Bayesian Networks Exposee

Eric Buitinck, s4070267
Max de Grauw, s4648501
Douwe van Erp, s4258126

September 27, 2018

1 Problem Domain

Student achievement is an important domain to understand in order to improve the level of education of a country. Portugal is an example of a European country with statistically high student failure rates in the secondary education system. By using data of demographic and social features and student grades, we can investigate which factors influence student performance. We will use publicly available data for this purpose in order to model the underlying causal structure for this domain by using a Bayesian network. Our main research questions are: How does the amount of free time directly affect student performance? What is of the greatest influence to student performance among the following variables: family support, (free) school support and paid support? What are some concrete recommendations we can give to improve performance (according to this data)?

2 Data

For our project we will be using the student performance dataset[1], found on the UCI machine learning repository. This dataset consists of instances of secondary school students from Portugal and information about these students and their grades. It is actually a combination of two datasets aiming to measure student performance in two subjects, Portuguese and mathematics, with one dataset for each. We have chosen to use those instances that occur in both datasets. The sets are merged by adding a single variable, Subject, that represents the original dataset and the relevant subject. The overlapping data consists of 382 instances and 34 attributes, of which one is added manually and three attributes are the final results: three sequential grades throughout the year. We have narrowed the attribute set down to the following: school, mother's education, father's education, mother's job, father's job, home to school travel time, number of past class failures, extra educational support from school, family educational support, extra paid classes within the course subject, extra-curricular activities, ambition for higher education, with a romantic relationship, quality of family relationships, free time after school, going out with friends, current health status, number of school absences, subject (Math or Portuguese), and final grade, leaving us with 22 attributes. We have also added a custom variable, IQ, that seems relevant to the data set, as per the assignments. The meaning of each of these variables is explained on the linked UCI site.

3 Application

We will implement our Bayesian Network as a linear model, where linear regression functions shall be used to represent the probability distributions at each node. We will implement our Bayesian Network in Python, using the *pgmpy* package, which specializes in creating and using Bayesian Networks. Specifically we want to determine: the direct effect of free time on the final student grade, and the effects of family support, school support and paid support on the final grade.

4 Initial causal diagram

The following page includes the cause diagram indicating our progress and idea for the project so far.

[1] https://archive.ics.uci.edu/ml/datasets/student+performance

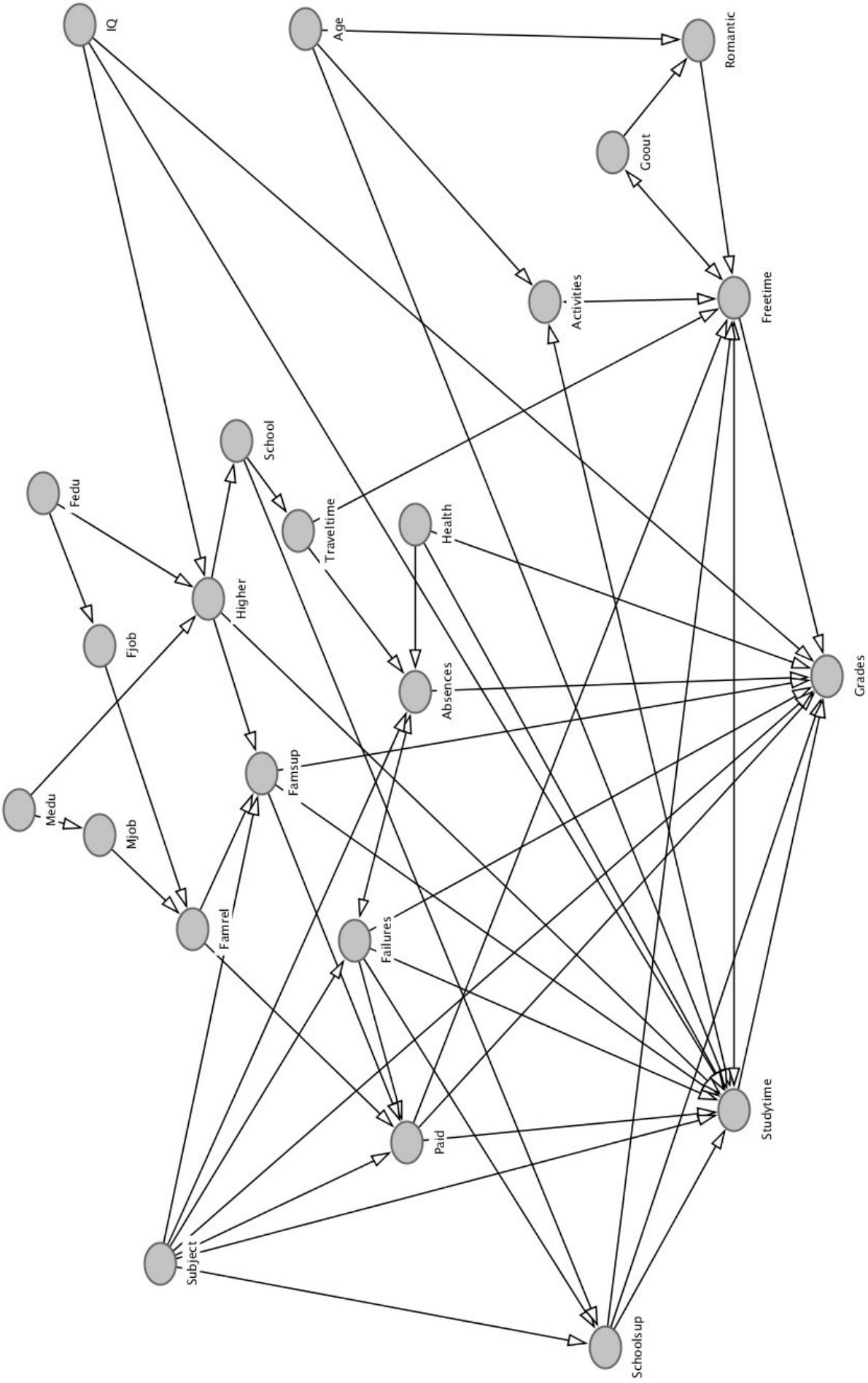

Figure 1: A graph of the causal diagram for the data. Note that the IQ node is added and not included in the data.