*mathematics*

MDPI

*Article*

# Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey

**Babek Erdebilli [1,*] and Burcu Devrim-İçtenbaş [2,†]**

[1] Department of Industrial Engineering, Ankara Yıldırım Beyazıt University, Ankara 06010, Turkey
[2] Department of Industrial Engineering, Faculty of Engineering, Ankara Science University, Ankara 06200, Turkey; burcu.ictenbas@ankarabilim.edu.tr
* Correspondence: berdebilli@ybu.edu.tr
† Former name: B. D. Rouyendegh.

**Abstract:** Predicting medical waste (MW) properly is vital for an effective waste management system (WMS), but it is difficult because of inadequate data and various factors that impact MW. This study's primary objective was to develop an ensemble voting regression algorithm based on machine learning (ML) algorithms such as random forests (RFs), gradient boosting machines (GBMs), and adaptive boosting (AdaBoost) to predict the MW for Istanbul, the largest city in Turkey. This was the first study to use ML algorithms to predict MW, to our knowledge. First, three ML algorithms were developed based on official data. To compare their performances, performance measures such as mean absolute deviation (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and coefficient of determination (R-squared) were calculated. Among the standalone ML models, RF achieved the best performance. Then, these base models were used to construct the proposed ensemble voting regression (VR) model utilizing weighted averages according to the base models' performances. The proposed model outperformed three baseline models, with the lowest RMSE (843.70). This study gives an effective tool to practitioners and decision-makers for planning and constructing medical waste management systems by predicting the MW quantity.

**Keywords:** adaptive boosting; ensemble machine learning; gradient boosting machine; sustainability; random forests; waste prediction

**MSC:** 90-11



**Citation:** Erdebilli, B.; Devrim-İçtenbaş, B. Ensemble Voting Regression Based on Machine Learning for Predicting Medical Waste: A Case from Turkey. *Mathematics* **2022**, *10*, 2466. https://doi.org/10.3390/math10142466

Academic Editors: Victor Leiva and Ioannis G. Tsoulos

Received: 9 May 2022
Accepted: 11 July 2022
Published: 15 July 2022



## 1. Introduction

There is increasing urban growth due to industrial and economic progression, and this has led to increasing trends in population and population density. As a result, there has been an increase in the amount of medical waste (MW) generated as the number of hospitals, veterinarians, clinics, and other health institutions has increased [1,2]. According to the United States Environmental Protection Agency (USEPA) and the World Health Organization (WHO), MW is a hazardous waste category because it contains potentially deadly pathogens, chemical anticancer agents, and hazardous and radioactive wastes [2,3]. Furthermore, cutting and sharp materials can cause a slew of issues for those who work with them [3]. Up to 25% of medical waste (by weight) has been deemed infectious [4]. Because of the possible public health concerns and environmental pollution, the collection and disposal of medical waste is a serious issue, particularly in poor nations [5]. Now, the coronavirus disease 2019 (COVID-19) pandemic has made the handling of medical waste and the performance of waste management systems more important than ever [6]. There is a need to predict the amount of MW accurately to decide the proper disposal options and plan the operational capacities of recycling, storage, disposal, and transportation [1,7–9].

In prior research, several different models, such as statistical models, data mining, sample surveys, time series models, artificial intelligence, and ML algorithms, have been

applied in order to make a prediction regarding the total quantity of MW that would be present in the future. The details of related studies are given in Table 1 regarding data sources, methods, and performance measures. For instance, Tesfahun et al. [10] utilized multiple linear regression (MLR) to estimate the hospital healthcare waste generation rate in Ethiopia. They collected data from 24 hospitals, and they constructed predictive models for waste generation with respect to different types of hospitals. The multiple linear regression equation developed by Çetinkaya et al. [11] to estimate medical waste in Aksaray City, Turkey used age classification and the gross national product in United States dollars (USD) as input parameters. They concluded that these input parameters were significant for medical waste estimation. Idowu et al. [12] applied MLR to estimate the rate of waste generated from two healthcare facilities in Logos, Nigeria. Their study indicated that the number of patients and bed capacity were significant parameters for MW quantity. Al-Khatib et al. [13] developed three equations to predict general hospital waste, hazardous solid hospital waste, and total solid hospital waste. They achieved high coefficient of determination (R-squared) scores for all three equations. From other studies that used MLR models, [13] presented and proved that assumptions of MLR models which must be checked. Bdour et al. [14] constructed an MLR model to generate waste quantities in Jordan. The study indicated that the number of beds and the number of patients were significant parameters and that private hospitals generated less waste than public hospitals. Sabour et al. [15] developed a mathematical model to generate infectious hospital waste quantities in Iran. Korkut [16] constructed a linear model to estimate the amount of waste from hospitals in İstanbul, utilizing population as the only input parameter. Ceylan et al. [1] compared linear regression (LR), support vector regression (SVR), grey modeling (GM), and autoregressive integrated moving average (ARIMA) models to estimate medical waste generation in İstanbul. The ARIMA (0,1,2) model was selected as the most fitting model for predicting MW generation. However, this study did not find significant parameters that affected MW generation. Chauhan and Singh [17] employed an ARIMA model to forecast healthcare waste generation. Golbaz et al. [3] applied many neuron- and kernel-based machine learning methods as well as MLR for predicting hospital solid waste generation. They concluded that from the MLR analysis, hospital ownership type and the number of hospital staff were significant parameters that affected MW generation, but they indicated that conventional MLR methods required more complex modeling when the number of input parameters increased. Artificial neural network (ANN)-based models can provide predictions with lower errors than those provided by MLR. Karpušenkaitė et al. [18] tried to forecast medical waste by using scarce official data in ANN, MLR, partial least squares (PLS), and support vector machine (SVM) methods and by using short and extra-short datasets. The ANN method failed with this small dataset. Jahandideh et al. [2] applied MLR and ANN to estimate medical waste amounts as well as waste types in Iran using hospital type, hospital capacity, and bed occupancy as parameters. The authors concluded that ANN outperformed MLR in dealing with the nonlinearity between dependent and independent variables. Karpušenkaitė et al. [19] developed a hybrid model using the coefficients generated by moving average (MA) and Holt's method in the regression equation. They proved that the waste generation rate was positively correlated with the numbers of inpatients and outpatients. Thakur and Ramesh [20] compared MLR and ANN to estimate the MW generation rates in India. The ANN model outperformed the MLR model.

To sum up, MLR models achieved good model performances (R-squared > 0.80) and provided the most significant input parameters. However, conventional MLR methods have deficiencies. For example, they are not able to predict the MW amount when using large numbers of input variables that need complex modeling, and they use many assumptions that are not easy to meet in real life. ML and DL approaches such as ANN, many neuron- and kernel-based ML algorithms, and SVM outperformed MLR on many statistical measures [2,3,18,20] because of their capacity to solve the nonlinear connection between input and output variables. Despite this, these approaches lacked substantial input factors,

which hindered their ability to estimate the rate of medical waste creation, and deep learning algorithms such as ANN did not perform well with limited data. Using time-based data on MW amounts, some researchers utilized time series modeled as different ARIMA models [1,21]. Time-series analysis has limitations: it requires large datasets to detect seasonality, several factors influence MW generation, and data contain outliers and missing values. Thus, predicting MW generation is a regression problem, not a time-series one [22]. According to a number of studies, classical and ML-based algorithms for time-series data competed, just as for prediction issues [21,22].

**Table 1.** Literature overview.

| Reference | Data Source | Methods | Performance Measures |
|---|---|---|---|
| [1] | Official data | LR, SVR, GM (1,1), and ARIMA | R-squared, RMSE, MAD, MAPE |
| [2] | Surveys | MLR, ANN | MAE, RMSE, R-squared |
| [3] | Survey | MLR and neuron- and kernel-based methods | R-squared, MSE |
| [10] | Survey | MLR | R-squared |
| [11] | Survey | MLR | MBE, MAPE |
| [12] | Survey | MLR | R-squared |
| [13] | Survey | MLR | R-squared |
| [14] | Survey | MLR | R-squared |
| [15] | Survey | MLR | R-squared |
| [16] | Official data | LR | R-squared |
| [17] | Official data | ARIMA | MAPE, MAE |
| [18] | Official data | ANN, MLR, PLS, and SVM | R-squared, RMSE, MAE, MAPE |
| [19] | Official data | MA, Holt's | MAPE |
| [20] | Survey | MLR, ANN | MAE, RMSE, R-squared |

**MLR**: multiple linear regression; **LR**: linear regression; **SVR**: support vector regression; **ARIMA**: autoregressive integrated moving average; **ANN**: artificial neural network; **PLS**: partial least squares; **SVM**: support vector machine; **MA**: moving average; **Holt's**: Holt's exponential smoothing; **R-squared**: coefficient of determination; **RMSE**: root mean square error; **MAE**: mean absolute error; **MAD**: mean absolute deviation; **MAPE**: mean absolute percentage error; **MBE**: mean bias error.

ML algorithms can be used to estimate MW amounts in order to discover trends, patterns, and changes with greater accuracy than traditional regression analysis, as previous research has demonstrated [9]. Furthermore, most studies estimating MW quantity have not included the most substantial input factors, which may be critical information for an effective medical waste management system. Most research for estimating MW quantity has relied on surveys and questionnaires because of the absence of a historical MW database, particularly in developing nations, although this may result in inaccurate projections due to the lack of actual data [3,9,14,18,23,24]. At this point, MW prediction has problems such as limited data and many parameters affecting the amount of MW, so there is a need for more powerful algorithms, such as ensemble methods based on ML algorithms, to handle these problems. However, to the best of our knowledge, no investigations have been undertaken utilizing the random forest (RF), gradient boosting machine (GBM), and adaptive boosting (AdaBoost) algorithms to predict MW amounts, nor any utilizing ensemble methods based on ML algorithms, which are considered to represent a better approach than single algorithms [25–30]. Instead of using single machine learning algorithms, this study proposed ensemble voting regression (VR) algorithms based on machine learning to obtain better predictions of MW generation in İstanbul, the largest city in Turkey, utilizing official data. MW has a significant impact on both the environment and public health, since Istanbul has a population of more than 15 million people and is home to 17% of the hospitals, 20% of the bed capacity, and 54% percent of the private hospitals in Turkey [1,8]. The RF, GBM, and AdaBoost ML algorithms were chosen as base learners or ensemble members to form the ensemble VR to estimate the MW amount. These ML approaches were chosen for their outstanding performance in investigating tiny datasets and avoiding overfitting [31]. Furthermore, these methods do not entail assumptions that

are difficult to meet in real life, as conventional statistical methods and MLR methods do. The ensemble learning that was used in this study boosted performance, combining three base learners and utilizing voting regression to outperform traditional methods [32]. First, to enable more precise calculation, 17 years' worth of official data on MW amounts was downloaded from the Open Data Portal of the Istanbul Metropolitan Municipality rather than data collected through surveys or questionnaires [1]. Then, ML techniques were used to forecast the MW quantity. Finally, the most important input factors that had an impact on the quantity of MW were identified by comparing their performances using performance metrics such as RMSE, MAE, MAPE, and R-squared. The ensemble VR's weighted average was obtained by utilizing a ranking technique to merge these three fundamental ML algorithms. The proposed model's performance was compared with that of the three individual ML algorithms with the same performance measures.

The proposed ensemble approach contributes to the literature in two ways. First, utilizing real data with the RF, GBM and AdaBoost machine learning algorithms as a result of one of the implementation steps of this proposed method enabled determining the most important significant parameters and the best model among the single ML algorithms. Second, the proposed model had higher predictive accuracy than standalone ML models.

The following structure makes up the layout of the paper: Section 2 provides an explanation of the approach's specifics. The steps of the suggested model are detailed in Section 3. In Section 4, the outcomes of the suggested model are covered. Finally, Section 5 presents the study's results.

## 2. Methods

### *2.1. Machine Learning Algorithms*

ML is a subset of artificial intelligence (AI), which is the intelligence that allows computers to learn from data by combining computer science and statistical analysis methods to create algorithms that are "statistically proficient" [33]. There are two kinds of these algorithms: supervised and unsupervised. There are two categories of algorithms for supervised learning, classification and regression, based on how they determine the connections between prospective independent and dependent variables.

Ensemble techniques are ML algorithms that develop and combine many base models to improve the performance of regression problems, classification issues, and feature selection [32]. According to the learner generation process, ensemble techniques are of two types: parallel, represented by bagging (Breiman, 1996) [34], and sequential, represented by boosting [32]. In boosting, several training sets with sample size n are selected from the data using a sampling technique known as bootstrap sampling to ensure the independence of distinct sampling training sets. The ultimate model is determined by combining the predictions of all models [34]. Multiple base learners are independent of one another, since they are constructed simultaneously. The sequential ensemble type is characterized by the sequential construction of numerous learners, which improves the performance of the final model, since subsequent learners may avoid the faults of their predecessors.

#### 2.1.1. Random Forest (RF)

RF is a ML approach that combines numerous decision trees to solve classification and regression issues. Multiple decision trees are combined using bootstrap sampling, and their majority classification vote and average are allocated to regression [35,36]. RF is a robust technique for imbalanced, missing, and multicollinear data [32,35]. The analysis has two stages:

Stage 1: To start building a forest, a sample of the initial dataset (training data) is chosen and then replaced to make subdatasets. Then, regression trees are made according to these smaller sets of data. In the training stage, it is possible to change the number of variables (mtry) and number of trees (ntree).

Stage 2: After the model has been trained, a prediction can be made. First, all of the input variables for each regression tree are added up. Second, the final result is judged by taking the average of the predictions from each tree [36].

#### 2.1.2. Adaptive Boosting (AdaBoost)

AdaBoost is a popular algorithm that was invented by Freund and Schapire [37], and an illustration of the approach can be found in Algorithm 1 [32]. The exponential sort of loss function is applied, and weights and classifiers are constructed through the use of forward stagewise additive modeling.

**Algorithm 1** The AdaBoost algorithm [32]

Input: Dataset $D = \{(x_1, y_1),\ (x_2, y_2), \ldots, (x_m,\ y_m)\}$;
The weight of each training set sample produces establishes a weight vector $Z_t$;
The base learning algorithm is denoted as $L$;
The number of learning rounds denoted as $T$.
Process:
$D_1(i) = 1/m$ % The weight distribution is initialized
For $t = 1, 2, \ldots, T$ :
$h_t = L(D, D_t)$; % Train a base learner $h_t$ from $D$ utilizing distribution $D_t$
$\varepsilon_t = P_{r_{i\sim D_i}}[h_t(x_i \nmid y_i)]$; % The error of $h_t$ is measured
$\propto_t = \frac{1}{2} ln \frac{1-\varepsilon_t}{\varepsilon_t}$;
$D_{t+1}(i) = \frac{D_t(i)}{sum(Z_t)} \times \begin{cases} \exp(-\alpha_t) & if\ h_t(x_i) = y_i \\ \exp(\alpha_t) & if\ h_t(x_i) \neq y_i \end{cases}$ % Revise the distribution, where $Z_t$ denotes the normalization factor that enables $D_{t+1}$ to be a distribution.
end.
Output: $F_{(x)} = \text{sign} \sum_{t=1}^{T} a_t h_t(x)$

#### 2.1.3. Gradient Boosting Machine (GBM)

GBM is a boosting technique that was proposed by Friedman in 2001 [38]. It is also known as gradient boosted regression trees (GBRT), multiple additive regression trees (MART), and [36]. It comprises decision trees that are trained sequentially, modifying the negative gradients with each iteration to learn from the decision trees [37].

A dataset is represented as $\{x_i, y_i\}$, $i$ = 1,2 … $N$, where $x_i$ represents a set of features and $y_i$ represents the label. $\Psi$ ($y$, $F(x)$) is the loss function. The GBM algorithm has five steps, as follows [32]:

Step 1: $\beta$ denoted as an initial constant value, and it is obtained as:

$$F_0(\mathrm{x}) = \text{argmin}\ _{\beta} \sum_{i=1}^{N} \Psi(y_i, \beta). \tag{1}$$

Step 2: For the number of iterations $k$ = 1,2 … $K$, the gradient loss function is:

$$y_i^* = \frac{\partial \Psi(y_i, F(x_i))}{\partial \Psi F(x_i)}_{F(x)=F_{k-1}(x)}, \ i = \{1,\ 2 \ldots N\}. \tag{2}$$

Step 3: The parameter $\theta_k$ is computed by employing the method of least squares, and the initial model $h(x_i; \theta_k)$ is constructed by fitting sample data in the following way:

$$\theta_k = \text{argmin}\ _{\theta,\beta} \sum_{i=1}^{N} [y_i^* - \beta h(x_i; \theta)]^2 \tag{3}$$

Step 4: The new model weight is determined by minimizing the loss function:

$$\gamma_k = \text{argmin}_{\gamma} \sum_{i=1}^{N} \psi(y_i, F_{k-1}(x) + \gamma h(x_i; \theta_k)) \tag{4}$$

Step 5: The model is optimized as:

$$F_k(x) = F_{k-1}(x) + \gamma_k h(x_i; \theta_k). \quad (5)$$

This loop executes until a predetermined number of iterations is reached or convergence conditions are satisfied.

### *2.2. Voting Regressor (VR)*

A voting ensemble is a machine learning ensemble methodology that uses many methods in lieu of a single model to increase the system's performance. This approach can be applied to both classification and regression issues by combining the results of numerous methods. For regression issues, the ensembles for which are referred to as voting regressors (VRs), the estimators of all models are averaged to get a final estimate [25]. There are two approaches to awarding votes: average voting (AV) and weighted voting (WV). In the case of AV, the weights are equivalent and equal 1. A disadvantage of AV is that all of the models in ensemble are accepted as equally effective; however, this situation is very unlikely, especially if different machine learning algorithms are used. WV specifies a weight coefficient to each ensemble member. The weight can be a floating-point number between 0 and 1, in which case the sum is equal to 1, or an integer starting at 1 denoting the number of votes given to the corresponding ensemble member [39].

## 3. Proposed Model

### *3.1. General Context*

An ensemble voting regression that utilized RF, GBM and AdaBoost to estimate MW amounts was developed in this work and is presented in Figure 1. The superiority of the proposed model was due to combining ML algorithms to provide the capability to effectively predict MW amounts. The proposed system was also compared with standalone ML models and provided significant parameters that affected the MW amount based on these algorithms. All analyses were conducted with Python 3.8.6 and PyCharm on a desktop computer with the following specifications: Intel(R) Core (TM) i7-10750H CPU @ 2.60 GHz, 0.59 GHz processor, 32.0 GB RAM, 64-bit operating system, x64-based processor. Based on the scikit-learn library, random.seed(), a Python function, was used to ensure the reproducibility of the dividing process. An ensemble voting regression has two phases, namely preparing the data and building the model.

### *3.2. Phase 1: Preparing the Data*

Step 1: Official data was obtained. Istaç Company is a subsidiary of İstanbul Metropolitan Municipality, and it is responsible for regularly collecting MW from around İstanbul [1,16]. Information regarding MW in relation to waste amount by year, waste type, and district was retrieved from the Open Data Portal Istanbul Metropolitan Municipality Department [40], and the crude birth rate (CBR), gross domestic product (GDP), number of hospitals (NH), and number of beds available at the hospitals (NB), were extracted from the Population Information Table [41] by the Turkish Statistics Institute for the years 2004–2020. The two tables were merged in an Excel worksheet by year. The dataset comprised five variables: MW, NH, NB, CBR and GDP.

Step 2: The variables that could potentially affect the amount of MW were selected as input variables. The variables used in the model were analyzed using descriptive statistics and graphical methods for describing the dataset. The descriptive statistics of the variables, including mean, standard deviation, and min and max values, were calculated. Outliers are exceptional observations that occur between normal observations and can lead to erroneous interpretations; hence, boxplots were used to display the distribution of variables and identify outliers [42]. Furthermore, pair plots were constructed to describe the dataset before the modeling phase, since they show both the relationship between two variables and the distribution of the single sample.

Step 3: The preprocessing of data was performed. This is recognized as a crucial stage in ML modeling, because real-world data contain features such as inconsistency, inaccuracy, and incompleteness, which lead to the production of erroneous findings [43]. Therefore, it is essential that data be preprocessed prior to modeling. Data preprocessing consisted of cleansing, instance selection, normalization, transformation, feature extraction, and feature selection [9]. The winsorization technique is used to handle outliers by reducing extreme outliers to a specified percentile of the data [44]. In this study, this method was used to replace outliers with the 25th and 75th percentiles of data for the corresponding variables. In order to handle missing values, several methods were used, namely changing with mean/median/mode, omitting rows, changing values to match earlier or later values, and using regression methods to make the right decision [44].

Step 4: For training and evaluating samples, datasets were separated into training (80% of samples) and testing (20% of samples) sets.

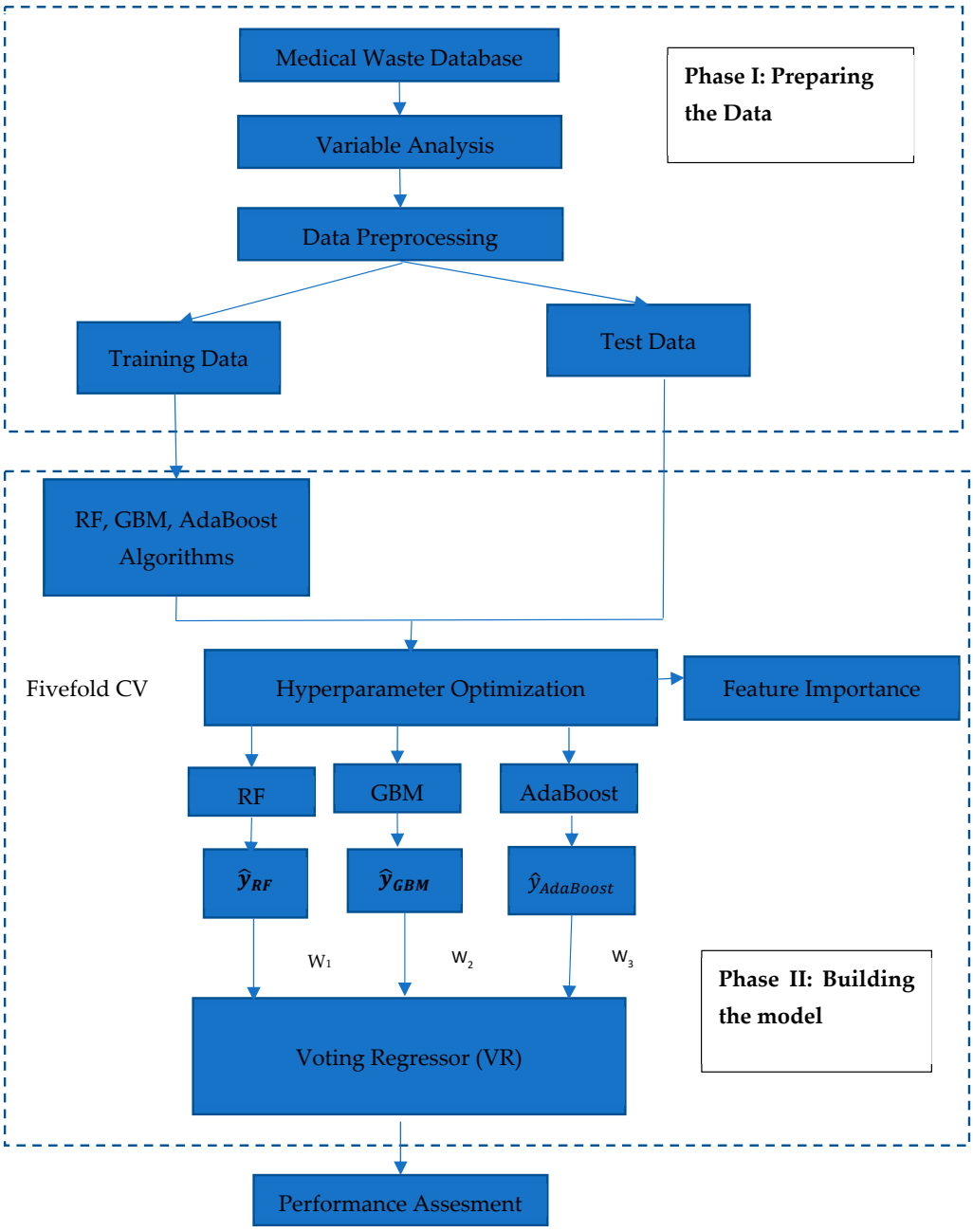


**Figure 1.** The process of the proposed model.

### *3.3. Phase 2: Building the Model*

Step 1: The single machine learning algorithms RF, GBM, and AdaBoost were fitted using training data. The parameters (hyperparameters) of these algorithms that were used in this study were set by default in the training phase [45]. The selected parameters' default values were:

- For RF, the number of trees in the forest (n_estimators) was 100, the minimum number of samples demanded to split an internal node (min_samples_split) was 2, the maximum depth of the tree (max_depth) was None, and the maximum features of each tree (max_features) was 1.
- For GBM, the shrinkage coefficient of each tree (learning_rate) was 0.1, the maximum depth of the tree (max_depth) was 3, the number of trees (n_estimators) was 100, the subsample ratio of training samples (subsample) was 1.0.
- For the AdaBoost model, the maximum number of estimators (n_estimators) was 50, learning rate (learning_rate) was 1.0.

Step 2: After computing, from the training and test data, performance measures based on parameter default values, the best combinations of hyperparameters can be found by the grid search, randomized search, and Bayesian optimization methods. The randomized search and Bayesian optimization methods are better when the number of combinations is large and the number of iterations is small. Grid search was used in this study to determine the best set of hyperparameters because there were not many combinations [46]. The K-fold CV approach is widely used in finding the best hyperparameter combination [31]. In this study, the fivefold CV approach was employed, in which the training set was separated into five equal subsamples, one of which was used as the validation set and the remaining four of which were utilized as the training subset. Five iterations were executed, i.e., until each subsample was used as a validation set. The five validation sets' average was computed, and the results were used to determine the ideal hyperparameter settings.

Step 3: The performances of the models for each method were analyzed with the use of test data, and the relevance of the characteristics was figured out. On the basis of the test data, the outcomes of the predictions made by the RF, GBM, and AdaBoost algorithms were computed.

Four metrics, MAE, RMSE, MAPE, and R-squared, as shown in Equations (6)–(9) [9,18], were used in order to compare the performances of the single ML models:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \tag{6}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - x_i)^2}{n}} \tag{7}$$

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - x_i)^2}{\sum_{i=1}^{n} (y_i - \overline{x}_i)^2} \tag{8}$$

$$\mathrm{MAPE} = \frac{1}{n} \frac{\sum_{i=1}^{n} |y_i - x_i|}{y_i} \times 100 \tag{9}$$

where $x_i$ represents the predicted value for the $i$th observation, $y_i$ represents the actual value for the $i$th observation, $\overline{x}_i$ represents the average of predicted values, and $n$ represents the number of observations. Models with higher R-squared values were more successful than those with lower R-squared values.

Step 4: The three algorithms were combined to construct an ensemble voting regressor (VR) using weighted averages based on the standalone ML performance. Then, the trained ensemble VR was fitted. $\hat{y}_{RF}$, $\hat{y}_{GBM}$, and $\hat{y}_{\mathrm{AdaBoost}}$ denote the predictions of each single algorithm in Figure 1. In order to adjust weights, a ranking method, that is, a form of weighted voting, was used instead of average voting. The method uses a ranking to reveal

the number of votes that every ensemble has in the weighted average. For instance, the best model has three votes, the second best, two votes, and the worst model, one vote if there are three ensemble learners. $w_1$, $w_2$ and $w_3$ represent the votes based on performance. Last, the performance of the proposed model in predicting MW amounts was obtained and compared with that of the standalone ML algorithms.

## 4. Results

### *4.1. Data Acquisition*

As a result of combining two tables [40,41], the dataset comprised 17 rows and 5 columns. The MW quantity data for the years between 2004 and 2020 are shown in Figure 2, which depicts a progressive increase in the amount of medical waste.

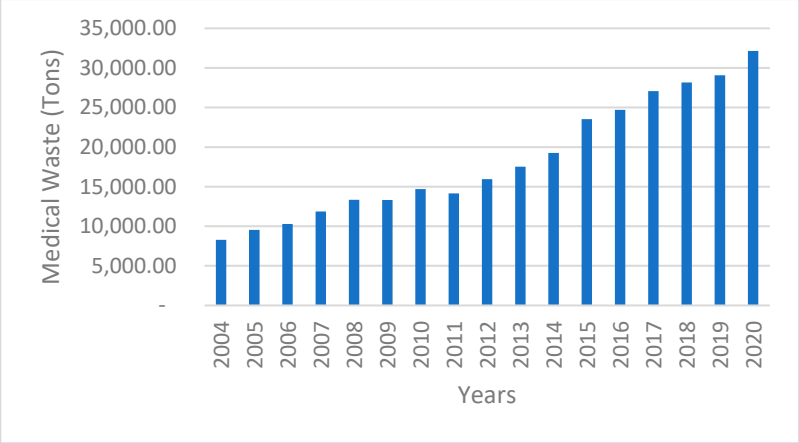


**Figure 2.** Annual MW amount of İstanbul [40].

### *4.2. Variable Analysis*

This study's input variables that might impact MW were chosen based on literature. In earlier research, MW quantity relied on number of patients, medical specialty, number of beds, and department size [2,47]. Socioeconomic factors such as CBR and GDP were selected as input variables in addition to NH and NB [25]. MW was a dependent variable, while the other variables were independent variables. Table 2 provides descriptive data for each variable.

**Table 2.** Statistics that are descriptive of both the input and the output variables.

| Category | Description | Mean | Standard Deviation | Min | Max |
|---|---|---|---|---|---|
| MW | Amount of the medical waste per year | 18,407.8 | 7579.8 | 8279.3 | 32,143.8 |
| NH | Total hospital per year | 221 | 15.5 | 198 | 238 |
| NB | Total beds per year | 33,522.3 | 3441.2 | 28,958 | 40,697 |
| CBR | The ratio of the number of live births during a year to the average population in that year, expressed per 1000 persons | 15.7 | 1.4 | 12.3 | 17 |
| GDP | Total monetary value of all final goods and services produced (and sold on the market) within a country during a year. | 39,334.1 | 22,085.7 | 14,795 | 86,798 |

For detecting outliers, the boxplots of the dependent and independent variables are given in Figure 3 [42]. CBR and NB had clear outlier observations. The winsorization method was used to replace outliers with the 25th and 75th percentiles of data for these variables.

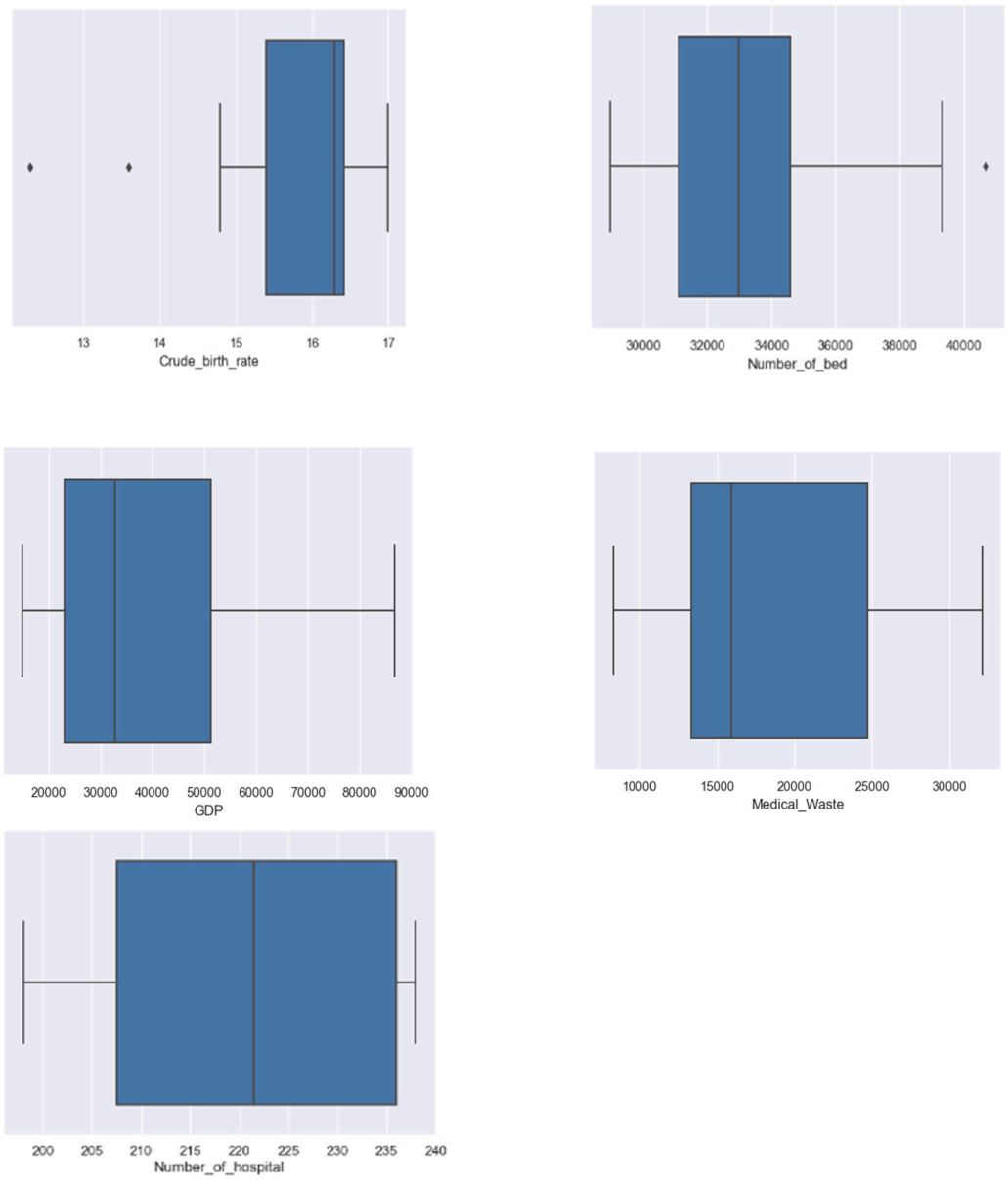


**Figure 3.** Boxplots representing the variables of interest, inputs, and outputs used in the development of the models.

The pair plots for variables are given in Figure 4. From this graph, it is obvious that MW had positive relationships with NH, NB, and GDP and a negative relationship CBR. There was a positive correlation among the predictor variables NH, NB, and GDP, so multicollinearity was not violated, which is one of the assumptions of MLR that must meet. From the histograms, it was observed that all of the variables had skewed distributions, with CBR having an even more skewed distribution. The fact that the ML algorithms used in this study did not require any assumptions could provide great convenience to users.

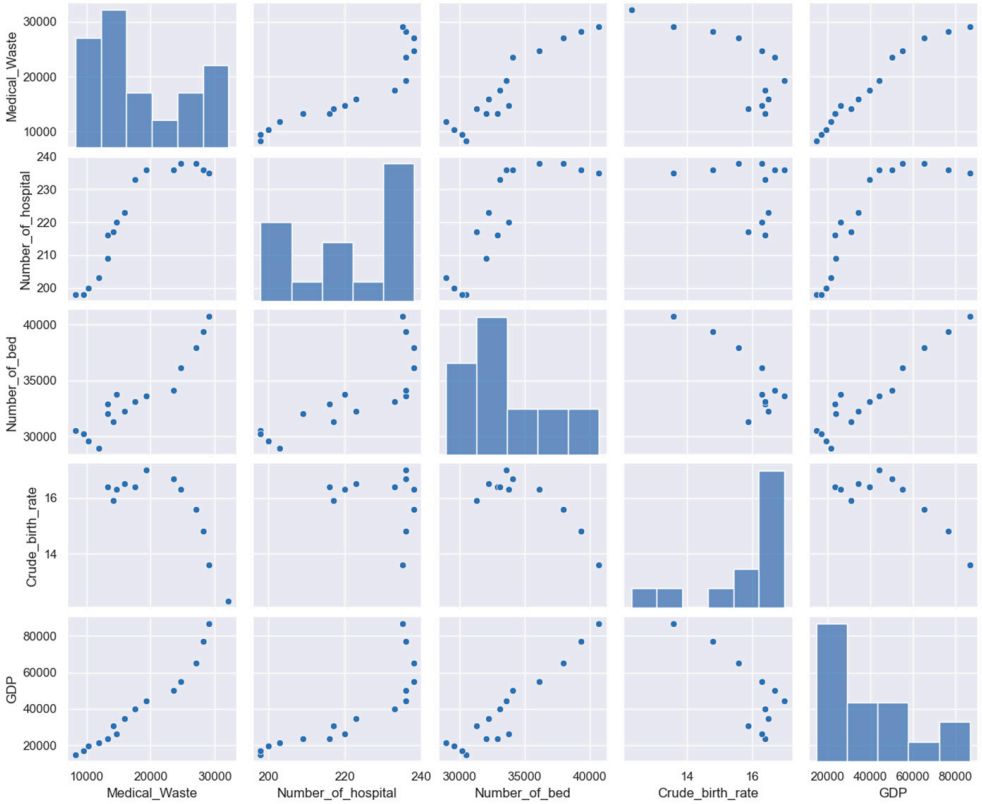


**Figure 4.** Pair plots for input and output variables.

### *4.3. Data Preprocessing*

Each data collection required a unique and distinct preprocessing method. Null values and outliers had to be processed prior to modeling, since NB, NP, CBR, and GDP had missing values and CBR and NB had outliers. The missing values of CBR were replaced with the median value because of its skewed distribution, whereas those of NB, NP, and GDP were replaced with their prior values because of an apparent upward tendency over duration.

### *4.4. Hyperparameter Optimization*

After computing, from the training and test data, performance measures based on parameter default values, the best combinations of hyperparameters were found. The tuned hyperparameters, hyperparameter values, and optimal values for the single machine learning algorithms are shown in Table 3.

**Table 3.** Results of hyperparameter optimization.

| Algorithm | Hyperparameters | Meanings | Search Values | Optimal Values |
|---|---|---|---|---|
| RF | max_depth | Maximum depth of tree | (5, 8, None) | 5 |
| | max_features | Maximum features of each tree | (3, 5, 15) | 3 |
| | n_estimators | Number of trees | (200, 500) | 500 |
| | min_samples_split | Minimum number of samples for leaf nodes | (2, 5, 8) | 2 |
| GBM | learning_rate | Shrinkage coefficient of each tree | (0.01, 0.1) | 0.1 |
| | max_depth | Maximum depth of tree | (3, 8) | 8 |
| | n_estimators | Number of trees | (500, 1000) | 1000 |
| | subsample | Subsample ratio of training samples | (1, 0.5, 0.7) | 0.5 |
| AdaBoost | learning_rate | Shrinkage coefficient of each tree | (0.01, 0.1) | 0.01 |
| | n l estimators | Number of trees | (500, 1000) | 1000 |

### *4.5. Comparison of Single ML Algorithms*

Performance measures of each ML algorithm were calculated and are shown in Table 4.

**Table 4.** Prediction results of single ML algorithms.

| Models | MAE | RMSE | R-Squared | MAPE |
|---|---|---|---|---|
| RF | 1093.342 | 868.734 | 0.959 | 0.050 |
| GBM | 1332.665 | 1117.195 | 0.939 | 0.064 |
| AdaBoost | 3349.578 | 2698.408 | 0.615 | 0.153 |

Regarding RMSE, RF obtained the greatest performance with 868.7344, followed by GBM and AdaBoost with 1117.1954 and 2684.4044, respectively. The performance success was achieved in the same order with respect to R-squared, MAE, and MAPE, with values of 0.959, 0.9391, and 0.6155; 1093.3421, 1332.6659, and 3349.5788; and 0.05, 0.0648, and 0.1531, respectively. Based on the findings, RF and GBM could be said to be good models for predicting MW amounts. Considering all performance measures, RF was the best model, as it had the lowest MAE, RMSE, and MAPE scores and the highest R-squared score. This means that RF and GBM could be used to predict future MW amounts even with a small dataset. Although AdaBoost has practical advantages, such as low implementation complexity and only a single tuning parameter, it did not perform well here.

Comparing these results with those of other studies utilizing MLR methods as measured by the performance criterion R-squared, the RF model was more successful here than in some studies [3,14,18,20] and nearly reached the very good MLR results of other studies in the literature [2,11–13]. As mentioned before, MLR entails many assumptions. Our dataset was not suitable for the use of MLR because of its multicollinearity. Comparing the RF model with the ARIMA model, while it gave better results than those of ARIMA in [17], it was less successful than ARIMA in study [1]. The problem with ARIMA models is that when the dataset is small, it is not long enough to capture the seasonality; furthermore, ARIMA models do not provide significant parameters. When the RF model was compared with ANN, a deep learning method, RF gave better performance results than the ANN in [2] but not better results than the ANN in [20]. ANN is not an efficient method for small datasets such as those in this study.

Comparing this study's results with those obtained for models dedicated to İstanbul using R-squared as a performance criterion, [1] achieved an R-squared value of 0.9888 using ARIMA, and [16] achieved a value of 0.9918 using LR, whereas this study achieved a value of 0.9599 using RF. Ceylan et al. [1] tried to estimate MW amounts using a time-series model, but this study did not propose the significant factors. Korkut [16] obtained a high score using only one input variable, namely population. However, knowledge of the most significant parameters affecting MW amounts will be helpful in strategic plans to make decisions. The machine algorithms in this study provided the degree of importance of parameters that affected the amount of MW. The significance of each input variable was computed using the RF, GBM, and AdaBoost methods depicted in Figure 5. The order of relevance for the RF algorithm was NB > GDP > NH > CBR, that for the GBM algorithm was GDP > NH > CBR > NB, and that for the AdaBoost algorithm was NH > NB > GDP > CBR.

Regarding the single ML algorithms, a comparative analysis of each variable is shown in Figure 6. While NB was the most critical factor for the RF algorithm, it was the least significant factor for the GBM algorithm. However, NH had a similar degree of significance for all ML algorithms in our study.


**Figure 5.** The feature importance of each single ML algorithm.

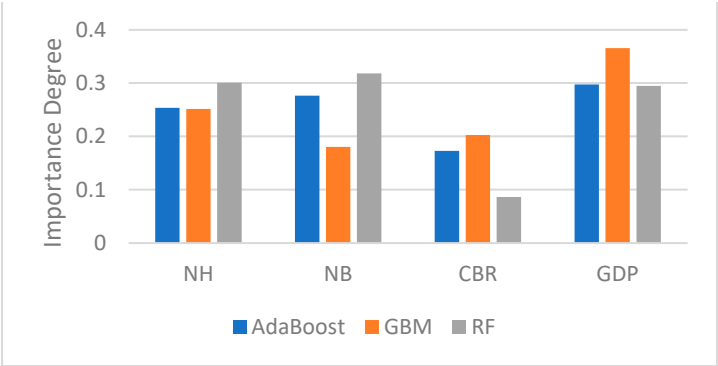


**Figure 6.** Degree of significance of each variable with relation to individual ML algorithms.

The total degree of significance for each variable is depicted in Figure 7. GDP was the most influential input variable on the amount of MW Daskalopoulos et al. [48], Dissanayaka and Vasanthapriyan [23], and the authors of [11] found that there was a substantial association between the volume of municipal garbage and the gross domestic product. Increasing a city's GDP or prosperity may result in a rise in garbage and medical waste production. This study revealed that NH was the second, and NB, the third, most significant variable in determining the quantity of MW, which was in accord with earlier research [3,32]. NB and NH are responsible for the production rate of infectious waste. The least important factor for the single ML algorithms was CBR.

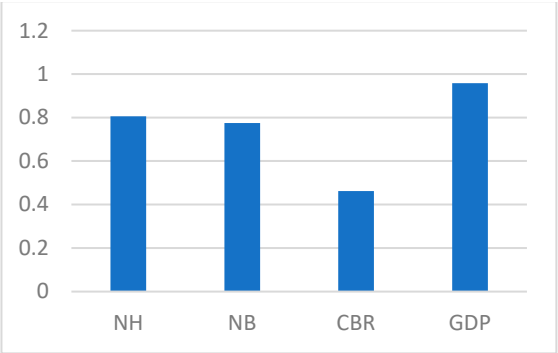


**Figure 7.** Total importance degree for each variable.

### *4.6. Performance Assessment*

In order to construct the ensemble VR, three base models, RF, GBM, and AdaBoost, were combined with weighted voting. According to performance measures, RF had three votes, GBM had two votes, and AdaBoost had only one vote. Using these weights, a voting regressor was fitted, and the performance measures of RMSE, MAE, MAPE, and R-squared were calculated, as given in Table 5. The results were compared with those of the standalone ML algorithms.

**Table 5.** Performance assessment of the proposed model.

| | MAE | RMSE | R-Squared | MAPE |
|---|---|---|---|---|
| Ensemble VR | 843.702 | 922.042 | 0.970 | 0.057 |

The proposed ensemble VR algorithm outperformed the standalone ML models in all the performance measures except for MAPE, with an MAE of 843.702, an RMSE of 922.042, and an R-squared of 0.970. Of the single models, RF had the smallest MAPE of 0.050, while the proposed model had 0.057. Since there were no previous studies utilizing the VR method, it is not possible to make comparisons with previous studies. However, the performance of the proposed method was closer to that of the algorithms that give better results in the literature compared with other standalone machine learning algorithms for predicting MW amounts [1,2,11–13,16]. As stated before, MW estimation is a challenging task, since data are limited and many factors can affect the amount of MW. The results of this study implied that the proposed model for estimating the MW amount had good performance by utilizing little real data as well as providing most significant input variables. Furthermore, this algorithm did not require assumptions, such as those required by statistical models, that are difficult to meet in real life.

## 5. Conclusions

From a sustainability point of view, both planning and developing waste management solutions for medical waste for future facilities depend on an accurate estimate of MW, especially in megacities such as Istanbul, where MW could have large effects on the environment and public health. Therefore, MW generation was predicted using three ML algorithms: RF, GBM, and AdaBoost. The performances of these algorithms were measured using MAE, RMSE, MAPE, and R-squared. At this step, the most important input parameters were also given. Then, these ML algorithms built the proposed ensemble VR, which used the ranking method to assign weights based on how well each single ML model did. All of these experiments were conducted with official data from the past 17 years, with the dependent variables being NH, NB, CBR, and GDP. RF did better than the other standalone ML algorithms, and GDP was the most substantial input variable for predicting the quantity of MW. The suggested ensemble VR fared better than individual ML methods.

The primary drawback of this research was that the dataset was relatively small and limited. The data on MW amounts are incomplete, as are those on other aspects, such as medical waste type, economic and social information, and the type of health institution producing the MW, and these may impact the MW amount in İstanbul, the most crowded city in Turkey. The appropriateness of the standalone and suggested ensemble VR algorithms for prediction could be determined using a larger database containing other factors that may in the future have a substantial impact on the amount of MW generated in İstanbul. Another limitation for this study was using only one hyperparameter tuning method, grid search, since we had little data. When we have more data, other hyperparameter tuning methods, such as Bayesian hyperparameter optimization and random search, and deep learning approaches will be utilized in future research. Choosing the relative weights for each ensemble member, such as random search or other optimization methods, remains a challenging task to solve in future studies.

This study will help practitioners and decision-makers create an effective medical waste management system by selecting the best algorithms for predicting MW volume and supplying key input factors. The suggested ensemble approach may lead to future prediction research in numerous fields, not limited to estimating MW.

**Author Contributions:** Conceptualization, methodology, formal investigation, and writing—original draft preparation, B.D.-İ.; review, validation, editing, and supervision, B.E. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Ceylan, Z.; Bulkan, S.; Elevli, S. Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: A case study for megacity Istanbul. *J. Environ. Health Sci. Eng.* **2020**, *18*, 687–697. [CrossRef]
2. Jahandideh, S.; Jahandideh, S.; Asadabadi, E.B.; Askarian, M.; Movahedi, M.M.; Hosseini, S.; Jahandideh, M. The use of artificial neural networks and multiple linear regression to predict rate of medical waste generation. *Waste Manag.* **2009**, *29*, 2874–2879. [CrossRef]
3. Golbaz, S.; Nabizadeh, R.; Sajadi, H.S. Comparative study of predicting hospital solid waste generation using multiple linear regression and artificial intelligence. *J. Environ. Health Sci. Eng.* **2019**, *17*, 41–51. [CrossRef]
4. Shinee, E.; Gombojav, E.; Nishimura, A.; Hamajima, N.; Ito, K. Healthcare waste management in the capital city of Mongolia. *Waste Manag.* **2008**, *28*, 435–441. [CrossRef]
5. Nie, L.; Qiao, Z.; Wu, H. Medical Waste Management in China: A Case Study of Xinxiang. *J. Environ. Prot.* **2014**, *5*, 803–810. [CrossRef]
6. Tirkolaee, E.B.; Abbasian, P.; Weber, G.-W. Sustainable fuzzy multi-trip location-routing problem for medical waste management during the COVID-19 outbreak. *Sci. Total Environ.* **2020**, *756*, 143607. [CrossRef]
7. Uysal, F.; Tinmaz, E. Medical waste management in Trachea region of Turkey: Suggested remedial action. *Waste Manag. Res.* **2004**, *22*, 403–407. [CrossRef]
8. Birpınar, M.E.; Bilgili, M.S.; Erdoğan, T. Medical waste management in Turkey: A case study of Istanbul. *Waste Manag.* **2009**, *29*, 445–448. [CrossRef]
9. Nguyen, X.C.; Nguyen, T.T.H.; La, D.D.; Kumar, G.; Rene, E.R.; Nguyen, D.D.; Chang, S.W.; Chung, W.J.; Nguyen, X.H.; Nguyen, V.K. Development of Machine Learning—Based Models to Forecast Solid Waste Generation in Residential Areas: A Case Study from Vietnam. *Resour. Conserv. Recycl.* **2021**, *167*, 105381. [CrossRef]
10. Tesfahun, E.; Kumie, A.; Beyene, A. Developing models for the prediction of hospital healthcare waste generation rate. *Waste Manag. Res.* **2015**, *34*, 75–80. [CrossRef]
11. Çetinkaya, A.Y.; Kuzu, S.L.; Demir, A. Medical waste management in a mid-populated Turkish city and development of medical waste prediction model. *Environ. Dev. Sustain.* **2019**, *22*, 6233–6244. [CrossRef]
12. Idowu, I.; Alo, B.; Atherton, W.; Al Khaddar, R. Profile of medical waste management in two healthcare facilities in Lagos, Nigeria: A case study. *Waste Manag. Res.* **2013**, *31*, 494–501. [CrossRef]
13. Al-Khatib, I.A.; Fkhidah, I.A.; Khatib, J.I.; Kontogianni, S. Implementation of a multi-variable regression analysis in the assessment of the generation rate and composition of hospital solid waste for the design of a sustainable management system in developing countries. *Waste Manag. Res.* **2010**, *34*, 225–234. [CrossRef] [PubMed]
14. Bdour, A.; Altrabsheh, B.; Hadadin, N.; Al-Shareif, M. Assessment of medical wastes management practice: A case study of the northern part of Jordan. *Waste Manag.* **2007**, *27*, 746–759. [CrossRef] [PubMed]
15. Sabour, M.R.; Mohamedifard, A.; Kamalan, H. A mathematical model to predict the composition and generation of hospital wastes in Iran. *Waste Manag.* **2007**, *27*, 584–587. [CrossRef]
16. Korkut, E.N. Estimations and analysis of medical waste amounts in the city of Istanbul and proposing a new approach for the estimation of future medical waste amounts. *Waste Manag.* **2018**, *81*, 168–176. [CrossRef]
17. Chauhan, A.; Singh, A. An ARIMA model for the forecasting of healthcare waste generation in the Garhwal region of Uttarakhand, India. *Int. J. Serv. Oper. Inform.* **2017**, *8*, 352. [CrossRef]
18. Karpušenkaitė, A.; Ruzgas, T.; Denafas, G. Forecasting medical waste generation using short and extra short datasets: Case study of Lithuania. *Waste Manag. Res.* **2016**, *34*, 378–387. [CrossRef]
19. Karpušenkaitė, A.; Ruzgas, T.; Denafas, G. Time-series-based hybrid mathematical modelling method adapted to forecast automotive and medical waste generation: Case study of Lithuania. *Waste Manag. Res.* **2018**, *36*, 454–462. [CrossRef]

20. Thakur, V.; Ramesh, A. Analyzing composition and generation rates of biomedical waste in selected hospitals of Uttarakhand, India. *J. Mater. Cycles Waste Manag.* **2017**, *20*, 877–890. [CrossRef]
21. Papacharalampous, G.; Tyralis, H.; Koutsoyiannis, D. Univariate Time Series Forecasting of Temperature and Precipitation with a Focus on Machine Learning Algorithms: A Multiple-Case Study from Greece. *Water Resour. Manag.* **2018**, *32*, 5207–5239. [CrossRef]
22. Pavlyshenko, B.M. Machine-Learning Models for Sales Time Series Forecasting. *Data* **2019**, *4*, 15. [CrossRef]
23. Dissanayaka, D.; Vasanthapriyan, S. Forecast municipal solid waste generation in Sri Lanka. In Proceedings of the 2019 International Conference on Advancements in Computing, Malabe, Sri Lanka, 5–7 December 2019; pp. 210–215. [CrossRef]
24. Meleko, A.; Adane, A. Assessment of Health Care Waste Generation Rate and Evaluation of its Management System in Mizan Tepi University Teaching Hospital (MTUTH), Bench Maji Zone, South West Ethiopia. *Ann. Rev. Res.* **2018**, *1*, 555566. [CrossRef]
25. Kilimci, Z.H. Ensemble Regression-Based Gold Price (XAU/USD) Prediction. *J. Emerg. Comput. Technol.* **2022**, *2*, 7–12.
26. Yang, N.-C.; Ismail, H. Voting-Based Ensemble Learning Algorithm for Fault Detection in Photovoltaic Systems under Different Weather Conditions. *Mathematics* **2022**, *10*, 285. [CrossRef]
27. Phyo, P.-P.; Byun, Y.-C.; Park, N. Short-Term Energy Forecasting Using Machine-Learning-Based Ensemble Voting Regression. *Symmetry* **2022**, *14*, 160. [CrossRef]
28. Pham, T.A.; Vu, H.-L.T. Application of Ensemble Learning Using Weight Voting Protocol in the Prediction of Pile Bearing Capacity. *Math. Probl. Eng.* **2021**, *2021*, 1–14. [CrossRef]
29. Bhuiyan, A.M.; Sahi, R.K.; Islam, R.; Mahmud, S. Machine Learning Techniques Applied to Predict Tropospheric Ozone in a Semi-Arid Climate Region. *Mathematics* **2021**, *9*, 2901. [CrossRef]
30. Bayahya, A.Y.; Alhalabi, W.; Alamri, S.H. Older Adults Get Lost in Virtual Reality: Visuospatial Disorder Detection in Dementia Using a Voting Approach Based on Machine Learning Algorithms. *Mathematics* **2022**, *10*, 1953. [CrossRef]
31. Liang, W.; Luo, S.; Zhao, G.; Wu, H. Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. *Mathematics* **2020**, *8*, 765. [CrossRef]
32. Li, Y.; Chen, W. A Comparative Performance Assessment of Ensemble Learning for Credit Scoring. *Mathematics* **2020**, *8*, 1756. [CrossRef]
33. Gutierrez, G. Artificial Intelligence in the Intensive Care Unit. *Crit. Care* **2020**, *24*, 101. [CrossRef]
34. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
35. Byeon, H. Exploring Factors for Predicting Anxiety Disorders of the Elderly Living Alone in South Korea Using Interpretable Machine Learning: A Population-Based Study. *Int. J. Environ. Res. Public Health* **2021**, *18*, 7625. [CrossRef]
36. Ahmad, M.; Kamiński, P.; Olczak, P.; Alam, M.; Iqbal, M.; Ahmad, F.; Sasui, S.; Khan, B. Development of Prediction Models for Shear Strength of Rockfill Material Using Machine Learning Techniques. *Appl. Sci.* **2021**, *11*, 6167. [CrossRef]
37. Freund, Y.; Schapire, R.E. Experiments with a New Boosting Algorithm. In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 148–156.
38. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
39. How to Develop a Weighted Average Ensemble With Python. Available online: https://machinelearningmastery.com/weighted-average-ensemble-with-python/#:~{}:text=Weighted%20average%20or%20weighted%20sum%20ensemble%20is%20an%20ensemble%20machine,related%20to%20the%20voting%20ensemble (accessed on 14 May 2022).
40. İstanbul Metropolitan Municipality Department Open Data Portal. Available online: https://data.ibb.gov.tr/en/dataset/tibbi-atik-miktari/resource/474d4b1c-6cd4-4626-8a69-d9c269d3809 (accessed on 6 June 2021).
41. The Official Website of Turkish Statistics Institute. Available online: https://biruni.tuik.gov.tr/medas/ (accessed on 10 June 2021).
42. Schwertman, N.C.; Owens, M.A.; Adnan, R. A simple more general boxplot method for identifying outliers. *Comput. Stat. Data Anal.* **2004**, *47*, 165–174. [CrossRef]
43. Ramírez-Gallego, S.; Krawczyk, B.; García, S.; Woźniak, M.; Herrera, F. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing* **2017**, *239*, 39–57. [CrossRef]
44. Kwak, S.K.; Kim, J.H. Statistical data preparation: Management of missing values and outliers. *Korean J. Anesthesiol.* **2017**, *70*, 407–411. [CrossRef]
45. Scikit-Learn. Available online: https://scikit-learn.org/ (accessed on 3 June 2022).
46. A Practical Introduction to Grid Search, Random Search, and Bayes Search. Available online: https://towardsdatascience.com/a-practical-introduction-to-grid-search-random-search-and-bayes-search-d5580b1d941d (accessed on 3 June 2022).
47. Awad, A.R.; Obeidat, M.; Al-Shareef, M. Mathematical-Statistical Models of Generated Hazardous Hospital Solid Waste. *Int. J. Environ. Waste Manag.* **2004**, *39*, 315–327. [CrossRef]
48. Daskalopoulos, E.; Badr, O.; Probert, S. Municipal solid waste: A prediction methodology for the generation rate and composition in the European Union countries and the United States of America. *Resour. Conserv. Recycl.* **1998**, *24*, 155–166. [CrossRef]