Report, Bayesian Networks

By: Lieuwe Meijdam, Pratik Kulkarni, Krutika Dhavale

Problem domain

Drug use/consumption is something that has been around for a long time and is a serious problem globally. Drug consumption includes numerous risk factors. A number of factors are correlated with initial drug use including psychological, social, personal, and individual factors. In this project a Bayesian Network will be build focusing on the relationship/ predictability between personality traits and drug usage. using the Revised NEO Five-Factor Inventory (NEO-FFI-R) [1], Barratt Impulsiveness Scale version 11 (BIS-11) [2] and Impulsivity Sensation Seeking scale(ImpSS) [3].

The “Big 5” personality traits are according to psychologists the most adaptable and comprehensive to understand human individual differences [1]. The personality profile of an individual plays an important role in becoming a drug addict. In this project a Social Science area related public Drug consumption (quantified) Data Set that contains records for 1885 respondents collected through an online survey in 2016 [4] will be used. The survey contains information on multiple traits were 10 out of 12 will be used in this project. Next to those traits, the drugs usage level of 18 different drug types were questioned. 4 out of 18 were chosen to include in this project: Caffeine a mild drug that most will not even consider a drug; Cannabis a semi socially accepted relaxation drug; Meth a highly addictive stimulant drug and Semeron a fake drug used to identify “over claimers”. The GitHub for this project can be referred to for all coding [5].

Dataset description

Ranges and levels indicated correspond to the original state of the dataset

Variable name	**Variable type**	**Levels**	**Range**	**Description**
Age	Ordinal	6		Age group of survey respondents.
Gender	Categorical	2		Male or Female
Education	Ordinal	9		Highest finished degree level of the respondent.
Nscore	Continuous		-3.46436, 3.27393	Nscore, Neuroticism; calm, confident vs. anxious, sensitive [1].
Escore	Continuous		-3.46436, 3.27393	Escore, Extraversion; reserved, thoughtful vs. spontaneous/fun-loving [1].
Oscore	Continuous		-3.27393, 2.90161	Oscore, Openness to experience; routine/cautious vs. inventive/curious [1].
Ascore	Continuous		-3.46436, 3.46436	Ascore, Agreeableness; uncooperative/callous vs. compassionate/trusting [1].
Cscore	Continuous		-3.46436, 3.46436	Cscore, Conscientiousness; organized/careful vs. chaotic/careless [1].
Impulsive	Continuous		-2.555245, 2.90161	Impulsive; the measure of impulsivity for a person using the BIS-11 [2].
Sensation seeking	Continuous		-2.07848, 1.92173	A score based upon the Impulsive Sensation Seeking scale [3].
Caffeine	Ordinal	7	0-6	A level that indicates when the drug was last used (frequency). 0 never used vs. 6 yesterday [4].
Cannabis	Ordinal	7	0-6	A level that indicates when the drug was last used (frequency). 0 never used vs. 6 yesterday [4].
Meth	Ordinal	7	0-6	A level that indicates when the drug was last used (frequency). 0 never used vs. 6 yesterday [4].
Semeron	Ordinal	7	0-6	A level that indicates when the drug was last used (frequency). 0 never used vs. 6 yesterday [4].

Preprocessing of the dataset

The dataset in its original form (how it was distributed) was heavily preprocessed. The original author of the dataset had preprocessed the dataset in such a way that it was fully numerical in order to fit with certain use cases. For example all of the ordinal/categorical data were binned and transformed into a normalized form(for detailed information about this refer to the original paper [1]). Numerical values in the dataset that were already numerical from origin were also preprocessed in a similar way, therefore losing their original value.

For the use in the context of this project a few design choices were made in order to shape the data to a format that was convenient. This with the help of several exploratory data analysis techniques to first create the insights that were needed to understand the data and its properties. Although every variable has been considered separately, it can mostly be broken down in the following rules:

- Originally discrete values with ordinal ordering were transformed back to discrete.
- If the variable was numerical and it was originally (roughly) normally distributed it was also transformed back to its original value.
- Numerical values whose original distribution was not (roughly normal) were skipped.
- Categorical variables with two values and no ordering were binarized.
- Drugs levels were all transformed to numeric values, while keeping ordering in mind.

To automate this preprocessing as much as possible the Python programming language was used. The documented code can be found on GitHub for reference [5.1].

The Exploratory Data Analysis was also conducted with the Python programming language. This process mainly existed out of tasks like: Inspecting data types, checking for missing values, plotting distributions, plotting correlation matrices. During this phase it was detected that several classes suffered from a problematic class imbalance, in the case of the Semeron variable it was thus bad that it was deemed unusable (more on the other variables later on). The documented code can be found on GitHub for reference [5.2].

Network

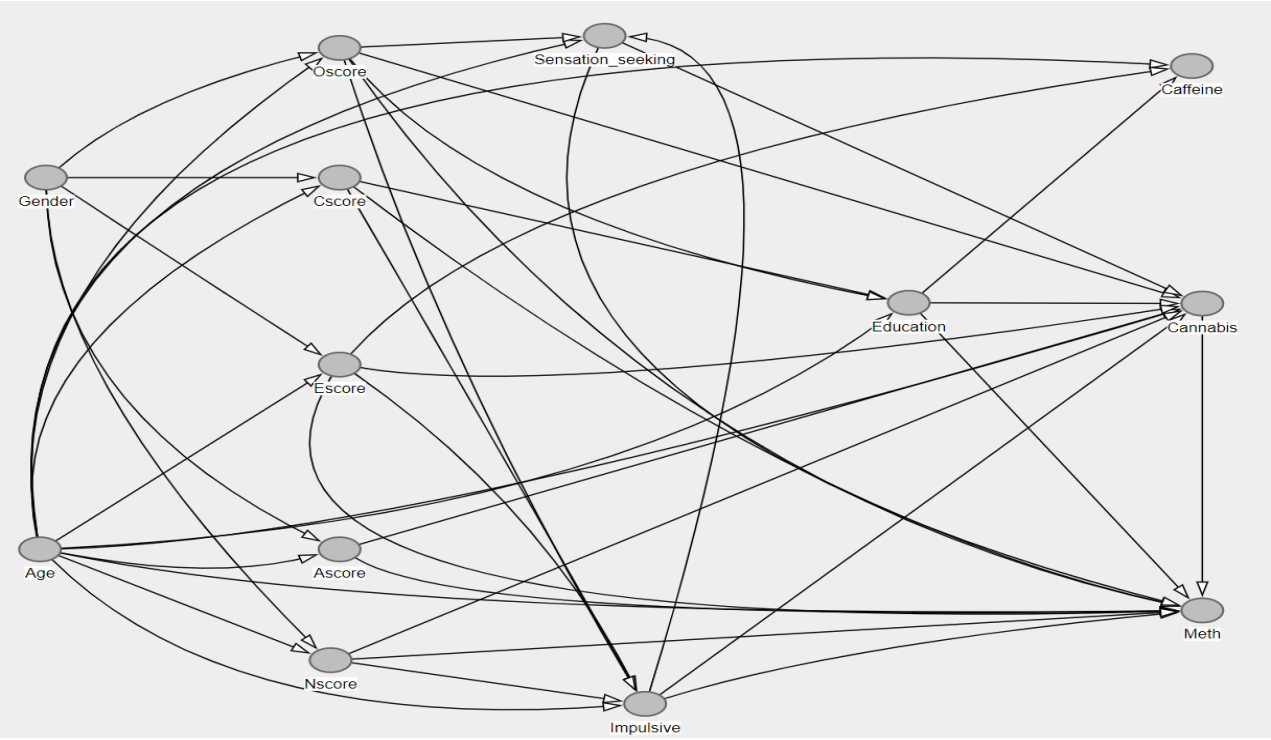

Figure 1 Bayesian network "DAG"

Designing the network

Dagitty.net [6] was used to create and edit the Bayesian network. Within the browser-based environment, it was possible to generate the code for a DAG. Edges were drawn between nodes based on our initial real-world knowledge (or bias even) about drug usage and its relations with other factors. For this, relations between variables were assumed considering either what society/academia has preconceived notions about, or what was thought the relation should be. The initial design was created by ordering the variables in a fashion of time. However, given the nature of the variables this was not fully applicable, more on this later (assumptions and practicalities). Next to this knowledge and information the model was further designed with the help of statistical tests, however rationality and human knowledge were always the leading factors.

Assumptions and practicalities

Due to the very complex and somewhat difficult to grasp nature of the dataset and its variables a few key assumptions will be discussed shortly:

- Since the dataset contains three different test measures for personality (Big 5, ImpSS, Bis-11) the choice was made to set the Big 5 scores first in order of the network followed by the other two tests. Meaning that the Big 5 cannot be influenced by the other two but the other way around is possible. Reasoning for this was that the Big 5 captures more aspects of personality.
- To avoid cycles and latent variables (between scores) all Big 5 scores were considered independent. This also makes sense when reasoning about it (but that can be said about there being connections between the scores as well). So, this is also a choice to reduce complexity.
- Age and gender influence all Big 5 scores, this was based upon literature [7].

Testing the network

Depending on the data types that are present beneath a Bayesian network a certain approach should be chosen in order to test and model it properly. Since the dataset of this project consists of a combination of both discrete and continuous variables a Structural Equation Model (SEM) was chosen (an approach suited for continuous data). Note that it also would have been possible to bin all the continuous variables and/or make use of one hot encoding to be able to use an approach suited for discrete data. The reason this was not chosen however was because information would have been lost with this approach.

Since the nature of a SEM requires all data to be continuous, the data had to be further processed. It is not possible to simply transform regular categories to a meaningful numerical mapping. It is however possible to make use of the ordinal nature of the present variables to create a meaningful mapping. The technique that was used to accomplish this was with the creation of a “polychoric correlation matrix”. With the combination of the matrix, the dataset, and the Bayesian network graph it was possible to create a model and conduct tests on it. (More on this under “Implementation details”)

The approach to validate and update the graph/model was to conduct tests of implied independence on the network and its underlying data. When conducting these tests, the goal was to find implied independencies that scored below a certain threshold of the P-value “0.05” (this indicates that there is strong evidence against the implied independence). However as stated before whenever there was a serious doubt about the validity of the result the deciding factor was human intuition. To give an example, one of the tests implied that there was strong evidence against the independence of age and gender (which would make no sense if you think about it rationally). The most serious/ interesting tests were applied and experimented with, not with all of them to avoid overfitting of the model. As a final step the AIC and BIC comparison statistics were used to compare multiple versions of the model. The process/results of testing can be found in the documented R-code on GitHub [5.3].

Implementation details

Inference task 1, calculating path coefficients:

In this first inference task the goal is to use the structure of the network graph to learn something about the relations (edges) between the different graph nodes. To do this the goal will be to calculate path coefficients that display the influence and the strength of it between nodes.

Lavaan is a package used to estimate multivariate statistical models, including structural equation models. This package was used to model the Structural Equation Model (SEM) [8].

As already briefly discussed in the previous section a SEM model was created using the data, the polychoric correlation matrix and the graph. This model was used for testing purposes but that is not the only thing it is going to be used for. The hybrid discrete/ continuous model will also be used to calculate the path coefficients. Since the model was already created and tested this is fairly simple to achieve by telling Lavaan to plot the graph and extract/calculate its coefficients based upon the fitted model. The R-code to execute this can be referred to on GitHub [5.4].

Inference task 2, prediction:

For the implementation of the prediction algorithm the R-package “Bnlearn” [9] was used. Bnlearn is an R-package that enables users to do various things with the graphical structure of Bayesian networks, in this case inference in the form of prediction.

Unfortunately, it will not be possible to use the exact same model/ package to make predictions on drug usage of new unseen “cases”. Since Bnlearn does not allow the use of a hybrid model like used in the previous section, a SEM based upon pure numerical data will be created (so no polychoric matrix will be used). In terms of preprocessing the ordinal nature of the categories will again be used to transform all variables to the expected data type. After completing this step, a numerical dataset and model that work with this dataset are acquired.

Now that the dataset and the model are ready the processed data is split into two separate sets: a “training” set and a “testing” set where 75% of the data is used to fit the model and the other 25% for testing. Keeping in mind that class imbalances were found during the Exploratory Data Analysis [5.2] stage these two sets will be made to have matching distributions in terms of the data. The reason for splitting into these two sets is to be able to test how well the model generalizes to unseen data (data records that were not used to fit the model). The R-code to execute this can be referred to on the GitHub repository [5.5].

Application of network

Inference task 1, Path Coefficients:

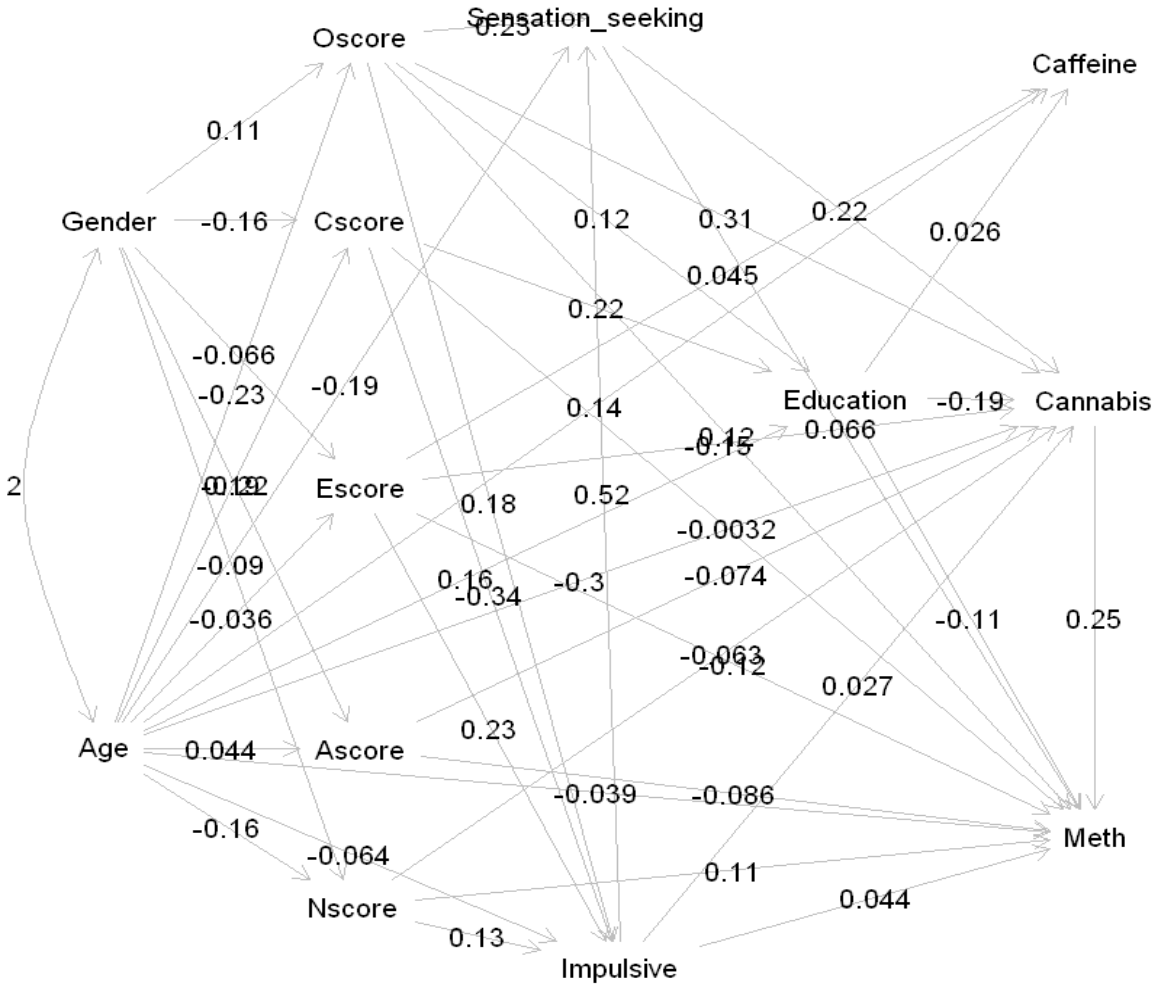

Figure 2 Bayesian Network with path coefficients

This graph can be used to explain and reason further about relations between variables and especially the strength of the relations. A few reasonably expressive relations that were found particularly interesting will be listed below (This with regard to space):

Relation	Description
Cannabis -> Meth (positive relation)	This relation seems to strengthen the belief that Cannabis is very often a so-called "gateway" drug. Next to this belief it also seems to agree with a variety of studies [10].
Oscore -> Cannabis, Meth (positive relation)	This further strengthens/agrees with our initial beliefs that people that are more open to new experiences are more susceptible to drug use.
Education -> Cannabis, Meth (negative relation)	Because this coefficient is negative this relation implies that when the education variable is high (finished a higher level of education) the chance of using meth and cannabis frequently is lower. Or in reverse: High drug usage lowers chance of finishing a higher level of education.

There were also a few relationships that were less significant/expressive than expected:

Relation	Description
Sensation seeking -> Meth (small negative relation)	Based on prior assumptions about the world it was expected that sensation seeking behavior would be a possible cause for meth/cannabis usage. This because these drugs were believed to bring some sort of sensation. Interesting is the fact that this assumption does hold for cannabis.
Escore -> Caffeine (small negative relation)	It is a well-known fact that caffeine is a widely used and socially accepted drug. For this reason, it was already assumed that it would be hard to draw an interesting causal relation because it is so general. There was however reasoning that a very energetic person would be less likely to use caffeine. The test statistics also agreed that the chance of the variables being independent was small. The relation coefficient however is very small.
Ascore -> Meth, Cannabis (small negative relation)	Based upon prior assumptions there was an expectation for a strong coefficient for this path because there was reasoned that people that agree very easily might be more likely tricked into drugs at some point.

An interesting observation that can be seen in the path coefficients is the lack of expressive relations between the “Caffeine” node and the other nodes in the network (except for Age). If the other drug nodes (Cannabis and Meth) are inspected, it can be seen that these variables suffer less from this issue hence making the model more expressive for these variables.

Inference task 2, Prediction:

For the prediction itself the package Bnlearn [9] was used (the same as for the implementation). The Bnlearn package provides the user with multiple options on how to conduct to prediction. In this project two of those options were explored namely: “Parents” and “Bayes-lw”, a brief explanation of both:

Parents: When predicting a value, the algorithm will use only the parent nodes of the node that is predicted. The relevant parent variables will be extracted from the dataset based upon what the fitted model specifies. A local probability distribution will be made and used to predict.
Bayes-lw: This approach does not only compute based upon the parent nodes but on all the present variable nodes in the network. These nodes will be used to compute an average likelihood weighting based on a number of samples.

In order to measure how accurate, the model predicts a function was used which compares the model’s predicted value against the actual value. To express this into a score a very straightforward approach was used: Calculate the difference between all predictions and their actual values and take the average of them. *Mean of 10 random initializations of the train and test set were used [4.5].

Prediction algorithm	Caffeine	Cannabis	Meth
Parents*	3.6145	0.9303	1.4424
Bayes-lw*	3.6153	0.9775	1.4410
Difference*	0.0010	0.1454	0.0014

Discussion/ Conclusion

When looking back at the results in terms of the research topics it can be divided in the network itself, the path coefficients, and the prediction. Starting with the network itself, after thorough testing and reasoning about the networks layout a satisfying network was found which also indicated a proper fit. That having said, the network will always have a bias from the ones that designed it. The key assumption was that the network would reach its best expressiveness with the combination of both statistical testing and human knowledge/perception of the world. The network however has certain limitations in its modeling. One of the limitations is the design choice of not allowing connections between big five scores to avoid introducing latent variables and/or introducing cycles which would go against the DAG specifications. During testing and reasoning about the model it was also concluded that not all variables are present (e.g. growing up environment, traumatizing experiences) to properly model the relations, which leads to another limitation of the network.

When inspecting the path coefficients, it can be seen that the network is not able to reason well with certain variables (e.g. “caffeine” which has nearly no expressive connections). The results of the coefficients however were able to substantiate beliefs from both own perception and results found in academia, which indicates that certain important connections were successfully captured.

The prediction using the network shows potential for cannabis and meth but suffers from instability between different initializations, Cannabis shows best results which corresponds to what is described in the original paper [4]. Caffeine is hard to predict since the network is not very expressive about this variable (see path coefficients). There does not seem to be a significant difference between using the entire network or only the parent nodes for prediction. That being said it is believed that the true potential of the Bayesian network does not lie purely in prediction but giving a person with domain knowledge a tool to help reason about certain cases.

In short, the most important findings are: The evidence that leads to the belief that there are variables missing to fully model all relations properly; the network being able to substantiate a variety of meaningful relations by showing their coefficients and the prediction showing promise for cannabis and meth but lacking enough information for stability. The key limitations are the design of the network limiting the interrelations of the “Big 5” and the absence of more expressive variables. Ideas for further projects extending on this idea could be the introduction of more domain knowledge in the model design; the introduction of more feature rich data and a comparison with a wider range of prediction algorithms.

References

[1] R.R.M.C., & P.T.C. (2004). A contemplated revision of the NEO Five-Factor inventory. Personality and Individual Differences, https://doi.org/10.1016/S0191-8869(03)00118-1

[2] Stanford, M. S., Mathias, C. W., Dougherty, D. M., Lake, S. L., Anderson, N. E., & Patton, J. H. (2009). Fifty years of the Barratt Impulsiveness Scale: An update and review. Personality and Individual Differences, 47(5), https://doi.org/10.1016/j.paid.2009.04.008

[3] Fernández-Artamendi, S., Martínez-Loredo, V., Fernández-Hermida, J. R., & Carballo-Crespo, J. L. (2016). The Impulsive Sensation Seeking (ImpSS): Psychometric properties and predictive validity regarding substance use with Spanish adolescents. Personality and Individual Differences, 90, https://doi.org/10.1016/j.paid.2015.11.003

[4] E.F., A.K.M., E.M.M., V.E., & A.N.G. (2017). The Five Factor Model of personality and evaluation of drug consumption risk. Cornell University, https://arxiv.org/abs/1506.06297

[5] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment

[5.1] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment/blob/main/Data_Preprocessing.ipynb

[5.2] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment/blob/main/EDA.ipynb

[5.3] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment/blob/main/Network_Building_Testing.ipynb

[5.4] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment/blob/main/Path_Coefficient_Estimating.ipynb

[5.5] GitHub source for this project, https://github.com/prat8897/BayesianNetworkAssignment/blob/main/Network_Prediction.ipynb

[6] Johannes Textor, Benito van der Zander, Mark K. Gilthorpe, Maciej Liskiewicz, George T.H. Ellison. Robust causal inference using directed acyclic graphs: The R package 'dagitty'. International Journal of Epidemiology 45(6):1887-1894, 2016.

[7] Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the big five across the life span: Evidence from two national samples. Psychology and Aging, https://doi.org/10.1037/a0012897

[8] Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48(2), 1-36, http://www.jstatsoft.org/v48/i02/

[9] Marco Scutari, Ph.D., bnlearn - an R package for Bayesian network learning and inference, https://www.bnlearn.com/

[10] U.S. Department of Justice, W.N., & J.W. (2018). Is Cannabis a Gateway Drug? https://www.ncjrs.gov/pdffiles1/nij/252950.pdf