Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

UNCOVERING UNDERGRADUATE BEHAVIORAL PATTERNS: A STUDY OF ACADEMIC AND SOCIAL PROFILES ACROSS FOUR UNIVERSITIES USING CHI-SQUARE AND K-MODES CLUSTERING

Salman Ahmad[1], Habib Ullah Khan[*2], Atta Ullah[3]

[1,*2,3] *Department of Statistics, University of Malakand, Chakdara, Dir Lower, KPK, Pakistan*

[1]salmanahmadstatistics1961@gmail.com, [*2]habibullah858@uom.edu.pk, [3]attaullahaarslan34@gmail.com

DOI: https://doi.org/10.5281/zenodo.16917054

Keywords
Behavioral Patterns, CGPA, Chi-Square, K-Modes Clustering, Social Media, Study Hours

Article History
Received: 19 May, 2025
Accepted: 24 July, 2025
Published: 21 August, 2025

Corresponding Author: *
Habib Ullah Khan

Abstract
Quantitative understanding of relationships between students' behavioral patterns and academic performances is a significant step towards personalized education. This study investigates behavioral patterns among undergraduate students across four universities, focusing on various behaviors like hours spent on social media, internet consultation during academic tasks, CGPA, daily study hours, preferred entertainment, and preferred mode of social interaction. Data were collected through a structured questionnaire comprising twelve categorical variables. Chi-square tests identified several significant associations, including the relationship between CGPA and daily study hours, as well as between preferred entertainment type and social interaction mode, indicating clear behavioral dependencies. K-modes clustering, with the optimal number of clusters determined via the elbow method, yielded five distinct student profiles. Cluster profiling revealed patterns such as academically focused students with limited leisure activities, socially active students with moderate academic engagement, and balanced profiles combining both strong academic performance and diverse leisure interests. These findings offer valuable insights into how behavioral tendencies correlate with academic outcomes, contributing to targeted educational strategies that address both performance enhancement and student well-being.

INTRODUCTION

In recent years, the increasing integration of digital technologies into students' daily lives has profoundly influenced their academic performance, social behavior, and lifestyle choices. Undergraduate students, in particular, navigate a complex interplay of academic responsibilities, social interactions, and recreational activities, all of which can shape their educational outcomes and overall well-being. Variables such as the number of hours spent on social media, reliance on the internet for academic assignments, cumulative grade point average (CGPA), daily study hours, preferred entertainment formats, and modes of social interaction offer valuable insights into the behavioral and learning patterns of university students. Understanding these patterns is crucial for educational researchers, as it can inform institutional

policies, enhance academic support systems, and guide targeted interventions.

The present study focuses on identifying and analyzing patterns within these variables by collecting primary data from undergraduate students across four universities (University of Swat, University of Malakand, University of Shangla, and University of Shiringal). These universities are located in district Swat, Dir lower, Shangla and Dir upper respectively. These universities were selected from a pool of eight universities:

- University of Swat
- University of Shangla
- University of Buner
- University of Agriculture, Swat
- University of Malakand
- University of Shiringal
- University of Chitral
- Abdul Wali Khan University, Timergara

Universities were strategically selected based on their geographical locations to ensure broad regional representation. This selection criterion was aimed at capturing the heterogeneity of student populations influenced by varying socio-cultural and environmental contexts. By including institutions from diverse geographic areas, the research sought to encompass a wide spectrum of academic, social, and behavioral patterns. This approach enhanced the representativeness of the sample, thereby increasing the validity and generalizability of the study's findings on student behavior.

The selection of the variables was intentional, given their relevance to both academic engagement and lifestyle balance. Social media usage, for instance, has been associated with both positive academic support and potential distraction, depending on the context and intensity of use. Similarly, preferred modes of entertainment and social interaction may reflect underlying personality traits, coping strategies, or stress management approaches, while study-related behaviors such as hours spent studying and the use of the internet for assignments directly relate to academic performance indicators such as CGPA.

To uncover underlying associations between these categorical variables, the Chi-square test of independence was employed. This statistical method is particularly suitable for determining whether there are significant relationships between pairs of categorical variables, making it an essential precursor to more complex pattern recognition methods. In this context, Chi-square analysis allowed for the identification of statistically significant links such as whether higher CGPA is associated with specific study patterns or whether entertainment preferences vary significantly with social media usage habits.

Following the examination of variable associations, the study applied K-modes clustering to categorize students into distinct behavioral and academic profiles. K-modes clustering, an extension of the well-known K-means algorithm designed for categorical data, was selected for its capacity to efficiently handle non-numeric variables while minimizing information loss through data transformation. This method enabled the grouping of students into homogenous clusters based on shared behavioral patterns, offering a richer and more interpretable understanding of the data than would be possible through univariate or bivariate analyses alone.

The optimal number of clusters was determined using the elbow method, which plots the cost function (within-cluster dissimilarity) against the number of clusters to identify the point where additional clusters provide diminishing returns in explanatory power. In this study, the elbow plot indicated that five clusters best represented the data structure, balancing interpretability and complexity.

By integrating Chi-square testing for association with K-modes clustering for pattern discovery, this research bridges statistical inference and unsupervised learning to provide a nuanced understanding of student behavioral profiles. The findings have potential applications in academic advising, mental health support, and policy formulation aimed at promoting balanced and effective study habits among undergraduates.

Literature Review

The investigation of behavioral patterns among undergraduate students, encompassing variables such as social media usage, study hours, preferred entertainment, social interactions, and their associations with academic outcomes like Cumulative Grade Point Average (CGPA), has gained prominence in educational research. This literature review synthesizes key studies, focusing on the impact of digital and social behaviors on academic success,

statistical associations (e.g., via Chi-square tests), and pattern identification through clustering methods. Drawing from recent systematic reviews and empirical studies, it highlights how these factors interrelate to inform targeted interventions for student well-being and performance.

Social media usage has been extensively studied for its dual role as a facilitator of learning and a potential distraction. A study on secondary school students revealed high daily usage (2-4 hours on average), primarily on platforms like Facebook and WhatsApp, but found no significant influence of usage frequency on academic achievement in subjects like Accounting ($F(3, 146) = 1.948$, $p \geq 0.05$) (Oguguo et al., 2020). However, gender moderated this effect, with females achieving higher mean scores (76.23 vs. 71.12 for males; $t(2, 148) = -1.994$, $p \leq 0.05$), suggesting contextual differences in how social media affects focus and performance (Oguguo et al., 2020). In university settings, motives for social media use (e.g., information-seeking or entertainment) were shown to indirectly impact GPA through daily time spent, with excessive usage linked to lower academic outcomes among Vietnamese students (Cuong et al., 2025). Broader reviews confirm mixed effects: while social media can enhance interpersonal relations and well-being, heavy reliance often correlates negatively with GPA and study habits (Chandrasena & Ilankoon, 2022; Al Mosharrafa et al., 2024; Connolly, n.d.). For instance, increased Facebook activity was negatively associated with GPA in a cohort of Danish undergraduates ($r_S = -0.15$, $p < 0.05$) (Kassarnig et al., 2018).

Study hours and related habits emerge as strong predictors of CGPA. A causal analysis of socio-academic factors demonstrated significant direct effects of study hours and class attendance on CGPA ($p < 0.01$), with group study also contributing indirectly (Hosen et al., 2025). Systematic reviews underscore prior academic achievement and study behaviors (e.g., time on task, self-regulation) as the most influential variables, appearing in 69% of predictive studies, often yielding high accuracy in regression models (up to 93% at course level) (Alyahyan & Düştegör, 2020; Hellas et al., 2018). Behavioral patterns, including psychological attributes like motivation and anxiety, further modulate these associations; for example, self-efficacy in preparing study schedules correlated positively with higher CGPA clusters (Talib et al., 2023). Chi-square tests have been employed for feature selection, identifying significant relationships between categorical variables such as demographics and performance (e.g., in comparing greedy and information gain methods) (Talib et al., 2023). During the COVID-19 era, models incorporating study habits predicted grades with enhanced accuracy, emphasizing adaptive behaviors (Asim et al., 2024).

Social interactions, another behavioral facet, show homophily effects: students with higher-performing peers in call/text networks exhibited better CGPA (mean peer GPA correlation: $r_S = 0.25$, $p < 0.001$), while proximity to low performers was detrimental (Kassarnig et al., 2018). Entertainment preferences, though less studied, align with broader leisure behaviors; excessive non-academic activities (e.g., social media as entertainment) were linked to reduced study time and lower CGPA (Han, 2023; Connolly, n.d.).

Clustering methods, such as K-means and fuzzy clustering, have been pivotal in profiling student behaviors without predefined labels. A study on 140 undergraduates applied K-means to weekly data, yielding three clusters (low, average, high performance) based on learning styles and self-efficacy, with week-9 behaviors best predicting outcomes (accuracy improved by 15% over grade-based models) (Talib et al., 2023). Systematic reviews report clustering as an unsupervised technique for grouping students by engagement patterns, often preceding classification (e.g., K-means for at-risk identification), with social and behavioral data enhancing model robustness (Alyahyan & Düştegör, 2020; Talib et al., 2023; Hellas et al., 2018). Fuzzy clustering has also been used to assess psychological health impacts on performance, revealing clusters where low study engagement correlates with poorer CGPA (Han, 2023). In network-based analyses, clustering social ties identified performance homophily, with supervised models achieving 57.9% accuracy in classifying low/moderate/high performers (Kassarnig et al., 2018).

The literature indicates clear dependencies between behavioral variables and academic success, with social media often posing risks, study habits providing protective effects, and clustering enabling nuanced

profiling. Future research should integrate real-time data for dynamic interventions. Gaps remain in entertainment and interaction modes, aligning with the current study's focus.

Methodology

Data Collection

The present study employed a quantitative, cross-sectional design to examine patterns in undergraduate students' academic and social behaviors. Primary data were collected from four universities, targeting undergraduate students across diverse disciplines. A structured, self-administered questionnaire was developed to capture responses on twelve categorical variables related to key aspects of students' academic and social lives. These variables included daily hours spent on social media, frequency of Internet consultation when completing assignments, cumulative grade point average (CGPA), daily study hours, preferred type of entertainment, and preferred mode of social interaction, among others. The selection of these variables was guided by the objective of identifying behavioral and lifestyle patterns that may be associated with students' academic performance and social engagement. Each variable was formulated in a categorical format to facilitate non-parametric statistical analysis and clustering based on discrete patterns.

Rationale for Variable Selection

The variables were chosen to represent a balanced combination of academic habits (e.g., study hours, assignment strategies), technology use (e.g., social media engagement, Internet consultation), and leisure/social behaviors (e.g., entertainment preferences, social interaction modes). These aspects are widely recognized in educational and behavioral research as influential in shaping students' academic outcomes and social development. The intention was to capture multidimensional lifestyle profiles that could be meaningfully analyzed for both association and clustering.

Data Analysis Procedures

Data analysis was conducted in two main stages:

Assessment of Associations Using Chi-Square Tests

To explore potential relationships among categorical variables, Pearson's Chi-square test of independence was applied to all relevant variable pairs. This test was chosen because it is specifically designed to evaluate whether an observed frequency distribution differs significantly from an expected distribution under the assumption of independence. By doing so, the analysis could identify statistically significant associations that might reflect underlying behavioral or academic linkages between variables. The use of Chi-square tests prior to clustering ensured that the dataset retained meaningful variables and that potential redundancy or noise from unrelated variables was minimized.

Pattern Identification Using K-Modes Clustering

Following the association analysis, K-Modes clustering was employed to identify distinct groups of students exhibiting similar behavioral patterns. K-Modes was selected instead of K-Means because the dataset consisted entirely of categorical variables. Unlike K-Means, which uses Euclidean distance and requires numerical input, K-Modes utilizes a simple matching dissimilarity measure and updates cluster centers using the mode of each variable, making it well-suited for categorical data analysis. This approach allowed for the grouping of students based on shared lifestyle and academic characteristics without requiring arbitrary numerical conversions.

Determination of Optimal Cluster Number

To determine the appropriate number of clusters, the elbow method was applied using the clustering cost (sum of dissimilarities within clusters) as the metric. The clustering cost was calculated for a range of cluster counts, and the results were plotted against the number of clusters. The "elbow point," where the marginal reduction in clustering cost began to diminish, indicated the optimal number of clusters. In this study, the elbow point was observed at five clusters, suggesting that this configuration provided the best balance between model complexity and explanatory power.

Software and Implementation

All statistical analyses were conducted using Python, employing the scikit-learn and kmodes libraries for clustering and the scipy library for Chi-square testing. Data visualization was performed using matplotlib and seaborn to aid interpretation of results.

Ethical Considerations

Participation in the study was voluntary, and respondents were assured of the confidentiality and anonymity of their responses. No personally identifying information was collected, and the data were used solely for academic research purposes.

Frequency Tables and Bar Plots

The following table shows the frequencies of the responses to various questions in the questionnaire, by the respondents.

Category	Subcategory	Frequency	Percent	Valid Percent	Cumulative Percent
Gender	Female	255	36.00%	36.00%	36.00%
	Male	453	64.00%	64.00%	100.00%
Institution	University Of Swat	177	25.00%	25.00%	25.00%
	University Of Malakand	177	25.00%	25.00%	50.00%
	University Of Shiringal	177	25.00%	25.00%	75.00%
	University Of Shangla	177	25.00%	25.00%	100.00%
Online Learning Platforms	Coursera	109	15.40%	15.40%	15.40%
	Khan Academy	96	13.60%	13.60%	29.00%
	Code Academy	50	7.10%	7.10%	36.00%
	Others	453	64.00%	64.00%	100.00%
Online Courses Taken	1-2	208	29.40%	29.40%	29.40%
	3-4	117	16.50%	16.50%	45.90%
	More than 5	84	11.90%	11.90%	57.80%
	None	299	42.20%	42.20%	100.00%
Internet Consultation	Rarely	125	17.70%	17.70%	17.70%
	Sometimes	315	44.50%	44.50%	62.10%
	Often	121	17.10%	17.10%	79.20%
	Always	147	20.80%	20.80%	100.00%
Study Material Preference	Reading	99	14.00%	14.00%	14.10%
	Video Lectures	264	37.30%	37.30%	51.40%
	Both	283	40.00%	40.00%	91.40%
	None	61	8.60%	8.60%	100.00%
Social Media Platform	Facebook	270	38.10%	38.10%	38.10%
	Instagram	157	22.20%	22.20%	60.30%
	Twitter	55	7.80%	7.80%	68.10%
	Others	226	31.90%	31.90%	100.00%
Time Spent on Social Media	1-2 hours	237	33.50%	33.50%	33.50%
	3-4 hours	258	36.40%	36.40%	69.90%
	5-6 hours	121	17.10%	17.10%	87.00%

	More than 7 hours	92	13.00%	13.00%	100.00%
Preferred Social Interaction	Physical Meetings	239	33.80%	33.80%	33.80%
	Online Interactions	153	21.60%	21.60%	55.40%
	Both	204	28.80%	28.80%	84.20%
	None	112	15.80%	15.80%	100.00%
Social Media Platform for Interaction	WhatsApp	539	76.10%	76.10%	76.10%
	Facebook	64	9.00%	9.00%	85.20%
	WeChat	42	5.90%	5.90%	91.10%
	Others	63	8.90%	8.90%	100.00%
Preferred Online Communication Platform	WhatsApp	539	76.10%	76.10%	76.10%
	Facebook	64	9.00%	9.00%	85.20%
	WeChat	42	5.90%	5.90%	91.10%
	Others	63	8.90%	8.90%	100.00%
Preferred Entertainment	Sports and Games	230	32.50%	32.50%	32.50%
	Social Media	222	31.40%	31.40%	63.80%
	Hobby Activities	119	16.80%	16.80%	80.60%
	Extra Learning	137	19.40%	19.40%	100.00%
Paid Entertainment Platforms	Netflix	98	13.80%	13.80%	14.10%
	Amazon Prime	82	11.60%	11.60%	25.70%
	Others	201	28.40%	28.40%	54.10%
	None	325	45.90%	45.90%	100.00%
Study Hours Per Day	1-2 hours	236	33.30%	33.30%	33.30%
	3-4 hours	245	34.60%	34.60%	67.90%
	4-5 hours	130	18.40%	18.40%	86.30%
	More than 6 hours	97	13.70%	13.70%	100.00%
C.GPA	2.0-2.5	68	9.60%	9.60%	9.60%
	2.51-3.0	118	16.70%	16.70%	26.30%
	3.1-3.50	270	38.10%	38.10%	64.40%
	3.51-4.0	252	35.60%	35.60%	100.00%

Table 1: Survey Data Summary

Interpretation of Table 1

255 (36%) respondents out of 708 were female while 453 (64%) were male. According to the collected data 109 (15.4%) students use Coursera, 96 (13.6%) use Khan Academy, 50 (7.1%) use Code Academy, and 453 (64%) use other platforms for online learning purposes. 208 (29.4%) students have taken 1-2 online courses, 117 (16.5%) students have taken 3-4 online courses, 84 (11.9%) students have taken more than 5 online courses and 299 (42.2%) have not taken any online courses. 125 (17.7%) respondents consult the internet rarely, 315 (44.5%) respondents consult the internet sometimes, 121 (17.1%) respondents consult

the internet often and 147 (20.8%) respondents consult the internet always, while doing their homework or assignments. 100 (14.1%) respondents prefer reading, 264 (37.3%) respondents prefer video lectures, 283 (40%) respondents prefer both and 61 (8.6%) respondents do not prefer any of the mentioned material for studying online. 270 (38.1%) respondents prefer Facebook, 157 (22.2%) respondents prefer Instagram, 55 (7.8%) respondents prefer Twitter and 226 (31.9%) respondents prefer other social media platforms. 237 (33.5%) respondents spend 1-2 hours, 258 (36.4%) respondents spend 3-4 hours, 121 (17.1%) respondents spend 5-6 hours and 92 (13%) respondents spend more than 7 hours on social media daily. 239 (33.8%) respondents prefer physical meetings and gatherings, 153 (21.6%) respondents prefer online interactions, 204 (28.8%) respondents prefer both and 112 (15.8%) respondents do not prefer any of the mentioned interactions as social interaction with friends and fellows. 539 (76.1%) respondents use Whatsapp, 64 (9%) respondents use Facebook, 42 (5.9%) respondents use WeChat and 63 (8.9%) respondents use other platforms for connecting and socializing. 230 (32.5%) respondents prefer sports and games in their free time, 222 (31.4%) respondents prefer to spend free time on social media, 119 (16.8%) respondents spend it doing their hobby and 137 (19.4%) respondents use their free time to learn something extra. 98 (13.8%) respondents have paid subscription for Netflix, 82 (11.6%) respondents have paid subscription for Amazon Prime, 201 (28.4%) respondents have paid subscription for other platforms and 325 (45.9%) respondents do not have any subscription for entertainment platforms. 236 (33.3%) respondents study for 1-2 hours daily, 245 (34.6%) respondents study for 3-4 hours daily, 130 (18.4%) respondents study for 4-5 hours daily and 97 (13.7%) respondents study for more than 6 hours daily. CGPA of 68 (9.6%) respondents lie in the range 2.0-2.5, CGPA of 118 (16.7%) respondents lie in the range 2.51-3.0, CGPA of 270 (38.1%) respondents lie in the range 3.1-3.50 and CGPA of 252 (35.6%) respondents lie in the range 3.51-4.0.

Following are bar plots that show the distribution of various variables of across different groups.

Bar Plots

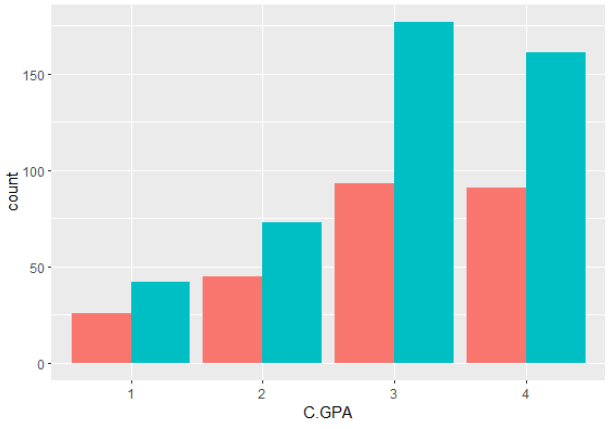

Figure 1: Gender versus C.GPA

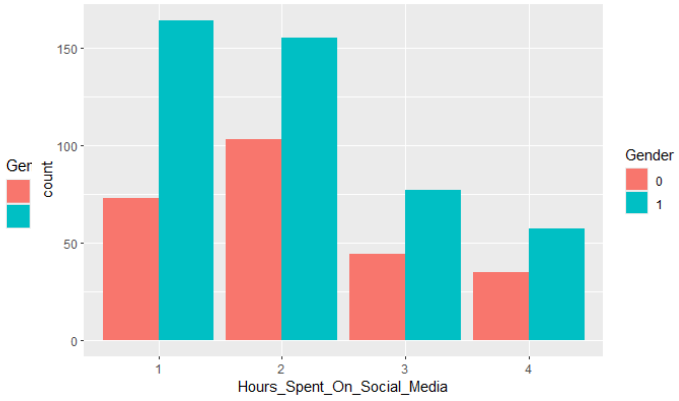

Figure 2: Gender versus Hours spent on social media

Figure 3: Gender versus number of online courses taken

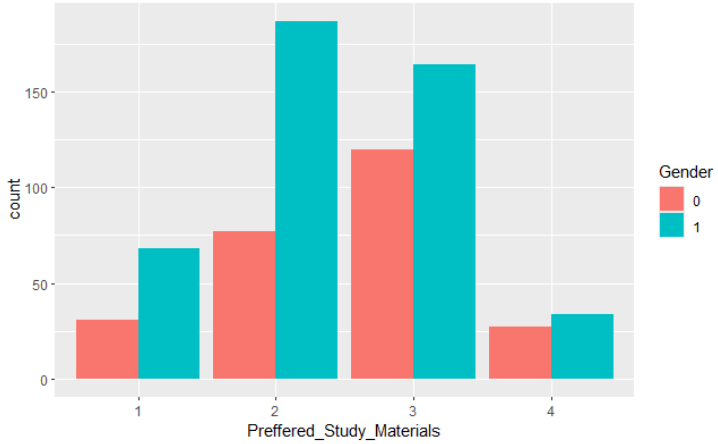

Figure 4: Gender versus preferred materials for study

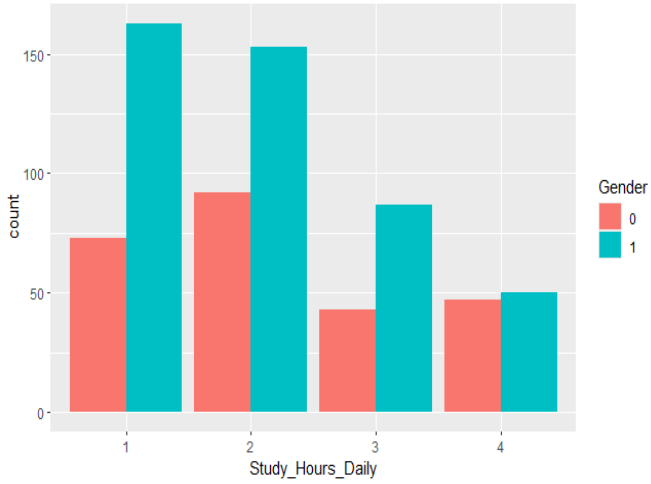

Figure 5: Gender versus preferred entertainment type

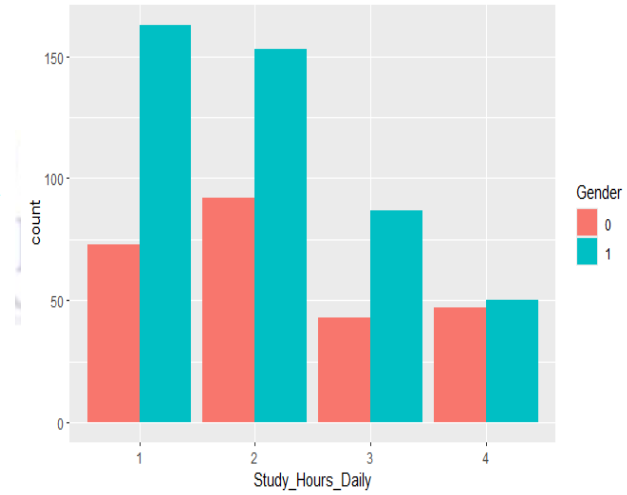

Figure 6: Gender versus study hours daily

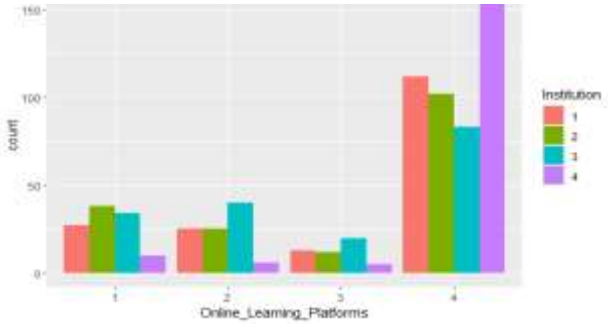

Figure 7: Preferred Online Learning Platform versus Institution

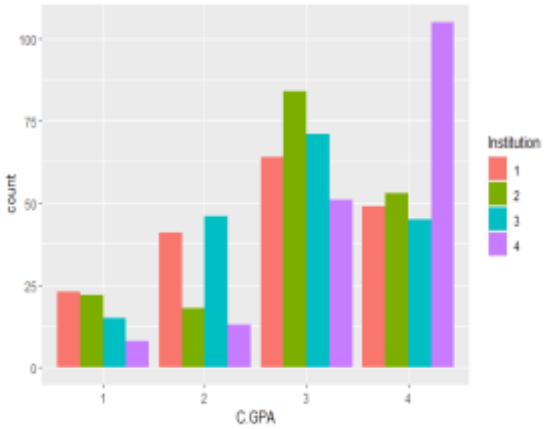

Figure 8: C.GPA versus Institution

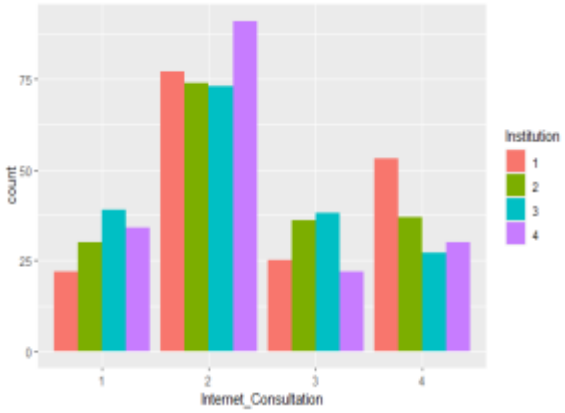

Figure 9: Internet Consultation Frequency versus Institution

Fig 1-6 show the distribution of various variables (CGPA, Hours spent on social media, Online courses taken, preferred study material, preferred entertainment type, study hours daily) across the two gender categories (0 : female, 1 : male). The x-axis represents the categories of the respective variables. More ever, Fig 7-9 show the distribution of some variables (CGPA, Internet consultation frequency, Preferred online learning platform) across the four universities (1: University of Swat, 2: University of Malakand, 3: University of Shiringal, 4: University of Shangla).

Chi-Square tests

Variable Pair	p-value	Cramer's V	Strength
Gender × Study Hours	0.019	0.118	Weak
Gender × Preferred Study Materials	0.0048	0.135	Weak-Moderate
Study Hours × C.GPA	4.02e-6	0.140	Moderate
Hours on Social Media × C.GPA	0.010	0.101	Weak
Entertainment Type × Study Hours	1.3e-5	0.135	Weak–Moderate
Online Courses Taken × Learning Platforms	1.47e-6	0.144	Moderate

Table 2: Chi-Square tests between various pairs of variables

Table 14 shows the results of Chi-square association analysis which revealed some statistically significant relationships between the studied variables, with varying effect sizes shown by the Cramer's V. A weak association was observed between Gender and Study Hours ($p = 0.019$, Cramer's $V = 0.118$), indicating that while gender plays a role in determining study duration, the association is relatively small. Gender and Preferred Study Materials showed a slightly stronger association ($p = 0.0048$, Cramer's $V = 0.135$), indicating that study material preferences may vary moderately between the two genders. The relationship between Study Hours and Cumulative GPA (C.GPA) was statistically significant and of moderate strength ($p = 4.02 \times 10^{-6}$, Cramer's $V = 0.140$), showing that differences in study duration are meaningfully linked

to academic performance. Hours spent on social media and C.GPA showed a weak association ($p = 0.010$, Cramer's $V = 0.101$), suggesting that while social media use is statistically related to GPA, its effect may be limited. A weak to moderate association was also seen between preferred entertainment type and study hours ($p = 1.3 \times 10^{-5}$, Cramer's $V = 0.135$), showing that entertainment preferences may influence or correlate with study time. Finally, the relationship between online courses taken and online learning platforms used showed the strongest effect in the dataset ($p = 1.47 \times 10^{-6}$, Cramer's $V = 0.144$), suggesting that the type of learning platform chosen is meaningfully related to the number of online courses completed.

Overall, while all the associations tested were statistically significant, the effect sizes were mostly weak to moderate, showing that although patterns exist, they are not strong predictors in isolation. These results provide a useful foundation for further exploration, such as clustering, to uncover more complex interaction patterns between the variables.

K-Modes Clustering

Table: Cluster Profiling Summary (K-Modes Clustering Results)

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Gender	Male	Male	Female	Male	Male
Institution	University of Shangla	University of Shiringal	University of Swat	University of Shiringal	University of Malakand
Online Learning Platforms	Others	Others	Others	Others	Others
Online Courses Taken	None	1-2	None	1-2	None
Internet Consultation	Sometimes	Sometimes	Sometimes	Often	Sometimes
Preferred Study Materials	Video Lectures	Both	Both	Both	Video Lectures
Preferred Social Platforms	Facebook	Facebook	Others	Instagram	Facebook
Hours Spent on Social Media	1-2	3-4	3-4	3-4	1-2
Preferred Social Interaction	Physical meetings and gatherings	Both	Physical meetings and gatherings	Physical meetings and gatherings	Physical meetings and gatherings
Online Platforms for Social Interaction	WhatsApp	WhatsApp	WhatsApp	WhatsApp	WhatsApp
Preferred Entertainment Type	Sports and Games	Social Media	Social Media	Sports and Games	Social Media
Platform with Paid Subscription	None	Others	None	Others	None
Study Hours Daily	1-2	3-4	3-4	3-4	1-2
C.GPA	3.51-4.0	3.1-3.50	3.1-3.50	3.1-3.50	3.1-3.50

The K-modes Clustering showed five different student segments with unique behavioral and academic profiles.

Cluster 1 is predominantly make students from the University of Shangla, characterized by minimal online course engagement, occasional internet consultation, and a preference for video lectures as study materials. Their daily social media usage is 1-2 hours (mainly Facebook). They prefer physical social gatherings and sports and games for entertainment and have high academic performance (C.GPA in range 3.51-4.0). Cluster 2 is also male dominated, mainly from University of Shiringal, and shows more active online academic engagement, with majority taking courses in the range 1-2 and higher daily study hours (3-4). These students prefer both traditional (Reading) and video based study materials, spend more time on social media (3-4 hours, mostly on Facebook), and show balanced preference for both online and physical interactions. Their entertainment choice leans toward social media and they often have access to paid content platforms.

Cluster 3 is a unique one being the only female majority group, primarily from University of Swat. These students take no online courses but show similar study patterns to cluster 2 in terms of daily study hours, hours spent on social media and study materials (both reading and video based materials). They engage with other social platforms other than Facebook and Whatsapp, prefer physical interactions, and spend 3-4 hours daily on social media, with social media based entertainment dominating their leisure time. Cluster 4, male students from University of Shiringal, are among the most digitally engaged, with frequent internet consultations and mostly taking online courses in range 1-2. They maintain balanced study material preferences, dedicate 3-4 hours to studying, and engage heavily with Instagram. Their entertainment preferences shifts toward sports and games, and they are more likely to use platforms with paid subscriptions. Finally, Cluster 5 contains students from the University of Malakand who exhibit relatively low academic engagement meaning does not prefer online courses, 1-2 hours of study daily, and a preference for video lectures. They predominantly use Facebook, spend less time on social media (1-2 hours daily), and prefer social media based entertainment without paid platform usage.

Overall, the clustering highlights distinct patterns in gender, institutional background, study habits, online learning engagement, and social preferences. The differences between clusters suggest that institutional affiliation and gender are closely tied to variations in academic behavior, digital engagement, and entertainment preferences, with potential implications for targeted interventions to improve learning outcomes and engagement.

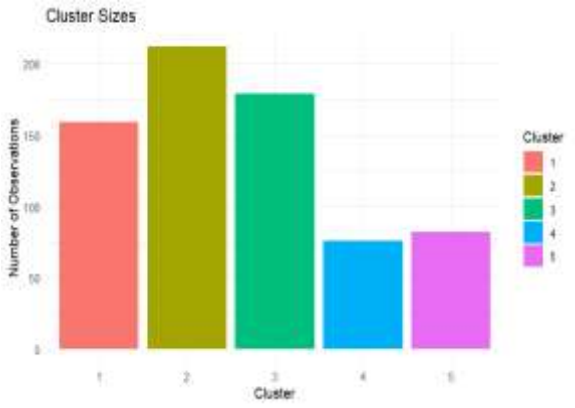

Figure 10: Sizes of the Clusters

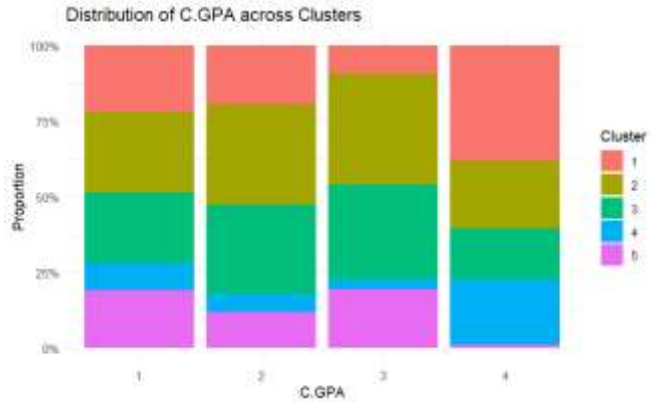

Figure 11: Distribution of Gender Across Cluster

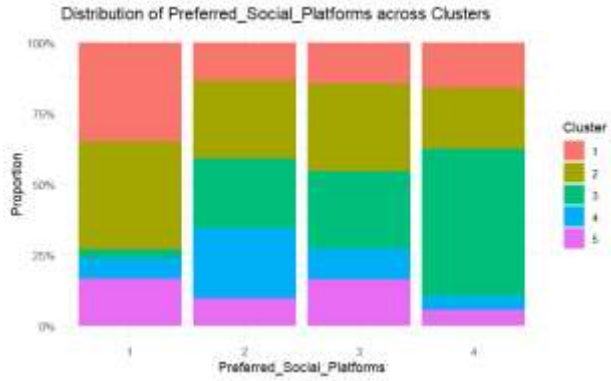

Figure 13: Distribution of Preferred Social Platforms Across Cluster

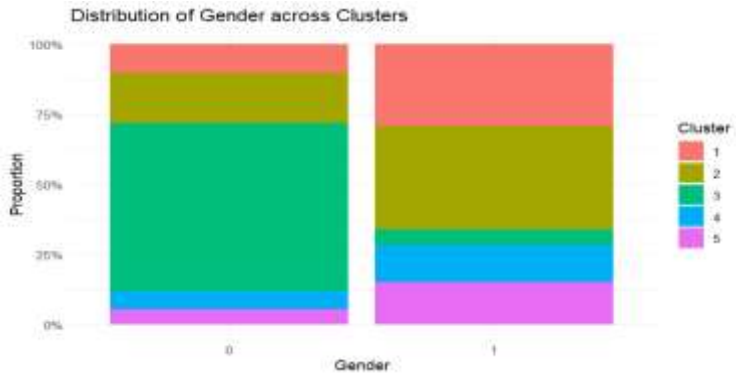

Figure 12: Distribution of CGPA Across Cluster

Fig 10 shows the sizes of different clusters. Cluster 2 is largest one containing 212 instances while Cluster 4 is the smallest one containing only 76 instances. Cluster 1 contains 159 instances, Cluster 3 contains 179 instances and Cluster 5 contains 82 observations.

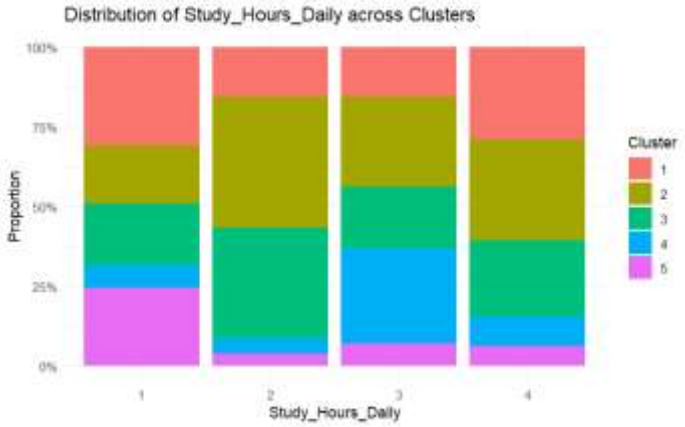

Figure 14: Distribution of Study Hours Daily Across Cluster

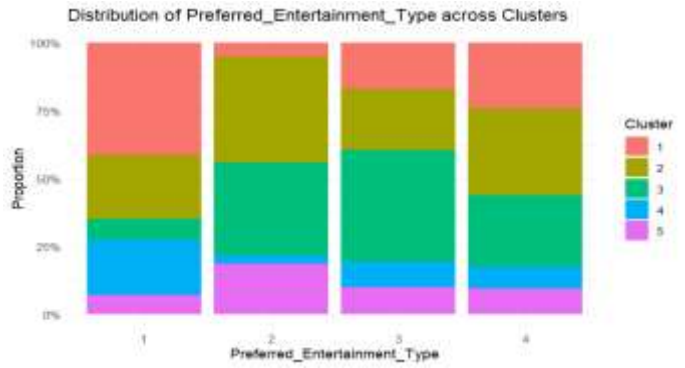

Figure 15: Distribution of Preferred Entertainment type Across Cluster

Fig 11-15 show the distribution of some of the important variables (gender, CGPA, preferred social platforms, study hours daily, preferred entertainment type) across the five clusters.
The x-axes contain the categories of the respective variable.

i. Gender (**0**: female, **1**: male)
ii. CGPA (**1**: 2.0-2.50, **2**: 2.51-3.0, **3**: 3.1-3.50, **4**: 3.51-4.0)
iii. Preferred social platforms (**1**: Facebook, **2**: Instagram, **3**: Twitter, **4**: Others)
iv. Study hours daily (**1**: 1-2, **2**: 3-4, **3**: 4-5, **4**: More than 6)
v. Preferred entertainment type (**1**: Sports and games, **2**: Social media, **3**: Doing hobby, **4**: Extra learning)

Association Rule Mining within K-Modes Clusters: Summary & Insights
Association rules for each cluster were extracted with parameters:

i. Minimum Support: 10%
ii. Minimum Confidence: 60%

The total dataset was divided into 5 clusters using K-Modes. Below are the top rules for each cluster along with interpretations.

Cluster 1
Sample size: 159 observations
Top rules:

Rule	Support	Confidence	Count	Interpretation
GPA = 3.51-4.0	0.604	0.604	96	Majority of students in this cluster had a perfect GPA
Social Media Hours = 1-2	0.616	0.616	98	Most spent minimal time on social media
Online Learning Platform = Others	0.748	0.748	119	Heavy preference for platforms other than Coursera, CodeaAcademy and Khan Academy
Social Interactions = Physical meetings and gatherings	0.792	0.792	126	Most students preferred face to face social interactions
Gender = Male	0.836	0.836	133	Strongly skewed towards male group

Implication: High-performing, low social media usage, gender-biased cluster with a dominant preference for platforms category "Others"

Cluster 2
Sample size: 212 observations
Top rules:

Rule	Support	Confidence	Count	Interpretation
Social Interactions = Physical meetings and gatherings	0.741	0.741	157	Preference for face to face interaction
Gender = Male	0.783	0.783	166	Predominantly male gender.
Preferred Study Materials = Reading	0.118	1.00	25	Reading materials strongly associated with this cluster.
Online Courses Taken = None → Gender = Male	0.108	0.885	23	No online courses tied to gender.

Implication: Male dominated cluster having strong preference for reading materials and face to face interactions with no preference for online courses.

Cluster 3
Sample size: 179 observations
Top rules:

Rule	Support	Confidence	Count	Interpretation
Preferred Social Platform = 4	0.654	0.654	117	High reliance on platforms other than Facebook, Instagram and Twitter
Social Interactions = Physical meetings and gatherings	0.793	0.793	142	Majority interact face to face
Gender = Female	0.855	0.855	153	Gender majority is reversed compared to cluster 2.
Preferred Entertainment Type = Sports and games	0.101	1.000	18	Entertainment preference for sports and games uniquely defines this cluster.

Implication: Strong identity cluster—social platform, gender, and entertainment preference align closely.

Cluster 4
Sample size: 76 observations
Top rules:

Rule	Support	Confidence	Count	Interpretation
Preferred Entertainment Type = Sports and games	0.618	0.618	47	Majority enjoy a outdoor sports rather than spending time on social or other entertainment types.
Online Learning Platform = Others	0.711	0.711	54	Clear platform preference for other platforms
GPA = 3.51-4.0	0.711	0.711	54	High GPA cluster again.
Gender = Male	0.789	0.789	60	Strong gender skew.

Implication: Small but high-performing cluster with distinct entertainment and gender preference.

Cluster 5
Sample size: 82 observations
Top rules:

Rule	Support	Confidence	Count	Interpretation
Online Learning Platform = Others	0.610	0.610	50	Preference for other learning platforms
Paid Subscription Platform = none	0.622	0.622	51	Many does not have any paid subscription
GPA = 3.1-3.50	**0.634**	**0.634**	**52**	**Slightly lower GPA distribution.**
Study Hours Daily = 1-2	0.707	0.707	58	Most study only 1-2 hour daily.

Implication: A moderate-GPA cluster with minimal study hours, having no interest in paid entertainment platforms.

Clustering Method Validation (Elbow Method)

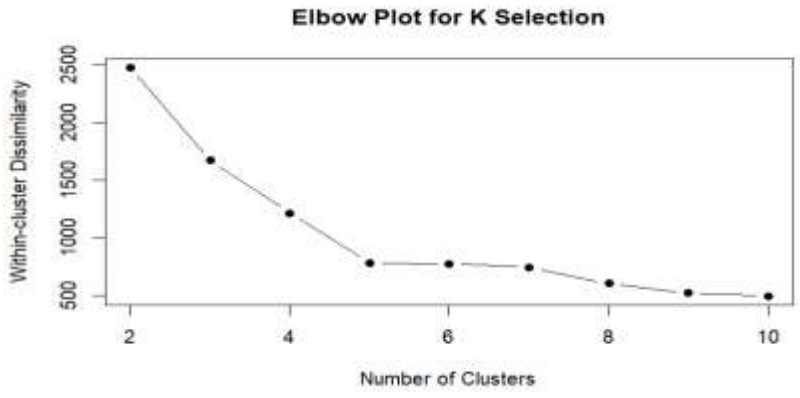

Figure 16: Elbow Plot showing optimum number of clusters

The plot in fig 16 shows an elbow plot which is used to identify the optimum number of clusters for K-modes clustering. The elbow plot shows the relation between the number of clusters and the within-cluster dissimilarity, which measures how close the data points within each cluster are. As the number of clusters increase 2 to 5, there is steep decline in the dissimilarity. After that the slope of the line flattens showing no significant decrease occurs after 5. Which suggest that, the optimum number of cluster is 5, minimizing the within cluster dissimilarity and avoiding unnecessary complexity.

References

[1]. Al Mosharrafa, R., Akther, T., & Siddique, F. K. (2024). Impact of social media usage on academic performance of university students: Mediating role of mental health under a cross-sectional study in Bangladesh. Health Science Reports, 7(1), Article e1788. https://doi.org/10.1002/hsr2.1788

[2]. Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. International Journal of Educational Technology in Higher Education, 17, Article 3. https://doi.org/10.1186/s41239-020-0177-7

[3]. Asim, M., Raza, I., Rauf, H. T., Ikram, M., Majid, A., & Shoaib, M. (2024). SAPPNet: A spatial attention-based predictor for enhanced student performance prediction in online education. Scientific Reports, 14, Article 75242. https://doi.org/10.1038/s41598-024-75242-2

[4]. Chandrasena, P. P. C. M., & Ilankoon, I. M. P. S. (2022). The impact of social media on academic performance and interpersonal relations among health sciences undergraduates. Journal of Education and Health Promotion, 11, Article 117. https://doi.org/10.4103/jehp.jehp_603_21

[5]. Connolly, C. (n.d.). The impact of social media on student well-being and academic performance. Medium. https://medium.com/@ciaranpconnolly/the-impact-of-social-media-on-student-well-being-and-academic-performance-f7941a468992

[6]. Cuong, T. V., Khai, N. T., Oo, T. Z., & Józsa, K. (2025). The impact of social media use motives on students' GPA: The mediating role of daily time usage. Education Sciences, 15(3), 317. https://doi.org/10.3390/educsci15030317

[7]. Han, H. (2023). Fuzzy clustering algorithm for university students' psychological fitness and performance detection. Heliyon, 9(8), Article e18550. https://doi.org/10.1016/j.heliyon.2023.e18550

[8]. Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: A systematic literature review. In Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (pp. 175-199). Association for Computing Machinery. https://doi.org/10.1145/3293881.3295783

[9]. Hosen, M. B., Ahmed, S., Akter, B., & Anannya, M. O. (2025). Impact, causation and prediction of socio-academic and economic factors in exam-centric student evaluation measures using machine learning and causal analysis [Preprint]. arXiv. https://arxiv.org/abs/2506.12030

[10]. Kassarnig, V., Bjerre-Nielsen, A., Mones, E., Lehmann, S., & Lassen, D. D. (2018). Academic performance and behavioral patterns. EPJ Data Science, 7, Article 10. https://doi.org/10.1140/epjds/s13688-018-0138-8

[11]. Oguguo, B. C. E., Ajuonuma, J. O., Azubuike, R., Ene, C. U., & Atta, F. O. (2020). Influence of social media on students' academic achievement. International Journal of Evaluation and Research in Education, 9(4), 1000-1009. https://doi.org/10.11591/ijere.v9i4.20638

[12]. Talib, N. I. M., Majid, N. A. A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using K-means clustering and support vector machine. Applied Sciences, 13(5), Article 3267. https://doi.org/10.3390/app13053267

[13]. Cao, Y., Gao, J., Lian, D., Rong, Z., Shi, J., Wang, Q., Wu, Y., Yao, H., & Zhou, T. (2017). *Orderness predicts academic performance: Behavioral analysis on campus lifestyle* [Preprint]. arXiv. https://arxiv.org/abs/1704.04103

[14]. Dinh, T., Hauchi, W., Fournier-Viger, P., Lisik, D., Ha, M.-Q., Dam, H.-C., & Huynh, V.-N. (2024). Categorical data clustering: 25 years beyond K-modes [Preprint]. arXiv. https://arxiv.org/abs/2408.17244

[15]. Dorman, K. S., & Maitra, R. (2020). An efficient K-modes algorithm for clustering categorical datasets (OTQT) [Preprint]. arXiv. https://arxiv.org/abs/2006.03936

[16]. Murty, M. R. (2013). Cluster analysis on different data sets using K-modes and K-prototype algorithms. *ResearchGate*. https://www.researchgate.net/publication/259174077

[17]. Sharma, N., & Gaud, N. (2015). K-modes clustering algorithm for categorical data. *International Journal of Computer Applications, 127*(17), 1–6. https://doi.org/10.5120/ijca2015906708

[18]. Wikipedia. (2025). Pearson's chi-square test. In *Wikipedia, the free encyclopedia*. Retrieved August 17, 2025, from https://en.wikipedia.org/wiki/Chi-squared_test