Bayesian Belief Network

A look at college student data and the interaction between sleep, missed classes, GPA, depression, anxiety, stress, and happiness.

By Sarah Fullerton, Nethmin Liyanage, Asanti Gemada

Our Data:

- Our goal was to create a Bayesian network using the built in R data set *SleepStudy*.

  College students, STEM majors in particular, have this inevitable journey that has many good parts to it but also many stressful, hard and sleep depriving lengths to make ends meet.

  Our group in particular felt closely related to the idea that the average amount of sleep we as college students get a night can directly affect our missed classes, GPA, and levels of DAS (depression, anxiety, and stress). Depending on how these are affected, these three variables can directly affect our happiness in good or bad ways (lower/higher it).

  That is, we believe that average sleep will indirectly affect happiness levels in college students.

  With this, we created the following Bayesian Belief Network, more commonly known as a Bayesian Network based on data collected about college students.

Setting up our data:

- Here we chose to direct our focus on five of the data set variables to more focus our findings.
- These variables are Average Sleep, GPA, Happiness, DAS (Depression, Anxiety, Stress) Score, and Classes Missed.
- With these variables, we wanted to look at how Average sleep can affect a college student's GPA, Classes Missed, and overall DAS which in turn may lead to what we signify as a "good" amount of Happiness and a "bad" amount of Happiness.

A peek at our data:

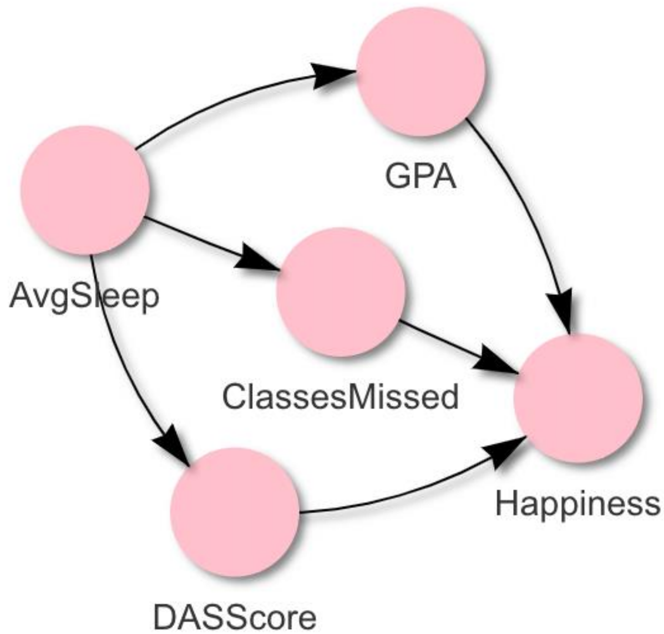
	AverageSleep <dbl>	GPA <dbl>	Happiness <dbl>	DASScore <dbl>	ClassesMissed <dbl>
1	7.18	3.60	28	15	0
2	6.93	3.24	25	4	0
3	5.02	2.97	17	45	12
4	6.90	3.76	32	11	0
5	6.35	3.20	15	46	4
6	9.04	3.50	22	50	0

Our Network:

In R, we used the packages *bnlearn*, *bnstruct*, and *visNetwork* to create a super fun and dynamic Bayesian Belief Network.

This network looks at the directed graph between the node Average Sleep to the three nodes DASScore, Classes Missed, and GPA, all to the node Happiness.

Running this network in R is recommended for the full experience. A static version displays as follows:



Creating our Boolean Variables:

To make observations about our data set, we turned our variables into Boolean variables. To do this, we converted each variable into a Boolean with value 0 or 1 using the information below:

Booleans were designated “low” if values fell *below* the median for that variable. Consequently, booleans were designated “high” if values fell *above* variable median.

Low AverageSleep: (value of 1)

AverageSleep ≤ median(AverageSleep)

i.e. AverageSleep ≤ 8

High GPA: (value of 1)

GPA ≥ median(GPA)

i.e. GPA ≥ 3.3

High DASScore: (value of 1)

DASScore ≥ median(DASSCore)

i.e. DASScore ≥ 27

High ClassesMissed: (value of 1)

ClassesMissed ≥ median(ClassesMissed);

i.e. ClassesMissed ≥ 27

High Happiness: (value of 1)

Happiness ≥ median(Happiness)

I.e. Happiness ≥ 27

Conversely, we assigned values the Boolean variables High AverageSleep, Low GPA, Low DASScore, Low ClassesMissed, and Low Happiness as 0.

Findings:

Using our data and Boolean variables, we calculated corresponding conditional probabilities: This work was done in our R and Excel documents.

Conditional Probability Table based on Happiness:

C,G,D	HA=1	HA = 0	P(H = 1 | col A)	P(H = 0 | col A)
HHH	13	20	0.051383399	0.079051383
LHH	9	14	0.035573123	0.055335968
HLH	17	35	0.067193676	0.138339921
LLH	4	15	0.015810277	0.059288538
HHL	16	9	0.063241107	0.035573123
LHL	19	8	0.075098814	0.031620553
HLL	36	13	0.14229249	0.051383399
LLL	16	9	0.063241107	0.035573123

To draw some conclusions from our data, we looked at situations with our main 3 variables: ClassesMissed (C) , GPA (G) , & DASScore (D) against the factor Happiness (HA). There were nine possible situations for a student to be in, depending on if they had a high (=1) or low (=0) score in one of these categories. For example: HHH is when all three variables are high, or when a student has C = 1, G = 1, or D = 1. These 9 situations had two levels: Happiness = 1 or Happiness = 0. 13 students were under the HHH category and had a Happiness score of 1. The probabilities were calculated dividing that number by the total number of participants, 253 to find the percentage of students in that subcategory.

We found that there were *two situations* with a larger percentage: HLL with a Low happiness score and HLH with a High Happiness score. 14% of the students reported being unhappy while also having a high amount of classes skipped, a low gpa, and high DASScore. This makes sense, as these students are probably under a lot of stress. But another 14% of the students reported being happy despite also having a high amount of classes missed and low gpa. This group of students reported a low DassScore, so we can conclude that is why they *may feel happier* than other students who *also have a lower gpa and skip classes*.

Conditional Probability Table based on Sleep:

C,G,D	AS=1	AS = 0	P(AS = 1 | col A)	P(AS = 0 | col A)
HHH	9	24	0.035573123	0.09486166
LHH	13	10	0.051383399	0.039525692
HLH	26	26	0.102766798	0.102766798
LLH	9	10	0.035573123	0.039525692
HHL	10	15	0.039525692	0.059288538
LHL	12	15	0.04743083	0.059288538
HLL	29	20	0.114624506	0.079051383
LLL	16	9	0.063241107	0.035573123

This table shows the conditional probability that a student has a High or Low sleep score based on the 3 variables ClassesMissed, GPA, and DASScore. We found that *there was an equal amount of students under the HLH category between having a high and low amount of sleep*, So the amount of sleep did not seem to impact students in that category. HLL, which has a high amount of missed classes, low gpa, and low DASScore had the largest percentage of students whose AvgSleep is high.

References

Onyper, S., Thacher, P., Gilbert, J., Gradess, S., "Class Start Times, Sleep, and Academic Performance in College: A Path Analysis," April 2012; 29(3): 318-335.

Scutari M (2010). “Learning Bayesian Networks with the bnlearn R Package.” *Journal of Statistical Software*, **35**(3), 1–22. doi:10.18637/jss.v035.i03.

Franzin A, Sambo F, Camillo Bd (2017). “bnstruct: an R package for Bayesian Network structure learning in the presence of missing data.” *Bioinformatics*, **33**(8), 1250–1252. doi:10.1093/bioinformatics/btw807.

Almende B.V. and Contributors, Thieurmel B (2022). “visNetwork: Network Visualization using ‘vis.js’ Library.” *R package version 2.1.2*. https://cran.r-project.org/package=visNetwork