UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Evaluating A Human-Machine Student Intervention Framework in Higher Education from Legacy Data

by

Herkulaas Morkel van Eyssen Combrink

Supervisor: Prof. V. Marivate
Co-Supervisor: Prof. B. Rosman

Submitted in partial fulfilment of the requirements for the degree of
Philosophiae Doctor in Computer Sciences

in the

Faculty of Engineering, Built Environment and Information Technology
University of Pretoria

June 2023

"It is not about what education can do for me, but what I can do for others through my education" – HMvE Combrink

SUMMARY

Evaluating A Human-Machine Student Intervention Framework in Higher Education from Legacy Data

by

Herkulaas Morkel van Eyssen Combrink

Supervisor: Prof. V. Marivate
Co-Supervisor: Prof. B. Rosman
Department: Computer Sciences
University: University of Pretoria
Degree: Philosophiae Doctor
Keywords: Bayesian Networks, Deep Learning, Education, Framework, Machine Learning, Multi-Armed Bandits, Reinforcement Learning, Synthetic Data

This study aims to answer the question: “Can education be abstracted in a framework and conceptually studied for a student intervention process?” To do so, it designs and tests a student intervention framework that uses a systems approach and complexity theory to learn from different student contexts and recommend and apply systemic interventions for students. The framework is based on the assumption that education is a complex system that involves multiple interacting elements, such as students, teachers, curriculum, policies, and resources. The specific interventions and their impact on student success are not assessed in this work, as they depend on the expertise in the education domain. Rather, this work focuses on the framework that can be used to apply and evaluate different contexts and interventions. The study applies complexity theory and systemic intervention theory to understand the lens and the methods for studying the system. It also explains the education system in South Africa as the context of the study. One of the main challenges of the study is data sharing and data handling. To address this challenge, the study generates synthetic data from openly available tabular data and evaluates its conditional interdependence using different machine learning classification tasks. Then, it applies the same methods to a real-world education

dataset from the University of the Free State. The study contextualizes the student intervention framework as a multi-armed bandit (MAB) stateless reinforcement learning problem and tests its performance and viability using probabilistic models. The results show that the probabilistic models yield the best results with the minimum required fine-tuning, and that they scale well to the real-world dataset. The results also indicate that the framework is viable for student intervention recommendations within the context of contextual bandits. The study concludes with a discussion of the implications and limitations of the framework, and suggests areas for future research.

OPSOMMING

Evaluating A Human-Machine Student Intervention Framework in Higher Education from Legacy Data

by

Herkulaas Morkel van Eyssen Combrink

Supervisor: Prof. V. Marivate
Co-Supervisor: Prof. B. Rosman
Department: Computer Sciences
University: University of Pretoria
Degree: Philosophiae Doctor
Keywords: Bayesian Networks, Deep Learning, Education, Framework, Machine Learning, Multi-Armed Bandits, Reinforcement Learning, Synthetic Data

Hierdie studie het ten doel om die vraag te beantwoord: “Kan onderwys in 'n raamwerk geabstraheer en konseptueel bestudeer word vir 'n studente-intervensieproses?” Om dit te doen, ontwerp en toets dit 'n studente-intervensieraamwerk wat 'n stelselbenadering en kompleksiteitsteorie gebruik om van verskillende studentekontekste te leer en sistemiese intervensies vir studente aan te beveel en toe te pas. Die raamwerk is gebaseer op die aanname dat onderwys 'n komplekse stelsel is wat verskeie interaktiewe elemente behels, soos studente, onderwysers, kurrikulum, beleid en hulpbronne. Die spesifieke intervensies en hul impak op studentesukses word nie in hierdie werk beoordeel nie, aangesien dit afhang van die kundigheid in die onderwysdomein. Inteendeel, hierdie werk fokus op die raamwerk wat gebruik kan word om verskillende kontekste en intervensies toe te pas en te evalueer. Die studie pas kompleksiteitsteorie en sistemiese intervensieteorie toe om die lens en die metodes vir die bestudering van die stelsel te verstaan. Dit verduidelik ook die onderwysstelsel in Suid-Afrika as die konteks van die studie. Een van die hoofuitdagings van die studie is datadeling en -hantering. Om hierdie uitdaging aan te spreek, genereer die studie sintetiese data uit oop beskikbare tabeldata en evalueer dit die voorwaardelike

interafhanklikheid met behulp van verskillende masjienleer klassifikasietake. Dan pas dit dieselfde metodes toe op 'n werklike onderwysdatastel van die Universiteit van die Vrystaat. Die studie kontekstualiseer die studente-intervensieraamwerk as 'n multi-arm bandiet (MAB) staatlose versterkingsleerprobleem, en toets sy prestasie en lewensvatbaarheid met behulp van waarskynlikheidsmodelle. Die resultate toon dat die waarskynlikheidsmodelle die beste resultate lewer met die minimum vereiste fynafstelling, en dat hulle goed skaal na die werklike datastel. Die resultate dui ook daarop dat die raamwerk lewensvatbaar is vir studente-intervensieaanbevelings binne die konteks van kontekstuele bandiete. Die studie sluit af met 'n bespreking van die implikasies en beperkings van die raamwerk, en stel gebiede vir toekomstige navorsing voor.

KAKARETSO

Evaluating A Human-Machine Student Intervention Framework in Higher Education from Legacy Data

ka

Herkulaas Morkel van Eyssen Combrink

Supervisor: Prof. V. Marivate
Co-Supervisor: Prof. B. Rosman
Department: Computer Sciences
University: University of Pretoria
Degree: Philosophiae Doctor
Keywords: Bayesian Networks, Deep Learning, Education, Framework, Machine Learning, Multi-Armed Bandits, Reinforcement Learning, Synthetic Data

Thuto e e dira go araba potso: “A thuto e ka abstrakwa mo sebokeng le go ithutelwa ka tsela ya konseptuele go dira tsamaiso ya go tsaya karolo ya baithuti?” Go dira jalo, e dira le go leka seboka sa go tsaya karolo ya baithuti se se dirisang tsela ya ditiragalo le teori ya boiteko go ithuta go tswa mo dikonteksteng tsa baithuti tse di farologaneng le go neela baithuti ditlamelo tsa ditiragalo le go di dirisa. Seboka se se theilweng mo go bonaleng gore thuto ke tiragalo e e boitekang e e akaretsang ditiragalo tse di farologaneng, jaaka baithuti, barutabana, kurikulamo, melao, le dikgwebo. Ditlamelo tse di kgethegileng le maatla a a nang le one mo go atlegeng ga baithuti ga a lekolwa mo tiro eno, ka gore e a itshetlela boitseanape mo lefelong la thuto. Go feta fa, tiro eno e fokotsa mo sebokeng se se ka dirisiwang go dirisa le go leka dikontekste tse di farologaneng le ditlamelo. Thuto e e dirisa teori ya boiteko le teori ya go tsaya karolo ya ditiragalo go tlhalosa lenaneo le metotso ya go ithuta tiragalo. E tlhalosa le tiragalo ya thuto mo Aforika Borwa jaaka kontekste ya thuto. E nngwe ya ditshwanelo tse di kwa godimo tsa thuto ke go abelana le go laola data. Go rarabolola ditshwanelo tse, thuto e e dira data e e dirilweng go tswa mo data ya tabulere e e sa tlhokomelwang le go e leka go tswa mo go bonaleng gore e na le maemo a a farologaneng

go dirisa metotso ya go ithuta ya meesene e e farologaneng ya go kgaoganya. Ka morago, e dirisa metotso e e tsamaisanang le data ya thuto ya nnete mo Aforika Borwa mo Yunibesithing ya Freistata. Thuto e e dira go bonala seboka sa go tsaya karolo ya baithuti jaaka mathata a a nang le diatla tse di farologaneng (MAB) a a sa nang le maemo a go ithuta go rarabolola. Dipoelo tsa ditlhotlhomiso di bontsha gore metotso ya go tlhoka go itse pele e e neela dipoelo tse di kwa godimo, le go tlhoka go fetola sentle. Gape, dipoelo di arogana sentle le data ya nnete. Dipoelo di bontsha le gore seboka se se a tsamaya go neela baithuti ditlamelo tsa ditiragalo mo konteksteng ya diatla tse di farologaneng. Thuto e e tlhagisa ka poledisano ya maemo le ditshwanelo tsa seboka, le go neela dikarolo tsa dithuto tse di tlang.

DECLARATION

I declare that the thesis, which I hereby submit for the degree Philosophiae Doctor at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

ETHICS STATEMENT

The author, whose name appears on the title page of this thesis, has obtained, for the research described in this work, the applicable research ethics approval, University of Pretoria (UP) (ethics number EBIT/19/2022) and the University of the Free State (UFS) (ethics number UFS-HSD2022/0195/22).

The author declares that s/he has observed the ethical standards required in terms of the University of Pretoria's Code of Ethics for Researchers and the Policy guidelines for responsible research.

HMvE Combrink

2023

ACKNOWLEDGEMENTS

This project could not have been possible without the assistance of God, my supervisors, friends, and family.

I am indebted to my supervisors for their patience, wisdom, generosity, and knowledge. Prof. Vukosi Marivate, for the endless support and wisdom across various domains, for abundantly sharing your knowledge, for your mentorship and guidance, for your critical perspective and for the insight you provided throughout this process. Prof. Benjamin Rosman, for your support, critical perspective, guidance, and sharing knowledge which made the thesis possible and succinct. Thank you both.

To my colleagues, Philippe Burger, Martin Clark, Jacques Maritz, Anneri Muller, Ronald Sinra, and Sivuyile Nzimeni, your support on this journey was invaluable and I appreciate you all.

I would like to acknowledge my daughter, Anchelle Kamogelo Combrink, for your support, love, kindness and care. I truly appreciate your existence and value as a person.

To my grandfather Herkulaas Morkel van Eyssen Combrink, o robale ka kagiso, kea go rata ntate mogolo. To my father, Herkulaas Combrink, thank you for your love and your support throughout this journey. To my sister and brother in law, Daleen Oppermann and Selwin Oppermann, thank you both for your love and support in this journey.

To my friends, family, and people who assisted in this journey Judith Klins, Stella Naledi Klins, Alexander Sandile Klins Combrink, Michael Thabo Klins Combrink, Shiela Combrink, Albert Combrink, Mike Johnson, Carin Johnson, Inneke Zulch, Bobby Mitchley, Baphumelele Masikisiki, Lubabalo Saba, Warren Dennis Allison, Rhys Vermaas, Julia West, Natasha Wort, Didintle Jeanette Matlhoko, Amantle Portia Matlhoko, Hanta Henning, and Elmar Henning, Anne Guillaume, thank you for all your support, advice, care, and contributions.

LIST OF ABBREVIATIONS

AA	Academic Advising
AI	Artificial Intelligence
AIC	Akaike Information Criterion
A-Step	Tutorial Programmes at the UFS
BIC	Bayesian Information Criterion
BN	Bayesian Network
CI	Confidence Interval
CM	Confusion Matrix
CTL	Centre for Teaching and Learning
Cumsum	Cumulative Sums
DAG	Direct Acyclic Graph
DT	Decision Trees
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
HEI	Higher Education Institutions
kNN	k-Nearest Neighbours
LR	Logistic Regression
MAB	Multi-Armed Bandit
ML	Machine Learning
MLER	Machine Learning in Education Research
POPIA	Protection of Personal Information Act
RL	Reinforcement Learning
SD	Standard Deviation
UFS	University of the Free State
UFS101	First Year Seminar at the UFS
ULD	Unit for Language Development
UP	University of Pretoria

List of Figures

List of Tables

"The journey of a thousand miles begins with one step" - Lao Tzu

TABLE OF CONTENTS

CHAPTER 1 CONTRIBUTIONS, PROBLEM STATEMENT, AND OUTLINE OF THESIS

1.1 Introduction

The main question of this study was what would a framework look like if higher education was abstracted and conceptually studied to provide insight into an autonomous student intervention system? To address this question, a systems approach was used to develop, test, and evaluate a student intervention framework for the education domain by drawing from complexity theory and systemic intervention theory. Several technical approaches were used to generate synthetic data, measure the synthetic data's utility, and design and test an environment simulating the student intervention system as a multi-armed bandit problem. Next, the contributions of this thesis will be outlined.

1.2 Contributions of this thesis

This study gave rise to research outputs in education, data science, and philosophy. The first contribution was to outline the theories that overarch this study namely complexity theory and systemic intervention. Thereafter, synthetic data was generated for the education domain, which was then tested on a real-world dataset. Finally, systemic intervention in education was abstracted into a multi-armed bandit problem. Complexity theory as a framework outlines the importance of defining the boundaries and concepts within a system, rather than understanding a particular component within a system. Systemic intervention is rooted in complexity theory but outlines how to provide an intervention from within a system. All of these concepts were applied to the education domain using a systems approach. Before this study there was no:

1. Literature available outlining challenges, gaps, and areas for future research for machine learning in education (MLER). This study conceptualised these challenges and added to literature in this emerging field;

2. Applied literature on methods for producing synthetic education tabular data. This study explored these approaches and mapped and evaluated the methods for producing synthetic tabular data for education using a deep learning and a probabilistic model approach;
3. Applications of using machine learning classification tasks to measure the utility of synthetic tabular data for the education domain. This study made use of machine learning classification tasks to test the conditional interdependence of synthetically generated tabular education data; and
4. Environments abstracting the systemic intervention within an education system as a multi-armed bandit problem. This study created this environment and evaluated the results by comparing the overall cumulative reward for three algorithms, unexplored in this environment.

1.3 Problem statement

In South African educational institutions, the problems of student throughput rates and student retention rates persist (Moodley and Singh, 2015; Masutha, 2022). Student retention rate is when a student remains in the institution despite poor academic performance in the previous year, while a student throughput rate refers to a student's progress towards graduation (Botha, 2016; Manik and Ramrathan, 2021). Various student support strategies have been explored to improve both student throughput and retention rates, such as social and academic support from the institution, and high-impact practices that encourage student agency (San and Gou, 2023). However, the needs of students vary depending on their context, including factors such as socio-economic status, language, and support structures. Implementing interventions to support students is complicated due to the vast and complex needs of students and the availability of interventions from the institution. Real-time analytics to identify and evaluate the effectiveness of interventions are time-consuming and costly, even with dedicated staff. Thus, there is a need for a transparent and ethical student support framework to design smart systems that can adapt to different contexts. This study delved into a framework that was designed to generate synthetic data from a limited sample of a real-world dataset from South African higher education. Additionally, this framework was used to construct a multi-armed bandit (MAB) problem, which served as a map for the educational intervention framework in higher education.

1.4 Outline of the thesis

The study encapsulated in this thesis applies contexts specific to the field of education to the principles inherent in computer science, with both domains being indispensable to the research. As a result, two fundamental components are constantly in overplay with one another, the education domain, and the applied computer science domain. This translates into the background chapter, with Chapter 2 focusing on the complexity theory and the system intervention theory used in the framework. Chapter 3 is based on synthetic data generated from education tabular data and its associated utility is evaluated using different machine learning classification tasks. Chapter 4 applies the components of synthetic data generation to a South African university dataset, specific to the University of the Free State, Economic and Management Sciences. Chapter 5 explores a multi-armed bandit problem, which represents a stateless action reward feedback loop to simulate an autonomous decision maker in a variety of different contextual education problems and evaluated the different conceptual problems that may arise with this approach. Chapter 6 concludes the contributions and recommendations from this study. In the next section, the publications, research outputs, and awards from this thesis will be outlined.

1.5 Publications, research outputs, and awards from this thesis

Conference and workshop presentations

- Combrink, H.M.v.E, Marivate, V. and Rosman., 2020. Agent Based Artificial Intelligence for Timeous Student Support Strategies. *Flexible Futures*, University of Pretoria [Accessed on 19 November 2022: https://www.youtube.com/watch?v=Tz2fCwGedaw]
- Combrink, H.M.v.E, Marivate, V. and Rosman., 2022. Exploratory Data Analytics for Higher Education Tabular Datasets: An Open-Source Application for Institutional Practitioners. *SAAIR Learner Analytics Institute*, University of the Western Cape [Accessed on 19 November 2022: https://www.youtube.com/watch?v=1iY6qRzgXHw]
- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022. A Framework for Undergraduate Data Collection Strategies for Student Support Recommendation Systems in Higher Education. arXiv preprint arXiv:2210.10657.
- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2020. Multi-Armed Bandit Problems In Education. *BRICS Young Scientist Forum*.

- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022. Artificial Intelligence In Education. *BRICS Young Scientist Forum*.

Datasets and repositories

- Combrink, H.M.v.E., Carr, E. de Wet, Katinka. Marivate, V. Rosman, B., 2023. South Africa Education Data and Visualisations. University of the Free State. Dataset. https://doi.org/10.38140/ufs.22081058.v3
- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022. Github repository Exploratory Data Analytics in Education. Available on GitHub https://github.com/dsfsi/Higher_Education_EDA

Journal articles

- Combrink, H.M.v.E., Marivate, V. Masikisiki, B., Technology-Enhanced Learning, Data Sharing, and Machine Learning Challenges in South African Education. *Education Sciences*. 2023, 13, 438. https://doi.org/10.3390/educsci13050438
- Combrink, H.M.v.E., Carr, E. de Wet, Katinka. Marivate, V. Rosman, B., 2023. South Africa Education Data and Visualisations. University of the Free State. Dataset. *Data in Brief* https://doi.org/10.38140/ufs.22081058.v3
- Combrink, H.M.v.E, Marivate, V. and Rosman, B., 2022. Comparing Synthetic Tabular Data Generation Between a Probabilistic Model and a Deep Learning Model for Education Use Cases. *arXiv preprint arXiv:2210.08528*.
- Combrink, H.M.v.E, Marivate, V. and Rosman, B., 2022, December. Reinforcement learning in education: A multi-armed bandit approach. In International Conference on Emerging Technologies for Developing Countries (pp. 3-16). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35883-8_1

Awards and achievements

- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2023. Nominated as one of the top 100 Artificial Intelligence early state projects in the world, the second South African project for this nomination to the International Research Centre on Artificial Intelligence under the auspices of UNESCO**.** https://ircai.org/top100/entry/a-framework-for-designing-student-support-strategies-in-higher-education-institutions-using-artificial-intelligence/
- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2020. Representing South Africa in the category Artificial Intelligence in the 5th Annual BRICS Young Scientists Consortium.

- Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022. Representing South Africa in the category Artificial Intelligence and Computer Learning 7th Annual BRICS Young Scientists Consortium.
- Combrink, H.M.v.E., 2021. African Institute of Mathematical Sciences (AIMS) Cohort 0 Data Science.
- Combrink, H.M.v.E., 2021. Microsoft Doctoral Symposium DSAA 2021.

In the next chapter, the theoretical underpinnings of the thesis will be outlined.

CHAPTER 2 EDUCATION AS A COMPLEX SYSTEM: A FRAMEWORK FOR SYSTEMIC INTERVENTIONS

2.1 Introduction

The contribution of this chapter is to orientate the reader within the systemic approach and landscape. To outline some of the challenges associated with systemic interventions in higher education and to propose a human-machine student intervention framework for higher education. This has not been done prior using this approach. To achieve this, there were three overarching objectives in this chapter:

1. To outline the approach, theories, and theoretical lenses used in the study, including a systems approach, complexity theory, and systemic interventions within complex systems;
2. To abstract complex education systems so that they may be studied in relation to an education system; and
3. To define a student intervention framework associated with machine learning research in the South African education context.

The purpose of this chapter is to outline the approach and theories used to frame this study and thesis. Furthermore, this chapter also outlines some of the challenges associated with machine learning in education research (MLER), which has never been done for the South African context before. Finally, this chapter outlines the framework to study systemic interventions within complex education systems, which is a novel interdisciplinary field. The work in this chapter was published in Combrink *et al.*, (2023a), Combrink *et al.*, (2023b), and Combrink *et al.*, (2022a), and reference is made to this research output in this chapter. In the sections to follow, the systems approach, complicated systems, and complexity theory will be outlined.

2.2 Overview of Systems, systems that are complicated, and complex systems

A systems approach to problem solving has been extensively used across various domains to unpack the complexities and nuances required to contextually address studying a system (Agbo *et al.*, 2019; Brackett *et al.*, 2019). A systems approach is a framework to understand contexts within a system and to systemically think around components that work together and how to identify and address challenges that may be present within a system. A systems approach is an iterative process involving several conceptual steps to be addressed to unpack problems related to a system (Cennamo and Kalk, 2019). By its very definition, a system represents a set of processes or components that function together to create a collective (Scoones *et al.*, 2020). Therefore, a system consists of different parts, each with its own complexities and functions.

To apply a systems approach, a few fundamental concepts need to be concerned at all times. Firstly, the problem within a system needs to be defined. Sometimes the context of studying a system is to identify and define the problem itself, but as far as possible, the problems and challenges that are intended to be addressed needs to be outlined. This can also include the problem of identifying the actual problem if this is not known prior. In the context of higher education, there are several well-articulated issues, and if the specific set of issues are not properly outlined, the arguments around the exact exploration might be convoluted under the vast amounts of known challenges within and between the various systems. Next, defining the constraints of the system is vital for the success of implementing this approach. Constraints include contexts and concepts outside purely empirical processes. For higher education, this includes the context of the system, participants within the system, and potential constraints that are present within the system. Thereafter, the goals and objectives of the exploration need to be outlined in the context of the system. Once established, the measurements and metrics to evaluate these need to be developed. Then, potential alternatives may be explored, and the alternatives need to be evaluated using the initial models so that these two approaches (the first and the alternative) may be compared to one another (Peterson and Davie, 2007; Hermans, 2019). Finally, a decision needs to be taken regarding the system based on this exploration as a final outcome of the study (Figure 2.1).

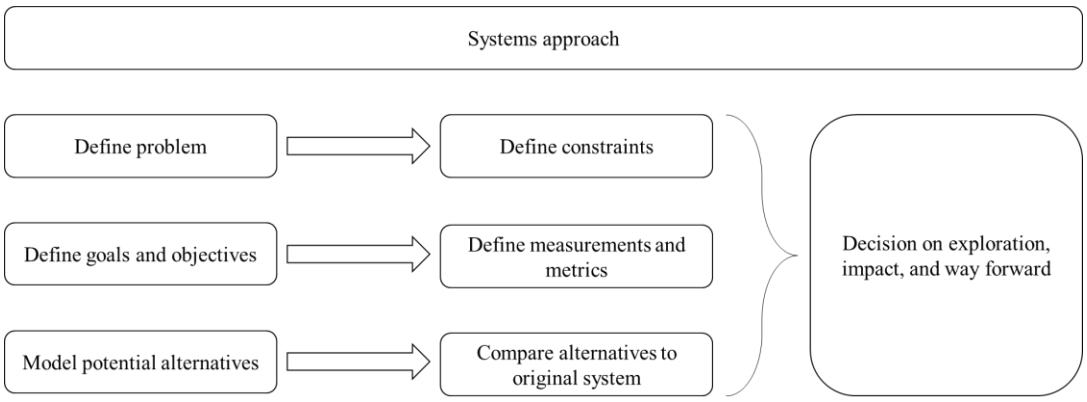


Figure 2-1 Outline of the systems approach.

In the systems approach, anomalies within a system can be used to study behaviour, relationships, interactions, and how underlying structures impact the collective. In other words, if there are unique components within a system, an anomaly within these components can be isolated and studied to provide insight into what the significance of that specific part is in relation to the system itself (Bosch *et al*., 2021). To this extent, a system in its simplest form can consist of processes and/or components that work together to function as a collective (Figure 2.2).

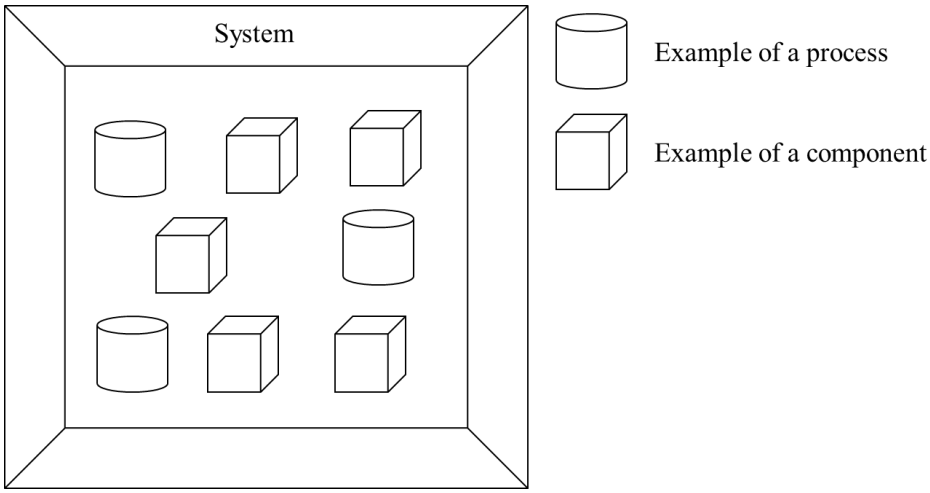


Figure 2-2 An example of a system.

An example of a system can be a classroom within higher education, filled with students, a lecturer and the material to be presented for that specific day. All the components within the classroom are important, but some components have a far more significant impact on the learning experience

than others. If a part of the classroom is missing, the impact of the fault can be studied in relation to the functioning of the entire system, in this case, how effectively a classroom would work in the absence of a content specific expert, like a lecturer. A component in this context may be one of the actors in the system, like a lecturer, and a process in this instance can be an assignment students do to gain a deeper understanding of the content. The system is thus everything involved in learning, consisting of the components (physical students and lecturer) and processes (assignments, learning, and reading etc.). However, within a classroom setting the diversity between individuals and lecturers also add a layer of complexity to this system. Within the context of a system, some systems may have more intricate complexities than others, such as the addition of socioeconomic factors that impact the access certain students had prior to attending higher education. This example, and others like it add different complexities to the initial classroom system, and in this case shall be referred to as complicated systems (Tonetto and Saurin, 2021; Pickard and Beasley, 2022).

A complicated system in this instance can be defined as a system that requires a high degree of specialisation to construct, implement, and understand all the various components within the system (Suyatinov, 2020). An example of a complicated system is the classroom that also includes the software used in teaching, the learning management system for packaging content, best teaching practices present in the classroom but also a contextual understanding of the students, their needs, pedagogy, educational frameworks, and student support strategies including academic and non-academic support to enhance student learning. Complicated systems in this instance can be used to differentiate between a system that requires limited stakeholders, processes, or components to function from one that requires a high degree of specialisation. This difference allows an understanding of both simpler systems and complicated systems. The commonalities between a system and a complicated system are that they both are confined by one overarching goal or outcome and work closely together to achieve this outcome. In the context of a system and a complicated system, the outcome is student success, by increasing throughput and retention rates, but the level of complexity differs significantly if all of the out of classroom variables, such as socioeconomic status, prior learning, and capabilities are factored into the understanding of the problem (Figure 2.3).

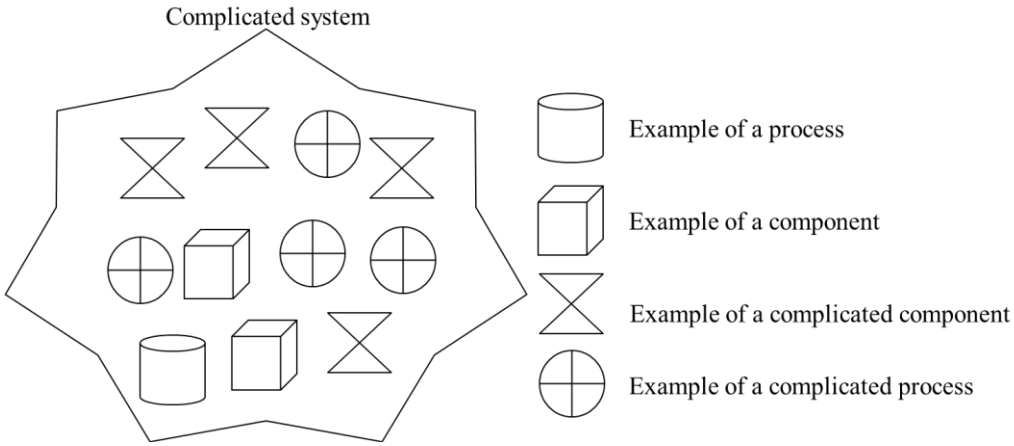


Figure 2-3 An example of a complicated system.

Understanding that systems may be either simpler, or complicated, provides some context when different systems are compared to one another within a large system. A large system may consist of a combination of systems and complicated systems. When a large system contains parts, processes, simple systems, and/or complicated systems as components, then that large system is called a complex system (Ladyman *et al.*, 2013) (Figure 2.4).

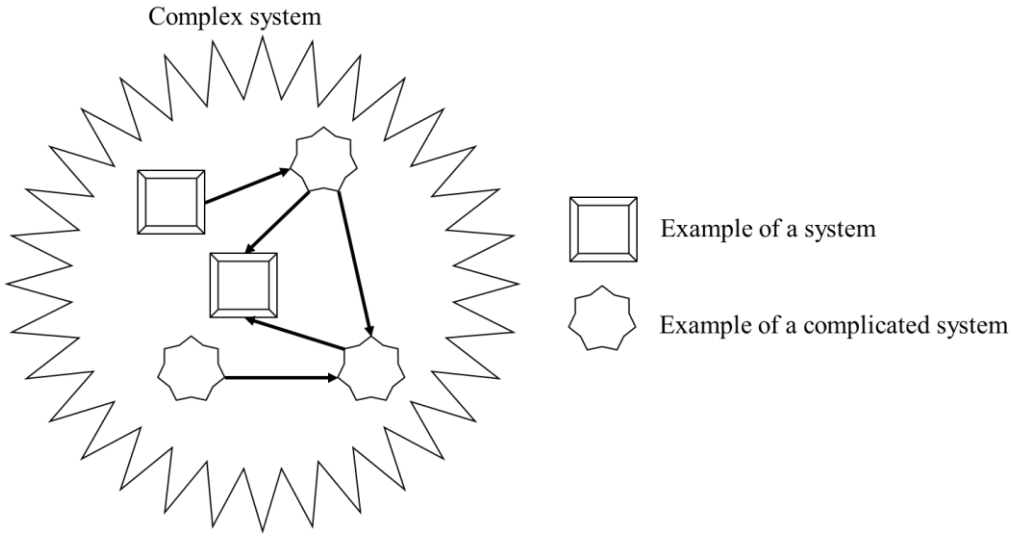


Figure 2-4 An example of a complex system.

Analysing a complex system necessitates the abstraction of concepts, with an emphasis on understanding the strength of the relationships between the various parts, processes, or components within the subsystems. This approach prioritizes the interconnectedness of the system as a whole

over the impact of any individual part within a subsystem (Svítek, 2015). In other words, understanding a complex system will require measuring the strength of the relationship between the underlying systems rather than trying to define the impact of a specific variable within a system to its outcome (Magee, 2004). In addition to this, a systems approach is required to study a complex system but with the emphasis on the strength of the relationships between variables. In the higher education context, there are several systems and complicated systems within institutions that influence a students' learning and learning outcomes, and in the next section, education as a complex system will therefore be outlined.

2.2.1 Education as a complex system

Student throughput rate and student retention rates are two fundamental concepts that keep education researchers engaged because they are important indicators that can be used by institutions to measure student success (Botha, 2018). Student throughput rate refers to: "…the rate at which a cohort successfully completes a qualification within the stipulated timeframe for that qualification" (Botha, 2016). Student retention, on the other hand, refers to a student's ability to remain in the system (Manik and Ramrathan, 2021). In this study, the term 'student' will be used to refer to a person partaking in education (Combrink and Oosthuizen, 2022).

If we consider education, throughput rate could literally mean the percentage of students that passed a particular year of study, and retention would be the percentage of students that remain in the system the following year if they are supposed to return (Marthers *et al.*, 2015). Retention will not be used for students completing a journey, which is the end goal of an education system (Lourens and Bleazard, 2016). Furthermore, both student retention and student throughput rate can be expressed as a percentage and aid decision makers within an institution with policies, decision making or outline areas that require improvement such as a particular subject students struggle with (relative to the throughput and retention rates of that subject). As a result, monitoring throughput and retention rates can serve as a litmus test to identify particular processes within subjects or across the curriculum for students that require an intervention (Janse van Vuuren, 2020). This monitoring and evaluation can either be in the form of tracking a specific subject, such as mathematics or physical sciences, or to track a variety of variables that impact overall academic success, like tracking class and online attendance and using attendance to identify a student at risk

of failing a specific year of study (van Zyl and Ngwenya, 2020). There have been several studies that outlined what types of factors should be tracked that can be used to determine, monitor, and/or evaluate student success, but most of these processes are reliant on the inputs of a person within an institution aiding the student with a particular challenge (Vondrell and Sweeney, 1989; Pritchard and Wilson, 2003; Murray *et al.*, 2008; Huberts *et al.*, 2022). Additionally, these studies are also mostly context bound with specific interventions linked to a particular region, policy, or cultural norm that might not be applicable elsewhere. In other words, the way in which students receive support will differ between institutions, departments within an institution and/or between different countries. Another added layer of complexity is that learning in and of itself is complicated and will thus differ between individuals. What makes education challenging to understand in relation to what is needed to promote student success is the diversity of processes, components, systems, and complicated systems that make up learning. It is for this reason that education should be seen as a complex system because the relationships between stakeholders, policies, contexts, and different environments have a significant impact on student throughput and retention rates (Lemke and Sabelli, 2008). Another justification for education as a complex system resides in the idea that individual components within a particular system cannot be quantified to the final outcome (Jacobson *et al.*, 2019). For example, the impact of a first year orientation programme and its relation to the final examination mark of a student in their third year cannot be determined, but rather that the relationship between orientations and student success should be studied.

A further viewpoint of complexity is to define boundaries of systems so that the strength of the relationship or the importance of different components within a complex system may be defined (Ghaffarzadegan *et al.*, 2017). Studying a phenomenon relies heavily on applying techniques to analyse complex datasets and provide insights into underlying factors and outline the value of the data itself. These methodologies thus have many overlapping concepts and often share similar approaches to predicting, categorising, and filtering through complex information to gain the desired insight. However, as these approaches become more complex and intricate, it becomes increasingly challenging to contextualise them without a perspective rooted in complexity theory as an adaptation to the systems approach. As a result, the justification of education as a complex system will be used throughout this study as it provides a framework that can assist in

understanding the complexities associated with education as a complex system. In the next section, systemic interventions within complex systems will be explained.

2.3 Systemic interventions within a complex system

The progress of technology has facilitated a connection between individuals and systems, resulting in environments that can adapt to the needs of people without human intervention (Shemshack, and Spector, 2020; Qian *et al.,* 2020). This level of autonomous optimisation was initially only present in industrial processes but has become a part of everyday life through technological devices such as computers, smartphones, and smart devices (Szłapczyński and Ghaemi, 2019; Martin *et al.,* 2020; Hall *et al.,* 2021). For the education context this is important as several authors have noted that the people within education institutions alone are not enough to deal with the increasing demand of challenges students face within institutions (du Plessis and Mestry, 2019; Brunetti *et al.,* 2020; García-Peñalvo, 2021; van der Rijst *et al.,* 2022). For this reason, more and more education institutions have turned to the use of technological tools to assist with scaling interventions that can aid student success (Rotar, 2022). However, certain interventions as part of broader education are harder to implement reliably and require more expertise. When it comes to the education sector, implementing self-learning systems requires domain-specific knowledge, or there may be consequences for users (Abu-Salih, 2021). Complexity theory and systemic intervention are theoretical frameworks that can help us understand and evaluate educational systems as complex phenomena. These frameworks highlight the dynamic and nonlinear interactions among the various elements of the system, such as teachers, students, curriculum, policies, and resources. They also enable us to critically examine the assumptions, values, and perspectives that shape our understanding of the system and its problems. By applying these frameworks, we can explore the multiple dimensions and perspectives of the system and identify the opportunities and challenges for improving its performance and outcomes. This focus is possible because less attention is given to the details and inner workings of each specific component or process within the system, but rather to view the relationships between variables as the point of study (Rothgang and Lageman, 2022).

In the context of interventions, some problems require a more immediate and urgent response, like a medical intervention, and others require more repetition and contextual support, such as a series

of tutorial classes to reinforce a concept. In this specific example, the immediate impact of the medical intervention can be studied as the consequences might be more significant in relation to the outcome of the student in the short term. However, as a system grows in size and complexity, isolated events become less useful in understanding the system as a whole, such as the impact of a specific tutorial in relation to the final mark of a student. It is therefore more important in the context of different systems to have a well-defined general student intervention framework than to have a list with specific interventions for each type of specific problem because the context and understanding for stakeholders in other parts of the system, or in different systems may be lost (Midgley and Rajagopalan, 2020). In other words, knowing how to specifically handle a student that is in need of a specific medical treatment is very important at a process level, but in terms of the entire system, and more importantly, when systems interact with other systems, it becomes useless.

In reality, specific interventions within a system are important, but the framework relating to the part this plays within a complex system is just as important (Midgley and Lindhult, 2021). This perspective highlights the importance of understanding the interplay between different actors and their impact on the system (Foote *et al.*, 2021). From individual schools to larger higher education institutions, education is made up of complex entities that are interconnected in intricate ways (Gates *et al.*, 2021). For example, the relationship a student has to their lecturer, peers, the support services and their families have a significant impact on their ability to transition into higher education, and the ability to transition has been associated with academic success, but this alone is not a predictor of academic success as there are multiple factors at play that impact a student's academic success.

Implementing interventions in education for students based on their individual needs requires a deep understanding of these complexities and the relationships between the entities involved in their learning. For example, implementing an intervention framework within education for assessing student performance requires understanding of the complex interplay between lecturers, students, administrators, and the social and academic support students have access to (Midgley and Rajagopalan, 2020). Such a system must be designed to take into account the many different factors that influence student success, including family background, socioeconomic status, and

cultural factors, in addition to the pedagogical and academic factors that influence student success. Additionally, student intervention frameworks must be designed to adapt and evolve as the needs of the students and the research within the field changes over time (Ulrich, 2012). In other words, if new research emerges suggesting a different way to support students that promote student success, the framework, technologies, and systems involved in the interventions for students should be robust enough to factor in this context. Therefore, understanding the complexities of complex systems is critical for developing effective intervention frameworks. By taking a holistic perspective that focuses on relationships between stakeholders, researchers and practitioners can develop frameworks that are responsive, adaptive, and effective in meeting the needs of the complex education system. In the context of complex systems, intervention recommendations are important.

When implementing interventions within a complex system like education, it is crucial to have specific interventions linked to tangible outcomes, such as referring a student to a tutorial program or a healthcare worker (van den Hurk *et al.*, 2019; Shaw *et al.*, 2020). In the context of education, a variable, such as student attendance, or the final mark for a subject at the end of a semester is a representation of a complicated system within a complex system (Kearney, 2021). In other words, the overall attendance a student obtained for a subject in higher education is a representation of the hours the student put into their studies, the student engagement level of the subject matter, a culmination of prior context related to the subject matter, the overall attendance of the course, the time management, goal setting, and academic planning of the student, the access to resources, the time the student spent on traveling between higher education and home, the food security, socioeconomic status, support networks, grit, willingness to learn, and the emotional maturity of the student, among others. Simply put, a simplistic outcome variable in the context of education is the representation of a complex system because of all the factors involved in that specific outcome variable. Therefore, understanding the boundaries of implementation within a broader system is vital because implementing systemic interventions within education requires domain-specific expertise within the context of a complex system (Larrabee Sønderlund *et al.*, 2019). Education-specific surveys have shown a hyper-focus on unique use cases for educational implementation, rather than assessing general system-based interventions for education to be successful (Amissah *et al.*, 2019; York *et al.*, 2019). This poses a problem as there is currently a deficit in the

fundamental understanding of what is needed to assist a student when a problem arises within the context of education.

One problem area could be based on the epistemic vagueness associated with the fundamental purpose of education research (Platz and Platz, 2021). Some researchers argue that the purpose of education research is to identify gaps within current systems and to develop models (Harris and Patton, 2019; Granić and Marangunić, 2019). Others suggest that it should focus on pedagogy and its development, while some propose that education research should correlate, assess, and evaluate specific interventions within a narrow and specific context (Hardman, 2019). The insights from these types of studies are valuable from the context of that specific use case, but may not be applicable in another context. For example, if a study conducted in China, identified education challenges within a middle school context for first-year robotics, and recommendations were implemented to improve student outcomes, the validity of these recommendations and strategies for students residing in other countries like Egypt, Brazil, or South Africa is questionable.

Therefore, it is necessary to have a domain-specific understanding of complex education systems when implementing interventions to address specific problems. The focus should be on developing systemic interventions that take into account the relationships between stakeholders and micro-systems, rather than narrow, specific interventions that may not be applicable in other contexts. This will require a fundamental shift in the way education research is conducted and the development of domain-specific expertise to address complex education systems that implement interventions. In the sections to follow, systemic interventions, the impact and some challenges facing the higher education system, and the human-machine student intervention framework for higher education will be outlined.

2.3.1 Systemic interventions and domain specific knowledge

The advent of information and communication technologies has allowed people to connect with information and each other more easily. These technologies have also facilitated the development of information management systems that can store and retrieve specific information (González-Zamar *et al.*, 2020). Learning management systems (LMS) are tools that have been developed to manage teaching and learning content in the education sector. They can also assess students and

identify students at risk of failing, especially when analytics is involved (Bradley, 2021). Recent advancements have enabled the development of machine learning (ML), deep learning (DL), and reinforcement learning (RL) algorithms on standard platforms that make these platforms function to incorporate the user experience and tailor make this experience using these technologies. ML involves performing mathematical and algorithmic tasks on structured data to perform a specific task on a test set of data (Hart *et al.,* 2021). DL is a branch of ML that involves artificial neural networks and is particularly well suited to large amounts of data and particular structure within the data (see Chapter 3). RL, on the other hand, is a type of computational learning that focusses on autonomous decision making (Sutton and Barto, 2018) (see Chapter 5). It involves creating a virtual environment with specific rules and policies, and an agent that can take certain actions within the environment to learn and optimise rewards. Artificial intelligence (AI) systems can use one or more of these algorithms to learn from data, take actions, and optimise rewards. However, building such systems requires not only technical expertise, but also domain-specific knowledge.

As it currently stands, several education institutions have collaborative entities that enable cooperation between experts and practitioners across education (Bidandi *et al.,* 2022). However, these entities are yet to align their efforts to support the upliftment of technologies that require expertise from systems-based education researchers (Kezar, 2005). At the moment, systems-based thinking and systems-based research is in a deficit (Monat *et al.,* 2020; Majeed *et al.,* 2022). The lack of systems-based research has led to the development of technologies and processes that do not form part of a larger system, and as a result, are harder to scale. Frameworks are harder to implement because domain specific experts are not always informing the system level engineers innovating such technologies (Conboy and Carroll, 2019). As a result, system level thinking is required within the education domain. Therefore, it is imperative for education researchers to collaborate with experts in systems theory, data, and computer sciences, among others.

In addition to these factors, there are challenges associated with the data needed for developing student interventions (Combrink *et al.,* 2023). Currently, there is no centralised repository that has a consolidated consensus based on the success of student interventions. For example, just like health sciences as a faculty has a consensus and methodological techniques to ensure that a specific intervention assists a particular health outcome, so too should education interventions be decided

upon based on evidence, context, and informed by expertise. This type of domain-specific knowledge will not only become important for people building AI systems, but will, in part, be vital for the development and deployment of such systems within the South African context and elsewhere.

As such, systemic interventions require an understanding of the context, landscape and available interventions within education to be successful. To implement these within an autonomous system requires further understanding into how the system itself will behave over time, to what extent the system will adapt, and what the impact of changes in the data would mean to the effectiveness of implementing such a system. To overcome these challenges, the use of autonomous learning technologies may pose a solution because such a technology needs to learn from a specific context and apply what it had learned from the context within a particular environment (Han, 2019; Fierro-Saltos *et al.*, 2020; Kang, 2021). As such, the use cases, environments and interventions will differ from one institution to the next, but the overarching framework to learn from these contexts will be similar. To identify an overarching framework for South Africa, the basic and higher education system first needs to be contextualised. In the context of South Africa, the education system faces a variety of challenges, and in the next section, some of these will be outlined.

2.4 Basic education in South Africa

2.4.1 Outline of the basic education system

The basic education system in South Africa follows a 12-year curriculum ranging from grade 1 to 12 (Maddock and Maroun, 2018). Each year increases in complexity, as students move through the different phases associated with basic education. The three phases of basic education are the foundation phase (grades 1 – 3), intermediate phase (grades 4 – 6), and senior phase (grades 7 – 9). The later grades in high school from grade 10 onwards are known as the further education and training phase and is sometimes grouped into the senior phase in discussion. Each grade takes approximately 10 months to complete over a 12 month period (including holidays), and the school curriculum typically starts at grade 1, the year the student turns 7 years of age (Hart and Laher, 2015).

In South Africa, the Gini coefficient – a measurement of wealth inequality – is one of the highest in the world (Posel *et al.*, 2020). This indicates that there is a significant wealth divide within the society, which can also be a sign of widespread poverty. As a result, the education system too has a divide in terms of the poverty and wealth distribution within the education system itself (Francis and Webster, 2019). To address the needs of schools with varying levels of financial resources, the government uses a quintile system on a scale of 1 to 5 to classify schools based on their financial and socioeconomic needs. In South Africa, the quintile system can be used as a proxy to outline this wealth divide within the education sector and illustrate to what extent the sector is divided based on this wealth gap. Quintile 1 schools require the most support as they are typically located in areas with poor infrastructure and lack specialised educators. In contrast, Quintile 5 schools have little to no state support and rely on third-stream income to cover operational costs and salaries of educators (Ogbonnaya and Awuah, 2019). One major difference between quintiles is the level of autonomy that school management has in implementing changes to the education system. Quintile 1 – 3 schools are heavily reliant on government policy and teaching implementation plans. In contrast, Quintile 4 and 5 schools have more autonomy and can procure additional tools and technologies to support teaching and learning, as long as they meet basic policy requirements (Sayed *et al.*, 2020).

This creates a disparity in access to technology and digital devices for students in different quintiles. Students in Quintile 4 and 5 schools are more likely to use digital devices as part of their daily learning or have access to them at home, while students in Quintile 1 – 3 schools may have limited access to functional computers within the school context, and a virtually non-existent digital asset based at home apart from a smart phone (Makalima *et al.*, 2023). Although the landscape of the digital divide in South Africa is slowly changing, there are still significant challenges that need to be addressed in relation to access to educational technologies. Another major difference between quintiles is the ability of schools to incorporate new curricula. Quintile 4 and 5 schools can usually afford to pay for additional skills such as data science, robotics, and analytics, which exposes students to more diverse subject areas. In contrast, Quintile 1 – 3 schools may not have the resources to incorporate additional subjects into the curriculum (Maistry and Africa, 2020).

There is also a significant difference in the home environment and level of support for learning outside of school (Venter, 2022.). Students from Quintile 1 – 3 schools are more likely to come from families where the guardians are part of the general labour force and may be illiterate, which can limit their ability to reinforce knowledge and upskill their children at home. In contrast, students in Quintile 4 and 5 schools are more likely to come from families in urban areas with higher levels of education and may have access to additional resources outside of school (Hossain, 2021). The quintile system is an attempt to provide support to schools with varying levels of financial resources, but it still creates significant disparities in access to resources and opportunities for students. The population densities also differ between the different provinces in South Africa, of which there are nine (Deumert, 2010). Depending on the population differences between the provinces, the relative number of students within each province per quintile does differ. Approximately 87% of the schools in South Africa can be classified as Quintile 1 – 3 schools (Motala and Carel, 2019). Certain provinces, like the Eastern Cape and KwaZulu-Natal have a disproportionate number of Quintile 1 – 3 schools as compared to the number of Quintile 4 and 5 schools.

The education outcomes related to the overall pass rate, and the academic averages, the student throughput and retention rates among Quintile 1 – 3 schools, are significantly lower than Quintile 4 and 5 schools. This creates a further divide in terms of the level of education, and this lack of quality education impacts the students entering higher education (Van der Berg, 2008; Burnett, 2021). In the next section, some of the challenges entering higher education from the basic education system will be outlined.

2.4.2 Challenges in higher education in South Africa based on basic education

In the context of this study, higher education includes all tertiary education institutions ranging from small private education institutions to large universities within South Africa. Due to the challenges students face in the basic education system, by the time they enter higher education, there are shortcomings in their ability to transition to the new environment (Mgaga and Scholes, 2019). This is further amplified when comparing a student who came from a Quintile 1 school and a student who went through a Quintile 5 school environment (Pellicer and Piraino, 2019). As a result, there are several interventions that higher education institutions have implemented to

address these challenges, like sufficient orientation programmes, first-year seminars, summer and winter schools, as well as capstone subjects prior to university (Strydom and Loots, 2020).

Despite these initial strategies, student throughput and retention rates are still problematic (Olivier *et al.*, 2020). Due to this, several studies have been conducted to investigate ways to improve student success, including the promotion of student engagement, the types of academic and non-academic support that may be provided to students, as well as the potential impact of these interventions on academic success (Bond *et al.*, 2020). For example, studies have been conducted to investigate the comparative effect and impact of a shift in policy, mentoring, and peer support on student throughput and retention rates (Cranfield *et al.*, 2021; Wang and Hofkens, 2021). As promising as these studies are in illustrating a measurable effect between a student intervention and an increase in student success, the exact contribution of the intervention to student success will vary among individuals from different contexts (Brewer *et al.*, 2019). This is also further complicated when assessing the same intervention between students who are studying different subject matter, have different socio-economic statuses, and who have different levels of emotional maturity, to name only some factors (Gallagher and Savage, 2020). As a result, interventions to assist student success are institution-specific and need to be investigated based on the needs of the students and the challenges they face. To advance research in this field thus requires an understanding of the available data and the challenges that different students face.

This data needs to be freely shared and made freely available in order for researchers to be able to use sophisticated technologies that use machine learning, for example, to identify and recommend interventions to students. Unfortunately, the current data sharing policies, Protection of Personal Information Act (POPIA), and data transfer agreements between institutions, researchers and policy makers are very strict, and the rate of these investigations, the red tape surrounding such investigations, and the solutions to assist students are very difficult to implement (Malherbe, 2021). In the next section, data sharing challenges within South Africa to conduct research investigating student interventions will be outlined.

2.5 Challenges in data sharing in education

Access to data can be complicated because different countries and academic institutions have varying policies and regulations. The General Data Protection Regulation (GDPR) in Europe has faced criticism from researchers because of the confusion it creates regarding consent and data sharing (Jia *et al.*, 2021). This leads to conflicts between regulations and research ethics because the importance of understanding the context surrounding different research entities, including those that are involved with human participants' demographics, gender, lifestyle, diet, and general behaviour, is emphasised in these debates (Gefenas *et al.*, 2021). Some researchers are therefore hesitant to share data due to the possibility of violating GDPR which will lead to severe legal consequences (Combrink *et al.*, 2023).

Similarly, South Africa has implemented protection legislation, POPIA, and ethical policies to safeguard personal information (Konig, 2021). In higher education, these policies are under strict control and significant legal accountability measures have been put in place, such as the clear consequences to data owners, data custodians, and data handlers should there be a breach of any kind regarding the data (Swales *et al.*, 2022). At its core, POPIA provides protection for distinct personal information, including race, ethnic origin, political persuasions, and more. However, concerns have been raised regarding potential conflicts with existing data protections and research ethics as the fear leads to a lack of data sharing, which in turn makes researchers reluctant to engage with research involving such data. Some researchers have argued that POPIA is a positive development for data privacy as it places restrictions on the distribution and sharing of data, which affects access to and sharing of data (Adams *et al.*, 2021). This may compromise the practice of data sharing without affecting the generation of deidentified data. There also exists a lack of certainty over what type of data may and may not be shared, even in the event of informed consent. For example, under the regulations, a participant may have the right to withdraw their information from the study at any point. However, in certain instances, this becomes very difficult if, for example, images are used as training data and the models as well as the training data were made publicly available. Therefore, although the strict regulations are intended to safeguard personal information and protect members of the public from potential exploitation, the innovations and research that require training data are stifled.

The importance of the training data is not to have data that can link to a specific individual, but rather to serve as the training data to perform a specific machine learning task, such as a classification task. For this type of classification, personal information such as the name or student number, or contact information of the person, is not important (Chandler, 2020). The argument is that some of these policies and regulations impose restrictions on data access and sharing, delaying academic research, especially in areas that require large amounts of training data, such as machine learning for educational research (MLER). Other fields that require significant amounts of data to develop innovative technologies are also affected. Furthermore, the type of data publicly available cannot be used as training data as the information is either presented as meta-data and what is required for MLER is raw data, in many instances (Combrink *et al.*, 2023). Furthermore, the dataset required for MLER varies depending on the context of the research being conducted, but in most cases, anonymity of the research participants is maintained. For example, if the research is focused on predicting a student's academic performance, the required dataset typically consists of student grades, attendance, and other variables collected, without revealing their identities, while adhering to POPIA. The challenge comes in with the data collection itself as the triangulation of the data, prior to the research, requires the context and unique identities of the students. This means that there will be a process of unmasking and masking to get the training dataset ready (Kaneko and Bollegala, 2022). Unfortunately, there is no universally accepted masking and unmasking policy specific to the sharing of higher education data in South Africa (Uleanya and Ke, 2019).

Similarly, if MLER aims to automate assessment methods, the dataset required includes answers and scores, without disclosing the identities of the students (Luan and Tsai , 2021). In this instance, there will also be a need to unmask and mask the data prior to the training dataset used in the MLER, but for the MLER itself, personal information is not required. However, what this masking and unmasking entails and what the correct procedures are in handling this information is still not known. Another challenge with data sharing is the vague interpretation and implementation of these regulations on international entities that have an operational function on learning technologies, such as learning management systems.

Blended learning is an integral part of higher education, exacerbated by the COVID-19 pandemic (Rasheed *et al.*, 2020). Institutions of higher learning had to shift to a combination of synchronous and asynchronous learning (Ogbonna *et al.*, 2019). This meant that more teachers and lecturers had to incorporate the use of learning management systems and other technologies to integrate and share content to their students. The challenge is that most of the organisations that own the software have an almost unrestricted access to the data, allowing them to analyse and even sell data to improve user experience. The regulation of this is highly complex as auditing this information is very difficult. This concept is not just applicable to learning management systems but also any internet-compatible devices or software (Combrink *et al.*, 2023).

To overcome these challenges, synthetic data may pose a solution (see Chapter 3). It is therefore important in the creation and motivation of a framework that deals with systemic interventions that synthetic data, the system itself, as well as the system-to-system level interactions consider the level of abstraction required for the framework to be a representation of the complex system. It is for this reason that, to conceptualise a generalisable framework, these contexts are incorporated and that this level of abstraction is required. In the next section, the proposed framework for this study will be outlined. Each component of the framework represents a different context to be explored in the dissertation.

2.6 A framework for human-machine student interventions in higher education

Consider that an intervention recommended by a system is bound by a starting time, data collection, and a subsequent recommendation (Collins and Cockburn, 2020). In other words, if an autonomous system makes a recommendation, the former requires context in the form of data collected for a specific period of time, bound by the context of the intervention. For example, data collected within the first week of university will work well for an at-risk prediction within the first week but might not be relevant for an event a year later. Subsequently, there are several variables that may have an impact on the overall academic success of the student despite the time it is bound by, such as comprehension in the language of instruction (Xie and Curle, 2022). In different studies, it has shown that a mastery of the language of instruction is linked to comprehension, and as a result, linked to overall academic performance (Cervetti and Hiebert, 2019; Jackson *et al.*, 2020). There have also been studies associating this level of comprehension with abstract concepts

such as mathematics, further strengthening the argument that there is prior knowledge that has an impact on the overall academic performance of a student (Leung *et al.*, 2019; Stoffelsma and Spooren, 2019). Some of the interventions would thus require a specific amount of data prior to the recommendation itself being made.

Not all recommendations would require the same amount of context; as a result, there is a time step attached to a recommendation within the context of higher education data. That time step indicates which data should be collected during this time for the proposed intervention. If we draw on the theory of systemic intervention, as outlined earlier in section 2.1, the purpose of proposing an intervention within a system draws on the boundaries that delineate the system (Midgley and Rajagopalan, 2020; Midgley and Lindhult, 2021). It is for this reason that the time bound step is also associated with the parameters of the data collected within this context so that all the information is bound by a significant and intentional time, signify a specific set of variables from within the system, and that the intervention is specific enough to be recommended within the context of the entire system (Figure 2.8).

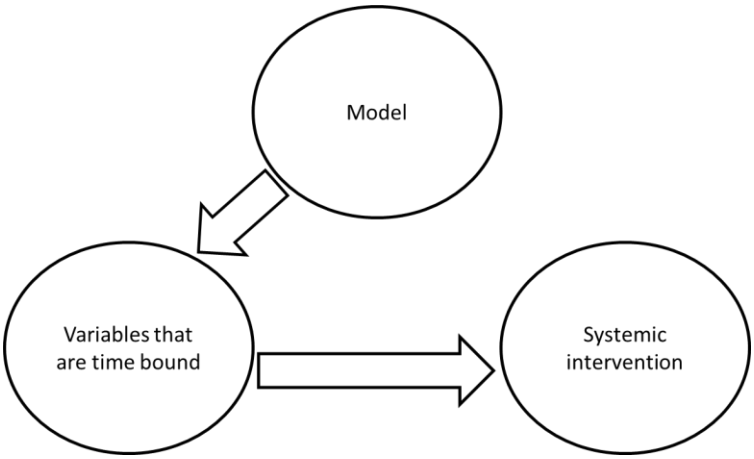


Figure 2-5 Time dependency of interventions

Given that a time dependency is associated with an intervention, interventions need to be customisable depending on the time associated with the intervention, because the impact and the value of the intervention might change over time (Foote *et al.*, 2021). For example, if a new alternative with higher impact on student success for a tutorial programme is discovered, the system would need to be adaptable enough to factor in these new parameters without having to be

reconfigured. Drawing from the context outlined in this chapter, an autonomous human-machine intervention framework requires raw data to start the process. This is justified as the raw data for every context will be different, and a “one size fits all” dataset to generating synthetic data will be detrimental as contextual information will be lost. It is therefore important that any framework intended to consider context, and learn from a specific environment, will require datasets from that environment, for that environment.

Next, to handle the vast amounts of data required to effectively implement MLER as well as mitigate the risks associated with breaching personal data, synthetic data is required to be generated. The synthetic data needs to contain a target variable, otherwise called a class variable, that considers the outcome of a particular student. This class variable will then be used as a label for such a system to start the intervention process. What is important is that the class variable association is performed on the synthetic data as well. Once a class variable is assigned to the raw and synthetic data, all models require an evaluation process to determine to what extent the classification of the class variable has been conducted. Thereafter, the accuracy of these models in terms of generating the data need to be determined.

To generate synthetic data for education is not yet common practice, and this will be explored in the next chapter. Once the initial steps have been completed to identify the time, collect the data, generate the synthetic data, and label the data with a class variable, the way in which the system learns over time needs to be explored. Once this information is complete, the correctly labelled student information may enter a system that can learn over time (Figure 2.9).

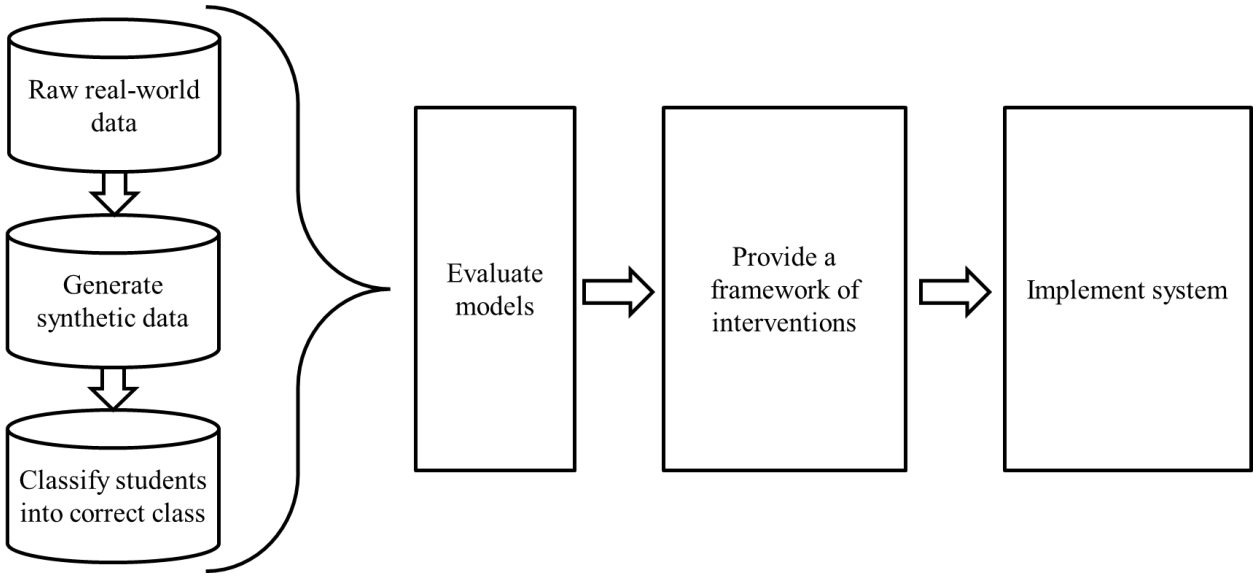


Figure 2-6 A proposed human-machine student intervention framework for higher education

A vital component of this framework is to incorporate a complex systems approach to understanding the impact of testing such a system. Firstly, it is important that such a framework is not concerned about a specific intervention, but rather, whether the system can label an intervention that works, and identify it from an intervention that does not work. Another important difference is assuming that not all interventions will impact the students in the same way. This means that the framework and the subsequent experiments to measure the framework should factor in the important contexts of the system. The last factor for such a system is to incorporate the human-machine part of the framework. It is assumed that at each part of the system from the priming to the setup of the system, as well as assessing which intervention best fits a problem, that human input will be required. This input is vital for the contextualisation of the framework across various contexts. This thesis will thus be concerned with the approach used and each of the experiments are intended to provide clarity on the particular process within the framework.

2.7 Conclusion

The purpose of this chapter was to outline the human-machine student intervention framework for higher education. This chapter also served to provide a high-level overview of some of the most pressing system-level challenges faced in this domain, including data sharing, a need for contextually relevant data, and an approach that studies education, rather than the specific interventions itself. This chapter also serves as the theoretical backbone for the rest of the thesis in

terms of the principles outlined. Although education researchers are exploring a variety of different approaches influencing student throughput and retention rates, there are still gaps associated with the types of interventions that can specifically impact student success. In addition to this, the focus of this work addresses the gaps associated with system-level approaches in higher education. In the next chapter, synthetic data generation and its evaluation will be explored.

CHAPTER 3 SYNTHETIC DATA GENERATION AND ITS EVALUATION FOR EDUCATION TABULAR DATA

3.1 Introduction

The contribution of this chapter is to generate synthetic tabular data using a Bayesian Network (BN) and a Generative Adversarial Network (GAN), and to use different evaluation metrics to test the utility of the synthetic data using various machine learning classification tasks, which has not yet been done for the education context. To address this aim, the following objectives were set out:

1. To generate synthetic education tabular data using both a GAN and two BNs; and
2. To measure utility of the original and synthetic data using machine learning classification tasks.

The work in this chapter came from the research output Combrink *et al.,* 2022a. In the sections to follow, machine learning, deep learning, probabilistic models and the different types of evaluation criteria will be explained.

3.2 Overview of synthetic tabular data

Synthetic data is information that has been computationally generated from a pre-existing dataset (Figueira and Vaz, 2022). Tabular data is a common data type consisting out of any number of samples in the form of rows, and features in the form of columns (Shwartz-Ziv and Armon, 2022). In the context of tabular structured data, there are several use cases were generating and using synthetic data can speed up the rate of digital innovation and collaboration in the education domain (Vie *et al.*, 2022). In addition, synthetically generating tabular data in the context of education can help overcome some of the challenges associated with data sharing in higher education institutions (HEIs), particularly in South Africa, where privacy laws and regulations may limit the sharing of personal information (refer to data sharing challenges discussed in Chapter 2).

Education data often contains sensitive information, such as students' names, addresses, and academic records, which must be protected by law. Synthetic data therefore provides a way to

create representative datasets that can be shared without violating privacy laws or risking the exposure of sensitive information (Ghatak and Sakurai, 2023). Furthermore, synthetic data is also useful in use cases where existing data may be hard to come by. Once generated, synthetic data can be used to conduct experiments without the need to go through extensive gatekeeping mechanisms to perform the necessary research. This allows for the development of scalable solutions that require large datasets, making it possible to conduct research on systems that can process large amounts of information. Another important factor to consider is that data must be grounded in reality, otherwise the research findings may not generalise to real data, and its application with real-world context.

The education context involves collecting and using various types of information, such as student demographics, academic performance, attendance, and course enrolment data, among others. One of the most commonly used information types in this setting is tabular data that contains some of the aforementioned information (Grundkiewicz *et al.*, 2019). Synthetic education tabular data is thus a valuable tool for all types of education researchers, including data analysts, policymakers, entrepreneurs, and machine learning and/or software engineers innovating technologies in this domain. However, as outlined in Chapter 2, there are gaps and obstacles involved in working with the data available for education research – specifically machine learning in education (MLER) – and synthetic data can help alleviate this problem. As mentioned, synthetic data is also useful when large amounts of information are required for a specific computational model to work, and more so when the data are not readily accessible or behind a wall of red tape (Cheng and Yu, 2019). In the sections to follow, the ways in which synthetic data can be generated and how the utility of synthetic data can be evaluated will be explored.

3.3 Generating synthetic data

3.3.1 Machine learning

There are several ways to generate synthetic data, such as by learning the parameters of a generative model, and then sampling once these have been learned, as in supervised learning (Dahmen and Cook, 2019). Supervised learning algorithms are systems where the input and output data are provided for a specific context (van Engelen and Hoos, 2020). This means that the information is already in a computer-readable format and labelled within the given context. In

supervised learning, the labelling especially plays an important role in stratifying the information into what is referred to as a training and testing set. The training set is the information from which the model learns the features of the dataset in order to apply them to the test set in a very specific way. Supervised learning makes use of the idea that the training data are complete, and the context of the training data is applied to a test dataset. Simply put, if researchers want to use a supervised learning approach to solve a problem, then they will need to have data on the problem and data on the use case they want to apply the supervised learning problem to (Muhammadb and Yan, 2015).

One of the subsets of machine learning (ML) is supervised learning classification. ML classification is a supervised learning task where the model aims to label a new datapoint within a dataset based on prior context gained from the training data (Kotsiantis *et al.*, 2006). For instance, a ML classification task can be used to classify a set of variables with an unknown class variable, based on context gained from complete data from a different dataset. Examples of ML classification tasks include classifying whether a student will pass or fail, based on variables within the dataset. Additionally, because supervised learning is a form of ML, it can also be used to derive context if there is potential conditional interdependence leading to utility of the synthetic data (Samitas *et al.*, 2020).

Conditional interdependence refers to the relatability between variables within a dataset and how strong the associations between data in a dataset are (Cruz and Wishart, 2006; Montaño-Gutierrez *et al.*, 2017). By using this context, ML can be used to determine a form of utility in the underlying variables by evaluating the performance of the model on the synthetic data (Smith *et al.*, 2020; Watzel *et al.*, 2020; Zhang *et al.*, 2020; Topuz *et al.*, 2023; Buggineni *et al.*, 2024). If the performance from the classification was high, a high degree of interdependence was present between the variables as the model learned from the training data. There are thus a variety of problems that can be solved with this approach, but in this context, classification is important (Birhane, 2021; Li *et al.*, 2022). Table 3.1 illustrates some algorithms commonly used in classification problems.

Table 3-1 Commonly used classification algorithms and their functions

Name of Algorithm	Example
Logistic Regression (LR)	Using logistic regression to detect students at risk of failing.
Decision Trees (DT)	Applying decision trees to detect student dropout profiles.
k-Nearest Neighbours (kNN)	Using k-nearest neighbours to classify student cohorts.

Classification is the ability to label a subset of data points within a dataset into a class or group. There are several types of ML algorithms used for classification tasks, each with their own limitations and underlying processes. The correct algorithm needs to fit the correct problem, and, at times (with each dataset), it may be challenging to define the details as to which algorithm fits the solution to the problem in the best possible way (Lee and Shin, 2020). Different types of ML algorithms commonly used in classification tasks will be outlined next.

3.3.1.1 Logistic Regression

Logistic Regression (LR) is a statistical method widely employed in machine learning to predict the probability of a categorical dependent variable, based on one or more independent variables (Alsariera *et al*., 2022). The dependent variable in LR is categorical, meaning it can take on a limited number of discrete values. These values often represent different categories or classes, and their nature may vary across datasets and use cases. The fundamental output of LR is the probability that a given set of variables belongs to a particular class, typically denoted as 0 or 1 (Dreiseitl and Ohno-Machado, 2002). This characteristic makes LR particularly suitable for binary classification tasks. For instance, in an educational setting, LR could be utilized to predict whether a student is likely to pass or fail based on various factors such as attendance, grades, and study hours. LR accomplishes this classification task by computing probabilities using a sigmoid function (Alloghani *et al*., 2020). The sigmoid function, also known as the logistic function, maps any real-valued number into a range between 0 and 1, making it ideal for interpreting the model's output as a probability. The mathematical representation of the sigmoid function is as follows (equation 3.1):

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{3.1}$$

where $f(x)$ is the sigmoid function used to determine the probabilities within the function. The exponential term e is the base of the natural logarithm, and the negative sign ensures that the function maps high positive inputs to values close to 1 and high negative inputs to values close to 0. This property of the sigmoid function is crucial for the binary classification capability of LR. The sigmoid function has a distinct sigmoid curve, often called the S-curve (Assuah *et al.*, 2022). The curve of the function is a visual representation of how the function maps a real number between 0 and 1. To this extent, the sigmoid function is non-linear (Cawley *et al.*, 2006). This property allows for the mapping of more complex functions, beyond linear associations. (Figure 3.1).

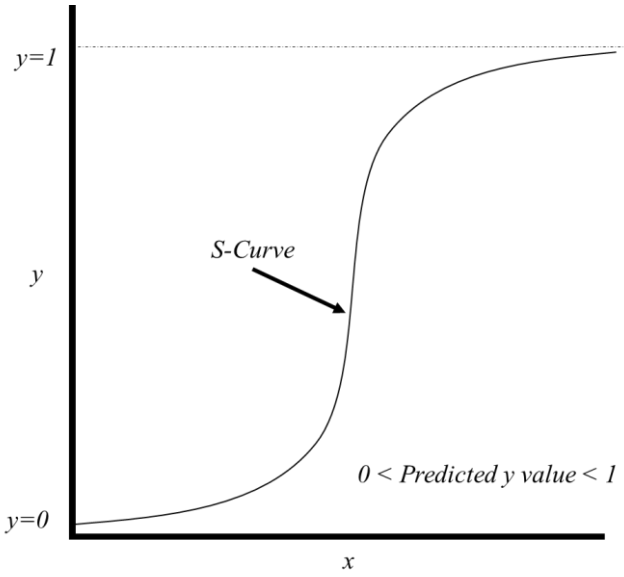


Figure 3-1 Illustration of logistic regression S-curve

As mentioned earlier, the sigmoid function transforms numerical values (in this case, the independent variables) into an expression of probability as a value between 0 and 1 (Chi *et al.*, 2021). This transformation is crucial in determining the class of a given value. To distinguish if a certain value belongs to 0 or 1, a certain threshold is introduced (Date *et al.*, 2021). In the context of LR, this threshold divides all data into two groups, labelled as 0 or 1. For instance, if the threshold

is set at 0.6, values below 0.6 are assigned to class "0", while those above are assigned to class "1" (Jadhav *et al.,* 2020). For example, if the threshold value is 0.6, then everything below 0.6 would be ascribed to "0", and everything above 0.6 would be assigned to the "1" class. The sigmoid function thus maps a real value to a value between 0 and 1 based on the features within the data (Heinze-Deml and Meinshausen, 2021). Then, a decision is made as to which class will be assigned. To this extent, the LR algorithm predicts the probability of data points belonging to a particular class variable, based on the feature variables converted to a probability (Lee and Shin, 2020). Another important concept in LR is weights and bias which are learned from training data. This is done through a cost function, which is a measurement of how close the models predictions are to the actual outputs. For an LR, a logarithmic function is used to represent the cost (Cawley *et al.,* 2006). By using a logarithmic function allows the algorithm to make use of gradient decent, an optimisation algorithm that finds the optimal output. Once the cost function is known, the weights and bias are factored into the gradient decent to minimise the cost and maximise the best parameters for making the prediction. The term weight in the context of a LR is represented by a real number. Each weight is associated with a specific input feature in the dataset, influencing the impact of that feature on the predictive model's decisions. In the context of tabular data, the features would be the columns representing different variables within the dataset. Weight in this instance thus defines the importance of a particular variable to the classification decision (Muhammad and Yan, 2015). On the other hand, bias, also referred to as the intercept, is also a real number which is added to the weighted inputs. The trade-off between weights and bias in LR establishes the probability of a specific input's association with a particular class variable. During training, LR weights assess the impact of individual features, and bias informs the adjustment to the different predictions.

The concept that influences the model coefficients of an LR is the maximum likelihood estimation by a generalised method least square. Maximum likelihood is calculated using multiple estimations of all regression coefficients (Yağcı, 2022). This maximum likelihood estimation effectively transforms the LR from a logistic model to a linear model. An Odds Ratio (OR) is introduced into the model, which quantifies the likelihood of an event occurring in one class compared to another, making it suitable for binary classification tasks (Dreiseitl and Ohno-Machado, 2002). The OR represents the number of times the probability of a specific event will occur given an increase or decrease in the unit change of an independent variable(s). In this context, if the OR is above a certain

value, it is assumed that the probability of a change in the dependent variable due to a change in the independent variable is more likely (Zeineddine *et al*., 2021 Conversely, if the OR is below a certain value, it represents the probability of a decreasing change due to an increase in the independent variable. A confidence interval (CI) is calculated for each predictor used, creating a range with an α probability. The linear combination can be defined as (equation 3.2):

$$z = X * w + b \ , \tag{3.2}$$

where X is the input features corresponding to a datapoint, w the vector corresponding to class labels, b the bias term as a scalar value, and z the linear combination of features and weights. Then, the cost function is to be determined. There are several types of cost functions, but for LR, different kinds of logarithmic functions may be used. Then, a gradient decent optimisation algorithm is used to determine the gradient of the cost function with the current weights and biases. Thereafter, the weights and bias are updated. The two hyperparameters that are set initially in LR is thus the α and n. A loop is introduced, until n where the linear combination and sigmoid function are calculated. The α is used to determine the gradient of the cost function, in the context of the weight and bias, respectively. The overarching algorithm of a LR is summarised below (algorithm 1).

Algorithm 1: Logistic Regression

1. **Initialize** w and b
2. **Set** α and n
3. **For** i in 1 **to** n
4. **For** each training example (x, y):
5. Calculate linear combination: $z = X * w + b$
6. Calculate sigmoid function: $f(x) = \frac{1}{1+e^{-x}}$
7. Determine cost function using the maximum likelihood estimation
8. Calculate gradients of the cost function (α)
9. Update weights: $w \leftarrow w - \alpha$
10. Update bias: $b \leftarrow b - \alpha$
11. **End for**
12. **End for**
13. Calculate OR

The LR algorithm is iteratively executed until a predetermined number of instances have been completed. Upon the completion of the training phase, the model, now equipped with learned parameters, is capable of classifying new data points based on the features of a class variable. To illustrate the application of LR in an educational context, consider a binary classification task. This task uses features, which in the case of tabular data, are represented by the columns in the dataset. The classification task aims to categorise a student into one of two categories (pass or fail), based on these features. For this specific task, let's assume that the classification is determined by whether a student passes or fails and that this is the outcome variable for the classification task. The model is then provided with training data, which in this instance, is a tabular dataset with known outcomes corresponding to the features. This methodology allows for the prediction of a student's success or failure based on various factors such as attendance, assignment grades, and participation in class discussions. The model can then be used to inform interventions or support strategies to help students who are predicted to struggle. A practical example of the use of LR in education was conducted by Assuah *et al*., (2022). Next, k-Nearest Neighbours as a classifier will be explained.

3.3.1.2 k-Nearest Neighbours

k-Nearest Neighbours (kNN) as an algorithm can be used for classification tasks (Cunningham and Delany, 2021). Similarly, to LR, kNN as a supervised learning algorithm is feature based, which means it relies on features extracted from the input data to perform the classification task. kNN, however, differs from LR in some of the following significant ways. Firstly, kNN measures similarities between features in the underlying data to make predictions on which class a set of features belongs to, whereas LR makes use of weighted sums to calculate the probability of features belonging to a specific class. Secondly, the decision boundaries for kNN are non-linear and are based on the distribution of datapoints, whereas LR's decision boundary is based on a linear function of the features (Zeineddine *et al*., 2021). Finally, another major difference between kNN and LR are that the hyperparameters differ where LR is concerned over α and N and kNN uses the number of neighbours (k) and a distance metric (measured in Euclidian distance in a feature space). Due to the differences between kNN and LR, an assumption cannot be made as to which algorithm would work best for a particular dataset without testing. For example, while kNN might be more suitable for datasets where the information is non-linear and the relationships

between the underlying data might be more complex, LR is more suitable for datasets with a lot of noise and outliers (Cunningham and Delany, 2021). With kNN, features are presented in a feature space, and the distances between the spaces are mapped using Euclidean distance in the feature space. The expression of these distances can be calculated based on the relative proximity of the datapoints and the labelled class variables (algorithm 2).

Algorithm 2: k-Nearest Neighbour

1. **Initialize** dataset, query point, *k*
2. **Set** *X*, *Y*, *x*
3. Calculate the distance between the query point and each data point in the dataset:
4. **For** *i* = 1 **to** *m* do
5. Compute Euclidean distance $d(y, x)$
6. **End for**
7. **Sort** the Euclidean distances in ascending order to find the nearest neighbours
8. **Select** the *k* data points with the smallest distances to the query point
9. Determine the most frequent class label among the kNN
10. **Return** the majority class label as the predicted class for the query point

In algorithm 2, a set of labelled data points are presented in a feature space, each data point representing different features and a specific class label. The set variables for kNN can be defined as *X*, *Y*, *x* where *X* represents the training data, *Y* represents the labels given to the data, and *x* represents the unknown sample. The query point is the specific data point for which a class label should be determined. The value of *k* represents the number of neighbours to consider, given the query point. Distances in the algorithm refers to the Euclidian distance in the feature space, measured in units between the query point and all the data points within the given dataset in the function. The nearest neighbours are the Euclidean distances between the data points within the dataset, within a list, sorted in ascending order. The kNN are the closest *k* data points to the query point. Then finally, the class label is the label that appears most frequently within the number of *k* to the query point. kNN provides a simplified solution to classifications but may also be computationally expensive when large datasets are used (Figure 3.2).

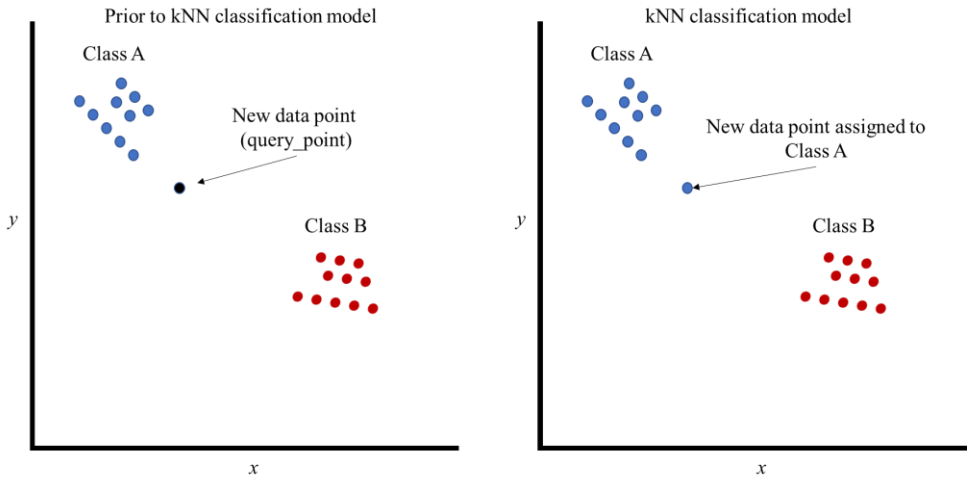


Figure 3-2 Illustration of a kNN classification based on Euclidian distance

In the context of ML classifications, there are several other types of classifiers and classification algorithms, of which a decision tree in the context of a classification task will be discussed next.

3.3.1.3 Decision Tree

A Decision Tree (DT) is a non-parametric supervised learning method that can be used for classification tasks (Kadiyala and Kumar, 2018). A DT performs by partitioning a dataset into smaller subsets of data, based on features. These subsets are then further stratified into smaller subunits based on features. Each split is known as a node, branch, or leaf. A node is the point by which a dataset is divided into its subset of data. The function of the DT is to choose the best split in the data at each node (Magee and de Weck, 2004). When a branch is split at a node, the resulting branches result in leaf nodes which represent the various types of predicted labels (Figure 3.3). As illustrated in Figure 3.3, the different colours represent different class labels, and the decision to split the tree is based on features within the data.

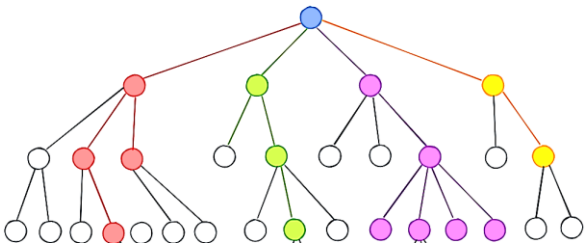


Figure 3-3 Illustration of a Decision Tree algorithm subdividing datasets.

There are similarities between a DT and LR in that both are supervised learning algorithms, and both can provide insight into feature importance. In the case of LR, larger weights to features can indicate stronger feature influence, and with DT, features appearing near the top of the tree are used for early splits, and on average, are more important features in the dataset. Despite these similarities, there are significant differences between the two. Firstly, the model structure between a DT and LR is different. A DT makes decisions based on the values of features, whereas LR uses a linear combination of features which is converted to a logistic function producing probabilities on which the LR performs classifications. Secondly, similarly to kNN, DT is sensitive to outliers which may create bias splits in the trees (Bhattacharya *et al.*,2017). Ultimately, a DT algorithm functions by dividing the data into smaller and smaller subsets until a specific stop criterion is met (algorithm 3).

Algorithm 3: Decision Trees

1. **Initialize** dataset, features for class *k,* target features
2. **Set** *X, Y, x*
3. **Select** the best feature to stratify the dataset into smaller subsets of data based on the features associated with the class variable
4. Create a decision node for the selected feature by finding the discrete function $f(A)$
5. **If** the splitting metric > threshold, **then** label *Y* **as** $f(A)$
6. **For** each outcome of $f(A)$:
7. Create subtree = subset of *X, Y, x*
8. Connect the node for to a subtree
9. **End for**
10. **End if**
11. Create the subsets of data based on the feature's values
12. **For** each subset:
13. Attach the resulting subtree to the decision node created in step 6
14. **End for**
15. **Return** the decision node

In algorithm 3, the dataset represents the training data, where each point contains information on the features and a specific class variable. The features in this instance are used to determine the subsets of the data. Thereafter, algorithm 3 makes use of the feature context to determine the target feature and decide based on the partitioning of the data into its smaller subsets (splits). A stop criterion is implemented so that the loss at each node is minimised. Each node of data is split into two, and the decision to choose a specific child, or sub-node of information, is based on features

in the data. In the instance of the algorithm, *X* represents the training data, *Y* represents the labels given to the data, and *x* represents the unknown sample. The decision to use these three algorithms was based on literature searches and it was concluded that these are the most commonly used for classification tasks of tabular data (Hagenauer and Helbich, 2017). It must be noted that identifying the specific algorithm to best fit a problem is an important task, but it was not the focus of this study. In ML classification tasks, the goal is to predict the class of an observation based on its features or variables (Tixier *et al.*, 2016). To this extent, each of the aforementioned ML algorithms requires a class variable to learn from and apply the learned context to. However, to generate synthetic data, deep learning approaches have been favoured over ML approaches, which will be further explored in section 3.3.2. In the next section, the concept of a generative adversarial network in the context of deep learning will be briefly outlined.

3.3.2 Deep learning and Generative Adversarial Networks (GANs)

Deep learning is a subset of ML that makes use of artificial neural networks to process large and complex data sets (LeCun *et al.*, 2015). Neural networks are designed to function similarly to an abstraction of the neurons in a functional human brain, by using an intricate network of nodes, linked together between different layers and connections to transform and analyse data (Dong *et al.*, 2021). These aforementioned layers are hidden and consist out of a series of neural networks linked together to define the parameters. As data processing takes place, the information is transformed from one hidden layer to the next and the data are transformed. The final results is an output layer that provides the context of the process. As beforementioned, deep learning models typically have an input and output layer, with several hidden layers in-between the input and output layer (Ganaie *et al.*, 2022). As information is passed through a deep learning algorithm, if the network was designed correctly and the problem is solvable, the information is passed between neurons and processed at each one, so that the hidden layers to provide information to an output layer (Figure 3.4).

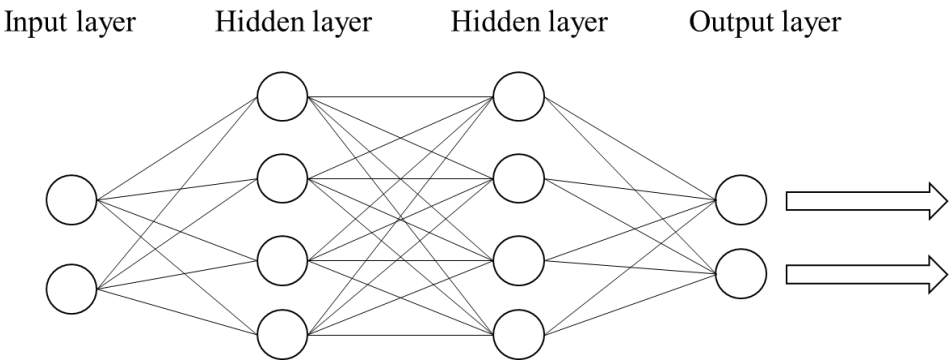


Figure 3-4 Example of a neural network.

By analysing the patterns and relationships within the data, deep learning algorithms can learn to accurately classify class variables within a dataset. Deep learning has been used in generating synthetic data because of its accuracy (Nikolenko, 2021). One of the most popular supervised deep learning models to generate synthetic data is the use of Generative Adversarial Networks (GANs) (Xu, 2020). GANs have been used to generate synthetic data, especially in creating large training datasets, and have been shown to work either very well, or not at all, at generating useful synthetic data for a specific task, such as ML classification using labelled training data (Engelmann and Lessmann, 2021). To this extent, the current use of GANs to generate synthetic tabular data for the education context is unexplored.

GANs are a type of deep learning model that consists of two neural networks: a generator network and a discriminator network. Figure 3.5 displays the general architecture of a GAN. The process begins by generating random latent variables. These latent variables are not directly associated with any specific feature in the real data, instead, they represent different patterns that the generator can learn to map different datapoints. The generator then produces multiple instances of these variables based on their observed ranges. The discriminator receives both the simulated and original data, and its role is to distinguish between the two. The simulated and original data are passed through a condition function to evaluate the model's ability to differentiate between them. The condition function represents a neural network which takes both the simulated and real data into account to evaluate the model's capability to distinguish between the two, which assists in improving the model's ability to distinguish between the synthetic and real data. If the discriminator classifies the data as synthetic, the synthetic data generation component of the model

is fine-tuned, and new synthetic data is produced and fed back into the discriminator. This iterative process continues until the discriminator cannot distinguish between the real and synthetic data produced by the model (Creswell *et al.*, 2018).

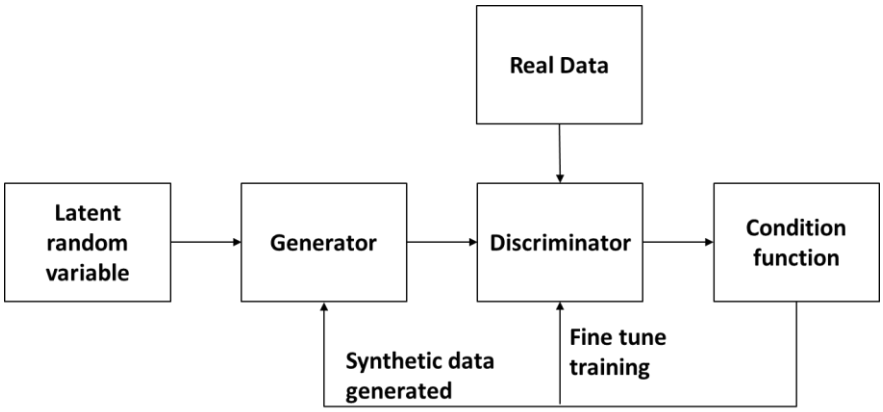


Figure 3-5 General structure of a Generative Adversarial Network (GAN).

One commonly used method to train GANs is to minimise the Jensen-Shannon divergence, which is a statistical measure of the difference between two probability distributions (Lou *et al.*, 2022). Specifically, in the context of GANs, the two distributions that are compared are the distribution of the original data and the distribution of synthetic data generated by the generator network. By minimising the Jensen-Shannon divergence between these two distributions, the goal is to make the synthetic data distribution as similar as possible to the original data distribution. This means that a representative sample of the original dataset needs to be used to generate a synthetic representation of the information. This also means that the similarity of the synthetic data will be based on the distance between the probability distributions between the original and synthetic data, indicating that their central tendencies will be similar. It also implies that the function of the discriminator network is to establish a Nash equilibrium between the two networks (Baasch *et al.*, 2021).

If we consider a GAN, the Nash equilibrium refers to a state where neither the generator network nor the discriminator network can enhance its performance because neither can unilaterally improve on the performance of the model (Bynagari, 2019). Simply put, this state is reached when improvements of the synthetic data are no longer observed compared to the performance of the

original data. Ultimately, the objective of a GAN is to achieve this equilibrium, in which the generator produces synthetic data that is indistinguishable from the original data in terms of its probability distributions. If we consider that a generator network is (G) and a discriminator network is (D), we can write the notation as (equation 3.3):

$$min_G max_D(V)_{(D,G)} = E_x{\sim}p_{data}(x)\left[\log_{(D(x))}\right] + E_z{\sim}p_z(z)\left[\log_{\left(1-D(G(z))\right)}\right], \qquad (3.3)$$

where G represents the output of the generator network, given input noise z, D represents the output of the discriminator network, given the original dataset x, p_{data} is the distribution of original data, and p_z is the distribution of input noise z. In other words, in the context of a GAN the expression $min_G max_D(V)_{(D,G)}$ signifies the adversarial training between two neural networks, the G and D. The D's objective is to perfect its skill in distinguishing real data from synthetic data (Goodfellow *et al.*,2020). On the other hand, the G's goal is to trick the discriminator by creating data that closely resembles real data. This competitive process continues in iterations, with both the generator and the discriminator improving over time. The generator becomes better at creating realistic data, and the discriminator improves at identifying real and generated data. The process ends when the generator produces data that is indistinguishable from real data, and the discriminator classifies all data as real with a certain probability (indicating it can no longer differentiate between real and generated data). The p_{data} is expressed in a loss function and is the expected value of the logarithm representing the discriminator network output from the original dataset to categorise the original data as non-synthetic (Aggarawal *et al.*, 2021). The p_z represents the loss function of the synthetic data created by the generator network which promotes the production of synthetic data that is indistinguishable from the original dataset. The $minG$ and $max_D(V)$ function to minimise the weights of the loss functions for the generator network while the discriminator network updates its weights to maximise the weights for the same function. Both networks are trained simultaneously, and the function stops executing once the equilibrium is reached (Hong *et al.*,2019). Once the GAN is trained and the equilibrium is reached, the GAN can be used to generate large quantities of synthetic data based on the original dataset (Sarwat *et al.*, 2022). One of the advantages of a GAN is that it has been associated as a method to identify relationships between variables without statistically modelling them. In other words, the

relationships needed to generate complex data can be learned directly from the data itself. Furthermore, it has been shown that GANs can produce data that have an indistinguishable central tendency between the synthetic data and the original data. In the next section, balanced and unbalanced data will be outlined.

3.3.2.1 Balanced and unbalanced data

In a classification task, balanced data refers to a dataset in which the number of instances for each class is roughly equal or proportional (Cano *et al.*, 2016). In other words, each class in the dataset has an equal or similar representation between the different classes. For example, if there is a tabular education dataset that contains 100 students, with two different classes, pass and fail, with their accompanying student performance data and half of the sample are students that pass, and the other half fail with a precise 50% split between pass (n_1 = 50) and fail (n_2 = 50), then one may consider the dataset balanced. If the dataset has an uneven distribution between the class variables, in this case, between pass (n_1 = 90) and fail (n_2 = 10), then the dataset is unbalanced. Balanced datasets are important considerations for certain classification algorithms, as well as for datasets of a certain size. For instance, considering the previous example, an unbalanced distribution between pass (n_1 = 9950) and fail (n_2 = 50) might lead to a much poorer performing model than a balanced dataset. This is because in an unbalanced dataset, such as with the previous example, if the model just predict pass, then the model will be correct 99.5% of the time, without having to assess the features present within the dataset.

Balanced data is thus important in classification tasks because it ensures that the performance of the classifier is not biased towards any particular class (Jadhav *et al.*, 2022). If one class is significantly overrepresented in the dataset, the classifier may become biased towards that class and produce inaccurate results. In terms of tabular education data, this might pose a problem as the dataset received is unbalanced and there might be instances where the information is heavily skewed toward one class in the dataset. Several approaches have been applied to deal with the balanced and unbalanced data problem, but the most common approaches include over- and under sampling and generating new instances of a specific class (Gottschall *et al.*, 2009; Al-Masni *et al.*, 2020).

In the over- and under sampling technique, the minority class should be over sampled, and the majority class under sampled (Vuttipittayamongkol and Elyan, 2020). There are instances where the generation of synthetic data might not be easy, especially in the absence of certain information. For example, assuming one has a tabular education dataset with 100 students and two classes, but the entire dataset contains students that passed (n_1 = 100). In this example, there are no students failing, but the possibility exists. If this dataset was used to train a classification task to differentiate between pass and fail, then the data would be biased toward pass because the information is unbalanced and contains no reference to the class fail (n_2 = 0). Although balanced data has several advantages over unbalanced data in classification tasks, there are disadvantages associated with balanced data.

Collecting a balanced dataset can be more challenging since it requires equal representation from each class. As mentioned before, even though over- and under sampling or generating synthetic data can solve this problem, this can lead to additional costs and efforts to obtain a representative sample from each class. Another challenge is that the performance of a balanced classifier may degrade if the class distribution in the population being studied differs significantly from the dataset. This can result in a model that is less accurate when applied to new data outside of the dataset's population. Therefore, while balanced data can lead to more accurate and fairer results, it also comes with some challenges that need to be addressed to ensure its effectiveness in classification tasks (Heinze-Deml and Meinshausen, 2021).

Unbalanced data can still lead to accurate predictions if the classifier is trained properly, with the right parameters and datasets (Kerstens and van de Woestyne, 2014). Even in certain use cases where the minority class may have fewer instances, it can still contain valuable information that can be used to develop a robust model. However, it is important to note that these advantages of unbalanced data may be limited by the extent of the class imbalance and the training strategies used to develop the model. While it may be easier to collect an unbalanced dataset since one class may be more prevalent, it can lead to less accurate predictions due to bias towards the majority class (Viloria *et al.*, 2020). In a particular sense, the training efficiency of an unbalanced classifier may also be impacted.

Training efficiency in this instance refers to the ability of a classification model to effectively learn and generalise from a given dataset (Chi *et al.*, 2021; Date *et al.*, 2021). The effect of training efficiency is seen when the training data from one dataset can be used on another dataset from the same or similar context, either within a given population or between different populations. In most cases, specific training data are needed for each specific population, but in some instances, the same training data can be used between different populations.

A classification model with good training efficiency should be able to learn from the data without requiring an excessive amount of data (Na *et al.*, 2022). Efficient training also helps to reduce the risk of overfitting, which occurs when a model becomes too complex and starts to fit to the noise in the training data rather than the underlying patterns (Zhang *et al.*, 2021). It is therefore important to consider training efficiency, how balanced the data are, and the type of algorithm that is used to train models on large datasets, especially for the education context. Given this context, however, there are other forms of generating accurate synthetic data. One of these is a form of probabilistic modelling. In the next section, Bayesian networks (BNs) will be explored as a form of probabilistic modelling to generate synthetic data.

3.3.3 Bayesian networks (BN)

A BN is a type of probabilistic model that can be used to represent the relationships between variables and the probabilities of the variable distribution within the context of a probabilistic network, different from the types of networks previously discussed (Derks and de Waal, 2020). These networks can be graphically represented and computationally modelled from either data or experts around a certain topic and are highly effective at understanding relationships between variables (Joubert and de Waal, 2020). A BN works with discrete variables where the probabilistic distribution is dependent on the conditional probabilities of a given category within a set of variables (de Waal *et al.*, 2016). The general probabilistic structure implies that an independent probability is denoted by $P(x)$ and a conditional probability denoted as the $P(x|y_{i\ldots\ldots}y_n)$. Therefore, x is a function of independent probability, and y a function of conditional probability. As such, probabilities (conditional and independent) can be denoted by the following equation (equation 3.4):

$$P(x_1, x_2 \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid y(x_i)). \quad (3.4)$$

In the equation, $P(x_i)$ has a set of parent variable denoted by $y(x_i)$, which allows the annotation to define a joint probability given by $P(x_i \mid y(x_i))$. Each term in $P(x_i \mid y(x_i))$ represents the probability of variable (x_i) given its parent variables, and the product is over all variables in the network. For the purpose of explaining a BN, let us consider an example that uses the following variables such as Exam Level, IQ Level, Marks, Aptitude Score, and their probability of leading to final Admission. This is not a real example, but rather a conceptual example to explain how a BN can work in the context of education. In this BN example, Exam Level's probability is denoted as $P(el)$, IQ Level's probability is denoted as $P(iql)$, Aptitude Score's conditional probability given IQ Level is denoted as $P(as|iql)$, Marks' conditional probability given IQ Level and Exam Level is denoted as $P(xm|el, iql)$, and Admission's conditional probability given Marks is denoted as $P(a|m)$. By using these probabilities, we can create a probabilistic model to calculate a candidate's admission score (equation 3.5):

$$P(el, iql, as, m, a) = P(el) * P(iql) * P(as|iql) * P(m|el, iql) * P(a|m). \quad (3.5)$$

Once the probabilistic distributions are known, these networks can be represented using a direct acyclic graph (DAG), for example, compiling all of the aforementioned variables into a DAG (Figure 3.6).

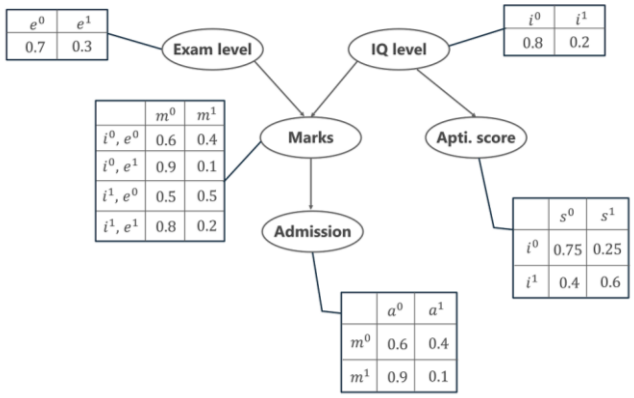


Figure 3-6 Example of a DAG and probabilistic structure for education[1].

In the aforementioned example, it is important to note that the example was just used to illustrate how a BN is structured, and not that the variables, such as IQ and exam level are binary. In reality,.a BN is much more complex and more discrete variables are represented within such a network. Both the structure (DAG) and probability distribution of a BN can be learned from either data or from an expert (de Waal *et al.*, 2016). There are two things that need to be learned in the context of a BN, namely the probability distributions and the structure of the DAG. One way to learn the structure of a BN is through a constraint-based approach (Pavlin *et al.*, 2019). The constraint-based approach in a BN is supported by the conditional independence theorem, which states that if two variables are conditionally independent given a set of variables, then there exists no edge between them in the network (Natori *et al.*, 2015). To identify these relationships between variables, statistical tests such as the chi-square test and G-test are used to determine whether the observed data deviates significantly from the expected values under the null hypothesis of conditional independence (McHugh, 2013; Berrett and Samworth, 2021).

In the context of the chi-square test, the null hypothesis for the chi-square test is that the observed data follows an expected distribution, while the alternative hypothesis is that the observed data

[1] https://uol.de/en/lcs/probabilistic-programming/webchurch-andopenbugs/example-5-bayesian-network-student-model [last accessed 15 October 2022]

differs significantly from the expected distribution. The test statistic for the chi-square test can be noted as (equation 3.6):

$$X_i^2 = \sum \frac{(O_i - E_i)^2}{E_i}, \tag{3.6}$$

where X_i^2 is the chi-square test statistic, i is the number of categories, and O_i is the number of observations within the different categories. Furthermore, E_i is the expected frequency in category. As a result, the expected frequency can be noted as (equation 3.7):

$$E_i = \frac{\left(\sum_{j=1}^{n} X_{ij}\right)\left(\sum_{j=1}^{n} X_i\right)}{n}, \tag{3.7}$$

where n is the total number of observations, X_{ij} is the observed values for i in observation j and X_i is the total number of observations in category i. In other words, E_i represents the expected value for a specific row. The terms i in observation j refer to the row and column indices of the dataset, respectively. To this effect, a comparison between the observed frequencies with the expected frequencies are conducted to determine whether there is a significant difference between them. The test statistic X_c^2 follows a chi-square distribution, in the number of columns in the contingency table. In other words, the test statistic, which measures the difference between observed and expected frequencies, follows a chi-square distribution. The specific chi-square distribution it follows depends on the degrees of freedom, which are determined by the number of rows and columns in the contingency table (Das *et al.*,2022). This is a fundamental concept in understanding the chi-square test of independence or homogeneity. To reject the null hypothesis at a significance level a, a comparison with the calculated value of X_c^2 is performed with a critical value from the chi-square distribution table. If the calculated value of X_c^2 is greater than the critical value, the null hypothesis is rejected, and it can be concluded that there is evidence of a significant association between the variables.

On the other hand, the G-test, also known as the log-likelihood ratio test, is used to compare the observed frequencies of two or more groups with the expected frequencies (Tsamardinos and Borboudakis, 2010). The test statistic for the G-test is given by (equation 3.8):

$$G = 2\sum_{i=1}^{k} O_i ln\frac{O_i}{E_i}, \tag{3.8}$$

where O_i is the observed frequency for the data, E_i is the expected frequency under the null hypothesis, and k is the number of data groups being compared. To perform the G-test, calculating the expected frequencies for each group based on the null hypothesis is performed. Thereafter, the test statistic is performed. Finally, the test statistic and the critical value from the chi-square distribution are compared and the null hypothesis is rejected if the test statistic exceeds the critical value. If the null hypothesis is rejected, it is an indication that there is a presence of a dependent relationship between the variables and no edge is added between them in the BN. There are several other types of constraint-based approaches used in the construction of a BN, but in this study, the focus was not to optimise the approach to construct a BN or DAG (Van Beek and Hoffmann, 2015). To this extent, another approach to model the structure of a BN is to use a statistical test to determine which variables are independent of each other given other variables, known as a score-based method (Kersting and De Raedt, 2001).

A score-based method assigns a score to each possible network structure based on how well it fits the data (Anderson, 2019). The score is calculated by using a measurement such as the Bayesian Information Criterion (BIC) or Akaike Information Criterion (AIC) which balances model complexity with goodness of fit (Edwards *et al.*, 2010). BIC can be defined as (equation 3.9):

$$BIC = -2\log(L) + k(\log(n)), \tag{3.9}$$

where L is the likelihood of the model given the data, k is the number of parameters in the model, and n is the sample size. This equation demonstrates that models fitted using the BIC can be represented by the likelihood ratio. In other words, it quantifies the probability that a specific model could generate the observed data, given the constraints of its outputs and the potential

scenarios derived from the data. (Bae *et al.*,2016). The advantage of this approach is that it penalises complex models to prevent overfitting, which is common if the discrete variables are oversimplified (Beretta *et al.*, 2018). On the other hand, AIC can be used to balance model complexity to a best fit, and can be written as (equation 3.10):

$$AIC = -2\,(\log(L)) + 2k\,, \tag{3.10}$$

AIC and BIC can both be used for balancing complexity to models (Alguilera-Rueda *et al.*, 2020). To interpret the results, a small value of AIC indicates that the model optimally fit the data which means that the results can be scored based on the fit to the data. The highest score of the various results indicates the best fit for the BN based on the relationship between child and parent nodes. Ultimately, this process is repeated for each node and between different node pairs in the BN to determine the best fit based on these scores. This process is repeated until the structure of the BN is complete (Lu *et al.*, 2021).

Another method used to determine the structure of a BN, is what is known as a hybrid method (Zheng *et al.*, 2023). The hybrid method uses elements of the constraint-based method and the score-based method to determine the structure of the BN. Both are used to determine the conditional interdependence of the relationships between variables in the data and using this context to provide a basis for the structure of the BN (de Waal and Yoo, 2018). Due to these approaches, it is possible to derive a BN structure from data. Once the structure of the BN is known, it is then possible to calculate the probability distributions of the nodes using conditional probability tables (Kaikkonen *et al.*, 2021).

Conditional probability tables of a particular node within a BN provides the conditional probability of that node, given its parent nodes. One method to determine this is by means of the maximum likelihood estimation. The maximum likelihood estimation is calculated by tallying the different frequencies from different variables within the context of a given dataset. To do so, a likelihood function is used where (equation 3.11):

$$L(Z|X) = f(X|Z)\,, \tag{3.11}$$

is calculated based on the maximum likelihood estimate for Z, which represents the parameters to be estimated from the observed data X, so that (equation 3.12):

$$\hat{Z} = argmaxL(Z|X)\,, \tag{3.12}$$

can be expressed as a log function where (equation 3.13):

$$\log L(Z|X) = logf(X|Z)\,, \tag{3.13}$$

can be used to derive the maximum likelihood estimate, which can be denoted as (equation 3.14):

$$\hat{Z} = argmaxlogL(Z|X)\ . \tag{3.14}$$

To this extent, the value for Z represents the value that determines the most probable distribution for the data, given Z. To start the process, Z is set to zero for each instance. This does imply that the variables need to have the ability to be stratified into a representation of discrete variables to account for the probability within each node of the network (Scanagatta *et al.*, 2019). For example, instead of having variables within a table represent a range between 1 – 1000 (meaning the numbers of a specific variable may range from 1 to 1000), the ranges need to be converted to discrete variables to become categories representing the information. In other words, instead of having a numeric value between 1 and 1000 within a specific variable (which implies that there could be 1000 different numbers in the dataset for that variable), the data can be stratified into five categories instead to provide 1 – 199, 200 – 399, 400 – 599, 600 – 799, and 800 – 1000. This means that the frequencies will tally the group distribution, instead of the number of times a particular value arises.

Once the conditional probability tables are constructed for the entire BN, estimating predictions within the network and making inferences on the data is possible. Unlike a GAN, the BN builds

its network based on conditional probabilities and the relationship between variables within the BN is known. An example of this is a BN learned from a few datapoints in which the network structure and parameters were learned from a few variables (Xu, 2020). Once the BN (its structure and parameters) were learned, an infinite amount of synthetic data could be produced. Synthetic tabular data can therefore be generated using different methods, models and approaches. Some of these approaches include random number generation, other machine and deep learning approaches, and different types of probabilistic modelling, but in this study, these will not be discussed as they were not as common in the generation of synthetic tabular data (Nikolenko, 2021). Ultimately, whichever approaches are used to generate the synthetic data, evaluating the effectiveness of the data that was generated compared to the original data requires several different evaluation criteria and techniques. In the section to follow, an outline for the evaluation of synthetic data will be given.

3.3.4 Evaluation of synthetic data

To evaluate how effectively synthetic tabular data was generated, several methods need to be used to compare the original dataset to the synthetically generated one. In the section below, an outline of methods used to evaluate synthetic data are explained. The first set of evaluation criteria is in the form of descriptive statistics to measure central tendencies and the distribution of the data. To this extent, the aim is looking at central tendencies of the data and what the slope of the tendency is.

3.3.4.1 Kurtosis

One of the methods that we can use to compare the distribution of variables within a dataset is the use of kurtosis (Balanda and MacGillivray, 1988). Kurtosis is a measurement that is used to measure the sharpness or “tailedness” of a frequency-distribution curve. Kurtosis is a statistical measure used to describe the degree to which scores cluster in the tails or the peak of a frequency distribution. The peak is the tallest part of the distribution, and the tails are the ends of the distribution (Balanda and MacGillivray, 1988). There are three types of kurtoses: mesokurtic (Kurt = 0), leptokurtic (Kurt >0), and platykurtic [(Kurt <0), (Figure 3.7)].

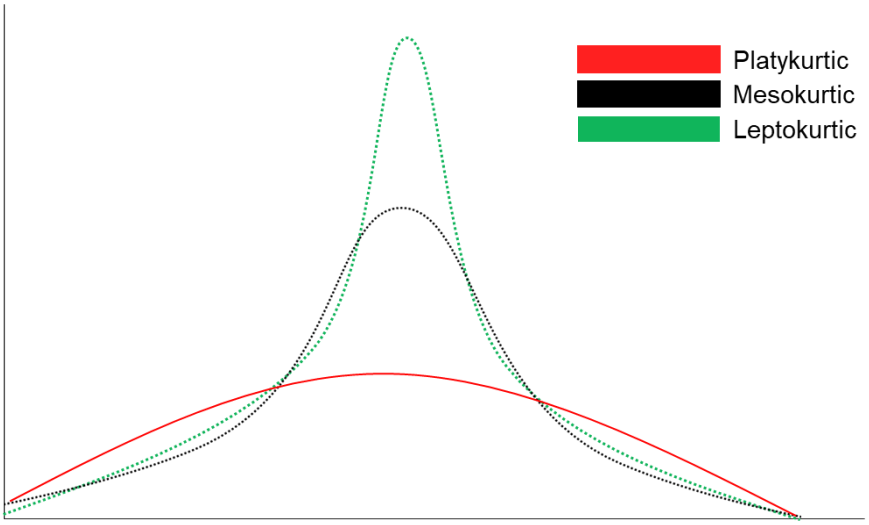


Figure 3-7 General properties of kurtosis frequency distributions.

.

As such, kurtosis can be shown as (equation 3.15):

$$Kurtosis = \frac{\mu_4}{\sigma^4}, \tag{3.15}$$

where μ_4 represents the fourth central and σ the standard deviation. The fourth central moment is a measure of the heaviness of the tails of a distribution. It is calculated as the expected value of the random variable's deviations from the mean, raised to the fourth power. This measure provides insights into the degree of outlier presence in a distribution. Simply put, it quantifies the extent of extreme values (outliers) in the data set. The higher the fourth central moment, the more outlier-prone the distribution. On the other hand, the standard deviation is a measurement of the deviations of units away from the mean. In the context of kurtosis, σ can be calculated from the following (equation 3.16):

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}, \tag{3.16}$$

where x and μ represent differences in the means and n the number of samples. The second method of evaluating the synthetic data against the original data is to use a comparison of means in the form of a t-test.

3.3.4.2 t-test

A t-test is a statistical test used to compare the means of two datasets. It allows one to determine whether there is a significant difference between the means of the two datasets or whether they are similar. In the context of synthetic data generation, a t-test can be used to compare the means of the original dataset and the synthetic dataset generated. By comparing the means of the two sets of variables between different datasets, the central tendencies of the data for their similarity may be calculated, indicating that the datasets are similar given a specific variable in terms of their similarity of means (Kruschke, 2013). To perform a t-test, the means and standard deviations of the two datasets need to be determined. Thereafter, the t-value may be determined as follows (equation 3.17):

$$t = \frac{\widehat{x_1} - \widehat{x_2}}{\sigma_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad , \tag{3.17}$$

where $\widehat{x_1} - \widehat{x_2}$ represent the two different sample means, n_1 and n_2 the two different sample sizes, and σ_p the pooled the standard deviation. The pooled standard deviation is a weighted average of standard deviations from two or more independent groups (Eickhoff *et al.*, 2023). If the t-value is greater than a predetermined critical value, the null hypothesis that the means are the same is rejected, concluding that the datasets are significantly different. If the t-value is less than the critical value, the null hypothesis cannot be rejected, concluding that the datasets are similar on the basis of central tendencies. By performing a t-test on the synthetic dataset and the original dataset, one can determine if the synthetic data accurately represents the original dataset. If the distribution of the two datasets is similar, one can conclude that the synthetic dataset is a good representation of the original dataset's central tendencies and distribution. However, if the distribution is significantly different, it indicates that the synthetic dataset is not a good representation of the original dataset and may need to be adjusted on the basis of the model.

Within the context of comparing datasets, several other statistical measures exist to compare two datasets to one another, such as ANOVA, MANCOVA, and z-tests, to name a few, but measuring the central tendencies does not equate to the data being identical. For example, a dataset may have the same mean, and same standard deviation, but can be fundamentally different, as follows (Figure 3.8):

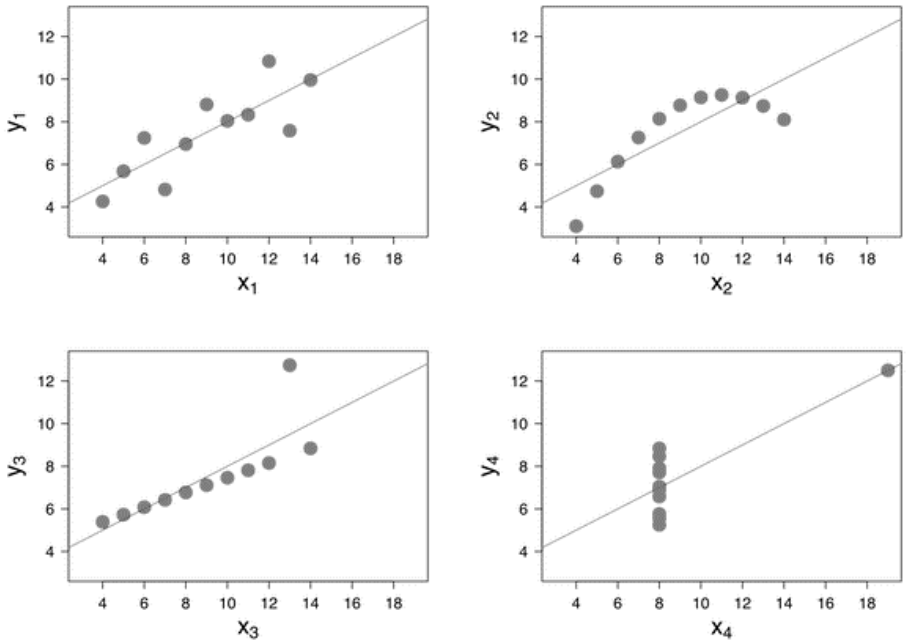


Figure 3-8 Representation of datasets with the same mean and same standard deviation that are fundamentally different (Aggarwal and Ranganathan, 2016).

Given this example, measuring central tendency alone is not enough to compare datasets as there are additional trends associated in the information that need to be considered. To add a different evaluation, use of cumulative sums to identify trends in the data may be used.

3.3.4.3 Cumulative sum and density

Another evaluation method that can be used to measure a synthetic dataset generated from an original dataset is to compare the cumulative sums of the variables within the dataset. Cumulative sum, or cumsum for short, is a mathematical operation that calculates the running total of a sequence of numbers (Canh *et al.*, 2019). The method involves adding up the values of each part in the sequence and accumulating them into a new sequence. The first part in the new sequence is equal to the first elements in the original sequence, the second part in the new sequence is the sum

of the first and second element in the original sequence, the third part in the new sequence is the sum of the first three elements in the original sequence, and this cumulative pattern continues until the sequences are complete. Cumsums thus provide a method by which patterns in the data across various samples can be compared to one another. These trends and patterns have been used in anomaly detection and to detect outliers within datasets. In addition to this, the density of the data may also be determined.

In the context of data analysis, density refers to the distribution of data points within a dataset. The measure of density is based on how closely data points are clustered within a specific range of values (Yang and Delpha, 2022). To compare two datasets with one another, their density curves may be plotted (equation 3.18):

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{3.18}$$

where $f(x)$ is the estimated probability density at point x with n number of observations and h range within each of the datapoints, known as the bandwidth parameter. K then represents what is called the kernel function which is centered around a datapoint for each bandwidth parameter. There are several different kernel functions, but in the context of this study, the Gaussian kernel function (Yin *et al.*, 2011) will be outlined (equation 3.19):

$$K(x) = \frac{1}{n\sigma\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{1}{2}\left(\frac{x_i - x}{\sigma}\right)^2} \quad . \tag{3.19}$$

The Gaussian kernel function assists with transforming some input data into a higher-dimensional feature space where the datapoints can be separated, according to some kind of a parameter. These parameters are scalar, meaning that the quantity that is described by its size alone, and lacks a direction. Measuring the conditional interdependence of synthetically generated data might be useful to compare the real to synthetic data, but it might not be useful for its utility. In the next section, measuring utility in the context of synthetic data will be outlined.

3.3.5 Measuring utility of synthetic data

In this context, the term utility denotes the efficacy of synthetic data in training a machine learning model (Stephens *et al.*, 2022). Unlike evaluating synthetic data in terms of its distribution, being able to measure its utility is a useful method to determine how viable a synthetic dataset can be for a supervised machine learning task. According to Buggineni *et al.*, (2024), utility of a synthetic dataset can be measured using similarity metrics, regression error, clustering, visual inspection or classification accuracy. Therefore, the utility of a synthetic dataset can be assessed by performing a classification task on both the real and synthetic data, followed by the application of a confusion matrix to evaluate the classification accuracy on both datasets to determine if the synthetic data can effectively train a classifier (discussed later in this Chapter, see section 3.3.5.2). Moreover, alternative measures like learning curves can be employed to illustrate the learning rate and efficiency of the classifier, thereby offering insights into the model's training on synthetic data (discussed later in this Chapter, see section 3.3.5.3). This form of utility evaluation has found application in diverse fields, including healthcare and material science (Emam *et al*., 2021 Buggineni *et al.*, 2024). It is important to note that while there exist several methods for evaluating synthetic data, such as conditional predictive impact (CPI), these evaluation metrics do not exclusively test for utility and were not employed in this study. The authors acknowledge the significance of conducting a CPI investigation, but only subsequent to establishing the utility of a synthetic dataset. In the subsections to follow, machine learning classification tasks to test the utility of synthetic data, and both a confusion matrix and learning curves to evaluate model accuracy and visualise learning rates, will be discussed.

3.3.5.1 Machine learning classification tasks

In the context of tabular education data, the class can be the outcome of a particular student criteria, such as student performance at the end of a semester like pass or fail, given all the variables that led to the class classification. These features or variables may have a conditional interdependence, meaning that the value of one variable may depend on the value of another variable within the particular dataset. To test utility, a classification task may be performed on the original data and on the synthetic data (Liu *et al.*, 2010; Wickramaratna, 2010; Montaño-Gutierrez *et al.*, 2017). If the utility is similar between the original data and the synthetic data, then the accuracy scores will be similar, if measured. If the accuracy of the synthetic data surpasses that of the original data, it

can suggest a couple of possibilities. One possibility is that there was sufficient conditional dependence between the variables. This would result in a high utility, and the model could benefit from the increased sample size. Alternatively, the synthetic data might have simplified a certain relationship of interest. On the other hand, if the accuracy scores for the synthetic dataset are lower, this could indicate a different scenario. It might suggest that the conditional dependence of the synthetic data was not as strong as in the original dataset and would thus have a lower utility. This could mean that the relationship between the variables did not effectively translate into the synthetic dataset, but that the degree of this utility may be measured. If the performance is low, then the model failed to learn an underlying relationship in the data which would also result in the dataset having a low utility. In other words, the accuracy of the model in predicting the class of observations can be used as an indicator of the degree of utility given the features in the data. This is therefore a proxy for the utility of the dataset to be used to train a supervised classifier. To evaluate the performance of each of the datasets in this study, accuracy was determined by means of a confusion matrix (CM) applied to different algorithms.

3.3.5.2 Confusion matrix (CM)

By definition, a CM is a calculation used to visualise the accuracy of a classification algorithm by illustrating the relationship between actual and predicted outcomes in a classification task (Haghighi *et al.*, 2018), (Figure 3.9).

Predicted results			
Actual Results		0	1
	0	True Negative = TN	False Positive = FP (Type I error)
	1	False Negative = FN (Type II error)	True Positive = TP

Figure 3-9 Confusion matrix relational diagram.

Outlined above (Figure 3.9), the “1” represents the class the model is trying to predict, and in this context could mean “pass”. On the other hand, the “0” represents the negative class, or the class that is not the primary focus of the model. To this extent, a CM can be used to further evaluate

how accuracy was affected. For an effective evaluation, three different contexts need to be considered for the model score, namely: recall, precision, and an F1-Score (Goutte and Gaussier, 2005). Precision is a measurement identifying how accurate the model is at identifying a specific characteristic and can be broadly defined as the proportion of true positive (TP) classifications within the sum of true positive and false negative (FN) predictions as such (equation 3.20):

$$Precision = \frac{TP}{TP + FP} \quad , \tag{3.20}$$

On the other hand, recall is a measurement that identifies the percentage of the predictions that are correctly classified. To clarify, recall can be defined as the proportion of true positive (TP) classifications from within the sum of true positive (TP) and false negative (FN) predictions so that (equation 3.21):

$$Recall = \frac{TP}{TP + FN} \quad , \tag{3.21}$$

An F1-Score is a weighted average of the true positive rate (recall) and precision so that (3.22):

$$F1\ score = \frac{2(Recall * Precision)}{Recall + Precision} \quad , \tag{3.22}$$

The efficiency of the classification model is based on a calculation of the accurate and inaccurate predictions for each of the predicted outcomes, in which the proportion for the sum of the true positive (TP) and true negative (TN) classifications are extracted from the sum total of the true positive (TP), true negative (TN) and type I (FP) and type II (FN) errors (equation 3.23):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \text{ ,} \tag{3.23}$$

The precision, recall, F1-Score and accuracy of a classification task can be calculated using a variety of different training and test splits in the data and as a comparison between two separate datasets. If a dataset contains 100 students, all their student performance data, categorised into one of two classes, pass (n = 50), and fail (n = 50), one can use a random 70% of the dataset as training data, to classify the remaining 30% of one's dataset. These results can then be used to construct the confusion matrix and interpret the result.

Another approach is to use the data from one entire dataset to predict the classes in another. For example, if one has a tabular education dataset of 100 students with two classes, and another dataset with 10 students with the same variables, then one can use the 100-student dataset as the training data, and the 10-student dataset to test the results. In this example, the model will be looking at generalisation, and the overall performance of the model would be negatively affected. A contributing factor to the precision, recall, F1-Score and accuracy of a classification task, is how balanced the data are. However, the confusion matrix only provides the overall scores at the end of an analysis. To understand how the algorithms performing the ML task learned over time, a learning curve may be used.

3.3.5.3 Learning curves

One way to represent training efficiency is through the use of a learning curve (Amari, 1993). A learning curve is a visual representation of the performance of a machine learning model over time as it learns from new examples. Typically, the learning curve plots the training error and the validation error as a function of the number of training examples used. In terms of training efficiency, the learning curve can provide insights into how efficiently the model is learning from the data as the training sample size increases (Figure 3.10).

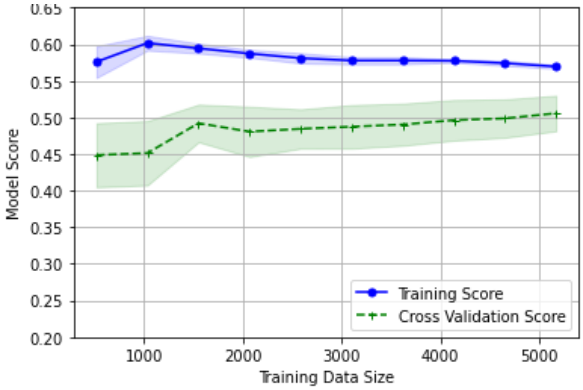


Figure 3-9 Example of a learning curve.

In addition to this, the slope of the learning curve can indicate how quickly the model is improving with additional training examples. Sudden changes in a learning curve can indicate a sudden shift in the variable distributions or homogeneity of the data (Figure 3.11).

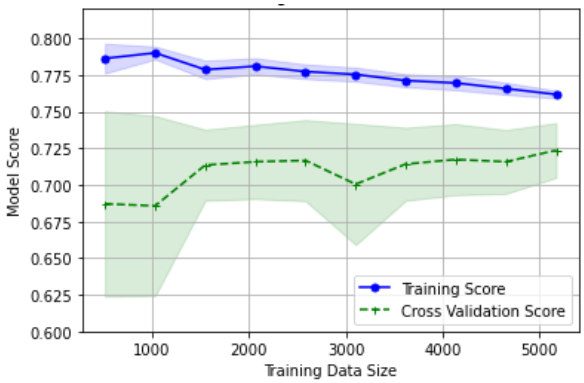


Figure 3-10 Slope of a learning curve.

If the slope of the learning curve is steep, this indicates that the model is learning quickly and efficiently from the data. On the other hand, if the slope of the learning curve is shallow, this indicates that the model is not learning as quickly and may require more training examples or more complex features to improve its performance (Figure 3.12).

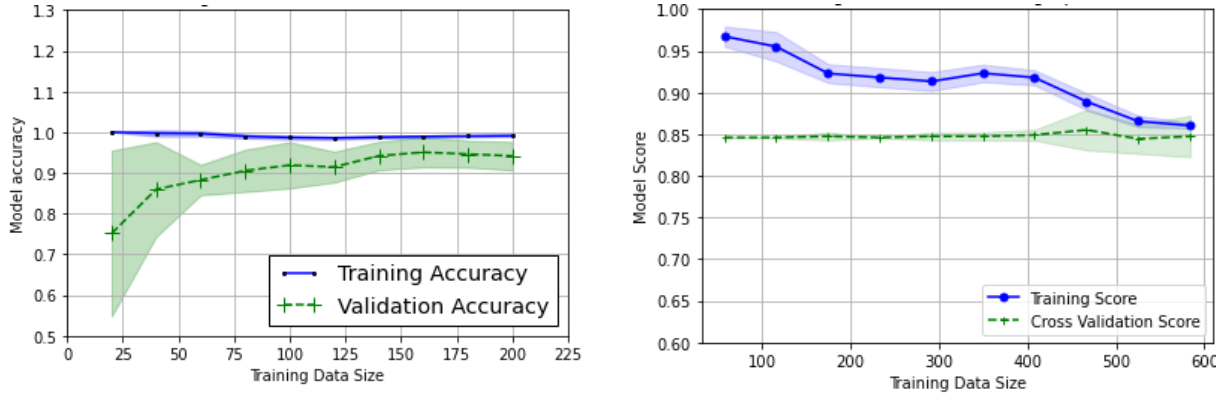


Figure 3-11 (Left) an example of adequate learning, (Right) an example of a model not learning due to underlying issues in the training data.

Additionally, the learning curve can also reveal information about overfitting and underfitting, which can impact the training efficiency of a model (Figure 3.13).

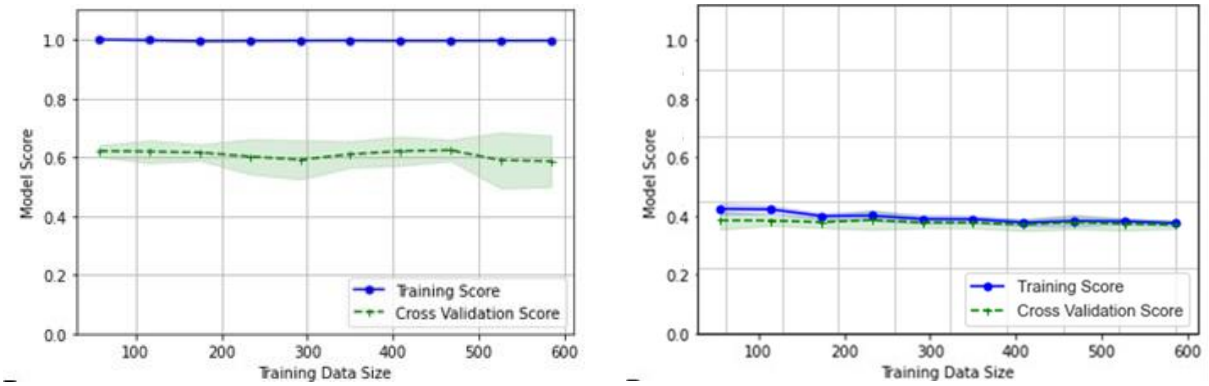


Figure 3-12 (Left) an example of overfitting, (Right) an example of underfitting.

Overfitting is a common problem in ML where a model becomes too complex and learns to fit the noise in the training data, resulting in poor performance on data outside the training data (Ying, 2019). In a learning curve, overfitting would look like a large gap between the training and cross validation score curve. The training curve would show high accuracy or low loss, while the cross-validation score curve would show lower accuracy or higher loss, indicating that the model is overfitting to the training data. As the model's complexity increases, the training curve will continue to improve while the cross-validation score curve will start to plateau and eventually decrease. One way to address overfitting is to use more data for training, which can help the model to learn more robust patterns and reduce the risk of overfitting.

Underfitting occurs when the model is too simple and cannot capture the underlying patterns in the data, resulting in poor performance on both the training and testing data (Cunningham and Delany, 2021). On a learning curve, underfitting would appear as a plateau or a convergence of both the training and testing curves at a low score. This indicates that the model is not complex enough to capture the patterns in the data and needs to be trained with more features or a more complex model architecture. Increasing the complexity of the model can help to address underfitting by allowing it to capture more patterns in the data, but too much complexity can lead to overfitting. All figures in this section are mere illustrations of the different types of learning curves that can be expected, and the different types of interpretations to them.

To this extent, there is always a balance that needs to be reached between overfitting and underfitting a model, by either increasing the complexity of the data, decreasing the complexity of the data, and/or increasing the sample size of the training data. Given that all of the aforementioned models require good and substantial data to build models and evaluate models for the education context, the purpose of this chapter was to generate and evaluate synthetic tabular data for the education context by comparing synthetic data generated from a GAN and a BN.

3.4 Methodology

The experiments in this section involved testing and evaluating synthetic data generated from an open source dataset for the education domain (https://github.com/dsfsi/Higher_Education_EDA/tree/main/opendata [last accessed 15 October 2022]). This dataset is publicly available and is based on public domain data. The dataset from an engineering student cohort was used, and all the context from the original dataset was kept. This dataset contained only a few variables, namely: gender, grade point average of the first year (CGPA100), grade point average of the second year (CGPA200), grade point average of the final year (CGPA300), and overall grade point average. Variables within the dataset were transformed into discrete categorical variables from their original data types for the BN, and only the class variable was converted to discrete variables for the GAN (called the “Class variable” in the final dataset). For all the feature variables, not the class variable, each numerical variable was stratified into three categories based on an even 33.33%, given the range of the variable (and not the available data), so that (equation 3.24):

$$Distribution = \frac{range\ of\ the\ variable}{3}. \quad (3.24)$$

This means that only discrete variables were used and measured within the dataset for the BN, and a combination of discrete and continuous variables for the GAN. The purpose of this distinction was to simplify the data for the purpose of the comparison between the synthetic and original data. The overall grade point average was feature engineered to be discrete into one of three classes, namely: fail, pass, and pass with a distinction. The data were unbalanced, with a distribution of data containing students that failed (n = 302; 9.97%), passed (n = 1659; 54.77%), and passed with distinction (n = 1068; 35.26%) (Table 3.2).

Table 3-2 Distribution of class variables used in the experiment.

Name of class	**Distribution (%)**
Fail	302 (9.97%)
Pass	1,659 (54.77%)
Pass with distinction	1,068 (35.26%)

The synthetic tabular data was generated from the aforementioned datasets using a GAN (see Chapter 3, section 3.2.2) and a BN (see Chapter 3, section 3.3.3), totalling 10,000 different students' information in synthetic data that was compared to the original data. The hyperparameters for the GAN were not fine-tuned, and the chosen hyperparameters in the GAN was based on the default settings for synthetic tabular data as described by the publications in a survey conducted by Figueira and Vaz, (2022) and the systematic review of tabular data as described by Hernandez *et al.,* (2022). In the case of GAN, the epochs and batch size were set to 300 and 500, respectively. The epochs of 300 was based on general consensus for tabular data, as outlined in Figueira and Vas, (2022). Epochs refers to one complete run of the method through an entire dataset. In other words, a complete epoch involves running every training sample in the dataset, updating all weights and biases, going through all batch sizes. The batch size of 500, which is large for a GAN, was used because the complexity of the data was low. In the context of batch

size, this hyperparameter refers to the number of samples the GAN processes before updating the internal model parameters. The choice of epochs and batch size will depend on the complexity of the data and the computational resources available to complete the task. For tabular data, epochs can be set to greater than 400 due to the dimensionality of the data, but for more complex data types, the epochs will be set lower, depending on the computational resources (Xu, 2020). For the GAN, the use of log frequency for discrete variables in conditional sampling were enabled, with an embedding size of random samples passed through the generator set at 128. The log frequency was enabled because it could handle uneven distributions within the discrete variables. The embedding size of 128 was decided upon based on the library specifications and setting the embedding size to 128 means that each piece of random noise input to the generator is a 128-dimensional vector. The size of both the output samples for each of the residuals and discriminator were set to 256. This choice in residuals is double that of the embedding size. For the BN, both a constraint-based method and a score-based method was used, by applying a BIC (see Chapter 3, section 3.3.3). The synthetic dataset for the GAN and the BN was generated from an original dataset containing 3,029 different samples. The choice in sample size was to ensure that ample training data was represented within the dataset.

To measure the utility of the synthetic data, three classification algorithms namely LR, DT, and kNN were used to predict the class variable of the original data and in each of the datasets that were synthetically generated by the GAN and the BNs (see Chapter 3, sections 3.3.1 and 3.3.4). This study was not focused on improving the accuracy of the models by fine tuning hyperparameters, therefore the hyperparameters for each model was kept constant throughout all the experiments. As such, the hyperparameters for the classifiers were based on the education classification tasks as described by Kotsiantis *et al.*, (2006), Lau *et al.*, (2019), and Pallathadka *et al.*, (2023). For LR, inverse regularisation strength was set at 1, with no class weight and a fit intercept with an intercept scaling value of 1. LR also had no L1 ratio, maximum iteration value set at 100, and verbose set at 0, and the warm start function disabled for the model. The choice over a smaller inverse regularisation strength allows the model to generalise better to unseen data (Cawley *et al.*, 2006). It was set to 1 to balance the trade-off between underfitting and overfitting, but for better results, further experimentation is required. The class weight parameter is determined by how balanced the data are – generally for unbalanced data a weight is determined but, in this

case, the weight was removed. For the maximum number of iterations, a low value might cause the solver to stop before fully converging, while a high value might cause longer training times. It was set to 100 as a balance between training time and model performance and this number is often predetermined based on experimentation, but as indicated the hyperparameters chosen were based on prior studies (Ertekin *et al.,* 2007). Verbose refers to the amount of information the algorithm provides during training. If verbose is set at 0, no output is provided during training and if verbose is set at larger than 0, more detail is provided the higher the number. The higher the verbose, the slower the training and the choice to keep verbose at 0 was to speed up the analysis.

For DT, no maximum depth was specified. The minimum impurity decrease for DT was set at 0. Furthermore, DT had a minimum leaf sample size at 1, and a minimum sample split set at 2, with a minimum weight fraction for each leaf set at 0. Maximum depth refers to the number of nodes that need to be specified within a tree until all the leaves are prune. The choice not to specify the maximum depth has the potential to lead to complex tress that overfit the data. The reason why no maximum depth was chosen is because the dataset is known and is not as complex (Roshanski *et al.,* 2023), however, for complex datasets it is vital to specify a maximum depth otherwise it will lead to overfitting (Irvin *et al.,* 2021). The choice to set the leaf size to 1 was to optimise better classifications with the tabular dataset, however, with larger more complex datasets and tabular datasets with more features, the leaf size needs to be fine tuned to a better number using a variety of techniques to fine tune such as the correct cross validation procedures (Kirchner *et al.,* 2006). A minimum sample split of 2 means that at least two samples are required for a node to split, which is the smallest possible value for this parameter, allowing the tree to grow as deep as possible (Song and Ying, 2015). Setting the minimum weight fraction for each leaf at 0 means there's no constraint on the weights of the instances that a leaf node can contain, giving the model the freedom to make very specific classifications (Olson *et al.,* 2018).

For kNN, the leaf size was set at 30, with the Minkowski ($p = 2$) as the chosen metric. The number of metric parameters were set to none with $k = 5$. Finally, the weighting of kNN was calibrated to uniform, as the dataset was not complex. The Minkowski metric is a type of p-norm distance used in the kNN algorithm. It is a generalisation of other distances. When $p = 1$, it becomes Manhattan distance, and when $p = 2$, it becomes Euclidean distance. The choice of distance metric depends

on the nature of the data. If the features have the same scale or units, Euclidean distance (which is a special case of Minkowski distance) is the best choice, such as the case with the education tabular data used in this study. If the features have different scales or units, Manhattan distance might be more appropriate, but in this study was not considered (Maghari, 2018). The number of *k* is the number of neighbours the model considers. Although there are several metrics to use to choose this number, it is generally accepted to use an odd number of *k* (so that the classifier choose a class in difficult cases), as well as a number that is not too large as larger numbers require more computational resources (Zhang *et al.*, 2017).

Due to the simplicity of the data, a lack of hyperparameter tuning for each experiment, there is a risk that the data will be overfit in all class predictions as the variables will represent an oversimplification of the underlying data (see Chapter 3, section 3.3.5.1). Despite this, the overall accuracy will still provide an indication of the interdependence in the synthetic data between the tabular data generated from the GAN and BN. To this extent, the Precision, Recall, Accuracy and an F1-Score were evaluated for each of the algorithms (see Chapter 3, section 3.3.5.1). The two BN structure learning methods were compared to one another using a confusion matrix (see Chapter 3, section 3.3.5.2). If there were no differences between the original and synthetic data for the GAN and BN, then the synthetic data generated using these two models would be similar. The learning rate as well as an evaluation of the models in terms of a fit for the data (overfitting or underfitting) was performed and visualised using a learning curve (see Chapter 3, section 3.3.5.3). Once all the experiments were performed, the best fit models, synthetic data generation pipeline, and algorithms were used to describe the best fit to generate synthetic tabular data using the open-source dataset. In the next section, the results will be given, first, by outlining the experiments conducted using the GAN, and secondly, the experiments conducted with the BN.

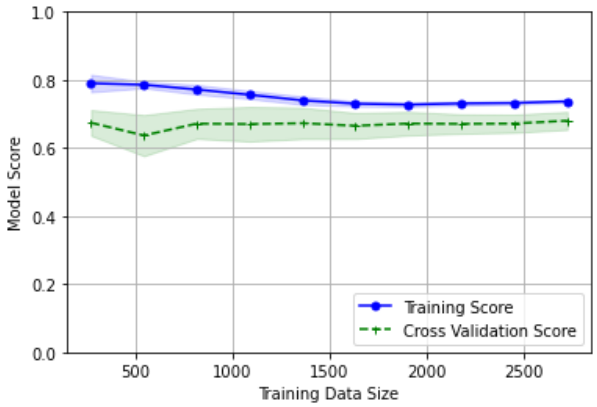
3.5 Results and discussion

3.5.1 Measuring utility on original data

The utility was measured using the three different algorithms. Firstly, a series of experiments were performed on the original data. This baseline served as the starting point of comparisons between different combinations of training and testing data. The first algorithm compared on the original testing and training data without synthetic data was kNN (overall accuracy = 76.95%). Even

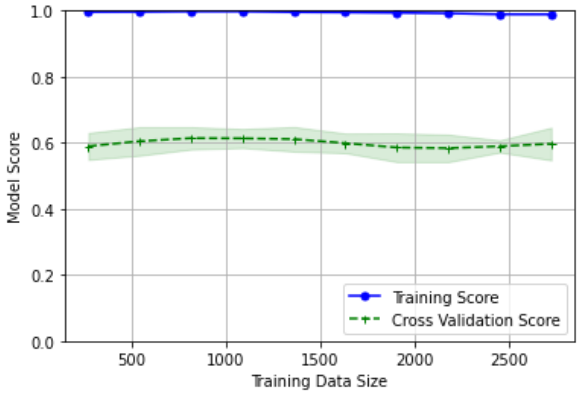
though the training score decreased, the cross-validation score increased over time. Based on these results, more data can yield better results (Figure 3.14). Next, the DT (overall accuracy = 65.1%) was performed on the original data. Based on the learning curve, the model overfitted the data. The last classification task performed on the original data was LR (overall accuracy = 62.01%). This model underfit the data. A further observation is the failure of the LR to identify class variable 1. This can be due to the binary nature of the LR choosing the two most represented class variables (class variable 2 (pass) and class variable 3 (pass with distinction)) as the data were unbalanced. Observations made on the original data in terms of the utility indicated that kNN was the best performing algorithm followed by DT and LR. DT overfit the data, and LR underfit the data. One possible explanation for the overfit data could be the simplicity of the dataset, and the lack of large training data used in the classifications, and a feature in the original data that skews the classification task. Another observation made was that class variable 3 (pass with distinction) had the highest precision for kNN and DT, whereas class variable 2 (pass) was the highest for LR. In terms of recall and the F1-scores between the algorithms, kNN (recall = 85%, F1-score = 72%), DT (recall = 83%, F1-score = 72%), and LR (recall = 69%, F1-score = 67%) scored the highest for class variable 2 (pass). Between the three classifiers, kNN performed the best, whereas DT overfit the data and LR underfit the data. Based on these results, The highest precision for predicting class 3 variables (pass with distinction) was achieved by kNN and DT. This means that these two algorithms were more accurate in predicting students who would ‘pass with distinction’ and made fewer mistakes (false positives) compared to LR. kNN and DT were better at identifying students who would class 2 variables (pass), capturing a higher percentage of students who did indeed class 2 variables (pass). In terms of F1-score, kNN and DT outperformed LR for predicting class 2 variables (pass). This suggests that they achieved a better balance between precision and recall when predicting class 2 variables (pass). In terms of the utility, the sensitivity (recall) was highest for kNN for class 2 variables (pass). Next, the same experiments were performed to measure the utility on the GAN generated synthetic tabular data.

(a) kNN learning curve, accuracy, precision, recall, F1-score on original data



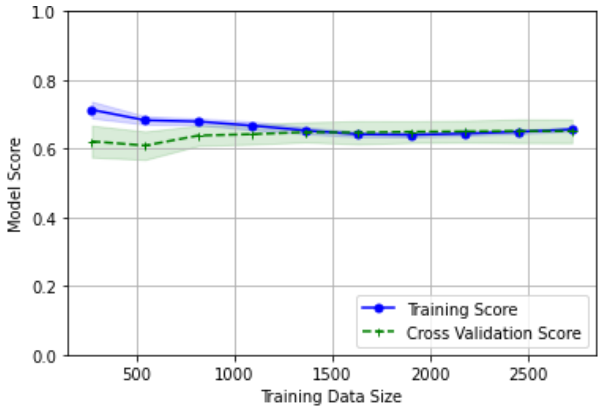
Class variable	Precision	Recall	F1-score	Support
Fail	0.61	0.37	0.46	302
Pass	0.77	0.85	0.8	1659
Pass with distinction	0.81	0.76	0.78	1068
Accuracy	76.95%			

(b) DT learning curve, accuracy, precision, recall, F1-score on original data



Class variable	Precision	Recall	F1-score	Support
Fail	0.47	0.03	0.06	302
Pass	0.64	0.83	0.72	1659
Pass with distinction	0.68	0.55	0.61	1068
Accuracy	65.1%			

(c) LR learning curve, accuracy, precision, recall, F1-score on original data.



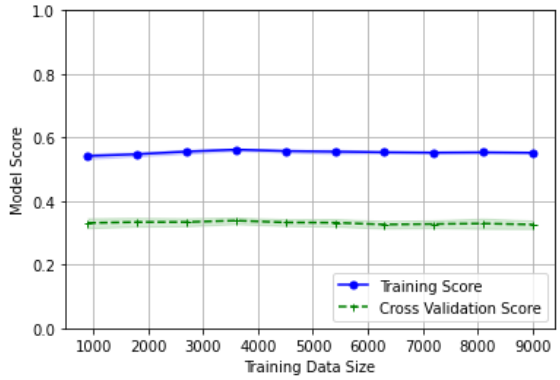
Class variable	Precision	Recall	F1-score	Support
Fail	0	0	0	302
Pass	0.65	0.69	0.67	1659
Pass with distinction	0.58	0.69	0.63	1068
Accuracy	62.01%			

Figure 3-13 Learning curves and confusion matrix for kNN, DT, and LR on the original data.

3.5.2 Measuring utility on GAN generated synthetic tabular data

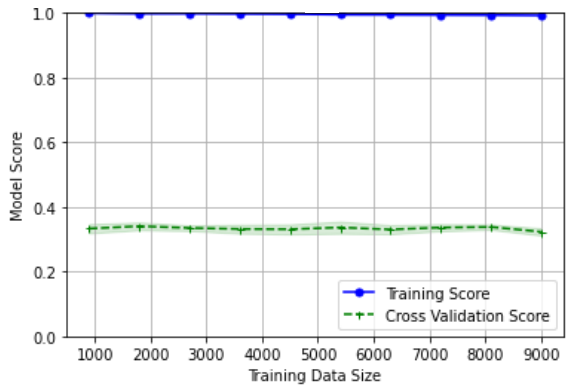
The kNN (overall accuracy = 37.06%) algorithm overfit the data (Figure 3.15). Similarly, the kNN classification task performed on the original data, and the GAN generated synthetic tabular data performed best on class variable 3 (precision = 38%, recall = 51%, and F1-score = 44%). However, there was a significant decrease in the correct classifications based on the synthetic data. Another observation is the stability of the training and cross-validation score, neither of which had improved performance on the larger dataset. In this study increasing the size of the data set did not result in a better prediction of the classifier. For the DT (overall accuracy = 36.59%), the training score was nearly 100%. In addition to this, similarly to the kNN algorithm, the model overfited the data. Like the results of the kNN, an increase in training data size did not improve the accuracy. The findings for LR (overall accuracy = 37.33%) underfit the data. In the original dataset, the LR also underfit the data, but unlike the original dataset, poorer performance of precision, recall, and the F1-score were observed. The overall utility decreased with synthetic data generated with a GAN. In the original dataset, kNN and DT demonstrated superior precision in predicting the outcome of class variable 3 (pass with distinction). However, upon application to the synthetic dataset, a significant decrease in performance was observed. The LR algorithm had the highest precision for the prediction of class variable 2 (pass) in the original dataset, also illustrated a decrease in performance when applied to the synthetic dataset. Regarding the recall and F1-score metrics for the prediction of class variable 2 (pass), kNN and DT outperformed LR in the original dataset. Nevertheless, a decline in these metrics was noted for all three algorithms when they were applied to the synthetic dataset. Moreover, the synthetic data led to overfitting in the kNN and DT, and underfitting in the LR model. These observations suggest that the synthetic data may not have adequately captured the complexity and nuances inherent in the original dataset, thereby leading to the poorer performance of the algorithms. The last set of experiments performed, compared the error rates of the training data and test data expressed as a relative percentage of the dataset. To do so, the training data consisted of the original dataset, and the test data consisted out of the GAN generated synthetic tabular dataset (Figure 3.16).

(a) kNN learning curve, accuracy, precision, recall, F1-score on GAN generated synthetic tabular data



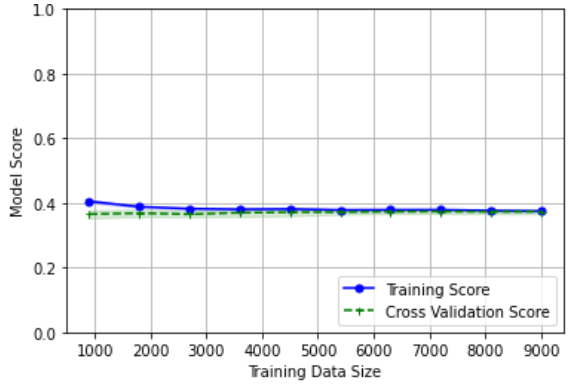
Class variable	Precision	Recall	F1-score	Support
Fail	0.29	0.04	0.06	2806
Pass	0.37	0.49	0.42	3548
Pass with distinction	0.38	0.51	0.44	3646
Accuracy	37.06%			

(b) DT learning curve, accuracy, precision, recall, F1-score on GAN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Fail	0.27	0.06	0.09	2806
Pass	0.37	0.38	0.38	3548
Pass with distinction	0.37	0.59	0.46	3646
Accuracy	36.59%			

(c) LR learning curve, accuracy, precision, recall, F1-score on GAN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Fail	0	0	0	2806
Pass	0.37	0.52	0.43	3548
Pass with distinction	0.38	0.52	0.44	3646
Accuracy	37.33%			

Figure 3-14 Learning curves and confusion matrix for kNN, DT, and LR on GAN generated synthetic tabular data.

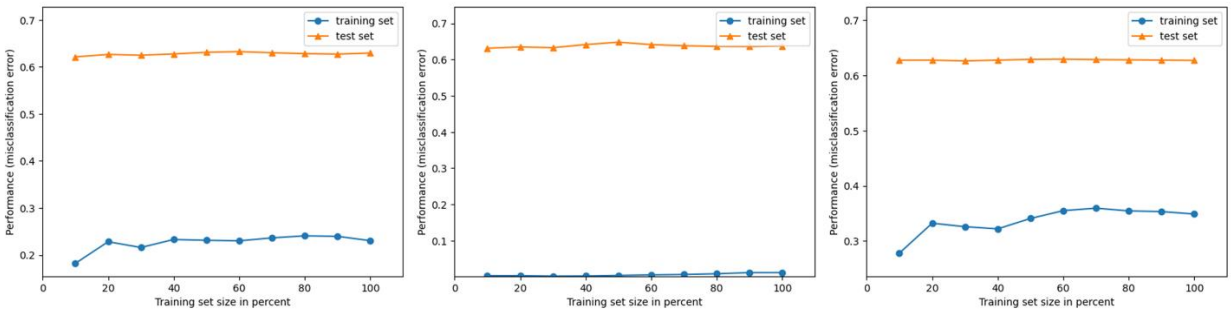


Figure 3-15 Overall performance of kNN, DT and LR with original data as training set and synthetic data as test set.

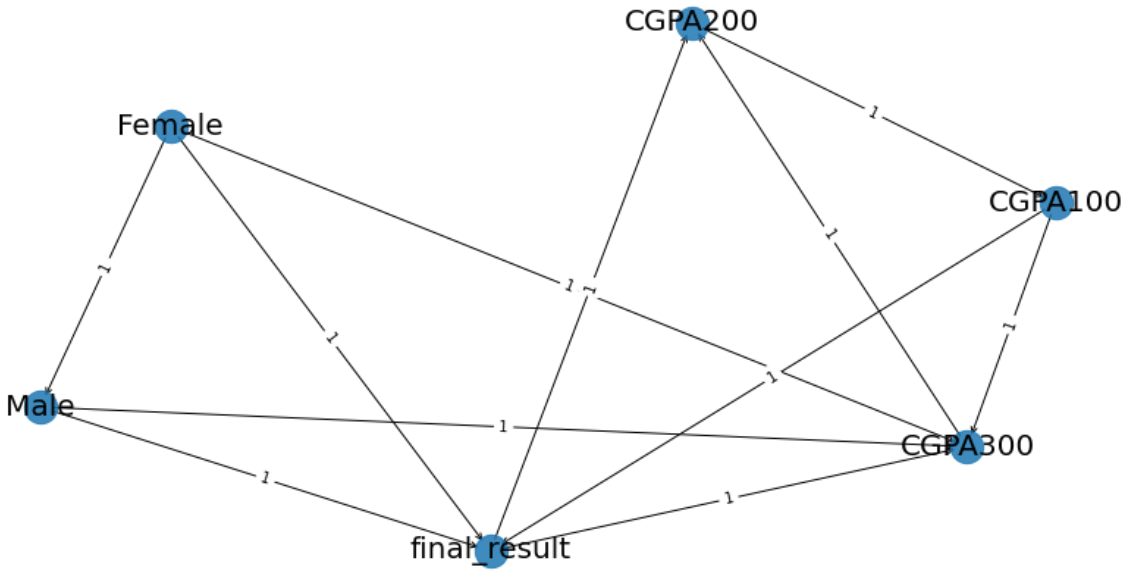
These findings indicate that there were error rate differences between the two datasets, with a much greater relative error in the testing data. These findings indicate that the algorithms failed to recognise the test data as similar to the original data on the basis of the utility between the variables in the datasets. For a GAN to optimally function, a deep understanding of domain specific characteristics are needed. Furthermore, the GANs performance can be improved with extensive experimentation on the various hyperparameter fine tuning, exploring different types of network architectures, and measuring the outputs changes to different loss functions used (Alarsan and Younes, 2021). The hyperparameters in question are specific to the algorithms tuning that is executed prior to the implementation of the algorithm. Some of these include adjusting the batch size, the learning rate, number of epochs and the latent dimensions. Secondly, the network architecture is specific to the way in which the neural network is constructed with respect to both the generator and discriminator. For the architecture adjustments, the optimal number of layers, the type of layer, an exploration into the optimal sigmoid threshold and the optimal normalisation techniques should be explored. The last series of experiments that may be further conducted relate to the best loss function after the most optimal hyperparameters and network architectures were established. These may include investigating the effect of binary cross entropy, Wassertein distance, or the squared error between the binary labels and the discriminator outputs (Zhang *et al.*, 2023). The optimal tuning of parameters, architecture and loss of the GAN was not explicitly part of the focus of the study. The authors note that the performance of the algorithm may be greatly improved with this level of finetuning, however, this extensive deep learning fine tuning

is grounds for further study. In the next section, the results from the synthetic data generated from the BN will be discussed.

3.5.3 BN structure learning from data

To learn the structure of the BN, both a score-based method and a constraint-based method were used (see Chapter 3, section 3.3.3). When learning a BN from data, the way in which the variable pairs are associated to one another will make a difference to the probabilistic distribution in the network and the inference that can be drawn from them (Daly *et al.*, 2011). In both instances when the BN was learned from data, gender had different associations to the outcome. In the first instance, the constraint based DAG outlined the female gender was linked to both male, CGPA300 and final result. For the score-based method, the female gender did not lead to any outcome variable in the BN. Moreover, while CGPA100, CGPA200, and CGPA300 were interconnected, the sequence of these associations did not follow a linear pattern in either of the DAGs. This observation underscores the complexity and non-linearity inherent in the relationships among these variables. It must be noted that while a DAG is a powerful method visualisation for trying to understand complex systems, there are challenges inherent in DAGs learned from data, including the misinterpretation and misrepresentation of causal relationships (Luthfi *et al.*, 2018). The absence of a direct link between two variables does not rule out a causal effect. The effect could be mediated through other variables in the network, or it could be that the data does not adequately represent all relevant scenarios, as described in Daly *et al.*, (2011). Therefore, caution must be exercised to avoid causal misinterpretation. Not explored in this study, but noted are methods and domain knowledge are essential in validating any inferred causal relationships when a DAG is constructed (Figure 3.17).

DAG of constraint-based BN



DAG of score-based BN

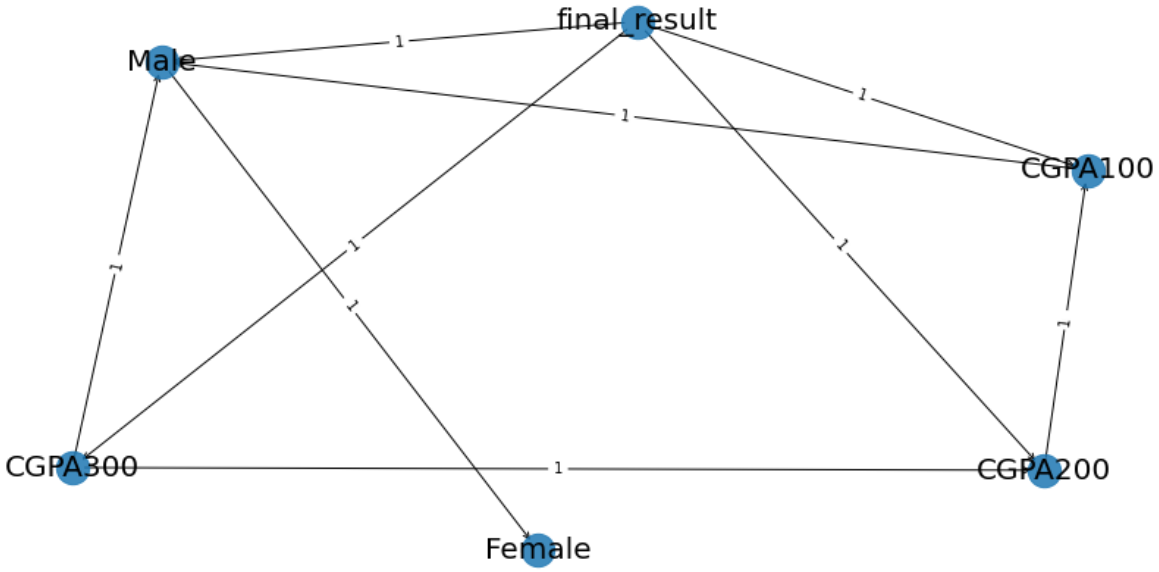


Figure 3-16 DAG structure learning results from a score-based and constraint-based methods.

When both of the DAGs were compared, there were observed differences in the direction and the relationships between the variables (Figure 3.18). In total, there were four connected node similarities between the two structures (Figure 3.18: lines in blue between Male and CGPA100 and CGPA300). In a study comparing a constraint-based method and a score based method, it was determined that the structures do differ if the underlying data were either not correctly classified, or if the underlying information lacks a latent variable that might enhance the strength of the relationship between the variables for a score-based and constraint-based method to have similar results (Werhli *et al.*, 2006). In several studies, it has been found that structure learning becomes problematic when the discrete variables are closely related to one another (Uusitalo, 2007; Zhou *et al.*, 2014; Kabir, and Papadopoulos, 2019). In both instances of the structure learning, the algorithms could create the structure, but there were differences observed between them.

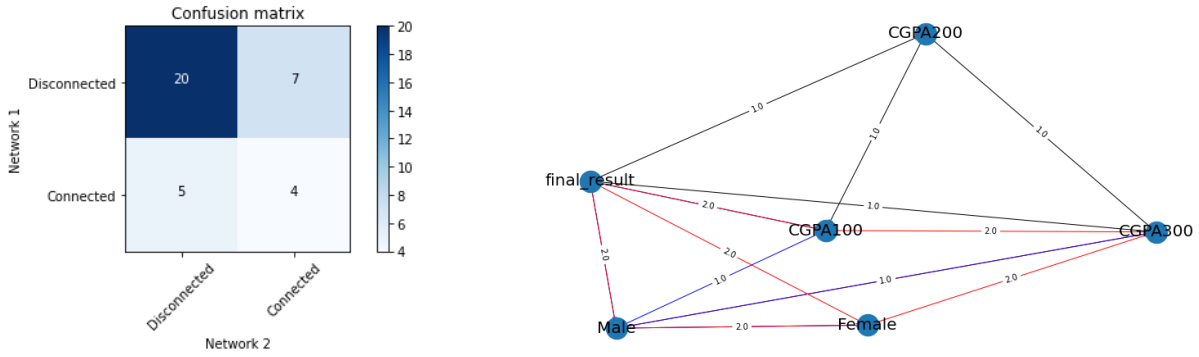


Figure 3-17 Comparison between score-based and constraint-based DAG from the data.

Despite the observed differences between these DAGs, it was important to measure the utility. Even though the structures are not identical, if synthetic data can be generated from the with a high enough utility, then it is still useful for the education context. Next, to measure the utility of the data generated by the probabilistic model in both instances, the same machine learning tasks were performed on the data. In the next section, the results from the classification will be provided for the raw discrete data, the DAG from the score-based method, and the DAG from the constraint-based method.

3.5.4 Measuring utility on raw discrete tabular data

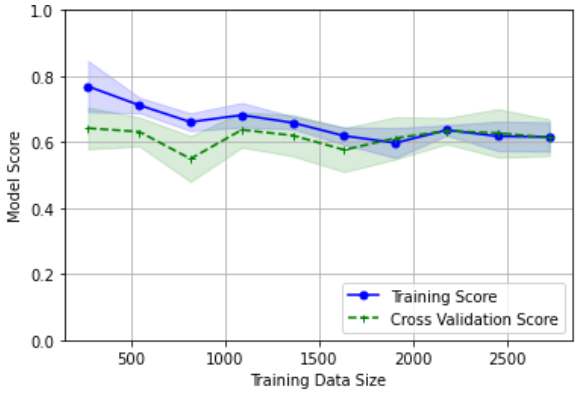
All continuous variables were feature engineered to be discrete so that a raw discrete dataset could be created without any continuous variables. The first set of results related to classifications were performed on the raw discrete dataset – the first of which included kNN (overall accuracy = 61.69%) (Figure 3.19), and the second set of results for the DT (overall accuracy = 66.75%) performed better for class 2 variables, unlike the results observed on the continuous raw data. The last classification that was performed on the raw discrete dataset was LR (overall accuracy = 63.91%).

The shape of all three classifications indicates that there was a closer tendency for models to underfit the model, however, learning the shape and classification of the data did take place. The overall performance of the classifiers for the discrete raw data was poorer than the raw continuous data. A possible explanation for this was the lack of contextual categories in the raw discrete data that were created during the feature engineering process. In a study performed on DAGs, it was found that more complex data with more variables within each parent and child node, yields better results in terms of the complexities associated with the variable pairs (Talvitie *et al.*, 2019). In the context of a classification task, this has the potential to alleviate some of the underfit results observed thus far. The next set of utility using a classification task was performed on the synthetic data generated from the DAG created from the constraint-based method.

3.5.5 Measuring utility on constraint-based DAG tabular data

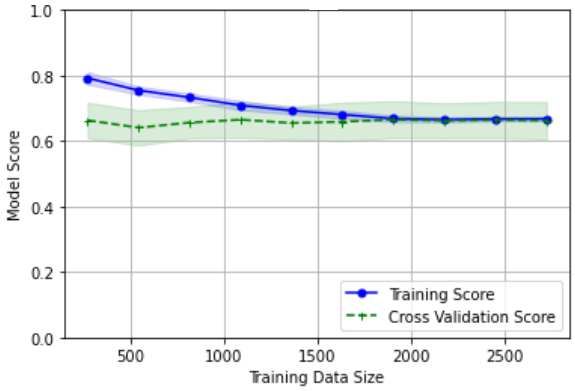
The first experiment measured kNN (overall accuracy = 47.78%) was outperformed by the DT (overall accuracy = 56.73%) and LR (overall accuracy = 55.69%) (Figure 3.20). When comparing the synthetic tabular data between the GAN, the constraint-based method produced better results. Despite the better accuracy scores, all the models underfit the data. In different studies performed on underfit data it was recommended that increasing data complexity can assist in alleviating this problem (Ghasemian *et al.*, 2019; Bashir *et al.*, 2020). Since the data generated only contained, at most, three variables in each parent of child node, the data might be too simplified for the model to learn the appropriate context. Similar observations have been made on probabilistic models looking at its impact on classification tasks and predictions.

(a) BN raw data kNN learning curve, accuracy, precision, recall, F1-score on original data



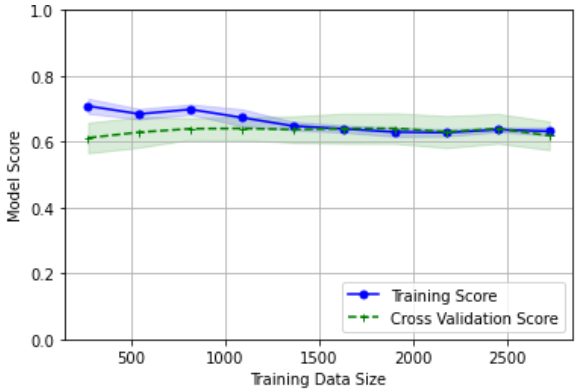
Class variable	Precision	Recall	F1-score	Support
Fail	0.26	0.18	0.22	302
Pass	0.62	0.82	0.71	1659
Pass with distinction	0.74	0.43	0.55	1068
Accuracy	61.69%			

(b) BN raw data DT learning curve, accuracy, precision, recall, F1-score on original data



Class variable	Precision	Recall	F1-score	Support
Fail	0.64	0.69	0.68	302
Pass	0.7	0.71	0.7	1659
Pass with distinction	0.63	0.79	0.7	1068
Accuracy	66.75%			

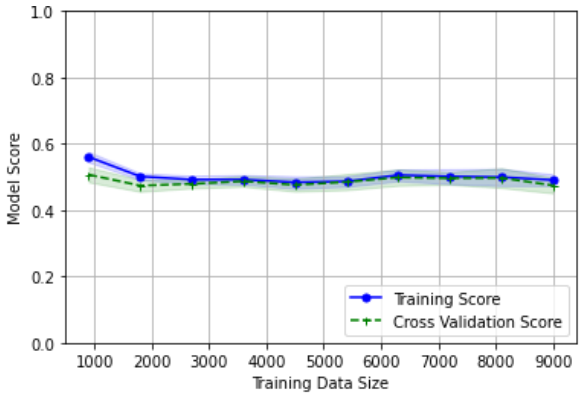
(c) BN raw data LR learning curve, accuracy, precision, recall, F1-score on original data



Class variable	Precision	Recall	F1-score	Support
Fail	0	0	0	302
Pass	0.65	0.73	0.69	1659
Pass with distinction	0.62	0.67	0.64	1068
Accuracy	63.91%			

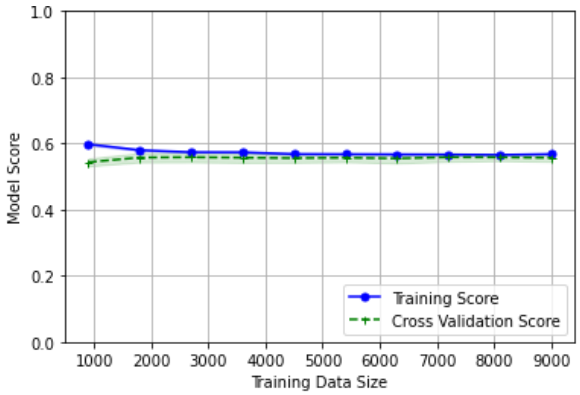
Figure 3-18 Learning curves and confusion matrix for kNN, DT, and LR on original discrete data.

(a) BN constraint based synthetic data kNN learning curve, accuracy, precision, recall, F1-score



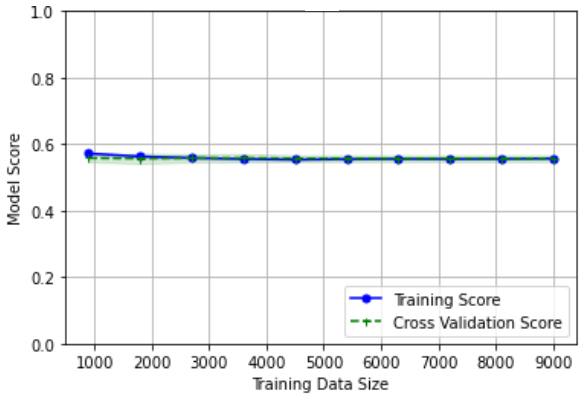
Class variable	Precision	Recall	F1-score	Support
Fail	0.19	0.25	0.22	1585
Pass	0.55	0.65	0.59	4899
Pass with distinction	0.56	0.33	0.41	3516
Accuracy	47.48%			

(b) BN constraint based synthetic data DT learning curve, accuracy, precision, recall, F1-score



Class variable	Precision	Recall	F1-score	Support
Fail	0.38	0.01	0.02	1545
Pass	0.56	0.79	0.66	4896
Pass with distinction	0.58	0.5	0.53	3559
Accuracy	56.73%			

(c) BN constraint based synthetic data LR learning curve, accuracy, precision, recall, F1-score



Class variable	Precision	Recall	F1-score	Support
Fail	0	0	0	1545
Pass	0.55	0.78	0.65	4896
Pass with distinction	0.57	0.5	0.53	3559
Accuracy	55.69%			

Figure 3-19 Learning curves and confusion matrix for kNN, DT, and LR on constraint-based DAG synthetic tabular data.

Similarly, the error rates of the training data and test data expressed as a relative percentage of the dataset was conducted using each algorithm where the training data and test data differed. In this instance, the training data contained the discrete raw dataset, and the test data contained the constraint-based DAG synthetic data (Figure 3.23).

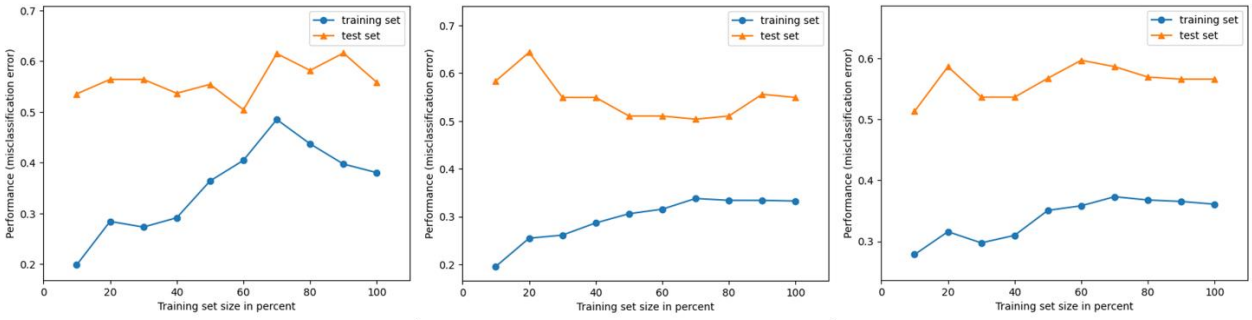


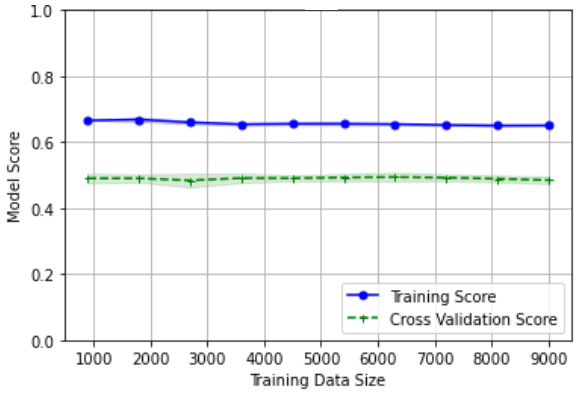
Figure 3-20 Overall performance of kNN, DT and LR with original data as training set and constraint-based DAG synthetic data as test set.

These results indicate that although the training data had higher error rates for the raw discrete data, the overall error rates were less than the GAN produced synthetic tabular data. This indicates that the utility of the synthetic data generated from the DAG was greater than that observed in the GAN. In the last section, the score-based DAG synthetic data experiments will be illustrated and discussed.

3.5.6 Measuring utility on score-based DAG synthetic tabular data

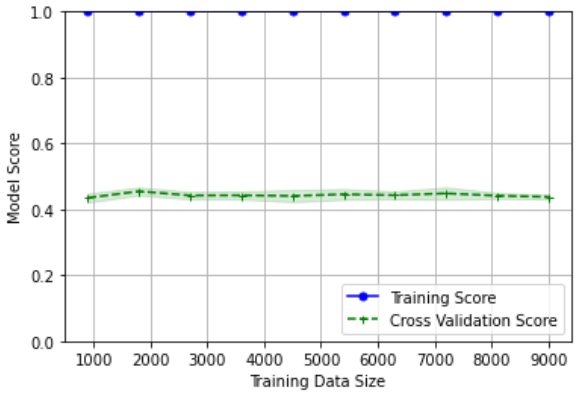
In the last experiments, kNN (overall accuracy = 66.78%) slightly overfit the data, whereas DT (overall accuracy = 100%) overfit the data (Figure 3.22). LR (overall accuracy = 58.31%) underfit the data in a similar way as the experiments performed on the synthetic data generated from the constraint-based DAG. Between the score-based methods and the constraint-based methods, the score-based methods had a better overall accuracy. Another observation was that the score-based methods were more prone to overfit the data.

(a) BN score based synthetic data kNN learning curve, accuracy, precision, recall, F1-score



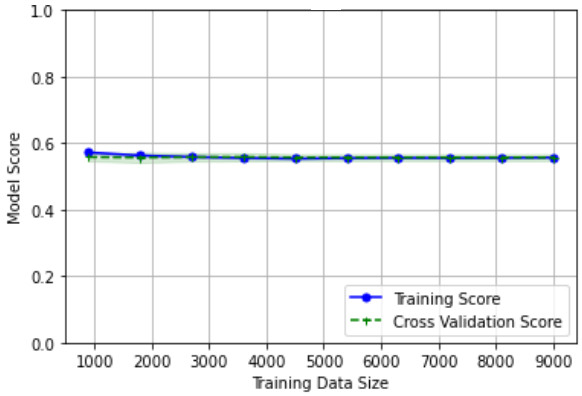
Class variable	Precision	Recall	F1-score	Support
Fail	0.52	0.36	0.42	302
Pass	0.68	0.82	0.74	1659
Pass with distinction	0.68	0.52	0.59	1068
Accuracy	66.78%			

(b) BN score based synthetic data DT learning curve, accuracy, precision, recall, F1-score



Class variable	Precision	Recall	F1-score	Support
Fail	1	1	1	302
Pass	1	1	1	1659
Pass with distinction	1	1	1	1068
Accuracy	100%			

(c) BN score based synthetic data LR learning curve, accuracy, precision, recall, F1-score



Class variable	Precision	Recall	F1-score	Support
Fail	0	0	0	302
Pass	0.6	0.78	0.68	1659
Pass with distinction	0.54	0.45	0.49	1068
Accuracy	58.31%			

Figure 3-21 Learning curves and confusion matrix for kNN, DT, and LR on score-based DAG synthetic tabular data.

In all three instances, it was easier for the models to predict the structure and relationship between the training data than the test data (Figure 3.23). None of the models were a best fit, however, the error rates did indicate that more complexity is required in the underlying data to fit the models best. Another observation are the sharp increases and decreases in the misclassification errors of kNN. The experiment's outcome could be attributed to the choice of hyperparameters for kNN. The selection of *k* was not thoroughly investigated in this study making it difficult to assess its optimal value. The authors, however, suggested that a different *k* value and different weighting between the class variables might have mitigated the issue observed in the experiment. A small *k* value can make the model sensitive to noise, while a large *k* can make it computationally expensive and potentially less accurate. The weighting scheme, on the other hand, determines how much influence each neighbour has on the prediction (Cunningham and Delany, 2021). Therefore, different combinations of *k* and weighting schemes can significantly impact the performance of kNN (Zhang *et al.*, 2017).

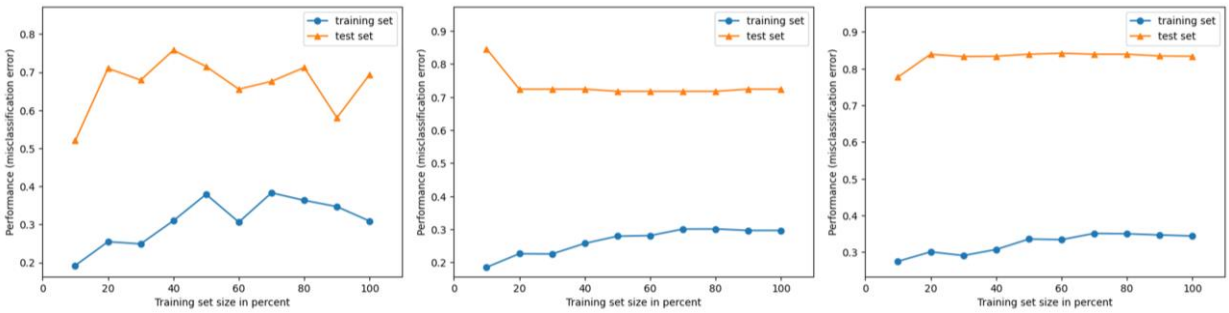


Figure 3-22 Overall performance of kNN, DT and LR with original data as training set and score-based DAG synthetic data as test set.

Based on the findings measuring utility, the best results were obtained from the raw, unprocessed original dataset for all three algorithms. The second highest were the experiments performed on the discrete raw dataset. The results from the GAN indicated that measuring utility using kNN and DT were prone to overfitting. An explanation for this is that the GAN was not optimised appropriately. It is important that in future studies, an emphasis is placed on the hyperparameter finetuning of a GAN to produce synthetic data that has a high utility. Using LR, for the GAN and both DAG generated synthetic data, the model underfit the data. On average, the models fit to DAG produced synthetic data (both constraint-based, and score-based) were prone to underfit the

data in all instances. Despite the structure learning for both the constraint-based, and score-based having a misrepresentation of causal relationships, the outcomes when the utility was measured still outperformed the GAN. Although not explicitly measured in this study, this observation is an indication that data with some degree of conditional independence can lead to a higher utility. For machine learning in Education this is useful because synthetic data can fast track innovation as none of the privacy and sample size issues are present when using this type of data. The utility is also a useful tool for machine learning classifiers, and large amounts of training data is therefore possible to synthetically generate using the approaches proposed in this thesis.

3.6 Conclusion

In this chapter, the utility of data was evaluated using classifiers, underscoring the importance of aligning the appropriate algorithm with the problem at hand in classification tasks. It was observed that binary classification models are not efficacious when dealing with tasks involving more than two class variables. The experiments demonstrated that synthetic data, generated using a GAN and a constraint-based DAG, yielded the lowest accuracy, precision, recall, and F1-scores when a classifier was employed to measure utility. Specifically, for data produced by the GAN, kNN and DT overfit the data, while LR underfit the data. The highest utility of synthetic data was observed with score-based methods. Based on these findings, it is recommended that synthetic tabular data, intended for educational tabular data, should make use of score-based DAGs with an increased number of variable classes and added complexity in the data. This approach is anticipated to address the underfitting challenges observed in the experiments. Furthermore, it is advised that the same hyperparameters and algorithms be employed between the original and synthetic data, provided that utility is being measured. Although GANs did not perform adequately, future research can benefit this area of exploration to yield better results. While the objective of this chapter was not to optimise the classification tasks used, the use of a classifier proved to be a valuable tool in measuring utility, given that the same parameters are used between testing the same class variables between different training data. In the next Chapter, the principles delineated in Chapter 3 will be applied to a real-world dataset from the University of the Free State, South Africa.

CHAPTER 4 REAL-WORLD SYNTHETIC DATA GENERATION: THE UNIVERSITY OF THE FREE STATE CASE STUDY

4.1 Introduction

The contribution of this chapter was to apply the context learned from Chapter 3 on a University of the Free State dataset, which has never been done before. Based on the insight gained from Chapter 3, a score based Bayesian Network (BN) was constructed to generate the synthetic tabular data. The utility of the synthetic data was measured using a k-Nearest Neighbours (kNN), Decision Tree (DT), and Logistic Regression (LR) algorithm. To apply the context learned from Chapter 3 on a real world dataset, the following objectives were set out:

1. Feature engineer continuous variables to discrete variables;
2. Construct a direct acyclic graph (DAG) and determine the parameters using a score based method; and
3. Measure and visualise the accuracy, precision, recall, and F1-scores of the real-world and synthetic data using a kNN, DT, and LR algorithm.

In the sections to follow, the context of this dataset and university, the results, and the findings of these will be outlined.

4.2 University of the Free State context

The University of the Free State (UFS) is situated in central South Africa. This university has three main campuses, a variety of smaller satellite campuses, and a student cohort of more than 33 000 in any given year (Combrink and Oosthuizen, 2020). There are seven primary faculties within the UFS. One of these is the faculty of Economic and Management Sciences (EMS), located across two campuses namely the Bloemfontein campus (BFN) and the QwaQwa (QQ) campus (Combrink and Oosthuizen, 2022). The EMS faculty offers degree programmes ranging from undergraduate to PhD in five primary disciplines, namely Industrial Psychology, Public Management, Business Management, Economics and Finance, and Accounting (Coetzee *et al.*,2021). The UFS has put in place several initiatives and interventions to assist student throughput and retention rates by means

of academic and non-academic support (Strydom *et al.,* 2010). Among these support systems are academic literacy support in the form of a unit for language design (ULD), student transition in the form of a first-year seminar (UFS101), academic tutorial programmes (A-Step), and Academic Advising (AA), which are all situated in the Centre for Teaching and Learning (CTL). AA offered in the CTL is performed on a consultation basis and includes time management advice, goal setting, as well as referral to appropriate student support initiatives within the institution should it be needed (Tiroyabone and Strydom, 2021). In addition, the UFS offers services in the form of counselling and medical services to students, should they require medical and psychological support. These interventions are also recommended from other support structures, such as creative writing, from writing centres, and academic literacy support (van Aardt, 2019). Furthermore, the UFS has a feeding scheme and programme to support food insecure students (Ruswa and Gore, 2022). Within faculties, academic support in the form of the academics themselves (in-person academic consultations) as well as curriculum-specific academic advice to track credit load is offered (Schoeman *et al.,* 2021). Although there are other forms of support to students than what is offered by student academic services and the support from specific residences within the institution, these are for specific sub-groups of students within the institution and do not apply to all students. To contextualise the interventions available at the UFS for all students, the support available to all students at the UFS can be broadly summarised to represent the aforementioned interventions (Table 4.1).

Table 4-1 Types of support available to all students at the UFS.

Type of support	**Name of intervention**
Academic (assisting in understanding concepts, improve learning, improve student engagement)	First year seminar, Orientation, ULD, A-Step, AA, academic consultations, faculty-specific academic advising
Non-academic (assisting non-academic needs that influence the students' journey)	First year seminar, Orientation, AA, academic consultations, University feeding schemes, medical services, student counselling services

4.2.1 Impact of interventions on student success

Several existing studies assess the impact certain interventions have on student success (Kinkle 2020; Baus *et al.*, 2021; Kreth *et al.*, 2021; Stone, 2021; Eudy and Brooks, 2022). Although these results are bound by specific contexts, and the implementation of these interventions will differ depending on the training and the type of person implementing the intervention, for the purposes of this section the assumption is made that these interventions are universal and universally effective. It is not to say that these interventions will or will not work, and the focus and emphasis of this chapter is not in assessing the extent to which these interventions work. Drawing on the continued extent of implementing complexity theory within the context of a systems approach, the focus of this section outlines how well synthetic data can be generated from a real-world South African context. Any data from any higher education institution could have been used to prove the concepts outlined in generating synthetic tabular data. The transfer of insights across institutions may not exhibit strong generalisability, but the capacity to create synthetic data from higher education establishments will perform effectively within the established framework. From this perspective, once synthetic data can be generated with a high degree of utility, it can then be used to contextualise which students require an intervention, and to what extent an intervention may assist the student population. In other words, the data from a specific institution is used and synthetic data generated from that context so that ample amounts of information may be generated for further exploration, measurements, testing, and model development.

Consequently, if researchers do not know the impact of a specific intervention to an education outcome, or what the contribution of a specific intervention is in the context of student success, then no system will be able to infer a better outcome. For example, if a system recommends a tutorial to a medical emergency, and the impact of seeing a clinician to academic performance is not known, then the outcome of the recommendation will not be useful to the student because this connection and association is not known in the domain. Drawing from the insights gained from Chapter 3, two assumptions can be made that the raw data that is used represents a series of linked systems (indicating that a DAG created from this type of data will represent a complex system) and that the synthetic data might not yield as high a utility as the raw data (depending on the sample size of each). For example, considering the context from Chapter 2, the academic performance of a student, which is a percentage, is a representation of a variety of different datapoints, variables,

and contexts, aggregated into the overall performance for a student. As such, when a DAG is created in an education context, each variable represents a variety of datapoints that may not be collected that contribute toward that variable (such as student support, prior learning, motivation, food security etc. on academic performance). Given that the system needs to use real-world data to create a synthetic data representation on which the systems may be built, there is a need to outline the time associated with the intervention (see Chapter 2, section 2.9). The time associated with the intervention refers to the duration used to collect the datapoints that inform the intervention and the timing of the intervention recommendation itself. To this extent, the desired outcome measured is if the qualification was obtained. For this reason, the purpose of this chapter is to apply the framework to a real-world dataset and evaluate the results in terms of whether or not a qualification obtained can be predicted. If the qualification obtained can be predicted, then those who will not obtain their qualification would thus require an intervention. The challenge resides in knowing if there are appropriate interventions for each use case, and how to identify them. These are not discussed in this thesis, but it important to consider for future research. In the next section, the methodologies used to outline this process will be discussed.

4.3 Methods

4.3.1 Research design

The experiments conducted in this section are based on a real-world dataset, and concepts from Chapters 2 and 3 are applied to the dataset including feature engineering on the real-world data, synthetic data generated from the real-world data, the utility of the dataset evaluated using a classification task based on the parameters of the real-world dataset (see Chapter 2, section 2.9, figure 2.5, see Chapter 3 sections 3.3.3, 3.3.5). These concepts include taking the real-world (primary) dataset, which is from the University of the Free State, constructing a BN from the data using a score-based method, then generating synthetic data (secondary dataset) using this method, and then evaluating the utility using three different classifiers to test the results. The total sample of synthetic data was 100 000 students. The results were illustrated in terms of the precision, recall, F1-scores and overall accuracy, and the learning rates were illustrated using a learning curve.

For the context of this dataset, the timeframe of the intervention would then represent data collected at specific points during a student's undergraduate journey, and the intervention would

thus be implemented prior to their final semester. If this is the case, then the intervention would have to be created at this stage. It must be emphasised that the abstraction is that some data are associated with a certain time interval, and that there needs to be an appropriate intervention based on this. Not all interventions, types of data, and time intervals will be similar. In other words, the purpose of the subsequent experiments in this particular study was to create synthetic data. This data was derived from variables representing various systems, which ultimately determined whether a student achieved their qualification or not. This also means that if such data can be synthetically generated, then it will be the training data for an intervention at the last stage of a student, prior to obtaining their qualification. Any student intervention requires a contextual approach to outline the time, data, and interventions needed for that specific set of students that fall within a particular set of categories. In the next section, the ethics, data, and evaluations used will be outlined.

4.3.2 Ethical clearance

Ethical clearance was obtained from both the University of Pretoria (UP) (ethics number EBIT/19/2022) and the University of the Free State (UFS) (ethics number UFS-HSD2022/0195/22). All of the data handling and procedures executed on this chapter complied with the legal and ethical data handling procedures outlined by the study protocol (**Appendix A** and **Appendix B**).

4.3.3 Data

Each of the variables within the dataset represents a complex set of processes to obtain the variables. For example, each variable will be a cumulative number, but contributes towards this variable will not be used, like an AP score, which represents the cumulation of all school marks into a single number. These variables are used in learning and educational analytics to make predictions on students (Janse van Vuuren, 2020). Due to the complexities associated with student success and what these marks represent, it is important to note that each of these variables represent a system, and the combination of them represents a complex system which we are abstracting to the most common eight variables used in student prediction. The class variable, in this instance representing if a qualification was obtained, will be the target variable used to measure the utility.

The data used in the creation and evaluation of the synthetic data as well as the context from which the simulations were derived, were data obtained from EMS faculty for both the BFN and QQ campuses from 2012 – 2021. It spanned all degree programmes for all departments within the faculty. Included in the dataset were variables (Table 4.2) related to whether the qualification was obtained as well as the specific department (Table 4.3).

Table 4-2 Variables from the real-world dataset.

Name of variable	**Number of discrete categories**
AP Score	3
English Mark	6
Department Code	5
1st Year Credits Enrolled	4
1st Year Credits Obtained	4
3rd Year Credits Enrolled	16
3rd Year Credits Obtained	16
Qualification obtained	2

The variables were chosen based on literature found in studies that outline important academic variables used in student classification tasks and student prediction models (Imran *et al.*, 2019; Lau *et al.*, 2019; Alyahyan and Düştegör, 2020; Coussement *et al.*, 2020; Namoun and Alshanqiti, 2020; Zeineddine *et al.*, 2021; Yağcı, 2022). The AP score is a prior to university admission score that is calculated from the final marks of a student when they leave high school (Bengesai and Pocock, 2021). The English mark is the final English mark students obtained when they left high school (Rauchas, *et al.*, 2006). The department code indicates which department the student forms a part of in the context of EMS. For the first year, and the third year, the number of credits the student enrolled for, and the number of credits the student passed at the end of that year was factored in. The data obtained represents 3 year degrees. These degrees require a minimum of three years to complete. The second year credits were omitted as there are contexts such as students transferring between faculties and universities, changing of degree programmes and a different

credit load to the first and third year of study. Each subject within a degree programme has a certain number of credits associated with it, and a certain number of credits need to be passed in order for a student to progress through the academic journey. Finally, if the qualification was obtained or not was indicated as a binary variable.

Table 4-3 Qualification obtained per department.

Department	Qualification not obtained	Qualification obtained	Total
Industrial psychology	1425	468	1893
Public administration	135	360	2946
Economic and management sciences	2723	223	495
Business management	576	1289	1865
Accounting	688	1007	1695

4.3.4 Evaluation

A synthetic dataset was created and evaluated for the BN (see Chapter 3, sections 3.2.3.1 – 3.2.3.6). The same hyperparameters as discussed in Chapter 3 was on the datasets. However, different training and test datasets were evaluated in different combinations between the real-world dataset and the synthetic dataset (see Chapter 3, sections 3.5.2 – 3.5.7). To this extent, the training and the test data differed between the experiments (Table 4.4).

Table 4-4 Different training and test data used on real-world dataset.

Name of experiment	Name of training dataset	Name of testing dataset
Accuracy of the raw discrete data	Raw discrete dataset	Raw discrete dataset
Accuracy of the synthetic discrete data	Synthetic discrete dataset	Synthetic discrete dataset
Accuracy of different training and test data	Raw discrete dataset	Synthetic discrete dataset

Both the learning curves and misclassification error rates will be reported on and interpreted to understand the learning rate and model performance for the various training and testing dataset pairs. In the next section, the results and discussion will be outlined on the real-world data.

4.4 Results and discussion

The DAG generated based on the score based method indicated that qualification obtained relied on three variables, namely: department code, 3rd year credits enrolled for, and 3rd year credits obtained (Figure 4.1). Furthermore, English mark and AP score are associated to one another. The structure learned using the score based method for these variables were sequentially correct. In other words, a student needed to pass their first year credits before their third year, and their third year before they obtained their qualification. The two associations that were important to note is that English mark and AP score is associated with the third year credits enrolled for indicating that AP score and English mark have an association with the third year credits, and not the association with first year as indicated in certain studies (Zeineddine *et al.*, 2021; Thompson *et al.*, 2022).

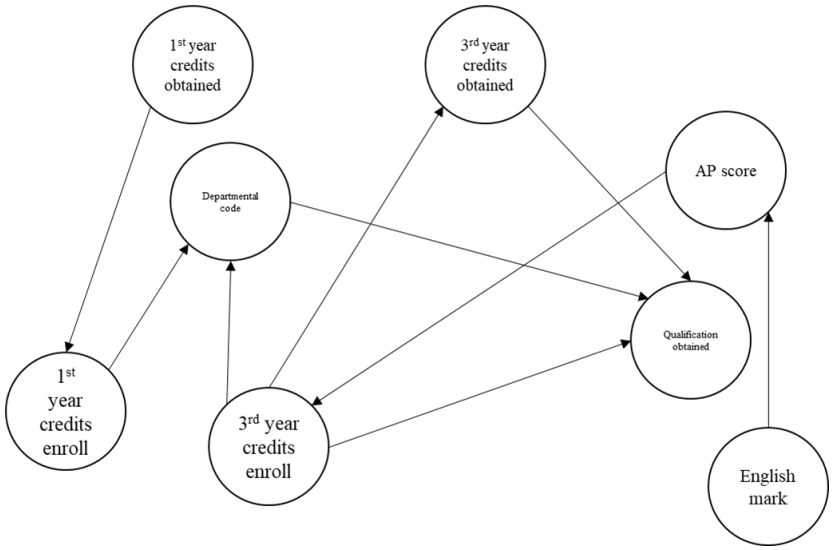


Figure 4-1 DAG of the discrete education dataset used from the UFS.

Another observation based on the DAG learned from the data was that the department code was directly associated in the structure of the DAG to qualification obtained. In other words, a high AP score and low English mark, as well as low credit score and high credit load for one academic

department might not have the same probability and might not serve as the same indicator between different departments. This does imply that the department a student is a part of makes a difference to the association between the variables. This observation has been noted in several studies that have indicated that the specific degree programme people study towards have nuanced differences in terms of predicting the outcome of student success (Beaulac and Rosenthal, 2019; Canning *et al.*, 2019; Alsariera *et al.*, 2022; Baashar *et al.*, 2022; Pallathadka *et al.*, 2023). There was an incorrect association on the basis of the structure namely first year credits obtained. The association was not detected based on a specific metric, but rather that a student enrols for credits prior to obtaining them. In the instance of the data, both the credits enrolled for and obtained were the same because a student could not progress to their final year of study, if all their first year credits were not obtained. In reality, the first year credits enrolled for will be placed prior to the credits obtained, and this misplacement was identified on the DAG itself.

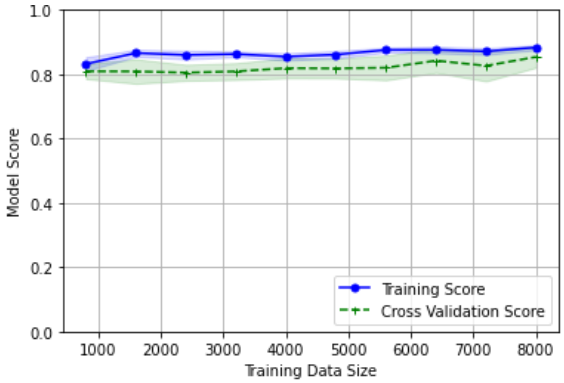
A possible explanation for this is in the underlying data, containing the same credits for both these variables. To put it differently, for a student to have made it to their third year, their first year credits would be passed. In future studies, variable associations like this should not be included in the data as it might skew the results. Next, the classification tasks were performed on the raw discrete data to determine the utility (Figure 4.2). Based on the results kNN (overall accuracy = 90.22%), DT (overall accuracy = 100.00%), and LR (overall accuracy = 85.92%) were better at predicting if a student obtained their qualification, than a student who did not when comparing the precision (kNN = 90.00%, DT = 100.00%, LR = 88.00%), recall (kNN = 98.00%, DT = 100.00%, LR = 95.00%), and F1-scores (kNN = 94.00%, DT = 100.00%, LR = 91.00%). The DT overfit the data, which was not an ideal outcome despite the high prediction value. Of the three algorithms, kNN and LR had high overall accuracy scores associated with the information. Next, experiments performed on the synthetic data measuring the utility indicated that DT (overall accuracy = 100.00%) overfit the data (Figure 4.3). In these experiments, kNN (overall accuracy = 80.24%), and LR (overall accuracy = 70.28%) learned their context after approximately 25 000 students. This means that the overall precision (kNN = 82.00%, LR = 73.00%), recall (kNN = 93.00%, LR = 92.00%), and F1-score (kNN = 87.00%, LR = 81.00%) for class variable 1 remained the same and consistent after 25 000 student observations. The same can be said for the precision (kNN = 75.00%, LR = 53.00%), recall (kNN = 52.00%, LR = 20.00%), and F1-scores (kNN = 62.00%, LR

= 29.00%) for class variable 2. Precision as a measurement indicates the relative fractions of the data where the retrieved classifications are relevant, indicating that more class 1 variables were relevant based on the classification. Recall on the other hand is an indication that the class 2 variables had a higher error rate in terms of the classifications, especially for LR. This is an interesting observation as LR is a binary classifier, and considering that only two class variables were used, LR performed the poorest in the classification task. In these experiments, DT overfit the data and unfortunately overfit data will struggle to fit new data with the same variables. In the context of complex systems, the synthetic data and utility for all of the score based synthetic data were higher than the previous results (see Chapter 3, section 3.5).

The more complex dataset was achieved with an increased sample size of the training data, and more categories within the variables used were included within the raw discrete tabular data, as compared to the datasets used in Chapter 3 to construct the BN. The training data was increased to 100 000 samples. The increased dataset could show the impact of the training on these sample sizes and illustrated when the models sufficiently learned from this data. Despite these increases, DT still overfit the data. While increasing the sample size generally improves the model's ability to learn, it can also increase the risk of overfitting, especially if the data includes noise or outliers. The DT might have learned these unnecessary details too well. Including more categories within the variables can increase the complexity of the model. DTs can create very complex decision boundaries, which can lead to overfitting if not properly controlled. The increased complexity of the dataset, with more categories and larger sample size, can make the model more prone to overfitting. The model might have captured the noise in the data instead of the actual patterns. To overcome these issues, different algorithms may be used and fine-tuned to improve the algorithm on the underlying data. Lastly, experiments were performed using the real-world raw discrete synthetic dataset as the training data, and the synthetic data as the test data. All of the experiments had an overall miscalculation error between 30% - 70% for all three algorithms used. A few observations indicated different misclassification errors at different stages of the training and testing data (Figure 4.4). The first is the difference observed between the training score and the test score. The training score performed better over time since this real world data retained a context within the variables from which the model could learn over time. In other words, as the training sample size increased, so too did the misclassification error rate decrease. Unfortunately,

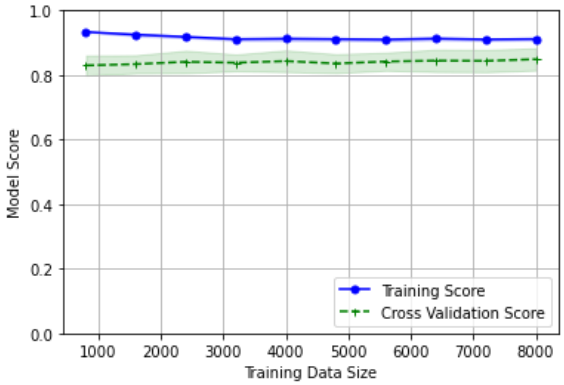
the test misclassification error rate did not improve over time, neither did it decrease. Both the training and testing score were within 10% of one another, which was much less than observer in Chapter 3, where an average difference of 30% was observed between the training and the testing misclassification errors. This is an indication that the models could learn context from the raw and synthetically generated tabular data with the more complex dataset, better than the simple datasets used in Chapter 3.

(a) kNN learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



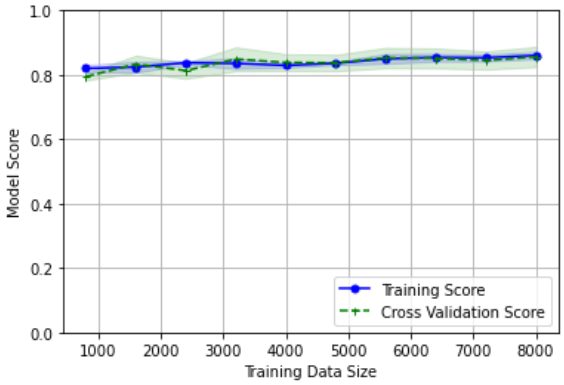
Class variable	Precision	Recall	F1-score	Support
Yes	0.89	0.96	0.92	6804
No	0.81	0.62	0.71	2090
Accuracy	87.77%			

(b) DT learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Yes	0.91	0.98	0.94	6804
No	0.92	0.68	0.78	2090
Accuracy	91.02%			

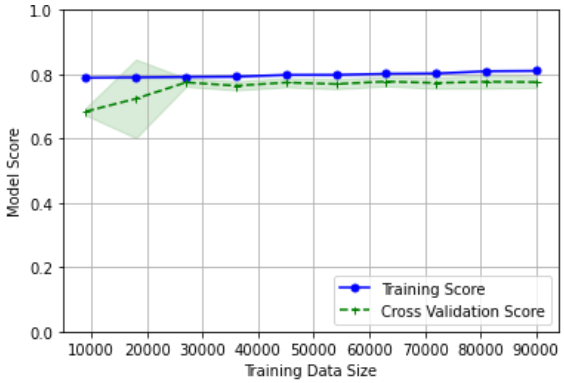
(c) LR learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Yes	0.88	0.95	0.91	6804
No	0.79	0.58	0.67	2090
Accuracy	86.41%			

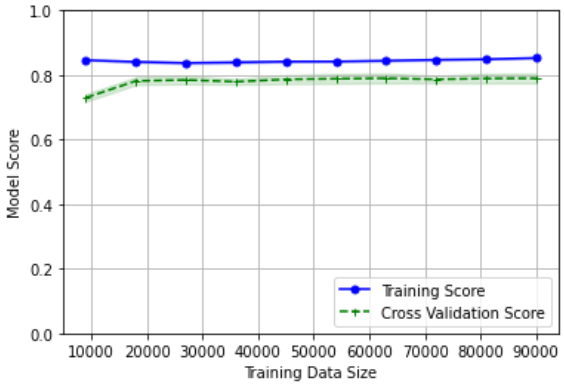
Figure 4-2 Learning curves for kNN, DT, and LR on raw discrete dataset.

(a) kNN learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



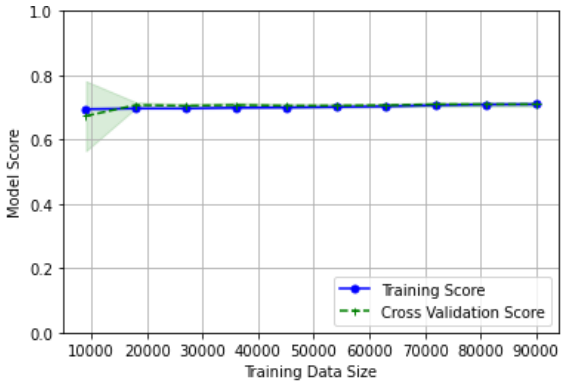
Class variable	Precision	Recall	F1-score	Support
Yes	0.73	0.92	0.82	69656
No	0.56	0.22	0.31	30344
Accuracy	81.09%			

(b) DT learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Yes	0.86	0.94	0.9	69656
No	0.83	0.64	0.72	30344
Accuracy	85.19%			

(c) LR learning curve, accuracy, precision, recall, F1-score on BN generated synthetic tabular data



Class variable	Precision	Recall	F1-score	Support
Yes	0.73	0.92	0.82	69656
No	0.56	0.22	0.31	30344
Accuracy	70.98%			

Figure 4-3 Learning curves for kNN, DT, and LR on synthetic score based discrete dataset.

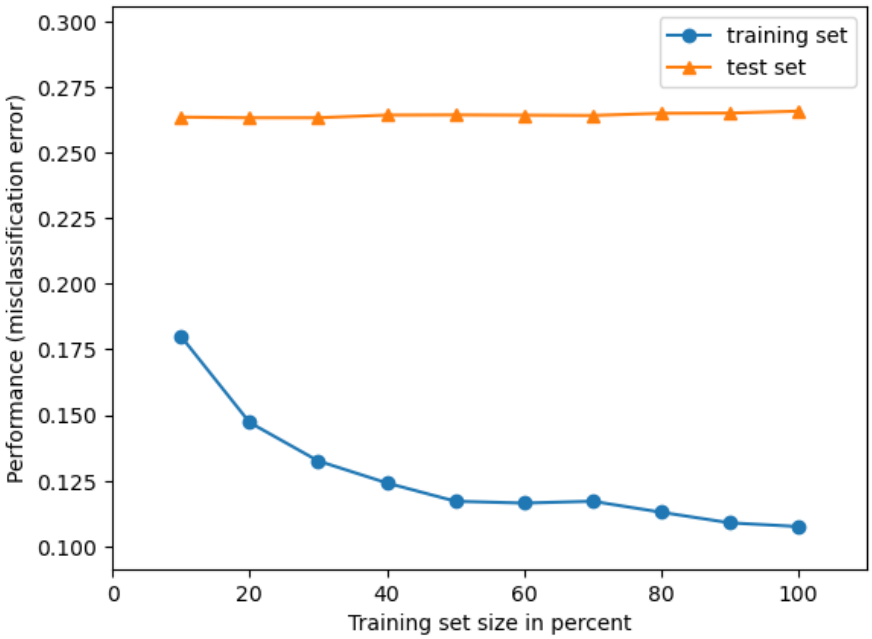


Figure 4-4 Misclassification errors between raw training data and synthetic test data for kNN.

Within the framework of DT that overfitted the data as per the learning curve, a comparable issue was noticed between the raw discrete tabular data for training and the synthetic data used for testing (Figure 4.5).

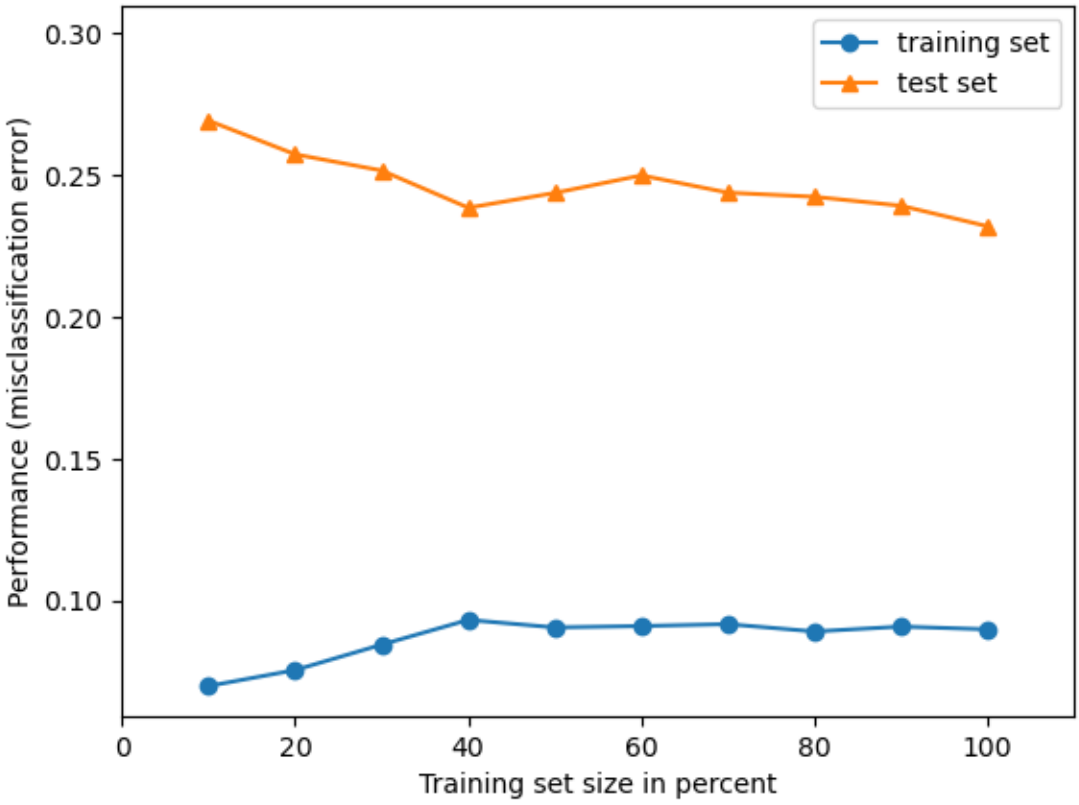


Figure 4-5 Misclassification errors between raw training data and synthetic test data for DT.

For the LR, a similar pattern was observed as the kNN algorithm, but at a much lower threshold (Figure 4.6). Even though the training data accuracy increased over the duration of the dataset, the test data's performance decreased by an observable increase in misclassification errors over time. Patterns like this has been observed in literature with large training datasets where the quality of the underlying data may be improved (Van Beek and Hoffmann, 2015).

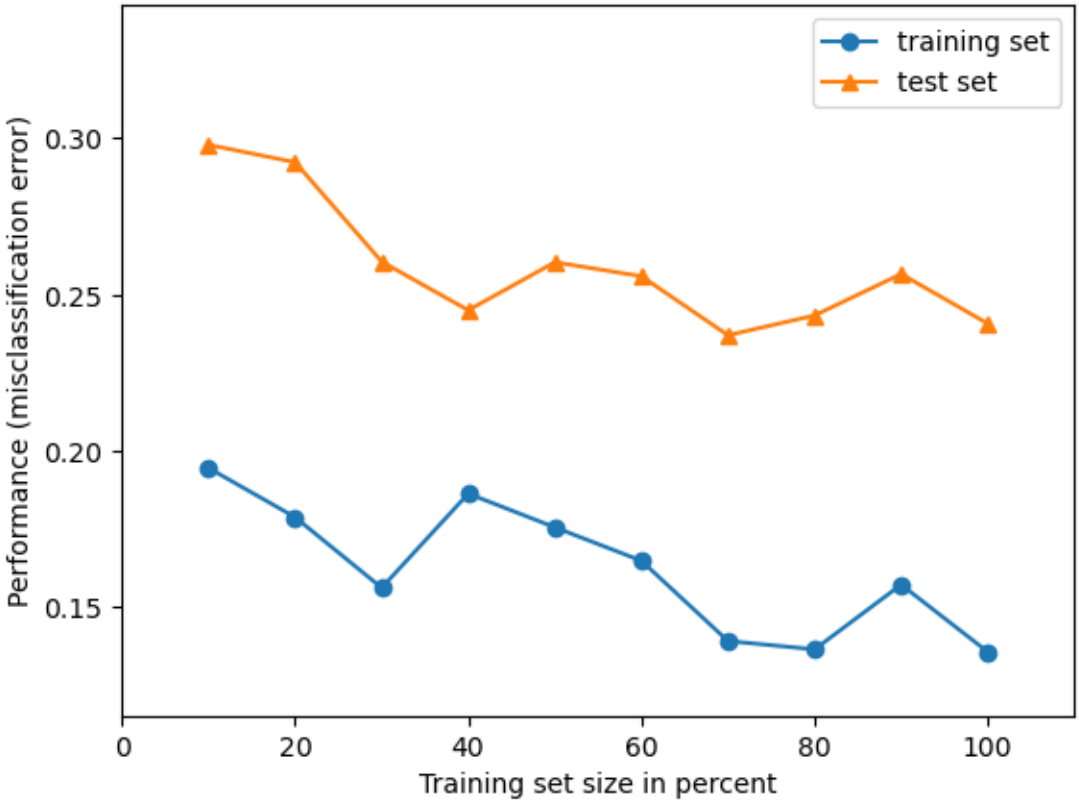


Figure 4-6 Misclassification errors between raw training data and synthetic test data for LR.

An overarching observation is that the score based DAG with parameters learned from data did produce data with an overall accuracy greater than 70% for all three algorithms. There was therefore a translatability in terms of the utility between the variables in the underlying raw data, and the ML algorithms could learn the context from the underlying raw real-world data and apply it to synthetic data.

4.5 Conclusion

Generating synthetic data is important for models that require large datasets. The performance of the learning curves between synthetic and real-world data illustrated that synthetic data can be used to simulate a real world context from a probabilistic perspective. Although the fine tuning of this approach was not explicitly explored, the application of the context and methodology as outlined in Chapter 3, scaled to the data used in Chapter 4. This chapter further illustrated that a score based structured learning methodology can be applied on real-world higher education data

to produce synthetic data that is similar to the original data. This was observed by the similarities in the performance between the different experiments.

As illustrated in this chapter, BNs are useful tools to generate synthetic data for the education context. Although not explored in this study, are the important roles and associations that a BN can make in terms of inference and contextual understanding within a complex system based on the context learned from data. In terms of a complex system, and the implementation of the framework, it is recommended that the right procedure for generating synthetic data is used.

Currently (2023) there are no specific guidelines for the generic generation of synthetic higher education data, albeit that there are institutions that share deidentified real-world data. This context allows for the development of such guidelines that are in compliance with POPIA, GDPR, and institutional policies so that an infinite amount of synthetic higher education data may be generated to advance machine learning in education research (MLER). Furthermore, it is vital that the synthetic data is created from an original dataset representing that context to ensure that the correct context is captured in the associations between the variables. In the context of a complex system, the BN can provide useful ways to generate synthetic data representing the complex system, but it is noted that a BN accounts for discrete variables, and there is still a need to investigate models of producing synthetic data that account for continuous variables within tabular higher education data. As computational power increases so too can the complexity of these models, and the data they represent within a system.

The significance of these results illustrates the importance of being able to represent a complex education system into a probabilistic model, like a BN because a BN can generate an infinite amount of synthetic tabular data for the education context and be used to probabilistically quantify inferences based on the network. Inference was not discussed in this thesis, but inference is an important feature of a BN, and can assist in institutional decision making. Inference allows for the probabilities of anomalies within an education system, and a BN can assist education decision makers with understanding what influences particular parts of the education system that are fundamentally complex to model. Another important observation is that the methods and use case described in this chapter is repeatable to other faculties within the same institution as well as to

other education (basic, and higher education) institutions. For the latter to work, the data should be used from that specific region or context. What this allows for is a probabilistic model that contextualises the context and the probabilities associated with students within that context. In the next chapter, how contexts that intervene with students can learn over time from a complex system perspective will be performed.

CHAPTER 5 A MULTI-ARMED BANDIT APPROACH TO AUTONOMOUS LEARNING SYSTEMS IN EDUCATION

5.1 Introduction

The contribution of this chapter was to simulate the education recommendation context as a multi-armed bandit problem, which has not been done prior for the education domain. This contribution is justified by a need to conceptualise a decision maker that can learn from data to provide student recommendations in education, across different contexts. The aim of this chapter was to investigate the application of student recommendations in the education context by abstracting education recommendations into a multi-armed bandit problem. To address this aim, the following objectives were set out:

1. To design and develop an environment simulation that mimics the complexities of systemic interventions as a multi-armed bandit problem;
2. To compare the performance of different algorithms commonly used in multi-armed bandit problems in terms of its reward functions over the number of episodes per simulation;
3. To illustrate different student recommendation reward functions over the number of episodes of the agents learning rate; and
4. To identify and discuss the challenges and limitations of using a multi-armed bandit problem in the education context using the aforementioned approach.

Different multi-armed bandit algorithms were implemented and evaluated within the simulation to optimise the cumulative reward for the agent. The work in this chapter came from the publication by Combrink *et al.*, 2022b, specific to multi-armed bandits in education. The results obtained from the experiments were analysed to determine the effectiveness of the recommended interventions; the latter's impact on the implications for student performance, given the intervention was also assessed.

5.2 Reference to institutional support needs

In the education context, and specifically in higher education, student data across institutions are kept in computer readable formats, typically in the form of tabular data (see Chapter 2, section 2.2.1). Such data comprise student demographics, their learning materials, academic success, attendance, as well as analytics related to their behaviour. As outlined in Chapter 2, there are a myriad of challenges associated with student transition in South African higher education, leaving higher education institutions to face student-related socio-economic problems that impact their academic progress and academic achievement. Unlike findings presented in the literature related to first-world countries, the gravity and scale of the challenges faced by the average South African student range from literacy and English comprehension gaps to food insecurity and abject poverty (Dominguez-Whitehead and Sing, 2015; Mirata *et al.*, 2020). Unfortunately, these challenges affect students' academic success, and ignoring these problems do not serve students, academic faculty staff, academic support endeavours, academic programmes, or the institutions facing these challenges. Due to the scale of the problem, the staff compliment of higher education institutions alone is insufficient to identify and implement the interventions required to address all the challenges students face. Several strategies have been explored in recommending interventions for students (see Chapter 2, section 2.2.2), including recommender systems, discussed in the section that follows.

5.2.1 Recommender systems

A recommender system is a dynamic machine learning algorithm that uses data and context to make recommendations within a specific system (Quijano-Sánchez *et al.*, 2020; Wu *et al.*, 2022). A recommender system can be described as an information filtering system for users, based on their preferences, needs, and/or contexts. At the core of it, a recommender system is implemented in instances where there are a lot of different items, contexts, products etc. to choose from. With a recommender system, the user has the ability to find a specific item that they are searching for faster than, for example, having to read through the entire inventory of items to choose from. As mentioned, finding relevant information is the task at hand, and this context can be applied to physical and digital contexts such as search engines, content recommendation and/or product recommendations, to name a few. In the context of educational interventions, the student is positioned as the user, while the item could be an educational tool or intervention that has the

potential to assist a student's academic success. Examples of such systems include collaborative filtering, which considers user context, and content-based filtering, which focuses on item context (Srifi *et al.*, 2020). Collaborative filtering relies on knowledge about the user input, or what users find relevant, to recommend interventions. The drawback of this approach is that such systems may introduce conformation bias because the recommendation is generic and generalised. This means that recommendations for users who require nuanced or unique items may not be recommended. It may also introduce bias by only recommending a set of specific items, resulting in a skewed recommendation system. This means that collaborative filtering looks for similarities between users and bases the recommendation on features shared between users. For the education context, one way to view collaborative filtering is that once different categories of students, or risk profiles of students are established, common features between these risk categories may be grouped so that recommendations may be made on the basis of shared features between users. Content-based filtering, on the other hand, emphasises the specific user's experience. In the context of education, content-based filtering could be a system that collects data on how students experience certain recommendations and as such, records those findings and uses this content for recommendations. However, the recommendations garnered from a system that treats all users as identical may not scale well in diverse settings where students, cultures, and contexts differ, of which South Africa is but one example (Nassar *et al.*, 2020). As a result, a hybrid approach that combines elements from both methods has been explored to determine whether it may feasibly provide more effective recommendations (Khanal *et al.*, 2020). Recommender systems have been a successful tool for making recommendations in society at large (Milano *et al.*, 2020). However, in the context of very specific use cases, such as recommending an explicit educational intervention to a student, these systems have been far less successful due to some of the systems' fundamental problems. These include, first of all, the cold start problem, where no information about users or items is available in the system when it starts for the first time (Natarajan *et al.*, 2020). A second issue is data sparsity, as not all relevant information about users or items may be collected (Batmaz *et al.*, 2019). Scalability, where the system needs to cope with a growing number of students and the supporting interventions, is another problem. Furthermore, if there is a change in recommendation, where the same problem might require a different intervention between two students due to a variety of social and demographic factors, it might also cause issues as the right features between the students may not be captured to best fit the problem (Milano *et al.*, 2020).

Finally, a lack of data, which can hinder effective recommendations, in addition to changes in user preferences over time, are problematic (Figure 5.1).

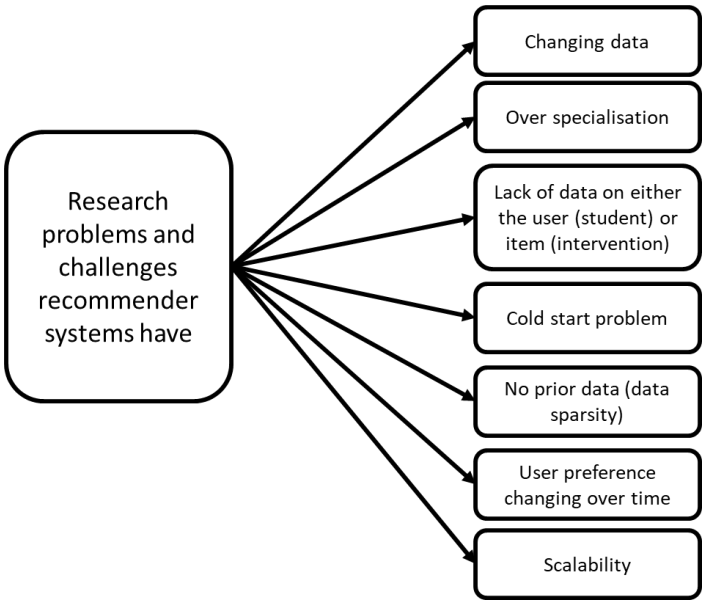


Figure 5-1 Common research problems associated with recommender systems[2].

The cold start problem occurs when there is no prior information about a user or item available and the system wants to make a recommendation. This problem can also occur when a system is initialised or starting up for the first time. The cold start problem is prevalent in the education setting, as it may occur when a new student enters the system by transferring from another institution or changing from a specific degree programme in one faculty to another degree programme in a different faculty or when a novel student support strategy is being implemented by an institution. Based on where the data are missing in the system, three types of cold start problems generally need to be considered (Lu *et al.*, 2020). The first is the entry of a new user with no previously existing data into the system (data sparsity), and the second comprises entering a new item into the system. The third relates to the system being started for this first time if there are no prior data collected about the student and/or problem (essentially, initialisation of the system). Data sparsity refers to instances where not all relevant information about users (students) is collected within the system. The sparsity of data can result in mislabelling of interventions

[2] source: https://iopscience.iop.org/article/10.1088/1742-6596/1717/1/012002

(items) in the education context. This poses challenges for accurate recommendations. Scalability is also a major concern for recommender systems; increasing numbers of students and interventions may challenge the system's ability to fit the right recommendation to the correct user. Studies have shown that, when a certain threshold or sample size is reached (depending on the algorithm and system), the results may not be desirable, as there is a tendency to inertia with the recommendations (Kiran *et al.*, 2020). For example, while the overall accuracy of a recommender system increases with a larger training sample, the system will eventually only recommend specific items based on specific features, and if not, new data are collected about a user or item, then this too can cause a problem within the recommendations. Probably the most difficult challenge to deal with is a change in user preferences or needs over time. An intervention that might have been successful for a student in year one of study might be redundant in the student's following study year. Students' needs, and the interventions required to address these, may change over time, and the system should therefore be able to adapt and support these changes to provide relevant recommendations. As such, a different approach is needed to solve this problem. As outlined in Chapter 3 (section 3.2.1), supervised learning requires prior data to solve a problem, like a classification task. However, as explained above, data about recommending an intervention to a particular student might not always be available. It is important to note that although a recommender system might be directly trained on real institutional data, all departments might not have enough information to effectively run the recommender system, hence the need for synthetic data to do batch learning. Despite this, there are still challenges that recommender systems pose as dynamic filtering systems as outlined prior, and in such cases, reinforcement learning might provide a solution (Chiu *et al.*, 2021). A fundamental distinction exists between the subcategories in ML whereby ML can be categorised as either unsupervised learning, supervised learning or reinforcement learning. Reinforcement learning involves an autonomous decision-making agent – in the form of an algorithm – that makes choices based on actions and receives rewards based on those choices within a specific environment (Alloghani *et al.*, 2020). This approach is commonly referred to as reinforcement learning, which employs principles akin to classical conditioning, but which is applied within a decision-making framework.

5.2.2 Reinforcement learning

Reinforcement learning considers an interaction between an agent and a preexisting environment (Sutton and Barto, 1999). An agent navigates through the environment by taking actions and learns from rewards that arise based on the action the agent takes and the effects that they have on the environment. Additionally, the configuration of the current environment is also known as a state. The agent takes an action, and once the action is taken, the new state is initiated. The purpose of a reinforcement learning algorithm is to learn a policy which is mapped from states to actions, to maximise the cumulative rewards (Figure 5.2).

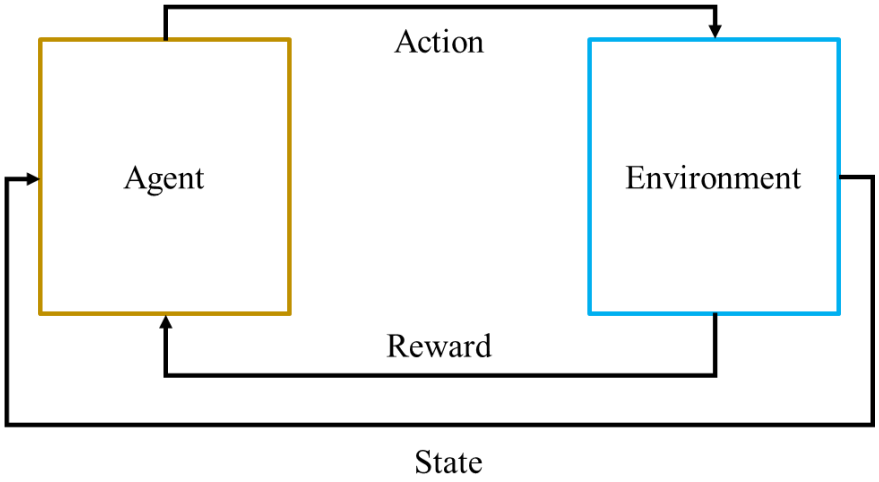


Figure 5-2 Overview of a traditional reinforcement learning algorithm (Sutton and Barto, 1999).

Seen in a machine learning context, the above-mentioned model allows for reinforcement learning that enables autonomous decision-making to maximise a reward within a system (Shin *et al.*, 2019). Over time, the policy converges only if the correct learning algorithm for the agent is used. As outlined before, reinforcement learning can be considered a viable replacement for traditional recommender systems, as the former comprises an autonomous decision-maker learning how to better perform its actions over time. According to Sutton and Barto, (1999), this type of autonomous decision-making can be represented as a tuple denoted by $\mu = (S, A, T, R, \gamma)$, where:

- S is a set of all the states;
- A is a set of all the possible actions that can be taken by the agent;

- T denotes the transition from one state to another and can be shown as the transition from S_t to S_{t+1} given a certain A so that T: $S_t * A * S_{t+1} \rightarrow [0, 1]$ from state 0 to 1, or from state s to state s'. This can further be shown as a function of $T(s, a, s')$ representing the probability of an A leading from state s to state s', given a specific A;
- R is the reward function, $R(s, a, s')$, for the same state action pairs to show the transition between states leading to a specific reward; and
- γ is the discount factor, bounded such that $\gamma \in [0, 1]$. The discount factor is denoted by γ, which determines the present value that is used for future rewards. A value of 0 makes the agent "myopic" (only considering immediate rewards), while a value close to 1 makes the agent consider future rewards more heavily.

A reinforcement learning problem can be formulated using a Markov decision process (MDP). An action-value is what can be expected as a reward for taking a certain action within the MDP, and this is also referred to as the value of a state-action pair (Sutton and Barto, 1999). The value of a given state is equal to the action with the highest value of the reward for an optimal action in a specific state, with the included discount factor which is multiplied by the next states value from the Bellman Equation and can be written as (equation 4.1):

$$V(s) = \sum_{a \in R} \pi(s, a) R(s, a) + \gamma \sum_{s\prime} T\left(s' \middle| s, \pi(s)\right) V(s'), \tag{4.1}$$

In this equation, $V(s)$ is the specific value associated to a specific state, and $V(s')$ the value associated with the following state, given an action a. The notation $T\left(s' \middle| s, \pi(s)\right)$ is the transition probability. It denotes the probability of transitioning from state s' to the current state s and taking an action $\pi(s)$. In the context of $V(s')$, this is the value added to state s'. The specific reward function $R(s, a, s')$ represents the reward given after action a is taken in state s. The probability of taking a specific action based on certain states is called a policy, and is denoted by π. The Bellman Equation (Sutton and Barto, 1999), and can thus be written as (equation 4.2):

$$Q^{\pi}(s,a) = R(s,a,s') + \gamma \sum_{s' \in S} T\left(s' \middle| s, \pi(s)\right) V^{\pi}(s'). \tag{4.2}$$

The Bellman Equation is required in dynamic programming to solve a MDP problem. As noted earlier, reinforcement learning solves the problem of finding the overall optimal reward function, so that there is a maximum value of reward given the state action pairs. In other words, reinforcement learning allows the system to choose an *a*, where *R* is given per step. This in turn provides the return *G*, a sum of the discounted *R* per episode to maximise the cumulative reward based on the highest *R* for each episode. However, if we consider the forms of data described in this dissertation, which is primarily tabular data, the traditional student recommendation problem becomes difficult to abstract. For example, if a describes a recommendation intervention, and each student is described as a s, there will be too many actions to navigate through given each student. As such, a novel approach not clearly defined in literature is required.

To this extent, if the reinforcement learning problem is applied to education, we can define the environment as follows. The *R* is the reward an agent needs to take within a given environment that can represent a utility function; in this case, it is the recommendation function. Instead of focusing on a specific intervention such as a tutorial, a set of learning material, or a context specific intervention, *a* should represent a type of recommendation action within a particular context. That way, the focus is on the framework of the reinforcement learning instead of on the impact of the specific intervention. Next, instead of representing a student as an *a*, students represent a s, with each student corresponds to a single state. The reward function can be designed so that the rewards represent recommendations that were successful in a specific context, were partially successful, or unsuccessful. To clarify, in our simulation, we won't be examining the effects of specific interventions, such as recommending a student to attend a tutorial. Instead, we'll be looking at whether a recommendation was successful or not based on certain criteria. This approach simplifies the understanding of the framework's implementation and allows us to focus on the reward that comes from the recommendation, rather than the specific recommendation given to a student.

In this context, the statistics are independent of the problem and remain constant. Given these parameters, a category of reinforcement learning algorithms that don't rely on state, known as the multi-armed bandit (MAB) problem, are suitable for this situation. This class can be applied to the abstracted educational scenario described above.

5.2.3 Multi-Armed Bandits (MABs)

MABs consider an agent that is an autonomous decision-maker within a specific environment, and that this environment is stateless – that is, there are no state transitions within the environment (Figure 5.3).

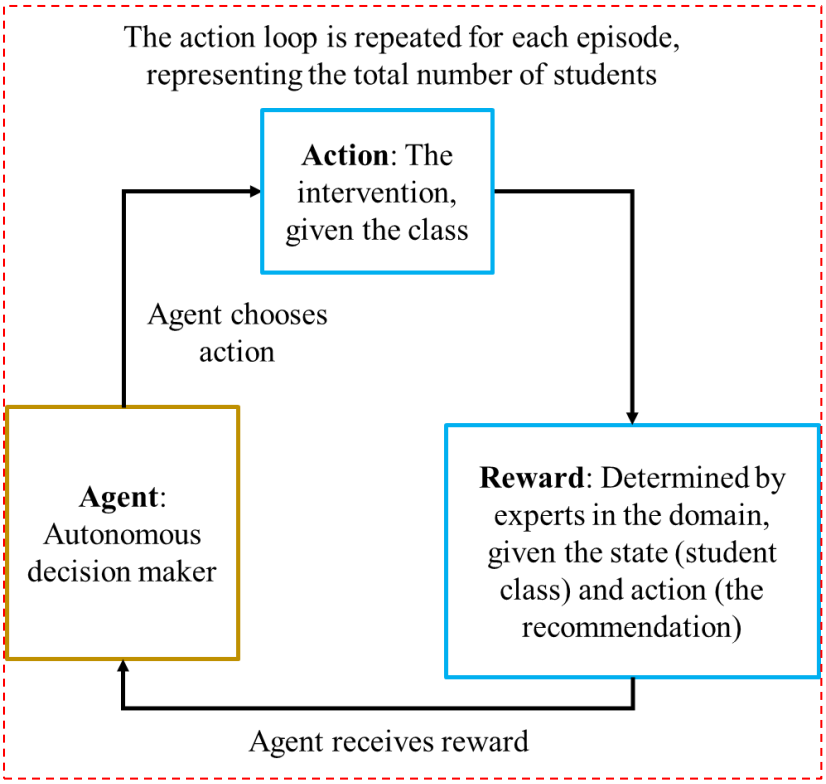


Figure 5-3 Schematic representation of the student intervention MAB environment.

In the MAB problem, an agent needs to make a choice to pull an 'arm' within a system where there are multiple arms to choose from (Morijiri *et al.*, 2022). In this example, each arm within the system represents a recommendation. The number of arms is customisable depending on the problem that needs to be solved. For the purposes of this study's experiments, and to illustrate the

cumulative rewards, this example makes use of four arms (recommendations) in the environment. The purpose of the MAB is to solve the problem of choosing the arm within a given context that yield the highest rewards overall (Marković *et al.*, 2021). For this to be practical within the education context, a few assumptions need to be made. The first assumption is that there is a pre-existing function that categorises the student into a class; the algorithm needs to choose the correct intervention to fit the class, also referred to as a contextual bandit. Secondly, the assumption is that we can add different values to the reward to illustrate different levels of impact so that the agent can learn to recommend better choices over time. The final supposition assumes that a perfect feedback system that collects the right data from students, and that provides that context to the system in a seamless way, is in place (Figure 5.4).

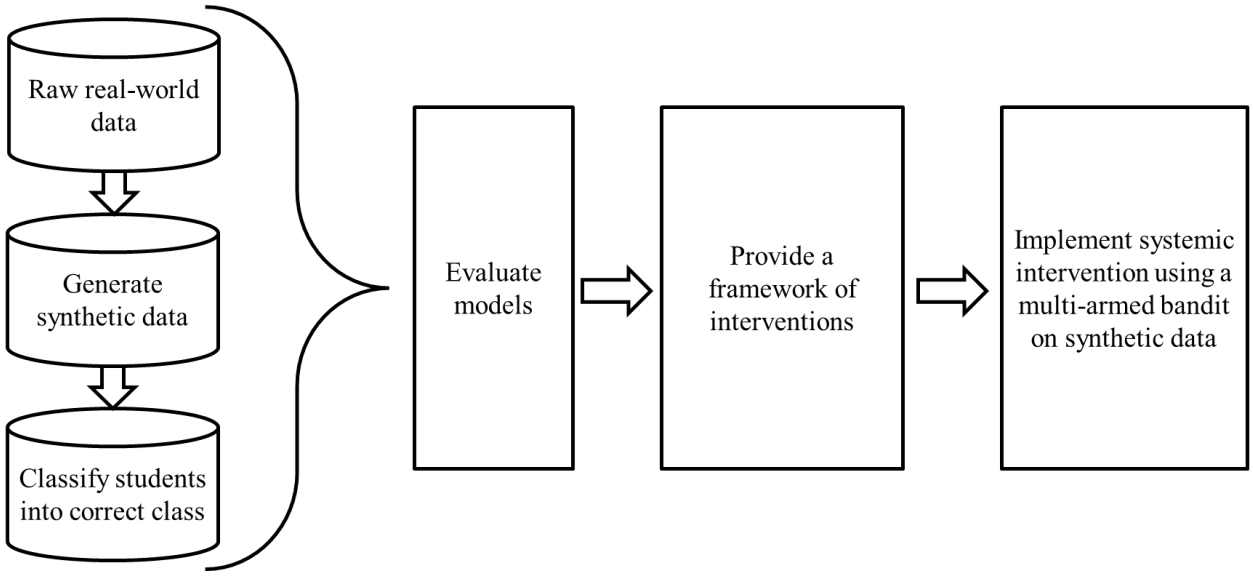


Figure 5-4 Autonomous student intervention framework for education (see Chapter 2, section 2.9).

Depending on how the environment is designed, MAB algorithms do not suffer from the cold start problem, because prior learning, such as in contextual bandits, and the balancing of exploration and exploitation takes place. However, the algorithms must still implement some form of prior training and learning, potentially on synthetic data generated (see Chapters 3 and 4). Furthermore, the action of the agent representing the types of interventions is fully customisable and does not suffer from the change in data problem, as long as there is a non-zero probability for exploration (Wan *et al.*, 2021). However, the non-zero exploration probability may take a while after initiation to detect a change, depending on the sample size and number of examples. This means that, if such

a system was built at scale, it would be tailormade for different institutions but there still needs to be a batch learning process, preferably on legacy data or synthetic data, prior to initialisation. The aim of this part of the study was to contextualise and simulate the cumulative reward within the aforementioned education environment for an intervention recommendation problem in the education context. The purpose was not to optimise the algorithm, but rather to interpret the impact of this approach on the problem's framework.

5.3 Methods

The MAB simulations will be outlined in three phases, with the first representing the environment based on the aforementioned parameter discussions, the second on the algorithms used, and the third phase on how the information will be visualised for the simulations.

5.3.1 Phase I: MAB student intervention environment

5.3.1.1 Assumptions of environment

The purpose of this section is to introduce the assumptions and the design of the simulation that will be used to study the effects of different interventions on the academic performance of students in a problem-solving context. This section is based on the outline described by Combrink *et al*., (2022b). The simulation used a multi-armed bandit (MAB) algorithm to model the agent's learning process and decision making. The agent had to choose among different types of conceptual interventions to assist the students. The environment was designed to be scaled to a student at a time, rather than all students within a system. The interventions can be correct, partially correct, incorrect, or unknown, depending on how well they match the student's needs and preferences. The agent received feedback in the form of rewards, which depended on the student's category and the intervention's type and how well the intervention worked for the student to improve on student success. The agent will aim to maximise the cumulative rewards over time by learning from the feedback and updating its own policy about the interventions and its impact. The simulation made a few assumptions about the environment and the system. The first is that the class of student has already been determined and that context enters the system. This means that the agent knows the student's category, which represents how they perform academically with or without an intervention. The student's category also represents the environment that influences

their learning outcomes and behaviours, given the intervention. In the context of a MAB problem, as described earlier is known as a contextual bandit problem (Rosman *et al.*, 2016)

The second assumption is that the simulation will use four categories of students, based on how they respond to different interventions. These categories are outlined in the next section, but are conceptual in nature and do not define a specific intervention, like a tutorial etc. Rather, these categories represent what would be current and incorrect for an agent to recommend within a system.

The third assumption is that the simulation does not focus on how well recommending a specific domain-specific intervention works. This means that the agent does not have to learn the exact actions that correspond to each intervention type for each student category. Instead, the simulation focuses on whether the agent can learn the general effectiveness of each intervention type for each student category. This means that the agent only has to learn which intervention type is the best for each student category, and not the specific details of the intervention. For example, the agent does not have to learn whether a student needs more practice, feedback, or motivation, but only what would happen to the agents own learning if they are faced with the correct, partially correct, incorrect, or unknown intervention recommendation to a student.

Another assumption is that the reward functions are known for the purpose of evaluating the agents learning. The Last assumption is that interventions only focus on maximising the reward. This means that the agent does not consider any other factors or objectives when choosing an intervention, such as the student's satisfaction, engagement, or well-being. The agent only cares about the reward, which is a representation of how well the interventions work given the parameters of the simulation. The agent will try to choose the intervention that has the highest expected reward for each student category. In other words, what is considered instead is whether an effective intervention can be made, as well as the effect thereof on the cumulative rewards. For example, the different arms of the MAB represents either a good or a bad intervention and it is up to the algorithm to learn which interventions are good or not. In the next section, the outcomes of the different interventions used in this simulation are defined.

5.3.1.2 Defining the Different Interventions and Classes

In this study, we propose a model that simulates how an agent within a system can recommend educational interventions to students based on their needs and preferences. Even though we are describing these recommended interventions in the context of students, the agent will have a MAB problem for each student, so that the different types of interventions are recommended to the various different needs students face. Educational interventions are actions or strategies that aim to improve the learning outcomes and well-being of students, such as tutoring, feedback, mentoring, counselling, or peer support, to name a few. These interventions can vary in their effectiveness depending on the type and level of the problem faced by the student, such as academic, behavioural, social, or emotional. It is important for the agent to learn from the feedback of the students and adapt its recommendations accordingly. To do so, we frame the different interventions and classes for a MAB problem. The MAB problem models a situation where the agent has to choose one of several possible interventions for each student, without knowing the exact reward of each intervention in advance. The reward is a measure of how well the intervention matches the student's problem and helps them achieve their goals. For the experiments we will measure the cumulative rewards, but it must be kept in mind that this learning will take place for each student. The system's objective is to maximise the cumulative reward over time by exploring and exploiting the best interventions for each student, separately as the system learns over time.

To simplify the simulation, we assume that there are four different classes of students, each with a different problem and a different optimal intervention. We also assume that there are four different types of interventions that the agent can recommend, each with a different reward function. The reward functions are predetermined, but the agent does not know them in advance. It has to learn them through trial and error. We incorporated the known outcome to evaluate how the agent learns over time, without this knowledge. The first type of intervention represents the incorrect intervention. If the agent chooses this, there will be no reward. This means that this intervention does not match the problem it is trying to address. For example, if the student has a behavioural or socio economic problem, and the agent recommends a tutoring session, this would be an incorrect intervention. The second type of intervention represents the unknown intervention. If the agent chooses this, there will be an unknown consequence; in other words, instances where an intervention is recommended without the outcome being known. As we define the different

classes of students, an unknown consequence could either represent a weak reward, or no reward, depending on the class of student. For example, if the student has a social problem, and the agent recommends a counselling session, this could be an unknown intervention, as the outcome may depend on the quality and suitability of the counsellor, or some latent problem the student may or may not have, which may or may not impact the outcome.

The third type of intervention represents the correct and ideal intervention. If the agent chooses this, there will be the highest reward based on the actions taken. This means that this intervention matches the problem faced by the student and helps them achieve their goals. For example, if the student has an academic problem, and the system recommends a feedback session to that specific problem, this would be a correct and ideal intervention. The correct intervention can only be known once it has impacted the outcome of the student, and extensive research is needed to fine tune these results.

Finally, the fourth type represents the partially correct intervention. If the agent chooses this, there will be a reward, but one that is weaker than the ideal recommendation. This means that this intervention may work for some students, but not for all students. For example, if the student has an emotional problem, and the system recommends a peer support session, this could be a partial intervention, as the outcome may depend on the compatibility and availability of the peer.

All reward functions for all interventions are predetermined but the agent does not know this context because the agent is learning the rewards to given actions taken for different classes. This is so that the agent's performance may be evaluated by looking at the cumulative rewards. In this simulation, four different intervention recommendations will be given to students by the system. This means that the agent can take one of four actions to solve the MAB problem. Instead of attaching a specific type of intervention to a specific type of problem, this simulation involves abstracting the concept of a student simulation to the concept of its likelihood to work. In reality, these interventions and their use cases need to be tested on a case-by-case basis (Table 5.1).

Table 5-1 MAB arm representation in the context of student recommendations.

Name	Name	Description
Incorrect	Arm 1	The incorrect intervention
Unknown	Arm 2	An intervention with an unknown consequence
Correct	Arm 3	The correct intervention
Partially correct	Arm 4	An intervention that may work, but not in all instances

As mentioned prior, four classes of students will also be represented in the simulation, and in the next section we will outline these.

5.3.1.3 Defining the Class of Students in the Simulation

The simulation used four categories of students to represent how they may perform academically with or without an intervention. The intervention can be correct, partially correct, incorrect, or unknown, depending on the student's needs. The simulation will assume certain probabilities for each student category and intervention type and use them to explore the effects of different interventions on the system and the problem-solving process. The four student categories are: Category 1, who will succeed regardless of the intervention; Category 2, who will succeed if they receive the correct intervention, and have a 70% chance of succeeding with the partially correct intervention and a 50% chance with the incorrect or unknown intervention; Category 3, who will succeed if they receive the correct intervention, and have a 50% chance of succeeding with the partially correct intervention and a 25% chance with the unknown intervention, but will fail with the incorrect intervention; and Category 4, who will only succeed 50% of the time with the correct intervention, and will fail with any other intervention. Each student category represents a different environment that influences their learning outcomes and behaviours. The simulation will help to identify the best intervention for each student category and to evaluate the impact of the intervention on the system and the problem-solving process. In the context of the MAB problem, this also means that each class of student represents its own environment (Table 5.2).

Table 5-2 Summary of simulation parameters in the experiment.

Category number	Environment number	Likelihood to pass for incorrect	Likelihood to pass for unknown	Likelihood to pass for correct	Likelihood to pass for partially correct
1	1	100%	100%	100%	100%
2	2	50%	50%	100%	70%
3	3	0%	0%	100%	25%
4	4	0%	0%	50%	0%

For the purpose of explaining the results, there is a balanced number of students between the classes and for each environment. As such, for the simulations, the episode number will be set to 500, representing 500 students per class category of student (n = 2000 students). Phase II of the simulation is described next, with an emphasis on the algorithms used.

5.3.2 Phase II: MAB Algorithms Used in the Simulation

In this simulation, three different approaches to decision-making algorithms for the MAB were used namely: a random agent, epsilon-greedy, and upper confidence bound (UCB). The three approaches represent three different ways in which a MAB would make decisions within the environment over the number of specified episodes. Firstly, the random agent, is one that takes random actions in the simulation. The random agent was included to understand what the consequences would be if a system had to randomly implement decisions that led to student intervention recommendations within this specific context. In other words, what would the outcome be if decision were just made that were random, uninformed, and did not learn from any context over the number of episodes. The second algorithm, epsilon-greedy, explores the trade-off between exploration and exploitation (Umami and Rahmawati, 2021). In this instance, exploration refers to the MAB trying out different options to learn some context about their potential rewards. Exploitation on the other hand represents actions that yield the highest number of estimated rewards, given the current knowledge within the learning process of the algorithm. In other words, the agent implementing this strategy takes an action with an unknown return or reward. Once the

action is taken, the agent is then given the reward for the action and repeats the process. The epsilon part of the algorithm is the percentage of times an agent decides to choose an arm with an unknown outcome. If epsilon is too high, it will behave like a random agent; if epsilon is too low, the agent will be too conservative. What this algorithm tries to achieve is to maximise the reward function by exploring and exploiting actions with known and unknown reward functions. For the simulation, epsilon was set at 5%, representing an exploration of 5% within the simulation where the agent would explore. The last algorithm used is upper confidence bound (UCB). With UCB, an exploration/exploitation framework like epsilon-greedy is used. However, unlike epsilon greedy, the trade-off is calculated and updated as the agent learns more from the environment (Li *et al.,* 2020). This means that UCB learns more the more it interacts with the environment, and so maximises the reward functions from the environment over time. To do so, UCB has the goal to balance the trade-off between exploration and exploitation in an optimal manner. UCB assigns a confidence bound to each arm's estimated reward by making use of uncertainty. In this instance, the uncertainty is represented by the upper percentile of a confidence interval. UCB then chooses the arm with the highest upper confidence bound which favours arms that could yield the highest potential rewards. As the algorithm learns, it refines its estimates and the upper confidence bound becomes more specific as decisions are made over time. Many better MAB algorithms, such as Bayesian Policy Reuse and Temporal Difference Learning (used in a full reinforcement learning problem), can be used to solve the MAB problem (Rosman *et al.,* 2016). For the purpose of this simulation, however, the emphasis was placed on the impact of learning, knowing context, and testing the framework, rather than optimising the MAB decision-making algorithm. The following section discusses phase III of the simulation, specific to the visualisation of the simulation, for the contextual representation of the results.

5.3.3 Phase III: Visualisation Illustrating the MAB Simulations

To visualise the data from the simulations, both the distribution of the chosen recommendations and the cumulative rewards, plotted over the number of episodes, were shown. To obtain a distribution of the possibilities within the experiments, each of the three algorithms were repeated 10 000 times. That is, there were 10 000 runs per algorithm, for 500 episodes per run, to illustrate the full distribution of the possible outcomes of the algorithms and to show potential confidence intervals within the cumulative rewards as the breadth of the algorithmic distributions were shown.

In this context, we are dealing with the outcomes of three distinct algorithms, and a random agent, each repeated 10,000 times over 500 episodes. The purpose of these repetitions is to capture the full spectrum of potential outcomes, thereby providing a comprehensive overview of the algorithmic distributions. The visualisation of this data is achieved through a combination of bar charts and a cumulative reward graph. The bar charts, four in total, are positioned adjacent to the main visualisation. Each bar chart represents one of the four 'arms' that can be 'pulled' in the algorithmic process.

Complementing these bar charts is the cumulative reward graph. This graph plots the cumulative rewards over the number of episodes, providing a visual narrative of the algorithm's performance over time. The mean line serves as a reference point, indicating the average reward at any given episode. However, the true power of this graph lies in its ability to illustrate the range of possible outcomes, which will be shown using a shaded representation surrounding the mean line, represented by an upper and lower limit. These limits are derived from the outcomes of all 10,000 runs, encapsulating the breadth of the algorithmic distributions. This interval serves as a visual indicator of the reliability of the mean line, providing context and allowing for a more nuanced interpretation of the data. The combination of these visual elements - the bar charts and the cumulative reward graph - creates a powerful tool for understanding the behaviour and performance of the algorithms. By presenting the data in this way, we can gain insights that would be difficult, if not impossible, to obtain through numerical analysis alone. However, it's important to remember that these visualisations are merely tools. They do not provide answers, but rather guide our understanding and interpretation of the data. In the next section the results will be illustrated.

5.4 Results and discussion

The first simulation illustrated the outcomes of deploying a random agent (Figure 5.5).

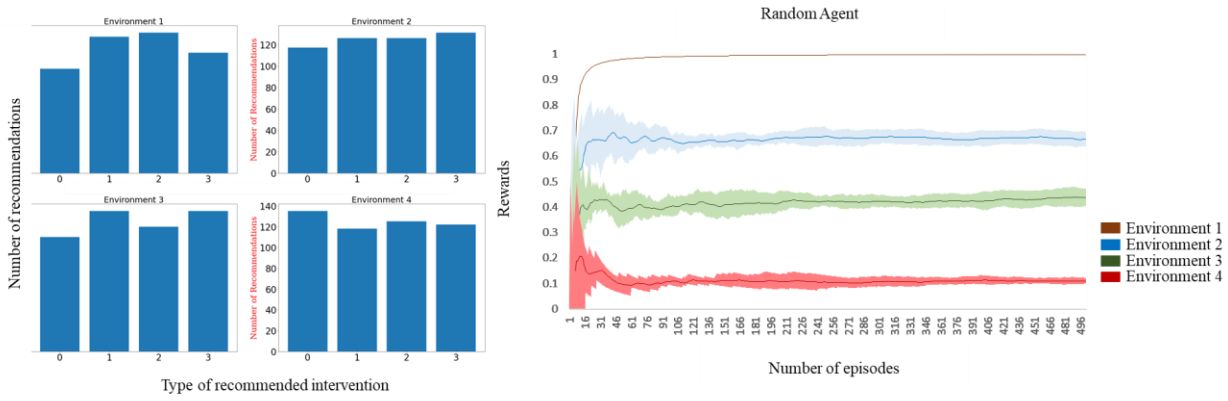


Figure 5-5 Random agent recommending intervention simulation.

As seen in the diagram, students who will pass regardless of the intervention (Environment 1) remained unaffected. Within the experiments the agent did not learn. It may seem like learning took place; however, the confidence intervals were as a result of a smoothing artifact. Across all the environments, the agent chose a random intervention for each of the four options. There were significant consequences for Category 2 (Environment 2) (average 65%; CI 62 – 69%), Category 3 (Environment 3) (average 42%; CI 38 – 45%), and especially Category 4 (Environment 4) (average 10%; CI 8 – 12%) students. What this simulation illustrates is that the consequences of random interventions will not affect students who will pass no matter the intervention. The effect on students who require a particular intervention will be influenced, and only approximately a quarter of the cohort in this simulation will get the correct intervention. In a real-world example, the number of recommended interventions will determine the distribution of the interventions across a cohort. The more interventions there are, the greater the potential harm will be for the different classes of students. Fewer interventions will yield a lesser impact in such a system. In the second simulation, the epsilon-greedy algorithm was implemented (Figure 5.6).

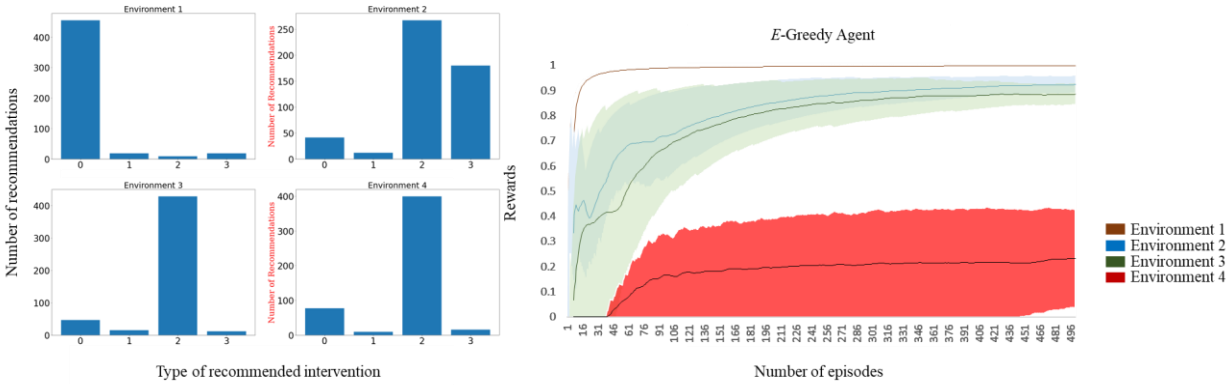


Figure 5-6 Epsilon greedy agent recommending intervention simulation.

As expected, Environment 1 was unaffected by whichever intervention was recommended. Environment 4, representing Category 4 students (average 20%; CI 2 – 35%), was affected most. Interestingly, both Environment 2 and 3 (average 80%; CI 75 – 90%) had similar results in terms of the overall cumulative rewards. For Environment 2, the distribution of interventions chosen ranged from the correct intervention to the partially correct intervention. This could be due to the likelihood of getting a seemingly optimal reward for the distribution of Category 2 students from Environment 2. Another observation specific to Environment 4 pertained to the range of consequences potentially associated with exploring interventions if the class requires a specific intervention. Although the upper limit of recommendations was 35%, the lower limit was 2%, indicating that there are instances where a random agent may have outperformed an agent exploring interventions where consequences were attached. What is important to note about these experiments is that the convergence rate can be improved with better hyperparameter design, and in this case, a better decay function for epsilon as it changes over time. Epsilon is a parameter often used in reinforcement learning algorithms, including Multi-Armed Bandits (MABs). It represents the exploration rate, i.e., the probability of choosing a random action instead of the one that the model believes to be the best. A decay function is used to decrease epsilon over time, allowing the model to explore the environment widely in the early stages of training and then gradually focus more on exploiting the best actions. A well-designed decay function can balance exploration and exploitation effectively, leading to better performance. The purpose of this part of the simulation was not to optimise the algorithm but rather to illustrate fundamental differences within different approaches taken for this specific MABs.

For the last experiment, UCB outperformed the previous two algorithms, and the ranges for each of the environments (2, 3, and 4) were closer to the average score. This is an indication of the consistency in the results over a 10 000 run simulation (Figure 5.7).

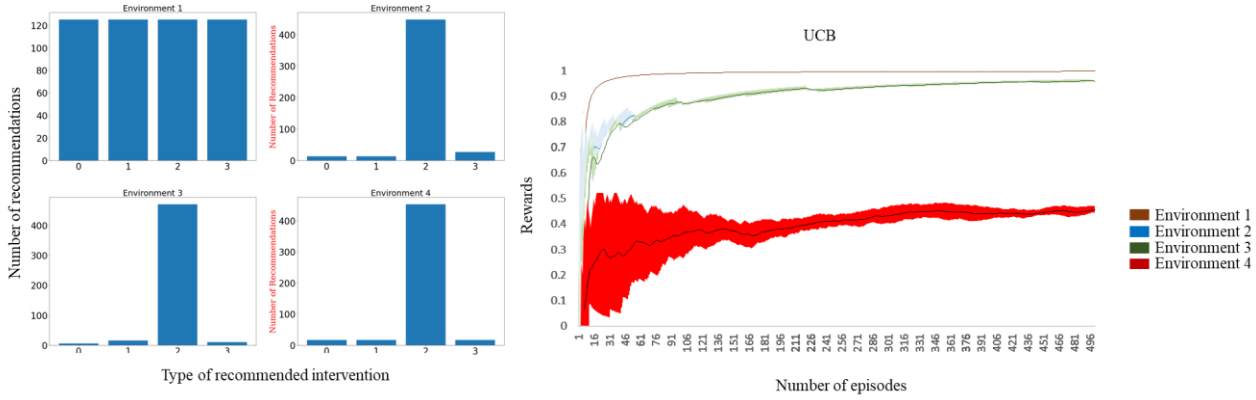


Figure 5-7 UCB agent recommending intervention simulation.

For the UCB agent, optimal hyperparameters can be designed to reach a convergence sooner in the experiment (Li *et al.*, 2020). As stated before, the purpose of these experiments was to apply these three algorithms to MABs to illustrate the impact of them on these environments, and further studies can optimise the hyperparameters as well as refine the environments.

Student retention and dropout remain a universal problem within higher education. Considerable attention has been given to addressing this problem, but it remains unsolved. Several strategies have been implemented to address this problem, including the use of automated systems to address the needs of students. As promising as these approaches are, nuanced approaches that change over time are needed to address students' continuous needs. Furthermore, various support and student intervention strategies have been implemented, but the use case of these is bound to specific contexts. As such, these do not always scale to all institutions or address the broad scope of the challenges higher learning institutions face. This chapter exemplified how such an adaptive framework may be implemented over time, but also that there are considerable areas that require further investigation for such systems to work well within education implementing a system using MAB. Moreover, the types of interventions that address specific problems students face differ

vastly, depending both on the context and the students' needs. On the basis of the experiments performed, a few important observations were made; these need to be considered for any system that uses an autonomous decision-maker that learns from the data over time. The first important observation is that the consequences of an agent acting randomly does not learn over time. Additionally, from this particular outcome is the comparison of Environment 4 from the epsilon-greedy agent to the same environment for the random agent. Hyperparameter fine tuning would assist with this problem. This observation is an indication that, in certain instances, a system can be used to make recommendations, while this should not be done in others. The challenge would then be to rather identify which interventions can or cannot be solved by a recommendation. Secondly, as seen in across all the experiments, students that will pass no matter the intervention were unaffected by the algorithms. This too is an important observation. As with the previous observation, the challenge lies in the identification of these students and there would be a need to explore more sophisticated algorithms that can take action by probing participants and gathering additional data to make a decision. The third observation highlights that the initial number of episodes require leeway to explore. As outlined in Chapter 3, this can potentially be performed based on an off-policy or batch-learning basis. These types imply that the learning takes place offline with data about the system prior to implementing the real-world steps. What is further promising about this potential approach is the use of synthetic data prior to the implementation, as outlined in Chapter 3. If an institution has limited available student cohort data, this approach might be a viable option, as the synthetic data produced can be useful in generating enough contextual data to simulate enough of the student cohort to test the system and the simulations before system implementation.

5.5 Considerations for practical system implementation

The purpose of this section is to highlight considerations when practically implementing the framework within an education context. These include, but are not limited to, a student data simulator, and a systemic intervention system. A student data simulator (SDS) addresses the challenge of generating infinite synthetic data from a specific student cohort. Such a system alleviates the fear of sharing personal information, while having the ability to generate endless synthetic data that can be used as training data for a machine learning task in the education context. Another important feature of such a system is the ability to infer probabilities within a network to

gain insight into the likelihood of a particular student or situation to occur. The systemic intervention system (SIS) on the other hand is a system that can scale student support within online systems based on the best way to address a challenge a student faces. Such a system would run concurrently while students interact with an online platform like a learning management system. An SIS would use specific features and context obtained to learn which recommendations best fit the student problem. To elaborate on the practical considerations for system implementation, the two aforementioned systems (SDS and SIS) will be used as examples to illustrate the challenges these systems need to overcome, starting with an SDS.

5.5.1 Student data simulator (SDS)

The first system that can be built from this framework is a SDS system. An SDS is a system that can both model and generate data representing a student or multiple students within a specific education context (see Chapters 3 and 4). The data generated can range from simple to complex and should include as much context about the student so that the user of such a student simulator may tailor the data needs to their requirements. Such a student simulator is useful for synthetic data generation, probabilistically modelling data within an institution, and can be used for sharing large amounts of student data build from training data originating within that institution. As noted earlier (see Chapter 2, section 2.3.3), data sharing poses major challenges in the machine learning for education domain, and such a system may provide a viable solution to this problem. As such, the following steps outline what is needed to create such a system:

Step 1: Regulatory approval to use the data for an SDS

For the student data simulator to be contextually relevant there needs to be a small sample of real-world data from hence the probabilistic models are built. For this step to be implemented correctly, the data, its context and the required ethics are needed. Even though the SDS can generate infinite data, the training data to build the SDS still requires ethical approval and the correct data masking to adhere to all the requirements from the required data regulatory bodies so that it can be used in such a system.

Step 2: Define the computational resources needs for the system

Training a probabilistic model like a Bayesian network (BN) structure and parameters from data, is computationally expensive and requires a lot of resources to build. As a result, there is a need to balance the number of students to effectively train the model, especially in low resource or low data contexts. For example, a first year education cohort may have 400+ students, but a first year philosophy cohort only 15. The number of variables to use and the requirements are determined by the complexity of the data, and the type of BN that will be used to build the data. According to Joubert and De Waal (2020), when building a complex BN like an activity-based travel demand network, a rigorous process is needed to regularly test the accuracy of the models. This testing, in addition to the building, requires computational resources and depending on the complexity of the data and the BN methods used, the computational resources need to match that.

Step 3: Feature engineer the data

For certain use cases, this will mean that there will be context that is lost as a result of this type of feature engineering. For example, a student mark would need to be presented as a category rather than a specific percentage. As such, the feature engineering used to test the variables and the number of categories for each node needs to be predetermined. The specific choice over the best method to use will depend on the data and the use case. This does mean that the feature engineering would require different approaches to test the best fit for the specific system within a specific institution. Additionally, the distribution of the data used in the training may be unbalanced (see Chapter 3, section 3.3.2.1). As a result, different BNs are required to be built to test the best fit BN for the problem (see Chapter 3, section 3.5.4).

Step 4: Build, evaluate, and compare different BN during the pilot phase

As outlined in Chapter 3 (see section 3.3.3), different scoring methods and approaches can be used to build a BN when it learns from data to generate the structure and learn the parameters. As a result, the initial pilot phase should include different BN built from data using these different approaches. Based on the study, an initial sample size of a minimum of 100 samples is needed to build the network, and the training data for the BN should therefore be based on a minimum of 100 deidentified, feature engineered samples representing a student cohort. Once the different BN have been built, the models need to be validated. The first validation is based on the similarities

between the BN, specific to the number of connected and disconnected nodes (see Chapter 3, section 3.5.4). Once the networks have been validated, the best fit BN needs to be presented to experts within the domain for further validation. The purpose of this validation is to reach a consensus over the BN that best learned from the data, based on expert context. This includes the involvement of people with expertise in the variables associated with the construction of the BN. Once a consensus is reached, synthetic data needs to be generated from the BN of choice to be validated.

Step 5: Test the utility of the BN using a ML classifier

Once the synthetic data has been generated using the aforementioned BN, different ML classifiers need to be fine-tuned and tested to evaluate the utility of the data. In addition to this, how well the classifier performed at all class predictions needs to be assessed. As noted in Chapter 3 (see sections 3.5.5 – 3.5.7) certain classes may yield a better accuracy than others. There could be a variety of reasons for different class variables performing better in the classification task (see Chapter 4, section 4.4), of which the learning curves can provide context into whether or not the model is overfitting, underfitting or learning at a sufficient rate from the training data. Once the correct training data features have been engineered, the BN has been built, the utility has been tested, and a consensus is reached between experts on the best BN to use for the SDS, the user interface may be designed, and the system may be deployed.

Step 6: Design considerations for implementation

A challenging factor is the user interface for such a system. The decision makers, researchers and users that can find meaning in the inferences generated from such a system need to be able to navigate such systems without being an expert user. This is because the depth and level of insight that can be generated by using inference in a BN, does require contextual understanding into the underlying data. For example, if a student simulator was built on all demographic, academic, learning-management system, and socio-economic data about a student, a user of such a system needs to have well defined and well-articulated questions in order to derive meaning from such a system. This means that expert level-training to use and navigate around a student simulator is required, and an investigation to optimise the level of understanding to manage such a system, is needed. Another challenge to overcome relates to the visualisations needed to properly monitor

and evaluate such a system. For a student simulator to effectively work, different institutions need a data sharing framework and the legal instruments to ensure that the data, and the unintended consequences of sharing such data, are well understood. Although there are no foreseeable direct risks associated with the implementation of a student simulator, there still is a need to conduct such experiments to ensure that the benefits of having such a system are properly understood. Next, the practical considerations associated with a systemic intervention system (SIS) will be outlined.

5.5.2 Systemic intervention system (SIS)

A SIS is a system that considers some contextual knowledge about users within a system and uses that context to recommend an intervention within a system. Although the task is a recommender problem, the systemic intervention system does not work as an information filtering system. Instead, such a system uses information learned about a particular problem and a possible solution within a MAB context to recommend a particular intervention (see Chapter 5, section 5.2.3). Practically put, such a system would be presented within an online context, or within a platform where students engage online in a regular manner. In the next section the steps to building such an SIS system will be outlined.

Step 1: Define which signals the MAB learns from, including the actions the agent may take

The MAB would present certain content to the users. This content would be the intervention. The MAB would navigate which content recommended to users would be the most useful, based on a certain signal. This signal can be in the form of a combination of click data, and the duration students spent viewing the content that was recommended. These signals can provide the feedback the system uses to update the rewards. To this extent, a feedback loop that the users, in this case, the students, provide on a recommendation made that added meaning, may be useful. Another signal that the agent can have is from the academic advisors or educators within the system. If a series of recommendations are made, then the agent can present the lessons learned to a group of internal users within an education institution. The expert users can then provide feedback to the system, and the reward functions can thus be updated. A drawback with both of these approaches is that the turnaround time for this type of feedback might take too long and the agent might only update a particular recommendation after having implemented the same strategy for a while. As a result, there is a third type of solution to this problem, and that is to collect real time data on the

interventions themselves. The signals, the approach and the actions the agent takes needs to be decided upon prior to the development of the SIS as these design choices will influence all areas of the MAB.

Step 2: Reach consensus on the student interventions and their impact

Suppose a system recommends a student to seek academic advising, for example. If the frequency of students entering academic advisors was monitored, then the system can look at a relative proportional increase of students going for academic advising as a result of the recommendations. The problem with this is measuring the impact of a student going to an academic advisor. For example, if an SIS recommends 500 students to seek academic advising, how will the impact of the advising be measured? Although not perfect, there are measurements that can be put in place to quantify these in real time, such as putting in a measurement or model that can measure impact based on some criteria, that is then collected and given to the agent as a signal. Another problem a student systemic intervention system can solve is figuring out which interventions best fit different problems. Much like a clinical trial, a MAB can also be used in the form of AB testing to measure the impact and efficiency of different types of interventions that students require within a complex education system. If students are presented with different types of content, for example, given a specific educational challenge, the MAB can learn which recommendations were the most useful to the students. Thus, a consensus needs to be reached between experts within a particular context over which recommendations the agent may use as actions within the system.

Step 3: Choosing the best MAB and mapping unintended consequences

Choosing the Best MAB to fit the problem requires testing of the learning rate, cumulative rewards, and the approaches used to optimise the system, including the signals used. As such, a variety of different MAB approaches may be taken. One recommendation to contextualise the SIS to a specific context, is to learn from prior data in the form of a contextual bandit or an algorithm like Bayesian Policy reuse (Rosman *et al*., 2016). Once decided upon, the different algorithms are required to be tested in a simulated environment prior to live testing. During this time, it would be useful to map out the potential unintended consequences of implementing such a system. The

mapping exercise should include what to do if either the system, student or any person within the institution is faced. The following scenarios and guiding questions may assist with the mapping:

- Suppose an incorrect recommendation is made to a student that does not address the problem the student is facing. How will these students be identified? How will these problems be addressed?
- Suppose there is a negative consequence towards a student as a direct result of the recommendation made to a student. Who will take accountability for this problem, the engineers that built the system or the panel of experts that agreed upon the recommendation the agent in the system was allowed to make? What data should the system collect to learn from this context and prevent further harm? What are the procedures that should be in place, to prevent any further harm to students?
- If the recommendation does not address the challenges the student faces, where should the student be referred to?
- If the system fails due to physical problems like internet connectivity, what measures should be in place to mitigate risk and prevent bottlenecks in terms of recommendations?
- How will the impact of the recommendations within the system be measured?

The guiding questions mentioned earlier are not the only ones considered in this mapping exercise. However, they hold significant importance when dealing with systems that necessitate the collaboration of human judgment and computer-based decision making. Due to this, a student systemic intervention system can thus be used to recommend interventions to students, as well as be designed in a way to learn which interventions work best for different problems a student faces within a complex education system. Such a system can add immense value to an education institution at both learning what works best for students and which interventions work best to fit a particular problem students face.

5.6 Conclusion

To create effective smart systems within higher education institutions, it is crucial to design models that are generalisable and adaptable to the similarities and differences among students. This requires a combination of education theory, human intervention, adaptability, and human-machine interactions embedded within the system. Analytical and computational models, including analytics, and reinforcement learning algorithms, need to be evaluated as part of the fundamental

process of establishing such a system. However, it is important to note that technology alone cannot define the boundaries and scope of interventions and processes within the education context. Understanding how to implement interventions within complex systems is vital, and coupling the science of systems learning with domain-specific knowledge is essential for success. While the rapid advancement of technology may hold promise in improving student throughput and retention rates, prioritising our fundamental understanding of concepts is crucial for responsible implementation of autonomous learning in education.

CHAPTER 6 CONCLUSION

In this thesis, the main question “Can education be abstracted in a framework and conceptually studied for a student intervention process?” was addressed throughout all the chapters. The thesis started by framing the theories and approaches to address the question. The main contributions of this thesis included synthetic data generation practices, measuring utility using machine learning classification tasks, framing higher education as a complex system within the lens of systemic intervention, applying principles of synthetic data generation to a real world dataset, designing experimentation for systemic intervention as a multi-armed bandit problem and the conceptual design of the aforementioned conceptual contributions to real world application. To this extent, how a systems approach and complexity theory in the context of higher education are applied was outlined. This approach (as outlined in Chapter 2) provided the systemic outline to unpack the student intervention framework, which included elements of complexity theory, systemic intervention, synthetic data generation and an autonomous decision maker into the overall framework. These theories were contextualised to a known system theory, specifically systemic intervention. For each of the subsequent chapters, these theories were applied and all experiments within each chapter represented the complex system in a simplified manner, which would have been otherwise impossible if not for the application of these theories.

In Chapter 2, challenges faced in the education domain with regard to available data, data sharing, systemic interventions and challenges in the education context was outlined. In Chapter 3, the challenge of data sharing, and a lack of available data was addressed by means of generating synthetic tabular data for the education context. In Chapter 3, different methods to generate synthetic tabular data using different models were explored. Furthermore, in Chapter 3, the utility of these models was measured using a classification task. Based on these findings, a Bayesian Network using Bayesian information criterion as the score based method to generate the direct acyclic graph was the best model to generate synthetic data. From this chapter it was observed that the models overfit and underfit the data. To address both of these issues, it was recommended that more complexity is introduced into the underlying data when discrete variables are used as well as a larger synthetic dataset is used when testing the models. From the experiments, when the original data are used as the training data, with the synthetic data as the test set, the models could make

some predictions but, in most cases, it was not ideal. A markable difference was observed between data synthetically generated using probabilistic models against the deep learning models. Probabilistic models did perform better overall, but the authors note that additional studies incorporating better hyperparameter fine tuning would have the potential to improve the performance of the deep learning models. Across all models used, a score based method produced the best results, and this informed the application of this context to the real-world dataset.

In Chapter 4, these aforementioned concepts from Chapter 3 were applied to a real-world dataset, the University of the Free State, Economics and Management Sciences. The results obtained from the synthetic data demonstrate that this approach can accurately generate synthetic student data, effectively capturing the complexity of the education system. Another important observation is that the experiments conducted also provided insight into the underlying associations within the data, which are useful observations in understanding how complex systems relate to one another using data. These inferences are a biproduct of using a probabilistic model, and although not explored in this thesis, highlights an area for future research in this domain. Although the complexities of these systems were not fully explored in this thesis, how the framework applies to the education context was outlined and tested.

Lastly, in Chapter 5, the education student interventions recommendation problem was abstracted as a stateless multi-armed bandit problem. Based on these findings, it is suggested that researchers and scholars need to know the impact of an intervention before rolling-out such a system at scale to prevent some of the unintended consequences such as the potential harm of recommending an intervention that might negatively impact a student. Another important contribution was the impact of discovering the potential harms associated with an autonomous decision maker exploring interventions if the impact of the intervention or the type of intervention the student requires is not known.

In conclusion, this research has illustrated the use of complexity theory and systemic intervention in understanding and studying higher education as a complex system. The intricate interplay of numerous components within the educational system, from students and lecturers to resources and policies, was examined through this lens. This approach has provided a holistic perspective,

revealing the system's emergent properties and the often non-linear effects of systemic interventions such how useful it is to understand utility, and what systems like a multi-armed bandit are capable of if provided the correct context. However, the inherent complexity and context-dependency of educational systems necessitate ongoing monitoring and adjustment of interventions. The insights gained are invaluable for the future of educational practice that can lead to better refinement of training and ultimately an adjustment to policy to address data sharing issues, studying systemic interventions within complex systems and the upskilling required within the education system to make use of the potential emerging technologies within this domain.

6.1 Limitations of the study

Understanding complex systems and how the relationships between complicated systems within a complex system interact with one another remain a challenge. Oversimplifying a complex system allows for the understanding of the system, but with this understanding comes a loss in context and a reduction in the specific details that make up a system. To this extent, the data used in the experiments from this thesis were specific to a faculty within a South African university. If more data, and more complicated data were used, deeper insights would be gained from applying the approaches in this thesis. It is therefore recommended that with each complex system level study conducted, that the smaller subunits consisting of the systems themselves are well understood. This means that studies that focus on the context of education in all its forms and complexities from a systems level perspective needs to be studied. As the metrics and science of exploring the impact of complex systems increase, so too will the understanding of how to implement these in the context of a framework. It is therefore also recommended that as the fundamental thought processes surrounding complex systems increase, so too should the supporting science and integrating the philosophies to the systems being built are vital for better systems and better associations to be created within such systems. This includes more studies on synthetic data for education, multi-armed bandit problems for education, and better data to perform more sophisticated experiments. The last limitation of the study is the acknowledgement that interdisciplinary fields combining education with data science is not yet common practice, and there still remains a need to frame problems in this domain so that an entire range of scholars can contribute toward this emerging scientific field.

6.2 Next steps

The findings of this study highlight a need to explore the GANs ability to generate synthetic tabular data for the education context in more detail. As outlined in the thesis, the GANs performance of generating synthetic tabular data for the education context can be improved with extensive research on the various hyperparameters that can be fine-tuned, exploring different types of network architectures, and measuring the outputs changes to different loss functions used. Grounds for such an extensive deep learning study is justified by the successes other researchers have had in the domain by fine tuning tabular data and image data. Based on the promising results of this study, synthetic data generated in education should be explored to improve better data sharing practices and increase the number of studies in this domain. Better data sharing practices and better ways of generating synthetic data means that more people will have access to large amounts of information to develop better models for education. This includes the incorporation with more complex datasets and data types for education. Furthermore, different types of classification tasks should be explored for the education domain. This includes applying classification tasks trained on text data from student assignments, tabular data that includes complex variables such as attendance, food security data, and safety data, among others. A viable data sharing framework for generating and sharing synthetic data for education should be created so that researchers in the education domain can collaborate and innovate new ways of supporting students based on data. The framework could incorporate synthetic data generated using a combination of synthetic data generated through deep learning and probabilistic models. Included in this process is the establishment of complete Bayesian Networks to represent these complex systems. One important feature of Bayesian Networks is the ability to construct inferences based on the networks themselves so that the probabilities of nodes within the networks may be established. These probabilities can assist not only in the generation of synthetic data, but also aid university decision makers understand the extent of their student cohort and the probabilities associated with their student cohorts across various departments so that causality and trends within the data may be explored to provide better insight into how institutions may respond to the needs of students. Another important area for exploration is the implementation of the student intervention framework on a small scale no risk student cohort using a multi-armed bandit problem. The cohort should be selected as a low-to-no risk students that receive interventions from a system based on their data. This type of quasi-experiment should first be piloted in a controlled environment and

then ultimately scaled to a small real-world example within an institution. To do so, several ethical, and governance strategies need to be in place to ensure that the experiments on human participants are conducted in a safe and responsible manner. Determining how such a system will learn over time is vital for the success of such an exploration so that these concepts may be applied to other reinforcement learning algorithms in this domain. The authors note that there are several environments and scenarios that need to be created to effectively study different algorithms in the education domain. This includes creating environments that can contextualise the implementation of specific interventions in education. For example, if a systemic intervention is going to prompt a student, what type of actions will the agent take, how will the reward function be measured, and how will the system be designed is still not known. A large amount of conceptual work is still required to abstract and conceptualise the extent to which reinforcement learning may be applied in the education domain, across all spheres. This also includes having agents and people work together to solve certain tasks that can be scaled to large cohorts of students, where the number of people to implement the intervention may be limited. To unpack and explore these concepts require expertise in philosophy, education, social science, computer and data science, and management, among others. The types of explorations and resources needed to conduct these experiments cannot be understated, but neither can the importance of this work.

6.3 Implications for future research in this domain

The thesis highlighted the significance of incorporating various domains into the realm of education. An important feature of this work is the emphasis education researchers need to place on system based education research projects. This includes questions in the education domain such as:

- To what extent can an intervention improve academic performance?
- Which interventions improve academic performance, given a specific challenge?
- Which methodologies should be employed in education to monitor student progression ethically?

Although these questions alone are not the epitome of possibilities in this domain, they are fundamental stepping stones in understanding the full extent of what a system should and should not do in the context of providing an intervention to students for the education domain. Another important implication is the type of information that is shared about the education

domain for researchers. In a time where data governance and the sharing of data is under severe scrutiny, innovation should not be stifled because of this. Instead, synthetic data may pose a viable solution for researchers to explore the possibilities of innovation without the risk of leaking personal information. To this extent, a governing body that explores the masking and unmasking of education data for machine learning in education research may be established so that data sharing is done responsibly. This could also mean that probabilistic models may be used as an effective blueprint for higher education data sharing. Finally, it is noted that this type of work requires an integrated, multidisciplinary approach by incorporating various domains.

REFERENCES

1. Abu-Salih, B., 2021. Domain-specific knowledge graphs: A survey. Journal of Network and Computer Applications, 185(1), p.103076. https://doi.org/10.1016/j.jnca.2021.103076
2. Adams, R., Adeleke, F., Anderson, D., Bawa, A., Branson, N., Christoffels, A., Vries, J.D., Etheredge, H., Flack-Davison, E., Gaffley, M. and Marks, M., 2021. POPIA code of conduct for research. South African Journal of Science, 117(5-6), pp.1-12. https://dx.doi.org/10.17159/sajs.2021/10933
3. Agbo, F.J., Oyelere, S.S., Suhonen, J. and Adewumi, S., 2019. November. A systematic review of computational thinking approach for programming education in higher education institutions. In Proceedings of the 19th Koli Calling International Conference on Computing Education Research 1(1), pp. 1-10. https://doi.org/10.1145/3364510.3364521
4. Aggarwal, R. and Ranganathan, P., 2016. Common pitfalls in statistical analysis: The use of correlation techniques. Perspectives In Clinical Research, 7(4), p.187. https://doi.org/10.4103/2229-3485.192046
5. Aggarwal, A., Mittal, M. and Battineni, G., 2021. Generative adversarial network: An overview of theory and applications. International Journal of Information Management Data Insights, 1(1), p.100004. https://doi.org/10.1016/j.jjimei.2020.100004
6. Aguilera-Rueda, V.J., Cruz-Ramírez, N. and Mezura-Montes, E., 2020. Data-driven Bayesian Network learning: A bi-objective approach to address the bias-variance decomposition. Mathematical and Computational Applications, 25(2), p.37. https://doi.org/10.3390/mca25020037
7. Alarsan, F.I. and Younes, M., 2021. Best selection of generative adversarial networks hyper-parameters using genetic algorithm. SN Computer Science, 2(4), p.283. https://doi.org/10.1007/s42979-021-00689-3
8. Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A. and Aljaaf, A.J., 2020. A systematic review on supervised and unsupervised machine learning algorithms for data science. Supervised and Unsupervised Learning for Data Science, pp.3-21. https://doi.org/10.1007/978-3-030-22475-2_1

9. Al-Masni, M.A., Kim, D.H. and Kim, T.S., 2020. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Computer Methods and Programs in Biomedicine, 190(1), p.105351. https://doi.org/10.1016/j.cmpb.2020.105351
10. Alsariera, Y.A., Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A.A. and Ali, N.A., 2022. Assessment and evaluation of different machine learning algorithms for predicting student performance. Computational Intelligence and Neuroscience, 2022. https://doi.org/10.1155/2022/4151487
11. Alyahyan, E. and Düştegör, D., 2020. Predicting academic success in higher education: literature review and best practices. International Journal of Educational Technology in Higher Education, 17, pp.1-21. https://doi.org/10.1186/s41239-020-0177-7
12. Amari, S.I., 1993. A universal theorem on learning curves. Neural networks, 6(2), pp.161-166. https://doi.org/10.1016/0893-6080(93)90013-M
13. Amissah, M., Gannon, T. and Monat, J., 2020. What is systems thinking? Expert Perspectives from The WPI Systems Thinking Colloquium of 2 October 2019. https://doi.org/10.3390/systems8010006
14. Anderson, B., 2019. Using Bayesian networks to perform reject inference. Expert Systems with Applications, 137(1), pp.349-356. https://doi.org/10.1016/j.eswa.2019.07.011
15. Assuah, C.K., Sabtiwu, R., Armah, R.B., Abedu, G. and Awulo, F., 2022. Pass or Failure of Students in The "WASSCE" Mathematics Mock Examination: The Binary Logistic Regression Model. International Journal of Education, Learning and Development, 10(4), pp.38-56. https://www.eajournals.org/
16. Baasch, G., Rousseau, G. and Evins, R., 2021. A conditional generative adversarial network for energy use in multiple buildings using scarce data. Energy and AI, 5(1), p.100087. https://doi.org/10.1016/j.egyai.2021.100087
17. Baashar, Y., Alkawsi, G., Mustafa, A., Alkahtani, A.A., Alsariera, Y.A., Ali, A.Q., Hashim, W. and Tiong, S.K., 2022. Toward predicting student's academic performance using artificial neural networks (ANNs). Applied Sciences, 12(3), p.1289. https://doi.org/10.3390/app12031289
18. Bhattacharya, G., Ghosh, K. and Chowdhury, A.S., 2017, November. kNN classification with an outlier informative distance measure. In International Conference on Pattern

Recognition and Machine Intelligence (pp. 21-27). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69900-4_3

19. Bae, H., Monti, S., Montano, M., Steinberg, M.H., Perls, T.T. and Sebastiani, P., 2016. Learning Bayesian networks from correlated data. Scientific Reports, 6(1), pp.1-14. | https://doi.org/10.1038/srep25156
20. Balanda, K.P. and MacGillivray, H.L., 1988. Kurtosis: a critical review. The American Statistician, 42(2), pp.111-119. https://doi.org/10.1080/00031305.1988.10475539
21. Batmaz, Z., Yurekli, A., Bilge, A. and Kaleli, C., 2019. A review on deep learning for recommender systems: challenges and remedies. Artificial Intelligence Review, 52(1), pp.1-37. https://doi.org/10.1007/s10462-018-9654-y
22. Baus, C.A., Lunsford, D. and Valdes, K., 2021. Factors influencing student success in a graduate clinical neuroscience course: a survey study. The American Journal of Occupational Therapy, 75(Supplement_2), 1(1), pp.7512505188p1-7512505188p1. https://doi.org/10.5014/ajot.2021.75S2-RP188
23. Beaulac, C. and Rosenthal, J.S., 2019. Predicting university students' academic success and major using random forests. Research in Higher Education, 60(1), pp.1048-1064. https://doi.org/10.1007/s11162-019-09546-y
24. Bengesai, A.V. and Pocock, J., 2021. Patterns of persistence among engineering students at a South African university: A decision tree analysis. South African Journal of Science, 117(3-4), pp.1-9. http://dx.doi.org/10.17159/sajs.2021/7712
25. Beretta, S., Castelli, M., Gonçalves, I., Henriques, R. and Ramazzotti, D., 2018. Learning the structure of bayesian networks: A quantitative assessment of the effect of different algorithmic schemes. Complexity, 2018. https://doi.org/10.1155/2018/1591878
26. Berrett, T.B. and Samworth, R.J., 2021. USP: an independence test that improves on Pearson's chi-squared and the G-test. Proceedings of the Royal Society A, 477(2256), p.20210549. https://doi.org/10.1098/rspa.2021.0549
27. Bashir, D., Montañez, G.D., Sehra, S., Segura, P.S. and Lauw, J., 2020. An information-theoretic perspective on overfitting and underfitting. In AI 2020: Advances in Artificial Intelligence: 33rd Australasian Joint Conference, AI 2020, Canberra, ACT, Australia, November 29–30, 2020, Proceedings 33 (pp. 347-358). Springer International Publishing. https://doi.org/10.1007/978-3-030-64984-5_27

28. Bidandi, F., Ambe, N.A. and Mukong, C., 2022. Collaboration and partnerships between higher education institutions and various stakeholders: Case study, the University of the Western Cape, South Africa. https://doi.org/10.21203/rs.3.rs-1371503/v1
29. Birhane, A., 2021. Algorithmic injustice: a relational ethics approach. Patterns, 2(2), p.100205. https://doi.org/10.1016/j.patter.2021.100205
30. Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O. and Kerres, M., 2020. Mapping research in student engagement and educational technology in higher education: A systematic evidence map. International Journal of Educational Technology in Higher Education, 17(1), pp.1-30. https://doi.org/10.1186/s41239-019-0176-8
31. Bosch, J., Olsson, H.H. and Crnkovic, I., 2021. Engineering ai systems: A research agenda. Artificial Intelligence Paradigms for Smart Cyber-Physical Systems, 4(1), pp.1-19. https://doi.org/10.4018/978-1-7998-5101-1.ch001
32. Botha, R.J., 2016. Postgraduate throughput trends: A case study at the university of Ghana. Journal of Social Sciences, 49(1-2), pp.58-66. https://doi.org/10.1080/09718923.2016.11893597
33. Botha, R.J., 2018. Student throughput trends on postgraduate level: An African case study. The Independent Journal of Teaching and Learning, 13(2), pp.53-66. https://hdl.handle.net/10520/EJC-12224d59f2
34. Brackett, M.A., Bailey, C.S., Hoffmann, J.D. and Simmons, D.N., 2019. RULER: A theory-driven, systemic approach to social, emotional, and academic learning. Educational Psychologist, 54(3), pp.144-161. https://doi.org/10.1080/00461520.2019.1614447
35. Bradley, V.M., 2021. Learning Management System (LMS) use with online instruction. International Journal of Technology in Education, 4(1), pp.68-92. https://doi.org/10.46328/ijte.36
36. Brewer, M.L., van Kessel, G., Sanderson, B., Naumann, F., Lane, M., Reubenson, A. and Carter, A., 2019. Resilience in higher education students: A scoping review. Higher Education Research and Development, 38(6), pp.1105-1120. https://doi.org/10.1080/07294360.2019.1626810
37. Brunetti, F., Matt, D.T., Bonfanti, A., De Longhi, A., Pedrini, G. and Orzes, G., 2020. Digital transformation challenges: strategies emerging from a multi-stakeholder

approach. The TQM Journal, 32(4), pp.697-724. https://doi.org/10.1108/TQM-12-2019-0309

38. Buggineni, V., Chen, C. and Camelio, J., 2024 Enhancing Manufacturing Operations with Synthetic Data: A Systematic Framework for Data Generation, Accuracy, and Utility. Frontiers in Manufacturing Technology, 4, p.1320166. https://doi.org/10.3389/fmtec.2024.1320166
39. Burnett, C., 2021. A national study on the state and status of physical education in South African public schools. Physical Education and Sport Pedagogy, 26(2), pp.179-196. https://doi.org/10.1080/17408989.2020.1792869
40. Bynagari, N.B., 2019. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. Asian Journal of Applied Science and Engineering, 8, pp.25-34. https://pdfs.semanticscholar.org/5e2b/175b3a92ff36494d97c0055cd0a2f6df5f28.pdf
41. Canh, N.P., Wongchoti, U., Thanh, S.D. and Thong, N.T., 2019. Systematic risk in cryptocurrency market: Evidence from DCC-MGARCH model. Finance Research Letters, 29(1), pp.90-100. https://doi.org/10.1016/j.frl.2019.03.011
42. Canning, E.A., Muenks, K., Green, D.J. and Murphy, M.C., 2019. STEM faculty who believe ability is fixed have larger racial achievement gaps and inspire less student motivation in their classes. Science Advances, 5(2), p. 4734. https://doi.org/10.1126/sciadv.aau4734
43. Cano, A., Nguyen, D.T., Ventura, S. and Cios, K.J., 2016. ur-CAIM: improved CAIM discretization for unbalanced and balanced data. Soft Computing, 20, pp.173-188. https://doi.org/10.1007/s00500-014-1488-1
44. Cawley, G., Talbot, N. and Girolami, M., 2006. Sparse multinomial logistic regression via bayesian l1 regularisation. Advances in neural information processing systems, 19. https://proceedings.neurips.cc/paper/2006/hash/b22b257ad0519d4500539da3c8bcf4dd-Abstract.html
45. Cennamo, K. and Kalk, D., 2019. Real-world instructional design: An iterative approach to designing learning experiences. Routledge. Available at: https://books.google.co.za/books?id=xT33DwAAQBAJ&lpg=PP1&ots=gSvO4fZxRE&dq=A%20systems%20approach%20iterative&lr&pg=PR4#v=onepage&q=A%20systems%20approach%20iterative&f=false

46. Cervetti, G.N. and Hiebert, E.H., 2019. Knowledge at the center of English language arts instruction. The Reading Teacher, 72(4), pp.499-507. Available at: https://www.jstor.org/stable/26801637
47. Chandler, L., 2020. A Legal Comparative Analysis of Automated Decision-Making and Reasonableness in Administrative Law. Available at: SSRN 3817043. http://dx.doi.org/10.2139/ssrn.3817043
48. Cheng, L. and Yu, T., 2019. A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems. International Journal of Energy Research, 43(6), pp.1928-1973.https://doi.org/10.1002/er.4333
49. Chi, H., Zhang, Y., Tang, T.L.E., Mirabella, L., Dalloro, L., Song, L. and Paulino, G.H., 2021. Universal machine learning for topology optimization. Computer Methods in Applied Mechanics and Engineering, 375(1), p.112739. https://doi.org/10.1016/j.cma.2019.112739
50. Chiu, M.C., Huang, J.H., Gupta, S. and Akman, G., 2021. Developing a personalized recommendation system in a smart product service system based on unsupervised learning model. Computers in Industry, 128(1), p.103421. https://doi.org/10.1016/j.compind.2021.103421
51. Coetzee, J., Neneh, B., Stemmet, K., Lamprecht, J., Motsitsi, C. and Sereeco, W., 2021. South African universities in a time of increasing disruption. South African Journal of Economic and Management Sciences, 24(1), pp.1-12. http://dx.doi.org/10.4102/sajems.v24i1.3739
52. Collins, A.G. and Cockburn, J., 2020. Beyond dichotomies in reinforcement learning. Nature Reviews Neuroscience, 21(10), pp.576-586. https://doi.org/10.1038/s41583-020-0355-6
53. Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022a. Comparing Synthetic Tabular Data Generation Between a Probabilistic Model and a Deep Learning Model for Education Use Cases. arXiv preprint arXiv:2210.08528.
54. Combrink, H.M.v.E., Marivate, V. and Rosman, B., 2022b, December. Reinforcement learning in education: A multi-armed bandit approach. In International Conference on Emerging Technologies for Developing Countries (pp. 3-16). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35883-8_1

55. Combrink, H.M.v.E., Marivate, V. and Masikisiki, B., Technology-Enhanced Learning, Data Sharing, and Machine Learning Challenges in South African Education. Education. Sciences. 2023, 13, 438. https://doi.org/10.3390/educsci13050438
56. Combrink, H.M.v.E., Carr, E. de Wet, Katinka. Marivate, V. and Rosman, B., 2023b. South Africa Education Data and Visualisations. University of the Free State. Dataset. Data in Brief https://doi.org/10.38140/ufs.22081058.v3
57. Herkulaas, H.M.v.E., and Oosthuizen, L.L., 2020. First-year student transition at the University of the Free State during Covid-19: Challenges and insights. Journal of Student Affairs in Africa, 8(2), pp.31-44. https://doi.org/10.24085/jsaa.v8i2.4446
58. Combrink, H.M.v.E., and Oosthuizen, L.L., 2022. Strengthening online teaching and learning by closing the feedback loop. Southern African Review of Education with Education with Production, 27(1), pp.57-76. https://hdl.handle.net/10520/ejc-sare_v27_n1_a4
59. Conboy, K. and Carroll, N., 2019. Implementing large-scale agile frameworks: challenges and recommendations. IEEE Software, 36(2), pp.44-50. https://doi.org/10.1109/MS.2018.2884865
60. Coussement, K., Phan, M., De Caigny, A., Benoit, D.F. and Raes, A., 2020. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. Decision Support Systems, 135(1), p.113325. https://doi.org/10.1016/j.dss.2020.113325
61. Cranfield, D.J., Tick, A., Venter, I.M., Blignaut, R.J. and Renaud, K., 2021. Higher education students' perceptions of online learning during COVID-19—A comparative study. Education Sciences, 11(8), p.403. https://doi.org/10.3390/educsci11080403
62. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B. and Bharath, A.A., 2018. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, 35(1), pp.53-65. https://doi.org/10.1109/MSP.2017.2765202
63. Cruz, J.A. and Wishart, D.S., 2006. Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2(1), p.117693510600200030.
64. Cunningham, P. and Delany, S.J., 2021. k-Nearest neighbour classifiers-A Tutorial. ACM Computing Surveys (CSUR), 54(6), pp.1-25. https://doi.org/10.1145/3459665

65. Cunningham, P. and Delany, S.J., 2021. Underestimation bias and underfitting in machine learning. In Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers 1, 1(1), pp. 20-31. Springer International Publishing. https://doi.org/10.1007/978-3-030-73959-1_2
66. Dahmen, J. and Cook, D., 2019. SynSys: A synthetic data generation system for healthcare applications. Sensors, 19(5), p.1181. https://doi.org/10.3390/s19051181
67. Daly, R., Shen, Q. and Aitken, S., 2011. Learning Bayesian networks: approaches and issues. The knowledge engineering review, 26(2), pp.99-157. https://doi.org/10.1017/S0269888910000251
68. Das, B.K., Jha, D.N., Sahu, S.K., Yadav, A.K., Raman, R.K. and Kartikeyan, M., 2022. Chi-Square Test of Significance. In Concept Building in Fisheries Data Analysis (pp. 81-94). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-4411-6_5
69. Date, P., Arthur, D. and Pusey-Nazzaro, L., 2021. QUBO formulations for training machine learning models. Scientific reports, 11(1), p.10029. https://doi.org/10.1038/s41598-021-89461-4
70. De Waal, A. and Yoo, K., 2018, July. Latent variable Bayesian networks constructed using structural equation modelling. In 2018 21st International Conference on Information Fusion (FUSION), 1(1), pp. 688-695. IEEE. https://ieeexplore.ieee.org/abstract/document/8455240
71. de Waal, A., Koen, H., de Villiers, P., Roodt, H., Moorosi, N. and Pavlin, G., 2016, July. Construction and evaluation of Bayesian networks with expert-defined latent variables. In 2016 19th international conference on information fusion (fusion),1(1), pp. 774-781. IEEE. Available at: https://ieeexplore.ieee.org/abstract/document/7527965?casa_token=AZN6DMhlXpYAAAAA:L4CKTbyxKdw7et_2GeaqfwC1W0asgGKbQHkyA1a_p35BS7LkuML0WsloCrlJh1aX6Xdcn0JVIArQ
72. Derks, I.P. and De Waal, A., 2020. A taxonomy of explainable Bayesian networks. In Artificial Intelligence Research: First Southern African Conference for AI Research, SACAIR 2020, Muldersdrift, South Africa, February 22-26, 2021, Proceedings 1, 1(1), pp.

220-235. Springer International Publishing. https://doi.org/10.1007/978-3-030-66151-9_14

73. Deumert, A., 2010. Tracking the demographics of (urban) language shift–an analysis of South African census data. Journal of Multilingual and Multicultural Development, 31(1), pp.13-35. https://doi.org/10.1080/01434630903215125
74. Dominguez-Whitehead, Y. and Sing, N., 2015. International students in the South African higher education system: A review of pressing challenges. South African Journal of Higher Education, 29(4), pp.77-95. https://hdl.handle.net/10520/EJC182455
75. Dong, S., Wang, P. and Abbas, K., 2021. A survey on deep learning and its applications. Computer Science Review, 40(1), p.100379. https://doi.org/10.1016/j.cosrev.2021.100379
76. Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5-6), pp.352-359. https://doi.org/10.1016/S1532-0464(03)00034-0
77. Du Plessis, P. and Mestry, R., 2019. Teachers for rural schools–a challenge for South Africa. South African Journal of Education, 39(1), pp. 13-27. https://doi.org/10.15700/saje.v39ns1a1774
78. Edwards, D., De Abreu, G.C. and Labouriau, R., 2010. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC Bioinformatics, 11(1), pp.1-13. https://doi.org/10.1186/1471-2105-11-18
79. Eickhoff, J., Zaborek, J., Chen, G., Sahasrabuddhe, V.V., Ford, L.G., Szabo, E. and Kim, K., 2023. A systematic review and pooled analysis of hypothesized versus observed effect sizes in early phase cancer prevention clinical trials. Cancer Prevention Research, pp.CAPR-23. https://doi.org/10.1158/1940-6207.CAPR-23-0060
80. Engelmann, J. and Lessmann, S., 2021. Conditional Wasserstein GAN-based over sampling of tabular data for imbalanced learning. Expert Systems with Applications, 174(1), p.114582. https://doi.org/10.1016/j.eswa.2021.114582
81. El Emam, K., Mosquera, L., Jonker, E. and Sood, H., 2021. Evaluating the utility of synthetic COVID-19 case data. JAMIA open, 4(1), p.ooab012. https://doi.org/10.1093/jamiaopen/ooab012

82. Ertekin, S., Huang, J., Bottou, L. and Giles, L., 2007, November. Learning on the border: active learning in imbalanced data classification. In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management (pp. 127-136). https://doi.org/10.1145/1321440.1321461
83. Eudy, C. and Brooks, S., 2022. Factors impacting student success in a fundamentals course of an associate degree nursing program. Teaching and Learning in Nursing, 17(1), pp.11-16.https://doi.org/10.1016/j.teln.2021.05.004
84. Fierro-Saltos, W., Sanz, C., Zangara, A., Guevara, C., Arias-Flores, H., Castillo-Salazar, D., Varela-Aldás, J., Borja-Galeas, C., Rivera, R., Hidalgo-Guijarro, J. and Yandún-Velasteguí, M., 2020. Autonomous learning mediated by digital technology processes in higher education: A systematic review. In Human Systems Engineering and Design II: Proceedings of the 2nd International Conference on Human Systems Engineering and Design (IHSED2019): Future Trends and Applications, September 16-18, 2019, Universität der Bundeswehr München, Munich, Germany, 1(1), pp. 65-71. Springer International Publishing. https://doi.org/10.1007/978-3-030-27928-8_11
85. Figueira, A. and Vaz, B., 2022. Survey on synthetic data generation, evaluation methods and GANs. Mathematics, 10(15), p.2733. https://doi.org/10.3390/math10152733
86. Foote, J., Midgley, G., Ahuriri-Driscoll, A., Hepi, M. and Earl-Goulet, J., 2021. Systemic evaluation of community environmental management programmes. European Journal of Operational Research, 288(1), pp.207-224. https://doi.org/10.1016/j.ejor.2020.05.019
87. Francis, D. and Webster, E., 2019. Poverty and inequality in South Africa: Critical reflections. Development Southern Africa, 36(6), pp.788-802. https://doi.org/10.1080/0376835X.2019.1666703
88. Gallagher, S.E. and Savage, T., 2020. Challenge-based learning in higher education: an exploratory literature review. Teaching in Higher Education, 7(1), pp.1-23. https://doi.org/10.1080/13562517.2020.1863354
89. Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. and Suganthan, P.N., 2022. Ensemble deep learning: A review. Engineering Applications of Artificial Intelligence, 115(1), p.105151.https://doi.org/10.1016/j.engappai.2022.105151

90. García-Peñalvo, F.J., 2021. Avoiding the dark side of digital transformation in teaching. An institutional reference framework for eLearning in higher education. Sustainability, 13(4), p.2023. https://doi.org/10.3390/su13042023
91. Gates, E.F., Walton, M., Vidueira, P. and McNall, M., 2021. Introducing systems-and complexity-informed evaluation. New Directions for Evaluation, 2021(170), pp.13-25. https://doi.org/10.1002/ev.20466
92. Gefenas, E., Lekstutiene, J., Lukaseviciene, V., Hartlev, M., Mourby, M. and Cathaoir, K.Ó., 2022. Controversies between regulations of research ethics and protection of personal data: informed consent at a cross-road. Medicine, Health Care and Philosophy, pp.1-8. https://doi.org/10.1007/s11019-021-10060-1
93. Ghaffarzadegan, N., Larson, R. and Hawley, J., 2017. Education as a complex system. Systems Research and Behavioral Science, 34(3), p.211. https://doi.org/10.1002/sres.2405
94. Ghasemian, A., Hosseinmardi, H. and Clauset, A., 2019. Evaluating overfit and underfit in models of network community structure. IEEE Transactions on Knowledge and Data Engineering, 32(9), pp.1722-1735. https://doi.org/10.1109/TKDE.2019.2911585
95. Ghatak, D. and Sakurai, K., 2023, January. A Survey on Privacy Preserving Synthetic Data Generation and a Discussion on a Privacy-Utility Trade-off Problem. In Science of Cyber Security-SciSec 2022 Workshops: AI-CryptoSec, TA-BC-NFT, and MathSci-Qsafe 2022, Matsue, Japan, August 10–12, 2022, Revised Selected Papers, 1(1), pp. 167-180. Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7769-5_13
96. González-Zamar, M.D., Abad-Segura, E., López-Meneses, E. and Gómez-Galán, J., 2020. Managing ICT for sustainable education: Research analysis in the context of higher education. Sustainability, 12(19), p.8254. https://doi.org/10.3390/su12198254
97. Gottschall, J., Peinke, J., Lippens, V. and Nagel, V., 2009. Exploring the dynamics of balance data—movement variability in terms of drift and diffusion. Physics Letters A, 373(8-9), pp.811-816. https://doi.org/10.1016/j.physleta.2008.12.026
98. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM, 63(11), pp.139-144. https://doi.org/10.1145/3422622

99. Goutte, C. and Gaussier, E., 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In Advances in Information Retrieval: 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005. Proceedings 27, 1(1), pp. 345-359. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_25

100. Granić, A. and Marangunić, N., 2019. Technology acceptance model in educational context: A systematic literature review. British Journal of Educational Technology, 50(5), pp.2572-2593. https://doi.org/10.1111/bjet.12864

101. Grundkiewicz, R., Junczys-Dowmunt, M. and Heafield, K., 2019, August. Neural grammatical error correction systems with unsupervised pre-training on synthetic data. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 1(1), pp. 252-263. https://doi.org/10.18653/v1/W19-4427

102. Hagenauer, J. and Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications, 78(1), pp.273-282. https://doi.org/10.1016/j.eswa.2017.01.057

103. Haghighi, S., Jasemi, M., Hessabi, S. and Zolanvari, A., 2018. PyCM: Multiclass confusion matrix library in Python. Journal of Open Source Software, 3(25), p.729. https://doi.org/10.21105/joss.00729

104. Hall, B.L., Taylor, C.J., Labes, R., Massey, A.F., Menzel, R., Bourne, R.A. and Chamberlain, T.W., 2021. Autonomous optimisation of a nanoparticle catalysed reduction reaction in continuous flow. Chemical Communications, 57(40), pp.4926-4929. https://doi.org/10.1039/D1CC00859E

105. Han, H., 2019. Design and implementation of web-based English autonomous learning system. International Journal of Emerging Technologies in Learning (Online), 14(6), p.18. https://doi.org/10.3991/ijet.v14i06.9718

106. Hardman, J., 2019. Towards a pedagogical model of teaching with ICTs for mathematics attainment in primary school: A review of studies 2008–2018. Heliyon, 5(5), p.e01726.https://doi.org/10.1016/j.heliyon.2019.e01726

107. Harris, J.C. and Patton, L.D., 2019. Un/doing intersectionality through higher education research. The Journal of Higher Education, 90(3), pp.347-372. https://doi.org/10.1080/00221546.2018.1536936

108. Hart, G.L., Mueller, T., Toher, C. and Curtarolo, S., 2021. Machine learning for alloys. Nature Reviews Materials, 6(8), pp.730-755. https://doi.org/10.1007/s12525-021-00475-2

109. Hart, S.A. and Laher, S., 2015. Perceived usefulness and culture as predictors of teachers attitudes towards educational technology in South Africa. South African Journal of Education, 35(4). https://doi.org/10.15700/saje.v35n4a1180

110. Heinze-Deml, C. and Meinshausen, N., 2021. Conditional variance penalties and domain shift robustness. Machine Learning, 110(2), pp.303-348.DOIhttps://doi.org/10.1007/s10994-020-05924-1

111. Hermans, T., 2019. Translation in systems: Descriptive and systemic approaches explained. Routledge. Available at: https://books.google.co.za/books?hl=en&lr=&id=gnCdDwAAQBAJ&oi=fnd&pg=PT8&dq=A+systems+approach+explained&ots=MWppl7Zm_K&sig=c4hEvF3kpE-1AOEM8dJEqftuISc&redir_esc=y#v=onepage&q=A%20systems%20approach%20explained&f=false

112. Hernandez, M., Epelde, G., Alberdi, A., Cilla, R. and Rankin, D., 2022. Synthetic data generation for tabular health records: A systematic review. Neurocomputing. https://doi.org/10.1016/j.neucom.2022.04.053

113. Hossain, M., 2021. Unequal experience of COVID-induced remote schooling in four developing countries. International Journal of Educational Development, 85(1), p.102446. https://doi.org/10.1016/j.ijedudev.2021.102446

114. Huberts, L.C., Schoonhoven, M. and Does, R.J., 2022. Multilevel process monitoring: A case study to predict student success or failure. Journal of Quality Technology, 54(2), pp.127-143. https://doi.org/10.1080/00224065.2020.1828008

115. Hong, Y., Hwang, U., Yoo, J. and Yoon, S., 2019. How generative adversarial networks and their variants work: An overview. ACM Computing Surveys (CSUR), 52(1), pp.1-43. https://doi.org/10.1145/3301282

116. Imran, M., Latif, S., Mehmood, D. and Shah, M.S., 2019. Student academic performance prediction using supervised learning techniques. International Journal of Emerging Technologies in Learning, 14(14). https://doi.org/10.3991/ijet.v14i14.10310

117. Irvin, J., Zhou, S., McNicol, G., Lu, F., Liu, V., Fluet-Chouinard, E., Ouyang, Z., Knox, S.H., Lucas-Moffat, A., Trotta, C. and Papale, D., 2021. Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. Agricultural and forest meteorology, 308, p.108528. https://doi.org/10.1016/j.agrformet.2021.108528
118. Jackson, J.K., Huerta, M. and Garza, T., 2020. A promising science and literacy instructional model with Hispanic fifth grade students. The Journal of Educational Research, 113(2), pp.79-92. https://doi.org/10.1080/00220671.2020.1728734
119. Jacobson, M.J., Levin, J.A. and Kapur, M., 2019. Education as a complex system: Conceptual and methodological implications. Educational Researcher, 48(2), pp.112-119. https://doi.org/10.3102/0013189X19826958
120. Jadhav, A., Mostafa, S.M., Elmannai, H. and Karim, F.K., 2022. An Empirical Assessment of Performance of Data Balancing Techniques in Classification Task. Applied Sciences, 12(8), p.3928. https://doi.org/10.3390/app12083928
121. Janse van Vuuren, E.C., 2020. Development of a contextualised data analytics framework in South African higher education: Evolvement of teacher (teaching) analytics as an indispensable component. South African Journal of Higher Education, 34(1), pp.137-157. https://hdl.handle.net/10520/EJC-1e1a0881b0
122. Jensen, F.V., 2009. Bayesian networks. Wiley Interdisciplinary Reviews: Computational Statistics, 1(3), pp.307-315. https://doi.org/10.1002/wics.48
123. Jia, J., Jin, G.Z. and Wagman, L., 2021. The short-run effects of the general data protection regulation on technology venture investment. Marketing Science, 40(4), pp.661-684. https://doi.org/10.1287/mksc.2020.1271
124. Joubert, J.W. and De Waal, A., 2020. Activity-based travel demand generation using Bayesian networks. Transportation Research Part C: Emerging Technologies, 120(1), p.102804. https://doi.org/10.1016/j.trc.2020.102804
125. Kabir, S. and Papadopoulos, Y., 2019. Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. Safety Science, 115(1), pp.154-175. https://doi.org/10.1016/j.ssci.2019.02.009

126. Kadiyala, A. and Kumar, A., 2018. Applications of python to evaluate the performance of decision tree-based boosting algorithms. Environmental Progress and Sustainable Energy, 37(2), pp.618-623. https://doi.org/10.1002/ep.12888
127. Kaikkonen, L., Parviainen, T., Rahikainen, M., Uusitalo, L. and Lehikoinen, A., 2021. Bayesian networks in environmental risk assessment: A review. Integrated Environmental Assessment and Management, 17(1), pp.62-78. https://doi.org/10.1002/ieam.4332
128. Kaneko, M. and Bollegala, D., 2022, June. Unmasking the mask–evaluating social biases in masked language models. In Proceedings of the AAAI Conference on Artificial Intelligence , 36(11), pp. 11954-11962. https://doi.org/10.1609/aaai.v36i11.21453
129. Kang, P., 2021, April. Design of English Autonomous Learning System Based on Digital Signal Processing. In Journal of Physics: Conference Series 1881(4), p. 042065. IOP Publishing. https://doi.org/10.1088/1742-6596/1881/4/042065
130. Kearney, C.A., 2021, August. Integrating systemic and analytic approaches to school attendance problems: Synergistic frameworks for research and policy directions. In Child and Youth Care Forum, 50(1), pp. 701-742. Springer US. https://doi.org/10.1007/s10566-020-09591-0
131. Kerstens, K. and van de Woestyne, I., 2014. Comparing Malmquist and Hicks–Moorsteen productivity indices: Exploring the impact of unbalanced vs. balanced panel data. European Journal of Operational Research, 233(3), pp.749-758. https://doi.org/10.1016/j.ejor.2013.09.009
132. Kersting, K. and De Raedt, L., 2001. Towards combining inductive logic programming with Bayesian networks. In Inductive Logic Programming: 11th International Conference, ILP 2001 Strasbourg, France, September 9–11, 2001 Proceedings 11, 1(1), pp. 118-131. Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-44797-0_10
133. Kezar, A., 2005. Redesigning for collaboration within higher education institutions: An exploration into the developmental process. Research in Higher Education, 46(1), pp.831-860. https://doi.org/10.1007/s11162-004-6227-5
134. Khanal, S.S., Prasad, P.W.C., Alsadoon, A. and Maag, A., 2020. A systematic review: machine learning based recommendation systems for e-learning. Education and Information Technologies, 25(1), pp.2635-2664. https://doi.org/10.1007/s10639-019-10063-9

135. Kinkle, R.M., 2020. Factors influencing student success in associate degree respiratory therapy programs. Health Professions Education, 6(3), pp.343-353. https://doi.org/10.1016/j.hpe.2020.06.003
136. Kiran, R., Kumar, P. and Bhasker, B., 2020. DNNRec: A novel deep learning based hybrid recommender system. Expert Systems with Applications, 144(1), p.113054. https://doi.org/10.1016/j.eswa.2019.113054
137. Konig, J., 2021. The clock is ticking for compliance with POPIA. Without Prejudice, 21(2), pp.21-22.https://hdl.handle.net/10520/ejc-jb_prej_v21_n2_a11
138. Kotsiantis, S.B., Zaharakis, I.D. and Pintelas, P.E., 2006. Machine learning: a review of classification and combining techniques. Artificial Intelligence Review, 26(3), pp.159-190. https://doi.org/10.1007/s10462-007-9052-3
139. Kreth, Q., Spirou, M.E., Budenstein, S. and Melkers, J., 2019. How prior experience and self-efficacy shape graduate student perceptions of an online learning environment in computing. Computer Science Education, 29(4), pp.357-381. https://doi.org/10.1080/08993408.2019.1601459
140. Kirchner, K.A.T.R.I.N., Tölle, K.H. and Krieter, J., 2006. Optimisation of the decision tree technique applied to simulated sow herd datasets. Computers and electronics in agriculture, 50(1), pp.15-24. https://doi.org/10.1016/j.compag.2005.07.002
141. Kruschke, J.K., 2013. Bayesian estimation supersedes the t test. Journal of Experimental Psychology: General, 142(2), p.573. https://doi.org/10.1037/a0029146
142. Ladyman, J., Lambert, J. and Wiesner, K., 2013. What is a complex system?. European Journal for Philosophy of Science, 3(1), pp.33-67. https://doi.org/10.1007/s13194-012-0056-8
143. Larrabee Sønderlund, A., Hughes, E. and Smith, J., 2019. The efficacy of learning analytics interventions in higher education: A systematic review. British Journal of Educational Technology, 50(5), pp.2594-2618. https://doi.org/10.1111/bjet.12720
144. Lau, E.T., Sun, L. and Yang, Q., 2019. Modelling, prediction and classification of student academic performance using artificial neural networks. SN Applied Sciences, 1(1), pp.1-10. https://doi.org/10.1007/s42452-019-0884-7
145. LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep Learning. Nature, 521(7553), pp.436-444. https://doi.org/10.1038/nature14539

146. Lee, I. and Shin, Y.J., 2020. Machine learning for enterprises: Applications, algorithm selection, and challenges. Business Horizons, 63(2), pp.157-170. https://doi.org/10.1016/j.bushor.2019.10.005

147. Lemke, J., & Sabelli, N. (2008). Complex systems and educational change: Towards a new research agenda. Educational Philosophy and Theory, 40(1), 118–129. https://doi.org/10.1111/j.1469-5812.2007.00401.x

148. Leung, W.T.V., Tam, T.Y.T., Pan, W.C., Wu, C.D., Lung, S.C.C. and Spengler, J.D., 2019. How is environmental greenness related to students' academic performance in English and Mathematics?. Landscape and Urban Planning, 181(1), pp.118-124. https://doi.org/10.1016/j.landurbplan.2018.09.021

149. Li, P., Sofuoglu, S.E., Aviyente, S. and Maiti, T., 2022. Coupled support tensor machine classification for multimodal neuroimaging data. Statistical Analysis and Data Mining: The ASA Data Science Journal, 15(6), pp.797-818. https://doi.org/10.1002/sam.11587

150. Li, Y., Hu, Q. and Li, N., 2020. A reliability-aware multi-armed bandit approach to learn and select users in demand response. Automatica, 119(1), p.109015. https://doi.org/10.1016/j.automatica.2020.109015

151. Liu, H., Liu, L. and Zhang, H., 2010. Ensemble gene selection for cancer classification. Pattern Recognition, 43(8), pp.2763-2772. https://doi.org/10.1016/j.patcog.2010.02.008

152. Lourens, A. and Bleazard, D., 2016. Applying predictive analytics in identifying students at risk: A case study. South African Journal of Higher Education, 30(2), pp.129-142. https://hdl.handle.net/10520/EJC191693

153. Lu, N.Y., Zhang, K. and Yuan, C., 2021, May. Improving causal discovery by optimal bayesian network learning. In Proceedings of the AAAI Conference on Artificial Intelligence, 35(10), pp. 8741-8748. https://doi.org/10.1609/aaai.v35i10.17059

154. Lu, Y., Fang, Y. and Shi, C., 2020, August. Meta-learning on heterogeneous information networks for cold-start recommendation. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1(1), pp. 1563-1573. https://doi.org/10.1145/3394486.3403207

155. Luan, H. and Tsai, C.C., 2021. A review of using machine learning approaches for precision education. Educational Technology and Society, 24(1), pp.250-266. https://www.jstor.org/stable/26977871

156. Luo, J., Huang, J., Ma, J. and Li, H., 2022. An evaluation method of conditional deep convolutional generative adversarial networks for mechanical fault diagnosis. Journal of Vibration and Control, 28(11-12), pp.1379-1389. https://doi.org/10.1177/1077546321993563

157. Luthfi, A., Janssen, M. and Crompvoets, J., 2018, June. A causal explanatory model of Bayesian-belief networks for analysing the risks of opening data. In International Symposium on Business Modeling and Software Design (pp. 289-297). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-94214-8_20

158. Maddock, L. and Maroun, W., 2018. Exploring the present state of South African education: Challenges and recommendations. South African Journal of Higher Education, 32(2), pp.192-214. https://hdl.handle.net/10520/EJC-ed4f7161d

159. Magee, C. and de Weck, O., 2004. Complex system classification. http://hdl.handle.net/1721.1/6753

160. Maghari, A., 2018. Prediction of student's performance using modified KNN classifiers. Prediction of Student's Performance Using Modified KNN Classifiers. In The First International Conference on Engineering and Future Technology (ICEFT 2018) (pp. 143-150). Available at SSRN: https://ssrn.com/abstract=3704733

161. Maistry, S.M. and Africa, I.E., 2020. Neoliberal stratification: The confounding effect of the school poverty quintile ranking system in South Africa. South African Journal of Education, 40(4), pp. 4-23. https://doi.org/10.15700/saje.v40n4a1872

162. Majeed, B.H., Jawad, L.F. and ALRikabi, H.T., 2022. Computational Thinking (CT) Among University Students. International Journal of Interactive Mobile Technologies, 16(10), pp. 1-15. https://doi.org/10.3991/ijim.v16i10.30043

163. Makalima, C., Gwala, Y., Makasi, L., Baza, A. and Lwanga, A.M., 2023, April. Co-designing an Integrated Digital Education Portal for the Eastern Cape Rural Learners. In Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, 1(1), pp. 1-7. https://doi.org/10.1145/3544549.3583839

164. Malherbe, J., 2021. The Protection of Personal Information Act: Its effect on Clinical Practice and Health Research. South African Journal of Occupational Therapy, 51(2), pp.2-3. http://dx.doi.org/10.17159/sajs.2021/9490

165. Manik, S. and Ramrathan, L., 2021. Institutional Leadership Efforts Driving Student Retention and Success: A Case Study of the University of KwaZulu-Natal, South Africa. Student Retention and Success in Higher Education: Institutional Change for the 21st Century, 1(1), pp.109-131. https://doi.org/10.1007/978-3-030-80045-1_6

166. Marković, D., Stojić, H., Schwöbel, S. and Kiebel, S.J., 2021. An empirical evaluation of active inference in multi-armed bandits. Neural Networks, 144(1), pp.229-246. https://doi.org/10.1016/j.neunet.2021.08.018

167. Marthers, P., Herrup, P. and Steele, J., 2015. Consider the costs of student attrition. Enrollment Management Report, 19(6), pp.1-5. https://doi.org/10.1002/emt.30090

168. Martin, F., Chen, Y., Moore, R.L. and Westine, C.D., 2020. Systematic review of adaptive learning research designs, context, strategies, and technologies from 2009 to 2018. Educational Technology Research and Development, 68, pp.1903-1929. https://doi.org/10.1007/s11423-020-09793-2

169. Masutha, M., 2022. Highs, lows and turning points in marginalised transitions and experiences of noncompletion amongst pushed dropouts in South African higher education. Education sciences, 12(9), p.608. https://doi.org/10.3390/educsci12090608

170. McHugh, M.L., 2013. The chi-square test of independence. Biochemia medica, 23(2), pp.143-149. https://doi.org/10.11613/BM.2013.018

171. Mgaga, P. and Scholes, M.C., 2019. Does tertiary education in South Africa equip professional foresters for the future?. Southern Forests: a Journal of Forest Science, 81(4), pp.377-385. https://doi.org/10.2989/20702620.2019.1615230

172. Midgley, G. and Lindhult, E., 2021. A systems perspective on systemic innovation. Systems Research and Behavioral Science, 38(5), pp.635-670. https://doi.org/10.1002/sres.2819

173. Midgley, G. and Rajagopalan, R., 2020. Critical systems thinking, systemic intervention, and beyond. Handbook of Systems Sciences, 1(1), pp.1-51. https://doi.org/10.1007/978-981-13-0370-8_7-1

174. Milano, S., Taddeo, M. and Floridi, L., 2020. Recommender systems and their ethical challenges. Ai and Society, 35(1), pp.957-967. https://doi.org/10.1007/s00146-020-00950-y

175. Mirata, V., Hirt, F., Bergamin, P. and van der Westhuizen, C., 2020. Challenges and contexts in establishing adaptive learning in higher education: findings from a Delphi study. International Journal of Educational Technology in Higher Education, 17(1), pp.1-25. https://doi.org/10.1186/s41239-020-00209-y

176. Monat, J., Amissah, M. and Gannon, T., 2020. Practical applications of systems thinking to business. Systems, 8(2), p.14. https://doi.org/10.3390/systems8020014

177. Montaño-Gutierrez, L.F., Ohta, S., Kustatscher, G., Earnshaw, W.C. and Rappsilber, J., 2017. Nano Random Forests to mine protein complexes and their relationships in quantitative proteomics data. Molecular Biology of The Cell, 28(5), pp.673-680. https://doi.org/10.1091/mbc.e16-06-0370

178. Moodley, P. and Singh, R.J., 2015. Addressing student dropout rates at South African universities. Alternation (Durban). http://hdl.handle.net/10321/1648

179. Morijiri, K., Mihana, T., Kanno, K., Naruse, M. and Uchida, A., 2022. Decision making for large-scale multi-armed bandit problems using bias control of chaotic temporal waveforms in semiconductor lasers. Scientific Reports, 12(1), p.8073. https://doi.org/10.1038/s41598-022-12155-y

180. Motala, S. and Carel, D., 2019. Educational funding and equity in South African schools. South African schooling: The enigma of inequality: A study of the present situation and future possibilities, pp.67-85. https://doi.org/10.1007/978-3-030-18811-5_4

181. Muhammad, I. and Yan, Z., 2015. Supervised machine learning approaches: a survey. ICTACT Journal on Soft Computing, 5(3), p. 15. https://doi.org/10.21917/ijsc.2015.0133

182. Murray, K.T., Merriman, C.S. and Adamson, C., 2008. Use of the HESI admission assessment to predict student success. CIN: Computers, Informatics, Nursing, 26(3), pp.167-172. https://doi.org/10.1097/01.NCN.0000304781.27070.a7

183. Na, S., Heo, S., Han, S., Shin, Y. and Lee, M., 2022. Development of an artificial intelligence model to recognise construction waste by applying image data augmentation and transfer learning. Buildings, 12(2), p.175. https://doi.org/10.3390/buildings12020175

184. Namoun, A. and Alshanqiti, A., 2020. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. Applied Sciences, 11(1), p.237. https://doi.org/10.3390/app11010237

185. Nassar, N., Jafar, A. and Rahhal, Y., 2020. A novel deep multi-criteria collaborative filtering model for recommendation system. Knowledge-Based Systems, 187(1), p.104811. https://doi.org/10.1016/j.knosys.2019.06.019

186. Natarajan, S., Vairavasundaram, S., Natarajan, S. and Gandomi, A.H., 2020. Resolving data sparsity and cold start problem in collaborative filtering recommender system using linked open data. Expert Systems with Applications, 149(1), p.113248. https://doi.org/10.1016/j.eswa.2020.113248

187. Natori, K., Uto, M., Nishiyama, Y., Kawano, S. and Ueno, M., 2015. Constraint-based learning Bayesian networks using Bayes factor. In Advanced Methodologies for Bayesian Networks: Second International Workshop, AMBN 2015, Yokohama, Japan, November 16-18, 2015. Proceedings 2 (pp. 15-31). Springer International Publishing. https://doi.org/10.1007/978-3-319-28379-1_2

188. Nikolenko, S.I., 2021. Synthetic data for deep learning, 174(1). Springer Nature. ISBN: 978-3-030-75178-4 https://doi.org/10.1007/978-3-030-75178-4

189. Olson, R.S., Cava, W.L., Mustahsan, Z., Varik, A. and Moore, J.H., 2018. Data-driven advice for applying machine learning to bioinformatics problems. In Pacific Symposium on Biocomputing 2018: Proceedings of the Pacific Symposium (pp. 192-203). https://doi.org/10.1142/9789813235533_0018

190. Ogbonna, C.G., Ibezim, N.E. and Obi, C.A., 2019. Synchronous versus asynchronous e-learning in teaching word processing: An experimental approach. South African Journal of Education, 39(2), pp.1-15.https://hdl.handle.net/10520/EJC-168a98cd12

191. Ogbonnaya, U.I. and Awuah, F.K., 2019. Quintile ranking of schools in south africa and learners'achievement in probability. Statistics Education Research Journal, 18(1), pp.106-119.https://doi.org/10.52041/serj.v18i1.153

192. Olivier, E., Morin, A.J., Langlois, J., Tardif-Grenier, K. and Archambault, I., 2020. Internalizing and externalizing behavior problems and student engagement in elementary and secondary school students. Journal Of Youth and Adolescence, 49(1), pp.2327-2346. https://doi.org/10.1007/s10964-020-01295-x

193. Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K., 2023. Classification and prediction of student performance data using various machine learning algorithms. Materials Today: Proceedings, 80(1), pp.3782-3785. https://doi.org/10.1016/j.matpr.2021.07.382

194. Pavlin G, Jousselme AL, de Villiers JP, Costa PC, Laskey K, Mignet F, de Waal A. Online system evaluation and learning of data source models: A probabilistic generative approach. In2019 22th International Conference on Information Fusion (FUSION) 2019 Jul 2, 1(1), pp. 1-10. IEEE. https://ieeexplore.ieee.org/abstract/document/9011396

195. Pellicer, M. and Piraino, P., 2019. The effect of nonpersonnel resources on educational outcomes: Evidence from South Africa. Economic Development and Cultural Change, 67(4), pp.907-934. https://doi.org/10.1086/700103

196. Peterson, L.L. and Davie, B.S., 2007. Computer networks: a systems approach. Elsevier. Available at: https://books.google.co.za/books?hl=en&lr=&id=pspnGbHFGLcC&oi=fnd&pg=PP1&dq=A+systems+approach+explained&ots=70bFTKX-RR&sig=tC9lC5NGPzNtHki0m30iPcpjS_Y&redir_esc=y#v=onepage&q=A%20systems%20approach%20explained&f=false

197. Pickard, A.C. and Beasley, R., 2022, July. Engineering Complicated Systems Still Needs Systems Engineering and Thinking. In INCOSE International Symposium, 32(1), pp. 721-736). https://doi.org/10.1002/iis2.12960

198. Platz, M. and Platz, M., 2021. Epistemic Aims of School Education. Good Relationships in Schools: Teachers, Students, and the Epistemic Aims of Education, 1(1), pp.9-26. https://doi.org/10.1007/978-3-662-64137-8_2

199. Posel, D., Casale, D. and Grapsa, E., 2020. Household variation and inequality: The implications of equivalence scales in South Africa. African Review of Economics and Finance, 12(1), pp.102-122. https://hdl.handle.net/10520/EJC-1d056a3683

200. Pritchard, M.E. and Wilson, G.S., 2003. Using emotional and social factors to predict student success. Journal of College Student Development, 44(1), pp.18-28. 10.1353/csd.2003.0008

201. Qian, C., Zheng, B., Shen, Y., Jing, L., Li, E., Shen, L. and Chen, H., 2020. Deep-learning-enabled self-adaptive microwave cloak without human intervention. Nature Photonics, 14(6), pp.383-390. https://doi.org/10.1038/s41566-020-0604-2

202. Quijano-Sánchez, L., Cantador, I., Cortés-Cediel, M.E. and Gil, O., 2020. Recommender systems for smart cities. Information Systems, 92(1), p.101545. https://doi.org/10.1016/j.is.2020.101545

203. Rasheed, R.A., Kamsin, A. and Abdullah, N.A., 2020. Challenges in the online component of blended learning: A systematic review. Computers and Education, 144(1), p.103701.https://doi.org/10.1016/j.compedu.2019.103701

204. Rauchas, S., Rosman, B., Konidaris, G. and Sanders, I., 2006. Language performance at high school and success in first year computer science. ACM SIGCSE Bulletin, 38(1), pp.398-402. https://doi.org/10.1145/1124706.1121467

205. Rosman, B., Hawasly, M. and Ramamoorthy, S., 2016. Bayesian policy reuse. Machine Learning, 104(1), pp.99-127. https://doi.org/10.1007/s10994-016-5547-y

206. Roshanski, I., Kalech, M. and Rokach, L., 2023. Automatic Feature Engineering for Learning Compact Decision Trees. Expert Systems with Applications, 229, p.120470. https://doi.org/10.1016/j.eswa.2023.120470

207. Rothgang, M. and Lageman, B., 2022. Systems Analysis in Evaluation: The unfulfilled promise. Journal for Research and Technology Policy Evaluation, (53), pp.181-191. https://doi.org/10.22163/fteval.2022.556

208. Rotar, O., 2022. Online student support: A framework for embedding support interventions into the online learning cycle. Research and Practice in Technology Enhanced Learning, 17(1), pp.1-23. https://doi.org/10.1186/s41039-021-00178-4

209. Ruswa, A.S. and Gore, O.T., 2022. Rethinking student poverty: perspectives from a higher education institution in South Africa. Higher Education Research and Development, 41(7), pp.2353-2366. https://doi.org/10.1080/07294360.2021.2014409

210. San, C.K. and Guo, H., 2023. Institutional support, social support, and academic performance: mediating role of academic adaptation. European Journal of Psychology of Education, 38(4), pp.1659-1675. https://doi.org/10.1007/s10212-022-00657-2

211. Samitas, A., Kampouris, E. and Kenourgios, D., 2020. Machine learning as an early warning system to predict financial crisis. International Review of Financial Analysis, 71(1), p.101507. https://doi.org/10.1016/j.irfa.2020.101507

212. Sarwat, S., Ullah, N., Sadiq, S., Saleem, R., Umer, M., Eshmawi, A.A., Mohamed, A. and Ashraf, I., 2022. Predicting Students' Academic Performance with Conditional Generative Adversarial Network and Deep SVM. Sensors, 22(13), p.4834. https://doi.org/10.3390/s22134834

213. Sayed, Y., Motala, S., Carel, D. and Ahmed, R., 2020. School governance and funding policy in South Africa: Towards social justice and equity in education policy. South African Journal of Education, 40(4). https://doi.org/10.15700/saje.v40n4a2045

214. Scanagatta, M., Salmerón, A. and Stella, F., 2019. A survey on Bayesian network structure learning from data. Progress in Artificial Intelligence, 8(1), pp.425-439. https://doi.org/10.1007/s13748-019-00194-y

215. Schoeman, M., Loots, S. and Bezuidenhoud, L., 2021. Merging Academic and Career Advising to Offer Holistic Student Support: A University Perspective. Journal of Student Affairs in Africa, 9(2), pp.85-100. https://doi.org/10.24085/jsaa.v9i2.3700

216. Scoones, I., Stirling, A., Abrol, D., Atela, J., Charli-Joseph, L., Eakin, H., Ely, A., Olsson, P., Pereira, L., Priya, R. and Van Zwanenberg, P., 2020. Transformations to sustainability: combining structural, systemic and enabling approaches. Current Opinion in Environmental Sustainability, 42(1), pp.65-75. https://doi.org/10.1016/j.cosust.2019.12.004

217. Shaw, L., Kiegaldie, D. and Farlie, M.K., 2020. Education interventions for health professionals on falls prevention in health care settings: a 10-year scoping review. BMC Geriatrics, 20(1), pp.1-13. https://doi.org/10.1186/s12877-020-01819-x

218. Shemshack, A. and Spector, J.M., 2020. A systematic literature review of personalized learning terms. Smart Learning Environments, 7(1), pp.1-20. https://rdcu.be/dxHYK

219. Shin, J., Badgwell, T.A., Liu, K.H. and Lee, J.H., 2019. Reinforcement learning–overview of recent progress and implications for process control. Computers and Chemical Engineering, 127(1), pp.282-294. https://doi.org/10.1016/j.compchemeng.2019.05.029

220. Shwartz-Ziv, R. and Armon, A., 2022. Tabular data: Deep learning is not all you need. Information Fusion, 81, pp.84-90. https://doi.org/10.1016/j.inffus.2021.11.011

221. Smith, A.W., Rae, I.J., Forsyth, C., Oliveira, D.M., Freeman, M.P. and Jackson, D.R., 2020. Probabilistic forecasts of storm sudden commencements from interplanetary shocks using machine learning. Space Weather, 18(11), p.e2020SW002603.https://doi.org/10.1029/2020SW002603

222. Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130. https://doi.org/10.11919/j.issn.1002-0829.215044

223. Srifi, M., Oussous, A., Ait Lahcen, A. and Mouline, S., 2020. Recommender systems based on collaborative filtering using review texts—a survey. Information, 11(6), p.317. https://doi.org/10.3390/info11060317

224. Stephens, M.E., O'Neal, C.M., Westrup, A.M., Muhammad, F.Y., McKenzie, D.M., Fagg, A.H. and Smith, Z.A., 2022. Utility of machine learning algorithms in degenerative cervical and lumbar spine disease: a systematic review. Neurosurgical Review, pp.1-14. https://doi.org/10.1007/s10143-021-01624-z

225. Stoffelsma, L. and Spooren, W., 2019. The relationship between English reading proficiency and academic achievement of first-year science and mathematics students in a multilingual context. International Journal of Science and Mathematics Education, 17(1), pp.905-922. https://doi.org/10.1007/s10763-018-9905-z

226. Stone, D.C., 2021. Student success and the high school-university transition: 100 years of chemistry education research. Chemistry Education Research and Practice, 22(3), pp.579-601. https://doi.org/10.1039/D1RP00085C

227. Strydom, F. and Loots, S., 2020. The student voice as contributor to quality education through institutional design. South African Journal of Higher Education, 34(5), pp.20-34.https://hdl.handle.net/10520/ejc-high-v34-n5-a2

228. Strydom, F., Kuh, G. and Mentz, M., 2010. Enhancing success in South Africa's higher education: Measuring student engagement. Acta Academica, 42(1), pp.259-278. https://hdl.handle.net/10520/EJC15471

229. Sutton, R.S. and Barto, A.G., 1999. Reinforcement learning. Journal of Cognitive Neuroscience, 11(1), pp.126-134. https://doi.org/10.1162/089892999563184

230. Sutton, R.S. and Barto, A.G., 2018. Reinforcement learning: An introduction. MIT Press. Available at:

https://books.google.co.za/books?hl=en&lr=&id=uWV0DwAAQBAJ&oi=fnd&pg=PR7&dq=reinforcement+learning+sutton&ots=mivJq2Z4k1&sig=UuonntKgBaJxFI1QlIoXAdvvTbw&redir_esc=y#v=onepage&q=reinforcement%20learning%20sutton&f=false

231. Suyatinov, S.I., 2020. Educational Laboratory Complex for the Study of Complicated Systems. In ITM Web of Conferences, 35(1), p. 01018. EDP Sciences. https://doi.org/10.1051/itmconf/20203501018
232. Svítek, M., 2015. Towards complex system theory. Neural Network World, 25(1), p.5. https://doi.org/10.14311/NNW.2015.25.001
233. Swales, L., Thaldar, D. and Donnelly, D.L., 2022. Why research institutions should indemnify researchers against POPIA civil liability. South African Journal of Science, 118(3-4), 12(1), pp.22-24. https://hdl.handle.net/10520/ejc-sajsci-v118-n3-a9
234. Szłapczyński, R. and Ghaemi, H., 2019. Framework of an evolutionary multi-objective optimisation method for planning a safe trajectory for a marine autonomous surface ship. Polish Maritime Research, 26(4), pp.69-79. https://doi.org/10.2478/pomr-2019-0068
235. Talvitie, T., Eggeling, R. and Koivisto, M., 2019. Learning Bayesian networks with local structure, mixed variables, and exact algorithms. International Journal of Approximate Reasoning, 115(1), pp.69-95. https://doi.org/10.1016/j.ijar.2019.09.002
236. Thompson, G., Aizawa, I., Curle, S. and Rose, H., 2022. Exploring the role of self-efficacy beliefs and learner success in English medium instruction. International Journal of Bilingual Education and Bilingualism, 25(1), pp.196-209. https://doi.org/10.1080/13670050.2019.1651819
237. Tiroyabone, G.W. and Strydom, F., 2021. The development of academic advising to enable student success in South Africa. Journal of Student Affairs in Africa, 9(2), pp.1-16. https://doi.org/10.24085/jsaa.v9i2.3656
238. Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B. and Bowman, D., 2016. Application of machine learning to construction injury prediction. Automation in Construction, 69(1), pp.102-114. https://doi.org/10.1016/j.autcon.2016.05.016
239. Tonetto, M.S. and Saurin, T.A., 2021. Choosing fall protection systems in construction sites: Coping with complex rather than complicated systems. Safety Science, 143(1), p.105412. https://doi.org/10.1016/j.ssci.2021.105412

240. Topuz, K., Davazdahemami, B. and Delen, D., 2023. A Bayesian belief network-based analytics methodology for early-stage risk detection of novel diseases. Annals of Operations Research, 6(1), pp.1-25. https://doi.org/10.1007/s10479-023-05377-4

241. Tsamardinos, I. and Borboudakis, G., 2010. Permutation testing improves Bayesian network learning. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21, pp. 322-337. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15939-8_21

242. Uleanya, C. and Ke, Y., 2019. Review of preparedness of rural African communities nexus formal education in the fourth industrial revolution. South African Review of Sociology, 50(3-4), pp.91-103. https://doi.org/10.1080/21528586.2019.1639074

243. Ulrich, W., 2012. Operational research and critical systems thinking—an integrated perspective: Part 1: OR as applied systems thinking. Journal of the Operational Research Society, 63(1), pp.1228-1247. https://doi.org/10.1057/jors.2011.141

244. Umami, I. and Rahmawati, L., 2021. Comparing Epsilon greedy and Thompson sampling model for multi-armed bandit algorithm on marketing dataset. Journal of Applied Data Sciences, 2(2). https://doi.org/10.47738/jads.v2i2.28

245. Uusitalo, L., 2007. Advantages and challenges of Bayesian networks in environmental modelling. Ecological Modelling, 203(3-4), pp.312-318. https://doi.org/10.1016/j.ecolmodel.2006.11.033

246. van Aardt, P., 2019. Using students' creative writing towards decolonising an Academic Literacy curriculum. Journal of Decolonising Disciplines, 1(2). https://doi.org/10.35293/jdd.v1i2

247. Van Beek, P. and Hoffmann, H.F., 2015. Machine learning of Bayesian networks using constraint programming. In Principles and Practice of Constraint Programming: 21st International Conference, CP 2015, Cork, Ireland, August 31--September 4, 2015, Proceedings 21, pp. 429-445. Springer International Publishing. https://doi.org/10.1007/978-3-319-23219-5_31

248. van den Hurk, A., Meelissen, M. and van Langen, A., 2019. Interventions in education to prevent STEM pipeline leakage. International Journal of Science Education, 41(2), pp.150-164. https://doi.org/10.1080/09500693.2018.1540897

249. Van der Berg, S., 2008. How effective are poor schools? Poverty and educational outcomes in South Africa. Studies in Educational Evaluation, 34(3), pp.145-154. https://doi.org/10.1016/j.stueduc.2008.07.005

250. van der Rijst, R.M., Lamers, A.M. and Admiraal, W.F., 2022. Addressing student challenges in transnational education in Oman: the importance of student interaction with teaching staff and Peers. Compare: A Journal of Comparative and International Education, 5(1), pp.1-17. https://doi.org/10.1080/03057925.2021.2017768

251. Van Engelen, J.E. and Hoos, H.H., 2020. A survey on semi-supervised learning. Machine Learning, 109(2), pp.373-440. https://doi.org/10.1007/s10994-019-05855-6

252. van Zyl, A., Dampier, G. and Ngwenya, N., 2020. Effective institutional intervention where it makes the biggest difference to student success: The University of Johannesburg (UJ) integrated student success initiative (ISSI). Journal of Student Affairs in Africa, 8(2), pp.59-71. https://doi.org/10.24085/jsaa.v8i2.4448

253. Venter, L., 2022. A systems perspective on early childhood development education in South Africa. International Journal of Child Care and Education Policy, 16(1), pp.1-25. https://doi.org/10.1186/s40723-022-00100-5

254. Vie, J.J., Rigaux, T. and Minn, S., 2022, September. Privacy-Preserving Synthetic Educational Data Generation. In Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption: 17th European Conference on Technology Enhanced Learning, EC-TEL 2022, Toulouse, France, September 12–16, 2022, Proceedings, pp. 393-406. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-16290-9_29

255. Viloria, A., Lezama, O.B.P. and Mercado-Caruzo, N., 2020. Unbalanced data processing using oversampling: Machine Learning. Procedia Computer Science, 175, pp.108-113. https://doi.org/10.1016/j.procs.2020.07.018

256. Vondrell, J.H. and Sweeney, J.M., 1989. Independent study: Using learning style assessment to predict student success. The Journal of Continuing Higher Education, 37(1), pp.5-7. https://doi.org/10.1080/07377366.1989.10401157

257. Vuttipittayamongkol, P. and Elyan, E., 2020. Improved overlap-based undersampling for imbalanced dataset classification with application to epilepsy and parkinson's

disease. International Journal of Neural Systems, 30(08), p.2050043. https://doi.org/10.1142/S0129065720500434

258. Wan, Y., Xian, J. and Yan, C., 2021. A contextual multi-armed bandit approach based on implicit feedback for online recommendation. In Knowledge Management in Organizations: 15th International Conference, KMO 2021, Kaohsiung, Taiwan, July 20-22, 2021, Proceedings 15, pp. 380-392. Springer International Publishing. https://doi.org/10.1007/978-3-030-81635-3_31
259. Wang, M.T. and Hofkens, T.L., 2020. Beyond classroom academics: A school-wide and multi-contextual perspective on student engagement in school. Adolescent Research Review, 5(1), pp.419-433. https://doi.org/10.1007/s40894-019-00115-z
260. Watzel, T., Kürzinger, L., Li, L. and Rigoll, G., 2020. Synchronized Forward-Backward Transformer for End-to-End Speech Recognition. In Speech and Computer: 22nd International Conference, SPECOM 2020, St. Petersburg, Russia, October 7–9, 2020, Proceedings 22, pp. 646-656. Springer International Publishing. https://doi.org/10.1007/978-3-030-60276-5_62
261. Werhli, A.V., Grzegorczyk, M. and Husmeier, D., 2006. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. Bioinformatics, 22(20), pp.2523-2531. https://doi.org/10.1093/bioinformatics/btl391
262. Wickramaratna, K.J., 2010. DS-ARM: An Association Rule Based Predictor that Can Learn from Imperfect Data, Doctoral dissertation, University of Miami. https://scholarship.miami.edu/esploro/outputs/991031447570102976
263. Wu, S., Sun, F., Zhang, W., Xie, X. and Cui, B., 2022. Graph neural networks in recommender systems: a survey. ACM Computing Surveys, 55(5), pp.1-37. https://doi.org/10.1145/3535101
264. Xie, W. and Curle, S., 2022. Success in English medium instruction in China: Significant indicators and implications. International Journal of Bilingual Education and Bilingualism, 25(2), pp.585-597. https://doi.org/10.1080/13670050.2019.1703898
265. Xu, L., 2020. Synthesizing tabular data using conditional GAN, Doctoral dissertation, Massachusetts Institute of Technology. https://hdl.handle.net/1721.1/128349

266. Yağcı, M., 2022. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), p.11. https://doi.org/10.1186/s40561-022-00192-z

267. Yang, J. and Delpha, C., 2022. An incipient fault diagnosis methodology using local Mahalanobis distance: Detection process based on empirical probability density estimation. Signal Processing, 190(1), p.108308. https://doi.org/10.1016/j.sigpro.2021.108308

268. Ying, X., 2019, February. An overview of overfitting and its solutions. In Journal of Physics: Conference Series, 1168(1), p.022022. IOP Publishing. https://doi.org/10.1088/1742-6596/1168/2/022022

269. Yin, J.B., Li, T. and Shen, H.B., 2011. Gaussian kernel optimization: Complex problem and a simple solution. Neurocomputing, 74(18), pp.3816-3822. https://doi.org/10.1016/j.neucom.2011.07.017

270. York, S., Lavi, R., Dori, Y.J. and Orgill, M., 2019. Applications of systems thinking in STEM education. Journal of Chemical Education, 96(12), pp.2742-2751. https://doi.org/10.1021/acs.jchemed.9b00261

271. Zeineddine, H., Braendle, U. and Farah, A., 2021. Enhancing prediction of student success: Automated machine learning approach. Computers and Electrical Engineering, 89, p.106903. https://doi.org/10.1016/j.compeleceng.2020.106903

272. Zhang, S., Li, X., Zong, M., Zhu, X. and Wang, R., 2017. Efficient kNN classification with different numbers of nearest neighbors. IEEE transactions on neural networks and learning systems, 29(5), pp.1774-1785. https://doi.org/10.1109/TNNLS.2017.2673241

273. Zhang, X., Li, J., Cai, Z., Zhang, L., Chen, Z. and Liu, C., 2021. Over-fitting suppression training strategies for deep learning-based atrial fibrillation detection. Medical and Biological Engineering and Computing, 59(1), pp.165-173. https://doi.org/10.1007/s11517-020-02292-9

274. Zhang, Y., Huang, L., Liu, Y., Chen, Q., Li, X. and Hu, J., 2020. Prediction of mortality at one year after surgery for pertrochanteric fracture in the elderly via a Bayesian belief network. Injury, 51(2), pp.407-413. https://doi.org/10.1016/j.injury.2019.11.029

275. Zhang, Y., Zaidi, N., Zhou, J. and Li, G., 2023. Interpretable tabular data generation. Knowledge and Information Systems, pp.1-29. https://doi.org/10.1007/s10115-023-01834-5

276. Zheng, H., Xie, W., Ryzhov, I.O. and Xie, D., 2023. Policy Optimization in Dynamic Bayesian Network Hybrid Models of Biomanufacturing Processes. INFORMS Journal on Computing, 35(1), pp.66-82. https://doi.org/10.1287/ijoc.2022.1232

277. Zhou, Y., Fenton, N. and Neil, M., 2014. Bayesian network approach to multinomial parameter learning using data and expert judgments. International Journal of Approximate Reasoning, 55(5), pp.1252-1268. https://doi.org/10.1016/j.ijar.2014.02.008

APPENDIX A: ETHICAL CLEARANCE: UNIVERSITY OF THE FREE STATE

UNIVERSITY OF THE FREE STATE
UNIVERSITEIT VAN DIE VRYSTAAT
YUNIVESITHI YA FREISTATA
UFS UV

GENERAL/HUMAN RESEARCH ETHICS COMMITTEE (GHREC)

07-May-2022

Dear Mr Herkulaas Combrink

Application Approved

Research Project Title:
Evaluating A Human-Machine Student Intervention Framework in Higher Education from Legacy Data

Ethical Clearance number:
UFS-HSD2022/0195/22

We are pleased to inform you that your application for ethical clearance has been approved. Your ethical clearance is valid for twelve (12) months from the date of issue. We request that any changes that may take place during the course of your study/research project be submitted to the ethics office to ensure ethical transparency. furthermore, you are requested to submit the final report of your study/research project to the ethics office. Should you require more time to complete this research, please apply for an extension. Thank you for submitting your proposal for ethical clearance; we wish you the best of luck and success with your research.

Yours sincerely

Dr Adri Du Plessis
Chairperson: General/Human Research Ethics Committee

Dr Adri du Plessis
Digitally signed by Dr Adri du Plessis
Date: 2022.05.07 17:18:11 +02'00'

205 Nelson Mandela Drive
Park West
Bloemfontein 9301
South Africa

P.O. Box 339
Bloemfontein 9300
Tel: +27 (0)51 401 9337
duplessisA@ufs.ac.za
www.ufs.ac.za

APPENDIX B: ETHICAL CLEARANCE: UNIVERSITY OF PRETORIA

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

24 April 2022

Reference number: EBIT/19/2022

Mr HMv Combrink
Department: Computer Science
University of Pretoria
Pretoria
0083

Dear Mr HMv Combrink,

FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY

Your recent application to the EBIT Research Ethics Committee refers.

Approval is granted for the application with reference number that appears above.

1. This means that the research project entitled "Evaluating a human-machine student intervention framework in higher education from legacy data" has been approved as submitted. It is important to note what approval implies. This is expanded on in the points that follow.
2. This approval does not imply that the researcher, student or lecturer is relieved of any accountability in terms of the Code of Ethics for Scholarly Activities of the University of Pretoria, or the Policy and Procedures for Responsible Research of the University of Pretoria. These documents are available on the website of the EBIT Research Ethics Committee.
3. If action is taken beyond the approved application, approval is withdrawn automatically.
4. According to the regulations, any relevant problem arising from the study or research methodology as well as any amendments or changes, must be brought to the attention of the EBIT Research Ethics Office.
5. The Committee must be notified on completion of the project.

The Committee wishes you every success with the research project.

Prof K.-Y. Chan
Chair: Faculty Committee for Research Ethics and Integrity
FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY