Masters Dissertation

Developing a Bayesian Network Model to Predict Students' Performance Based on the Analysis of their Higher Education Trajectory

---

Thabo Victor Ramaano (1134539)

Supervised by **Dr Ashwini Jadhav** and **Prof. Ritesh Ajoodha**

University of the Witwatersrand, Johannesburg
School of Computer Science and Applied Mathematics

August 22, 2024

Declaration

I, **Thabo Ramaano** with student number **1134539** declare that this research report is my own work. This work is submitted for the degree of Master of Science in Computer Science (Research) at the University of the Witwatersrand, Johannesburg. This work has not been submitted to any other higher learning institution for degree or examination purposes.

Signed by	Thabo Ramaano
Signature	
Date	22 August 2024

Abstract

The Admission Point Score (APS) metric, utilised as a response to admit prospective students for an academic course, may appear effective in determining student success. In reality, almost 50% of students admitted to a science programme in a higher education institution failed to meet all the requirements necessary to complete the programme during the period of 2008 and 2015. This had a direct impact on the overall graduation throughput. Thus, the focus of this research was geared towards the adoption of a probabilistic graphical approach to advocate its mechanism as a viable alternative to the APS metric when determining student success trajectories at a higher education level. The purpose of this approach was to provide higher education institutions with a system to monitor students' academic performance en-route to graduation from a probabilistic and graphical point of view. This research employed a probability distribution distance metric to ascertain how close the learned models were to the true model for varying sample sizes. The significance of these results addressed the need for knowledge discovery of dependencies that existed between the students' module results in a higher education trajectory that spans three years.

Keywords— Bayesian network; higher education trajectory, hill-climbing structure learning algorithm, Kullback-Leibler divergence, variable elimination inference algorithm

Contents

1 Introduction **4**

2 Related Work **9**

2.1 Impact of the Current Admission Techniques in Determining the Success of Students . . . 11

2.1.1 APS Metric as an Admission Indicator . . . 11

2.1.2 Grade Point Average as an Admission Indicator . . . 12

2.1.3 Admission Techniques Summary . . . 12

2.2 Machine Learning Approach to Predicting Student Success . . . 13

2.2.1 Feature Selection for Maximising the Predictive Accuracy . . . 13

2.2.2 Conceptual Framework . . . 14

2.3 Influence of Probabilistic-Graphical Models in Predicting Student Success . . . 15

2.3.1 Naive Bayes Model for Predicting Student Success . . . 17

2.3.2 Bayesian Model for Predicting Student Success . . . 17

2.3.2.1 Bayesian Model Structure . . . 17

2.3.2.2 Bayesian Model Parameter Estimation . . . 19

2.3.2.3 Bayesian Model Inference . . . 19

2.4 Future Work Based on Related Work . . . 19

3 Methodology **21**

3.1 Data Construction . . . 22

3.2 Probabilistic Graphical Model: Bayesian Model . . . 22

3.3 Dynamic Bayesian Model . . . 23

3.3.1 Construction of a Ground Truth Bayesian Probabilistic Model . . . 24

3.3.2 Construction of a Learned Bayesian Probabilistic Model . . . 26

3.4 Bayesian Parameter Learning using Bayesian Estimation . . . 26

3.5 Kullback-Leibler (KL) Divergence Metric for Comparing Probability Distributions . . . 27

3.6 Inference for Bayesian Models . . . 27

3.6.1 Exact Inference . . . 29

3.6.1.1 Variable Elimination . . . 29

4 Results and Discussion **32**

4.1 KL Divergence Probability Distribution Comparison . . . 32

4.2 Variable Elimination Inference Results . . . 35

4.2.1 Test Case 1: Student Admission 35
4.2.2 Test Case 2: Post-Admission Monitoring 36
4.2.3 Test Case 3: Monitor Student Completion Status 38
4.2.4 Test Case Results Summary 40
4.3 Results as they Pertain to the Research Questions 40
4.3.1 Research Question 1 40
4.3.2 Research Question 2 40
4.4 Research Limitation 41

5 Conclusion **42**
5.1 Contribution 43
5.2 Future Work 43

References **46**

Chapter 1

Introduction

South African higher education institutions are facing growing concerns over high dropout rates and low graduation throughputs amongst students [Temoso and Myeki 2022]. According to a report on higher education and skills in South Africa conducted by Statistics South Africa, it has been established that the on-time graduation rate for students pursuing three-year degree programmes across all higher education institutions in South Africa stood at less than 30% [Maluleke 2017].

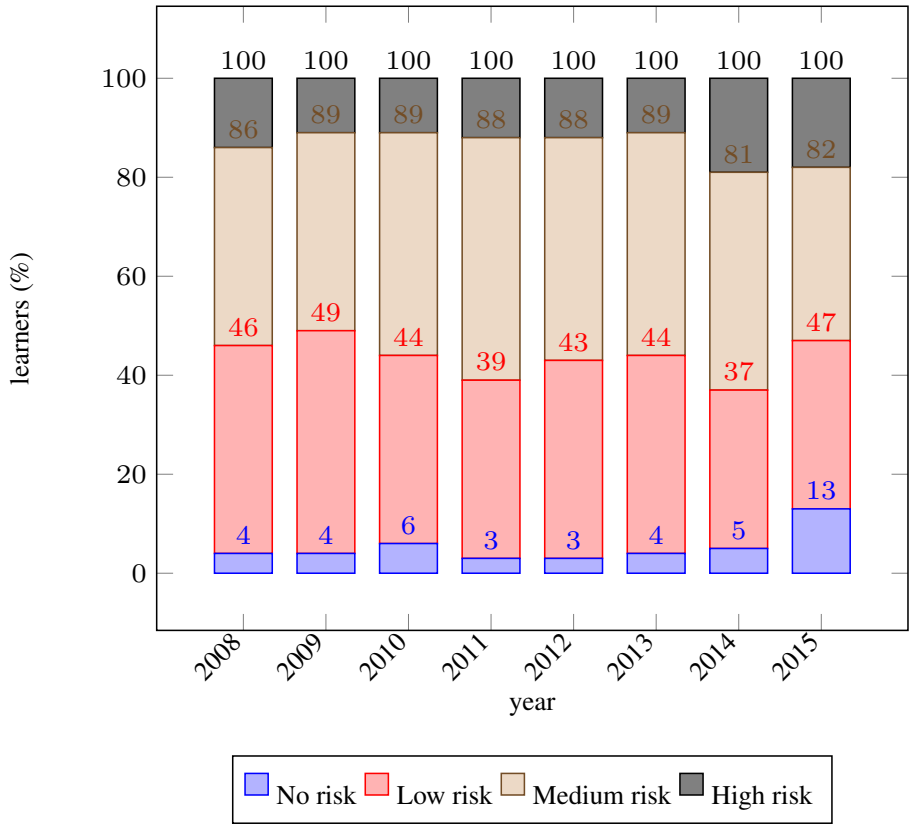

Figure 1.1: Percentage of students (categorised as no-risk, low-risk, medium-risk and high-risk) registered at a higher education institution between 2008 and 2015 adopted from Ajoodha [2022].

The Admission Point Score (APS) metric is currently adopted by South African higher education institutions as an admission indicator to admit prospective students. As a result, this admission metric has not yielded the desired results of achieving a plausible graduation throughput.

According to Figure 1.1, almost 50% of students admitted for degree studies failed to complete all the necessary requirements for their respective academic courses when considering medium risk and high risk students. To cater for this, an improved mechanism is required to supplement the current APS metric with the purpose of improving the overall on-time graduation throughput.

The higher education trajectory of students is the journey they undertake to acquiring their academic qualification in the number of years stipulated for the qualification. As such, the trajectory can be deemed a success (student has obtained a qualification) or failure (student has discontinued studies). In simpler terms, it can be thought of as the progression through the higher education pipeline [Haas and Hadjar 2020]. The illustration in Figure 1.2 is a graphical representation of a higher education academic trajectory that spans three years and depicts both the success and failure that emanates in a student's higher education trajectory.

According to Haas and Hadjar [2020], the topic of academic trajectory of students in higher education, en-route to obtaining their academic qualifications, is an under-researched issue despite its relevance in determining student success. In addition, Haas and Hadjar [2020] provided the illustration in Figure 1.3 that depicts the three levels of predictors, namely: macro-level, meso-level and micro-level in the higher education domain. The aforementioned higher education predictors have the potential of shaping a student's higher education trajectory. This study will make use of academic performances to determine student success trajectories which is an attribute located in the micro-level of predictors.

A number of studies have made significant advancements in the domain of predicting student performance using a variety of predictive algorithms. Results of particular interest are those emanating from studies using probabilistic graphical models [Bekele and Menzel 2005; Moradi and Khanteymoori 2012; Nudelman *et al.* 2019]. The aforementioned studies provide useful methodologies and concepts required for this research. These include: the construction of a Bayesian model [Moradi and Khanteymoori 2012]; real world application of using a Bayesian model in an education environment [Bekele and Menzel 2005] and taking advantage of the graphical nature of Bayesian modelling to get a deeper understanding of the relationships that influence the success of a student [Nudelman *et al.* 2019].

The limitations of the aforementioned predictive algorithms lie in their inability to explore the students' higher education trajectory from first year to final year. An analysis of this nature may provide researchers and educators with valuable insights to optimize the probability of on-time graduation by monitoring students' higher education academic journey.

The purpose of this research is to employ a Bayesian probabilistic approach to predict a student's higher education trajectory and provide higher education institutions with a simplified probabilistic-graphical view for admitting new students. A model of this nature is useful in assisting higher education institutions monitor students' academic trajectory with the aim of employing guidance tools to assist them in graduating on time.

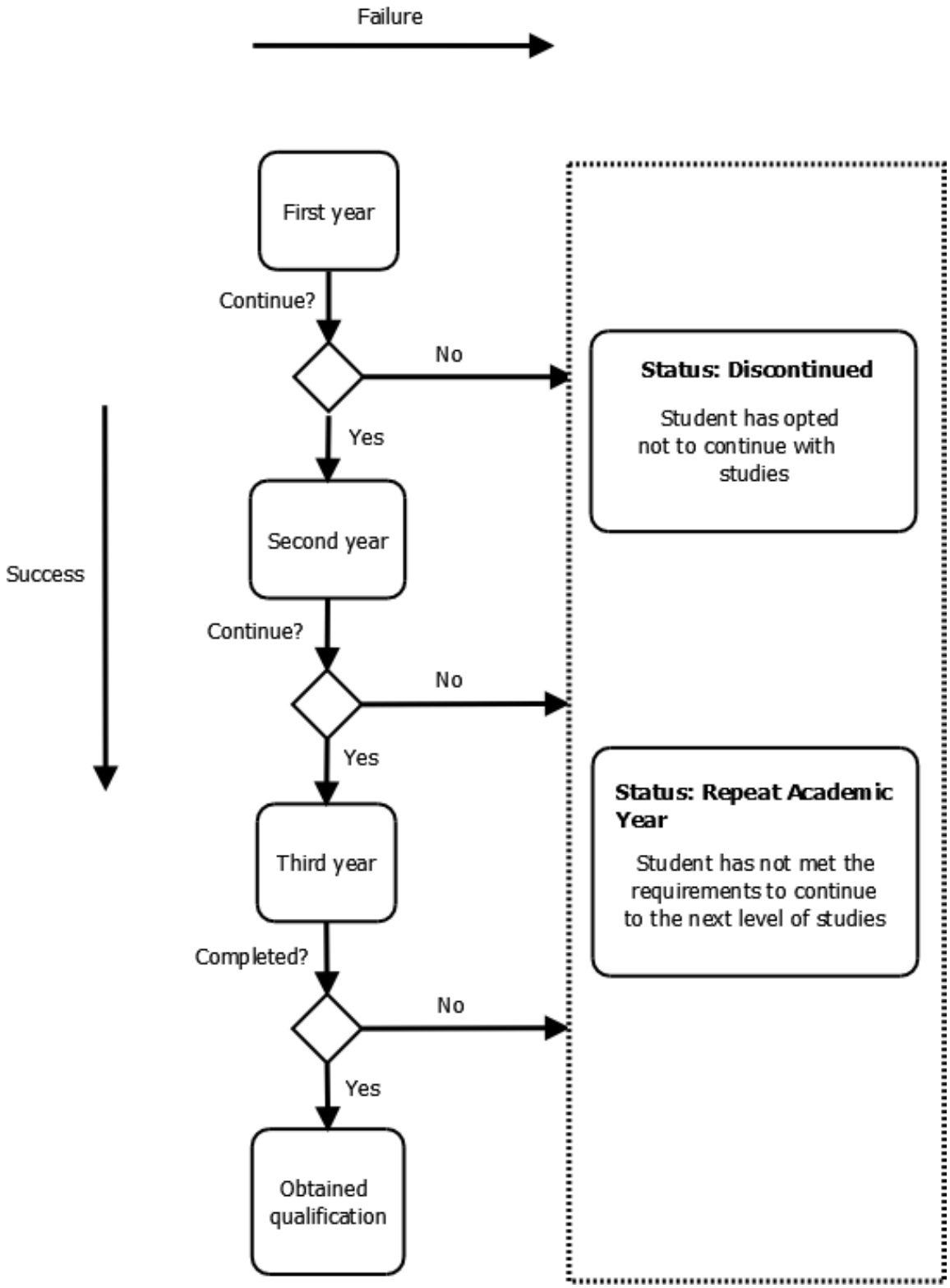

Figure 1.2: Higher education trajectory of an academic qualification that spans three years depicting both the success and failure as it pertains to the higher education trajectory of students. A success is represented by the downward facing arrow and a failure is represented by the horizontal facing arrow.

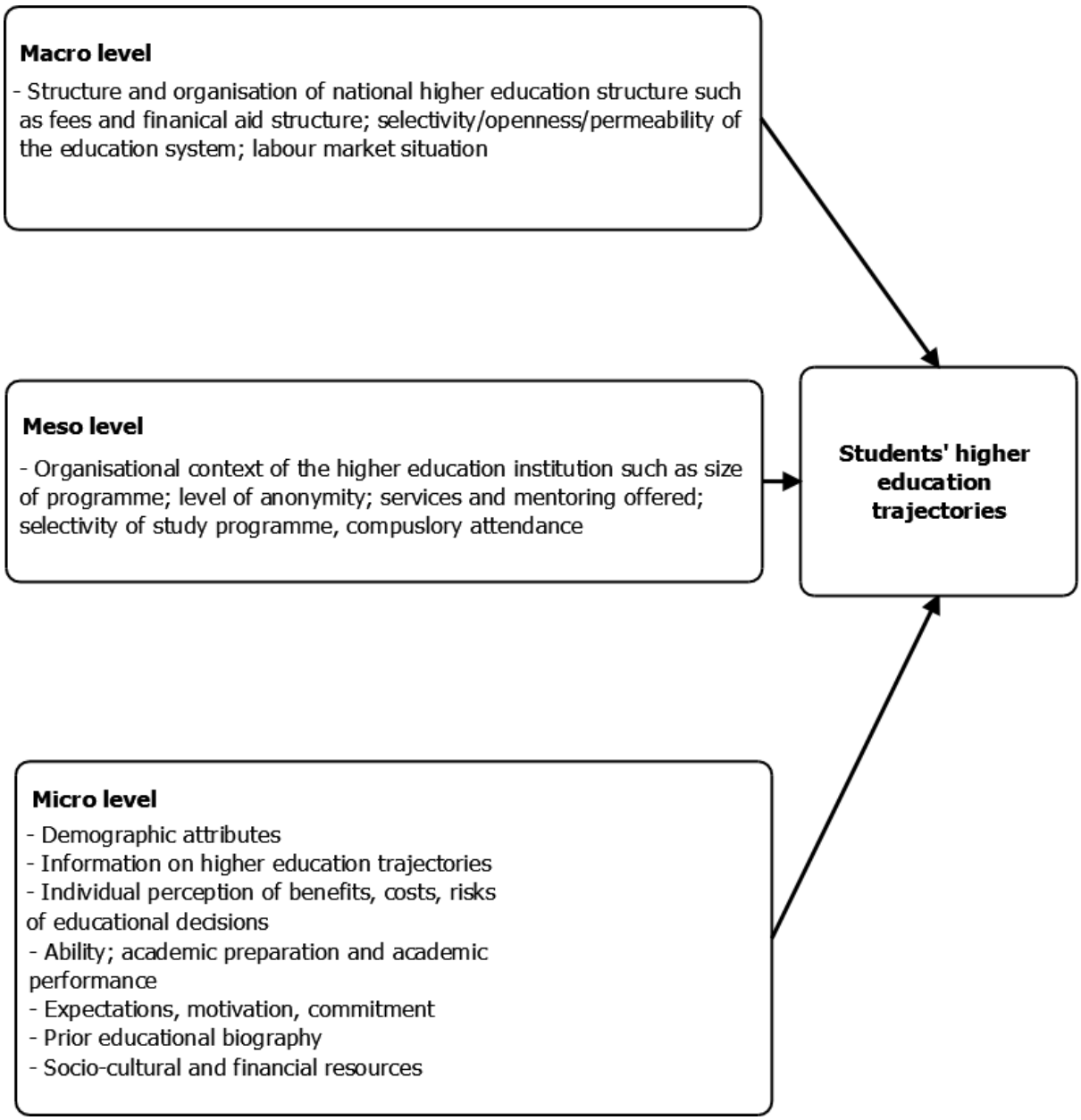

Figure 1.3: Higher education trajectory predictors categorised into three different levels, namely: macro-level, meso-level and micro-level adapted by Haas and Hadjar [2020].

When performing this research the following research questions are proposed: (1) What would constitute a viable course of action to undertake in order to determine student success trajectories? (2) To what extent do the results from the probabilistic graphical model provide an effective alternative to determining student success trajectories as compared to using the APS metric?

In order to answer the proposed research questions, this research experimented with secondary and higher education academic scores of students from computer science and mathematics majors. Various Bayesian models were learned using a popular score-based structure learning algorithm known as the hill-climbing structure learning algorithm. In addition, this

research implemented the Bayesian Estimator to learn the parameters of the model. The resulting learned Bayesian models were compared with a model developed using expert knowledge (known as the ground truth model), using a probability distribution statistical distance measure in order to ascertain how well the Bayesian model was able to learn the relationships between course modules in a higher education trajectory that spans three years.

In addition, a graphical view of the computed statistical distance between the learned model with the ground truth model was presented to provide an illustration of the impact of the learned Bayesian model structures relative to the ground truth model for increasing sample sizes. Thus, the resulting learned models are used to determine student success trajectories in a higher education trajectory that spans three years.

Results from this research will contribute to further the application of Bayesian probabilistic modelling in the domain of predicting students' success upon analyses of their trajectory through higher education. In addition, this research will provide higher education administrators with a mechanism to employ guidance tools to assist students with monitoring their higher education progression, hence, optimising their chances of graduating on time.

The structure for the remainder of this research document will begin with covering related work that have made significant contributions in using machine learning to predict student performance. Section 3 will provide an outline of the procedure used in performing the experiment of this research. The findings from this research will be outlined and discussed in Section 4. Finally, Section 5 will conclude with a summary of the research, contributions from this research and recommendations for possible future work.

Chapter 2

Related Work

The task of predicting student performance in an effort to identify students at risk of not obtaining their academic qualification on time has become an urgent requirement for educational institutions [Mahmoud Abu Zohair 2019]. It is important that higher education institutions employ effective measures to improve student completion rates in response to the low on-time student graduation rate alluded to in the study by [Maluleke 2017].

Results presented in Table 2.1 represent a summary of past studies that made useful contributions in employing Bayesian modelling to predict the performance of students. Furthermore, the results presented in the table make mention of the data that was used to perform the experiments; the features used in the development of the Bayesian model; the performance metrics used in evaluating the performance of the developed Bayesian models and the outcome from the performance metrics in the domain of predicting the success of students in school.

Furthermore, the features represented in Table 2.1 form the nodes in the construction of the respective Bayesian models. A relationship is established between the nodes during the learning Bayesian process illustrating the dependencies and independencies that exist between the nodes in the model. Furthermore, it is worth noting that the common features of interest amongst the observed studies in Table 2.1 are gender [Bekele and Menzel 2005; Mathye and Ajoodha 2018; Moradi and Khanteymoori 2012] and academic course grades [Bekele and Menzel 2005; Kustitskaya *et al.* 2020; Mathye and Ajoodha 2018; Moradi and Khanteymoori 2012; Nudelman *et al.* 2019] which form part of the micro-level predictors of predicting a student's higher education trajectory as illustrated in Figure 1.3.

This section will begin by exploring the impact of admission metrics, that are currently used to admit students for higher education studies, in an effort to determine student success. This section will proceed to exploring studies that have made significant contributions in utilising machine learning models to aid in predicting the success of students. Finally, This section will conclude with studies that have utilised a probabilistic-graphical approach to determine the success of students and providing insights in the factors that affect the success of students.

Table 2.1: Bayesian learning approaches with their associated features; performance metrics and outcomes from the performance metrics determined from past studies when predicting student success.

Author	**Data**	**Features (nodes)**	**Performance Metrics**	**Value**
Bekele and Menzel [2005]	514 data records of high school students were considered	Gender; group work attitude; interest for mathematics; achievement motivation; self-confidence; shyness; English performance and mathematics performance	Accuracy	64%
Kustitskaya *et al.* [2020]	129 data records of student performance and final grades for the academic course: Probability and Statistics	Exam final grade; number of lectures attended in a certain week; number of practicums attended in a certain week; number of points awarded for achievement in the class in a certain week; quiz score; e-test score and an indicator of student persistent.	Weighted F-score Sensitivity	$> 80\%$ $> 85\%$
Mathye and Ajoodha [2018]	Higher education students' data records from the faculty of science	Grade; calendar year; year of study; gender; race; course; final grade outcome; progress outcome type description; country; citizenship; language; nationality; permit type description and institution description	Accuracy	81.25%

Moradi and Khantey-moori [2012]	A sample of 500 data records of higher education students	Employment status; gender; semester; teacher and score	Accuracy	66%
Nudelman *et al.* [2019]	783 data records of first year higher education computer science students	Province of secondary school; secondary school quintile; academic literacy and quantitative literacy results; financial aid status; science results; advanced mathematics attempt results; English results; work ethic; attitude and at-risk status	F1-Score MCC Sensitivity Specificity	84.14% 61.98% 66.67% 92.42%

2.1 Impact of the Current Admission Techniques in Determining the Success of Students

It has been well-established that to admit prospective students for higher education studies, in response to the increase in access to higher education institutions [Cross and Carpentier 2009], a viable admission technique is required to determine students that are most likely to complete their academic course successfully. This section will primarily focus on the contributions of admission metrics that are currently used to seek out candidates for higher education studies.

2.1.1 APS Metric as an Admission Indicator

In South African, the APS metric is one of the performance measures utilised by higher education institutions to seek out candidates for admission for higher education studies. According to a study by Ajoodha [2022], almost 50% of students failed to complete the requirements of their academic course in the science stream.

Furthermore, an interesting result emanating from Ajoodha [2022] showed that relying solely on the employment of the APS score to determine the success of a student in higher education did not yield the desired student success rate as compared to incorporating biographical and individual attributes to the predictive model. Table 2.2 illustrates this point with results of the predictive accuracy of the aforementioned models. The predictive accuracy becomes greater when incorporating both biographical and individual attributes as opposed to relying solely on the APS score and the combination of the APS score with the grades for English and Mathematics as predictors for determining the success of students in higher education [Ajoodha 2022].

Table 2.2: Predictive accuracies resulting from a study by Ajoodha [2022] for determining learner vulnerability for different combinations of features as opposed to using the APS score solely.

Attributes	**Machine learning model predictive accuracy**		
	Multilayer perception;	**Logistic regression;**	**Naive Bayes**
APS score	30%	34%	28%
APS score; English grades; Mathematics grades	50%	44%	38%
APS score; English grades; Mathematics grades; Biographical information; Individual attributes	85%	83%	82%

2.1.2 Grade Point Average as an Admission Indicator

The Grade Point Average (GPA) metric is a common admission criteria administered by higher education institutions because of its perceived importance in measuring the academic outcome of students in higher education level [Almarabheh *et al.* 2022]. A study conducted by Almarabheh *et al.* [2022] experimented with both science test results and English test results in addition to the secondary school GPA to predict the higher education outcome of students admitted to a medical programme. The study found that the science test results proved to be a good predictor of performance for medical students as compared to the other predictors considered [Almarabheh *et al.* 2022].

2.1.3 Admission Techniques Summary

Results produced from studies in these sections provide an indication that solely depending on the secondary school GPA or APS score is not viable as the only indicator to admit students for higher education studies. It is important to introduce valid and reliable admission tools [Almarabheh *et al.* 2022] in order to improve in the intake of prospective students for studying at higher education institutions.

2.2 Machine Learning Approach to Predicting Student Success

Results from Tables 2.1 and 2.2 place significant emphasis on the predictive capabilities attached to using machine learning models to determine student success. In order to make use of the predictive capabilities of machine learning models, it is important to select the right features that have an effect on determining the success of students in higher education by employing feature selection techniques.

2.2.1 Feature Selection for Maximising the Predictive Accuracy

A popular feature selection technique utilised by past studies in order to enhance the predictive capabilities of a machine learning model is the information gain (entropy) method. Mngadi *et al.* [2020] attempted to use information gain (entropy) to determine the importance of each feature used in the study and their contribution in the event of determining the risk profile of students.

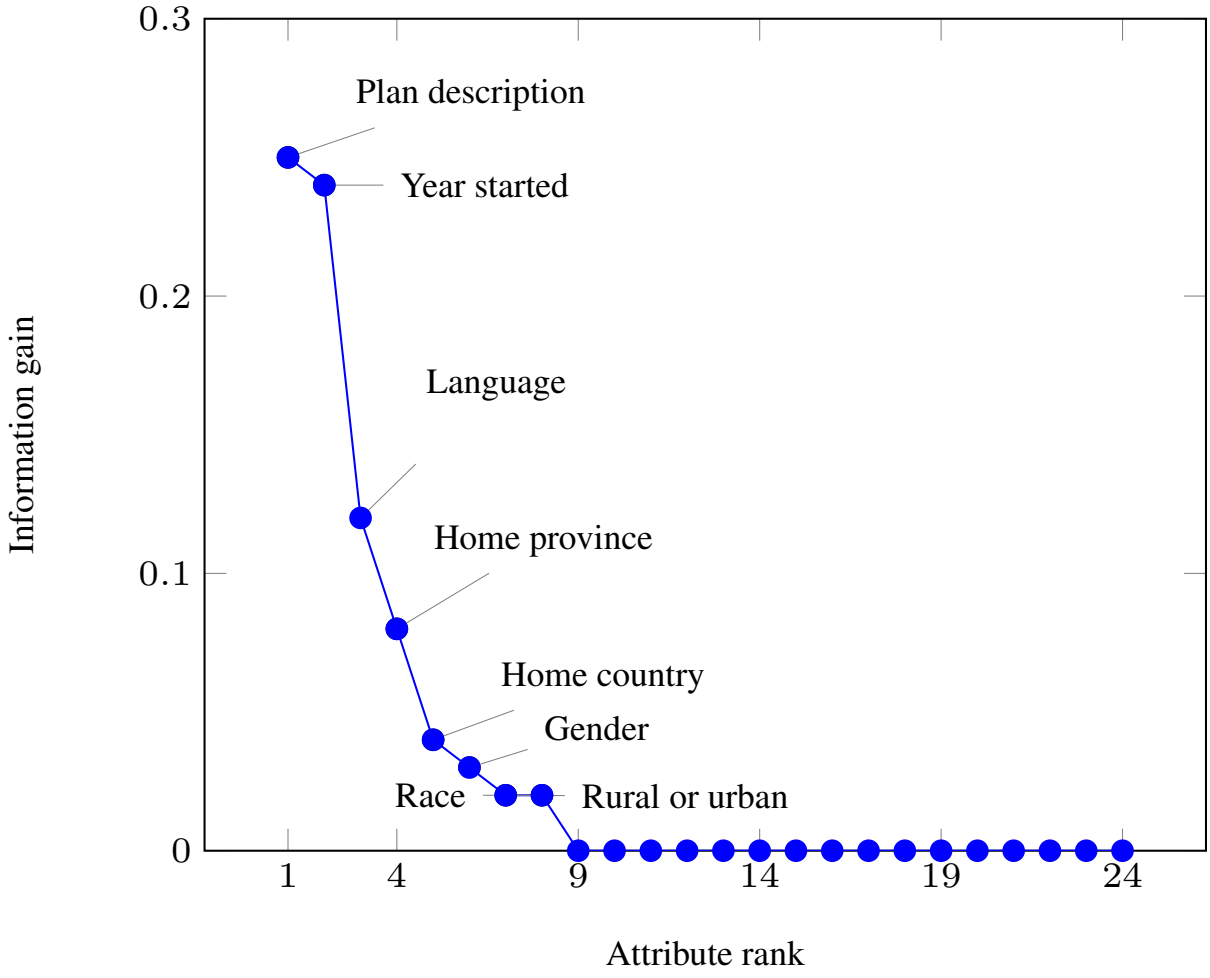

Figure 2.1: Attribute ranking from the most important attributes to the least important attributes based on the computed information gain (entropy) when determining the risk profile of students adapted from Mngadi *et al.* [2020].

Attributes with the highest information gain (entropy) are considered important attributes. In this study, the important features were both individual and biographical data as presented in the graph in Figure 2.1 [Mngadi *et al.* 2020]. By using the information gain (entropy) feature selection technique, the random forest model trained in this experiment produced the greatest accuracy of 85% which was significantly better than the accuracy produced from solely using the APS score as illustrated in Table 2.2.

Furthermore, there are numerous other studies that have utilised feature selection techniques to select the best features for their respective machine learning models. In particular, Aissaoui *et al.* [2020] experimented with a variety of feature selection techniques including: cforest; random forest; LMG; First; Last; Multivariate adaptive regression splines(MARS) and Boruta. Furthermore, Aissaoui *et al.* [2020] developed multiple multivariate linear regression models from the resulting top five attributes from each method respectively. It was found that the best multivariate linear regression model used in determining student performance resulted from applying the MARS method [Aissaoui *et al.* 2020].

In addition to the computational aspect of selecting important features, another methodology applied in seeking out important features is to approach experts in the education field to assist in determining the most important factors that affect the success of students. One such study that applied this approach was Bekele and Menzel [2005]. The study mentioned above utilized expert opinions and literature reviews to select the top eight attributes for constructing a Bayesian model aimed at predicting performance in Mathematics. This effort resulted in a 64% accuracy rate [Bekele and Menzel 2005].

2.2.2 Conceptual Framework

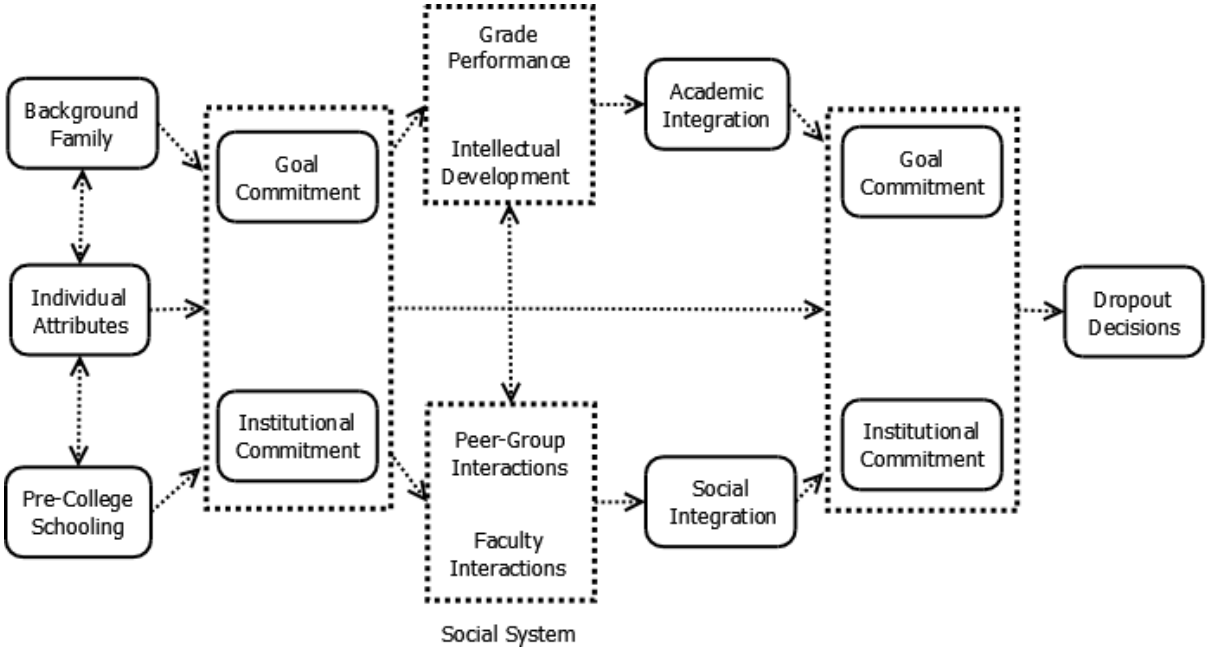

Figure 2.2: Conceptual framework proposed by Tinto [1975] presenting a system for using student and institutional attributes to determine the dropout status of students.

The results from Ajoodha [2022] has shown that a student's APS score alone is not a viable admission metric to determine their success in higher education, hence, additional attributes are required to supplement the current admission criteria to determine the success of a student in higher education. The conceptual framework, depicted in Figure 2.2, proposed by Tinto [1975], promotes a methodology of grouping student attributes into three groups, namely: background, individual and pre-college/school attributes. This framework has been successfully adopted by previous studies [Abed *et al.* 2020; Ajoodha 2022; Mngadi *et al.* 2020] to determine the success of students in their higher education studies.

The results produced in Table 2.3 illustrate a grouping of attributes from past studies that experimented with machine learning models on a defined set of attributes to determine student success. The attributes from these studies are grouped into the three groups proposed in the conceptual framework by Tinto [1975]. The introduction of other attributes into a predictive model, in addition to the APS score, has the potential of improving the predictive accuracies of the proposed machine learning model with the overall aim of determining student success in higher education.

Furthermore, it is evident from the resulting performance measures for each study that attribute selection plays a significant role in the predictive capabilities of a machine learning model. Thus, this provides an indication that the inclusion of more attributes (provided that these attributes make a significant contribution in predicting student performance) in a machine learning model increases the predictive capabilities of a machine learning model as opposed to depending on the APS score as the sole admission metric for determining student success.

The limitations of the non-probabilistic graphical models depicted in Table 2.3 lie in their inability to produce a graphical model. A graphical model would depict the dependent relationships that exist between the student results in a higher education trajectory that spans the length of an academic course. Thus, the graphical model and associated probability distributions associated with the nodes of the graphical model would be applied to infer the probability of success in a certain model given prior results within the trajectory.

For this reason, the results that are of particular interest to this research are those related to studies concerned in the implementation of probabilistic graphical approaches in order to determine the success of students in higher education as presented in Table 2.1 [Bekele and Menzel 2005; Kustitskaya *et al.* 2020; Mathye and Ajoodha 2018; Moradi and Khanteymoori 2012; Nudelman *et al.* 2019].

2.3 Influence of Probabilistic-Graphical Models in Predicting Student Success

The probabilistic-graphical nature of the associated Bayesian models have contributed to understanding the probability of success or failure for students based on the relationships that are formed between student attributes in an academic course using concepts in graph theory and probability theory.

This section will serve to unpack contributions made from applying probabilistic-graphical models in the domain of determining student success in higher education. Upon unpacking

Table 2.3: Grouping attributes from past studies into background, individual and pre-college/schooling as per the conceptual framework defined by Tinto [1975] to observe attribute group selection with model performance.

Author	Background	Individual	Pre-college/schooling	Best Model	Model Performance
Aissaoui *et al.* [2020]	✓	✓		Multivariate linear regression	4.368 (RMSE) 0.103 (R^2) 3.301 (MAE)
Buraimoh *et al.* [2021]	✓	✓	✓	Classification and Regression Tree	0.86 (Accuracy) 0.86 (F1-score) 0.97 (AUC)
Abed *et al.* [2020]	✓	✓	✓	Naive Bayes	0.69 (Accuracy)
Al Mayahi and Al-Bahri [2020]			✓	Support Vector Classifier	0.88 (Accuracy)
Nedeva and Pehlivanova [2021]	✓	✓	✓	Multilayer Perceptron	0.974 (Accuracy) 0.974 (F1-score) 0.0190 (MAE) 0.0802 (RMAE) 0.0719 (RAE) 0.222 (RRSE)

these contributions, this section will proceed to identifying further contributions to be made that will be attempted in this research.

2.3.1 Naive Bayes Model for Predicting Student Success

The naive Bayes model has proven itself as a simple but effective classification machine learning model when predicting student success. For example, Abed *et al.* [2020] experimented with six machine learning models to develop a recommendation system for students, aiming to provide them with advice on their higher education trajectory. Based on the six machine learning models, the naive Bayes model resulted as the best model achieving the greatest accuracy of 69.18% [Abed *et al.* 2020].

The naive Bayes model may seem a viable machine learning model to determine student success from a probabilistic-graphical approach based on its simple yet effective nature. Unfortunately, for the purposes of this research, the naive Bayes model is limited by its independence assumption.

The conditional independence assumption of a naive Bayes model suggests that all the attributes in a Bayesian model are conditionally independent of each other for a given class label for each given instance as illustrated in the diagram of a simple naive Bayes model in Figure 2.3 [Ajoodha 2022]. The diagram in Figure 2.3 shows that for a given class, nodes X_1 to X_n are conditionally independent. A Bayesian model that is not restricted by the independence assumption ensures that there is a deeper understanding of the relationship that exists between the student attributes that may not have been apparent when implementing a naive Bayes model.

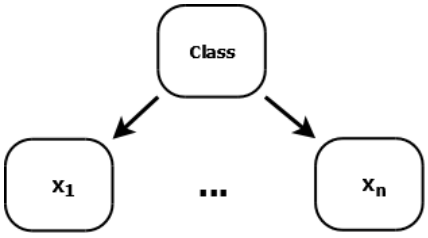

Figure 2.3: A simple representation of a naive Bayes model where node features X_1 to X_n are conditionally independent for a given class adapted by Ajoodha [2022].

2.3.2 Bayesian Model for Predicting Student Success

The independence assumption attached to the naive Bayes model limits the learning of relationships that may exist between the student's attributes. This section will introduce previous work that have employed Bayesian models that are not restricted by the independence assumption in learning student attribute relationships that contribute to determining student success.

2.3.2.1 Bayesian Model Structure

The Bayesian model structure assists in visualising and identifying the relationships between the student attributes in predicting student success. The selected attributes are represented by the nodes in the developed Bayesian model.

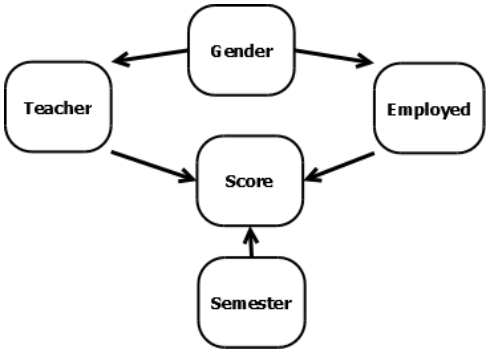

Figure 2.4: Resulting Bayesian model structure used in predicting student scores adapted by Moradi and Khanteymoori [2012].

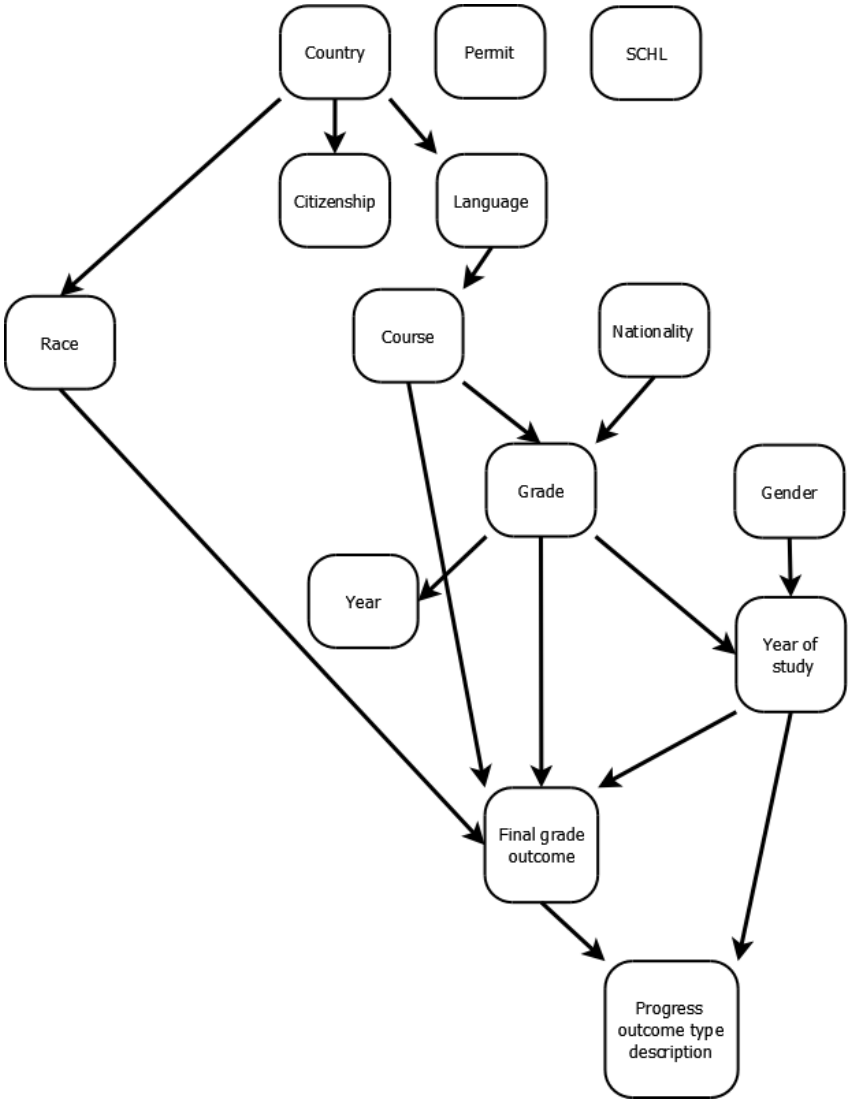

Figure 2.5: Resulting Bayesian model structure used in visualising students' biographical and enrollment attributes to predict student performance adapted by Mathye and Ajoodha [2018].

As opposed to placing reliance on expert knowledge to learn a Bayesian model structure, it is possible to learn the structure of a Bayesian model through employing one of a variety of structure learning algorithms. For instance, Moradi and Khanteymoori [2012] made use of different algorithms including K2 and BDeu (Bayesian Dirichlet equivalent uniform) to construct a Bayesian model structure. An example of a Bayesian model structure resulting from Moradi and Khanteymoori [2012] has been illustrated in Figure 2.4.

Furthermore, it is possible to have some of the selected student attributes not present in the final Bayesian model structure. The student attributes not selected for the final model have no relationship with any of the other attributes. For instance, the Bayesian model structure developed by Mathye and Ajoodha [2018] experimented with fourteen biographical and enrollment student attributes and it was found that two of the attributes, permit type description (Permit) and institution description (SCHL) were presented with no relationships with the other attributes based on the illustrated results in Figure 2.5.

2.3.2.2 Bayesian Model Parameter Estimation

The nodes representing each random variable (student attribute) of the Bayesian model have an associated continuous probability distribution presented in a continuous probability table (CPT). Nudelman *et al.* [2019] used the training data to compute prior probabilities for each random variable and thereafter the conditional and posterior probabilities were computed. In order to account for the immeasurable latent variables present in the construction of the Bayesian model, Nudelman *et al.* [2019] opted to use the Expectation Maximization (EM) algorithm for the purpose of estimating the parameters of the developed Bayesian model.

There are a variety of other parameter estimation algorithms that are used for the purpose of computing the associated probability distribution for each random variable. Mathye and Ajoodha [2018] employed the Bayesian parameter estimation method which functions by formulating the prior probability distribution on a random variable θ and uses the dataset to compute the posterior probability of the random variable θ.

2.3.2.3 Bayesian Model Inference

Bayesian inference allows for the possibility of employing a Bayesian model B for the purpose of computing the marginal probability $P(N = n)$, where n represents the sample size, for each node $v \in B$ and each possible instantiation v [Rama *et al.* 2012]. In the study by Mathye and Ajoodha [2018], a Junction Tree algorithm was used to infer students' grade based on their demographic attributes. Moradi and Khanteymoori [2012] experimented with a number of inference algorithms including likelihood sampling; logic sampling and clustering to query the developed Bayesian model in order to determine a student's grade.

2.4 Future Work Based on Related Work

The resulting Bayesian model structures depict a graphical understanding of the relationships between the student attributes that lead to identifying at-risk students [Kustitskaya *et al.* 2020; Nudelman *et al.* 2019] and predict student success [Mathye and Ajoodha 2018; Moradi and Khanteymoori 2012]. This may appear effective at face value, however, more work is required

to further unpack results emanating from the application of Bayesian models and possibly improve on the predictive accuracy associated with determining a student's risk. [Nudelman *et al.* 2019].

Taking into consideration results from this section and the need to further unpack these results, this research will attempt to explore the prediction of student success trajectories using a Bayesian approach in an academic course that spans three years. The design of the Bayesian model will be a dynamic model showing relationships between course modules for mathematics and computer science between first year and third year for a three year trajectory. In addition, certain elements from previous work including the choice of Bayesian estimator and score-based function will be incorporated in this research.

With that said, the next section will provide an in-depth methodology to conduct the experiment for this research including: the data construction; the development process of a probabilistic-graphical model; the usage of a statistical distance measure to compare the probability distributions of the probabilistic graphical models and conclude with the choice of inference algorithm to compute the posterior probability distributions.

Chapter 3

Methodology

The work studied by past researchers, in utilising Bayesian modelling to determine student success, employ variables limited to only a single time period. This suggests that the studies do not model the students' higher education trajectory from their first year to their final year of studies. As such, these studies are limited in their capability to predict a successful higher education trajectory thereby improving the overall graduation throughput. The approach of this research is to employ a dynamic Bayesian model where the student results in mathematics and computer science are present in all the time slices of the model. As opposed to previous studies, the variables in this model are not static as they change for each time period. The non-static nature of these variables makes it possible to observe the change in results in these modules as students progress in their higher education trajectory that spans a period of three years with the goal of predicting a successful trajectory.

Furthermore, this research attempted to learn a variety of Bayesian models using the hill-climb structure learning algorithm and compare their learned probability distributions with the ground truth model using the Kullback-Leibler (KL) divergence metric. The significance of a methodology of this nature is to obtain a Bayesian model that is as close as possible to the ground truth model by measuring the statistical distance in their respective associated probability distributions. Thus, the obtained Bayesian model is used in determining student success trajectories in response to a call to further unpack results emanating from Nudelman *et al.* [2019] by further observing higher education student trajectories.

The outline of the methodology will begin with discussing data construction in order to get the data ready for the construction of the Bayesian model. Next, the development of the ground truth Bayesian model and learned models will be discussed along with algorithms used for structure learning; parameter estimation and inference. Finally, the probability distribution comparison between the ground truth model and learned models using the KL divergence metric will be discussed along with the observed effect of the computed average KL divergence with the increasing sample size of the dataset.

3.1 Data Construction

The data used for this research was synthetic data. Previous studies observed biographical and individual attributes in addition to the academic performance as per the conceptual framework proposed by Tinto [1975]. The aim of this study was to predict the higher education trajectory of students purely from an academic performance point of view. As such, the attributes of importance for this research was the academic results from computer science and mathematics majors. The academic results are for grade 12; first year; second year and third year. Thus, this highlights the student's academic higher education trajectory for a three year period.

3.2 Probabilistic Graphical Model: Bayesian Model

This research proceeded to develop a probabilistic graphical model, namely, a Bayesian model. Bayesian models are graphical models represented by both nodes and edges Rama *et al.* [2012]. The nodes in the Bayesian model are represented by the student's marks whereas the edges are represented by the conditional dependencies between the student's marks in a higher education trajectory that spans three years. According to Ajoodha and Rosman [2018], two random variables X and Y are said to be conditionally independent of each other if following the introduction of a known set Z, then knowledge of X provides no information about Y as illustrated in Figure 3.1.

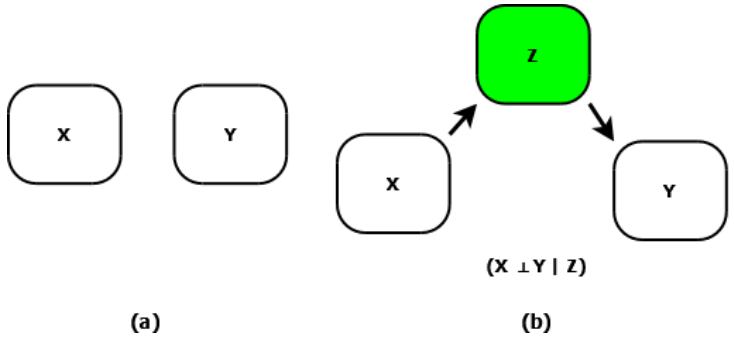

Figure 3.1: An illustration depicting the conditional independence of two variables in a Bayesian model. (a) Represents the two conditional independent random variables X and Y. (b) Represents the addition of a set Z to confirm the conditional independence assumption $(X \perp Y|Z)$ of the two random variables X and Y.

The graphical property of a Bayesian model lies in its directed acyclic nature. The directed acyclic nature of the Bayesian model graph implies that cycles in the graph are absent. The edges are represented by directional arrows as illustrated in the example of the resulting Bayesian model from a study by Bekele and Menzel [2005] in Figure 3.2.

Due to the acyclic nature of a Bayesian model, its associated graph is referred to as a directed acyclic graph (DAG). In addition to the directed acyclic nature of Bayesian models, there is also the conditional probability tables [Rama *et al.* 2012] associated with each node in the Bayesian model. The development of a Bayesian model is based on the Markov condition which states that every variable in a Bayesian model is independent of its non-descendants non-parents given its parents and leads to its joint probability distribution, given by the equation

$P(X_1, ..., X_n) = \prod_i P(X_i|parent(X_i))$, where $X_i, i = 1, 2..., n$, represents the independent variable in a Bayesian model [Cozman 2000].

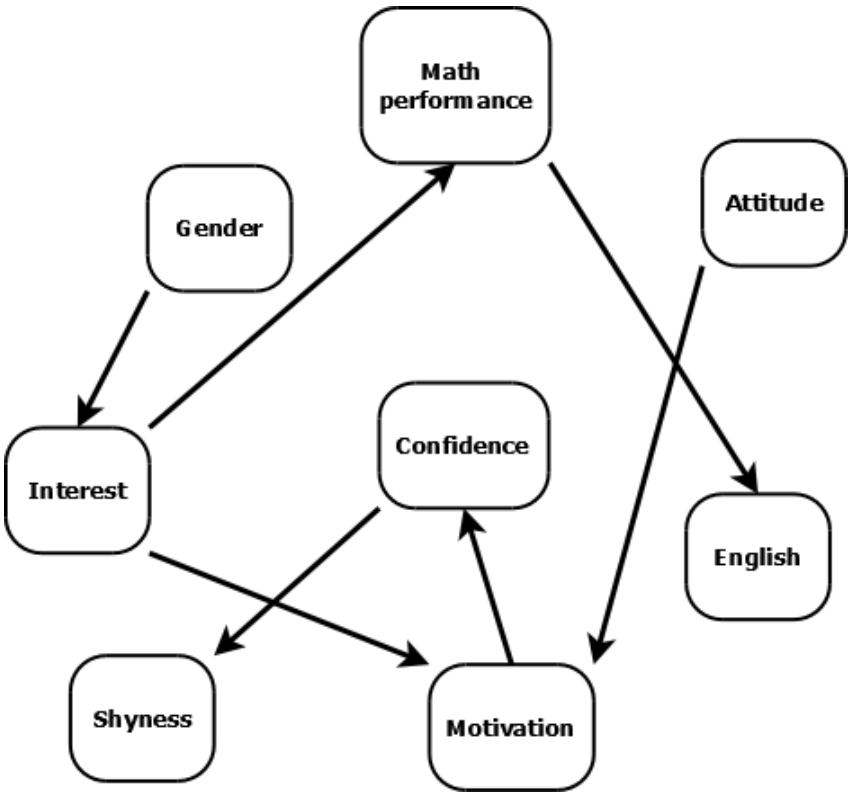

Figure 3.2: Example of a Bayesian model resulting from a study by Bekele and Menzel [2005].

The random variables that make up the Bayesian models in this section are static. This implies that they do not change. The Bayesian models in this section serve as an entry point to the specialised form of a Bayesian model which is the dynamic Bayesian model. The next section extends from static random variables and discusses random variables that change over time which are essential for the purposes of this research given that the concept of higher education trajectories are introduced. The variables chosen for this research are time-dependent, hence, the decision for the application of a dynamic Bayesian model to model the trajectory of a student in a three year degree.

3.3 Dynamic Bayesian Model

A dynamic Bayesian model describes a system that evolves with time [Mihajlović and Petkovic 2001]. As such, the static nature of the standard Bayesian model, discussed in Section 3.2, becomes a dynamic model as it now models motive forces [Mihajlović and Petkovic 2001].

The Markov assumption and time invariant assumption (defined at-length in the study by Ajoodha and Rosman [2018]) compactly represent a trajectory over time and are used to unroll the transition model (contained in the two-time slice Bayesian model along with the initial distribution) into a dynamic Bayesian model [Ajoodha and Rosman 2018].

The graphical representation of a dynamic Bayesian model is made up of variables occurring at multiple time slices with edges emanating within each time slice and different time slices. According to Ajoodha and Rosman [2018], intra-time-slice edges are edges that occur within each time slice whereas inter-time-slice edges are edges that occur between time slices. Furthermore, edges that occur between the same variables in different time slices are known as persistent edges and the variables are known as persistent variables [Ajoodha and Rosman 2018]. An example of a dynamic Bayesian model, adapted from Ajoodha and Rosman [2018], unrolled over three time slices is illustrated in Figure 3.3.

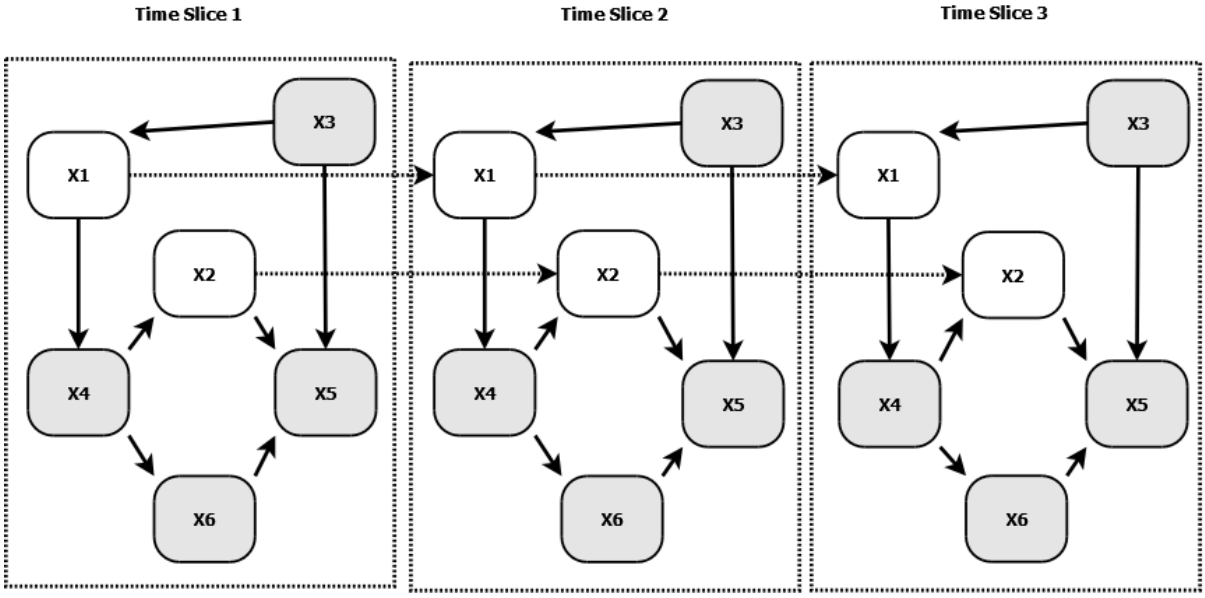

Figure 3.3: An example of a dynamic Bayesian model unrolled over three time slices. The shaded variables are observable whereas the unshaded variables are persistent variables as they are connected to each other in adjacent time slices using the dotted edges. This illustration has been adapted by Ajoodha and Rosman [2018].

Furthermore, according to Mihajlović and Petkovic [2001], the probability distribution function associated with a dynamic Bayesian model system is represented by the equation $P(X,Y) = \prod_{t=1}^{T-1} P(x_i|x_{t-1}) \prod_{t=0}^{T-1} P(y_i|x_t)P(x_0)$. The description of the parameters of the dynamic Bayesian equation are as follows: T represents the time boundary; X = (x_0,..,x_{T-1}) represents the sequence of T hidden-state variables; Y = (y_0,..,y_{T-1}) represents the sequence of T observable variables; $P(x_i|x_{t-1})$ represents the state transition probability distribution function; $P(y_i|x_t)$ represents the observation probability distribution function and $P(x_0)$ represents the initial state distribution.

3.3.1 Construction of a Ground Truth Bayesian Probabilistic Model

The development of the ground truth model for this research was through expert knowledge and from literature. The developed ground truth model, illustrated in Figure 3.4, is represented as a dynamic Bayesian network with four time slices. Time slice 0 represented the student's grade 12 results whereas time slices 1, 2 and 3 represented the student's results for first, second and third year respectively in a higher education trajectory that spans three years.

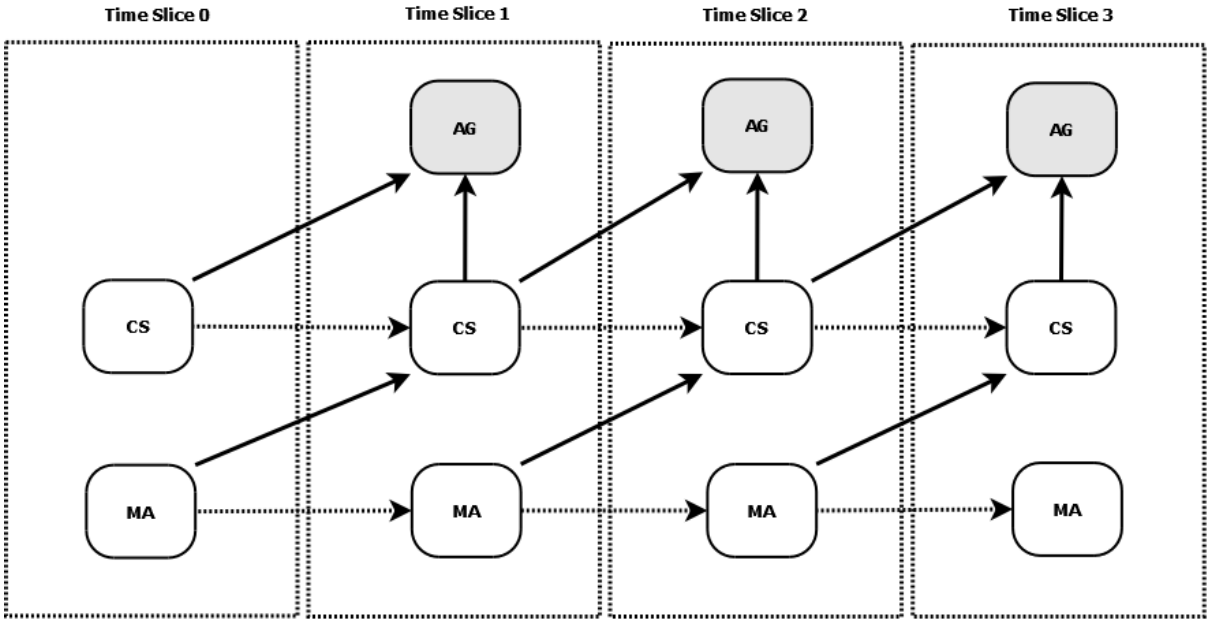

Figure 3.4: The ground truth Bayesian model structure for this research with four time slices. Time slice zero represents the grade 12 results while time slices 1, 2 and 3 represents the student's higher education results for first, second and third year respectively. The computer science and mathematics scores (unshaded) represent the persistent variables due to the dotted persistent edges whereas the final aggregate score (shaded) represents the observable variable.

The course codes MA, CS and AG represented the final mathematics, computer science and aggregate score results respectively for a student in a higher education trajectory that spans three years. There are relationships formed between the academic scores within the time slices and between adjacent time slices. The dotted lines are indicative of the inter-time slice persistent edges. The computer science and mathematics scores are referred to as persistent variables whereas the final aggregate score is the observable variable.

This research has considered only results for mathematics; computer science and the aggregate results for the purpose of modelling the change in results for the three year degree. Based on the availability of data, it is possible to consider other student related factors, for example, students' mode of learning for the duration of their degree. However, due to limitations in the dataset, this research is limited to only the students' results.

Upon forming relationships between adjacent time slices, it is important to note that the variables (states) represented by the student's mathematics; computer science and aggregate scores, in the dynamic Bayesian model in Figure 3.4 satisfy the Markovian condition such that the state at time slice t depends only on the states at its immediate time slice and current time slice [Mihajlović and Petkovic 2001]. In addition to the valid relationships between nodes within each time slice, the valid relationships formed between nodes in different time slices are those that exist in adjacent time slices and must be from a lower time slice l to a higher time slice $l + 1$ adjacent to the previous time slice l as illustrated in the example in Figure 3.5.

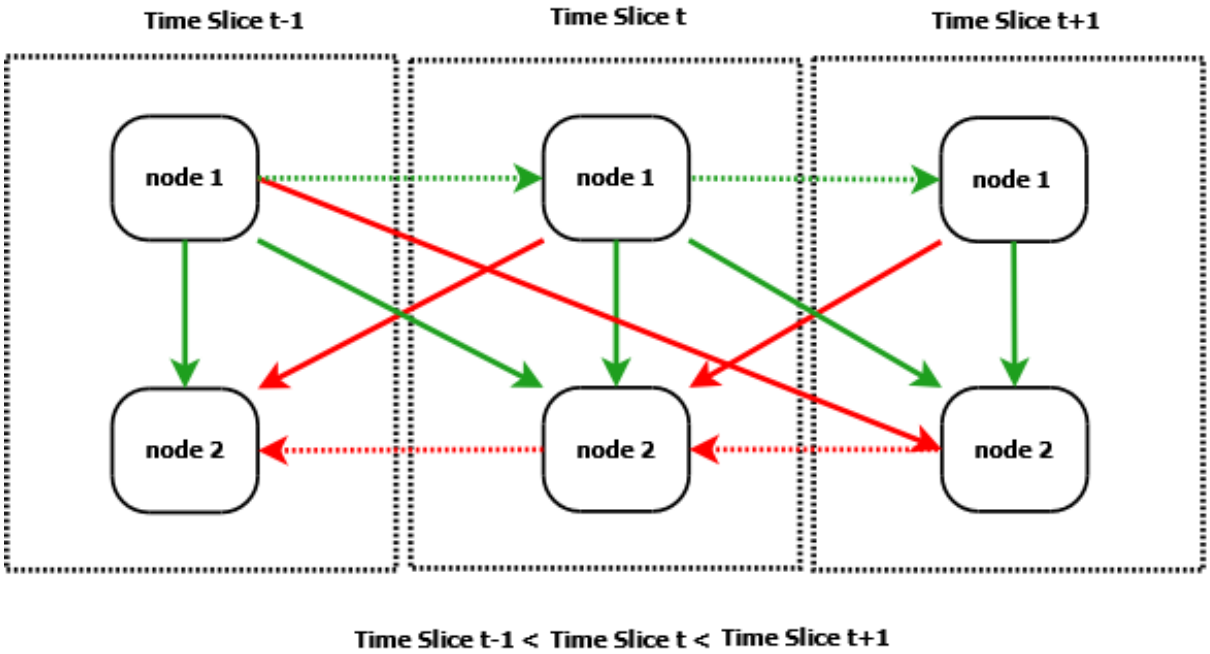

Figure 3.5: Example of an illustration of the Markovian condition for a dynamic Bayesian model from time slice t-1 to time slice t+1. The green directional arrows represent the valid relationships between the nodes whereas the red directional arrows represent the invalid directional arrows as it pertains to the satisfaction of the Markovian condition.

3.3.2 Construction of a Learned Bayesian Probabilistic Model

This stage of the research methodology attempted to learn the structures of a variety of Bayesian models using the popular score-based hill-climbing structure learning algorithm. According to Gámez *et al.* [2011], learning a Bayesian model is an NP-hard problem which justifies the usage of a heuristic search algorithm such as the hill-climbing structure learning algorithm.

The score-based hill-climbing structure learning algorithm was implemented using the following score functions: BDeu score; BIC score; BDs score; K2 score and AIC score in order to learn the structure of the Bayesian model. The importance of the score function lies in its ability to determine how well the data fits to the model [Zhou 2011].

In addition, score-based algorithms put forward a criterion in which the structure of the Bayesian model can be evaluated on the proposed dataset [Zhou 2011]. As such, all possible Bayesian model structures are searched and scored and the Bayesian model structure with the highest possible score is chosen as the best structure for the Bayesian model [Zhou 2011].

3.4 Bayesian Parameter Learning using Bayesian Estimation

This research utilised the Bayesian estimation estimator in an attempt to learn the parameters of the Bayesian model. In order to observe the impact of learning the associated probability distributions for the Bayesian model with different sample sizes of the dataset, the parameters of the Bayesian model are learned for different sizes of the sample dataset and compared with the ground truth model using a statistical distance metric.

Bayesian estimation considers the parameters (θ) as random variables as opposed to considering them as unknown constants [Mathye and Ajoodha 2018]. A prior probability distribution $p(\theta)$ is formulated and is used to compute the posterior probability $p(\theta|D)$ from the dataset D [Mathye and Ajoodha 2018]. According to Ajoodha and Rosman [2018], the values of the parameter θ updates as more data is obtained over time.

3.5 Kullback-Leibler (KL) Divergence Metric for Comparing Probability Distributions

Upon constructing the ground truth model and learning the Bayesian model from data, a KL divergence is computed to compare the probability distributions associated with the ground truth model with that of the learned models. The KL divergence computes the error made in learning the Bayesian model by measuring the difference in the probability distributions between the learned model and the ground truth model [Moral *et al.* 2021].

The computation of the KL divergence (denoted as $KL_{G,L}$), adapted from Moral *et al.* [2021], between the ground truth model (G) and the learned model (L), defined on the same set of variables X = (x_1,x_2...,x_n), is presented in Equation 3.1.

$$\begin{aligned} KL_{G,L} &= \sum_x p^G(x) log(\frac{p^G(x)}{p^L(x)}) \\ &= \sum_x p^G(x)[log(p^G(x)) - log(p^L(x))] \\ &= \sum_x p^G(x) log(p^G(x)) - \sum_x p^G(x) log(p^L(x)) \end{aligned} \tag{3.1}$$

The purpose of the KL divergence statistical measure is to determine the statistical difference between the probability distribution associated with the ground truth model (G) which is represented by $p^G(x)$ and the probability distribution associated with the learned model (L) which is represented by $p^L(x)$.

3.6 Inference for Bayesian Models

The inference process for Bayesian models involves computing the probability of the variable of interest based on the condition of a given evidence or set of fixed variables [Tubía *et al.* 2023]. The aim of inference is to compute the posterior probability $P(X|E)$ of the variable of interest (state variable) X given the observable evidence E as illustrated in Bayes theorem in Equation 3.2.

The probability of the evidence variable E given the state variable X represented as $P(E|X)$ is known as the likelihood. The probability of the state variable X, represented as $P(X)$, is known as the prior probability whereas the probability of the evidence E, represented as $P(E)$, is known as the marginal probability. As such, the combination of these parameters give rise

to the equation in Bayes theorem (Equation 3.2) which is the basis for the computation of the posterior probability distribution $P(X|E)$ in inference calculations.

$$P(X|E) = \frac{P(E|X).P(X)}{P(E)} \tag{3.2}$$

Several inference algorithms exist to compute the posterior probability of the state variable given the evidence. Inference algorithms are categorised into two categories, namely: exact inference algorithms and approximate inference algorithms. The examples of the two categories of inference algorithms are represented in the illustration in Figure 3.6 adapted by Grim and Felzenszwalb [2022]. The inference algorithms colour-coded in green represent the exact inference algorithms whereas the approximate inference algorithms are colour-coded in yellow.

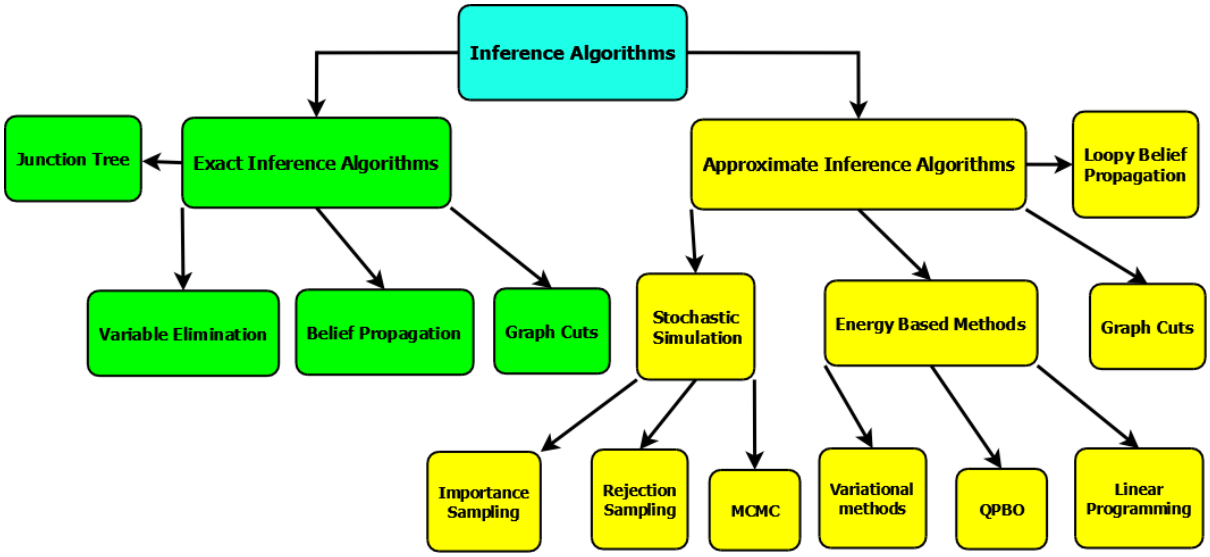

Figure 3.6: Examples of inference algorithms categorised into exact inference algorithms and approximate inference algorithms adapted by Grim and Felzenszwalb [2022].

This research implemented an exact inference algorithm known as variable elimination to compute the posterior distributions. The inference results aimed to justify the reasoning for Bayesian probabilistic modelling to predict student performance as opposed to the APS metric currently employed to gauge the performance of a student.

The reason for implementing the variable elimination inference algorithm is due to its simplistic nature [Ren, Dongping *et al.* 2022]. It is considered a fundamental probabilistic inference algorithm when employed on Bayesian models [Aslay *et al.* 2021]. The inference computations performed in this research are fairly simple test cases to illustrate the advantage of probabilistic-graphical modelling to determine student success as opposed to placing reliance solely on the APS metric. As such, for the purposes of this research, variable elimination is the sufficient choice for inference computation.

3.6.1 Exact Inference

For many real world applications, exact inference is known to be tractable and is limited by its performance which is deemed exponential in the worse case [Tubía *et al.* 2023]. According to Grim and Felzenszwalb [2022], for arbitrary joint distributions defined over a larger group of random variables, exact inference is known to be intractable.

This category of inference takes an evidence e and initial variable v and uses them to compute the likelihood of e ($P(e)$) and the marginal distribution ($P(V = v)$) or posterior distribution ($P(V = v|e)$) of v [Rama *et al.* 2012]. The purpose behind the implementation of exact inference algorithms is to manipulate the joint distribution in order to reduce the overall computational complexity [Grim and Felzenszwalb 2022]. The exact inference algorithm discussed and employed in this research is the variable elimination inference algorithm.

3.6.1.1 Variable Elimination

According to Ren, Dongping *et al.* [2022], variable elimination is an exact reasoning algorithm with the ability to solve complex reasoning associated with Bayesian models. The complexity associated with using the variable elimination methodology lies in the order of its elimination with the detection of the optimal elimination order classified as an NP-hard problem [Ren, Dongping *et al.* 2022].

The algorithm for executing the variable elimination operation, adapted from Simonsson and Mostad [2018], has been presented in the pseudocode in Algorithm 1. The operations of importance in the variable elimination procedure is the factor multiplication and factor marginalization in which the detailed definitions can be found in the study by Simonsson and Mostad [2018].

According to Algorithm 1, the procedure for the variable elimination algorithm requires a set of conditional probability distributions (CPDs) that are used to construct the factors (also referred to as potentials) $\Phi = \{ \phi_1, ..., \phi_N \}$. The operations that are possible on the factors are marginalisation (summation) of a variable and factor multiplication. Other requirements to implement the variable elimination algorithm include the query variables X to compute the posterior distribution; the observable variables y and the fixed elimination variables Z.

The fixed elimination variables stored in Z are ordered. According to Ren, Dongping *et al.* [2022], in terms of the computational cost of the elimination method, the cost for elimination is considered to be $\delta = \sum_{i=1}^{n} \delta(Z_i)$ where $\delta(Z_i) = |Z| \prod_{i=1}^{l} |X_i|$. Hence, $\delta(Z_i)$ is the referred to as the elimination cost of the variable Z [Ren, Dongping *et al.* 2022]. The computation of the query involves summing out the variables in Z one at time [Aslay *et al.* 2021]. Upon completing the factor multiplication and marginalisation of the variable Z procedures, the algorithm returns the new factors Φ with the variables from Z removed from the initial factors.

According to Tubía *et al.* [2023], the computation of variable elimination is contributed by eliminating all variables stored in Z one at a time after an ordering and operation has been performed over the potentials. The potentials refers to the local distributions of the variable and the parents of the variable [Tubía *et al.* 2023]. An example illustration of the variable elimination process, illustrating the elimination of two variables, has been presented in Figure 3.7 adapted by Tubía *et al.* [2023].

Algorithm 1 Variable elimination algorithm adapted by Simonsson and Mostad [2018]

Require: a set $\prod$ of CPDs associated with the Bayesian model, the query variables X in which to compute the posterior distribution and the evidence variables y which are considered the observed variables

1. Using the CPDs contained in the set $\prod$, the factors $\Phi = \{ \phi_1, ..., \phi_N \}$ are constructed
2. The variable set is partitioned into query X, evidence y and elimination Z
3. **for** all i = 1,...,N **do**
4. Replace ϕ_i with $\phi_i[Y = y]$
5. **end for**
6. From the variables contained in Z, the ordering $Z_1, ..., Z_k$ are fixed
7. **for** i = 1,...,k **do**
8. $\Phi^{'} \leftarrow \{\phi \in \Phi : Z_i \in Scope(\phi)\}$
9. $\Phi^{''} \leftarrow \Phi \backslash \Phi^{'}$
10. $\psi \leftarrow \prod_{\phi \in \Phi^{'}} \phi$ - This is the factor multiplication procedure
11. $\rho \leftarrow \int_{Z_i} \psi(\cdot) dZ_i$ - This is the factor marginalization procedure
12. $\Phi \leftarrow \Phi^{''} \cup \{\rho\}$
13. **end for**
14. **return** Φ

The illustration in Figure 3.7 is a representation of the graphical procedure of eliminating the variables coherence and difficulty. It is clear from the diagram that the order of elimination is the variable coherence first followed by the variable difficulty. The variable elimination procedure terminates when there is no longer variables to eliminate and returns the variables with the absence of the eliminated variables.

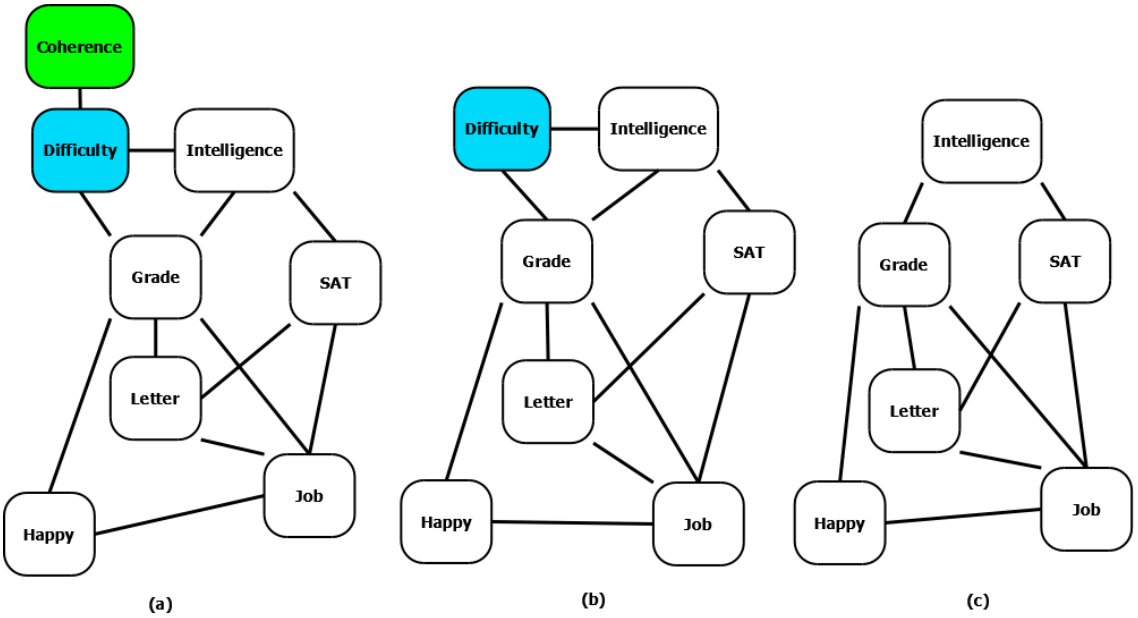

Figure 3.7: An example illustration of the variable elimination process adapted by Tubía *et al.* [2023] where (a) represents the variables *coherence* and *difficulty* in the list of variables before they are eliminated; (b) represents the elimination of the variable *coherence* and (c) represents the elimination of the variable *difficulty*.

The next section will present and discuss the results from experiments conducted from this methodology including the KL divergence results; variable elimination inference results; results as they pertain to answering the proposed research questions and the limitations of this research.

Chapter 4

Results and Discussion

The expectation from implementing the outlined methodology is to approach the prediction of student success trajectories from a probabilistic-graphical point of view by employing Bayesian probabilistic modelling. The hill-climbing structure learning algorithm is employed to learn the structures of a variety of Bayesian models and comparing the probability distribution associated with the learned models and the ground truth model using the KL divergence statistical measure. The final model selected is as close to the ground truth model based on their associated probability distributions in order to use it to determine student success.

The ground truth model illustrated by Figure 3.4 represented the true higher education trajectory developed based on expert knowledge. The structure learning algorithms employed to learn the learned models to compare with the ground truth model included: the k2 score-based model; the BDeu score-based model; the BDs score-based model and the BIC score-based model. The Bayesian estimation estimator was employed to learn the probability distributions associated with the nodes of the learned Bayesian models and produce a continuous probability table (CPT).

This section will explore the outcomes emanating from the results by implementing the outlined methodology of this research. Thus, these results formulate the answers for the proposed research questions. In addition, this section will discuss the limitations associated with implementing this research.

4.1 KL Divergence Probability Distribution Comparison

The aim of employing a statistical distance measure is to be able to ascertain how well the learned models learned by comparing these models with a true model. The statistical distance measure employed in this research was the KL divergence metric. An average KL divergence was computed between the developed Bayesian ground truth model and the various learned models, constructed using the hill-climbing structure learning algorithm, to compare the probability distribution associated with the academic modules for the students when constructing the Bayesian probabilistic model.

Table 4.1: Average KL divergence comparing the probability distributions associated with the ground truth model and various learned models for varying sample sizes of the dataset.

Sample size	**Bayesian models**			
	BDs score model	**BDeu score model**	**BIC score model**	**K2 score model**
100	0.33597	0.30192	0.31801	0.28345
1000	0.16072	0.14104	0.16332	0.15022
2000	0.12809	0.12760	0.13123	0.12502
3000	0.12277	0.10872	0.11997	0.10979
4000	0.09476	0.10293	0.10861	0.10519
5000	0.09191	0.10134	0.10338	0.09668
6000	0.08964	0.09275	0.10101	0.09757
7000	0.08769	0.09726	0.09630	0.09026
8000	0.08893	0.09490	0.09480	0.09472
9000	0.08629	0.09244	0.09243	0.09228

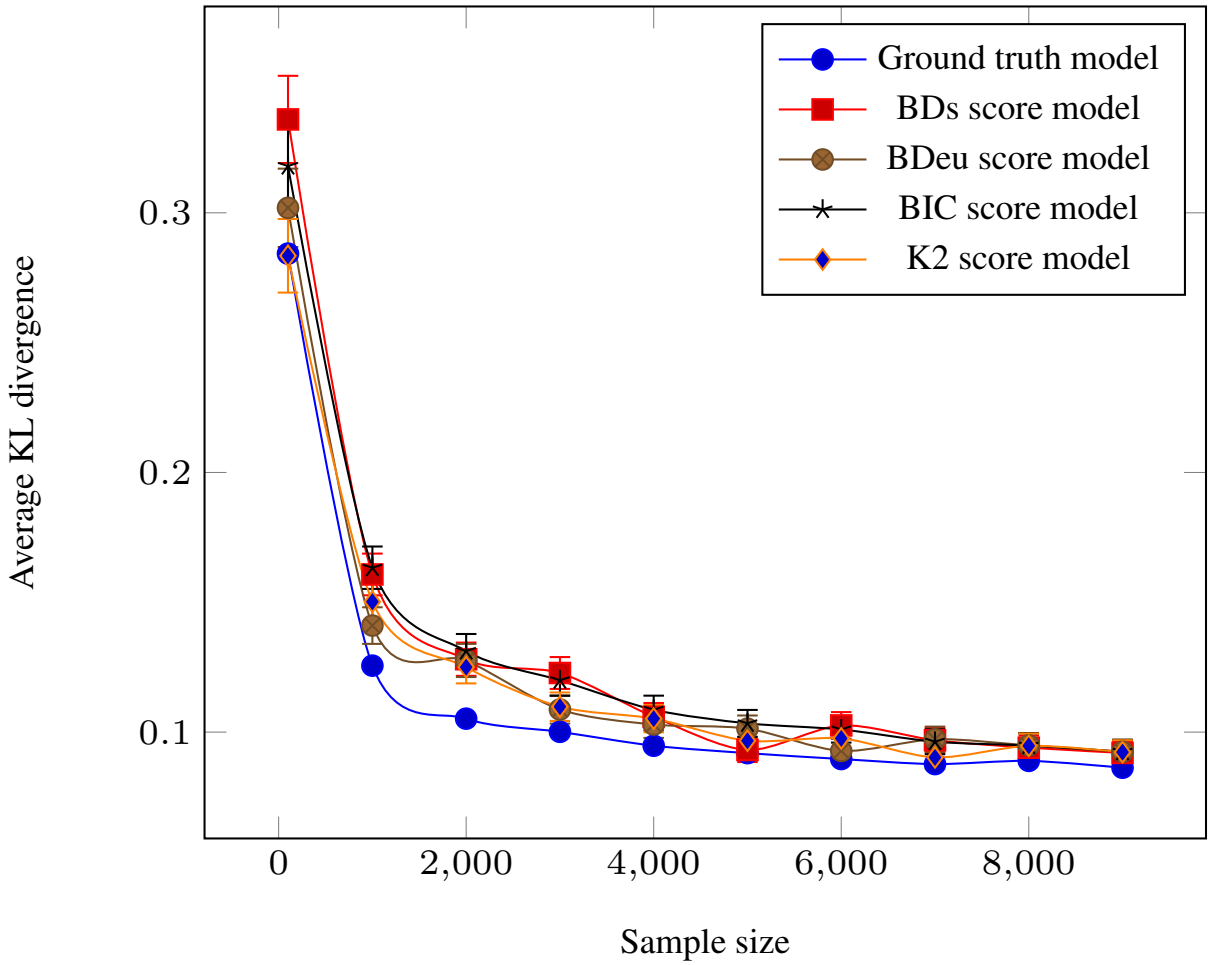

Figure 4.1: Average KL divergence comparing the probability distributions associated with the ground truth model and various learned models with increasing sample size with the addition of error bars (5% error).

The results presented in Table 4.1 illustrate that the average KL divergence decreases as the sample size increases. This suggests that the probability distributions computed during the learning process for the various learned models closely resembles the ground truth model for large sizes of the sample dataset. Furthermore, based on the results presented in Table 4.1, the BDs score-based Bayesian model resulted as the Bayesian model that closely resembled the ground truth model by a small margin when compared to the other learned Bayesian models.

The resulting graph illustrated in Figure 4.1 graphically represents the impact of the sample size of the dataset with the computed KL divergence metric relative to the ground truth model represented by the blue curve. For relatively large sample sizes, the average KL divergence reaches its minimum for all Bayesian models considered including: the BDs score-based model; the BDeu score-based model; the BIC score-based model and the K2 score-based model.

The learned probability distributions associated with each of the aforementioned learned Bayesian models constructed using the hill-climbing structure learning algorithm has shown to be closer to the ground truth model considering the KL divergence plot in Figure 4.1. Thus, these results provide an indication that it is plausible to consider the hill-climbing structure learning algorithm to learn a Bayesian model to predict student success trajectories.

In addition to representing the probability distribution difference between the various learned model and the ground truth model, Figure 4.1 has incorporated error bars, using 5% error, to each of the points in the curves of the various learned Bayesian models. The addition of error bars represent the variability of the computed average KL divergence and are an indication of the error made in computing the KL divergence for increasing sample sizes of the dataset for each learned Bayesian model.

An interesting observation in Figure 4.1 is that as the sample size increases the lengths of the error bars get smaller for all Bayesian models considered. This is an indication that the learned model gets better at learning and resembling the true model as the sample size increases.

Previous literature studied in Section 2 that modelled student results, particularly literature concerned with the employment of probabilistic graphical models, utilised either the naive Bayes model or the static Bayesian model. The random variables of the naive Bayes model operated under the independence assumption whereas the random variables of the static Bayesian model do not change over time. This research took it a step further by exploring random variables that change over time, hence, the employment of a dynamic Bayesian model.

The advantage of implementing the dynamic Bayesian model over the static Bayesian model is the ability to model the trajectory of students' results over the duration of the degree. In addition, the inclusion of the KL divergence computation provides clarity on the learned model that best resembles the ground truth model when comparing their respective probability distributions. In the case of this research, based on the results presented in Table 4.1 and Figure 4.1, the BDs score-based Bayesian model was found to resemble the ground truth model in comparison to the other learned models.

Given that the BDs score-based Bayesian model closely resembled the ground truth model based on the computed average KL divergence values in Table 4.1, the following subsection will

attempt to employ the BDs score-based Bayesian model in order to perform inference computations using the variable elimination inference algorithm. This is done so as to illustrate the impact of probabilistic-graphical modelling on higher education trajectories on student admission; student monitoring and to gauge the graduation status of students as opposed to relying solely on the APS metric.

4.2 Variable Elimination Inference Results

This research employed the variable elimination inference algorithm to illustrate the significance of probabilistic-graphical modelling on determining student success as opposed to the APS metric. This is performed by computing the posterior probability distribution of a test case given the evidence variables. The test cases in the following subsections depict student admission; student monitoring post-admission and gauge graduation status using a probabilistic-graphical approach by employing the learned BDs score-based Bayesian model.

4.2.1 Test Case 1: Student Admission

Consider a student with grade 12 results as depicted in Table 4.2. The results depicted are for mathematics and computer studies. The aim was to use these given evidence variables to compute the posterior probability distribution for first year success in higher education studies in order to determine the admission status of the student. The variable to query is the final aggregate score for first year.

Table 4.2: Student results used as evidence variables in the variable elimination inference algorithm for test case 1.

Subject	**Outcome**
Grade 12 mathematics	Pass
Grade 12 computer studies	Fail

The results from the variable elimination inference algorithm performed on the student's results in Table 4.2 when using the BDs Bayesian model are depicted in the form of a table (Table 4.3) and a graph (Figure 4.2). The results suggests that a student with the results (evidence variables) as depicted in Table 4.2 will most likely pass the first year as the majority class distribution is the *pass* class with a probability of 0.8301. The next class is the *fail* class with a probability of 0.1374 and the lowest class is the *pass with distinction* class with a probability of 0.0325. These results suggest that the student is less likely to pass first year with a distinction.

Table 4.3: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 1 depicting the probability distribution for a student's performance in first year.

First year final aggregate score	Φ**(First year final aggregate score)**
Fail	0.1374
Pass	0.8301
Pass with distinction	0.0325

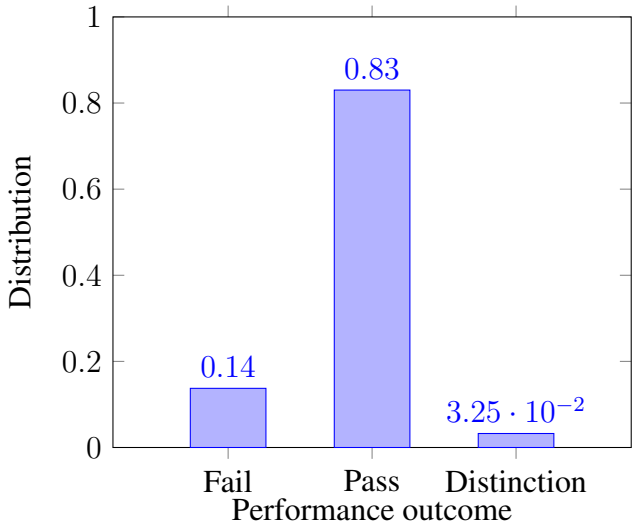

Figure 4.2: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 1 depicting the probability distribution for a student's performance in first year.

4.2.2 Test Case 2: Post-Admission Monitoring

Consider a student with results as depicted in Table 4.4. The results depicted are for grade 12 mathematics and computer studies; first year mathematics and computer science and first year aggregate score. The aim was to use these given evidence variables to compute the posterior probability distribution for second year success in higher education studies as a means to further monitor a students' academic progress in their higher education trajectory. The variable to query is the final aggregate score for second year.

The results from the variable elimination inference algorithm when using the BDs Bayesian

Table 4.4: Student results used as evidence variables in the variable elimination inference algorithm for test case 1.

Subject	**Outcome**
Grade 12 mathematics	Pass
Grade 12 computer studies	Fail
First year mathematics	Pass
First year computer science	Pass with distinction
First year aggregate score	Pass

model are depicted in the form of a table (Table 4.5) and a graph (Figure 4.3). The results suggests that a student with the results (evidence variables) as depicted in Table 4.4 will most likely pass the second year as the majority class distribution is the *pass* class with a probability of 0.7885. The next class is the *fail* class with a probability of 0.1936 and the lowest class is the *pass with distinction* class with a probability of 0.0179. These results suggest that the student is less likely to pass second year with a distinction, however, the student is likely to pass the second year of study.

Table 4.5: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 2 depicting the probability distribution for a student's performance in second year.

Second year final aggregate score	Φ**(Second year final aggregate score)**
Fail	0.1936
Pass	0.7885
Pass with distinction	0.0179

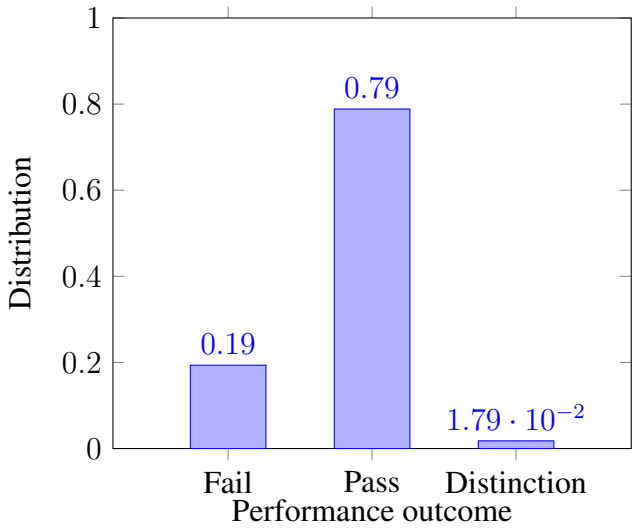

Figure 4.3: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 2 depicting the probability distribution for a student's performance in second year.

4.2.3 Test Case 3: Monitor Student Completion Status

Consider a student with results as depicted in Table 4.6. The results depicted are for grade 12 mathematics and computer studies; first year mathematics and computer science; first year aggregate score; second year mathematics and computer science and second year aggregate score. The aim was to use these given evidence variables to compute the posterior probability distribution for third year success in higher education studies as a means to gauge the completion status of the student. The variable to query is the final aggregate score for third year.

Table 4.6: Student results used as evidence variables in the variable elimination inference algorithm for test case 3.

Subject	**Outcome**
Grade 12 mathematics	Pass
Grade 12 computer studies	Fail
First year mathematics	Pass
First year computer science	Pass with distinction

First year aggregate score	Pass
Second year mathematics	Pass
Second year computer science	Pass
Second year aggregate score	Pass

Table 4.7: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 3 depicting the probability distribution for a student's performance in third year.

Third year final aggregate score	Φ**(Third year final aggregate score)**
Fail	0.2021
Pass	0.7632
Pass with distinction	0.0348

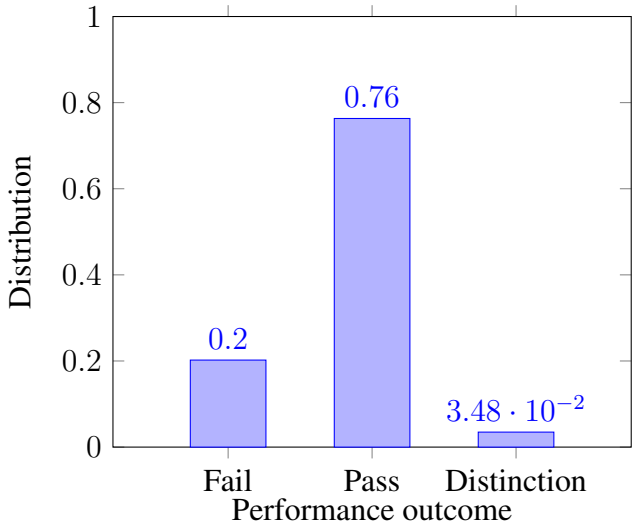

Figure 4.4: Results from the variable elimination inference algorithm using the BDs score-based Bayesian model for test case 3 depicting the probability distribution for a student's performance in third year.

The results from the variable elimination inference algorithm when using the BDs Bayesian model are depicted in the form of a table (Table 4.7) and a graph (Figure 4.4). The results suggests that a student with the results (evidence variables) as depicted in Table 4.6 will most likely pass the third year as the majority class distribution is the *pass* class with a probability of 0.7632. The next class is the *fail* class with a probability of 0.2021 and the lowest class is the *pass with distinction* class with a probability of 0.0348. These results suggest that the student is less likely to pass third year with a distinction, however, the student is likely to pass the third year of study.

4.2.4 Test Case Results Summary

The aim of the three test cases depicted in this section was to illustrate the significant impact of probabilistic-graphical models in determining student success. The three test cases depicted were student admission; post-admission monitoring and gauge student completion. The outcome from these test cases were the probabilistic distributions from the three classes of performance outcome including: fail, pass and pass with distinction.

The results further emphasize the importance of viewing a student's performance in a simplified probabilistic-graphical point of view. Higher education institutions can employ this probabilistic-graphical approach to improve student admission and monitoring post-admission thereby improving the overall graduation throughput.

The next subsection will make use of the results produced from this research to answer the research questions proposed for this research

4.3 Results as they Pertain to the Research Questions

4.3.1 Research Question 1

The graph illustrated in Figure 4.1 represented the computed average KL divergence between a variety of learned models and the ground truth model for varying sample sizes of the dataset. A methodology of this nature, which resulted in the results produced in this research, constitutes a viable course of action as it suggests that the learned probability distribution associated with the random variables of the developed Bayesian models can be compared with that of the ground truth model using a statistical distance metric in order to ascertain how well the Bayesian models were able to learn relative to the ground truth model.

Given this result, it is plausible to make use of these learned models to determine student success trajectories by using inference algorithms, such as the variable elimination inference algorithm employed in this research, to compute the associated performance distribution of the student.

4.3.2 Research Question 2

The advantage of using a Bayesian probabilistic model is that it provides an effective simplified probabilistic view of a student's outcome as opposed to using the APS metric. The graphical

nature of the developed Bayesian model depicts the dependencies and independencies that exist between a student's results in a higher education trajectory that spans three years.

Furthermore, the results from the variable elimination inference algorithm performed to compute the posterior distribution based on the given evidence highlights a probabilistic view of determining student success as opposed to depending solely on the APS metric. The examples performed in the test cases in the variable elimination inference section illustrated the performance distribution for fail; pass and pass with distinction.

This suggests that a probabilistic view of students' academic progress has the potential of assisting higher education institution administrators gauge their potential for admittance to an academic programme based on the their predicted academic trajectory and monitor their academic progress post admission stage on-route to graduation. This is in line with the aim of improving the overall on-time graduating rate.

The following subsection will discuss the limitations for conducting the experiment for this research.

4.4 Research Limitation

The results for this research are based solely on the contents of the synthetic dataset generated for this research. Thus, it has not been tested with a dataset that represents the real world student environment which would be beneficial in confirming these results.

Experimenting on real world data is beneficial in gaining more insights into a student's higher education trajectory in a practical environment so as to put in place interventions in higher education institutions to maximise the probability of graduating on time.

Chapter 5

Conclusion

The purpose of this research was to approach determining the success of student trajectories from a probabilistic-graphical point of view by employing Bayesian probabilistic modelling. The rationale for this was as a result of growing concerns over high dropout rates and low graduation throughputs in higher education institutions [Temoso and Myeki 2022]. In addition, the purpose was to find an alternative to the APS metric currently utilised by South African higher education institutions as a student admission requirement.

The features used in constructing the Bayesian model results emanated from computer science and mathematics majors. The ground truth model, depicted in Figure 3.4, was developed from expert knowledge and used as a base model to compare against when learning a Bayesian model. The ground truth model developed was a dynamic model with four time slices. Time slice 0 represented the grade 12 results of a student whereas time slices 1,2 and 3 represented the higher education journey of a student in first year; second year and third year in a higher education trajectory that spanned a trajectory of a period of three years.

This research explored Bayesian models implemented using various score functions of the popular hill-climbing structure learning algorithm. The score functions included: BDs score; BDeu score; BIC score and K2 score. In addition, the parameters of the learned models were computed using the Bayesian Estimation estimator for different sizes of the sample dataset.

The results from this research presented the computed average KL divergence between the probability distributions of the various learned Bayesian models with the ground truth model for increasing sample size. The results of the computed average KL divergence was presented in Table 4.1. In terms of a graphical representation, the observed effect in Figure 4.1 displayed an exponential decrease in the average KL divergence with increasing sample size for each learned Bayesian model relative to the ground truth model. A representation of this nature suggests that for large sample datasets, the learned models gets better at learning and closely resembles the ground truth model.

Furthermore, this research implemented an exact inference algorithm known as variable elimination. The purpose of implementing the inference algorithm was to compute the posterior distribution of a set query based on given evidence with the outcome of the algorithm presenting a probabilistic view of a student's aggregate score distribution.

The test cases presented in this research was to further emphasize the significance of the probabilistic-graphical approach to predicting student success as opposed to relying solely on the APS metric. The test cases included: student admission; post-admission monitoring and gauge student completion.

The outcome from the KL divergence results are indicative of how well the Bayesian models learned when compared with the ground truth model by comparing their associated probability distributions. The outcome from the inference algorithm are indicative of an alternative to viewing the higher education trajectory of a student and using it to compute the probability distribution of a student's final results as opposed to relying solely on the APS metric.

The main takeaway from this research is that the development of a dynamic Bayesian model provides a clear picture of a student's higher education trajectory for a three year degree. In comparison, static Bayesian models would only view a student's performance at a single point in time. In addition to the APS metric (currently used to admit students for higher education studies), higher education administrators would benefit from a complete view of students' projected results for the degree of their choice in order to make the final admission decision.

5.1 Contribution

As alluded to by Nudelman *et al.* [2019], the significance of employing a probabilistic-graphical model lies not only on its predictive capabilities, but also on the graphical nature that is prominent in gaining information and understanding associations that exist between the student attributes in predicting the success of students. As such, this research provides a simplified probabilistic-graphical view of predicting student success trajectories using Bayesian probabilistic modelling with the intention of improving the overall on-time graduation throughput.

5.2 Future Work

There are opportunities for researchers to test the results from this research by using real world data. Testing the results of this research using real world data would provide more insights in determining student success in a practical environment. Various other structure learning techniques can be employed in order to produce Bayesian models that can be compared with the ground truth model in terms of their associated random variable probability distributions as presented in the results of this research.

References

[Abed *et al.* 2020] Tasneem Abed, Ritesh Ajoodha, and Ashwini Jadhav. A prediction model to improve student placement at a south african higher education institution. pages 1–6, 01 2020.

[Aissaoui *et al.* 2020] Ouafae Aissaoui, Yasser Madani, Lahcen Oughdir, Ahmed Dakkak, and Youssouf EL ALLIOUI. *A Multiple Linear Regression-Based Approach to Predict Student Performance*, pages 9–23. 01 2020.

[Ajoodha and Rosman 2018] Ritesh Ajoodha and Benjamin Rosman. *Influence modelling and learning between dynamic bayesian networks using score-based structure learning*. PhD thesis, The University of the Witwatersrand, 2018.

[Ajoodha *et al.* 2020] Ritesh Ajoodha, Ashwini Jadhav, and Shalini Dukhan. Forecasting learner attrition for student success at a south african university. pages 19–28, 09 2020.

[Ajoodha 2022] Ritesh Ajoodha. Identifying academically vulnerable learners in first-year science programmes at a South African higher-education institution. *South African Computer Journal*, 34:120 – 148, 12 2022.

[Al Mayahi and Al-Bahri 2020] Khalfan Al Mayahi and Mahmood Al-Bahri. Machine learning based predicting student academic success. In *2020 12th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 264–268, 2020.

[Almarabheh *et al.* 2022] Amer Almarabheh, Mohamed Hany Shehata, Abdulrahman Ismaeel, Hani Atwa, and Ahmed Jaradat. Predictive validity of admission criteria in predicting academic performance of medical students: A retrospective cohort study. *Frontiers in Medicine*, 9, 2022.

[Aslay *et al.* 2021] Cigdem Aslay, Martino Ciaperoni, Aristides Gionis, and Michael Mathioudakis. *Query the model: precomputations for efficient inference with Bayesian Networks*, 2021.

[Bekele and Menzel 2005] Rahel Bekele and Wolfgang Menzel. A bayesian approach to predict performance of a student (bapps): A case with ethiopian students. pages 189–194, 01 2005.

[Buraimoh *et al.* 2021] Eluwumi Buraimoh, Ritesh Ajoodha, and Kershree Padayachee. Application of machine learning techniques to the prediction of student success. pages 1–6, 04 2021.

[Cozman 2000] Fabio Cozman. Generalizing variable elimination in bayesian networks. *Workshop on Probabilistic Reasoning in Artificial Intelligence*, 11 2000.

[Cross and Carpentier 2009] Michael Cross and Claude Carpentier. 'new students' in south african higher education : institutional culture, student performance and the challenge of democratisation. *Perspectives in Education*, 27:6–18, 2009.

[Grim and Felzenszwalb 2022] Anna Grim and Pedro Felzenszwalb. *Message Passing Dynamics of Belief Propagation Algorithms*. PhD thesis, Brown University, 2022.

[Gámez *et al.* 2011] José Gámez, Juan Mateo, and Jose Puerta. Learning bayesian networks by hill climbing: Efficient methods based on progressive restriction of the neighborhood. *Data Mining and Knowledge Discovery*, 22:106–148, 05 2011.

[Haas and Hadjar 2020] Christina Haas and Andreas Hadjar. Students' trajectories through higher education: a review of quantitative research. *Higher Education*, 79, 06 2020.

[Kustitskaya *et al.* 2020] Tatiana Kustitskaya, Alexey Kytmanov, and Mikhail Noskov. Student-at-risk detection by current learning performance indicators using bayesian networks. 04 2020.

[Mahmoud Abu Zohair 2019] Lubna Mahmoud Abu Zohair. Prediction of student's performance by modelling small dataset size. 16:18, 12 2019.

[Maluleke 2017] Risenga Maluleke. *Education Series Volume V Higher education and skills in south Africa, 2017*. Technical report, Pretoria, 2017.

[Mathye and Ajoodha 2018] Macdaline Raisibe Mathye and Ritesh Ajoodha. *A Bayesian approach to predicting student performance using biographical and enrollment observations*. 09 2018.

[Mihajlović and Petkovic 2001] Vojkan Mihajlović and M. Petkovic. Dynamic bayesian networks: A state of the art. *Europhysics Letters (epl)*, 01 2001.

[Mngadi *et al.* 2020] Noluthando Mngadi, Ritesh Ajoodha, and Ashwnini Jadhav. *A Theoretical Model to Predict Undergraduate Learner Attrition using Background, Individual, and Schooling Attributes*. 08 2020.

[Moradi and Khanteymoori 2012] Parham Moradi and Alireza Khanteymoori. Predict student scores using bayesian networks. *Procedia - Social and Behavioral Sciences*, 46:4476–4480, 12 2012.

[Moral *et al.* 2021] Serafín Moral, Andrés Cano, and Manuel Gómez-Olmedo. Computation of kullback–leibler divergence in bayesian networks. *Entropy*, 23:1122, 08 2021.

[Napitupulu and Walanda 2018] Mery Napitupulu and Daud Walanda. Relevance of admission system on students' grade point average: A case study. 01 2018.

[Nedeva and Pehlivanova 2021] Veselina Nedeva and Tanya Pehlivanova. Students' performance analyses using machine learning algorithms in weka. *IOP Conference Series: Materials Science and Engineering*, 1031:012061, 01 2021.

[Nudelman *et al.* 2019] Zachary Nudelman, Deshendran Moodley, and Sonia Berman. *Using Bayesian Networks and Machine Learning to Predict Computer Science Success*, pages 207–222. 01 2019.

[Rama *et al.* 2012] Siva Rama, Krishna Jaladi, and Dharmaiah Devarapalli. Bayesian network and variable elimination algorithm for reasoning under uncertainty. *IJCSIT*, 3:5227–5230, 12 2012.

[Ren, Dongping *et al.* 2022] Ren, Dongping, Guo, Jianxi, and Hao, Xiaoli. Bayesian network variable elimination method optimal elimination order construction. *ITM Web Conf.*, 45:01012, 2022.

[Simonsson and Mostad 2018] Ivar Simonsson and Petter Mostad. *Exact inference in Bayesian networks and applications in forensic statistics*. PhD thesis, Chalmers University of Technology and University of Gothenburg, 2018.

[Temoso and Myeki 2022] Omphile Temoso and Lindikaya Myeki. Estimating south african higher education productivity and its determinants using färe-primont index: Are historically disadvantaged universities catching up? *Research in Higher Education*, 64:1–22, 05 2022.

[Tinto 1975] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975.

[Tubía *et al.* 2023] Marta Alonso Tubía, Concha Bielza, and Pedro Larrañaga. Facilitating the inference interpretation in bayesian networks. Universidad Politécnica de Madrid, 2023.

[Yedidia *et al.* 2003] Jonathan Yedidia, William Freeman, and Yair Weiss. *Understanding belief propagation and its generalizations*, volume 8, pages 239–269. 01 2003.

[Zhou 2011] Yang Zhou. Structure learning of probabilistic graphical models: A comprehensive survey. *ArXiv*, abs/1111.6925, 2011.