AI-Driven Roommate Matching for University Accommodation: A Clustering-Based Approach

KADDACHE Mohammed El Amine
university of ain temouchent
department of computer science
Master: cybersecurity and artificial intelligence (CYSIA)
kaddache66@gmail.com

ELMEGUENNI Nabil
university of ain temouchent
department of computer science
Master: cybersecurity and artificial intelligence (CYSIA)
nazj4@gmail.com

***Abstract* :—University accommodation significantly influences student well-being and academic performance, yet traditional room allocation methods often neglect lifestyle compatibility. This paper presents an AI-driven roommate matching system developed for the National Programming and Artificial Intelligence Competition (NPAIC'24) at the University Center El Bayadh. We employed unsupervised machine learning techniques, specifically K-Means and Hierarchical Clustering, to group 4,000 synthetic student profiles based on 25 behavioral and demographic features. After preprocessing and dimensionality reduction using PCA, K-Means achieved optimal clustering with K=8 (Silhouette Score: 0.0789), while Hierarchical Clustering identified 5 distinct groups (Silhouette Score: 0.0347). Our findings demonstrate that algorithmic matching can effectively identify compatible roommate pairs, potentially reducing social friction and enhancing residential satisfaction within the Smart Campus framework.**

***Index Terms* :—Artificial Intelligence, K-Means Clustering, Hierarchical Clustering, University Accommodation, Roommate Matching, Smart Campus**

I. Introduction

The evolution of higher education infrastructure increasingly relies on Artificial Intelligence to optimize services beyond the classroom. While AI applications in academic analytics and campus logistics have gained traction [1], student accommodation—a critical determinant of well-being and retention—remains underutilized in this transformation. Traditional hostel allocation methods, predominantly manual or rule-based, often result in incompatible roommate pairings, leading to interpersonal conflicts and diminished academic outcomes.

Recent studies underscore the importance of lifestyle compatibility in shared living environments. Misalignments in sleep schedules, cleanliness standards, and social habits contribute to elevated stress levels among students [2]. The challenge is particularly acute for Algeria's expanding higher education sector, where the "Digital Algeria 2030" initiative prioritizes service digitalization through platforms like PROGRES [3]. However, these systems currently lack intelligent decision-making capabilities for personalized student services.

This paper addresses this gap by presenting an AI-driven roommate matching system developed for NPAIC'24. Our approach leverages unsupervised clustering algorithms to analyze multidimensional student profiles, enabling data-driven compatibility assessments. By grouping students with similar behavioral patterns and lifestyle preferences, we aim to minimize dormitory conflicts and foster harmonious residential communities.

The remainder of this paper is structured as follows: Section II reviews relevant literature; Section III details our methodology; Section IV presents experimental results; and Section V concludes with implications and future directions.

II. Related Work

A. Smart Campus Initiatives

The Smart Campus paradigm conceptualizes universities as integrated ecosystems managed through IoT, big data, and AI technologies. Research demonstrates AI's efficacy in various university operations: Artificial Neural Networks (ANN) predict campus transportation demand [4], while Lasso Regression models optimize catering services and reduce food waste by up to 30% [5]. These applications validate AI's capacity to manage complex, non-linear institutional data.

B. AI in Residential Services

Within accommodation management, AI adoption has progressed from administrative automation to predictive analytics:

Operational Efficiency: AI-powered chatbots and automated booking systems reduce administrative burdens by approximately 40%, streamlining check-in processes and maintenance requests [6].

Predictive Matching: The StayMate platform exemplifies advanced roommate matching through hybrid algorithms combining rule-based filtering with collaborative filtering. By analyzing personality metrics (e.g., Big Five traits) and behavioral data (sleep patterns, study habits), such systems significantly mitigate roommate conflicts [2].

Resource Optimization: Decision Tree algorithms predict occupancy rates with 95% accuracy, enabling dynamic pricing and energy-efficient IoT integration for climate control [6].

C. Algorithmic Matching Techniques

Clustering algorithms are particularly suited for roommate matching due to their ability to identify natural groupings in unlabeled data. K-Means partitions students into K clusters by minimizing intra-cluster variance, while Hierarchical

Clustering builds nested groupings via agglomerative (bottom-up) or divisive (top-down) strategies. Both methods have been successfully applied to student segmentation tasks in educational contexts.

III. Methodology

A. Dataset Generation

Given the absence of real-world student accommodation datasets, we synthesized a comprehensive dataset of 4,000 student profiles. Each profile comprises 25 features spanning demographic, academic, behavioral, and lifestyle dimensions (Table I).

TABLE I
Feature Categories in Synthetic Dataset

Category	Features
Demographic	Age, Height, Weight
Academic	Bac GPA, Bac Stream, Major
Lifestyle	Sleep Schedule, Cleanliness Level
Health	Smoking, Chronic Conditions, Allergies
Social	Introversion Scale, Guests Frequency
Preferences	Room Temperature, Study Environment

Data generation employed NumPy's random functions with seed initialization (seed=42) to ensure reproducibility. Categorical variables (e.g., majors, sleep habits) were sampled from realistic distributions, while continuous variables (e.g., age, GPA) were drawn from appropriate uniform or normal distributions. The dataset intentionally overrepresents healthy students (75% with no chronic conditions) to reflect typical university populations.

B. Data Preprocessing

Effective clustering requires homogeneous feature scaling. Our preprocessing pipeline consisted of:

Encoding: Categorical features (17 variables) were transformed using Label Encoding, converting text labels to numerical representations suitable for distance-based algorithms.

Normalization: All 25 features underwent Min-Max Scaling, mapping values to the [0, 1] range:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

This prevents features with larger magnitudes (e.g., Home Distance: 5–800 km) from dominating distance calculations over smaller-scale features (e.g., Cleanliness Level: 1–5).

C. Dimensionality Reduction

To mitigate the curse of dimensionality and enhance computational efficiency, we applied Principal Component Analysis (PCA) with 95% variance retention. PCA linearly transforms the original 25 features into 23 orthogonal principal components, capturing 96.49% of dataset variance. This reduction eliminates redundant information while preserving the data's intrinsic structure.

D. Clustering Algorithms

1) K-Means Clustering: K-Means partitions data into K clusters by iteratively:

1) Initializing K centroids randomly
2) Assigning each sample to its nearest centroid (Euclidean distance)
3) Recalculating centroids as cluster means
4) Repeating until convergence

We determined the optimal K using two validation metrics:

- **Elbow Method:** Plots inertia (within-cluster sum of squares) against K. The "elbow point" indicates diminishing returns in cluster compactness.
- **Silhouette Score:** Measures cluster cohesion and separation, ranging from -1 (poor) to +1 (excellent):

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{2}$$

where $a(i)$ is the mean intra-cluster distance and $b(i)$ is the mean nearest-cluster distance.

2) Hierarchical Clustering: Agglomerative Hierarchical Clustering builds a dendrogram by:

1) Treating each sample as a singleton cluster
2) Merging the two closest clusters (Ward linkage: minimizes variance)
3) Repeating until all samples form a single cluster

We extracted the optimal flat clustering by cutting the dendrogram at the height maximizing the Silhouette Score across K = 2–10 clusters.

E. Implementation

The system was implemented in Python 3.10 using:

- **Data Handling:** Pandas, NumPy
- **Machine Learning:** Scikit-learn (MinMaxScaler, PCA, KMeans, AgglomerativeClustering)
- **Visualization:** Matplotlib, Seaborn, SciPy (dendrogram)

IV. Results and Discussion

A. Dimensionality Reduction

PCA successfully reduced the feature space from 25 to 23 dimensions while retaining 96.49% of variance. Figure 6 visualizes the dataset in the first three principal components (PC1, PC2, PC3), which collectively explain approximately 13.5% of variance. The visualization reveals no obvious natural clustering, justifying the application of algorithmic methods.

B. K-Means Clustering Results

1) Optimal K Selection: The Elbow Method (Fig. 2) shows inertia declining smoothly from 13,080 (K=2) to 10,800 (K=10), with a noticeable inflection around K=5–6. The Silhouette Score analysis (Fig. 2) identifies K=8 as optimal, achieving a score of 0.0789—indicating modest but meaningful cluster separation.

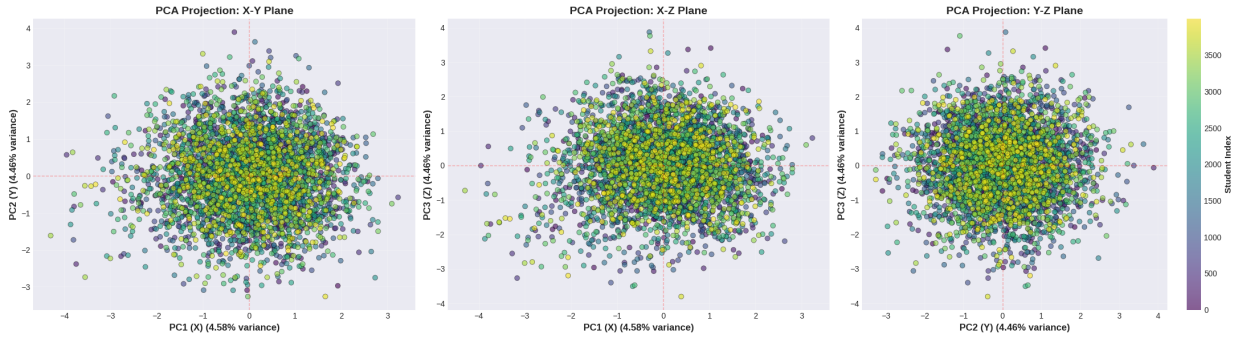

Fig. 1. 3D PCA projection showing student distribution across the first three principal components. Color gradient represents Student ID. The lack of distinct visual clusters highlights the complexity of the dataset.

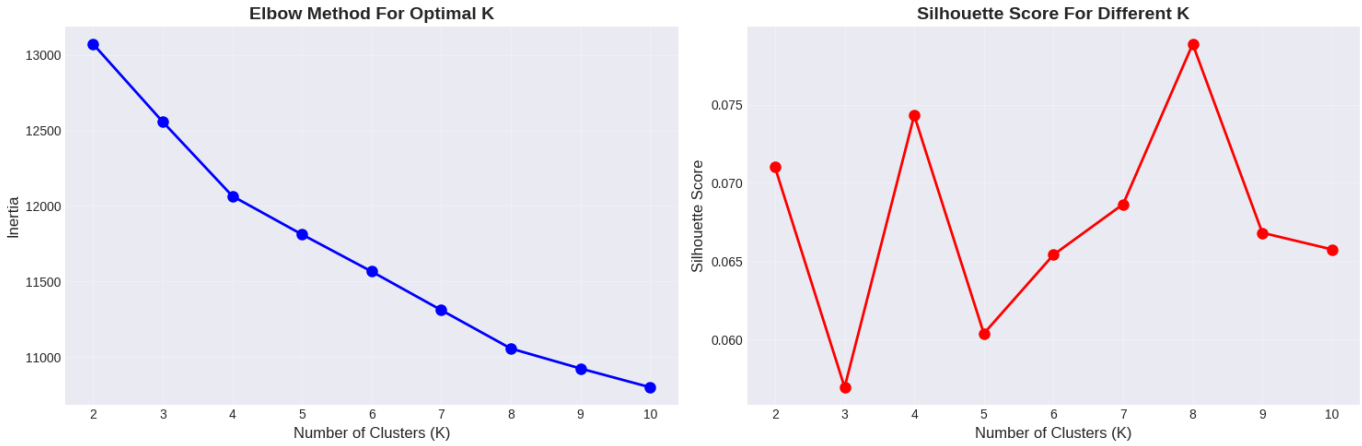

Fig. 2. K-Means validation metrics. Left: Elbow Method showing inertia decay. Right: Silhouette Scores peaking at K=8, selected as the optimal configuration.

TABLE II
K-MEANS CLUSTER PROFILES (K=8)

Cluster	Size	Sleep	Study Env.	Intro.
0	482	Early Bird	Music	5.65
1	493	Early Bird	Group	5.36
2	507	Night Owl	Group	5.58
3	511	Night Owl	Music	5.43
4	521	Night Owl	Group	5.72
5	503	Early Bird	Silence	5.50
6	530	Night Owl	Music	5.68
7	453	Early Bird	Group	5.73

2) Cluster Characteristics: K-Means with K=8 produced relatively balanced clusters (453–530 students each). Table II summarizes key cluster profiles:

Notably, clusters differentiate primarily along sleep schedules (Early Bird vs. Night Owl) and study environments (Silence, Music, Group). Introversion scales remain relatively homogeneous (5.36–5.73), suggesting balanced extroversion-introversion distributions across clusters. Figure 5 visualizes cluster distributions in PCA space.

C. Hierarchical Clustering Results

1) Dendrogram Analysis: The complete dendrogram (Fig. 3) reveals hierarchical relationships among students. Applying a cut-off threshold at distance 8.7, we identified 5 optimal clusters (Silhouette Score: 0.0347).

2) Cluster Characteristics: Hierarchical Clustering produced less balanced clusters (549–1,117 students), with Cluster 2 being the largest. Figure 4 shows the Silhouette Score optimization, peaking at K=5.

Table III contrasts cluster profiles:

TABLE III
HIERARCHICAL CLUSTER PROFILES (K=5)

Cluster	Size	Sleep	Smoke	Clean.
0	930	Night Owl	No	2.85
1	824	Early Bird	No	2.98
2	1,117	Early Bird	No	3.09
3	549	Night Owl	Yes	3.12
4	580	Early Bird	No	2.98

Cluster 3 uniquely captures smoking students with Night Owl tendencies—a critical distinction for accommodation allocation to prevent conflicts with non-smokers. Figure **??** illustrates the clustering structure.

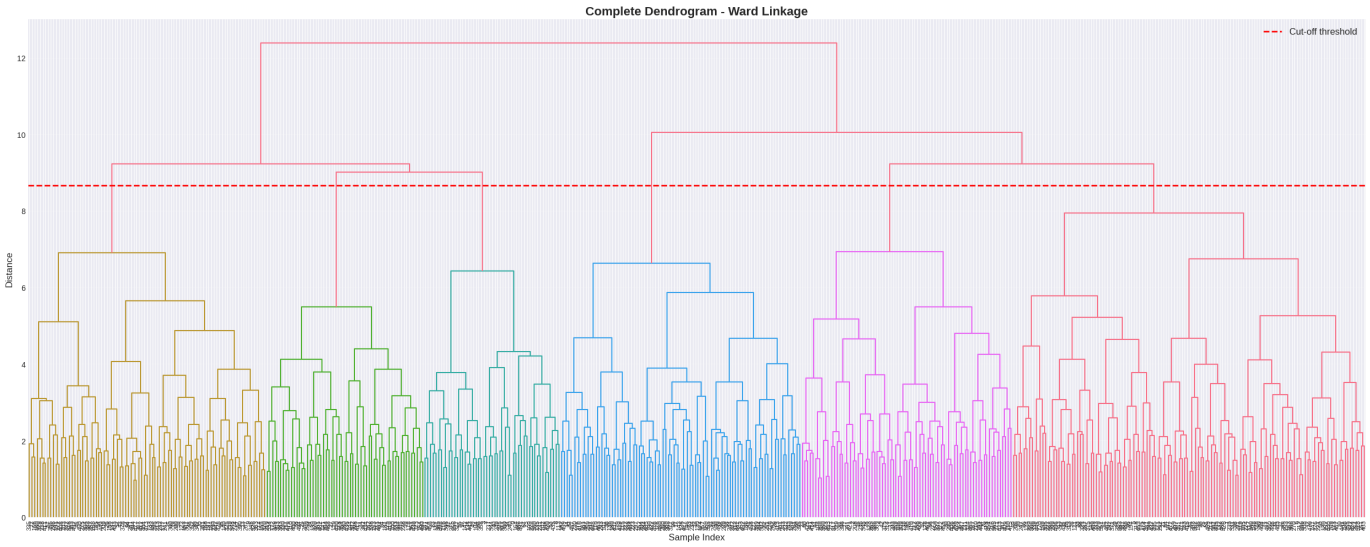

Fig. 3. Complete linkage dendrogram (Ward method). The red dashed line indicates the optimal cut-off threshold, yielding 5 clusters. Color-coded branches represent the identified groups.

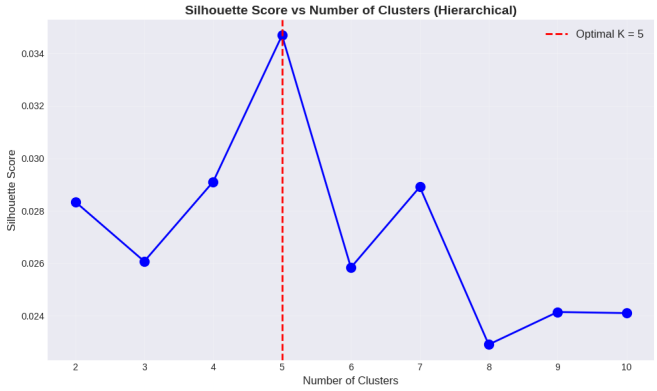

Fig. 4. Silhouette Score optimization for Hierarchical Clustering, identifying K=5 as optimal.

D. Comparative Analysis

- **Cluster Quality:** K-Means (0.0789) outperforms Hierarchical Clustering (0.0347) in Silhouette Score, indicating better-defined clusters.
- **Balance:** K-Means produces more evenly distributed groups, simplifying room allocation logistics.
- **Interpretability:** Hierarchical Clustering's dendrogram provides intuitive hierarchical relationships, while K-Means offers clearer separation along lifestyle dimensions.
- **Smoking Detection:** Hierarchical Clustering's Cluster 3 effectively isolates smokers—a critical feature for policy-compliant allocations.

E. Practical Implications

Our system enables universities to:

1) **Automate Matching:** Replace manual allocation with algorithmic assignments based on compatibility scores.
2) **Reduce Conflicts:** Pair students with aligned sleep schedules, cleanliness standards, and social preferences.
3) **Enhance Satisfaction:** Improve residential well-being, potentially boosting retention rates.
4) **Policy Enforcement:** Automatically segregate smokers to designated dormitories.

F. Limitations

- **Synthetic Data:** Real-world validation with actual student surveys is necessary.
- **Low Silhouette Scores:** Modest scores suggest overlapping clusters, possibly due to high-dimensional feature space or lack of strong natural groupings.
- **Static Features:** Current model ignores evolving preferences; longitudinal tracking could refine matches.

V. Conclusion and Future Work

This paper presented an AI-driven roommate matching system leveraging K-Means and Hierarchical Clustering to optimize university accommodation. Using a synthetic dataset of 4,000 students with 25 behavioral features, we demonstrated that unsupervised learning can effectively segment students into compatible groups. K-Means (K=8) achieved superior cluster quality, while Hierarchical Clustering (K=5) provided interpretable hierarchical structures and successfully isolated smoking populations.

Our work contributes to the Smart Campus ecosystem by showcasing practical AI applications beyond academic analytics. By replacing rule-based allocation with data-driven matching, universities can foster harmonious residential communities, directly impacting student well-being and academic success.

Future Directions:

- **Real-World Deployment:** Collect actual student data through surveys integrated with PROGRES.

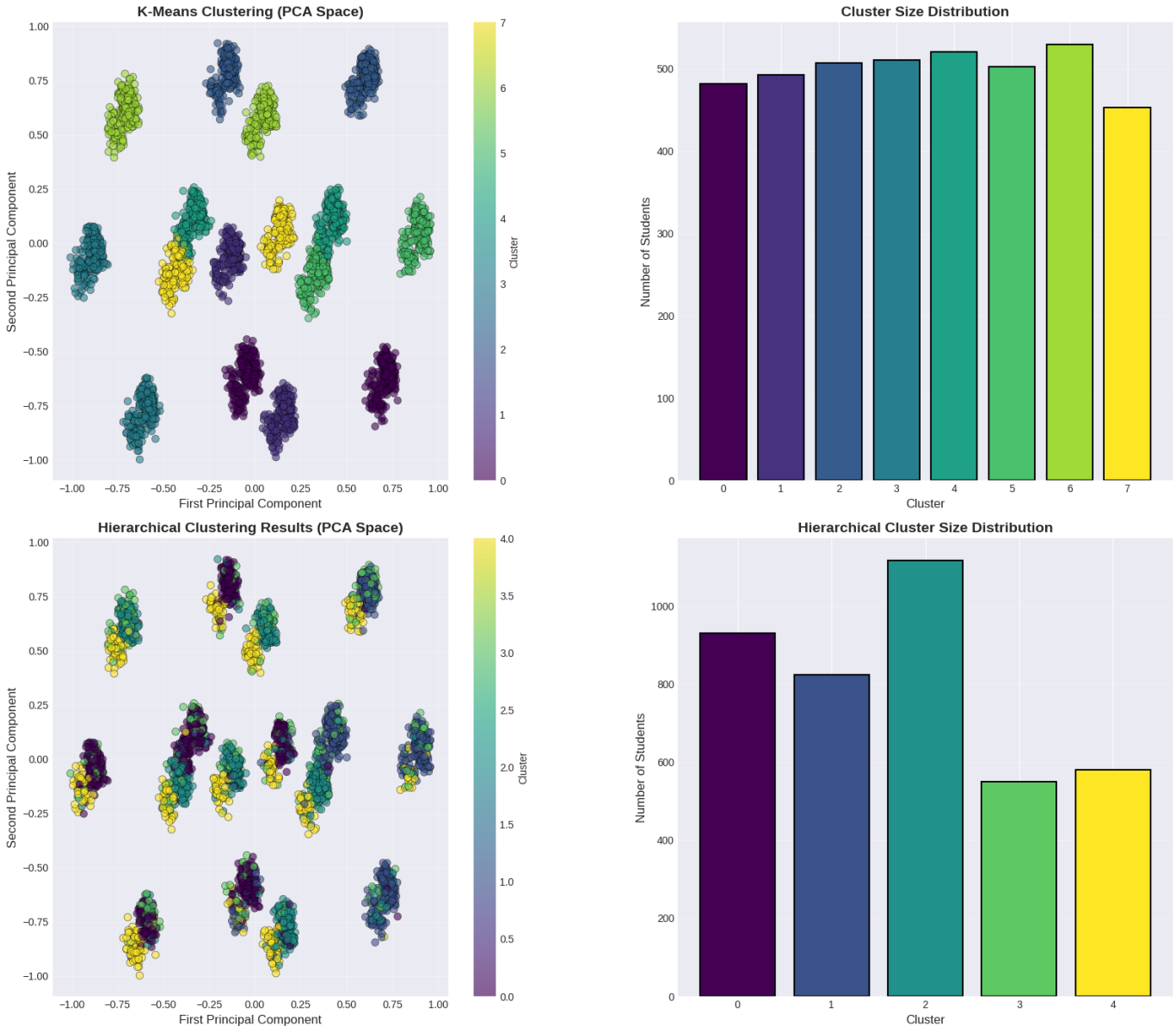

Fig. 5. K-Means and Hierarchical clustering visualization. Top-left: PCA space projection with cluster assignments. Top-right: Cluster size distribution showing balanced partitioning. Bottom-left: 3D cluster visualization. Bottom-right: Average feature heatmap revealing subtle inter-cluster differences.

- **Hybrid Models:** Combine clustering with preference-based algorithms (collaborative filtering) for personalized recommendations.
- **Feedback Loops:** Implement post-assignment surveys to refine models based on roommate satisfaction metrics.
- **Deep Learning:** Explore autoencoders for nonlinear dimensionality reduction and improved feature representation.
- **Scalability:** Test algorithms on larger datasets (10,000+ students) across multiple universities.

By integrating AI into accommodation services, Algerian universities can advance toward the Digital Algeria 2030 vision, delivering personalized, efficient, and student-centric residential experiences.

VI. Acknowledgment

We extend our thanks to the University Center of Nour El-Bachir, El Bayadh, for this kind hospitality, warm reception, and the excellent organization of the First Edition of the National Competition in Programming and Artificial Intelligence. We also thank the esteemed professors for their valuable contributions and our fellow students for their spirit of fair competition.

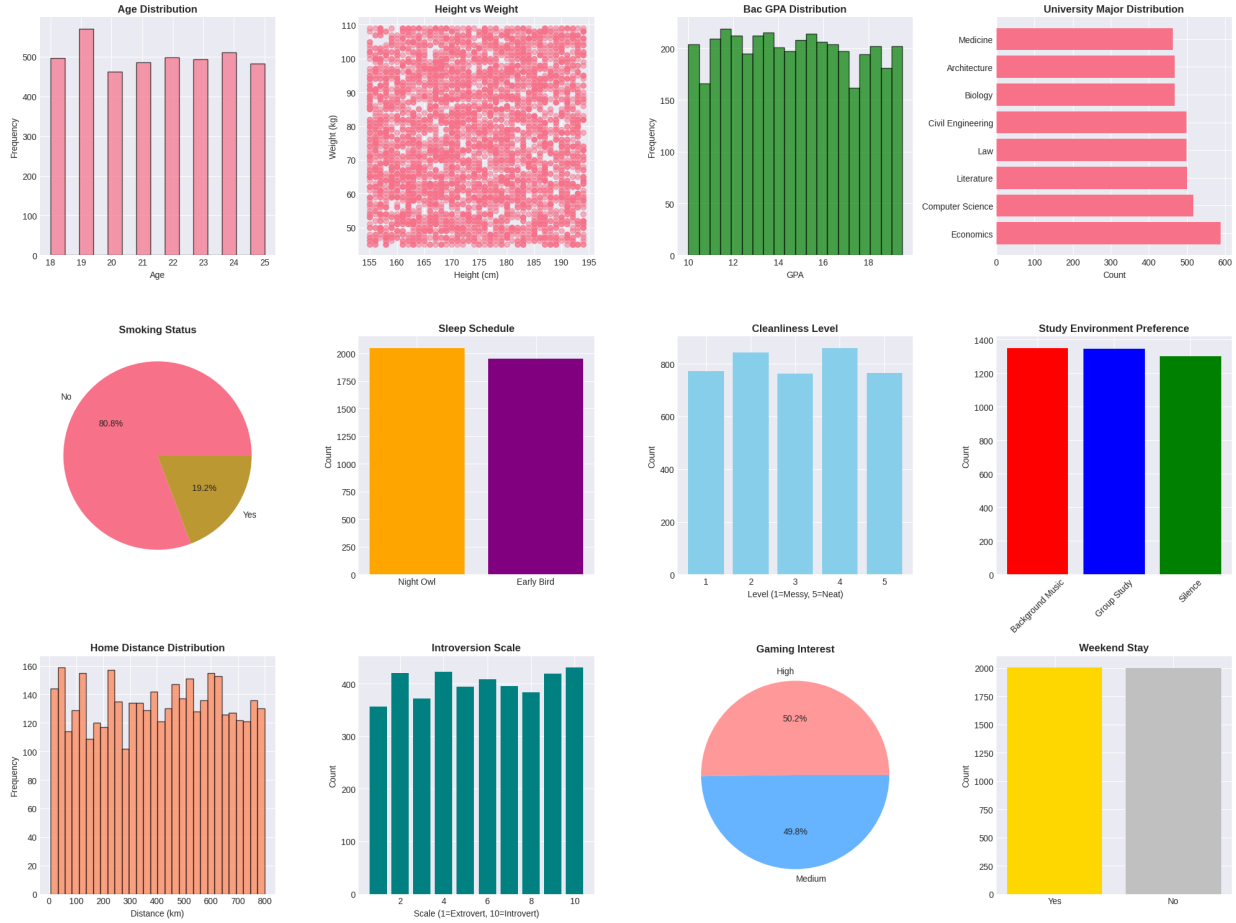

Fig. 6. General and detailed statistics for each column of the dataset

REFERENCES

[1] C. Virtuales, "The artificial intelligence in higher education: a paradigm shift?" *Campus Virtuales*, 2024. [Online]. Available: http://www.uajournals.com/campusvirtuales/journal/27/9.pdf

[2] IJARSCT, "Staymate: Smart roomate solutions," *International Journal of Advanced Research in Science, Communication and Technology*, 2024. [Online]. Available: https://ijarsct.co.in/Paper27028.pdf

[3] O. A. J. Index, "Artificial intelligence and the transformation of higher education in algeria," *OAJI*, 2024. [Online]. Available: https://oaji.net/pdf.html?n=2024/9211-1765094510.pdf

[4] MDPI, "Campus shuttle bus route optimization using machine learning," *Sustainability*, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/13/1/225

[5] M. D. P. I, "Reducing food waste in campus dining: A data-driven approach," *Sustainability*, 2024. [Online]. Available: https://www.mdpi.com/2071-1050/17/2/379

[6] IJSDR, "Ai-enhanced hostel management and booking system: An innovative approach," *International Journal of Scientific Development and Research (IJSDR)*, 2024. [Online]. Available: https://ijsdr.org/papers/IJSDR2504349.pdf