Assignment 1 report
Bayesian Networks and Causal Inference

Floris Rossel (s1010794), Andro Radičević (s1102974), Mario Tsatsev(s1028415)

December 9th, 2022

1 Introduction

High quality education should be available to everyone, however, many factors affect the ability of students to perform well in their programmes. Unequal backgrounds and circumstances can result in varying degrees of student performance. To give everyone equal chances of succeeding, it is important to understand the factors that have a big effect on student performance. This paper will primarily focus on determining the effect of various studying habits on the final grade of the student, with limited measured influence of student background. Our project will use a publicly available data set, from which we will construct a Bayesian network model to find the major factors influencing student performance.

2 Data

The dataset used in our research was the "Higher Education Students Performance Evaluation Dataset" The data was collected from the Faculty of Engineering and Faculty of Educational Sciences students in 2019.[2]

2.1 Raw Data

The dataset consists of 145 samples of 33 attributes each. Each attribute is divided in several classes represented as integers, with some exceptions. The number of classes wary with each attribute. The description of each attribute is provided in the appendix2:

2.2 Preprocessing

During the preprocessing steps, we removed most variables not related to the study habits of students. We kept only a limited number of variables related to the student background.
The variables removed are, according to the table 2, 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20 and 32.
For each remaining variable an ordering was defined when possible. In addition, a category was removed if no samples were found containing it. An example of that is attendance to classes where the value "*never*" had zero occurrences. Other variables were turned into binary variables. This processing was necessary for the data to be compatible with the chosen methods, and to prevent the loss of information which can occur when defining ordered variable as categorical. The exact way each variable was processed is provided in the appendix3:

3 Methods

In order to find the causal pathways that influence the students' performance, we implemented a structural equation model (SEM) in R programming language using several packages; daggity[3], lavaan[1] and tidyverse[4]. We decided to start by creating a simple DAG (directed acyclic graph) that contains very few variables that we thought were the strongest predictors in relation to the grade based on theoretical knowledge. We iteratively added new variables and arrows to the DAG based on conditional independence test results for each addition. The workflow went as follows:

1. Add a couple of variables (1 to 3) to the DAG without arrows that we thought were strong predictors.
2. Test the model using linear regression with localTests and the polychoric correlation matrix for the relevant variables, and order the results such that the implied conditional independences with the lowest p-values are on top.
3. Add arrows for the variable pairs that have low p-values (< 0.05) for their conditional independence test and can be causally explained with domain specific knowledge.
4. Iterate step 2 and 3 until there are no pairs left that can be causally explained with domain specific knowledge.
5. Fit the model to the data with the lavaan package.
6. Remove connections that have very low coefficients (< 0.1)
7. Redo step 2, and add different arrows if the p-value for the pair where we removed the arrow was too low. If the new arrow(s) again got a very low coefficient, we removed it and stopped going to step 2.
8. Go back to step 1, until we added all variables that we found most relevant for student performance. We ignored variables that describe the student background, like parental status, salary and transportation method.

4 Final model

Our final model (Figure 1) contains the relevant variables and causal connections that our methods produced. As we can see, the factors that directly affect the students' grades are 'class attendance', 'study hours', 'listening', 'taking notes', 'preparation to midterm exams (when)', 'cumulative grade point average of last semester', 'age' and 'sex'. All of these, except for sex, make sense. However, the data strongly suggested that sex had a big effect on the grade. We could not come up with a good explanation for this effect, but we left it in as the effect was quite strong.

The Path coefficients of the fitted model can be found in the appendix1. They are used to gauge the correlation between variables and can also imply causal effects. Most of them confirm conventional knowledge about good study habits. For example, there is a very strong correlation between studying continuously during the semester and getting a better grade. Other, weaker correlations are also informative as they describe effects that although small are also not negligible, and combine together to noticeably affect the final grade.

Overall, we can confidently conclude that study habits, as represented by most of the variables, affect the grade. We also found that age and sex had a big correlation with these intermediate factors, as well as the grade. There were also some effects recorded that we were confident are due to errors in data collection procedures and do not imply causation, such as correlation between age and sex. They were not added as arrows in the final model.

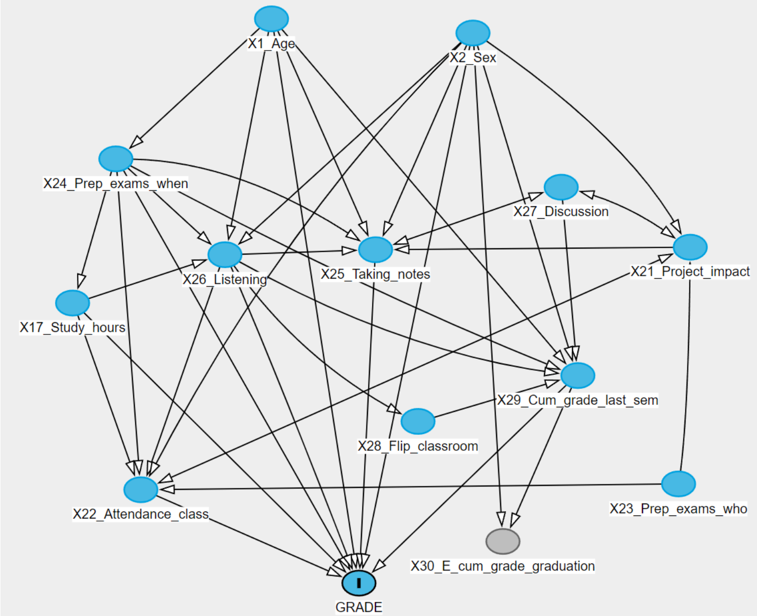

Figure 1: Final DAG

5 Discussion

We are very satisfied with our final model for the following reasons:

- The model is not too complex, and all arrows have some logical theoretical explanation that most people can probably understand.
- The p-values for the conditional independence tests are all not too low, which means that the data does not strongly suggest that some implied conditional independences are really wrong.
- The model does not contain regression coefficients that are too low, and each regression result is significant ($p < 0.05$), meaning that each arrow has a meaningful effect in the data. 1

We think that our systematic workflow for expanding and fine tuning our graph has contributed to getting such a good result. The process behind adding connections was quite conservative and therefore did not lead to a too complex model.

References

[1] Yves Rosseel. "lavaan: An R Package for Structural Equation Modeling". In: *Journal of Statistical Software* 48.2 (2012), pp. 1–36. DOI: 10.18637/jss.v048.i02.

[2] Boran Sekeroglu, Kamil Dimililer, and Kubra Tuncal. "Student Performance Prediction and Classification Using Machine Learning Algorithms". In: Mar. 2019, pp. 7–11. ISBN: 978-1-4503-6267-2. DOI: 10.1145/3318396.3318419.

[3] Johannes Textor et al. "Robust causal inference using directed acyclic graphs: the R package 'dagitty'". In: *International Journal of Epidemiology* 45.6 (2016), pp. 1887–1894. DOI: 10.1093/ije/dyw341.

[4] Hadley Wickham et al. "Welcome to the tidyverse". In: *Journal of Open Source Software* 4.43 (2019), p. 1686. DOI: 10.21105/joss.01686.

A Appendix

Table 1: Regression results

Regressions	**Estimate**	**Standard error**	**z-value**	$\mathbf{P}(>\lvert\mathbf{z}\rvert)$
GRADE ∼ X17_Study_hours	-0.477	0.091	-5.244	0.000
X22_Attendance_class ∼ X17_Study_hours	0.440	0.091	4.857	0.000
X26_Listening ∼ X17_Study_hours	0.453	0.083	5.457	0.000
GRADE ∼ X1_Age	-0.409	0.066	-6.239	0.000
X24_Prep_exams_when ∼ X1_Age	-0.396	0.076	-5.191	0.000
X25_Taking_notes ∼ X1_Age	-0.199	0.072	-2.773	0.006
X26_Listening ∼ X1_Age	0.170	0.066	2.578	0.010
X29_Cum_grade_last_sem ∼ X1_Age	0.208	0.078	2.649	0.008
X25_Taking_notes ∼ X21_Project_impact	-0.268	0.071	-3.767	0.000
GRADE ∼ X22_Attendance_class	0.235	0.073	3.228	0.001
X21_Project_impact ∼ X23_Prep_exams_when	0.254	0.076	3.360	0.001
X22_Attendance_class ∼ X23_Prep_exams_when	0.142	0.063	2.256	0.024
GRADE ∼ X24_Prep_exams_when	0.876	0.133	6.572	0.000
X17_Study_hours ∼ X24_Prep_exams_when	0.700	0.059	11.793	0.000
X22_Attendance_class ∼ X24_Prep_exams_when	-1.122	0.109	-10.270	0.000
X25_Taking_notes ∼ X24_Prep_exams_when	0.461	0.080	5.770	0.000
X26_Listening ∼ X24_Prep_exams_when	-0.813	0.087	-9.349	0.000
X29_Cum_grade_last_sem ∼ X24_Prep_exams_when	0.267	0.087	3.054	0.002
GRADE ∼ X25_Taking_notes	-0.255	0.063	-4.031	0.000
X22_Attendance_class ∼ X26_Listening	-0.790	0.081	-9.744	0.000
X25_Taking_notes ∼ X26_Listening	0.394	0.081	4.865	0.000
X28_Flip_classroom ∼ X26_Listening	0.186	0.078	2.370	0.018
X29_Cum_grade_last_sem ∼ X26_Listening	0.276	0.089	3.089	0.002
X29_Cum_grade_last_sem ∼ X27_Discussion	0.255	0.069	3.707	0.000
X29_Cum_grade_last_sem ∼ X28_Flp_clssrm	-0.193	0.071	-2.725	0.006
GRADE ∼ X29_Cum_grade_ls_	0.189	0.062	3.039	0.002
X30_E_cum_grade_graduation ∼ X29_Cm_grd_ls_	0.679	0.057	11.886	0.000
GRADE ∼ X2_Sex	0.928	0.081	11.397	0.000
X21_Project_impact ∼X2_Sex	0.207	0.076	2.725	0.006
X22_Attendance_class ∼ X2_Sex	-0.546	0.071	-7.643	0.000
X25_Taking_notes ∼ X2_Sex	0.371	0.076	4.866	0.000
X26_Listening ∼ X2_Sex	-0.442	0.061	-7.281	0.000
X29_Cum_grade_last_sem ∼ X2_Sex	0.394	0.081	4.869	0.000
X30_E_cum_grade_graduation ∼ X2_Sex	0.188	0.057	3.297	0.001

Table 2: Raw data

ID	Variable name	Type	Number of levels
0	Student ID	Categorical	Unique ID given to each student
1	Student Age	Categorical	(1: 18-21, 2: 22-25, 3: above 26)
2	Sex	Categorical	(1: female, 2: male)
3	Graduated high-school type	Categorical	(1: private, 2: state, 3: other)
4	Scholarship type	Categorical	(1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full)
5	Additional work	Categorical	(1: Yes, 2: No)
6	Regular artistic or sports activity	Categorical	(1: Yes, 2: No)
7	Do you have a partner	Categorical	(1: Yes, 2: No)
8	Total salary if available	Categorical	(1: USD 135-200, 2: USD 201-270, 3: USD 271-340, 4: USD 341-410, 5: above 410)
9	Transportation to the university	Categorical	(1: Bus, 2: Private car/taxi, 3: bicycle, 4: Other)
10	Accommodation type in Cyprus	Categorical	(1: rental, 2: dormitory, 3: with family, 4: Other)
11	Mother's education	Categorical	(1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
12	Father's education	Categorical	(1: primary school, 2: secondary school, 3: high school, 4: university, 5: MSc., 6: Ph.D.)
13	Number of sisters/brothers (if available)	Categorical	(1: 1, 2:, 2, 3: 3, 4: 4, 5: 5 or above)
14	Parental status	Categorical	(1: married, 2: divorced, 3: died - one of them or both)
15	Mother's occupation	Categorical	(1: retired, 2: housewife, 3: government officer, 4: private sector employee, 5: self-employment, 6: other)
16	Father's occupation	Categorical	(1: retired, 2: government officer, 3: private sector employee, 4: self-employment, 5: other)
17	Weekly study hours	Categorical	(1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)
18	Reading frequency (non-scientific books/journals)	Categorical	(1: None, 2: Sometimes, 3: Often)
19	Reading frequency (scientific books/journals)	Categorical	(1: None, 2: Sometimes, 3: Often)
20	Attendance to the seminars/conferences related to the department	Categorical	(1: Yes, 2: No)
21	Impact of your projects/activities on your success	Categorical	(1: positive, 2: negative, 3: neutral)
22	Attendance to classes	Categorical	(1: always, 2: sometimes, 3: never)
23	Preparation to midterm exams 1	Categorical	(1: alone, 2: with friends, 3: not applicable))
24	Preparation to midterm exams 2	Categorical	(1: closest date to the exam, 2: regularly during the semester, 3: never)
25	Taking notes in classes	Categorical	(1: never, 2: sometimes, 3: always)
26	Listening in classes	Categorical	(1: never, 2: sometimes, 3: always)
27	Discussion improves my interest and success in the course	Categorical	(1: never, 2: sometimes, 3: always)
28	Flip-classroom	Categorical	(1: not useful, 2: useful, 3: not applicable)
29	Cumulative grade point average in the last semester (/4.00)	Categorical	(/4.00): (1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
30	Expected Cumulative grade point average in the graduation (/4.00)	Categorical	(1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
31	Course ID	Categorical	Unique ID given to each course
O	Grade	Categorical	(0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)

Table 3: Processed data

ID	Variable name	Type	Number of levels
1	Age	Ordered	(1: 18-21, 2: 22-25, 3: above 26)
2	Sex	Categorical	(1: female, 2: male)
17	Weekly study hours	Ordered	(1: None, 2: <5 hours, 3: 6-10 hours, 4: 11-20 hours, 5: more than 20 hours)
21	Impact of your projects/activities on your success	Ordered	(2: negative, 3: neutral, 1: positive)
22	Attendance to classes	Categorical	(1: always, 2: sometimes)
23	Preparation to midterm exams 1	Categorical	(1: alone, 0: not alone)
24	Preparation to midterm exams 2	Ordered	(3: never, 1: closest date to the exam, 2: regularly during the semester)
25	Taking notes in classes	Ordered	(1: never, 2: sometimes, 3: always)
26	Listening in classes	Ordered	(1: never, 2: sometimes, 3: always)
27	Discussion improves my interest and success in the course	Ordered	(1: never, 2: sometimes, 3: always)
28	Flip-classroom	Categorical	(1: useful, 0: not useful)
29	Cumulative grade point average in the last semester (/4.00)	Ordered	(1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
30	Expected Cumulative grade point average in the graduation (/4.00)	Ordered	(1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49)
O	Grade	Ordered	(0: Fail, 1: DD, 2: DC, 3: CC, 4: CB, 5: BB, 6: BA, 7: AA)