



# Machine learning tool-based prediction and forecasting of municipal solid waste generation rate: a case study in Guwahati, Assam, India

T. Singh<sup>1</sup> · R. V. S. Uppaluri<sup>2</sup>

Received: 16 July 2022 / Revised: 11 October 2022 / Accepted: 29 October 2022 / Published online: 13 November 2022

© The Author(s) under exclusive licence to Iranian Society of Environmentalists (IRSEN) and Science and Research Branch, Islamic Azad University 2022

## Abstract

Integrated large-scale solid waste management (SWM) policies are the need of the hour to design, develop and sustain SWM models. An accurate prediction and forecasting of municipal solid waste generation (MSWG) rate are essential for such advanced strategies. The primary objective of this study is to examine the criticality of demographic and socio-economic parameters for the fair prediction and forecasting of the MSWG rate. Machine learning (ML) models were formulated by mapping solid waste quantities at the municipal level with socio-economic and demographic variables of Guwahati city. Tree-based ML algorithms, namely decision tree (DT), random forest (RF) and gradient boosting (GB), were applied to build the models with 1936 data size. The moving average (MA) approaches were adapted for the forecasting of the MSWG rate. Model validation resulted in a root mean square error, RMSE (3.01), mean absolute error, MAE (2.86) and coefficient of determination,  $R^2$  (0.99) for the GB model and correlation coefficient ( $r$ ) of 0.82 between observed and predicted values and thereby resulted in best performance in conjunction with DT and RF. With the exponential MA, the forecasted RMSE and  $R^2$  for GB, RF and DT were 2.12, 3.63 and 4.22; and 0.981, 0.972 and 0.967, respectively. However, with a model accuracy of 97%, the computation time for GB model (19.18 min) exhibited maximum due to its high complexity. The overall methodology involved developing effective tools to aid in regional SWM and planning through the integration of data sources in the public domain, pre-processing and modelling from diverse sources.

**Keywords** Municipal solid waste · Modelling · Machine learning algorithms · Prediction · Forecasting

## Abbreviations

$R^2_{DT}$	Coefficient of determination for decision tree
$RMSE_{DT}$	Root mean square error for decision tree model
$R^2_{RF}$	Coefficient of determination for random forest
$RMSE_{RF}$	Root mean square error for random forest model
$R^2_{GB}$	Coefficient of determination for gradient boosting
$RMSE_{GB}$	Root mean square error for gradient boosting model
$EMA_{DT}$	Exponential moving average for decision tree
$SMA_{DT}$	Simple moving average for decision tree

$WMA_{DT}$	Weighted moving average for decision tree
$EMA_{RF}$	Exponential moving average for random forest
$SMA_{RF}$	Simple moving average for random forest
$WMA_{RF}$	Weighted moving average for random forest
$EMA_{GB}$	Exponential moving average for gradient boosting
$SMA_{GB}$	Simple moving average for gradient boosting
$WMA_{GB}$	Weighted moving average for gradient boosting

## Introduction

Global solid waste management (SWM) programme necessitates upon the important criterion of the quantification of generated and collected solid waste. Municipal solid waste generation (MSWG) and its impacts on humankind and the environment are global issues. Presently, 1.3 billion tons/year of MSW is generated daily on a global scale and is expected to peak about 2.2 billion tons/year by 2025 (Palacio et al. 2019, World Bank, 2018). Rapid urbanization and

Editorial responsibility: J Aravind.

✉ R. V. S. Uppaluri  
ramgopal@iitg.ac.in

<sup>1</sup> Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati 781039, India

<sup>2</sup> Department of Chemical Engineering, Indian Institute of Technology Guwahati, Guwahati, 781039, India



enhanced consumption patterns profusely contribute towards such complex challenges associated with waste generation (Guerrero et al. 2013). In the SWM programme, the quantity of generated and collected solid wastes is a significant criterion (Kamaraj et al. 2020; Fu et al. 2015; Tchobanoglous & Kreith 1994). However, due to unplanned waste collection and inadequacy of strategic planning policies, the modelling-based SWM did not yet mature from a real-world application perspective. Therefore, to develop and implement an effective SWM system, it is crucial to assure upon a meticulous and definite prediction methodology of the MSWG rate (Azadi & Karimi-Jashni 2016; Yu et al. 2015; Ghinea et al. 2016).

Several factors and their highly complex combinations contribute towards the MSWG rate of a city or a village. These primarily refer to geographical location, population, literacy rate, food habits, culture and beliefs, number of households (HH) in a specific geographical area and various economic constraints, viz. gross district domestic product (GDDP) and employment status (Tauqeer et al. 2022a, 2022b). Furthermore, HH solid waste is highly heterogeneous and is widely dependent on the socio-economic status of the HH (Khalil et al. 2022; Miezah et al. 2015; Sankoh et al., 2012). Socio-economic and demographic variables that measure economic affluence and life style parameters have been found to enhance the MSWG rate. Notably, these refer to population income (Sankoh et al., 2012), gross domestic product (GDP), consumption expenditure level (Daskalopoulos et al. 1998) and purchasing power parity (Abbas et al. 2022; Zhu & Atikur Rahman 2020). All these relatively influenced the MSWG rate. Also, little prior art infers that socio-economic factors such as education influence, age (Debrah et al. 2021) and GDDP (Grazhdani 2016) potentially influence the MSWG rate. Additionally, occupation type (Abbas et al. 2022; Zhu & Atikur Rahman 2020) and employment status (Sankoh et al., 2012) have been indicated to be important markers.

A precise estimate of the MSWG rate is necessary for the design, development and commissioning of an integrated SWM system. To do so, in a concise study area, primary data are collected in terms of alternate MSWG rates and associated temporal trends. However, inadequate resources, inefficient and ineffective management, and a lack of measurement infrastructure in many developing nations eventually result in an incomplete historical data of the MSWG rate (Dyson & Chang 2005). Such data usually comprise of several outliers, missing information, noises, etc. Hence, the MSWG rate prediction through standard approaches that consider timely variations of the MSWG is highly challenging. Therefore, to address this problem effectively, new techniques are to be customized for the efficient prediction of SWM rate and thereby enable future generation capacities.

Sustainable waste management necessitates upon the accurate estimation of MSWG. Such quantification methodologies can be used as a foundation to improvise waste management laws, environmental impact assessments, social and economic cost evaluations, waste management system designs and planning of related infrastructure such as collection points, recycling facilities, landfills and incinerators (Hoque and Rahman 2020; Ma et al. 2020). Due to the inherent complexity of several variables, it is challenging to estimate MSWG with an acceptable precision (Abbasi and El Hanandeh 2016; Beigl et al. 2008).

Conventionally, either practical models or complex technologies have been used to project the MSWG (Beigl et al. 2008). The majority of MSW prediction models have been overly straightforward and application-focused. For instance, Sun et al. (2018) used a category estimation approach to forecast the creation of MSW for the vast Tokyo Metropolis. In order to optimize Shanghai's MSW recycling network, Lv et al. (2020) evaluated the quantity of future MSWG using binary linear regression. A number of studies have also been conducted that are solely concerned with the reliability and accuracy of the adopted methodologies for MSW rate estimation.

The available literature for the prediction of the SWG rate refers to a wide range of empirical and abstract modelling techniques (Molina-Gómez et al. 2021). These can be as simple as application-based pedagogies or as complex as sophisticated technologies with thorough academic insights (Chung 2010). In the prior art, more than 63 studies have been addressed for the MSWG rate prediction from 1980 to 2020. These can be classified into seven major categories from a methodological perspective. Briefly, these can be analysed as: – (i) detailed statistical approach based on surveys (Even et al. 1981; Lohani & Hartono 1985); (ii) regression analysis (Abu Qdais et al. 1997; Araiza-Aguilar et al. 2020; Denafas et al. 2014; Shamshiry et al. 2014); (iii) material flow model (Hu et al. 2010; Huang et al. 2013; Noufal et al. 2020; Schiller et al. 2010); (iv) trend analysis using time series (Ali Abdoli et al. 2012; Katsamaki et al. 1998; Navarro-Esbrí et al. 2002; Rimaityte et al. 2012); (v) artificial intelligence (AI) modelling approaches (Abbasi & El Hanandeh 2016; Hannan et al. 2015; Kolekar et al. 2016; Noori et al. 2009a, b); (vi) artificial neural network (ANN) (Adamović et al. 2018; Jalili Ghazi Zade & Noori, 2008; Noori et al. 2009a, b; Shahabi et al. 2014); and (vii) machine learning (ML) approaches (Abbasi et al. 2013; Abdoli et al. 2011; Johnson et al. 2017; Abbasi et al., 2016). However, each method has its own strengths and limitations.

While statistical models are common and useful, they are severely constrained with presumptions such as independence and normality of the input variables. Therefore, they could not resolve complex issues (Kumar et al. 2018). Despite indicating ambiguity and subjectivity, the system



dynamic modelling approach proved to be effective to describe the causation and relationships between variables in a system (Xiao et al. 2020). In few cases, more appropriate techniques such as input–output analysis and material flow analysis have been applied.

In the past two decades, a gradual increase in the application of AI was witnessed for the successful prediction and optimization of MSW using complex nonlinear processes with multidimensional and noisy data (Abdallah et al. 2020). For instance, Abbasi and El Hanandeh (2016) developed a prediction model and estimated monthly MSWG quantities for the following six years using novel AI technologies such as support vector regression and support vector machine (SVR/SVM), adaptive neuro-fuzzy inference system (ANFIS), ANN and k-nearest neighbours (kNN). Using ANN and decision trees (DT), Kannagara et al. (2018) suggested a study framework to forecast the amount of MSWG. Later, the authors examined the results of various models for the determination of subjectivity in the findings and their dependence upon a few crucial criteria. As a result, AI methodologies have been considered to be effective for the prediction of MSWG rate.

In addition, relevant ML techniques have been analysed for their efficacy towards the challenging problem of MSWG rate prediction and forecasting. While several algorithms may appear promising, the linear models have efficacy in terms of robustness. Based on a flow chart-like tree structure, the DT can predict the dependent variables using relevant degrees of freedom. The algorithm falls under the category of supervised ML. DT regression is often used for continuous dependent problems. Using the DT model, Rathod et al. (2020) predicted the MSWG rate by integrating domestic MSW quantities with demographic and socio-economic variables of 200 regions around Akola city (Maharashtra). The model performed with a root mean square error, RMSE of 0.1747 and an adjusted coefficient of determination,  $R^2$  of 0.5044, respectively. However, the model suffered overfitting issues.

Having been developed for about 20 years, the random forest (RF) can be used to handle both classification and regression issues. The RF follows the bootstrap aggregating (bagging) process (Breiman, 1996). High precision, good handling of missing values and a large number of attribute dimensions have been the potential benefits of the RF. For instance, the RF forecasting of Vietnam's MSWG yield provided the best results (Nguyen et al. 2021). The model performed with an  $R^2$  value greater than 0.96 and a mean absolute error (MAE) of 121.5–125.0. However, the computation time has been poor, and hence, the model would be ineffective for the real-time predictions.

The gradient boosting (GB) technique has been an efficient ensemble learner and is based on DT (Chen and Guestrin 2016). The GB algorithm has been renowned for

handling huge numbers of tuneable hyper-parameters, its ability to handle missing values and its many unique features including parallelization, remote computing and cache optimization (De Clercq et al. 2020). The algorithm has been proved to be highly effective in forecasting particulate matter ( $PM_{2.5}$ ), fire danger and heavy metals in comparison with other methods (Tauqeer et al. 2022a, b; Bhagat et al. 2021; Michael et al. 2021; Xiao et al. 2021). However, it has been rarely used to predict MSW. One article addressed the GB model based on organic fraction of the MSW projection (Adeogba et al. 2019). The authors improved the prediction accuracy by combining the outputs of weak models to form a single consensus model and obtained an  $R^2$  score of 0.766 and 0.899 for garden and food waste models, respectively. However, the model lacked good error accuracy (about 11%) as it underwent overfitting. Such poor performance can be resolved by applying L1 and L2 regularization penalties or hyper-parameter optimization (HPO).

Based on weekly waste generation data in Mashhad city (Iran), few authors reported the effectiveness of the feed-forward ANN model to estimate MSWG volume (Jalili Ghazi Zade & Noori, 2008). The authors obtained 0.746 correlation coefficient ( $r$ ) and 3.18% error for an ANN with 16 neurons in the hidden layer for the effective prediction of weekly waste volume. Another article reported that population and standard of living index parameters are important factors to influence MSWG rate (Daskalopoulos et al. 1998). With such a basis, the authors used a regression model to correlate GDP and mean standard of living in the UK and USA as strong correlating factors for the MSWG rate prediction. Despite reporting MSWG rate prediction, the above-mentioned prior art did not address forecasting, an important tool for the sustainable development of the smart city ecosystem. Few authors applied a seasonal autoregressive and integrated moving average (sARIMA) method for MSWG rate prediction (Navarro-Esbrí et al. 2002). Their results affirmed a medium- to long-term prediction that performed well for a minimum of 2–3 years of predictions with a mean relative error lower than 5% of the accuracy level. Gradient boosting regression trees (GBRT), an ML technique, were applied to predict the MSWG rate associated with more than 750,000 HH in New York city (Kontokosta et al. 2018). With an  $R^2$  of 0.87, the model was able to predict the total weekly MSWG rate. However, data limitations constrained the predictive efficacy of the model.

A primary issue with MSWG prediction model application is in terms of the complexity of the real-time system. This translates into a dependable dataset of waste categorization and intensive information as input factors and variables to the modelling system. On the contrary, modelling using multivariate analysis (more than one independent variable), including ANN, fuzzy logic systems, genetic algorithm, system dynamics and multiple regression analysis, may at times



foster complexities associated with the diverse interactions with the variables. Thereby, the validation of the prediction model becomes difficult (Kolekar et al. 2016). Moreover, decision-making and planning SWM is inevitable due to appropriate planning and operational strategies (Ali et al. 2022). Hence, the challenge for the accurate prediction of MSWG needs to be addressed along with an efficient forecasting strategy.

Despite affirming numerous statistical and computational approaches for MSWG rate prediction, very little prior art convey the targeted application of ML approaches such as DT, RF (Ghanbari et al. 2021) and GB (Adeogba et al. 2019). These methods were mostly studied individually but not collectively. In this regard, the ML algorithms have been promising to facilitate better accuracy for even smaller datasets. These algorithms have been proved to be efficient for MSWG prediction and as a future prospect for advanced research (Ghanbari et al. 2021). ML models can effectively tackle overfitting issues that translate into high training but lower test scores. This is due to the model learning trained data in an extensive framework. Such models being unfit to perform precisely on smaller databases have not been proved yet from the perspective of generalized strategies. To overcome the overfitting issue, multiple ML algorithms can be used for a wider range of working principles and a comparative investigation can be targeted to determine a model for the chosen problem. To predict the MSWG rate by applying a series of ML algorithms, no attempts have been made till date. Moreover, limited studies have been reported for MSWG prediction using the GB predictive forecasting model. However, it is a fact that GB is susceptible to overfitting issues and no studies have been reported to overcome the issue.

To address overfitting, the DT can be considered as one of the promising options of ML algorithms. In general, the ML algorithms facilitate predictive models with precision, ease in estimations and stability (Navada et al. 2011). DT-based methods (Johnson et al. 2017) have been efficiently and effectively adapted for waste modelling using cross-sectional data (Beigl et al. 2004). These models require minimum data transformations (Ali & Ahmad 2019). Adapting the DT model, support vector machines (SVM) and ANN algorithms, (Kavyanifar et al. 2020) predicted the rate of waste production per capita in coastal areas of Hormozgan province, Iran. For training and model validation purposes, the authors used 75% and 25% of data, respectively. The RMSE and MAE were considered as validation parameters. The investigations revealed that the predictive error of the DT approach was lower than both SVM and ANN methods.

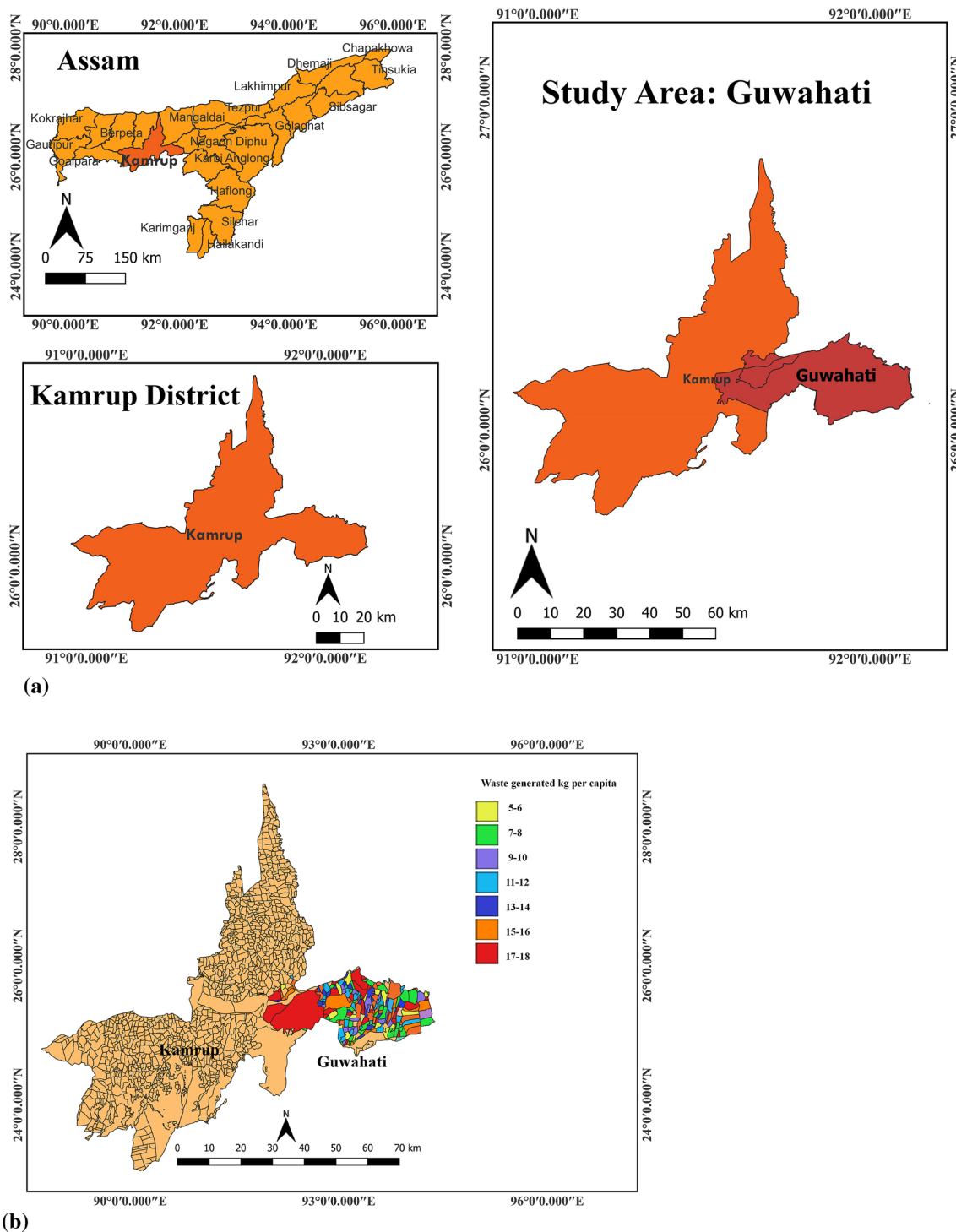
It is well known that the RF and GB machine learning algorithms have been widely used towards classification and predictive modelling of structured datasets (Adeogba et al. 2019). In this regard, RF and kNN algorithms were deployed

for MSWG prediction by few authors (Dissanayaka & Vasanthapriyan 2019; Nguyen et al. 2021). For the prediction of weekly MSWG rate in New York city, the GB was effective with a good  $r$  value of 0.82 (Johnson et al. 2017). Thus, GB, RF and DT can be considered to be potential ML algorithms for city wide MSWG prediction and forecasting.

With the proven limited scope of ML algorithms, three ensemble techniques, namely DT, RF and GB algorithms, have been considered in this article to develop a robust model and facilitate a comparative analysis and long-term model performance evaluation for the MSWG prediction in Guwahati city of north-east India. Guwahati, the capital of Assam state in north-east India, is one of the fast growing cities in India (Ahmed et al. 2021). Thereby, it can be considered as a representative city that needs appropriate and well-planned smart city strategies. Thereby, the city was selected to showcase the efficacy of chosen ML algorithms for MSWG rate prediction. To analyse the data, DT has been used as a first model that deployed ML as a nonparametric algorithm to model data separation limitations based on the learning decision rules on the input characteristics of the model. Following this, RF and GB were considered as the second and third approaches as predictive models for the exploration of associated correlations among socio-economic parameters. Moreover, to address the model overfitting issues, the chosen models were customized to serve better towards MSWG rate prediction. In this study, socio-economic factors, namely population (POP), GDDP, literate population (LP), total HH,  $HH_{Size}$  and worker population (WP), have been considered along with time to influence the MSWG rate prediction. The forecasting was carried out using simple moving average (SMA), weighted moving average (WMA) and exponential moving average (EMA) approaches.

## Study area

Guwahati, one of the oldest urban centres of India, occupies the most dominant position in administration and trade in the entire north-east India. The city lies between  $25^{\circ}5' N$  to  $26^{\circ}12' N$  latitude and  $91^{\circ}34' E$  to  $91^{\circ}5' E$  longitude and covers a large part of the Kamrup Metropolitan district of Assam (Fig. 1(a)). The spatial expansion of Guwahati has been due to the huge population explosion over a short period of time. According to the census reports, its population increased from 43,615 in 1951 to 9,68,549 in 2011. The average density of the population of the city is as high as 4468 persons per  $km^2$  as per the 2011 census. The city has presently 60 municipal wards and recently included a new ward no. 61 (Narengi ward) in the Kamrup Metropolitan district (Census 2011–19). The city centre, old city, business centres as well as the newly developed residential areas affirm highest population density. This is indicative of the



**Fig. 1** a Location map of the study area, b ward-wise waste generation per capita of Guwahati city (data source: GMC, 2018)

growing compactness and complexities in the land use of the city with time (Bhattacharyya, 2001). The highly heterogeneous functional characteristics of the city have led to the excessive population influx and rapid urban development. Like any other growing city, the lifestyle of the people is in a phase of transition due to rapid urban development.

This resulted in high waste generation from varied sources in the city.

Guwahati shares 1.94% of Assam’s geographical area with a GDP per capita of ₹280,650 (US\$3,700) in 2020–21. These demographics indicated it as one of the most economically vibrant regions in north-east India (Ministry of

Statistics and Programme Implementation, MOSPI Report 2020–21). The city reportedly generates about 550 TPD on a daily basis (Central Pollution Control Board, CPCB Report, 2019). The recent garbage dumpsite for the city is Chandrapur, an ecologically sensitive area in the suburbs of Guwahati. The dumpsite is within 400 m of the Kolong river, 500 m of the Amchang Wildlife Sanctuary and 2 km of the Pobitora Wildlife Sanctuary (home to many one-horned rhinoceroses). Hence, the site is opined to violate conservation and management rules implying landfill sites in terms of their location with respect to habitation clusters, forest areas, protected areas, water bodies and historical places. The previous dumpsite for the city was in Boragaon (close to Deepor Beel Wildlife sanctuary). Since 2006, about 95% of this waste has been dumped at the Boragaon site. Deepor Beel is a wetland recognized by the International Ramsar Convention and important bird habitat. The dumpsite's proximity to the sanctuary violated both MSW (Management and Handling) Rules 2000 and the Wetland (Conservation and Management) Rules 2010, both for polluting a protected site. Further, the Guwahati Municipal Corporation (GMC) and the non-governmental organizations (NGOs) could not resolve the primary issues of the irregularities in garbage collection from the doorsteps of the city's residents. Despite taking several proactive measures and steps for the improvement of solid waste scenario, the GMC has a highly challenging issue to collect and dispose the solid waste. The issue is becoming chaotic and translates into an alarming situation. Hence, a proper planning is required in this aspect. Without such appropriate planning and management, such vast quantities of waste are bound to hinder the sustainable development of the city and may even lead to civil unrest and strife due to inadequate and imbalanced emphasis on the quality lifestyle for one and all.

Figure 1(b) shows the MSWG per capita in the Guwahati region (Data source: GMC, 2019). It has been observed that the standard deviation of MSW generation per capita of municipalities with a lower population density (rural municipalities) was lower than that of the areas with higher population density (urban municipalities). Thus, the MSWG per capita in urban areas has a wider range. For instance, the city with an area of 14.98 km<sup>2</sup> generates 160 tonnes waste per day for a population density of 2223 (33,305 persons). Incidentally, the ward with highest population density only possessed an area of 25.97 km<sup>2</sup> but for 49,021 persons ward. With respect to the annual MSWG per capita, the increment was approximately threefold and got enhanced from 2.84 kg per person (1991) to 6.21 kg per person (2019).

## Materials and methods

Python version 3.8 was deployed to analyse the data and implement the prediction models. To automate loading, pre-processing and integration of data, python scripts were developed. NumPy, Scikit-learn, Pandas and Matplotlib library packages have been considered to implement modelling efforts on a i7-4790 CPU @ 3.60 GHz processor configuration.

Figure 2 outlines the overall methodology being adopted in this article. The figure depicts a data collection and pre-processing phase and a modelling and analysis phase in the overall methodology. For the preparation and transformation of raw secondary data into processed data for appropriate modelling and analysis, several data pre-processing steps were needed. These refer to the extraction of data (secondary), data loading into proper data structures, derivation of socio-economic factors, data transformation, outliers removal through filtering mechanisms and finally integration of data into consolidated datasets. Finally, after retrieving the combined datasets, the data were fed to ML algorithms and for the respective functionalities of the training, validation and forecasting of the alternate models.

## Data consolidation and socio-economic parameters selection

The consolidated data for the conducted investigations include POP, GDDP, GDDP<sub>per capita</sub>, LP, WP and HH information Table 1. For these data, the best data sources correspond to the Indian census data (a decennial census performed nationwide), and for GDDP, the annual data were collected from the MOSPI. Also, the GDDP<sub>per capita</sub> was obtained by dividing the country's GDDP with its total population. For HH data, the number of HH and HH sizes were considered. Also, the number of HH was further sub-categorized into good, living and dilapidated conditions. Accordingly, HH<sub>Size</sub> was classified as 1–3, 4–8 and above 9.

An extensive prior art in the chosen field highlighted socio-demographic factors and datasets as the most significant contributors to MSWG rate (Benítez et al. 2008; Ghinea et al. 2016; Johnson et al. 2017; Noori et al. 2009a, b; Tchobanoglous & Kreith 1994). The population (POP) has been unarguably a dominant parameter to critically influence the total waste generation rate. Hence, it has been regarded as a key variable to customize the performance of long-term time series prediction models (Hockett et al. 1995; Katsamaki et al. 1998; Navarro-Esbrí et al. 2002; Rimaityte et al. 2012). Most socio-economic factors that influence the MSWG rate such as education level (Debrah et al. 2021), employment status (Bandara et al. 2007), total HH and HH<sub>Size</sub> (Suthar & Singh 2015) have been provided

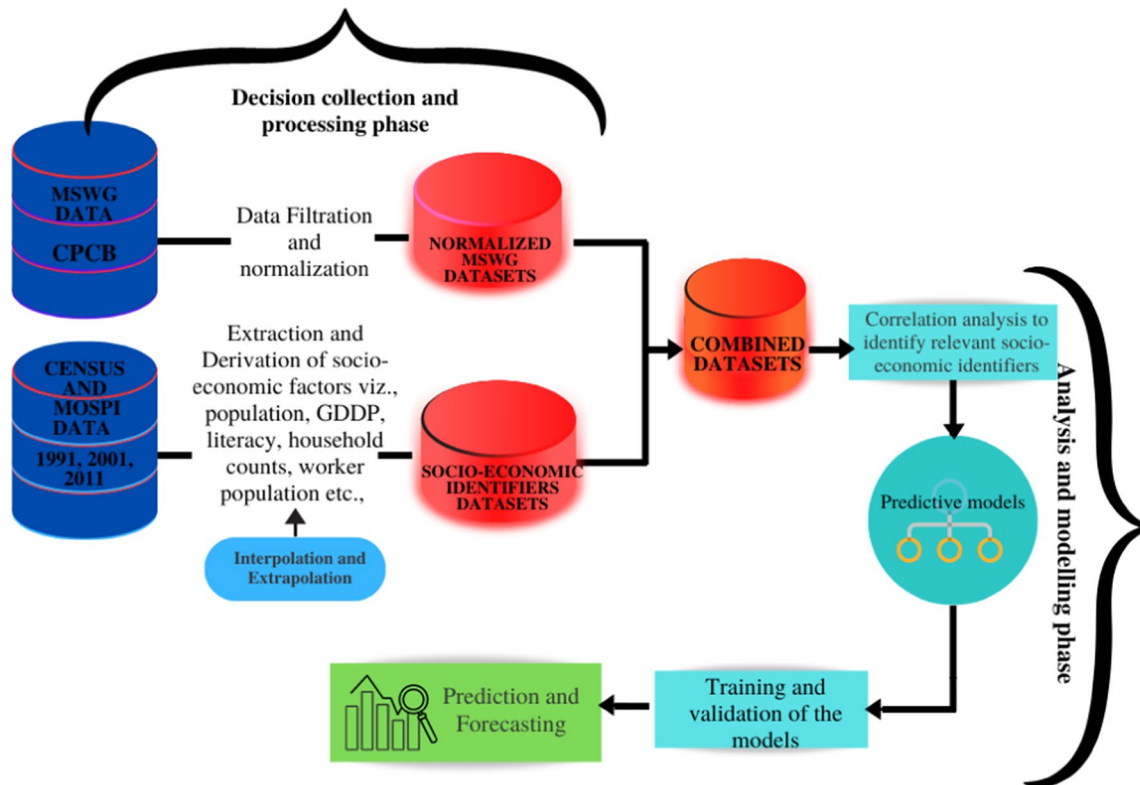


Fig. 2 Overall MSWG prediction and forecasting methodology using ML algorithms

**Table 1** Data sources for the independent and dependent variables in the conducted study

Parameters	Data sources	Available years
<i>Independent variables</i>		
POP	Census data	1991–2011
LP	Census data	1991–2011
GDDP	MOSPI data	1991–2011
HH data (HH count: <i>Good, Living, Dilapidated</i> and sizes: <i>HH 1–3, 4–9 and above 9</i> )	Census data	1991–2011
WP (main, marginal and non-WP)	Census data	1991–2011
<i>Dependent variable</i>		
MSWG	CPCB	1991–2016

in the national census data and at 10-year intervals. The annual data for GDDP (Grazhdani 2016) were obtained from MOSPI at the municipal level. This was as per the norm that the MOSPI data at subdivisions referred to that of the municipality level. The same was verified through a comparison of POP projections reported in waste generation and census data. Further, such level-based data classification in the census data evaluates the competence of the desired socio-economic factors in various alternate methods. Also, over the years, this level-based data varied

substantially. Accordingly, the socio-economic factors were interpreted in such a way that they can be invariably computed using data structures in the census data of 1991, 2001 and 2011. To evaluate the annual values from 1991 to 2011, linear interpolation and extrapolation were adopted. To provide adequate consistency for short-length data imputation issues, linear interpolation has been reported (Sun & Chungpaibulpatana 2017). Additionally, an analysis of the relevant socio-economic factors in the census reports affirmed progressive but not random variations.

In this article, the MSWG rate data for validation were procured from the CPCB, Government of India. The CPCB presents annual consolidated reports being developed with the forwarded reports of the SWM from the local bodies and then to the State Pollution Control Boards. Thereby, CPCB is responsible for coordination and assigns resources for waste management programmes deployed by multilevel municipal governance bodies. Thus, CPCB information with respect to waste generation rates is available in the public domain. Despite generating a compiled waste generation data from the municipalities, the data can have errors due to the self-reporting strategies. Thereby, such data may not be free from errors due to inaccurate records, over- or under-evaluated waste quantity data and misinterpreted survey records. For this reason, since obtained process data reconciliation (PDR)

and data validation are prone to possess erroneous information and outliers for the carried investigations, it was assumed that to describe cross-sectional differences among municipal bodies, the data possessed adequate quality and variation. Thereby, the data can be applied for the development of better predictive models.

To systematically assist modelling efforts, the preliminary screening of socio-economic factors was targeted between MSWG rate and socio-economic factors. To do so, Pearson correlation coefficient ( $r$ ) being applied in an earlier investigation as a ranking technique to select ML algorithm features was adopted (Iguyon & Elisseff 2003). Thereby, the  $r$  between each variable pair was calculated using the expression:

$$r = \frac{N(\sum xy) - (\sum x)(\sum y)}{\sqrt{[N \sum x^2 - (\sum x)^2][N \sum y^2 - (\sum y)^2]}} \quad (1)$$

where  $r$  is the correlation coefficient between variables  $x$  and  $y$ , and  $N$  is the number of observations. Thus, using correlation coefficients, the criticality and sensitivity of the socio-economic parameters have been addressed to considerably improve upon the prediction accuracy.

### Pre-processing of the data

To prepare and transform the collected data into refined data, various data pre-processing steps were needed for appropriate modelling and subsequent analysis. These refer to the structures deriving socio-economic factors, transforming data, filtering to remove outliers and finally integration of the data into combined datasets.

For the detection and removal of outliers from datasets, interquartile range (IQR) filtering was used. Outliers can result into inaccurate estimates or significant changes in waste generation patterns due to the divergence of the data from the municipal bodies. The upper and lower limits of the valid data range per municipality were determined as Upper Limit =  $Q_3 + IQR \times 1.5$  and Lower Limit =  $Q_1 - IQR \times 1.5$ . In these expressions,  $Q_1$  and  $Q_3$  correspond to the first (or lower) and third (or upper) quartiles of the collected data, respectively. Subsequently, the data points that lie beyond and below the corresponding upper and lower quartiles were considered as outliers and were filtered.

### Machine learning techniques

Regression has been a dominant and major statistical practice in ML and can be applied in economics, psychology, geography and so forth (Dietz et al. 1997; Sammut and Webb, 2017). To evaluate the dependence or relation between random

variables of interest and thereby deduce mathematical functions, regression analysis (RA) is applied in the form of regression models that explain the pertinent data correlation characteristics (Dietz et al. 1997; Wisniewski & Rawlings 1990). RA requires two real-valued variables being represented as target/dependent ( $y$ ) and independent variables ( $x$ ). The primary objective of RA is to map a function such that  $y = f(x) + \epsilon$ , where  $\epsilon$  corresponds to the error (Draper and Smith 1981; Freund and Wilson 1998). Multiregression problems arise due to the occurrence of more than one independent variable. For such systems, regression expression transforms to  $y = f(x_1, x_2, x_3, \dots, x_n) + \epsilon$  where  $(x_1, x_2, x_3, \dots, x_n) \in x$ .

Several independent variables (namely POP, LP, GDDP and  $GDDP_{percapita}$ ,  $HH$ ,  $HH_{size}$  and  $WP$ ) and one target variable (MSWG) have been considered in this study. To estimate the mathematical functions, regression algorithms was applied to evaluate the best-fit model that represents MSWG as a function of all the considered seven core independent variables. Mathematically, the objective function is expressed as:

$$MSWG = f(\text{POP}, \text{LP}, \text{GDDP}, \text{GDDP}_{percapita}, \text{HH}, \text{HH}_{size}, \text{WP}) + \epsilon \quad (2)$$

### Decision tree

Tree-based models are most popular among ML models. This is due to their iterative divide-and-conquer nature. In such models, nonparametric method is followed to recognize complex patterns associated with the classification tasks that have the properties associated with several types of pattern and a large number of attributes (Kannangara et al. 2018). Such models are efficient and easy to implement but with intensive computation. A set of highly explicable and logical (if-then) conditions is constructed by the tree-based ML models. This was achieved recursively by subdividing the decision space into smaller sub-spaces using training data presented to the decision process in the form of a tree. Thus, the algorithms can discreetly implement feature selection and can be applied for classification and regression of datasets with larger variables (Quinlan 1999; Solano Meza et al. 2019). Recently, DT techniques have been applied on a wider scale to predict the waste generation rate and for long-term waste prediction (Breiman 1996; Breiman et al. 2017).

Numerous techniques exist to construct a tree. However, in this article, regression tree (RT) was constructed using a famous yet most widely applied framework called classification and regression trees (CART). CART was first introduced by Brieman's research group (Breiman et al. 2017). The CART involves the specific logical tests being conducted for the entire datasets ( $D$ ) that exist at the root or internal node of the tree. Thereby, DT partitions the test data into two groups  $D_1$  and  $D_2$  such that the partitions mutually satisfy the minimization of overall sum of squared error, SSE:





$$SSE = \sum_{i \in D_1} (y_i - \bar{y}_{D_1})^2 + \sum_{j \in D_2} (y_j - \bar{y}_{D_2})^2 \quad (3)$$

where  $\bar{y}_{D_1}$  and  $\bar{y}_{D_2}$  refer to the mean predictions of the training set for  $D_1$  and  $D_2$ , respectively. The process is iterated until the convergence criterion are met with no more possible splits. Finally, the decision nodes (last nodes) yield the predicted values of the targeted variables (Hastie et al., 2011; Johnson et al. 2017; Kannangara et al. 2018).

### Random forest

Using random sub-spaces and bagging approaches the research group of Nisbet and Breiman (Nisbet et al., 2009), Breiman et al. (2017) were the first to delineate upon the RF model-based bagging algorithm for classification purposes. As one of the most dominant ML algorithms, the RF incorporates an ensemble of tree predictors. Thereby, for all trees in the forest, each tree relies on the values of an arbitrary vector sampled individually and with the same distribution. Thereby, RF generates the final output as an average of all tree predictions (Shi & Horvath 2006).

### Gradient boosting

GB model demonstrates the ability to model complex non-linear relationships between variables. The algorithm functions on the concept of the DT regression models. Promising performance to achieve higher prediction accuracy than that being achieved through conventional time series approaches has been reported (Johnson et al. 2017). DT utilizes an approach to bifurcate and thereby customize various linear models to fit each region (Breiman et al. 2017; Hastie et al., 2011). Thereafter, the process is recursively performed through the determination of the split point (maximum deduction for residual sum of squares, RSS) at each stage. Such a procedural methodology generates a single tree-like structure that best depicts the underlying correlation between variables in a dataset.

In this study, two ensemble procedures have been used. These are the extensions of bagging (RF) and boosting (GB), respectively. Bagging (also known as bootstrap aggregation) and boosting are two widely used ensemble learning paradigms (models that result in a combination of multiple simple models) in an ML algorithm. The core idea behind these DT-based ensemble strategies is to build several DTs and consolidate their predictions through the average (in regression) or voting (in classification) methodologies. Hence, while the variance gets reduced, the prediction accuracy gets enhanced. Thereby, bagging allows the construction of individual DT and allocates equal weight to all the DTs. On the other hand, in due course of the boosting of the new DT, the performance of the prior ones gets influenced and a

weight-based assignment on the trees' performance supports the computation procedures (Friedman 2001; Sutton 2005).

Table 2 lists the most frequently used approaches for the prediction and forecasting of the SWM. Such comparative analysis will be useful for the researchers to choose appropriate algorithm for the mentioned problem. In the traditional approaches such as statistical and material flow models, the application of the heterogeneous data and minimization of error and uncertainties are not possible. Hence, prediction and forecasting accuracies could not be enhanced. As a viable alternative, the ML has an added advantage to include all possible information in the database and field survey data. Thus, the ML-based methods, due to non-priori selection of the variable, are more successful than the traditional methods. The methodology outlined in the article will assist in the decision-making process to accurately estimate the MSWG rate. Thereby, the limitation of traditional approaches for MSWG rate could be overcome for the effective planning of SWM facilities. Also, among ML, except DT, RF and GB, other approaches have limitations in terms of identification and inclusion of other influencing explanatory variables, sensitivity to outliers, lower accuracy, etc. However, the tree-based models have the advantage to consider available information in a particular field. Hence, greater possibilities exist to enhance the prediction accuracy of the model.

### Model training

For unseen or new data samples, signifying poor performances, the ML algorithms are susceptible to the overfitting issue. Hence, model testing is mandatory with the unseen data. To do so, the entire dataset was divided into training and testing datasets. While the training set was used to build the model, the overfitting was evaluated with the testing set. In general, the relevant prior art used 80:20 or 60:40, 90:10, 85:15 (Pao & Chih 2006) ratios of the training to testing datasets. The trained models have been evaluated through the testing of new data for which no model existed previously. To do so, the training and test datasets were split as 70:30 ratio in this work. This was justified with the observation that the chosen ratio yielded a minimum average model error.

### Model testing and validation

To analyse the regression algorithms' capabilities for the set objective, the trained models have been tested with new dataset being not utilized for the training phase. The validation process has been based on a two-step hierarchy, namely a parameter optimization (PO) phase and a training or prediction phase. During PO, for all the tree models, to infer appropriate parameters for the data with tenfold cross-validation, a

**Table 2** Advantages and limitations of the traditional and different ML algorithms used in SWM

Approach	Advantages	Limitations	References
Statistical	Reliability of results	Cannot be applied to heterogeneous data	Kumar and Samadder, (2020)
Material flow model	Reduces the likelihood of risks, downtime and additional costs	Overlooks errors and uncertainties	Huang et al., (2013); Noufal et al., (2020)
SVR/SVM	Appropriate for small-sized problems with low generalization error	Hypersensitive to outliers and kernel selection	Abbasi et al., (2022)
k-means	Fast and easy implementation with fewer parameters	Hypersensitive to noise and outliers	Du et al. (2022)
ANFIS	Combines the concept of ANN and fuzzy logic	Unsuitable for features with higher dimensions	Younes et al., (2015)
kNN	Outlier insensitive	High computation and lower accuracy	Nguyen et al., (2021)
RF	Feature selection is not necessary and model variance is reduced	Insensitive to attribute data with varying values	Ghanbari et al., (2021); Nguyen et al. 2021
DT	High observability and efficiency	Overfitting issues and insensitive to feature correlation	Kavyanifar et al., (2020)
GB/GBRT	Train faster on larger datasets and accuracy more	Computationally expensive	Adeogba et al., (2019)
CNN	Feature extraction ability	Requires parameter tuning and huge datasets	Lin et al., (2021)
ANN	High robustness and fit for any nonlinear relationship	Requires multiple parameters	Kavyanifar et al., (2020)
ARIMA	Better performance for long-term forecast	Computationally expensive	Yu et al. (2015)

grid search was implemented simultaneously. The following step involved the application of parameters inferred from the optimization step and subsequent comparison through error computation strategies being summarized as follows.

The train and the test scores have been evaluated to estimate the model's accuracy. The predictive performance of the ML model was quantified in terms of mean square error (MSE), MAE and coefficient of determination ( $R^2$ ) (Breiman et al. 2017):

$$\text{MAE} = \frac{\sum_{i=1}^n (Y_i - X_i)^2}{n} \quad (4)$$

where  $Y_i$ = predicted value,  $X_i$ = true value and  $n$ = total number of data points

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2 \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \quad (6)$$

where  $n$  is the number of observations,  $\hat{Y}_i$  is the model predicted value,  $Y_i$  is the actual (data) value and  $\bar{Y}$  is the mean

value of waste generation rate. The SSE was also evaluated as a percentage value and was achieved from the RMSE. For both training and testing datasets, the SSE and  $R^2$  were determined and affirmed performance indices. Due to the adjustment of model parameters and associated structure, the training error has been usually found to be lower than the testing error.

### Model forecasting

Using moving average (MA) technique, the forecasted MSWG was determined to understand pertinent data trends using alternate best-fit models. Thus, the predictions represent an average trends of any subset of numbers. The MA is appropriate for the forecasting of long-term trends and can be calculated for any time period. In this article, three MA approaches, namely SMA, WMA and EMA, have been considered for their incorporation into the long-term forecasting prediction models (Droke, 2001).

The SMA is a simple and straightforward technical indicator and is evaluated as the summation of the recent data points in a given dataset and their division with the total number of time periods. Thereby, the SMA is best expressed as:

**Table 3** Conditioned and integrated datasets summary for training and testing studies

Dataset	Dependent variable	Independent variable(s)	Number of data points
MSWG	CPCB (1991–2016)		376
POP		Census Report (1991–2011)	312
Literacy		Census Report (1991–2011)	312
GDDP		Ministry of Statistics and Programme Implementation (1991–2011)	312
HH (count and size)		Census Report (1991–2011)	312
WP		Census Report (1991–2011)	312

$$SMA = \frac{(A_1 + A_2 + \dots + A_n)}{n} \tag{7}$$

where  $A$  = average in period and  $n$  = number of time periods

The WMA involves the assignment of heavy weights to the more current data points due to their greater relevance than the data that corresponds to the distant past. Thus, different weights have been assigned at diverse points of the sample window during WMA. Mathematically, the WMA is the complexity of the data with a fixed weighting function and is expressed as:

$$WMA = \frac{\sum_{i=1}^n W_i * D_i}{\sum_{i=1}^n W_i} \tag{8}$$

where  $W$  = weighting value,  $D$  = data values and  $n$  = number of time periods.

EMA customizes greater weight assignment to the most recent data points. Thereby, EMA is more responsive in the forecasting methodology and involves the following four steps:

*Step 1:* Compute the SMA over a particular time period.

*Step 2:* Calculate the EMA weighting multiplier (also known as the "exponential smoothing") as: thereby, for any time period  $t$ , the smoothed value  $S_t$  is found by computing:

$$S_t = \alpha y_{t-1} + (1 - \alpha)S_{t-1} \quad 0 < \alpha \leq 1, t \geq 3 \tag{9}$$

In the above basic equation of exponential smoothing, the constant or parameter  $\alpha$  is termed as the smoothing constant and  $y$  refers for the original observation.

*Step 3:* Use both smoothing factor and previous EM, to evaluate the current value.

*Step 4:* Assign higher weight to recent data points using the expression:

$$EMA_c = \left( \left[ C_T X \left( \frac{s}{1+N} \right) \right] \right) + EMA_p X \left( \left[ 1 - \left( \frac{s}{1+N} \right) \right] \right) \tag{10}$$

where  $EMA_p$  = EMA for the previous time period,  $EMA_c$  = EMA for the current time period,  $s$  = smoothing,  $N$  = number of time periods and  $C_T$  = Current time period.

## Results and discussion

### Data Characteristics

#### Data Attributes

Table 3 summarizes the datasets prior to the pre-processing phase. Prior to this, the MSWG data from CPCB constituted 408 data points. Also, census and MOSPI data referred to 360 data points. After pre-processing, the CPCB, census and MOSPI data were reduced to 376 and 312, respectively. The MSWG rate of Guwahati city has been increasing steadily since 1991 (Fig. S1) and affirmed a steady but not a substantial enhancement in the waste generation pattern. This is in agreement with the literature reported hypothesis that the POP rise and habitat improvisation together contribute to the proportional enhancement in the MSWG rate. To be specific, the annual volume of MSWG in Guwahati city increased approximately twofold from 1.48 MT (metric ton) in 1991 to 2.13 MT in 2016.

After data conditioning and integration, the socio-economic parameters were reduced to 312 data points for each parameter. Figure S2(a–g) depicts the respective socio-economic factor variation with year along with the MSWG rate. The plots depicting the MSWG rate along with population parameters, namely POP and LP, illustrated a steep increase in the MSWG rate and a positive trend (Fig. S2 (a-b)). In the recent past, due to economic growth, a significant increase in urban POP enhanced the dramatic urban growth and altered land use to thereby increase the MSWG rate (Salmon & Gholamalifard, 2006).

Figure S2(c) and (d), respectively, depicts the MSWG trend along with GDDP and  $GDDP_{per\ capita}$  for Guwahati city.

**Table 4** Correlation matrix data of MSW and socio-economic parameters

	MSWG	POP	LP	GDDP	<i>GDDP</i> <sub>per capita</sub>	HH	<i>HH</i> <sub>Size</sub>	WP
MSWG	1							
POP	0.87	1						
LP	0.82	0.65	1					
GDDP	0.76	0.69	0.68	1				
<i>GDDP</i> <sub>per capita</sub>	0.72	0.64	0.61	0.75	1			
HH	0.85	0.84	0.54	0.64	0.66	1		
<i>HH</i> <sub>Size</sub>	0.81	0.65	0.51	0.61	0.58	0.66	1	
WP	0.74	0.68	0.71	0.75	0.76	0.61	0.56	1

The figures clearly illustrate that for 2002 and 2006–2008, a highly nonlinear variation exists for the GDDP and *GDDP*<sub>per capita</sub>. However, the MSWG rate enhanced steadily during this period but enhanced steeply in 2010–2011. These variations have been due to the pertinent impact of globalization past the Indian economy and its critical influence on the city's income demographics. In this article, the GDDP was considered as an indicator of economic growth due to the fact that it is likely to be sensitive to affirm financial capacity for the payment towards environmental improvement.

The comparative trends of MSWG rate and HH indicators are shown in Fig. S2 (e) and (f), respectively. All figures affirmed that while the HH counts and sizes enhanced steadily, the MSWG also enhanced steadily but with a steep and nonlinear increase in 2010. The correlation conveys that with the total increase in HH counts and *HH*<sub>Size</sub>, the MSWG rate increases. Further, the HH in good condition affirms more waste generation than those with livable and dilapidated conditions. Similarly, HH sizes with 4–8 members direct more waste generation than *HH*<sub>Size</sub> 1–3 and above 9. Figure S2(g) depicts the employment scenarios for Guwahati city along with the MSWG rate. The data confirm more non-WP in the city. These include students, pensioners, dependents, beggars, vagrants, etc. However, for this parameter, a declining trend can be seen since 2001, along with a marginal reduction trend in the MSWG rate. Despite all this, the steep and nonlinear enhancement in MSWG rate existed in 2010. This is due to the strong variation in habitat and lifestyle parameters. The MSWG rate can be analysed to be positively correlated with the employment status statistics. However, more solid waste has been generated by the main WP than the marginal due to better relaxation in the economic constraints for the former case.

### Correlation analysis of the data

The correlation analysis was primarily conducted to determine and rank the socio-economic parameters with high correlation indices (with dependent variables). Thereby, those parameters have been isolated with weak or no correlation with each other. Such an analysis would be helpful to obtain

the most relevant information on the modelling efforts. To do so, a correlation matrix that estimates the correlation index between each variable pair has been determined and analysed.

Table 4 summarizes the correlation matrix for MSWG and mentioned socio-economic parameters. The first column in the table confirms that among all parameters, the POP was found to have the maximum  $r$  of 0.87. Figure S3 shows this positive correlation between POP and MSW to infer that the waste generation rate has been positively influenced with the population. The positive trends of GDDP-MSWG and WP-MSWG have been consistent with the inferences indicated in the relevant prior art (Bandara et al. 2007). Also, LP-MSWG, GDDP-MSWG, *GDDP*<sub>per capita</sub>-MSWG, HH, *HH*<sub>Size</sub> and WP with respect to MSWG have all affirmed positive correlation. Further, GDDP-WP, HH-*HH*<sub>Size</sub>, GDDP-*GDDP*<sub>per capita</sub> and LP-WP parameter pairs have been observed to be significantly correlated with one another. Thereby, to subdue the multicollinearity issues, such model pairs have been avoided in due course of the modelling effort. Scattered plots of POP, LP, GDDP, HH counts and sizes versus the MSW further illustrate such pairing effect (Fig. S3 (a-m)).

Following this, HH, LP and *HH*<sub>Size</sub> can be analysed to possess higher correlation coefficients of 0.85, 0.82 and 0.81, respectively. Moreover, the number of HH and its sizes were found to be positively correlated with MSWG. Thereby, this observation validated the hypothesis that the size of a city influences per HH waste quantities or considered socio-economic factors. Similarly, the WP also exhibited a positive correlation with respect to the MSWG rate. Altogether, the table reflected an overall positive correlation on all the variables with the lowest *GDDP*<sub>per capita</sub> ( $r=0.72$ ).

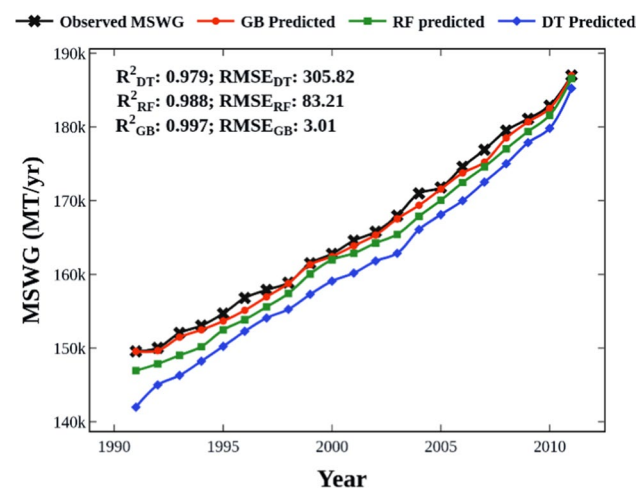
### Modelling and prediction of MSWG

According to the ranking, the socio-economic and demographic factors were first incorporated one by one in the model. Table S1 summarizes the RMSE variation for training and testing cases for various combinations of considered socio-economic and demographic factors in the model. In the case studies, error analysis was obtained using a stepwise

LR-based modelling approach. For the GB, the training RMSE reduced to 198.04 to 154.19 for the mentioned socio-economic parameters but not  $HH_{size}$  and  $WP$ . Corresponding trends in the testing model RMSE indicated a reduction from 111.60 to 104.56 without considering  $HH_{size}$  and  $WP$ . Similar inferences can be obtained for the RMSE trends of RF and DT models and for both training and testing model cases. In summary, the training error reduced with the inclusion of more and more socio-economic parameters. This occurred due to adequate enhancement in the model complexity. However, the testing error ceased to reduce for the testing model case that refers to the model performance evaluation for new data after considering  $HH$  parameter. Thus, it can be inferred that the MWSG prediction rate has been sensitively influenced with POP, LP, GDP,  $GDDP_{per\ capita}$  and  $HH$  and inclusion of other additional socio-economic parameters did not improvise the model performance.

To assess upon the developed model accuracy, a comparison of secondary and simulated data was made for the considered observation period (1991–2011). In general, the obtained results indicate good accuracy of the modelling effort and infer upon its validity for the forecasting of the MWSG. For the Guwahati city, Fig. 3 depicts a comparative trend analysis plot for the secondary data and DT, RF and GB model predicted data. The results indicate that in comparison with the DT, the GB and RF fairly tracked historical developments in the MWSG patterns.

The trained models must be tested with new data (usually unseen data to the model), thereby analysing the capabilities



**Fig. 3** Predicting results depicting GB, RF and DT performance to predict MWSG

**Table 5** Performance comparisons of the models

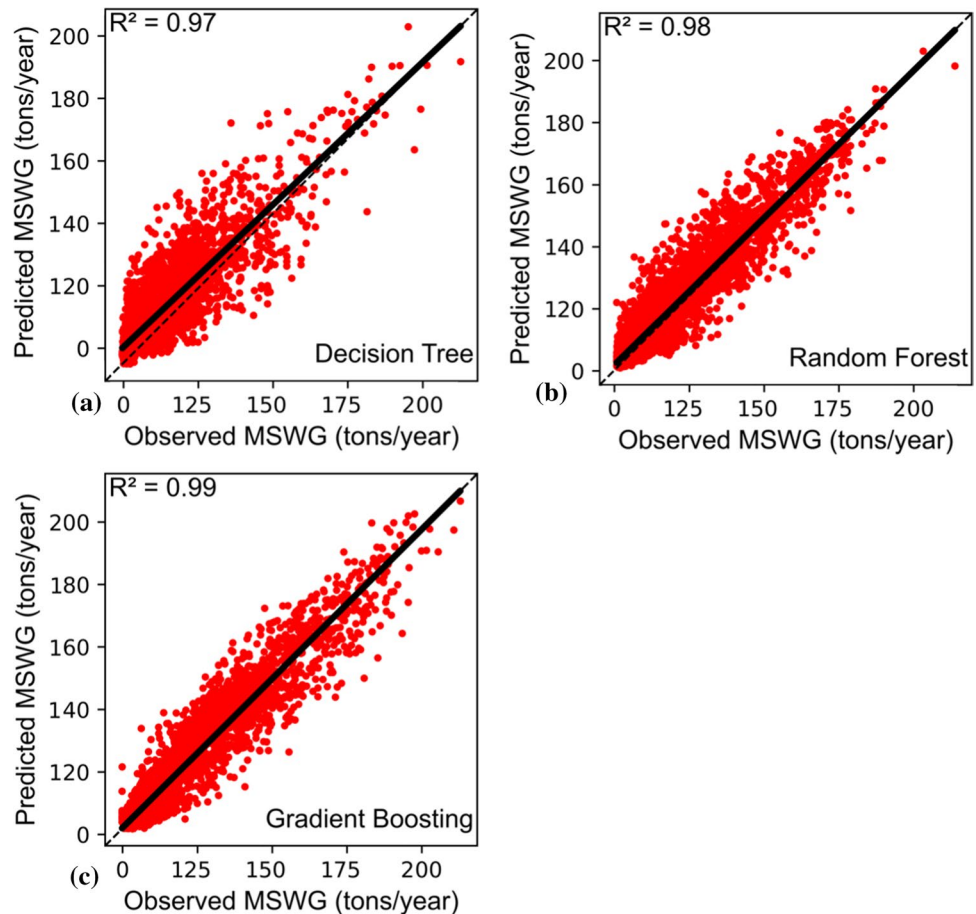
Algorithms	Train score	Test score	R <sup>2</sup>	RMSE	MAE	IOA
DT	0.86	0.81	0.97	325.82	302.20	0.45
RF	0.90	0.85	0.98	83.21	74.84	0.72
GB	0.94	0.91	0.99	3.01	2.86	0.94

of the studied regression algorithms. With this, confidence can be ensured with respect to the model performance. For this purpose, the total dataset was partitioned into two parts (70% deployed for training and 30% for the testing purpose). Thereby, a grid search was performed to infer appropriate parameters for the data with tenfold cross-validation for all the models simultaneously. Important observations and findings of the carried out investigations have been presented as follows.

For the considered DT, RF and GB tree-based models, quality results have been obtained. The training and testing case studies for DT, RF and GB, indicated scores of 1.00, 0.98 and 0.99 and 0.92, 0.95 and 0.98, respectively. The maximum depth of DT, RF and GB during the training phase were 28, 57 and 48, respectively. Among these, the DT produced the best train score. However, the variation between train and test scores has been observed higher for the DT value of 0.6 in comparison with the RF (0.2) and GB (0.1) models. This is due to the fact that since DT is susceptible to handle overfitness associated with the application of greedy algorithms, the optimal tree may not be found for few cases. In other words, the DT model learnt the training data too well and could not be therefore generalized to meet the desired needs. On the contrary, RF and GB have been capable to fix the overfitness issue that exists in the DT and this is confirmed in the observed trends. Figure 3 depicts the MWSG rate prediction by DT, RF and GB models in conjunction with the secondary data of Guwahati city. The trends affirmed closer vicinity of data trends of RF and GB in comparison with the DT. Further, among all models, the GB exhibited a promising performance with an R<sup>2</sup> and model error (calculated using data partitioned ratio) value of 0.997 and 3.01, respectively.

Few statistical indicators have been used to assess upon the performance of the investigated models DT, RF and GB. Table 5 and Fig. 4(a-c) exhibit and illustrate the relevant findings and the fitness charts. As the model with lowest complexity in the tree-based ML algorithms, the DT performed average with a good linear fitness relationship (R<sup>2</sup>=0.97). Based on the DT, the ensemble learning algorithms RF and GB exhibited good performances. In this study, GB performed superior to other models with R<sup>2</sup>=0.99 and as a state-of-the-art ML model, it is good with promising features such as parallelization, handling sparse data and avoiding overfitting. With an R<sup>2</sup> of 0.98, RF also performed well. As a result, GB can be used for further predictions, and other models can be considered as supplementary. Further, the index of agreement (IOA) for the models have been calculated as the ratio between the MSE and the potential error.

**Fig. 4** a-c Scatter plot between actual and predicted values from a DT, b RF and c GB



$$IOA = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (|P_i - \bar{P}| + |O_i - \bar{O}|^2)}, 0 \leq IOA \leq 1 \tag{11}$$

where  $O_i$  is the observation value,  $P_i$  is the forecast value,  $\bar{O}$  is the average observation values and  $\bar{P}$  is the average forecast values.

A perfect match indicates an IOA value of 1, whereas 0 refers to no agreement at all. The RMSE is a measurement of goodness of fit. Also, lower RMSE means a lower error or better fit. In the considered tree-based models, DT exhibited a low train (0.86) and test score (0.81) and affirmed a  $R^2$ , RMSE, MAE and IOA of 0.97, 325.82, 302.20 and 0.45, respectively. GB resulted in high train (0.94) and test score (0.91) and reduced RMSE (3.01) and was therefore able to provide an accurate prediction as discussed in Table 5. Along with the train and test scores, the RMSE (83.21) and MAE (74.84) have been better for RF. However, these are high in comparison with other models. The overall evaluation and comparison of the three ML algorithms confirm that GB is the best-suited ML model.

Figure 5(a-c) depicts accuracy and loss graphs to illustrate the performance of the suggested model. The accuracy and loss have been represented as the y-axes in these plots that considered the sample size (percentage) as the x-axis. The number of training cycles being adopted for the complete dataset or a percentage of training sets has been represented as the x-axis. For a variation in train size, the y1-axis for the accuracy varies as DT (0.65–0.93), RF (0.72–0.97) and GB (0.81–0.98) and the loss curves (y2-axis) reduced for DT (580.01–325.82), RF (482.55–83.21) and GB (138.09–3.01). A deeper introspection into the accuracy graph reveals that in a given set of small sample during the initial stages prompted the curve grew quickly for each model. Thereby, DT, RF and GB affirmed accuracies of 93%, 97% and 98%, respectively. Also, both accuracy curves confirmed an upward tendency with percentage samples. In fact, the first through third sample sizes support the scenario of a significant expansion. The growth rate then gradually declined until it achieved a stationary profile. Similar to this, the loss measurements convey upon the model limitations and hence its subpar performance. The statistics indicate that the loss (error) has been decreasing, and hence, higher model performance existed with steady increase in the fitness. Even though the larger time frame indicated that there

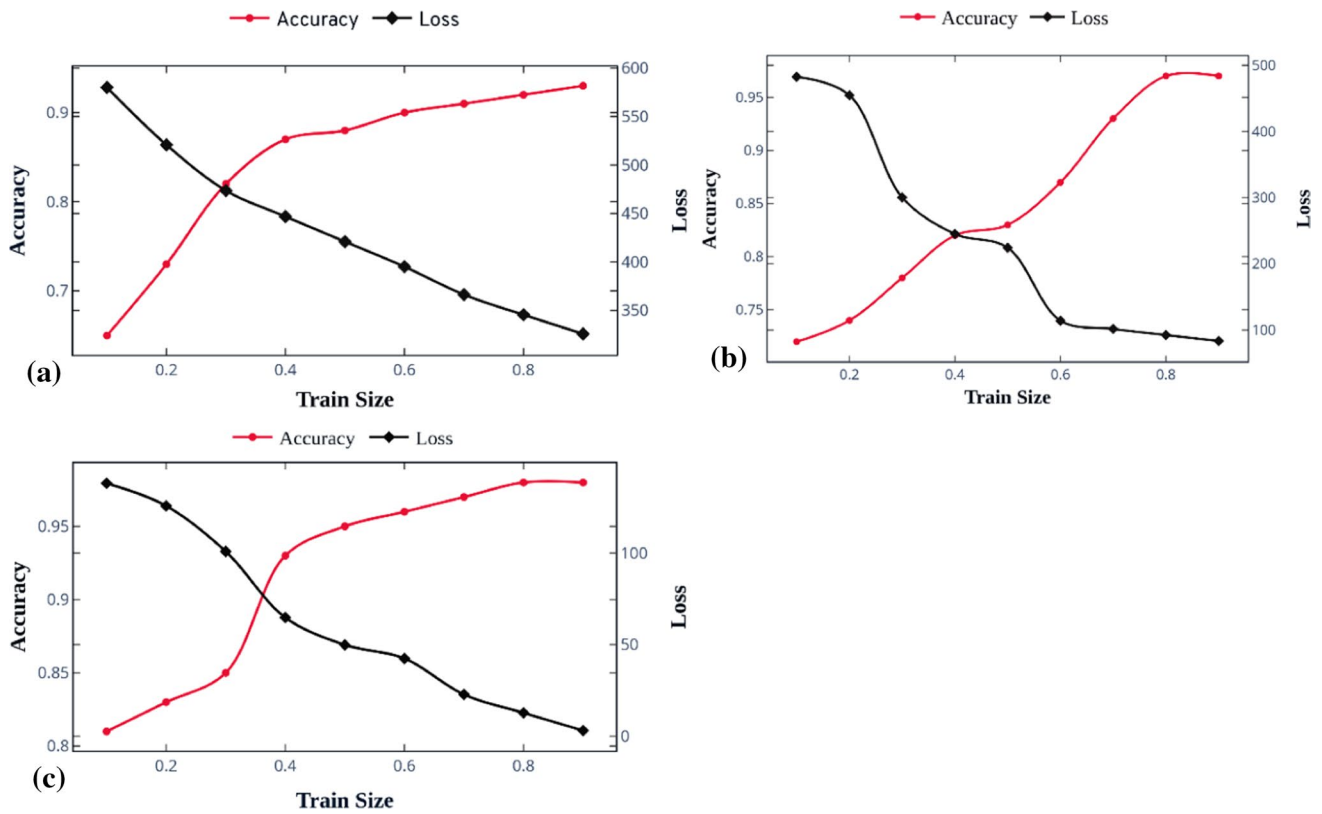


Fig. 5 (a-c) Accuracy and loss curves of DT, RF and GB

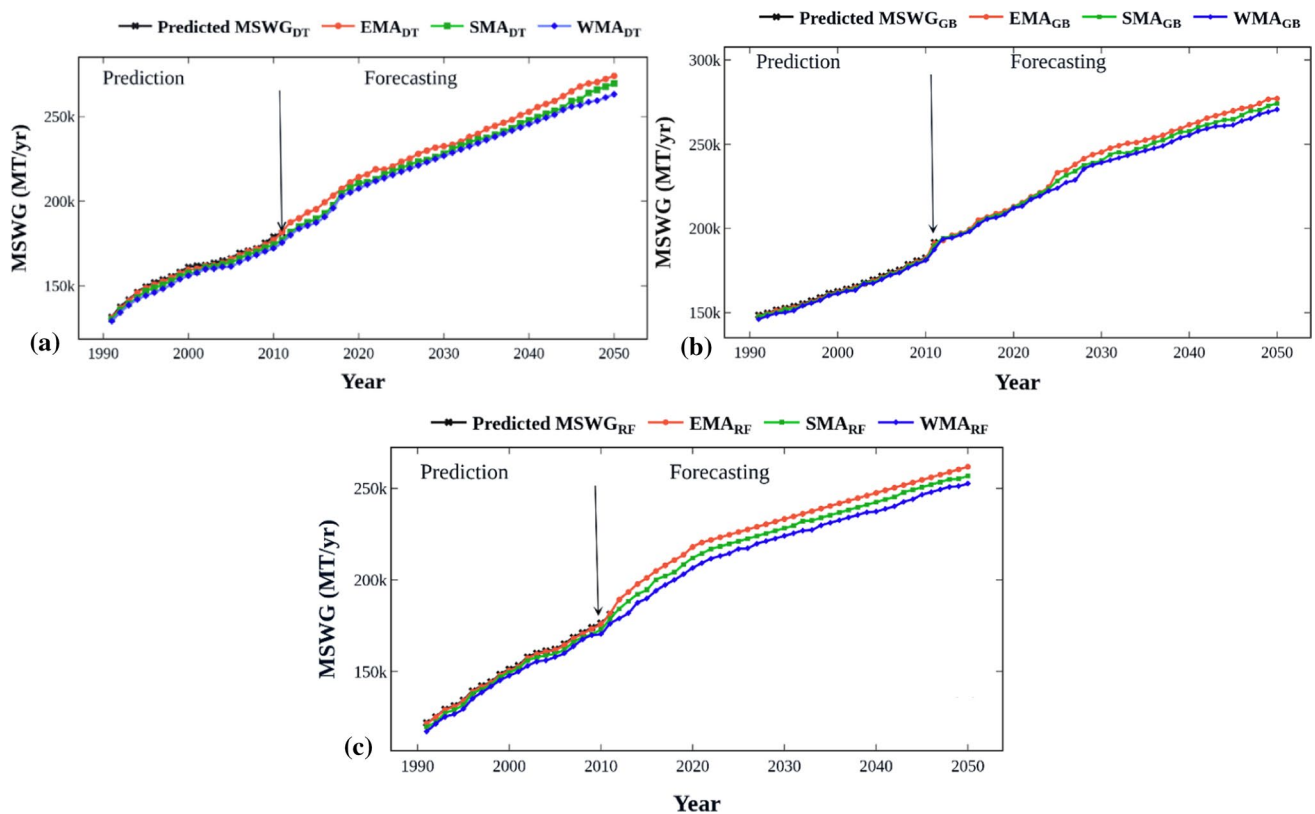
have not been many peaks and valleys, the reduction in loss over time inferred that the model is successfully adapted and performed well.

Involving many attributes, the DT is a nonparametric technique being often applied for the recognition of complex patterns in even smaller datasets. In this study, the DT algorithm reduced the uncertainty associated with the identification of an unknown pattern. Compared to this, the RF algorithms constitutes multiple DTs in which the DT model forms an ensemble with bagging. Hence, the RF model comparatively reduced the data variance and prevented the DT model’s greater dependence on highly influential variables. Lastly, the boosting approach enabled the addition of newer models for their sequential ensembling. In essence, boosting attacks the bias–variance trade-offs through the initiation with a weak model (for instance only with a few splits in the DT model) and subsequent sequential boosting in its performance through the continued effort of the model to build new trees. Thereby, the GB’s performance efficiency enhanced in comparison with the RF and DT.

**MSWG forecasting using statistical models**

MA methodology has been considered in this work for the MSWG forecasting using statistical models. This was

targeted for noise minimization and outliers removal. MA is a statistical tool being adopted for the forecasting and analysis of time series data. The method functions through the acquisition of long-range correlations (Molugaram, & Rao, 2017). For the GB, RF and DT, Fig. 6(a-c), respectively, illustrates the forecasted MSWG rate plots for the Guwahati city in the year range of 2012–2050. Three MA methods, namely EMA, SMA and WMA, have been considered to obtain useful insights and analysis of the forecasted MSWG rate trends for the DT, RF and GB. For GB, in the year range of 2012–2050, the MSWG rate has been forecasted to alter as 194,152–277,035, 192,750–274,105 and 193,580–270,646 MT/year for EMA, SMA and WMA cases, respectively. Similarly, for RF, the corresponding forecasted rates were 189,231–261,759, 184,209–256,749 and 178,973–252,646 MT/year. For DT, it was 187,477–274,012, 181,704–269,549 and 179,934–263,157 MT/year, respectively. For all models, it was observed that the EMA could better forecast the MSWG rate followed with SMA and then WMA. Also, to better evaluate the model accuracy, the MAE and RMSE evaluations were also considered along with the correlation coefficient. Overall, a positive correlation is apparent between waste generation and time. From the figures, the projection statistics convey that by the year 2050, approximately 277,035 MT of waste would be generated in



**Fig. 6** Predicted and forecasted MSWG trends of EMA, SMA and WMA cases for **a** GB **b** RF and **c** DT algorithms

the city and the daily per capita waste generation in the city is expected to increase by approximately 20%.

Figure 7 a-d presents a comparison of the train and test scores for the models. For the GB, the training score of 0.94 was obtained and for the RF and DT, these were 0.90 and 0.86, respectively. Similarly, the test scores for GB, RF and DT were 0.91, 0.90 and 0.81, respectively. Figure 7 b, c and d illustrates a comparison of MAE, RMSE and  $R^2$  for the models, respectively. The RMSE values for DT, RF and GB are 325.82, 83.21 and 3.01, respectively. The MAE results show a similar trend and were 302.20, 74.89 and 2.88, respectively, for DT, RF and GB. The  $R^2$  value for GB, RF and DT were 0.997, 0.988 and 0.979, respectively.

Thus, the results affirmed that the GB performed better than the RF and DT. The RMSE is a measure of the goodness of the fit. Lower MSE indicates lower error and hence better fit. The overall estimate and comparative study of all considered tree-based algorithms inferred that the GB is the best-suited algorithm to predict the MSWG rate. Also, while RF and GB provided best results for all parameters, the RF at times repeated closer data prediction for few sets. On the contrary, the GB did not indicate such issues.

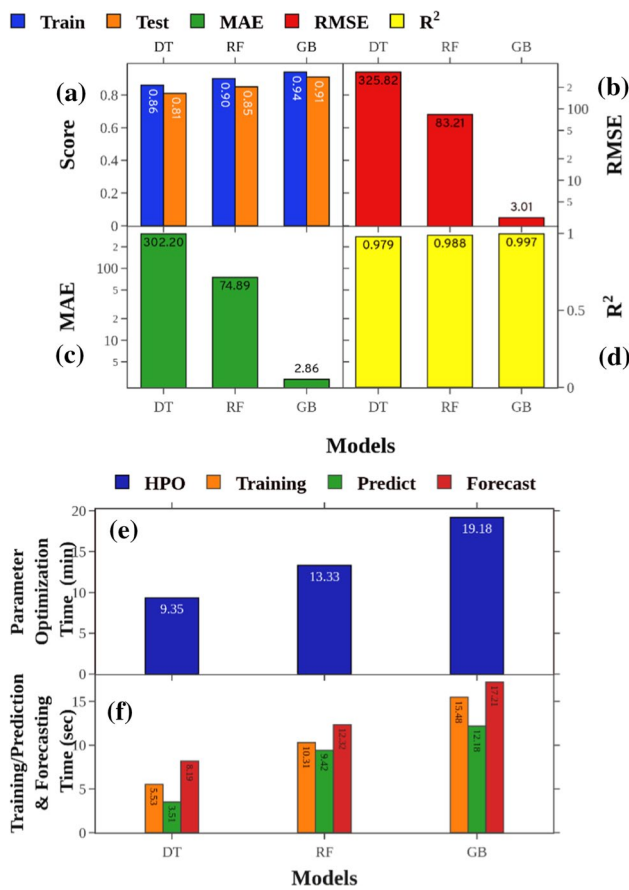
The forecasted data have been compared with the primary data obtained from the GMC. The obtained data referred to the MSWG pertinent rate for 2016–2019. Table 6

summarizes the model error in conjunction with the primary data. For model validation, the MSWG data source was collected from GMC for the period of 2016–2019. It can be easily understood that the GB indicated lowest forecasting errors (1.22 – 1.57%, year ranging 2016–2019) in comparison with the RF and DT. Thus, while RF and DT models confirmed almost similar error values, the RF was better stationed to predict the MSWG rates. The same was also affirmed by its  $R^2$  values. Henceforth, the validation confirms the better performance of GB and RF.

Compared to the literature reported error values, the GB model was efficient to better predict the MSWG rate trends (Johnson et al. 2017). However, the DT model (with forecasting error range of 2.57–2.55%) did not perform well in comparison with the GB and RF. This can be attributed to the ability of the bagging and boosting algorithms and thereby learn complex nonlinear behaviour and subsequent optimization of model parameters. Due to its nonparametric nature, the DT model can build only a single model for a given training data. On the contrary, the bagging (RF) and boosting techniques (GB) can produce multiple trees that efficiently combine their output. Hence, better optimization could be targeted to find the best model.

The ML solution needs to be highly reliable and accurate. To resolve the issues associated with the achievement of





**Fig. 7** (a-f) Comparative depiction of (a) train and test scores, b RMSE, c MAE, d R<sup>2</sup> a HPO and b train, prediction and forecasting time of all the tree-based models

higher accuracy, the appropriate selection of several parameters is mandatory for the subsequent training of a model. To achieve these tasks, HPO is often targeted. The HPO is an important step towards the customized real-time application of the ML solution methodology. Often, HPO involves significant time consumption towards effective training of the model and appropriate selection of correct parameters to achieve best accuracy score of the tested model. This work customized grid search-based determination of best parameters for each considered model. This involved the creation of several runs using different parameters with specified

transformations and estimators. For the transformation step, the parametric combination yielding best results was chosen.

Along with the criticality of the HPO computation time, Fig. 7 (e-f) depicts the computational time demand for each considered model. The PO for less complex computational models is less than 10 min, whereas the PO for more complex high-end ML algorithms is around 20 min. The GB was observed to be computationally expensive with 19.18 min due to associated training complexities. On an interval, the training time was close to 5 s for all three tree-based models. Thus, DT being a computational model with lower complexity demanded only 9.35 min of computational time for the parametric optimization. For both DT and RF, the training time has been lower than 15 s. Comparatively, the GB took 15.48 s for the training process. In the forecasting investigations, the computational times were similar to those being reported for the training scenarios. The assessment of computation time conveys no advantage for the high-end ML algorithms in terms of the accurate outcomes. On the other hand, dependencies of the computational time can be observed on dataset size and available computational power.

During forecasting studies, the socio-economic and demographic factors were incorporated one by one according to their rank in the model. Thereby, POP, LP, GDP, GDDP<sub>percapita</sub> and HH have been observed to improve the model performance considerably. Table 7 summarizes and compares the forecasting performance of the models for their optimal complexities in terms of RMSE and R<sup>2</sup> for all three MA approaches. The optimal GB and DT models for EMA approach indicated an RMSE and R<sup>2</sup> values of 2.12, 4.22 and 0.981, 0.967, respectively. The obtained trends were similar to that being discussed for the training cases.

### Generalization ability and model applications

The tested tree-based models have been inferred for their ability to predict in general the MSWG rates (see Table 5). In this article, higher R<sup>2</sup> values was observed for both RF and GB models during the training and testing phases. Thereby, it can be inferred that the developed ML models were neither overfitted nor were overtrained. In general, modelling encounters two common issues (i) model that involves too well-fit training data and (ii) too low fit testing data. To mitigate the overfitness issues, resampling procedure

**Table 6** Forecasting error data of alternate models

Year	MSWG	GB forecasted MT/year	GB model Error (%)	RF forecasted MT/year	RF Model Error (%)	DT forecasted MT/year	DT Model Error (%)
2016	207,521	204,985	1.22	203,120	2.12	202,172	2.57
2017	209,812	206,722	1.47	206,115	1.76	205,448	2.07
2018	212,825	208,825	1.87	207,543	2.48	206,433	3.00
2019	213,761	210,387	1.57	209,215	2.12	208,300	2.55

**Table 7** Forecasting performance of optimally trained models in terms of RMSE and R<sup>2</sup>

Approach	GB		RF		DT	
	R <sup>2</sup>	RMSE	R <sup>2</sup>	RMSE	R <sup>2</sup>	RMSE
EMA	0.981	2.12	0.972	3.63	0.967	4.22
SMA	0.977	3.83	0.966	5.72	0.956	7.81
WMA	0.951	5.91	0.959	8.21	0.942	11.45

**Table 8** Best data summary for MSWG prediction using ML algorithms

S. No	Authors	ML Models	Dataset	Dataset partition ratio		Prediction performance		Inference
				Training	Testing	R <sup>2</sup>	RMSE	
1	Kannangara et al. (2018)	ANN DT	1553 Data from 2001 to 2014	80	20	0.72 0.54	20 23	ML can produce accurate models for prediction if sufficient socio-economic explanatory variables are given
2	Dissanayaka and Vasanthapriyan (2019)	LR ANN RF	Data from 2009 to 2017	80	20	0.6973 0.9923 0.9608	2706.78 622.086 1500.12	Influential variables such as literacy, expenditures have not been studied to increase the efficiency in waste management
3	Johnson et al. (2017)	GB	232 (Historical data from 2005 to 2011)	80	20	0.88	21.6	Weather conditions as an external feature would further improve the robustness of the model
4	Rathod et al. (2020)	DT	Data collected from 200 regions of Akola city	70	30	0.5044	0.1747	The model suffered overfitting issues
5	Nguyen et al. (2021)	kNN RF	189 (2015–2017)	80	20	0.96 0.97	202.3 201.6	Accuracy of the model was low due to the size and diversity of the data, including factors like lack of data at lower administrative levels
6	This study	DT RF GB	1660, 376 Data from 1991 to 2011 (Census of India, CPCB)	70	30	0.979 0.988 0.997	325.82 83.21 3.01	RF and GB were not overfitted due to high R <sup>2</sup> values during both training and testing The RMSE for the GB model shows significantly better results than RF and DT

was incorporated. This involve tenfold cross-validation and was applied in the modelling effort. However, in this study, during predictive analysis for MSWG, the R<sup>2</sup> values was found to be lower than 0.85 during the testing phase despite indicating very high R<sup>2</sup> values (> 0.90) during the training phase. The comparative study shows that GB exhibits best performance among all the tree-based algorithms (a train and test score of 0.94 and 0.91, respectively). The RMSE, MAE and R<sup>2</sup> among the tree-based algorithms as predicted from GB are 3.01, 2.86 and 0.997, respectively. For the prediction of the MSWG rate, the ANN and DT models affirmed testing phase R<sup>2</sup> values of 0.72 and 0.54, respectively (Kannangara et al. 2018). However, the authors

divided the dataset into 60:40, 70:30, 80:20, 85:15 and 90:10 training to testing ratios. Thereby, each partition ratio generated a total of hundred arbitrary partitions.

Few authors reported good prediction performance of the RF model in terms of R<sup>2</sup> (0.96), RMSE (1500.12) (Dissanayaka & Vasanthapriyan 2019) and R<sup>2</sup> (0.97), RMSE (201.6) (Nguyen et al. 2021). However, the DT models in both cases exhibited lower accuracy and influential socio-economic variables have not been considered. With a smaller data size of 232, Johnson et al., (2017) reported the prediction model accuracy of GB model of R<sup>2</sup> (0.88) and RMSE (21.6). With good performance metrics of RMSE (0.174), the DT model suffered overfitting issues. Thus, an efficient testing phase

modelling effort strategy has not been reported in the literature. Hence, the reported trends affirmed the lower  $r$  values. Contrary to this, this work reported good correlation values even during the testing phase and hence exhibited critical findings. Table 8 conveys a comparison and contrasting of the data. In summary, this study demonstrated the efficient performance of chosen ML algorithms for the MSWG rate prediction using smaller datasets.

### Practical implications

The outcome of this study has numerous practical implications for researchers, policymakers or environmental protection groups. Foremost, the estimated MSWG in any region can be implied to investigate urban metabolism in order to apply and develop circular economy concept. Urban metabolism is broadly used to delineate upon a city's ecosystem in terms of its consumption of material, food, energy and water to support its growth and reproduction, and to generate products and by-products (e.g. GHG, pollutants and waste) (Lu et al. 2021). The amount of waste generated is a critical metric for understanding the urban eco-metabolism and especially in the industrial sector (Zhang et al. 2018). It is also a useful index to apprehend the efficiency of a circular economy system (MacArthur, 2013) that attempts to repurpose certain waste resources for more circular applications.

Second, the predicted MSWG rate can be employed to catalyse an evidence-based policy-making. For instance, it can be used to plan a region's waste management capacity, such as landfill space and existing and expected 3R capacities. In due course of the implementation of this activity, planners frequently encounter issues such as lack or limitations of data. Policymakers can create adequate arrangements such as incentives for recyclers and penalties for polluters. This can be set based on the severity of the problem and waste management capabilities. Subsidies, tax deductions and low-cost land usage have all been widely used in the past to assist recyclers in terms of their enhanced profitability. The predicted result can also be utilized to coordinate inter-regional planning and coordination. One of the most significant pieces of information for such policy research and development activities will be a very good estimation study of MSWG.

Lastly, the MSW generated can be employed in a variety of public engagement activities. The urgency of the problem can be better perceived by the general public by exhibiting the capacity of recycling and landfills, as well as the MSWG. As a consequence, it may be more effective in persuading stakeholders to avoid the not in my back yard (NIMBY) mentality (Bao et al. 2019) and to actively pursue a circular economy (López Ruiz et al. 2020). Such an estimate will provide a longitudinal dataset that conceptualizes the trend of the SWM performance. Thereby, it will likely assist

people to achieve a virtual circle between built environment development and natural environment protection. However, to do so, it shall be performed on a regular basis.

### Summary of the studies

MSWG data at the regional level is critical to develop effective SWM planning. However, many regions and especially emerging ones, lack reliable data. This study used limited, publicly available data and adequate data analytics to predict MSWG in Guwahati city. Seven factors such as POP, GDDP, GDDP<sub>per capita</sub>, HH, HH<sub>Size</sub>, LP and WP were adopted. The results of the data analysis show that these factors can explain the majority of the variations in MSWG, with coefficients of determination ( $R^2$ ) of 0.75 or higher. To describe the variation of MSWG rate, despite their nationwide availability, the socio-economic and demographic parameters from census data were not sufficient. These data can be used to develop waste management strategies such as monitoring the urban metabolism of input (e.g. materials, energy) and output (e.g. waste), and planning waste management facilities (e.g. recycling plants or landfills). This approach in predicting municipal waste can be used as a guide for other regions by considering their own development and environmental issues.

This research also contributes to MSWG estimation methods. Three tree-based models, namely DT, RF and GB, as popular and powerful ML models were chosen and compared based on their strengths and drawbacks. GB and RF delivered promising results in comparison with DT. GB exhibited better performance among the two algorithms ( $R^2$  and RMSE values of 0.997 and 3.11, respectively). The DT approach showed higher errors (3.00% for the year 2018) and low  $R^2$  values (around 0.967–0.942) in comparison with GB and RF models. The GDDP data appear to be critically required to effectively predict the MSWG rate. Moreover, the RMSE for EMA-GB forecast results (2.12) was the best and outperformed the RF (3.63) and DT (4.22) models.

This study has its share of limitations. At first, it is based on a small number of data points, despite the best efforts being made to collect data. Second, while it is permissible to extrapolate data-scarce locations from data-rich ones, individual characteristics of each smaller region have yet to be taken into account in the estimation. Finally, projecting future waste generation from current data is necessarily inexact. As a result, researchers should take a dynamic approach to monitor modelling outcomes, and thereby carry out needful adjustments. Therefore, the most important reason for such estimation projects is to use the results in real-life situations. This research illuminates potential viability of developing tools to aid regional waste planning by gathering, pre-processing, integrating and modelling publicly available data



from varied sources for achieving waste management goals. Future studies can be conducted by considering other socio-economic factors and address environmental and planning policy and management issues which are reliably accessible at the urban municipal level. Henceforth, waste models can be built with better prediction and modelling performances.

## Conclusion

An accurate prediction of MSWG rate is critical for sustainable and efficient MSW management. With this as a primary objective, the widely used ML models have been proved for their efficacy. Despite such thematic prior art, the article for the first time assessed the performance of GB with RF and DT for MSW rate prediction. Also, MA approaches have been deployed for long-term forecasting of the MSWG rate. The modelling effort involved the needful integration of prominent explanatory factors. Indian census programme data, MOSPI and CPCB data further supplemented the annual MSWG datasets.

The model performance has been assessed in terms of five metrics, namely  $r$ ,  $R^2$ , RMSE, MAE and IOA. Several inferences have been deduced from the modelling studies. Firstly, POP, HH counts, LP, GDDP indicated strong positive correlation index. Secondly, among all algorithms, the RF and GB with response train and test scores of 0.90 and 0.94, respectively, performed well and among these two, the GB had a model accuracy of 97%. Thirdly, critical novelty referred to the overfitting issues of the DT structures being mitigated using the grid-based search tenfold cross-validation (a hyper-tuning approach) and forecasting with the MA-based approach. Thereby, this article uniquely contributed to overcome overfitness issues of DT to model nonlinear data and lower learning speed with the HPO for the RF model. Overall the GB model ( $r=0.94$ ,  $R^2=0.99$ ,  $RMSE=3.01$ ,  $MAE=2.86$  and  $IOA=0.94$ ) surpassed the RF model ( $r=0.90$ ,  $R^2=0.98$ ,  $RMSE=83.21$ ,  $MAE=74.84$  and  $IOA=0.72$ ) and DT ( $r=0.82$ ,  $R^2=0.97$ ,  $RMSE=325.82$ ,  $MAE=302.20$  and  $IOA=0.45$ ) model performance. Fourthly, the  $R^2$  and RMSE values for EMA-GB and SMA-GB were 2.12 and 3.83, respectively, and confirmed their statistical distinction. On the contrary, the RMSEs for EMA-DT and WMA-DT were 4.22 and 11.45, respectively, and henceforth affirmed indifference among the models. Thus, the GB exhibited the best performance to predict MSWG rate for Guwahati city.

In this article, the comprehensive major findings are as follows. Firstly, the HPO fine-tuned tree-based ML models had higher test scores and confirmed superior model predictive accuracy values. Secondly, hyper-tuned DT exhibited linear prediction and the RF model performance improved with the enhancement of the learning speed. Thirdly, the

input data have its own limitations due to survey-based restrictions. Hence, for any city, further data in terms of ward-wise information will further complicate model performance and the HPO has been anticipated to meet such stringent needs. Thus, the suggested methodology is generic in nature and can be applied suitability for any city and for much complex input datasets through an appropriate modification of influential parameters. Also, the accuracy graph-based analysis confirmed that for a given set of small samples, each model curve grew quickly during the initial stages, and the DT, RF and GB had accuracies of 93%, 97% and 98%, respectively. Fourthly and finally, the quest for generalized application of ML algorithms for MSWG rate prediction has been complimented to target GB for its prediction speed and accuracy to handle complex datasets with greater ease.

The research confirmed that the estimation of MSWG rate is very crucial for the subsequent system planning of MSW management from both short- and long-term perspectives. Using statistical data in different cities, the presented generic approach can be applied for MSWG rate prediction in any city in the world. To do so, influencing parameters need to be investigated through careful introspection and analysis of the complex data. Best findings of this work can foster the Guwahati metropolitan region to initiate and develop an integrated decentralized community-based SWM approach, enhanced recycling and composting practices (bio-waste) for the subsequent realization of circular bio-economy. According to this study, HH with higher incomes produces more waste. Therefore, the study conveyed the need for the implementation of decentralized community-based SWM strategies through micro-management strategies in the municipality. For this, cooperative working culture between government and other local entities is the need of the hour. With such strategies, MSWM can gradually reduce associated environmental pollution and health risks in the city.

This study showcased the added value of the ML and MA approaches to model historical data and thereby effectively serve the dynamic needs of the solid waste industry. The limited historical datasets at a municipal level are characteristic terms that are to be addressed by MSWG modelling strategies. However, in this article, the ML methods have been to be effective and outperformed well with respect to the classical statistical methods. Moreover, the ML models simultaneously minimized prediction uncertainty and enhanced prediction vigour. Besides these, the ML models can provide insight into decision-making for stakeholders and open pathways for the systematic improvisation of solid waste environmental policies. The critical findings of the article serve as a baseline or benchmark information on the MSW problem in the residential areas of the city. Such useful analysis will assist in the formal understanding of the critical issues of the SWM. Since solid waste assessment



is a continuous process, the periodic assessment strategies will successfully support the associated dynamic challenges.

Future studies need to target more complex survey-based data along with a wider variety of demographic and socio-economic data of MSWG to supplement and verify ML algorithm's competence in terms of updation, retraining and CPU time marginalization. Such research efforts could also target the examination of alternate optimization algorithms to effectively train the tree-based models for MSWG rate prediction. The proposed model could also be strengthened by choosing the most critical input elements from the dataset using dimensionality reduction techniques such as principle component analysis (PCA). In summary, the carried investigations and critical findings provide new hope and horizon for the generic representation and resolution of MSWG rate prediction and forecasting problem using tree-based ML models.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s13762-022-04644-4>.

**Acknowledgements** The authors thankfully acknowledge the Centre for the Environment, Indian Institute of Technology Guwahati, India, Guwahati Municipal Corporation (GMC), Assam, India, Office of the Registrar General & Census Commissioner, India (ORGI), Central Pollution Control Board (CPCB) and Ministry of Statistics Programme and Implementation (MOSPI) for providing requisite data for carrying out this research.

**Authors Contribution** TS was involved in data extraction, data pre-processing, investigation, modelling, writing the original draft, writing—reviewing and editing, and validation. Ramagopal VSU was responsible for conceptualization, supervision, validation, visualization, and writing—reviewing and editing.

**Funding** No external funding was obtained from any external funding agency.

**Data and Material availability** Not applicable.

## Declarations

**Conflict of interests** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Abbas MA, Iqbal M, Tauqeer HM, Turan V, Farhad M (2022) Micro-contaminants in wastewater. In *Environ Micropollut*, Elsevier. <https://doi.org/10.1016/B978-0-323-90555-8.00018>
- Abbasi M, El Hanandeh A (2016) Forecasting municipal solid waste generation using artificial intelligence modelling approaches. *Waste Manag* 56:13–22
- Abbasi M, Abduli MA, Omidvar B, Baghvand A (2013) Forecasting municipal solid waste generation by hybrid support vector machine and partial least square model. *Int J Environ Res* 7(1):27–38
- Abbasi M, Dehban H, Farokhnia A, Roozbahani R, Bahreinimotlagh M (2022) Long-term streamflow prediction using hybrid SVR-ANN based on Bayesian model averaging. *J of Hydrol Engg* 27(11):05022018
- Abdallah M, Talib MA, Feroz S, Nasir Q, Abdalla H, Mahfood B (2020) Artificial intelligence applications in solid waste management: A systematic research review. *Waste Manag* 109:231–246
- Abdoli MA, Falahnezhad M, Behboudian S (2011) Multivariate econometric approach for solid waste generation modeling: Impact of climate factors. *Environ Engg Sci* 28(9):627–633
- Abu Qdais HA, Hamoda MF, Newham J (1997) Analysis of residential solid waste at generation sites. *Waste Manag Res* 15(4):395–406
- Adamović VM, Antanasijević DZ, Ristić M, Perić-Grujić AA, Pocaĳt VV (2018) An optimized artificial neural network model for the prediction of rate of hazardous chemical and healthcare waste generation at the national level. *J of Mat Cycl Waste Manag* 20(3):1736–1750
- Adeogba E, Barty P, O'Dwyer E, Guo M (2019) Waste-to-Resource Transformation: Gradient Boosting Modeling for Organic Fraction Municipal Solid Waste Projection. *ACS Sustain Chem Engg* 7(12):10460–10466
- Ahmed IA, Dutta DK, Baig MRI, Roy SS, Rahman A (2021) Implications of changes in temperature and precipitation on the discharge of Brahmaputra River in the urban watershed of Guwahati. *India Environ Monit Assess* 193(8):1–21
- Ali SA, Ahmad A (2019) Forecasting MSW generation using artificial neural network time series model: a study from metropolitan city. *SN Appl Sci* 1(11):1–16
- Ali RA, Nik Ibrahim NNL, Ghani WAK (2022) Utilization of process network synthesis and machine learning as decision-making tools for municipal solid waste management. *Int J of Environ Sci Tech* 19(3):1985–1996
- Ali Abdoli M, Falah Nezhad M, Salehi Sede R, Behboudian S (2012) Longterm forecasting of solid waste generation by the artificial neural networks. *Environ Prog Sustain Energ* 31(4):628–636
- Araiza-Aguilar JA, Rojas-Valencia MN, Aguilar-Vera RA (2020) Forecast generation model of municipal solid waste using multiple linear regression. *Glob J of Environ Sci Manag* 6(1):1–14
- Azadi S, Karimi-Jashni A (2016) Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: A case study of Fars province. *Iran Waste Manag* 48:14–23
- Bandara NJGJ, Hettiaratchi JPA, Wirasinghe SC, Pilapiiya S (2007) Relation of waste generation and composition to socio-economic factors: A case study. *Environ Monit Assess* 135(1–3):31–39
- Bao Z, Lu W, Chi B, Yuan H, Hao J (2019) Procurement innovation for a circular economy of construction and demolition waste: Lessons learnt from Suzhou, China. *Waste Manag* 99:12–21
- Beigl P, Lebersorger S, Salhofer S (2008) Modelling municipal solid waste generation: A review. *Waste Manag* 28(1):200–214. <https://doi.org/10.1016/j.wasman.2006.12.011>
- Beigl, P, Wassermann, G, Schneider, F, & Salhofer, S. (2004) Forecasting Municipal Solid Waste Generation in Major European Cities 9th International Congress on Environmental Modelling and Software 2: 1–6
- Benítez SO, Lozano-Olvera G, Morelos RA, de Vega CA (2008) Mathematical modeling to predict residential solid waste generation. *Waste Manag* 28:S7–S13
- Bhagat SK, Tung TM, Yaseen ZM (2021) Heavy metal contamination prediction using ensemble model: case study of Bay sedimentation. *Australia J of Hazard Mat* 403:123492
- Bhattacharyya M (2001) Street Food Vending in Guwahati: A Survival Strategy. *J of Soc Sci* 5(1–2):127–131
- Breiman L (1996) Bagging predictors. *Machine learning* 24(2):123–140



- Breiman L, Friedman J H, Olshen R A, Stone C J (2017) Classification and regression trees. Routledge. <https://doi.org/10.1201/9781315139470>
- Chen, T., & Guestrin, C. (2016) Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785–794
- Chung SS (2010) Projecting municipal solid waste: The case of Hong Kong SAR. *Resour, Conserv Recycl* 54(11):759–768
- Daskalopoulos E, Badr O, Probert SD (1998) 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* 24(2):155–166
- De Clercq D, Wen Z, Fei F, Caicedo L, Yuan K, Shang R (2020) Interpretable machine learning for predicting biomethane production in industrial-scale anaerobic co-digestion. *Sci Total Environ* 712:134574
- Debrah JK, Vidal DG, Dinis MAP (2021) Raising awareness on solid waste management through formal education for sustainability: A developing countries evidence review. *Recycl* 6(1):1–21
- Denafas G, Ruzgas T, Martuzevičius D, Shmarin S, Hoffmann M, Mykhaylenko V, Ludwig C (2014) Seasonal variation of municipal solid waste generation and composition in four East European cities. *Resour, Conserv Recycl* 89:22–30
- Dietz EJ, Mendenhall W, Sincich T (1997) A second course in statistics: regression analysis. *J Am Stat Asso* 92(438):797
- Dissanayaka, D. M. S. H., & Vasanthapriyan, S. (2019) Forecast municipal solid waste generation in Sri Lanka. In: 2019 international conference on advancements in computing (ICAC), pp 210–215, IEEE.
- Draper NR, Smith H (1998) Applied regression analysis, vol 326. Wiley
- Droke C (2001) Moving averages simplified, vol 96. Marketplace Books
- Du X, Niu D, Chen Y, Wang X, BiZ, (2022) City classification for municipal solid waste prediction in mainland China based on K-means clustering. *Waste Manag* 144:445–453. <https://doi.org/10.1016/j.wasman.2022.04.024>
- Dyson B, Chang NB (2005) Forecasting municipal solid waste generation in a fast-growing urban region with system dynamics modeling. *Waste Manag* 25(7):669–679
- Even JC, Arberg P, Parker JR, Alter H (1981) Residential waste generation - A case study. *Resour Conserv* 6(3–4):187–201
- Freund R, Wilson W (1998) Regression analysis: statistical modeling of a response variable. Academic Press, San Diego, CA
- Friedman J (2001) Greedy Function Approximation : A Gradient Boosting Machine Author ( s ): Jerome H. Friedman *Annals Stat* 29(5):1189–1232
- Fu HZ, Li ZS, Wang RH (2015) Estimating municipal solid waste generation by different activities and various resident groups in five provinces of China. *Waste Manag* 41:3–11
- Ghanbari F, Kamalan H, Sarraf A (2021) An evolutionary machine learning approach for municipal solid waste generation estimation utilizing socioeconomic components. *Arabian J Geosci* 14(2):1–16
- Ghinea C, Drăgoi EN, Comăniță ED, Gavrilăscu M, Câmpăean T, Curteanu S, Gavrilăscu M (2016) Forecasting municipal solid waste generation using prognostic tools and regression analysis. *J Environ Manag* 182:80–93
- Grazhdani D (2016) Assessing the variables affecting on the rate of solid waste generation and recycling: An empirical analysis in Prespa Park. *Waste Manag* 48:3–13
- Guerrero LA, Maas G, Hogland W (2013) Solid waste management challenges for cities in developing countries. *Waste Manag* 33(1):220–232
- Hannan MA, Mamun AA (2015) A review on technologies and their usage in solid waste monitoring and management systems: Issues and challenges. *Waste Manag* 43:509–523
- Hastie T, Tibshirani R, Friedman JH, Friedman JH (2011) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York, pp 1–758
- Hockett D, Lober DJ, Pilgrim K (1995) Determinants of per capita municipal solid waste generation in the southeastern United States. *J Environ Manag* 45(3):205–217
- Hoque MM, Rahman MTU (2020) Landfill area estimation based on solid waste collection prediction using ANN model and final waste disposal options. *J Clean Prod* 256:120387
- Hu M, van der Voet E, Huppés G (2010) Dynamic Material Flow Analysis for Strategic Construction and Demolition Waste Management in Beijing. *J Ind Ecol* 14(3):440–456
- Huang T, Shi F, Tanikawa H, Fei J, Han J (2013) Materials demand and environmental impact of buildings construction and demolition in China based on dynamic material flow analysis. *Resour, Conserv Recycl* 72:91–101
- Iguyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
- Johnson NE, Janiuk O, Cazap D, Liu L, Starobin D, Dobler G, Ghandehari M (2017) Patterns of waste generation: A gradient boosting model for short-term waste prediction in New York City. *Waste Manag* 62:3–11. <https://doi.org/10.1016/j.wasman.2017.01.037>
- Kamaraj M, Ramachandran KK, Aravind J (2020) Biohydrogen production from waste materials: benefits and challenges. *Int J Environ Sci Tech* 17(1):559–576
- Kannangara M, Dua R, Ahmadi L, Bensebaa F (2018) Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. *Waste Manag* 74:3–15
- Katsamaki A, Willems S, Diamadopoulos E (1998) Time Series Analysis of Municipal Solid Waste Generation Rates. *J Environ Engg* 124(2):178–183
- Kavyanifar B, Tavakoli B, Torkaman J, Taheri AM, Orkomi AA (2020) Coastal solid waste prediction by applying machine learning approaches (Case study: Noor, mazandaran province, iran). *Caspian J Environ Sci* 18(3):227–236
- Khalil, M., Iqbal, M., Turan, V., Tauqeer, H. M., Farhad, M., Ahmed, A., & Yasin, S. (2022) Household chemicals and their impact. In: *Environ Micropollutants*, Elsevier, pp 201–232
- Kolekar KA, Hazra T, Chakrabarty SN (2016) A Review on Prediction of Municipal Solid Waste Generation Models. *Procedia Environ Sci* 35:238–244
- Kontokosta CE, Hong B, Johnson NE, Starobin D (2018) Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. *Comp, Environ Urban Syst* 70:151–162
- Kumar A, Samadder SR (2020) Performance evaluation of anaerobic digestion technology for energy recovery from organic fraction of municipal solid waste: A review. *Energy* 197:117253
- Kumar A, Samadder SR, Kumar N, Singh C (2018) Estimation of the generation rate of different types of plastic wastes and possible revenue recovery from informal recycling. *Waste Manag* 79:781–790
- Lin K, Zhao Y, Tian L, Zhao C, Zhang M, Zhou T (2021) Estimation of municipal solid waste amount based on one-dimension convolutional neural network and long short-term memory with attention



- mechanism model: a case study of Shanghai. *Sci Total Environ* 791:148088. <https://doi.org/10.1016/j.scitotenv.2021.148088>
- Lohani BN, Hartono DM (1985) Estimation of solid waste generation rates in the city of Bandung. *Indonesia Waste Manag Res* 3(2):103–117
- Lu JW, Zhang S, Hai J, Lei M (2017) Status and perspectives of municipal solid waste incineration in China: A comparison with developed regions. *Waste Manag* 69:170–186
- Lu W, Lou J, Webster C, Xue F, Bao Z, Chi B (2021) Estimating construction waste generation in the Greater Bay Area, China using machine learning. *Waste Manag* 134:78–88
- Lv J, Dong H, Geng Y, Li H (2020) Optimization of recyclable MSW recycling network: A Chinese case of Shanghai. *Waste Manag* 102:763–772
- Ma QX, Shan H, Zhang CM, Zhang HL, Li GM, Yang RM, Chen JM (2020) Decontamination of face masks with steam for mask reuse in fighting the pandemic COVID-19: experimental supports. *J Med Virol* 92(10):1971–1974
- MacArthur E Foundation. (2013) Towards the circular economy. *J of Ind Ecol* 23–44
- Michael Y, Helman D, Glickman O, Gabay D, Brenner S, Lensky IM (2021) Forecasting fire risk with machine learning and dynamic information derived from satellite vegetation index time-series. *Sci Total Environ* 764:142844
- Miezah K, Obiri-Danso K, Kádár Z, Fei-Baffoe B, Mensah MY (2015) Municipal solid waste characterization and quantification as a measure towards effective waste management in Ghana. *Waste Manag* 46:15–27
- Molina-Gómez NI, Díaz-Arévalo JL, López-Jiménez PA (2021) Air quality and urban sustainable development: the application of machine learning tools. *Intl J Environ Sci and Tech* 18(4):1029–1046
- Molugaram K, Rao GS, Shah A, Davergave N (2017) Statistical techniques for transportation engineering. Butterworth-Heinemann
- Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011) Overview of use of decision tree algorithms in machine learning. *Proceedings - 2011 IEEE Control and System Graduate Research Colloquium ICSGRC 2011*: 37–42
- Navarro-Esbrí J, Diamadopoulos E, Ginestar D (2002) Time series analysis and forecasting techniques for municipal solid waste management. *Resour, Conserv Recycl* 35(3):201–214
- Nguyen XC, Nguyen TTH, La DD, Kumar G, Rene ER, Nguyen DD, Nguyen VK (2021) Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam. *Resour, Conserv Recycl* 167(2020):105381
- Nisbet, R., Elder, J., & Miner, G. D. (2009) *Handbook of statistical analysis and data mining applications*. Academic press
- Noori R, Abdoli MA, Ameri Ghasrodashti A, Jalili Ghazizade M (2009a) Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: A case study of Mashhad. *Environ Progress Sustain Eng* 28(2):249–258
- Noori R, Abdoli MA, Farrokhnia A, Ghaemi A (2009b) Solid waste generation predicting by hybrid of artificial neural network and wavelet transform. *J Environ Stud* 35(49):25–30
- Noufal M, Yuanyuan L, Maalla Z, Adipah S (2020) (2020) Determinants of household solid waste generation and composition in Homs city. *J of Environ and Public Health, Syria*
- Palacio JCEscobar, Santos JJCS, Renó MLG, Júnior JCF, Carvalho M, Reyes Arnaldo Martín Martínez, Orozco DJR (2019) Municipal Solid Waste Management and Energy Recovery. In: Al-Bahadly Ibrahim H (ed) *Energy Conversion - Current Technologies and Future Trends*. IntechOpen. <https://doi.org/10.5772/intechopen.79235>
- Pao HT, Chih YY (2006) Comparison of TSCS regression and neural network models for panel data forecasting: Debt policy. *Neural Comp Appl* 15(2):117–123
- Pinka Sankoh F, Yan X, Conteh MH (2012) A Situational Assessment Of Socioeconomic Factors Affecting Solid Waste Generation And Composition In Freetown. *Sierra Leone J Environ Prot* 03(07):563–568
- Quinlan JR (1999) Simplifying decision trees. *Int J Human Comp Stud* 51(2):497–510
- Rathod T, Hudnurkar M, Ambekar S (2020) Use of Machine Learning in Predicting the Generation of Solid Waste. *PalArch's J Arch Egypt/egyptology* 17(6):4323–4335
- Rimaityte I, Ruzgas T, Denafas G, Račys V, Martuzevicius D (2012) Application and evaluation of forecasting methods for municipal solid waste generation in an Eastern-European city. *Waste Manag Res* 30(1):89–98
- Ruiz LAL, Ramón XR, Domingo SG (2020) The circular economy in the construction and demolition waste sector—A review and an integrative model approach. *J Clean Prod* 248:119238
- Salmon Mahini A, Gholamalifard M (2006) Siting MSW landfills with a weighted linear combination methodology in a GIS environment. *International J Environ Sci Tech* 3(4):435–445
- Sammur C, Webb GI (eds) (2017) *Encyclopedia of machine learning*. Springer Science & Business Media
- Schiller F, Raffield T, Angus A, Herben M, Young PJ, Longhurst PJ, Pollard SJT (2010) Hidden flows and waste processing—an analysis of illustrative futures. *Environ Tech* 31(14):1507–1516
- Shahabi H, Keihanfard S, Ahmad BB, Amiri MJT (2014) Evaluating Boolean, AHP and WLC methods for the selection of waste landfill sites using GIS and satellite images. *Environ Earth Sci* 71(9):4221–4233
- Shamshiry, E., Mokhtar, M. Bin, & Abdulai, A. (2014) Comparison of Artificial Neural Network (ANN) and Multiple Regression Analysis for Predicting the amount of Solid Waste Generation in a Tourist and Tropical Area—Langkawi Island. *Environ Engg (BCEE)*, 161–166
- Shi T, Horvath S (2006) Unsupervised learning with random forest predictors. *J of Comp and Graph Stat* 15(1):118–138
- Solano Meza JK, Orjuela Yepes D, Rodrigo-Illarri J, Cassiraga E (2019) Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks. *Heliyon* 5(11):e02810
- Sun N, Chungpaibulpatana S (2017) Development of an Appropriate Model for Forecasting Municipal Solid Waste Generation in Bangkok. *Energy Procedia* 138:907–912
- Sun L, Fujii M, Tasaki T, Dong H, Ohnishi S (2018) Improving waste to energy rate by promoting an integrated municipal solid-waste management system. *Resour, Conserv Recycl* 136:289–296
- Suthar S, Singh P (2015) Household solid waste generation and composition in different family size and socio-economic groups: A case study. *Sustain Cities Soc* 14(1):56–63
- Sutton CD (2005) Classification and regression trees, bagging, and boosting. *Handb Stat* 24:303–329. [https://doi.org/10.1016/s0169-7161\(04\)24011-1](https://doi.org/10.1016/s0169-7161(04)24011-1)
- Tauqeer HM, Turan V, Iqbal M (2022a) Production of safer vegetables from heavy metals contaminated soils: the current situation, concerns associated with human health and novel management strategies. *Advances in Bioremediation and Phytoremediation for Sustainable Soil Management*. Springer, Cham, pp 301–312
- Tauqeer HM, Turan V, Farhad M, Iqbal M (2022b) Sustainable agriculture and plant production by virtue of biochar in the era of climate



- change. *Managing Plant Production Under Changing Environment*. Springer, Singapore, pp 21–42
- Tchobanoglous, G., & Kreith, F. (1994) *Integrated Solid Waste Management: Engineering Principles and Management Issues*. In: *Bulletin of Science, Technology & Society*
- Wisniewski M, Rawlings JO (1990) *Applied Regression Analysis: A Research Tool*. *J Oper Res Soc* 41(8):782
- World Bank Annual Reports (2018) Retrieved from <https://openknowledge.worldbank.org/handle/10986/30326/9781464812965.pdf>
- Xiao S, Dong H, Geng Y, Tian X, Liu C, Li H (2020) Policy impacts on Municipal Solid Waste management in Shanghai: A system dynamics model analysis. *J Clean Prod* 262:121366
- Xiao Q, Liang F, Ning M, Zhang Q, Bi J, Kebin He Yu, Lei Y L (2021) The long-term trend of PM2.5-related mortality in China: The effects of source data selection. *Chemosphere* 263:127894. <https://doi.org/10.1016/j.chemosphere.2020.127894>
- Younes MK, Nopiah ZM, Basri NA, Basri H, Abushammala MF, KNA M, (2015) Solid waste forecasting using modified ANFIS modeling. *J Air Waste Manag Asso* 65(10):1229–1238
- Yu Y, Huang QP, Ma XY, He JH (2015) Prediction of urban waste disposal based on ARIMA model. *Appl Mech Mater* 768:707–713. <https://doi.org/10.4028/www.scientific.net/AMM.768.707>
- Jalili Ghazi Zade, M., & Noori, R. (2008) Prediction of municipal solid waste generation by use of artificial neural network: A case study of Mashhad. *Int J of Environ Res* 2(1): 13–22
- Zhang Y, Lu W, Wing-Yan Tam V, Feng Y (2018) From urban metabolism to industrial ecosystem metabolism: a study of construction in Shanghai from 2004 to 2014. *J Clean Prod* 202:428–438
- Zhu L, Atikur Rahman KM (2020) Impact of purchasing power parity and consumption expenditure rise on urban solid waste generation in China. *Int J Asian Soc Sci* 10(9):458–470

## Reference to datasets

- [https://censusindia.gov.in/digitallibrary/Archive\\_home.aspx](https://censusindia.gov.in/digitallibrary/Archive_home.aspx)  
[https://cpcb.nic.in/uploads/MSW/Waste\\_generation\\_Composition.pdf](https://cpcb.nic.in/uploads/MSW/Waste_generation_Composition.pdf)  
<http://microdata.gov.in/nada43/index.php/catalog>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.