Youth Mental Health Clustering

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

June 13th, 2025

1 Cover Letter

The method and the result section for both domain-specific clustering and composite clustering was revised to improve clarity and strengthen the interpretive flow of the results. In the earlier version, the writing primarily described outcomes without consistently linking them to contextual relevance or theoretical framing. In this revision, the results were restructured to more clearly articulate how each cluster reflects psychosocial conditions and to better explain the significance of cluster cohesion, separation, and cross-domain alignment. The revised narrative more directly connects findings to the stated hypotheses, resulting in a more coherent and contextually grounded interpretation of adolescent mental health risk across the home, social, and physical health domains.

2 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.

- **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.

3 Methods

3.1 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.

3.2 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.

3.3 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.

4 Result

4.1 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.

4.1.1 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.

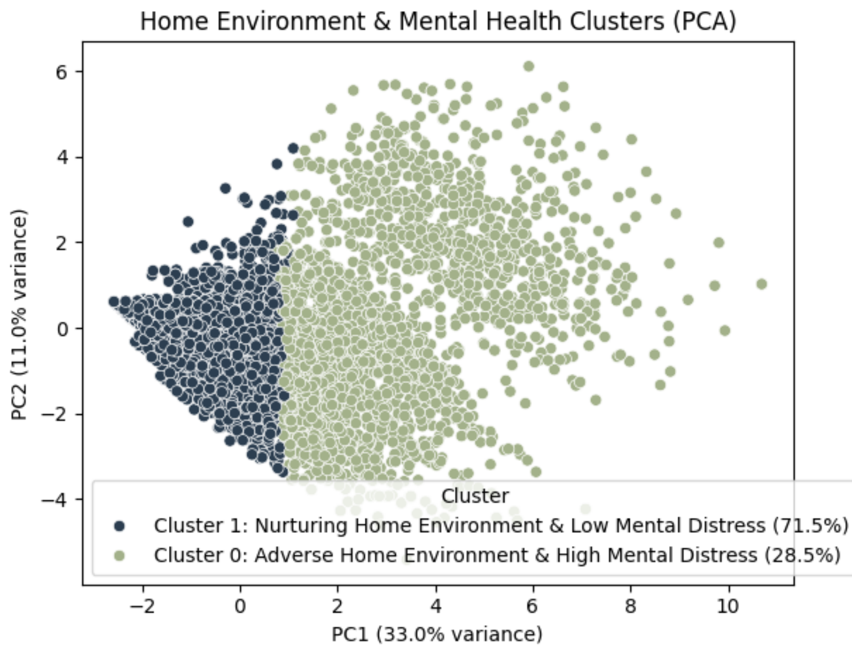

Figure 1: PCA visualization of clusters in the home environment domain

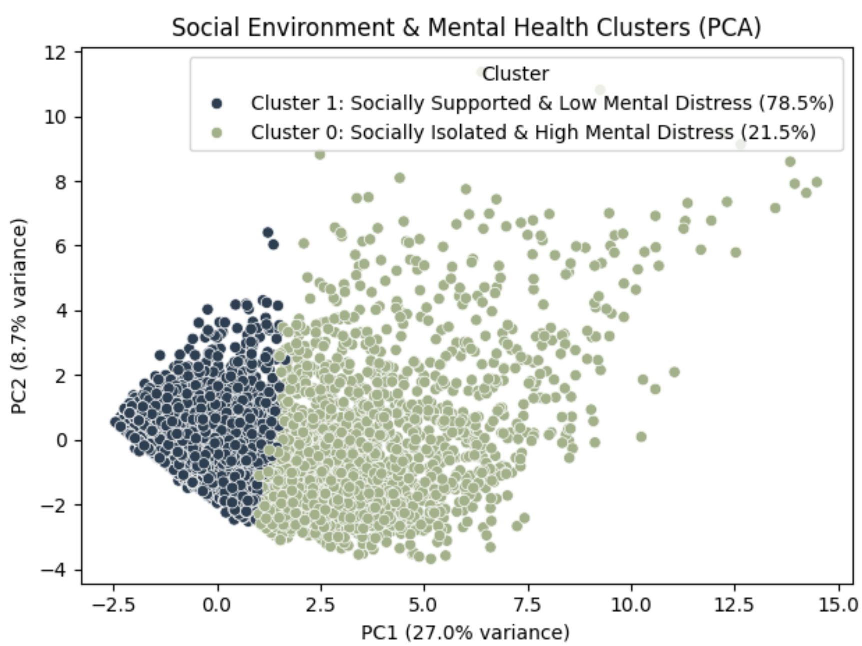

Figure 2: PCA visualization of clusters in the social environment domain

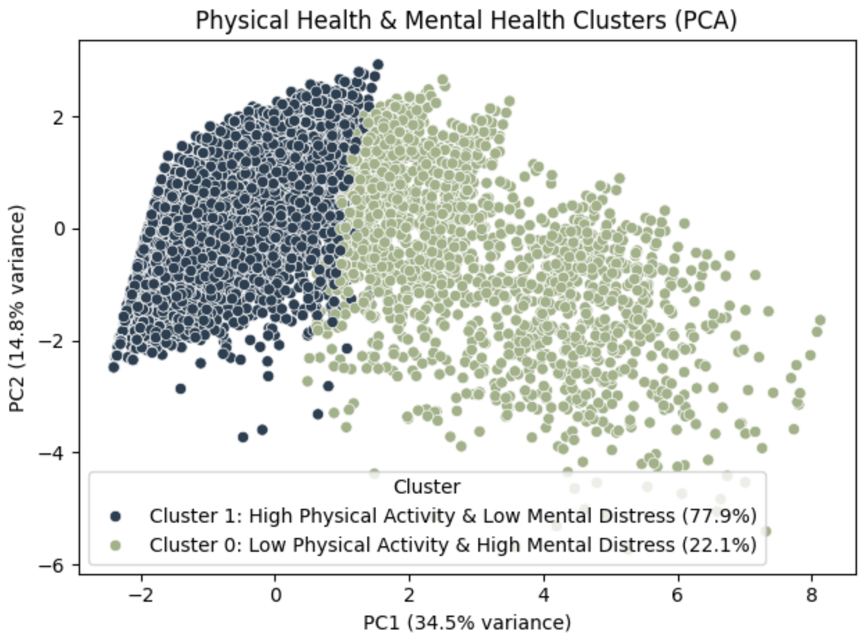

Figure 3: PCA visualization of clusters in the physical health domain

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.

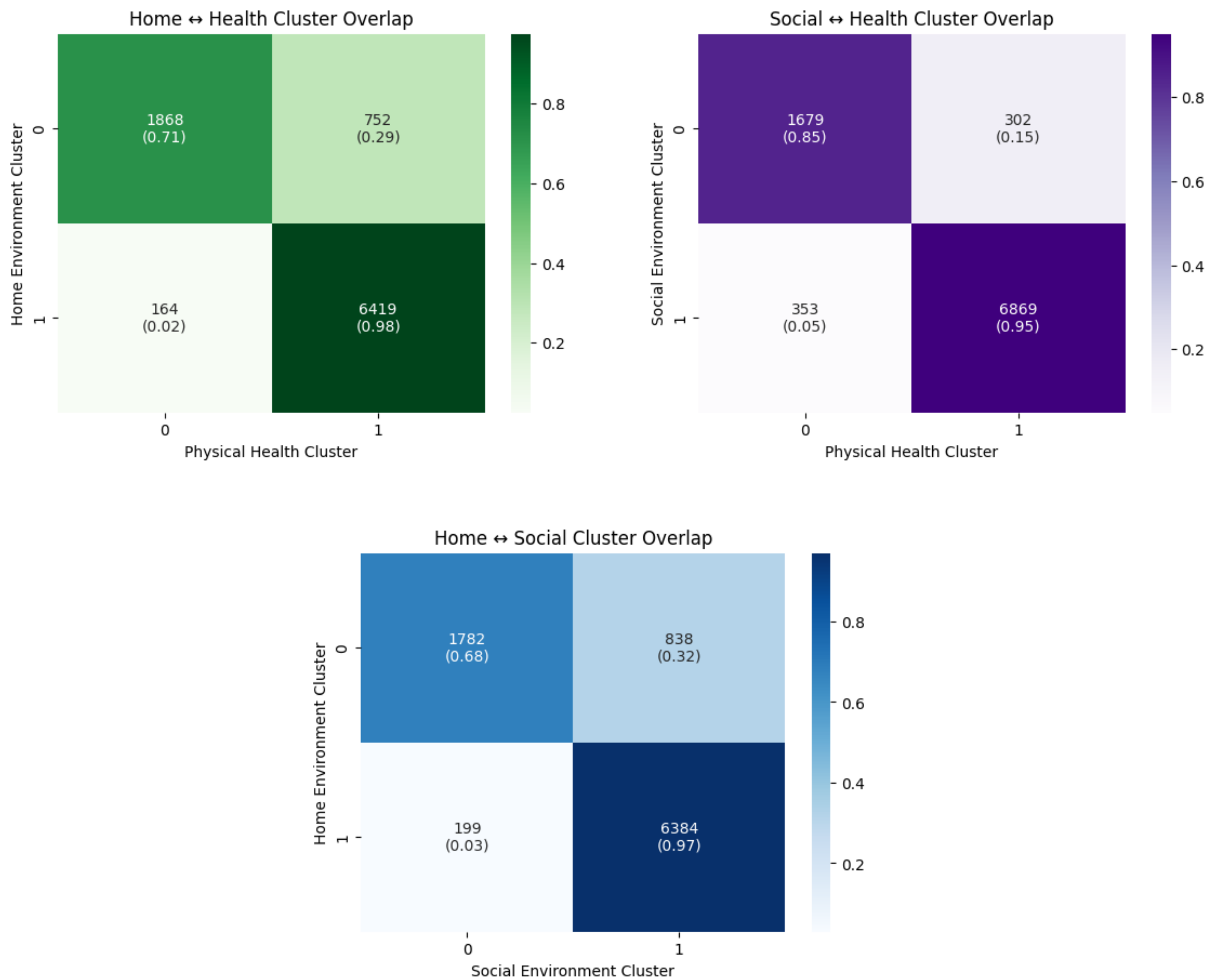

Figure 4: Cross-Domain Cluster Overlap Heatmaps

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.

4.2 Composite Clustering

In response to feedback highlighting the limitations of our original k=4 solution, particularly the incoherence of Cluster 2 and the superior silhouette score at k=2, we re-evaluated the composite clustering model using k=2. As before, the full dataset integrated all domain-specific variables and underwent z-score normalization. We performed dimensionality reduction with PCA (retaining two components), and then applied KMeans clustering with k=2 to the standardized data.

Although the elbow method had previously suggested diminishing inertia gains after k=4 or 5, we prioritized silhouette-based validation in this revision. The silhouette analysis supported k=2 as the highest-performing configuration in terms of cohesion and separation. However, we found that even with k=2, internal coherence varied sharply between the two resulting clusters.

4.3 PCA Projection

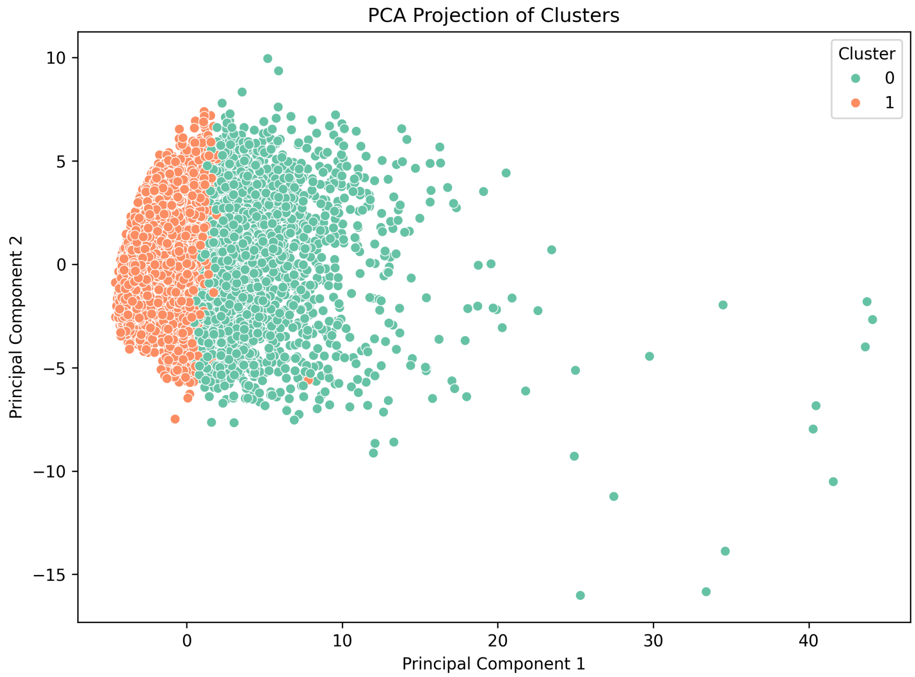

Figure 5: PCA Projection of KMeans Clusters (k = 2, Full Dataset)

The PCA projection of the two-cluster KMeans model offers a visual summary of how individuals are distributed across the first two principal components of the full dataset. As shown in Figure 5, Cluster 1 (orange) forms a relatively tight, compact grouping concentrated on the left side of the plot. In contrast, Cluster 0 (green) is more dispersed, occupying a broader region across both principal components.

This spatial pattern aligns with earlier findings: Cluster 1 corresponds to a low-risk, well-regulated group, whereas Cluster 0 captures a more diffuse population characterized by elevated mental health risk. The partial separation along the first principal component indicates that the clustering algorithm was able to detect a meaningful axis of variation—likely reflecting an underlying risk gradient—but not a complete partition of the sample.

The presence of multiple outliers within Cluster 0 further supports the interpretation that emotionally distressed students are not easily clustered and exhibit high variability in how risk manifests.

While the PCA projection confirms that the model captures real structure in the data, it also highlights the limitations of applying centroid-based clustering to a population with significant internal heterogeneity.

4.4 Cluster Interpretation

Figure 6: Top 20 Most Distinctive Features by Average Cluster Importance in the Composite model

Figure 6 displays the average values of the 20 most distinctive variables across the two clusters identified by the KMeans algorithm. Each row represents a survey question, and each column corresponds to a cluster. The heatmap reveals clear behavioral contrasts, especially in areas related to substance use, early experimentation, and sexual behavior.

Cluster 0 represents a higher-risk profile, with elevated averages across nearly every domain of concern. Members of this group reported more frequent use of marijuana (Q46, Q48), alcohol (Q42, Q44), and e-cigarettes (Q36), and they tended to initiate these behaviors at younger ages (Q32, Q41, Q47). They also reported more sexual partners (Q58, Q59), higher likelihood of substance use before sex (Q60), and lower engagement in protective practices such as condom use (Q61) and verbal consent (Q94). Other markers, like texting while driving (Q11) and inconsistent breakfast habits (Q75), suggest additional indicators of impulsivity or dysregulation. While this group doesn't map onto a single domain of risk, it consistently scores higher across a range of behavioral indicators.

Cluster 1, by contrast, appears to represent a lower-risk or well-regulated group. Individuals in this cluster consistently reported later initiation of substance use, lower recent use, fewer sexual partners, and greater likelihood of engaging in protective behaviors. They were also more likely to eat breakfast regularly and reported less exposure to family conflict (Q89). While not entirely risk-free, this group shows a much narrower behavioral risk profile, with many indicators concentrated near the minimum values.

Taken together, the heatmap reinforces the idea that one cluster reflects more diffuse but elevated engagement across multiple forms of risk, while the other reflects more delayed or normative behavior. The clearest separation appears in lifetime marijuana use, age at first intercourse, and sexual risk practices—suggesting that these variables may be central to the model's internal differentiation.

4.5 Mental Health Risk

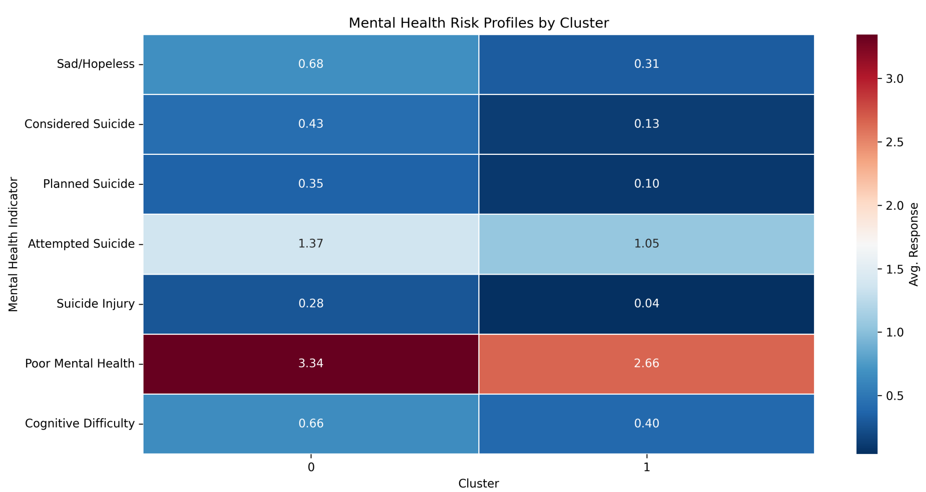

Figure 7: Mental Health Risk Profiles by Cluster (Average Response Levels)

The heatmap summarizes the average scores for seven mental health indicators across the two clusters. Consistent with previous findings, Cluster 0 shows higher average responses on every item, suggesting greater psychological distress and elevated risk across multiple dimensions: Suicide-related indicators: - Consideration (0.43), planning (0.35), attempts (1.37), and injury (0.28) are all markedly higher in Cluster 0 than in Cluster 1, where values are near the floor. - Poor mental health (Q84) stands out with a cluster average of 3.34, compared to 2.66 in Cluster 1 — the single largest difference. - Cognitive difficulty (Q106) and sadness/hopelessness (Q26) are also higher in Cluster 0, indicating emotional and functional struggles.

In contrast, Cluster 1 reports lower distress across all indicators, particularly in suicidal thoughts and behavior, where averages are near zero. This cluster aligns with the earlier behavioral interpretation of a lower-risk, well-regulated group.

This pattern reinforces the interpretation that Cluster 0 concentrates psychosocial vulnerability, though even within this group, variability remains. The elevated levels of suicide attempts and emotional distress underscore the clinical relevance of the profiles, and suggest that clustering on behavioral data successfully captured meaningful gradients in mental health status.

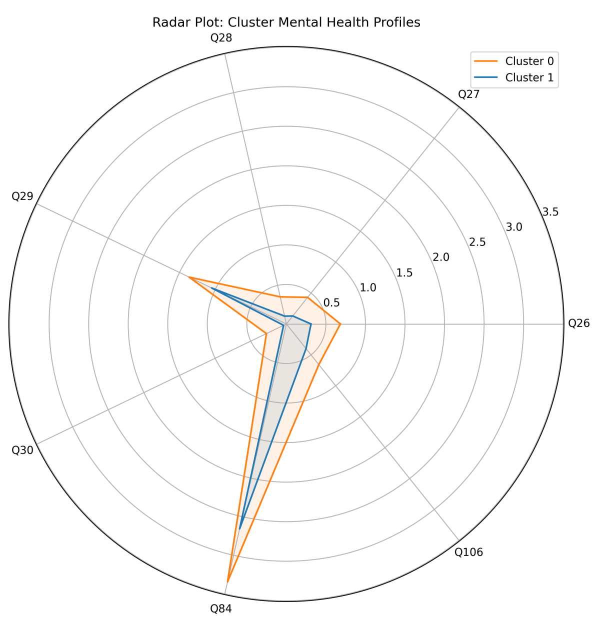

Figure 8: Radar Plot of Cluster Mental Health Profiles (For Selected Mental Health Indicators)

The radar plot reinforces the heatmap findings, visually summarizing mental health differences across clusters. Cluster 0 consistently scores higher on all indicators, especially poor mental health (Q84) and suicide attempts (Q29), forming a distinct outward profile of elevated distress. Cluster 1's tighter shape reflects lower and more uniform responses across items, aligning with a lower-risk mental health profile.

4.6 Cluster Summary

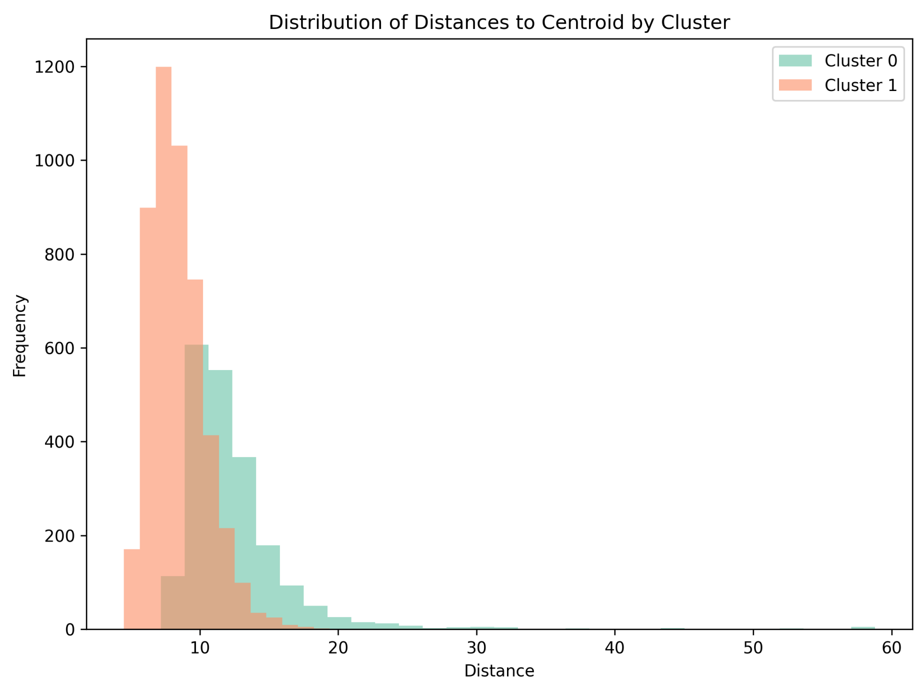

Figure 9: Distribution of Distances to Cluster Centroid

The distance distribution plot illustrates internal cohesion within each cluster. Cluster 1 (orange) shows a tight, narrow distribution centered around lower distances, indicating strong within-cluster similarity. In contrast, Cluster 0 (blue) has a wider spread with a longer tail, reflecting greater internal variability. This supports prior findings that the high-risk cluster is less well-defined, encompassing a broader range of psychosocial profiles.

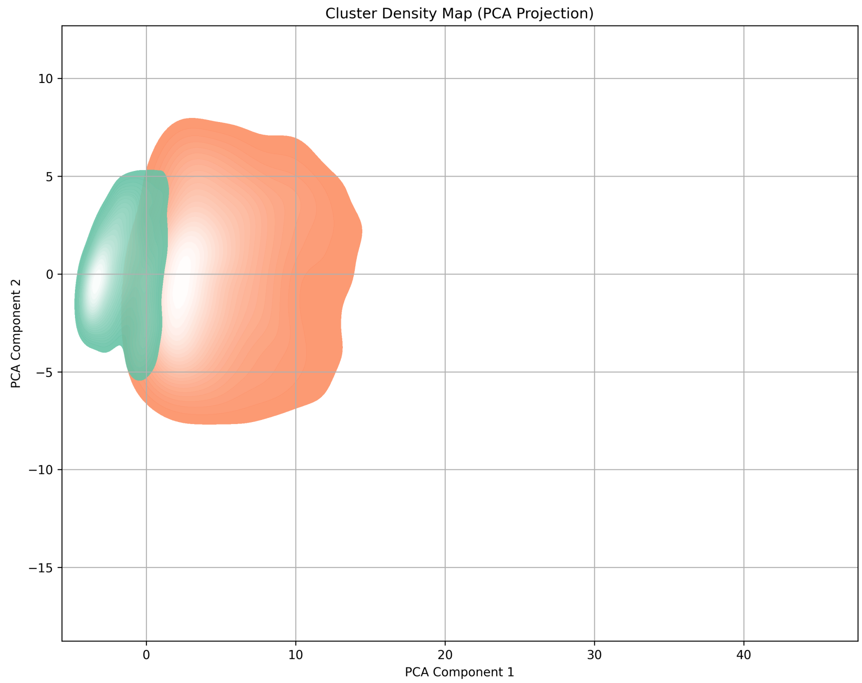

Figure 10: Density map of composite clusters in PCA-reduced space

This density map helps visualize how tightly or loosely the two clusters are distributed in reduced PCA space. Cluster 1 (blue), which we've consistently interpreted as the well-regulated, low-risk group, shows up as a dense and compact area—suggesting strong internal cohesion and shared characteristics. Cluster 0 (orange), in contrast, is much more spread out, with a less defined shape and broader coverage across both components. That pattern lines up with everything else we've seen: the high-risk group is more variable, harder to pin down, and doesn't cluster as neatly. This plot reinforces the idea that emotional and behavioral distress among adolescents doesn't follow a single trajectory—it takes many forms, and that complexity shows up clearly here.

4.7 Reliability Checks

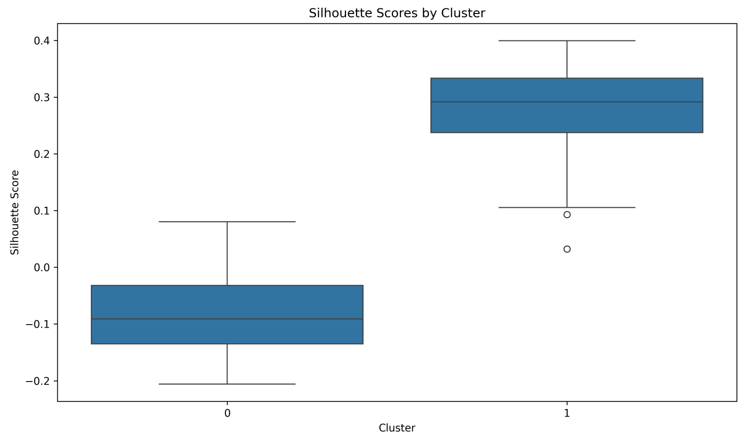

Figure 11: Cluster Validation Using Silhouette Scores

The silhouette scores reflect a clear difference in cluster cohesion. Cluster 1 shows consistently positive values, centered around 0.3, suggesting it forms a well-defined and cohesive group. In contrast, Cluster 0 has lower and often negative scores, indicating many members are not clearly grouped and may lie closer to the boundary between clusters. This aligns with earlier findings suggesting that students in Cluster 0 — those with elevated risk — are harder to characterize cleanly.

Additional metrics support this interpretation. The average silhouette score across all students is 0.175, which is modest and reflects the difficulty of cleanly partitioning this population. The Calinski–Harabasz Index of 666.06 indicates decent cluster separation relative to compactness, while the Davies–Bouldin Index of 2.995 suggests moderate overlap between the clusters. Overall, the model captures meaningful structure but also highlights the statistical fuzziness of high-risk profiles.

4.8 DBSCAN Results

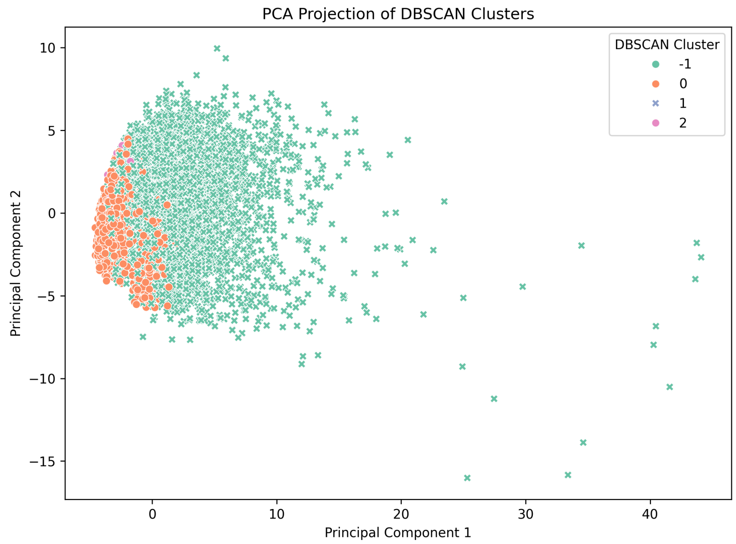

Figure 12: DBSCAN PCA Projection Plot

In response to reviewer feedback suggesting a density-based alternative to KMeans, we applied DBSCAN to the full dataset. Unlike KMeans, DBSCAN defines clusters based on point density, requiring two hyperparameters: the neighborhood radius (eps) and the minimum number of points (min_samples) needed to form a cluster.

We systematically explored over 30 parameter combinations, varying `eps` between 1.5 and 3.5 and `min_samples` between 2 and 6. Across all runs, DBSCAN consistently identified either no clusters (all points marked as noise) or a small number of very compact clusters alongside a large set of noise points. In our most stable solution (e.g., eps = 2.7, min_samples = 4), DBSCAN identified three clusters and 4,440 noise points, with a silhouette score of -0.064; which is consistent with weak separation and significant class overlap.

Importantly, the large noise group exhibited elevated scores on suicide attempts, sadness, and poor mental health, mirroring the diffuse high-risk cluster from our KMeans analysis. These findings suggest that emotionally distressed adolescents do not form a high-density cluster, and may instead be best conceptualized as a set of heterogeneous, loosely affiliated individuals. DBSCAN's treatment of these students as "noise" supports our conclusion that standard clustering methods struggle to capture the complexity of high-risk youth profiles.

5 Discussion

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.

Our revised composite clustering analysis incorporated direct feedback about model validity and cohesion. Moving from the original four-cluster KMeans solution to a streamlined two-cluster model improved silhouette metrics and clarified boundaries for a well-regulated, low-risk group. However, it also revealed persistent challenges in defining high-risk adolescents as a cohesive cluster.

Cluster 0, which consistently captured students with the highest mental health risk, remained internally diffuse—both in the original k=4 model and in this revised k=2 model. This recurring pattern highlights a broader insight: adolescents with elevated psychosocial distress do not form a well-bounded group. Instead, they display varied levels and types of distress, coping styles, and contextual influences, all of which challenge fixed-profile clustering approaches.

To probe this issue further, we applied DBSCAN, a density-based algorithm capable of detecting irregularly shaped clusters and excluding noise. The model identified only a few small clusters while labeling nearly all students—and especially those with high mental health risk—as noise. While counterintuitive, this result supports our interpretation: at-risk students are not densely grouped in the feature space and are better conceptualized as heterogeneous outliers rather than a single, unified subgroup.

While the two-cluster KMeans model still offers a practical, risk-aligned classification framework, our analyses emphasize the limitations of rigid or low-k clustering when applied to complex adolescent data.

6 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.

Ultimately, this project underscores both the strengths and shortcomings of unsupervised learning in applied mental health research. Well-regulated populations may be easily profiled, but high-risk students require nuanced, adaptable approaches—statistically and clinically—to be meaningfully understood and supported.

Future work should explore soft clustering methods, individualized predictive models, or hybrid frameworks that can better account for the diverse profiles and experiences embedded within psychosocial distress. These approaches may offer better tools for capturing the variability observed in high-risk adolescents and open new avenues for building targeted intervention strategies grounded in data.