Student depression in India

Lorenzo Pellegrino
Giorgio Scavello

Main objectives

01

Investigate the behaviour and fitting methods of models with latent variable.

02

Answer the 4 queries formulated by taking inspiration from related literature.

Data preparation

- Root causes
- Concauses
- Symptoms

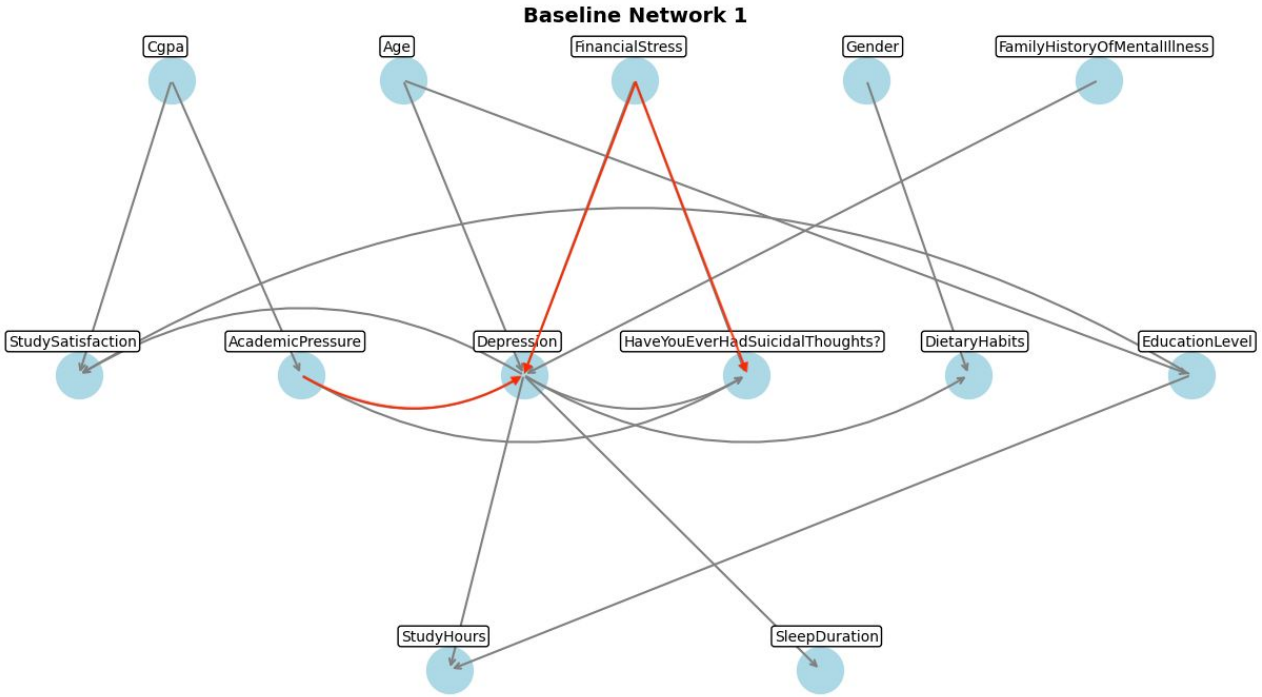
Then we followed a preparation pipeline for adapting the data to pgmpy constraints about Bayesian Networks:

Cleaning ⇒ Enrichment/Specific Issues ⇒ Binning

- Cgpa
- Age
- Financial stress
- Gender
- Family history of mental illness
- Academic pressure
- ***Depression***
- Education level
- Study satisfaction (*)
- Dietary habits
- “Have you ever had suicidal thoughts?“
- Study hours
- Sleep duration

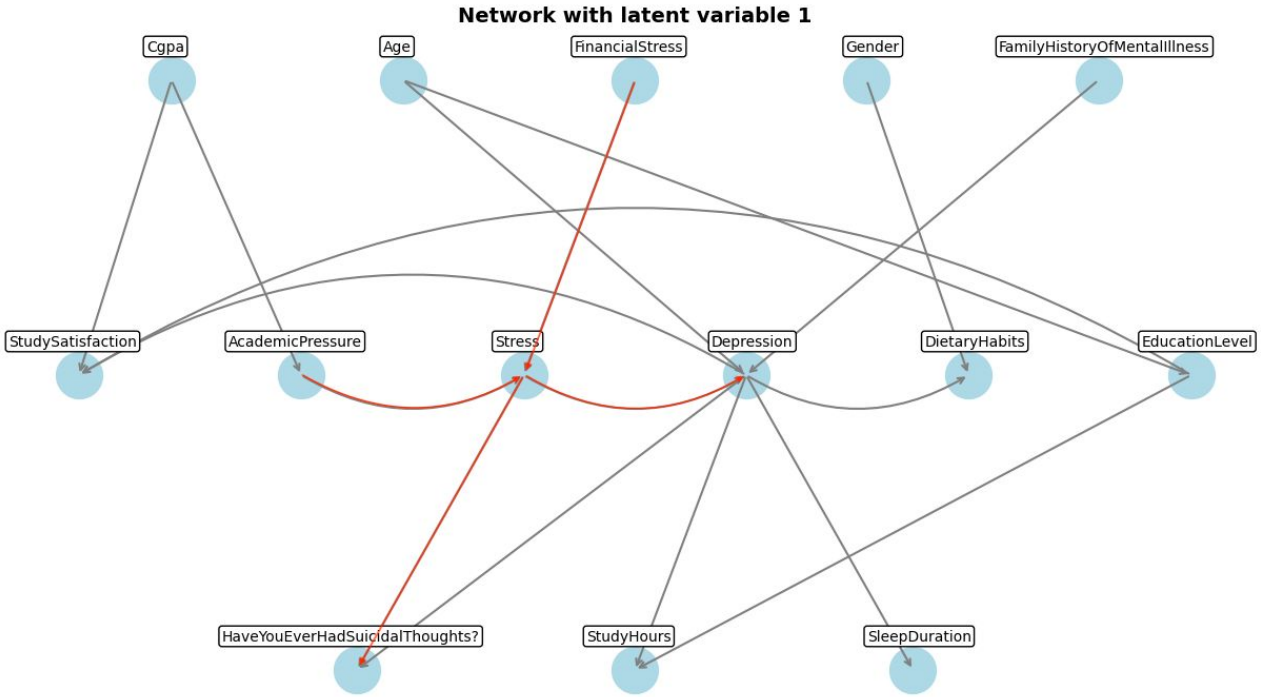
(*) We tried a parallel experiment with the attribute as concause but this one ended up with better results.

Baseline Network



- Constructed with correlation methods, independence tests, maximization of the BIC score and related literature
- Fitted with `BayesianEstimator`

Network with latent variable



- The latent variable *Stress* introduces conditional independencies and increased sparsity
- Fitted with `ExpectationMaximization`

Models’ scores

The models have been evaluated on various scores:

Scores	Baseline Network	Network with latent variable
Number of parameters	710	580
Log-likelihood (*)	-392002.222	-392300.6
BIC score (*)	-395635.378691	-395268.536827
Prediction accuracy on “Depression”	0.843598	0.843419

(*) implemented an approximation for handling latent values

Queries

Probability of depression given control variables

Depression probability varied with Age but remained unaffected by Gender.

Probability of depression given high Stress

In the baseline model, extreme financial and academic stress was linked to higher depression risk. The second model gave similar results, but querying was harder due to unordered latent variable categories in pgmpy.

Probability of depression given Education Level

Students who had completed High School or Undergraduate studies showed a higher probability of depression compared to others.

Probability of depression symptoms given depression

The results aligned with expectations, except for two variables: sleep disruption showed no difference between depressed and non-depressed students, while depressed students were more likely to study less than two hours.

Conclusions

- The project enabled a deeper analysis of results from previously published studies using another model, reinforcing the previously found results.
- We found that while latent variable models can be useful, in simple frameworks like this one, a baseline model is often sufficient to answer most questions.
- A valuable direction for future research would be to evaluate the impact of university psychological services on students' well-being.