The use of Static and Dynamic Bayesian Networks in Educational Data Mining Processes

Beatriz Schindler[1], *Lilia Costa*[1], *Anderson Ara*[2]
[1] Federal University of Bahia (UFBA)
[2] Federal University of Paraná (UFPR)

Abbreviated abstract: The eminent use of systems and softwares to manage academic records makes it possible to collect large volumes of data. Therefore, data mining techniques are an increasingly useful support. For the analytical processing stage, Bayesian Networks (BN) were explored in the application to data from UFBA students. The first objective was to predict the dropout rate of students, seeking to identify associated factors. The second objective was to study the change in performance over the semesters.

Related publications:

– SOUZA, V. F. et al, Revista Brasileira de Informática na Educação, p. 519-546 (2021).
– NAGARAJAN, R. et al, Springer (2013).

UFBA
Universidade
Federal da Bahia

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

Problem, Data, Previous Works

1

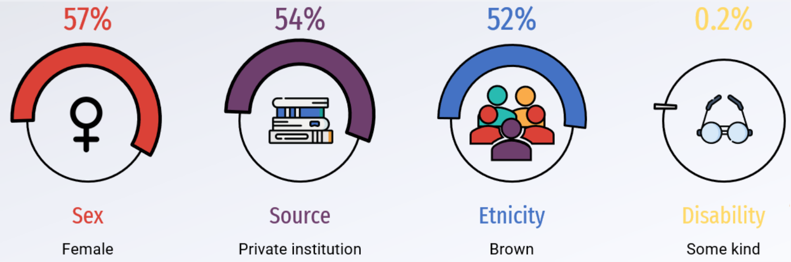
- Dropouts are related to the undergraduate courses, the teaching institution, the student and results from socioeconomic factors external to the institution (Brasil, 1997; Cruz, 2019). As such, it’s a complex problem.
- Being able to **predict**, based on their characteristics, whether a student will drop out or not is useful for designing interventions to reduce drop-out rates.
  - 13535 incoming students from 2010 to 2013 at UFBA. 25% reserved for validation and 75% to training.

2

- Competence in performing evaluative activities, necessary for the evolution of their formation, can be taken as performance.
- Does student performance remain the same **over time**? Are there **conditions** that reflect difficulty in improving performance level?
  - 13287 incoming students from 2005 to 2013 at UFBA, students who had attended at least 8 semesters continuously, with complete information.

Static Bayesian Networks

Approximate descriptions of the data



Dynamic Bayesian Networks

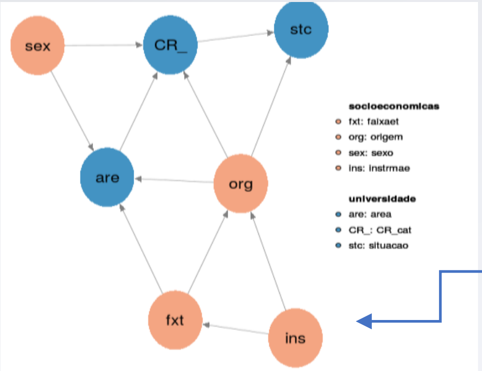
UFBA
Universidade
Federal da Bahia

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

Methods

- The Bayesian network is a method that represents the joint distribution of variables through conditional distributions in a graph structure. It can be used for predictions without sacrifice the interpretability of the model.
  - The dependence structure can be estimated by a machine learning algorithm (Hill-Climbing).

1



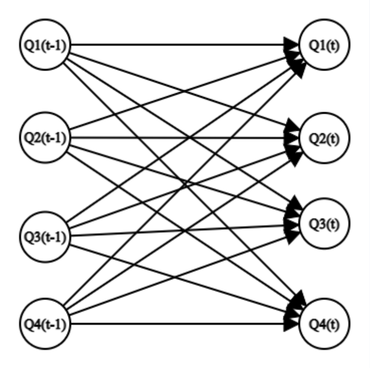
Estimated structure for the prediction of drop-out (“stc” node)

2

- Assuming homogeneity over time.
- Supposing that the process is a first order Markov chain.

$$P(Y_{it}|Y_{i1}, Y_{i2}, \ldots, Y_{it-1}) = P(Y_{it}|Y_{it-1}).$$

- The coefficient of school performance was standardized by the entrance class and discretized by quartiles based on previous works.
- A transition cumulative logistic model (without proportionality of the odds) was used to estimate the probabilities.



UFBA
Universidade
Federal da Bahia

3rd Conference on
Statistics and
Data Science
Salvador, Brazil (online)
October 28-30, 2021

Results and Conclusions

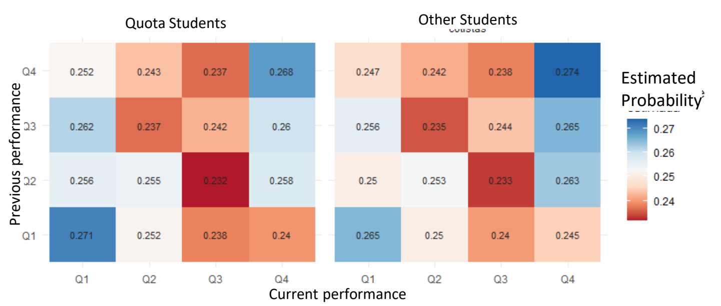
1 Drop-out problem

- Direct influence of student origin and performance on dropout.
- Indirect influence of other variables.
- Differences were found by course knowledge area.
- Drop-out probability
  - **lower** dropout probabilities for students from **public** institutions.
  - **higher** dropout probabilities for students with **lower performances**.

The model had an accuracy of 82.61%, 60.94% of sensibility and 90.74 of specificity.

2 Performance over time problem

- The model considering the influence of the past performance and being or not a quota student as only covariate had the best fit.
- Thus, the transition probabilities must have different values by the covariate.



- Quota students are more likely to remain in the lowest performing category than non-quota students;

UFBA
Universidade
Federal da Bahia

3rd Conference on
Statistics and Data Science
Salvador, Brazil (online)
October 28-30, 2021