SN Computer Science (2022) 3:184
https://doi.org/10.1007/s42979-022-01068-2

ORIGINAL RESEARCH

Prediction Tool on Fine Particle Pollutants and Air Quality for Environmental Engineering

Aparna S. Varde[1,2] · Abidha Pandey[3] · Xu Du[4]

Received: 25 May 2021 / Accepted: 19 December 2021 / Published online: 7 March 2022

Abstract

This article focuses on the research, design and implementation of a prediction tool for air quality to estimate pollutant concentrations, contributing to environmental engineering. It addresses prediction of fine particle air pollutants of diameter less than 2.5 μm (particulate matter 2.5), their concentration being substantially influenced by urban traffic. We collect worldwide multicity data from health-related public sources on which mining is performed using classical data mining/machine learning paradigms: association rules, clustering and classification. Challenges include adapting appropriate techniques based on data, and capturing subtle domain-specific aspects. The prediction tool is built using knowledge discovered by mining, leveraging health standards, catering to novice, intermediate and expert users. The prediction output is accurate, efficient, interpretable and useful as evident from our experiments. The tool is helpful for urban decision support. This work is beneficial in developing software systems such as intelligent tutors, mobile device apps and smart city tools. It contributes to smart environment, mobility and living, making a positive impact on smart cities and sustainability. In this work, we claim that classical computational paradigms in their fundamental form can be adapted to solve environmental engineering problems, with easy comprehension, as per the Occam's razor principle that advocates simplicity. This article constitutes applied research: using computational techniques to solve domain-specific problems. Future work includes exploring models in deep learning such as CNN and Bi-LSTM, and considering different types of pollutants as well as other sources besides multicity traffic data, to conduct further studies. This would address additional challenges with enhancements.

Keywords Applied research · Decision support · Environmental engineering · Health informatics · Machine learning · Urban studies

Introduction

The recent progress with smart cities and sustainability mandates a fundamental concern toward the environment. A major facet of the environment is air quality since it influences human health, flora and fauna, quality of cities and other aspects [1, 2]. In this work, we address the issue of air pollution. Our particular focus is on traffic data and the related impacts on the environment. To mitigate the effects of air pollution, it is useful to analyze its causes. It is helpful to make the public aware of the estimated extent of pollution in various regions, so they can make decisions about their residential lifestyles and relocation, and can take measures to reduce pollution. Likewise, it is important to provide urban agencies with detailed knowledge about the causes and effects of pollution in the respective areas, for supporting their decision-making on legislative policies that can minimize pollution and counterbalance its harmful effects. Moreover, it is good to offer useful data and analysis on air quality and its pertinent factors to environmental engineers to conduct further studies in related fields to achieve clean air goals. Given this motivation, the objectives of our work entail the research and development of a novel prediction

✉ Aparna S. Varde
vardea@montclair.edu

[1] Dept. of Computer Science, Environmental Science and Management PhD Program, Montclair State University, Montclair, NJ, USA

[2] Visiting Researcher at Max Planck Institute for Informatics, Saarbrücken, Germany

[3] Department of Computer Science, Montclair State University, Montclair, NJ, USA

[4] Department of Earth and Environmental Science, Environmental Science and Management PhD, Montclair State University, Montclair, NJ, USA

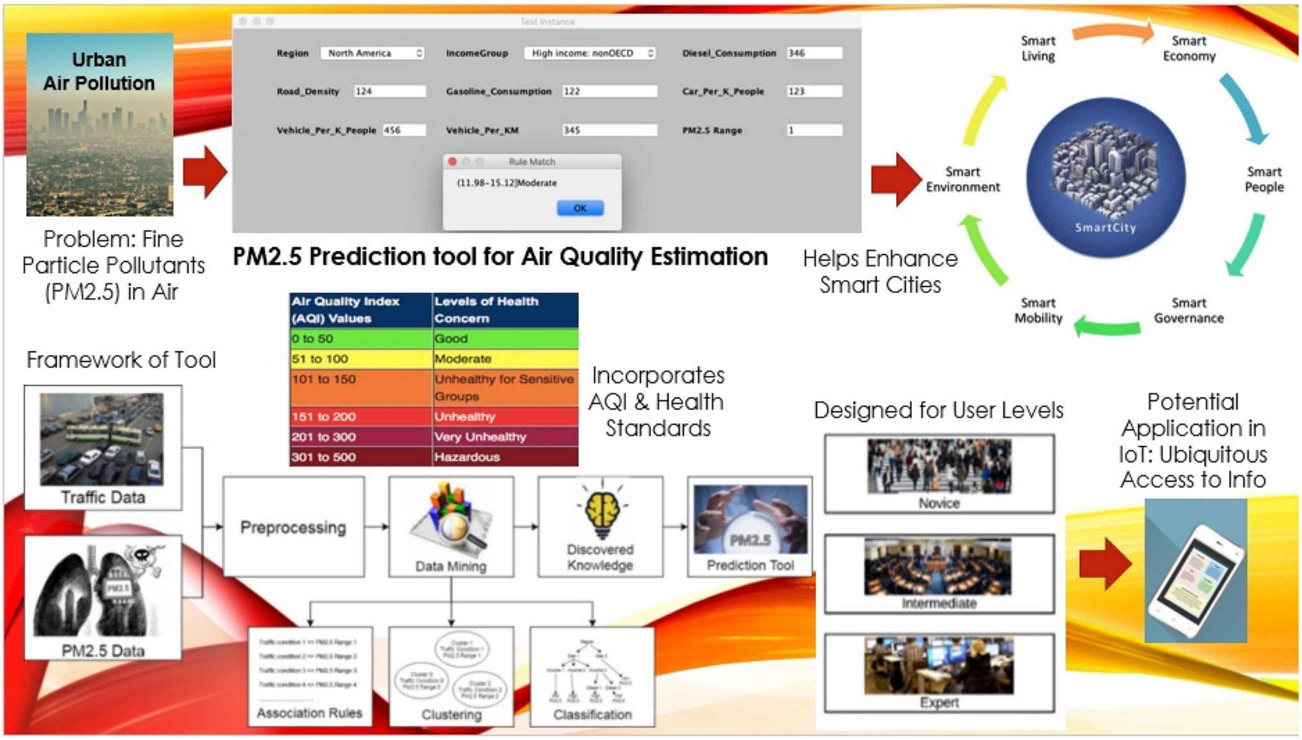
tool for fine particle air pollutants (PM2.5, i.e., particulate matter 2.5 with diameter less than 2.5 µm) based on the analysis of related Web data and incorporating health standards accepted worldwide, such that it would help various users: the common public, urban agencies as well as environmental engineers. Figure 1 is a snapshot of the welcome page of this tool that serves as its graphical abstract. The focus of this work is on fine particle air pollutants and traffic data, given that these are considered to be the most dangerous pollutants [3] and that traffic is their major source [4, 5]. The methods of this study are applicable to other pollutants and data sources as well. In air quality studies, high concentrations of PM2.5 can cause acute long-term damage to health, e.g., cardiovascular and respiratory diseases. This is proved by multiple studies [3, 6, 7]. The studies indicate that these particles have the ability to penetrate the defense mechanism of the human respiratory system and move into the lungs. The finer the particles, the deeper they penetrate.

Figure 2 presents an overview of this process as adapted from the literature [8] and depicted in introductory screens of our prediction tool. Traffic is a significant source of PM2.5 pollutants. An interesting study [5] shows that traffic generates PM2.5 with tailpipe and non-tailpipe emissions, i.e., secondary pollutants, road dust, brake/tire particles etc. Researchers accept that traffic volume, congestion and other related conditions can influence PM2.5, especially in urban areas with higher traffic [4, 9]. Due to negative effects of PM2.5 on human health, environmental engineers find it useful to follow regulations and standards for health. In the USA, the Clean Air Act requires the Environmental Protection Agency (EPA) to standardize national ambient PM2.5 to a certain level [10]. Recent PM2.5 standards have been revised in 2012. Health standards state that PM2.5 amounts should be less than 12 µg/m^3 for a 24-h period. Thus, the air quality index (AQI) system is adjusted to fit new standards and is made available via the Web [1] for public access.

Table 1 shows how the AQI system defines the effects of PM2.5 at different concentration ranges [1]. These ranges are based on average values per 24 h. EPA sets these National Ambient Air Quality Standards (NAAQS) using medical data [10]. The index values are numerical ranges for easier comprehension while revised breakpoints pertain to actual concentrations.

These serve as standards in our work, widely established via the Web, and typically accepted worldwide. We now define the problem addressed herewith. The focus of our problem is the development of a prediction tool for PM2.5 air pollutants, the first of its kind (to the best of our knowledge). This problem is divided into two parts:

1. Analyze PM2.5 data from the Web on traffic and related aspects to study causes of PM2.5 pollutants, and estimate air quality based on health standards.



Region: North America
IncomeGroup: High income: nonOECD
Diesel_Consumption: 346
Road_Density: 124
Gasoline_Consumption: 122
Car_Per_K_People: 123
Vehicle_Per_K_People: 456
Vehicle_Per_KM: 345
PM2.5 Range: 1

(11.98–15.12]Moderate

Air Quality Index (AQI) Values	Levels of Health Concern
0 to 50	Good
51 to 100	Moderate
101 to 150	Unhealthy for Sensitive Groups
151 to 200	Unhealthy
201 to 300	Very Unhealthy
301 to 500	Hazardous

Fig. 1 Prediction tool welcome page snapshot (graphical abstract)

2. Develop an interactive prediction tool for air quality estimation considering given parameters, catering to novice, intermediate and expert users.

The time horizon of this tool is the same as that of the AQI-based health standards. Its space horizon is worldwide due to multicity data analysis. In addressing the sub-problem of part 1, we obtain traffic and related data from worldwide repositories, stored in publicly available Web sources. Based on the data, we aim to answer research questions such as:

1 a) Which parameters are significant in causing the presence of PM2.5 in air?

1 b) How can one parameter in the traffic data potentially affect another?

1 c) What impact does the predicted air quality have on health standards?

1 d) What is the similarity between various regions in the multicity traffic data with respect to air pollution?

The findings from part 1 are significant in the sub-problem of part 2, i.e., detailed design and implementation of the prediction tool. An important feature of this tool is that it should be easily usable by various levels of users. This includes novice users such as urban residents who form the common public; intermediate users such as urban planners: management agencies, lawmakers; and expert users such as urban researchers, namely earth scientists, environmental engineers, professionals in data mining, machine learning and Web data management. Thus, the levels are based on the extent of knowledge of the users. Given this, we aim to address three main issues in this sub-problem:

- *Convenient access* The tool must be easily available to all levels of users, so they can get its benefits. It is important to determine user levels, as users may not know in advance whether they are novice, intermediate or expert. Moreover, not all users have unlimited access to the Internet and mobile devices. We should incorporate their needs.
- *Convenient user input* To use the tool, users must input some data. However, not all users have the required data and some of them may not understand how to use the data. Hence, the tool needs to provide some guidance on entering input data for estimation of air quality.
- *Convenient result interpretation* For users without much professional knowledge, the impact of the estimated output may be hard to understand. Thus, we need to make the estimation interpretable from their standpoint. On the other hand, some users may be interested in the detailed analysis that led to the results. The outputs of the estimation should be conveyed to them in more detail.

Considering these two sub-problems, we proceed to outline the rest of this applied research article. While PM2.5 analysis has been conducted by other researchers (and will be surveyed in our related work section), this article actually focuses on deploying classical data mining techniques in their simplest form for this process to facilitate comprehension in line with the Occam's razor principle [11] of preferring simpler solutions to problems. We also develop a novel user-friendly tool for air quality estimation based on different targeted user levels. Accordingly, the main contributions of this applied research article are as follows:

1. Investigating PM2.5 air pollutants by analysis of multicity traffic data to estimate air quality via globally accepted health standards with the AQI-based time horizon and worldwide space horizon.
2. Proving that classical data mining approaches are adaptable in their fundamental form to solve interesting problems in environmental engineering for easy comprehension by users, as per the Occam's razor principle.
3. Developing a user-friendly prediction tool for air quality estimation catering to novice, intermediate and expert users.
4. Outlining applications of this prediction tool in the development of relevant software systems as well as decision support in urban studies with a broader impact on smart cities.

The layout of the rest of this article is as follows. "Related work" presents a literature survey of related work in the area. Section "Methods and materials" describes our methods and materials in involved in the research and development pertaining to this prediction tool. "Evaluation and results" outlines the tool evaluation and results, including experimentation, system demonstration (demo) and user surveys. "Applications and discussion" provides applications and discussion with respect to development of software systems and decision support for urban users. "Conclusions, discussion and future work" states the conclusions and future work.

Related Work

Data Mining and Environmental Engineering

Environmental engineers and urban agencies seek to mitigate the negative effects of urbanization [12]. This extends to the corporate world, e.g., some companies seek cloud computing solutions for a greener environment [13]. Driving forces of urbanization are multiple according to related research [14], e.g., higher employment rate and better social service. Since multiple factors are involved, there is need for pertinent analysis and prediction to enhance urban planning. Data management and mining along with Web analysis and machine learning are used in numerous multidisciplinary

Fig. 2 Damage caused by PM2.5 air pollutants

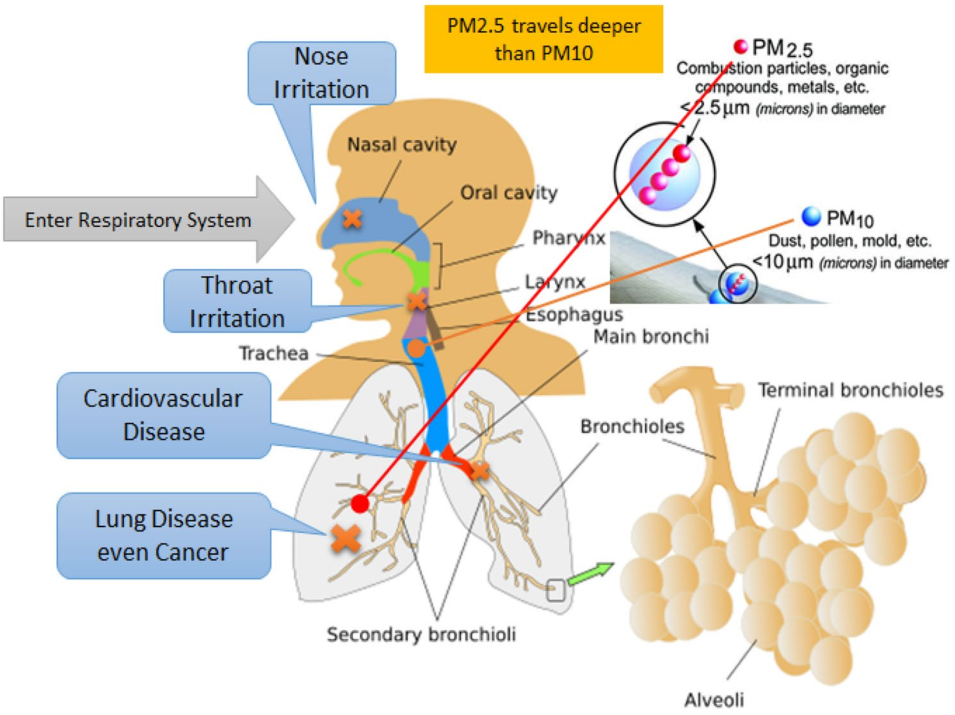


Table 1 AQI health standards for pm2.5 (revised breakpoints in µg/m^3, 24-h. average)

AQI categories	Index values	Revised breakpoints
Good	0–50	0.0–12.0
Moderate	51–100	12.1–35.4
Unhealthy for sensitive groups	101–150	35.5–55.5
Unhealthy	151–200	55.5–150.4
Very unhealthy	201–300	150.5–250. 4
Hazardous	301–400 401–500	250.5–350.4 350.5–500

studies, including those in scientific domains [15–20]. There is research on using the medical markup language MML, a Web-based standard for analysis of healthcare data [21], including that stored on the cloud. Data analytics with respect to urban sprawl and related issues is conducted [12, 22, 23], e.g., sprawl-causing factors, spatial data analysis on GIS (geographic information system) data, relocation based on multicity data analysis, etc. Data mining on social media and structured data sources is executed for analyzing pollution, incorporating public feedback [24]. There is research on personal protective equipment for SARS-Cov-2 [25] that can be crucial in the light of the recent COVID-19 pandemic. Pollution can be relevant here too. There are methods for accessing pollutant concentrations. Multiple linear regression models utilize fine temporal scale data with hourly concentrations to offer a fairly generic estimate of air pollutants [26]. Land use regression models that gauge relationships between land use change and air quality change prove functional for assessing their impact on the regional air quality [27]. Some studies have been conducted on various facets of PM2.5 pollutants [3–7, 9] as discussed in our introduction section.

Our work in this article is orthogonal to such studies. To the best of our knowledge, none of these works provides an interactive user-friendly prediction tool for air quality estimation, especially with respect to fine particle air pollutants and their health impacts, catering to various user levels. We address these issues in our work: targeting novice, intermediate and expert users; and discussing applications in software development tools, and urban decision support.

Analysis of Fine Particle Pollutants

Yang et al. [28] conduct sequential pattern mining with generalized sequential pattern (GSP) algorithm on the PM2.5 concentrations for three significant metropolitan areas of China around the four seasons in 2015. They find that pollution distribution is more substantial in autumn and winter due to the influences of both the monsoon system and winter indoor heating policy of China. This research focuses on the PM2.5 spatial distributions among the urban areas, which mainly contributes to the pollution transfer research, but not the actual prediction of PM2.5 concentration. Bai et al. [29] utilize the random forest method to improve the accuracy of PM2.5 prediction in eastern China. They compare the accuracy of random forest prediction with different parameter settings. Their results indicate that a simplified version with reduced parameters has the highest accuracy. The PM2.5 prediction model in this research utilizes ground-measured PM2.5 concentration data as significant parameters. The high density of PM2.5 monitoring stations in eastern China provides sufficient data. However, for the areas that lack specific data, their methodology may need some improvement. Lin et al. [30] build an air quality model to predict PM2.5 concentration based on publicly available OpenStreetMap (OSM) data and geo-spatial data mining without expert knowledge for air quality predictor selection. Their approach utilizes the PRISMS–DSCIC (Pediatric Research using Integrated Sensor Monitoring Systems–Data and Software Coordination and Integration Center) infrastructure as the data source and analytics platform. They conduct random forest mining to identify the influence of different map features on PM2.5 concentration. They use a simple k-means method to set the groups for monitoring stations. They define a buffer to map features as the influence area to PM2.5. This research conducts spatial mining in a novel way. However, data sources with such excellent conditions may not be available for other regions, which presents a potential pitfall.

Dias and Tchepel [31] study individual traffic-related PM2.5 exposure by utilizing trajectory data mining and mobile phone global positioning system (GPS) data. The researchers accordingly build an exposure model called the GPS-based exposure model to traffic-related air pollution (ExPOSITION). The mobile GPS data provides detailed time-location information for each individual. They implement the trajectory data mining and geo-spatial analysis algorithms to identify individual locations in several microenvironments (commuting, indoor, and outdoor areas). Ambient PM2.5 concentration data, outdoor/indoor infiltration factor, and possible indoor emission source data form inputs to a probabilistic approach to estimate the indoor exposure of PM2.5. This research shows the power of data mining on handling novel data sources such as personal mobile phone GPS data. It mitigates the gap between ambient PM2.5 concentration and real personal exposure. However, this research only validates the authors' simulation results by comparing it with other research results. Lary et al. [32] conducted a global scale PM2.5 concentration modeling and estimation with multiple machine learning methods. The primary data source is remote sensing data. In this research, the authors identify the critical features for PM2.5 concentrations in several selected hot spots. They mention the limitation of such immense scale data collection. The data continuity issues of different remote sensing data sources, such as different resolutions and data gaps, can be resolved with newer and more extensive coverage of more advanced satellites. However, the lack of training data for some areas needs consideration while conducting global scale PM2.5 research. That could present a potential roadblock. Min et al. [33] utilized k-means clustering analysis to find out suitable locations for air pollution monitoring stations. The researchers utilize geographic variables related to PM2.5 concentrations to separate the target area (Seoul, Korea) into several clusters with different conditions. The decrease in overall deviation (DiD) method limits the number of clusters. The clusters, which pertain to lack of monitoring stations, are useful to set new ones. The application of data mining on spatial analysis in this work is novel. However, this research only analyzes a limited set of data from specific regions. Geographic variables would need reselection since the analysis only reflects the metropolitan characteristics of Seoul. Li et al. [34] applied trajectory mining to identify potential PM2.5 sources in different seasons of Beijing. This research utilizes the data of pollutant concentration and air mass to reveal the origin and transfer of air pollutants. The researchers find that the concentrations of these pollutants are highly influenced by near surface air masses (summer and autumn) and high altitude air masses (winter and spring) in different seasons. They utilize clustering analysis to identify the pollutant sources of PM2.5. The research shows that the regional transport of pollutants has a significant impact on Beijing air quality. However, this research only utilizes monitoring station data as the PM2.5 data sources, which presents a space limitation. Introducing remote sensing data could improve accuracy and coverage.

While several works analyze PM2.5 data as per the studies surveyed herewith, they have their respective limitations as we have identified and stated alongside each piece of research. While some of these studies are localized, our study has a global horizon. An important claim in our work follows the Occam's razor principle of preferring simpler theories and hypotheses over complex ones [11]. Accordingly, we prove that classical data mining techniques in their basic form can be suitably adapted to solve interesting domain-specific problems with high efficiency and accuracy, while also facilitating easier comprehension by various users. In addition to staking this claim that constitutes a novel aspect of this article, we also develop a novel PM2.5 prediction tool that is user friendly and caters to three levels, i.e., novice, intermediate and expert, providing convenient access, inputs and result interpretation for air quality estimation, taking into account globally accepted health standards.

Smart Cities and Decision Support

There is much research recently in smart cities along with decision support for sustainable living [2, 35, 36]. Smart city characteristics such as smart environment and smart mobility are outlined in Ref. [35], along with the features for each characteristic. For example, energy efficiency, wildlife conservation, and green computing are some features of smart environment. The work in [37] proposes a novel approach called "morphing to the mean" that smooths out the expected energy demand curve and thus reduces demand fluctuation in smart cities. Works in [38–40] exemplify the importance of smart mobility, smart governance and smart living, respectively, by research in: better object detection to enhance automated transportation, transparency within governance through public feedback in legislative decision-making, and development of useful tools that contribute toward twenty-first century education. Notable sources often consider Singapore to the highest-ranked smart city worldwide [35, 41]. Other cities high in the ranking include Helsinki and Zurich. Ranking criteria are related to aspects such as intelligent lighting, fast data transmission speeds ranging TB/s, and systems geared toward achieving carbon neutrality. In Amsterdam, canal lights and street lamps automatically adjust their brightness levels based on pedestrian usage. Barcelona has bus routes designed to optimize energy efficiency using data analytics. New York City has the Link NYC system [42] with efficient kiosks to replace 9000–13,000 payphones. This embodies a free Wi-Fi initiative in NYC and is part of a global infrastructure.

Our work fits into the scope of such initiatives for smart cities. The prediction tool based on PM2.5 analysis estimates air quality, thus helping users to make decisions about urban policies, potential lifestyle adjustments, pollution control measures, etc. It thereby contributes to smart environment and other smart city characteristics (as discussed more in the section on applications). The paradigm of smart cities relates to Internet of things (IoT). Aspects such as sensors for smart connectivity are pertinent in IoT [43, 44]. The work in [45] describes data analytics in fog-based situations for efficient provisioning of resources in crowd sensing. Our work also contributes to this arena. The PM2.5 prediction tool can be useful in systems such as mobile apps. It is potentially useful to build healthcare apps for mobile devices, among other things, as explained later, in the section on applications.

Methods and Materials

Data Mining for PM2.5 Analysis

Data mining is the process of discovering novel, interesting and useful patterns and trends from large volumes of data [46, 47]. It relates to machine learning in artificial intelligence due to techniques involved that often simulate human reasoning and analysis, as adapted for machines. We claim that such techniques have the potential to analyze traffic conditions and PM2.5 concentrations in our research. We propose a solution to the problem of PM2.5 analysis by combining fundamental techniques: association rules, clustering and decision tree classifiers, into a predictive analytical framework to study the relationships between traffic factors and air pollution in multicity data. This framework for PM2.5 analysis appears in Fig. 3.

To conduct the predictive analysis, we need to obtain data on PM2.5 and urban traffic conditions. Data collection occurs from public Web sources such as online databases of the World Bank [48] and the World Health Organization [49]. The data in these repositories is of the order of megabytes (MB). Preprocessing occurs on the data after which we select relevant data parameters for further analysis. These appear in Table 2.

The raw data on these parameters undergo preprocessing with the methods of attribute selection and discretization to make them suitable for executing the data mining methods in our analytical framework. We transform numerical variables of PM2.5 into several discrete ranges based on the health impacts using the Web-based EPA standards for AQI (as outlined in Table 1). The transformed data are used for all the data mining processes in the framework of Fig. 3. The mining techniques of association rules, cluster analysis and decision tree classification are deployed as follows.

Association Rule Mining

The method of association rule mining finds relationships of the type $\alpha \Rightarrow \beta$ (α implies β), i.e., how one parameter affects another. The classical Apriori algorithm [50] for association rule mining is adapted for PM2.5 analysis. This works by finding rules based on frequent patterns. Considering this with respect to our dataset D (multicity traffic data), and a frequency threshold τ (experimental value), Apriori works to find item sets that are subsets of at least τ transactions in D, where each transaction is defined as follows. It is an instance of the multicity traffic data with the parameters: year, region, income group, PM2.5 range, road density (km road per 100 km land), per capita diesel, energy and gas consumption, motor vehicles and passenger cars per 1000 people, and vehicles per km road. Given this, in Apriori, frequent subsets are incremented one item at a time (candidate generation), and groups of candidates are tested with the data. The algorithm stops once no more successful increments occur. Apriori generates candidate item sets I of length $(n+1)$ from item sets of length n. It prunes candidates of infrequent sub-patterns. It outputs item sets F with frequent patterns. In our context, each candidate item is a combination of the given parameters with their values; frequent item sets are those higher than threshold τ, e.g., road density = "…" along with PM2.5 concentration = "…" may be frequent. Apriori works only with categorical data. Since our dataset has numerical values, these are converted to categorical values using filters.

Algorithm 1: Deploying Apriori for Association Rules from Traffic Data

```
Input: Multicity Traffic Dataset $D$, Frequency Threshold $\tau$
    1.  $I_n$: candidate item set in traffic data of size $n$
    2.  $F_n$ : frequent item set in traffic data of size $n$
    3.  $F_1$ = {frequent 1-items};  //individual parameters in traffic data
    4.  For ($n = 1$; $F_n$ !=$\varnothing$; $n$++) do begin
            a.  $I_{n+1}$ = candidates generated from $F_n$;
            b.  For each transaction $T \varepsilon D$ do
                Increment count of all candidates in $I_{n+1}$ from $T$
            c.  $F_{n+1}$ = candidates in $I_{n+1}$ with threshold $\tau$
    5.  End
Output: $\cup_n F_n$   // frequent combinations of candidates in traffic data
```

The pseudocode for Apriori as applicable to our work is shown here as Algorithm 1 (deployed from [46, 50]). Two important aspects are rule confidence (conf) and rule support (sup), called interestingness measures as they help find interesting rules [50]. In $\alpha \Rightarrow \beta$, rule confidence is the conditional probability of β occurring given that α occurs (probability of β given α, i.e., number of transactions with α and β divided by number of transactions with α). Rule support is the probability of both α and β occurring in the dataset (number of transactions with α and β divided by the total number of transactions). These measures appear in Eqs. (1) and (2), where T is a transaction, while T_α and T_β are transactions containing α and β, respectively. P refers to probability and # refers to number of (transactions), i.e., count. Hence, $\#T$ is the total number of transactions in the dataset, i.e., all transactions in the multicity traffic data here.

$$\mathrm{Conf}(\alpha \Rightarrow \beta) = P(\beta|\alpha) = \#\left(T_\alpha \cap T_\beta\right)/\#T_\alpha, \tag{1}$$

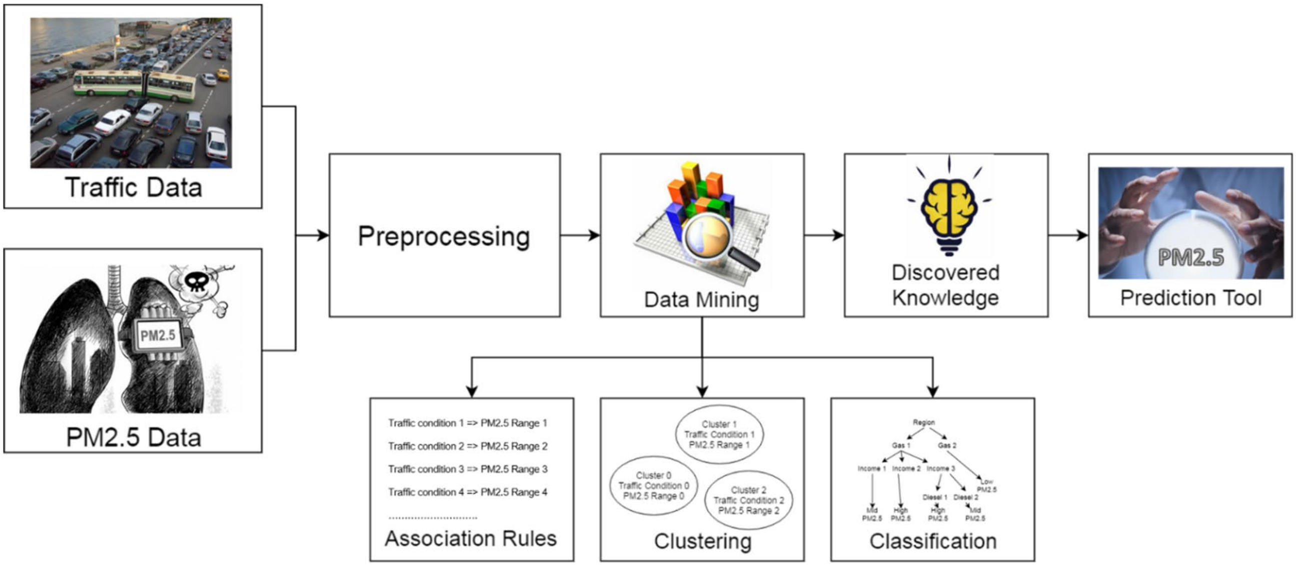

Traffic Data
PM2.5
PM2.5 Data
Preprocessing
Data Mining
Discovered Knowledge
PM2.5
Prediction Tool
Traffic condition 1 => PM2.5 Range 1
Traffic condition 2 => PM2.5 Range 2
Traffic condition 3 => PM2.5 Range 3
Traffic condition 4 => PM2.5 Range 4
....
Association Rules
Cluster 1 Traffic Condition 1 PM2.5 Range 1
Cluster 0 Traffic Condition 0 PM2.5 Range 0
Cluster 2 Traffic Condition 2 PM2.5 Range 2
Clustering
Region
Gas 1
Gas 2
Income 1
Income 2
Income 3
Low PM2.5
Diesel 1
Diesel 2
Mid PM2.5
High PM2.5
High PM2.5
Mid PM2.5
Classification

Fig. 3 Analytical framework for PM2.5 with data mining methods

Table 2 Parameter selection from worldwide multicity data

Series code	Series name
PM2.5	PM2.5 pollution, mean annual exposure (μg/m^3)
IS.ROD.DNST.K2	Road density (km of road per 100 km land area)
IS.ROD.DESL.PC	Road sector diesel fuel consumption per capita (kg of oil equivalent)
IS.ROD.ENGY.PC	Road sector energy consumption per capita (kg of oil equivalent)
IS.ROD.SGAS.PC	Road sector gasoline fuel consumption per capita (kg of oil equivalent)
IS.VEH.NVEH.P3	Motor vehicles (per 1000 people)
IS.VEH.PCAR.P3	Passenger cars (per 1000 people)
IS.VEH.ROAD.K1	Vehicles (per km of road)

$$\mathrm{Sup}(\alpha \Rightarrow \beta) = P(\alpha \wedge \beta) = \#\left(T_\alpha \cap T_\beta\right)/\#T. \quad (2)$$

These interestingness measures help to find useful rules. Thresholds are defined for minimum confidence (minconf) and minimum support (minsup) [50] to retain rules with confidence and support values greater than these thresholds. We define the thresholds as suitable experimental values incorporating domain knowledge.

The reason to deploy association rules in our predictive analytical framework is that they help understand causes of PM2.5 pollutants, and impacts urban traffic conditions on each other as well as on PM2.5 ranges. This can aid in studying the effects of parameters, e.g., if gas consumption in a certain region is in a specific range (low, very low etc.) to what extent does that cause the presence of PM2.5 in air (in terms of health impact being moderate, hazardous…) and so on. More details on this, with reference to our experimental data and challenges, appear in the section on evaluation. On a side note, we would like to mention that a possible disadvantage of Apriori is its high complexity which is $O(2^n)$ where n is the number of items in the item set, i.e., total number of data samples. Although this complexity is exponential, it is not a major concern for our data in this work, which is of the order of MB, and does not constitute big data. For our data, comprehending the associations between the parameters is more important and hence we conduct association rule mining using the classical Apriori algorithm.

Cluster Analysis

Clustering is a data mining technique that places objects in groups based on their similarities. The main idea is to have high intra-cluster similarity and low inter-cluster similarity. The notion of similarity is often domain dependent.

We apply the fundamental k-means algorithm [51] for clustering. This is explained as follows with reference to our work. Consider multicity traffic dataset D with n objects having numeric values $\left(x_1, x_2 \dots x_n\right)$. In this context, the n objects are the respective instances within D with the given parameters (as already stated for association rules). Given these, a suitable value of k is chosen as the desired number of clusters (k is an experimental value). Based on this, the algorithm partitions $\left(x_1, x_2 \dots x_n\right)$ into k clusters, so as to minimize the intra-cluster distance (ICD) in Eq. (3).

$$\mathrm{ICD} = \sum_{i=1}^{k} \sum_{x=1}^{n} x \varepsilon C_i \left(x - \mu_i\right)^2. \quad (3)$$

Here each C_i is a cluster, μ_i is its mean and x is any element of the cluster. For example, x could be: (108.63, 95.90, 39.33 …) corresponding to values of the parameters: diesel consumption, gas consumption, road density, etc., for a given instance of dataset D. Using the chosen value for k, the algorithm finds a partition of k clusters to optimize the partitioning criterion as per ICD. Since it needs to compute means, this algorithm works only with numerical data.

Steps of k-means as relevant to our work are summarized in Algorithm 2 (applied from [46, 51]). We use clustering in our framework since it is useful to reveal groups with similar PM2.5 ranges and other parameters; it provides at-a-glance displays for easy comprehension, e.g., a given cluster may contain instances mostly from high-income group countries regardless of their vehicles per km: indicating at-a-glance that income range is a significant factor in PM2.5 analysis.

Algorithm 2: Applying k-means for Clustering over Traffic Data

```
Input: Multicity Traffic Dataset $D$, Number of clusters $k$
  1. Partition n objects $(x_1, x_2 \dots x_n)$ into $k$ nonempty clusters $C_{i\ (i = 1\ to\ k)}$
  2. For $(i = 1; i \le k; i{+}{+})$ do begin
Compute $\mu_i$ = mean point of cluster $C_i$
Assign each object to (new) cluster with nearest mean point
  3. If no new assignments, then STOP
Else Go to Step 2
Output: final $C_{i\ (i = 1\ to\ k)}$ as $k$ clusters   // Groups of similar traffic data values
```

We prefer to use k-means since most of the data on PM2.5 and related aspects have numerical attributes. There are a few categorical attributes, which are converted to numerical values by using filters. For simplicity, we use the standard Euclidean distance [46] as the notion of similarity for the PM2.5 data. The clustering execution details along with challenges and the execution results are available in our evaluation section. The complexity of k-means is $O(nkt)$ where n is the number of objects, i.e., data samples, k is the number of clusters and t is the number of iterations. Thus, it can be noticed that k-means is a fairly simple algorithm, which is another reason for selecting it as a preferred method in our approach.

Decision Tree Classification

The process of classification in data mining predicts a target, based on learning from existing data. Decision trees used as classifiers generate specific paths for prediction purposes. A decision tree starts with a root, and creates nodes with paths, leading to leaves that depict outcomes or decisions. To learn these paths, a process typically followed is decision tree induction [46, 47]. In our work, we use the popular J4.8 algorithm [47, 52], a Java version of the C4.5 algorithm [53] for induction of decision trees. The resulting decision trees are used as classifiers to predict targets, in our data, PM2.5 ranges. We explain the C4.5 algorithm propounded by Quinlan [53], considering our data. The C4.5 algorithm learns decision trees from pre-labeled training data D with correctly labeled instances as inputs. In our work, D is pre-labeled multicity traffic data with PM2.5 ranges provided for each instance (with traffic parameters as stated in association rules). Given this, C4.5 follows a top-down, recursive strategy. First, it selects a suitable attribute for a root node. In our context, an attribute is a parameter of traffic data. Given this, C4.5 then creates a branch for each possible attribute by splitting instances into subsets, one for each branch. It repeats this process recursively for each branch, using only the instances that reach the branch. To select the best attribute for the root and subsequently the other nodes, it uses a heuristic of choosing the purest nodes. A popular criterion for this is information gain; C4.5 uses a normalized version of it called gain ratio [53]. We explain information gain using entropy: average rate at which information is produced by a stochastic source of data [53, 54]. Entropy is measured as negative logarithm of its probability mass function calculated using Eq. (4), where $\left(P_1, P_2 \ldots P_n\right)$ is the probability distribution of the data for n instances, i.e., in our work, the instances of multicity traffic data with the given parameters.

$$\text{Entropy}\left(P_1, P_2 \ldots P_n\right) = -\sum_{i=1}^{n} P_i \log_2 P_i. \tag{4}$$

Thus, entropy can be used to compute the information given by an individual attribute A or the entire dataset D using the probability distribution. Equations (5) and (6) express this, where Info(A) and Info(D) represent the information given by attribute A and the entire dataset D, respectively.

$$\text{Info }(A) = \text{Entropy}\left(P_1(A), P_2(A) \ldots P_n(A)\right), \tag{5}$$

$$\text{Info}(D) = \text{Entropy}\left(P_1(D), P_2(D) \ldots P_n(D)\right). \tag{6}$$

Thus information gain of attribute A, i.e., InfoGain(A), is defined as information given by the dataset minus that given by attribute A alone, as seen in Eq. (7) next.

$$\text{InfoGain}(A) = \text{Info}(A) - \text{Info}(D). \tag{7}$$

When node is pure, this measure should be zero and when impurity is maximum (i.e., all classes equally likely), this measure should be maximal. These properties are satisfied by information gain, making it a good choice for a purity measure. Yet, it is biased toward attributes with a large number of values. This may result in overfitting by selecting an attribute with many values, non-optimal for prediction. To counterbalance this, gain ratio is used to reduce the bias of information gain [53]. Gain ratio considers the number and size of branches in choosing an attribute and incorporates intrinsic information of a split, i.e., entropy of distribution of instances into branches. Hence, the gain ratio of an attribute reduces as intrinsic information gets larger. Gain ratio is formulated in Eq. (8), where IntrinsicInfo(A) denotes the intrinsic information of the attribute A. Gain ratio is thus normalized information gain.

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{IntrinsicInfo}(A)}. \tag{8}$$

This gain ratio is used by C4.5 as the purity criterion to choose the best node for splitting. Thus, the attribute with the highest gain ratio is selected as the root, and at each level the attribute with the highest gain ratio at that point becomes the best choice for the node. C4.5 recursively uses the same procedure until it reaches the leaves that represent decisions. It has the following choices for a termination criterion [53].

- If all instances belong to the same class, C4.5 creates a leaf node for the decision tree indicating that it should choose that class.
- If none of the attributes provides any information gain, C4.5 creates a decision node higher up the tree using the expected value of the class.
- If an instance of a previously unseen class is found, C4.5 again creates a decision node higher up the tree using the expected value.

Given this description of C4.5, the J4.8 algorithm [52], a Java implementation of C4.5, is outlined next as Algorithm 3 with respect to its application in our work (utilized from [47, 53]). Based on this, decision trees are output by J4.8 using multicity traffic data as training data D, and the results are useful for classifying test data. Thus, by following the path of a decision tree and matching a new instance with the existing data, a classification target can be predicted.

The reason for decision tree classifiers in our analytical framework is that PM2.5 ranges can be estimated based on new inputs by following the closest matching tree path from the data used for learning, e.g., a tree path may indicate that if region = "East Asia and Pacific", income group = "high" and vehicles per km = "0–50", then PM2.5 range = "low". If a new instance with similar parameters is entered, it can be estimated that the air quality in that instance is "good", based on this tree path and the corresponding health standards for "low" PM2.5 concentrations (from Table 1). If an exact path is not found in any new instance, we output an approximate match within a given threshold (experimental value). More challenges are discussed in the evaluation section. Note that we prefer decision trees to other classifiers such as random forests and neutral networks [46, 47], since it is important for us to trace the path from the root to a leaf in the prediction tool to find a full or partial match and convey the predicted response accordingly. This is less feasible with other classifiers, especially with the neural network which is a black box. Additionally, the tree paths also facilitate easier comprehension, which is useful in our tool since we cater to various user levels and provide glimpses of our analysis to them. Moreover, our experimentation reveals that for our data, decision trees also provide higher (or equally high accuracy) compared to other classifiers while also executing faster than others, thus providing the highest efficiency. Details on this appear in our evaluation section. Algorithm 3 summarized the utilization of decision tree classifiers in this work.

<u>Algorithm 3: Utilizing J4.8 for Decision Tree Classifiers on Traffic Data</u>

```
Input: Multicity Traffic Dataset $D$, Target Attribute (PM2.5 range)
    1. Begin J4.8($D$)
    2. Tree $T$ = {}
    3. If ($D$ is *pure* || another termination criterion) then STOP
    4. For each attribute $A \varepsilon D$ do    // each parameter of traffic data
        a. Calculate gain ratio (A)
    5. $A_{best}$ = Best attribute as per gain ratio(A)   // most significant traffic parameter to predict PM2.5 range
    6. $T$ = Form a decision node that tests $A_{best}$ in root
    7. $D_v$ = Derived subsets from $D$ using $A_{best}$
    8. For all $D_v$ do begin
        a. Tree $T_v$ = J4.8($D_v$)
        b. Affix $T_v$ to respective branch of $T$
        c. End
    9. End
Output: Tree $T$   // with classification target PM2.5 range
```

The reasons for choosing J4.8 are: it is a very popular implementation of the classical C4.5, which uses gain ratio; and it is user friendly, written in Java. Since our prediction tool is programmed in Java, we prefer this algorithm. The complexity of this algorithm is $O(n \log_2 n)$, where n is the number of data samples. Since this complexity is linear/logarithmic and does not head toward being quadratic or exponential, it serves as an added plus in selecting this algorithm in our approach, since we prefer the simplicity. Classification results are presented in the evaluation section, with results of association rule mining and clustering. Based on the results of all these data mining methods, we design the prediction tool for air quality estimation.

Prediction Tool Development

An important aspect of the prediction tool is categorization of user levels. This occurs to distinguish between users as novice, intermediate or expert. The tool interface is developed such that it first interacts with the users to determine their level and then proceeds with providing the functions accordingly. A snapshot of the welcome screen of the tool appears in Fig. 1.

User Level Categorization

We develop a User Level Questionnaire containing multiple choice questions on environmental engineering, especially air pollution. This is embedded into the tool and is offered to users at the entry point. If users declare that they are novice, they are exempt from the questionnaire. Else, this determines their level. Examples of questions are as follows.

<u>User Level Questionnaire (Partial Snapshot)</u>

(i) What is the difference between PM2.5 and PM10 particles?

A) Diameter ranges only *B) Health impacts only*

C) Both of these *D) Neither of these*

(ii) How do you expand the acronym AQI in environmental engineering?

A) Assessed Quantity ID *B) Air Quality Index*

C) Atmosphere Quest Input *D) None of these*

(iii) Why has the OECD organization been founded?

A) For all European nations *B) For car exports*

C) For economic cooperation *D) All of these*

Accordingly, we offer 25 multiple-choice questions. Users that score 20 points and above are labeled as expert, those between 10 and 19 points are intermediate and those with 9 points and below are novice. Our domain experts in environmental engineering consider this determination of user levels appropriate.

System Architecture

The prediction tool architecture with respect to user perspectives is illustrated in Fig. 4. Its details are explained for convenient access, convenient user input and convenient result interpretation at various user levels.

Convenient access: Various users have different needs from this prediction tool. For example, expert users would be interested in detailed visualization of attributes affecting PM2.5 concentrations while novice users would more likely be interested a quick at-a-glance access. Accordingly, the users are guided through the prediction tool to select its appropriate options for different types of analyses with various levels of detail. We also provide two versions of access: online/off-line. The tool Website offers desktop and mobile versions. This ensures access for most online users. For users with limited access to the Internet, an off-line installation package is provided through the download link on the website. Thus, all users can have convenient access.

Convenient user input: In this prediction tool, the users' input of traffic conditions is an essential step to get the predicted results. In this step, any user should be able to identify correct data and provide inputs precisely. Hence, a user-friendly graphical user interface (GUI) is designed for convenient user input. A basic tutorial is provided for all the first time users (novice, intermediate and expert) to ensure the learning of operations. Each input term has explanation along with a local file. We provide links to underlying databases with a menu-driven interface to help users get desired data for inputs. This is particularly useful to novice and intermediate users. For novice users (and intermediate users if needed), there are help files for guiding them to understand meanings of attributes. Such support functions can be turned off by expert users, as needed.

Convenient result interpretation: To connect PM2.5 ranges with health impacts, we convert numerical PM2.5 data into ranges based on EPA standards. The results include not only the predicted PM2.5 ranges but also the related health impacts. For example, instead of stating that the predicted PM2.5 range is x µg/m^3, the system states that the range is "safe", "moderate" etc. from a health standpoint. This is based on AQI mapping by using EPA standards from Table 1. There are functions to share information, catering to the needs of different users. Novices would most probably be interested in knowing safety levels, while experts would be more interested in actual ranges corresponding to those levels with analytical details. Suitable functions are provided to experts for viewing snapshots of data mining analyses based on traffic conditions. Thus, expert users (and perhaps intermediate users) can refer to these for further details on result interpretation. Novice users are provided with simple views, mainly from health angles.

Accordingly, the prediction tool is implemented based on this system architecture. This tool has already incorporated data mining results based on the PM2.5 analytical framework. Since these results are accurate (see evaluation section), they are used to design the tool. Its components are as seen in the figure: websites, online version and off-line version for convenient access; help files, relevant GUIs for convenient user input; and online/off-line explanation files, with result screens. All these components aid convenient result interpretation.

Evaluation and Results

Data Mining Experiments

The fundamental techniques of association rules, clustering and classification are combined into the predictive analytical framework of PM2.5 as described earlier. The well-known tool called WEKA (Waikato Environment for Knowledge Analysis) [52] is used in our work. WEKA offers Java implementations of various data mining/machine learning algorithms. Its open source Java code is integrated into our analytical framework in the appropriate parts. The PM2.5-related data are converted into a format called ARFF (attribute relation file format) [47, 52] suitable for WEKA. In addition, domain-specific aspects are incorporated within the data by using AQI standards and health impacts from Table 1. Thus, instead of stating that a given data instance in a certain numeric range, the tool states that the instance is "good", "very good", etc. as per these standards. All the experiments in this project have been executed on laptop computers in our Data Science Laboratory; these computers are Dell E5470 machines with Intel ® Core ™ i5-6300U @ 2.40 to 2.50 GHz, 256 GB SSD and 8 GB RAM running Windows 10.

Experiments with Association Rules

For association rule mining, the Apriori algorithm [50] has been utilized as shown in Algorithm 1. Since Apriori works with categorical data, numerical data are converted to categorical by using the "Discretize" filter of WEKA [52]. A challenge in this execution includes incorporating domain knowledge in rule mining, especially with respect to health

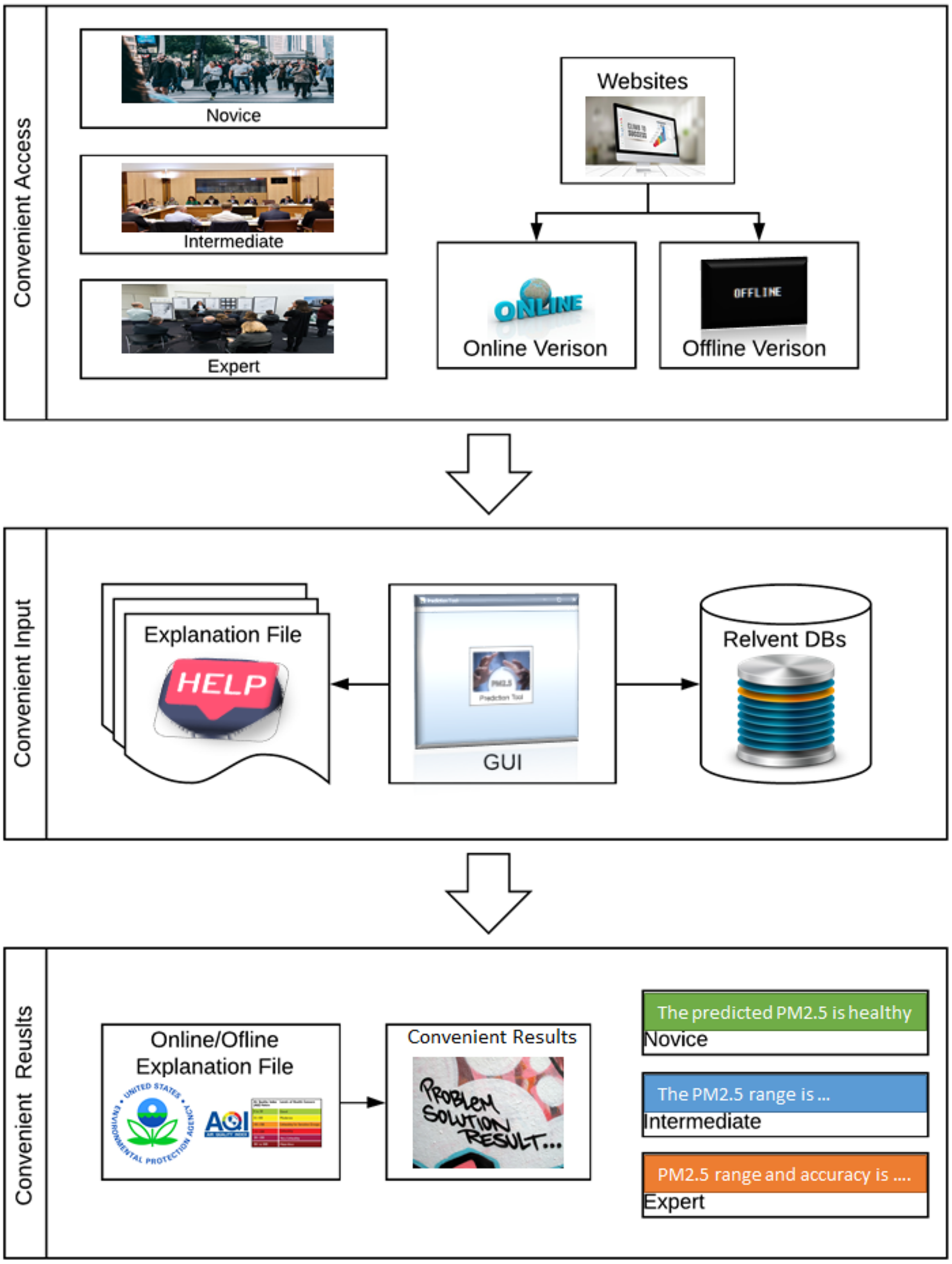


Fig. 4 Illustration of system architecture from user perspectives

standards. It is important to find rules that have significant domain-specific impacts, even if they may not be very frequent. On the other hand, finding too many rules can be overwhelming since some of them may be uninteresting. Thus, we need to strike a good balance between these two extremes. We address this challenge empirically by altering values of rule confidence and support, using our perception guided by domain knowledge, such that we select interesting rules. Thus, we do not rely solely on the actual values of confidence and support, but make adjustments based on the knowledge of environmental engineering as well as health standards, to output useful results. Based on these experiments, association analysis reveals interesting results. It is found that PM2.5 concentrations are quite strongly related to regions. Income parameters can significantly influence other traffic parameters. A few examples of interesting association rules are shown next.

PARTIAL SNAPSHOT OF ASSOCIATION RULE MINING OUTPUT

1. Year=’2005’ ^ ‘Region=Europe & Central Asia’ ^ Vehicles_per_km=’Very Low’ => PM2.5_Class= Good

2. Gasoline_Consumption = ‘Very Low’ ^ Road_Density=’Very Low’ ^ Cars_Per_K_People=’Low’ => PM2.5_Class=’Moderate’

3. Income_Group=’High Income Non-OECD’ ^ Diesel_Consumption= ‘High’ ^ Road_Density=’Moderate’ => PM2.5_Class=’Moderate’

4. Region=’Middle East & North Africa’ ^ Income_group=’Lower Middle’ ^ Road_Density=’Moderate’ => PM2.5_Class= ‘Moderate Potential Damage’

5. Region=’Europe & Central Asia’ ^ Income_Group=’High Income OECD’ ^ Road_Density=’Moderate’ => PM2.5_Class =’Very Good’

In these rules, the terms “very good”, “good”, “moderate”, “moderate potential damage”, etc., are used to describe PM2.5 ranges as per EPA health standards and AQI values. The terms “very good” and “good” are for PM2.5 ranges well within and close to safe EPA standards, respectively. The term “moderate” describes slightly higher than safe EPA ranges while “moderate potential damage” describes PM2.5 ranges with greater than recommended EPA standards. OECD refers to Organization for Economic Cooperation and Development [54] an inter-governmental organization founded to stimulate economic progress and world trade; most OECD countries are typically developed nations. Based on this interpretation, the first rule indicates that in 2005, for Europe and Central Asia, if vehicles per km are in the “very low” attribute range, the PM2.5 class is within the “good” range, i.e., safe EPA limits. The third rule states that in situations with high-income non-OECD backgrounds where the diesel consumption is high and the road density is moderate, the PM2.5 range is moderate with respect to health standards. Likewise, we can interpret other rules. Such rules can help to answer the research questions 1b) and 1c) in the introduction: pertaining to traffic parameters affecting each other, and air quality impacts with respect to health standards, respectively.

Experiments with Clustering

Clustering is conducted on the PM2.5 data using *k*-means [51] as shown in Algorithm 2. Most of our data are numerical as required by *k*-means. Non-numerical data are converted numerical using a “Nominal to Binary” filter of WEKA [52]. An important challenge here is to select a suitable value for *k*, such that we do not have too many clusters getting very specific leading to lack of inference about general qualities, nor do we have very few clusters causing too much of generalization. This is particularly significant for data on health and safety issues. We handle this challenge by consulting domain experts while assessing cluster quality, thereby making suitable alterations to values of *k*, and getting a desirable number of clusters. The minimum description length (MDL) principle [55], often used in clustering, is also useful in addressing this challenge and finding relatively optimal values for *k*, in our experiments. This proves to be helpful in our context. Accordingly, we obtain useful results, summarized with a partial snapshot of the output in Table 3.

The clustering results lead to interesting observations. Cluster0 has relatively low traffic conditions, yet its PM2.5 range is above EPA defined standards. It leads to the inference that in these countries, traffic volume may not be a major source of PM2.5, and there could be other causes of fine particle pollutants. Note that income of this cluster is

Table 3 Partial snapshot of clustering output from PM2.5 analysis

Parameter	Cluster0	Cluster1	Cluster2	Cluster3
Income group	Upper middle income	High-income OECD	High-income non-OECD	High-income OECD
Diesel consumption	108.63	416.61	208.70	266.42
Gas consumption	95.90	341.07	286.03	186.48
Road density	39.33	140.83	149.42	97.52
Cars per *K* people	120.54	493.28	234.14	290.32
vehicles/*K* people	151.99	588.38	288.69	345.04
Vehicles per km	37.90	50.23	86.21	27.88
PM2.5 range	(15.2–18.43]	(− inf to 5.85]	(21.76–inf]	(5.85–11.98]

the lowest, which could imply that low-income regions may give high PM2.5 ranges due to lack of economic regulations. Cluster2 has the highest PM2.5 ranges, yet it is not the highest traffic indicator. It probably leads us to deduce that countries in this cluster may have other major PM2.5 sources and/or poor regulations of automobile emissions. Cluster1 and Cluster3 both have PM2.5 ranges within safe EPA standards. While traffic indicators of Cluster1 are higher than those of Cluster3, its PM2.5 ranges are reduced to half that of Cluster3. This could therefore imply that regulatory mechanisms of countries in Cluster3 are better than those in Cluster1. Thus, it is observed that the actual traffic emissions do not have a direct proportionality with the PM2.5 concentrations. Other factors: income groups, alternative pollutant sources, regulatory mechanisms play a significant role. Clustering can be quite useful in answering research questions 1a) and 1d) in the introduction, regarding: significance of parameters causing PM2.5 presence, and similarity between various regions based on air pollution.

Experiments with Decision Tree Classifiers

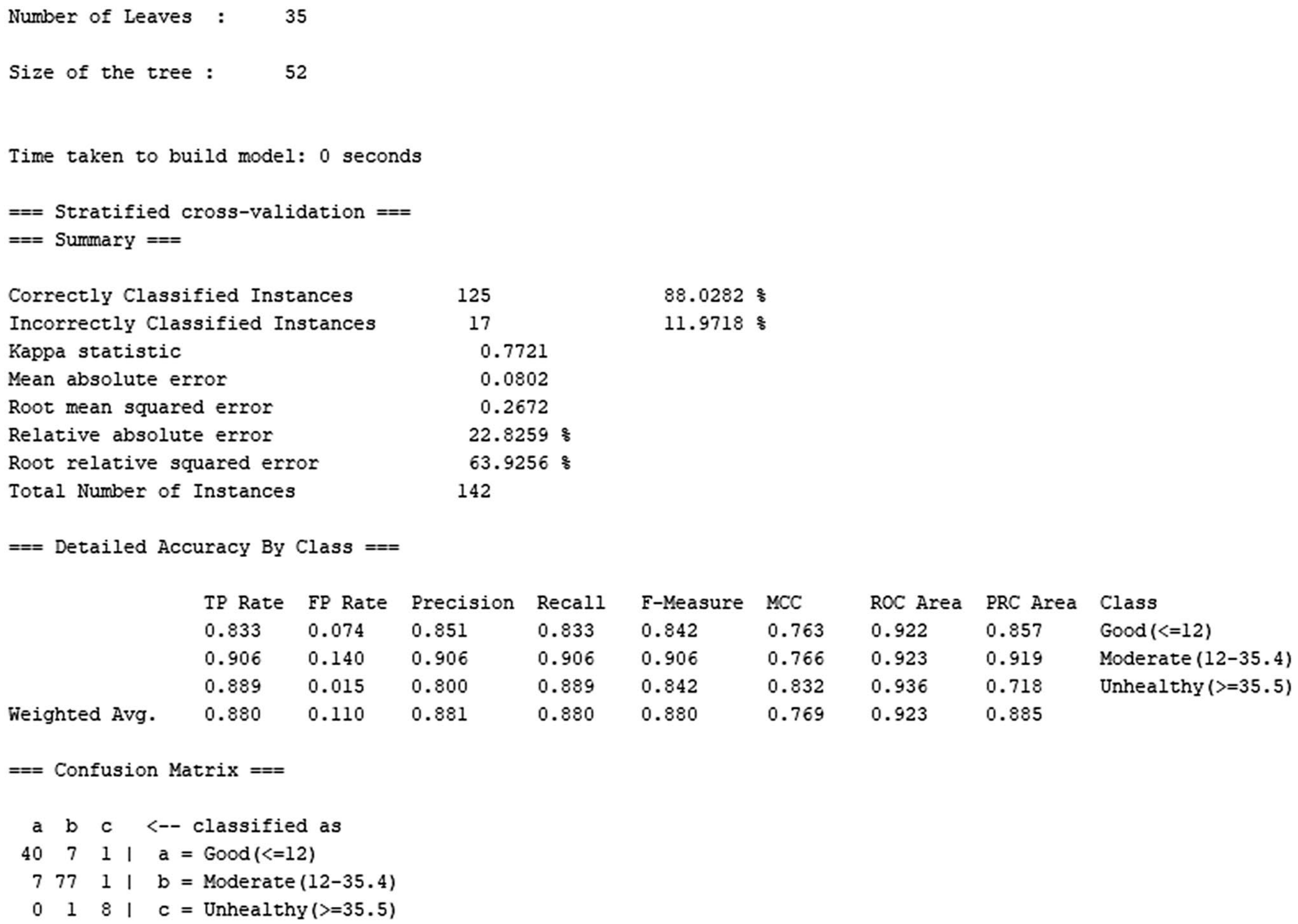
Classification is conducted with decision trees to predict PM2.5 ranges. We use J4.8 for decision tree induction [52, 53], as shown in Algorithm 3. Although this works with numerical data, we convert the data into categorical ranges, for better interpretability. This is because the classification target is the PM2.5 range and it is better to see this grouped into discrete categories rather than a continuous value, as categories help to relate easily with health standards. The main challenge in decision tree classification here is to maintain a trade-off between the simplicity of the tree and the completeness of its paths for matching with new data. The Occam's razor principle, attributed to the philosopher William of Ockham, states that "Simpler solutions are more likely to be correct than complex ones" and thus it is advisable to "shave away" unnecessary assumptions [11]. Adhering to this principle that we advocate in our entire work herewith, we would have a preference for smaller, shorter and simpler trees. Yet, we take into account another aspect: if the tree is too simple, it is less likely to give a complete match when it encounters new data; thus, the closest partially matching path would be used as the predicted output. This can adversely affect prediction accuracy (and can lead toward preferring more intricate trees). Hence, to tackle this challenge, we conduct multiple executions of decision tree classifiers, using its parameters to prune unnecessary branches if and only if they do not negatively affect classification accuracy, leveraging the needs of the domain. This is done by using domain-specific validation sets while learning, in addition to training sets. Trees with the highest classification accuracy on validation sets are used as outputs of the PM2.5 data analysis, for design of the prediction tool.

PARTIAL SNAPSHOT OF DECISION TREE CLASSIFIER OUTPUT

```
Region = East Asia & Pacific

| Gasoline_Consumption <= 427.7

| | IncomeGroup = High income: nonOECD: (18.43 to 21.755]

| | IncomeGroup = High income: OECD: (21.755 to inf]

| | IncomeGroup = Low income: (18.43 to 21.755]

| | IncomeGroup = Lower middle income: (11.98 to 15.12]

| | IncomeGroup = Upper middle income

| | | Diesel_Consumption <= 114.38: (21.755 to inf]

| | | Diesel_Consumption > 114.38: (11.98-15.12]

| Gasoline_Consumption > 427.7: (0 to 5.845]
```

A partial snapshot of decision trees from our classifier experiments is shown here. These decision trees are found to give high accuracies heading toward approximately 90% in classifying unseen test data. Multiple experiments are conducted; and average classification accuracy values obtained are close to that range. Hence, these trees are used for further work and are presented here. In this excerpt of decision trees, numbers in (] brackets indicate predicted values of PM2.5 ranges. By analyzing several paths, it is seen that the "Region" attributes have strong effects on PM2.5 ranges. It is also found that PM2.5 pollutants are highly associated with local conditions. An interesting and surprising observation is that high gasoline and diesel consumptions do not essentially lead to higher PM2.5 ranges. The justification for this upon further study in environmental engineering is that such high consumption often associates with very good economic conditions and pollutant regulations. Thus, though diesel and gas consumption is high, regulatory mechanisms lower the PM2.5 concentrations. Decision tree classifiers can thus help to answer research questions 1a) and 1c) in the introduction based on: parameter significance in presence of PM2.5, and impact of air quality on health (by mapping the predicted PM2.5 ranges to their corresponding health standards as per the AQI table). Furthermore, some surprising observations such as that stated here propel further studies.

Note that we use decision trees in this work for multiple reasons: convenient traversal of paths to find a full or partial match, high prediction accuracy as well as high learning efficiency. To emphasize this, we present results with comparative studies using other classifiers [46, 47]. These results include a summary of our comparative evaluation. In Fig. 5, we include an example of an execution with decision tree classifiers. We can see that the time required to build the model is negligible (0 s) while the accuracy obtained on unseen



```
Number of Leaves  :     35

Size of the tree :      52


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         125               88.0282 %
Incorrectly Classified Instances        17               11.9718 %
Kappa statistic                          0.7721
Mean absolute error                      0.0802
Root mean squared error                  0.2672
Relative absolute error                 22.8259 %
Root relative squared error             63.9256 %
Total Number of Instances              142

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.833    0.074    0.851      0.833    0.842      0.763    0.922     0.857     Good(<=12)
                 0.906    0.140    0.906      0.906    0.906      0.766    0.923     0.919     Moderate(12-35.4)
                 0.889    0.015    0.800      0.889    0.842      0.832    0.936     0.718     Unhealthy(>=35.5)
Weighted Avg.    0.880    0.110    0.881      0.880    0.880      0.769    0.923     0.885

=== Confusion Matrix ===

  a  b  c   <-- classified as
 40  7  1 |  a = Good(<=12)
  7 77  1 |  b = Moderate(12-35.4)
  0  1  8 |  c = Unhealthy(>=35.5)
```

Fig. 5 Example of decision tree execution

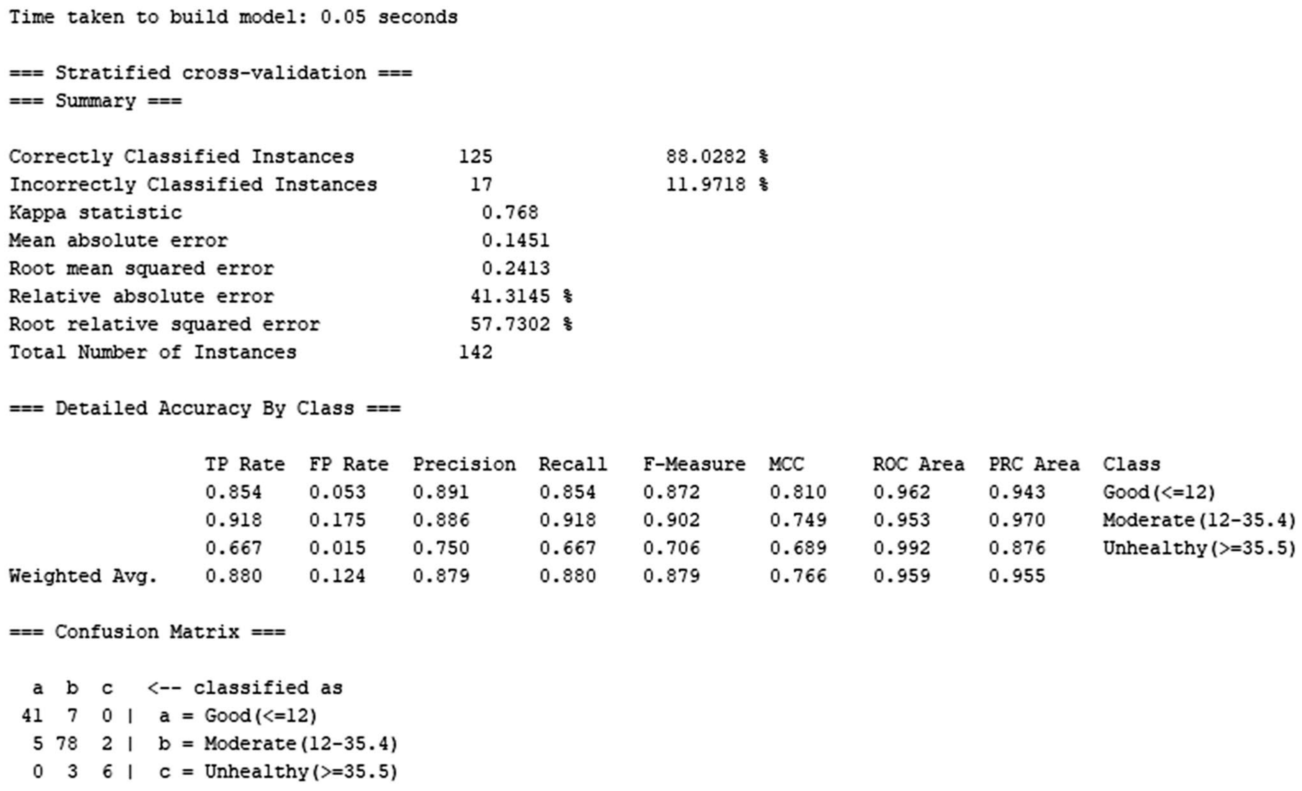
test data is more than 88% here. It also has the advantage of easy interpretation (as seen from the partial snapshot of the classifier output shown earlier). Likewise, Fig. 6 presents an example of execution on the same data with a random forest classifier for comparison. It is clear from this figure that the time taken to build the classifier is somewhat higher while the accuracy is similar to that of a decision tree. Moreover, it lacks the benefit of facilitating interpretation as easily as a decision tree. In Fig. 7, we present an example of execution with the artificial neural network classifier (ANN) on the very same data, for additional comparison. Note that this ANN actually gives lower classification accuracy than the decision tree while also requiring longer execution time. Furthermore, the ANN itself is a black box making any interpretation much harder, which poses issues in our context.

Comparative studies such as these are conducted with other classifiers as well. Based on various results obtained (as seen in the examples here), we find decision trees the most suitable for use within the prediction tool due to accuracy, efficiency and interpretation. The results from the analysis with all the data mining techniques here serve as the basis for building the prediction tool. We present a demo of this tool herewith.

Discussion on Experiments

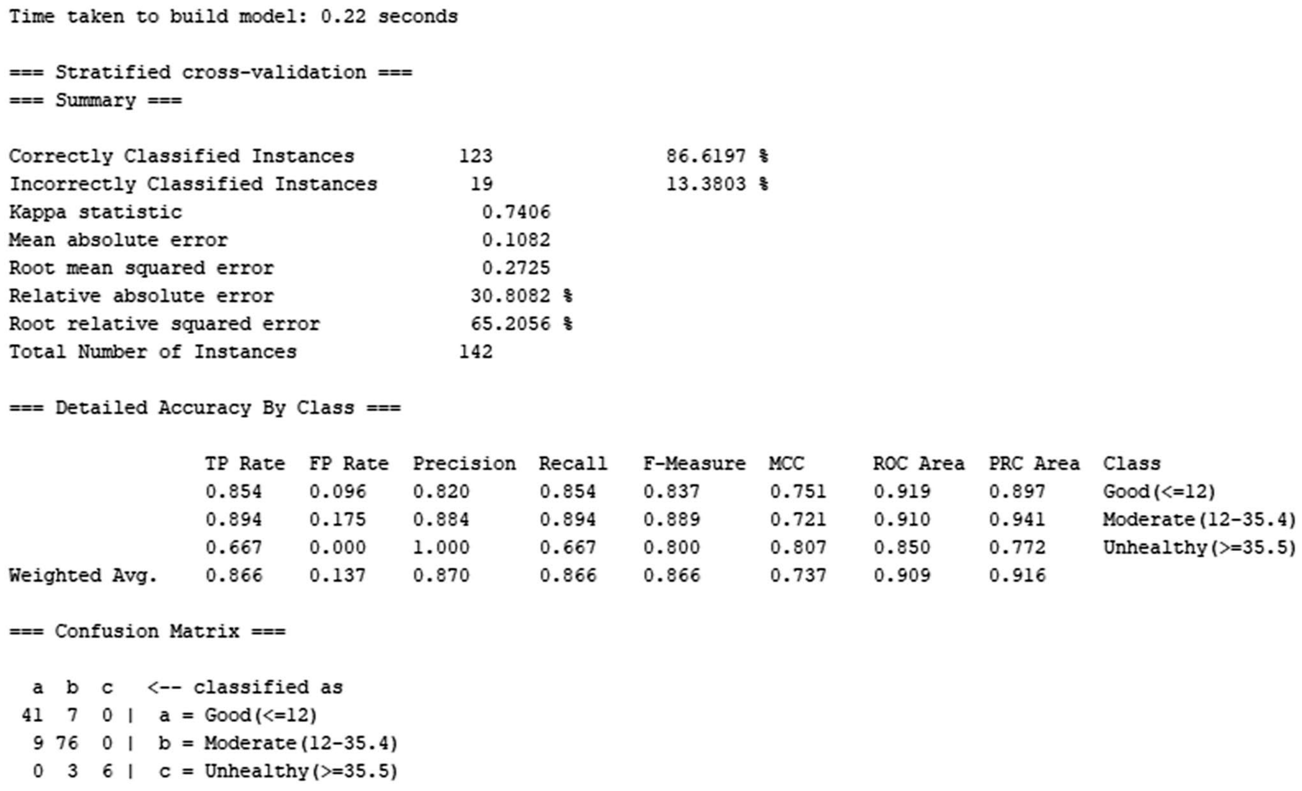
As evident from the results, high accuracy is obtained with the data mining and machine learning models used in this work, this being corroborated by comparative studies as well. In general it is to be noted that machine learning methods are non-parametric and do not require distributional approach and assumptions. However, to achieve the specification of hyper-parameters, they do require some fine-tuning through which controlling is possible. Considering these aspects, the specific hyper-parameters addressed in the experiments herewith include the following.

- Number of clusters in *k*-means: Since the “*k*” in the *k*-means algorithm is often crucial, we have experimented with values here ranging from 2 to 10 clusters. While 2 has been considered quite low with not too much distinction between the groups, 10 has been considered a little too high with not enough generalization, and neither



```
Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         125               88.0282 %
Incorrectly Classified Instances        17               11.9718 %
Kappa statistic                          0.768
Mean absolute error                      0.1451
Root mean squared error                  0.2413
Relative absolute error                 41.3145 %
Root relative squared error             57.7302 %
Total Number of Instances              142

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.854    0.053    0.891      0.854    0.872      0.810    0.962     0.943     Good(<=12)
                 0.918    0.175    0.886      0.918    0.902      0.749    0.953     0.970     Moderate(12-35.4)
                 0.667    0.015    0.750      0.667    0.706      0.689    0.992     0.876     Unhealthy(>=35.5)
Weighted Avg.    0.880    0.124    0.879      0.880    0.879      0.766    0.959     0.955

=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  7  0 |  a = Good(<=12)
  5 78  2 |  b = Moderate(12-35.4)
  0  3  6 |  c = Unhealthy(>=35.5)
```

Fig. 6 Random forest execution for comparative study



```
Time taken to build model: 0.22 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         123               86.6197 %
Incorrectly Classified Instances        19               13.3803 %
Kappa statistic                          0.7406
Mean absolute error                      0.1082
Root mean squared error                  0.2725
Relative absolute error                 30.8082 %
Root relative squared error             65.2056 %
Total Number of Instances              142

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.854    0.096    0.820      0.854    0.837      0.751    0.919     0.897     Good(<=12)
                 0.894    0.175    0.884      0.894    0.889      0.721    0.910     0.941     Moderate(12-35.4)
                 0.667    0.000    1.000      0.667    0.800      0.807    0.850     0.772     Unhealthy(>=35.5)
Weighted Avg.    0.866    0.137    0.870      0.866    0.866      0.737    0.909     0.916

=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  7  0 |  a = Good(<=12)
  9 76  0 |  b = Moderate(12-35.4)
  0  3  6 |  c = Unhealthy(>=35.5)
```

Fig. 7 Execution of ANN for comparative study

of these values has given us truly meaningful results. The most interesting results as per the interpretations of our domain experts have been obtained with $k = 4$ due to which we have presented those results here.

- Clustering seeds: The seed in clustering needs to be altered in every iteration for greater randomization and hence we have adhered to that principle to obtain more robust results. For every value of k, at least five different runs have been conducted with different clustering seeds, to retain selected ones, for documentation and study.
- Confidence in Apriori: The confidence of the association rules, i.e., the probability of β occurring given that α occurs in the rule $\alpha \Rightarrow \beta$, is a significant threshold in studying the impact of the parameters on each other. Hence, we have fine-tuned the confidence threshold to a large extent to understand parameter impacts in a reasonable manner. The range of alteration of confidence values includes 50–100%, in steps of 5% for each iteration. The results for association rules presented herewith are close to the 70% range, for the most part.
- Support in Apriori: The support for the association rules, i.e., the probability of α and β both occurring in the entire dataset, is another crucial threshold since it helps to understand the extent to which both the parameters (on the left and right side of a given rule) occur in the dataset. Hence, we have altered this value substantially, analogous to the value of confidence, keeping it in the same range, i.e., 50–100% in steps of 5% for each iteration. The results for association rules shown herewith depict support in the approximate range of 80%.
- Number of trees in J4.8: Since decision tree classification yields several possible trees based on different runs, we have considered the number of trees as an important hyperparameter. Hence, we have derived more than ten trees for every combination of attributes, and retained those that seemed most meaningful from a domain-specific angle.
- Maximum tree depth in J4.8: Given the fact that very short trees can be too generic thus not allowing us to draw clear inferences about the data, while trees with much higher depth could potentially lead to overfitting, we have preferred medium ranges for trees. Hence, we have conducted pruning of trees with depth greater than 6, in general, allowing the maximum tree depth to range from 3 to 6. It has been observed that we have obtained the most useful trees with a maximum depth of 4, considering the attributes analyzed, and incorporating the domain perspectives.
- Learning rate in J4.8: Analogous to tree depth, we have maintained the learning rate in the middle range. A very high learning rate would lead to fast results, but would possibly not classify the target accurately, while a very low learning rate would make the training too slow and inefficient. We have thus selected learning rates in the middle range based on the values provided by the software. We have conducted experiments with several learning rate values, considering at least five runs for each combinations of attributes, to achieve greater robustness in the learned hypothesis.

As per the above discussion, we have included hypertuning in our machine learning models, based on which exhaustive experimentation has been conducted, the interesting results being synopsized herewith. Note that *k*-fold cross-validation has been used in all our experiments for additional robustness, in addition to using separate training and test sets. The values of k have been varied from 4 to 10, since these are the common values typically used in machine learning. The results presented here constitute some examples of all the evaluations performed. Please see Figs. 5, 6 and 7 that include the results with the stratified (*k*-fold) cross-validation summary as shown therein. These results offer the hypothesis testing that is needed to use the learned hypothesis for building the prediction tool. We now present a system demonstration of this tool.

System Demo of Prediction Tool

The PM2.5 prediction tool is implemented in Java by using results from the predictive analysis with association rules, clustering and decision tree classifiers. This tool is suitable for novice, intermediate and expert users. Initial interaction between users and the tool can be as follows, presenting a step toward the convenient access issue.

Initial Interaction Example 1:

System: Please select your level (Novice / Intermediate / Expert / Unknown)

User: Novice

System: Please enter parameters in input screen …

Initial Interaction Example 2:

System: Please select your level (Novice / Intermediate / Expert / Unknown)

User: Unknown

System: Please take the questionnaire to find your level …

Questionnaire Output: 22/25

System: Your level is Expert. Please enter parameters in input screen …

We now present a few screenshots of the system demo. In Fig. 8, we portray an example of visualization results with multiple parameters simultaneously via the method of scatterplots. These enable the comparison of one parameter with respect to another in sufficient detail, useful to experts. It is useful for the analysis of trends, e.g., it can provide an intricate view on how the parameter of “cars per k people” relates to “diesel consumption” and “gasoline consumption” by observing the respective regions on the scatter plots.

Novice users may not be interested in such details. They may simply need to see the overall ranges. Thus, Fig. 9 shows an example of visualizing PM2.5 ranges with respect to their corresponding health levels on a table and a chart. It includes the information on one common screen, fit for easier comprehension. This gives novice users a brief idea of the distributions of PM2.5 ranges as per health standards without traversing intricate details. Likewise, for intermediate users, there are medium levels of detail provided in the tool.

Although we identify user levels and tool interactions as mentioned herewith, the disclaimer is that there can be differences. Novice users may at times want detailed analysis to gain further insights into the area of study; conversely expert users may sometimes need to see a simple view with brevity. Consequently, they can select appropriate functions from a top level menu to choose the extent of detail. Yet, note that our design as presented here is actually found to be suitable by real users (as seen in the tool assessment later). This aspect of the design and demo caters to the Convenient Access feature in the problem definition. We now illustrate the features of convenient input and results.

In Fig. 10, we present an example of specific parameters entered in an input screen. Note that expert users can enter parameters directly while intermediate users can click on the (?) which is next to the drop-down menu for each parameter, to select values. Novice users would have an additional help screen that pops up after clicking on the (?) for guidance. The outputs are also customized based on the type of users. For the given input in Fig. 10, the output to novice users is a message generalized as a cautionary advice as shown in Fig. 11. For intermediate users, the message has additional details on air quality and PM2.5 ranges along with the cautionary advice as illustrated in Fig. 12.

The example in Fig. 13 shows input conditions entered directly by an expert user (without parameter selection) that lead to the prediction of a safe PM2.5 range. In addition to estimating the PM2.5 range numerically, there is a message given to the user, “Much less than the 12 μg/m^3 regulation, minimal negative impact”. This takes into account the air quality standards of health as prescribed by EPA which would be useful to users in the expert category. Likewise, Fig. 14 shows an example of input conditions entered by an expert that give prediction of a moderate PM2.5 range.

The health impacts, predicted by this tool during the air quality estimation, are helpful with respect to result interpretation. This method is better than simply estimating the output as PM2.5 ranges, since specifically indicating the corresponding health impacts is more meaningful.

There are other screens available in the prediction tool that allow users to see more details on the data mining analyses leading to the results. These screens build upon on the simple excerpts of association rule mining, clustering and classification results (already shown earlier) and are more illustrative. An example of such a screen appears in Fig. 15 for a partial snapshot of a corresponding decision tree.

Note that the leaves of this tree appear in green with root, branches and intermediate nodes in various shades of brown, thus being in line with illustrative appeal for interpretation. Moreover, the leaves appear at different levels, indicating the lengths of the respective paths traversed to reach them. For example, it is clear that if the region is East Asia and Pacific and the gasoline consumption is greater than 427.7 units, the corresponding PM2.5 range (i.e., the leaf) would be in the range of 0–5.85 μg/m^3 so the decision is reached in three levels. On the other hand, the path is relatively longer for gasoline consumption being less than or equal to 427.7 units with income group being upper middle income, where the decision on PM2.5 range (leaf) is further based on diesel consumption and is reached in five levels. Moreover, such illustration facilitates easy comprehension of partial matches as well. Accordingly, the respective screens for other examples are available by clicking options from the tool’s menu, to provide further knowledge on the causes and effects of pollutants and to enable users in understanding the precise reasons that lead to the corresponding estimations.

Likewise, if the users need to understand the reasons for health impacts based on PM2.5 prediction levels, the AQI table, with air quality index values and the corresponding levels of health, provides easy comprehension. Figure 16 shows this standardized AQI table [1] with color coding as displayed in our tool in a user-friendly manner. Note that the time horizon for the prediction in the tool depends on the duration of these AQI values as per their health levels. In other words, if these mappings are valid for the next 5 years, then the estimation provided by the prediction tool is valid for that duration as well. The space horizon of this tool is worldwide since we analyze multicity data from across the globe and use globally accepted AQI standards along with their health impacts.

All such illustrative examples including the prediction inputs/outputs, visualizations, reasoning behind predictions, and corresponding AQI standards portray the convenient user input and result interpretation in the prediction tool as per our problem definition. We now proceed further with the tool assessment.

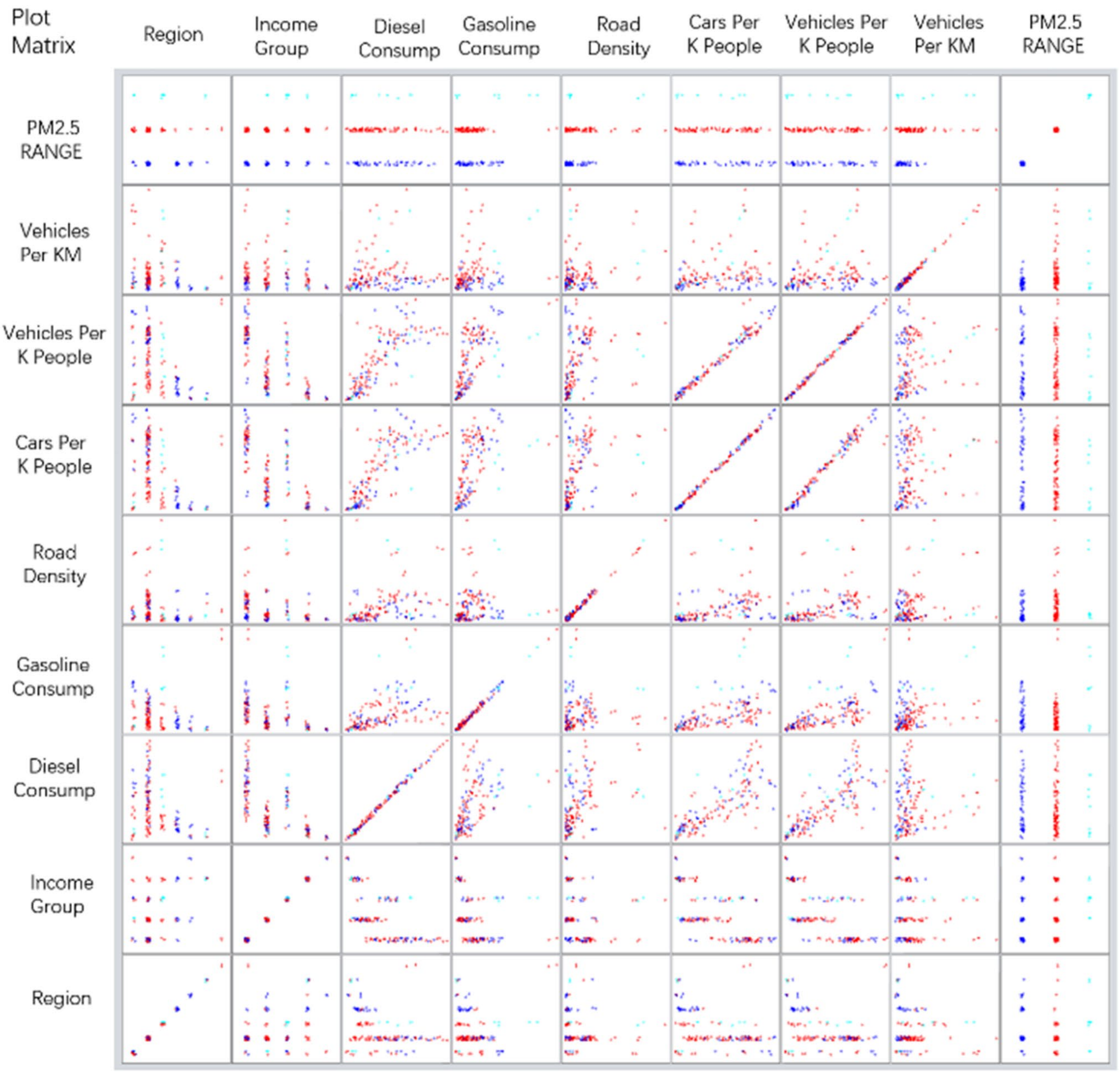


Fig. 8 Detailed visualization with scatter plots (for experts)

Tool Assessment by Targeted Users

The PM2.5 prediction tool has undergone assessment by various users: new students, environmental engineering researchers, working professionals, city residents etc. There are two questionnaires. The User Level Questionnaire (shown earlier) determines their level: novice, intermediate, expert. Furthermore, the Tool Reception Questionnaire, as shown next, serves to obtain their feedback on the performance of the prediction tool.

Tool Reception Questionnaire

For each of the following questions, please select your response as: strongly disagree/disagree/neutral/agree/strongly agree.

Q1. The tool has facilities for convenient inputs.

Q2. The tool is efficient in estimation.

Q3. The outputs with messages are useful.

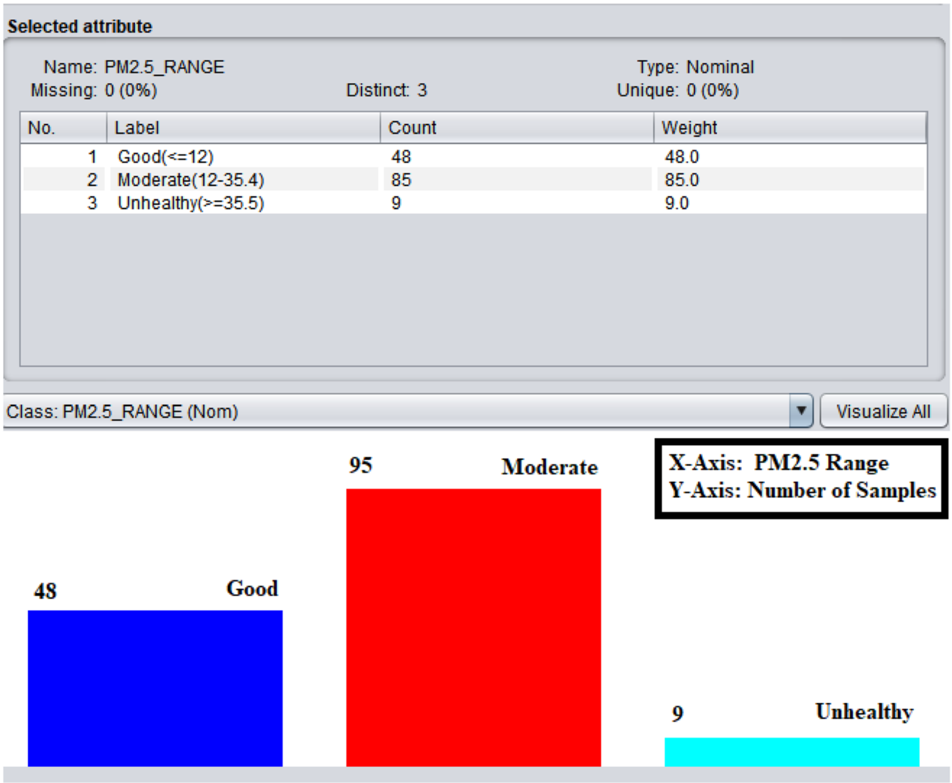
Q4. Visualization of parameters is helpful.

Q5. Overall, the tool is user friendly.

Q6. This tool motivates me to study the area further.

Considering these assessment results, we summarize observations for positive responses by the targeted users over all the questions, i.e., percentage of "strongly agree"

Fig. 9 Simple table and chart visualization (for novice users)

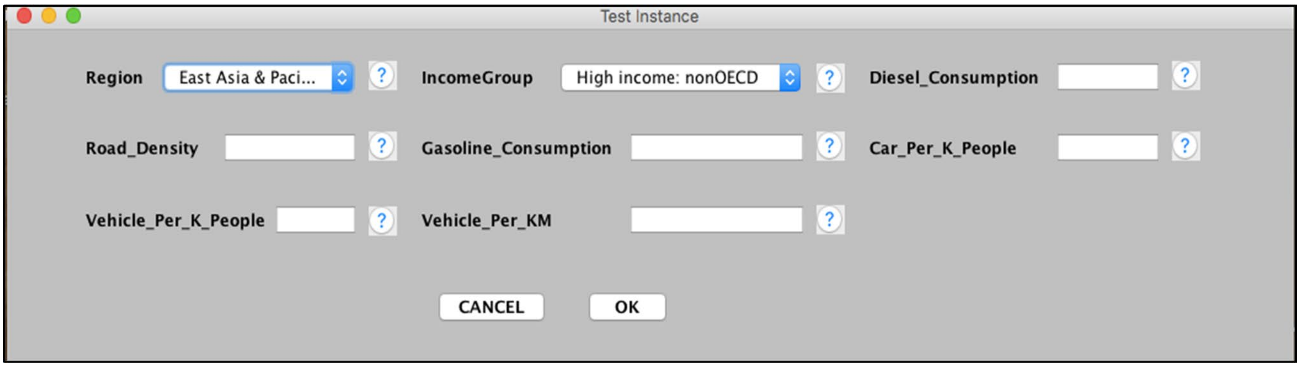


Selected attribute

Name: PM2.5_RANGE Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Good(<=12)	48	48.0
2	Moderate(12-35.4)	85	85.0
3	Unhealthy(>=35.5)	9	9.0

Class: PM2.5_RANGE (Nom)

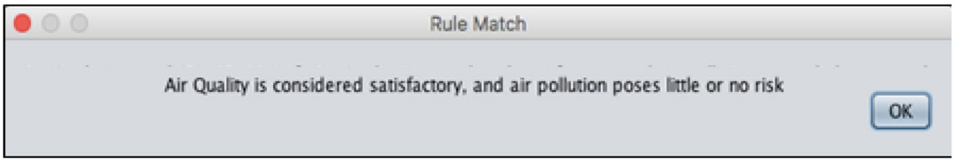


Test Instance

Region: East Asia & Paci...
IncomeGroup: High income: nonOECD
Diesel_Consumption
Road_Density
Gasoline_Consumption
Car_Per_K_People
Vehicle_Per_K_People
Vehicle_Per_KM

Fig. 10 Input screen with example entry

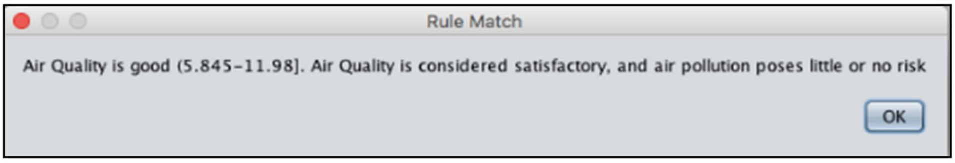
Fig. 11 Example of predicted output for novice user



Rule Match

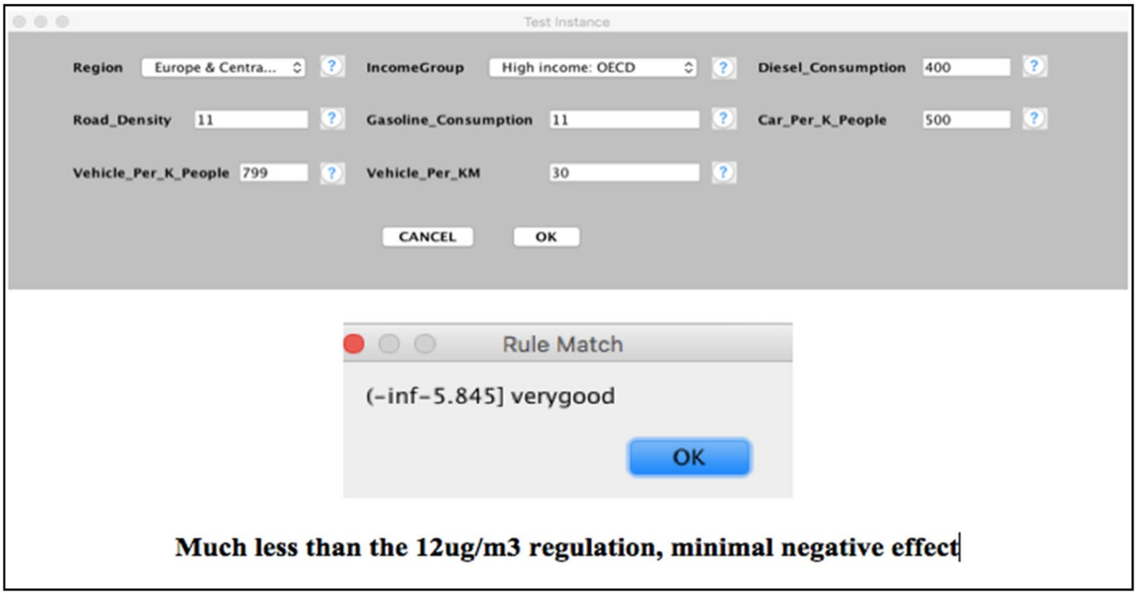
Air Quality is considered satisfactory, and air pollution poses little or no risk

Fig. 12 Example of predicted output for intermediate user



Rule Match

Air Quality is good (5.845-11.98]. Air Quality is considered satisfactory, and air pollution poses little or no risk

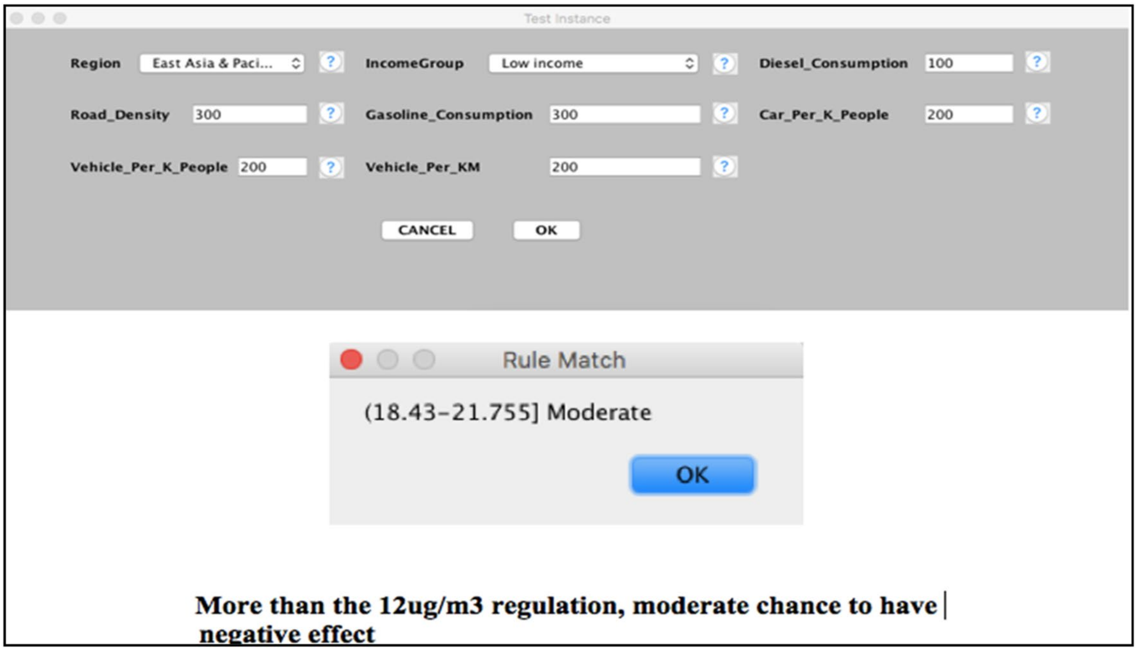


Region: Europe & Centra...
IncomeGroup: High income: OECD
Diesel_Consumption: 400
Road_Density: 11
Gasoline_Consumption: 11
Car_Per_K_People: 500
Vehicle_Per_K_People: 799
Vehicle_Per_KM: 30

(-inf-5.845] verygood

Much less than the 12ug/m3 regulation, minimal negative effect

Fig. 13 Example of expert user input with predicted output for safe PM2.5 range



Region: East Asia & Paci...
IncomeGroup: Low income
Diesel_Consumption: 100
Road_Density: 300
Gasoline_Consumption: 300
Car_Per_K_People: 200
Vehicle_Per_K_People: 200
Vehicle_Per_KM: 200

(18.43-21.755] Moderate

More than the 12ug/m3 regulation, moderate chance to have negative effect

Fig. 14 Example of expert user input with predicted output for moderate PM2.5 range

and "agree" responses, with their averages. These appear in Table 4 herewith. As per the table, we can see that the average percentage of responses for "strongly agree" is 29.33%, while that for "agree" is 52.67%, giving a combined average of 82% as a positive response measure over all the questions. It is thus evident that users in all categories are satisfied with the performance of the prediction tool.

In addition, expert users evaluate the accuracy of the tool to assess how well the air quality estimation occurs, taking into account PM2.5 prediction. Since they are domain experts in environmental engineering, they are able to assess the effectiveness of the estimation provided by the tool with respect to accuracy. Hence, expert users take a tool accuracy questionnaire with on the accuracy of the tool itself as well

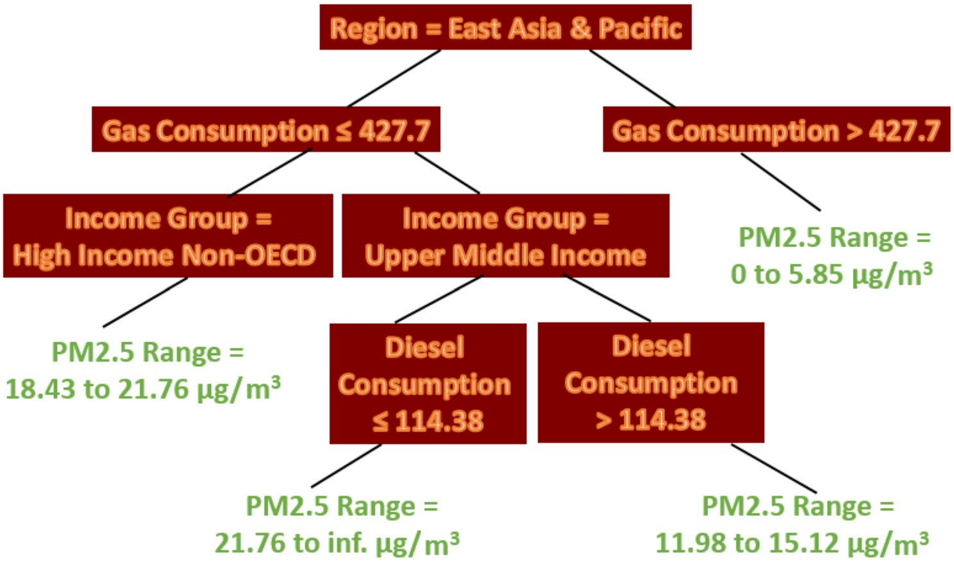
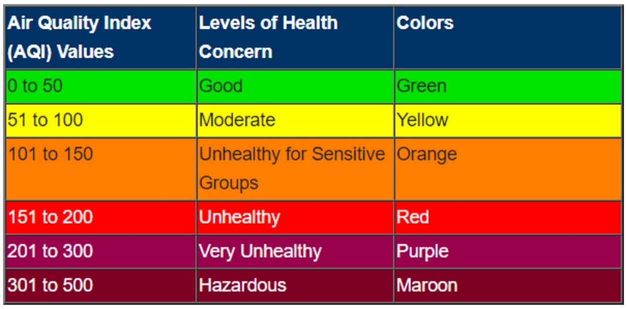

Region = East Asia & Pacific
Gas Consumption ≤ 427.7
Gas Consumption > 427.7
Income Group = High Income Non-OECD
Income Group = Upper Middle Income
PM2.5 Range = 0 to 5.85 µg/m³
PM2.5 Range = 18.43 to 21.76 µg/m³
Diesel Consumption ≤ 114.38
Diesel Consumption > 114.38
PM2.5 Range = 21.76 to inf. µg/m³
PM2.5 Range = 11.98 to 15.12 µg/m³

Fig. 15 Illustrative partial snapshot of decision tree example



Air Quality Index (AQI) Values	Levels of Health Concern	Colors
0 to 50	Good	Green
51 to 100	Moderate	Yellow
101 to 150	Unhealthy for Sensitive Groups	Orange
151 to 200	Unhealthy	Red
201 to 300	Very Unhealthy	Purple
301 to 500	Hazardous	Maroon

Fig. 16 AQI values and health impacts with levels of concern

Table 4 Positive responses for all questions by targeted users

User level	Strongly agree (%)	Agree (%)
Novice	32	51
Moderate	23	53
Expert	33	54
Average	29.33	52.67

as its usefulness (based on accuracy). The questions are as follows.

Tool Accuracy Questionnaire

Q7. How accurate is the prediction tool? (very accurate, accurate, medium, fairly accurate, inaccurate).

Q8. Based on the accuracy, this tool is: (extremely useful, useful, somewhat useful, not so useful, not at all useful).

The responses to this tool accuracy questionnaire appear in the following figures, i.e., Figures 17 and 18, respectively on Questions 7 and 8 herewith. These responses indicate that the prediction tool is accurate/very accurate, and useful/extremely useful, considering the assessment of all the expert users as observed in the figures herewith. More specifically, around 53% of the experts indicate that the tool is "very accurate", while around 42% consider it as "accurate".

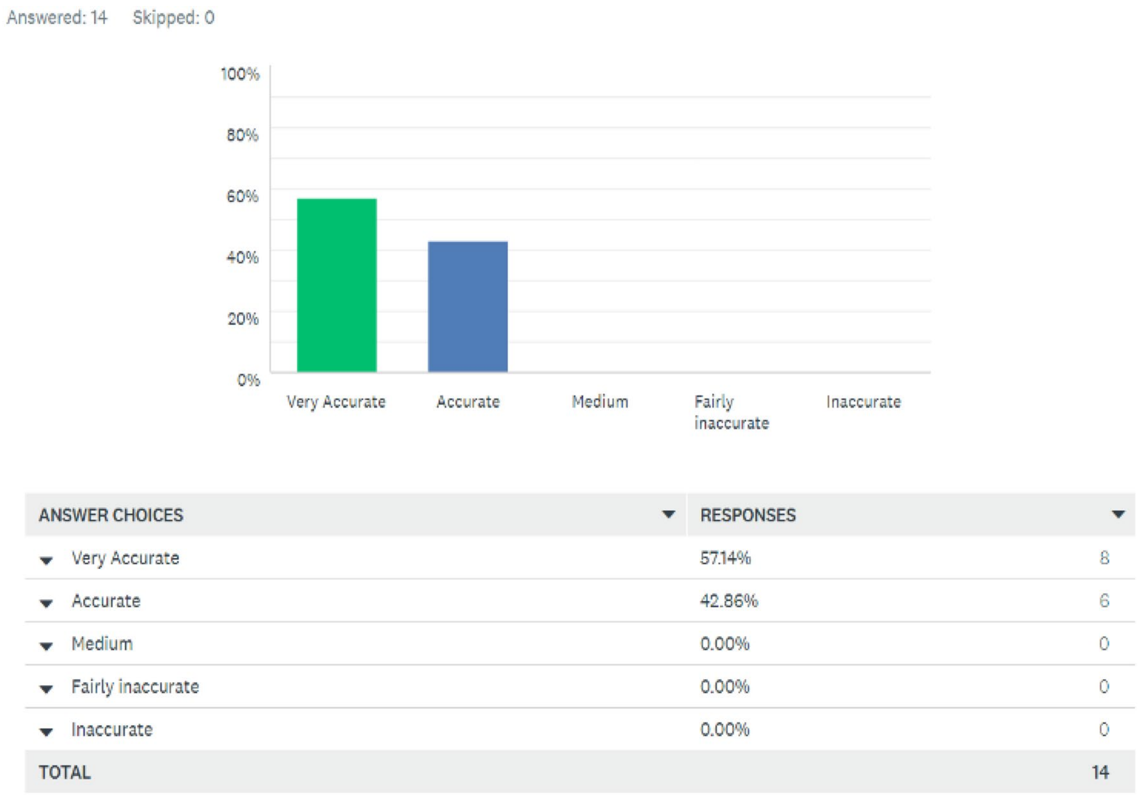
Likewise, high scores are obtained for the responses to the tool's usefulness based on its accuracy, where a large majority considers the tool "useful" and a considerable majority finds it "extremely useful". Hence, we can infer that this tool is accurate in air quality estimation incorporating health impacts as indicated by domain experts in environmental engineering. This is corroborated by the fact that the PM2.5 analytical framework gives classification accuracies close to the range of around 90% for unseen test data (subsection on decision tree classifier experiments). This is the reason that the results are used to design the prediction tool, giving high accuracy.

The interactive menu-driven user-friendly GUI of this tool is very appealing to the targeted users. Some users indicate that this tool motivates them to conduct further studies, which is an added plus. Moreover, this tool is very efficient with prediction speeds in the order of microseconds, which makes it quite useful for ubiquitous access. We now proceed with its applications.

Applications and Discussion

Developing Software Systems with Prediction Framework

We envision a few applications of the framework in our prediction tool, as found useful in development of pertinent software systems in interesting applications. We focus on



Answered: 14 Skipped: 0

ANSWER CHOICES	RESPONSES	
Very Accurate	57.14%	8
Accurate	42.86%	6
Medium	0.00%	0
Fairly inaccurate	0.00%	0
Inaccurate	0.00%	0
TOTAL		14

Fig. 17 Responses to accuracy of prediction tool

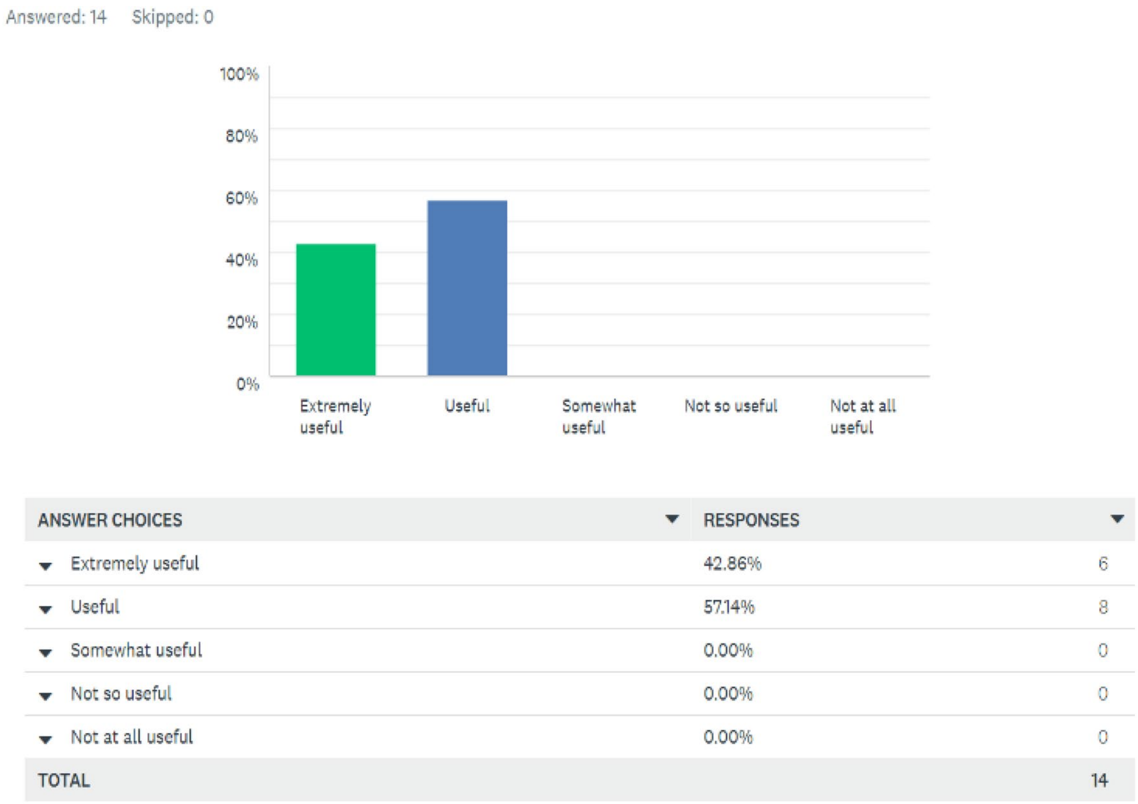
intelligent tutoring systems, mobile devices and smart city tools as illustrated in Fig. 19 and discussed next.

Intelligent tutoring systems: In fields such as data mining and environmental management, a tool for pollution prediction can potentially be good to embed within an intelligent tutoring system (ITS) or computer-based tutor (CBT). An ITS or CBT that has been in the literature for a long time [56] is a computer-based system for tutoring by incorporating the features of artificial intelligence, typically designed to simulate or assist human tutors. There are ITS applications in various areas, e.g., ELM-ART for learning LISP [57]. In line with many such systems, embedding the functionality of our prediction tool within an ITS can be advantageous to both students and teachers. Convenient components of this tool can help students acquire relevant knowledge of fine particle pollutants and their impacts on health. The functions for comprehension and assistance can provide great help to educators in conveying and assessing the details of the course material in the concerned fields.

An ITS can exemplify the applications of knowledge discovery in scientific areas. In environmental engineering, it can help in urban planning and related areas such as urban simulation. It can be useful to design case studies and experimental testbeds for learning principles through practice. Our prediction tool would be useful here for presenting the causes and effects of pollution, particularly since it caters to novice, intermediate and expert users. Studies on sustainability can potentially be the focus of such intelligent tutoring systems if used for education from K-12 through university curricula, to make students understand the importance of the environment.

Mobile devices: As is widely known, the IoT refers to the extension of Internet connectivity to various physical systems. Endowed with the Internet connectivity and hardware, these systems can communicate with others through the Internet, and can be remotely monitored [43, 44]. In line with IoT, a mobile application (app) is a computer program developed to run on devices such as smartphones and tablets [58]. There is recent work in this area [45, 59, 60] on aspects such as IoT enhancement, mobile device research and app development.

Our prediction tool could foster building suitable apps. Though Websites on computers provide convenient access to the tool, there is often the need for ubiquitous access through mobile devices. This motivates development of mobile apps to predict air quality and related aspects, e.g., pertinent healthcare apps that relate PM2.5 prediction to health advisory outputs. Users could potentially access such apps as easily as they access weather apps today. For instance, there has been much interest recently on health apps and related



Answered: 14 Skipped: 0

ANSWER CHOICES ▼	RESPONSES ▼	
▼ Extremely useful	42.86%	6
▼ Useful	57.14%	8
▼ Somewhat useful	0.00%	0
▼ Not so useful	0.00%	0
▼ Not at all useful	0.00%	0
TOTAL		14

Fig. 18 Responses to usefulness of tool (based on accuracy)

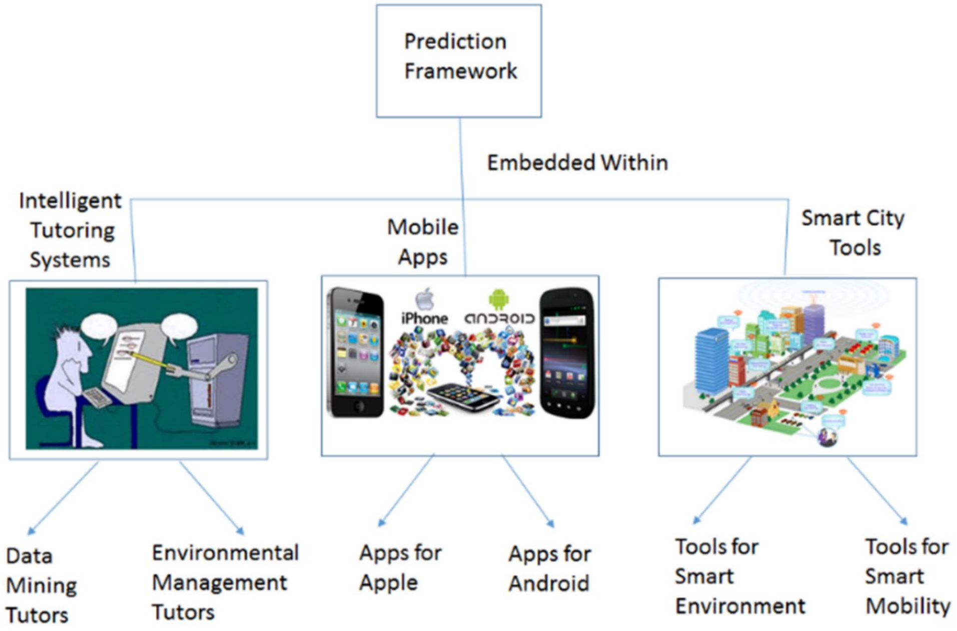


Fig. 19 Software systems with prediction framework: intelligent tutors, mobile apps and smart city tools

research on COVID-19, e.g., [59] in geographical tracking. There could be further research on similar lines pertaining to PM2.5 and its health impacts with advisory apps.

Any apps adapting our prediction tool within them are expected to support the Android and/or Apple platforms. Users on mobile platforms typically expect extremely fast inputs and outputs. In addition, storage space on mobile devices is usually limited. These issues thus pose challenges. Some challenges related to minor speed and storage enhancement issues can be addressed using existing techniques in mobile computing along with adequate implementation. Since our prediction tool is highly efficient with speeds in µs, it is very convenient for adaptation in mobile computing.

Challenges of a bigger magnitude are likely to call for further research. This could be with reference to domain-specific issues, unresolved space problems, enhanced efficiency needs and scalability to newer models (e.g., iPhone 8 to iPhone X). Thus, the framework of our prediction tool potentially used within mobile apps is likely to motivate further research. Hence, in addition to the direct impact of embedding it within mobile apps using state-of-the-art technology, there is the broader impact of inspiring more research and development in the mobile app area.

Smart city tools: In recent years, there is tremendous interest in development of smart cities. A smart city is often identified by characteristics [2, 35] as follows.

- *Smart environment* Green energy, pollution control, sustainable resources …
- *Smart people* Twenty-first century education, level of qualification, cosmopolitan society …
- *Smart governance* Transparency and open data, e-government, public and social services, participation in decisions …
- *Smart mobility* Local accessibility, sustainable and innovative transport, safe systems, ICT (information and communication technology) …
- *Smart economy* Entrepreneurship, economic image and trademarks, innovative spirit …
- *Smart living* Health conditions, housing, amenities, individual safety, cultural facilities …

We claim that our prediction tool can be useful within specific smart city applications, e.g., the development of more advanced large-scale prediction systems. For example, the smart city of Amsterdam has canal lights that automatically brighten and dim depending on pedestrian usage [2, 36]. These are striking aspects of traffic systems. Such advanced features are useful while collecting data for prediction, thus making the analysis of the pollution more sophisticated. It would be very interesting to predict PM2.5 ranges and estimate AQI levels in places where the traffic mechanisms are more environment friendly. Comparative studies can be conducted based on the adaptation of our prediction tool, considering data from highly ranked smart cities such as Amsterdam and Copenhagen in Europe, versus cities lower in the ranking that are aiming to get better. Tailor-made smart city tools can potentially be designed for such studies, to automate the comparisons and derive interesting conclusions. Hence, our prediction framework can be particularly useful in tools for smart environment and smart mobility due to its prospective deployment as described herewith that would help to green the environment and also make transportation systems more effective.

Hence, the prediction tool in this article can be a prospective addition to large-scale smart city tools that can embed it for specific applications. Our prediction tool makes a significant impact on smart environment, by helping to understand specific causes of fine particle pollutants, thereby aiding their mitigation. It also helps in air quality estimation based on health impacts, thus contributing to smart living for promoting sustainable lifestyles. Overall, our PM2.5 prediction tool has the broader impact of contributing to smart cities analogous to works such as [37, 38] and improving urban sustainability.

Decision Support for Urban Users

Various urban user groups can benefit from the prediction tool that can assist decision-making in numerous facets. Targeted user groups include urban residents (city dwellers) in the novice category, urban planners (e.g., management agencies) as intermediate and urban researchers (e.g., environmental engineers) as expert users. Others can be in one of these categories, e.g., new students in environmental engineering as novices, multidisciplinary researchers as intermediate/expert (based on User Level Questionnaire). All these users would benefit from the tool as portrayed in Fig. 20 and explained next.

Urban residents: The urban residents or regular city dwellers are directly influenced by air pollution. The benefits to this group are twofold since they can: adjust present lifestyles; and prepare for prospective relocations. They can potentially use this prediction tool to estimate pollutant concentrations in their vicinity. The results can inform them about corresponding health effects. This can help them make decisions about adjusting their present lifestyles in a given urban area, e.g. getting more frequent health checkups if pollutant concentrations are not in a safe range. If they contemplate a move to a different area, the knowledge of conditions therein and estimation of PM2.5 ranges can be useful to them. The tool can help them make decisions about whether to relocate. If they decide to relocate, they can be prepared based on air quality estimation there, e.g., they can aim to use greater public transport (for environmental greenness) or buy more cars (for their own convenience), considering how that affects pollution.

It can be seen that this positively influences the smart environment characteristic of smart cities [2, 35], which includes green energy and clean air as its priorities. To some extent, it affects smart mobility, as it addresses traffic issues, by propelling decisions of public versus private transport.

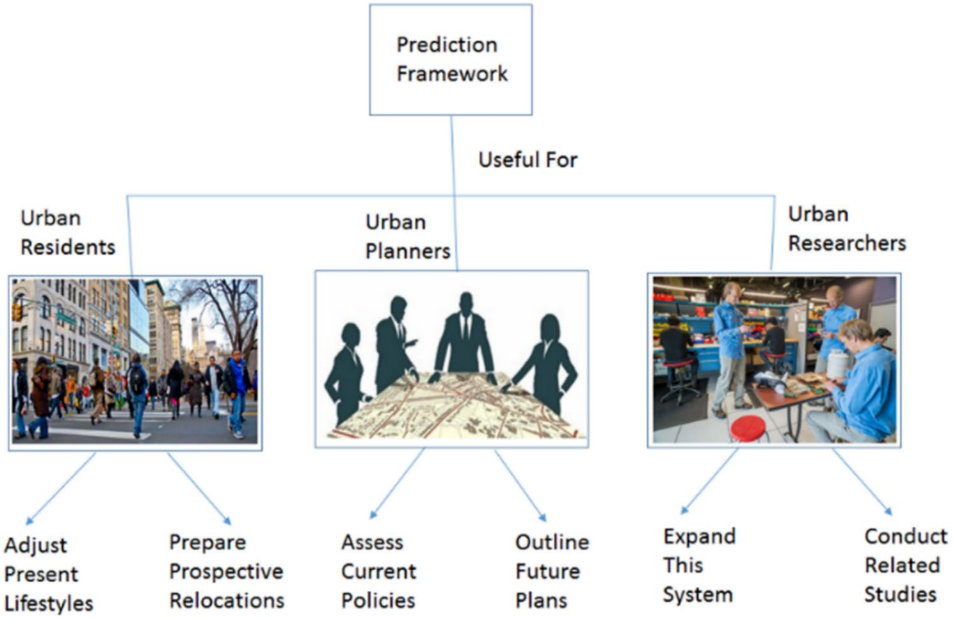


Fig. 20 Decision support for urban users with prediction tool

Urban planners: The urban planners, such as lawmakers and management agencies form the user group responsible for outlining local laws and managing the concerned urban area. These local laws and policies could pertain to the regulation of air quality and urban traffic facilities. Urban planners can benefit from the knowledge about the relationships between PM2.5 ranges and traffic conditions since it would primarily help them to: assess current issues; and outline future plans. The prediction tool herewith can provide information on how the current traffic conditions influence PM2.5 ranges and hence air quality. Given this information, lawmakers could introduce policies to adjust traffic conditions and related regulations such that they minimize pollutant concentrations. Furthermore, if proposed plans in any department lead to changes in traffic conditions, this prediction tool could estimate its influence on air quality in advance. This would help the urban agencies re-evaluate their plans and make decisions accordingly, catering to good health and safety, and thereby heading considerably toward greater public satisfaction.

This has direct impacts on the smart living characteristic of smart cities [2, 35], which entails health issues among other things. It also indirectly affects smart governance, which incorporates greater transparency through public feedback, tools for decision support etc. Our prediction tool can thus aid decision support in urban planning, thereby contributing to smart governance.

Urban researchers: The urban researchers, i.e., environmental engineers, earth scientists and other related professionals constitute the user group that deals with the long-term impacts of this work. They can further the research herewith and provide benefits to public welfare. They can utilize our prediction tool, mainly in two ways: expand this system; and conduct related studies. The tool in this work has the potential to expand with more data and other parameters of study. Urban researchers can gather other suitable data sets and reuse our predictive analytics framework to conduct estimation of pollutants besides PM2.5 (e.g., PM10: pollutants of diameter < 10 μm). They can use factors besides traffic conditions to predict air quality, thus adding more parameters in the estimation. In addition, urban researchers can conduct studies on related issues through the knowledge gained from this prediction tool and its impact on decision support. For example, the causes of pollution outlined here along with its health effects can motivate further research on the prevention of health hazards. This research can be supplementary to works such as [20, 59] that overlap computational research with healthcare. Researchers can potentially address issues such as the respiratory impacts of pollutants in highly affected urban areas, and develop suitable precautionary devices, e.g., high quality protective masks for people and enhanced filters for car emissions. Some of this work can advance beyond the state of the art. Our prediction tool would thus help to provide estimations that propel further research in related areas.

This has broader impacts on smart cities [2, 35]. Studying causes of pollution and discovering new trends from large volumes of data helps in working toward mitigating pollution and its harmful effects. This influences the smart environment characteristic. Since health issues are addressed in the prediction tool, this would motivate healthcare professionals to work with environmental researchers to counterbalance effects of pollution and improve living conditions. This could include developing stronger masks and filters.

Hence, this affects the smart living characteristic, mainly due to its health impacts.

Conclusions, Discussion and Future Work

In this work, we design a prediction tool for fine particle air pollutants, namely PM2.5, based on data mining over urban traffic conditions. Association rules and clustering help in understanding the causes and effects of pollution, while decision tree classifiers play a major role in predicting the target, i.e., PM2.5 range. Our research in this paper fits in the broader context of metaheuristic approaches to solve problems in the overall area of Environmental Engineering. In conclusion, we present a discussion of other approaches in this general category, and thereafter highlight our own contributions in proper context.

Some researchers [61] use an ensemble approach with three machine learning techniques to predict PM2.5 presence in the Greater London locale. Everyday predictions are made in a grid-based manner with the deployment of random forest (RF), gradient boosting machine (GBM), and artificial neural network (ANN) classifiers. Hyper-parameter tuning for the classifiers leverages grid search with mean square error and cross-validation. Input variables for all classifiers include temperature, barometric pressure, cloudiness, wind speed, wind direction, day of week, day of year, traffic counts, population density, land use type, distance to Heathrow, distance to water etc. In this study, RF and GBM outperform ANN. The final ensemble model provides prediction accuracy high enough to depict a strong ability to predict PM2.5 levels in Greater London. While this is an interesting study, it only considers classifiers; while our work in this paper adapts other machine learning methods such as association rules and clustering, in addition to classification to draw interesting inferences, including some rather surprising observations about PM2.5 levels and associated traffic conditions that constitute significant findings. Also, while their study focuses on a fixed geographical location, we consider a multicity situation with a global context.

Health investigation for the year 2015 is conducted in a study [62] via high spatio-temporal resolution to estimate PM2.5 impacts in the USA. This study considers a spatial resolution of 1 km, through a concentration–response curve obtained by meta-analysis of long-term PM2.5 exposure, taking into account novel observations at high and low PM2.5 levels. They calculate alterations in the number of fatalities considering the overall population, county level baselines, and reduction of PM2.5 levels by 20% and 40%, respectively. This helps to gauge the influence of PM2.5 on extreme situations that could be dangerous to the extent of being fatal. Important findings of their study include the fact that fatalities would reduce by 47,775 and 96,787 for 20% and 40% decrease in annual PM2.5 levels, respectively. This research offers substantial health benefits anticipated by lowering PM2.5 levels, including those regions fairly low PM2.5 already. It is very interesting work, however, they do not actual predict the PM2.5 levels themselves based on other parameters but rather use the PM2.5 concentrations directly to study health impacts. Our work in this paper goes a step beyond, in the sense of actually estimating the PM2.5 levels in addition to their health concerns, by discovering knowledge from existing multicity data on traffic parameters via machine learning techniques adapted for environmental studies.

In a recent piece of research [63], the authors address survival analysis, considering the cases of people who have experienced PM2.5 related events, with a focus on time-varying exposure. Time-invariant or slowly varying factors are kept under control. The authors predict the influence of annual air pollution on survival, with a focus on the Medicare cohort in Northeast USA during the time period 2000 to 2013. They employ a hybrid prediction methodology via the utilization of satellite remote sensing. The granularity of the analysis is at the level of zip codes of the concerned population. They control linear and quadratic terms for both age as well as year. The population of focus is of the age group 65–66, to disregard older Medicare subjects, e.g., in those exposed to higher pollution exposure in 2000. They observe that 1,254,706 fatalities occur in this cohort with the PM2.5 level being 10.5 μg/m^3 on an average. This result offers considerable evidence for causality of the association. It is an excellent recent study on cause–effect analysis. However, the subjects analyzed are too specific and span only a limited region as well as time period. Our research extends to a global scale with a time horizon that is proportional to the timeline of the data collected from sources such as the World Health Organization and World Bank. It is extendable to any other scale and horizon since the techniques are generic, and our framework is reusable. Moreover, we build an actual tool for the PM2.5 analysis and air quality prediction taking into account health standards imposed by the EPA o/f the USA, accepted worldwide. The tool is found useful by targeted users, and its framework can be embedded in other related tools.

In another recent study [64], correlations of long-term exposure to pollutants and hospital admissions are explored with respect to cardiovascular and respiratory diseases. The subjects of study constitute Medicare users in the USA. This study is inspired by the fact that it would have a utility value in risk assessment and effect modification. The researchers in this work devise a doubly robust additive hazard model abbreviated as DRAHM for evaluating the impacts of chronic PM2.5 exposure on hospital admissions. They consider illnesses such as stroke, myocardial infarction (MI), lung cancer, heart failure, pneumonia, and chronic obstructive pulmonary disease (COPD) for Medicare subjects during the time frame 2000 to 2013. Their model, i.e., DRAHM unbiased under the condition that one of the following is appropriately obtained: inverse probability weight (IPW) model or outcome regression model. PM2.5 levels in this study are derived using erstwhile high-resolution prediction models via machine learning techniques, and are spatio-temporally averaged with zip code granularity. The parameters considered as covariates here encompass demographic and socioeconomic factors. Data is analyzed incorporating gender, age, ethnicity, etc. The analyses are re-conducted for observations yielding less than the federal standard of 12 mg/m^3 for PM2.5 since those correspond to highly sensitive situations. The outcomes of this study indicate that chronic exposure to air pollution, especially with the presence of PM2.5 particles, augments the risks associated with cardiovascular and respiratory diseases on an additive scale, thereby increasing hospital admissions. Such studies provide further inspiration for our work, particularly due to their emphasis on the potentially hazardous effects of PM2.5 from a health perspective. In our work, we thrive on classical data mining and machine learning techniques with the claim that these can be used to solve challenging problems pertaining to the environment. We thus do not build a completely new model, instead we amalgamate the techniques of association rules, clustering and classification into an interesting predictive framework for PM2.5 analysis and air quality estimation as per globally accepted health standards. Progressing a step beyond some works in the literature, we actually provide a user-friendly, interactive tool for PM2.5 prediction with convenient access for novice, intermediate and expert users. Furthermore, our prediction framework itself is reusable and has the potential for further utilization within other pieces of software, e.g., mobile apps and intelligent tutors systems.

On the whole, this paper is orthogonal to the existing literature, and provides the 2 cents in the area of machine learning research along with environmental studies. Our work has some notable impacts as emphasized next. This paper constitutes “Applied Research” and makes the following contributions that we present as the highlights/findings of our work.

1. Designing among the first prediction tools for air quality, based on fine particle pollutants, leveraging health impacts (to the best of our knowledge).
2. Addressing a space horizon for prediction that is global due to multicity data analysis along with a time horizon that is specific to the duration of the AQI-based health impacts.
3. Discovering useful relationships between urban traffic and PM2.5 data that constitute findings, e.g.,
   (a) traffic volume per se, is not directly proportional to PM2.5 emissions,
   (b) "region" attributes have significant effects on PM2.5 concentrations,
   (c) high gasoline and diesel consumption does not always cause unsafe PM2.5 ranges,
   (d) economic conditions highly influence the presence of PM2.5 in air.
4. Addressing challenges in the execution of classical mining techniques over PM2.5 data, by combining empirical and theoretical approaches, harnessing domain knowledge
5. Obtaining high accuracy in prediction as evident from surveys conducted by domain experts in environmental engineering, further corroborating the validation obtained by achieving around 90% accuracy in classifying unseen test data in our experimental evaluation.
6. Achieving prediction efficiency of the order of microseconds, useful for ubiquitous applications.
7. Catering well to novice, intermediate and expert users of the tool, as evident from 82% positive responses on tool reception questions in surveys.
8. Identifying specific causes of pollution through association rules and clustering, along with predicting future levels using decision trees, thus taking steps toward pollution mitigation.
9. Outlining suitable applications within intelligent tutoring systems, mobile apps and smart city tools to highlight the usefulness of our prediction framework in developing other software systems.
10. Describing the benefits of this prediction tool in decision support for targeted users such as urban residents, urban planners and urban researchers.
11. Proving the claim that classical data mining paradigms in their fundamental form can be adapted to solve problems in environmental engineering with simplicity, accuracy and efficiency, thus advocating the Occam’s razor principle.
12. Making positive impacts on smart environment, mobility and living characteristics of smart cities.

This applied research can propel future work overlapping computer science areas such as data mining, machine learning and the Web, with domains such as environmental engineering, earth science and urban studies. We outline some open areas for further research as follows.

1. Conducting in-depth research on the environmental issues emerging from this study, e.g., mitigating the air pollution based on understanding its causes (through outcomes of association rules, clustering and decision

trees here), thus aiming to improve air quality with respect to health standards.
2. Exploring other techniques in data mining, e.g. fuzzy k-means, convolutional neural networks (CNN) and bidirectional long short-term memory (Bi-LSTM) to conduct advanced studies on related issues.
3. Addressing pollutants other than PM2.5, e.g. PM10 pollutants with larger diameters, to comprehend their impacts on air quality, and conducting comparative research beyond the studies in the literature.
4. Targeting data sources other than multicity urban traffic data from publicly available Web sources, for analyzing the potential impacts of other data on air pollution, with the goal of building decision support systems that would be useful in the future.
5. Mining opinions on pollution control from Web data in social media to fathom the reactions of the public on environmental issues, and using the results of the mining for decision-making on urban policies, to achieve transparence in governance as expected in smart cities.
6. Performing topic modeling by analyzing textual data from the Web pertaining to climate change, global warming and related issues that pertain to pollution and air quality.
7. Developing some of the software systems outlined here, i.e., intelligent tutors, mobile apps and smart city tools based on this prediction framework by suitably embedding its functionality within the concerned systems, and also conducting further research there, e.g., in healthcare apps for mobile devices and in large-scale decision support systems for smart cities.
8. Adapting the outcomes of this study to update relevant curricula in universities as well as in *K*-12 schools for making students more aware of the importance of safeguarding the environment via suitable measures that everyone can follow, thus constituting a broader impact of this work.

This applied research article would be beneficial to data mining, machine learning, intelligent systems, Web data analysis, environmental engineering and urban studies as well as related areas, e.g., healthcare and IoT. The methods and materials used here are reproducible to conduct studies with other data sources. We envisage that our work would encourage advanced research in common areas between computer science and domains such as environmental engineering. This work makes broader impacts on smart cities and urban sustainability.

Acknowledgements Aparna Varde is supported by grants from NSF "MRI: Acquisition of a High-Performance GPU Cluster for Research and Education" Award Number 2018575, and "MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross-Disciplinary Human-Centered Research and Education at Montclair State University", Award Number 2117308. In addition, she has support through the Faculty Scholarship Program and the Doctoral Faculty Program at Montclair State University. During her sabbatical in the autumn 2021 semester, she is a visiting researcher in the Databases and Information Systems group of Dr. Gerhard Weikum at the Max Planck Institute for Informatics, Saarbrücken, Germany. Abidha Pandey has been funded by a Research Assistantship and thereafter a Teaching Assistantship from the Department of Computer Science. Xu Du has received a Doctoral Assistantship from the Environmental Management Ph.D. Program. We would like to thank Dr. David Brown, Professor Emeritus at Worcester Polytechnic Institute for his valuable feedback on this work.

Author Contributions The author contributions are approximately in the order in which their names appear in this article.

Funding Aparna Varde acknowledges grants from NSF "MRI: Acquisition of a High-Performance GPU Cluster for Research and Education" Award Number 2018575, and "MRI: Acquisition of a Multimodal Collaborative Robot System (MCROS) to Support Cross-Disciplinary Human-Centered Research and Education at Montclair State University", Award Number 2117308. She is also supported by a Faculty Scholarship Program and a Doctoral Faculty Program from Montclair State University (MSU). Abidha Pandey has been supported by a Research Assistantship as well as a Teaching Assistantship from MSU. Xu Du has been supported by a Doctoral Assistantship from MSU.

Availability of Data and Materials The data used for this analysis is from publicly available sources such as the World Health Organization (WHO). The software developed in this work, namely, the prediction tool is on GitHub and can be made available to interested users upon request.

Code Availability The authors of this work retain the ownership of the code. The tool resulting from this work is free and open source and the authors can provide access upon request.

Declarations

Conflict of interest The authors have no conflicts of interest/competing interests to declare to the best of their knowledge.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication The authors give consent to publish this article.

References

1. AQI-Revised PM2.5 AQI breakpoints. http://aqicn.org/faq/2013-09-09/revised-pm25-aqi-breakpoints/
2. IEEE Smart Cities, https://smartcities.ieee.org/
3. Pope C III. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. J Am Med Assoc (JAMA). 2002;287(9):1132.
4. Batterman S. Temporal and spatial variation in allocating annual traffic activity across an urban region and implications for air quality assessments. Transp Res. 2015;41:401–15.

5. Zikova N, Wang Y, Yang F, Li X, Tian M, Hopke P. On the source contribution to Beijing PM2.5 concentrations. Atmos Environ. 2016;134:84–95.
6. Dockery D. Acute respiratory effects of particulate air pollution. Annu Rev Public Health. 1994;15(1):107–32.
7. Rom W, Samet J. Small particles with big effects. Am J Respir Crit Care Med. 2006;173(4):365–6.
8. Particulate Air Pollution Associated with Kidney Disease, http://www.frackcheckwv.net/2017/09/25/particulate-air-pollution-associated-with-kidney-disease
9. Pant P, Harrison R. Estimation of the contribution of road traffic emissions to particulate matter concentrations from field measurements: a review. Atmos Environ. 2013;77:78–97.
10. EPA National Ambient Air Quality Standards (NAAQS) for particulate matter (PM) | particulate matter (PM) pollution. https://www.epa.gov/pm-pollution/2012-national-ambient-air-quality-standards-naaqs-particulate-matter-pm
11. B. Duignan, "Occam's Razor Philosophy" Encyclopedia Britannica, updated 2020. https://www.britannica.com/topic/Occams-razor
12. Hamidi S, Ewing R. A longitudinal study of changes in urban sprawl between 2000 and 2010 in the United States. Landsc Urban Plan. 2014;128:72–82.
13. Pawlish M, Varde A. The DevOps paradigm with cloud data analytics for green business applications. ACM SIGKDD Explor. 2018;20(1):51–9.
14. Nagy R, Lockaby B. Urbanization in the southeastern United States: Socioeconomic forces and ecological responses along an urban-rural gradient. Urban Ecosyst. 2010;14(1):71–86.
15. Li X, Gar-On Y. Data mining of cellular automata's transition rules. Int J Geogr Inf Sci. 2004;18(8):723–44.
16. Miller H, Han J. Geographic data mining and knowledge discovery. 1st ed. Boca Raton: Taylor & Francis; 2009.
17. Nica A, Suchanek FM, Varde AS. Emerging multidisciplinary research across database management systems. ACM SIGMOD Rec. 2010;39(3):33–6.
18. Rajasekar U, Weng Q. Application of association rule mining for exploring the relationship between urban land surface temperature and biophysical/social parameters. Photogramm Eng Remote Sens. 2009;75(4):385–96.
19. Suchanek FM, Varde AS, Nayak R, Senellart P. The hidden Web, XML and the semantic Web: Scientific data management perspectives. In: Proceedings of ACM EDBT, Uppsala, Sweden; 2011. p. 534–537.
20. Xue W, Li Q, Xue Q. Text detection and recognition for images of medical laboratory reports with a deep learning approach. IEEE Access. 2019;8:407–16.
21. Tancer J, Varde A.S. The deployment of MML for data analytics over the cloud. In: Proceedings of IEEE ICDM workshops, Vancouver, Canada, 2011. p. 188–195.
22. Du X, Varde AS. Mining multicity urban data for sustainable population relocation. Int J Comput Electr Autom Control Inf Eng. 2015;9(12):2530–7.
23. Pampoore-Thampi A, Varde AS, Yu D. Mining GIS data to predict urban sprawl. In: Proceedings ACM KDD Bloomberg Track, New York City; 2014. p. 118–125.
24. Du X, Emebo O, Varde AS, Tandon N, Nag Chowdhury S, Weikum G. Air quality assessment from social media and structured data: pollutants and health impacts in urban planning. In: Proceedings of IEEE ICDE (workshops), Helsinki, Finland, May 2016. p. 54–59.
25. Jinia A, Sumbul N, Meert C, Miller C, Clarke S, Kearfott K, Matsuzak M, Pozzi S. Review of sterilization techniques for medical and personal protective equipment contaminated with SARS-CoV-2. IEEE Access. 2020;8:111347–54.
26. Vlachogianni A, Kassomenos P, Karppinen A, Karakitsios S, Kukkonen J. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci Total Environ. 2011;409(8):1559–71.
27. Aguilera I, Eeftens M, Meier R, Ducret-Stich R, Schindler C, Ineichen A, Phuleria H, Probst-Hensch N, Tsai M, Künzli N. Land use regression models for crustal and traffic-related PM2.5 constituents. Environ Res. 2015;140:377–84.
28. Yang G, Huang J, Li X. Mining sequential patterns of PM2.5 pollution in three zones in China. J Clean Prod. 2018;170:388–98. https://doi.org/10.1016/j.jclepro.2017.09.162.
29. Bai K, Li K, Chang N, Gao W. Advancing the prediction accuracy of satellite-based PM2.5 concentration mapping: a perspective of data mining through in situ PM2.5 measurements. Environ Pollut. 2019;254:113047. https://doi.org/10.1016/j.envpol.2019.113047.
30. Lin Y, Chiang Y, Pan F, Stripelis D, Ambite J, Ecekel S, Harbre R. Mining public datasets for modeling intra-city PM2.5 concentrations at a fine spatial resolution. ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems; 2013. https://doi.org/10.1145/3139958.3140013.
31. Dias D, Tchepel O. Modelling of human exposure to air pollution in the urban environment: a GPS-based approach. Environ Sci Pollut Res. 2013;21(5):3558–71. https://doi.org/10.1007/s11356-013-2277-6.
32. Lary D, Faruque F, Malakar N, Moore A, Roscoe B, Adams Z, Eggelston Y. Estimating the global abundance of ground level presence of particulate matter (PM2.5). Geospat Health. 2014;8(3):611. https://doi.org/10.4081/gh.2014.292.
33. Min K, Kwon H, Kim K, Kim S. Air pollution monitoring design for epidemiological application in a densely populated city. Int J Environ Res Public Health. 2017;14(7):686. https://doi.org/10.3390/ijerph14070686.
34. Li D, Liu J, Zhang J, Gui H, Du P, Yu T, Wang J, Lu Y, Liu W. Identification of long-range transport pathways and potential sources of PM2.5 and PM10 in Beijing from 2014 to 2015. J Environ Sci. 2017;56:214–29. https://doi.org/10.1016/j.jes.2016.06.035.
35. TU Wien. "European Smart Cities", Tech. Rep., Vienna University of Technology, Vienna, Austria (2015)
36. Wikipedia on Smart Cities. https://en.wikipedia.org/wiki/Smart_city
37. Alamaniotis M. Morphing to the mean approach of anticipated electricity demand in smart city partitions using citizen elasticities. In: Proceedings of IEEE International Smart Cities Conference (ISC2), Kansas City, MO, USA; 2018. p. 1–7.
38. Pandey A, Puri M, Varde AS. Object detection with neural models, deep learning and common sense to aid smart mobility. In: Proceedings of IEEE ICTAI, Volos, Greece; 2017. p. 859–863.
39. Puri M, Du X, Varde AS, de Melo G. Mapping ordinances and tweets using smart city characteristics to aid opinion mining. In: Proceedings of W3C's WWW (Comp. Vol.), Lyon, France; 2018. p. 1721–1728.
40. Kaluarachchi A, Roychoudhury D, Varde AS, Weikum G. SITAC: discovering semantically identical temporally altering concepts in text archives. In: Proceedings of ACM EDBT Demo Track, Uppsala, Sweden; 2011. p. 566–569.
41. IMD Business School, Lausanne Switzerland. Smart City Index 2020: Singapore, Helsinki and Zurich triumph in global smart city index (2020). https://www.imd.org/smart-city-observatory/smart-city-index. Accessed May 2021.
42. LinkNYC. https://en.wikipedia.org/wiki/LinkNYC. Accessed May 2021.
43. Amazon Web Services, AWS IoT. https://aws.amazon.com/iot/. Accessed May 2021.
44. Wikipedia on IoT, https://en.wikipedia.org/wiki/Internet_of_things. Accessed May 2021.

45. Arkian H, Diyanat A, Pourkhalilia A. MIST: Fog-based data analytics scheme with cost-efficient resource provisioning for IoT crowd-sensing applications. J Netw Comput Appl. 2017;82:152–65.
46. Han J, Kamber M, Pei J. Data mining: concepts and techniques. 3rd ed. Burlington: Morgan Kaufmann; 2012. (**ISBN 0123814790**).
47. Witten IH, Frank E, Hall MA. Data mining: practical machine learning tools and techniques. 3rd ed. Burlington: Morgan Kaufmann; 2011.
48. World Bank Data. http://data.worldbank.org/indicator/EN.ATM.PM25.MC.M3. Accessed May 2021.
49. World Health Organization. http://www.who.int/gho/countries/en/. Accessed May 2021.
50. Agrawal R, Imielinski T, Swami A. Mining association rules between sets of items in large databases. SIGMOD Rec. 1993;22(2):207–16.
51. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, 1967. p. 281–297.
52. WEKA: Waikato Environment for Knowledge Analysis, Univ. of Waikato, New Zealand. https://www.cs.waikato.ac.nz/ml/weka/. Accessed May 2021.
53. Quinlan J. C4.5: programs for machine learning. 1st ed. San Francisco: Morgan Kaufmann; 1993.
54. OECD countries. https://www.oecd.org/about/. Accessed May 2021.
55. Rissanen J. Modeling by shortest data description. Automatica. 1978;14(5):465–658.
56. Nivana H. Intelligent tutoring systems: an overview. Artif Intell Rev. 1990;4(4):251–77.
57. ELM-ART: Episodic learner model, the adaptive remote tutor for LISP. http://art2.ph-freiburg.de/Lisp-Course. Accessed May 2021.
58. Technopedia-Mobile Apps. www.techopedia.com/definition/2953/mobile-application-mobile-app. Accessed May 2021.
59. Boulos MNK, Geraghty EM. Geographical tracking and mapping of coronavirus disease COVID-19. Int J Health Geogr. 2020. https://doi.org/10.1186/s12942-020-00202-8.
60. Varghese C, Varde AS, Du X. An ordinance-tweet mining app to disseminate urban policy knowledge for smart governance. In: Proceedings of I3E, Skukuza, South Africa (online conference), vol. 2; 2020. p. 389–401.
61. Schwartz J, Danesh Yazdi M, Kuang Z, Dimakopoulou K, Beevers S, Barratt B, Katsouyani K. Predicting fine particulate matter (pm2.5) in the greater London area: an ensemble approach using machine learning methods. Environ Epidemiol. 2019;3:355–6.
62. Vodonos A. Estimation of excess mortality due to long-term exposure to PM25 in continental United States using a high-spatiotemporal resolution model. Environ Res. 2021. https://doi.org/10.1016/j.envres.2021.110904 (**PMID: 33636186**).
63. Schwartz JD, Yitshak-Sade M, Zanobetti A, Di Q, Requia WJ, Dominici F, Mittleman MA. A self-controlled approach to survival analysis, with application to air pollution and mortality. Environ Int. 2021. https://doi.org/10.1016/j.envint.2021.106861 (**ISSN 0160-4120**).
64. Yazdi MD, Wang Y, Di Q, Wei Y, Requia WJ, Shi L, Sabath MB, Dominici F, Coull BA, Evans JS, Koutrakis P. Long-term association of air pollution and hospital admissions among medicare participants using a doubly robust additive hazards model. Am Heart Assoc. 2021;143(16):1584–96.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.