Investigating student depression in India using Bayesian Networks

Lorenzo Pellegrino, Giorgio Scavello
Master's Degree in Artificial Intelligence, University of Bologna
{lorenzo.pellegrino2, giorgio.scavello}@studio.unibo.it

April 8, 2025

Abstract

The goal of this mini-project is to investigate the possible causes of student depression in India starting from previous works on the same topic ((Deb et al. 2016), (Liu and Wang 2024)), but using different methods which could lead to different results. As the variables in the chosen dataset showed significant signs of correlation, we thought it was ideal to model the data using Bayesian Networks.
The result of the study were as expected except for the following: a latent variable is not necessary to model simple tasks as this one, while it would be more interesting for more complex models; in our dataset depressed students showed a probability of disrupted sleep similar to the one of students which are not depressed, contrarily to what expected.

Introduction

Domain

Depression is the most common mental disorder, affecting over 300 million people globally ((WHO) and others 2017). Students are particularly vulnerable due to academic stress and life transitions, increasing the risk of self-harm and suicide. In India, student suicides rose by 70% from 2011 to 2021 (Maji et al. 2025). This issue can be analyzed using a Bayesian Network, drawing from similar studies (Liu and Wang 2024), despite their use of customer satisfaction methods.

Aim

- investigate the effects of introducing a latent variable in a Bayesian Network;
- compare different methods for fitting Bayesian Networks present in the pgmpy library;
- answer the hypothesis inspired by other papers on similar topics ((Liu and Wang 2024), (Deb et al. 2016)).

Method

Pgmpy (Ankan and Textor 2024) was used to implement Bayesian Networks and execute queries. The Tschuprow index and χ^2 tests identified variable relationships, though uncertainty in the Depression–Study Satisfaction link led to fitting two models for each type (baseline and latent). Model selection relied on the BIC score to penalize complexity and prevent overfitting. Additional criteria, such as average d-separation, sparsity score, and Depression classification accuracy, also informed the best-fitting model.

Results

The baseline models were likely superior due to their lower fitting times while maintaining similar metric and query performance. Thus, a MAP estimator is preferable to Expectation-Maximization, given the latter's high computational cost. Two queries designed to verify symptoms of sleep disorders and low concentration (WHO 2023) yielded unexpected results.

Model

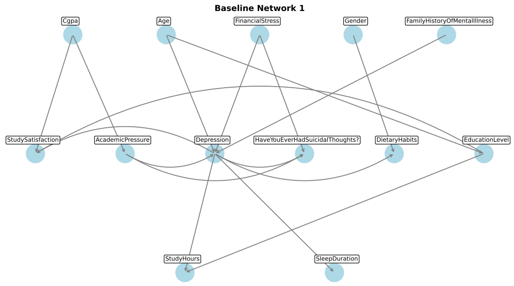

Figure 1: Baseline model

Most nodes were self-explanatory, but some continuous variables (age (UN 1982), CGPA (Commission and others 2018), education level, and study hours) were binned following guidelines when available to simplify Bayesian Network modeling. Certain variables (gender, suicidal thoughts, family history of mental illness, and depression) were encoded as binary, while others were divided into 4–6 bins to preserve data characteristics.
Conditional distributions were estimated using Maximum Likelihood Estimation (Koller and Friedman 2009). A MAP estimator with an uninformative prior was typically preferred to handle sparse categorical data, though some models exhibited overfitting due to low sample counts in certain categories. For models with the latent variable 'Stress',

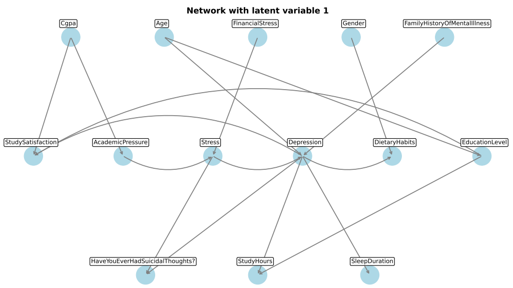

Figure 2: Model using the latent variable ‘Stress’

Expectation-Maximization was used instead, as MAP estimation is infeasible with missing data, leading to longer fitting times.

Model construction was primarily guided by variable correlations. Since all variables were categorical after binning, the Tschuprow index was used to measure relationships, as it accounts for differing bin counts, which could bias the χ^2 test. While correlation values were generally low, relationships were deemed significant using a Bonferroni-corrected χ^2 test with a 0.05 threshold. The direction of relationships was determined based on all the previously cited papers. However, not all significant correlations were directly modeled to prevent overfitting. Instead, the *Markov Blanket* of each node was used to capture indirect connections.

Analysis

Experimental setup

This project aimed to analyze the impact of the introduction of a latent variable by comparing the indices defined in between the baseline model and the latent variable model. Additionally, it compared Expectation-Maximization, pgmpy's only method for fitting latent variable models, with the MAP estimator, evaluating the trade-off between fitting time and model performance.

The queries, inspired by (Deb et al. 2016) and (Liu and Wang 2024), focused on:

- the probability of depression given control variables (Gender, Age);
- the probability of depression given Education Level;
- the probability of depression given high Stress (both models);
- the probability of depression symptoms (Suicidal Thoughts, low and high Study Hours, low or extreme Sleep Duration, unhealthy Dietary Habits, low Study Satisfaction) given depression;

Results

For the query on depression given Age and Gender, the probability changed only with age, remaining constant across genders. This aligns with findings from previous studies (Deb et al. 2016), which indicate that younger students have a higher likelihood of experiencing depression. A similar trend was observed for education level, where students who had completed High School or Undergraduate studies exhibited a higher probability of depression compared to others. This pattern held true in both models.

From the third query on the baseline model, it was evident that students experiencing extreme stress (financial or academic) were more likely to be depressed. The latent variable model also reflected different probabilities for varying stress levels. However, pgmpy's fitting method for latent variables does not inherently order categorical variables, making direct interpretation challenging. As a result, querying the baseline model was necessary, even if computationally more demanding.

An unexpected result emerged regarding sleep disruption: the probability remained the same for both depressed and non-depressed individuals. Another surprising finding was that depressed students had a lower probability of studying for less than two hours compared to non-depressed students. This query aimed to model the probability of low concentration among depressed individuals (WHO 2023), though it may not have been the optimal approach. Depressed students tend to study more than students which are not depressed, this could be expressed by a low concentration also, as students with depression could require more time to perform the same task than their peers. This should be investigated furthermore. The fourth query also confirmed that depressed students were less likely to be satisfied with their studies compared to their non-depressed peers.

Given the similarity in findings between the baseline and latent variable models, the baseline model is preferable in such cases, as it requires significantly less computational effort than the Expectation-Maximization-based latent variable model.

Conclusion

This model provided a broad understanding of the potential causes of student depression among Indian students. Building on previous research, it further explored the likelihood of depression across different student groups during their academic journey. A valuable direction for future research would be to assess the impact of university psychological services and whether they contribute to an improved quality of life for students. The main limitation of the study was the dataset, as it did not contain the total amount of information to investigate every query we were interested in, such as the one on symptoms of depression.

Links to external resources

- *Link to the original dataset*

References

[Ankan and Textor 2024] Ankan, A., and Textor, J. 2024. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research* 25(265):1–8.

[Commission and others 2018] Commission, U. G., et al. 2018. Ugc guidelines on adoption of choice based credit system.

[Deb et al. 2016] Deb, S.; Banu, P. R.; Thomas, S.; Vardhan, R. V.; Rao, P. T.; and Khawaja, N. 2016. Depression among indian university students and its association with perceived university academic environment, living arrangements and personal issues. *Asian journal of psychiatry* 23:108–117.

[Koller and Friedman 2009] Koller, D., and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

[Liu and Wang 2024] Liu, X., and Wang, J. 2024. Depression, anxiety, and student satisfaction with university life among college students: a cross-lagged study. *Humanities and Social Sciences Communications* 11(1):1–11.

[Maji et al. 2025] Maji, S.; Jordan, G.; Bansod, S.; Upadhyay, A.; Deevela, D.; and Biswas, S. 2025. Student suicide in india: An analysis of newspaper articles (2019–2023). *Early Intervention in Psychiatry*.

[UN 1982] UN. 1982. Provisional guidelines on standard international age classifications.

[(WHO) and others 2017] (WHO), W. H. O., et al. 2017. Depression and other common mental disorders: Global health estimates (accessed 11 february 2021 from https://apps. who. int/iris/bitstream/handle/10665/254610. Technical report, WHO-MSD-MER-2017.2-eng. pdf.

[WHO 2023] WHO. 2023. Depressive disorder (depression). `https://www.who.int/news-room/fact-sheets/detail/depression`. Accessed: 2025-03-31.