JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Explainable Bayesian Network Recommender for Personalized University Program Selection

Philippe Boribo Kikunda [1, 2, 3,*], Jérémie Ndikumagenge [2], Longin Ndayisaba [2], and Thierry Nsabimana [2]

[1] Computer Science Department, Faculty of Sciences, Université Catholique de Bukavu (UCB); PO Box 285; Bukavu, Democratic Republic of the Congo; e-mail: kikunda.boribo@ucbukavu.ac.cd
[2] Doctoral school of the University of Burundi, Center for Research in Infrastructure, Environment and Technology (CRIET), University of Burundi, Bujumbura, Burundi; e-mail: jeremie.ndikumagenge@ub.edu.bi; longin.ndayisaba@ub.edu.bi; thierry.nsabimana@ub.edu.bi
[3] Management Computer Department, Institut Supérieur Pédagogique de Bukavu (ISP/Bukavu), PO Box 854, Bukavu, Democratic Republic of the Congo
* Corresponding Author: Philippe Boribo Kikunda

Abstract: In a context where students face increasingly complex academic choices, this work proposes a recommendation system based on Bayesian networks to guide new baccalaureate holders in their university choices. Using a dataset containing variables such as secondary school section, gender, type of school, percentage obtained, age, and first-year honors, we have constructed a probabilistic model capturing the dependencies between these characteristics and the option chosen. The data is collected at the Catholic University of Bukavu, the Official University of Bukavu, and the Higher Institute of Education of Bukavu, preprocessed and then used to learn the structure via the hill-climbing algorithm with the BIC score using R's bnlearn tool. The model enables us to estimate the probability that a candidate will choose a given stream, depending on their profile. The approach has been validated using metrics such as BIC, cross-validation, and bootstrap and offers a good compromise between interpretability and predictive performance. The results highlight the potential of Bayesian networks in constructing explainable recommendation systems in the field of academic guidance. The system produces orientation probability maps for each candidate, which can be used by enrollment service advisers, as well as an ordered list of options relevant to the candidate's profile. With a remarkable performance on a test sample of precision@k=0.85, recall@k=0.61, ndcg=0.8, and Map=0.88, it constitutes an effective lever for reducing the risk of being misdirected in universities in South-Kivu, in the Democratic Republic of Congo.

Keywords: Bayesian Network; Educational Data Mining; Hill Climbing Structure Learning; Personalized Recommendation; Probabilistic Graphical Model; Recommender System; Sensitivity Analysis.

Received: May, 9th 2025
Revised: June, 2nd 2025
Accepted: June, 8th 2025
Published: June, 11th 2025

1. Introduction

University guidance plays a crucial role in students' academic and professional careers. However, in many developing countries, students encounter major difficulties when choosing their university courses [1]. The lack of personalized academic advice or guidance units, limited access to information, and the influence of non-academic factors all contribute to guidance choices that are often poorly aligned with learners' real aptitudes and interests. In South Kivu, for example, the option choice is influenced by parents, friends, or the stream that offers the experience of a good job after studies. Other candidates are motivated solely by the race to obtain a degree by any means necessary [2]. This misdirection directly impacts performance, dropout rates, and the waste of educational resources [3]. Therefore, developing decision support systems based on reliable data is imperative to improve university guidance.

While several universities worldwide are adopting intelligent recommendation systems to guide students [4], this technology is still underdeveloped in the Congolese education system, particularly in the universities of South Kivu. The use of traditional methods is limited

to a handful of universities. What's more, traditional approaches from the educational sciences, psychology, and sociology are often based on intuition or subjective criteria. They cannot deal effectively with the growing complexity of students' academic, demographic, and behavioral data [5]. It is in this sense that intelligent, automated and scalable solutions capable of effectively assisting academic guidance decisions are enviable.

Intelligent recommendation systems, often based on artificial intelligence, generally fall into five main categories:

- Collaborative filtering (CF) recommender systems rely on techniques based on models (clustering, regression, etc.) or memory (user-element interactions) [6].
- Content-based (CB) recommender systems depend on information's characteristics [7].
- Knowledge-based recommender systems (KB) adopt techniques based on ontology, cases, or constraints [8].
- Demographic-based recommender systems (DF) depend on users' demographic data [9].
- Hybrid recommender systems represent a combination of two or more filtering techniques [10].

Several researchers [11]–[18] have integrated these approaches in realizing recommender systems for university majors, focusing on course-level recommendations. However, these approaches face specific challenges in the field of education, such as the cold-start problem, the absence or incompleteness of data, and the poor interpretability of results. In a context where the transparency of decisions is crucial, particularly for teachers and administrators, explicability becomes a fundamental criterion.

In this context, Bayesian Networks (BNs) are a particularly suitable approach. They make it possible to model the causal relationships between different variables (school, personal, etc.) while effectively managing missing data. Furthermore, BNs offer in-valuable interpretability thanks to their graphical structure, which explicit the dependencies between factors. Thus, the authors in [19] show that BN constitute a transformative tool in educational recommendation systems by remedying the limitations of traditional algorithms, promoting innovation, and personalizing learning experiences. They enable a more dynamic approach to recommendations, improving the adaptability and effectiveness of educational content delivery. Unlike black-box models like some deep learning algorithms, BNs allow recommendations to be explained, encouraging their acceptance in sensitive educational settings.

The main objective of this study is to design a recommendation system based on BNs capable of suggesting an optimal academic orientation from students' academic, demographic, and institutional data. The hypothesis is that the Bayesian approach will generate accurate recommendations, even in the presence of incomplete data, while retaining sufficient interpretability to be used as an academic decision-support tool.

The rest of the work is organized as follows. The next Section is dedicated to the literature review presenting the theory on Bayesian networks and a series of research works on this approach in education. The second Section describes our methodology for designing our recommendation system based on Bayesian networks. The fourth Section presents the main results obtained, followed by a discussion of current results in this field. Finally, the sixth Section presents a conclusion with some future perspectives.

2. Literature Review

Academic guidance is a major issue for education systems, as it directly influences academic success and professional integration. Several approaches have been proposed for predicting or recommending a course of study, including rule-based recommendation systems, classification techniques, and artificial intelligence.

Bayesian networks have been widely used to model uncertainties and make decisions. A guidance recommendation system is an area where uncertainty comes into play. Several studies and research projects have used this approach to provide a decision support tool. In this Section, we give a general overview of Bayesian networks and then look at some of the work that has used them in education.

2.1. Bayesian Networks

Bayesian networks are probabilistic graphical models that depict variables and their conditional dependencies using directed acyclic graphs (DAGs). Each node in the graph corresponds to a variable, while the edges represent the probabilistic relationships between these

variables [20]. This representation allows the joint probability distribution over all variables to be expressed as the product of the conditional distributions of each node given its parents.

If an edge is directed from node X to node$Y$, then X is referred to as the parent node of Y. In this context, X can be interpreted as exerting a probabilistic influence on Y—X may be viewed as the cause, and Y as the effect. However, this cause-effect relationship is not deterministic but rather probabilistic. Bayesian networks can compute the joint probability distribution of a set of random variables that describe a given phenomenon. For a list of random variables$X_1, X_2, \dots, X_n$, the joint probability is computed using Equation (1).

$$P(X_1, X_2, \dots, X_n) = P(X_n|X_{n-1}, \dots, X_1) \cdot P(X_{n-1}|X_{n-2}, \dots, X_1) \dots P(X_2|X_1) \cdot P(X_1) = \prod_{i=1}^{n} P\big(X_i|Parents(X_i)\big) \quad (1)$$

This holds under the condition that $Parents(X_i) \subseteq \{X_1, X_2, \dots, X_{i-1}\}$ which can be guaranteed by numbering the nodes in a topological order that respects the partial ordering defined by the DAG. Equation (1) faithfully represents the domain if each node is conditionally independent of all its non-parent predecessors, given its parent nodes. Constructing a Bayesian network involves two main steps: learning the network's structure and its parameters.

Bayesian networks (BNs) have become a powerful tool in education, offering innovative solutions to various challenges in teaching, learning, and educational research. These probabilistic models are particularly effective in dealing with uncertainty, capturing complex inter-variable relationships, and enabling personalized interventions. BNs have been used in educational settings to model student behavior, forecast academic performance, and identify the factors that influence learning outcomes [21].

BNs are especially well-suited to environments with incomplete or uncertain data, making them ideal for modeling complex educational contexts where data may be missing or ambiguous. Moreover, they provide interpretable results, essential for educators and policymakers making informed, data-driven decisions [22].

2.2 Related works

Bayesian networks have emerged as a foundational tool in educational research, with one of the earliest and most influential applications being student modeling. These probabilistic graphical models facilitate nuanced assessments of learners' knowledge state, skills, and developmental trajectories. For instance, Bayesian networks have been used in intelligent tutoring systems (ITS) to provide personalized feedback and adaptive learning experiences. The Andes tutoring system for Newtonian physics is a notable example, where BNs are employed to perform long-term knowledge assessment, plan recognition, and predict student actions [23].

Similarly, BNs have been applied to model student performance across multiple assessments, providing a holistic view of learning progress. This approach goes beyond traditional scoring systems by capturing the relationships between different assessments and identifying areas where students may need additional support [24].

In this spirit of providing intelligent teaching, other researchers have focused their studies on using BNs to detect students at risk of failing at universities. This allows timely interventions, such as additional support or personalized learning plans, to improve student outcomes [25].

Personalized learning represents another pivotal arena in which BNs have demonstrated considerable impact. BNs can recommend tailored learning paths and resources by modeling individual student preferences and learning behaviors. This approach has been shown to enhance innovation in education by moving beyond traditional recommendation algorithms that often perpetuate established preferences.

For example, a study on mobile learning and ethnomathematics used Bayesian networks to examine how cultural elements influence math learning. The results suggested incorporating cultural examples into mobile learning apps can improve student engagement and performance [26].

Moreover, BNs have been widely used to predict academic performance and identify factors influencing student success. A study using PISA data demonstrated that BNs can

achieve high accuracy (86.2%) in predicting scientific success, highlighting the importance of family-related variables in academic outcomes [21].

Similarly, BNs have been applied to model the relationship between pre-enrollment achievement and first-year university performance in STEM fields (Science, Technology, Engineering and Mathematics. This approach has been particularly useful in identifying at-risk students and informing early interventions [27].

Other authors in [28] try to solve the problem of rare data encountered by building a recommendation system by adopting a Bayesian approach. For example, the authors in [29] develop a Bayesian network that allows to characterize the relationship between knowledge points, to diagnose the mastery of a knowledge point by learners by monitoring their learning behaviors and their tests, and to adjust the planning of the learning content according to the differences between learners.

Bayesian networks have revolutionized various aspects of education, from student modeling and cognitive diagnosis to personalized learning and academic performance prediction. Their ability to handle uncertainty, provide interpretable results, and enable personalized interventions makes them a valuable tool for educators and researchers. As the field continues to evolve, integrating Bayesian networks with emerging technologies and their application to new domains will further enhance their impact on education.

However, few studies have focused on African educational contexts, especially in South Kivu in the Democratic Republic of Congo, where data is often heterogeneous, incomplete, and difficult to model. Our work sets itself apart by applying a Bayesian network to a real dataset from a Congolese university while considering contextual variables such as the type of secondary school, gender, or Section attended. Also, most of these works presented in the literature review are limited to predicting the probability of the event given a most likely observation. However, in this study, we propose an algorithm capable of ranking the probabilities of possible events by considering the quantity of data used by the network to calculate this probability.

3. Proposed Method

In Fig.1, we present the steps of the methodological procedure we followed to obtain our BNs. These steps can be divided into preprocessing and model-building operations. In the end, when we already have the BNs built, we can use them for the conditional calculation. An automatic filtering algorithm then processes the results to give the candidate a list of possible choices.

3.1. Preprocessing operations

Descriptive statistical analyses were first carried out to understand the input variables' distribution better. The following can be seen in Figure 2:

- In Figure 2a, the analysis reveals a female predominance in certain faculties, such as economics and law, while boys predominantly attend technical faculties such as agronomy, computer science, and medicine. This gender disparity may indicate a gender orientation as early as secondary school.
- In Figure 2d, most students are between 18 and 21, which is the typical age for university entry. However, the presence of older applicants may reflect less linear educational paths or resumed studies. Particularly in the Faculty of Medicine, we observe a high proportion of outliers by age compared with other faculties in Figure 2b.
- In Figure 2c, the analysis shows a strong concentration of high percentages among students from 'itfm' followed by those from college, while those from 'school complex' and other schools obtain low marks on average in the first year at university. This indicates a gap in the quality of pre-university preparation.
- In Figure 2e, the analysis reveals that those who studied biochemistry mainly went into medicine and agronomy. On the other hand, those who had done the business section tended to go into economics and computing. We also note that many students from the pedagogy section go on to all faculties.
- Finally, in Figure 2f, most students from the biochemistry section obtain good grades compared with the other sections.

These distributions justify the need for a recommendation system capable of taking into account variables such as gender, type of school of origin, section, and past performance to guide students toward courses suited to their profile.

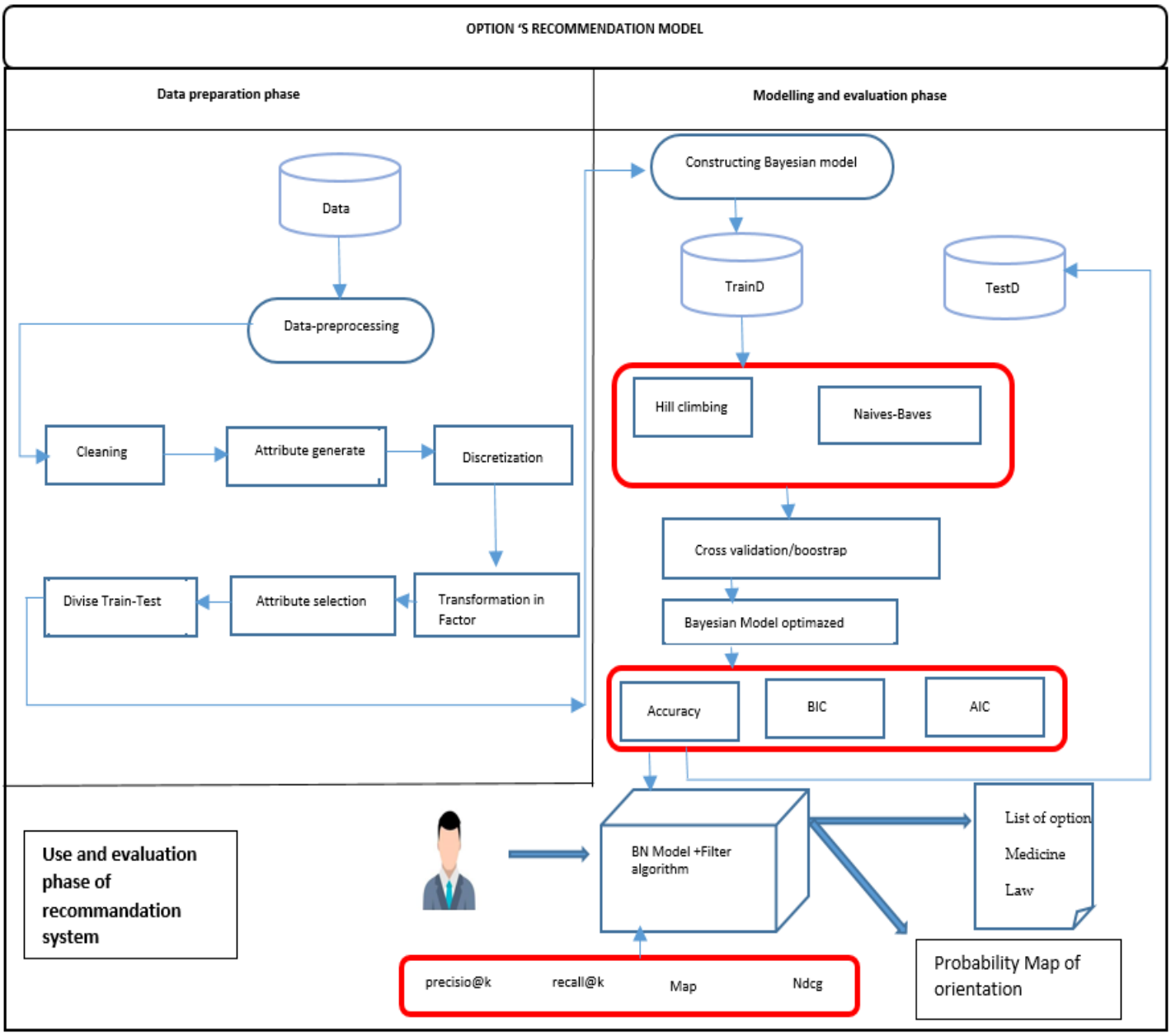

Figure 1. The framework of constructing Bayesian model

In order to learn an efficient Bayesian model, several preparatory steps were carried out:

- Encoding of categorical variables: all textual variables were converted into categories for processing by the model.
- Discretization of continuous variables:
- The secondary percentage was discretized into four classes:
  Passable (50-55%), Good (55-65%), Very Good (65-75%), Excellent (75-90%).
- The first year percentage has been broken down into four classes: Passable (50-55%), Good (55-65%), Very Good (65-75%), and Excellent (75-90%)
- Age was broken down into three categories: young (16-18 years), normal (18-22 years) and advanced (22-30 years).
- Data cleaning: Remove duplicates and manage missing values.

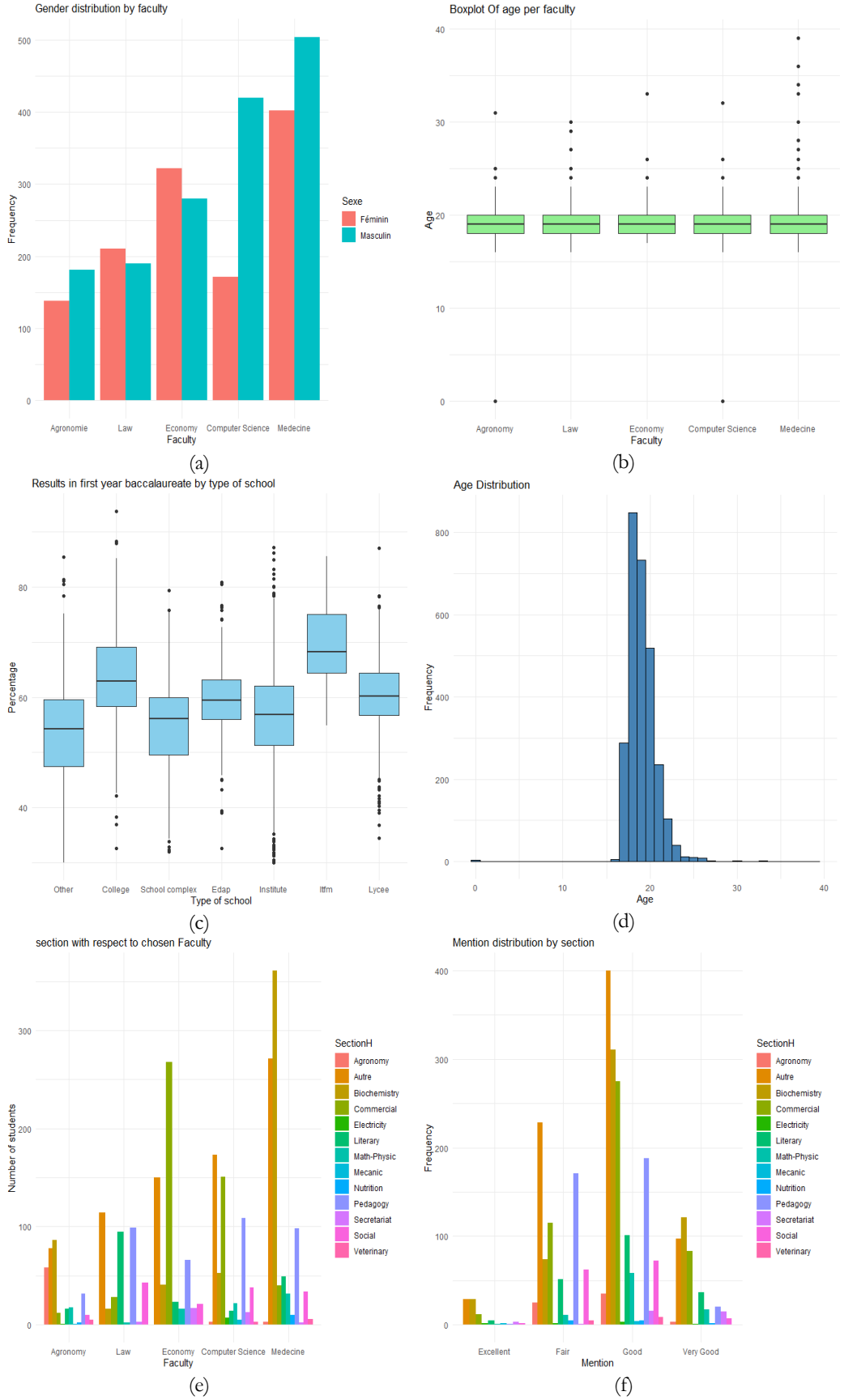

Figure 2. Distributions (**a**) Gender Distribution by Faculty, (**b**) Box of Age by Faculty, (**c**)Results in first-year baccalaureate by type of school, (**d**) Age distribution, (**e**) Section concerning chosen Faculty Distribution, (**f**) Mention Distribution by Section

3.2. Processing operations

After these preprocessing operations, the Bayesian network is constructed. Once the optimized Bayesian network has been obtained, it is used to predict the study program that offers the highest probability given the candidate's profile. The level will make this prediction of success, i.e., Passable, Good, Very Good, Excellent, and the probabilities of the choice of each study program will be calculated. These results will be arranged in a two-dimensional table and then sorted by our proposed algorithm.

3.3. Algorithm for estimation of the structure

The basic idea of this algorithm is to start with an initial solution, then explore neighboring solutions and move towards the best one if it improves the objective (the score).

Algorithm 1. Hill Climbing

```
INPUT: Initial_state: starting state
OUTPUT: best_state: local optimum
Begin
    best_state ←initial_state
    repeat
        Neighbor← best_neigbor of best_state(according to evaluation function)
        If evaluation(neighbor)<= evaluation(best_state) then
            return best_state // local optimum reached
        Else
            Best_state←neighbor
        End if
    Until convergence or stopping condition
end
```

A neighbor of a Bayesian graph is a structure obtained from the current structure by making an elementary modification, such as adding an edge, removing an edge or reversing an edge. Each neighbor must meet the non-cyclicity condition (DAG). The Equation (2) calculates BIC score as follows:

$$BIC = \log L - \frac{k}{2}\log n \tag{2}$$

Where $\log L$ is the model's log-likelihood; k is the total number of parameters; and n is the sample size.

Once the structure of the Bayesian network has been estimated, it can be used to calculate the probability of an event given an observation. Several algorithms can be used: inference by enumeration, eliminating variables, clustering algorithms, direct sampling methods, inference by Monte Carlo Markov chain simulation (MCMC), etc.

3.4. Algorithm for estimating conditional probabilities

To estimate the conditional probabilities of a node in a Bayesian network, we determine its Conditional Probability Table (CPT), that is, the probability distribution of the node given all possible combinations of the states of its parent nodes. When a complete dataset is available (i.e., all variables in the network are fully observed), the conditional probabilities can be estimated using relative frequencies, a method known as parametric learning.

Let X be the target node, and $Pa(X)$ the set of its parent nodes. For each possible configuration pa of the parents, the conditional probability is calculated using the Equation (3).

$$P(X = x|Pa(X) = pa) = \frac{N(X = x, Pa(X) = pa)}{N(Pa(X) = pa)} \tag{3}$$

Where $N(X = x, Pa(X) = pa)$ is the number of instances in the data where $X = x$ and $Pa(X) = pa$; $N(Pa(X) = pa)$ is the number of instances where the parents take the configuration pa.

This method is simple and intuitive, but it requires sufficient data points for each configuration of the parent variables to ensure reliable estimates.

When the dataset contains missing values or unobserved variables, the Expectation-Maximization (EM) algorithm can be used to estimate the CPTs. The algorithm iteratively alternates between two steps:

- E-step (Expectation): Estimate missing data or the posterior distribution of hidden variables based on the current parameters.
- M-step (Maximization): Update the parameters (CPTs) to maximize the expected likelihood from the E-step.

Finally, without sufficient data, expert knowledge can be used to estimate the conditional probabilities manually. Experts provide subjective probability assessments based on domain knowledge, which are then encoded into the CPTs.

3.5. Comparison of Approaches

In the context of solving our problem, which consists of developing a recommendation system for a university option, the choice of machine learning method is crucial, both for the quality of the results obtained and for the interpretability and robustness of the model. This section compares several approaches commonly used in machine learning, such as decision trees and Naïve Bayes versus Bayesian networks.

Table 1. Predictive accuracy of faculty assignments using different models.

Model	Accuracy
Naïve Bayes	0.4781
Decision Tree	0.4579
Bayesian Network	0.6987

As shown in Table 1, the Bayesian network performs better than the decision tree and Naïve Bayes models. This is attributed to its ability to effectively model the probabilistic dependencies between variables, resulting in higher predictive accuracy. Additionally, Bayesian networks offer enhanced interpretability, a key advantage in educational decision-support contexts. In addition, the Bayesian network behaves very well when faced with incomplete data, allowing the expert's knowledge to be taken into account in the form of probabilities and thus offering a better understanding of the decision-making process.

4. Results and Discussion

In this section we will present the Bayesian network obtained using the escalation algorithm on the sample. Our sample includes 2818 pieces of information on the type of school of origin, gender, age, the section taken at secondary school, the Faculty chosen, the percentage obtained at the end of secondary school tests, and the percentage obtained in the first year at university. First, we present the structure of the Bayesian network, followed by the conditional probability table for the 'Faculty' node.

4.1. Bayesian network structure

After carrying out the preprocessing operations detailed in the Procedure section, we applied the escalation algorithm to discover the Bayesian network structure shown in Figure 3. This network was constructed to model the probabilistic relationships between the students' characteristics.

In Figure 3, the nodes represent the variables, and the directed arcs indicate conditional dependencies. We observe that the choice of Faculty at university is directly influenced by gender, type of school of origin, percentage in secondary school leaving exams, age, and section. This node will attract our attention, as it is the one on which we make our inferences in our queries to calculate the probability of success of a candidate based on the profile they present.

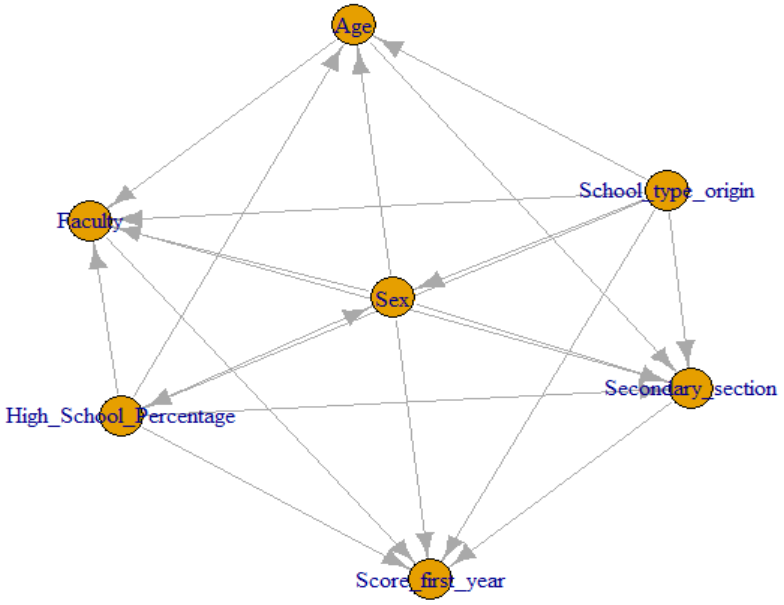

Figure 3. Bayesian network using the escalation algorithm

We then applied cross-validation and used the boostrap to detect persistent arcs. After optimization, we found the following structure at Figure 4.

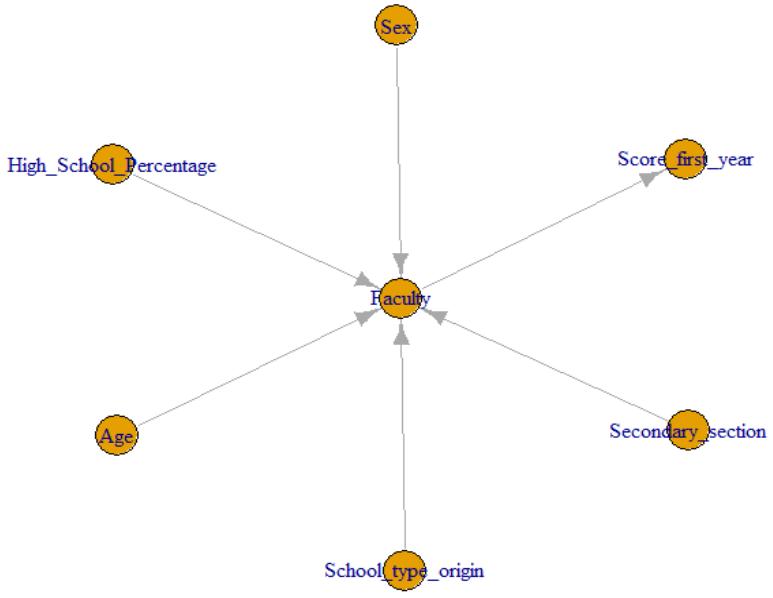

Figure 4. Bayesian network structure optimized

4.2. Evaluation of the BN Model

The model evaluation is done on a test data set that was not used to train the model. We calculated the model accuracy using Bayesian information criterion (BIC) and Akaike information criterion (AIC). As shown in Table 2, the optimized model presents higher values than the optimized model.

Table 2. Metrics for Bayesian model evaluation.

Model	BIC	AIC	Accuracy
Model at Figure 3	-105820	-41319.65	0.63
Model at Figure 4	-159961	-56676.31	0.70

4.3. Conditional probability table for the Faculty node

The faculty node is influenced by five variables, namely the type of secondary school the student attended, gender, age, final secondary school percentage, and academic Section. The number of conditional probabilities required at this node is determined by combining all possible values for each influencing variable.

Let n be the total number of conditional probabilities to be computed, with:

- F is the number of categories of the variable Faculty,
- T is the number of terms for the variable Type of school,
- S is the number of terms for the variable Sex,
- P is the number of terms for the variable Percentage,
- Se is the number of items in the variable Section

The total number of conditional probabilities is then given by:

$$n = F \times T \times S \times P \times Se \tag{4}$$

Given that F=5, T=7, S=2, P=4, and Se=13, we obtain n=10920.

Table 3. Partial view of the conditional probability table of the Faculty node.

School_type_orig	Faculty	Sex	Age	Section	% secondary	Prob
College	Medicine	M	young	biochemistry	excellent	0.666631
School complex	Medicine	M	young	biochemistry	excellent	0.2
Edap	Medicine	M	young	biochemistry	excellent	0.2
Institute	Medicine	M	young	biochemistry	excellent	0.999634
…	…	…	…	…	…	…

It is therefore difficult to present everything here, but we can present just a part of it, for example the conditional probability table for the Faculty node, with College as the school type, Medicine as the selected Faculty, and Section as the university choice, as shown in Table 3. The Equation (5) computes this probability.

$$Prob(Faculty = Medicine | School\_type\_origin, Sex, Age, High\_School\_Percentage, Section) \tag{5}$$

We can see, for example, the probability of choosing Medicine, given that the candidate has done the Biochemistry section at an institute-type secondary school, is male, no more than 18 years old, and had a percentage of at least 75% in the secondary school leaving exams, is 0.99. This probability is calculated using the Equation (6).

$$Prob(Faculty = Medicine | School\_type\_origin = institute, Sex = M, Age = young, High\_School\_Percentage = Excellent, Section = Biochemistry) = 0.99 \tag{6}$$

This probability table can be converted into the form of a map shown in Figure 5. On the x-axis, we see the different options symbolized by A = Agronomics, E = Economics, D = Law, C = Computer Science, and M = Medicine. On the x-axis at the top, we find two types of information: a line for the percentage levels at the secondary level and another line for the percentage level expected to be achieved in the first year at university.

On the y-axis, on the left, we find gender, and on the right, we find the type of school (institute, college, school complex, ITFM, etc.). Given the combination of information on these different axes, the blue colors indicate a high probability. If it is white, this means that the probability is very low.

For example, the blue rectangle is in the bottom left-hand corner. This rectangle is found in the combination of Medicine, Male, and Female, Excellent and Weak, and coming from a 'Lycee '-type school. This can be interpreted as follows: regardless of the options chosen at secondary school, if you come from a 'Lycee' school and achieved a high percentage at secondary level, the probability of going into Medicine and obtaining a 'passable' percentage is high. This shows that the percentage alone is insufficient to decide on a career.

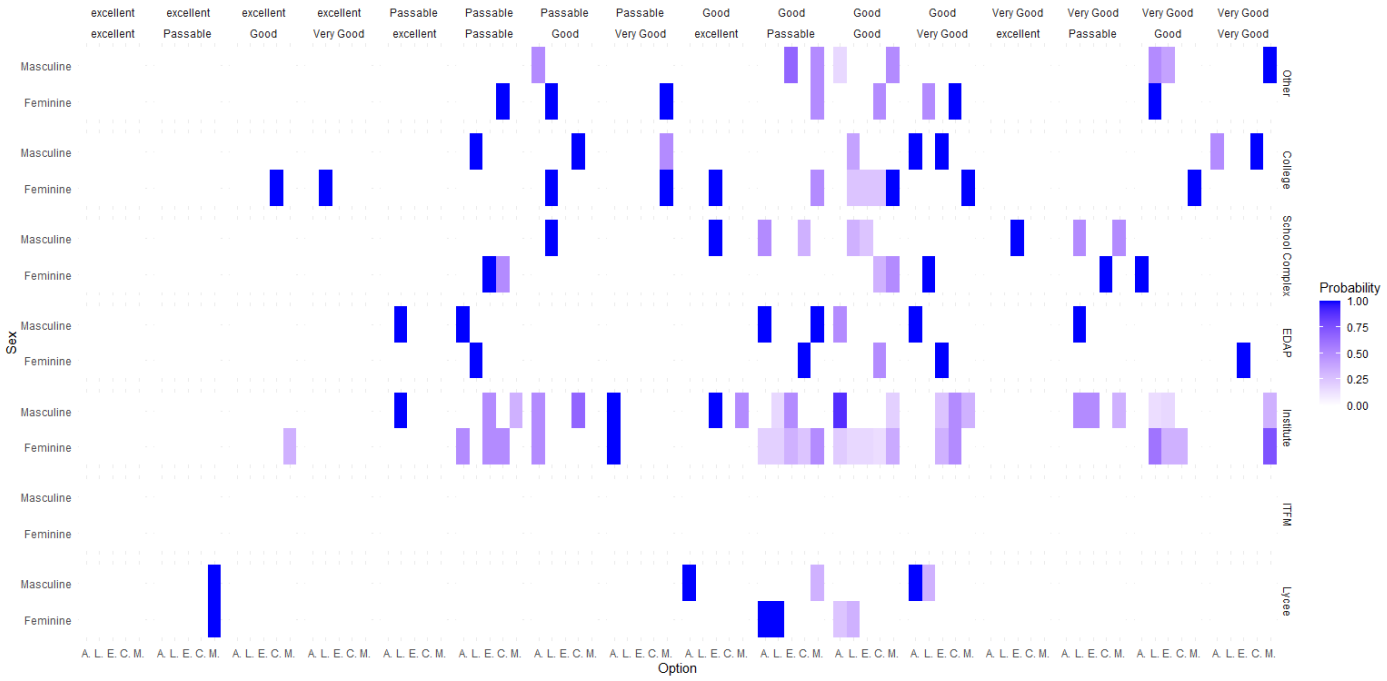

Figure 5. Choice of option according to the school of origin, percentage at secondary school, expected percentage at university, and gender

4.4. Algorithm proposed for arranging resultats of recommendation

After calculating the conditional probabilities of the event given the observation by the Bayesian network constructed, we store the data in a two-dimensional t table. To display the result of the recommendation, we propose the following algorithm:

- scan each line of the Table 4;
- detect the maximum in that row;
- record the corresponding column name;
- sort then the complete list of results;
- display the results in descending order.

In pseudocode, we present the algorithm below.

Algorithm 2. Arrange List Recommendation

```
INPUT: Array1[4][5]: real     // values of array
          faculty[5]: strings["Agro","Law","Economy","computer science", "medicine"]
OUTPUT: results: list of pair (name of column, max_value); I, j, index_max :integer;
Max_value: real
 1:    begin
 2:    array1←array[[0.2,0.2,0.2,0.2],[0.16,0.16.0.0,0.67],[0.0,0.0,0.49,0.0,0.49],
 3:                              [0.99,0.0,0.0,0.0,0.0]]
 4:    Results←[]
 5:    For I =0 to 3 do
 6:        Max_val←0
 7:     For j=1 to 4 do
 8:          If tableau[i][j]>max_val then
 9:                Max_value←Tableau[i][j]
 10:               Index_max← j
 11:         End if
 12:    End For
 13:    For each (name, Val) in results
 14:         print name+…. +val
 15:  end
```

4.5. Evaluation Metrics for recommendation system

To evaluate our recommendation system, the following metrics were used:

- Precision@k: proportion of the first k items recommended that are relevant
- Recall@k: proportion of relevant items found in the first k items.
- Ndcg@k (normalized Discounted Cumulative Gain): considers the position in the ranking. The higher a relevant element is, the better.
- MAP (Mean Average Precision): average precision calculated at each position where a relevant item is found.

For k=1 and as number of elements (max 2, i.e., number of expected options), our system performed Precision@1=0.85, Recall@1=0.61, Ndcg@=0.8, Map=0.88. These values show that the system performs well and is potentially suitable for the recommendation task for which it was designed. The precision@k should, in principle, exceed the accuracy of the Bayesian model used in an autonomous predictive framework, which was 70%.

4.6. Application to Examples and Sensibility Analysis

Let us consider the case of a student with the following profile:
Example 1:

- Type of school: College
- Sex: Male
- Age: Normal (18–22)
- Humanities section: Biochemistry
- Percentage in secondary school: Very Good (65–75)

We use the Bayesian network to generate a list of recommended university options for this student by varying the expected level of success in their first year. The following four questions are analyzed:

Question 1. Can this student achieve a passable percentage at university, and in which option? To answer this, we compute:

$$Prob(Faculty|Typeschool = college, Sex = M, Age = normal, \text{High\_School}\_Percentage = very\ Good, Section = Biochemistry, Score\_first\_year = Passable) \quad (7)$$

Based on the data from the first row of Table 4 there is no faculty for which this student will likely achieve only a passable grade. The uniform probability value of 0.2 is due to the default Bayesian estimation that avoids zero probability by smoothing in the absence of supporting data.

Table 4. Conditional probability values are based on the candidate profile (Example 1).

	Agro	Law	Economy	Computer Science	Medicine
Eq. (7)	0.2	0.2	0.2	0.2	0.2
Eq. (8)	0.16	0.16	0.0	0.0	0.67
Eq. (9)	0.0	0.0	0.49	0.0	0.49
Eq. (10)	0.99	0.0	0.0	0.0	0.0

Question 2. Can this student achieve a good percentage at university, and in which option? We compute:

$$Prob(Faculty|Typeschool = college, Sex = M, Age = normal, High\_School\_Pourcentage = very\ Good, Section = Biochemistry, Score\_first\_year = Good) \quad (8)$$

The highest probability in this case is for Medicine, suggesting this option for achieving a good result.

Question 3. Can this student achieve a very good percentage (65–75) at university, and in which option? We compute:

$$Prob(Faculty|Typeschool = college, Sex = M, Age = normal, High\_School\_Percentage = very\ Good, Section = Biochemistry, score\_first\_year = very\ Good) \quad (9)$$

The options most likely to lead to a very good result are Economics and Medicine.

Question 4. Can this student achieve an excellent percentage (>75), and in which option? We compute:

$$Prob(Faculty|Typeschool = college, Sex = M, Age = normal, High\_School\_Percentage = very\ Good, Section = Biochemistry, Score\_first\_year = excellent) \quad (10)$$

In this case, Agronomy is the most probable choice for achieving an excellent percentage. Thus, based on all computed conditional probabilities, this student's recommended list of study options is Agronomy, Medicine, Economics, Computer Science, and Law. This result is also illustrated in Figure 6.

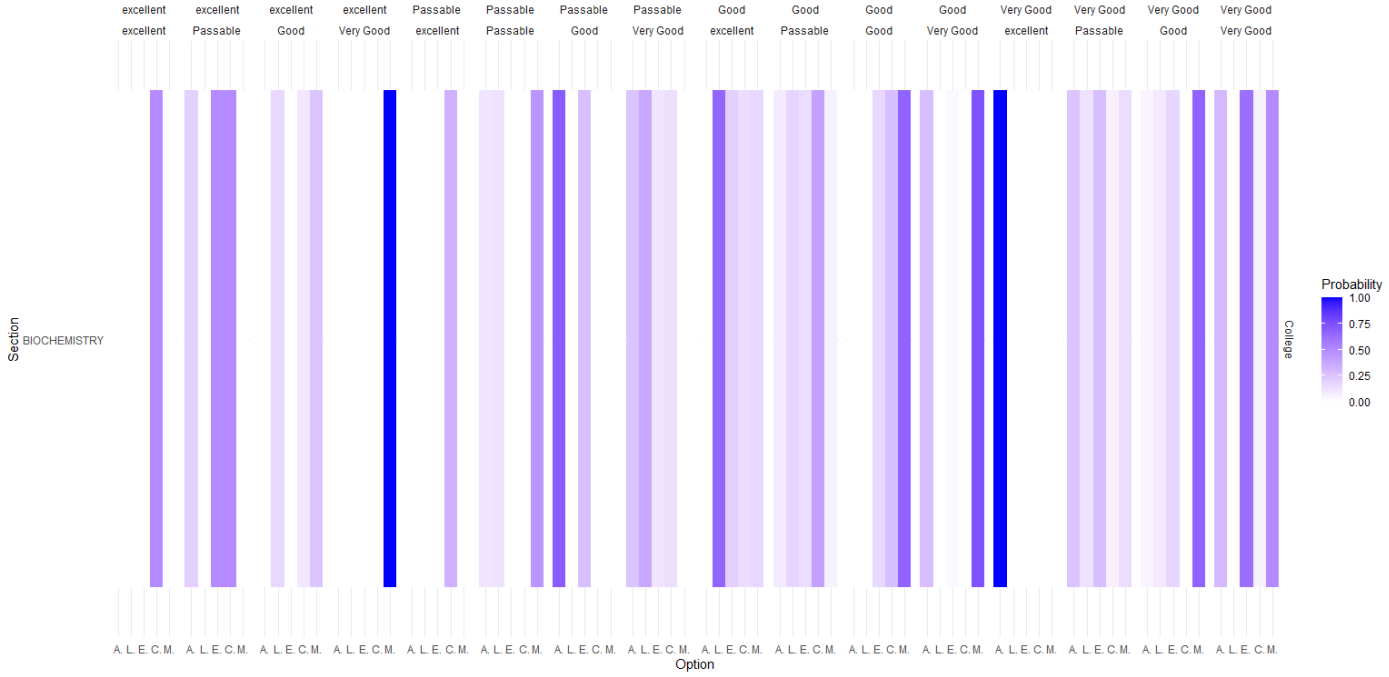

Figure 6. Probability map for the choice of profile1 option

In Figure 6, the x-axis represents the five options symbolized by A = Agronomy, L = Law, E = Economy, C = Computer Science, and M = Medicine. The first line at the top corresponds to secondary school performance (e.g., Excellent, Passable, Good, Very Good), while the second line corresponds to expected university performance levels. On the y-axis, section is displayed on the left, and type of school is on the right. For Example 1, the student is associated with the Very Good label. In this region, Agronomy is highly likely to be selected, followed by Medicine (associated with an expected good result), and then Economics. To analyze the impact of variations in the input profile, we modify certain attributes from Example 1 and record the recommended options. The outcomes of these variations are summarized in Table 5.

The sensitivity analysis shows that certain variables strongly influence the output probabilities. For instance, comparing Example 1 with Example 4 shows that a higher secondary percentage increases the likelihood of being recommended for Medicine. In contrast, changing the type of school (Example 1 vs. Example 3) has minimal impact on the recommendation for Medicine.

Meanwhile, modifying the Section from Biochemistry to Commercial (Example 1 vs. Example 2) and changing the gender from Male to Female (Example 1 vs. Example 5) significantly shifts the probability toward Computer Science. This analysis shows that students

from the Biochemistry section are more likely to be recommended for Medicine or Agronomy. A higher percentage score further strengthens the recommendation for Medicine, which aligns with prior findings [30], [31].

Table 5. Profile variations and recommendation outcomes.

Example	Type of School	Sex	Age	Section	Percent	Recommended Options
1	College	Male	18–22	Biochemistry	Very Good	Agronomy, Medicine, Economics, CS, Law
2	College	Male	18–22	Commercial	Very Good	CS, Economics, Law, Medicine, Agronomy
3	Institute	Male	18–22	Biochemistry	Very Good	Medicine, Agronomy, Economics, CS, Law
4	College	Male	18–22	Commercial	Excellent	Medicine, Agronomy, CS, Economics, Law
5	College	Female	18–22	Biochemistry	Very Good	CS, Medicine, Agronomy, Economics, Law

On the other hand, students from the Commercial Section tend to be recommended for Economics or Computer Science. If they also come from a College-type school, the likelihood of Economics increases slightly. Although the Bayesian network provides recommendations based on high probability estimates derived from observed data, it does have limitations. First, the model does not consider personal preferences, motivations, or career aspirations. For instance, a high probability for Economics does not imply it is the most suitable program for a student interested in Law. As shown in Figure 6, a student with the same profile as Example 1 but with an average percentage could still succeed in Law with an excellent first-year performance. When a student's personal goals align with the model's recommendations, the results are optimal.

Another key limitation concerns potential bias in the dataset. The recommendations may not reflect broader contexts if certain schools or academic streams are underrepresented, as is the case for 'ITFM' in Figure 5. Therefore, continuous model updates are necessary as more data becomes available.

5. Discussion

The results of applying the Bayesian network in the context of university guidance show an interesting ability to model the conditional dependencies between students' socio-academic characteristics and their choice of university options. This probabilistic approach made it possible not only to predict the most likely choices for a new student but also to analyze the relative influence of each variable, such as the type of institution, the section taken, or the percentage obtained in the state exam.

Using robust estimation techniques such as BIC and cross-validation enhanced the reliability of the predictions. However, we observed that some unstable arcs revealed by the bootstrap should be removed to avoid misinterpretations. This highlights the importance of not over-learning the structures, particularly when the data is incomplete or unbalanced. Compared with traditional recommendation systems (collaborative or content-based filtering), the Bayesian network has the advantage of being interpretable and integrating a priori knowledge, including in the form of an ontology. It, therefore, offers a transparent and explainable solution, which is essential in a field as sensitive as educational guidance.

However, there are still limitations. One of the significant challenges is the availability of complete and balanced data. Some variables did not have enough representative levels, which sometimes led to the exclusion of data in the calculation of scores or zero probabilities. In addition, the learned structures are highly dependent on the parameter settings which justifies the importance of a sensitivity analysis. Finally, although the model has been shown to perform well, its use in practice will require a user-friendly interface for educational advisers and tests on other cohorts to assess its robustness and generalizability.

6. Conclusions

This work proposed an approach based on Bayesian networks to model and predict the choice of a university option based on characteristics derived from students' secondary education. Experimental results show that this approach can capture the dependency relationships between variables in an explicable way while providing a coherent probabilistic recommendation system. There are several methods for creating a recommendation system. Among these methods, the hybrid approach still shows better results than other approaches. An innovation in this study is that we have created a hybrid recommendation system based on the Bayesian network. This work thus provides a methodological framework for further studies on using Bayesian networks in the recommendation system, particularly in the Congolese education system.

The practical contribution of this study lies in its potential to support educational institutions and counselors in guiding students toward the most suitable academic options. The model's ability to generate probabilistic predictions by map or algorithm allows for data-driven decision-making, offering an evidence-based approach for students when making critical career choices. Then, when applying for admission and enrolment, an orientation probability card may be printed out for the student. However, given that other relevant variables such as the candidate's interest, career objective, and other environmental variables (parents 'profession, parents' level of education, family standard of living, etc.) were not taken into account, the results obtained should not be considered as a stand-alone solution but rather as a tool to support decision-making.

Thus, one of the potential developments of this study would be to develop ontology-based models to provide even more precise and contextual recommendations. By structuring knowledge hierarchically and relationally, ontology would enable us to understand better the complex interrelationships between the factors influencing student choice (such as academic backgrounds, interests, and career aspirations). This would allow further personalizing recommendations based on contextual data and shared knowledge in a specific area of expertise.

Another potential development of this study lies in extending the predictive approach by transposing it into a web-based system. By integrating this model into an online interactive platform, students could easily consult personalized recommendations on their academic orientation. Such a system could collect real-time information on student profiles, update predictions based on new data (other options, for example), and provide detailed advice on pathway options through a user-friendly interface.

Finally, an expert system could be designed to automate decision-making in academic guidance contexts. Such a system could be based on decision rules derived from the Bayesian model developed in this study, enriching them with an interface allowing academic advisors to interact with the system's recommendations, thus adjusting the results according to the specific needs of students. The solution presented in this study will limit the risk of Congolese students being misdirected at university, which causes failure and dropout in the early years.

Author Contributions: Conceptualization: Philippe Boribo Kikunda; Methodology: Philippe Boribo Kikunda, Thierry Nsabimana and Longin Ndayisaba.; Software: Thierry Nsabimana.; Validation: Philippe Boribo Kikunda., Jérémie Ndikumagenge and Longin Ndayisaba.; Formal analysis: Philippe Boribo Kikunda.; Investigation: Philippe Boribo Kikunda.; Resources: Philippe Boribo Kikunda; Data curation: Philippe Boribo Kikunda; Writing—original draft preparation: Philippe Boribo Kikunda.; Writing—review and editing: Philippe Boribo Kikunda.; Visualization: Thierry Nsabimana and Longin Ndayisaba.; Supervision: Jérémie Ndikumagenge and Thierry Nsabimana. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: data is available at https://github.com/philbor87/article.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] Z. Keshf and S. Khanum, "Career Guidance and Counseling Needs in a Developing Country's Context: A Qualitative Study," *Sage Open*, vol. 11, no. 3, p. 21582440211040120, Jul. 2021, doi: 10.1177/21582440211040119.

[2] M. Poncelet and P. Kapagama Ikando, "Private Higher Education in a Post-Abdication State: (In)Governance and Inequality in the Democratic Republic of Congo," in *Private Higher Education and Inequalities in the Global South: Lessons from Africa, Latin America and Asia*, E. Gérard, Ed. Cham: Springer Nature Switzerland, 2024, pp. 223–268. doi: 10.1007/978-3-031-54756-0_7.

[3] P. Perchinunno, M. Bilancia, and D. Vitale, "A Statistical Analysis of Factors Affecting Higher Education Dropouts," *Soc. Indic. Res.*, vol. 156, no. 2–3, pp. 341–362, Aug. 2021, doi: 10.1007/s11205-019-02249-y.

[4] A. Pantoja-Vallejo and B. Berrios-Aguayo, "TIMONEL: Recommendation System Applied to the Educational Orientation of Higher Education Students," in *Innovation in Information Systems and Technologies to Support Learning Research*, M. Serrhini, C. Silva, and S. Aljahdali, Eds. Springer International Publishing, 2020, pp. 14–26. doi: 10.1007/978-3-030-36778-7_2.

[5] C. R. I. McAdams and V. A. Foster, "Promoting the Development of High-Risk College Students Through a Deliberate Psychological Education-Based Freshman Orientation Course," *J. Freshm. Year Exp. Students Transit.*, vol. 10, no. 1, pp. 51–72, 1998.

[6] J. Ben Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The Adaptive Web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 291–324. doi: 10.1007/978-3-540-72079-9_9.

[7] D. Wang, Y. Liang, D. Xu, X. Feng, and R. Guan, "A content-based recommender system for computer science publications," *Knowledge-Based Syst.*, vol. 157, pp. 1–9, Oct. 2018, doi: 10.1016/j.knosys.2018.05.001.

[8] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artif. Intell. Rev.*, vol. 50, no. 1, pp. 21–48, Jun. 2018, doi: 10.1007/s10462-017-9539-5.

[9] L. Wenzhe and S. V Grigorev, "Implementation of the Demographic-Based Recommendation Algorithm Using Big Data," in *2022 11th International Conference on Information Communication and Applications (ICICA)*, Jun. 2022, pp. 1–5. doi: 10.1109/ICICA56942.2022.00007.

[10] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," *User Model. User-adapt. Interact.*, vol. 12, no. 4, pp. 331–370, May 2002, doi: 10.1023/A:1021240730564.

[11] M. Amane, K. Aissaoui, and M. Berrada, "ERSDO: E-learning Recommender System based on Dynamic Ontology," *Educ. Inf. Technol.*, vol. 27, no. 6, pp. 7549–7561, Jul. 2022, doi: 10.1007/s10639-022-10914-y.

[12] O. EL AISSAOUI and L. OUGHDIR, "A learning style-based Ontology Matching to enhance learning resources recommendation," in *2020 1st International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET)*, Apr. 2020, pp. 1–7. doi: 10.1109/IRASET48871.2020.9092142.

[13] S. Assami, N. Daoudi, and R. Ajhoun, "Ontology-Based Modeling for a Personalized MOOC Recommender System," in *Information Systems and Technologies to Support Learning*, Á. Rocha and M. Serrhini, Eds. Springer International Publishing, 2019, pp. 21–28. doi: 10.1007/978-3-030-03577-8_3.

[14] L. Romero, C. Saucedo, M. L. Caliusco, and M. Gutiérrez, "Supporting self-regulated learning and personalization using ePortfolios: a semantic approach based on learning paths," *Int. J. Educ. Technol. High. Educ.*, vol. 16, no. 1, p. 16, Dec. 2019, doi: 10.1186/s41239-019-0146-1.

[15] S. Bhaskaran, R. Marappan, and B. Santhi, "Design and Analysis of a Cluster-Based Intelligent Hybrid Recommendation System for E-Learning Applications," *Mathematics*, vol. 9, no. 2, p. 197, Jan. 2021, doi: 10.3390/math9020197.

[16] J. C. S. Perez, R. F. Manrique, O. Mariño, M. L. Vásquez, and N. Cardozo, "A course hybrid recommender system for limited information scenarios | Journal of Educational Data Mining," *J. Educ. Data Min.*, vol. 14, no. 3, pp. 162–188, May 2022, doi: 10.5281/zenodo.7304829.

[17] N. Kamal, F. Sarker, A. Rahman, S. Hossain, and K. A. Mamun, "Recommender System in Academic Choices of Higher Education: A Systematic Review," *IEEE Access*, vol. 12, pp. 35475–35501, May 2024, doi: 10.1109/ACCESS.2024.3368058.

[18] K. K. San, H. H. Win, and K. E. E. Chaw, "Enhancing Hybrid Course Recommendation with Weighted Voting Ensemble Learning," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 337–347, Jan. 2025, doi: 10.62411/faith.3048-3719-55.

[19] Z. Tagdimi, I. Amzil, Y. Jdidou, and S. Aammou, "Enhancing Innovation Through Bayesian Networks in Recommender Learning Paths," in *Technological Tools for Innovative Teaching*, IGI Global Scientific Publishing, 2023, pp. 277–291. doi: 10.4018/979-8-3693-3132-3.ch014.

[20] F. V Jensen and T. D. Nielsen, Eds., "Learning the Structure of Bayesian Networks," in *Bayesian Networks and Decision Graphs: February 8, 2007*, New York, NY: Springer, 2007, pp. 229–264. doi: 10.1007/978-0-387-68282-2_7.

[21] H. A. KARABOĞA and İ. DEMİR, "Examining the factors affecting students' science success with Bayesian networks," *Int. J. Assess. Tools Educ.*, vol. 10, no. 3, pp. 413–433, Sep. 2023, doi: 10.21449/ijate.1218659.

[22] S. Saeedi, D. Božanić, and R. Safa, "Strategic Analytics for Predicting Students' Academic Performance Using Cluster Analysis and Bayesian Networks," *Educ. Sci. Manag.*, vol. 2, no. 4, pp. 197–214, Dec. 2024, doi: 10.56578/esm020402.

[23] C. Conati, A. Gertner, and K. VanLehn, "Using Bayesian Networks to Manage Uncertainty in Student Modeling," *User Model. User-adapt. Interact.*, vol. 12, no. 4, pp. 371–417, May 2002, doi: 10.1023/A:1021258506583.

[24] J. Xu and N. Dadey, "Using Bayesian Networks to Characterize Student Performance across Multiple Assessments of Individual Standards," *Appl. Meas. Educ.*, vol. 35, no. 3, pp. 179–196, Jul. 2022, doi: 10.1080/08957347.2022.2103134.

[25] J. Tarbes, P. Morales, M. Levano, P. Schwarzenberg, O. Nicolis, and B. Peralta, "Explainable Prediction of Academic Failure Using Bayesian Networks," in *2022 IEEE International Conference on Automation/XXV Congress of the Chilean Association of Automatic Control (ICA-ACCA)*, Oct. 2022, pp. 1–6. doi: 10.1109/ICA-ACCA56767.2022.10006086.

[26] J. D. Johnson, L. Smail, D. Corey, and A. M. Jarrah, "Using Bayesian Networks to Provide Educational Implications: Mobile Learning and Ethnomathematics to Improve Sustainability in Mathematics Education," *Sustainability*, vol. 14, no. 10, p. 5897, May 2022, doi: 10.3390/su14105897.

[27] M. Baranyi, K. Gal, R. Molontay, and M. Szabo, "Modeling Students' Academic Performance Using Bayesian Networks," in *2019 17th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, Nov. 2019, pp. 42–49. doi: 10.1109/ICETA48886.2019.9040067.

[28] X. Wang, V. Hoo, M. Liu, J. Li, and Y. C. Wu, "Advancing legal recommendation system with enhanced Bayesian network machine learning," *Artif. Intell. Law*, Nov. 2024, doi: 10.1007/s10506-024-09424-8.

[29] X. Feng, "Research on the Design of Recommendation System for Learning Methods Based on Bayesian Networks," in *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Nov. 2023, pp. 1239–1243. doi: 10.1109/ACAIT60137.2023.10528544.

[30] L. M. Mwandigha, P. A. Tiffin, L. W. Paton, A. S. Kasim, and J. R. Böhnke, "What is the effect of secondary (high) schooling on subsequent medical school performance? A national, UK-based, cohort study," *BMJ Open*, vol. 8, no. 5, p. e020291, May 2018, doi: 10.1136/bmjopen-2017-020291.

[31] R. Igić, "Preparations of Students for Enrollment in Medical Schools," *J. Med. Educ. Curric. Dev.*, vol. 11, p. 23821205241264696, Jan. 2024, doi: 10.1177/23821205241264698.