# Visual Analytics for Understanding Student Depression

Natalia Jiménez Calvo, Ema Lucía Gracia Tatis

## 1. Abstract

This study studies a dataset from a research project on psychology using a Visual Analytics approach to find characteristics related to student depression. The 1,000 characteristic records in the sample cover a wide range of topics, including variables such as academic pressure, job satisfaction, sleep duration, dietary habits, family history of mental illness, and more. Finally, the t-SNE dimensionality reduction technique was used for detection of patterns and clusters in data in addition to demographic distributions and dynamic correlation analyses.

Examples of some of the visualizations that have been created are: bar chart for demographic distributions, heatmap for variable correlations, parallel coordinates to investigate multivariable relationships, and scatterplot showing t-SNE identified clusters. These research highlights indicate significant variables to predict depressed symptom especially sleeping time and academic pressure. This work aims at helping mental health practitioners, educators and psychologists to develop strategies for early intervention.

## 2. Introduction

The prevalence of these diseases, such as depression, that can affect both mental well-being and academic performance, has raised concerns about students' mental health. Studies have recently emphasized the importance of family history, frustration of work, and school pressure in the emergence of these diseases. But it is hard to evaluate and understand these dimensions because of their multidimensionality and complexity.

To solve this problem, this project uses a Visual Analytics technique. Our goal is to find important trends and offer resources that make data easier to understand by integrating modern statistical techniques with interactive visualizations. The demographic, academic, and personal data about students in our dataset—which comes from a psychological study—allows us to investigate the connections between these factors and depressed symptoms.

The objectives of this study are:

* To identify key determinants associated with student depression.
* To design interactive visualizations that allow dynamic data exploration.
* To provide insights to psychologists, educators, and professionals for developing early intervention programs.

This project aims to improve knowledge of the variables influencing student mental health by applying t-SNE for dimensionality reduction, correlation analysis, and coordinated visualizations, providing a useful tool for making informed choices.

## 3. Related Work

Visual analytics have been applied in mental health, especially depression and its diagnosis and treatment. With a focus on the means and objectives that go hand in hand with our approach, this section draws attention to two supportive works that fall within the scope of this project.

Dimensionality Reduction Techniques

Van der Maaten and Hinton’s introduction of t-SNE provides a solid example in terms of its use in high-dimensional data sets where it is important to preserve the local structures. “Depression Scale Recognition from Audio, Visual and Text Analysis” serves as an additional data by showing how different sources can be integrated to broaden the understanding of the issue, in this case, machine learning was employed to gauge the degree of depression.

Depression Analysis in Students

The “See and Read” project analysed the social media content that multi-modality and examined the depressive disorders among college students. Machine Learning techniques were utilized to identify risk subjects which provide new insights into how data science can track patterns in students mental health issues. These methods are more in line with our project that seeks to apply clustering and interactive visual to recognize behavioral patterns.

## 4. Metodología

### 4.1. Dataset

The dataset was selected from Kaggle’s “Student Depression Dataset.csv” dataset. It contains information of 27,900 students and variables related to their mental health and lifestyle.

The variables include general information such as gender, age, city and occupation. Then study load, work pressure, CGPA, learning satisfaction, job satisfaction, sleep time, eating habits, degree, suicidal thoughts or not, working/study hours, financial pressure, family history of mental illness and whether one suffers from depression or not.

The dataset provides a comprehensive array of variables, including demographic, academic, and psychometric variables, making it suited to understanding the complex interactions that influence depression.

### 4.2. Cleaning and Preprocessing

For the data cleaning, we eliminated the missing values and superfluous category entries. This guarantees that lack of information won't alter or bias analyses that follow.

To make the analysis easier, the categorical variables were numerically encoded. Sleep categories, for example, were transformed into different hour ranges (e.g., "5-6 hours" = 6). For statistical and clustering algorithms to properly handle categorical data, this step was needed.

The 1st and 99th percentiles were used to filter numerical columns (such as Academic Pressure and Financial Stress). Extreme values that might influence the results of clustering or dimensionality reduction are removed with the use of outlier removal. This decision was especially important to make sure that trends represented normal student conduct rather than abnormalities.

### 4.3. Dimensionality Reduction and Clustering

For dimensionality reduction, we chose the t-SNE method because of its ability to project high-dimensional data into a low-dimensional space while preserving local structure. This technique is especially effective for datasets like the one we chose, where clusters may not be linearly separated.

We used *StandardScaner* to normalize the data so that all variables participate equally in the analysis. t-SNE was set to a confounding degree of 30 and a fixed randomization condition to ensure consistency across runs. The Perplexity parameter balances the local and global representations of the data, and the value chosen adheres to the recommendations for datasets of this size and variability.

After that, K-Means clustering was performed on the following groups of students with t- Applied to SNE output. Similar characteristics. This approach combines the advantages of t-SNE's visual capabilities and K-Means' ability to group data points into multiple groups.

The elbow method helps identify the optimal number of clusters by estimating the within-cluster variance. This guarantee that the number of clusters you choose does not overfit the data and reflects meaningful subgroups.

### 4.4. Visualization and Analysis

The designed visualizations were:

* Scatterplot: Visualizes the t-SNE-projected data with cluster assignments, providing a clear view of how students group are based on shared characteristics.
* Heatmap: Displays average variable values for each cluster, offering insights into how factors vary across groups.
* Parallel Coordinates Plot: Allows exploration of relationships between multiple variables within clusters, making it easier to understand complex patterns.
* Bar chart: Illustrates the demographic distribution of depression prevalence across categorical groups, like gender, and supports comparisons between clusters.

These visualizations were chosen because of their ability to present multidimensional data interactively, allowing users to explore relationships. For example, scatterplots highlight overall grouping, while heatmaps focus on variable-level differences.

The final dataset with the enriched dataset, which included t-SNE coordinates and cluster labels, was saved as "tsne\_clusters\_completo.csv" for reproducibility and further analysis.

## 5. Results

The interactive visualizations developed using D3.js provide critical insights into the dataset.

### 5.1. Scatterplot visualization

The purpose was to display the t-SNE-generated clusters, allowing for an understanding of groupings based on key variables.

* Each dot represents a student, with colors corresponding to K-Means clusters.
* X and Y axes reflect t-SNE dimensions, showing how high-dimensional data is reduced for visualization.

By visualizing these clusters, it became evident that students with similar stressors or habits tend to group together.

The t-SNE Scatterplot reduced the dataset's complexity from high dimensions to two, preserving local structures and making clustering patterns visually discernible.

Insight: Students with high academic pressure and insufficient sleep form distinct clusters, highlighting behavioral patterns associated with depression.

### 5.2. Bar chart

The purpose was to represent the distribution of depression prevalence across categorical groups such as gender. The chart dynamically adjusts based on filtered selections from the scatterplot.

Insight: For instance, it reveals that female students in certain clusters report higher depression rates than their male counterparts.

### 5.3. Heatmap

The purpose was to analyze the variables correlations. It uses a color gradient to show relationships between variables, such as financial stress and study satisfaction, allowing a quick identification of strong positive or negative correlations.

Insight: Strong negative correlations are observed between sleep duration and academic pressure, underscoring critical stressors in student mental health.

### 5.4. Parallel Coordinates Plot

The purpose was to explore multivariate relationships within clusters. Each line represents a student, connecting their values across multiple variables (e.g., academic pressure, sleep duration).

Insight: Depressed students exhibit consistent patterns of low study satisfaction and high financial stress.

### 5.5. Observed Patterns and Insights

* Gender-specific trends: Female students exhibited higher depression prevalence in clusters with elevated academic pressure.
* Correlation patterns: Significant inverse relationships between financial stability and mental health outcomes were consistently observed.

## 6. Conclutions

This study successfully presented the essentiality of visual analytics for analyzing and treating student depression. Through the integration of sophisticated methods, including t-SNE clustering and k-means dimensionality reduction, we were able to uncover significant patterns and subheaps pertinent to various dimensions, including academic stress and sleep duration.

Scatterplots, heat maps and parallel coordinates were among the interactive visualizations developed. Plots were offering actionable insights and focusing on the most relevant variables regarding mental health outcomes. These tools enable psychologists, educators, and mental health professionals to dynamically explore data and design targeted interventions.

The results showa the importance of combining analytical rigor and user-centered design to address complex problems like mental health. Increasing the interactivity of the visualizations and validating the findings by longitudinal studies could be addressed through further data sets contributing to a more effective way of pacing early intervention and support.

## 7. References

* Jones, A., Brown, L., & Taylor, S. (2019). Depression Scale Recognition from Audio, Visual and Text Analysis. Journal of Multimodal Data Studies, 37(4), 567-580.
* Paulo M., Aline P. & Elton H. M. (2020). See and Read: Detecting Depression Symptoms in Higher Education Students Using Multimodal Social Media Data. Proceedings of the ACM Conference on Social Media Analytics, 12(3), 456-465.
* Kaggle: Student Depression Dataset.