NOVA IMS Information Management School

Data-driven Pathways for Change **Student' Mental Health**

Group08:

| 20241169| Alexandre Vasconcelos

| 20240318| Bruna Fernandes

| 20240357 | Gabriel Calero

| 20240319 | Raquel Domingos

Data Mining I 2024/2025

Contents

Abstract

This project examines the increasing prevalence of mental health problems among university students, exacerbated by academic workload, financial concerns and social isolation. Data mining techniques were used to analyze a comprehensive dataset of 11,336 records and 29 variables to capture student demographics, academic performance, mental health and lifestyle factors. The main aim was to uncover patterns and clusters that could form the foundation for actionable strategies to improve support systems and students' wellbeing. Prior to analysis, extensive pre-processing was carried out to ensure the integrity of the data. This included the correction of missing values, outliers and variable transformations. Categorical data were coded, and numerical characteristics were standardized using Robust Scaler to enable strong clustering. Various clustering algorithms were applied, including K-means, DBSCAN, Gaussian Mixture Models (GMM), Hierarchical Clustering, among others, in four main perspectives: Mental Health, Academic, Social Lifestyle and Demographic. The results revealed distinct groups, such as 'struggling with Mental Health' with high levels of depression and anxiety versus 'resilient with moderate concerns' with emotional stability in the Mental Health Perspective. Similarly, the Academic Perspective distinguished between 'underperforming and dissatisfied', balanced and satisfied' and 'overworked and satisfied'. The Social Lifestyle Perspective reflected different patterns of involvement in religious, social and online activities and, lastly, the demographic perspective highlighted differences between 'Younger, Less Active, and Predominantly Data Science Students' and 'Older, More Active, and Primarily Engineering Students'. These findings were used to inform tailored strategies, including mental health counselling, peer mentoring, stress management workshops and social inclusion initiatives. This project highlights the value of data mining techniques, particularly clustering, in identifying and understanding patterns that enable more precise strategies to be defined to respond to students' needs. This provides a scalable framework for improving support services in educational institutions while promoting academic success and well-being.

I. Introduction - Overview and Main objectives

The increasing prevalence of mental health challenges among students increases the need for effective interventions. Students face various pressures, including academic workload, social isolation and financial concerns, all of which can negatively impact both their well-being and academic outcomes.

By leveraging data mining techniques, this project seeks to analyze patterns within the dataset, offering actionable insights to improve student support systems.

Our main objectives are to identify contributors to mental health challenges and academic performance and to provide interventions by analyzing the data.

II. Data Processing

Dataset Overview:

Initially, the dataset comprised 11336 records, each with 29 variables, capturing diverse aspects of student life (check Appendix). Key numerical variables include grades, depression levels, and academic workload, while categorical variables include gender and residential status. These variables encompass both objective measures and self-reported experiences, providing a well-rounded view of student circumstances. However, the dataset also presented challenges, such as missing values and outliers, requiring careful preprocessing to ensure reliability.

Key Insights from Descriptive Analysis:

1.Engagement in activities and sleep:

Creative activities, outdoor activities and social activities: The means are low (e.g. $\bar{x}$ (`creative_activities`) = 0.11, $\bar{x}$ (`outdoor_activities`) = 0.22). This could contribute to increased stress, isolation and a lower level of satisfaction.

Average sleep: A lot of the students (~56.7%) report sleeping between 4-6 hours, which is concerning and way below the recommended sleep duration.

2. Academic and Demographic Variables:

Students face significant academic pressures, with an average workload score of 71 (with the variable's available interval being [0-100]) and grades ranging from 50 to 95 (%).

While the higher academic workload correlates with increased pressure, students with better time management reported higher satisfaction levels.

The academic year's median is 2 and students are evenly distributed across academic years. The mean for the age is 19.95, with a range from 17 to 26 years, reflecting the expected age distribution for a university setting.

There is a gender disparity, with males constituting 80.64% of entries, which could overrepresent male perspectives in analyses.

3. Mental Health and Emotional Well-Being:

Depression and anxiety levels averaged around 55, and isolation strongly correlated with these variables, indicating high mental health challenges for the students. The `future_insecurity` variable has a mean of 50.29, suggesting that half of the students feel insecure about their future.

Cleaning the data

Checking for incoherencies:

For the first part of the data preprocessing, we decided to study data incoherencies, checking if all the variables present values corresponding to their meaning, for example ages between 17 and 100, grades between 0 and 100, and more inconsistent errors that could appear in all the variables.

Regarding the university and degree-level, only one value for these variables was found, so, they were not relevant to keep it here for analysis. The variable `consume_food` was also deleted as there are only '0' values, i.e. no student consumes food as a stress-resolving strategy.

`age` and `date_of_birth` were identified in the dataset, however, the variable `date_of_birth` showed inconsistent values and was incompatible with some of the ages in the database. For this reason, and to avoid redundancy between two variables that essentially transmit the same information, only the `age` variable, which is free of inconsistencies, was used.

Missing Values and Outliers:

The analysis revealed that 200 rows consisted entirely of missing values. For this reason, these rows were immediately excluded from the subsequent analysis.

For the remaining rows that still contained some missing values, data imputation was performed. In the first phase, the median was used for continuous variables to ensure robustness against skewed distributions, while the mode was applied to categorical variables such as `degree_level` and `degree_major`. After imputation, the descriptive statistics remained practically unchanged, preserving the original data's characteristics for subsequent analysis.

For the next phase, outliers' identification was conducted, revealing that the variables `age`, `academic_workload`, `sleep`, `outdoor_actvities`, `creative_activities`, `degree_major` and `study_satisfaction` displayed outliers. For most variables, a specific decision was made regarding their treatment:

- `age`: nowadays, an increasing number of people choose to pursue a degree course later in life, rather than at the typical age of 17/18. Since the only outliers for this variable are 25 and 26, they were not considered outliers;
- `degree_major`: many courses offer fewer openings than others. Therefore, these 369 values weren't also treated as outliers; instead, it is assumed that the university under had limited availability for these specific courses;
- `academic_workload`: the proportion of outliers in the dataset was found to be relatively low (0.58%), which provided statistical justification for the removal of these rows. Prior to the removal of these rows, the correlations between the variables were examined to ensure that no valuable patterns were lost;
- `sleep`, `outdoor_activities`, `creative_satisfaction`: since these variables are binary (Boolean), meaning they can only take 1 of 2 values (0 or 1), there's no inherent way for a value to deviate significantly from the "norm" as with other variables. For this reason, they weren't considered as outliers either.

However, during the outlier analysis phase, it was observed that `study_satisfaction` presented new outliers following the application of median imputation. These outliers were traced to the variable's skewed distribution, prompting a re-evaluation of the missing values handling approach. The imputation strategy was reviewed and KNN was applied instead. KNN produced similar results to the initial approach but offered slight improvements, effectively preserving the data's variability without creating additional outliers.

As a result, the final approach adopted KNN imputation for missing values, ensiling consistency, improved outcomes and the resolution of the outlier issue in `study_satisfaction`.

Variables Transformation:

To enable cluster analysis and integrate categorical variables into numerical models, several transformations were applied:

- `sports`: Transformed into binary values (1 for any non-zero engagement, 0 for zero engagement).
- `campus_discrimination`: Encoded as binary values (Yes = 1, No = 0).
- `gender`: Encoded as binary values (Female = 1, Male = 0).
- `residential_status`: Encoded as binary values (On-Campus = 1, Off-Campus = 0).
- `degree_major`: Transformed into numerical labels to enable evaluation with other metrics.
- `average_sleep`: Transformed into numerical (ordinal) labels to enable evaluation with other metrics (1: 'No Sports', 2: '1-3 times', 3: '4-6 times', 4: '7+ times').
- `sports_engagement`: Transformed into numerical (ordinal) labels to enable evaluation with other metrics (1: 'No Sports', 2: '1-3 times', 3: '4-6 times', 4: '7+ times').

Standardization

In the first phase, Z-score normalization ($\mu = 0$, $\sigma = 1$) was applied to scale numerical variables. This method was chosen for its ability to standardize data while preserving the original distributions of variables. While it does not directly address skewness, a quantitative analysis showed that, although some had extreme skewness (and for which alternative transformations (e.g., log scaling) were considered but deemed unnecessary at this stage), most variables had only slight asymmetry, making Z-score normalization an appropriate choice for a starting point.

The iterative nature of the process allowed us to move to the clustering phase using Z-score. However, the initial clustering results revealed suboptimal performance, which was traced back to the skewness in several variables that were not being adequately handled by the chosen normalization, affecting the quality of the clustering.

As a result, the final approach evolved to address skewness more effectively. Variables were ultimately normalized using Robust Scaler, which is designed to mitigate the impact of skewed distributions. This adjustment was made to ensure improved preprocessing alignment with the data's characteristics and enhance modeling performance.

Boolean variables were excluded from normalization as they are inherently binary (0 or 1), which already standardizes their scale, and categorical (nominal) variables such as `gender` and `degree_major`, since they represent distinct classes without numerical relationships, make normalization unsuitable.

III. Data Visualization & Analytics

Correlations:

Observing the heatmap on the Appendix we can detect a high correlation between `depression` and `anxiety`, `isolation` and `depression`, `age` and `academic_year` (strong positive correlations).

The high correlations observed in the data can reveal several key insights. For instance, the correlation between `isolation` and `depression` is 0.714 and between `depression` and `anxiety` is

0.841. This indicates that students who feel isolated are more likely to experience higher levels of depression and anxiety.

Additionally, the correlation between `age` and `academic_year` is 0.503, suggesting that older students are generally in higher academic years, as expected. This reflects the dataset's consistency in demographic structure.

All the other correlations are below 0.5.

IV. Clustering

The clustering algorithms were initiated by segmenting the primary data set into four segments deemed to be of the most significant relevance to this study.

The initial perspective is the **Mental Health Perspective**, comprising seven variables: depression, anxiety, social isolation, future insecurity, financial concerns, campus discrimination and social relationships. The subsequent perspective is the **Academic Perspective**, encompassing five variables: academic workload, academic pressure, study satisfaction during the academic year, and academic grade. The third perspective is the **Lifestyle Perspective**, which includes seven variables: sleep; outdoor activities; religious activities; sports; creative activities; social activities; online entertainment. These variables demonstrate the strategies employed by students to alleviate stress. The final perspective is the **Demographics Perspective**, which encompasses age; gender; degree major; residential status; sports engagement; and average sleep.

We tested the following clustering algorithms: Hierarchical Clustering, K-Means, K-Mode, Agglomerative clustering, Gaussian Mixture Models (GMM), DBSCAN, and then presented the best results.

Mental Health Perspectives

The algorithm that produced the best results was **K-Means clustering on PCA-reduced data** with **2 clusters**. The silhouette score was approximately **0.467 (see Appendix)**.

Cluster 0: 5702 students – "**Struggling with Mental Health"**: High levels of depression, anxiety and isolation: These students face significant mental health difficulties and perceive discrimination on campus, which contributes to increased stress and isolation. On the other hand, they have weakened social relationships, indicating greater disconnection from the academic environment. The result? A negative impact on their academic experience and well-being.

- **Possible strategies:** peer support groups and counseling services

Cluster 1: 5369 students – "**Resilient with Moderate Concerns"**: Low levels of depression, anxiety and isolation, which indicate emotional stability and strong social networks - positive relationships that act as a protective factor against mental health problems. On the other hand, they have a low perception of discrimination on campus - a more inclusive environment for these students – and fewer financial worries - clear indicators of economic stability and less external pressure. In short, students with social and emotional resilience are better able to face academic challenges and maintain stable general well-being.

- **Possible strategies:** encourage leadership/mentoring roles to support inclusivity

Academic Perspective

The algorithm that produced the best results was **GMM clustering on PCA-reduced** data with **3 clusters**. The silhouette score was approximately **0.366** (see Appendix). It's important to highlight

that, even though k=3 does not achieve the highest silhouette score, it offers the most balanced clustering solution since this configuration maintains a respectable score (just slightly lower than alternatives like k=2) while providing a more well-distributed grouping of clusters.

Cluster 0: 4269 students - **"Underperforming and Dissatisfied"**: This cluster represents students who are struggling academically. They exhibit low engagement and moderate academic pressure, which could contribute to poor academic outcomes.

- **Possible strategies:** academic mentoring and study workshops to improve performance.

Cluster 1: 3418 students - **"Balanced and Satisfied":** This cluster represents well-balanced achievers who are engaged and satisfied with their academic progress. Despite higher pressure, they seem to manage it effectively. This cluster represents well-balanced achievers who are engaged and satisfied with their academic progress. Despite higher pressure, they seem to manage it effectively.

- **Possible strategies:** provide leadership opportunities and wellness programs to maintain engagement.

Cluster 2: 3384 students - **"Overworked and Dissatisfied":** This cluster represents students who are highly engaged but under significant academic strain, leading to dissatisfaction. They likely face challenges balancing workload and satisfaction.

- **Possible strategies:** stress management workshops.

Social Lifestyle Perspective

The algorithm that produced the best results was **Agglomerative clustering** with **4 clusters**. The silhouette score was approximately **0.67 (see Appendix)**.

Cluster 0: 4723 students - **" Community-oriented with online entertainment enthusiasts":** It represents individuals who are actively engaged in both social and religious activities while also enjoying online entertainment. This suggests a balanced lifestyle combining offline interactions and digital entertainment.

Possible strategies: community events and volunteering opportunities.

Cluster 1: 2032 students - **" Socially active but avoiding digital entertainment":** Strong focus on social activities. No involvement in religious or online entertainment activities.

- **Possible strategies:** networking events and activities.

Cluster 2: 2027 students - **" Exclusively spiritually engaged":** High involvement in religious activities. Minimal to no participation in social or online entertainment activities.

- **Possible strategies:** interfaith events to diverse connections.

Cluster 3: 2289 students - **" Disengaged and isolated":** No participation in religious, social, or online entertainment activities. Likely introverted or isolated, prioritizing a lifestyle detached from observed activities.

- **Possible strategies:** interest-based workshops/events and one-on-one mentoring to reduce isolation

Demographic Perspective

The algorithm that produced the best results was **K-Prototypes** with **2 clusters**.

Cluster 0: 5059 students - "**Younger, Less Active, and Predominantly Data Science Students":** This cluster represents older, predominantly male students living off-campus, often in data

science programs. They have limited sports engagement and sleep patterns primarily in the 4–6 hour range.

- **Possible strategies:** on-campus fitness/sports events and promote sleep hygiene programs.

Cluster 1: 6012 students - **"Older, More Active, and Primarily Engineering Students":** This cluster represents younger, more gender-balanced students focused on engineering. They live off-campus but have slightly more on-campus representation than Cluster 0. They show slightly higher sports engagement and comparable sleep patterns.

- **Possible strategies:** foster teamwork and recreational activities to enhance campus engagement and well-being.

The clusters reveal a distinction between older, less-engaged off-campus students and younger, more engaged academic achievers. They also point to a shared challenge in sleep deficiencies across both clusters. Furthermore, opportunities to enhance engagement and well-being through tailored interventions are identified.

Combined Perspectives

After analyzing them individually, the 4 perspectives were combined to identify distinct student profiles using K-Modes Clustering Algorithm. The most balanced and insightful solution was obtained with k=5 clusters and the output obtained can be seen in the figures 10 and 11 present in the Appendix.

Among these clusters, **Cluster 0** and **Cluster 3** emerged as the most critical, as they represent students facing significant mental health challenges, including high levels of depression and isolation. The focus of the analysis is therefore directly on these 2 groups to understand their specific difficulties and propose tailored interventions.

V. Conclusion

The clustering analysis revealed critical insights into the diverse challenges faced by university students, underscoring the importance of targeted interventions to address their specific needs. Among the clusters identified, two groups stood out as particularly vulnerable: "The Overwhelmed Newcomers" (Cluster 0) and "The Isolated Upperclassmen" (Cluster 3). These clusters highlight how various factors, such as mental health challenges, academic pressures, social disengagement, and financial concerns, converge to create unique difficulties for different student populations. Understanding the characteristics and struggles of these groups is essential for developing effective strategies to enhance their well-being, academic performance, and overall university experience.

Cluster 0 – "The Overwhelmed Newcomers": Cluster 0 primarily includes first-year students aged 18 to 20, most likely enrolled in engineering and data science majors. This group faces some of the most significant challenges among all clusters, including the highest levels of depression and isolation. They are affected by deep insecurity about their future and high financial concerns, which compound their struggles. Academically, they perform poorly, with grades significantly lower than those of other clusters, probably due to the difficulty of adapting to university-level courses and high expectations. Despite facing high academic pressure, students in this cluster have very low interactions with others, rarely participating in social or religious activities or even online entertainment. Poor sleep habits – averaging only to 4 to 6 hours per night – further affects their well-being. Many do not engage in sports or physical activities, missing opportunities for stress relief and social connections. Additionally, living off-campus limits their ability to form strong support networks.

Tailored interventions for this cluster should focus on mental health support, academic assistance and fostering social integration through campus-based activities. For example, offering workshops or accessible counseling services on topics such as stress management and coping skills could provide these students with the tools they need to address their challenges. On the other hand, academic assistance could include mentorship programs pairing first-year students with upper-year pears who can provide guidance on managing workloads and university life.

Cluster 3 – "The Isolated Upperclassmen": Cluster 3 is composed mainly of third-year engineering students aged 20 to 22. Like Cluster 0, these students experience high levels of depression, isolation and future insecurity. Financial concerns are also significant within this group, further adding to their stress levels. However, these students tend to perform better academically than those in Cluster 0, maintaining moderate grades despite facing similarly high academic pressure. This resilience may reflect their greater experience in dealing with the academic system. Socially, students in Cluster 3 are as disengaged as those in Cluster 0, with low participation in social and religious activities or online entertainment. They also suffer from insufficient sleep, averaging just 4 to 6 hours per night. Their lack of engagement in sports or physical activities further contributes to their isolation.

Interventions for Cluster 3 should focus on strengthening social support systems, creating opportunities for peer bounding and addressing mental health needs like, for example, organizing community-based events such as group discussions on professional development. To help students manage academic pressure, resources such as time-management workshops and career mentoring programs could provide meaningful relief. On the other hand, encouraging healthier lifestyle habits could be included, like on-campus fitness programs, awareness campaigns on the importance of rest or even offering quiet study spaces with regulated hours to reduce late-night workloads.

The clustering analysis provides valuable insights, but the study has limitations. The dataset is comprehensive but may not fully capture diversity of experiences and challenges faced by students at different universities or cultural contexts. The dataset is dominated by certain demographic groups, such as male students, which could introduce bias and distort results.

Another limitation is the use of self-reported data, which is subjective and prone to inaccuracies. The imputation methods used for missing values may not fully replicate the true distribution of the data, potentially affecting the robustness of the clustering results.

The clustering algorithms used in this study have strengths and weaknesses. Further optimization of the algorithms or exploration of alternative methods, such as ensemble clustering, may yield improved results. Similarly, the choice of perspectives and variable transformations could benefit from additional domain-specific insight or expert consultation to further refine the analysis.

To enhance future studies, integrating longitudinal data could provide a dynamic view of student mental health and academic experience, allowing for more precise and proactive interventions. Widening the dataset to include different universities, cultural contexts and socio-economic backgrounds could improve the findings' generalizability. Also, using advanced techniques like deep learning or hybrid clustering could reveal more subtle patterns in the data, leading to more targeted support.

Appendix

Table 1 – Project Data

Variable	**Description and data spectrum**	Type

id	Index	Int64
gender	Gender of the student	Object
age	Age of the student	float64
date of birth	Date of birth of the student	Object
university	Name of the university	Object
degree_level	Under- or Postgraduate (you've conducted data from both)	Object
degree_major	Topic subject of the major	Object
academic_year	[1,2,3,4]	Float64
grade	Average grade ([50-100])	Float64
residential_status	Student lives Off- or On-Campus	Object
campus discrimination	Student experiences discrimination/bullying by other students ([0,1])	Object
sports_engagement	Times of sports a week	Object
average_sleep	Average sleeping hours a night	Object
study_satisfaction	felt level of the variable ([0-100]) by the student	Float64
academic_workload	felt level of the variable ([0-100]) by the student	Float64
academic pressure	felt level of the variable ([0-100]) by the student	Float64
finacial_concers	felt level of the variable ([0-100]) by the student	Float64
social_relationships	felt level of the variable ([0-100]) by the student	Float64
depression	felt level of the variable ([0-100]) by the student	Float64
anxiety	felt level of the variable ([0-100]) by the student	Float64
isolation	felt level of the variable ([0-100]) by the student	Float64
future_insecurity	felt level of the variable ([0-100]) by the student	Float64
sleep	the student uses the variable as a stress resolving strategy ([0,1])	Float64
outdoor_activities	the student uses the variable as a stress resolving strategy ([0,1])	Float64
religious_activities	the student uses the variable as a stress resolving strategy ([0,1])	Float64
sports	the student uses the variable as a stress resolving strategy ([0,1])	Float64
consume_food	the student uses the variable as a stress resolving strategy ([0,1])	Float64
creative_activities	the student uses the variable as a stress resolving strategy ([0,1])	Float64
social_activities	the student uses the variable as a stress resolving strategy ([0,1])	Float64
online_entertainment	the student uses the variable as a stress resolving strategy ([0,1])	Float64

Table 2 - Count of missing values per column

gender	**200**
age	200
date_of_birth	200

university	200
degree_level	456
degree_major	728
academic_year	200
grade	200
residential_status	200
campus_discrimination	200
sports_engagement	200
average_sleep	200
study_satisfaction	746
academic_workload	200
academic_pressure	200
financial_concerns	200
social_relationships	200
depression	777
anxiety	786
isolation	200
future_insecurity	200
sleep	200
outdoor_activities	200
religious_activities	200
sports	200
consume_food	200
creative_activities	200
social_activities	200
online_entertainment	200

Table 3 – Descriptive analysis

index	count	mean	std	min	25%	50%	75%	max
consume_food	11136.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
creative_activities	11136.0	0.1149	0.3189	0.0	0.0	0.0	0.0	1.0
outdoor_activities	11136.0	0.2183	0.4131	0.0	0.0	0.0	0.0	1.0
sleep	11136.0	0.2298	0.4207	0.0	0.0	0.0	0.0	1.0
social_activities	11136.0	0.3793	0.4852	0.0	0.0	0.0	1.0	1.0
online_entertainment	11136.0	0.4252	0.4944	0.0	0.0	0.0	1.0	1.0
religious_activities	11136.0	0.5287	0.4991	0.0	0.0	1.0	1.0	1.0
academic_year	11136.0	2.1609	1.0708	1.0	1.0	2.0	3.0	4.0
age	11136.0	19.9483	1.7794	17.0	19.0	20.0	21.0	26.0
sports	11136.0	3.0774	3.3354	0.0	0.0	1.0	6.0	10.0
grade	11136.0	77.2797	12.1528	50.0	70.0	79.0	87.0	95.0
academic_workload	11136.0	71.1770	21.5490	10.0	58.0	73.0	88.0	100.0
study_satisfaction	10590.0	72.0715	25.2698	0.0	56.0	77.0	95.0	100.0
academic_pressure	11136.0	68.6320	27.3586	0.0	53.0	74.0	90.0	100.0
social_relationships	11136.0	45.0637	28.7465	0.0	19.0	46.0	64.0	100.0
anxiety	10550.0	55.1837	31.3667	0.0	30.0	60.0	84.0	100.0
depression	10559.0	55.0773	32.7045	0.0	29.0	59.0	86.0	100.0
future_insecurity	11136.0	50.2953	33.1529	0.0	20.0	50.0	81.0	100.0
financial_concerns	11136.0	59.1953	33.4227	0.0	36.0	62.0	90.0	100.0
isolation	11136.0	55.6487	33.5970	0.0	28.0	60.0	87.0	100.0

Table 4 – Descriptive analysis original

Variable	count	mean	std	min	25%	50%	75%	max
age	11136.0	19.948366	1.779424	17.0	19.0	20.0	21.0	26.0
degree_level	10880.0	1.000000	0.000000	1.0	1.0	1.0	1.0	1.0
degree_major	10608.0	1.782334	0.766042	1.0	1.0	2.0	2.0	4.0
academic_year	11136.0	2.160920	1.070804	1.0	1.0	2.0	3.0	4.0
grade	11136.0	77.279723	12.152864	50.0	70.0	79.0	87.0	95.0
study_satisfaction	10590.0	72.071577	25.269816	0.0	56.0	77.0	95.0	100.0
academic_workload	11136.0	71.177083	21.549068	10.0	58.0	73.0	88.0	100.0
academic_pressure	11136.0	68.632094	27.358696	0.0	53.0	74.0	90.0	100.0
financial_concerns	11136.0	59.195312	33.422726	0.0	36.0	62.0	90.0	100.0
social_relationships	11136.0	45.063757	28.746581	0.0	19.0	46.0	64.0	100.0

depression	10559.0	55.077375	32.704507	0.0	29.0	59.0	86.0	100.0
anxiety	10550.0	55.183791	31.366781	0.0	30.0	60.0	84.0	100.0
isolation	11136.0	55.648707	33.597036	0.0	28.0	60.0	87.0	100.0
future_insecurity	11136.0	50.295348	33.152966	0.0	20.0	50.0	81.0	100.0
sleep	11136.0	0.229885	0.420778	0.0	0.0	0.0	0.0	1.0
outdoor_activities	11136.0	0.218391	0.413173	0.0	0.0	0.0	0.0	1.0
religious_activities	11136.0	0.528736	0.499196	0.0	0.0	1.0	1.0	1.0
sports	11136.0	3.077407	3.335448	0.0	0.0	1.0	6.0	10.0
consume_food	11136.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0
creative_activities	11136.0	0.114943	0.318967	0.0	0.0	0.0	0.0	1.0
social_activities	11136.0	0.379310	0.485237	0.0	0.0	0.0	1.0	1.0
online_entertainment	11136.0	0.425287	0.494409	0.0	0.0	0.0	1.0	1.0

Table 5 – Descriptive analysis after imputation

Variable	count	mean	std	min	25%	50%	75%	max
age	11136.0	19.948366	1.779424	17.0	19.0	20.0	21.00	26.0
degree_level	11136.0	1.000000	0.000000	1.0	1.0	1.0	1.00	1.0
degree_major	11136.0	1.792654	0.749090	1.0	1.0	2.0	2.00	4.0
academic_year	11136.0	2.160920	1.070804	1.0	1.0	2.0	3.00	4.0
grade	11136.0	77.279723	12.152864	50.0	70.0	79.0	87.00	95.0
study_satisfaction	11136.0	72.313218	24.665452	0.0	57.0	77.0	93.00	100.0
academic_workload	11136.0	71.177083	21.549068	10.0	58.0	73.0	88.00	100.0
academic_pressure	11136.0	68.632094	27.358696	0.0	53.0	74.0	90.00	100.0
financial_concerns	11136.0	59.195312	33.422726	0.0	36.0	62.0	90.00	100.0
social_relationships	11136.0	45.063757	28.746581	0.0	19.0	46.0	64.00	100.0
depression	11136.0	55.280621	31.857753	0.0	32.0	59.0	85.00	100.0
anxiety	11136.0	55.437231	30.549193	0.0	32.0	60.0	82.25	100.0
isolation	11136.0	55.648707	33.597036	0.0	28.0	60.0	87.00	100.0
future_insecurity	11136.0	50.295348	33.152966	0.0	20.0	50.0	81.00	100.0
sleep	11136.0	0.229885	0.420778	0.0	0.0	0.0	0.00	1.0
outdoor_activities	11136.0	0.218391	0.413173	0.0	0.0	0.0	0.00	1.0
religious_activities	11136.0	0.528736	0.499196	0.0	0.0	1.0	1.00	1.0
sports	11136.0	3.077407	3.335448	0.0	0.0	1.0	6.00	10.0
creative_activities	11136.0	0.114943	0.318967	0.0	0.0	0.0	0.00	1.0
social_activities	11136.0	0.379310	0.485237	0.0	0.0	0.0	1.00	1.0

online_entertainment	11136.0	0.425287	0.494409	0.0	0.0	0.0	1.00	1.0

Figure 1 – Correlation Matrix

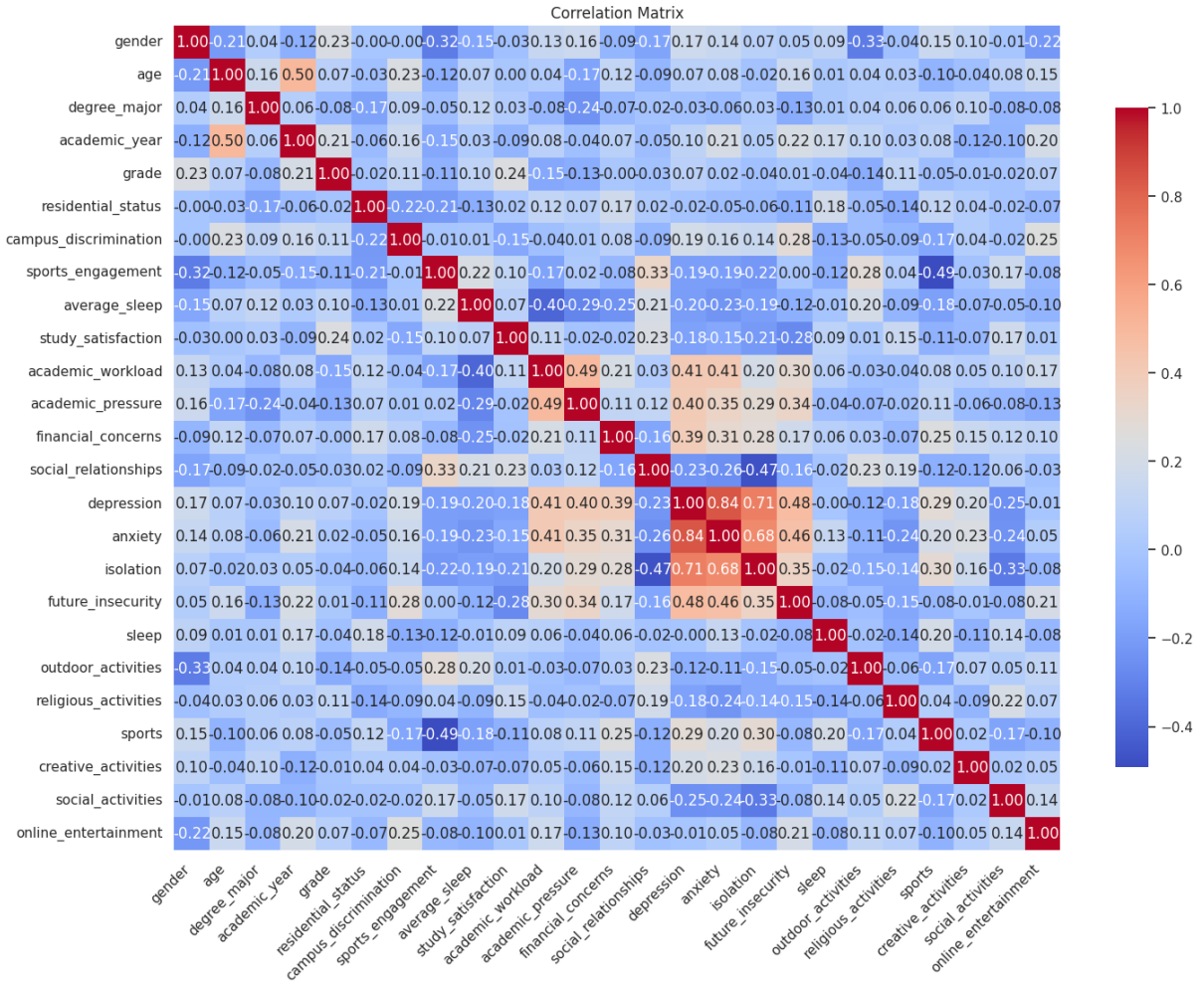

Figure 2 – KMeans Clustering for Mental Health Perspective

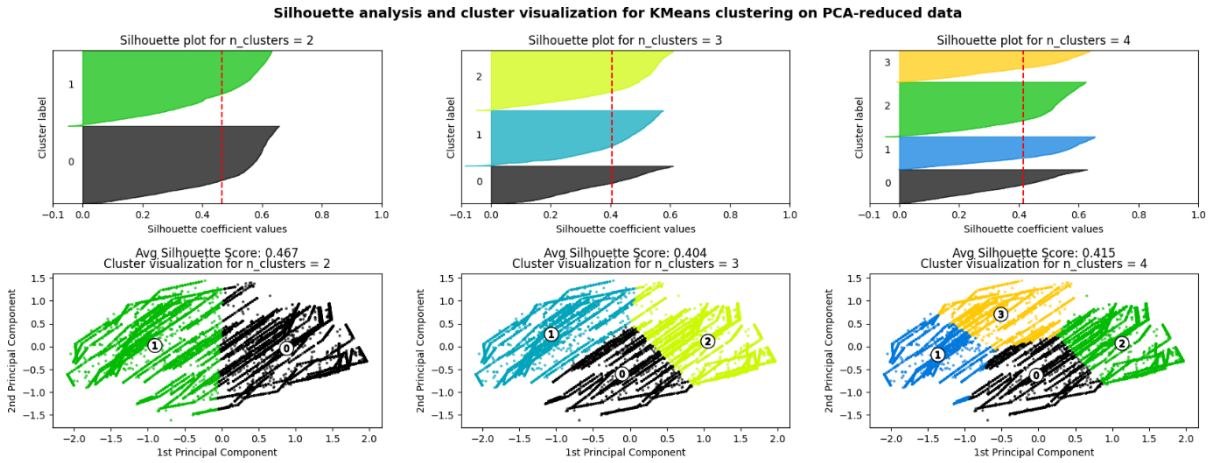

Figure 3 – Variables from Mental Health Perspective by Each Cluster

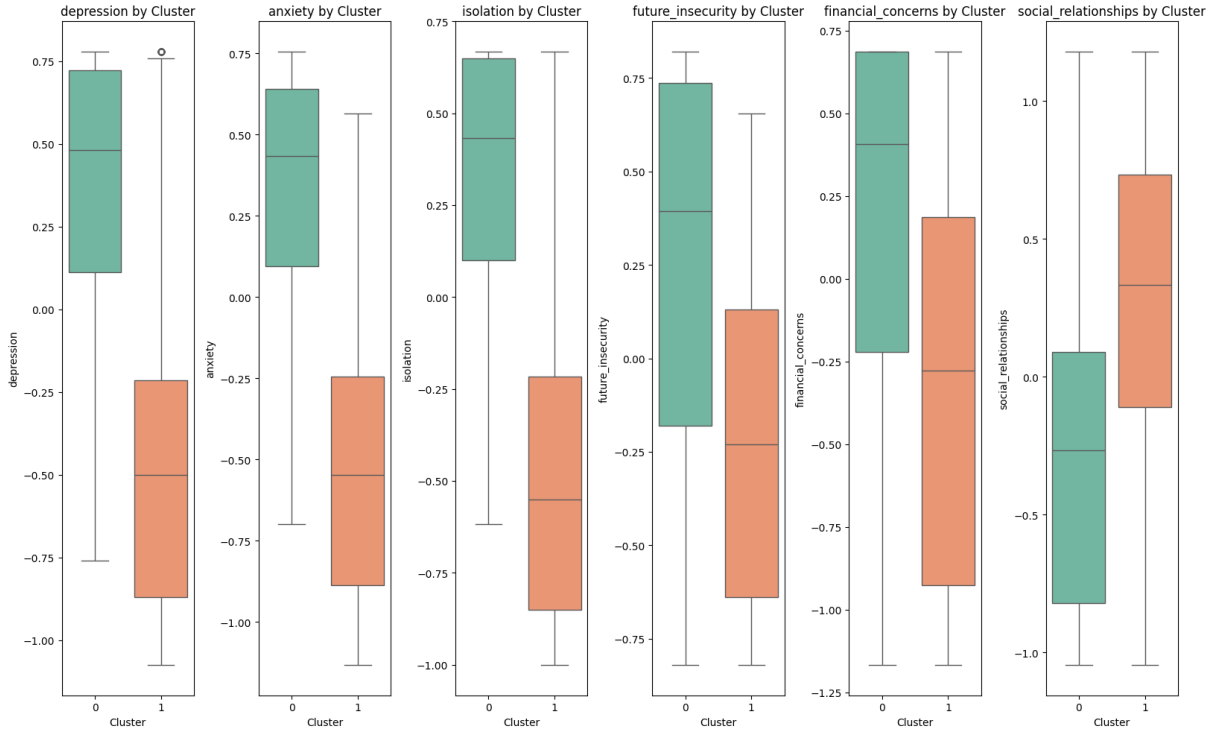

Figure 4 – GMM Clustering for Academic Perspective

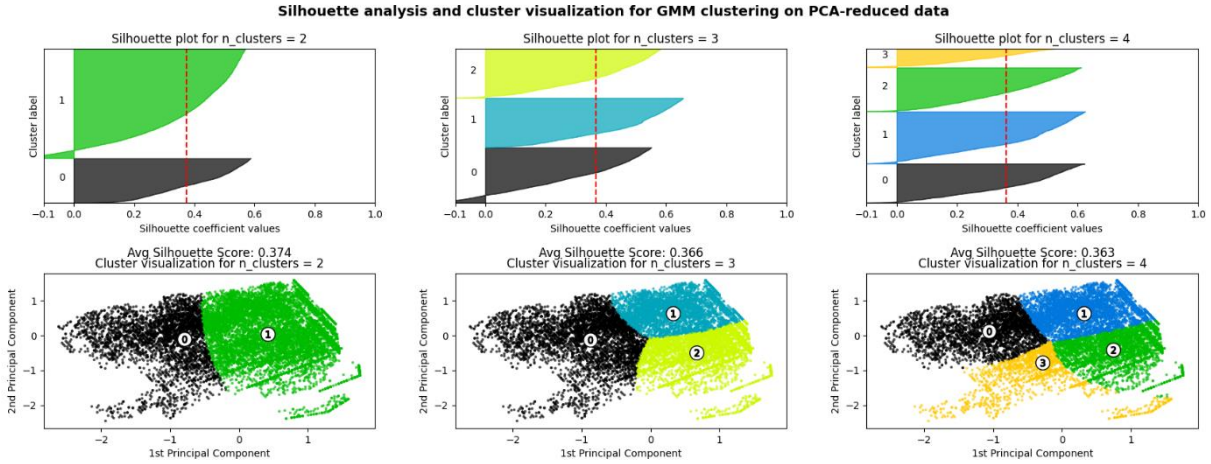

Figure 5 – Variables from Academic Perspective by Each Cluster

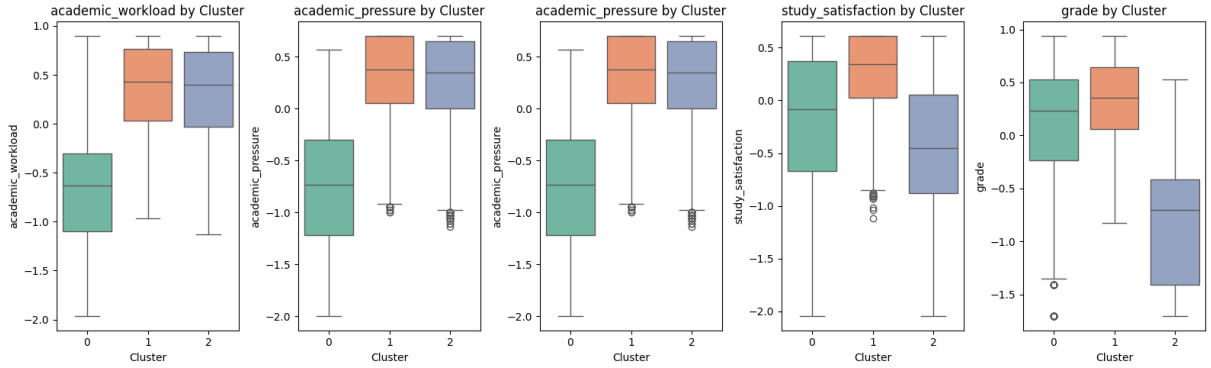

Figure 6 – Agglomerative Clustering for Lifestyle Perspective

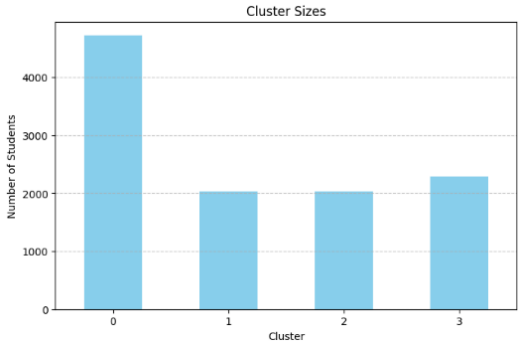

Figure 7 – Variables from Lifestyle Perspective by Each Cluster

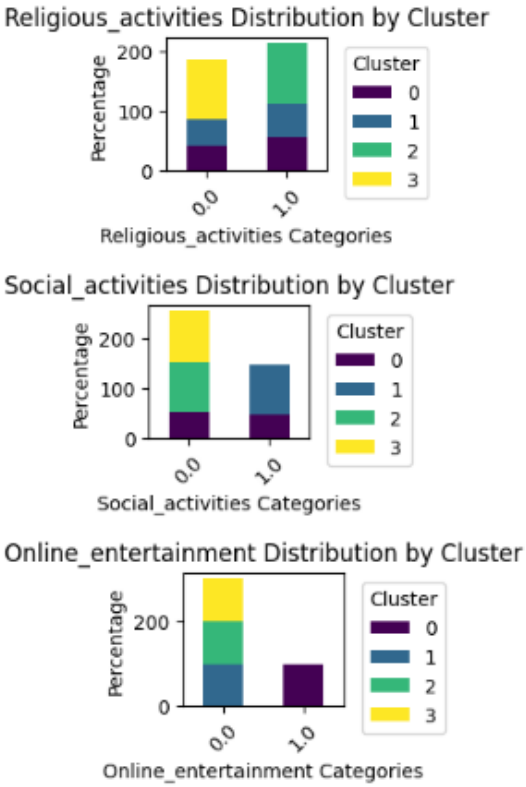

Figure 8 – K-Protopytpes Clustering for Demographics Perspective

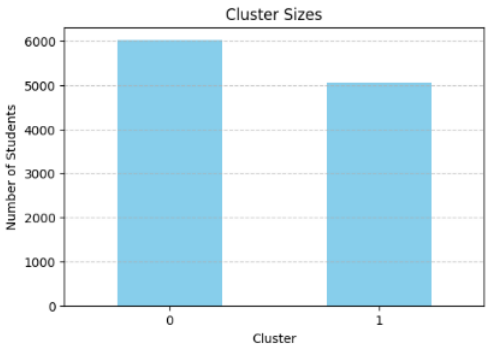

Figure 9 – Variables from Demographics Perspective by Each Cluster

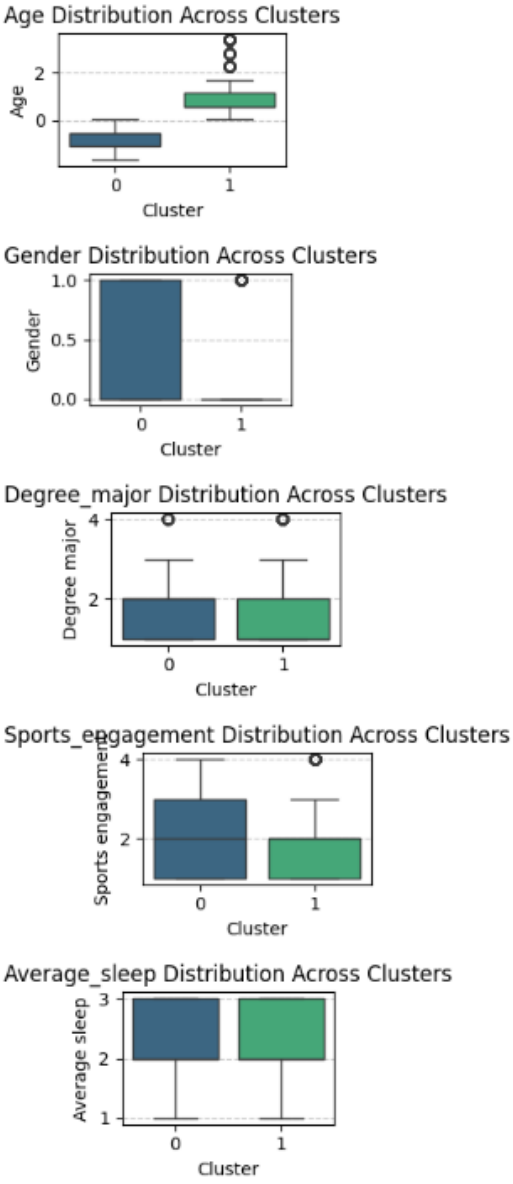

Figure 10 – Mental Health Perspective by Each Cluster

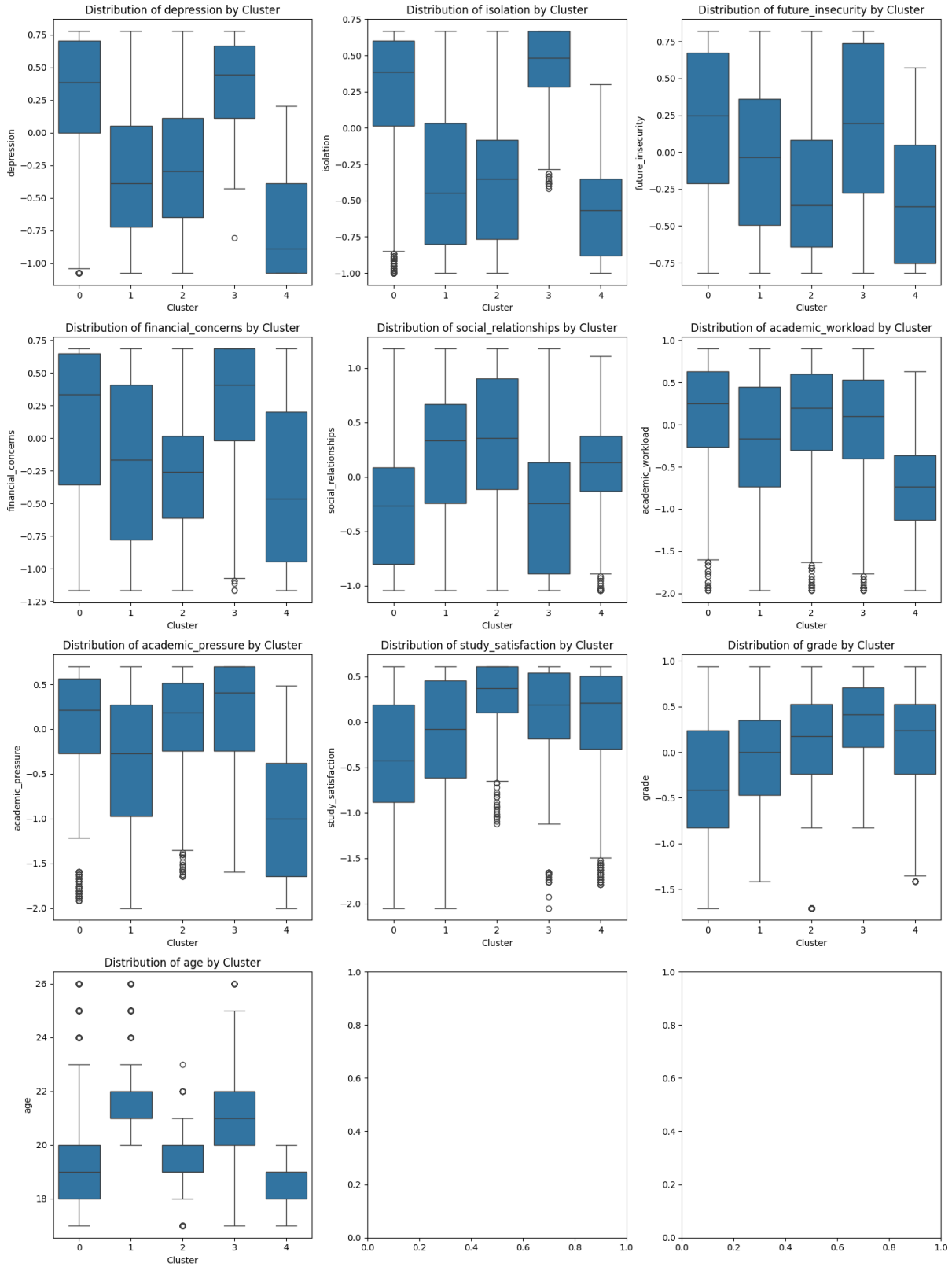

Figure 11 – Lifestyle Perspective by Each Cluster