82 and 0.81, with a training and testing data ratio of 80%:20%. On the other hand, Linear Regression has an error value of 0.83 and 0.82 at a ratio of 80%:20%. The analysis discussed in this study can be a reference for predicting and taking the necessary actions to prevent an increase in the volume of waste in DKI Jakarta Province.

## CCS CONCEPTS

• Computing methodologies • Machine learning

## KEYWORDS

Waste Transportation, Machine Learning, Linear Regression, Random Forest Regression

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [Permissions@acm.org](mailto:Permissions@acm.org).

ICONETSI, September 21–22, 2022, Tangerang, Indonesia

© 2022 Association for Computing Machinery.

ISBN 978-1-4503-9718-6/20/09...\$15.00

<https://doi.org/10.1145/3429789.34298xx>

## ACM Reference format:

Eka Pramudianzah, Yaya Sudarya Triana and Rahmat Budiarto. 2022. Analysis of Waste Transportation Volume in DKI Jakarta Province using Linear Regression and Random Forest Regression. In Proceedings of International Conference on Engineering and Information Technology for Sustainable Industry (ICONETSI 2022), September 21-22, 2022, Tangerang, Indonesia. ACM, New York, NY, USA, (NUMBER OF PAGES) pages.

## 1 Introduction

The waste problem in Jakarta, the capital of Indonesia, cannot be adequately solved because of high density of the population and at the same time continuously experience population growth from year to year. Thus, the tonnage of waste is increasing due to difficulties in managing waste piles [1]. Furthermore, the indifference of Jakarta residents to the environment leads to the cause of the tonnage of waste in Jakarta, which continues to increase yearly. Sources of waste are dominated by household, tourism, industrial, and marine waste in the small islands of Jakarta. The waste will usually end up in the Final Disposal Site (FDA), Temporary Disposal Site (TDS) and the sea or river [2]. Garbage thrown into the sea or rivers causes the garbage to end up on the beach so that it can cause the index of beaches/small islands in Indonesia to be "extremely dirty" and lower the visitor acceptance index. As a result, plastic waste dominates the tourism sector in Jakarta, with a percentage value of 83.86% of the total waste [3]. This fact is further strengthened because the coastal environment in Indonesia's seas has been polluted by plastic waste, especially in Jakarta Bay, where the highest microplastic contamination is 37.440 - 38.790/kg resulting from dry weight particles [4].

Based on data from the National Waste Management Information System, Jakarta's waste in 2020 was 3,054,812.22 tons, while the amount in Jakarta in 2021 reached 3,083,437.85 tons [5]. In addition, from October to December 2021, the volume of waste transported from the Jakarta River reached 121,433.53 m3. Even though the tonnage of waste always increases yearly, people in Jakarta still live on polluted riverbanks. Therefore, the Jakarta Provincial Government's efforts to reduce the volume of

waste in rivers are to budget IDR 1 trillion to normalize rivers and reservoirs in Jakarta and build a waste processing facility, namely the Intermediate Treatment Facility (ITF) [6]. This effort is very effective because, on February 7, 2021, the Provincial Government of Jakarta Province succeeded in reducing the volume of waste in the Jakarta River to 153.98 tons or 436 m<sup>3</sup>. The volume of this waste has decreased considerably compared to the volume on September 22, 2020, which amounted to 707.46 tons or around 2,033 m<sup>3</sup>. Therefore, in assisting the Jakarta Provincial Government in reducing the volume of waste from the river, a prediction method for the volume of waste transportation from Jakarta Rivers is required.

Kumar et al. [7] carry out research on the prediction of waste generation from residential areas in Vietnam using six machine learning algorithms, including Cubist Algorithm, k-Nearest Neighbors (k-NN), Artificial Neural Network (ANN), Linear Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). The data for the experiments are taken from the Vietnam National Bureau of Statistics that covers eight independent variables with aspects of consumption, economy, demographics, and waste generation in residential areas, while the dependent variable is the amount of waste collected daily. The results of this study indicate that the k-NN and Random Forest algorithms have good predictive abilities with R2 metric prediction values of 0.96 and 0.97, respectively; and the error value indicated by the mean absolute error (MAE) metric is 121.5 and 125.0, respectively. These results indicate that the k-NN and Random Forest algorithms are very suitable for assisting the Vietnamese Government in predicting waste generation by improving their recycling practices.

Another study was conducted by Kannangara *et al* [8] on predicting the generation and diversion of regional municipal solid waste in Canada using two machine learning algorithms,

i.e.: Decision Tree and ANN. The dataset used has been integrated with annual residential solid waste generation and paper transfer data from 220 cities in Ontario, Canada, along with socioeconomic data and demographic data based on the Canadian census program. The result of the study shows that the ANN produces a good accuracy value of 72%, while the decision tree algorithm only produces an accuracy value of 54%.

The main contribution of this research is a scheme to analyze and predict the volume of rivers waste transportation in Jakarta to assist the corresponding local agency in planning and managing the river waste in efficient manner.

## 2 Research Methodology

Figure 1 shows the proposed research design. It consists of four main modules, i.e.: Data Preprocessing, which includes data transformation, data cleansing, data selection, and data scaling; Split Validation; Prediction using Linear Regression (LR) and Random Forest Regression (RFR); and Validation of the prediction models and results. This study uses data from the Open Data Jakarta website from January to June 2021. The details of each module are explained in the following subsections.

### 2.1 Dataset Description

The data collected from the Open Data Jakarta website are grouped into 6 datasets by month (January, February, March, April, May, and June). Each dataset has a different number of rows, i.e.: 54343 for January data, 49252 for February data, 54715 for March data, 53100 for April data, 54870 for May data, and 53130 for June data. In addition, each dataset has 8 attributes, i.e.: month, point of waste collection location, sub-district, region, length or area of waste handling, unit length/area, date, and volume of waste per day in m<sup>3</sup> [9].

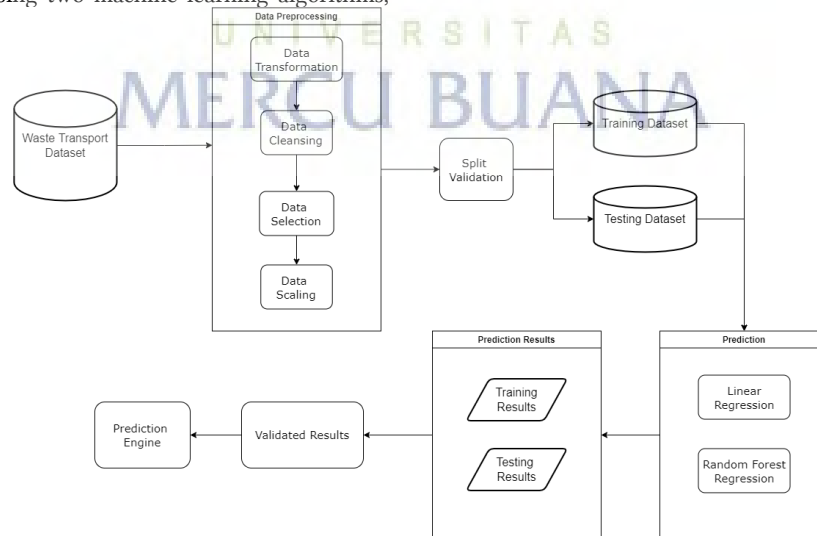


Figure 1. Research Flowchart in Predicting the Volume of Waste Transportation

## 2.2 Data Preprocessing

Preprocessing is several steps needed before entering the machine learning modeling stage. The dataset will be prepared first, such as changing values and data types, cleaning data from missing or outlier values, selecting data, and scaling data values. [8].

### 2.2.1 Data Transformation

The transformation carried out in this study is to change the naming of attributes that were previously written in Indonesian to English. In addition, the next transformation is also carried out on the data values, precisely on the attribute of the volume of waste per day ( $m^3$ ), which has the missing value converted into mean value.

### 2.2.2 Data Cleansing

At this stage, the data cleaning process is carried out on the value of outliers on the attribute volume of waste per day ( $m^3$ ). In addition, it removes attributes that are not needed during the research, i.e.: month attribute.

### 2.2.3 Data Selection

Data selection is the stage of selecting attributes to serve as independent and dependent variables. In this study, the attribute date is the independent variable, and the attribute volume of waste per day ( $m^3$ ) is the dependent variable.

### 2.2.4 Data Scaling

The last preprocessing carried out in this research is data scaling. Data scaling aims to overcome inconsistent data by making the data in the same range so that the resulting prediction results are accurate [10]. The scaling used in this study is a standard scaler where the data range for the average value is 0, and the standard deviation is 1 [11].

## 2.3 Split Validation

Split Validation is a process of selecting attributes for the independent and dependent variables and dividing randomly the data into two parts, namely training and testing data, based on the percentage of testing sizes that have been defined. In this study, the independent variable or predictor variable was selected based on the date of waste collected from several locations. In contrast, the volume of waste per day was the dependent or response variable [12].

## 2.4 Linear Regression

Linear Regression is a machine learning algorithm based on a supervised learning scheme that can predict between one or more independent variables ( $x$ ) and one dependent variable ( $y$ ) [13]. Both variables (independent and dependent variables) must consist of numerical data. The regression analysis output will conclude whether the independent variable has a linear relationship with the dependent variable. The basic formula for linear regression is represented in (1) [13].

$$Y = a + bX \quad (1)$$

Where:

$Y$ : forecast value (dependent variable).

$a$ : intercept.

$b$ : slope.

$X$ : number of periods (Independent variable).

## 2.5 Random Forest Regression

Random Forest is one of the ensemble learning methods that can perform several tasks such as classification and regression. The ensemble learning method is based on the premise that it will produce much higher performance than other machine learning algorithms to solve real equation search problems by controlling predictive values to avoid overfitting [14]. The Random Forest Regression Algorithm is very suitable for non-linear modelling relationships between variables because it can overcome complex

relationships between variables so that it will not be affected by multicollinearity [15]. The formula for Random Forest Regression is represented in (2) [16].

$$H(x) = \frac{1}{T} \sum_{t=1}^T h_t(x) \quad (2)$$

Where:

$T$ : number of regression trees.

$h_i(x)$ : output of the  $i$ -th regression trees ( $h_i$ ) on sample  $x$ .

## 2.6 Mean Absolute Error

Mean Absolute Error (MAE) is a metric that can measure the absolute error of an average value between the predicted value and the actual value [17]. The formula for the mean absolute error is represented in (3) [18].

$$MAE = \frac{1}{n} \sum_{i=1}^n |P_i - O_i| \quad (3)$$

Where:

$P_i$  and  $O_i$ : estimated and observed values.

$n$ : number of samples.

## 2.7 Mean Squared Error

Mean Squared Error (MSE) is a metric that can measure the difference in the average value obtained from the number of squared errors divided by the total number of predicted values, so this metric is very suitable to be used when weighting larger values for larger error values [19]. The formula for the mean squared error is represented in (4) [8].

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (4)$$

Where:

$n$ : number of samples

$\hat{Y}_i$ : predicted values

$Y_i$ : observed values

## 2.8 K-Fold Cross Validation

K-Fold Cross Validation is a cross-validation technique using a subset of data that has been randomized to test the success rate of a machine learning algorithm and can also be used in choosing which algorithm produces accurate accuracy values [20]. The randomized data subset consists of training and testing data where the subset has a predefined number of subsets ( $k$  value) [21]. Each  $k$  iteration is used for one testing data, and the rest is used as training data. The process is repeated until all data has been evaluated.

## 3 Result and Discussion

Differentiating the prediction results based on data distribution and data validation, each using outlier values and without using outlier values in the dependent variable. The reason why this experiment is distinguished based on the use of outlier values in the dependent variable is because by differentiating the experiment to determine the characteristics between the independent and dependent variables and the data points that affect the distribution of the relationship between the 2 variables used (independent variable from the date attribute and the dependent variable from attribute volume of waste per day) from the prediction results. The reason why this experiment is distinguished based on the use of outlier values in the dependent variable is because by differentiating the experiment to determine the characteristics between the independent and dependent variables and the data points that affect the distribution of the relationship between the 2 variables used (independent variable from the date attribute and the dependent variable from attribute volume of waste per day) from the prediction results.

In addition, the prediction results will also be divided into six different months (from January to June), with the predicted output in the form of the volume of waste transported from



January to June 2022. In this section, the prediction process is carried out with two algorithms, i.e.: Linear Regression and Random Forest Regression.

### 3.1 Split Validation Results

This section experiments with datasets from January to June. Each of these datasets will be divided into training and testing data with a total proportion of 80% to 20%. Training data is used to train the machine learning algorithms, and testing data is used after passing through the training process. In addition, experiments were also carried out on Linear Regression to see whether the independent variables had a positive linear relationship or not. This experiment is distinguished based on 2 differences in the dependent variable: without considering outlier and with considering outlier values.

Table 1, Table 2, and Table 3 exhibit the predicted results of waste transportation without outliers resulted in prediction error values per month transportation. The result is that May is the month with a lower error rate than the other five months, but when considering on the algorithm, Random Forest Regression provides much lower error values than Linear Regression with error values of 0.82 and 0.81.

Table 1. Split Validation Results without Outliers (January- February)

Algorithms	Data Train		Data Test		Data Train		Data Test	
	January				February			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear Regression	1.57	2.98	1.59	3.07	0.97	1.16	0.96	1.16
Random Forest Regression	1.56	2.97	1.59	3.08	0.97	1.16	0.96	1.16

Table 2. Split Validation Results without Outliers (March-April)

Algorithms	Data Train		Data Test		Data Train		Data Test	
	March				April			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear Regression	0.92	1.15	0.91	1.12	0.90	1.09	0.91	1.12
Random Forest Regression	0.92	1.15	0.91	1.12	0.90	1.09	0.91	1.12

Table 3. Split Validation Results without Outliers (May-June)

Algorithm	Data Train		Data Test		Data Train		Data Test	
	May				June			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Linear Regression	0.83	0.82	0.83	0.82	1.17	1.74	1.17	1.71
Random Forest Regression	0.82	0.81	0.82	0.81	1.17	1.73	1.16	1.71

Next, experiments using outliers in the dependent variable are carried out. The results obtained in Table 4, Table 5, and Table

6 show that June 2021 has an error rate of metrics MAE and MSE relatively low compared to the other five months. However, when considering on the algorithm, the Linear Regression dominantly provides lower prediction results compared to the Random Forest Regression with an error rate of 1.83 for the MAE metric in the training data and 1.81 for the MAE metric in the testing data. On the other hand, in the MSE metric for the training data, the Random Forest Regression has a low error rate compared to the Linear Regression, which is 10.37, while the prediction rate for the MSE metric in the testing data provides the same result, i.e.: 10.14.

Table 4. Split Validation Results using Outliers (January – February)

Algorithm	Data Train		Data Test		Data Train		Data Test	
	January				February			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	3.20	19.12	3.30	21.72	2.27	25.97	2.26	20.67
RFR	3.20	19.10	3.30	21.73	2.27	25.97	2.26	20.67

Table 5. Split Validation Results using Outliers (March – April)

Algorithm	Data Train		Data Test		Data Train		Data Test	
	March				April			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	2.07	11.73	2.08	12.26	2.14	15.32	2.24	38.04
RFR	3.20	19.10	2.08	12.28	2.14	15.31	2.24	38.03

Table 6. Split Validation Results using Outliers (May – June)

Algorithm	Data Train		Data Test		Data Train		Data Test	
	May				June			
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	2.02	12.13	2.02	11.30	1.83	10.38	1.81	10.14
RFR	2.03	12.08	2.03	11.25	1.84	10.37	1.82	10.14

### 3.2 Validation Results

Having done experimenting with data distribution based on a percentage of 80% training data and 20% testing data, the next step is to validate the prediction results using the k-fold cross-validation method. In addition, the validation process used in this study uses 3 different fold value scenarios, including 5-fold, 10-fold, and 15-fold. One of the advantages of using multiple k-fold value scenarios is to avoid bias and overfitting in the prediction results [22]. This experiment will also be differentiated based on 2 differences in the dependent variable: without using outlier values and using outlier values.

Table 7, Table 8, and Table 9 show the validation results related to waste transportation per month without outlier values in the dependent variable. It can be concluded that the May dataset provides the lowest error values compared to the other five months. Even the error values obtained from the three k-fold values based on MAE and MSE metrics have the same ones obtained in the previous experiment (experiment with data

distribution without outlier values). However, from the algorithm's perspective, Random Forest Regression obtain the lowest error values compared to Linear Regression where Random Forest Regression has error values of 0.82 and 0.81, while Linear Regression has error values of 0.83 and 0.82.

Table 7. Validation Results without Outliers (January–February)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		January				February			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	1.57	2.98	1.59	3.07	0.97	1.16	0.96	1.16
	10	1.57	2.98	1.59	3.07	0.97	1.16	0.96	1.16
	15	1.57	2.98	1.59	3.07	0.97	1.16	0.96	1.16
RFR	5	1.56	2.98	1.58	3.07	0.97	1.16	0.96	1.16
	10	1.56	2.98	1.56	3.07	0.97	1.16	0.96	1.16
	15	1.56	2.98	1.56	3.07	0.97	1.16	0.96	1.16

Table 8. Validation Results without Outliers (March – April)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		March				April			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	0.92	1.15	0.90	1.12	0.90	1.09	0.91	1.12
	10	0.92	1.15	0.90	1.12	0.90	1.09	0.91	1.12
	15	0.92	1.15	0.90	1.12	0.90	1.09	0.91	1.12
RFR	5	0.92	1.15	0.91	1.13	0.90	1.09	0.91	1.13
	10	0.92	1.15	0.91	1.13	0.90	1.09	0.91	1.13
	15	0.92	1.15	0.91	1.13	0.90	1.09	0.91	1.13

Table 9. Validation Results without Outliers (May – June)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		May				June			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	0.83	0.82	0.83	0.82	1.17	1.74	1.16	1.71
	10	0.83	0.82	0.83	0.82	1.17	1.74	1.16	1.71
	15	0.83	0.82	0.83	0.82	1.17	1.74	1.16	1.71
RFR	5	0.82	0.81	0.82	0.81	1.17	1.74	1.16	1.71
	10	0.82	0.81	0.82	0.81	1.17	1.74	1.17	1.71
	15	0.82	0.81	0.82	0.81	1.17	1.73	1.17	1.71

After validating without outlier values, the next step is to validate with outlier values in the dependent variable on the waste volume transportation data. Table 10, Table 11, and Table 12 show that June 2021 has a reasonably low error rate in predicting compared to the other five months, for both, the MAE and MSE metrics. However, when considering on the algorithm, Linear Regression has good validation results compared to the Random Forest Regression algorithm.

In addition, the results of the three scenarios show that the k-fold value in the MAE metric with training data has the same predictive value as the results obtained in the previous experiment (experiment with data distribution using outlier values), i.e.: 1.83 for Linear Regression and 1.84 for Random Forest Regression. Furthermore, the same results were also obtained from the three scenarios of the k-fold value with a linear regression algorithm which produced a predictive value in the MSE metric of 10.14 on the testing data. Furthermore, the MSE metric with the Linear Regression algorithm also shows the same predictive value as the previous experiment (experiment with

data distribution using outlier values), which is 10.38. On the other hand, the Forest Random Regression algorithm shows a different predictive value from the previous experiment, i.e.: 10.39.

The difference in the predicted value is also found in the MSE metric with the testing data because the three scenarios of the k-fold value get a predictive value of 10.15. While the prediction value of 10.14 was obtained in the previous experiment (experiment with data distribution using outlier values). Furthermore, the three scenarios of the k-fold value in the MAE metric with training data in Linear Regression show different predictive values from the previous experiment because the predictive value from the validation results is 1.80, while the predictive value from the previous experiment is 1.81. Meanwhile, the Random Forest Regression algorithm with a 5-fold value provides different results from the other two k-fold scenarios, where the predicted value generated in the 5-fold scenario is 1.85, and the other two scenarios (10-fold and 15-fold) is 1.84.

Table 10. Validation Results using Outliers (January – February)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		January				February			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	3.20	19.12	3.38	21.69	2.27	25.97	2.24	20.69
	10	3.20	19.12	3.38	21.69	2.27	25.97	2.24	20.68
	15	3.20	19.12	3.38	21.69	2.27	25.97	2.24	20.68
RFR	5	3.20	19.13	3.38	21.77	2.27	25.98	2.25	20.79
	10	3.20	19.14	3.20	21.76	2.27	25.99	2.27	20.75
	15	3.20	19.13	3.20	21.76	2.27	25.99	2.27	20.75

Table 11. Validation Results using Outliers (March – April)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		March				April			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	2.07	11.73	2.09	12.26	2.14	15.33	2.30	38.03
	10	2.07	11.74	2.09	12.26	2.14	15.33	2.30	38.05
	15	2.07	11.74	2.09	12.27	2.14	15.33	2.30	38.03
RFR	5	2.07	11.74	2.09	12.31	2.14	15.33	2.29	38.15
	10	2.07	11.74	2.09	12.32	2.14	15.33	2.14	38.15
	15	2.07	11.74	2.09	12.30	2.14	15.33	2.14	38.13

Table 12. Validation Results using Outliers (May – June)

Algorithm	k	Data Training		Data Testing		Data Training		Data Testing	
		May				June			
		MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
LR	5	2.02	12.13	2.01	11.30	1.83	10.38	1.80	10.14
	10	2.02	12.13	2.01	11.30	1.83	10.38	1.80	10.14
	15	2.02	12.13	2.01	11.30	1.83	10.38	1.80	10.14

	5		3		0		8		4
RFR	5	2.03	12.1	2.03	11.3	1.84	10.3	1.85	10.1
			0		1		9		5
	1	2.03	12.1	2.03	11.3	1.84	10.3	1.84	10.1
			0		0		9		5
	1	2.03	12.1	2.03	11.2	1.84	10.3	1.84	10.1
			0		8		9		5

**3.3 Discussion**

This section describes whether the independent and dependent variables have a positive slope or not based on the prediction results shown in the Linear Regression graph using the May 2021 data because the prediction results with the May data have a lower error rate than the other five months without outlier values. In addition to the Linear Regression graphs, this section also discusses the prediction results in line graphs generated by the Random Forest Regression algorithm without outlier values.

Previously, the number of data in the May dataset was 54870 data. However, as many as 6096 data were indicated as outlier values and were omitted from the data, leaving only 48774 data. In addition, the correlation between the date attribute and the volume of waste per day in the May dataset is not good enough, which is only 0.017, indicating that the data used in the experiment is not suitable for implementation using Linear Regression algorithm. Figure 2 shows the correlation value between the date attribute and the waste volume attribute per day.

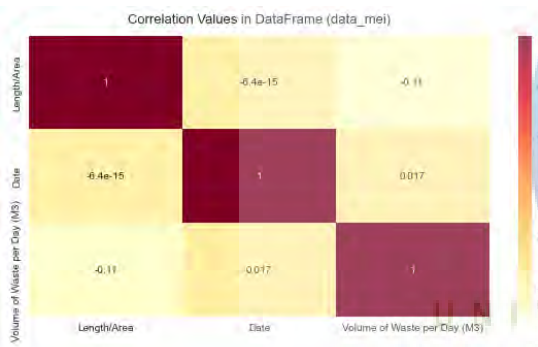


Figure 2. Correlation Values in May Dataset

Therefore, the linear graph of the experimental results without outlier values shows that the independent and dependent variables do not have a positive linear slope but a zero slope. The zero slopes are caused by the coordinates of the dependent variable (y) do not change when the coordinates of the independent variable (x) change.

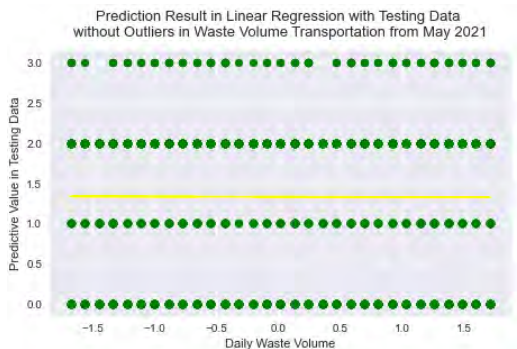


Figure 3. Prediction using Linear Regression for Testing Data without Outliers

Figure 3 shows the prediction results with linear graphs on experiments without outlier values for the May data with a linear

regression algorithm that does not have a positive linear slope but zero slopes (horizontal line). So, the discussion results based on experiments without outlier values using a linear regression algorithm on the waste transportation volume dataset in May still resulted in zero slopes, indicated by a horizontal line in testing data in Figure 3 lies in the distribution of data points and the location of the linear line because, without outlier values, the distribution of data points is spread over several parts horizontally followed by a horizontal line. This means that 6.096 data that have been removed from the dependent variable because they are indicated as outlier values are very influential in placing the data points along with the horizontal linear line.

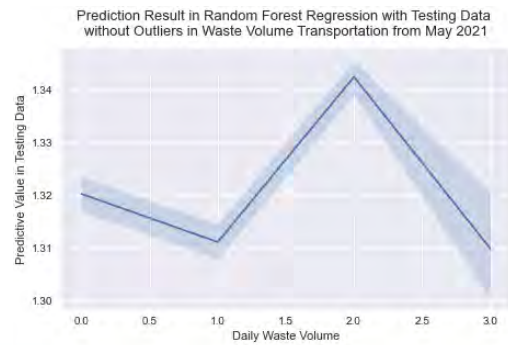


Figure 4. Prediction Result on Random Forest for Testing Data with Outliers

The subsequent discussion will explain the prediction results from Random Forest Regression without outlier values in the dependent variable based on May data. Figure 4 shows a downward line graph so that it can be concluded that the MAE and MSE metrics error rate in predicting the volume of waste transport data for May is very small. In addition, Figure 4 also shows that the tonnage of waste to be transported from the waters of the DKI Jakarta river in May 2022 will be much less than in 2021.

**4 Conclusion**

Experimental results showed that Random Forest Regression performed better compare to the Linear Regression, due to Random Forest Regression produces the lowest error value compared to Linear Regression. The error values for Random Forest Regression are 0.82 and 0.81, while the error values for Linear Regression are 0.83 and 0.82.

The problem of waste in the waters of the DKI Jakarta Province is still the biggest problem that must be faced together. Based on the results of research that has been carried out regarding the transportation of waste volumes from several rivers in DKI Jakarta within six months using the Linear Regression and Random Forest Regression algorithms, it is known that the Random Forest Regression has a reasonably good performance compared to Linear Regression due

In addition, among the six datasets used in this study, the May dataset provided the best performance. The May dataset produces the lowest error value for both Linear Regression and Random Forest Regression. This research provided useful information regarding the transportation of waste volumes and the use of Linear Regression and Random Forest Regression algorithms as a reference for further research on the transportation of waste volumes in the rivers of DKI Jakarta Province in the future.

**ACKNOWLEDGMENTS**

Authors thanks to Research & Technology Directorate of Universitas Mercu Buana for partially support on this research.

## REFERENCES

- [1] A. Brotosusilo, S. H. Nabila, H. A. Negoro, and D. Utari, "The level of individual participation of community in implementing effective solid waste management policies," *Global Journal of Environmental Science and Management*, vol. 6, no. 3, pp. 341–354, 2020, doi: 10.22034/gjesm.2020.03.05.
- [2] Y. A. Hidayat, S. Kiranamahsa, and M. A. Zamal, "A study of plastic waste management effectiveness in Indonesia industries," *AIMS Energy*, vol. 7, no. 3, pp. 350–370, 2019, doi: 10.3934/ENERGY.2019.3.350.
- [3] Y. Hayati, L. Adrianto, M. Krisanti, W. S. Pranowo, and F. Kurmiawan, "Magnitudes and tourist perception of marine debris on small tourism island: Assessment of Tidung Island, Jakarta, Indonesia," *Marine Pollution Bulletin*, vol. 158, no. June, p. 111393, 2020, doi: 10.1016/j.marpolbul.2020.111393.
- [4] P. Lestari and Y. Trihadiningrum, "The impact of improper solid waste management to plastic pollution in Indonesian coast and marine environment," *Marine Pollution Bulletin*, vol. 149, no. August, p. 110505, 2019, doi: 10.1016/j.marpolbul.2019.110505.
- [5] "SIPSN - Sistem Informasi Pengelolaan Sampah Nasional." <https://sipsn.menlhk.go.id/sipsn/public/data/komposisi> (accessed Jun. 06, 2022).
- [6] F. Hermawan, "Optimization Of Transportation of Municipal Solid Waste from Resource to Intermediate Treatment Facility with Nearest Neighbour Method (Study on six Sub District in DKI Jakarta Province)," *JOURNAL OF ENVIRONMENTAL SCIENCE AND SUSTAINABLE DEVELOPMENT*, vol. 1, no. 1, Dec. 2018, doi: 10.7454/jessd.v1i1.21.
- [7] X. C. Nguyen *et al.*, "Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam," *Resources, Conservation and Recycling*, vol. 167, no. July 2020, p. 105381, 2021, doi: 10.1016/j.resconrec.2020.105381.
- [8] M. Kannangara, R. Dua, L. Ahmadi, and F. Bensebaa, "Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches," *Waste Management*, vol. 74, pp. 3–15, 2018, doi: 10.1016/j.wasman.2017.11.057.
- [9] "Data Volume Pengangkutan Sampah di Kali Sungai Situ Waduk Tahun 2021 - Open Data Jakarta." <https://data.jakarta.go.id/dataset/data-volume-pengangkutan-sampah-di-kali-sungai-situ-waduk-tahun-2021> (accessed Jun. 08, 2022).
- [10] V. N. G. Raju, K. P. Lakshmi, V. M. Jain, A. Kalidindi, and V. Padma, "Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification," *Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020*, no. Icssit, pp. 729–735, 2020, doi: 10.1109/ICSSIT48917.2020.9214160.
- [11] S. Ray, "A Quick Review of Machine Learning Algorithms," *Proceedings of the International Conference on Machine Learning, Big Data, Cloud and Parallel Computing: Trends, Perspectives and Prospects, COMITCon 2019*, pp. 35–39, 2019, doi: 10.1109/COMITCon.2019.8862451.
- [12] W. M. Baihaqi, M. Dianingrum, K. A. N. Ramadhan, and T. Hariguna, "Linear regression method to model and forecast the number of patient visits in the hospital," *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, pp. 247–252, 2018, doi: 10.1109/ICITISEE.2018.8720979.
- [13] P. Dong, H. Peng, X. Cheng, Y. Xing, X. Zhou, and D. Huang, "A Random Forest Regression Model for Predicting Residual Stresses and Cutting Forces Introduced by Turning IN718 Alloy," *2019 IEEE International Conference on Computation, Communication and Engineering, ICCCE 2019*, pp. 5–8, 2019, doi: 10.1109/ICCCE48422.2019.9010767.
- [14] I. Ouedraogo, P. Defourny, and M. Vanclooster, "Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale," *Hydrogeology Journal*, vol. 27, no. 3, pp. 1081–1098, 2019, doi: 10.1007/s10040-018-1900-5.
- [15] H. Liu, Q. Li, D. Yu, and Y. Gu, "Air quality index and air pollutant concentration prediction based on machine learning algorithms," *Applied Sciences (Switzerland)*, vol. 9, no. 19, 2019, doi: 10.3390/app9194069.
- [16] Q. Quan, Z. Hao, H. Xifeng, and L. Jingchun, "Research on water temperature prediction based on improved support vector regression," *Neural Computing and Applications*, vol. 4, 2020, doi: 10.1007/s00521-020-04836-4.
- [17] X. Xie *et al.*, "Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land," *Ecological Indicators*, vol. 120, no. June 2020, p. 106925, 2021, doi: 10.1016/j.ecolind.2020.106925.
- [18] U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, and R. Irfan, "Efficient Water Quality Prediction Using Supervised Machine Learning," *Water (Basel)*, vol. 11, no. 1, pp. 1–14, 2019.
- [19] B. G. Marcot and A. M. Hanea, "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis?," *Computational Statistics*, vol. 36, no. 3, pp. 2009–2031, 2021, doi: 10.1007/s00180-020-00999-9.
- [20] A. Bayhaqy, S. Sfenrianto, K. Nainggolan, and E. R. Kaburuan, "Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naive Bayes," *2018 International Conference on Orange Technologies, ICOT 2018*, no. November 2019, 2018, doi: 10.1109/ICOT.2018.8705796.
- [21] M. Hasnain, M. F. Pasha, I. Ghani, M. Imran, M. Y. Alzahrani, and R. Budiarto, "Evaluating Trust Prediction and Confusion Matrix Measures for Web Services Ranking," *IEEE Access*, vol. 8, pp. 90847–90861, 2020, doi: 10.1109/ACCESS.2020.2994222.



## KERTAS KERJA

### Ringkasan

Kertas kerja ini merupakan tempat pengumpulan bahan material kelengkapan pembuatan sebuah artikel jurnal dengan judul Komparasi Metode Regresi Linear dan Regresi Random Forest Terhadap Volume Pengangkutan Sampah. Kertas kerja ini berisikan semua material-material hasil penelitian Tugas Akhir yang tidak dimuat atau disertakan pada artikel jurnal. Pada kertas kerja ini terdapat material seperti *literature review*, Analisis Perancangan, *Source Code*, *Dataset*, Tahapan Eksperimen serta Hasil Eksperimen Secara Keseluruhan.

Hasil penelitian ini merupakan hasil prediksi berupa analisis terhadap volume pengangkutan sampah dengan menggunakan algoritma Regresi Linear dan Regresi *Random Forest*. Diharapkan hasil penelitian ini dapat menjadi bahan evaluasi Pemerintah Provinsi DKI Jakarta dan Dinas Lingkungan Hidup dalam mengatasi permasalahan volume pengangkutan sampah di perairan sungai DKI Jakarta.

