Student Dropout Prediction

Zunaira Hasnain
Master's Degree in Artificial Intelligence, University of Bologna
zunaira.zunaira3@studio.unibo.it
May 28, 2025

Abstract

This mini project aims to develop and evaluate a Bayesian network model for predicting student dropout risk using a dataset from a higher education institution. The dataset includes demographic, socio economic, academic, and regional economic variables. Key features are selected and discretized. A directed acyclic graph (DAG) is constructed using domain knowledge, and conditional probability tables are estimated through Maximum Likelihood Estimation. The model is implemented in Python using the pgmpy library and tested through a series of inference queries. The results demonstrate the model's ability to highlight at risk student profiles and offer meaningful insights for designing targeted retention strategies.

Introduction

Domain

Student dropout is a persistent challenge in higher education, carrying academic, economic, and social consequences. Institutions must identify at risk students and understand the factors contributing to dropout. This project addresses the issue by analyzing demographic, academic, and socio economic data.

The dataset is collected from a Portuguese higher education institution, includes features available at enrollment such as age, international status, scholarship status, attendance mode, and regional economic indicators like unemployment and inflation rates. These variables are chosen for their relevance to both personal and environmental influences on student outcomes.

Using a probabilistic framework, the project models the conditional dependencies among variables with a Bayesian network, allowing for clear, interpretable reasoning. Beyond predicting dropout, the goal is to understand its underlying causes and support data driven interventions that improve student retention.

Aim

The project aims at the definition of a model capable of predicting student dropout risk using academic, demographic, and socio economic data from a higher education institution. The goal is for the model to support the identification of high risk student profiles and provide a basis for interpreting how various factors contribute to dropout likelihood. The project itself represents an attempt at probabilistic modeling of student retention and may serve as the foundation for larger scale interventions or future research on academic success and institutional support strategies.

Method

This project follows a structured three step process to develop and evaluate a Bayesian network for predicting student dropout risk. Each step focuses on a specific aspect of the workflow, from preparing the data to constructing the model and finally analyzing its behavior through inference.

- **Data Preparation and Discretization**
  The dataset was cleaned and relevant features were selected. Continuous variables such as age at enrollment, unemployment rate, and inflation rate were discretized into categorical bins to enhance interpretability. A final DataFrame was created using features like attendance mode, scholarship status, international status, and the discretized variables.
- **Model Construction and Parameter Learning**
  A directed acyclic graph (DAG) was defined based on domain knowledge using the pgmpy library. Conditional probability tables were estimated using the BayesianEstimator with a BDeu prior (equivalent sample size of 10). Structural validation included checks for acyclicity, correct independence assertions, Markov blankets, and polytree confirmation.
- **Inference and Analysis**
  Both exact inference (Variable Elimination) and approximate inference (Likelihood Weighted Sampling with 10000 samples) were performed. A series of predictive, diagnostic, and intercausal queries were run, and the outputs from both methods were compared to ensure consistency and interpretability. The results provided insights that support targeted intervention strategies.

Results

The model reveals that older students attending evening classes are among the most at risk groups for dropout. In contrast, international students who receive scholarships tend to have significantly lower dropout risk. Inflation appears to play a notable role in overall dropout trends. Additionally, intercausal analysis suggests that scholarships offer protective effects for domestic students, while international dropouts are less likely to receive such financial support.

Model

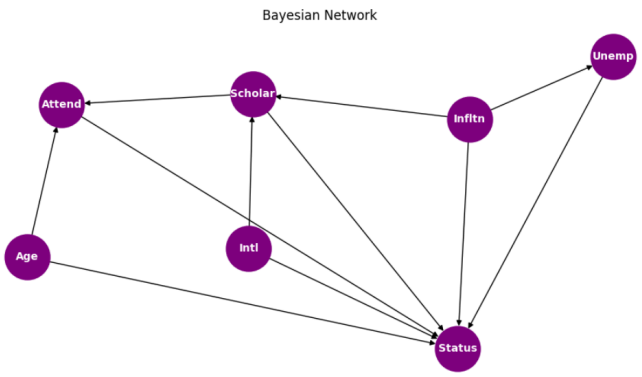

Figure 1: Bayesian Network Model

Figure 1 presents our Bayesian network across seven variables: Age, Attend, International, Scholar, Inflation, Unemployment, and Status, capturing how student life stage and economic factors drive outcomes. The model includes discretized variables to support interpretability. Specifically, Age (the student's age at enrollment) is categorized into four groups: Teen (under 20), Young Adult (20–29), Adult (30–39), and Older (40 and above). Unemployment is classified as Low (below 8%), Medium (8–11.9%), or High (12% and above), while Inflation is grouped into Deflation (below 0%), Stable (0–0.9%), and Rising (1% and above). Age directly influences both Attend (daytime vs. evening attendance) and Status (Dropout, Enrolled, Graduate), with Attend also affecting Status. International (international student) determines Scholar (scholarship holder) and independently impacts Status, while Scholar in turn shapes both Attend and Status. On the macroeconomic side, Inflation influences Scholar, Unemployment, and Status directly, with Unemployment also contributing to Status. This structure encodes our causal assumptions about how individual attributes, financial support, and economic conditions interact to influence student persistence.

Analysis

Experimental setup

To assess the performance and interpretability of the Bayesian network, four key inference queries were formulated and tested. Each query was designed to explore a different aspect of probabilistic reasoning, ranging from predictive and diagnostic to intercausal, and was followed by a contextual interpretation of the results. The objective was to determine whether the model could provide coherent and meaningful responses to real world questions based on the available data. The performed queries were:

1. How likely is a student to drop out if they are an international student and on a scholarship?
2. Does attending evening classes increase dropout probability for older students?
3. If a student dropped out, what is the probability that inflation was rising?
4. What is the probability that a student holds a scholarship given that they dropped out and are an international student?

To ensure the model's reliability, all Conditional Probability Tables (CPTs) were checked to confirm they summed to one, and the structure was validated as a directed acyclic graph (DAG). Additional consistency checks included local independence assertions and examination of each node's Markov blanket. Both exact inference (variable elimination) and approximate inference (likelihood weighted sampling) were used to compare posterior distributions and validate reasoning consistency.

Results

Inference outcomes aligned with domain expectations. International scholarship students had a low dropout rate (~24%), while older students attending evening classes faced the highest (~56%). Inflation periods corresponded with elevated dropout probability (~54%). Intercausal queries showed that about 36% of domestic graduates had received scholarships, compared to only 16% among international dropouts. Surprisingly, the model revealed that international students who dropped out were less likely to have received financial aid, challenging the assumption that they are generally better supported.

Conclusion

This study demonstrates the effectiveness of Bayesian networks in modeling dropout risk, enabling probabilistic interpretation of how demographic, academic, and economic factors interact. Key takeaways include the elevated risk for older evening learners and the protective effect of scholarships. One limitation lies in the scalability of the model, as the conditional probability tables grow exponentially with the number of parent variables and their states. Additionally, the static nature of the model limits its ability to capture temporal dynamics. Future work could explore dynamic Bayesian networks and more scalable approximate inference techniques.

Links to external resources

The dataset used in this project was obtained from Kaggle: https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data.

References

1. pgmpy Developers. (2023). pgmpy: Probabilistic graphical models using Python (Version 0.1.20) [Software]. https://pgmpy.org
2. NetworkX Developers. (2023). NetworkX: Network analysis in Python (Version 3.1) [Software]. https://networkx.org
3. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55