Data Description and Exploratory Analysis

The Student Stress Level dataset, sourced from kaggle, the dataset examines student stress levels and factors that may influence them. It includes 21variables and 1,100 observations collected from students aged 15 to 24 in Dharan, Nepal. The data was gathered through a survey conducted between June 2022 and October 2022, aiming to provide insights into the daily lives and stressors of students, .

The dataset consists of 21 integer variables with no missing values. The response variable, stress_level, has three categories (0, 1, and 2), representing different levels of stress among students, and it is well-balanced with approximately 33.91% labeled as 0, 32.55% as 1, and 33.55% as 2. Most variables range from 1 to 5, where 0–1 indicates low levels, 2–3 represents moderate levels, and 4–5 indicates high levels. However, some have distinct ranges, such as anxiety (0–21), self_esteem (0–30), and depression (0–27). The variable capturing mental_health_history is binary (0 or 1), and blood_pressure is categorized into three levels (1 = low, 2 = normal, 3 = high). The histograms show that many variables cluster around moderate or higher values, though a few are skewed toward the lower end.

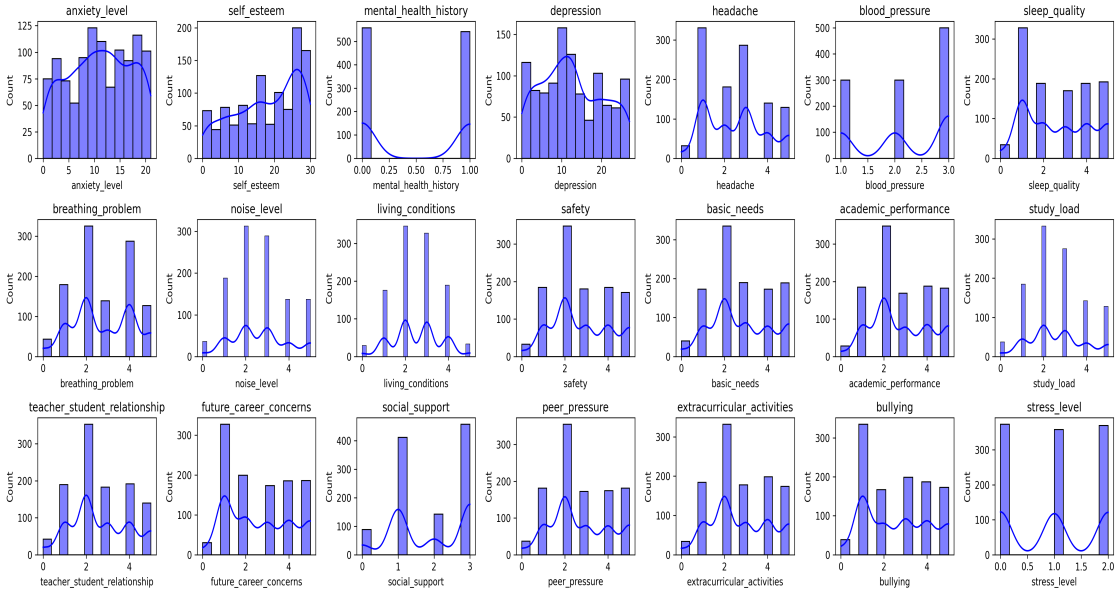

Figure 1: Distribution of all Variables

The correlation analysis shows that most variables are strongly associated (±0.57 to ±0.76) with the target variable, stress_level, while blood_pressure has a weaker correlation (0.39). For instance, self_esteem and social_support exhibit notably negative correlations with stress, suggesting that students with higher self-esteem or stronger support networks tend to experience less stress. Furthermore, many pairs of variables demonstrate moderate to high associations, such as bullying with anxiety, depression, peer_pressure, and safety.

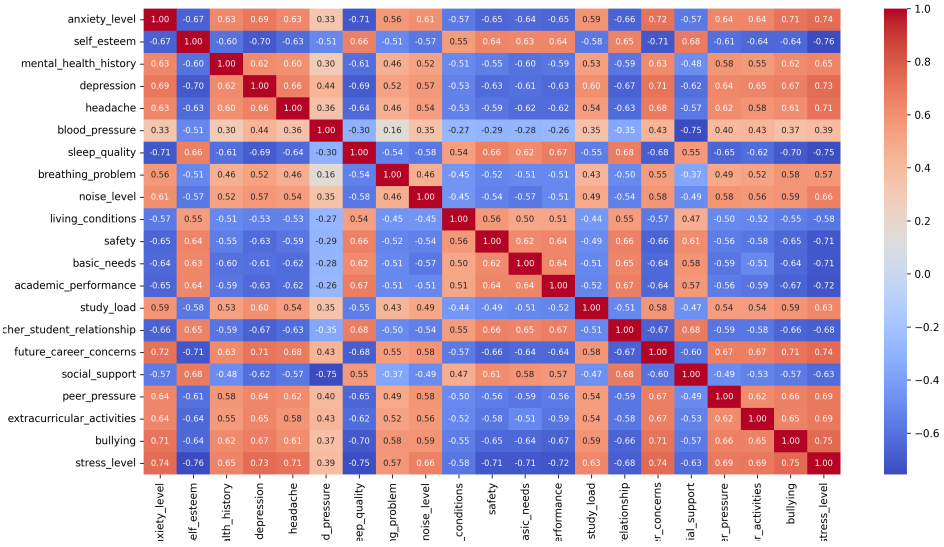

Figure 2: Correlation of all Variables

To keep the visualizations concise and focused, I selected the eight predictors with the strongest correlations to the target variable, stress_level. Including all 21 variables would make the pairplot overly complex and difficult to interpret. By narrowing down to these key predictors, we can better highlight their relationships with one another and their overall influence on stress levels. In the pair plot (Fighure 3), three variables, depression, anxiety_level, and self_esteem, show clear separation among the different stress categories. Students who report higher depression or anxiety_level generally appear in the high-stress category, while those with stronger self_esteem are more often in the low-stress category. These findings suggest that mental health factors, particularly anxiety, depression, and self_esteem, play a substantial role in determining student stress levels.

Figure 3: Pair Plot of Highly correlated Variables

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) identifies directions in the data that capture the most variation by examining the eigenvalues and eigenvectors of the covariance matrix Σ. Suppose $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of Σ, and $v_1, v_2, \ldots, v_p$ are the corresponding eigenvectors. The i-th principal component is defined as $W_i = v_i'(X - \mu)$, where X is the data vector and μ is its mean. The first principal component, W_1, captures the greatest possible variance in the data, while each subsequent component captures the next largest portion of variance under the constraint that it is orthogonal to the previous components. The fraction of the total variance explained by the i-th component is $\lambda_i / \mathrm{tr}(\Sigma)$, and keeping the first r components captures $\sum_{i=1}^{r} \lambda_i / \mathrm{tr}(\Sigma)$ of the total variance, guiding the choice of how many components to retain.

The elbow plot indicates that beyond two principal components, the increase in explained variance slows considerably, suggesting that two components might be sufficient. However, those two components account for only 65% of the total variance. To capture at least 80% of the total variance, eight principal components were selected, collectively explaining about 82% of the total variance.

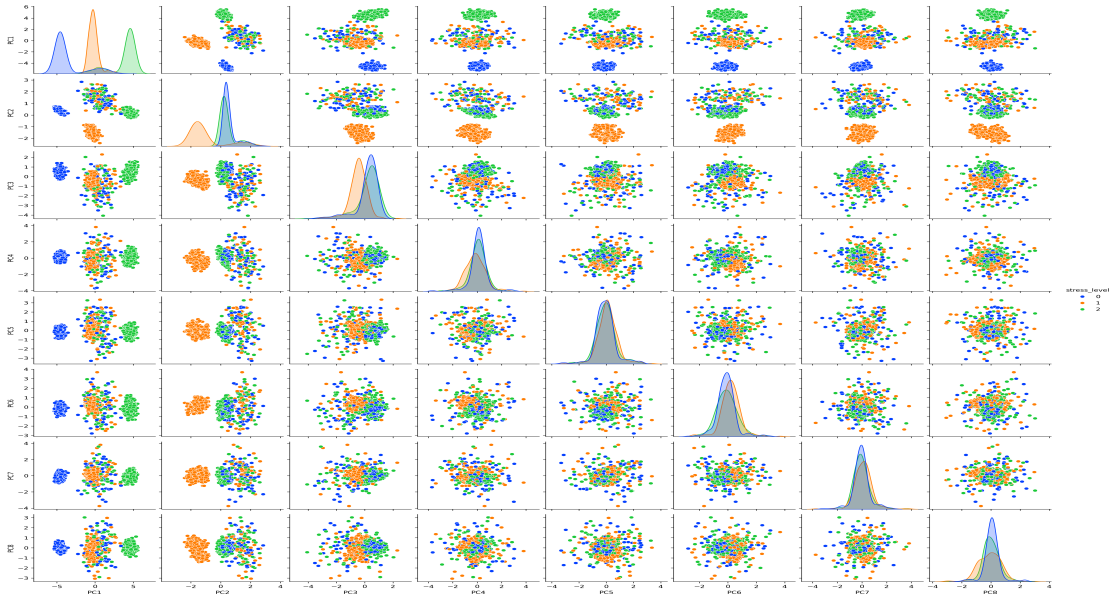

Figure 4: Pair Plot of Principal Components

Factor Analysis (FA)

In factor analysis, each p-dimensional observation $\mathbf{X}_i$ is modeled as

$$\mathbf{X}_i = \boldsymbol{\mu} + \Lambda \mathbf{U}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{U}_i$ is a q-dimensional latent factor vector and $\boldsymbol{\epsilon}_i$ is noise. Both $\mathbf{U}_i$ and $\boldsymbol{\epsilon}_i$ are assumed to be normally distributed with mean zero and covariance matrices $\mathbf{I}_q$ and $\boldsymbol{\Psi}$, respectively. As a result, $\mathbf{X}_i$ follows a normal distribution with mean $\boldsymbol{\mu}$ and covariance $\Lambda\Lambda' + \boldsymbol{\Psi}$. Because $q < p$, factor analysis effectively reduces the dimensionality of the data by capturing most of its variability in fewer latent factors. The model parameters $\boldsymbol{\mu}$, Λ, and $\boldsymbol{\Psi}$ are typically estimated via maximum likelihood, often using the expectation-maximization (EM) algorithm, which treats the latent factors $\mathbf{U}_i$ as "hidden" variables when calculating the complete-data log-likelihood.

When applying factor analysis, the Kaiser criterion (eigenvalues > 1) suggested two factors (Factor 1 = 11.91, Factor 2 = 1.20). However, an examination of explained variance indicated that the first four factors together accounted for 50% of the data variability, so four factors were ultimately retained. The choice of rotation method was guided by predefined domains, Psychological, Physiological, Social, Environmental, and Academic. Comparing Varimax and Promax showed that Promax aligned the four retained factors more closely with domain knowledge and better reflected real-world scenarios in which the factors are somewhat correlated.

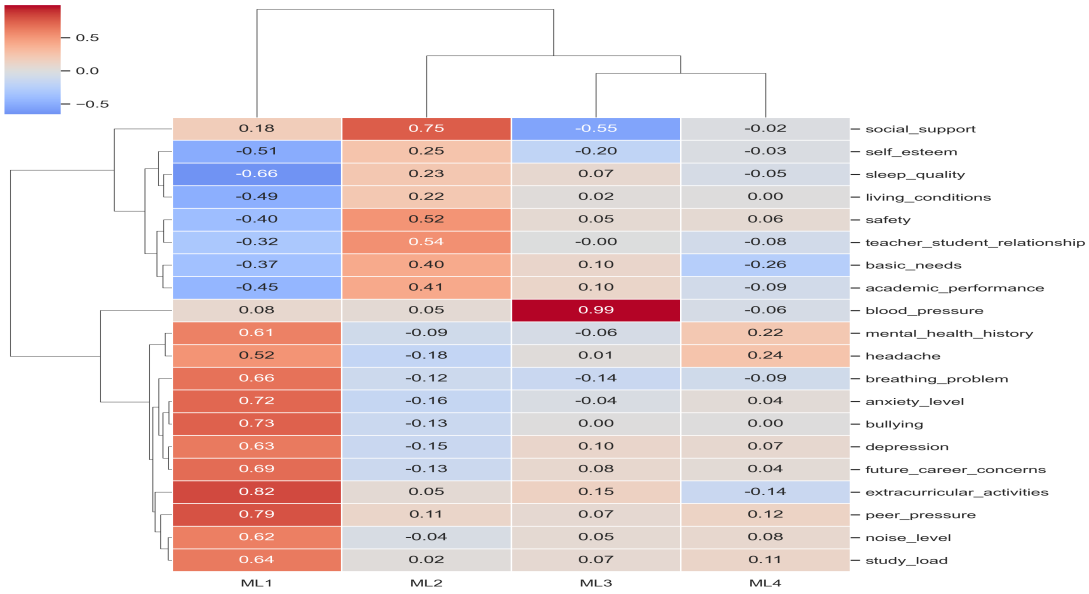

Figure 5: Factor Analysis Heatmap

KMeans Clustering Models

Both models were built using the k-means clustering algorithm, with the number of clusters set to three to align with the original target categories. In the first model, all 20 original variables served as predictors, while in the second model, the data were first reduced to eight principal components via PCA, and those components were then used as predictors. Despite the dimensionality reduction, both models achieved the same misclassification rate (0.13) and adjusted rand index (0.63), indicating that using eight principal components retained enough of the data's structure to match the performance of the full 20-variable model.

Model	Predictor	Cluster	MCR	ARI
Kmean_orig	20 variables	3	0.13	0.63
Kmean_pca	8 principal components	3	0.13	0.63

Table 1: Clustering Results: Comparison of K-means Models