Prediction Medical Vaste Utilizing Ensemble Machine Learning Algorithms: A Case from Turkey

Burcu Devrim-Içtenbaş  
Anadolu Üniversitesi: Anadolu Universitesi

babak daneshvar rouyendegh  (babek.erdebilli2015@gmail.com)  
Ankara Yildirim Beyazit Uni  https://orcid.org/0000-0001-8860-3903

Research Article

Keywords: medical waste, prediction, Random Forests (RF), Gradient Boosting Machine (GBM), 24. AdaBoost, ensemble machine learning

Posted Date: January 28th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-879280/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
1. ABSTRACT
2. It is an important task to predict medical waste (MW) estimation accurately for effective Waste
3. Management System (VMS). The main aim of this study was to compare three ensemble machine learning
4. algorithms to predict medical waste for Istanbul which is the biggest city in Turkey. There exists new
5. machine learning (ML) algorithms called Ensemble Machine Learning Algorithms that have shown
6. significant success in other disciplines yet have not been examined for MW. To bridge the literature gap,
7. in this study, for the first time, a total of three ensemble machine learning algorithms: Random Forests
8. (RF), Gradient Boosting Machine (GBM) and AdaBoost are developed to predict MW generation. To
9. employ this study. 17-years real data were obtained from Istanbul Metropolitan Municipality Department
10. Open Data Portal with the input variables namely number of hospitals, number of bed available at the
11. hospital, crude birth rate and Gross Domestic Product (GDP). 80% of the total database being used for
12. developing the models, whereas the rest 20% were used to validate the models In order to compare their
13. performances, 5-fold cross-validation was applied and performance measures (MAE, RMSE and R-
14. squared) were calculated in this study. Of the ensemble models, the RF model provided better performance
15. than those of other models with RMSE, MAE, and $R^2$ of 1194.2, 898.12, 0.95, respectively, whereas the
16. second best GBM accuracy with RMSE, MAE, and $R^2$ of 1290.76, 1160.43, 0.94, respectively. Although,
17. CatBoost was interpreted as the efficient model for small datasets among the Machine Learning algorithms,
18. was poorest accuracy with RMSE, MAE, and $R^2$ of 3349.57, 2698.4, 0.61. In addition, the findings revealed
19. that GDP and number of hospitals were the most important inputs for the predicting MW generation using
20. ensemble machine learning algorithms. These results will helpful for decision makers regarding both
21. planning and designing medical waste management systems in the future facilities in the sense of
22. sustainable management.
23. Keywords: medical waste, prediction, Random Forests (RF), Gradient Boosting Machine (GBM),
24. AdaBoost, ensemble machine learning
25. INTRODUCTION
26. There is an increasing trend in population and population density because of urban growth due to industrial
27. and economic advancements. This will also lead to an increase the medical instutions like the number of
28. hospitals, clinics and other health facilities and led to large amount of medical waste (MW) production
29. (Ceylan et al., 2020; Jahandideh et al., 2009). The World Health Organization (WHO) and US
30. Environmental Protection Agency (USEPA) reported that MW is hazardous waste type since its potentially
31. dangerous existence of different pathogens, hazardous and chemical anticancer agents and radioactive
32. wastes (Jahandideh et al., 2009; Golbaz et al., 2019). Also, cutting and sharp materials can cause many
33. problems to whom they related (Golbaz et al., 2019). Shinee (2008) stated that approximately 15% - 25%
34. (by weight) of medical waste is accepted infectious. The collection and disposal of this waste in other words
35. the management of medical waste is the critical issue especially in developing countries due to its potential
36. public health risks and environmental pollution risks (Niez et al., 2014). Nowadays, the performance of waste
37. management system handling the medical waste is more critical (vital) than ever because of the
38. epidemic outbreak of the novel Coronavirus (COVID-19) (Tirkolae et al., 2021). The accurate prediction
39. of this type waste quantity (amount) will be helpful to determine the suitable disposal methods and to plan
40. the recycling, storage, transportation and disposal operation characteristics (Uysal and Tınmaz, 2004;
41. Birpinar et al., 2009; Ceylan et al., 2020; Cuong Nguyen et al., 2021).
42. To predict the amount of the MW in the future, several models have been used including data mining, time
43. series models, sample surveys, statistical models, artificial intelligence and machine learning algorithms in
44. the previous studies. Most of the researchers have been utilized Multiple Linear Regression (MLR) to
45. predict the generation rates and physical properties of MW with important factors such as occupancy rate,
46. number of hospitals, number of beds, number of total patients (Bdour et al., 2007; Sabour et al., 2007; Idowu
47. et al., 2013; Al-Khatib et al., 2016; Çetinkaya et al., 2020). They achieved high model performances
48. ($R^2>0.80$) with their models and determined the most influential factors but traditional MLR methods has
49. assumptions that are difficult to meet in real life and they can hardly predict MW amount when increasing
the number of input variables requires more complex modelling. Due to modeling the non-linear relationship between the input and output variables, some studies compare the conventional statistical techniques such as MLR with machine learning algorithms such as ANN, SVM and several Neuron and Kernel-based machine learning methods (Jahandideh et al. 2009; Karpušenkaštė et al. 2016; Thakur and Ramesh 2018; Golbaz et al. 2019). The studies showed that machine learning algorithms gave better results than statistical techniques and these results were attributed to the ability of solving the non-linear relationship between input and output variables (Karpušenkaštė et al. 2016). However, significant input variables were not addressed in these methods which emphasizing its performance in predicting medical waste production rate. Some studies utilized time series modeling as different autoregressive integrated moving average (ARIMA) models to predict the MW generation when they have time-based MW amount data (Chauhan and Singh 2017; Ceylan et al. 2020). But MW prediction is a regression problem rather than a time series problem because of limitations of time series analysis: have long time historical data to capture seasonality; data have missing value and outlier; many external factors that affect MW generation (Pavlyshenko 2019). Some studies revealed that traditional algorithms and machine learning based algorithms for time series data seem to be equally competitive for prediction problems (Papacharalampous et al. 2018; Pavlyshenko 2019). Ensemble machine-learning methods such as Random Forest and Gradient Boosting Machine will detect patterns in the time series even if having small data (Pavlyshenko 2019). These previous studies shown that ML can be utilized for predicting MW generation in sense of higher flexibility that their ability detect patterns, trends and fluctuations more accurately according to conventional regression analysis (Nguyen 2021). Also, most of the studies for predicting MW generation didn’t performed to determine most significant input variables, which will be a useful information for effective Medical Waste Management system. On the other hand, lack of historical MW database in especially developing countries may cause the difficulties with understanding the current situation and forecasting medical waste generation (Nguyen et al. 2021; Karpušenkaštė et al. 2016; Disanayaka and Vasanthapriyan 2019). Therefore, most of studies for predicting medical waste generation have been utilized surveys and questionnaire (Bdour et al. 2007; Meleko and Adane 2018; Golbaz et al. 2019) but this may yield inaccurate predictions since not using actual data. The main aim of this study is to apply and compare different models for predicting MW generation for Istanbul, the largest city in Turkey. Over the 15 million population, it has 17% of hospitals, 20% of bed capacity, and 54% of private hospitals of Turkey that means a big impact on both health and the environment (Birpinar et al. 2009; Ceylan et al. 2020). For this purpose, ensemble machine learning algorithms Random Forests (RF), Gradient Boosting Machine (GBM), and AdaBoost were utilized to predict MW generation. These ensemble methods were chosen because of their excellent performances in sense of working well small datasets and avoiding the overfitting (Liang et al. 2020). First, actual data for MW amount for 17 years was obtained from the İstanbul Metropolitan Municipality Department Open Data Portal hence more accurate estimation can be achieved rather than data obtained from surveys or questionnaires (Ceylan et al. 2020). Next, these ensemble Machine Learning algorithms are applied to predict MW generation. Finally, their comprehensive performances are analyzed and compared with the performance measures MAE, RMSE and R-squared and most significant factors are determined that affecting MW generation.

Although the proven successful machine learning algorithms have been used to predict the municipal solid waste generation (MSW) widely (Disanayaka and Vasanthapriyan 2019; Abbasi and Hanandeh 2016; Johnson et al. 2017; Kumar et al. 2018; Nguyen et al. 2021), ML algorithms have not received the necessary attention to estimate the MW generation (Golbaz et al. 2019). Also, to the best of our knowledge, Random Forests (RF), Gradient Boosting Machine (GBM), and AdaBoost algorithms have not been used to predict the MW generation. This study contributes to the use of machine learning algorithms, especially ensemble machine learning algorithms, in more studies on medical waste estimation.

The layout of the paper structured as follows: The details of Random Forests, GBM and AdaBoost were explained in Section 2. Section 3 specifies the prediction model from four ingredients: Data Acquisition, Data Pre-processing, Hyperparameter Optimization and Performance Measures. The results of applications are discussed in Section 4. Finally, Section 5 presents conclusions, limitations of the study and future
101.2. ML METHODS

ML is a subset of Artificial Intelligence (AI) that is intelligence in which machines extract knowledge from the data. ML combined statistical analysis techniques with computer science in order to generate algorithms capable of “statistical learning” (Gutierrez 2020). ML algorithms are divided into two categories: supervised and unsupervised. Supervised learning algorithms to discover relationships between potential explanatory features and a known target outcome and divided into two categories namely classification and regression. In this study Ensemble Methods Random Forests, Gradient Boosting Machine and AdaBoost algorithms performed to predict the MW generation.

109.2.1 Ensemble Methods

Ensemble methods is a machine learning algorithm build and combine several base models to solve classification problems, regression problems, feature selection with excellent performance (Li and Chen 2020). It has two types namely a parallel method represented by Bagging and sequential method represented by Boosting based on the base learner generation process (Li and Chen 2020). Multiple base learners are constructed simultaneously that means that independent of each other and this feature lead to improve the performance of final model while in the sequential ensemble type multiple learners is constructed in sequence hence the model is improved by the next learners can avoid the errors of previous learners.

118.2.1. Bagging

Several the training sets have been chosen with a bootstrap sampling method which means to take n samples from the data with replacement that can ensure the independence of different sampling training sets and then final model are determined by combining the predictions from all the models (Breiman 1996). The algorithm (Breiman 1996) is given Figure 1.

```
123. Input: Dataset S= {(x₁, y₁), (x₂, y₂) …(xₙ, yₙ)};
    i. Base learning algorithm L;
    ii. Number of base learners m.
124. Process: 
    For j=1,2,…,m:
      i. Sⱼ=bootstrap(S); %Generate a bootstrap sample from S 
      ii. hⱼ=L(Sⱼ)  %Train a base learner hⱼ from the the bootstrap sample
126. end.
127. Output: H(x) = mean( h₁(x),…,hₘ(x))   % For regression studies
```

**Fig. 1** The Bagging Algorithm (Li and Chen 2020)

129.2.1 Random Forests (RF)

Random Forest (RF) is a machine learning algorithm that mainly is used in Classification and Regression problems which combines the output of multiple decision trees to reach a single result. After building multiple decision trees on different samples with bootstrap sampling method, takes their majority vote for classification and average in case of regression (Breiman 2001; Byeon 2021). RF is a robust algorithm for missing data, unbalanced data but not sensitive to multicollinearity (Breiman 2001; Li and Chen 2020). The analysis can be divided into two stages:

136. Stage 1: The bootstrap statistic technique is used to randomly sample from the initial data set (training data) for creating a sequence of sub-data sets then using regression trees based on these sub-data sets, the forest is built. Each tree is trained by choosing a set of variables at random and two important parameters namely the number of trees (ntree) and the number of variables (mtry) can be adjusted during the training stage.

140. Stage 2: A prediction can be made after the model has been trained. Input variables are evaluated for all regression trees first, and then the final output is calculated by measuring the average value of each individual tree’s prediction (Ahmad et al. 2021).

143.2.2 Boosting

The basic idea that firstly a weak classifier is constructed on the training set, each sample assigned a weight based on classification performance. If the sample classified correctly, the weight relatively small number
otherwise it will be a large number. Boosting is an iterative process to make samples with large weights by adding weak learners to the previous weak learners and the data weights are readjusted in order to obtain final strong classifier. Boosting algorithm fits this kind of ensemble models

\[ f(x) = \sum_{m=0}^{M} f_m(x) = f_0(x) + \sum_{m=1}^{M} \theta_m \phi_m(x), \]  

where \( f_0 \) represents the initial guess, \( \phi_m(x) \) denotes the base estimator at iteration \( m \) and \( \theta_m \) is the weight for \( m \)th estimator. The product of \( \theta_m \) * \( \phi_m(x) \) is the “step” at iteration \( m \). Most of the boosting algorithms can be viewed as to solve

\[ \{\theta_m, \phi_m\} = \arg \min_{\theta_m, \phi_m} \sum_{i=1}^{n} L(y_i, f^{(m-1)}(x_i) + \theta_m \phi_m(x_i)) \]  

at each iteration, where \( f^{m-1} \) represents the current estimation.

**AdaBoost**

Adaptive Boosting (AdaBoost) is a well-known algorithm among Boosting algorithms that will when the loss function is exponential type and the weights and classifiers are derived by means of forward stage-wise additive modeling. It is proposed by Freund and Schapire (Freund and Schapire 1996) and the algorithm is given in Figure 2 (Li and Chen 2020).

**GBM**

Gradient Boosting Machine (GBM) is one of the Boosting algorithms introduced by Friedman in 2001. The algorithm also known as Multiple Additive Regression Trees (MART) and Gradient Boosted Regression Trees (GBRT) (Friedman 2001). GBDT is a kind of ensemble model of decision trees, which are trained in sequence and GBDT learns the decision trees by fitting the negative gradients (also known

---

146. Otherwise it will be a large number. Boosting is an iterative process to make samples with large weights by adding weak learners to the previous weak learners and the data weights are readjusted in order to obtain final strong classifier. Boosting algorithm fits this kind of ensemble models.

149.

\[ f(x) = \sum_{m=0}^{M} f_m(x) = f_0(x) + \sum_{m=1}^{M} \theta_m \phi_m(x), \]  

where \( f_0 \) represents the initial guess, \( \phi_m(x) \) denotes the base estimator at iteration \( m \) and \( \theta_m \) is the weight for \( m \)th estimator. The product of \( \theta_m \) * \( \phi_m(x) \) is the “step” at iteration \( m \). Most of the boosting algorithms can be viewed as to solve

\[ \{\theta_m, \phi_m\} = \arg \min_{\theta_m, \phi_m} \sum_{i=1}^{n} L(y_i, f^{(m-1)}(x_i) + \theta_m \phi_m(x_i)) \]  

at each iteration, where \( f^{m-1} \) represents the current estimation.

**AdaBoost**

Adaptive Boosting (AdaBoost) is a well-known algorithm among Boosting algorithms that will when the loss function is exponential type and the weights and classifiers are derived by means of forward stage-wise additive modeling. It is proposed by Freund and Schapire (Freund and Schapire 1996) and the algorithm is given in Figure 2 (Li and Chen 2020).

**GBM**

Gradient Boosting Machine (GBM) is one of the Boosting algorithms introduced by Friedman in 2001. The algorithm also known as Multiple Additive Regression Trees (MART) and Gradient Boosted Regression Trees (GBRT) (Friedman 2001). GBDT is a kind of ensemble model of decision trees, which are trained in sequence and GBDT learns the decision trees by fitting the negative gradients (also known
in each iteration, (Ke et al. 2017). Different from AdaBoost which use the error of previous weak learner to adjust sample weight then iterates one by one while GBM adjusted sample data in the direction of the steepest descent is given by the negative gradient of the loss function to make algorithm converge globally (Li and Chen 2020).

Given a data set as \( \{x_i, y_i\}, i=1,2,...,N \), where \( x_i \) denotes a set of features and \( y_i \) represents the label. \( \Psi(y, F(x)) \) is the loss function. The steps of GBM are shown as follows (Li and Chen 2020):

Step 1: The initial constant value \( \beta \) is obtained as:

\[
F_0(x) = \arg\min_{\beta} \sum_{i=1}^{N} \Psi(y_i, \beta). \tag{3}
\]

Step 2: the gradient of loss function is written as:

\[
y_i^* = \left[ \frac{\partial \psi(y_i, F(x))}{\partial F(x)} \right]_{F(x)=F_{k-1}(x)}, i=1,2,...,N. \tag{4}
\]

Step 3: The initial model \( h(x_i; \theta_k) \) is formed by fitting sample data, the parameter \( \theta_k \) is calculated by using the least square method:

\[
i. \quad \theta_k = \arg\min_{\theta_k} \sum_{i=1}^{N} [y_i^* - \beta h(x_i; \theta)]^2 \tag{5}
\]

Step 4: The new weight of the model is expressed as follow by minimizing the loss function

\[
i. \quad \gamma_k = \arg\min_{\gamma} \sum_{i=1}^{N} \psi(y_i, F_{k-1}(x) + \gamma h(x_i; \theta_k)) \tag{6}
\]

Step: Optimized the model as

\[
1. \quad F_k(x) = F_{k-1}(x) + \gamma_k h(x_i; \theta_k). \tag{7}
\]

This loop is implemented until a specified number of iterations or convergence conditions are met.

**193.3. Prediction Model**

In order to compare of three machine learning algorithms for the prediction medical waste, the experiment consist of five steps respectively, data acquisition, data pre-processing, utilizing ensemble methods, hyper-parameter tuning and evaluation criteria. The prediction model experiment procedure given in Figure 3. All analyses were performed using Python version 3.8.6 on PyCharm based on the scikit-learn and XGboost libraries. The Python function \( \text{random.seed()} \) is used in order to ensure the reproducibility of the dividing process.
3.1 Data acquisition

The MW data used for 17-year in this study was obtained from Istanbul Metropolitan Municipality Department Open Data Portal (https://data.ibb.gov.tr/2.2). The data set were merged medical waste from Waste Amount by District, Year and Waste Type Table (https://data.ibb.gov.tr/dataset/31d85b21-32a9-4270-95d9-a1712a6567ea/resource/50036dfd-aea5-4f06-832f-f7020fdafaf5a/download/ilce-yl-ve-atk-turu-bazinda-atk-miktar-2021.xlsx) and number of hospitals, number of bed available at the hospital, crude birth rate and Gross Domestic Product from Population Information Table (https://biruni.tuik.gov.tr/medas/) the years between 2004-2020. The dataset consists of five variables: medical waste (MW), number of hospitals (NH), number of beds (NB), crude birth rate (CBR) and Gross Domestic Product (GDP). The dataset consists of 17 rows and 5 columns. The trend of dependent variable (MW) data the years between 2004-2020 is shown in Figure X. As can be seen from the Figure X, the amount of medical waste is increasing gradually.

Fig. 3 The process of prediction model

202.3.2 Input Variables

Fig 4. Annual MW generation of İstanbul (İstanbul Metropolitan Municipality Department Open Data Portal)
The descriptive statistics for each variable are given in Table 1. The input variables were selected based on literature and data availability which may influenced to the MW generation. In the previous studies, medical waste depends on several factors such as number of beds/patients, type of specialization, number of size of departments (Award et al. 2004; Jahandideh et al. 2009). In this study, besides number of hospitals (NH) and number of beds (NB) socioeconomic factors such as crude birth rate and GDP for MSW generation were selected as input variables (Dissanayaka and Vasanthapriyan 2019).

MW is an output variable (i.e., the dependent or predictive variable or target output), while the other variables were the inputs (i.e., the independent variables). The boxplots of the variables are given Figure 5 for presenting distribution of variables and detecting outliers (Schwertman et al. 2004). Outliers are data points in a dataset that are abnormal observations between normal observations and can lead to odd accuracy scores that can skew measurements because the results don’t present true results. The presence of outliers suggests the need for pre-processing of the data to improve the accuracy of the result or the need for more advanced methods. It is appearance that all variables except for CBR don’t have any outlier observations.

Table 1 Descriptive Statistics of the input and output variables used to develop models.

| Category | Description                                                                 | Mean  | Standard Deviation | Min    | Max    |
|----------|-----------------------------------------------------------------------------|-------|--------------------|--------|--------|
| MW       | Amount of the medical waste per year                                       | 18.407,8 | 7.579,8          | 8.279,3 | 32.143,851 |
| NH       | Total hospitals per year                                                    | 221   | 15,5               | 198    | 238    |
| NB       | Total beds per year                                                         | 33.522,3 | 3.441,2          | 28.958 | 40.697 |
| CBR      | The ratio of the number of live births during the year to the average population in that year. The value is expressed per 1,000 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 |

Medical Waste: 
NH: 
NB:
230.3.3 Data Pre-processing

Data pre-processing is a crucial step in the machine learning modelling since, real-world data is usually incomplete, inconsistent, inaccurate (contains errors or outliers), and often lacks specific attribute values/trends (Ramírez-Gallego et al. 2017). Analyzing this kind of data will lead produce misleading results so it is extremely important that data pre-processing before feeding it into model. Data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction and selection (Nguyen et al. 2021).

The challenge of data preparation is that each dataset is unique and different. Handling null values and outliers are the concepts before the modelling in this study since NP, NB, GDP and CBR have missing values. Values and CBR has one outlier.

Winsorize method that to set extreme outliers equal to a specified percentile of the data when desiring to retain the observations (Kwak and Kim 2017). All observations greater than the (Q3+1.5*IQR) percentile equal to the value at the (Q3+1.5 IQR) th percentile and all observations less than the (Q1-1.5*IQR) th percentile equal to the value at the (Q1-1.5*IQR) th percentile in this study.

There are many methods namely deleting row, replacing with mean/median/mode, assigning a unique category, replacing values previous or next value, predicting missing values like a regression method and using algorithms which support missing values like KNN or RF to handle missing values for making right decision (Kwak and Kim 2017). Due to having not enough data, the missing values have been filled with the previous values for the NP, NB, GDP variables since these variables have an increasing trend over the years while CBR’s missing values have been filled with the median value because of it has skewed distribution. Cross-validation (CV) is a resampling procedure used to create random subsets of samples for training data, the remaining data for testing data when a limited data (Nguyen et al. 2021). Five-fold CV method is utilized in this study that the original training set is divided into five equal size subsamples. A one part is chosen as the validation set, the remaining four subsamples are used as the training sub-set. This method is repeated five times until each sub-sample is used as a validation set. Next, the average accuracy of the five validation sets is to assign the optimal hyperparameter values. The procedure is shown in Figure 6. The data set is divided into two groups as training set (80% of samples) is used for training the model and testing set (20% of samples) is used for evaluating samples.

229. Fig. 5 Boxplots for input and output variables used to develop models
258.3.4 Hyperparameter Optimization
259. Instead of finding the best combination of hyperparameters manually, grid search, randomized search, Bayesian optimization and heuristic search methods are mainly utilized as hyperparameters search methods.
260. Grid search provides a list of all hyperparameter values and their possible combinations, evaluates all combinations, and selects values that provide the best results. To determine the best set of hyperparameters, Grid Search is used in this study. Hyperparameters which are tuned, hyperparameter values and optimal values in Random Forests (RF), Gradient Boosting (GBM) and AdaBoost algorithms, as shown in Table 2.

| Algorithm      | Hyperparameters          | Meanings                                         | Search Values     | Optimal Values |
|----------------|--------------------------|--------------------------------------------------|-------------------|----------------|
| Random Forests | 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]        | 2              |
|                | min_samples_split        | Minimum number of samples for leaf nodes          | [2, 5, 8]         |                |
| GBM            | learning_rate max_depth  | Shrinkage coefficient of each tree                | [0.01, 0.1]       | 0.1            |
|                | n_estimators             | Maximum depth of tree                             | [3, 8]            | 8              |
|                | subsample                | Number of trees                                   | [500, 1000]       | 500            |
|                |                          | Subsample ratio of training samples               | [1, 0.5, 0.7]     | 0.5            |
| AdaBoost  | learning_rate | Shrinkage coefficient of each tree | Number of trees |
|----------|----------------|-----------------------------------|----------------|
| n_estimators | [0.01, 0.1] | 1000 | [500, 1000] |

267.3.4 Performance Measures

268. To compare the performances of the ensemble machine learning models, three metrics have been used: namely the mean absolute error (MAE), the root mean square error (RMSE) and coefficient of determination ($R^2$) as shown in Equation (1-3) (Karpušenkaitė et al. 2016; Nguyen et al. 2021)

i. \[ MAE = \frac{\sum_{i=1}^{n}|y_i-x_i|}{n} \] (8)

ii. \[ RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i-x_i)^2}{n}} \] (9)

iii. \[ R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-x_i)^2}{\sum_{i=1}^{n}(y_i-x)^2} \] (10)

271. Where, \( n \) is the number of observations, \( x_i \) is the predicted of waste generation for \( i \)th observation, \( y_i \) is the actual value for \( i \)th observation and \( \bar{x}_i \) is the average of predicted values. The models with high \( R^2 \) values are better than the models with low \( R^2 \) values. Also the closer the value of \( R^2 \) is to 1 means better fitting, the better the effect of the model.

275.4.RESULTS

276.4.1 Overall Prediction Results

277. The prediction results of RF, GBM, AdaBoost ensemble algorithms were obtained on the test set data.

278. Performance measures namely MAE, RMSE and \( R^2 \) for each algorithm was calculated performance measures as presented in Table 3.

| Models   | MAE     | RMSE    | \( R^2 \) |
|----------|---------|---------|-----------|
| Random Forests RF | 1194.29 | 898.12  | 0.95      |
| GBM      | 1290.76 | 1160.43 | 0.94      |
| AdaBoost | 3349.57 | 2698.4  | 0.61      |

281. With regards the RMSE, RF achieves the best performance (898.12), GBM and AdaBoost second and third with the 1160.43 and 2698.4 respectively and shown in Fig 7. As seen in Figure 8 and Figure 9, the same performance order is achieved with respect to \( R^2 \) and MAE with values 0.95, 0.94, 0.61 and 1194.29, 284.1290.76, 3349.57 respectively. According to the results, RF and and GBM had good performances in the 285.predicting MW. On the whole, all performance measures considered, RF outperforms other machine learning algorithms GBM and AdaBoost with the lowest MAE and RMSE also the higher \( R^2 \) performance scores. ANN provides higher accurate results than linear regression and random forest models. These results indicated that, even with a small dataset, an adequate prediction model can be developed utilizing ensemble machine learning algorithms.
1. Fig. 7 RMSE values of different algorithms performed in this study

![Fig. 7 RMSE values of different algorithms performed in this study](image)

2. Fig. 8 MAE values of different algorithms performed in this study

![Fig. 8 MAE values of different algorithms performed in this study](image)

3. Fig. 9 R² values of different algorithms performed in this study

![Fig. 9 R² values of different algorithms performed in this study](image)

**291.4.2 Feature Importance of input variables**

The relative importance of input variables can be an important information to an effective medical waste management system. In this study, the importance of each input variable was obtained Random Forests (RF), Gradient Boosting Machine (GBM), AdaBoost algorithms which was shown in Figure 10 respectively. For the Random Forests (RF) algorithm, the rank of importance was NB>GDP>NH>CBR. Using GBM algorithm, the rank of importance degree was GDP>NH>CBR>NB. According the AdaBoost algorithm the rank of importance degree was NH>NB>GDP>CBR.
The feature importance of each ensemble machine learning algorithm was shown in Figure 11. For instance, the number of beds most significant factor in RF algorithm while the less significant variable for GBM algorithm. Number of hospitals, nearly the same importance degree for different algorithms in this study. Overall, according to total important degree.
Overall, according to total important degree for each variable as shown in Fig 12. GDP is the most influential factor that effect MW generation. Daskalopoulos et al (1998) and Dissanayaka and Vasanthapriyan (2019) concluded that the high correlation between GDP affected by the MSW and this study concluded that the high correlation is also valid between GDP and MW. There is a direct relationship between the increase GDP or wealth of a city and the increase in waste generation as well as medical waste generation. NH and NB are the second and third most important factors for this study for predicting MW generation and consistent with previous study results (Golbaz et al. 2019; Bdour et al. 2020). Obviously, NH and NB affect the production rate of infectious waste. CBR contributed least to the predictive models.

5. CONCLUSION

Accurate prediction of MW amount can be useful for both planning and designing medical waste management systems in the future facilities in the sense of sustainable management (Çetinkaya et al. 2019; Ceylan et al. 2020). Therefore, Ensemble Machine Learning Algorithms Random Forests (RF), Gradient Boosting Machine (GBM) and AdaBoost were utilized to predict MW generation. In order to compare their performances, 5-fold cross-validation was applied and performance measures (MAE, RMSE and R-squared) were calculated in this study. These experiments all have been implemented to a 17-year real-world data which was obtained Istanbul Metropolitan Municipality Department Open Data Portal with the input variables namely number of hospitals, number of bed available at the hospital, crude birth rate and Gross Domestic Product (GDP). Among the three ensemble machine learning Algorithms, RF outperforms the other algorithms while GBM and AdaBoost are ranked second and third. GDP is the most important input variables for predicting the MW amount.

There are some limitations in this study. Firstly, the data set is relatively small. Istanbul which is the biggest city in Turkey, the data on MW generation is still incomplete regarding the other factors such as social economic, health institutions type and medical waste type may affect the MW generation as well as effective waste management system. Although the ensemble machine learning algorithms work well with small data set, prediction performance will be better for a larger dataset. This is the main limitation on prediction of MW generation. In the future, the adequacy of these algorithms for prediction can be concluded with a larger database and including the other input variables such as number of doctors, type of medical institutions which may significant effect to MW generation. Second limitation of this study is utilizing only one hyperparameter tuning method called grid search. Other hyper-parameter tuning approaches like Random Search and Bayesian hyper-parameter optimization will be utilized in the future research.

The results of this study will help decision makers and practitioners for establishing an efficient medical waste management system like selecting the suitable algorithms for accurate estimation of MW as well as a further insight about the significant (most important) input variables. With this study, it is hoped that
Machine learning algorithms will increase in medical waste estimation models.

Author contribution

Dr Burcu Devrim-İçtenbaş: formal analysis, investigation, data curation, writing—original draft.
Assoc Prof Dr Babak Daneshvar Rouyendegh (B.Erdebili): conceptualization, validation, supervision.

Funding

No funds, grants, or other support were received.

Data Availability

Data supporting the findings are available from the corresponding author upon reasonable request.

Declarations

Ethical approval
Not applicable.

Consent to participate
Informed consent was obtained from all individual participants included in the study.

Consent to publish
All the authors give the publisher the permission to publish the research work.

Conflict of interest
The authors declare that they have no conflict of interest.

REFERENCES

Abbasi M, Hanandeh AE (2016) Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste Manag 56:13-22. https://doi.org/10.1016/j.wasman.2016.05.018.
Ahmad M, Kamiński P, Olczak P et al (2021) Development of Prediction Models for Shear Strength of Rockfill Material Using Machine Learning Techniques. Appl. Sci. 2021 11:6167. https://doi.org/10.3390/app11136167
Al-Khatib IA, Abu Fkhidah I, Khatib JJ, Kontogianni S (2016) 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 34:225–34. https://doi.org/10.1177/0734242X15622813
Award AR, Obeidat R, Al-shareef M (2004) Mathematical–statistical models of generated hazardous hospital solid waste. Int J Environ Waste Manage 2: 315–327. https://doi.org/10.1081/ese-120027524
Bdour A, Altrabsheh B, Hadadin N, Al-Shareif M (2007) Assessment of medical wastes management practice: a case study of the northern part of Jordan. Waste Manag 27:746–59. https://doi.org/10.1016/J.WASMAN.2006.03.004.
Birpinar ME, Bilgili MS, Erdoğan T (2009) Medical waste management in Turkey: A case study of İstanbul. Waste Manag. 29(1):445-8. doi: 10.1016/j.wasman.2008.03.015.
Breiman L (1996) Bagging predictors. Mach. Learn 24:123–140. doi:10.1007/BF00058655
Breiman L (2001) Random forests. Mach. Learn 45:5–32.
Byeon H (2021) 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 376.18 :7625. https://doi.org/10.3390/ijerph18147625
Ceylan Z, Bulkan S, Elelvi S (2020) Prediction of medical waste generation using SVR, GM (1,1) and ARIMA models: a case study for megacity Istanbul. J Environ Health Sci Engineer 18:687–697. https://doi.org/10.1007/s40201-020-00495-8
Chauhan A, Singh A (2017) An ARIMA model for the forecasting of healthcare waste generation in the Garhwal region of Uttarakhand. India Int J Serv Oper Informatics 8:352.
Demir A (2020) Medical waste management in a mid-populated Turkish city and development of medical waste prediction model. Environ Dev Sustain 22: 6233–6244. https://doi.org/10.1007/s10668-019-00474-6
Daskalopoulos E, Badr O, Probert SD (1998) Municipal solid waste: a prediction methodology for the...
387. generation rate and composition in the European Union countries and the United States of America. Resour Conserv Recycl 24 (2):155-166. https://doi.org/10.1016/S0921-3449(98)00032-9
388. Dissanayaka DMSh, Vasanthapriyan S (2019) Forecast municipal solid waste generation in Sri Lanka. In 390. 2019 International Conference on Advancements in Computing 210-215. doi: 391. 10.1109/ICAC49085.2019.9103421.
392. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Statist 29(5): 1189-393. 1232 doi: 10.1214/aos/1013203451
394. Freund Y, Schapire RE (1996) Experiments with a New Boosting Algorithm. In Proceedings of the 13th 395. International Conference on Machine Learning, Bari, Italy, 3–6 July 1996, 148–156.
396. Golbaz S, Nabizadeh R, Sajadi HS (2019) Comparative study of predicting hospital solid waste generation 397. using multiple linear regression and artificial intelligence. J Environ Health Sci Engineer 17:41–51.
398. https://doi.org/10.1007/s40201-018-00324-z
399. Gutierrez G (2020) Artificial Intelligence in the Intensive Care Unit. Crit Care 24(1):101. doi: 400. 10.1186/s13054-020-2785-y.
401. Idowu I, Alo B, Atherton W, Al KR (2013) Profile of medical waste management in two healthcare 402. facilities in Lagos, Nigeria: a case study. Waste Manag Res 31:494–501. 403. https://doi.org/10.1177/0734242X13479429
404. Jahandideh S, Jahandideh S, Asadabadi EB, Askarian M, Movahedi MM, Hosseini S, Jahandideh M (2009) 405. The use of artificial neural networks and multiple linear regression to predict rate of medical waste 406. generation. Waste Manag. 29(11):2874-9. doi: 10.1016/j.wasman.2009.06.027.
407. Johnson NE, Ianiuk O, Cazap D et al (2017) Patterns of waste generation: A gradient boosting model for 408. short-term waste prediction in New York City. Waste manag 62:3-11.
409. https://doi.org/10.1016/j.wasman.2017.01.037
410. Karpušenkaitė A, Ruzgas T, Denafas G (2016) Forecasting medical waste generation using short and extra 411. short datasets: Case study of Lithuania. Waste Manag Res 34(4):378-87. doi: 10.1177/0734242X16628977
412. Ke G, Meng Q, Finley T et al (2017) LightGBM: A highly efficient gradient boosting decision tree. Adv 413. Neural Inf Process Syst 30: 3147–3155
414. Kumar A, Samadder SR, Kumar N, Singh C (2018) Estimation of the generation rate of different types of 415. plastic wastes and possible revenue recovery from informal recycling. Waste Manag 79:781-790. doi: 416. 10.1016/j.wasman.2018.08.045.
417. Kwak SK, Kim JH (2017) Statistical data preparation: management of missing values and outliers. Korean 418. J of anesthesiology 70(4):407. doi: 10.4097/kjae.2017.70.4.407
419. Li Y, Chen WA (2020) Comparative Performance Assessment of Ensemble Learning for Credit Scoring. 420. Math 2020, 8:1756. https://doi.org/10.3390/math8101756
421. Liang W, Luo S, Zhao G, Wu H (2020) Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and 422. LightGBM Algorithms. Math 8:765. https://doi.org/10.3390/math8050765
423. Liang YC, Maimury Y, Chen AHL, Juarez JRC (2020) Machine Learning-Based Prediction of Air Quality. 424. Appl. Sci. 10:9151. https://doi.org/10.3390/app10249151
425. Meleko A, Adane A (2018) Assessment of Health Care Waste Generation Rate and Evaluation of its 426. Management System in Mizan Tepi University Teaching Hospital (MTUTH), Bench Maji Zone, South 427. West Ethiopia. Ann Rev Resear. 1: 555566. http://dx.doi.org/10.19080/ARR.2018.01.555566
428. Nie L, Qiao Z, Wu H (2014) Medical Waste Management in China: A Case Study of Xinxian. J Environ 429. Prot Ecol 5: 803-810. http://dx.doi.org/10.4236/jep.2014.510082
430. Nguyen XC, Nguyen TTH, La DD et al (2021) Development of machine learning-based models to forecast 431. solid waste generation in residential areas: A case study from Vietnam. Resour Conserv Recycl 432. 167:105381. https://doi.org/10.1016/j.resconrec.2020.105381
433. Papacharalampous G., Tyralis H, Koutsoyiannis D (2018) Univariate time series forecasting of temperature 434. and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water 435. Resour Manag 32:5207–5239. http://dx.doi.org/10.1007/s11269-018-2155-6
436. Pavlyshenko BM (2019) Machine-Learning Models for Sales Time Series Forecasting. Data 4(1):1–11. 437. https://doi.org/10.3390/data4010015
438. Ramirez-Gallego S, Krawczyk B, Garcia S, Woźniak M, Herrera F (2017) A survey on data preprocessing
for data stream mining: Current status and future directions. Neurocomputing 239:39-57.
https://doi.org/10.1016/j.neucom.2017.01.078

Sabour MR, Mohamedifard A, Kamalan H (2007) A mathematical model to predict the composition and
generation of hospital wastes in Iran. Waste Manag 27:584–7. https://doi.org/10.1016/J

Schwertman NC, Owens MA, Adnan R (2004) A simple more general boxplot method for identifying
outliers. Comput stat & data anal 47(1) :165-174. https://doi.org/10.1016/j.csda.2003.10.012

Shinee E, Gombojav E et al. (2008) Healthcare Waste Management in the Capital City of Mongolia. Waste
Manag 28: 435-444.

Thakur V, Ramesh A (2018) Analyzing composition and generation rates of biomedical waste in selected
hospitals of Uttarakhand, India. J Mater Cycles Waste Manag 20:877–90. https://doi.org/10.1007/s10163-

Tirkolaee EB, Abbasian P, Weber GW (2021) Sustainable fuzzy multi-trip location-routing problem for
medical waste management during the COVID-19 outbreak. Sci Total Environ 756:143607. doi:
10.1016/j.scitotenv.2020.143607

Uysal F, Tinmaz E (2004) Medical waste management in Trachea region of Turkey: suggested remedial
action. Waste Management & Research. Int J of ISWA 22(5):403-407. DOI: 10.1177/0734242x04045690.
https://data.ibb.gov.tr/en/dataset/31d85b21-32a9-4270-95d9-1712a6567ea/resource/50036dfd-aea5-4f06-
832f-f7020fd05ad/download/ilce-yl-ve-atk-turu-baznda-atk-miktar-2021.xlsx (accessed 06 June 2021)
https://biruni.tuik.gov.tr/medas/ (accessed 10 June 2021)