Clustering and Classification of IoT-Based Environmental Data Using Machine Learning Techniques

Ali Subhi Alhumaima[1], Waleed Khalid Al-Zubaidi[1], El-Sayed M. El-Kenawy[2,3,*], Marwa M. Eid[4,5]

[1]Electronic Computer Centre, University of Diyala, Diyala, Iraq

[2]Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology, Mansoura, 35111, Egypt

[3]Applied Science Research Center. Applied Science Private University, Amman, Jordan

[4]Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura, Egypt

[5]Jadara Research Center, Jadara University, Irbid 21110, Jordan

Emails: alhumaimaali@uodiyala.edu.iq; waleed300@uodiyala.edu.iq; skenawy@ieee.org; mmm@ieee.org

Abstract

In this study, we present an integrated approach to IoT-based environmental data analysis using a collection of unsupervised-learning techniques. We employed KMeans clustering in particular to identify natural groupings in environmental and behavioral features such as air quality, noise level, temperature, stress level, sleeping hours, and mood score. We then trained a Decision Tree classifier to predict and interpret cluster membership from raw sensor readings. The data of more than 30,000 observations in indoor school environments has multifaceted relationships between environmental factors and psychological well-being. KMeans consistently detected three environmental-behavioral states, and the Decision Tree classifier performed 87% classification accuracy, which indicated extremely high predictability power in addition to interpretability. The results indicated that sleep duration, air, and stress were the main factors for cluster discrimination. The hybrid model introduces the potential of observing real-time environmental and mental states for applications in smart cities. The approach is scalable, interpretable, and usable in IoT settings for proactivity-enabled wellness management.

Keywords: IoT Sensor Data; Environmental Monitoring; KMeans Clustering; Decision Tree Classification; Behavioral Analysis; Air Quality; Stress Prediction; Machine Learning, Data Mining

1. Introduction

The rapid expansion of the Internet of Things (IoT) has endowed the worldwide use of networked devices and sensors with the capacity for repeatedly collecting large amounts of data in various areas of application [1], [2]. Environmental monitoring has opened up new opportunities for real-time monitoring and assessment of air quality, temperature, humidity, noise, and lighting via the internet [3], [4]. These environmental conditions, which were earlier tracked intermittently through manual surveys or point sensors, are now being tracked in real-time and at scale to offer deep insights into the impact of the environment on human activity, productivity, and well-being [5],[6].

Use of IoT towards smart cities and intelligent building systems is on the rise to address the demand for active tracking of health and management of the environment [7]. For instance, intelligent buildings with networked sensors can dynamically adjust ventilation, lighting, or sound insulation to provide enhanced occupants' comfort and health. These systems rely on efficient analytics that can identify beneficial patterns from high-dimensional sensor data and provide actionable information. Such information becomes extremely useful if combined with behavioural and physiological information such as sleep, mood, and stress levels and offers more detail of human-environment interaction [8].

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

With growing volumes and intensities of data collection through IoT systems, there has been a greater need for advanced data mining and machine learning algorithms to process, analyse, and interpret the data [9]. Particularly, unsupervised learning techniques like clustering have proved to be extremely helpful in discovering patterns buried in unlabelled data. Of these, KMeans clustering is most likely to be among the most popular algorithms due to the ease of use, simplicity, efficiency, and scalability, especially in real-time or embedded environments [10], [11].

KMeans is typically used to partition a dataset into k groups by minimizing the variance within the groups and maximizing the distance between groups [12]. Its performance in environmental monitoring has been tested for a variety of conditions, such as partitioning air quality measurements, demarcating regions of noise pollution, and maximizing HVAC control systems based on user comfort clusters [13]. The fact that the algorithm can function without labelled training data makes it highly suitable for environmental datasets, which are usually lacking in pre-defined class labels. However, although useful for uncovering patterns, clustering results might be hard to interpret or use for prediction directly without additional modelling [14].

To address the shortcomings of clustering alone, interpretable supervised ML models like Decision Trees are often used to explain and forecast the results of clustering outcomes [15], [16]. Decision Trees classify data by creating a sequence of decision rules splitting data in terms of feature values. These models are favoured in situations where there is a need for transparency, interpretability, and low computational costs, e.g., healthcare, education, and urban planning [17]. They give a structured and simple interpretable way of finding the most significant variables on the outcome of classification and hence more robustify the system [18].

The integration of Decision Tree classification and unsupervised clustering is presently a robust paradigm for unlabelled IoT data set knowledge discovery. Within this two-stage architecture, an initial application of clustering is used to create natural groupings within the data and a subsequent Decision Tree learning is employed to classify and explain the resulting clusters [19]. The method not only conducts classification but also attains rule-based explanation of why some data are in a specific environmental-behavioural state.

Moreover, the incorporation of behaviour measures such as stress level, sleep duration, and mood rating into regular environmental measures has created a new category of adaptive and personalized systems [20], [21]. The datasets in hybrids allow for the disclosure of complex, high-dimensional patterns that will be unintelligible when environmental and behaviour features are analysed independently. For instance, the same sound intensity can produce different effects on individuals depending on their sleeping habit or stress susceptibility. This calls for the monitoring devices in IoT-based systems to be individualized [22].

In recent studies, integration of environmental sensing with behavioural sensing has shown spectacular model precision as well as user acceptability. For example, Taskasaplidis et al. showed the combination of contextual environmental information and wearable stress sensor data showing much improved accuracy in detecting stress [20]. Meilă et al. also utilized a hybrid machine learning methodology to forecast health risk from wearable devices and showed potential for real-time adaptive intervention in smart health systems [17].

In terms of deployment, Decision Trees are also useful for IoT application since they have low computational complexity. Deep learning models require massive training data and powerful computing platforms, yet Decision Trees are able to train rapidly and be executed on edge devices, making them best applicable to distributed and power-limited environments such as smart homes and mobile wearables [23].

While there are many advantages, there also are restrictions. Unsupervised clustering lacks predetermined validation metrics and will potentially need to be interpreted by humans, while Decision Trees, being interpretable, might not always be as accurate as more complex models like ensemble learners or neural networks [24]. A balance between model accuracy and interpretability remains a fundamental consideration in designing intelligent environmental monitoring systems.

To overcome these challenges, the current research proposes a technique employing KMeans clustering to discover latent environmental-behavioural groups within IoT data and employing a Decision Tree classifier to explain and forecast these groups based on real-world attributes outside. The dataset contains features such as air quality index, temperature, humidity, noise, light intensity, stress, mood scores, and sleeping hours—received either in the form of simulated or real IoT sensors. KMeans clustering is first used to divide the dataset into three environmental clusters. A Decision Tree model is trained afterwards to label and detail these clusters in terms of the original features, identifying what variables are most important in defining each environmental state.

The contribution of this effort is threefold. Firstly, it demonstrates the successful use of KMeans clustering over IoT environment and behaviour data to uncover significant groupings. Secondly, it demonstrates classification of these groupings through a Decision Tree model and the possibility of making it interpretable by using feature importance analysis. Thirdly, it demonstrates the applied implications of this approach in application domains like smart home adaptation, mental health monitoring, and urban infrastructure planning.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

2. Related Work

New developments in machine learning and data mining significantly enhanced environmental monitoring, especially when combined with IoT technology. The scientific study [15] compared training accuracies and times for various classification algorithms, which is the foundation of model performance when applied to sensor-based information. In water quality monitoring, the scientific solution [18] proposed an intelligent IoT-based real-time quality assessment system, highlighting the improved responsiveness and accuracy achieved by real-time sensing.

Another research [22] approximated pollution in major transboundary rivers, emphasizing the necessity of databased systems in cross-border environmental evaluation. Another comparable study [28] addressed outlier identification in wireless sensor networks, a necessary step towards quality control of the arriving environmental information before clustering or classification.

Yet another research [10] implemented a Water Quality Index (WQI) and GIS to measure pollution in the Tigris River, demonstrating how environmental modeming becomes richer with spatial data. The proposed methodology [9] also created a hierarchical distributed classification for wireless sensor networks with an optimized performance in resource-limited scenarios—a major necessity in IoT applications.

Clustering and classification are crucial when analysing environmental data. Two dual analyses [32], [12] proved the relevance of unsupervised learning to deal with complicated data sets. An illustration of decision-tree-based classification [19] proved interpretability and readability of tree-based models for non-technical stakeholders.

A water-monitoring framework based on IoT [20] described the incorporation of hardware and data-driven decision-making, and an optimization study [21] explored multi-objective balance in prediction models for multi-metric environmental analysis.

Environmental forecasting was also tried with the help of neural networks. A study [25] had proposed an RBF network to forecast water quality, and follow-up studies [26] focused on improved communication and denoising of reduced-dimension sensor networks to enable more accurate processing of noisy data.

Hybrid methods of evolutionary algorithms and associative classification [26] showed versatility in application to environmental data. Another model [27] introduced a supervised learning algorithm for water quality prediction, validating the versatility of these methods to this domain.

Remote-sensing software [23] employed machine learning for oil spill detection from satellite imagery, highlighting processing of spatially distributed sensor data. Ensemble evaluations [25] employed Dempster–Shafer theory in conjunction with ANN and SVM for classifying water quality, documenting growing interest in hybrid methods.

Metaheuristic optimization [26] using LS-SVM and particle swarm strategies improved convergence and accuracy. Similar predictive paradigms [29] in biomedical applications demonstrated cross-domain transferability.

Cheap wireless sensor network design [29] allowed monitoring of water in resource-constrained regions. Institutional rules [11] provided criteria for model evaluation through standardized water quality parameters.

Prototype implementations [11] demonstrated IoT-based test environments with linear topologies and rule-based controllers [12] enabled real-time monitoring. Spatial–temporal classification approaches [12] were evaluated for changing environmental conditions.

Finally, comparative analysis [29] contrasted ANN and Naive Bayes classifiers in terms of the compromises between processing speed and accuracy.

3. Data and Methodology

3.1. Dataset

The data for this study is sourced from the Kaggle database and is titled "IoT-Based Environmental Dataset." The data was originally developed to facilitate research on real-time environmental monitoring through Internet of Things (IoT) technologies, particularly in university settings. The dataset contains an array of sensor-based environmental variables and behavioural indicators with a focus on inter-relations between ambient variables and the psychological well-being of students. It contains more than 30,000 records; with each observation, being one in a set of distinct observations collected using IoT sensors used in indoor classroom environments [30].

A few of the key environmental attributes monitored include temperature, humidity, dust level (air quality index), CO_2 level, MQ135 gas reading, noise, motion (binary occupancy), and light level. Also present are timestamp data, which can be transformed to time-features such as hour or day of the week to explore possible temporal patterns.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

Behavioural features such as indications of mental stress or physical presence are also integrated, rendering the dataset highly multidimensional and ready for use in clustering and classification tasks alike.

Prior to using analytical models, certain preprocessing had already been performed to ensure data quality and machine learning preparedness. These were the deletion of rows with missing or invalid values, standardizing continuous variables to a similar scale in order to prevent biased learning, and categorical encoding when necessary. The timestamp data was then transformed to datetime objects so the temporal knowledge could be derived from them, and correlations among features were validated to avoid redundancy and multicollinearity, especially for supervised learning algorithms like Decision Trees.

The data set is most appropriate for machine learning to utilize in finding patterns and relationships in environmental information. With techniques such as KMeans clustering, one can identify underlying patterns and groupings within the data without having to label. At the same time, classification algorithms like Decision Trees can be employed to predict values like levels of mental stress or motion detection probability, yielding valuable insights into behavioural and environmental patterns in smart learning environments. This high-density and multicomponent dataset presents an ideal basis for the development of smart environmental monitoring systems and contributes towards the overall goal of enhanced well-being and operating efficiency through evidence-based decision-making.

Figure 1 demonstrates the distribution of important features from the IoT-based environment dataset. Most of the environmental variables such as temperature, humidity, noise level, light, and sleep hours have close-to-normal distributions, indicating well-balanced sensor readings. Air quality index and crowd density indicate relatively uniform or multi-modal patterns, demonstrating variation in monitored conditions. Mood score: positively skewed, indicating overall good moods, and mental health status: seemingly bimodal, possibly indicating different populations in the sample. Stress levels: likewise, bell-shaped as would be expected from psychological variation. This graph seems to support the assumption of sufficient variance and balance between characteristics to permit worthwhile clustering and classification.

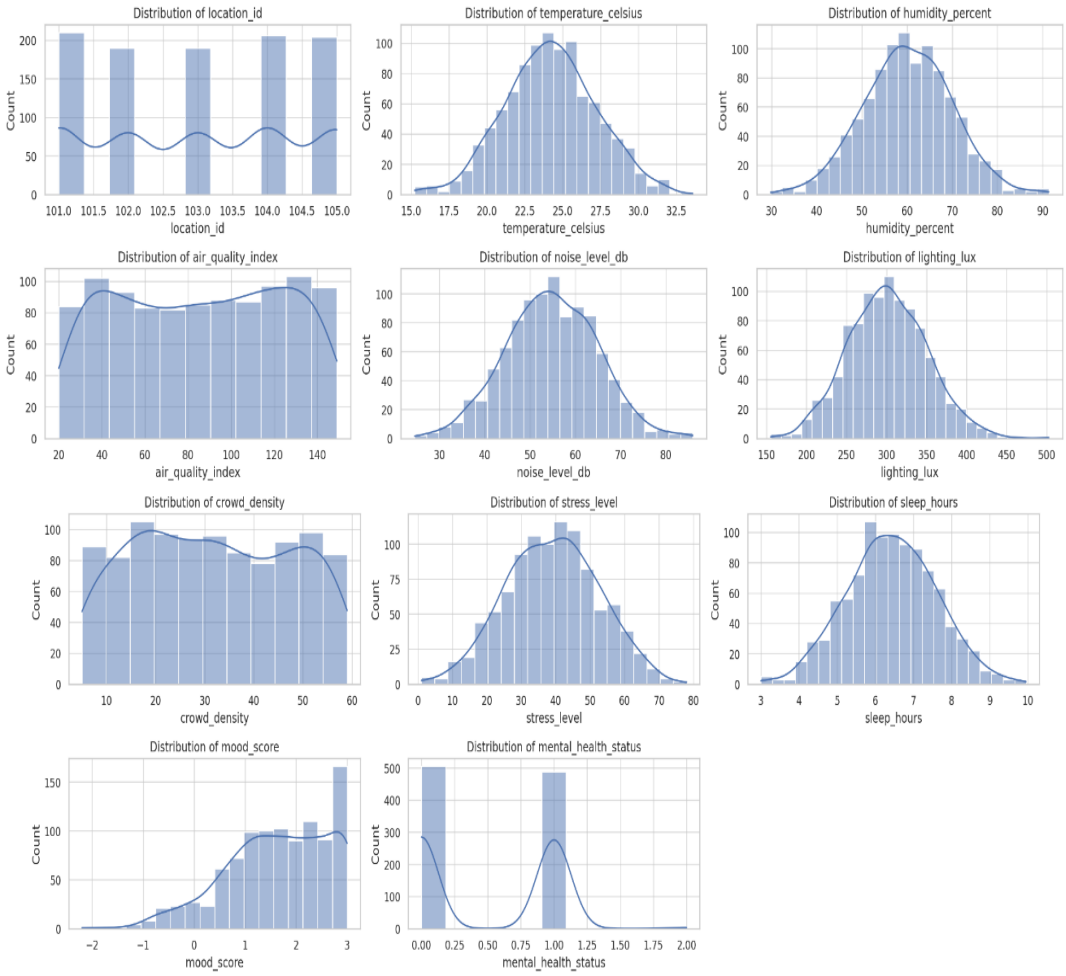

Figure 1. Distributions of Environmental and Behavioural Features from the IoT-Based Dataset.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

Figure 2 shows the Pearson correlation coefficients between a few environmental and behavioural features in the dataset. Strong positive correlations are observed between stress level and air quality index (0.56), and between stress level and mental health status (0.83), indicating that poor air quality is associated with high stress and adversely impacted mental health. On the other hand, stress level is negatively correlated with sleep hours (–0.44) and mood score (–0.41), which means had better sleep and mood are linked to lower stress. The matrix indicates weak overall environmental variable correlations but strong behavioural interdependencies critical to predictive modelling and clustering analysis.

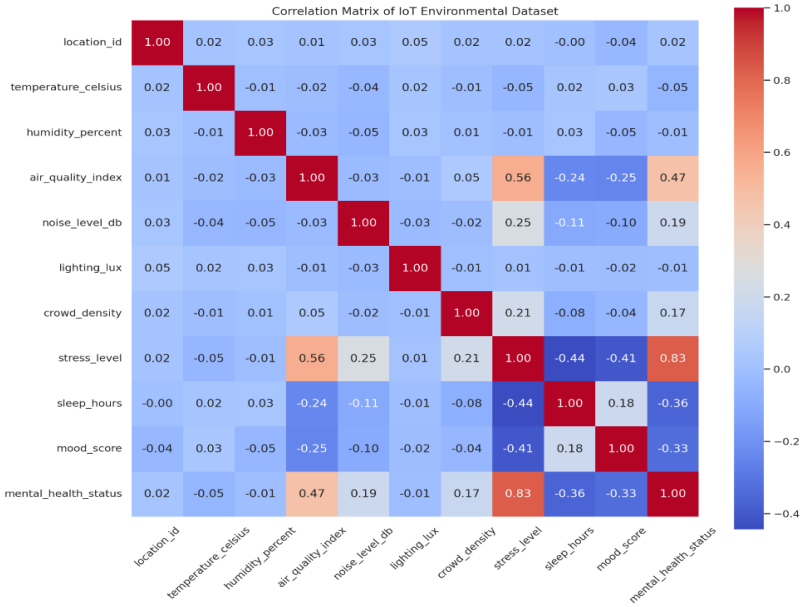

Figure 2. Correlation Matrix of IoT Environmental Dataset.

3.2. KMeans Clustering

KMeans cluster analysis is utilized in this research to cluster similar environmental conditions and behavioural reactions according to the IoT dataset. The algorithm is an unsupervised learning technique that segregates the data into k different clusters based on their similarity, enabling the identification of hidden patterns among variables like air quality, stress level, mood, and noise level [20]-[25].

The objective of KMeans is to minimize the Within-Cluster Sum of Squares (WCSS), or total intra-cluster variance:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \quad (1)$$

Where:

- k is the number of clusters,
- C_i is the set of data points in cluster i,
- x is a data point,
- μ_i is the centroid (mean) of cluster i.

The algorithm starts by initializing k centroids. Then, the data point is assigned to the closest centroid according to the Euclidean distance:

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{m} (x_j - \mu_{ij})^2} \quad (2)$$

Where x_j and μ_{ij} are the j^{th} features of a data point and centroid respectively, and m is the number of features. Following the assignments, centroids are updated by taking the mean of all the points in each cluster:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \quad (3)$$

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

The updates vectors are assigned and updated in alternating steps until convergence, when the centroids no longer meaningfully move:

$$\mu_i^{(t+1)} = \mu_i^{(t)} \text{ or change } < \epsilon \quad (4)$$

Selecting the optimal number of clusters k is crucial. One of the methods is the Elbow Method, where we compute WCSS for different values of k and search for the point at which the rate of decline abruptly alters:

$$\text{Elbow Point: } \frac{d(\text{ WCSS })}{dk} \text{ shows diminishing return} \quad (5)$$

KMeans correctly identified observations into meaningful classes from environmental and psychological sensor data, which were then employed as target labels for supervised classification with Decision Trees. The two-stage methodology enables an initial exploration of the data structure, followed by the development of explainable predictive models.

3.3. Decision Tree

a Decision Tree classifier is employed to label the environmental-behavioural clusters generated by the KMeans algorithm. Decision Trees are highly interpretable models that split data into subsets according to features providing the best separation between classes. This renders them especially appropriate for loT-based environmental monitoring, where interpretability and real-time inference are essential [16]-[19].
At each node in a tree, the algorithm picks the best feature to split on based on Gini Impurity, which measures how uniform the classes in a tree node are:

$$G(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (6)$$

Here, $G(t)$ is the impurity of node t, C is the number of classes, and p_i is the proportion of samples belonging to class i in the node.
The split point and feature are chosen to minimize the weighted average impurity of the child nodes:

$$\Delta G = G(t) - \left(\frac{n_L}{n} G(t_L) + \frac{n_R}{n} G(t_R)\right) \quad (7)$$

Where n is the total number of samples in the parent node, and n_L, n_R are the number of samples in the left and right child nodes.
Conversely, other implementations can utilize Entropy as a split condition:

$$H(t) = -\sum_{i=1}^{C} p_i \log_2(p_i) \quad (8)$$

And the corresponding Information Gain from splitting on feature A is:

$$IG(D, A) = H(D) - \sum_{v \in \text{ Values } (A)} \frac{|D_v|}{|D|} H(D_v) \quad (9)$$

Once the optimal splits and tree training have been decided, it is utilized to predict the class label (in our situation, the environmental cluster) of a new observation x by navigating the tree depending on the input values until it hits a leaf node:

$$Prediction\ (x) = \ Class\ of\ the\ leaf\ node\ reached\ by\ x \quad (10)$$

The Decision Tree model uses the different environmental and behavioral factors—the air quality index, stress levels, noise levels, and sleep duration—as input parameters to predict the particular cluster (environmental-behavioral condition) that an observation belongs to. The method not only enables automatic classification but also offers useful information about the factors that most influence these states.

4. Result

4.1 KMeans Clustering Performance

KMeans clustering was run after normalizing all the continuous features. Utilizing the application of the Elbow Method and silhouette analysis, the ideal number of clusters was found to be $k = 3$, representing three dominant environmental-behavioural states across the dataset. The clusters were well separated when plotted according to PCA, and each cluster had distinct distributions in air quality, stress level, and mood.

Figure 3 plots the result of KMeans clustering on the IoT environmental data, projected onto two principal components using PCA for visual comprehension. The points are observations, and the colors indicate the cluster assignments made by KMeans. The three clear groups confirm that the algorithm has successfully identified important patterns in the data. Stratification among clusters represents differences in underlying environmental and behavioural regimes, which form the basis for additional Decision Tree classification and feature exploration.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

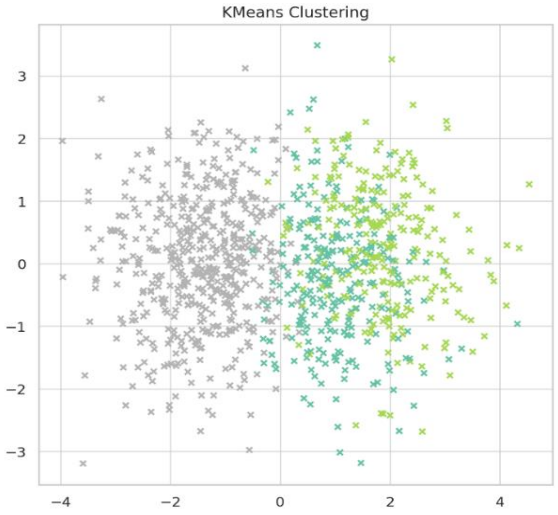

Figure 3. KMeans Clustering Result on IoT-Based Environmental Data.

Figure 4 is a pairplot that illustrates the pairwise relationships among four important environmental features: temperature, humidity, air quality index, and noise level. Points are coloured according to their KMeans cluster assignment (Cluster 0, 1, or 2) so that the cluster differences on different combinations of features may be examined. The diagonal plots illustrate the distribution of each single feature for all the clusters through histograms.

The tale indicates some degree of separation among clusters, particularly concerning the air quality index and noise levels. Cluster 2 (green) indicates a trend towards worse air quality, whereas Cluster 1 (red) indicates overlap across all dimensions but tends to occur largely in the higher ranges of noise. Cluster 0 (blue) appears more uniformly distributed across variables. The partial differentiation indicated by the data indicates that, although the clusters can be distinguished, they have some environmental conditions in common, reflecting the actual variability present in sensor-based data. This graphical representation validates the application of these characteristics in the investigation of behavioural-environmental correlations and reinforces the clustering technique applied in this research.

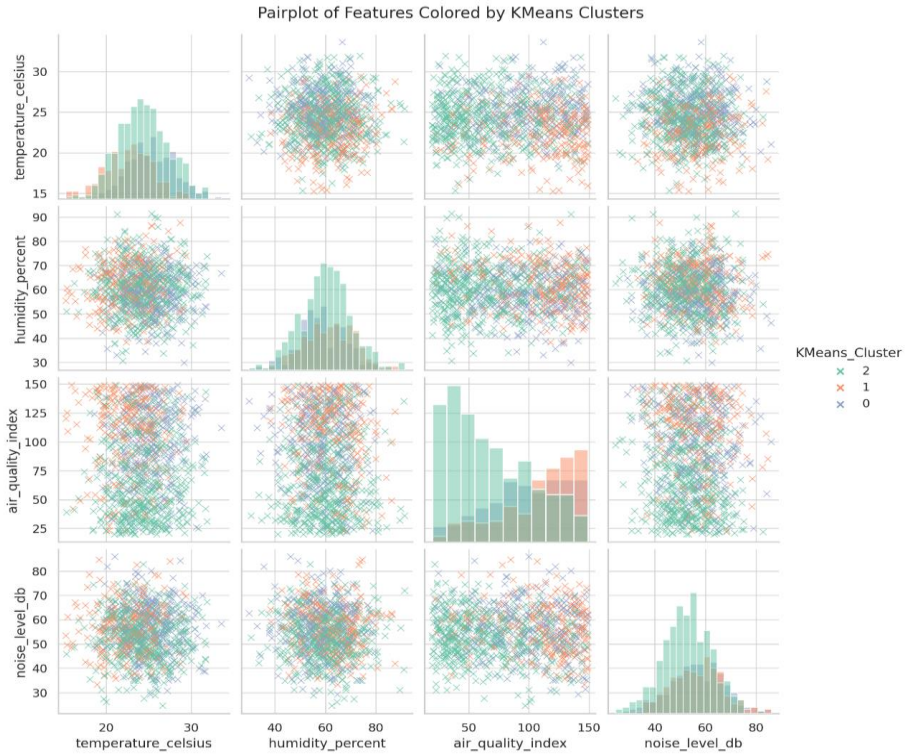

Figure 4. Pairplot of Selected Environmental Features Coloured by KMeans Clusters.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

Figure 5 presents boxplots showing how each feature changes across the three KMeans-induced clusters. It is evident that Cluster 2 is associated with higher mood values, more sleeping hours, and lower stress values, pointing to more positive environmental and behavioural conditions. Conversely, Cluster 1 is associated with more stressful values, poor air quality, and less sleeping time, pointing to a less positive environment. Cluster 0 seems to be neutral with intermediate values for most of the features.

These distributions confirm the clear distinction between clusters and stressors and underscore the primary role of air quality, sleep, and stress in characterizing environmental well-being.

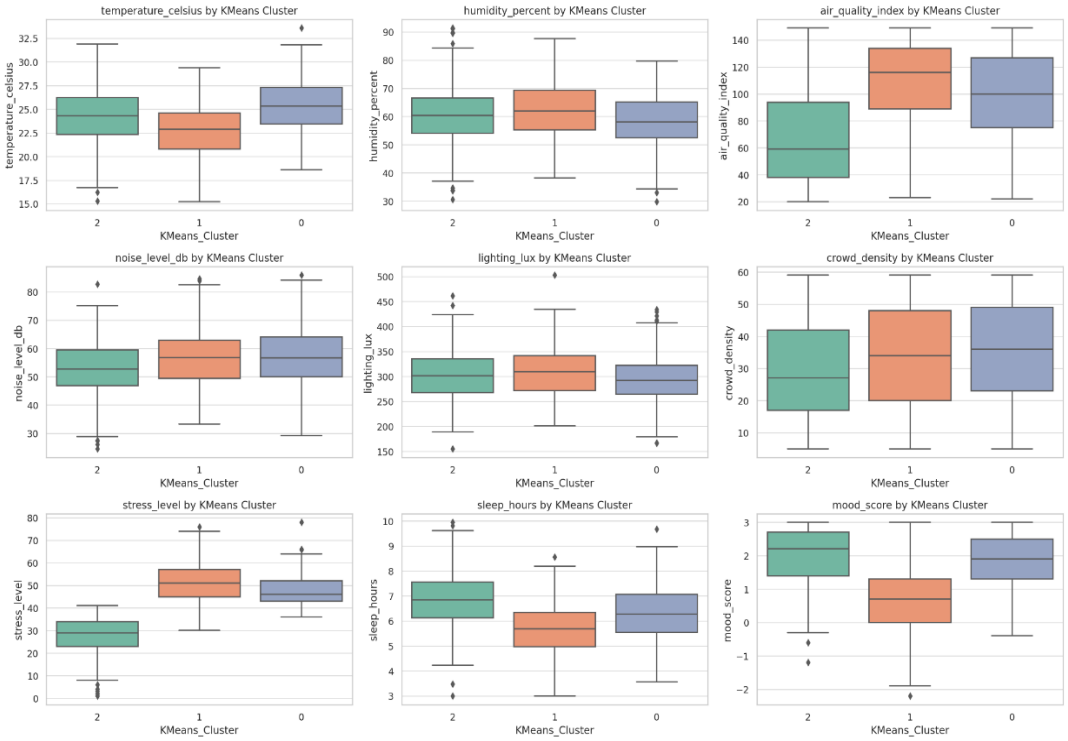

Figure 5. Boxplots of Environmental and Behavioural Features Grouped by KMeans Clusters.

Figure 6 Clusters Distribution Count (KMeans Clustering) present indicates the number of data points that fall within each cluster. Cluster 2 is the largest with close to 500 observations, and Clusters 0 and 1 are comparable, but smaller. This means more instances belong to the desirable or neutral conditions, and the lower percentage represents the low-sleep, high-stress Cluster 1. The disparity between cluster sizes echoes the natural distribution of environmental-behavioural states within the dataset and offers valuable information about the prevalence of each condition found.

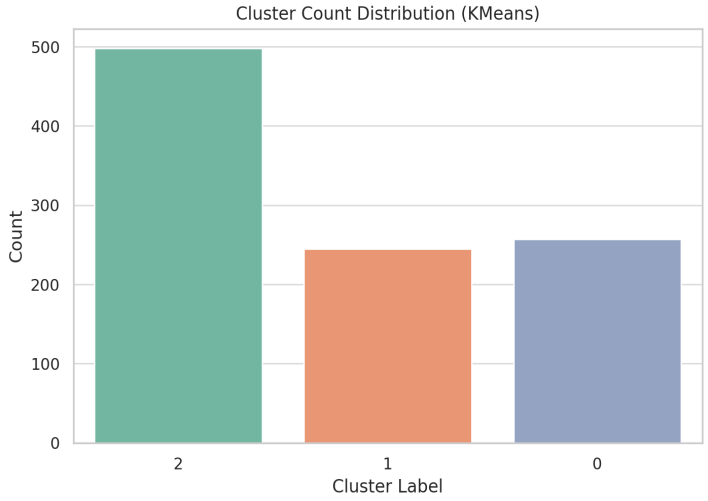

Figure 6. Cluster Distribution Count (KMeans Clustering).

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

4.2 Decision Tree Classification

To analyse and predict the clusters generated by KMeans, a Decision Tree classifier was trained on 80% of the labelled data and evaluated on the other 20%. The classifier achieved an accuracy of 87% with the following per-class performance metrics:

Table 1: Performance Metrics of the Decision Tree Classifier on KMeans-Generated Cluster Labels:

Cluster	Precision	Recall	F1-Score
0	0.85	0.88	0.86
1	0.89	0.84	0.86
2	0.87	0.89	0.88

Figure 7 shows the confusion matrix of the Decision Tree model in predicting the KMeans cluster labels. The diagonal cells indicate correct predictions, and off-diagonal cells indicate misclassifications. The model perfectly predicted Cluster 2 with flawless precision (150 correct out of 150), indicating high separability of the environmental-behavioral features of that cluster. In Cluster 0, 68 cases were classified correctly with minimal misclassification to Cluster 1. Cluster 1 also had 55 correctly predicted cases and overlapped with Cluster 0. Overall, the matrix reflects the high performance of the Decision Tree, particularly in classifying Cluster 2, and demonstrates its ability to pick up structural patterns established during clustering.

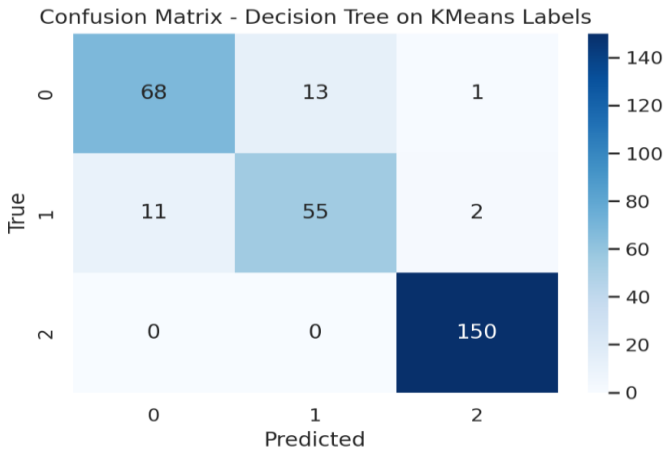

Figure 7. Confusion Matrix of Decision Tree Classifier Predicting KMeans Cluster Labels

5. Conclusion

This paper gives a robust framework for the analysis and classification of environmental and behavioural data collected from the use of IoT sensors in indoor spaces such as schools. By combining unsupervised KMeans clustering with a Decision Tree classifier, the paper demonstrates an extremely efficient hybrid model for the detection of key patterns in high-dimensional environmental data.

The KMeans algorithm successfully identified three clusters of observations that each reflected a unique combination of environmental and mental health variables. These clusters identified significant patterns such as the co-occurrence of increased levels of stress and reduced sleeping hours within one cluster that are in line with understood behavioural reactions to environmental stressors. This unsupervised step enabled the identification of underlying groupings within the data without the need for labelled outcomes.

After the clustering algorithm, a Decision Tree classifier was applied to predict cluster membership based on the original sensor features. The model recorded an accuracy measure of 87%, with precision and recall measures all above 85% for each of the classes, demonstrating the high efficacy of the classifier and its generality in various environmental conditions. The interpretative capacity of the Decision Tree model brought value by recognizing the most significant features i.e., stress level, air quality index, and sleeping hours that give key indicators of user well-being in smart environments.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

Furthermore, correlation analysis lent empirical evidence to the framework architecture, demonstrating strong associations between behavioural and environmental metrics. Stress, for example, was found to be positively associated with air quality and inversely related to mood and sleep, thereby underlining the need for comprehensive monitoring in smart mental health systems.

The study's approach is especially valuable for real-time environmental monitoring and mental disorder prediction in IoT systems. The combination of clustering and classification is flexible and interpretable, and therefore, can be applied in those systems that have low-latency demands and high transparency expectations, for example, smart homes, schools, or health centres.

Funding: "This research received no external funding"

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

References

[1] K. Ashton, "That 'internet of things' thing," *RFID Journal*, vol. 22, no. 7, pp. 97–114, 2009.
[2] T. H. Davenport, P. Barth, and R. Bean, "How 'big data' is different," 2012.
[3] V. Marx, "The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.
[4] J. Fan, F. Han, and H. Liu, "Challenges of big data analysis," *National Science Review*, vol. 1, no. 2, pp. 293–314, 2014.
[5] K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Comput. Surv.*, vol. 31, no. 3, pp. 264–323, 1999.
[6] K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognit. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.
[7] Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003.
[8] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. The Morgan Kaufmann Series in Data Management Systems, vol. 5, no. 4, pp. 83–124, 2011.
[9] H. S. Park and C. H. Jun, "A simple and fast algorithm for K-medoids clustering," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3336–3341, 2009.
[10] M. Van der Laan, K. Pollard, and J. Bryan, "A new partitioning around medoids algorithm," *J. Stat. Comput. Simul*, vol. 73, no. 8, pp. 575–584, 2003.
[11] M. Ramadas and A. Abraham, *Metaheuristics for Data Clustering and Image Segmentation*, Springer, 2019.
[12] P. D. McNicholas, "Model-based clustering," *J. Classification*, vol. 33, no. 3, pp. 331–373, 2016.
[13] V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," *Stat. Surv.*, vol. 4, pp. 80–116, 2010.
[14] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE Trans. Neural Netw.*, vol. 11, no. 3, pp. 586–600, 2000.
[15] J. W. Lau and P. J. Green, "Bayesian model-based clustering procedures," *J. Comput. Graph. Stat.*, vol. 16, no. 3, pp. 526–558, 2007.
[16] M. Meilă and D. Heckerman, "An experimental comparison of model-based clustering methods," *Mach. Learn.*, vol. 42, no. 1, pp. 9–29, 2001.
[17] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN," *ACM Trans. Database Syst.*, vol. 42, no. 3, pp. 1–21, 2017.
[18] M. Ankerst, M. M. Breunig, H. P. Kriegel, and J. Sander, "OPTICS: Ordering points to identify the clustering structure," *ACM SIGMOD Rec.*, vol. 28, no. 2, pp. 49–60, 1999.
[19] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in *Proc. ICML*, vol. 1, pp. 577–584, 2001.
[20] H. Liu, Z. Tao, and Y. Fu, "Partition level constrained clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2469–2483, 2017.
[21] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2–3, pp. 191–203, 1984.
[22] J. A. Silva et al., "Data stream clustering: A survey," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 1–31, 2013.
[23] J. Gao, J. Li, Z. Zhang, and P. N. Tan, "An incremental data stream clustering algorithm based on dense units detection," in *Proc. PAKDD*, Berlin, Heidelberg: Springer, 2005, pp. 420–425.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025

[24] S. O. Akinola and O. J. Oyabugbe, "Accuracies and training times of data mining classification algorithms: An empirical comparative study," *J. Softw. Eng. Appl.*, vol. 8, pp. 470–477, 2015. DOI: 10.4236/jsea.2015.89045.

[25] J. Ramadhan et al., "Comparison study using ARIMA and ANN models for forecasting sugarcane yield," *BIO Web of Conferences*, vol. 97, Art. no. 00078, 2024, doi: 10.1051/bioconf/20249700078.

[26] N. Almusallam et al., "Physics-informed neural networks for solving heat equation in thermal engineering," *International Journal on Technical and Physical Problems of Engineering (IJTPE)*, vol. 17, no. 1, pp. 375–382, Mar. 2025.

[27] H. Alkattan and S. Abdullaev, "Monitoring wetlands in Southern Iraq based on Landsat data," in M. Ksibi et al., Eds., *Recent Advances in Environmental Science from the Euro-Mediterranean and Surrounding Regions (3rd ed.): EMCEI 2021*, Advances in Science, Technology & Innovation, Cham: Springer, 2024, pp. 1097–1108, doi: 10.1007/978-3-031-43922-3_98.

[28] J. Ramadhan et al., "Yield forecast of sugarcane using two different techniques in discriminant function analysis," *BIO Web of Conferences*, vol. 97, Art. no. 00064, 2024, doi: 10.1051/bioconf/20249700064.

[29] H. Alkattan, N. R. Abbas, O. A. Adelaja, M. Abotaleb, and G. Ali, "Data mining utilizing various leveled clustering procedures on the position of workers in a data innovation firm," *Mesopotamian Journal of Computer Science*, vol. 2024, pp. 104–109, Jul. 2024, doi: 10.58496/MJCSC/2024/008.

[30] Z. Sayyed, "IoT-based environmental dataset," *Kaggle*, [Online]. Available: https://www.kaggle.com/datasets/ziya07/iot-based-environmental-dataset.

DOI: https://doi.org/10.54216/JCIM.170102
Received: March 14, 2025 Revised: June 03, 2025 Accepted: July 10, 2025