A Methodology to Develop an Outdoor Activities Recommender Based on Air Pollution Variables

Pablo Arévalo[1], Marcos Orellana[1](✉), Priscila Cedillo[2], Juan-Fernando Lima[1], and Jorge Luis Zambrano-Martinez[1]

[1] Laboratorio de Investigación y Desarrollo en Informática - LIDI,
Universidad del Azuay, Cuenca, Ecuador
{pablo.loja,marore,flima,jorge.zambrano}@uazuay.edu.ec
[2] Universidad de Cuenca, Cuenca, Ecuador
priscila.cedillo@ucuenca.edu.ec

Abstract. Nowadays, the world faces a high level of environmental pollution. This phenomenon has become a constant challenge for our society due to its negative impact on health and the increased risk of disease. Considering this problem, applications, techniques and methodologies are generated that seek to relate atmospheric pollutants to each other to predict the state of the air. On the other hand, recommendation systems are present in numerous decision-making methods to find trends in various fields. Consequently, this work presents a methodology for a recommender system that provides people with the best hours to perform outdoor activities according to the pollutants found in the environment. The results obtained were verified through an evaluation and thus be able to contribute to the creation of new recommenders based on the previous topics.

Keywords: Recommender systems · Air quality · Data mining · Air pollutants · Meteorological variables

1 Introduction

Currently, the world faces high environmental air pollution [16], which this phenomenon directly affects health and increases the risk of diseases in people [3,32]. Furthermore, the short-term and long-term adverse effects of inhaling air pollutants on the respiratory and cardiovascular systems have been widely documented and confirmed [16,28]. Hence, understanding the impact of pollutants and weather variables when doing outdoor activities is a constant challenge for our society [3,32]. Pollution is the result of combining multiple atmospheric variables. However, some of them are very disquieting because they have a more significant impact on human health, the predominant ones being: ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), particulate matter ($PM_{2.5}$). These pollutants are part of the "Criteria Pollutants", a classification given by the Environmental Protection Agency (EPA) to all contaminants that are present in any place due to their nature and origin [2,36].

J. Herrera-Tapia et al. (Eds.): TICEC 2022, CCIS 1648, pp. 171–185, 2022.
https://doi.org/10.1007/978-3-031-18272-3_12

At the same time, efforts to monitor and improve the air quality have been intensified, generating a wide field of research [27]. Monitoring the concentrations of atmospheric pollutants, calculating the pollution index and analysing the correlations and incidences between contaminants are the most relevant for an adequate air quality evaluation [27]. As a result of the monitoring, large volumes of data are available to develop tools, techniques, or methodologies that can relate the variables to each other [4,27,34]. It is essential to specify the best range of time in which an individual is exposed to less contamination. Doing activities in environments with high exposure to these pollutants affect health and performance, and this issue needs to be addressed as part of deep research.

Recommender systems are tools that provide suggestions about items that can be useful to the user. These recommendations are linked to a decision-making process focused on a particular issue [26]. Thus, this paper aims to develop a methodology for supporting the development of a recommender system based on atmospheric pollutants and meteorological variables through data mining techniques. In this case, the K-Means algorithm is considered as an unsupervised grouping method that allows objects to be grouped into k groups depending on whether they contain similarities. This technique does not require labeled attributes allowing the adaptation to new examples and its use of large data sets [9,12]. This document is structured as follows: Sect. 2 presents the related works, Sect. 3 describes each stage and result of the proposed methodology, Sect. 4 shows the evaluation of the proposed methodology, Sect. 5 discusses the threads of validity, and Sect. 6 presents the conclusions and future works.

2 Related Works

This section aims to guide among studies of artificial intelligence, data mining methodologies and techniques. Manohar et al. [17] apply the clustering analysis in Chennai, India to study awareness in the population to identify more predominant groups of questions and seek a reduction in air pollution. In other cases, the pollution forecast is performed using neural networks and the Dijkstra algorithm on the Hadoop MapReduce framework to search for a short route with less environmental pollution [30]. Likewise, Taneja et al. [35] present a study in Delhi, India, with techniques such as linear regression and multilayer perceptron, pollution trends to perform predictions through various emission sources such as vehicle emissions, industrial emissions, demolition, etc. [34].

Besides, other studies consider Relative Humidity (RH) and the air temperature as the concerting variables influencing the practice of multiple sports or physical activities. Thus, Chowdhury et al. [6] consider these two atmospheric variables to evaluate their influence on sports performance, such as the level of comfort for different sports disciplines. The results show that those parameters significantly impact sports performance for outdoor events.

A study developed in China reveal that pollutants such as $PM_{2.5}$, PM_{10}, SO_2, CO, and NO_2 are correlated with RH [39]. Similarly, another study focused on air quality shows that RH, air temperature, and wind speed are dominant factors influencing air quality due to their significant effects on the dispersion and transformation of pollutants [14]. In another study, Giri et al. [8] use meteorological conditions such as temperature, rain, humidity, atmospheric pressure, direction, wind speed, and concentrations of particulate matter and apply the Pearson's correlation coefficient to demonstrate the influence of PM_{10} and the RH, the wind speed and humidity. Thus, air temperature and RH are related to the heat index, indicating the relationship and impact on human health.

Recommendation systems are commonly found in most applications that detect trends in multiple sectors, such as: entertainment, news, and products' sales [15]. For example, recommendation systems are developed to extract educational data for predicting student learning outcomes [35]. Likewise, recommendation systems are created in the health field and well-being, for example, selecting the right person to practice a suitable sport based on parameters such as heart rate, speed, and height using k-means for grouping these data [1]. Public health care datasets are also considered to analyse the performance of different machine learning techniques to provide an intelligent health environment that evaluates health from multiple perspectives and recommends the most appropriate actions [31]. Collecting metrics of allergens and air quality from pollution stations, a recommender system is proposed to solve mobility problems of citizens by informing them which pedestrian routes minimize the time of exposure to allergens [7]. Finally, recommendation systems have different approaches, such as collaborative filtering, which seeks a method to emit recommendations considering user interactions in the past [37]. But the content-based approach learns to recommend items similar to what the user is interested in through a history [26,37].

3 Methodology

Most recommender systems usually apply techniques and procedures from other areas such as Human-Computer Interaction (HCI) or Information Retrieval (IR) [26]. Also, they use an algorithm that can be interpreted as an instance of a data mining technique based on its three main steps, which are executed sequentially: i) previous data processing, ii) data analysis, and iii) interpretation of results. This methodology was focused on five main activities based on the three previous steps. The first activity was performed with the data collection for both atmospheric pollutants and meteorological variables. Then, the preprocessing of the collected data to avoid outliers to provide a clean dataset was performed as the second activity. Subsequently, in the third activity, grouping techniques were applied to generate groups of items with similar characteristics. In the fourth activity, it was essential to have evaluation metrics to qualify the results obtained and determine the optimal time with less pollution. Finally in the five activity the results was obtained, as shown in Fig. 1. It should be noted that these activities described above focus on the creation of a content recommender since, in each of the step, contaminants are considered as the basis for issuing recommendations, unlike demographic or collaborative filtering recommenders that place in the first place the opinions or characteristics of the users [26].

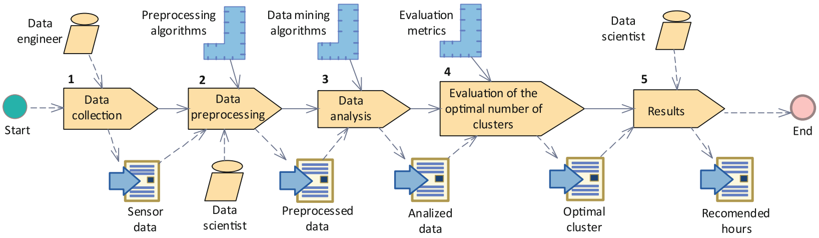

Fig. 1. Methodology for the outdoor activities recommender.

3.1 Data Collection

During 2018, a total of 52474 records were collected through the automatic monitoring station in Cuenca, Ecuador; the same one has a scope of four km, and stores these air pollution values by intervals of one minute, on the other hand, the meteorological variables are collected by intervals of 10 min [20], the Table 1 shows the descriptive values these variables.

Table 1. Collected pollutants and meteorological variable.

Description	Abbreviature	Measure
Pollutants		
Ozone	O_3	$\mu g/m^3$
Carbon monoxide	CO	mg/m^3
Nitrogen dioxide	NO_2	$\mu g/m^3$
Sulfur dioxide	SO_2	$\mu g/m^3$
Particulate matter	$PM_{2.5}$	$\mu g/m^3$
Meteorological variables		
Wind speed	$WINDSPEED\_AV$	m/s
Air temperature	$TEMPAIR\_AV$	$°C$
UV radiation (A)	UVA_AV	W/m^2
Global radiation	$RADGLOBAL\_AV$	W/m^2
Dew point	DP_AV	$°C$
Precipitation	$PREC\_SUM$	mm
Relative humidity	RH_AV	%

3.2 Data Preprocessing

Data of the real world needs to be preprocessed, as the data is often unclean, missing or inconsistent [26]. In the first instance, the data must have a unified frequency since there is a time discrepancy between the records of atmospheric pollutants and meteorological variables. For this, the DateTime attribute was used to average the records of atmospheric pollutants in periods of ten minutes.

In this way, the meteorological variables and atmospheric pollutants are records that can be linked [19]. A data cleaning process was performed to identify possible errors in the records that can occur due to sensor failures, power outages, computer system failures, and contaminant levels below detection limits, among others [4]. To mitigate these possible drawbacks, three methods were implemented: i) fill missing values with zeros and filter all data with values greater than zero, ii) isolate each variable, and iii) apply the Local Outlier Factor (LOF) algorithm, which establishes a limit so that all values are below it [23]. In detail, two regularized methods of missing imputation, Lasso and Ridge regression were considered to determine the number of forwarding and backward points needed to estimate the value of a missing data point; that is, if a point is outside the threshold, it is considered an outlier [23].

3.3 Data Analysis

In this activity, the data mining methods were selected and applied. For this reason, as a first point, it was necessary to determine the attributes to consider in the following tasks. Those tasks are essential to recommend the best time to perform outdoor activities. All collected air pollutants were selected as they have a higher impact (criterion pollutants) on human well-being [2,36]. Regarding the meteorological variables and according to previously analysed studies, relative humidity relative and air temperature were considered relevant attributes due to their direct impact on pollutants and outdoor activities [14,17,24,39].

It was essential to have numeric type columns to avoid errors when applying the grouping algorithm. Subsequently, an attribute derivation of the date type attributes was performed to obtain only the hours, discarding the dates, minutes, and seconds. The optimum time range was set from 5:00 to 22:00 outside this range were discarded because they biased the results when searching for the best result. Although there was no high pollution in these discarded hours, they can be dangerous due to the absence of pedestrians and vehicles [10]. Afterwards, it was necessary to adjust the range of values when dealing with attributes of different units and scales. For this, the Z-transform was used, subtracting the mean of the data from all the values, and later the result obtained was divided by the standard deviation. The distribution obtained from the data had a mean of zero and a variance of one [12]. This normalisation technique is shared and used in various areas, and compared to other tools, it preserves the original distribution of the data and, unlike other normalisation methods, it is less influenced by outliers [12,18].

Finally, the K-Means algorithm was applied, which divides a data set of N elements into k separate subsets. Initially it is required to specify the number of subsets [22]; each one of them is defined by its members and by its centroid. The centroid of each group is the point at which the sum of the distances of all elements in that group is minimized. The algorithm works by randomly selecting the centroids [26]. Subsequently, all elements are assigned to the subset whose centroid is closest to them. The new cluster centroid must be updated to account for items that were added or removed from the cluster. This operation continues until there are no more elements that change their membership in a cluster [26]. Although a grouping method is less precise than a classification method, it can be implemented as a preliminary step to reduce the number of candidates or distribute them among different recommendations [26]. The application of the methods described above was done in Rapidminer, an open source tool that has a data/text mining engine for integration in its own products.

3.4 Evaluation of the Optimal Number of Clusters

Evaluation metrics were essential to validate the obtained results in our previous activities. Firstly, the elbow method is used to determine the optimal number of clusters [38]. The method executed the K-Means on the dataset for a range of k values after each k value, calculated the Sum of Squared Errors (SSE). The goal was to choose a small value of k that still has a low SSE, represented by the elbow [21]. The optimal number of clusters obtained was $k = 8$, as shown in Fig. 2.

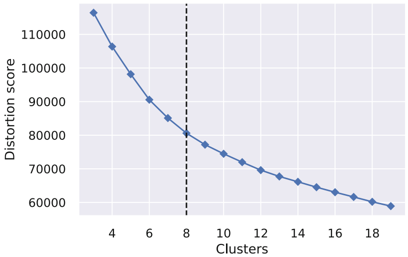

Fig. 2. The optimal number of clusters.

The method performed was graphical, facilitating the use of a large-scale dataset making it more desirable than other methods such as the silhouette index, which produces considerable computational overhead [38]. Subsequently, the best option of the result of the elbow method provided the most appropriate hours for outdoor activities as a recommendation. On the other hand, air pollutants were classified on six scales according to the health risk associated with air quality [14]. For this, it was necessary to use the air quality index, which indicates the possible impacts on health based on the actual concentrations of atmospheric pollutants in a range of 0 to 500 points [14]. Table 2 presents the quality standards for each of the input air pollutants used. These values were categorized according to the air quality index [14].

The meteorological variables were categorized using the heat and wind chill indices. The heat index is the temperature felt in the body when relative humidity is combined with temperature [11]. This index range was between 80 °F (26 °C) and 108 °F (42 °C), and it corresponding to RH_AV from 0% to 100%. Therefore, any value below that temperature range does not harm health since,

Table 2. Pollution index per pollutant.

Air quality	SO_2	NO_2	CO	O_3	$PM_{2.5}$
Excellent	0–50	0–40	0–2	0–100	0–35
Good	51–150	41–80	3–4	101–160	36–75
Light pollution	151–475	81–180	5–14	161–215	76–115
Moderate pollution	476–800	181–280	15–24	216–265	116–150
Strong contamination	801–1600	281–565	25–36	266–800	151–250
Severe contamination	>1601	>566	>37	–	>251

even if it has 100% humidity. The heat index does not indicate that it is dangerous, and their values correspond to shady places, so if they are exposed to direct sunlight, this index can increase by 15 °F, so it was considered to increase this range of 80 °F (26 °C) to 65 °F (18 °C) to cover this issue [11]. On the other hand, to determine the minimum temperature, the wind chill index was considered o refer to the loss of heat from the body to its environment during cold or windy days [13]. The available range is 5 °C to −50 °C [28]. Based on this, it was determined that the values that exceed 5 °C are suitable for outdoor activities since the risk due to low temperatures is minimised [13]. Table 3 summarise the ranges and health implication.

Table 3. Information about air temperature and RH.

TEMPAIR_AV	RH_AV	Health implication	Exerc.
Lees than 0	–	High risk of frostbite	No
0–5	0–100	Low risk of freezing	No
6–18	0–100	No harm to human health	Yes
19–26	0–100	No harm to human health/little probability	No
27–32	–	Fatigue is possible with prolonged exposure and/or physical activity	No
>32	–	High likelihood of heat cramps or heat exhaustion, and possible heat stroke	No

Data collected in Cuenca, Ecuador did not completely satisfy all the categories above. For this, the centroids in excellent air quality and allow outdoor activities were analyzed in each cluster.

3.5 Results

This subsection presents the results of the application of the proposed methodology. So, each cluster centroid was considered and the best option was chosen.

The centroids are a set of values that represent the behavior of the resulting clusters, in addition, depending on them, these groups can be categorized to issue recommendations [26]. As shown in Fig. 3, cluster number six presented an optimal result compared to the others after analysing each pollutant. This phenomenon can be observed in variables such as CO, SO_2, and $PM_{2.5}$. The centroids of this cluster were in the optimal range of air quality and allowed other clusters containing centroids with very high values to be quickly eliminated, as in the case of clusters number one, two, three, and seven.

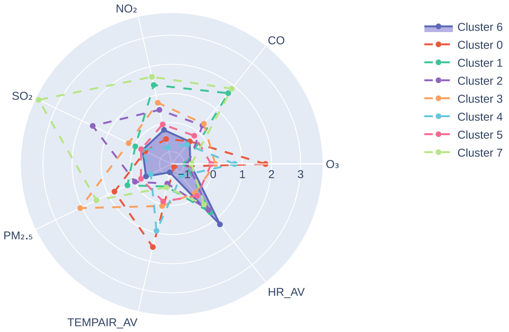

Fig. 3. The centroids of clusters.

Each centroid was denormalised to verify that these values allowed recommending the best hours for outdoor activities. This task was performed to observe their equivalence in the respective measurement units of each pollutant. Thus ratifying an excellent air quality and a temperature with humidity without risk to health. Although cluster six contained a high value in its centroid for RH_AV, this did not negatively influence having its centroid in $TEMPAIR\_AV$ in the optimal temperature range, and according to the heat index, there was no risk for activities outdoor. Likewise, this index allowed the elimination of clusters four and zero due to their high values at room temperature and possible complications in the open air such as fatigue.

Besides, considering O_3, cluster five was discarded. Once cluster 6 was determined as the best option, it is necessary to recommend the best time for outdoor activities. Kernel Density Estimation (KDE) smoothes observations with a Gaussian function, producing a continuous density estimate [22].

Figure 4 offers a comparison of the different hours for each cluster. In this case, the results determined that cluster six established that the best recommended hours to do outdoor activities cover a range from 5:00 to 8:00 and the night from 20:00 to 22:00. Table 4 shows the values of each centroid with the variables analysed. Concerning each atmospheric pollutant, SO, NO_2, CO, O_3, $PM_{2.5}$, the centroids, once denormalized, had values within excellent air quality. Therefore, there was a low level of air pollution, which is vital to avoid

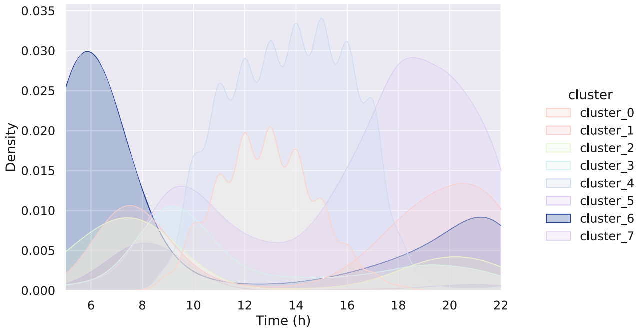

Fig. 4. KDE comparing each cluster to determine optimal time.

health damage. Otherwise, close to the noon-hour, atmospheric pollutants' values increased abruptly. This result was due to vehicular traffic, factories, and companies during working hours expelling pollutants and toxins into the environment in an uncontrolled manner, the same ones that, when found in the atmosphere, can combine, causing diseases of all kinds [33].

Table 4. Optimal clusters and related timestamps.

	Centroids	Denormalized centroids	Timestamp	Air quality
SO_2	−0.308	7.109	05:00–08:00 | 20:00–22:00	Excellent
NO_2	−0.223	15.328	05:00–08:00 | 20:00–22:00	Excellent
CO	−0.439	0.634	05:00–08:00 | 20:00–22:00	Excellent
O_3	−0.784	12.761	05:00–08:00 | 20:00–22:00	Excellent
$PM_{2.5}$	−0.44	7.093	05:00–08:00 | 20:00–22:00	Excellent
$TEMPAIR\_AV$	−1.127	12.167	05:00–08:00 | 20:00–22:00	Excellent
RH_AV	1.229	83.083	05:00–08:00 | 20:00–22:00	Excelent/Regular

Therefore, these hours were the worst time to perform activities outside and even worse if the person has respiratory or cardiac problems. Along with atmospheric pollutants, the centroid related to $TEMPAIR\_AV$ found a value of 12 °C, the optimal temperature to go outside. Therefore, concerning the thermal sensation index, there was no health risk due to low temperatures that can cause frostbite or diseases related to the respiratory system. On the other hand, considering the heat index, there was no risk of fatigue or sunstroke. As mentioned above, the RH_AV variable had a high relationship with the $TEMPAIR\_AV$. Although the value of the centroid is in an Excellent/Fair category, it has a $TEMPAIR\_AV$ that is outside the danger range of the heat index and does not have a negative effect. Based on these results analysed for each pollutant and atmospheric variable, it can be affirmed that the chosen cluster is the best option, and the recommended time to do activities outdoors that do not represent health risks is from 5:00 to 8:00 and from 20:00 to 22:00.

4 Evaluation of the Methodology

Evaluating this methodology as a case study, the validity and understanding of the methods presented in the above sections are essential. To perform this task, the steps proposed by Runeson and Höst [29] are followed for conducting and reporting this case study like a research in software engineering, were taken as a reference similar to any empirical study. The case studies, allow a more in-depth understanding of the phenomenon being studied in its current context.

4.1 Objectives and Research Questions

A case study investigates a flexible type in which planning is necessary for its development to be successful. From the Goal-Question Metric paradigm [5], the goal of this study is presented below:

- Evaluate the analysis phase of the proposed methodology based on atmospheric pollutants and meteorological variables.
- From the point of view of a Data scientist.
- Context: Research and development laboratory researchers.

The proposed methodology is based on the standard designed by Ricci [26] for recommenders in data mining, which focuses on any area. However, the presented methodology cannot be evaluated concerned another since no standard or similar examples of recommenders allow determining the best time for outdoor activities, considering atmospheric pollutants and meteorological variables as a basis [26]. The proposed research questions for the case study are:

- RQ1: Is the methodology presented to recommend outdoor activities based on pollutants and atmospheric variables perceived as useful and easy to use?.
- RQ2: Is there an intention to use this methodology in the future?.

4.2 Context and Survey Design

The context is represented by the proposed methodology of the outdoor activity recommender system that is evaluated. Researchers belonging to the Research and Development Laboratory were selected who have extensive knowledge in data mining techniques and work professionally in this same area. To answer the research questions, a survey based on the Technology Acceptance Model (TAM) has been included to provide evidence on the Perceived Ease Of Use (PEOU), Perceived Usefulness (PU), and Intention to Use (ITU) [25]. The Likert scale was considered, where one is regarded as a negative response, three as a neutral value and five as the highest value. These questions are shown in below list.

- PEOU1: The proposed methodology is simple and easy to follow.
- PEOU2: In general, the proposed methodology is easy to understand.
- PEOU3: The steps that must be followed to complete the methodology are clear and understandable.
- PEOU4: The proposed methodology is easy to learn.
- PEOU5: I think that the proposed methodology would be helpful as a basis for future projects.
- PU1: I consider that the proposed methodology would reduce the time and effort required to search for a recommender of outdoor activities.
- PU2: Generally, I consider the proposed methodology is useful.
- PU3: I consider the proposed methodology applicable when determining best time to perform outdoor activities.
- PU4: I think the methodology is detailed enough when guiding each of the steps focused on data science.
- PU5: This methodology provides an optimal method based on K-Means in the quest to determine the best time for outdoor activities.
- ITU1: If I were to provide training on air pollution and outdoor activities, I would consider this developed methodology.
- ITU2: If necessary, I would use this methodology in the future.
- ITU3: I would recommend the use of this methodology.

4.3 Results and Data Interpretation

For the case study to be successful, as a first task, an introduction was performed on the importance of the topic raised and the health benefit of being less exposed to atmospheric pollutants. Subsequently, the materials were delivered in digital format to the researchers, who performed the corresponding evaluation based on their reading and the model created in RapidMiner software. Once the data was obtained, the tabulations corresponding to each of the attributes were made: separate the questions according to the PEOU, the PU and the ITU. Table 5 shows the minimum and maximum values, and the means of the study variables, each of them consolidated (PEOU, PU, and ITU).

Table 5. Descriptive statistics for variables based on user perception.

Variable	Min	Max	Avg	Standard deviation	Error standard
PEOU	5.0	5.0	5.0	0.00	0.00
PU	4.0	5.0	4.6	0.40	0.28
ITU	4.0	5.0	4.67	0.33	0.23

The RQ raised can be answered based on the researchers who participated in the case study. The question seeks to perceive the usefulness of the methodology and ease of use. It can be concluded that it has been considered useful, based on the PU, where the mean has marked a value of 4.6 with a minimum of four and a maximum of five. Similarly, researchers found the methodology easy to use, with a 5/5 rating for PEOU. The second question focused on the ITU proposed methodology in the future. It averages of 4.67, implying that participants will consider the use of this methodology.

5 Threats to Validity

In this section we have analyzed the main threats that could interfere with the interpretation of the results. In relation to a validity of conclusion, there was a threat linked to the size and selection of the sample as well as the difficulty of finding data scientists. To mitigate this problem, expert researchers in this area belonging to the research laboratory were considered. In addition, due to the low significance of the sample size, experimental validation for a case study was eliminated. Similarly, they have proposed adding a more significant number of data scientists to corroborate the methodology with a larger sample in future work. Subsequently, considering an internal validity, a threat associated with the user's prior knowledge was presented, since they should have previously known the topics previously developed. To mitigate this threat, the selected researchers underwent initial training to state the research topic, its importance, and prior advances.

In the same way, there was a threat related to the validity of the survey presented. To mitigate this problem a Cronbach's alpha test of questions related to each subjective variable was performed to increase reliability. Being the minimum accepted threshold $\alpha = 0.70$, then, in PEOU $\alpha = 1$, in PU $\alpha = 0.937$, and ITU $\alpha = 0.75$. Validating the reality study applying TAM. Finally, to guarantee external validity, The participants were selected for convenience since they have a broad understanding of techniques, methods, and tools focused on data mining and know first-hand the analysed data of this study.

6 Conclusions

The use of the K-Means algorithm in developing our methodology resulted appropriate since it allows the use of large amounts of data on atmospheric pollutants and meteorological variables. However, these data must be normalised before their application to prevent larger-scale attributes from dominating the distances, affecting the results. For this purpose, it is required to have features that influence when doing outdoor activities depending on the location of the city. Atmospheric pollutants must be considered in their entirety since they mix within the environment, increasing their danger to health. The proposed methodology provides a solution to determine the best time for exercising outdoors. The selected time intervals are the best as there is no dangerous accumulation of these contaminants in the air, and the temperature and humidity do not represent a health hazard. Finally, based on the proposed case study, it is essential to indicate that a methodology is a valuable tool. Its ease of use allows its implementation without significant inconveniences. It can be used as a basis for future projects of contextualized recommenders, which consider the user's context, in this case the user's location with its respective air quality, humidity and temperature as the basis for the recommendations.

Acknowledgements. This work was partially supported by the Vice-Rector for Research of the University of Azuay for financial and academic support, as well as the entire staff of the Computer Science department and the Laboratory for Research and Development in Informatics - LIDI.

References

1. Abdulaziz, M., Al-motairy, B., Al-ghamdi, M., Al-qahtani, N.: Building a personalized fitness recommendation application based on sequential information. Int. J. Adv. Comput. Sci. Appl. **12**(1), 637–648 (2021). https://doi.org/10.14569/IJACSA.2021.0120173
2. Agency United States Environmental: Un resumen de la Ley de Aire Limpio (2007)
3. An, R., Zhang, S., Ji, M., Guan, C.: Impact of ambient air pollution on physical activity among adults: a systematic review and meta-analysis. Perspect. Public Health **138**(2), 111–121 (2018). https://doi.org/10.1177/1757913917726567
4. Arce, D., Lima, F., Orellana Cordero, M.P., Ortega, J., Sellers, C., Ortega, P.: Discovering behavioral patterns among air pollutants: a data mining approach. Enfoque UTE **9**(4), 168–179 (2018). https://doi.org/10.29019/enfoqueute.v9n4.411
5. Basili, V.R., Weiss, D.M.: A methodology for collecting valid software engineering data. IEEE Trans. Softw. Eng. **SE-10**(6), 728–738 (1984). https://doi.org/10.1109/TSE.1984.5010301
6. Chowdhury, A.S., Uddin, M.S., Tanjim, M.R., Noor, F., Rahman, R.M.: Application of data mining techniques on air pollution of Dhaka City. In: 2020 IEEE 10th International Conference on Intelligent Systems, IS 2020 - Proceedings, pp. 562–567 (2020). https://doi.org/10.1109/IS48319.2020.9200125
7. García-Díaz, J.A., Noguera-Arnaldos, J.Á., Hernández-Alcaraz, M.L., Robles-Marín, I.M., García-Sánchez, F., Valencia-García, R.: AllergyLESS. An intelligent recommender system to reduce exposition time to allergens in smart-cities. In: De La Prieta, F., Omatu, S., Fernández-Caballero, A. (eds.) DCAI 2018. AISC, vol. 800, pp. 61–68. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-94649-8_8
8. Giri, D., Venkatappa, K., Adhikary, P.: The influence of meteorological conditions on PM10 concentrations in Kathmandu Valley. Int. J. Environ. Res. **2**(1), 49–60 (2008). (ISSN: 1735-6865)
9. Google: Machine Learning, June 2022. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages
10. Guedes, I., Cardoso, C., Agra, C.: Emotional and insecurity reactions to different urban contexts—. GERN **2013**(1), 147 (2013)
11. Hass, A.L., Ellis, K.N., Mason, L.R., Hathaway, J.M., Howe, D.A.: Heat and humidity in the city: neighborhood heat index variability in a mid-sized city in the Southeastern United States. Int. J. Environ. Res. Public Health **13**(1), 117 (2016). https://doi.org/10.3390/ijerph13010117

12. Kotu, V., Deshpande, B.: Chapter 7 - clustering. In: Kotu, V., Deshpande, B. (eds.) Predictive Analytics and Data Mining, pp. 217–255. Morgan Kaufmann, Boston (2015). https://doi.org/10.1016/B978-0-12-801460-8.00007-0
13. Lankford, H.V., Fox, L.R.: The wind-chill index. Wilderness 'I&' Environ. Med. **32**(3), 392–399 (2021). https://doi.org/10.1016/j.wem.2021.04.005
14. Liu, Y., Wu, J., Yu, D., Hao, R.: Understanding the patterns and drivers of air pollution on multiple time scales: the case of Northern China. Environ. Manage. **61**(6), 1048–1061 (2018). https://doi.org/10.1007/s00267-018-1026-5
15. Lü, L., Medo, M., Yeung, C.H., Zhang, Y.C., Zhang, Z.K., Zhou, T.: Recommender systems. Phys. Rep. **519**(1), 1–49 (2012). https://doi.org/10.1016/j.physrep.2012.02.006
16. Mannucci, P.M., Franchini, M.: Health effects of ambient air pollution in developing countries. Int. J. Environ. Res. Public Health **14**(9), 1–8 (2017). https://doi.org/10.3390/ijerph14091048
17. Manohar, G., Devi, S., Rao, K.: A bi-level clustering analysis for studying about the sources of vehicular pollution in Chennai. Adv. Intell. Syst. Comput. **324**, 229–236 (2015). https://doi.org/10.1007/978-81-322-2126-5_26
18. Mohabeer, H., Soyjaudah, K.M., Pavaday, N.: Enhancing the performance of neural network classifiers using selected biometric features. In: SENSORCOMM 2011–5th International Conference on Sensor Technologies and Applications and WSNSCM 2011, 1st International Workshop on Sensor Networks for Supply Chain Management, pp. 140–144 (2011)
19. Orellana, M., Lima, J.F., Cedillo, P.: Discovering patterns of time association among air pollution and meteorological variables. In: Arai, K. (ed.) Advances in Information and Communication, pp. 205–215. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73103-8_13
20. Orellana, M., Salto, J., Cedillo, P.: Behavior analysis of atmospheric components and meteorological variables applying data mining association techniques. In: Arai, K. (ed.) Advances in Information and Communication, pp. 192–204. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73103-8_12
21. Pandey, A., Malviya, A.K.: Enhancing test case reduction by K-means algorithm and elbow method. Int. J. Comput. Sci. Eng. **6**(6), 299–303 (2018). https://doi.org/10.26438/ijcse/v6i6.299303
22. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
23. Peña, M., Ortega, P., Orellana, M.: A novel imputation method for missing values in air pollutant time series data. In: 2019 IEEE Latin American Conference on Computational Intelligence, LA-CCI 2019 (2019). https://doi.org/10.1109/LA-CCI47412.2019.9037053
24. Pezzoli, A., et al.: Effect of the environment on the sport performance: computer supported training - a case study for cycling sports. In: Cabri, J., Pezarat Correia, P., Barreiros, J. (eds.) Sports Science Research and Technology Support, pp. 1–16. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-17548-5_1
25. Rahimi, B., Nadri, H., Afshar, H.L., Timpka, T.: A systematic review of the technology acceptance model in health informatics. Appl. Clin. Inform. **9**(3), 604–634 (2018). https://doi.org/10.1055/s-0038-1668091
26. Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.): Recommender Systems Handbook. Springer, Boston (2011). https://doi.org/10.1007/978-0-387-85820-3
27. Rimensberger, N., Gross, M., Günther, T.: Visualization of clouds and atmospheric air flows. IEEE Comput. Graph. Appl. **39**(1), 12–25 (2019). https://doi.org/10.1109/MCG.2018.2880821

28. Rundell, K.W.: Effect of air pollution on athlete health and performance. Br. J. Sports Med. **46**(6), 407–412 (2012). https://doi.org/10.1136/bjsports-2011-090823
29. Runeson, P., Höst, M.: Guidelines for conducting and reporting case study research in software engineering. Empirical Softw. Eng. **14**(2), 131–164 (2009). https://doi.org/10.1007/s10664-008-9102-8
30. Sadiq, A., El Fazziki, A., Ouarzazi, J., Sadgal, M.: Towards an agent based traffic regulation and recommendation system for the on-road air quality control. SpringerPlus **5**(1), 1–19 (2016). https://doi.org/10.1186/s40064-016-3282-2
31. Sharma, R., Rani, S.: A novel approach for smart-healthcare recommender system. Adv. Intell. Syst. Comput. **1141**, 503–512 (2021). https://doi.org/10.1007/978-981-15-3383-9_46
32. Singla, S.: Air ality friendly route recommendation system. PhD Forum 2018 - Proceedings of the 2018 Workshop on MobiSys 2018 Ph.D. Forum, Part of MobiSys 2018, pp. 9–10 (2018). https://doi.org/10.1145/3212711.3212717
33. Swietlicki, E., Puri, S., Hansson, H.C., Edner, H.: Urban air pollution source apportionment using a combination of aerosol and gas monitoring techniques. Atmos. Environ. **30**(15), 2795–2809 (1996). https://doi.org/10.1016/1352-2310(95)00322-3
34. Taneja, S., Sharma, N., Oberoi, K., Navoria, Y.: Predicting trends in air pollution in Delhi using data mining. In: India International Conference on Information Processing, IICIP 2016 - Proceedings, pp. 1–6 (2017). https://doi.org/10.1109/IICIP.2016.7975379
35. Thai-Nghe, N., Drumond, L., Krohn-Grimberghe, A., Schmidt-Thieme, L.: Recommender system for predicting student performance. Procedia Comput. Sci. **1**(2), 2811–2819 (2010). https://doi.org/10.1016/j.procs.2010.08.006
36. Ubilla, C., Yohannessen, K.: Contaminación Atmosférica Efectos En La Salud Respiratoria En El Niño. Revista Médica Clínica Las Condes **28**(1), 111–118 (2017). https://doi.org/10.1016/j.rmclc.2016.12.003
37. Yu, L.: A35 - cloud storage-based personalized sports activity management in Internet plus O2O sports community. Concurr. Comput. **30**(24), 1–10 (2018). https://doi.org/10.1002/cpe.4932
38. Yuan, C., Yang, H.: Research on K-value selection method of K-means clustering algorithm. J. **2**(2), 226–235 (2019). https://doi.org/10.3390/j2020016
39. Zhou, H., et al.: Characteristics of air pollution and their relationship with meteorological parameters: Northern Versus Southern Cities of China. Atmosphere **11**(3), 253 (2020). https://doi.org/10.3390/atmos11030253