# Analysis of portuguese students' habits that influence education

## Michael Albarello, Matteo Nestola

Master's Degree in Artificial Intelligence, University of Bologna
{ michael.albarello, matteo.nestola }@studio.unibo.it

December 19, 2022

### Abstract

This mini-project aims to analyze the relationship between students' academic behaviour and other aspects of their life, i.e social aspects, free time habits etc. For these purposes, the *Alcohol Effects On Study* dataset, which link is given below, was found to be adequate; in particular, it consists of two parts, the first concerning the students' performance in mathematics and the second in Portuguese language. To exploit the totality of the given information, we decided to merge the two subsets to form a single bigger dataset that contains more complete information. By analyzing the data provided, we were able to construct a Bayesian network capable of computing a joint probability function for any variable by knowing the state of some other observations. We have found that students' academic careers and their private and social lives are highly correlated, often in an unexpected way, enough to make it interesting to further examine the Bayesian network.

The structure of the *jupyter notebook* was designed to ensure a smooth and easy reading with which, adding one brick at a time, even more complex concepts were covered.

## Introduction

### Domain

Dedicating a lot of time to study can be effective to obtain great results, but on the other hand, it is also good not to neglect yourself and your social and family relationships. According to such reasoning, the ideal goal of every student should be to find a proper equilibrium by which it is possible to achieve a good academic performance and still conduce a fairly active social life, without taking either one to extremes. Our project comes into play precisely in this context where, thanks to the model we created, we analyzed and compared which behaviors influence a student's school career and which do not. By doing so, we were able to give a 'measure' to the different attitudes of interest to understand how to dose them in order to achieve the desired results.

### Aim

As we said above, the purpose of our project work is to identify which behaviors and habits influence academic performance and to what extent. In particular, we can discern the following objectives:

- Constructing a Bayesian network so that we could perform probabilistic inference between the selected variables of interest as we have seen during the course and from the book (Ruiz-Pérez et al. 2019);

- Collecting data exploting different approaches to inference, such as *exact inference* and *approximate inference* and observe if there were different results in terms of conditional probability distributions;

- Trying to understand what knowledge arose from such collected information to figure out how features bind together;

## Method

In our project we have exploited many useful libraries, in particular:

- **pgmpy**, A powerful python library that enabled us to create the Bayesian network and perform operations on it, such as the estimation of the parameters and of the *conditional probability tables (CPTs)*, inference queries, both *exact* and *approximate* and all the rest of Bayesian network related tasks. Specifically, the used *pgmpy* packages, were:

  - BayesianNetwork,
  - TabularCPD,
  - ParameterEstimator,
  - BayesianEstimator,
  - VariableElimination,
  - ApproxInference.

- **networkx library**, which allowed us to to represent the model so that we could graphically analyze block by block and check their relationships;

- **pandas** and **numpy**, core libraries for data generation and manipulation;

- **matplotlib**, a comprehensive library for creating static, animated, and interactive visualizations in Python.

## Results

As we expected, students that fully committed themselves to their school career, not always achieved higher results, but inversely students that managed also to hang out with their friends or to experience extracurricular activities obtained similar results.
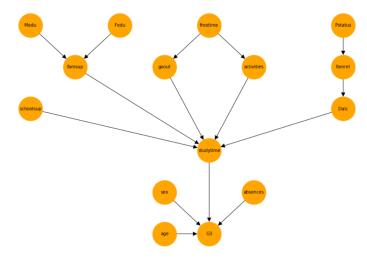
Figure 1: Bayesian network.

## Model

As we can see in the figure (Figure 1), there are fifteen features involved in the realization of our Bayesian network, which retain enough information to fully represent both social life and academic aspects. These nodes were extracted from the features of the original dataset after a careful examination and preprocessing steps to discretize values. At this point, by reading various papers, all referenced below, and by formulating our own detailed analysis, explained further in our notebook, we defined all of the connections between the various nodes of the network. Knowing the nodes and the influence exerted among them, we were able to build the Bayesian model and fit it with the dataset to compute the *Conditional Probability Tables* (*CPTs*), which in the notebook are all printed and represented in a graphical way.

## Analysis

### Experimental setup

As mentioned earlier, the analysis to be carried out through our network focuses on the relationships between features that serve as metrics of social life and academic performance. In particular, we selected as features of greatest interest those concerning the final grade, the study time, the extracurricular activities, the amount of alcohol consumption and the family support; among the most important we can spot:

- What is the probability distribution of the `final grade` given `studytime` and `activities`?

- What is the chance that a student receives `family support` given the `parental level of education`?

- What is a student's `studytime` given his `final grade`?

- What is the probability distribution of the amount of `consumed alcohol` given the time students `hang out with friends` and the amount of `studytime`?

## Results

The results of the previous queries are found to be in accordance with what was expected. Some interesting observations concern the query $P(FinalGrade|studytime, activities)$, whose graph shows that students' grades, given high values of 'studytime' and low values of 'activities,' are not much higher than in the case of intermediate values of 'studytime' and 'activities. Another interesting query is $P(Famsup|Fedu, Medu)$, which shows that the family's level of education has a high impact on the academic performances of the students

## Conclusion

From this it can be seen that studying as much as possible does not guarantee the achievement of higher scores with respect to the case of balanced amount of study time and social activities. Also, the results showed in some cases how the lack of data of some features was a limiting factor in being able to ask some questions to the model (i.e. there is only 1 student in the dataset who obtained 'Muito bom cum laude..' as a grade. As a matter of fact, the conclusions drawn from the analysis of the model were interesting because they revealed the importance of having to dose habits in order to find the perfect mix to succeed in both academic results and maintaining an active social life.

## Links to external resources

List of linkage to project resources:

- Github
- DATASET

Other resources found on the web that helped to understand the dataset provided and build the model:

- Grading System in Portugal
- How Working Parents Can Manage the Demands of School-Age Kids
- Correlation Between Parents' Education Level and Children's Success

## References

Ruiz-Pérez, I.; Ayala, F.; Puerta, J. M.; Elvira, J. L. L.; De Ste Croix, M.; Hernández-Sánchez, S.; and Vera-Garcia, F. J. 2019. A bayesian network approach to study the relationships between several neuromuscular performance measures and dynamic postural control in futsal players. *PLOS ONE* 14(7):1–18.