Youth Mental Health Clustering

Emily Han (Lead Author)
Yawen Dong, Erika Garza-Elorduy, Junyi Hui (Co-Authors)

June 13th, 2025

1 Introduction

As a transitional stage for human emotional and psychological development, adolescence is shaped by increasingly complex social and digital environments. In recent years, there has been a sharp rise in mental health challenges among highschool adolescents, including persistent sadness, anxiety, and suicidal ideation. Although factors including social pressures, family dynamics, and living conditions have been highlighted as central causes of adolescent distress, public disclosure often ignores the complex interplay of these factors.

Adolescents with similar levels of distress may experience vastly different risk environments. Clustering methods offer a way to uncover these hidden contextual differences by identifying subgroups with shared patterns of adversity and resilience. Domain-specific clustering isolates the influence of particular contexts (e.g., home, social, physical health), while composite clustering captures how multiple risk factors co-occur and interact.

Using 2023 Youth Risk Behavior Surveillance System (YRBSS) data, this study applied KMeans clustering to identify latent profiles of adolescents based on behavioral and contextual characteristics. The goal was to uncover distinct psychosocial subtypes that reflect varying risk patterns and inform more targeted, context-aware strategies for adolescent mental health support and offer potential implications for educators, policymakers, and clinicians seeking to design targeted prevention strategies that account for the complex and interconnected nature of adolescent experience.

It was hypothesized that:

- $\mathbf{H_1}$: Distinct subgroups of adolescents will emerge within each contextual domain (home, social, and physical health) when clustering is performed using domain-specific variables alongside mental health indicators.
- $\mathbf{H_2}$: In each domain, adolescents in clusters characterized by higher adversity (e.g., unstable home, low social support, or poor physical health) will report higher levels of psychological distress compared to their lower-risk counterparts.

- $\mathbf{H_3}$: Composite clustering using multidomain variables will yield more nuanced psychosocial profiles that differentiate adolescents not only by behavioral risk but also by severity and type of mental health burden.

2 Methods

2.1 Objective

This study applied clustering to responses from the 2023 Youth Risk Behavior Surveillance System (YRBSS) to examine adolescent mental health in relation to contextual risk factors. Clustering was performed within three distinct domains—home environment, social environment, and physical health—using domain-specific variables alongside shared mental health indicators to uncover subgroups reflecting contextually grounded distress patterns.

In addition, composite clustering was conducted on a broader set of variables spanning multiple domains (e.g., substance use, violence, resource access) to reveal more complex and multidimensional psychosocial profiles. This approach identified subgroups of adolescents who share multiple risk factors across different areas of life, helping to inform more targeted prevention efforts and guide the allocation of mental health and social support resources.

2.2 Data Preparation

The 2023 national YRBSS dataset (N = 20,103; 250 items) includes self-reported responses on mental health, behaviors, and contextual risk factors. For domain-specific clustering, subsets of variables were selected for the home, social, and physical health domains, each anchored by mental health indicators. For composite clustering, a broader selection of items spanning multiple domains—including violence, substance use, and resource access—was used to capture cross-cutting patterns. All variables were recoded to ensure higher values reflect greater adversity, and only complete cases were retained for analysis.

2.3 Analysis

Two complementary clustering strategies were used: domain-specific and composite. Domain-specific clustering was conducted separately for the home, social, and physical health domains. Each dataset included contextual variables from a single domain and a common set of mental health indicators, which were z-score standardized. KMeans clustering was applied, and the optimal number of clusters was determined using silhouette scores. A two-cluster solution (high vs. low risk) was selected for all domains, with cluster labels aligned for interpretability across domain (Cluster 0 = higher mental distress, Cluster 1 = lower mental distress).

Composite clustering used a broader selection of standardized variables across domains, including behavioral risks, contextual adversity, and mental health symptoms. As with the domain-specific models, all variables were standardized and preprocessed to ensure directional consistency—higher scores consistently reflected greater adversity or risk. KMeans clustering to assign participants to latent subgroups based on similarity in multidimensional risk profiles. Four-cluster solution (k = 4) was selected based on a combination of the elbow method and silhouette score analysis. This configuration provided a balance between granularity and interpretability, allowing for the identification of multiple multidimensional profiles while avoiding fragmentation or overfitting.

Principal Component Analysis (PCA) was used for visualizing cluster separation in reduced-dimensional space but did not inform the clustering directly. Descriptive statistics were computed to characterize each cluster's psychosocial profile.

3 Result

3.1 Domain-Specific Clustering Reveals a Consistent Risk–Distress Pattern

Across all three domains (home environment, social context, and physical health), a consistent pattern emerged. Adolescents exposed to greater contextual adversity were more likely to exhibit higher levels of mental distress. In each domain-specific clustering model, a two-cluster solution

was identified as optimal based on silhouette analysis, and the resulting clusters clearly differentiated between high-risk (Cluster 0) and low-risk profiles (Cluster 1).

Table 1: Descriptive statistics of clusters across domains

Domain	Cluster 0 Size (%)	Cluster 1 Size (%)	Avg Dist (C0)	Avg Dist (C1)	Silhouette Score
Home	28.5%	71.5%	6.12	3.80	0.311
Social	21.5%	78.5%	8.16	4.16	0.360
Health	22.1%	77.9%	4.89	3.57	0.329

As shown in Table 1, Cluster 0 in each domain represented the higher-risk group, while Cluster 1 reflected the lower-risk profile. Notably, Cluster 1 was substantially larger across all domains, encompassing approximately 72–79% of participants, indicating that most adolescents in the sample experienced relatively protective environments and lower psychological burden. In contrast, Cluster 0 contained a smaller subset with elevated adversity and mental distress. Average distances to the cluster centroids were consistently higher for Cluster 0, suggesting greater heterogeneity within the high-risk groups. Silhouette scores across domains ranged from 0.311 to 0.360, indicating moderate cluster separation. These values suggest that the two-cluster solutions captured meaningful distinctions, with some overlap between groups as expected in psychosocial data. The social domain showed the strongest cohesion, reflecting clearer differentiation in social support profiles. This underscore the importance of peer and social dynamics in differentiating adolescent mental health risk.

3.1.1 PCA-Based Visualization & Cross-Domain Overlaps

To better understand how different domains structure adolescent mental health risk, PCA visualizations were used to illustrate the separation between clusters. In all three contexts—home environment, social environment, and physical health—two distinct clusters consistently emerged as expected, separating adolescents into high- and low-risk profiles.

As shown in Figure 1, adolescents in Cluster 1, characterized by nurturing and stable home environments, formed a tightly concentrated grouping. In contrast, Cluster 0—representing adolescents

with adverse home conditions—was more dispersed, suggesting greater variability in household stressors and associated mental health burden.

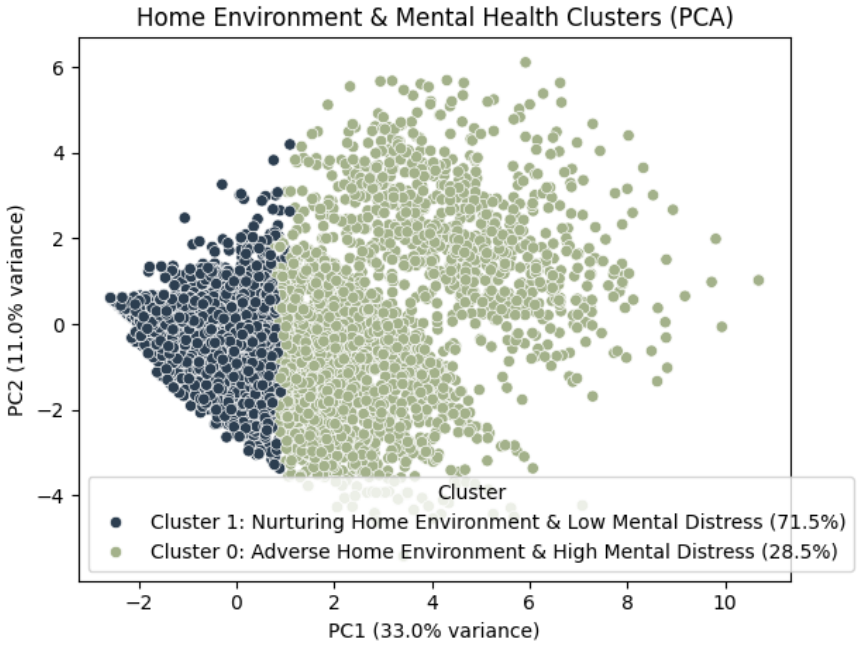

Figure 1: PCA visualization of clusters in the home environment domain

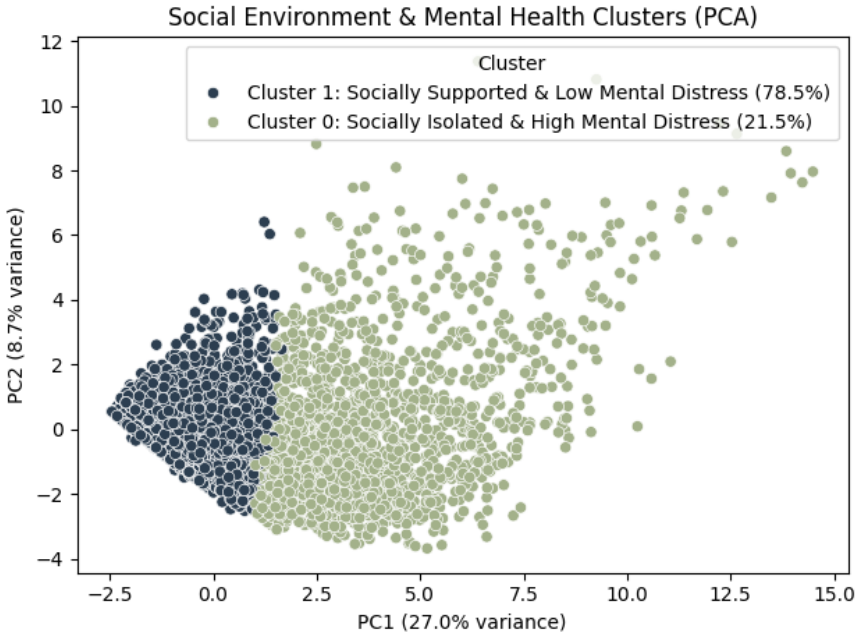

Figure 2: PCA visualization of clusters in the social environment domain

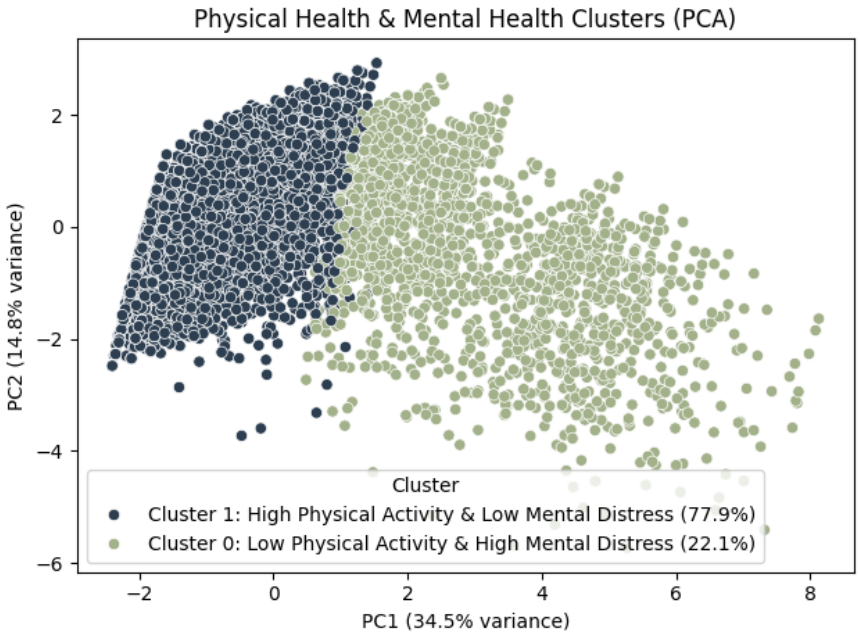

Figure 3: PCA visualization of clusters in the physical health domain

Figure 2 presents the clustering results for the social environment domain. As with the home context, Cluster 1 (characterized by strong social support) forms a relatively tight grouping. Interestingly, although Cluster 0 (social isolation) is still more dispersed than Cluster 1—as seen across all domains—its members are more tightly grouped compared to the Cluster 0 patterns in the other two domains. This is consistent with the social domain's higher silhouette score of 0.36 (see Table 1), suggesting that adolescents' social environment may create a clearer division in psychological risk profiles.

Figure 3 shows the PCA projection for the physical health domain. Compared to the home and social contexts, the separation between clusters is less visually distinct. Both Cluster 0 (low physical activity and high mental distress) and Cluster 1 (high physical activity and low distress) appear more dispersed. This more uniform spread is consistent with the average within-cluster distances shown in Table 1, where the physical health clusters had more comparable cohesion levels (Avg Dist: 4.89 for Cluster 0 and 3.57 for Cluster 1).

Taken together, these visualizations highlight a consistent pattern across all three domains: adolescents with greater contextual adversity tend to cluster separately from their lower-risk peers,

with these divisions aligning closely with levels of mental distress. However, the sharpness of this separation varies. The social environment domain produced the most well-defined clusters, both visually and quantitatively, suggesting that peer relationships and social support may influence adolescent mental health in more distinct ways in comparison. Notably, across all three domains, Cluster 1—representing adolescents in more supportive or protective environments—was consistently more tightly grouped, while Cluster 0 was more diffuse. This pattern reinforces the interpretation that adversity manifests in more heterogeneous ways, whereas protective contexts are associated with more consistent psychosocial outcomes.

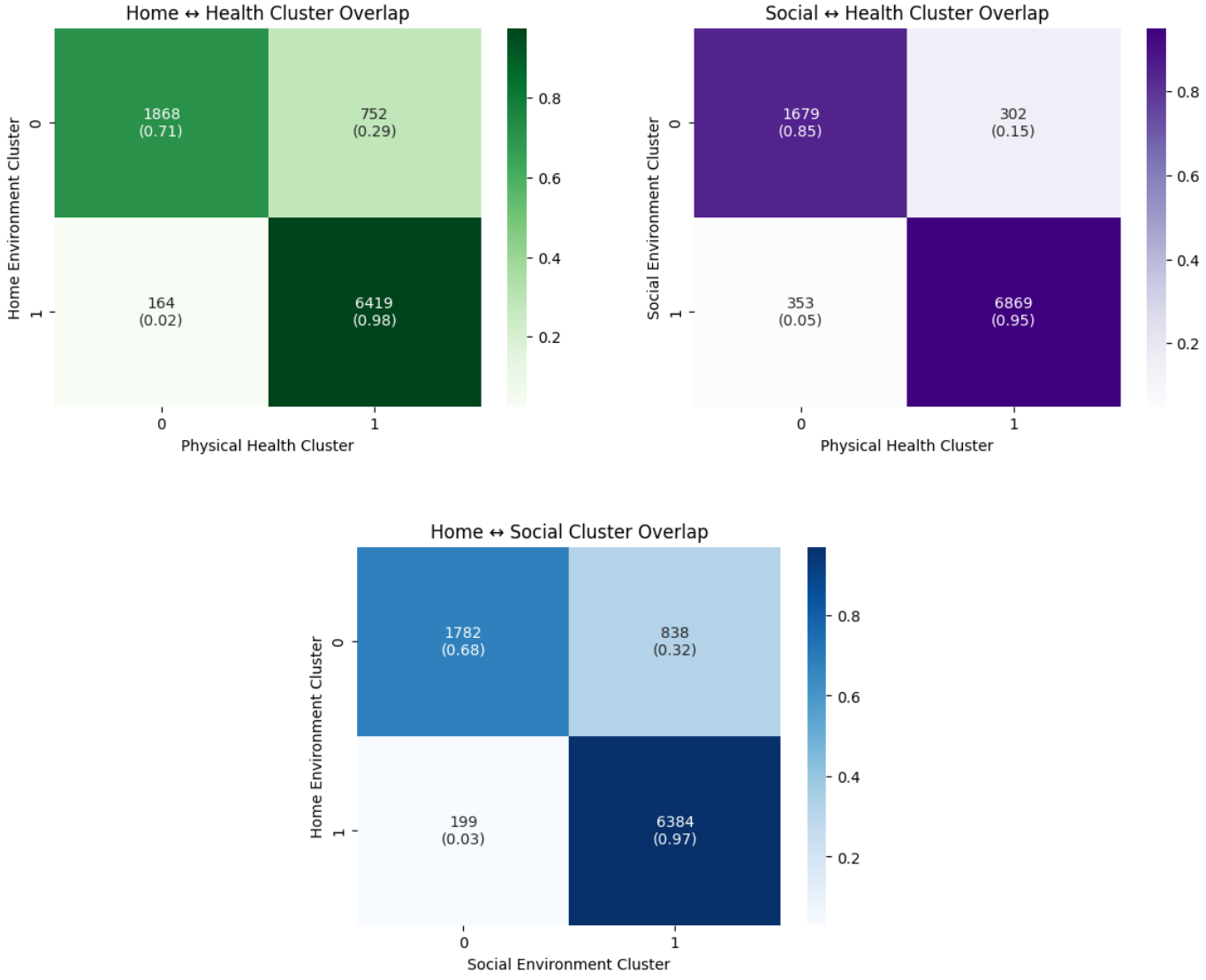

Figure 4: Cross-Domain Cluster Overlap Heatmaps

Figure 4 presents heatmaps depicting the overlap in cluster membership across home, social, and physical health domains. These cross-tabulations consistently reveals that adolescents identified as low-risk in one domain are very likely to be classified as low-risk in the others. For example, 97% of adolescents in the nurturing home cluster (Cluster 1) also belong to the socially supported cluster, and 98% align with the physically active group. Similarly, overlap between socially isolated and low-activity clusters (Cluster 0) is also pronounced. These findings complement the PCA visualizations, reinforcing that the same individuals tend to cluster together across different domains. This consistent alignment highlights the cumulative and interconnected structure of adolescent mental health risk—suggesting that protective or adverse environments rarely occur in isolation.

3.2 Composite Clustering — Start Here Eri <3

In the final phase of analysis, we constructed a full dataset by integrating all domain-specific variables into a single composite dataset. After merging, we applied z-score normalization to standardize all variables, ensuring that each feature contributed equally to the clustering process. We then performed dimensionality reduction using PCA, retaining the first two principal components to summarize the variance structure of the high-dimensional data. KMeans clustering was subsequently applied to the standardized dataset, with the number of clusters informed by earlier silhouette analyses conducted within individual domains.

Elbow Method We applied the elbow method to the full dataset in order to determine the optimal number of clusters (k) for KMeans clustering by plotting the model's inertia against various values of k. A "elbow" appears around k = 4 or 5, where the rate of decrease in inertia begins to level off.

Silhouette Score The highest score occurs at k = 2, suggesting the clearest separation between clusters at this point. However, as the elbow plot also indicated diminishing returns after k = 4 or 5, we selected k = 4 to allow for more nuanced differentiation among adolescent profiles, balancing cluster cohesion with interpretability.

3.3 PCA Projection

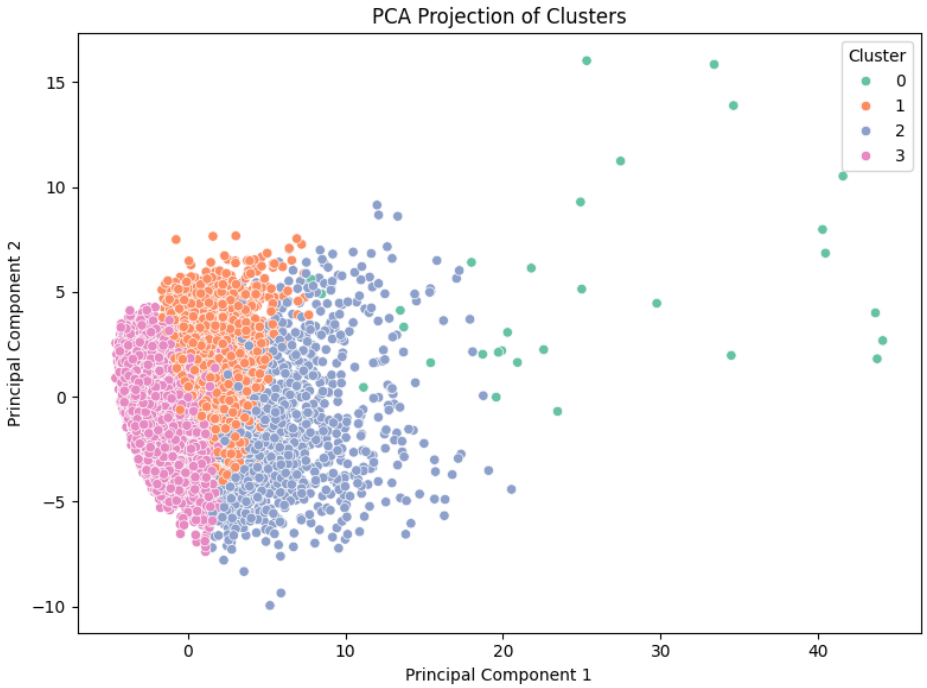

Figure 5: PCA visualization of clusters in the composite model

Figure 4: PCA Projection of KMeans Clusters (k = 4, Full Dataset) After choosing the optimal number of clusters, we visualize the KMeans clustering results (with k = 4) projected onto the first two principal components of the full dataset. Each point represents an individual, and colors indicate their assigned cluster. The plot shows partial separation among the four clusters, with Clusters 1, 2, and 3 forming relatively dense and distinct groupings, while Cluster 0 appears more dispersed and includes several outliers. This pattern suggests that most adolescents fall into well-defined psychosocial profiles, while a smaller subset (Cluster 0) may represent more atypical or mixed cases that do not cluster tightly with others.

3.4 Cluster Evaluation

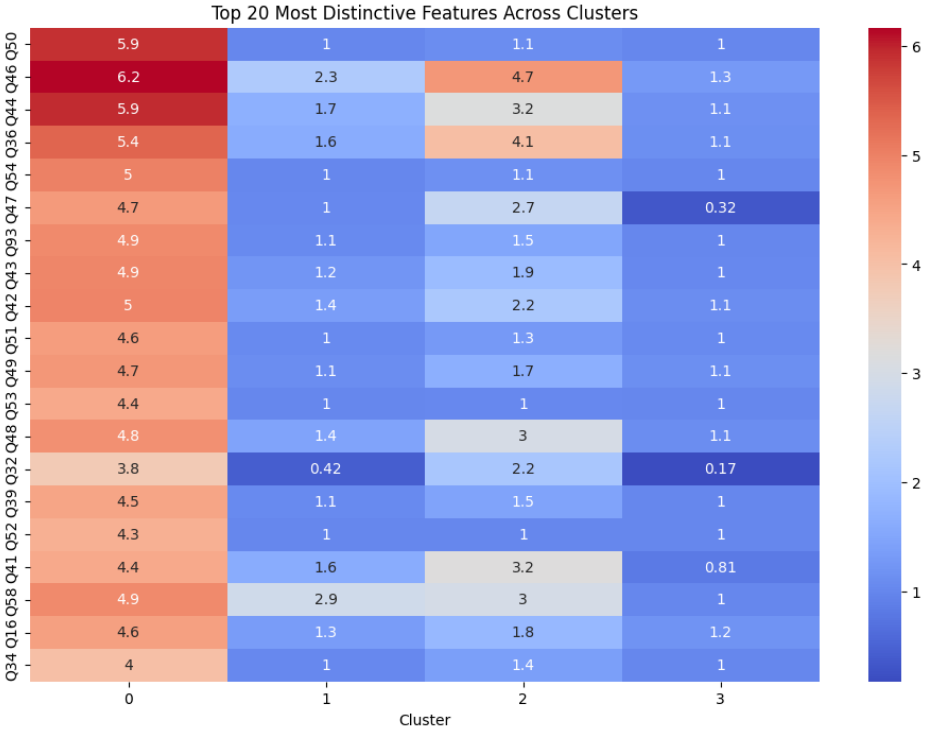

Figure 6: Top 20 features by average cluster importance in the composite model

Figure 5: Top 20 most distinctive feature across clusters (k = 4, Full Dataset)

Figure 5 displays the average values of the 20 most distinguishing variables across the four clusters derived from the full dataset. Each row represents a survey question (e.g., Q34, Q16), and each column corresponds to a cluster. The majority of questions centers on substance (tobacco products, alcohol, marijuana, and illicit drug) use, while early risk behavior initiation (Q32, Q41, Q47), physical aggression (Q16), and sexual behavior (Q58) are also covered. Color intensity reflects the average score for that variable within each cluster, with red indicating higher values and blue indicating lower values. From the heatmap, Cluster 0 showed the highest average scores across nearly all substance use variables. Members of this cluster also initiated risk behaviors early, as reflected in low age-of-onset values for cigarette (Q32) and alcohol use (Q41), and reported high levels of binge drinking (Q42-Q44), vaping (Q36), multiple sexual partners (Q58), and physical altercations (Q16). This group represents a concentrated "supercluster" of early and sustained risk. Cluster 1 displayed consistently low scores across most risk behaviors but showed relatively higher

engagement in more socially normative activities, such as marijuana use (Q46) and alcohol consumption (Q41, Q44), typically with delayed onset and minimal involvement in harder substances. Additionally, this group reported comparatively higher numbers of sexual partners (Q58) and a slight increase in e-cigarette use (Q36). Cluster 2 had relatively low overall scores across the risk behaviors, yet showed moderate involvement in certain socially accepted or accessible activities. Specifically, individuals in this cluster reported relatively high scores for recent and lifetime marijuana use (Q46, Q48), earlier experimentation with alcohol and marijuana (Q41, Q47), and occasional binge drinking (Q44). They also displayed increased levels of vaping behavior (Q36) and number of reported sexual partners (Q58), suggesting selective risk-taking in areas that may reflect social experimentation rather than pervasive high-risk patterns. Lastly, Cluster 3 was characterized by selective use of substances like marijuana (Q47) and cigarettes (Q32) but not broad lifetime experimentation with harder drugs. While their usage patterns were more concentrated, this cluster may indicate underlying emotional or mental health vulnerabilities not directly captured in the top 20 features but implied by their coping-oriented consumption patterns.

3.5 Mental Health Risk

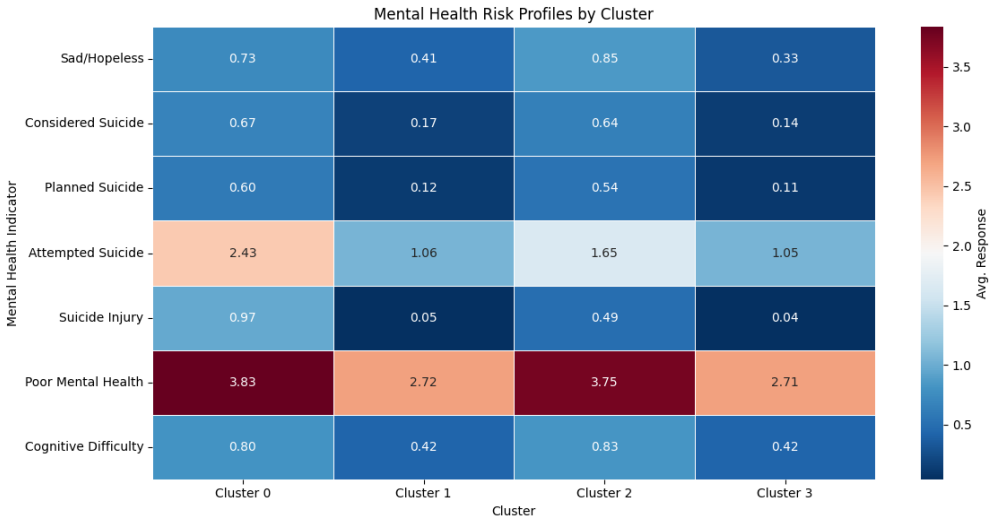

Figure 7: Cluster-wise summary of composite risk profiles

Figure 6: Mental Health Risk Profiles by Cluster (Average Response Levels) To explore more on mental health risk, we summarize the average responses for seven mental health indicators

(Q26-Q30, Q84, Q106) across the four identified clusters. Cluster 0 displays the highest levels of mental health risk, with elevated averages across all indicators. Most notably, this group reports the highest frequency of suicide attempts (2.43) and poorest overall mental health (3.83). They also score relatively high on suicide injury (0.97) and cognitive difficulty (0.80), marking them as the most vulnerable group. Cluster 2 also shows high psychological distress, with particularly high levels of poor mental health (3.75) and moderate-to-high values on suicide ideation, planning, and attempts. This group appears emotionally burdened despite lower involvement in broad-spectrum risk behaviors (as shown earlier), suggesting internalized risk or unmet mental health needs. Cluster 1 and Cluster 3 show lower average scores across most indicators. Cluster 3, in particular, has the lowest levels of suicide-related thoughts and behaviors, including the lowest scores on suicide consideration (0.14), planning (0.11), and injury (0.04). Similarly, Cluster 1 reports low values for suicidal ideation and relatively moderate cognitive difficulty, though its poor mental health score (2.72) suggests some emotional pressure.

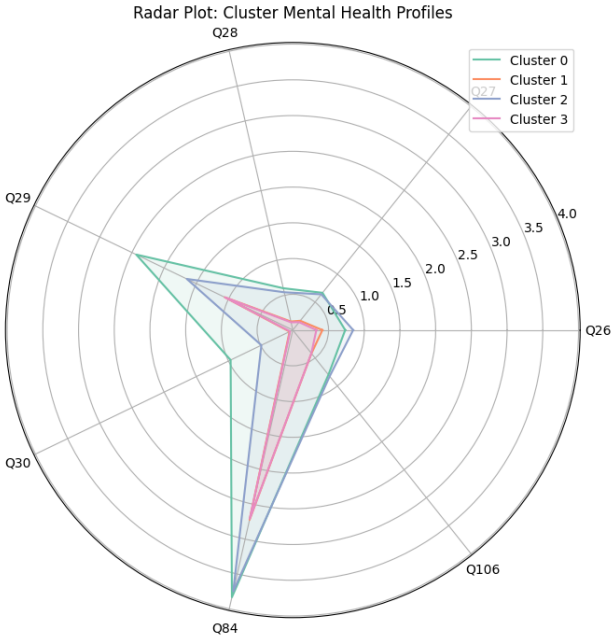

Figure 8: Radar plot comparing average domain scores across composite clusters

Figure 7: Radar Plot of Cluster Mental Health Profiles (For Selected Indicators)

The radar plot which visualizes average responses for selected mental health indicators across the four identified clusters shows similar results. Cluster 0 stands out with the highest overall values

on most items, particularly Q84 (poor mental health in the past 30 days) and Q29 (suicide attempt), highlighting its status as the highest-risk group. Cluster 2 also shows high levels on Q84, Q26 (emotional distress), and Q27 (suicidal ideation), suggesting internalized distress despite fewer behavioral risk factors. Clusters 1 and 3 display lower scores across nearly all indicators, indicating better mental health profiles, with Cluster 3 being the least affected.

3.6 Cluster Summary

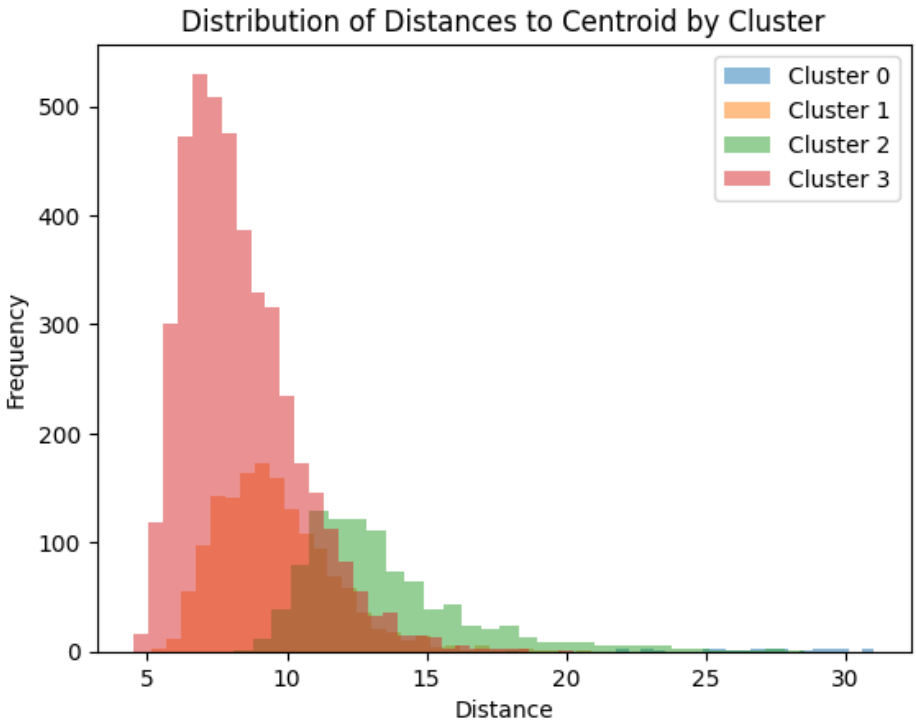

Figure 9: Distribution of mental health indicators across composite clusters

Figure 8: Distribution of Distances to Cluster Centroids Given the plot, Cluster 3 (red) has the shortest and narrowest distance distribution, suggesting a high degree of internal cohesion among its members. In contrast, Cluster 2 (green) exhibits a wider and more skewed distribution, indicating greater heterogeneity and weaker cohesion. Clusters 0 and 1 fall between these extremes, with Cluster 0 showing the broadest tail, possibly due to outliers or more complex internal structure.

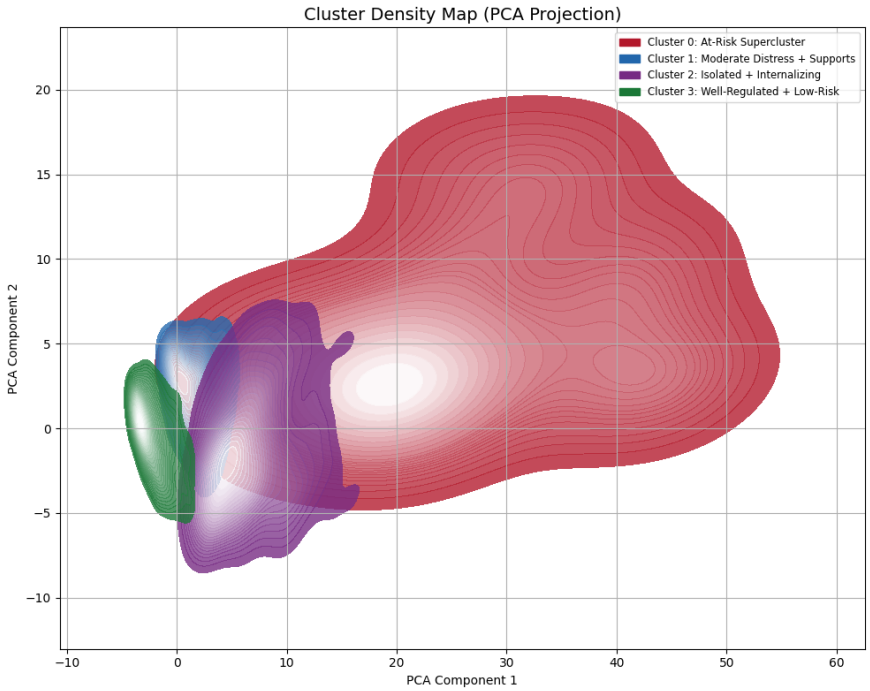

Figure 10: Density map of composite clusters in PCA-reduced space

Figure 9: Cluster Density Map The cluster density map visualizes the distribution of adolescents in a reduced two-dimensional space based on PCA. Cluster 0 is the most widely dispersed, indicating a high degree of heterogeneity among students with high risk behaviors and poor mental health. Cluster 3 is tightly concentrated, reflecting a well-regulated, low-risk group with consistently protective profiles. Cluster 1 shows moderate density, capturing students with some psychological distress but likely supported by stable social or behavioral factors. Cluster 2 occupies a distinct but less cohesive space, representing adolescents who are emotionally vulnerable and tend to internalize distress. Reliability Checks

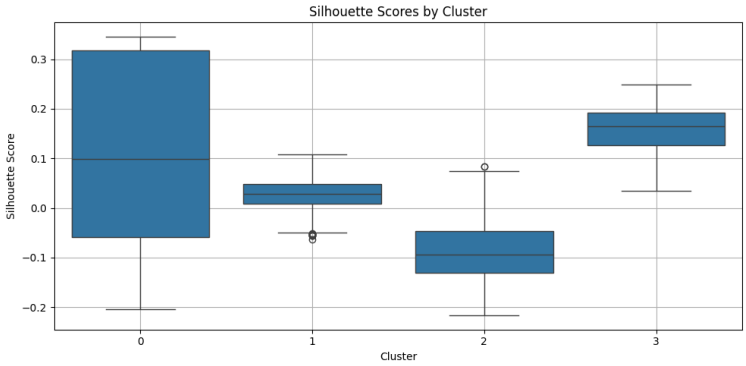

Figure 11: Silhouette score analysis across different cluster sizes

Figure 10: Cluster Validation Using Silhouette Scores Lastly, to evaluate the reliability of the clustering solution, we examined silhouette scores for each cluster. As shown in the boxplot, Cluster 3 demonstrated the highest and most consistent silhouette scores, indicating strong internal structure and clear differentiation from other clusters. Cluster 1 held moderate silhouette values, consistent with a group that is moderately cohesive but may have some overlap with other clusters. Cluster 0, while showing a broader distribution, maintained a generally positive silhouette range, suggesting that its larger and more heterogeneous membership still formed a valid group overall. Cluster 2 had the lowest silhouette scores, including many negative values, indicating a relatively weaker cohesion and potential misclassification for some individuals.

4 Discussion

** This is the revised draft for the domain clustering**

The findings of domain specific clustering support both H_1 and H_2 . First, consistent to H_1, distinct subgroups of adolescents emerged within each contextual domain—home, social, and physical health—when clustering was performed using domain-specific variables alongside mental health indicators. In all three cases, a two-cluster solution was optimal, separating adolescents into high- and low-risk profiles. PCA visualizations confirmed these patterns by illustrating consistent visual separation between clusters across domains, aligning with moderate silhouette scores and meaningful differences in cluster composition.

Furthermore, as suggested by H_2, adolescents in higher-adversity clusters (e.g., those with unstable home environments, low social support, or poor physical health) reported higher levels of psychological distress relative to their lower-risk counterparts. This was reflected not only in the descriptive summaries but also in the dispersion patterns seen in PCA space—where high-risk clusters (Cluster 0) were consistently more diffuse, suggesting greater heterogeneity in psychosocial burden. By contrast, Cluster 1 was more compact, indicating that supportive environments tend to be associated with more uniform and favorable mental health outcomes. Cross-domain heatmaps showed that low-risk and high-risk cluster memberships were highly overlapping across domains, reinforcing the idea that contextual adversity and protection are cumulative rather than isolated. These results highlight the interconnected nature of adolescents' experiences across home, social, and physical health contexts, emphasizing the need for integrated approaches to youth mental health support.

** Below was helen and yawen's draft** Based on the clustering results and validation analyses, our findings provide evident support for the proposed hypotheses. Consistent with H□, both the domain-specific and full-dataset clustering solutions revealed distinct psychosocial profiles among adolescents. The domain-specific analyses—focused separately on home, social, and physical health contexts—each produced meaningful two-cluster solutions, showing that mental health consistently co-varies with home stability, peer support, and physical activity. These context-specific profiles provided a primary understanding of how mental distress is embedded in particular environments. In the full-dataset clustering, this understanding was extended and refined into a more nuanced four-cluster solution. The four clusters—Cluster 0 (At-Risk Supercluster), Cluster 1 (Moderate Distress + Supports), Cluster 2 (Socially Isolated + Internalizing Risk), and Cluster 3 (Well-Regulated + Low-Risk)—are all proved to be behaviorally and spatially meaningful. The overlap across domain-specific and full-data solutions reinforces the consistency of certain patterns, such as the co-occurrence of social isolation, low physical activity, and emotional distress, as seen in Clusters 0 and 2. This provides clear support for H□, as adolescents with fewer protective contextual factors tended to report significantly higher frequencies of sadness, suicidal ideation, and poor mental health. For H□, the inclusion of contextual variables—such as bullying exposure, early sexual activity, and family environment—enhanced the interpretability and distinctiveness of the clusters. This was particularly evident in the heatmap of the top 20 most distinctive features

(Figure 5), where risk-related behaviors beyond mental health symptoms differentiated clusters. Finally, H□ is supported by multiple forms of visual and statistical validation. PCA scatterplots showed meaningful separation between clusters, while silhouette scores and cluster density plots revealed differences in internal cohesion and boundary clarity. Cluster 3, for example, was highly compact and well-defined, while Cluster 2 appeared more loosely defined and variable. These findings confirm that adolescent mental health profiles are structurally consistent across analytical approaches.

5 Conclusion and Future Directions

This study demonstrates the use of unsupervised learning in identifying distinct psychosocial profiles among adolescents by integrating mental health indicators with contextual variables including physical activity, social support, and home environment. The four-cluster solution revealed meaningful subgroups that varied in both behavioral risk and emotional vulnerability, offering a more comprehensive understanding of adolescent well-being than traditional variable-focused approaches. Moreover, future research could build on these findings by incorporating longitudinal data to examine how adolescents transition between clusters over time, or by linking these profiles to intervention outcomes could inform the development of targeted, data-driven strategies to support youth at varying levels of risk.