Cape Peninsula
University of Technology

THE APPLICATION OF BIG DATA ANALYTICS TO IMPROVE STUDENTS' PERFORMANCE IN SOUTH AFRICA

by

CLENCY SYNA NGARY NDZALUYA

Thesis submitted in fulfilment of the requirements for the degree
Doctor of Business Informatics

in the Faculty of Business and Management Sciences
at the Cape Peninsula University of Technology

Supervisor: Prof Michael Twum-Darko

District Six Campus, Cape Town
January 2023

DECLARATION

I, Clency Syna Ngary Ndzaluya, declare that the contents of this thesis represents my own unaided work, and that the thesis has not previously been submitted for academic examination towards any qualification. Furthermore, it represents my own opinions and not necessarily those of the Cape Peninsula University of Technology.

....

Signed **Date**

ABSTRACT

The thesis proposes a normative model for predicting students' performance at a selected tertiary Institution utilizing big data analytics (BDA). BDA can transform education by uncovering patterns, correlations, and other insights into learning outcomes, students' performances, and the effectiveness of learning and teaching practices. Therefore, the aim of this research was to explore how BDA and Machine Learning can be used within South Africa Higher Education and Training to improve students' performance and the challenges thereof. Adaptive structuration theory (AST) was used as a lens on the problem to first understand and interpret the embedded socio-technical processes. The Cross-Industry Standard Process for Data Mining (CRISP-DM) Model was used to devise a plan for the implementation of the machine learning algorithm to meet the objective of predicting students' performances. The algorithm was applied to raw data gathered from different cohorts of students on a structured postgraduate qualification. The implication of the output suggests the establishment of a physical Information Technology infrastructure that makes use of DBA capable of pulling information (structured and unstructured) into a presentation layer that allows for the application of ML on relevant students' data.

Key Words: Big Data, Big Data Analytics, Students performance, Machine Learning, Teaching and Learning.

ACKNOWLEDGEMENTS

I wish to thank:

- My Maker, Lord of lords, King of kings, through whom everything is made possible.
- My father, for his moral and financial support.
- Maida Felix for her support
- My sister for her encouragement and support
- Prof Michael Twum-Darko my supervisor for his guidance, motivation, and efforts.

May the Lord continue to lavish them all with his glory and blessings.

DEDICATION

Dedicated to Prof Michael Twum-Darko who taught me resilience and gave me the support and strength to finish the thesis.

Table of Contents

GLOSSARY

AST	Adaptive Structuration Theory
BDA	Big Data Analytics
EDM	Education Data Mining
ICT	Information Communication Technology
IHL	Institution of Higher Learning
ML	Machine Learning
MSE	Mean Squared Error
ST	Structuration Theory
SVM	Support Vector Machine

CHAPTER ONE: INTRODUCTION

1.1 INTRODUCTION

Education is considered a fundamental global building block for individual and societal development (Brende, 2015). As Lin (2020) and Pedros et al. (2022) highlight, "*education enables upward socioeconomic mobility*" and is key to full participation in the economy and society. Given education's role in preparing learners for socioeconomic participation and mobility, there have been growing concerns in South Africa around dissatisfaction with higher education outcomes and graduates' ability to contribute to society upon completion of their studies (Council on Higher Education, 2017; Ngalo-Morrison, 2017; Mlambo et al., 2021).

Specifically, inefficiencies have been cited regarding Institutions of Higher Learning (IHL) throughput rates, graduation rates, and ability to produce graduates ready for evolving workplace needs (Boughey, 2018; White et al., 2021). The South African higher education context faces multifaceted pressures, including escalating enrolment numbers, increasingly diverse student bodies, variable academic readiness upon entry, and rapidly shifting workplace skill demands that IHL struggle to continually realign curricula with (Gachago et al., 2013; Bozalek et al., 2013; Chetty et al., 2015; Khoza & Manik, 2016; Mlambo et al., 2021; White et al., 2021). In many instances, these dynamics contribute to suboptimal student academic performance and ongoing calls for enhancing teaching, learning, and graduate outcomes in South Africa (Nyamupangedengu, 2017; Tewari & Ilesanmi, 2020).

In response, IHL worldwide and in South Africa have invested substantially in varied strategies intended to catalyse improvements in student performance, including academic support initiatives, supplemental instruction efforts, and—of growing interest—leveraging technology to enhance learning and teaching (Botha, 2020; Van Zyl et al., 2020). For example, learning management systems (LMS) have been widely adopted by South African IHL to support pedagogical functions like assignment submissions, content delivery, assessments, and learning analytics (Gachago et al., 2013; Burtsev, 2021; Badaru & Adu, 2022).

However, research signals these technologies have yielded mixed outcomes, often failing to fulfil their primary objective of strengthening teaching and learning to boost student performance as anticipated (Ng'ambi et al., 2016; Gamede et al., 2022).

Meanwhile, big data analytics (BDA) has gained traction globally across sectors like business, healthcare, and education for its expansive potential to extract previously inaccessible insights from sizable, multifaceted data sources (Marr, 2016). Siemens & Long (2011) and Chaurasia et al. (2018) position BDA as an accelerator of disruptive innovation in higher education, while Picciano (2012) suggests learning analytics, a subset of BDA, enables unprecedented visibility into student learning processes. As Reyes (2015) summarizes, BDA can uniquely transform education operations, empower teaching effectiveness, remedy student performance issues, and propel data-driven decision-making.

While adoption is accelerating globally, BDA remains underexplored and underutilized within higher education in South Africa specifically (Strydom & Loots, 2018; Gamede, 2022). Existing research signals South African IHL generate expansive stores of student data from platforms like LMS. Still, inadequate analytics hampers extracting value from these assets to strengthen teaching and propel performance optimization efforts (Reyes, 2015). As such, an opportunity exists to address this gap by examining how BDA could be strategically leveraged using locally generated higher education datasets to model and predict student academic performance.

Deriving empirically validated, context-specific insights could significantly advance local IHL capabilities to harness BDA for data-driven teaching enhancements, targeted student support initiatives, and other innovations to improve academic outcomes. This study, therefore, explored BDA adoption potential in the South African higher education landscape by investigating relevant local datasets, evaluating predictive modelling techniques, and proposing an integrated BDA framework to guide analytics adoption efforts by

IHL seeking to boost student success. The following section expands on the research problem this study addresses.

1.2. RESEARCH PROBLEM

One of the main goals of Institutions of higher learning in South Africa is to improve students' performances through improved teaching and learning (Nyamupangedengu, 2017; Shay, 2017; Tewari & Ilesanmi, 2020). However, inadequate mining or analytics of Big Data has led to limited information not assisting teaching and learning structures to develop students' full potential (Reyes, 2015). Many South African institutions make use of various approaches and methods to achieve this goal. These include academic support centers, mentorship programs, and technology enhanced learning. The technological integration includes the use of software such as LMS and the analysis of data (Jordaan & Van der Merwe, 2015; Lemmens & Henn, 2016; Lourens & Bleazard, 2016; Burtsev, 2021; Prinsloo & Roberts, 2022).

Despite these methods and approaches, including the use of software, challenges persist in attempts to improve students' performances. The challenges persist either because the software is incorrectly applied relative to the available data or the selected software is not suitable, thereby giving results that do not help in improving students' performances (Arnold & Pistilli, 2012; Letseka et al., 2018; Buthelezi & Van Wyk, 2020). These challenges cause problems such as the inconsistent approach in using existing data as they evolve in many of these higher education institutions.

Most of these Institutions only consider the results of the students/learners and not the factors that influence the results. Thus, it is critical to employ an approach through a framework that encompasses factors of influence to enable forecast and to ensure consistency towards performance improvement. Figure 1.1 below illustrates the embedded socio-technical processes associated with the above-narrated phenomenon of inadequate analytics of Big Data leading to the limited information that can assist teaching and learning structures in developing students' full potential.

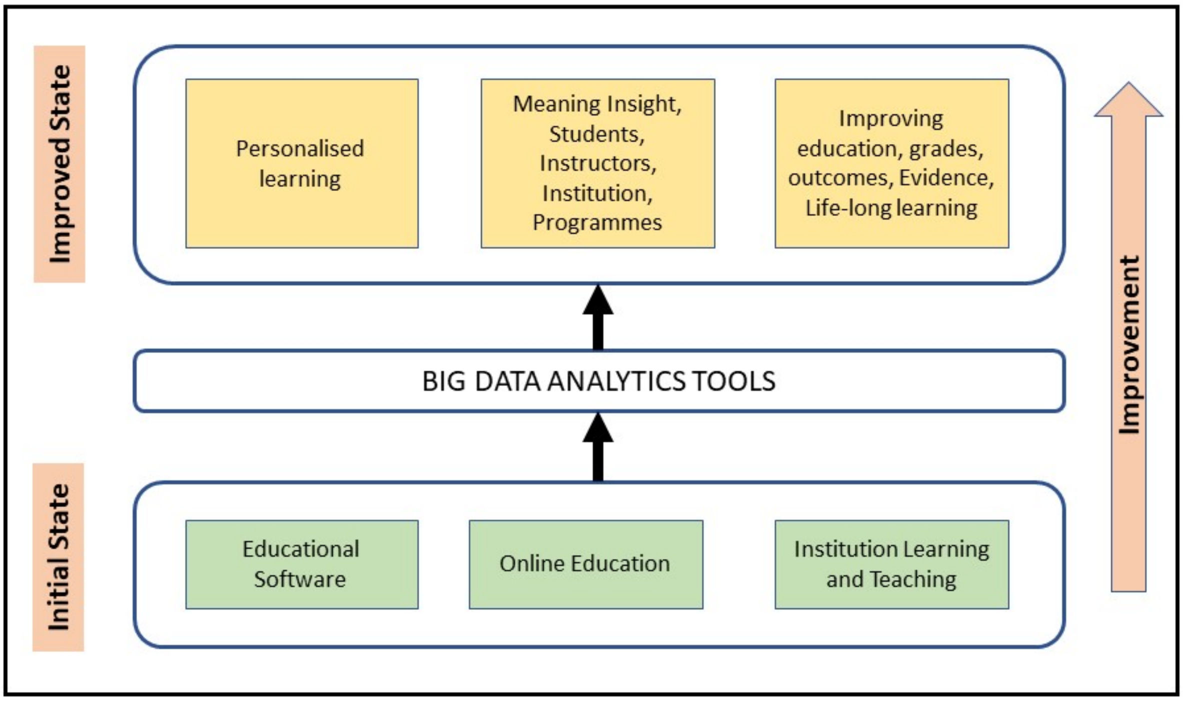

Improved State
Personalised learning
Meaning Insight, Students, Instructors, Institution, Programmes
Improving education, grades, outcomes, Evidence, Life-long learning
BIG DATA ANALYTICS TOOLS
Initial State
Educational Software
Online Education
Institution Learning and Teaching
Improvement

Figure 1.1: problem conceptualisation

1.3. RESEARCH OBJECTIVES AND QUESTIONS

1.3.1 Aim & Objectives

Given the stated problem, the aim of this research is to explore the use of big data analytics to address the challenges that institutions of higher learning face in their attempt to improve students' academic performance. The objectives to tease out the research aim are as follows:

a) To examine the current use of software to improve students' performance.
b) To investigate the nature of learning and teaching big data generated in the University.
c) To determine the relevance of teaching and learning big data in improving student performance.
d) To determine the relevance of data analytics, through machine learning, in predicting students' performance.

e) To propose a framework to guide the use of big data analytics in improving students' performance at Institutions of Higher Learning.

1.3.2 Research Questions

The research questions to address the objectives are:

a) How is software currently used to improve students' performances?
b) What is the nature of big data generated in the IHL, and how can it be used to improve students' performances?
c) How is current teaching and learning data generated being used by IHL?
d) How is big data analytics relevant for improving students' performances?
e) What framework can be used to guide IHL to improve students' performance?

1.4 OVERVIEW OF LITERATURE REVIEW

This section presents an overview of a review of literature related to the study, which includes students' performance, teaching and learning, software, and big data analytics. However, chapter 3 provides a detailed review of relevant literature and the gaps therein that underpin this research.

1.4.1 Students' performance

Student performances still do not reflect high academic success (Khosa et al., 2018) even though access to higher education has been made easier in the post-apartheid era in South Africa (Ng'ambi et al., 2016). Some of the reasons cited for this were:

- students needing personalized support,
- students having different levels of readiness and
- students under preparedness while entering IHL
- Internal factors (class, size, facilities)
- External factors (social life, family background) (Pillay, 2017; Van Den Berg, 2017).

Because of the failure to achieve high academic success, IHL across the world have set objectives to improve students' performance through improvement of teaching and learning. Although IHL has tried several approaches to improve students' performance, they have also now resorted to ICT (Baker, 2010). However, the results are varied of the use of ICT are varied. In the case of IHL in South Africa, the tools that have been implemented, such as LMS, have produced mixed results.

Big Data Analytics is now coming to the forefront, and IHL around the world are now reaping the benefits. IHL is now able to improve students' performance by customizing the curriculum for one student at a time (Papamitsiou & Economides, 2014; Arnold & Pistilli, 2012; Romero & Ventura, 2020).

1.4.2 Big Data

Big Data refers to datasets that are difficult to store, structured and unstructured, vast and complex. The datasets are also difficult to process using traditional methods and recent software technologies (Raghupathi & Raghupathi, 2014; Sharma, 2017). Big Data can be further characterised by volume, veracity, velocity, variety, and values ("5Vs") (Raghupathi, 2011, Zikopoulos & Eaton, 2014). This research looked at big data in education, as higher institutions of learning generate data from multiple sources, such as social media, email, or learning software, including blackboard.

1.4.3 Big Data Analytics

Big data in a vacuum are useless, and their relevance is established when they are used to derive information and insights that will drive the decision-making process (Gandomi & Haider, 2015). Processes such as Big Data Analytics are used to extract meaningful insights from the data. Indeed, Big Data Analytics (BDA) refers to both the massive collection of data and their analysis (Ali et al., 2013). The aim of the analysis is to discover patterns, intelligence, knowledge, or any other information that can enhance the decision-making process (Ali et al., 2013). Gandomi & Haider (2015) break down the BDA process into audio analytics, text analytics, social media analytics, video analytics, and predictive analytics (the focus of this research).

The BDA emergence has changed our society and is drawing attention from the public and technological experts. The amount of digital data continues to grow, and Big Data is in its infancy. The analysis and storage tools will continue to improve (Oracle, 2016; Rabella, 2016). One of the most visible applications of this new computer science field is the contextual advertisement, which makes it possible to offer each user the targeted proposals and elaborated proposals according to the personal traces of navigation (Manyika et al., 2011; Datamer, 2016).

Although Big Data Analytics has been used in several industries, such as the gaming Industry, network security, market and business, healthcare, telecommunication, sports, and education systems (Sharma, 2017), this research will look at Big Data Analytics in Education.

1.4.4 Big Data Analytics in Education

Enhancing education through data analytics is a growing focus, evidenced by organizations like the International Society of Education Data Mining and the Society of Learning Analytics Research. These societies publish free journals to aid institutions and policymakers in informing and enhancing education (Romero & Ventura, 2010).

The widespread use of digital devices, including computers, tablets, and smartphones, has expanded the applications of data analytics across various domains. In education, the collection and analysis of learning-related data are collectively known as "learning analytics" (Siemens & Long, 2011; Picciano, 2012), contributing to numerous research initiatives, industrial developments, and ethical considerations.

The integration of computers, Big Data, and digital technology in education has led to a significant transformation. Laptops and tablets now replace traditional tools in classrooms, generating substantial data on learning and teaching. Collaboration between technology companies and schools can leverage this data to enhance teaching methods, curricula, and support students facing challenges (Eggers, 2007; John & Wheeler, 2012).

Analytics plays a crucial role in higher education, offering instructors feedback, predicting student performance, and making recommendations. Predictive analytics can monitor and improve student outcomes, providing insights into their feelings during coursework. This approach enables the early identification of struggling students or ineffective teaching methods, fostering timely interventions (Romero & Ventura, 2010; Borray, 2017).

In the past, identifying students in need was challenging, relying on arbitrary choices by school leaders. However, continuous data analysis now allows for personalized learning experiences tailored to individual interests, knowledge, and intellectual capacities, reducing reliance on traditional indicators like exam results (Reschly & Christenson, 2006; Henard, 2009).

1.5 OVERVIEW OF UNDERPINNING THEORY

This section reviews the theory, adaptive structuration theory that underpins the research.

1.5.1 Background

Theory can be defined as a framework for observation and understanding, which informs and shapes what is seen and how it is seen. Hence, it can be said that theory informs our understanding of the phenomenon under investigation, a lens of sorts. It allows the researcher to make inferences between the theoretical and empirical, concrete, and abstract (Strauss & Corbin, 1994).

Rose & Scheepers (2001) explain that theory is used in Information communication technology (ICT) because it can assist in determining the interaction between technologies and organization, people, and personal contexts. The theory can also provide support in reconciling several theoretical perspectives. Finally, the theory is used in empirical studies because it helps deepen the knowledge about the phenomenon at hand (Broger, 2011; Heracleous, 2013).

Given the nature of the research questions and the envisaged output to derive a normative big data model to guide the improvement of students' performances, the phenomenon is a social construct that can be studied by using sociotechnical approaches through the lens of Adaptive Structuration Theory (AST). AST emanates from a social theory called structuration theory, which was established and applied by Anthony Giddens and later refined by Poole and DeSanctis to address the impact of technology. Thus, AST provides a deeper understanding of how improvement can be realized from the adaption of new technologies by organizations such as IHL (see section 1.5.3 below and Chapter 2).

1.5.2 Structuration theory and information systems research

To determine the interaction between information systems (IS) and people, organizations, and personal context, the theory has been widely used in the field of information systems research (Rose &Scheepers, 2001). Nevertheless, the author did not intend to use it in IS at its inception (Jones & Karsten, 2008). It was later readapted by IS researchers. Indeed, structuration theory was later adapted as the duality of Technology by Orlikowski (2000) and Adaptive structuration theory by (Poole & De Sanctis, 1994).

1.5.3 Adaptive structuration theory (AST)

To examine the interactions among groups and organizations with advanced technologies, Structuration Theory was reformed by John DeSanctis and Poole into Adaptive Structuration Theory. AST was expressed as "*the production and reproduction of the social systems through members' use of rules and resources in interaction*" (Jones & Karsten, 2003; Aktaruzzaman & Plunkett, 2016).

AST is a practical method designed for investigating advanced technologies in transforming an organisation. The transformation process is examined by AST from the perspectives of *"types of structures that are provided by the advanced technologies and the structures that actually emerge in human action as people interact with these technologies"* (Poole & De Sanctis, 1994). AST proposes that "*social structures provided by information technology can be described in two ways: structural features of the technology and the spirit of this feature set*" along with appropriation (Poole & De Sanctis, 1994).

The theory stresses the social aspects and critiques the technocentric view of the use of technology. Perceptions about how technology can be applied to activities and its role and utility are dynamically created by organizations and groups working while using information technology. These perceptions may extensively diverge across groups. Technology usage is influenced by these perceptions, which also facilitate impact on group results (Jones & Karsten, 2003; Aktaruzzaman & Plunkett, 2016).

1.5.4 Adaptive Structuration theory and big data in technology

AST was used in this research to examine the introduction of innovation technology such as big data analytics and demonstrate in what way the structures of big data infiltrated or can infiltrate the educational system (academic), influencing it. AST also demonstrated how big data's original intent is modified and influenced by the social structures of those academics. In summary, AST's appropriation model is useful for investigating the application and penetration of big data technology in our education system (Aktaruzzaman & Plunkett, 2016).

1.5.5 Conclusion

Despite the delay when compared to other industries, the education sector is now integrating Big Data and its benefits into its organization. This delay is mainly due to the perception of the collection and analysis of educational Big Data as an intrusion of governments attempting to monitor and study citizens. Others see it as an attempt by companies to identify and format their future clientele.

These fears and reticence are entirely acceptable. However, the quality of education is an essential factor in the successful development of a society. BDA holds the potential to help schools operate more efficiently, enable teachers to improve their methods, and prevent students from falling into school failure. It is, therefore, essential to exploit the opportunities offered by this technology to the maximum.

1.6 OVERVIEW OF RESEARCH APPROACH AND METHODOLOGY

This research utilized students' data to predict students' performance. This was done to derive a way to automate and improve students' performance from these predictions by identifying students at risk. Thus, the research sought to build a model that would perform the predictions. This was done by following the revised framework of CRISP-DM (Shearer, 2000).

The CRISP-DM Model has been used in education for data mining as educational data mining (EDM). This framework is relevant to this research because it proposed an approach that guided the predictability of students' performance as initially discussed. Hence, the steps of the CRISP-DM model were followed to help guide the model's predictability. The steps of the model included:

a) Domain Understanding: the objective of this step was to obtain clarity and relevant information required to achieve the objective.

b) Data Understanding: this step entailed collecting the data and checking for completeness and redundancy.
c) Data preparation: the data was prepared for analysis by cleaning and transforming it.
d) Application of machine learning: the correct machine learning algorithm was selected and then applied to the data.
e) Evaluation: this step entails interpreting the results from the application of machine learning.
f) Using the discovered knowledge: this involved applying the model to a performance system or simply documenting the discovered knowledge to transfer to interested parties.

1. 7. ETHICAL CONSIDERATION

The research adhered to Cape Peninsula University of Technology (CPUT) ethical requirements. Prior to primary data collection, a detailed explanation of the research purpose, methods, potential risks, benefits, and the voluntary nature of participation was provided to the university. It was emphasized that participation was entirely voluntary and that they could withdraw their data at any time without penalty.

To protect the privacy and confidentiality of the students, all collected data was anonymized. Personally identifiable information such as names, student identification numbers, and contact details were removed or replaced with unique identifiers. Anonymization techniques were employed to ensure that individual students could not be identified from the collected data.

Data collection focused on obtaining only the necessary information for the research objectives. The IHL was assured that their data would be used solely for research purposes and would not be shared or disclosed to any unauthorized individuals or entities. Any unused or unnecessary data was promptly deleted to minimize the potential risk of data breaches or misuse.

1.8. DELINEATION

The study focused on Institutions of Higher Learning, with a strong focus on technology. The case study is based on institutions of higher learning in the Western Cape. The institutions need to constantly evolve to keep up with market changes, which makes them a great opportunity for this research. Also, the research will be limited to IT students as this is where the research is based. Once that was done, the results of the algorithm were further applied to other departments.

1.9. OUTCOME

The outcome of the research is the accomplishment of the aim and objectives of the research. This entails proposing a big data analytics framework that can assist in improving students ‘performance. Also, the concept of BDA in education is still new, and this research will then assist in providing literature for it. The study will also provide a model to identify students at risk of failing by identifying the most appropriate machine learning algorithms.

1.10. SIGNIFICANCE OF THE STUDY

The significance of the study is that it will enhance how universities can improve teaching and learning to improve students’ performances through big data analytics via machine learning. This study also aims to be used as a reference for studies on students’ performance and to contribute to ongoing studies on machine learning and big data analytics.

1.11. THESIS ORGANISATION

This thesis is organized as follows:

Chapter 1 presents an overview of the thesis. This chapter introduced the need for BDA in education as IHL strives to improve students’ performances. This is then followed by establishing the problem statement, from which follows the research objectives, outcomes, and contributions.

Chapter 2 presents the underpinning theory. The chapter finally provided a conceptual framework derived from the preliminary reviewed literature and underpinning theoretical framework that served as further support to the theoretical lens to understand and interpret the socio-technical processes embedded in the phenomenon and how the laboratory work was designed and conducted.

Chapter 3 reviews current work relevant to Big Data Analytics and how it can be used to improve students' performances. Hence, the chapter begins by introducing the issue with students' performances, which then leads to the understanding of big Data Analytics and the role it can play in improving students' performance.

Chapter 4 presents the different stages of the CRISP-DM model and how they assisted in building a model that will perform the predictions.

Chapter 5 discusses the machine learning techniques used and their evaluation. The chapter describes the model selection, the training of the models, and the different results obtained.

Chapter 6 explains how the objectives were achieved, along with contributions, recommendations and further research emanating from this research.

1.12. SUMMARY

The chapter sets up the scene to discuss the field of study "*Big Data analytics to improve students' performance*". The questions asked here are pertaining to the aspects of big data analytics that can or are being exploited, in the process of improving students' performance, and the possible challenges. Through the lens of adaptive structuration theory, the appropriation of big analytics to improve students' performance was understood and interpreted.

The chapter introduced the research topic which was derived from the research problem and the underpinning aim and objectives of the research. It further conceptualised the research problem given the stated how and why of the phenomenon in the problem statement. To provide context to the research topic and the problem, the chapter also provides overview of the literature associated with the underpinning social theory and methodology, on previous and current research relevant to the research problem and indicating the significance of addressing the phenomenon.

Finally, the chapter provides a conceptual framework derived from the reviewed literature related to the area of this research, and underpinning theoretical framework that served as further support to the theoretical lens to understand and interpret the socio-technical processes embedded in the phenomenon and how the laboratory work was designed and conducted.

The next chapter deals with all the theoretical perspectives that assisted in interpreting the problem narrated in this chapter.

CHAPTER TWO: UNDERPINNING THEORY

2.1 BACKGROUND

The previous chapter introduced the research topic which was derived from the research problem and the underpinning aim and objectives of the research. This Chapter discusses the theoretical framework underpinning the research as a lens through which to understand and interpret the socio-technical processes embedded in the phenomenon. Drawing from Strauss & Corbin (1994), it is argued that a theoretical framework applied in empirical research is to observe and understand the phenomenon being study to inform and shape what is seen and how it has been observed. Hence it is agreeable that the social theory applied in this research informed the understanding of the phenomenon under investigation which allowed for making inferences between the theoretical and empirical, concrete, and abstract.

Rose & Scheepers (2001) further explain that social theories are used in Information communication technology (ICT) research because it can assist in determining the interaction between technologies and organisation, people, and personal contexts. The theory can provide support in reconciling several theoretical perspectives. Finally, theory is used in empirical studies because it helps to deepen the knowledge about the phenomenon at hand (Broger, 2011; Heracleous, 2013).

Given the nature of the research questions and the envisaged output to derive a normative big data model to guide the improvement of students' performances, the phenomenon as a social construct was studied by using sociotechnical approaches through the lens of Adaptive Structuration Theory (AST).

AST emanates from a social theory called structuration theory, which was developed and applied by Anthony Giddens (1984), and later refined by Poole and DeSanctis (1994). It can be argued that there are many social theories that can be used as lenses to study the sociotechnical processes embedded in many socially contrasted phenomena.

However, AST was found to be appropriate to study this phenomenon due to its components that offer meaning to mutually inclusive and iterative interplay between technology and the context of the phenomenon. Furthermore, it is not much of the IHL throughput rate resulting from the use of new technology. But how the results were affected by the adoption and adaption of the structures of technology and how IHL adapts its own structures to the technology.

2.2 SYNOPSIS OF STRUCTURATION THEORY

The original intent of Anthony Giddens crafting structuration theory was the understanding of the ontology of human society in the context of social actions and life (Jones & Karsten, 2003; Broger, 2011). According to Giddens (1984), social theories were under theorized because their use had shifted from ontological to epistemological approaches.

At that time, social theories were deemed appropriate for epistemological reasons. Furthermore, Giddens (ibid.) was also trying to address the deterministic view that claims that agents are constrained by their social structure. He also tried to address the deterministic extreme view, that is, the volunteerism view, which states that agents are not restricted by their social structures. This has led to him establishing the notions of dualism as in structure vs agent, macro vs micro.

Giddens (1984) was inspired by Karl Marx, who claimed that humans cannot make history independently of their own volition and choices. This drove Giddens to develop a theoretical framework that defines the role of agents or actors and how that agency creates social systems, which, in turn, influence the actor. Even though social theory was initially created with the aim of defining the ontology of human agents in actions and the social context, its use has now been readapted. Structuration Theory (ST) is now used by organizations and scholars in the context of administrative and social sciences, such as strategic management, entrepreneurship, communication, information technology, and organisational discourse. This adaptation of the theory to these numerous situations was done for several reasons.

2.2.1 Structuration theory and empirical work

Structuration Theory can assist with empirical work, with framing the research, and in the analysis and interpretation of the results. Giddens (1984) argues that ST is more beneficial when it contributes to solving empirically related problems and is of the view that ST can be used as a sensitizing device that would approach data in new ways and provide insights and scope for examining methodological challenges.

2.2.2 Structuration theory and Information System/Technology

Regarding information systems/Technology (IS/IT research, the construct of structure in ST provides an epistemological ground for data structuration and analyses based on routinized practices of social interactions. To determine IS/IT interaction with people, organisation, and personal context, The theory has been widely used in information systems' research (Rose &Scheepers, 2001). This is accomplished via understanding human-computer and human-human interactions, as well as how these interactions connect to social contexts (Ma, 2010). ST was readapted as Adapative structuration theory by (Poole & De Sanctis, 1994) and duality of Technology by Orlikowski (2000).

2.2.3 Structuration theory in organisation

Yates (1997) states ST was to understand organisational interactions as they shape values identity and capabilities. ST assist in analysing institution at microlevel business by providing a theoretical framework. That analytical framework can encompass the possibility of including institutional structures in the decision-making process. He framework can also include IT concepts such as Big Data Analytics, without forgetting the relevant actors (Yates, 1997; Albano et al., 2010). Within the context of this research, it will mean creating a theoretical framework for educational institutions, including the use of BDA, considering actors such as students and/or lecturers.

Furthermore, the reductionist's view argues that IT steers the enactment of its social structures. Thus, ST aims to reconcile the social construction of technology with this technological determinism. This is achieved by acknowledging the affect technology has on social structures which in turn is affected by these structures. (Yates, 1997; Toland & Yoong, 2010). So as BDA will constrain social structures, it will always be affected by it.

2.3 AGENT AND STRUCTURE

ST draws together two strands of thinking: hermeneutics and structuralist tradition (Rose & Scheepers, 2001). The hermeneutics tradition stresses the importance of human agent while The Structuralist tradition emphasises structure (constraint) as the focus (Rose & Scheepers, 2001). As Agent and structure influence each other, they are both scrutinised without giving prevalence to either one of them. ST examines the social and hermeneutics practices where agent and structure intersect (Rose & Scheepers, 2001; Gehman, 2008). Agent repetitively enacts, recreates, and develops the social structure which affects them both; Thus, Agent and structure are deemed mutually dependent (Rose & Scheepers, 2001). Agency refers to the "*capacity to make a difference*" additionally defined as transformative capacity (Giddens, 1984:14, Jones & Karsten, 2008). Structure is explained by Giddens (1984) as "*a set of rules and resources*" that are repeatedly involved in the social production that constrain and guide the action of agents.

2.3.1 Agency

Agency refers to human transformative capacity to make a difference. (Giddens, 1984:14, Jones & Karsten, 2008; Twum-Darko, 2014). Giddens (1984) denotes that human agent is not completely powerless over their circumstances but retain some transformational capability, albeit minor. This research is concerned with the transformational aspect of the agent.

An agent or IHL lecturer or employees in this case will be depending on structural concepts, such as BDA, to drive the transformational procedure and improve students' performance. Furthermore Also, agency is strongly linked to power. Power refers to the use of resources (Giddens, 1984:15). Regarding this research, agents will utilize the resources available to them within IHL to drive transformation and improve students 'performance. It is also important to note that human actions are recursive. That means that through interaction with structure they create and recreate these practices to the point that they become routinised. In the context of this research this will mean that the use of BDA will be used with IHL until it becomes ingrained to and routinised; thus, improving students' performance.

Furthermore, other characteristics or actors includes knowledgeability and reflexivity. Indeed, Giddens (1980) argues that, in the creation and recreation of social interaction, actors are possess a high level of awareness with regards to the knowledge they have and apply. Actors can also be reflective, meaning that they can observe and understand their actions even while enacting them. These reflective capabilities hence offer the possibility to enact changes. With regards to this research, human agents such as lecturers are reflective. Meaning as they are teaching, they have the possibility to reflect on it, and come up with new solutions to improve teaching and learning.

It is important that Actors' reflexivity and knowledge is sometimes constrained by the nature of actions, unconscious sources of motivation, the difficulty of articulating tacit knowledge, and unintended consequences of actions (Giddens, 1979:144).

2.3.2 Structure:

The social process relating to the reciprocal interaction between structural features of an institution and human agents is called structuration. This means that structure constrains and enables human agents' action, while being instantiated in human action themselves (Twum-Darko, 2014). Structures refers to rules and resources that repetitively involved within social interaction. Resources provides structure with a transformative capacity while providing agents with guidance and directions. Rules refer to practiced directives of social conducts that are accepted in social interaction. Rules sanction socially and mutually accepted practices that regulates human behaviour. This is made possible because rules constitute meaning and sanction misconducts (Twum-Darko, 2014).

2.4 DUALITY OF STRUCTURE

ST addresses the traditionalist dualistic belief that the social system is produced by agent or structure. ST describes a social system developed from both structure and agency without either one of them being predominant. This implies that structure will affect agent while agent also influences structure in return (Jones & Karsten, 2008; Twum-Darko, 2014).

The interplay of structure (rules and resources) and agency (action of knowledgeable and reflexive actors) creates patterns of interactions that will be institutionalised as standardised practices within organisations. Over time, the enactment and re-enactment of these established practices will constitute the organisation structural properties. These structures are drawn by actors in their dealings, while being reinforced by their continuous use.

The dimensions of the duality of structure are signification, domination, and legitimation. Signification relates to the procedures that produce meaning using interpretive scheme of knowledge. Through actors' interaction, this meaning is constituted and communicated. Legitimation is associated to the resources and norms that sanctions behaviour within an organisation. Norms are organisational rules or conventions governing or legitimating an appropriate conduct. Norms constitute organisational structures of legitimation. Finally, Power is a 'transformative capacity' of actors to transform the social and material world. This constitutes organisational structure of domination. However, actors always retain an inherent capacity to change a particular structure of domination.

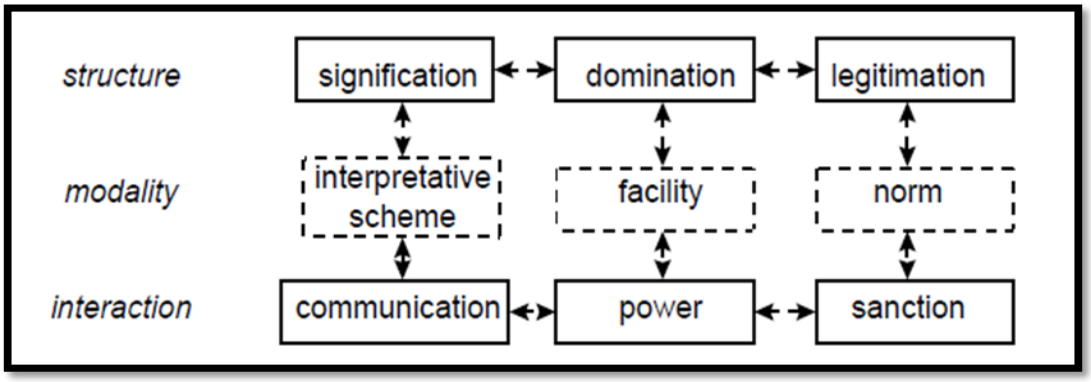


Figure 2.1: Dimensions of the duality of structure, Giddens (1984)

2.5 DUALITY OF TECHNOLOGY

As mentioned in the previous sections, ST was initially created to deal with social interactions. It was later readapted within the context of information system/ information technology. Indeed, Orlikowski came up with the concept of duality of technology that emphasis the technological influence and dynamic interplay between institutional structure, people and technology within an organisation Pham and Tanner (2015:6; Twum-Darko, 2014). This is illustrated in Figure 2.2 below. The figure displays some of the component of the duality of technology. This includes:

- Human agents which can be referred to users, decision-makers, and technology designers.
- Technology refers to material artifacts facilitating task execution in a working environment.
- Institutional properties of organisations – business strategies, structural arrangement, corporate governance, division of labour, culture, ideology, control mechanisms, standard operating procedures, expertise, communication patterns.

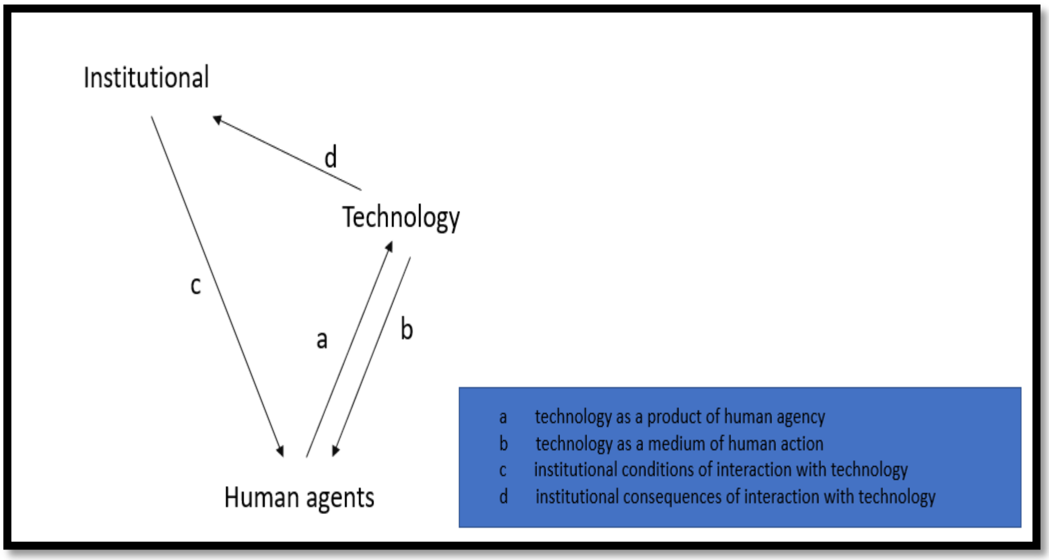

Institutional
d
Technology
c
a
b
Human agents
a technology as a product of human agency
b technology as a medium of human action
c institutional conditions of interaction with technology
d institutional consequences of interaction with technology

Figure 2 2: Enactment of the duality of technology (Orlikowski, 2000)

Orlikowslki considers the use of technology in relation to practice as exemplified in the repetitive creation and recreation of structure (Larsson, 2012:258). Information technology is thus conceptualised as the social product of human action within a particular structure (Orlikowski & Robey, 1991:151). Human action produces technology, according to Orlikowski (2000), which is also a by-product of that action. According to Veenstra, Melin & Axelsson (2014), the concept of duality of technology suggests that technology is both physically and socially constructed by actors, based on the meanings attached to it and the features they emphasise and use (Veenstra, Melin & Axelsson, 2014).

2.6 DUALITY OF TECHNOLOGY AND BIG DATA ANALYTICS

Technology is considered a force that has deterministic influence over organisational structures and properties (Orlikowski, 1992). Technology such as Big Data analytics have influence and impacts that are regulated by actors' action and organisation context (Brock & Khan, 2017). BDA would thus constitute a medium of human action. It will have tremendous impact on users, lecturers and students, standardising, restructuring, and improving teaching and learning. The duality of technology has been readapted in the following pictures.

However, in this research, the duality of technology and big data analytics highlights the complex interplay between technological resources and data analytics, and the ways in which technology both enables and constrains the use of data analytics.

On one hand, technology is a key enabler of big data analytics, as it provides the tools and resources needed to collect, process, and analyse large amounts of data. Without the right technology infrastructure and resources, it would be difficult to effectively use big data analytics to gain insights from data. On the other hand, technology can also constrain the use of big data analytics in several ways. The availability and cost of technology resources may limit the scope or scale of data analytics initiatives. Similarly, technological limitations or constraints may impact the accuracy or reliability of data analytics results.

Furthermore, the duality of technology and big data analytics can improve students' performance in several ways. On the one hand, technology can enable the use of big data analytics to personalise learning, identify at-risk students, improve student retention, and enhance the student experience. For instance, technology can provide the tools and resources needed to collect, process, and analyse large amounts of data on students' academic and non-academic experiences, which can inform the development of personalised learning materials or interventions to support at-risk students.

On the other hand, the duality of technology and big data analytics can also constrain the use of these tools to improve students' performance. For example, the availability and cost of technology resources may limit the scope or scale of data analytics initiatives, or technological limitations or constraints may impact the accuracy or reliability of data analytics results. Additionally, concerns about privacy and the ethical use of data may limit the types of data that can be collected and analysed, or the ways in which data analytics can be used to improve students' performance. Figure 2.3 shows how duality of technology was readapted in this research.

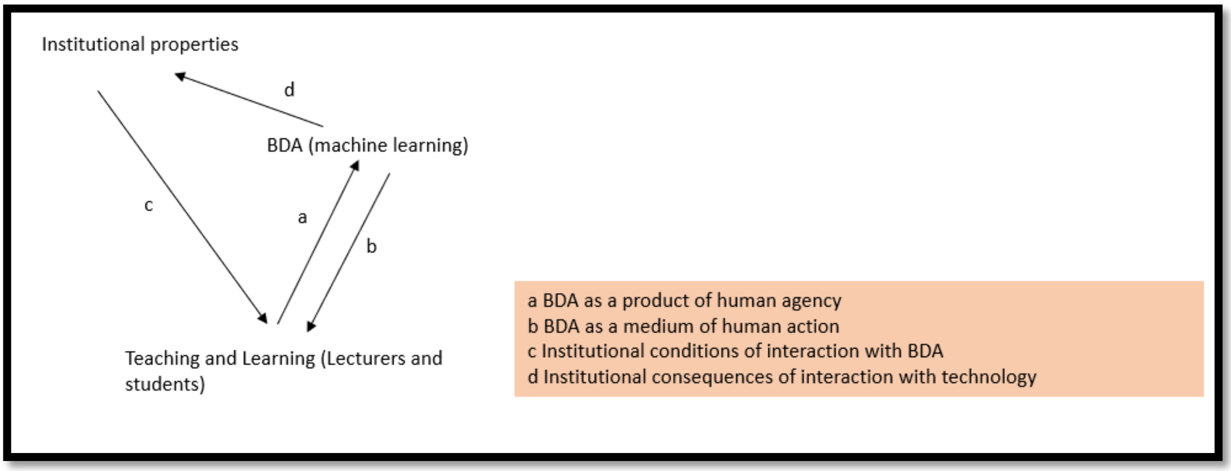

Institutional properties
d
BDA (machine learning)
c
a
b
Teaching and Learning (Lecturers and students)
a BDA as a product of human agency
b BDA as a medium of human action
c Institutional conditions of interaction with BDA
d Institutional consequences of interaction with technology

Figure 2.3: Adapted Duality of Technology for Big Data Analytics

The arrows in the above pictures represents different interplays:

a) Arrow (a) represent BDA as a production of human agency or actions. This means that BDA is considered an outcome such as design, development, modification, and appropriation. Big data analytics is a product of human agency in the sense that it is created and used by human beings to analyse and gain insights from data. Human beings decide what data to collect and analyse, how to collect and analyse it, and how to use the insights generated by data analytics.

In this sense, big data analytics is a product of human agency in the same way that other technologies or systems are created and used by humans to achieve specific goals or outcomes. However, it is important to recognise that big data analytics is not a neutral tool, and the way in which it is used and the insights it generates are shaped by the values, priorities, and biases of the humans who create and use it.

Finally, the role of human agency in the creation and use of big data analytics highlights the importance of considering the ethical and social implications of these tools, and of ensuring that they are used in a responsible and transparent manner.

b) Arrow (b) represents BDA as a medium of human action: this means BDA will facilitate actors’ actions. BDA will mediate and constrain actions drawing from interpretive schemes, norms, and facilities. Big data analytics can be understood as a medium of human action in the sense that it provides a means for humans to act based on the insights generated by data analytics. By analysing large amounts of data, big data analytics can provide insights into trends, patterns, and relationships that may not be immediately apparent to humans. These insights can inform decision-making and action-taking in a variety of contexts, such as business, healthcare, education, and more.

   For example, in a business context, big data analytics may be used to identify trends in customer behavior or to optimise supply chain management. In a healthcare context, big data analytics may be used to identify trends in patient outcomes or to inform the development of new treatments. In an educational context, big data analytics may be used to identify trends in student performance or to personalise learning experiences.

   Lastly, the role of big data analytics as a medium of human action highlights the ways in which these tools can inform and shape the actions of humans in a variety of contexts. It is important to recognise, however, that the insights generated by big data analytics are only as reliable and accurate as the data and algorithms used to generate them, and that the actions taken based on these insights are shaped by the values, priorities, and biases of the humans who use them.

c) Arrow (c) Institutional conditions of interaction with BDA. This entails institutional properties of the organisation will influence lecturers and students as their as they interact with BDA, hence have an impact on teaching and learning. Institutional conditions of interaction with big data analytics in an education environment refer to the structural and cultural frameworks that shape the use and adoption of big data analytics within educational settings. These conditions may include:

Policies and procedures: The policies and procedures governing the use of big data analytics within an educational institution can influence how these tools are used and adopted. For example, policies around data privacy and security may impact the types of data that can be collected and analysed, or the ways in which data analytics can be used to improve student outcomes.

Funding and resources: The availability of funding and resources, including technology infrastructure and personnel, can impact the use and adoption of big data analytics within educational settings. Institutions with more resources may be better able to implement and sustain data analytics initiatives, while institutions with limited resources may face challenges in this area.

Cultural norms and values: The cultural norms and values of an educational institution can influence the use and adoption of big data analytics. For example, an institution that values innovation and continuous improvement may be more likely to embrace data analytics as a tool to inform decision-making and improve student outcomes.

Ultimately, understanding the institutional conditions of interaction with big data analytics in an educational environment can help institutions to identify potential barriers to the adoption and use of these tools and to develop strategies to overcome these barriers. It can also help institutions to align the use of big data analytics with their goals and values and to ensure that these tools are used in a responsible and ethical manner.

d) Arrow (d) Institutional consequences of interaction with BDA. This implies that interaction with BDA impacts the structural properties of the institution, through the reinforcement and transformation of structures of signification. Institutional consequences of interaction with big data analytics in an educational environment refer to the impacts that the use of these tools has on the institution itself, including its structures, processes, and culture. Some potential consequences of interacting with big data analytics in an educational environment include:

Improved student outcomes: One potential consequence of interacting with big data analytics in education is the improvement of student outcomes, such as increased retention rates, higher grades, or improved graduation rates. By analysing data on students' academic and non-academic experiences, institutions can identify trends and patterns that may inform the development of personalised learning materials or interventions to support at-risk students.

Changes to organisational structures and processes: The use of big data analytics may also lead to changes in organisational structures and processes within educational institutions. For example, the adoption of data analytics may require the creation of new roles or responsibilities within the institution, or the restructuring of existing processes to accommodate data analytics initiatives.

Changes to institutional culture: Interacting with big data analytics may also lead to changes in the culture of educational institutions. For example, the use of these tools may shift the focus of decision-making from subjective judgments to data-driven approaches or may lead to a greater emphasis on continuous improvement and innovation.

Ultimately, the institutional consequences of interacting with big data analytics in an educational environment can be both positive and negative, and it is important for institutions to carefully consider the potential impacts of these tools on their structures, processes, and culture.

2.7 ADAPTIVE STRUCTURATION THEORY (AST)

Structuration Theory was reformed by DeSanctis and Poole into Adaptive Structuration Theory with the aim to study the interaction of groups and organisations with advanced technology. Adaptive structuration theory (AST) is a sociological theory that explains how individuals and groups use social structures to produce and reproduce their social worlds. The theory emphasises the role of human agency in shaping and adapting social structures, and it highlights the ongoing process of structuration, or the mutual influence between social structures and human action. AST was expressed as *"the production and reproduction of the social systems through members' use of rules and resources in interaction"* (Jones & Karsten, 2003; Aktaruzzaman & Plunkett, 2016).

AST is often used to study the use of information and communication technologies (ICTs) in organisations and society. AST suggests that ICTs are not neutral tools, but rather they are shaped by and shape the social structures and human agency within which they are used. AST has been applied to a wide range of social and organisational phenomena, including the use of social media, the adoption and diffusion of innovations, and the implementation and management of information systems. It is a useful framework for understanding how social structures and human agency interact and influence each other in dynamic and complex ways.

AST is a practical method designed for investigating advanced technologies in transforming organisation. The transformation process is examined by AST from the perspectives of "*types of structures that are provided by the advanced technologies and the structures that actually emerge in human action as people interact with these technologies*" (Poole & De Sanctis, 1994). AST proposes that "*social structures provided by information technology can be described in two ways: structural features of the technology and the spirit of this feature set"* along with appropriation (Poole & De Sanctis, 1994). There are several key components of adaptive structuration theory (AST):

Structures: AST defines structures as "the medium in which human activity is organised." Structures include both material and symbolic elements, such as physical artifacts, language, and social norms. Structures provide the framework for human action, but they are also shaped and adapted by human action.

Human agency: AST emphasises the role of human agency in shaping and adapting social structures. Human agency refers to the capacity of individuals and groups to act independently and to influence the world around them.

Structuration: AST highlights the ongoing process of structuration, or the mutual influence between social structures and human action. Structuration refers to the way in which social structures are both shaped by and shape human action, and it emphasises the dynamic and ongoing nature of this process.

Duality of structure: AST introduces the concept of the duality of structure, which refers to the idea that social structures are both constraining and enabling. Structures constrain human action by setting boundaries and limits, but they also enable action by providing the resources and opportunities needed to act.

Co-constitution: AST suggests that social structures and human agency are co-constituted, meaning that they are mutually constituting and influence each other. This means that social structures are not fixed or static, but rather they are continually shaped and adapted by human action, and vice versa.

Adaptability is an important concept in adaptive structuration theory (AST). AST suggests that social structures are not fixed or static, but rather they are continually shaped and adapted by human action. Therefore, the adaptability of social structures is an important factor in the ongoing process of structuration, or the mutual influence between social structures and human action. Adaptability refers to the ability of social structures to change and adapt in response to changing circumstances. This can involve both small-scale adjustments as well as larger-scale transformations. AST suggests that the adaptability of social structures is influenced by the human agency of the individuals and groups who use and shape these structures.

In the context of organisations and information systems, adaptability can refer to the ability of an organisation or system to respond to changes in the external environment, such as technological innovations or shifts in market conditions. AST would suggest that the adaptability of an organisation or system will be influenced by the social structures and human agency within the organisation or system, as well as the broader social structures in which it is embedded.

Finally, the theory stresses on the social aspects and critiques the technocentric view of the use of technology. Perceptions about how technology can be applied to activities and its role and utility are dynamically created by organisations and groups work while using information technology. These perceptions may extensively diverge across groups. Technology usage is influenced by these perceptions which also facilitate impact on group results ((Jones & Karsten, 2003; Aktaruzzaman & Plunkett, 2016).

2.8 AST AND BIG DATA ANALYTICS TECHNOLOGIES

2.8.1 AST and Big Data Analytics

Adaptive structuration theory (AST) can be used to understand the role of social structures and human agency in the context of big data analytics. Big data analytics involves the use of advanced technologies and techniques to process and analyse large and complex datasets to uncover insights and support decision making.

AST suggests that the use of big data analytics will be shaped by and shape the social structures and human agency within which it is embedded. For example, the adoption and diffusion of big data analytics within an organisation will be influenced by the goals, interests, and power dynamics of the stakeholders involved, as well as the broader social structures in which the organisation is embedded.

In addition, the decisions about which data to collect, how to analyse it, and how to use the insights generated will be influenced by the social structures and agency of the individuals and groups involved in the process. AST would also suggest that these decisions will be shaped by the goals, interests, and power dynamics of the parties involved, as well as the broader social structures in which they are embedded.

AST is used in this research to examine the introduction of innovation technology such as big data analytics and demonstrated the way the structures of big data influenced the educational system (academic). AST applied in this research also demonstrates how big data’s original intent is modified and influenced by the social structures of those academics (see figure 2.4 below). In summary, AST's appropriation model is fitted to investigate the application and penetration of big data Technology in our education system (Aktaruzzaman & Plunkett, 2016).

2.8.2 Conceptual Framework

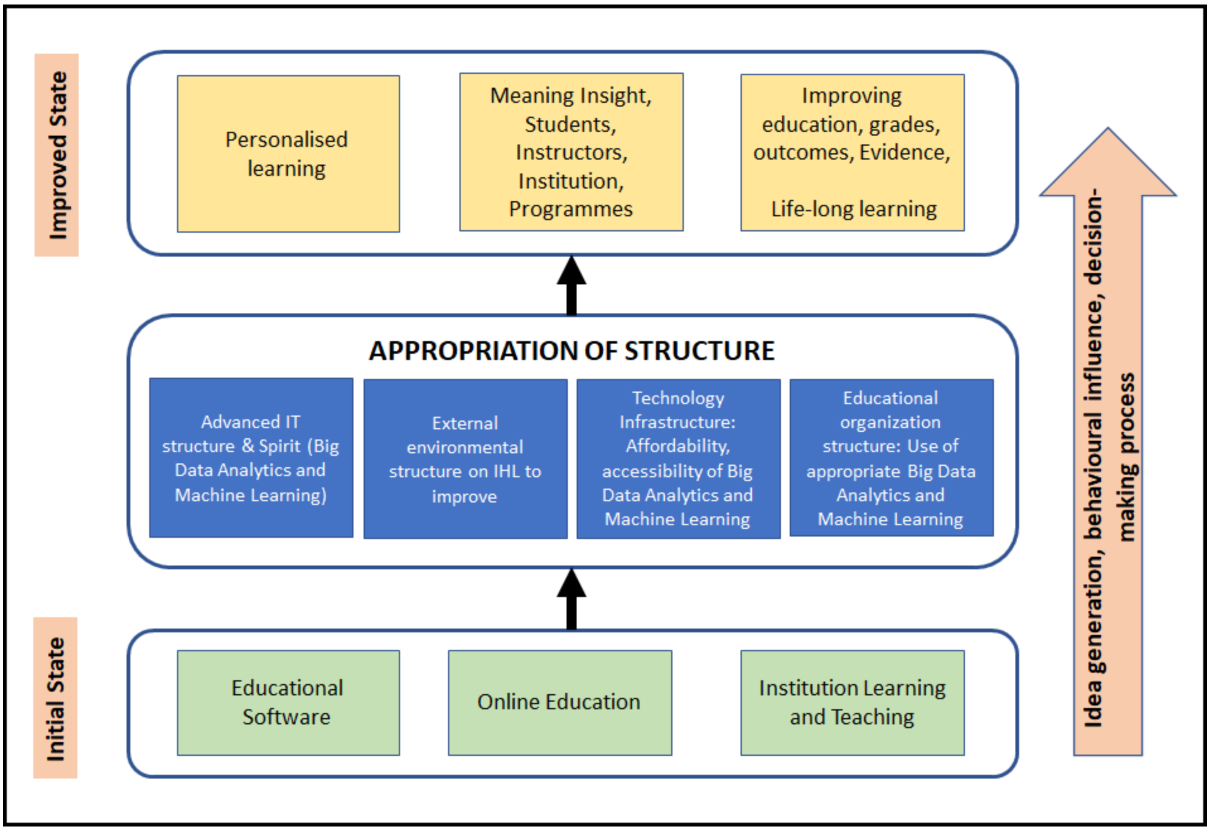

Improved State
Personalised learning
Meaning Insight, Students, Instructors, Institution, Programmes
Improving education, grades, outcomes, Evidence, Life-long learning
APPROPRIATION OF STRUCTURE
Advanced IT structure & Spirit (Big Data Analytics and Machine Learning)
External environmental structure on IHL to improve
Technology Infrastructure: Affordability, accessibility of Big Data Analytics and Machine Learning
Educational organization structure: Use of appropriate Big Data Analytics and Machine Learning
Initial State
Educational Software
Online Education
Institution Learning and Teaching
Idea generation, behavioural influence, decision-making process

Figure 2.4: Conceptual Framework

According to Poole & De sanctis (1994) advanced information technology contribution to social structures are made through structural features (rules and resources embedded in the system) and spirit (intended purpose and utilisation of the system) (Webb & LeRouge, 2009).

a) The structural features in this research represents Big Data Analytics and has been used to improve student performances. In the context of big data analytics and adaptive structuration theory (AST), advanced IT structures refer to the technological infrastructure and systems that are used to support the collection, storage, and analysis of large and complex datasets. Advanced IT structures can include hardware, software, and networks, as well as the data itself.

AST suggests that advanced IT structures are not neutral tools, but rather they are shaped by and shape the social structures and human agency within which they are used. In the context of big data analytics, advanced IT structures will be influenced by the goals, interests, and power dynamics of the stakeholders involved, as well as the broader social structures in which the organisation is embedded.

b) AST also introduces the concept of "spirits," which refers to the shared values, beliefs, and culture of a group or organisation. In the context of big data analytics, the spirits of an organisation or group will influence the adoption and use of advanced IT structures. For example, an organisation with a culture that values data-driven decision making is more likely to adopt and use advanced IT structures for big data analytics than an organisation with a culture that is more resistant to change.

Overall, AST suggests that advanced IT structures and spirits are interconnected and mutually constituting, and that they both influence and are influenced by the social structures and human agency within which they are embedded.

c) External environment structure in this research, refers to the working environment applying pressure on (IHL) to improve their curricula, learning and teaching to keep up with the marketplace. And to do so, IHL can make use of DBA to improve teaching and learning. The BDA tool use within this research Is machine learning. Machine learning ability to transform education and improve teaching and learning.

Also, when considering big data analytics and adaptive structuration theory (AST), the external environment refers to the broader social, economic, and technological context in which an organisation or group operates. The external environment includes factors such as market conditions, regulatory frameworks, and technological innovations that can influence the adoption and use of big data analytics.

AST suggests that the external environment structure is one of the factors that shapes and is shaped by the social structures and human agency within an organisation or group. In the context of big data analytics, the external environment structure can influence the adoption and use of advanced IT structures and the data collected and analysed.

For example, regulatory frameworks and data privacy laws can shape the types of data that an organisation is able to collect and analyse, while market conditions and technological innovations can influence the adoption and use of advanced IT structures. AST would suggest that the external environment structure will also be influenced by the social structures and human agency within the organisation or group, as well as the broader social structures in which it is embedded.

d) Technology infrastructure refers to the affordability and accessibility of BDA tools and technologies by IHL. In the context of big data analytics and adaptive structuration theory (AST), technology infrastructure refers to the hardware, software, and networks that are used to support the collection, storage, and analysis of large and complex datasets. Technology infrastructure can include servers, storage systems, databases, networking equipment, and analytics software, among other things.

AST suggests that technology infrastructure is shaped by and shapes the social structures and human agency within which it is used. In the context of big data analytics, technology infrastructure will be influenced by the goals, interests, and power dynamics of the stakeholders involved, as well as the broader social structures in which the organisation is embedded.

For example, the decisions about which technology infrastructure to use and how to deploy it will be influenced by the social structures and agency of the individuals and groups responsible for these tasks. Similarly, the use and maintenance of technology infrastructure will be influenced by the social structures and agency of the individuals and groups who use it on a day-to-day basis.

e) Education organisation structure is the use of appropriate Big Data Analytics by IHL. This may vary across institutions and support from IHL leadership is very important. IHL will have to "*balance conceptual technology education and the development of technology-specific skills*" (Webb & LeRouge, 2009).

Furthermore, the organisation structure of an educational institution refers to the way in which the institution is structured and the roles and responsibilities of the individuals and groups within it. Adaptive structuration theory (AST) suggests that organisation structure is one of the social structures that shape and are shaped by human agency.

In the context of big data analytics, the organisation structure of an educational institution will influence the adoption and use of advanced IT structures and the data collected and analysed. For example, the decision-making processes and power dynamics within an educational institution will shape the adoption and use of big data analytics, as will the roles and responsibilities of the individuals and groups involve in the process.

AST would also suggest that the organisation structure of an educational institution will be influenced by the external environment structure, as well as the broader social structures in which the institution is embedded. For example, regulatory frameworks and funding sources can shape the organisation structure of an educational institution, as can the goals and priorities of the stakeholders involve.

2.9 CHAPTER SUMMARY

The Chapter presented the underpinning theory Adaptive structuration theory. AST has been used in information systems to investigate the capability of advanced technologies to transform organisation. AST was applied to the problem conceptualisation to derive a conceptual framework that will drive the rest of the research.

CHAPTER 3: LITERATURE REVIEW

3.1 INTRODUCTION

The underpinning theory of this research was discussed in the previous chapter. The theoretical framework developed based on AST was discussed. Current work relevant to Big Data Analytics and how it can be used to improve students' performances is reviewed in this chapter. Hence, the chapter begins by introducing the issue with students' performances, which leads to the understanding of big Data Analytics and its contribution to improving students' performance.

3.2 STUDENTS' PERFORMANCE

Access to higher education has been made easier in the post-apartheid era in South Africa (Ng'ambi et al., 2016). However, Student performances still do not reflect high academic success (Khosa et al, 2017). Indeed, six percent of the South African population have university degrees, as reported by the Higher Education Department. This translates to just over 1.7 million people in the country holding degrees. South Africa's degree accomplishment is trailing several other countries also categorized as middle-income, although the country fairs better than other African countries. with 50-60% of first year students dropping out, University dropout rates in South Africa are reportedly incredibly high.

According to a report by the Department of Higher Education and Training, the pass rate for first-time entering university students in South Africa was just over 50% in 2018. This pass rate has been steadily declining over the past decade, and it is a source of concern for policymakers and educators.

The issue with low performance could be attributed to a wide variety of students needing personalised support, owing to different levels of readiness and under preparedness of students entering IHL (Pillay, 2017; Van Den Berg, 2017). Whether from disadvantaged or advantaged background skills like information literacy, academic writing, critical thinking, language skills, were found to be deficient (Khosa et al., 2017).

Furthermore, many factors affect or influence students' academic performances some of the factors are known while some are not known. The known factors can be divided between internal and external factors. Internal factors refer to class size, facilities, technologies use in the class, teacher's role, class environment. In South Africa, some of the following factors have been summarised as follows:

- Inadequate funding: Insufficient funding is a major challenge for higher education in South Africa. This can lead to inadequate resources and support for students, which can impact their performance (Ayuk & Koma, 2019; Kayembe & Nel, 2019; Prinsloo & Roberts, 2022).

- Inadequate support for students: Many students in South Africa struggle to afford the costs of higher education, including tuition fees, accommodation, and living expenses. This can lead to financial strain and stress, which can negatively impact their performance (Ayuk & Koma, 2019; Kayembe & Nel, 2019; Prinsloo & Roberts, 2022).

- A lack of access to quality education: Some students in South Africa do not have access to quality education, particularly those from disadvantaged backgrounds or from rural areas. This can impact their ability to succeed in higher education (Ayuk & Koma, 2019; Kayembe & Nel, 2019; Prinsloo & Roberts, 2022).

- Disparities in performance: There are significant disparities in the performance of higher education students in South Africa, with some groups performing better than others. This may be due to a range of factors, including differences in access to education and support, as well as socio-economic and cultural differences (Ayuk & Koma, 2019; Kayembe & Nel, 2019; Prinsloo & Roberts, 2022).

There have been several efforts made to improve the performance of higher education students in South Africa in recent years. Some of these efforts include:

- Increasing funding: The government has increased funding for higher education to improve the quality of education and support students. This includes initiatives such as the National Student Financial Aid Scheme (NSFAS), which grants financial assistance to students from low-income households (Ayuk & Koma, 2019).

- Improving support for students: The government and universities have also implemented initiatives to improve support for students, including tutoring programs, counseling services, and mentorship programs (Ayuk & Koma, 2019).

- Addressing disparities in access to education: as mentioned above there are significant disparities in the performance of higher education students in South Africa, with some groups performing better than others. Efforts have been made to address these disparities, including initiatives to increase access to education for historically disadvantaged groups, such as black students and students from rural areas (Ayuk & Koma, 2019).

- Improving the quality of education: The government and universities have also implemented initiatives to improve the quality of education, including efforts to attract and retain top faculty, invest in research, and modernise infrastructure (Ayuk & Koma, 2019).

- Information and communication technologies (ICTs) have also been used in South Africa with the aim to improve the performance of students. Here are a few applications:

- E-learning: Many universities in South Africa have adopted e-learning platforms, which allows students to submit assignments, grants them to access to course materials, and participate in online discussions. E-learning can help to improve access to education, particularly for students who may have difficulty attending traditional in-person classes (Mare & Mutezo, 2021; Mashau & Nyawo, 2021).

- Online tutoring and support: Some universities in South Africa have implemented online tutoring and support programs to help students who are struggling with coursework. This can include one-on-one tutoring sessions, as well as online resources and materials (Mare & Mutezo, 2021).

- Mobile learning: Many students in South Africa have access to mobile devices, and some universities have implemented mobile learning programs to take advantage of this. Mobile learning programs can include mobile apps and SMS-based systems, which can provide students with access to course materials and support on their mobile devices (Mayisela, 2013; Isaacs et al., 2019).

- Blended learning: Some universities in South Africa have implemented blended learning programs, which combine online and in-person learning. This can help to improve access to education and support for students, as well as providing more flexibility in the delivery of course materials. This also became more prominent in the post Covid era (Kayembe & Nel, 2019; Lentz & Foncha, 2021).

IHL have made use of several ways to improve students' performance, including the use of software or the analysis of big data on students (Baker & Inventado, 2014). Institutions of higher learning across the world have set for objectives to improve students' performance through improvement of teaching and learning. Due to the proliferation of software in the marketplace, Information communication technology (ICT) has been used in academical environment as well (Baker, 2010). Indeed, existing in an era with fast paced developing technologies, IHL have also decided to incorporate ICT to keep up with market changes. The results are varied.

Some Studies have shown a strong correlation between ICT use and both students' motivation and academic performances. While other studies show little to no correlation between ICT tools and students' performance (Castillo-Merino & Serradell-López, 2014; Mlambo etal., 2020). Also, on one end of the spectrum, scholars espouse the view that technology has already transformed universities. On the other end of the spectrum scholars are of the view that technology is disruptive, and universities have failed to cope with it (Gachago et al., 2013; Bozalek et al., 2013).

Big Data Analytics is now coming to the forefront, and IHL around the world are now reaping the benefits. IHL are now able to improve students' performance by customising the curriculum for one student at a time (Papamitsiou & Economides, 2014). With regards to this research, ICT refers to Big Data analytics. The next section introduces Big Data.

3.3 BIG DATA

The multiplication of data produced on the Internet is fostered by the development of Internet and social media (Kumari, 2016). This has led to the creation of new analysis tools. Big Data refers to the remarkable volume: the "*big*" volume of data emanating from the Internet (Hussien, 2020). Big Data can be retrieved in all trades: science, marketing, customer relations, transportation, health, education, etc (Smaya, 2022). A wind of revolution blows in everyone's back. The next hurricane is called Big Data! Unfortunately, too few are prepared (Wong & Hinnant, 2022). To understand the Big Data revolution, it is necessary to understand the stakes involved. This section will address the different stakes involved.

3.4 BIG DATA SYNOPSIS

The Big data emergence has extremely changed our society and will carry on drawing attentions from the public and technological experts (Smaya, 2022). Evidently, this era has been deemed a "*data flood*" era, demonstrated by the colossal volume of data emanating from multiple sources and the velocity at which it is generated (Hu et al., 2014; Callegaro & Yang, 2018). Indeed, Holst (2021) note that data has grown from two zettabytes in 2010 to sixty-four zettabytes in 2020 as per Figure 3.1 below. BD was devised to consider the significance of this data-outburst, which has led the data to be touted *"the new oil"*. Therefore, our society is expected to be transformed by Big Data (Gantz & Reinsel, 2012). There is an expected data tsunami.

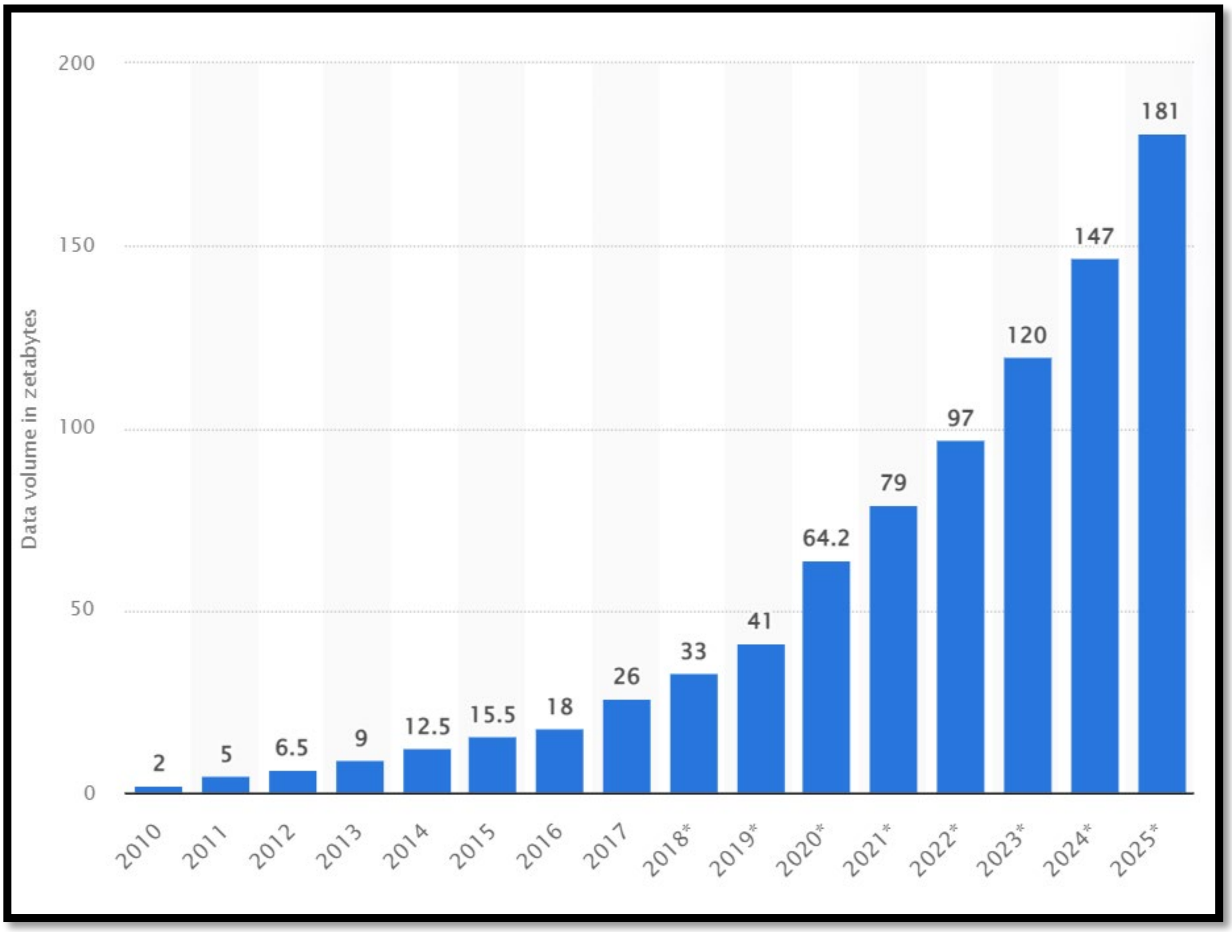
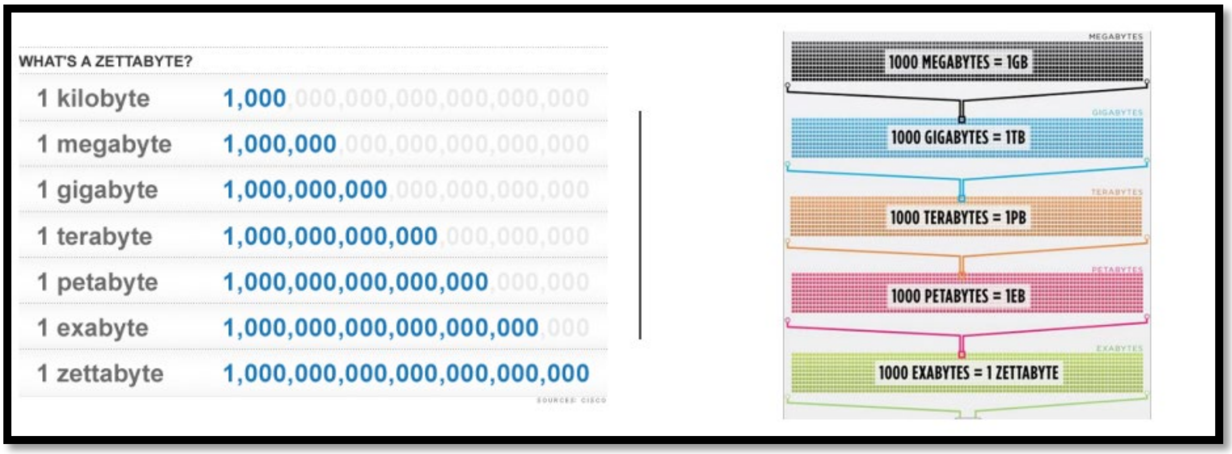


Figure 3.1 Data Volumes in Zettabytes (Holst, 2020)



WHAT'S A ZETTABYTE?

1 kilobyte	1,000,000,000,000,000,000,000
1 megabyte	1,000,000,000,000,000,000,000
1 gigabyte	1,000,000,000,000,000,000,000
1 terabyte	1,000,000,000,000,000,000,000
1 petabyte	1,000,000,000,000,000,000,000
1 exabyte	1,000,000,000,000,000,000,000
1 zettabyte	1,000,000,000,000,000,000,000

SOURCED: CISCO

Figure 3.2 Visualize a zettabyte (Savov, 2011)

This volume is more important since the expansion of the mobile Internet. Now, everyone is connected, 24/7 and constantly able to produce and receive digital data (Smaya, 2022). Moreover, the recent growth of connected objects that will lead to the new web 3. will only accentuate the significance of Big Data and its stakes. The cloud also has a great deal of responsibility in the Big Data revolution, making data storage easy and inexpensive (Ramachandra et al., 2022).

Designing a scalable big-data system encounters technical challenges due to big data uniqueness. Scalable big-data systems comprise collection of mechanisms and tools to extract load and improve dissimilar data while using parallel processing power to accomplish challenging analysis and transformations (Hu et al., 2014). Data emanates from varied sources and in increasing volume. This is the cause of increased complexity in sourcing and integrating data at scale from dispersed locations. For example, millions of accounts scattered globally produce over 175 million comprising tweets image, text, social relationship, and video (Kelly, 2013).

With regards to scalability, privacy protection, and fast retrieval, while offering performance and function, the systems (Big data) require storage and management of the collected heterogeneous and gigantic datasets. For instance, over thirty petabytes of data generated by users need to be stored, accessed, and analysed by Facebook (Kelly, 2013). It is essential for big data analytics to efficiently mine huge datasets at various echelons in Realtime or near. This would encompass visualisation, modelling, prediction, and optimisation - so that intrinsic possibilities may be discovered to enhance decision making and reap additional rewards and advantages.

The above-mentioned technological challenges require re-evaluation of traditional data management systems, extending from architectural principle to the particulars of implementation (Deepa et al., 2022). Nevertheless, relational database management systems (RDBMS) are the foundation on which traditional data management systems are mostly constructed. This makes these legacy systems inadequate while confronting the above-enumerated big-data challenges. Precisely, the incongruity between the evolving big-data and traditional RDBMS paradigm falls within the ensuing aspects.

From the viewpoint of data structure, RDBMSs cannot handle unstructured or semi-structured, while only offering support for structured data (Hu et al., 2014; Kune et al., 2016). From the standpoint of scalability, with the ever-increasing data volume, systems need to scale out with hardware commodities in parallel. RDBMSs expand with expensive hardware and cannot provide parallel processing, inappropriate to manage (Hu et al., 2014).

Numerous solutions for big data systems to tackle these challenges have been proposed in an ad-hoc manner. Cloud computing can be used as an infrastructure layer to Big Data Systems in order to meet certain infrastructure demands, such as elasticity, cost effectiveness and the ability to extend or reduce its capacity. Also, Distributed file systems and NoSQL databases are appropriate for persistent storage and management of immense datasets (Howard et al., 1988; Cattell, 2011).

Furthermore, to attain success (like website ranking) in processing group-aggregation tasks, a programming framework called MapReduce can be used (Dean& Ghemawat, 2008). Hadoop, the backbone in analysing big data, can be leveraged to integrate data storage, system management, data processing, and extra modules to construct a formidable system-level solution (White, 2012). Several big data applications can be built on these innovative platforms and technologies. Factoring the propagation of big-data technologies, a methodical framework is appropriate to apprehend the rapid development of big-data research and development and put the development in diverse frontiers in perspective (Hu et al., 2014).

3.4.1 Big data attributes

The following sections present Big Data attributes.

a) The “Vs” of big data

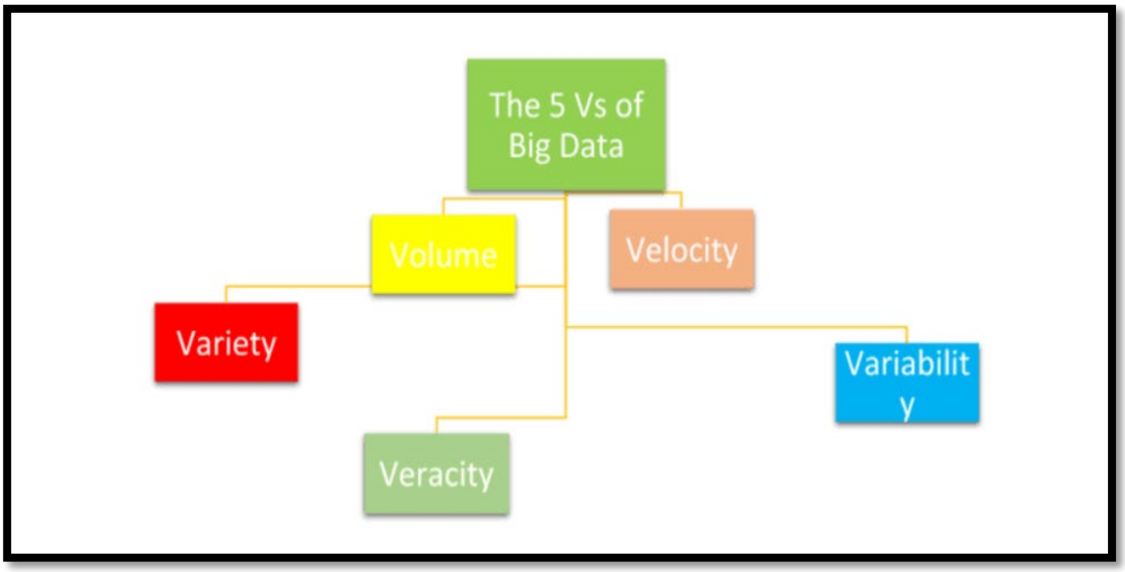


Figure 3.3: The five Vs of Big Data

Big data is conventionally defined by the “V” defined below:

- Volume: Big Data is an exceptional amount of data (Raghupathi, 2011, Zikopoulos & Eaton, 2011).

- Velocity: Big Data is a fast, real-time data processing (Raghupathi, 2011, Zikopoulos & Eaton, 2011).

- Variety: Big Data, it is varied data, taking diverse forms. So, an image, a video, a tweet, a like are datas. A simple trace left on a website following your visit, the famous cookies, or by one of your connected objects are datas (Raghupathi, 2011, Zikopoulos & Eaton, 2011).

- Veracity: Big Data poses the problem of the veracity of the data. Are they relevant, are they real? (Raghupathi, 2011, Zikopoulos & Eaton, 2011).
- Values: Big Data also raises the problem of knowing what added values bring this data. The sorting of the data is then indispensable. It is essential to select the data to be analysed, according to its activity and above all its objectives (Raghupathi, 2011, Zikopoulos & Eaton, 2011).

b) Traditional VS Big Data

Table 3.1 shows the difference between traditional vs Big Data discussed in section 3.1.

Table 3.1: Traditional VS Big Data

Attributes	Traditional data	Big Data
Volume	Gigabytes to terabytes	Petabytes to zettabytes
Organisation	Centralised	Distributed
Structure	Structured	Semi-structured and unstructured
Data model	Strict schema based	Flat schema
Data relationship	Complex interrelationships	Almost flat with few relationships

Big Data can be defined by its volume, as substantial volume of data is produced. It can also be defined by the Velocity at which it is produced, as we live in a digital age in which data is being generated at increasing pace. Big Data can be identified by its variety as the data is varied in nature, taking different forms (image, video, tweets) (Raghupathi, 2011; Zikopoulos & Eaton, 2011). Big Data can be defined as structured and unstructured datasets so large and complex, challenging to store and process using traditional methods and recent software technologies (Raghupathi & Raghupathi, 2014; Sharma, 2017).

c) Opportunities created by Big Data

Information is the fatal weapon of the 21st century (McNaught 2008; Azamfirei, 2016). The one who holds the information has the power. This is what we reiterate all the geopolitical experts and all the economists. Big Data is now the biggest source. The Big Data, to optimise its offer: it allows a complete analysis of the behaviour and the expectations of the consumer (Hofacker, 2016; Dinu & Radu, 2016). For example, Google Analytics can optimise its website by analysing real-time linked data: number of visits, browsing behaviour, bounce rate, number of pages read, clickthrough rate ...

a) The Big Data allows the optimisation of an e-commerce shop, train schedules or even a school program (Akter, 2016). With the analysis of the behaviors of Internet users according to the MOOCs (Massive open online course), it is indeed possible to analyse the interest of the courses and their contents (Porter, 2015; Al-Rahmi, 2019).

b) Big Data, to anticipate needs and demand: Big Data is the source of remarketing; this practice aims at displaying advertisements according to your navigation. Target, in the United States, thus manages to predict the future birth of pregnant women (Kuhn, 2023).

c) The Big Data, to optimise logistics and organisation: it allows, for example, following its sales in real time and thus optimising its management of stocks (Seyedan & Mafakheri, 2020).

These examples are only a handful of the opportunities that Big Data will offer. Companies, and not only, will have to be imaginative, organised and have an enormous sense of analysis to take full account of the phenomenon. From this mastery will result new uses that will upset our way of conceiving the Internet. This research will look at big data in education, as higher institutions of learning generate data from multiple sources, such as social media, email or learning software including blackboard.

3.5 BIG DATA IN EDUCATION

Higher learning institutions generate a diverse type of big data, which can be analysed to inform decision-making and improve teaching and learning practices. Some examples of the types of big data that are generated by higher learning institutions include:

- Student data: Higher learning institutions generate data on student performance, including grades, attendance, and engagement in class. This data can be used to track student progress and identify areas of strength and weakness (Murumba et al., 2017; Tasmin et al., 2020).

- Learning management system data: To provide course materials, monitor student progress, and facilitate communication between students and teachers, IHL often uses Learning Management Systems LMS. LMS data can include information on student activity, including the amount of time spent on course materials, the types of materials accessed, and participation in online discussions (Murumba et al., 2017; Matto, 2022).

- Student engagement data: Higher learning institutions may also generate data on student engagement, such as participation in extracurricular activities, use of campus resources, and interactions with faculty and staff (Matto, 2022).

- Institutional data: Higher learning institutions may also generate data on a wide range of institutional factors, including finances, enrolment, and student outcomes. This data can be utilized to improve institutional performance and inform decision-making (Tasmin et al., 2020).

Educational Big data can be collected at three different level (Fisher et al., 2020). The Data can be collected at Microlevel, which represent data created within seconds between actions that can capture data and multiple students.

a) In general, Microlevel Big Data are gathered automatically during students' interaction with their learning environment. These environments include MOOCs, simulations, intelligent tutoring systems, and games.

b) Mesolevel Big Data comprise a set of computerised written corpora gather during students' writing activities in their respective learning environment. These artifacts vary from students' assignments, social media interaction and/or writings from online discussion forums. The opportunities offered by Mesolevel Big Data include the capacity to capture students' progression in emotional states, social and intellectual abilities.

c) Macrolevel big data refer to data that is captured at the level of the institution. This includes admission data, student demographics, course enrolment, degree completion and campus data. This category of big data is usually collected over multiple years but updated infrequently.

It is necessary to note that although these categories are represented as distinct, they can intertwine (Fisher et al., 2020); there can be some form of overlaps. Social media data that might constitute mesolevel big data may have stamps that may qualify it as microlevel big data. This is not a challenge as it provides opportunities for better analysis. In the follow sections, the different categories will be probed further.

3.5.1 Microlevel Big Data

Educational Data that can occur within seconds is called Microlevel Big Data (MBD). Microlevel Big Data arise from interaction between students and data collection platforms. These platforms include simulations, Massive Open Online Courses (MOOCs), and intelligent tutoring systems. The student's interaction and the context in which it occurs is the information that comprises MBD. MBD is frequently employed to identify emotional states, cognitive strategies, or self-regulated learning behaviours (Ochoa & Worsley, 2016, Botelho et al., 2017).

a) Metacognitive and self-regulated learning Skills

Several researchers in the educational data mining community have probed metacognition and self-regulatory learning (Roll & Winne, 2015). These concepts explore student's capacity to regulate their learning processes, which is a crucial skill especially in less structured environments like MOOCs and LMS. The use of educational big data analytics approaches to analyse SRL generally involves modelling students' actions and processes performed within their learning environments. This serves to detect potential hurdles with the aim to improve students' learning, and to improve how system designers and developers can apply these ideas to enhance user interfaces.

Based on specific actions that students take and the software components they use, microlevel clickstream data provides detailed information about students' sequential and temporal patterns of behavior. According to Park et al. (2017), students' clickstream data on previewing and reviewing of course materials was used to develop and validate a way of measuring effort regulation. Students who made greater efforts to review the course materials passed the class more often.

According to Park et al. (2018), centred on student clickstream data in online courses with periodic deadlines, a measure of time management was developed and validated to identify student procrastination and the frequency of procrastination. Students who received “As” had significantly higher clickstream data and they were better at managing procrastination compared to students who received “Bs”.

b) Emotional States

Inferring non-cognitive concepts around engagement, motivation, and emotion is possible with microlevel data. One of the most extensively studied affective constructs has been confusion, frustration, engaged concentration and boredom. Various learning environments have been developed that utilise affect detection, puzzle games, including intelligent tutoring systems, and first-person simulations (Botelho et al., 2017; DeFalco et al., 2018; Hutt et al., 2019). As field observations become more common (trained individuals observing student behaviour during learning), multiple data sources are being combined so that they form a unique picture of students' behaviour.

c) Evaluating Student Knowledge

Microlevel clickstream data can be used to appraise knowledge inference (latent knowledge estimation) by comparing sets of correct and incorrect answers to problems. The most common methods are performance factors analysis (Pavlik et al., 2009), Bayesian knowledge tracing (Corbett & Anderson, 1995), and deep knowledge tracing (Khajah et al., 2016). Different frameworks are used by these methods to determine the level of mastery of specific skills by learners.

d) Using Data for Actionable Knowledge

To determine course of actions for students, such as when they disengage in online courses, big data can be used (Le et al., 2018). Also, to determine whether a student would stop working on a course, researchers at HarvardX analysed more than 2 million data points from more than 200,000 students during 10 MOOC courses. Interventions were then developed to improve student engagement using these detectors.

e) Challenges of Microlevel Big Data

Students often produce huge amounts of microlevel data; a single student may have thousands of data points. Using microlevel big data in education has been done in many ways. Analysis and Observation of phenomena that happen over a short amount of time has become possible. It is not uncommon to detect emotional state within a 20-second granularity (Botelho et al., 2017; DeFalco et al., 2018; Pardos et al., 2014), and the detectors created can be employed to analyse behavior over a yearly period (Pardos et al., 2014; Slater et al., 2016).

MBD can be relatively easy to collect, that is why most research make use of them. But the relative ease of collection does not mean there are no challenges and limitation. This is explained by the fact that MBD can lead to models that boost immediate prediction rather that models that can be persistent over time (Corbett & Anderson, 1995; Pardos et al., 2014).

3.5.2 Meso Level Big Data

Meso Level big data mainly refers to the written corpora. In the digital age, as academic writing shifts from paper to digital texts, it is increasingly possible to collect systematically collected student writing artifacts at scale. LMS, intelligent tutoring systems, website database, discussion forums, course assignments can provide the opportunity to have a large corpus of students' writing.

While Microlevel Big Data happens at a granularity of seconds, Meso Level Big Data happened within minutes to hours. Thus, the writing content produced by students may vary. For example, although a student may submit assignments weekly, that same student may engage in discussion forums daily. These various interactions increase the amount of data available.

Natural Language processing (NLP) which is an AI subfield specialising in the interactions between computers and human language can help automate the analysis of the data. NLP tasks includes the clusters of lexical, morphological, syntactic features and the patterns in student writing. The Coh-Metrix (McNamara & Graesser, 2012), for example, uses NLP to measure characteristics of texts related to discussion comprehension, including syntactical simplicity, narrative aspects, word concreteness, deep and referential cohesion. Likewise, the Word Count tool and Linguistic Inquiry

(Pennebaker et al., 2015) assesses psychological concepts such as leadership, confidence, emotional tone, and authenticity.

a) Supporting and Evaluating Cognitive Functioning

Through the use and analysis of Meso Level Big Data, it is possible to develop early warning systems that will assist in detecting students that are at risks of failing. These students would thus be catered to differently and could be provided with the additional support they need. This is possible because several research using Macrolevel have been conducted with regards to students' cognitive processes and functioning, their skills and knowledge and providing support to lecturers through auto grading.

The auto grading is built on inference of students cognitive functioning and skillset. This will help lecturers to focus their energy on students needing additional support. (Allen & McNamara, 2015 ; Yang, et al., 2015 ; Lan et al., 2015 ; Head et al., 2017 ; Allen et al., 2018 ; Crossley et al., 2018;). Furthermore, from this research, support systems were developed with the aim to provide feedback to students. This feedback thus serves to point learners toward the correct path (Price et al., 2016).

b) Supporting and Examining Social Processes

One of the benefits of big data is that it comes in various forms. Big Data can come in the form of videos, pictures, or social interaction on online discussion forums. This means that Data produced by student on these different platforms can be used to infer students' sentiments through their writings. This is vital because it can assist with establishing whether a student is struggling through his posts and written corpus. The written corpus data can also help in determining whether the teaching or the evaluation system is effective (Dzikovska et al., 2014; Hecking et al., 2016; Gelman et al., 2016; Scheihing et al., 2018; Cook et al., 2018).

c) Detecting Behavioural Engagement

Through the interaction of students on online platform data can be harnessed. The analysis of these data can lead to establishing the level of engagement and motivation of students. For example, some studies have been conducted that determined that students that are more engaged in a course will make use of personal pronouns in their interaction with the course. Sentiment analysis can also be conducted using the collected data. (Joksimović et al., 2015; Epp et al., 2017; Atapattu & Falkner, 2018).

d) Challenges of Meso Level Big Data

Meso Level data afford many opportunities to researchers. The data can help determine sentiment, motivation, and engagement. Furthermore, the textual analysis can help lecturers design courses with the aim to enhance students' engagement and facilitate peer-to-peer learning. But the applicability of the various tools emanating from this research may not have been extensively tested. In addition, students' engagement maybe affected by others factor than the ones inferred from an online platform. (Lan et al., 2015 ; Gelman et al., 2016 ; Slater et al., 2016 ; Atapattu & Falkner, 2018 ; Fesler et al., 2019).

3.5.3 Macro Level Big Data

Big data at the macrolevel are collected over long periods of time as opposed to the micro and meso levels. Student demographic and course enrolment details, course schedules and descriptions, campus living data, grade information, information about degree and major requirements are some of the university-wide institutional data. Students' demographic information, for example, is typically collected once and updated per students' request, in case changes may have happened. Nevertheless, such data can be used by administrators to improve administrative decisions, enhance student satisfaction, and boost college success, while also generating data-driven decision making.

a) Early-Warning Systems

It is important to develop early warning systems. These systems can be developed using a form of predictive analytics making use of decades of institutional data. The institution may have data that can help predict student dropout rate. The likelihood of dropout no longer needs to be established only when the student reaches out. By identifying these issues in advance, the institution can put into place pre-emptive measures. This not only increases students’ chances of success but also has found to be cost effective. The issues is then to determine appropriate data amongst institutional data that can lead to insight into enhancing students chances of success (Jayaprakash et al., 2014; Harrison et al., 2016; Chaturapruek et al., 2018).

Marist College In New York developed an early warning system that helped predict the students’ probability of failure based on academic standing, test scores, demographics, and LIMS session data. This help addressed course-level failure by creating predictive models to predict course failure (Jayaprakash et al., 2014). Students at risk of failure also received an email highlighting the fact that they were at risk of failure, and some of the steps to be undertaken to increase the likelihood of success (Jayaprakash et al., 2014; Harrison et al., 2016). it is necessary to note that these early warning systems objectives are to enhance students’ academic success. But in some of the studies it has been shown that receiving an email mentioning the potential risk of failures, led to students dropping out. These systems may sometimes have adverse impact (Jayaprakash et al., 2014).

b) Course Guidance and Information Systems

Predicting college level results from macrolevel data now is possible before students arrive on campus. With institutions having tremendous amount of data, systems have been developed to complement the early warning systems. Indeed, Hutt et al. (2018) made use of a national dataset to probe whether students would be likely to graduate within 4 years, using 166 features as predictor variables. Some of the variables included were student demographics, standardised test scores, institution-level graduation rates and academic achievement, and classification models.

There are also guidance systems that emanated from the macro level Data. Guidance systems make use of machine learning techniques to help students choose and select their courses based on certain categories. This will increase the chance of students' graduating. For instance, UC Berkeley in California made use of machine learning and historical enrolment data to suggest courses around their campuses based on students' interests (Pardos et al., 2019).

c) Challenges of Macrolevel Big Data

Institution have access to data that may improve students' performance. However, these institutions may not have the tools to utilise these data to enhance teaching and learning. Despite the benefits afforded by Macro level bid Data, there exists some challenges. Data collected in one institution may not be applicable in another. Also, Students goals not captured within the institutional data, may render the guidance systems limited. Finally, as in the case in the meso level, the results of these analyses may have adverse and unintended consequences and student outcomes.

3.6 BIG DATA ANALYTICS

Big data in a vacuum are useless and Its relevancy is established when used to derive information and insights that will drive the decision-making process (Gandomi & Haider, 2015). To extract meaningful insights from the data, processes such as Big Data Analytics (BDA) are used. Indeed, BDA refers to both the massive collection of data and their analysis (Ali et al., 2013). The aim of the analysis is to discover patterns, intelligence, knowledge, or any other information that can enhance the decision-making process (Ali et al., 2013). Gandomi & Haider (2015) break down the BDA process into text analytics, video analytics, audio analytics, social media analytics, and predictive analytics (the focus of this research).

BDA relates to the examination of large and complex datasets to unearth trends, patterns, and associations, especially referring to human behaviour and interactions. Big data analytics can be used to inform decision-making and support a wide range of activities, including marketing, risk management, and operations.

Big data analytics typically involves the use of specialised software and tools to process and analyse large datasets. These tools may include database management systems, data mining software, and statistical analysis software.

Big data analytics can be applied in a variety of sectors, including healthcare, finance, retail, and education. It has the potential to transform these sectors by providing organisations with the ability to make more informed decisions based on data-driven insights. However, it also raises significant ethical, legal, and social issues related to privacy, security, and bias.

The BDA emergence has extremely changed our society and is drawing attention from the public and technological experts. The amount of digital data continues to grow, and Big Data is in its infancy. The analysis and storage tools will continue to improve (Oracle, 2016; Rabella, 2016). One of the most visible applications of this new computer science field is the contextual advertisement which makes it possible to offer each user of the targeted proposals, elaborated proposals according to the personal traces of navigation (Manyika et al., 2011; Datamer, 2016).

Although Big Data Analytics has been used in several industries, such as gaming Industry, network security, market and business, healthcare, telecommunication, sports, education systems (Sharma, 2017) this research will look at Big Data Analytics in Education.

3.6.1 Big Data Analytics in Education

Improving education via data analytics is an ever-increasing area of focus of many scholars. Indeed, in that regard, societies such as, the international society of education data mining and the Society of learning analytics research have been created. Journals are published on these websites free of charge to assist institutions and policy makers in informing but mainly improving education (Romero & Ventura, 2010).

With the increase in the number of digital equipment, be it individual computers and all their variants (tablets, smartphones ...) or other connected objects (watches, glasses ...), the fields of application of these techniques have been open to all areas of human activity (health, transport, trade, culture, tourism ...) (Siemens & Long, 2011; Ali et al., 2013). Education is also heavily involved, so much so that all the techniques for collecting and analysing data relating to learning processes are described as "learning analytics"(Siemens & Long, 2011; Picciano, 2012). The practical applications of this type of approach are numerous, generate a lot of research, motivate many industrial developments, and pose important ethical questions.

Data analysis is now enabling education and school systems to be improved worldwide, thanks to Big Data and the proliferation of computers and digital technology in schools. Initially, computers were born in American universities. In the 1980s, they also began to appear in primary schools, colleges, and high schools. The school environment has enabled many students to learn about computers (Duderstadt et al., 2002; Marr, 2016).

Today, laptops and tablets are increasingly replacing white sheets and pens in classrooms. A large amount of data related to learning and teaching are generated from this digitisation. Schools and technology companies can now join forces to transform these data into pathways for developing better teaching methods, curricula, and remedying problems for students in difficulty (Eggers, 2007; John & Wheeler, 2012).

Analytics can improve higher education by providing feedback to instructors, by predicting students' performances and making recommendations to them. BDA can also assist in detecting undesirable students' behaviours, and contrasting courseware (Romero & Ventura, 2010). In the universities, lectures are by nature less interactive than in the lower levels of education. As a result, teachers receive little feedback on the effectiveness of their methods, until students pass or fail their exams (Anderson et al., 2014).

Predictive analytics can improve student outcomes by monitoring their performance and comparing them to those of the best students. More importantly, it can help improve school curricula by examining students' feelings during their coursework. A problem student, who does not progress enough, or a teacher criticised by many students can be detected and contacted by the administration (Borray, 2017).

Until now, identifying students in difficulty has always been a difficult task for school leaders. The choice of pupils to take charge of and who to provide additional assistance among the hundreds of individuals who pace a school or college has always been arbitrary (Vaugh, 2003; Marr, 2016). In the past, the main symptom of school failure was a drop in exam results. Today, thanks to the continuous analysis of the data of each student, it is possible to propose to each one a more personalised learning considering the centres of interest, the personal knowledge, and the individual intellectual capacities (Reschly & Christenson, 2006; Henard, 2009).

a) BDA in the education sector: improve student result

Improving outcomes for students is the overarching idea of using Big Data in education. Currently, answers to tests and assignments represent the only measure of the student's performance. However, each of the students creates a unique data trail throughout his or her life. However, each of the students creates a unique data trail throughout his or her life. However, each of the students creates a unique data trail throughout his or her life. To gain a better understanding of the student's behaviour and to create an optimum learning environment for students, it will be useful to analyse that data trail in real time. (Avella et al, 2016; Bhanu, 2018; Abdullah & Fahad, 2019).

Monitoring of student's behaviour, including how long it takes them to respond to a question, the resources they have used for exam preparation, their decisions to skip questions is possible thanks to the vast amount of data available in the education sector. This tracking of students' performances can be in giving each student instant feedback (Bhanu, 2018; Baig et al., 2020; Li & Jiang, 2021).

b) BDA in the education sector: customise programs

Customizable programs for each student can be developed with the help of big data. Regardless of the number of students at universities and colleges, tailored programmes may be developed for each student. Using what is known as "blended learning," a blend of internet and offline education, this may be achieved. (Bhanu, 2018; Abdullah & Fahad, 2019).

This enables students to take classes they're interested in and work at their pace while maintaining the possibility of being taught by professors on an offline basis. This is already apparent in the case of Massive Open Online Courses that are being developed worldwide. For example, only 400 students took the machine learning course taught by Andrew Ng at Stanford University. But it had attracted 100,000 students when the same course was delivered as a MOOC. (Joshi, 2017; Bhanu, 2018; Baig et al., 2020; Li & Jiang, 2021).

c) BDA in the education sector: reduce dropouts

The number of students dropping out of school and college would also be reduced as big data in the education sector would help improve the results of students. To predict future students' results, learning institutions can use predictive analytics on all the data collected. (Bhanu, 2018; Li & Jiang, 2021). To minimise the need for trial, and error, these predictions may be used to run scenario analysis of a course prior to its introduction into the curriculum. Indeed, With Big data, it is now possible to monitor and evaluate how students perform on the job market after graduation. future students would find this beneficial in their quest for the right learning institution and courses (Joshi, 2017; Bhanu, 2018; Abdullah & Fahad, 2019; Baig et al., 2020;).

d) Using Big Data to Improve Student Performance

From elementary school to university, big data is affecting education across the spectrum of learning. Big data systems help teachers and lecturers learn more about people's behaviour and form new conclusions, as the standard of technology and education is evolving (Bhanu, 2018; Baig et al., 2020; Li & Jiang, 2021).

Therefore, it is increasingly important for teachers and lecturers to understand the latest developments in education and data analytics. Today's teachers use big data technologies to find students' problems area, instead of relying on standardized tests to detect problems. Through adaptive learning, students can allocate more time in challenging subject areas while remaining in step with their classmates. (Joshi, 2017; Bhanu, 2018; Li & Jiang, 2021).

Educators are effectively using big data systems to monitor students' progress and likelihood of advancement, and to assess students accurately. Currently, there has been a significant improvement in the results achieved by learning institutions that have implemented big data systems to monitor and evaluate student performance. The development of educational plans has also been facilitated by technology to improve the student's engagement (Joshi, 2017; Abdullah & Fahad, 2019).

In recent years, several higher education institutions have embraced analytics as an essential component of their operations. Their main objective was to design a curriculum that captures the interest of prospective students and ensures their full engagement, maximal success in student retention and graduation rates as well as securing financial support from the government (Felton, 2016).

There are several institutions in universities and colleges across the USA that have already integrated BDA. The goal is to enhance the efficiency and effectiveness of core units within the institution, specifically focusing on student behaviour, activities, and providing necessary support and analytics to align university funds with their institutional objectives.

Arizona State University employed BDA with the primary goal of improving students' academic experience. The application of analytics has led to a 20% increase in the graduation rate (Zinshteyn, 2016; Attaran et al., 2018) at the university. The College Scheduler software is a specially designed tool that enables students to personally input their individual schedules, personal and academic data and transforms them into a dashboard to provide insights and analysis. The system considers the students' academic performance, as well as any sensitive information provided, ensuring confidentiality throughout the process.

The College Scheduler software is designed to meet individual needs and automatically suggests the necessary courses for them (Wells, 2016). This is beneficial because it prevents students from wasting time on courses that are not relevant to their majors. According to the findings of Zinshteyn (2016), College Scheduler software proved effective in both time and financial management. This approach could potentially increase college completion rates by over 3%.

Moreover, another university adopted this strategy successfully. the University of Maryland – College Park focuses on managing students' activities. By utilizing analytics to anticipate student achievement or lack thereof, they can enhance performance and provide proactive advising. The college has achieved positive outcomes by actively promoting minimized success disparities for poor students. By reducing the duration of graduation, it resulted in enhanced rates of graduating (Wells,2016).

By using predictive analytics, they can intervene with struggling students before it is too late. This analysis assists in identifying bottlenecks or other problems that could cause a student to drop out, such as difficult courses or pressing issues. As part of their approach, the university discovered through their research that late enrolment often resulted in poor performance for students (Attaran et al., 2018).

Consequently, they implemented a policy preventing last-minute class enrolments but still allow dropping classes within four days without penalty. To further enhance their efforts towards improving graduation rates by increasing success among enrolled individuals; University offers "intrusive advising" when necessary - providing guidance either regarding grades improvement methods or exploring alternative majors where applicable (Attaran et al., 2018).

Concordia University Wisconsin (CUW) has successfully implemented an analytics program to identify students who are at risk and provide them with assistance. Blackboard intelligence and learning analytics solutions are utilized to monitor the progress of students in their academic commitments, enabling early intervention for those who may be struggling.

Student advisors focus on student performance as well as various risk factors identified through data analysis using dashboards. This use of data analysis has significantly improved CUW's retention rate by 10%. In 2016, the university had a retention rate of 72%, but after implementing analytics along with increased involvement from faculty and administration, CUW was able to achieve an impressive retention rate of 82% within just one year.

Table 3.2: Applications of analytics in American colleges and universities (Attaran et al., 2018)

INSTITUTIONS	PROCESS TARGETED	OBJECTIVES	BENEFITS GAINED
MICHIGAN STATE UNIVERSITY	University advancement department	Identifying potential donors and providing deep insight into an individual alum's potential to give	Improved director and associate director productivity; improved visibility of donor patterns; improved overall user productivity; annual labour savings of US$34,434
CUW	Student performance management	Identifying at-risk students and helping them	Increased student retention rate to 82%, a 10% increase in 1 year
UNIVERSITY OF CALIFORNIA – SANTA BARBARA	University advancement department	Improving visibility of who will donate and repeat donors	Saved time and money; exponentially increased yearly revenue from donors
ARIZONA STATE UNIVERSITY	Student performance management	Improving students' course schedules	Graduation rates climbed by 20%
UNIVERSITY OF MARYLAND – COLLEGE PARK	Student performance management	Predicting student success or failure and intrusive advising	Helped narrow achievement gaps for minority and low-income students; improved graduation rates; shortened graduation time
MOUNT ST MARY'S UNIVERSITY IN EMMETSBURG	Student performance management	Predicting which incoming freshmen were unlikely to succeed	Boosted graduation rates by helping struggling students
GEORGIA STATE UNIVERSITY	Student performance management	Timely intervention for students with high chance of dropout	Graduation rates raised by six percentage points; eliminated achievement gaps for low-income and minority students
JOHNS HOPKINS UNIVERSITY	Student performance management	Flagging students who are missing assignments or skipping class	Helped increase graduation rates
UNIVERSITY OF TEXAS AT AUSTIN	User management and security	Timely and accurate intervention in security threats and incidence across the	Faster insight into anomalies; improved security posture; educed organizational

		distributed university network	risk; reduced incident investigation time
WASHBURN UNIVERSITY	Student performance management	Retention and graduation	Helped increase graduation rates
SINCLAIR COMMUNITY COLLEGE	Student enrolment management	Generate notification to send to individual student	25–33% increases in enrolments year over year
UNIVERSITY OF OREGON	Financial aid programme	Redesign the merit-aid programme to recruit high achievers more effectively	Enabled deeper insights into the behaviour of applicants who were accepted and offered merit aid, thus increasing the likelihood that these students would enrol; learned how much merit aid is needed in a financial aid package to make high-achieving, in-state students more likely to enrol
DELAWARE STATE UNIVERSITY	Student enrolment management	Identify students at risk and streamline best advising practices	Increased student retention rate to 70%; strived to improve 4-year graduation rate by 4 percentage points per year

South African IHL are increasingly acknowledging the global adoption of data analytics to uncover data-driven solutions for their challenges. Pre-admission screening tests were formerly utilized to draw in fresh applicants with strong academic credentials and predict student success (Cele, 2021).

These tests are increasingly being questioned as potential tools for social exclusion, particularly in South African higher education. These entrance exam results are no longer the accepted norms, even if they were once sufficient to predict both the duration of the study and the success of the students (Ningrum & Ekayani, 2019).

Also, the current student grade-linked data analytics is that it frequently leads to interventions that target the curriculum, the environment, or the student's cognitive ability too late in the learning process. Although Big Data analytics can enhance IIHL decision-making in South Africa, there are several impediments. These challenges include the following:

- a lack of resources, such as qualified staff, infrastructure constraints, and the right technology for analysis.
- a lack of knowledge pertaining to the techniques for data analysis and interpretation.
- IHL's readiness and capacity to handle Big Data (Cele, 2021).

Thus, as suggested by a study by Van Vuren (2020), it is important to contextualize approaches when implementing big data analytics with the above-mentioned challenges. The study stresses the importance of providing support tailored to students' specific needs and considering socio-economic factors. This is what the research is aiming to address.

3.6.2 BDA and Machine Learning

BDA within the context of this research refers to the tools used to process educational or other big data. Most of these tools make use of machine learning techniques. Applying machine learning to predict students' grades is becoming widely used, as there is a growing presence in the literature (Sweeney et al., 2016; Ren et al., 2017; O'Connell et al., 2018;). This is due to the availability of the data and techniques that increase the precision to which prediction can be done.

Within the first semester, early dropout rates have been recurrently researched (Aguiar et al., 2014; Chen et al., 2018; Gray et al., 2016; Zhang & Rangwala, 2018). For instance, it was established by Pardos et al. (2019) that failing courses in basic science (rather than computer science) was the cause of early dropouts in student trajectories. Thus, to increase retention rates focusing tutorial resources on these science courses was suggested. The work was then extended to model On-time versus over-time graduation.

3.6.2.1 Machine Learning and Big Data: definition and explanations

Computers can now learn without explicitly programmed. This is achieved through an artificial intelligence (AI) technology called machine learning (ML). But, to learn and grow, computers require data to analyse and train on (Janiesh et al., 2021). Therefore, Big Data is the core of ML, and in turn, machine learning is the technology that fully unlocks the potential of Big Data (Sun & Scanlon, 2019).

Machine Learning is not a novel concept. However, the exact definition is still confusing for many individuals. ML refers to the science of making predictions, discovering patterns from data based on data mining, statistics, predictive analysis, and pattern recognition (Tseng et al., 2020; Lavin et al., 2022; Ogunleye, 2022). The best-known machine learning algorithm is a perceptron. The first created was at the end of the 1950s.

In case where there is a need to discover insights from vast sets of differing and changing data, ML has been very effective. For Big data, ML has proven to be more effective in accuracy and speed, than the traditional methods (Zhou et al., 2017; Sarker, 2021). For instance, ML can discover possible fraud in a millisecond with transactional data on location, amount, and other social and historical data, (Oza, 2018; Lim et al., 2021; Walker, 2021; Seify et al., 2022).

To create models from data, ML includes several algorithms (Awad & Khanna, 2015). Contrary to traditional computer systems, ML systems do not follow instructions, rather, they learn from experience. This means that as the algorithm is exposed to more data, ML systems performance improve as they are "trained" (Mosqueira-Rey, 2022).

3.6.2.2 The different types of Machine Learning algorithms

ML algorithms can be classified as supervised, unsupervised and reinforcement learning.

a. With regards to supervised learning, data utilized for training is “tagged”. This helps the ML model in knowing what element or pattern to watch for in the data. The work conducted on the so-called tagged data is extended to untrained and unlabelled data. Thus, the model is trained to find the same elements or patterns on unlabelled data. The supervised learning algorithms include regression algorithms (numerical predictions) and classification algorithms (non-numerical predictions). One of these two models can be used, depending on the problem to be solved (Sah, 2020; Sarker, 2021).

b. Unsupervised learning, on the contrary, involves training the model on unlabelled data. The ML model goes through the data without any prior guidance with the aim to discover recurring trends or patterns. This method is usually used in fields like cybersecurity. Among the unsupervised models, we distinguish clustering algorithms (to find groups of similar objects), association (to find links between objects) and dimensional reduction (to choose or extract features) (Sah, 2020; Sarker, 2021).

c. A third approach is reinforcement learning. In this case, the algorithm learns by trying again and again to achieve a specific goal. He can try all kinds of techniques to achieve this. The model is rewarded if it gets close to the goal or penalised if it fails. By trying to get as many rewards as possible, it gradually improves. As an example, we can cite the AlphaGo program which triumphed over the Go game world champion. This program was trained by reinforcement (Sah, 2020; Sarker, 2021).

Table 3.3: Type of machine learning algorithms (Sah, 2020; Sarker, 2021)

Learning Types	Data processing tasks	Distinction norm	Learning algorithm
Supervised learning	Classification/regression /estimation	Computational classifiers	Support vector machine
		Statistical classifiers	Naïve Bayes
			Hidden Markov Model
			Bayesian networks
		Connectionist classifiers	Neural network
Unsupervised learning	Clustering/ prediction	Parametric	k-means
			Gaussian mixture model
		Nonparametric	Dirichlet process mixture model
			X-means
Reinforcement Learning	Decision making	Model-free	Q-learning
			R-learning
		Model-based	TD learning
			Sarsa Learning

The following illustrates the different algorithms and their potential benefits and limitations:

Table 3.4: Table 3.4 Pros and cons of machine learning algorithm (Sah, 2020;Kumar & Sowmya, 2021; Sarker, 2021)

Algorithm	Algorithm Type	Data Processing Tasks	Pros	Cons	Applications
Decisions Tree	**Supervised learning**	**Classification or regression**	• **Can handle both nominal and numerical attributes** • **Capable of dealing datasets that contain errors** • **Rapidly express complex alternatives clearly**	• **Sensitivity to the data set to irrelevant attributes and to noise.** • **Require the target attribute to have only distinct values.**	• **Image classification** • **Text categorisation**
K-means	Unsupervised learning	Clustering	• Computationally faster for big datasets if k is small. • Produce tighter clusters if the clusters are globular.	• Difficult to analyse k-value, the number of clusters. • Sensitive to scale	• Document classification • Customer segmentation
Support Vector Machine	Supervised Learning	Classification or regression	• Avoids overfitting • Easy to explicate results.	• For large datasets, training time is more. • Exhausting for data of different types.	• Pattern Recognition • Stock Market Forecasting
Neural Networks	Semi-supervised Learning	Classification or prediction	• Accurate predictive performance. • Have tolerance to noisy data.	• Determination of variable selection method. • Extrapolation.	• Driving • Fraud Detection
Naïve-Bayes Classifier	Supervised learning	Classification or regression	• Relatively simple and easy to use. • Not sensitive to missing and noisy data.	• Interprets all the data are equally important and independent. • Computational approximation is required.	• Spam filtering • Sentiment Analysis
K-nearest neighbour algorithm	Supervised learning	Classification or regression	• Easy to understand and interpret. • Gives more accuracy.	• Computationally expensive. • Usage of more memory.	• Face Recognition • Medical Imaging Data

The biological visual cortex architecture inspired Artificial neural networks (ANN). Deep Learning consists of a set of techniques that allow a neural network to learn through many layers to identify characteristics (Majaj & Pelli, 2018; Kindel et al., 2019; Alzubaidi et al., 2021). Many layers are hidden between the input and the output of the network. Each layer is constituted of artificial neurons. The data is processed by each layer and the results are transmitted to the next one.

Deep learning is a subset of ML while ML is a subcategory of AI. Visual recognition is among some of its most common application (Kindel et al., 2019; Brown, 2021). For instance, an algorithm is trained to identify faces from pictures from a camera. The algorithm will thus be able to detect a wanted individual in a crowd, identify the satisfaction rate when customers are going out of a store by spotting smiles. All this would of course, depend on the dataset provided etc. Also, the voice, the expression of a questioning, the tone, words, and affirmation can also be recognised by a set of algorithms (Lu & Yan, 2021; Sun et al., 2022; Tan, 2022).

On the other hand, this technique needs a lot of data to train and obtain sufficient success rates to be used. A Data Lake or Data Lake is essential to perfect the learning of Deep Learning algorithms. Deep learning demands additional computing power to work.

3.6.2.3 Machine Learning Use Case and applications

Machine Learning drives several modern and popular services. Some of the examples are the recommendation engines used by YouTube, Netflix, Spotify, Amazon (Gironacci, 2021). This also applies to web search engines like Google or Baidu. Twitter and Facebook, and other social networks make use of ML, as do voice assistants such as Alexa and Siri.

All these platforms collect data about users, to better understand them and improve their performance. Algorithms need to know what the viewer is looking at, what the Internet user clicks on, and what publications he reacts to on the networks. This way, they are then able to come up with better recommendations, answers, or search results (Portugal et al., 2015; Cena et al., 2020; Fayyaz et al., 2020; Tao et al.,2022).

Machine Learning systems also excel in the field of games. AI has already surpassed humans in the game of chess, checkers. She also manages to triumph over the best video game players like Starcraft or Dota 2 Togelius, 2019; Menon, 2021). ML is also employed for automatic linguistic translation, and for the conversion of spoken speech on the screen (speech-to-text) (Limbu, 2020; Vashisht et al., 2021; Madahana et al., 2022) . Another use case is sentiment analysis on social networks, also based on natural language processing (NLP) (Lappeman et al., 2020; Ahmed et al., 2022; Babu & Kanaga, 2022; Omuya et al.,2022).

ML is utilized additionally for the automatic classification and analysis of medical X-ray images (Ahmed et al., 2022; Alsaaidah et al., 2022; Erdaw & Tachbele, 2022). AI is proving to be very efficient in this area, occasionally even more than human experts in detecting anomalies or diseases. However, it cannot yet completely replace specialists given the stakes. Several companies have attempted to leverage machine learning to automatically review candidate resumes. However, training data biases lead to systematised discrimination against women or minorities (Lee et al., 2019; Castaneda et al., 2022).

Indeed, Machine Learning systems tend to favour candidates whose profile is like current candidates. They therefore tend to perpetuate and amplify the discrimination that already exists in the business world. This is a real problem, and Amazon, for example, preferred to cease its experiments in this area. Many companies are trying to combat bias in AI training data, such as Microsoft, IBM, or Google (Li, 2020).

Controversial facial recognition technology is also based on machine learning. However, here again, biases in the training data pose a serious problem. These systems are mostly trained on photos of white males, so their reliability is much lower for females and people of colour. This can lead to errors with terrible consequences. For example, innocent people have been mistaken for criminals and wrongly arrested (Lohr, 2018; Findley, 2020; Najibi, 2020) .

3.6.2.4 Machine Learning and Big Data

To take full advantage of the benefits of Big Data, traditional analytic tools are not powerful enough. The volume of data makes it impossible to carry out broad analysis; also, the correlation and relationships among the data are too critical for analysts to study all hypotheses (for value to be derived from this data). (Sivarajah et al, 2017; Hariri et al., 2019; Batko & Slezak, 2022).

Business intelligence and reporting tools that report on amounts, prepare accounts, or do SQL queries use basic analytical methods. Online analytical processing is a systematic extension of these basic analytical instruments requiring human intervention in determining the calculation to be carried out. (Lloyd, 2011).

To exploit the hidden opportunities of Big Data, ML is deemed the ideal solution (Zhou et al., 2017). Without the necessity to rely on human beings, this technology makes it possible to obtain value from huge and diverse data sources. Machine Learning is driven by data and is appropriate for the complexity of vast Big Data sources. ML can be employed on growing data sets Unlike traditional analytical tools. To obtain quality insights, a ML system can learn and grow by being fed more data. (Lavin et al., 2022; Tercan & Meisen, 2022). ML thus makes it likely to uncover patterns hidden in the data with more efficiency and effectiveness than human intelligence.

Machine learning and artificial intelligence would have been ineffective in the absence of big data. Data is the tool enabling Artificial intelligence to process and learn from the way humans think. The learning curve is accelerated by big data, which allows for the automated analysis of information (Muller et al., 2016). ML systems learns and become more accurate the more data it receives (Luan et al., 2020; Rahmani et al., 2021).

Artificial intelligence is now capable of learning by itself, without human assistance. The lack of available data and the inability to analyse vast amounts of data in seconds had limited progress until now (Xu et al, 2021). Nowadays, data is available any time and in real time. This allows AI and Machine Learning to transition to a data-driven approach (West & Allen, 2018). The technology is nimble enough now to retrieve and analyses massive datasets. Indeed, companies of all sectors are now aligning themselves with Amazon and Google in implementing AI solutions for their businesses.

3.6.2.5 Machine Learning and Big Data: Predictive analytics

To predict the probability of trends and financial results in a company, on basis of past data, predictive analytics uses data, algorithms, statistics, and ML techniques. Predictive analytics combine numerous disciplines and technologies such as data mining, statistical analysis, predictive modelling, and Machine Learning to forecast the future of businesses. For instance, the reactions of consumers or the consequences of a decision can be anticipated (Kumar & Garg, 2018; Zhang, 2020; Henrys, 2021; Rustagi & Goel, 2021; Wach & Chomiak-Orsa, 2021).

Predictive analytics help generate actionable insights from large datasets, so businesses can decide which direction to undertake and deliver a better customer experience. Many businesses are now able to use predictive analytics due to the increase in computing power, data, and the development of easier-to-use AI software and analytical tools like Salesforce Einstein.

According to a study conducted by Bluewolf (2016) with 1,700 Salesforce customers, 75% of companies that increase their investments in analytical technologies benefit from it. 81% of these Salesforce product users believe that using predictive analytics is the most important initiative in their sales strategy. Predictive analysis makes it possible to automate decision-making, and therefore increase the profitability and productivity of a company.

Machine learning and AI represent the next stage of data analysis (Kibria et al., 2017). Cognitive computing systems are constantly processing and learning about the business Landscape. The systems also intelligently predict consumer's needs, industry trends, and more (Da Costa et al., 2022). the level of cognitive applications is yet to be reached by few companies. Cognitive applications can be described by four characteristics: the capacity to extract ideas and reason, the capacity to evolve in expertise with each interaction, the process and understanding of unstructured data, and the capacity to speak, see, and hear to interact with humans in a natural way (Appel et al, 2017; Rao & Gudivada, 2018; Da Costa et al., 2022; Khurana et al., 2022). To do this, algorithmic processing of natural languages needs to be developed. It is important to note that research attempts to predictive students' performance. This will thus be included in predictive analytics.

3.7 Research Gaps

Several research gaps can be identified from the literature reviewed on using big data analytics to improve student performance in higher education. Firstly, while there are some studies exploring the use of learning analytics and educational data mining techniques, there is limited research specifically applying predictive analytics and machine learning algorithms to forecast and improve student academic outcomes. Most prior work focuses on descriptive and diagnostic analytics to understand factors influencing student performance. More research is needed on developing and validating predictive models that can accurately classify student success metrics like course grades, grade point averages, and on-time graduation.

Secondly, much of the existing research lacks sufficient rigor in terms of model performance benchmarks, testing on out-of-sample datasets, and statistical significance testing. Many studies report only overall accuracy measures on the training dataset, which can be misleading. There is a need for more thorough evaluation of predictive models using sound data science principles before deploying such systems in real-world educational settings.

Thirdly, while there are some examples of using analytics in Western and developed country contexts, there appears to be very limited published research on leveraging big data to enhance student outcomes in developing countries like South Africa. The few studies in African higher education highlight challenges around data quality, accessibility, infrastructure, and cultural differences that warrant further investigation within these contexts. More research situated in the South African university landscape could uncover additional considerations when applying data-driven approaches in this environment.

Finally, there are ethical implications of utilizing analytics in education that require further exploration. As models rely more on automating decisions that impact students, issues around transparency, accountability, and bias emerge. There is a responsibility to ensure predictive systems do not propagate historical disadvantages for marginalized groups. Research frameworks and governance policies co-developed with students and faculty could help address these concerns.

So, the key research gaps include a lack of predictive modelling research specifically focused on improving student performance, deficiencies in model evaluation rigor, a scarcity of studies grounded in developing world contexts like South Africa, and minimal research attention to the ethics of educational data mining. Addressing these interdisciplinary research gaps could significantly advance the use of big data analytics to enhance teaching, learning, and student success in higher education.

The current work uses the Big Data definition to classify machine learning difficulties to understand their origin. Furthermore, a range of machine learning methodologies are examined, and how they can tackle established obstacles is deliberated. This helps researchers to decide more intelligently which learning paradigm or approach to apply depending on the particular Big Data scenario. Additionally, it enables the identification of opportunities and research gaps in the field of machine learning using big data. As a result, this work facilitates and provides a thorough basis for further research.

3.8 SUMMARY

Big Data Analytics (BDA) can assist in providing teaching and learning structure that empowers and enhances students' performance. Also, the BDA has the potential to help Universities to operate more efficiently, enable teachers' effectiveness to improve their methods, and prevent students from falling into poor academic performance (Reyes, 2015). It is therefore essential to exploit the opportunities offered by this technology to improve and possibly, maximise output and performance (Marr, 2016; Kochetkov & Prokhorov, 2017).

Despite some delay compared to other industries, the education sector is now integrating Big Data and its benefits into its organisation. This delay is mainly due to the perception of the collection and analysis of educational Big Data as an intrusion of governments attempting to monitor and study citizens. Others see it as an attempt by companies to identify and format their future clienteles. These fears and reticence are entirely acceptable. However, the quality of education is an essential factor in the successful development of a society. BDA holds the potential to help schools operate more efficiently, enable teachers to improve their methods, and prevent students from falling into school failure. It is therefore essential to exploit the opportunities offered by this technology to the maximum.

CHAPTER 4: CROSS-INDUSTRY STANDARD PROCESS

4.1 INTRODUCTION

The previous chapter discussed current literature on the use of Big Data analytics concepts in Education to improve students' performance. This chapter discusses the phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM) that are relevant to this study: domain understanding, data understanding, data preparation, application of machine learning, evaluation, and the use of discovered knowledge (Kurgan and Musilek, 2006).

This research attempted to utilize students' data to predict students' performance. This is to derive a way to automate and improve student performance from these predictions by identifying students at risk. Thus, the research sought to build a model that would perform the predictions. This was done by following the revised framework of CRISP-DM (Shearer, 2000; Schröer et al., 2021).

4.2 Adaptive Structuration and CRISP-DM

Adaptive structuration theory is a theoretical framework that helps to explain how individuals and organizations make use of information and communication technologies (ICTs) to accomplish tasks and goals (Maley, 2020). The theory suggests that individuals and organizations are constantly adapting their use of ICTs to achieve their goals and that the way they use these technologies is shaped by the social and organizational structures in which they operate. It suggests that technology is not a passive tool but rather an active force that shapes social structures and practices (Calloway, 2010).

Adaptive structuration theory could be associated with the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework via the concept of "structuration." This concept refers to the ongoing process of negotiation and interpretation that occurs as individuals and organizations use ICTs to accomplish tasks and goals. It suggests that individuals and organizations must continually adapt their use of ICTs to make sense of the information and resources available to them and that this process of adaptation is influenced by the social and organizational structures in which they operate (Rains & Bonito, 2017).

In the context of data science, the CRISP-DM framework provides a structured approach for identifying and defining the goals of a data science project, evaluating the available data and resources, and developing and implementing a plan for collecting and analyzing data (Kristoffersen et al., 2019). By following this structured process, individuals and organizations can better navigate the complex and rapidly changing landscape of data science and make more informed decisions about how to use ICTs to accomplish their goals. This aligns with the concept of structuration in adaptive structuration theory, as it suggests that the CRISP-DM framework can help individuals and organizations adapt their use of ICTs to achieve their goals in a more efficient and effective manner.

In this sense, CRISP-DM can be seen as a way of using technology to structure and guide the process of data science in a way that is both efficient and effective. It can be thought of as an example of how technology is used to shape social practices and structures, as described by adaptive structuration theory.

4.3 THE CRISP-DM FRAMEWORK

CRISP-DM Model has been used in education for data mining as educational data mining (EDM). The figure below illustrates the initial model steps as proposed by Kurgan and Musilek (2006) and Martinez-Plumed et al. (2019). This framework is relevant to this research because it proposed an approach that guided the predictability of students' performance, as initially discussed in Chapter 2 Figure 2.3 (Adapted Duality of Technology for BDA) as a lens to understand and interpret the socio-technical processing between teaching and learning, Institutional properties, and BDA (machine learning). A minor revision was made to the CRISP-DM in the context of this research to include the application of machine learning. Figure 4.2 below illustrates the revised model of Figure 4.1, in which the steps are described as follows:

a) **Domain Understanding:** in this step of the model, goals are established. This entails seeking understanding, clarity, and relevant information that may be required. It also involves identifying the relevant stakeholders.

b) **Data Understanding:** entails the gathering of the relevant data. Then, the data is checked for completeness, whereby missing data is verified, as well as examined for redundancy. In this step, it also important to determine the usefulness of the data to contributing to the model that is going to be built.

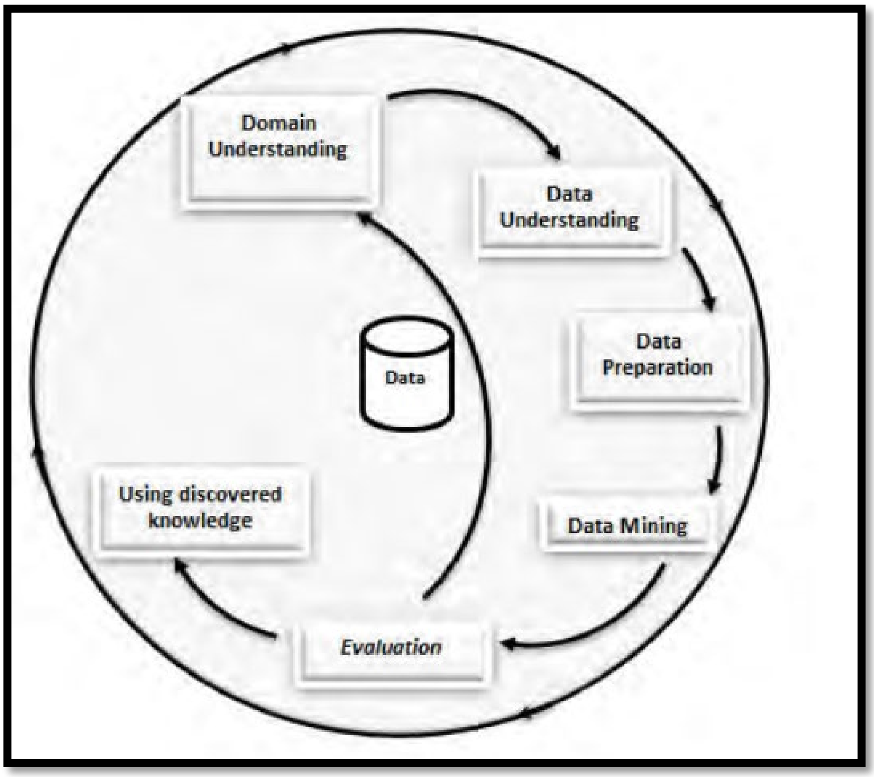


Figure 4.1: CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)

c) **Data preparation:** in this step, data is cleaned and transformed to prepare it for use in the model. The output of this step should be a dataset that is suitable for use in the next step.

d) **Application of machine learning:** this entails selecting amongst the different machine learning algorithms. Between regression, classification, and clustering). Once the correct models are selected, they are then applied to the data.

e) **Evaluation:** this step entails interpreting the results from the application of machine learning. This involves the discovery of new patterns. This step may also involve reconsidering prior steps to identify alternatives to improve the models.

f) **Using the discovered knowledge:** this may involve applying the model into a performance system, or simply documenting the discovered knowledge, to transfer to interested parties. These steps have been used in this research and they are discussed next.

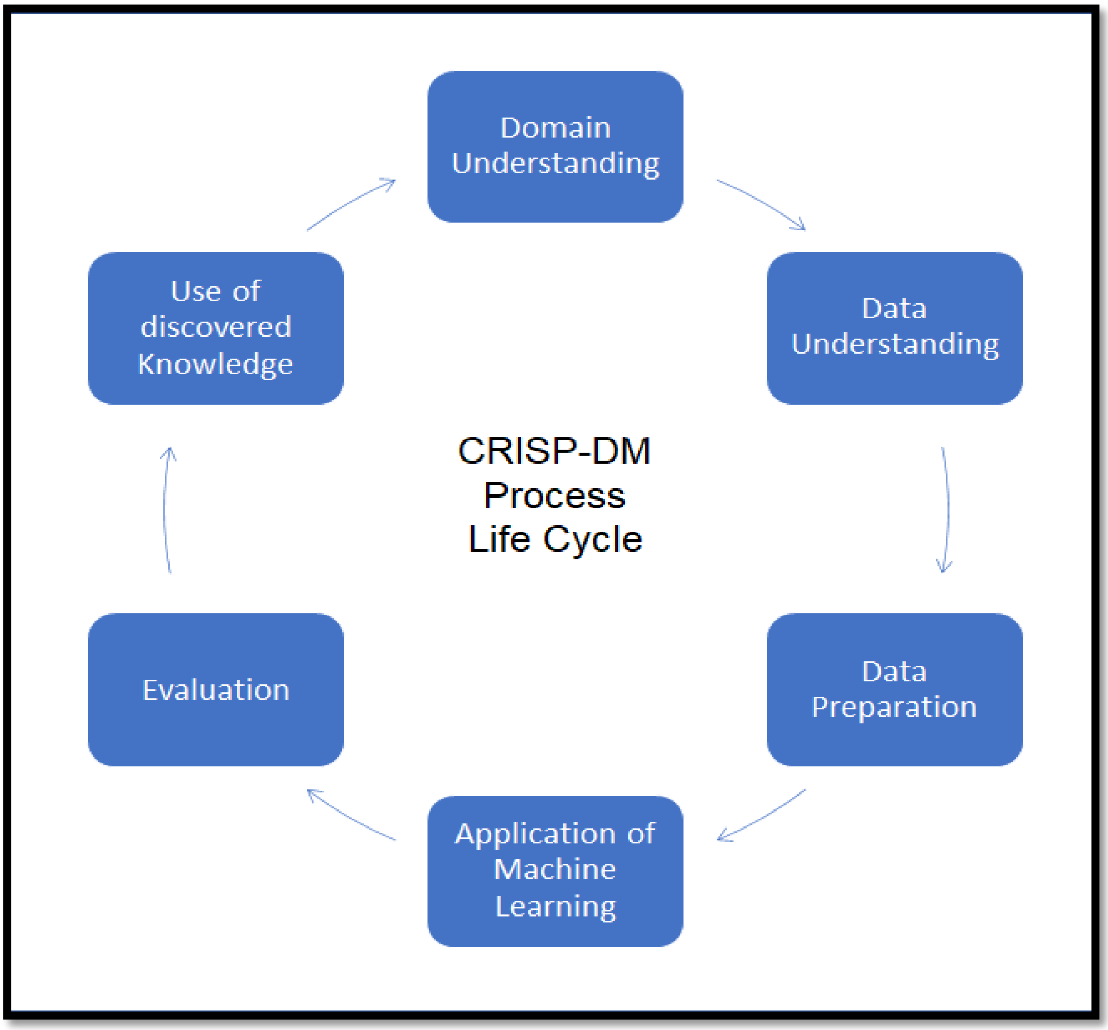

Domain Understanding
Data Understanding
Data Preparation
Application of Machine Learning
Evaluation
Use of discovered Knowledge
CRISP-DM Process Life Cycle

Figure 4.2: Revised CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)

The steps of the model can be summarised as per Figure 4.3 below. There is a domain understanding which revolves clarifying the problem statement and define the aim of the research. Then from the domain understanding, the process of gathering data is undertaken, prepared and made suitable for use by the selected machine learning algorithms, the algorithm is tested and evaluated and how the results can be used.

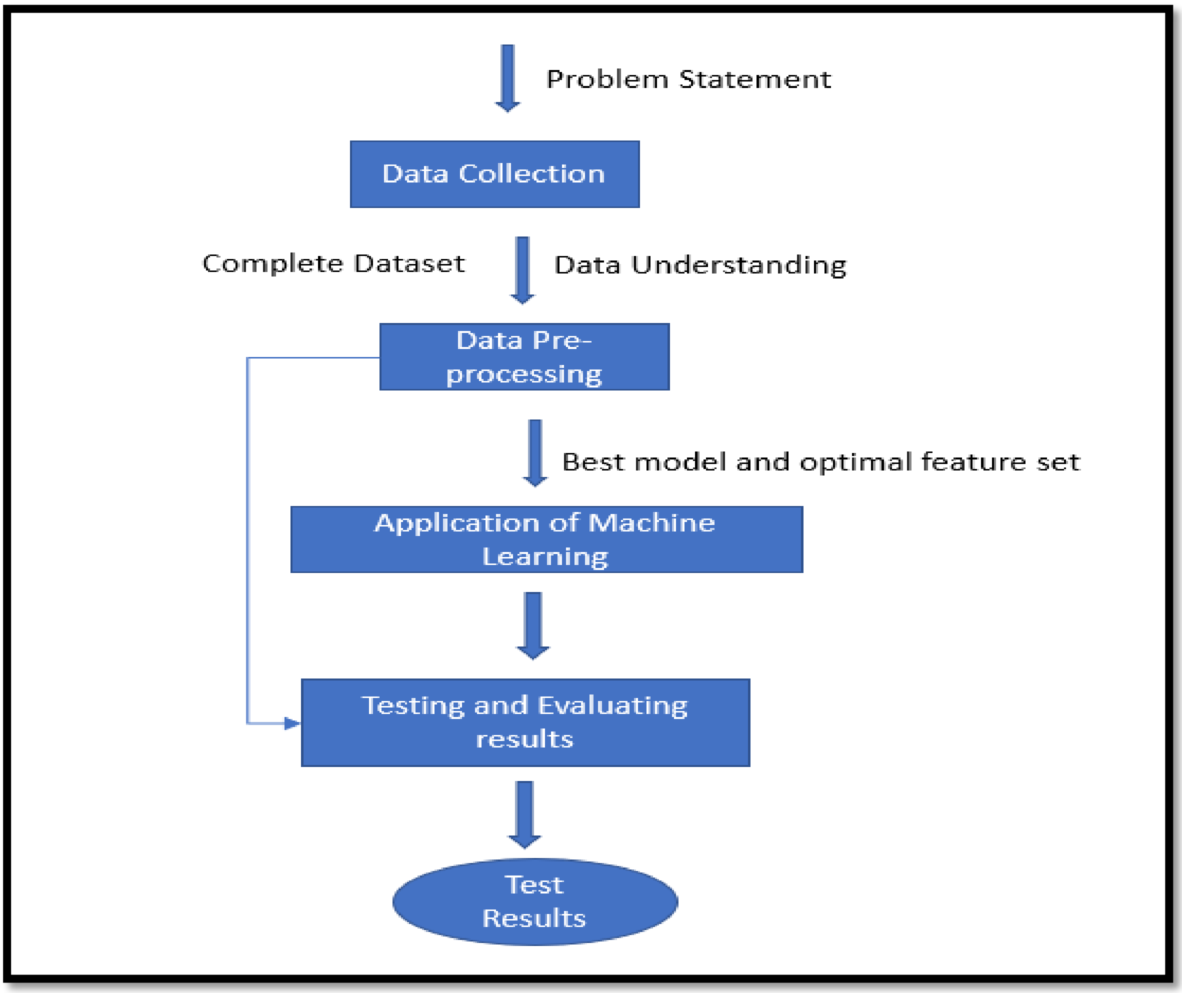


Figure 4.3: The application of machine learning to CRISP-DM and BDA

4.4 DOMAIN UNDERSTANDING

This step involved putting into focus the aim of the research which was to explore the use of big data analytics to address the challenges that institutions of higher learning face in their attempt to improve students' academic performance. This thus entailed developing a regression model that predicts students' final marks based on several inputs. The predicted mark will then determine whether the student is at risk of failing. The relevant stakeholders here were the students and the members of the faculty, and lecturers of the Master of Business Information Systems (MBIS).

4.5 DATA UNDERSTANDING

This step involved getting familiar with the data. In data science related problems, the data rather than the algorithm determines the approach. This means that data determines which algorithm is going to be used. This section explains the processing of gathering data. The data being used was provided by the institution. Students' systems such as LMS harvest data that can be used to improve students' performances and teaching and learning. Hence, analysis and prediction were done in this research with the MBIS subject, of which, the institution provided data. It is important to note that having access to more data would have been ideal for the research. More than data would have been preferred to make scalable predictions by taking into considerations different factors, varying across different faculties. However, a statistical sampling technique (this is explained below) was then used to augment the dataset. The complete data collection process is shown in Figure 4.4 below.

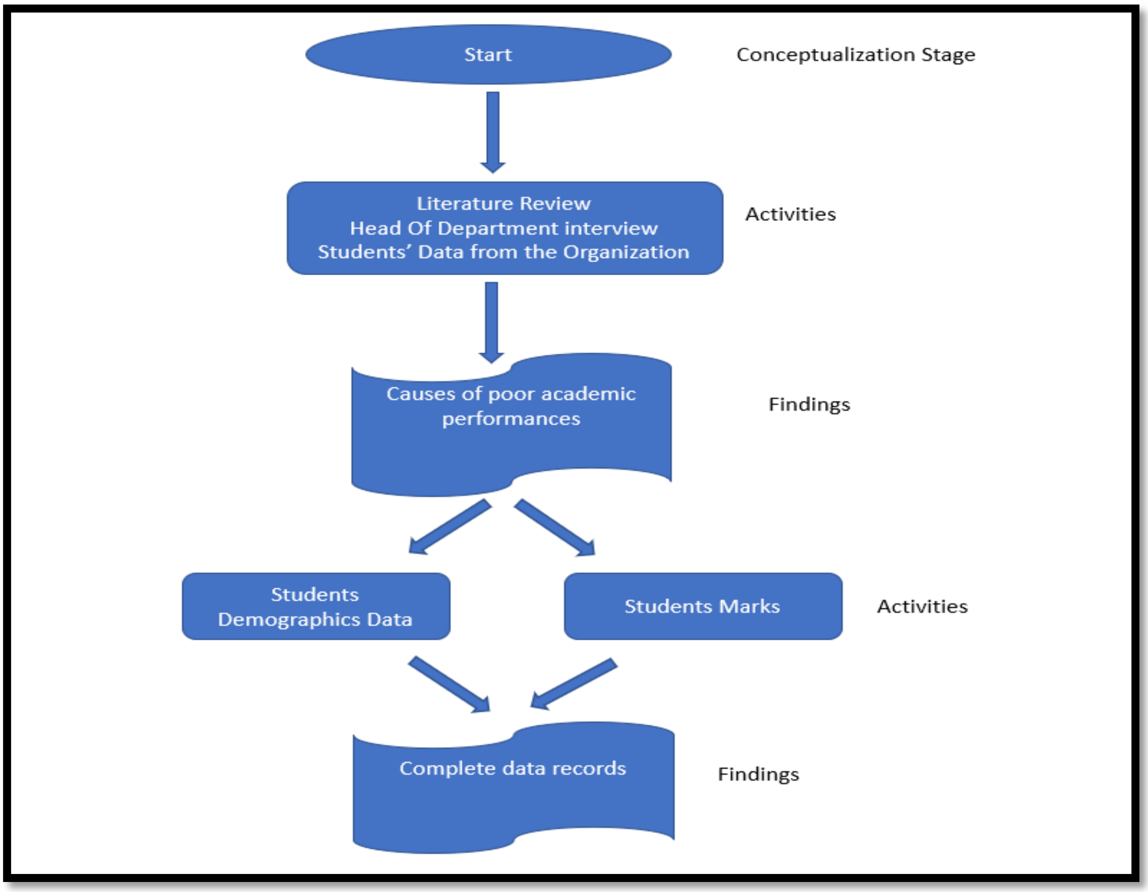
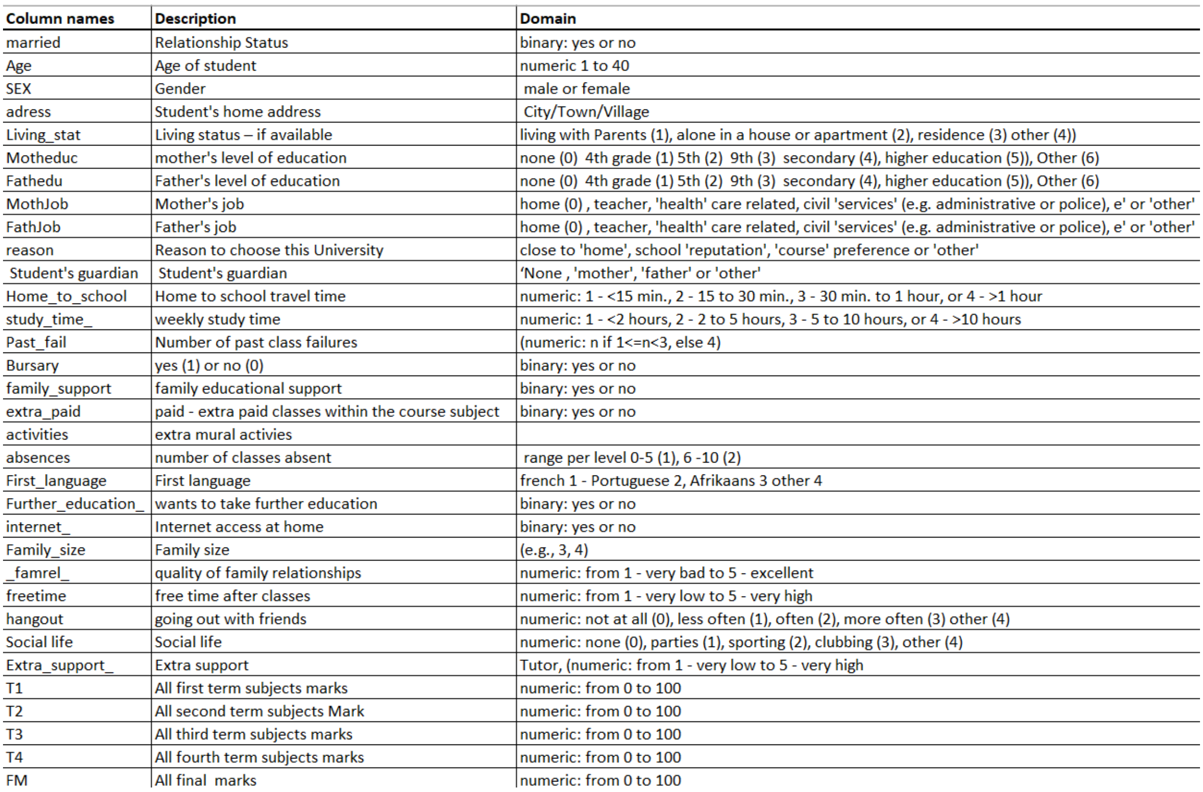

Start
Conceptualization Stage
Literature Review
Head Of Department interview
Students' Data from the Organization
Activities
Causes of poor academic performances
Findings
Students Demographics Data
Students Marks
Activities
Complete data records
Findings

Figure 4.4: Complete data gathering process

The first step of the data gathering process was the conceptualisation process which involved planning how to acquire the data. Then a preliminary literature review was undertaken to help assess the reasons of poor students' academic performance. These along with the interview with the head of department help assessed these causes. This preliminary research helped determined the relevant features that were demanded from the organisation. Students' demographics were cited as possible causes of poor performances. Factors like age, gender, relationship status, distance from university, access to internet were considered.

The dataset thus consisted of rows containing features about the students, the subject, and their grades. These properties ranged from their gender, their level of employment, their family background, demographics, housing. For each student there were binary values stating whether they work or not, and categorical values. Table 4.1 present the initial features (columns) in the dataset. After obtaining the dataset, the data preparation process ensued.

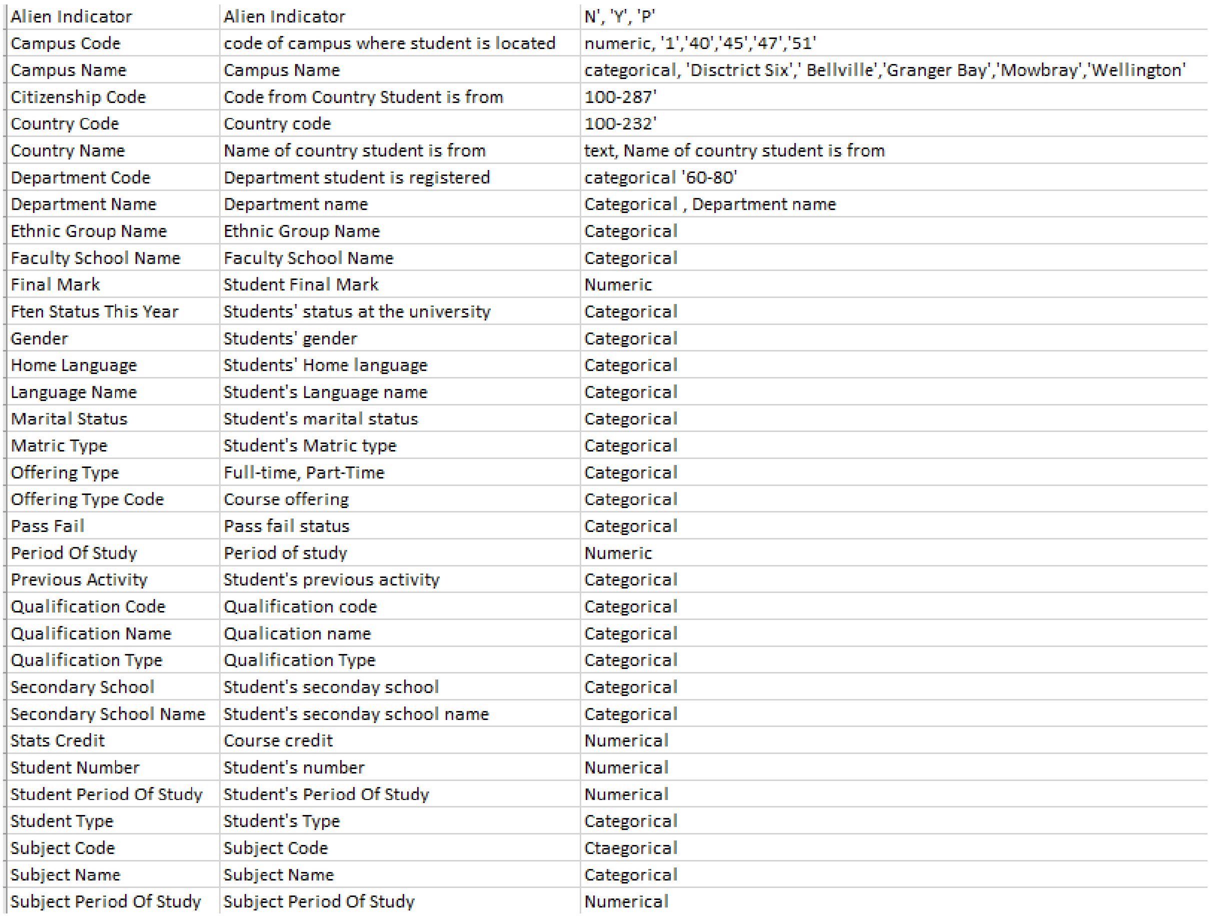
Table 4.1: Initial features of the dataset



Column names	Description	Domain
married	Relationship Status	binary: yes or no
Age	Age of student	numeric 1 to 40
SEX	Gender	male or female
adress	Student's home address	City/Town/Village
Living_stat	Living status – if available	living with Parents (1), alone in a house or apartment (2), residence (3) other (4))
Motheduc	mother's level of education	none (0) 4th grade (1) 5th (2) 9th (3) secondary (4), higher education (5)), Other (6)
Fathedu	Father's level of education	none (0) 4th grade (1) 5th (2) 9th (3) secondary (4), higher education (5)), Other (6)
MothJob	Mother's job	home (0) , teacher, 'health' care related, civil 'services' (e.g. administrative or police), e' or 'other'
FathJob	Father's job	home (0) , teacher, 'health' care related, civil 'services' (e.g. administrative or police), e' or 'other'
reason	Reason to choose this University	close to 'home', school 'reputation', 'course' preference or 'other'
Student's guardian	Student's guardian	‘None , 'mother', 'father' or 'other'
Home_to_school	Home to school travel time	numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
study_time_	weekly study time	numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
Past_fail	Number of past class failures	(numeric: n if 1<=n<3, else 4)
Bursary	yes (1) or no (0)	binary: yes or no
family_support	family educational support	binary: yes or no
extra_paid	paid - extra paid classes within the course subject	binary: yes or no
activities	extra mural activies	
absences	number of classes absent	range per level 0-5 (1), 6 -10 (2)
First_language	First language	french 1 - Portuguese 2, Afrikaans 3 other 4
Further_education_	wants to take further education	binary: yes or no
internet_	Internet access at home	binary: yes or no
Family_size	Family size	(e.g., 3, 4)
famrel	quality of family relationships	numeric: from 1 - very bad to 5 - excellent
freetime	free time after classes	numeric: from 1 - very low to 5 - very high
hangout	going out with friends	numeric: not at all (0), less often (1), often (2), more often (3) other (4)
Social life	Social life	numeric: none (0), parties (1), sporting (2), clubbing (3), other (4)
Extra_support_	Extra support	Tutor, (numeric: from 1 - very low to 5 - very high
T1	All first term subjects marks	numeric: from 0 to 100
T2	All second term subjects Mark	numeric: from 0 to 100
T3	All third term subjects marks	numeric: from 0 to 100
T4	All fourth term subjects marks	numeric: from 0 to 100
FM	All final marks	numeric: from 0 to 100

The secondary dataset was provided by the IHL once they were happy with the results. They provided more data for different cohorts and different years.

Table 4.2: Second dataset



Alien Indicator	Alien Indicator	N', 'Y', 'P'
Campus Code	code of campus where student is located	numeric, '1','40','45','47','51'
Campus Name	Campus Name	categorical, 'Disctrict Six',' Bellville','Granger Bay','Mowbray','Wellington'
Citizenship Code	Code from Country Student is from	100-287'
Country Code	Country code	100-232'
Country Name	Name of country student is from	text, Name of country student is from
Department Code	Department student is registered	categorical '60-80'
Department Name	Department name	Categorical , Department name
Ethnic Group Name	Ethnic Group Name	Categorical
Faculty School Name	Faculty School Name	Categorical
Final Mark	Student Final Mark	Numeric
Ften Status This Year	Students' status at the university	Categorical
Gender	Students' gender	Categorical
Home Language	Students' Home language	Categorical
Language Name	Student's Language name	Categorical
Marital Status	Student's marital status	Categorical
Matric Type	Student's Matric type	Categorical
Offering Type	Full-time, Part-Time	Categorical
Offering Type Code	Course offering	Categorical
Pass Fail	Pass fail status	Categorical
Period Of Study	Period of study	Numeric
Previous Activity	Student's previous activity	Categorical
Qualification Code	Qualification code	Categorical
Qualification Name	Qualication name	Categorical
Qualification Type	Qualification Type	Categorical
Secondary School	Student's seconday school	Categorical
Secondary School Name	Student's seconday school name	Categorical
Stats Credit	Course credit	Numerical
Student Number	Student's number	Numerical
Student Period Of Study	Student's Period Of Study	Numerical
Student Type	Student's Type	Categorical
Subject Code	Subject Code	Ctaegorical
Subject Name	Subject Name	Categorical
Subject Period Of Study	Subject Period Of Study	Numerical

4.6 DATA PREPARATION

Data for the MBIS subjects is released every year. For this research, data from years 2017 to 2019 were combined and cleaned up into one file. Once the data was merged it was then cleaned up again and pre-processed to remove redundancies and inconsistencies. Only after this was done, then the independent variable (the variable to be predicted) was decided. Various pre-processing tasks are discussed below and further expounded into in the next chapter – Chapter 5. The same process was applied to the secondary dataset as well.

4.6.1 Dealing with Missing, Incomplete, or Corrupted data

As mentioned in the previous section, in data science, data takes predominance over the algorithm. So, it is important to have clean data that will provide valuable insight. Not only does the presence of incomplete, missing, and corrupted data leads to wrong results, the data could also be in a format that the algorithm cannot read. For instance, machine learning algorithm do not read texts, so all texts in the datasets had to be transformed into numbers. Also, before doing any analysis on the data inconsistencies such as null values were removed, and some columns were dropped.

4.6.2 Feature normalisation and categorical data conversion:

Data available for this research came in different types. The different columns or features type included text, categories, number. The data available comes in various forms. Features may consist of numbers, dates, text, or as a defined. As algorithms vary in the way they handle data, most algorithms handle continuous numerical values. Therefore, in this research non-numerical data has been transformed several times. The transformation ensure that the data was in a format that the computer could understand. Number columns or features do not require transformation, but categorical and text data did require transformation.

Drawing from table 4.1 above it can be deduced that most data is categorical, i.e., the variables can be categorised and converted. A numerical representation had to be used to encode these categorical variables. *’’Home to school travel time*’’, ‘’*study time”* were example of categorical variables. These categorical variables were standardised in numerical format because they are required by machine learning algorithm to work properly.

4.6.3 Synthetic Minority Over-sampling Technique (SMOTE)

Several algorithms assume that classes are balanced and therefore construct the corresponding error function to maximise an overall accuracy rate. In the case of a non-representative dataset, the result would lead to biased predictions towards the minority class (overfitting). Many methods have been put forward to resolve the problem of class imbalance, including methods such as oversampling, under-sampling (Huang, 2015; Elreedy & Atiya, 2019; Gnip et al., 2021; Wang et al., 2021).

Synthetic Minority Over-sampling Technique (SMOTE) is a sampling technique based upon the idea of under-sampling and oversampling. When a subset of most of the class is chosen randomly with or without replacement substitution until each class has an equivalent number of perceptions. random elimination of some examples of the majority class to reduce their effect on the model. On the other hand, all examples of the minority class are kept. However, under-sampling the majority class may end up excluding important cases that provide information needed to differentiate between the two classes (Ratih et al., 2022; Yi et al., 2022).

Conversely, in oversampling, perceptions from the minority class is picked regardless of substitution until the two classes have an equivalent number of perceptions. generation of additional data (copies, synthetic data) of the minority class to increase their effect on the model. On the other hand, all cases of the majority class are kept, for example Random Over-Sampling, SMOTE or ADASYN. However, oversampling the minority class can lead to overfitting the model, since it will introduce duplicate instances of a set that is already small (Gonzalez-Cuautle et al., 2020; Buraimoh et al., 2021).

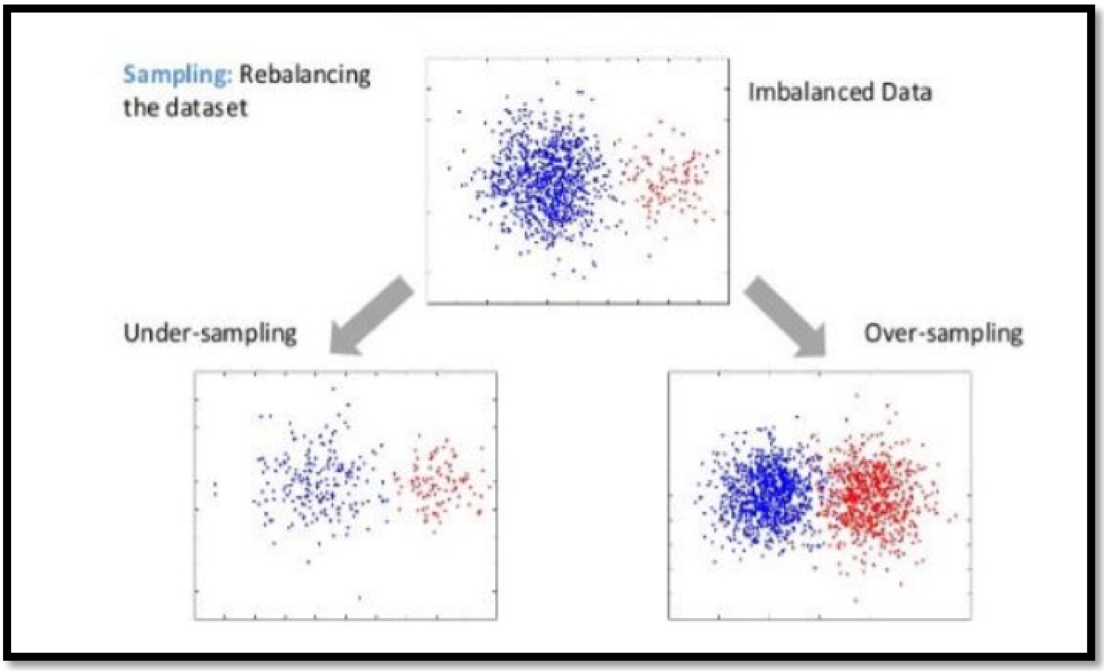


Figure 4.6: SMOTE (Rohit, 2017)

Figure 4.6 is a representation of under-sampling and oversampling. The SMOTE technique uses different ways to deal with imbalanced classes. As an oversampling technique, the SMOTE approach blends new information in the minority class, utilising the k-nearest neighbours' method to find observations that are similar to the existing observations. Instead of adding duplicates of existing observations to the datasets, a synthetic data point is engineered between the current data point and one of its k nearest neighbors. Thusly, this provides a broader decision set as opposed to smaller and dense decision set generated by the duplication in oversampling (Fujiwa et al., 2020; Koivu et al., 2020).

4.7 APPLICATION OF MACHINE LEARNING

Once all the relevant pre-processing has been done, the data was split in testing (portion of the data withheld for model evaluation, which would not have been seen by the model during the training phase) and training set. the whole dataset was divided into two subsets, with 20% of data in the testing, and 80% of observations in the training. to train and test the model Cross-validation was used. Cross validation is defined as *"a model validation technique for assessing* *how the result of a statistical analysis will generalise to an independent dataset"* (Fernandez, 2018; Xu & Goodacre, 2018).

The training set is employed by the machine to learn and discover patterns in the dataset, whereas the testing set is an independent dataset that assesses the performance of the generated model. the testing set also helps to acquire the performance characteristics such as accuracy (Xu & Goodacre, 2018). It is not ideal to test a model on the training set, because they would always return highest level of accuracy as the data was trained on.

The training phase include model selection and parameter tuning. Model selection is defined as both selecting a model and tuning and tweaking any parameters. Once the model was selected and it was trained. Training a model was conducted on a labelled dataset. Observations the training datasets comprised a set of outputs and inputs features. The dependent variable, which is the variable to be predicted, or the response variable is the output. The input features which are the predictor or explanatory variables are the independent variable. After training the model it was then tested on the test data set. Running a selected model on the training dataset and testing its performance on a test dataset is called cross validation. The initial results of the model training were poor so SMOTE was used to improve the results.

4.8 EVALUATION OF PREDICTION MODEL

Generalising beyond the data of the training set is the basic reason for prediction or forecasting models (Kappen et al., 2018; Westphal & Brannath, 2020). Data in the testing set for instance, will seldom be the same as the data in the training set. To assess the forecasting models some metrics are applied. Also, the coefficients acquired from the multiple regression model are assessed for the sign and the size. Higher value coefficients represent greater relevance to the model and vice versa (Silhavy et al., 2017; Emmert-Streib & Dehmer, 2019; Kalappan et al., 2021; Khan et al., 2022).

The overall error of the fit can be measured by calculating the Residual Sum of Squares:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

In addition to RSS, there are some other metrics that can be used:

Mean Squared Error

$$MSE = \frac{RSS}{n}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

R squared:

$$R^2 = 1 - \frac{RSS}{TSS}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

$R^2$is the coefficient of determination in this case.

Finally, Sklearn (the machine learning library in python) also has implementations of common error metrics to assess the fit of the model.

4.8 KNOWLEDGE DISCOVERY

This step involves designing a platform in which the analytics can be embedded. And show a prototype of sort of what the solution would look like. In this research this was also accomplished to give an idea of how students can be used to make the pridections.

The table 4.3 examines various elements and validates how the research approach aligns with standard data mining best practices. Key phases like domain understanding, data preparation, model building, and evaluation are highlighted. The choice to use the CRISP-DM framework provides a structured, validated process for the data mining effort to predict student performance. Careful attention is placed on data integrity and preprocessing before applying machine learning algorithms. Techniques like SMOTE and cross-validation demonstrate adherence to standards that reduce overfitting and improve generalizability.

Evaluation metrics quantify model accuracy while error and residual analysis uncover avenues for refinement. The prototyped solution brings the analytical discoveries into practical use via an integrated dashboard. Overall, the rigorous methodology produces actionable insights that can enable personalized interventions to help at-risk students.

Table 4.3 Comparison table based on proposed findings.

Elements	Description	Validation	Comparison Details
Domain Understanding	Defining research aims to predict student performance	Aligns with literature review and methodology	Goals focused on using big data analytics to improve student academic outcomes
Data Understanding	Obtaining student dataset from university	Data aligns with required features	Dataset spanned 3 years and contained academic, demographic, and social factors
Missing Data	Handling missing data in dataset	Ensures data integrity for analysis	Missing values were identified and removed prior to analysis
Feature Encoding	Converting categorical data to numerical	Required for machine learning algorithms	One-hot encoding used for categorical variables like gender, housing type
Class Imbalance	Using SMOTE to balance minority class	Improves model performance on imbalanced data	SMOTE helped address class imbalance between pass/fail grades

Data Splitting	80/20 split for training and testing	Standard practice for validation	Out-of-sample testing evaluates real-world effectiveness
Model Training	Training regression model on data	Key phase to discover patterns	Multiple regression models trained using sklearn packages
Model Selection	Choosing right ML algorithm	Regression suitable for numerical prediction	Linear regression and random forests evaluated for performance
Cross Validation	Testing model on unseen data	Evaluates real-world performance	5-fold stratified cross validation used
Model Evaluation	Metrics like RMSE, R2	Quantifies model accuracy	Additional metrics included Mean Absolute Error
Error Analysis	Identifying model weaknesses	Allows refinement and improvement	Analysed errors to improve feature selection and modelling
Knowledge Discovery	Documenting discoveries	Transfers insights to stakeholders	Discoveries allow personalized interventions for at-risk students

Overall Approach	Aligns with CRISP-DM framework	Provides validated standard process	Process aligns with best practices for data mining

4.9 SUMMARY

This chapters discusses the phases of the Cross-Industry Standard Process for Data Mining (CRISP-DM) and how they were used in this study to attempt to predict students' performance. The phases used in the research included : domain understanding; data understanding; data preparation; application of machine learning, evaluation, and the use of discovered knowledge (Kurgan and Musilek, 2006).

CHAPTER 5: MACHINE LANGUAGE TECHNIQUE AND PREDICTABILITY

5.1 INTRODUCTION

This research attempted to utilise students' data to predicts students' performance. This is to derive from these predictions a way to automate and improve students' performance by identifying students at risk. Thus, the research sought to build a model to perform the predictions. This was done by following the revised framework of CRISP-DM (Shearer, 2000).

The previous chapter discussed the steps undertaken in the application of the machine learning techniques. This chapters discusses the data understanding; data preparation; application of machine learning, evaluation of the machine learning algorithm.

5.2 OVERVIEW ON CLASSIFICATION OF MACHINE LEARNING ALGORITHMS

To perform sophisticated tasks, it is necessary for computers to make use of Machine learning, without any manual intervention (Nichols et al., 2019; Schmidt et al., 2019; Jovel & Greiner, 2021). Thus, a computer will be able to perform tasks performed by human beings such as voice recognition, cooking, driving (Kersting, 2018; Janiesch et al., 2021). More importantly, the computer will be able to perform tasks that human being is unable to perform (Pugliese et al., 2021; Sarker, 2021). This includes the analysis of complex and large datasets, weather forecasting.

Essentially, data science problems can be solved by Machine learning. Machine is defined as:

> *a concept to unify statistics, data analysis, machine learning and their associated methods to understand and analyse real phenomena" with data. (Cao, 2017)*

The undertaking of solving any data science problem cannot happen without properly categorising the problem. This will help ensure that the most appropriate algorithm is assigned to the problem (Sarker et al., 2021). The different machine learning categories can be found in the Table 5.1 below.

Table 5.1: Classification of Machine Learning Algorithms (Alzubi et al., 2018)

Classification Problem	**Is this A or B**
Anomaly Detection Problem	Find odd one out
Regression Problem	How much or how many?
Clustering Problem	• What is the structure behind the problem? • How is it organised?
Reinforcement Learning Problem	What should one do next?

The different categories as shown in Table 5.1 can be resolved by using the algorithm explained as follow:

a) Classification algorithm: in this case the output will classify in terms of categories. An example of this would be a problem trying to "classify" whether the phrases being written are either, English, French, Spanish. A classification algorithm is used to predict a categorical value, such as whether a student will fail or pass a course (Alzubi et al, 2018, Sarker, 2021). Classification algorithms are commonly used in education to predict students' grades or other measures of performance, based on data such as past grades, attendance, and engagement (Sankara, 2017; Tan, 2021). There are several other contexts in which a classification algorithm might be appropriate for predicting students' performance:

- Identifying at-risk students: A classification algorithm could be used to predict which students are at risk of failing a course, based on data such as past grades, attendance, and engagement. This could be used to identify at-risk students who may need additional support (Rawat et al., 2021; Yayla et al., 2021).

- Evaluating the effectiveness of teaching and learning practices: A classification algorithm could be used to predict which students are likely to pass or fail a course and compare the performance of students within different teaching and learning contexts. This could be used to evaluate the effectiveness of different teaching and learning practices and identify areas for improvement (Mounika & Persis, 2019; Li, 2022).

- Predicting students' future performance: A classification algorithm could be used to predict which students are likely to achieve certain academic milestones, such as graduating or obtaining a certain grade point average. This could be used to inform academic advising and support services and help students to achieve their academic goals (Khor, 2019; Rajak et al., 2020; Maura et al., 2021).

Classification algorithms are a useful tool for predicting students' performance in education, as they can be used to identify at-risk students, evaluate the effectiveness of teaching and learning practices, and predict students' future performance.

b) Anomaly Detection algorithm: Machine learning algorithms are also being used in fraud detection by credit companies (Thudumu et al., 2020; Huang et al., 2022). This is done by using anomaly detection, which refers to detecting observations, data points that deviate from a dataset's normal behaviour (Churova et al., 2021). Anomaly detection algorithms are used to identify unusual or unexpected patterns in data that may indicate a problem or issue. These algorithms are often used in a variety of contexts, such as cybersecurity, fraud detection, and quality control, to identify deviations from expected patterns that may indicate a problem or risk (Glaser et al., 2022; Yoshihara & Takahashi, 2022).

In the context of education, anomaly detection algorithms could be used to predict students' performance by identifying unusual patterns in student data that may indicate a problem or issue. For example, an anomaly detection algorithm could be used to identify students who are performing significantly below or above expectations, or who are experiencing a sudden decline in performance, to provide targeted support or interventions (Howlin et al., 2019; Lauria, 2021; Xie et al., 2022).

Anomaly detection algorithms may be appropriate for predicting students' performance in education in situations where it is desirable to identify unusual or unexpected patterns that may indicate a problem or issue (Salami et al., 2016; Guo et al., 2022). However, it is important to consider the limitations of anomaly detection algorithms, as they may not be effective at predicting students' performance in all situations. For example, they may be less effective at predicting students' performance in cases where there is a lack of data or where the data is highly variable. In such cases, other machine learning algorithms, such as regression or classification algorithms, may be more appropriate for predicting students' performance.

c) Regression algorithm: Regression is used to predict output based on a set of inputs. The output being predicted is usually a number. Since the objective is to predict student performance, specifically predicting student final mark to determine whether they will pass, this is the most appropriate category for this research (El Guabassi et al., 2021; Sarker, 2021). A regression algorithm is used to predict a continuous numerical value, such as a student's performance. Regression algorithms are commonly used in education to predict students' grades or other measures of performance, based on data such as past grades, attendance, and engagement. There are several contexts in which a regression algorithm might be appropriate for predicting students' performance (Strecht et al., 2015; Bag, 2020; Altamini et al., 2022):

- Predicting students' grades: A regression algorithm could be used to predict students' grades in a particular course or subject, based on data such as past grades, attendance, and engagement. This could be used to inform academic advising and support services and help students to achieve their academic goals.

- Evaluating the effectiveness of teaching and learning practices: A regression algorithm could be used to predict students' grades and compare the performance of students within different teaching and learning contexts. This could be used to evaluate the effectiveness of different teaching and learning practices and identify areas for improvement.

- Predicting students' future performance: A regression algorithm could be used to predict students' grades over time and identify trends and patterns in their performance. This could be used to inform academic advising and support services and help students to achieve their academic goals.

Regression algorithms are a useful tool for predicting students' performance in education, as they can be used to predict students' grades, evaluate the effectiveness of teaching and learning practices, and identify trends and patterns in students' performance over time.

d) Clustering algorithm: Clustering is part of unsupervised learning and aims to find and learn natural grouping. Clusters are then created based on the similarity of the structures within data (Xu & Tian, 2015; Rodriguez et al., 2019; Zhang, 2022). Clustering algorithms are commonly used in education to group students based on their performance or other characteristics, such as learning styles or personal interests. There are several contexts in which a clustering algorithm might be appropriate for predicting students' performance (Dutt et al., 2015; Nagesh & Satyamurty, 2018; Khayi & Rus, 2019; Mohamed et al., 2022):

- Grouping students for personalised learning: A clustering algorithm could be used to group students based on their performance, learning preferences, or other characteristics, to create personalised learning experiences for each group.

- Identifying at-risk students: A clustering algorithm could be used to group students based on their performance, attendance, or engagement, to identify at-risk students who may need additional support.

- Evaluating the effectiveness of teaching and learning practices: A clustering algorithm could be used to group students based on their performance and compare the performance of students within each group. This could be used to evaluate the effectiveness of different teaching and learning practices and identify areas for improvement.

Overall, clustering algorithms are a useful tool for predicting students' performance in education, as they can be used to group students based on their performance or other characteristics, and inform personalised learning experiences, identify at-risk students, and evaluate the effectiveness of teaching and learning practices.

e) Reinforcement algorithm: when a machine makes decision based on previous learning experiences it is classified as reinforcement algorithm. This consist of punishing undesired behaviours, while rewarding desired behaviours (Rolf et al., 2022; Singh et al., 2022). Through trial and error, the reinforcement agent is thus able to understand and interpret its environment. In Reinforcement learning an agent learns through receiving penalties or rewards for a course of actions and interacting with its environment and (Wells & Bednarz, 2021; Petrova-Dimitrova, 2022). It is commonly used in artificial intelligence and robotics to enable agents to learn and make decisions based on their experiences.

In the context of education, reinforcement learning could be used to predict students' performance by rewarding or punishing students based on their actions or outcomes (Zimmer et al., 2014; Khattak & Ahmad, 2018; Ausin, 2019). For example, a reinforcement learning algorithm could be used to predict which students are likely to achieve certain academic milestones, such as graduating or obtaining a certain grade point average, by rewarding students who achieve those milestones and punishing those who do not.

Reinforcement learning algorithms may be appropriate for predicting students' performance in education in situations where it is desirable to use rewards or penalties to shape student behaviour and outcomes. However, it is necessary to ponder the ethical implications of using reinforcement learning algorithms in education, as they may raise issues related to fairness and equity.

5.3 THE SELECTED AND APPROPRIATE MACHINE LANGUAGE ALGORITHM

It is not possible to determine which machine learning algorithm is *"more appropriate"* for predicting students' performance, as the appropriate algorithm will depend on the specific context and goals of the prediction task (Sarker, 2021).

ML algorithms differs in weaknesses and strengths. The algorithms are suitable for various prediction tasks. For example, a regression algorithm may be more appropriate for predicting a continuous numerical value, such as a student's grade, while a classification algorithm may be more appropriate for predicting a categorical value, such as whether a student will pass or fail a course (Sarker, 2021).

To determine the most appropriate machine learning algorithm for predicting students' performance, it is important to consider the type of prediction task, the available data, and the desired outcomes. It may also be helpful to experiment with different algorithms and compare their performance to identify the most effective approach (Strecht et al., 2015; Bag, 2020; Altamini et al., 2022).

As established in the previous section this research attempts to predict student performances, by predicting the final mark. By predicting the final mark, it would be easier to predict which student is at risk of failing. It was determined that regression is the most appropriate classification of the problems, and thus the most appropriate for the dataset at hand. However, in addition to regression, other algorithms deemed appropriate for regression problems such as support vector machines (SVM), random forest, decisions trees, were also used.

5.3.1 Regression Algorithm

Regression algorithms are used for prediction and forecasting, while also being used to determine causal relationships (Strecht et al., 2015; Bag, 2020; Altamini et al., 2022). The linear regression algorithm is considered supervised learning. With a regression algorithm the variable to be explained also called target variable (Y), the model attempts to predict using the predictive variables or explanatory variable (X). The aim is to obtain a cost or prediction function explaining the relationship between X and Y; the values of Y can be predicted from known values of X. A good example will be predicting student based on hour studies. This also establishes a causal relationship. It is also important to note that a regression algorithm is a ML model where the X variable can be qualitative or quantitative with the target variable (Y) being quantitative.

Regression is done using the equation of a straight line. A straight line can be described using the following equation:

Y= a+ Bx ……………………………………………….(1)

where a is the intercept of the line with the y-axis, and b is the gradient.

Using the equation above, it is possible to predict a Y value, for any given value of X. Thus, the relationships between the target or dependent variable Y and the set of explanatory or independent variables X are analysed through a regression model. Through an equation that predicts the values of the target variable as a linear combination of parameters, this relationship is established. There are two forms of regression algorithms. These are the multiple linear regression and simple linear regression (Jobson, 1991; Denis, 2018; Trunfio et al., 2022).

Table 5.2: Multiple regression models

Types of Linear Regression	**Number of Variables**
Simple Linear Regression	One explanatory or predictor variable
Multiple Linear Regression	Multiple predictor or explanatory variable

a) Simple linear regression model:

As shown in Table 2, a simple linear regression model only uses a single explanatory variable to help predict the target variable (Jobson, 1991; Denis, 2018; Trunfio et al., 2022).

Y= a+ bX ……………………………………………….(1)

With:

- a and b, the coefficients (y-intercept and slope) to be calculated
- Y, dependent random, the target variable
- X, the independent or explanatory variable,

b) multiple linear regression model:

A multiple regression model uses multiple explanatory or predictor variables (Jobson, 1991; Denis, 2018; Trunfio et al., 2022).

$$y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + a \text{ .... (2)}$$

where:

- y is the response variable which depends on the p predictor variables.

- a is the intercept, interpreted as the Y value as all predictor variables are equal to zero.

- b_j is the average effect on y of a one unit increase in x_j, assuming all other predictors remain fixed.

- As the research is making use of multiple predictors, this research will be focused on multiple linear regression model.

5.3.2 Decision Tree

As the research is dealing with a regression problem the most appropriate algorithm will thus be a regression algorithm. However, for regression problems other algorithms have been used as well: This includes Decision Tree. A decision tree is applied with the intent to classify future observations given a corpus of already labelled observations. Decision Trees partition the dataset into several sections on the basis of questions asked of the predictive variables. They can be described as upside down tree-like ML models. (Rokach & Maimon, 2005; Song & Ying, 2015; Venkatasubramaniam et al., 2017; Mienye et al., 2019).

To decide on a new input the model:

1. Starts at the root node, which is at the top of the tree.
2. Asks questions at each decision node about the attributes of the input.
3. Repeat 2. Until reaching a terminal node (also known as a leaf node) at the bottom of the upside-down tree.

Each terminal node in the tree contains a potential output for a given input. The value at a terminal node only becomes the output of our decision tree if the path of decisions on an input XX from the root node led to that particular terminal node.

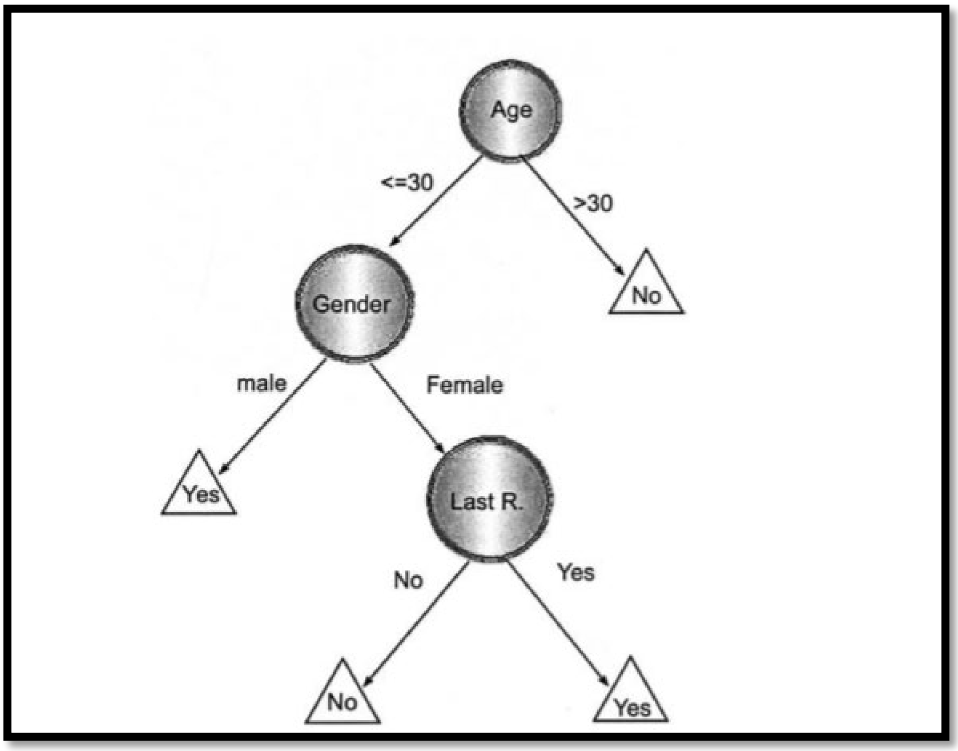


Figure 5.1: Example of Decision Trees (Rokach and Maimon, 2005)

Overfitting in Decision Trees: One of the attributes of decision trees is overfitting. Overfitting allows fitting details of the individual data points instead of the overall properties of the distributions they are drawn from. With overfitting it is very easy to go too deep in the tree. This issue can be addressed by using Random Forests (Al-Akhras et al., 2021).

5.3.3 Random Forest Regressor Model

A random forest is a powerful non-parametric algorithm built on decision trees and a model of an ensemble method. Random forest relies on aggregating the outputs of an ensemble of decision trees. the output is a mean prediction of the individual trees which are randomised. How do Random Forests Work? (Sarika et al., 2017)

1. Fitting data:

Taking note of the fact that N refers to the number of observations (rows) in the training dataset, and p the number of predictor variables (columns). The following is the typical algorithm for a Random Forest:

Bootstrapping: Drawing with replacement from the training dataset, randomly sample N observations. In this research an 85 to 15 percent split was used. 85 percent were used for training and 15 percent were used for testing. The training dataset constituted the N observations (Lee et al., 2020).

Use the N observations to grow a random forest tree as follows:

At each node:

(i) Select a random subset, m, of predictor variables, where m<p–√.

(ii) Pick the best variable/split-point among the selected predictor variables.

(iii) Divide data into two subsets based on the selected split.

(iv) Repeat until stopping criteria satisfied (e.g. minimum node sample size reached).

a) Repeat until desired number of random forest trees is reached.

Since the draw is done randomly, with replacement, from the training data during the bootstrapping step it is possible that:

b) Some data samples get resampled and thus reused when fitting different trees in the random forest.

c) Some data samples don't get sampled at all and thus do not get used in fitting the random forest.

This means that the dataset for which each tree is grown on, is slightly different, so random forests are less likely to overfit than decision trees.

2. Making Predictions:

To make a prediction, random forests combine multiple trees (Kern et al., 2019; Lee et al., 2020). with ensemble methods, the sum can be greater than the parts. This implies individual estimators doing the voting is not preferred to a dstvmajority vote among numerous estimators (with estimator representing trees). Scikit-Learn, a machine learning library, was used for the process of fitting a decision tree to the data. The RandomForestRegressor estimator was used. This was done by specifying the hyperparameters, some of the more important ones include:

(i). n_estimators: The total of trees to include in forest; 200 was entered.

(ii). min_samples_leaf: this represents the minimum samples needed at a leaf node. The specified minimum number of samples leaf was 1.

(iii). max_depth: The maximum depth of each forest tree (i.e., the total number of nodes between root and leaf node).

(iv). random_state: A number used to seed the random number generator. This ensures that the same tree is obtained each time the model.fit() is called - this particular hyperparameter is important in random forests since their training procedure is inherently random.

5.3.4 Support vector machine

Although Support vector machine (SVM) were initially used for classification problems, their used has now been extended to regression problems as well. The aim of a SVMs is to define an N-dimensional space hyperplane with N representing the number of features. The different data points will thus be distinctly classified by this hyperplane (Shmilovici, 2009; Zhang, 2012).

A Dataplane represents a decision boundary that will assist in classifying data points. Classes will be attributed to data points lying on each side of the hyperplane (Pradhan, 2012). to classify the data, It only needs a simple one-dimensional decision boundary (which is basically a line) when the dataset only has 2 features. However, the more features are added, the line needs to take on more dimensions, as a simple line will not suffice. A line only has one dimension. A flat shape has 2 dimensions. A 3D shape has 3. A shape for 4 or more dimensions is a hyperplane. In SVM, the hyperplane will always have one less dimension (−1) than the number of input features (p), or a total of (p−1) dimensions.

Using a (p−1) dimensional hyperplane, SVM aim to separate points. This means that the SVM will construct a decision boundary which will impute a label to points on the left and on the right side. When finding the separating hyperplane, the aim is to maximise the distance of the nearest points to the hyperplane. The data points which dictate where the separating hyperplane goes are called support vectors.

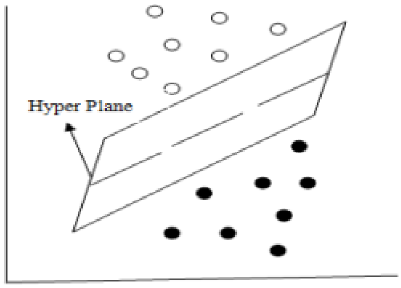


Figure 5.2: Hyperplane (Pradhan, 2012)

5.3.5 XGBRegressor

The gradient boosting technique is utilized in general and in this research to reinforce a model such as a decision tree that yields weak predictions. Gradient boosting may be applied to classification or regression or predictive modelling problems using a class of ensemble machine learning algorithms (Kaliappan et al., 2021). To adjust the prediction errors ensuing from prior models, trees are fit and added one at a time to the ensemble. Ensembles are constructed from decision tree models. boosting refers to a type of ensemble machine learning model.

5.4 PROBLEM DESCRIPTION AND ITS CONCEPTUALISATION

One of the main goals of Institutions of higher learning in South Africa is to improve students' performances through improved teaching and learning (Nyamupangedengu, 2017; Shay, 2017). However, inadequate mining or analytics of Big Data had led to insufficient information that can assist in providing teaching and learning structure that will empower and enhance students' full potential (Reyes, 2015). Many of South African institutions make use of various approaches and methods in an attempt to achieve this goal. Some of the approaches include the use of software and analysis of data (Jordaan & Van der Merwe, 2015; Lemmens & Henn, 2016; Lourens & Bleazard, 2016; Gašević, 2018; Prinsloo, 2018).

Despite these methods and approaches including the use of software, challenges persist in the attempts to improve students' performances. The challenges persist either because the software is incorrectly applied relative to the available data, or the selected software is not suitable, thereby giving results that do not help in improving students' performances (Lemmens & Henn, 2016; Prinsloo, 2018).

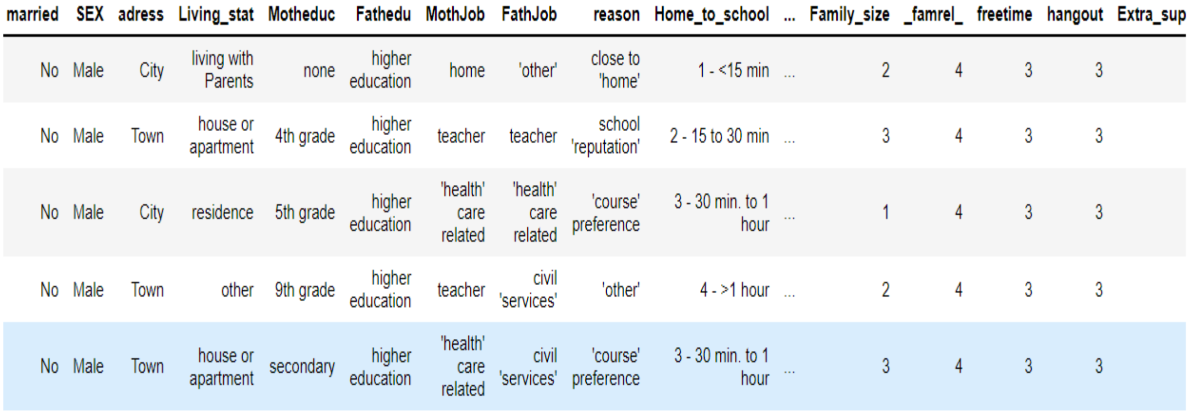
These challenges cause problems such as the inconsistent approach in using existing data as they evolve by many of these institutions and the consideration of only the results and not the factors that influence the results. Thus, it is critical to employ an approach through a framework, that encompasses factors of influence, enable forecast, and to ensure consistency, towards performance improvement.

5.5 THE APPLICATION OF ML ALGORITHM TO THE PROBLEM

5.5.1 Data Structure for modelling of final mark (FM):

There are 37 features in the dataset. Here are the columns that exist in the Dataset.

Table 5.3: Overview of the dataset on python



married	SEX	adress	Living_stat	Motheduc	Fathedu	MothJob	FathJob	reason	Home_to_school	...	Family_size	_famrel_	freetime	hangout	Extra_sup
No	Male	City	living with Parents	none	higher education	home	'other'	close to 'home'	1 - <15 min	...	2	4	3	3	
No	Male	Town	house or apartment	4th grade	higher education	teacher	teacher	school 'reputation'	2 - 15 to 30 min	...	3	4	3	3	
No	Male	City	residence	5th grade	higher education	'health' care related	'health' care related	'course' preference	3 - 30 min. to 1 hour	...	1	4	3	3	
No	Male	Town	other	9th grade	higher education	teacher	civil 'services'	'other'	4 - >1 hour	...	2	4	3	3	
No	Male	Town	house or apartment	secondary	higher education	'health' care related	civil 'services'	'course' preference	3 - 30 min. to 1 hour	...	3	4	3	3	

The data has both quantitative and qualitative data. As machine learning does not understand text, this means that this text will eventually be transformed into numbers before it is fed to an algorithm.

The following provide the meaning of some of the feature:

- Married: If a student is married or single.
- Home to school travel time: The amount of time taken by a student to reach home.
- Student's home address: If a student lives in a city or town.
- Living status: If a student lives with his/her parents or alone.
- T1: The grade of the student in the first subject.
- T2: The grade of the student in the second subject.
- T3: The grade of the student in the third subject.
- T4: The grade of the student in the fourth subject.

The target:

Final Mark (FM): This is the target label, which is the Final grade of the student that the algorithm is attempting to predict. Essentially the algorithm is fed data about students and predict the final mark based on the data entered. This will help lecturer to predict and identify students at risk of failing.

5.5.2 Data pre-processing & Cleaning of the data structure

(a) Unique values in each column validation

The objective here is to figure out how many unique values in each column and this analysis provides many benefits:

- to differentiate between Continuous and discrete columns that exist in the dataset. Examples of Discrete columns are ‘’Married’’, ‘’Further_education’’, ‘’Sex’’, ‘’Study_time’’. Examples of Continuous columns are ‘’T1’’ ,’’ T2’’ , ‘’T3’’, ‘’T4’’ , ‘’FM’’.

- to drop the columns that have one value only because these types of columns do not improve model building. Removing any such columns will thus reduce the number of features (columns). With the model showing that eight columns containing unique values were dropped: (“Student’s guardian”, ‘’Married’’, ‘’Qual’’,’’year’’).

b) Null values in each column validation

After removing the columns containing unique values, missing values in each column must be investigated. Here are some benefits of figuring out how many null values exist in each column.

- Any machine learning model will not accept null values (nan values), it will only accept numbers. Hence, if any column has a nan value, it will raise an error.

- Also, if a column has too many nan values it is best practice to delete that column because no accurate prediction can be made to replace these nan values with accuracy.

- On the other hand, if a column has a few nan values, these values can be replaced with some estimated number like the (mean, median, mode) of that column. All columns did not have nan values except the FM (Final mark) column which is the column to be predicted. The missing values inside the

FM column were replaced with the mode, the mode occurs most of the time compared to other values.

c) Checking the data type of each column:

Checking the datatype of each column will help with Identifying the non-numeric columns and changing them into numeric ones to be fed into the model. Again, machine learning models only accepts numerical values, hence, non-numerical values must be transformed.

It appears that most columns have an object (string) datatype; also, some columns appear to be numeric but are represented as an object. This will raise an error when applying some mathematical operations on those columns. these columns require transformation into a form that the model can accept. Furthermore, there are some discreet columns like “Gender” which have two unique values [“Female”, “Male”]. These values need to be changed into numbers, and the “get_dummies” function in python was used to handle this. After executing the code, object columns no longer exist.

Also, there were some punctuation marks embedded with the column’s names. Hence, these columns names were renamed into a more readable format. After doing the above cleaning to the data the final shape show that all the columns have numbers in them.

d) Exploratory data analysis (EDA)

In this step, visualisation was used on some columns to see their distribution to gain some insights about the data.

e) The distribution of our target column (FM)

In the Figure below, after plotting the histogram of (FM) column, the majority of the students’ scores are greater than 40 and less than 80.

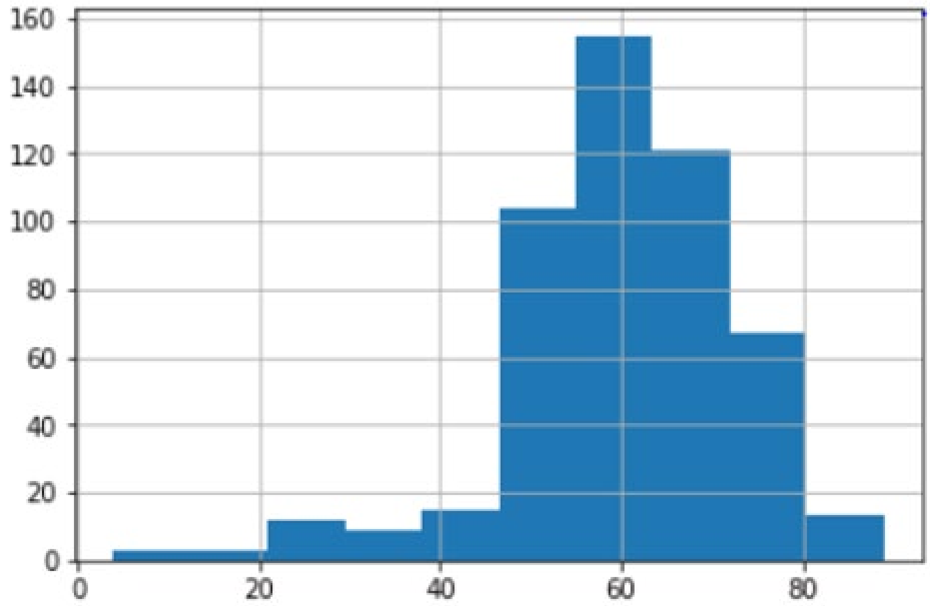


Figure 5.3: Histogram of the Final Mark

Are married students greater in number than single ones? It appears that most of the students are not married from the figure below.

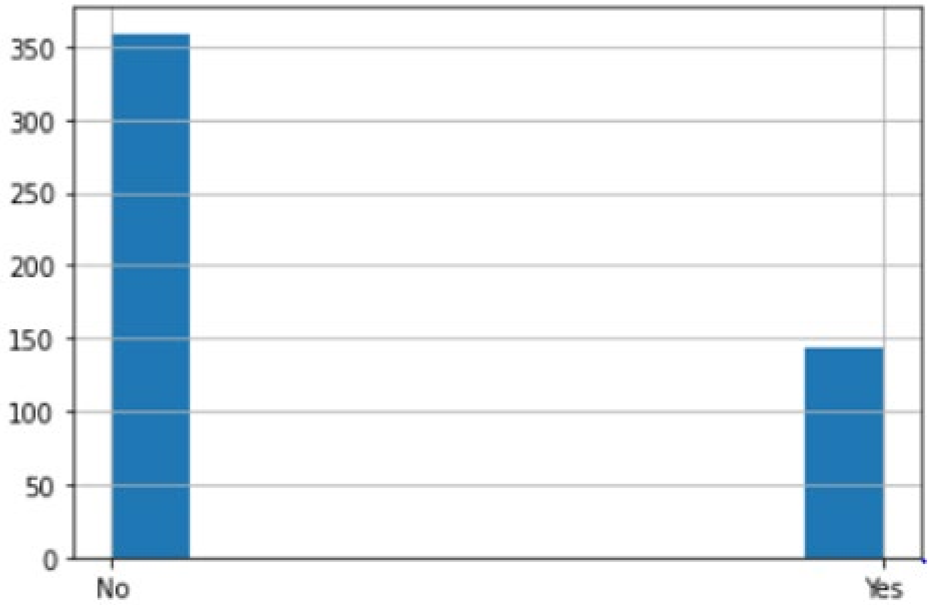


Figure 5. 4: Figure 5.6 Married Vs Single

From the figures below, student's scores lie between 50 to 75. The average score of married students is 60.69. The max score obtained by a married student is 88.0. The average score of Single students is 59.738. The max score obtained by a Single student is 89.0.

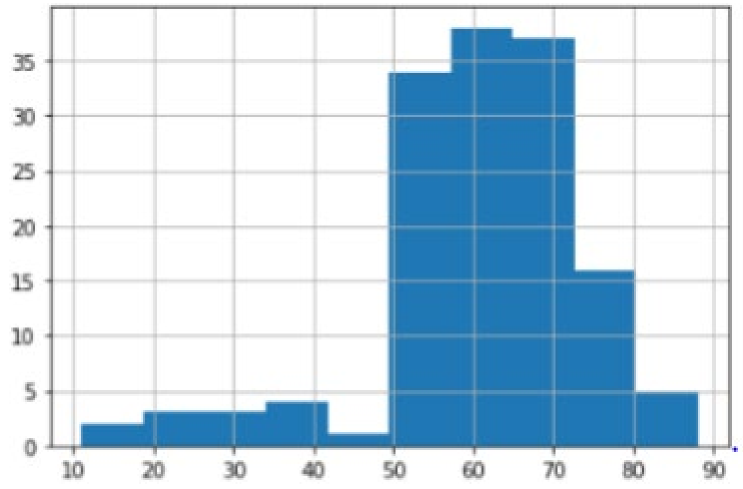


Figure 5.5: Distribution of Marks (married)

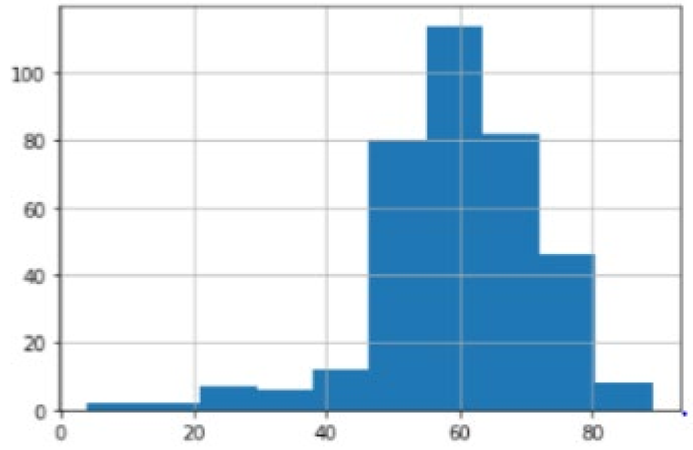


Figure 5.6: Distribution of Marks (single)

5.6 BUILDING THE MODEL TO PREDICT THE FINAL MARK

The first model built is a one that predicts the Final Mark of a student resting on all the features (the result of all the four exams). The first step included splitting the dataset into training / testing data which are (i) Two-Way Split and (ii) Three-Way Split.

5.6.1 Two-Way Split

When fitting a machine learning model to some data, the intention is to apply that model to forecasts/predicts on real-world data. Real-world data is unseen - it doesn't exist in the dataset - so to validate the model (check how well it performs), it needs to be tested on unseen data too.

However, gathering unseen data is not as simple as collecting it from outside the window and exposing it to the model: any new data would need to be cleansed, wrangled, and annotated just like the data in the dataset. The next best thing, then, is to simulate some unseen data, which can be done using the existing dataset by splitting it into two sets: one for training the model; and a second for testing it.

A model can be fitted using the training data. Subsequently, its accuracy can be assessed using the test set. Indeed, 20% of the data was set apart for testing while 80% was used for training. This implies that 20% of datapoints were in the test set, while 80% of data points or rows was allocated to the training set. A random selection is process is applied to these rows. This is done so that the mix of data in the train set is as close as possible to the mix in the test set. This research is making use of the two-way split.

5.6.2 Three-Way Split

Many academic works on machine learning talk about splitting the dataset into three distinct parts: train, validation, and test sets (winham et al., 2010). The idea here is that, as before, the training set would be utilized to apply the model to the observations. Thereafter, during the model tuning process where hyperparameters are tweaked and decisions on the dataset is made, the validation set is employed to test the model's performance. Once the model's performance on the validation set is acceptable, the previously unseen test set is brought out and used to deliver an unbiased evaluation of the model, that was initially fitted on the training data.

a) Caveats for using a validation set

On sufficiently small datasets, it may not be feasible to include a validation set for the following reasons, both of which should be intuitive:

- To accurately calculate model values, the model may need all possible data points.
- The uncertainty of the test set may be very high for small test sets. This causes differences in results from varying test sets.
- Evidently, further partitioning the training data into validation and training sets would remove precious observations for the training process.

5.6.3 Cross-Validation

To evaluate the effectiveness of a machine learning model based on unseen data, cross validation is a technique used. This entails splitting available information into test sets and training sets. The training set will thus be used for training learning, while the test sets will be used for evaluation. It involves dividing the available data into a training set and a test set, training the model on the training set, and evaluating the model on the test set (Little et al., 2017; Tennenholtz et al., 2018; Xu & Goodacre, 2018). This process is repeated multiple times, with the model being trained and evaluated on different splits of the data each time. Cross-validation can be subdivided as follows:

- K-fold cross-validation: data are randomly divided into k folds in the k-fold cross-validation technique. The model is trained and evaluated k times, and a different fold is used as a test set each time. The average of the k individual scores makes up the final evaluation score (Refaeilzadeh et al., 2009).

- Stratified k-fold cross-validation: data are split into k-folds in k-fold cross-validation. It is accomplished following a systematic or stratification method, which ensures that class distribution is similar for each fold of the cross-validation. The data are divided in kfolds according to a stratification method, which ensures class distribution is similar for each fold of the cross-validation. It is particularly useful in the case of an unbalanced distribution of classes (Prusty et al., 2022).

- Leave-one-out cross-validation: data is divided into k-folds in leave-one-out cross-validation, with k being the number of data points. A single data point is set apart as the test set each time the model is trained and evaluated (this happens k-times) (Cheng et al., 2017).

Cross-validation is a valuable technique for assessing the performance of a ML model. It offers a more robust estimate of the model performance. Cross-validation assists the model to be trained and evaluated on multiple different splits of the data. Finally, cross-validation is also helpful with preventing overfitting, as the model is not trained and evaluated on the same data (Refaeilzadeh et al., 2009).

Cross-validation is helpful in cases where the designer does not desire to use a validation set, or there is simply not enough data. K-fold cross validation is one of the most common versions of cross validation. During the training process, some proportion of the training data, say 10%, is held back, and effectively used as a validation set while the model parameters are calculated (Little et al., 2017; Tennenholtz et al., 2018; Xu & Goodacre, 2018).

In this research the available data was split into a test and a training set. to evaluate the algorithm's performance on unseen data, the test set was used. To train the algorithm, the training set was used. This led to algorithm being able to make predictions on the test set to evaluate how well the algorithm performed on unseen data.

5.7 MODEL EVALUATION

5.7.1 Introduction

Prediction models aim to generalize beyond the examples in the training set. The coefficients obtained from multiple linear regression model are checked for the size and sign. The bigger value coefficients represent higher relevance to the model and vice versa (Strecht et al., 2015; Bag, 2020; Altamini et al., 2022. If there is a positive or negative correlation between the dependent variable and each independent variable, the sign of the regression coefficient indicates this. For instance, the mean of the dependant variable increases when the independent variable value increases. This is denoting a positive coefficient, or positive correlation (El Guabassi et al., 2021; Sarker, 2021). A negative coefficient, however, trends in the opposite direction. The dependent variable tends to decrease as in the independent variable increases.

While holding constant other variables in the model, the coefficient value indicates how much the dependent variable's mean changes given changes in the independent variable (Silhavy et al., 2017; Emmert-Streib & Dehmer, 2019; Kalappan et al., 2021; Khan et al., 2022). The ability to hold the other variables constant is important because it allows to look at the effects of each variable in isolation.

The models were built simulating the options of having 2 marks, 3 marks and 4 marks. This helps predict student performances at various stages. Regression coefficients are tabulated in Table 5.4 below:

Table 5.4: Two marks

	Coefficient
Family_size	-1.20E-01
famrel	-2.66E+00
freetime	7.35E+11
hangout	-4.26E+11
Extra_support_	7.25E-02
Social_life_	-3.09E+11
T1	2.34E-01
T2	1.28E-01
married_No	4.09E+10
married_Yes	4.09E+10
SEX_Female	8.69E+10
SEX_Male	8.69E+10
adress_City	1.28E+10
adress_Town	1.28E+10
Living_stat_house or apartment	4.57E+10
Living_stat_living with Parents	4.57E+10
Living_stat_other	4.57E+10
Living_stat_residence	4.57E+10
Motheduc_4th grade	-8.24E+10
Motheduc_5th grade	-8.24E+10
Motheduc_9th grade	-8.24E+10
Motheduc_Other	-8.24E+10
Motheduc_higher education	-8.24E+10
Motheduc_none	-8.24E+10
Motheduc_secondary	-8.24E+10
Fathedu_4th grade	3.33E+10
Fathedu_5th grade	3.33E+10
Fathedu_9th grade	3.33E+10
Fathedu_higher education	3.33E+10
Fathedu_none	3.33E+10

	Coefficient
Fathedu_secondary	3.33E+10
MothJob_'health' care related	-2.34E+10
MothJob_'other'	-2.34E+10
MothJob_civil 'services'	-2.34E+10
MothJob_home	-2.34E+10
MothJob_teacher	-2.34E+10
FathJob_'health' care related	2.02E+10
FathJob_'other'	2.02E+10
FathJob_civil 'services'	2.02E+10
FathJob_teacher	2.02E+10
reason_'course' preference	4.13E+10
reason_'other'	4.13E+10
reason_close to 'home'	4.13E+10
reason_school 'reputation'	4.13E+10
reason_university fees	4.13E+10
Home_to_school_1 - <15 min	8.89E+09
Home_to_school_2 - 15 to 30 min	8.89E+09
Home_to_school_3 - 30 min. to 1 hour	8.89E+09
Home_to_school_4 - >1 hour	8.89E+09
study_time__2 - 2 to 5 hours	-5.19E+10
study_time__3 - 5 to 10 hours	-5.19E+10
study_time__4 - >10 hours	-5.19E+10
family_support_no	-1.31E+11
family_support_yes	-1.31E+11
extra_paid_no	-1.59E+10
extra_paid_yes	-6.87E+10
activities_no	5.76E+10
activities_yes	-4.24E+10
absences_0-5	-5.28E+09
absences_6-10.	-5.28E+09
absences_other	-5.28E+09

	Coefficient
First_language_Afrikaans	7.45E+10
First_language_English	7.45E+10
First_language_French	7.45E+10
First_language_Portuguese	7.45E+10
First_language_Xhosa	7.45E+10
Further_education__no	-4.24E+10
Further_education__yes	5.65E+10
internet__no	5.97E+10
internet__yes	-8.64E+10

By observing the coefficients, it is very clear that extra-curricular, language, parental education, studying time size have significant effects on final mark. The independent variables such as the use of internet shows a negative sign that states that these are negatively correlated with results as illustrated in Table 5:5 below:

Table 5.5: Three marks

Intercept: -5793200327.57596	Column1
	Coefficient
Family_size	-1.61E-01
famrel	-2.27E+00
freetime	-8.58E+10
hangout	5.76E+10
Extra_support_	-3.15E-01
Social_life_	2.81E+10
T1	1.31E-01
T2	4.80E-02
T3	2.39E-01
married_No	-2.61E+08
married_Yes	-2.61E+08
SEX_Female	-1.42E+09
SEX_Male	-1.42E+09
adress_City	-3.69E+08
adress_Town	-3.69E+08
Living_stat_house or apartment	4.28E+07
Living_stat_living with Parents	4.28E+07
Living_stat_other	4.28E+07
Living_stat_residence	4.28E+07
Motheduc_4th grade	-2.53E+09
Motheduc_5th grade	-2.53E+09
Motheduc_9th grade	-2.53E+09
Motheduc_Other	-2.53E+09
Motheduc_higher education	-2.53E+09
Motheduc_none	-2.53E+09
Motheduc_secondary	-2.53E+09
Fathedu_4th grade	1.92E+09
Fathedu_5th grade	1.92E+09
Fathedu_9th grade	1.92E+09

Intercept: -5793200327.57596	**Column1**
Fathedu_higher education	1.92E+09
Fathedu_none	1.92E+09
Fathedu_secondary	1.92E+09
MothJob_'health' care related	-5.52E+08
MothJob_'other'	-5.52E+08
MothJob_civil 'services'	-5.52E+08
MothJob_home	-5.52E+08
MothJob_teacher	-5.52E+08
FathJob_'health' care related	3.07E+08
FathJob_'other'	3.07E+08
FathJob_civil 'services'	3.07E+08
FathJob_teacher	3.07E+08
reason_'course' preference	1.60E+09
reason_'other'	1.60E+09
reason_close to 'home'	1.60E+09
reason_school 'reputation'	1.60E+09
reason_university fees	1.60E+09
Home_to_school_1 - <15 min	-3.05E+08
Home_to_school_2 - 15 to 30 min	-3.05E+08
Home_to_school_3 - 30 min. to 1 hour	-3.05E+08
Home_to_school_4 - >1 hour	-3.05E+08
study_time__2 - 2 to 5 hours	-1.89E+09
study_time__3 - 5 to 10 hours	-1.89E+09
study_time__4 - >10 hours	-1.89E+09
family_support_no	-1.08E+09
family_support_yes	-1.08E+09
extra_paid_no	-9.74E+08
extra_paid_yes	5.32E+09
activities_no	5.32E+09
activities_yes	3.07E+09
absences_0-5	1.97E+09

Intercept: -5793200327.57596	Column1
absences_6-10.	1.97E+09
absences_other	1.97E+09
First_language_Afrikaans	8.27E+07
First_language_English	8.27E+07
First_language_French	8.27E+07
First_language_Portuguese	8.27E+07
First_language_Xhosa	8.27E+07
Further_education__no	3.07E+09
Further_education__yes	-1.17E+09
internet__no	3.11E+09
internet__yes	-1.21E+09

5.8 BUILDING THE MODELS

The following sections demonstrate that the results of the algorithm were positive. However, first attempt of building the model with the existing data did not render positive results.

Another approach to enhance the model performance had to be followed:

- Increasing the size of the original data because it only contains 500 records.

- using a SMOTE algorithm that generates artificial data from the original one and this may lead to a better performance. The old data was passed to the SMOTE function, and it generates new artificial data. The number of records in the old data was 502. The number of artificial records that are generated is 10040. Now the total amount of data I have is 10542. New models were built with the new amount of data that were generated.

Thus, the results below show the mean squared error (MSE) the regression on score on both test and training data, before and after the SMOTE function was applied. This was done for three different models. For 2, 3, and 4 marks.

a) Model with Four marks

- **Random forest regressor**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.82	1.93
regr.score(X_train , Y_train)	0.97	0.998
regr.score(X_test , Y_test)	0.84	0.99

The model was trained using RandomForestRegressor using only T1, T2, T3, T4. The model returned 96.69% on training data before the SMOTE function, while retuning 99% after. On the test data it was 84% before SMOTE and 99% after SMOTE. Also, the Mean Squared Error (MSE) was 4.82 before the function and 1.93 after. It Is important that for four marks the results are positive before the SMOTE because it includes all four marks used in calculating the final mark (which is the variable being predicted) and 35% on test data highlighting an overfitting problem.

- **Support Vector Machine**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.88	1.92
regr.score(X_train , Y_train)	0.83	0.91
regr.score(X_test , Y_test)	0.86	0.93

The model was also trained on support vector machine (SVM). The model returned 83% on training data before the SMOTE function, while retuning 91% after. On the test data it was 86% before SMOTE and 93% after SMOTE. Also, the Mean Squared Error (MSE) was 4.88 before the function and 1.92.

- **XG Boost**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.95	1.80
regr.score(X_train , Y_train)	0.94	0.99
regr.score(X_test , Y_test)	0.76	0.99

The model was also trained on XG Boost. The model returned 94% on training data before the SMOTE function, while retuning 99% after. On the test data it was 76% before SMOTE and 99% after SMOTE. Also, the Mean Squared Error (MSE) was 4.95 before the function and 1.80. It is important to note also that for most the of the model, SVM performed the least. Random forest and XG Boost seem to be the most appropriate algorithms.

b) Model with two marks

- **Random forest regressor**

Evaluation criteria	Before Smote	After Smote
Mean squared error	10.12	1.95
regr.score(X_train , Y_train)	0.90	0.99
regr.score(X_test , Y_test)	0.39	0.97

The model was trained using RandomForestRegressor using only T1, T2. The model returned 90% on training data before the SMOTE function, while retuning 99% after. On the test data it was 39% before SMOTE and 97% after SMOTE. Also, the Mean Squared Error (MSE) was 10.12 before the function and 1.95 after. The model for two marks returns a higher MSE than for four marks. This is because the model has less marks to work with to predict the final mark. The results are also improved once the SMOTE function is applied.

- **XGBRegressor**

Evaluation criteria	Before Smote	After Smote
Mean squared error	10.60	1.81
regr.score(X_train , Y_train)	0.69	0.97
regr.score(X_test , Y_test)	0.66	0.99

The model was also trained on XG Boost. The model returned 69% on training data before the SMOTE function, while retuning 97% after. On the test data it was 66% before SMOTE and 99% after SMOTE. Also, the Mean Squared Error (MSE) was 10.6 before the function and 1.81. It is essential to also note that for most the of the model, SVM performed the least. Random forest and XG Boost seem to be the most appropriate algorithms.

- **Support Vector Machine**

Evaluation criteria	Before Smote	After Smote
Mean squared error	10.82	8.19
regr.score(X_train , Y_train)	0.45	0.52
regr.score(X_test , Y_test)	0.30	0.54

The model was also trained on support vector machine (SVM). The model returned 45% on training data before the SMOTE function, while retuning 52% after. On the test data it was 0.3% before SMOTE and 0.54% after SMOTE. Also, the Mean Squared Error (MSE) was 10.82 before the function and 8.19. SVM considerably underperformed compared to the rest of the algorithms.

c) Predicting marks using three marks (T1, T2, T3)

The same process was applied to fit the model for the model with three marks. The SMOTE algorithm was used to increase the number of records, to obtain better results and solve the overfitting problem.

- **Random forest regressor**

Evaluation criteria	Before Smote	After Smote
Mean squared error	7.94	2.79
regr.score(X_train , Y_train)	0.53	0.93
regr.score(X_test , Y_test)	0.46	0.98

The model was trained using RandomForestRegressor using only T1, T2, T3. The model returned 53% on training data before the SMOTE function, while retuning 93% after. On the test data it was 46% before SMOTE and 98% after SMOTE. Also, the Mean Squared Error (MSE) was 7.94 before the function and 2.79 after.

- **Support Vector Machine**

Evaluation criteria	Before Smote	After Smote
Mean squared error	7.43	3.21
regr.score(X_train , Y_train)	0.61	0.74
regr.score(X_test , Y_test)	0.53	0.73

The model was also trained on support vector machine (SVM). The model returned 61% on training data before the SMOTE function, while retuning 74% after. On the test data it was 0.53% before SMOTE and 0.73% after SMOTE. Also, the Mean Squared Error (MSE) was 7.43 before the function and 3.21. SVM considerably underperformed compared to the rest of the algorithms.

- **XG Boost**

Evaluation criteria	Before Smote	After Smote
Mean squared error	8.28	2.70
regr.score(X_train , Y_train)	0.81	0.98
regr.score(X_test , Y_test)	0.81	0.98

The model was also trained on XG Boost. The model returned 81% on training data before the SMOTE function, while retuning 98% after. On the test data it was 81% before SMOTE and 98% after SMOTE. Also, the Mean Squared Error (MSE) was 8.28 before the function and 2.7. The model performed better with three marks overall compared to two marks. Although the three marks model underperformed four marks.

5.9 THE APPLICATION OF ML ALGORITHM TO THE PROBLEM TO SECONDARY DATASET

The importance of obtaining a secondary dataset was to expose the algorithm to more data, and to more units of analysis. Initially the data was from a single cohort in Master in business management sciences. The secondary dataset however was richer it included data from both Postgraduate and Undergraduate students.

It also included from of Marketing students, Entrepreneurship, business computer application and several other units of analysis. This contributed to enhance the richness of the data, and to allow generalisation on much bigger scale. The second datasets provided had 36 instead 37 features from the previous datasets.

The data has both quantitative and qualitative data as the previous dataset. As machine learning does not understand text, this means that text was eventually transformed into numbers before it being fed to the algorithm. The same pre-processing steps that were applied to the prior datasets were also applied here and the below results were highlighted:

- **Random forest regressor**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.08	2.79
regr.score(X_train , Y_train)	0.86	0.93
regr.score(X_test , Y_test)	0.83	0.98

The model was trained using RandomForestRegressor. The model returned 86% on training data. On the test data it was 83%. Also, the Mean Squared Error (MSE) was 4.08.

- **Support Vector Machine**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.3	3.21
regr.score(X_train , Y_train)	0.81	0.74
regr.score(X_test , Y_test)	0.79	0.73

The model was also trained on support vector machine (SVM). The model returned 61% on training data before the SMOTE function, while retuning 74% after. On the test data it was 0.53% before SMOTE and 0.73% after SMOTE. Also, the Mean Squared Error (MSE) was 7.43 before the function and 3.21. SVM considerably underperformed compared to the rest of the algorithms.

- **XG Boost**

Evaluation criteria	Before Smote	After Smote
Mean squared error	4.38	2.70
regr.score(X_train , Y_train)	0.88	0.98
regr.score(X_test , Y_test)	0.87	0.98

The model was also trained on XG Boost. The model returned 81% on training data before the SMOTE function, while retuning 98% after. On the test data it was 81% before SMOTE and 98% after SMOTE. Also, the Mean Squared Error (MSE) was 8.28 before the function and 2.7. The model performed better with three marks overall compared to two marks. Although the three marks model underperformed four marks.

Conclusion

Given the above, it is conclusive from the above testing, remodelling and with the 3 marks model that Machine Learning with Bid Data Analytics is capable to predict the pass rate of learners and improve learning, teaching and assessment as illustrated by the conceptual framework of Figure 2.4. Hence Figure 5.7 below is the proposed general framework for the application of big data analytics to Improve students' performance in South Africa.

5.10 AST and the results of the machine learning algorithm

As indicated above, the results of the machine learning algorithm were considered positive, in predicting students’ performance. The implication of the results can be summarised by looking at them via the lens of Adaptive Structuration Theory. This can be done by considering the ways in which the technology in used or in practice creates and maintains social structures, and how these structures are shaping the outcomes of the algorithm. The following questions can be considered:

- How is the technology shaping the social structures that are being analysed/ predicted by the algorithm?
- How are these social structures influencing the outcomes of the algorithm?

How is the technology shaping the social structures that are being analysed/predicted by the algorithm?

The positive results obtained for the regression algorithms means that the machine learning algorithm can shape social structures the IHL in the following way:

Decision-making: the algorithms can be used to make decisions about students, such as detecting students with high risk of dropping out, assigning students to classes or interventions, or predicting students' grades. These decisions can shape the social structures of the IHL by influencing students' experiences and opportunities.

Personalisation: the algorithm can be used to personalise learning experiences for students, providing tailored content and recommendations based on students' interests and abilities. This can shape the social structures of the IHL by creating different learning paths and experiences for different students.

Assessment: Machine learning algorithms can be used to assess students' performance, providing automated feedback and grading. This can shape the social structures of the educational environment by influencing how students are evaluated and how they perceive their own abilities.

How are these social structures influencing the outcomes of the algorithm?

Social structures of the IHL influenced the outcomes of the algorithm in the following way:

Data: The data used to train and evaluate the algorithm were influenced by the social structures, such as the demographics of the individuals, the cultural context in which the data was collected, and any biases that were present in the data. These factors impacted the performance of the algorithm and the outcomes it produced. The variables that were used to predict students' performance included:

- **Previous academic performance**: Previous academic performance, such as grades or test scores, was a key predictor of future performance.

- **Demographic variables**: race, gender, age, or socio-economic status are demographic variables that were used to predict students' performance.

- **Attendance and engagement**: Attendance and engagement in class or other academic activities also related to students' performance.

- **Study habits and behaviours**: Factors such as study habits and behaviours, such as the amount of time spent studying or the use of study strategies, were also used to predict students' performance.

- **Personal and family characteristics**: Personal and family characteristics, such as motivation, support from family and friends, and self-esteem, were also related to students' performance.

Implementation: it is imperative to note that the implementation of a ML algorithm can be influenced by social structures, such as the organisational structures in which the algorithm is being used, the power dynamics between different groups of users, and the cultural context in which the algorithm is being applied. The social and institutional context may include factors such as:

- **The educational setting cultural and social norms**: The cultural and social norms of the educational setting, such as the expectations and values placed on students, may influence their performance.

- **The quality and accessibility of teaching and resources:** The quality and accessibility of teaching and resources, such as the expertise and experience of teachers and the availability of technology and materials, may also impact students' performance.

- **Students' social and economic context:** The social and economic context of the students, including factors such as their socio-economic status, family circumstances, and community resources, may also influence their performance.

- **The institutional policies and procedures:** The institutional policies and procedures, such as those governing grading and assessment, may also shape students' performance.

Interpretation: The interpretation of the results of a machine learning algorithm can be influenced by social structures, such as the expectations and assumptions of the individuals interpreting the results, the cultural context in which the results are being interpreted, and any biases that may be present in the interpretation process. These factors can impact how the results of the algorithm are used and the decisions that are made based on the results. For example, the inherent assumptions and hence interpretation, is that if the algorithm can predict students' performances, then it will be possible to predict students' performance. The choices and of the predicting variable is based on assumptions and interpretations.

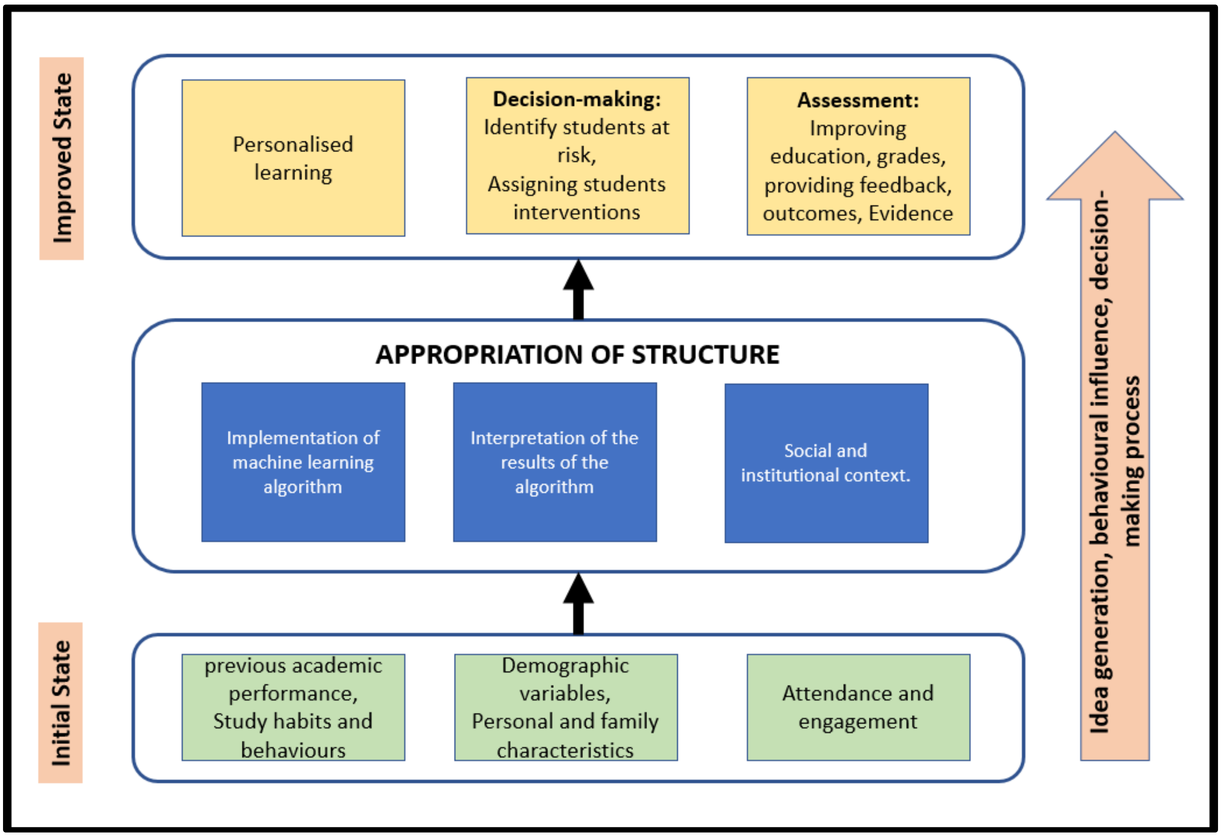

Improved State
Personalised learning
Decision-making: Identify students at risk, Assigning students interventions
Assessment: Improving education, grades, providing feedback, outcomes, Evidence
APPROPRIATION OF STRUCTURE
Implementation of machine learning algorithm
Interpretation of the results of the algorithm
Social and institutional context.
Initial State
previous academic performance, Study habits and behaviours
Demographic variables, Personal and family characteristics
Attendance and engagement
Idea generation, behavioural influence, decision-making process

Figure 5.7: Proposed General Framework

5.11 Discussion

This research aimed to utilize students' data to predict their academic performance, specifically their likelihood of passing or failing courses. The goal was to identify students at risk of failing so that targeted interventions can be provided. To accomplish this, machine learning techniques were applied to model and predicted student performance.

In addition to the general framework in figure 5.7, the contribution of this work is the development of machine learning models that can effectively predict students’ pass rates and final grades using variables related to demographics, family background, study habits, academic history, and more. Multiple models were built and evaluated, including regression algorithms like random forests, support vector machines, and XGBoost. The best performing models utilized ensemble methods like random forests and XGBoost.

A major component of this research was data understanding and preparation. The initial dataset contained 500 records and 37 features related to students. Extensive data cleaning and pre-processing was undertaken, including:

- Validation of unique values in each column
- Handling of missing values
- Ensuring all data was numeric rather than textual
- Exploratory data analysis using visualizations

To improve model performance, the SMOTE algorithm was applied to generate synthetic minority oversampling and balance the classes. This increased the number of records to 10,542. New models were built on this expanded dataset and showed significantly improved accuracy and lower error.

The models were trained to predict student final grades based on 2 exams scores, 3 exams scores, and 4 exams scores. This allowed performance prediction at multiple time points. Across metrics like R-squared, accuracy, and mean squared error, the ensemble methods of random forest and XGBoost outperformed other algorithms. For example, on 3 exams scores the random forest model improved from 53% accuracy to 93% accuracy after using SMOTE and tuning.

The table 5.6 compares model performance on key metrics like MSE, training score, and test score before and after using the SMOTE technique. Across multiple models and for predicting final marks based on varying number of subject marks, SMOTE consistently improved MSE, enhanced model fit and training accuracy, reduced overfitting, and helped models generalize better. This provides broad validation for the efficacy of using oversampling with synthetic minority data to boost model performance.

Table 5.6 Model Comparison

Model Feature/Metric	Before SMOTE	After SMOTE	Inference
MSE (Random Forest, 4 marks)	4.82	1.93	SMOTE improved performance by reducing MSE
Training Score (Random Forest, 4 marks)	0.97	0.998	SMOTE improved model fit on training data
Test Score (Random Forest, 4 marks)	0.84	0.99	SMOTE helped reduce overfitting
MSE (SVM, 4 marks)	4.88	1.92	SMOTE enhanced performance
Training Score (SVM, 4 marks)	0.83	0.91	SMOTE increased training accuracy

Test Score (SVM, 4 marks)	0.86	0.93	SMOTE helped model generalize better
MSE (Random Forest, 2 marks)	10.12	1.95	SMOTE substantially improved MSE
Training Score (Random Forest, 2 marks)	0.90	0.99	SMOTE enhanced training fit
Test Score (Random Forest, 2 marks)	0.39	0.97	SMOTE resolved overfitting issues
MSE (XGBoost, 3 marks)	8.28	2.7	SMOTE reduced error
Training Score (XGBoost, 3 marks)	0.81	0.98	SMOTE boosted training performance
Test Score (XGBoost, 3 marks)	0.81	0.98	Model generalized well after SMOTE

a) An Analysis of SMOTE and Its Significance:

The Synthetic Minority Over-Sampling Technique (SMOTE) is a commonly employed approach in the field of machine learning that aims to tackle the issue of imbalanced datasets (Soltanzadeh, & Hashemzadeh, 2021; Hussain et al., 2022). Imbalanced datasets manifest when there is a substantial disparity in representation between classes, typically with one class, often referred to as the minority class, being greatly underrepresented in comparison to the other class(es). The fundamental objective of SMOTE is to address the issue of class imbalance in datasets by generating synthetic samples for the minority class, thereby rebalancing the dataset.

The SMOTE algorithm operates by generating synthetic instances that are composed of amalgamations of pre-existing samples from the minority class. This process is achieved by selecting a sample from the minority class and identifying one or more of its nearest neighbours in the feature space. Subsequently, the system produces artificial instances by interpolating between adjacent samples along the line segments that connect them. The procedure successfully enhances the presence of the minority class inside the dataset.

Although the Synthetic Minority Over-sampling Technique (SMOTE) is widely recognized as a beneficial approach, it is important to acknowledge that it is not without its limitations and obstacles.

b) The potential for overfitting in a model.

The Synthetic Minority Over-Sampling Technique (SMOTE) has the potential to inject noise into the dataset due to the generation of synthetic samples that may not faithfully capture the genuine underlying distribution of the minority class. Overfitting can occur when a model demonstrates high performance on the training data but exhibits poor performance when applied to unseen data.
The phenomenon of information loss.

The synthetic samples produced by the Synthetic Minority Over-sampling Technique (SMOTE) are derived from the available data of the minority class, perhaps lacking the comprehensive representation of the entire range of variety within the minority class. The potential consequences of this phenomenon include a reduction in the amount of information available and the potential introduction of bias into the modelling process.

c) The Influence on Model Interpretability:

The use of synthetic samples into the analysis can introduce complexities in the interpretation of model predictions due to the potential lack of correspondence between these synthetic samples and real-world examples (Sauber-Cole & Khoshgoftaar, 2022). One of the key concerns in the field of computer science is the issue of scalability (Belgaum et al., 2021).

The computational cost of SMOTE can be significant, especially when working with datasets of considerable size. The augmentation of synthetic samples for each instance belonging to minority classes might substantially augment the overall size of the dataset (Temraz & Keane, 2022).

So, the influence of increased data on the outcomes of machine learning is significant. A larger dataset serves as a valuable source of information, enabling algorithms to uncover concealed insights, detect intricate patterns, and eventually enhance the accuracy and dependability of their predictions and classifications. This highlights the significance of acquiring data, ensuring its quality, and incorporating diverse data sets in the endeavour to develop more robust and efficient machine learning models.

d) The use of secondary dataset

A secondary student performance dataset was also obtained with new subjects and demographic groups. The best performing models still showed strong predictive capabilities on this dataset, demonstrating wider applicability. The secondary dataset had substantially more data.

Data plays a crucial role in the field of machine learning, serving as its fundamental resource. Data serves as the essential source that drives algorithms, enabling them to acquire knowledge, adjust their behaviour, generate forecasts, and make informed choices. The significance of data in the continuously developing field of machine learning and Big data cannot be overemphasized (Sarker, 2021).

The efficacy of machine learning models applied in this research was contingent upon the calibre, volume, and variety of the data on which they are trained. Increased availability of data allowed algorithms to analyse a wider range of information, enabling them to identify significant patterns more effectively.

Consequently, this reduced the likelihood of algorithms fitting excessively to irrelevant or random data, a phenomenon known as overfitting. The presence of high-quality data was essential for enabling algorithms to acquire knowledge from precise and inclusive instances, while the inclusion of diverse data facilitates the ability of models to adjust and respond to a broad range of circumstances and population groups.

The copious amount of data available from the secondary dataset offered a comprehensive collection of information, incorporating many events, circumstances, and intricacies that may have otherwise gone unnoticed in the initial, more limited dataset. In the context of this research, the secondary and larger dataset included a wider range of academic subjects, demographic characteristics, and behavioural features, so providing a comprehensive perspective on the various aspects that influence academic performance.

e) Analysing Patterns and Trends:

Machine learning algorithms exhibit high performance when applied to the identification and analysis of patterns and trends. They demonstrate exceptional proficiency in identifying repeating patterns and correlations within datasets (Hossain & Islam, 2023).

With an increased volume of data available, the algorithm possessed a heightened potential to discover complex and nuanced patterns that may have been difficult to identify within the smaller dataset. The capacity to discern subtle connections is of great significance, as it empowers algorithms to generate predictions or classifications with enhanced precision.

f) The potential for improved accuracy and reliability.

The pursuit of precision and dependability in machine learning is intricately linked to the abundance of available data. A more extensive dataset offers a more robust basis for training algorithms, enhancing their ability to effectively generalize to unfamiliar material (He et al., 2023).

Exposing algorithms to a wide range of instances mitigates the risk of overfitting, a prevalent issue characterized by models exhibiting high performance on training data but poor performance on unseen data. In situations such as medical diagnosis, where accuracy is of utmost importance, the presence of a large and comprehensive dataset can significantly impact the effectiveness of a diagnostic tool, differentiating between a valuable resource and a potentially hazardous undertaking.

5.12 SUMMARY:

The original data required a lot of pre-processing/cleaning to be fed into the model. As the data was not enough, some techniques were used to increase the size of data along with a secondary dataset with a substantial amount of data. Also, various models were used on the data, and based on the scores, the best ones were used in the future predictions. So, the key contributions are: (1) predictive modelling of student performance using machine learning, (2) model evaluation and tuning to maximize accuracy, (3) use of oversampling techniques to handle class imbalance, (4) demonstration of model efficacy across multiple datasets. The results indicate machine learning and big data analytics hold promise for predicting student outcomes.

CHAPTER 6: CONCLUSION AND RECOMMENDATIONS

6.1 INTRODUCTION

The core of this research is the use of Big Data analytics (specifically machine learning) concepts in Education to improve students' performance. This is discussed in Chapter 1, and subsequently elaborated in the following chapters. The results of the fieldwork and their analysis and interpretation were discussed in Chapter 5. The initial framework was discussed and revised as a general framework. The extent to which they addressed the objectives of the research are discussed in this chapter. Thus, this chapter concludes the research report as follows:

(a) Overview of the chapters.
(b) Research objectives revisited.
(c) Research contributions.
(d) Recommendations.
(e) Conclusion
(f) Limitations of the research and future research.
(g) Summary

6.2 OVERVIEW OF RESEARCH

The background to the research problem is discussed Chapter 1. The chapter indicated that although education is deemed an essential factor in the successful development of many societies there is a growing dissatisfaction with IHL throughput, and inability to contribute to the public good whether internationally and in South Africa. This is demonstrated by inefficiencies in IHL students' performances and graduation rates. Because of all the challenges that IHL faces the chapter discusses they are striving to improve throughput and output. Several approaches have been used by IHL including ICT tools such as LMS. The use of these various tools has led to mixed results. So IHL have had to consider other approaches such as the Big Data analytics. Finally, an overview of the rest of the thesis was given in the chapter.

Chapter 2 reviewed the research theoretical underpinning. Referring to the background and the problem statement, the position undertaken was that the phenomenon is a social reality that involves sociotechnical processes and as such it can be studied through the lenses of a social theory. Giddens' (1984) structuration theory (ST) and the adaptive structuration theory (AST) were deemed adequate to understand and interpret the phenomenon. Thus, an overview of structuration theory and its application in this research was provided in the chapter.

This led to Adaptive structuration theory (AST) as a theoretical lens to understand and interpret the embedded sociotechnical processes in improving pass rate at IHL. AST has been used in information systems research to investigate the capability of advanced technologies to influence the transformation of organisation or processes. AST was applied to the problem conceptualisation to derive a conceptual framework that guided the identification, customisation, and analysis of machine language algorithms used in the Big Data Analytics driving this research.

Chapter 3 reviewed current research work relevant to Big Data Analytics and how their conclusions, recommendations and gaps inform the attempt by this study to improve students' performances. Chapter 3 discussed that Student performances still do not reflect high academic success. In South Africa, it is estimated that between 50 % and 60 % of first year students drop out of university. Several factors were cited including students needing personalised support, owing to different levels of readiness and under preparedness of students entering IHL.

Chapter 3 also explains that Institutions of higher learning across the world have set for objectives to improve students' performance through improvement of teaching and learning. Due to the proliferation of software in the marketplace, ICT has been used in academical environment as well to keep up with market changes. The results are varied as some studies have shown a strong correlation between ICT use and both students' motivation and academic performances. While other studies show little to no correlation between ICT tools and students' performance.

Chapter 3 further explains that this has led to the considerations of other technologies such as BDA. The chapter then proceed to give an overview of big data and its attributes, and the nature of big data in education at different levels, macro, micro and meso levels. Then, the chapter discussed the properties of Big Data Analytics (BDA) to assist in providing teaching and learning structure that empowers and enhances students' performance. Furthermore, it discussed the potential of BDA to help Universities to operate more efficiently, enabling lecturers' effectiveness to improve their methods, and prevent students from falling into poor academic performance. The chapter also argued that it is essential to exploit the opportunities offered by this technology to improve and possibly, maximise output and performance of learners. Finally, the chapter established machine learning as the BDA tools in this research and gave overviews of machine learning algorithms.

Chapter 4 discussed the steps undertaken to utilise students' data to predicts students' performance. These steps were necessary to automate and improve students' performance by identifying students at risk. Thus, the chapter presented how this model was built through following each of the 6 stages of the revised framework of CRISP-DM (Shearer, 2000) used in educational data mining. from domain and data understanding to knowledge discovery. Chapter 4 applied the CRISP-DM stages as follows:

- For Domain Understanding goals were established and the relevant stakeholders were defined.
- For Data Understanding was collected and checked completeness and redundancy.
- For Data preparation data was cleaned and transformed to prepare it for use in the model. The output was a dataset suitable for use in the next step.

- For Application of machine learning a regression algorithm was deemed appropriate given the nature of the data to achieve the prediction. Once the correct models were selected, they were then applied to the data.

- For Evaluation, the results from the application of machine learning were interpreted. This involves the discovery of new patterns. The step also involved reconsidering prior steps as the initial poor results had to be improved.

- The discovered knowledge was thus documented with the aim to be transferred to interested parties.

Chapter 5 describes how the model was selected. A regression algorithm was deemed appropriate given the nature of the data, to achieve the prediction. RandomForestRegressor and XGBRegressor were then selected. Once selected, they were then applied to the data. Training was then conducted on the data, and the results evaluated. For Evaluation the results from the application of machine learning were interpreted. The original data required a lot of pre-processing/cleaning to be fed into the model. The size of data was not enough, this led to some poor results initially. A SMOTE technique was used to increase the size of data. Which contributed to improving the results.A secondary datasets was also used, on which the models were also applied. Finally, various models were used on the data, and based on the scores, the best ones were used in the future predictions.

Chapter 6 was a conclusive episode of the thesis. Thus, the overview of the content of the previous chapters was provided. Answers to questions defined in Chapter 1 were given based on the research content; and the research questions were revisited. Research contributions were provided on a theoretical, methodological, and practical level. Lastly, recommendations along with limitations were specified with further research.

6.3 RESEARCH OBJECTIVES REVISITED

One of the main goals of Institutions of higher learning in South Africa is to improve students' performances through improved teaching and learning (Nyamupangedengu, 2017; Shay, 2017). The research recognised that inadequate mining or analytics of Big Data had led to insufficient information that can assist in providing teaching and learning structure that will empower and enhance students' full potential (Reyes, 2015). Many of South African institutions make use of various approaches and methods in attempt to achieve this goal. The various approaches have provided mixed results this led institutions of higher education to consider BDA. Therefore, it was important to to explore the use of big data analytics to address the challenges that institutions of higher learning face in their attempt to improve students' academic performance. The problem was answered through the stated research objectives as provided in the following section.

6.3.1 Research Objective 1: To examine current use of software to improve students' performance

Institutions of higher learning across the world have set for objectives to improve students' performance through improvement of teaching and learning. Due to the proliferation of software in the marketplace, Information communication technology (ICT) has been used in academical environment as well. Indeed, existing in an era with fast paced developing technologies, IHL have also decided to incorporate ICT to keep up with market changes. The results are varied.

Some Studies have shown a strong correlation between ICT use and both students' motivation and academic performances. While other studies show little to no correlation between ICT tools and students' performance. Also, some scholars believe that technology have already transformed universities. However, some believe that technology is disruptive, and universities have failed to cope with it.

In the Case of IHL in South Africa, the tools that have been implemented such as LMS, learning management systems (LMS) to assess and manage students' academic activities, for improvement purposes. These have produced mixed results, from both poor and positive perspectives. In fact, many critiques alluded to the fact this software were not meeting their primary objectives which was to improve students' performances through improved teaching and learning in IHL. As a result, some institutions of higher learning seek different approaches.

In south Africa, there is potential for enhancing IHL decision-making through Big Data analytics, yet various obstacles hinder this progress. These challenges encompass factors like the organization's readiness and ability to handle large volumes of data, inadequate resources such as skilled personnel and suitable technology, lack of knowledge in utilizing effective analysis techniques for interpreting data insights, and individual traits playing a role as well.

Across the world, ICT tools such as big data analytics have been adopted and used to improve students' performances through improved teaching and learning. Indeed, educational organisations explore the use of big data to match students' career and goals with their studies, conduct evidence-based planning, and improve students' effort. Also, Big data Analytics can assist in discovering students' learning patterns which can be used to create teaching and learning programs specific to individuals. For instance, the case of Arizona State University, they facilitated their math courses using computer-based learning program. Teachers are no longer the only ones on board to monitor and assess students' progress.

6.3.2 Research Objective 2: To investigate the nature of teaching and learning big data generated in the University.

Section 3.4 discusses that educational big data can be collected at three different level. The Data can be collected at Microlevel, which represent data created within seconds between actions that can capture data and multiple students.

a) In general, Microlevel Big Data are gathered automatically during students' interaction with their learning environment. These environments include MOOCs, simulations, intelligent tutoring systems, and games.

b) Mesolevel Big Data comprise a set of computerised written corpora gather during students' writing activities in their respective learning environment. These artifacts vary from students' assignments, social media interaction and/or writings from online discussion forums. The opportunities offered by Mesolevel Big Data include include the capacity to capture students' progression in emotional states, social and intellectual abilities.

c) Macrolevel big data refer to data that is captured at the level of the institution. This includes admission data, student demographics, course enrolment, degree completion and campus data. This category of big data is usually collected over multiple years but updated infrequently.

It is necessary to observe that although these categories are represented as distinct, they can intertwine; there can be some form of overlaps. Social media data that might constitute meso-level big data may have stamps that may qualify it as microlevel big data. This is not a challenge as it provides opportunities for better analysis. In the follow sections, the different categories will be probed further. Section 4.4 discussed the data harvested for this research. Data harvested included students' marks, and demographics.

6.3.3 Research Objective 3: To determine the relevance of teaching and learning big data in improving student's performance.

Section 3.4 also explains that the data can be used to identify emotional states, cognitive strategies, or self-regulated learning behaviours. Educational Big data can also be used by administrators to improve administrative decisions, enhance student satisfaction, and boost college success, while also generating data-driven decision, and building early warning systems. With regards to this research the data collected was used to predict students' final marks, to help identify students at risk of failing. The data was used to develop a form early warning system.

For many South African institutions inadequate mining or analytics of Big Data had led to insufficient information that can assist in providing teaching and learning structure that will empower and enhance students' full potential. The adoption of Big Data in South Africa is rife with challenges such as universities' readiness and ability to handle large volumes of data, inadequate resources such as skilled personnel and suitable technology, lack of knowledge in utilizing effective analysis techniques for interpreting data insight Despite the above mentioned, Big Data's potential is still growing for South African Institutions

6.3.4 Research Objective 4: To determine the relevance of data analytics, through machine learning, in predicting students' performance.

Section 3.6.1 explains that from elementary school to university, big data is affecting education across the spectrum of learning. Big data systems help teachers and lecturers learn more about people's behaviour and form new conclusions, as the standard of technology and education is evolving. Therefore, it is increasingly important for teachers and lecturers to understand the latest developments in education and data analytics. Today's teachers use big data technologies to find students' problems area, instead of relying on standardized tests to detect problems. Through adaptive learning, students can allocate more time in challenging subject areas while remaining in step with their classmates.

Educators are effectively using big data systems to monitor students' progress and likelihood of advancement, and to assess students accurately. Currently, there has been a significant improvement in the results achieved by learning institutions that have implemented big data systems to monitor and evaluate student performance. The development of educational plans has also been facilitated by technology to improve the student's engagement.

Section 3.6.1 explains BDA can be used in the education sector to improve students' results. Currently, answers to tests and assignments represent the only measure of the student's performance. However, each of the students creates a unique data trail throughout his or her life. However, each of the students creates a unique data trail throughout his or her life. To gain a better understanding of the student's behaviour and to create an optimum learning environment for students, it will be useful to analyse that data trail in real time.

Monitoring of student's behaviour, including how long it takes them to respond to a question, the resources they have used for exam preparation, their decisions to skip questions is possible thanks to the vast amount of data available in the education sector. This tracking of students' performances can be in giving each student instant feedback.

Section 3.5.1 also explains that BDA in the education sector: customise programs. Regardless of the number of students at universities and colleges, tailored programmes may be developed for each student. Using what is known as "blended learning," a blend of internet and offline education, this may be achieved. This enables students to take classes they're interested in and work at their pace while maintaining the possibility of being taught by professors on an offline basis. This is already apparent in the case of Massive Open Online Courses that are being developed worldwide. For example, only 400 students took the machine learning course taught by Andrew Ng at Stanford University.

But it had attracted 100,000 students when the same course was delivered as a MOOC.

With regards to this research, data was used to predict students' performances. Being able to predict students results at different stages during the semester could help identify trends, and students at risk of failing. This would students seek more help sooner rather than later, with lecturers being to provide customised support at earlier stages.

6.4 RESEARCH CONTRIBUTIONS

the theoretical, methodological, and practical knowledge contribution this research has made to the disciplinary knowledge are discussed in this section. These contributions are generated through the application of machine learning to predict students' performance.

6.4.1 Theoretical Contributions

The use of underpinning theory to help achieve the results is the essence of the theoretical contribution by this research. Although there are other social theories that could have been used, however the dimensions of AST in relation to duality of technology provided the "clearest" theoretical lens that led to deeper understanding of the embedded sociotechnical processes associated with teaching and learning and learners' performance in relation to pass-rate. AST was used on the problem conceptualisation to derive a conceptual framework that guided the identification, customisation, and analysis of ML algorithms used in the Big Data Analytics that was driving this research.

AST was also used to interpret the outcome of the selected machine learning algorithm with the structured and unstructured data that provide convincing prediction of a learner's performance and pass-rate. As such the adoption of AST as the underpinning theory of the initial framework has led to the proposed general framework that will guide in determining the application of BDA to improve students' performance.

Furthermore, the use of BDA to improve students' performance is an emerging discipline in South Africa. The research also assisted with providing meaningful literature in the domain.

6.4.2 Methodological Contributions

The use of the experimental approach where the selected algorithm was simulated in the lab with data from a cohort instead of the traditional data collection is the methodological contributions in this research. The proposed revised CRISP-DM model was used to help guide the field work. The Crisp- DM model guided the following stages: domain understanding; data understanding; data preparation; application of machine learning, evaluation, and the use of discovered knowledge.

Therefore, the utilisation, application, and the combination of Adaptive Structuration Theory, through an initial framework of analysis, presented in chapter 2, and the CRISP- DM model provided the methodological contributions in this research. The research methodology used was instrumental in teasing out the problem and could serve as an example for other researchers. Finally, the research also identified the most appropriate algorithms for the predictions at hand.

6.4.3 Practical Contributions

The aim of this research was to explore the use of big data analytics to address the challenges that institutions of higher learning face in their attempt to improve students' academic performance. In Chapter 3, the aim was met through an in-depth literature. Various aspects of big data analytics were discussed in Chapter 3 where it was shown how BDA can contribute to Improving students' performances. Also, the research used BDA (machine learning) to predict students' performance, using regression algorithms. The use of the algorithm will then assist lecturers in determining which students are at risk of failing.

In addition, the conceptual framework of Figure 2.4 developed from the application of AST to the problem conceptualisation Figure 1.1 which guided the analysis of findings was affirmed in Chapter 5 as the general framework Figure 5.5 The contribution is meant to be normative, as the general framework is normative in nature. However, despite the normative nature of the output, it was to address any potential issue in Improving students' performance using BDA. The general framework can be used as a guide to facilitate the appropriation of BDA by Institution of Higher Learning.

6.5 RECOMMENDATIONS

To make use of BDA to assist in predicting students' performance it is important for universities to be educated on the different use cases of BDA in education. Also, it is imperative to identify potential barriers to adoption of BDA for IHL. Once that is determined a strategy can then be put in place to address the use of data analytics. Following the general framework derived from this research and the revised CRISP-DM can assist with this (figure 6.1). In chapter 2, it was stated that interaction with BDA impacts the structural properties of the institution, through the reinforcement and transformation of structures of signification.

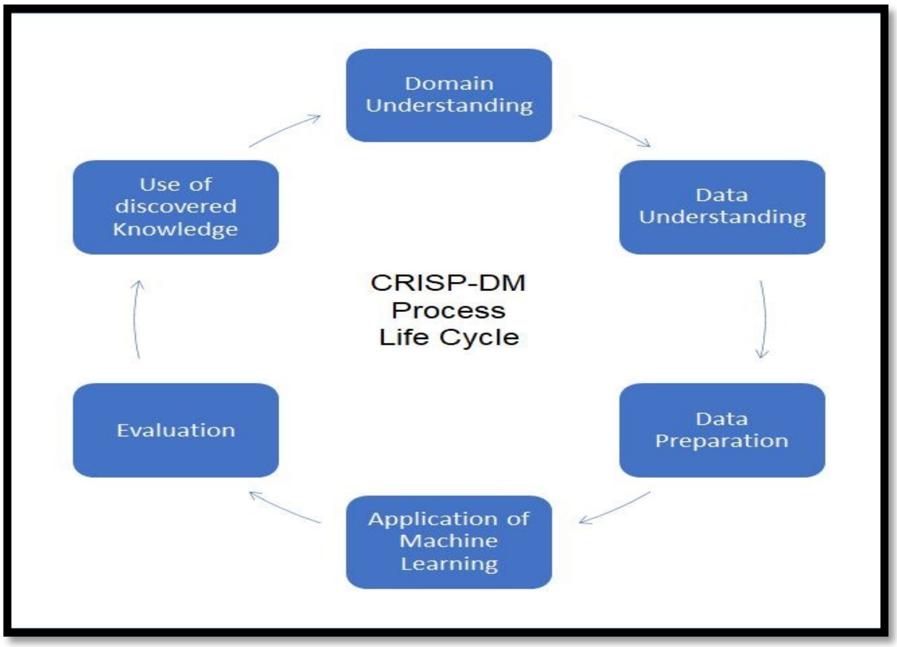

Domain Understanding
Data Understanding
Data Preparation
Application of Machine Learning
Evaluation
Use of discovered Knowledge
CRISP-DM Process Life Cycle

Figure 6.1: Revised CRISP-DM Process Life Cycle (Kurgan and Musilek, 2006)

In addition to the proposed general framework and the revised CRISP-DM model, the following recommendations can be made:

Data collection: decision making is strongly data driven in this current era. And in this digital era several tools are available. Collecting as much data, and as varied as possible will help improve the quality and accuracy and the predictions. Indeed, BDA requires huge amount data to be analysed for interesting pattern and so data should be collected on regular basis by IHL so that new and helpful discoveries could be made to enhance the process of predicting students' performances. For instance, in the context of this research, having students' behavioural data during testing could have helped improved the accuracy of the prediction. But because this data was not collected, it could not assist.

Model training: the concept of machine learning is for machine to learn by themselves. Once the algorithm is implemented, it is important to retrain the model on new data so that it will learn continue to learn and improve.

Data storage: For ease and access of use, it is important to a single repository for the data of the institution. This will make retrieval easy. Data can be stored on platforms that can accommodate for both structure and unstructured data (this is usually the case in NOSQL databases)
Further research on the CRISP-DM model by IHL can also provide guidance in IHL attempts to use BDA.

6.6 CONCLUSION

Education is considered a basic need for individuals. However low performance and inefficacies in graduation rates affect IHL throughput. It is as though IHL are failing to achieve their objectives. This has then caused IHL to look at other alternatives such as LMS which have produced mixed results. Across the world other IHL have now started looking at other technology such as BDA to improve students' performance. This research tried to understand the problem at hand via the use of AST. AST assisted in teasing the problem and helped derived conceptual framework.

The conceptual framework illustrates the appropriation of structure and the process to influence behaviour by IHL. This associated with the CRISP-DM helped provided an approach to the use of BDA in IHL using AST and the CRISP- DM Model, the research achieved its objectives to predict students' performance via an using a regression algorithm.

6.7 LIMITATIONS AND FURTHER RESEARCHES

The limitation of this research is mainly associated with the data harvested and the cohort. Indeed, the research had to make use of SMOTE statistical technique to increase the data at hand and to improve the results of the algorithm. Therefore, it will be more important to have more data spanning across a wider period, with several subjects to help improve the accuracy and veracity of the predications.

IHL also dealt with issues of privacy when it comes to the data, further work could focus on ensuring security and confidentiality, and privacy protection of the students' data. This will thus assist in streamlining all the delays and difficulties encountered in the process of collecting data. The scope of this study was focused on one cohort, MBIS. Further research could also be taking data from undergraduate and across different qualifications and different cohort for a comparative study.

Finally, further research in extending system abilities to predicting individual students' grades, monitoring students' learning patterns, and suggesting intervention methods applicable to stakeholders is also recommended.

REFERENCE LIST

Abdullah, M.A. and Fahad A.A., 2019. Enhancing Performance of Educational Data Using Big Data and Hadoop. International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 19 (2019) pp. 3814-3819. https://www.ripublication.com/ijaer19/ijaerv14n19_15.pdf

Abel, V.R., Tondeur, J. and Sang, G., 2022. Teacher perceptions about ICT integration into classroom instruction. Education Sciences, 12(9), p.609.

Ahmed, C., ElKorany, A. and ElSayed, E., 2022. Prediction of customer's perception in social networks by integrating sentiment analysis and machine learning. Journal of Intelligent Information Systems, pp.1-23.

Ahmed, S.B., Solis-Oba, R. and Ilie, L., 2022. Explainable-AI in Automated Medical Report Generation Using Chest X-ray Images. Applied Sciences, 12(22), p.11750.

Akter, S. and Wamba, S.F., 2016. Big data analytics in E-commerce: a systematic review and agenda for future research. Electronic Markets, 26(2), pp.173-194.

Al-Akhras, M., El Hindi, K., Habib, M. and Shawar, B.A., 2021. Instance reduction for avoiding overfitting in decision trees. Journal of Intelligent Systems, 30(1), pp.438-459.

Al-Rahmi, W., Aldraiweesh, A., Yahaya, N., Kamin, Y.B. and Zeki, A.M., 2019. Massive open online courses (MOOCs): Data on higher education. Data in brief, 22, pp.118-125.

Alsaaidah, B., Al-Hadidi, M.D.R., Al-Nsour, H., Masadeh, R. and AlZubi, N., 2022. Comprehensive Survey of Machine Learning Systems for COVID-19 Detection. Journal of Imaging, 8(10), p.267.

Ali, S., Jutla, D. N., & Bodorik, P. 2013. Engineering privacy for big data apps with the unified modeling language. In Big Data (BigData Congress), 2013 IEEE International Congress on (pp. 38-45). IEEE.

Altamimi, A.M., Azzeh, M. and Albashayreh, M., 2022. Predicting students' learning styles using regression techniques. Indonesian Journal of Electrical Engineering and Computer Science, 25(2), pp.1177-1185.

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L., 2021. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. Journal of big Data, 8(1), pp.1-74.

Alzubi, J., Nayyar, A. and Kumar, A., 2018, November. Machine learning from theory to algorithms: an overview. In Journal of physics: conference series (Vol. 1142, No. 1, p. 012012). IOP Publishing.

Appel, A.P., Candello, H. and Gandour, F.L., 2017. Cognitive computing: Where big data is driving us. In Handbook of Big Data Technologies (pp. 807-850). Springer, Cham.

Attaran, M., Stark, J. and Stotler, D., 2018. Opportunities and challenges for big data analytics in US higher education: A conceptual model for implementation. Industry and Higher Education, 32(3), pp.169-182.

Ausin, M.S., 2019, January. Leveraging deep reinforcement learning for pedagogical policy induction in an intelligent tutoring system. In In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019).

Avella, J.T., Kebritchi, M., Nunn, S.G. and Kanai, T., 2016. Learning analytics methods, benefits, and challenges in higher education: A systematic literature review. Online Learning, 20(2), pp.13-29.

Awad, M. and Khanna, R., 2015. Machine learning and knowledge discovery. In Efficient learning machines (pp. 19-38). Apress, Berkeley, CA.
Azamfirei, L., 2016. Knowledge is power. The Journal of Critical Care Medicine, 2(2), p.65.

Ayuk, P.T. and Koma, S.B., 2019. Funding, access and quality conundrum in South African higher education. African Journal of Public Affairs, 11(1), pp.176-195.

Babu, N.V. and Kanaga, E., 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. SN Computer Science, 3(1), pp.1-20.

Badaru, K.A. and Adu, E.O., 2022. Platformisation of Education: An Analysis of South African Universities' Learning Management Systems.
Research in Social Sciences and Technology, 7(2), pp.66-86.

Bag, A., 2020. A comparative study of regression algorithms for predicting graduate admission to a university.

Baig, M.I., Shuib, L. and Yadegaridehkordi, E., 2020. Big data in education: a state of the art, limitations, and future research directions. International Journal of Educational Technology in Higher Education, 17(1), pp.1-23.

Baker, R. S. J. D. 2010. Data mining for education. International encyclopedia of education, 7(3), 112-118.

Baker R.S. & Inventado P.S. 2014. Educational Data Mining and Learning Analytics. In: Larusson J., White B. (eds) Learning Analytics. Springer, New York, NY

Batko, K. and Ślęzak, A., 2022. The use of Big Data Analytics in healthcare. Journal of big Data, 9(1), pp.1-24.

Belgaum, M.R., Alansari, Z., Musa, S., Alam, M.M. and Mazliham, M.S., 2021. Role of artificial intelligence in cloud computing, IoT and SDN: Reliability and scalability issues. International Journal of Electrical and Computer Engineering, 11(5), p.4458.

Ben Youssef, A., & Dahmani, M. 2008. The impact of ICT on student performance in higher education: Direct effects, indirect effects and organisational change. HAL.

Bhanu, 2018. Big Data in Education. https://plopdo.com/2018/11/02/big-data-in-education/.

Department of Education. Retrieved from http://www.ed.gov/edblogs/technology/files/2012/03/edm-labrief.pdf
Borray, A., & Millichap, N. 2017. Trends and Technologies: iPASS.

Boughey, C. (2018). Using the curriculum to enhance teaching and learning. South African Journal of Science, 114(9-10): 1-3.

Bluewolf, 2016. The State of Salesforce Report.

Botha, A.J.M., 2020. A learning management system based framework for higher education quality programme review (Doctoral dissertation, University of Pretoria).

Boughey, C., 2018. Using the curriculum to enhance teaching and learning. South African Journal of Science, 114(9-10), pp.1-3.

Bozalek, V., Gachago, D., Alexander, L., Watters, K., Wood, D., Ivala, E. and Herrington, J., 2013. The use of emerging technologies for authentic learning: AS outh A frican study in higher education. British Journal of educational technology, 44(4), pp.629-638.

Brende, B. 2015. Why education is the key to development. World economic forum,7 July. Accessed online: https://www.weforum.org/agenda/2015/07/why-education-is-the-key-todevelopment/

Brock, V. and Khan, H.U., 2017. Big data analytics: does organizational factor matters impact technology acceptance?. Journal of Big Data, 4(1), pp.1-28.

Broger, D. 2011. Structuration theory and organization research. na.

Brown, S., 2021. Machine learning, explained. https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

Buraimoh, E., Ajoodha, R. and Padayachee, K., 2021, June. Importance of Data Re-Sampling and Dimensionality Reduction in Predicting Students' Success. In 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE) (pp. 1-6). IEEE.

Burtsev, V., 2021. ADOPTION OF LEARNING MANAGEMENT SYSTEMS AT SOUTH AFRICAN LEARNING INSTITUTIONS. In INTED2021 Proceedings (pp. 10818-10824). IATED.

Callegaro, M. and Yang, Y., 2018. The role of surveys in the era of “big data”. In The Palgrave handbook of survey research (pp. 175-192). Palgrave Macmillan, Cham.

Calloway, D., 2010. Adaptive Structuration Theory: Understanding How Advancing Technologies Drive Organisational Change. Weaverville, North Carolina, USA.

Cao, L., 2017. Data science: a comprehensive overview. ACM Computing Surveys (CSUR), 50(3), pp.1-42.

Castaneda, J., Jover, A., Calvet, L., Yanes, S., Juan, A.A. and Sainz, M., 2022. Dealing with Gender Bias Issues in Data-Algorithmic Processes: A Social-Statistical Perspective. Algorithms, 15(9), p.303.

Castillo-Merino, D., & Serradell-López, E. 2014. An analysis of the determinants of students' performance in e-learning. Computers in Human Behavior, 30: 476-484.

Cele, N., 2021. Big data-driven early alert systems as means of enhancing university student retention and success. South African Journal of Higher Education, 35(2), pp.56-72.

Cena, F., Rapp, A., Musto, C. and Semeraro, G., 2020. Generating recommendations from multiple data sources: A methodological framework for system design and its application. IEEE Access, 8, pp.183430-183447.

Chaurasia, S.S., Kodwani, D., Lachhwani, H. and Ketkar, M.A., 2018. Big data academic and learning analytics: Connecting the dots for academic excellence in higher education. International Journal of Educational Management, 32(6), pp.1099-1117.

Cheng, H., Garrick, D.J. and Fernando, R.L., 2017. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. Journal of animal science and biotechnology, 8(1), pp.1-5.

Chetty, R., Pather, S. and Condy, J., 2015. Challenges in higher education in South Africa. Telling Stories Differently: Engaging 21st century students through digital storytelling, pp.1-6.

Churová, V., Vyškovský, R., Maršálová, K., Kudláček, D. and Schwarz, D., 2021. Anomaly Detection Algorithm for Real-World Data and Evidence in Clinical Research: Implementation, Evaluation, and Validation Study. JMIR medical informatics, 9(5), p.e27172.

Council of Higher Education 2017. Learning to Teach in Higher Education in South Africa: An investigation into the influences of institutional context on the professional learning of academics in their roles as teachers. https://www.che.ac.za/sites/default/files/publications/PUB_HE%20Monitor%2014_20170401.pdf

Da Costa, R.L., Gupta, V., Gonçalves, R., Dias, Á., Pereira, L. and Gupta, C., 2022. Artificial Intelligence and Cognitive Computing in Companies in Portugal: An Outcome of Partial Least Squares—Structural Equations Modeling. Mathematics, 10(22), p.4358.

Datameer 2016. 5 Big Data Use Cases to Understand Your Customer Journey: Customer Analytics Ebook. https://www.datameer.com/pdf/Datameer-Customer-Analytics-ebook.pdf

David, Hazel & Ramos, Majayma & Anne, Justine & Valenzuela, Mae & Lourdes, Ma & Danganan, Catherine. (2022). Exemplifying the Implementation of the "No Child Left Behind Policy" on the Elementary Schools. 135-148.

Deepa, N., Pham, Q.V., Nguyen, D.C., Bhattacharya, S., Prabadevi, B., Gadekallu, T.R., Maddikunta, P.K.R., Fang, F. and Pathirana, P.N., 2022. A survey on blockchain for big data: approaches, opportunities, and future directions. Future Generation Computer Systems.

Denis, D., (2018). Simple and Multiple Linear Regression. 10.1002/9781119465775.ch9.

DeSanctis, G., & Poole, M. S. 1994. Capturing the complexity in advanced technology use: Adaptive structuration theory. Organization science, 5(2): 121-147.

Dhanaraj, R.K., Rajkumar, K. and Hariharan, U., 2020. Enterprise IoT modeling: supervised, unsupervised, and reinforcement learning. In Business Intelligence for Enterprise Internet of Things (pp. 55-79). Springer, Cham.

Dinu, D., Stoica, I. and RADU, A.V., 2016. Studying the Consumer Behavior through Big Data. Quality-Access to Success, 17.

Duderstadt, J. J., Atkins, D. E., & Van Houweling, D. E. 2002. Higher education in the digital age: Technology issues and strategies for American colleges and universities. Greenwood Publishing Group.

Dutt, A., Aghabozrgi, S., Ismail, M.A.B. and Mahroeian, H., 2015. Clustering algorithms applied in educational data mining. International Journal of Information and Electronics Engineering, 5(2), p.112.

Eggers, W. D. 2007. Government 2.0: Using technology to improve education, cut red tape, reduce gridlock, and enhance democracy. Rowman & Littlefield.

Elreedy, D. and Atiya, A.F., 2019. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Information Sciences, 505, pp.32-64.

El Guabassi, I., Bousalem, Z., Marah, R. and Qazdar, A., 2021, January. Forecasting Students' Academic Performance Using Different Regression Algorithms. In International Conference on Digital Technologies and Applications (pp. 221-231). Springer, Cham.

Emmert-Streib, F. and Dehmer, M., 2019. Evaluation of regression models: Model assessment, model selection and generalization error. Machine learning and knowledge extraction, 1(1), pp.521-551.

Erdaw, Y. and Tachbele, E., 2021. Machine learning model applied on chest X-ray images enables automatic detection of COVID-19 cases with high accuracy. International Journal of General Medicine, 14, p.4923.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A. and Kashef, R., 2020. Recommendation systems: Algorithms, challenges, metrics, and business opportunities. applied sciences, 10(21), p.7748.

Findley, B., 2020. Why racial bias is prevalent in facial recognition technology. Harvard Journal of Law and Technology.

Fischer, C., Pardos, Z.A., Baker, R.S., Williams, J.J., Smyth, P., Yu, R., Slater, S., Baker, R. and Warschauer, M., 2020. Mining big data in education: Affordances and challenges. Review of Research in Education, 44(1), pp.130-160.

Fischman, J. 2011. The rise of teaching machines. The Chronicle of Higher Education.

Fletcher, S. 2013. Special report. Scientific American, 309(2): 48-73.

Fujiwara, K., Huang, Y., Hori, K., Nishioji, K., Kobayashi, M., Kamaguchi, M. and Kano, M., 2020. Over-and under-sampling approach for extremely imbalanced and small minority data problem in health record analysis. Frontiers in public health, 8, p.178.

Gachago, D., Ivala, E., Backhouse, J., Bosman, J. P., & Bozalek, V. 2013. Towards a shared understanding of emerging technologies: Experiences in a collaborative research project in South Africa.

Gachago, D., Bozalek, V., & Ng'ambi, D. 2013. Transforming teaching with emerging technologies: Implications for higher education institutions. South African Journal of Higher Education, 27(2): 419-436.

Gamede, B.T., Ajani, O.A. and Afolabi, O.S., 2022. Exploring the adoption and usage of learning management system as alternative for curriculum delivery in South African higher education institutions during Covid-19 lockdown. International Journal of Higher Education, 11(1), pp.71-84.

Gandomi, A., & Haider, M. 2015. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2): 137-144.

Gantz, J., & Reinsel, D. 2012. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. IDC iView: IDC Analyze the future, 2007(2012): 1-16.

Gehman, J. 2008. Structuration theory summary. Retrieved April, 1, 2016.

Giddens, A. 1984. The construction of society. Cambridge: Polity.

Glaser, A.E., Harrison, J.P. and Josephs, D., 2022. Anomaly Detection Methods to Improve Supply Chain Data Quality and Operations. SMU Data Science Review, 6(1), p.3.

Gnip, P., Vokorokos, L. and Drotár, P., 2021. Selective oversampling approach for strongly imbalanced data. PeerJ Computer Science, 7, p.e604.

Gonzalez-Cuautle, D., Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, L.K., Portillo-Portillo, J., Olivares-Mercado, J., Perez-Meana, H.M. and Sandoval-Orozco, A.L., 2020. Synthetic minority oversampling technique for optimizing classification tasks in botnet and intrusion-detection-system datasets. Applied Sciences, 10(3), p.794.

Gironacci, I.M., 2021. Literature Review of Recommendation Systems. Transdisciplinary Perspectives on Risk Management and Cyber Intelligence, pp.119-129.

Guo, T., Bai, X., Tian, X., Firmin, S. and Xia, F., 2022. Educational anomaly analytics: features, methods, and challenges. Frontiers in big Data, 4, p.124.

He, H., Wang, Y., Qi, Y., Xu, Z., Li, Y. and Wang, Y., 2023. From Prediction to Design: Recent Advances in Machine Learning for the Study of 2D Materials. Nano Energy, p.108965.

Henard, F., & Mitterle, G. 2009. Quality guidelines in Higher Education" IMHE.

Heracleous, L. 2013. The employment of structuration theory in organizational discourse: Exploring methodological challenges. Management Communication Quarterly, 27(4), 599-606.

Hershkovitz, A., de Baker, R. S. J., Gobert, J., Wixon, M., & Pedro, M. S. 2013. Discovery with models: A case study on carelessness in computer-based science inquiry. American Behavioral Scientist, 57(10): 1480-1499.

Heymann, P., Bastiaens, E., Jansen, A., van Rosmalen, P. and Beausaert, S., 2022. A conceptual model of students' reflective practice for the development of employability competences, supported by an online learning platform. Education+ Training.

Hofacker, C.F., Malthouse, E.C. and Sultan, F., 2016. Big data and consumer behavior: Imminent opportunities. Journal of consumer marketing.

Hossain, M.A. and Islam, M.S., 2023. A novel hybrid feature selection and ensemble-based machine learning approach for botnet detection. Scientific Reports, 13(1), p.21207.

Huang, P.J., 2015. Classification of imbalanced data using synthetic over-sampling techniques. University of California, Los Angeles.

Huang, Q., Zheng, Z., Zhu, W., Fang, X., Fang, R. and Sun, W., 2022. Anomaly Detection Algorithm Based on Broad Learning System and Support Vector Domain Description. Mathematics, 10(18), p.3292.

Hussain, L., Lone, K.J., Awan, I.A., Abbasi, A.A. and Pirzada, J.U.R., 2022. Detecting congestive heart failure by extracting multimodal features with synthetic minority oversampling technique (SMOTE) for imbalanced data using robust machine learning techniques. Waves in Random and Complex Media, 32(3), pp.1079-1102.

Janiesch, C., Zschech, P. and Heinrich, K., 2021. Machine learning and deep learning. Electronic Markets, 31(3), pp.685-695.

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.

John, P., & Wheeler, S. 2012. The digital classroom: Harnessing technology for the future of learning and teaching. Routledge.

Joshi, N. 2017. 4 ways big data is transforming the education sector. https://www.linkedin.com/pulse/4-ways-big-data-transforming-education-sector-naveen-joshi/?originalSubdomain=es

Jones, M. R., & Karsten, H. 2008. Giddens's structuration theory and information systems research. MIS quarterly, 32(1): 127-157.

Hariri, R.H., Fredericks, E.M. and Bowers, K.M., 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. Journal of Big Data, 6(1), pp.1-16.

Henrys, K., 2021. Role of predictive analytics in business. Available at SSRN 3829621.

Holst, A., 2021. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2025. Statista, June.

Howlin, C.P. and Dziuban, C.D., 2019. Detecting Outlier Behaviors in Student Progress Trajectories Using a Repeated Fuzzy Clustering Approach. International Educational Data Mining Society.

Hussein, A.A., 2020. Fifty-Six Big Data V's Characteristics and Proposed Strategies to Overcome Security and Privacy Challenges (BD2). Journal of Information Security, 11(4), pp.304-328.

Isaacs, S., Roberts, N. and Spencer-Smith, G., 2019. Learning with mobile devices: A comparison of four mobile learning pilots in Africa. South African Journal of Education, 39(3).

Izzudin, M.I. and Judi, H.M., 2022. Personalised learning analytics promoting student's achievement and enhancing instructor's intervention in self-regulated meaningful learning. International Journal of Information and Education Technology, 12(11), pp.1243-1247.

Janiesch, C., Zschech, P. and Heinrich, K., 2021. Machine learning and deep learning. Electronic Markets, 31(3), pp.685-695.

Jobson, J.D., 1991. Multiple linear regression. In Applied multivariate data analysis (pp. 219-398). Springer, New York, NY.

Jovel, J. and Greiner, R., 2021. An Introduction to Machine Learning Approaches for Biomedical Research. Frontiers in Medicine, 8.

Kaliappan, J., Srinivasan, K., Mian Qaisar, S., Sundararajan, K. and Chang, C.Y., 2021. Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. Frontiers in Public Health, p.1319.

Kappen, T.H., van Klei, W.A., van Wolfswinkel, L., Kalkman, C.J., Vergouwe, Y. and Moons, K.G., 2018. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. Diagnostic and prognostic research, 2(1), pp.1-11.

Kayembe, C. and Nel, D., 2019. Challenges and opportunities for education in the Fourth Industrial Revolution. African Journal of Public Affairs, 11(3), pp.79-94.

Kayembe, C. and Nel, D., 2019. Challenges and opportunities for education in the Fourth Industrial Revolution. African Journal of Public Affairs, 11(3), pp.79-94.

Kern, C., Klausch, T. and Kreuter, F., 2019, April. Tree-based machine learning methods for survey research. In Survey research methods (Vol. 13, No. 1, p. 73). NIH Public Access.

Kersting, K., 2018. Machine learning and artificial intelligence: two fellow travelers on the quest for intelligent behavior in machines. Frontiers in big Data, 1, p.6.

Khan, M.A., Khan, R., Algarni, F., Kumar, I., Choudhary, A. and Srivastava, A., 2022. Performance evaluation of regression models for COVID-19: A statistical and predictive perspective. Ain Shams Engineering Journal, 13(2), p.101574.

Khattak, A. and Ahmad, A., (2018). Effects of Positive reinforcement on students' academic performance. 01. 220-225.

Khan, I., Ahmad, A.R., Jabeur, N. and Mahdi, M.N., 2021. An artificial intelligence approach to monitor student performance and devise preventive measures. Smart Learning Environments, 8(1), pp.1-18.

Khayi, N.A. and Rus, V., 2019. Clustering Students Based on Their Prior Knowledge. International Educational Data Mining Society.

Khodadadi, F., Dastjerdi, A.V. and Buyya, R., 2016. Internet of things: an overview. Internet of Things, pp.3-27.

Khor, E.T., 2019. Predictive models with machine learning algorithms to forecast students' performance.

Khosa, P., Pillay, R., & Dube, N. 2018. Inducting first-year social work students: Reflections on a discipline-specific approach to academic development. Social Work, 54(1): 111-132.

Khoza, S. B. and S. Manik. 2015. The recognition of 'digital technology refugees' amongst post
graduate students in a Higher Education institution. Alternation 17(2015): 190–208.

Khurana, D., Koli, A., Khatter, K. and Singh, S., 2022. Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications, pp.1-32.

Kibria, M.G., Nguyen, K., Villardi, G.P., Zhao, O., Ishizu, K. and Kojima, F., 2018. Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks. IEEE access, 6, pp.32328-32338.

Kindel, W.F., Christensen, E.D. and Zylberberg, J., 2019. Using deep learning to probe the neural code for images in primary visual cortex. Journal of vision, 19(4), pp.29-29.

King, J. and South, J., 2017. Reimagining the role of technology in higher education: A supplement to the national education technology plan. US Department of Education, Office of Educational Technology.

Kochetkov, O. T., & Prokhorov, I. V. 2017. The research of approaches of applying the results of big data analysis in higher education. In AIP Conference Proceedings(Vol. 1797, No. 1, p. 020008). AIP Publishing.

Koivu, A., Sairanen, M., Airola, A. and Pahikkala, T., 2020. Synthetic minority oversampling of vital statistics data with generative adversarial networks. Journal of the American Medical Informatics Association, 27(11), pp.1667-1674.

Kolowich, S. 2013. The new intelligence. Inside Higher Ed.

Lee, A. 1991. Integrating positivist and interpretivist approaches to organisational research, Organisation Science (2): pp 342- 365.

Kristoffersen, E., Aremu, O.O., Blomsma, F., Mikalef, P. and Li, J., 2019, September. Exploring the relationship between data science and circular economy: an enhanced CRISP-DM process model. In Conference on e-Business, e-Services and e-Society (pp. 177-189). Springer, Cham.

Kuhn, G., 2023. How Target Used Data Analytics to Predict Pregnancies. https://www.driveresearch.com/market-research-company-blog/how-target-used-data-analytics-to-predict-pregnancies/

Kumar, V. and Garg, M.L., 2018. Predictive analytics: a review of trends and techniques. International Journal of Computer Applications, 182(1), pp.31-37.

Kumar, V.P. and Sowmya, I., 2021. A Review on Pros and Cons of Machine Learning Algorithms.

Kumari, S., 2016. Impact of big data and social media on society. Global Journal for Research Analysis, 5, pp.437-438.

Kune, R., Konugurthi, P.K., Agarwal, A., Chillarige, R.R. and Buyya, R., 2016. The anatomy of big data computing. Software: Practice and Experience, 46(1), pp.79-105.

Lappeman, J., Clark, R., Evans, J., Sierra-Rubia, L. and Gordon, P., 2020. Studying social media sentiment using human validated analysis. MethodsX, 7, p.100867.

Lauría, E.J., 2021. Framing Early Alert of Struggling Students as an Anomaly Detection Problem: An Exploration. In CSEDU (1) (pp. 26-35).

Lavin, A., Gilligan-Lee, C.M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., Lange, D., Baydin, A.G., Sharma, A., Gibson, A. and Zheng, S., 2022. Technology readiness levels for machine learning systems. Nature Communications, 13(1), pp.1-19.

Lee, N.T., Resnick, P. and Barton, G., 2019. Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Brookings Institute: Washington, DC, USA.

Lee, T.H., Ullah, A. and Wang, R., 2020. Bootstrap aggregating and random forest. In Macroeconomic forecasting in the era of big data (pp. 389-429). Springer, Cham.

Lentz, G.S. and Foncha, J.W., 2021. Lecturer's Views on Using Blended Learning as an Intervention Programme for Teaching English Language Academic Writing to Cape Peninsula University of Technology (CPUT) First Year Students. Journal of English as an International Language, 16(1), pp.77-92.

Li, J. and Jiang, Y., 2021. The research trend of big data in education and the impact of teacher psychology on educational development during COVID-19: a systematic review and future perspective. Frontiers in Psychology, 12.

Li, M., 2020. To Build Less-Biased AI, Hire a More Diverse Team. Harvard Bus. Rev., Oct.

Li, Y., 2022. Similar Classification Algorithm for Educational and Teaching Knowledge Based on Machine Learning. Wireless Communications and Mobile Computing, 2022.

Lim, K.S., Lee, L.H. and Sim, Y.W., 2021. A review of machine learning algorithms for fraud detection in credit card transaction. International Journal of Computer Science & Network Security, 21(9), pp.31-40.

Limbu, S.H., 2020. Direct Speech to Speech Translation Using Machine Learning.

Lin, M.L., 2020. Educational Upward Mobility. Practices of Social Changes-Research on Social Mobility and Educational Inequality.
Int'l J. Soc. Sci. Stud., 8, p.25.

Little, M.A., Varoquaux, G., Saeb, S., Lonini, L., Jayaraman, A., Mohr, D.C. and Kording, K.P., 2017. Using and understanding cross-validation strategies. Perspectives on Saeb et al. GigaScience, 6(5), pp.1-6.

Lloyd, J., 2011. Identifying key components of business intelligence systems and their role in managerial decision making.

Lohr, S., 2018. Facial recognition is accurate, if you're a white guy. In Ethics of Data and Analytics (pp. 143-147). Auerbach Publications.

Lu, D. and Yan, L., 2021. Face detection and recognition algorithm in digital image based on computer vision sensor. Journal of Sensors, 2021.

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S.J., Ogata, H., Baltes, J., Guerra, R., Li, P. and Tsai, C.C., 2020. Challenges and future directions of big data and artificial intelligence in education. Frontiers in psychology, 11, p.580820.

Madahana, M., Khoza-Shangase, K., Moroe, N., Mayombo, D., Nyandoro, O. and Ekoru, J., 2022. A proposed artificial intelligence-based real-time speech-to-text to sign language translator for South African official languages for the COVID-19 era and beyond: In pursuit of solutions for the hearing impaired. South African Journal of Communication Disorders, 69(2), p.915.

Majaj, N.J. and Pelli, D.G., 2018. Deep learning—Using machine learning to study biological vision. Journal of vision, 18(13), pp.2-2.

Maley, L.B., 2020. Teaming at a Distance: The Work Experience on Global Virtual Teams (Doctoral dissertation, Antioch University).

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Hung Byers, A. 2011. The next frontier for innovation, competition, and productivity. McKinsey Global Institute Report.

Marr, B. 2016. Big Data and the Evolution of Education. Accessed online: http://data-informed.com/big-data-and-evolution-education/

Maré, S. and Mutezo, A.T., 2021. The effectiveness of e-tutoring in an open and distance e-learning environment: evidence from the university of south africa. Open Learning: The Journal of Open, Distance and e-Learning, 36(2), pp.164-180.

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M.J. and Flach, P., 2019. CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), pp.3048-3061.

Mashau, P. and Nyawo, J.C., 2021. The use of an online learning platform: a step towards e-learning. South African Journal of Higher Education, 35(2), pp.123-143.

Matto, G., 2022. Big Data Analytics Framework for Effective Higher Education Institutions. Tanzania Journal of Engineering and Technology, 41(1), pp.10-18.

Maurya, L.S., Hussain, M.S. and Singh, S., 2021. Developing classifiers through machine learning algorithms for Student placement prediction based on academic performance. Applied Artificial Intelligence, 35(6), pp.403-420.

Mayisela, T., 2013. The potential use of mobile technology: Enhancing accessibility and communication in a blended learning course. South African Journal of Education, 33(1), pp.1-18.

McNaught, C., 2008. Information literacy in the 21st century. In Encyclopedia of information technology curriculum integration (pp. 406-412). IGI Global.

Menon, A., 2021. The impact of Artificial Intelligence (AI) and Engines on Boardgames (Chess and Go).

Mienye, I.D., Sun, Y. and Wang, Z., 2019. Prediction performance of improved decision tree-based algorithms: a review. Procedia Manufacturing, 35, pp.698-703.

Mlambo, S., Rambe, P. and Schlebusch, L., 2020. Effects of Gauteng province’s educators’ ICT self-efficacy on their pedagogical use of ICTS in classrooms. Heliyon, 6(4), p.e03730.

Mlambo, V.H., Mlambo, D.N. and Adetiba, T.C., 2021. Expansion of higher education in South Africa: problems and possibilities.
J. Soc. Soc. Anthropol, 12, pp.30-40.

Mohamed Nafuri, A.F., Sani, N.S., Zainudin, N.F.A., Rahman, A.H.A. and Aliff, M., 2022. Clustering Analysis for Classifying Student Academic Performance in Higher Education. Applied Sciences, 12(19), p.9467.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J. and Fernández-Leal, Á., 2022. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, pp.1-50.

Mounika, B. and Persis, V., 2019. A Comparative Study of Machine Learning Algorithms for Student Academic Performance. International Journal of Computer Sciences and Engineering, 7(4), pp.721-725.

Müller, O., Junglas, I., Brocke, J.V. and Debortoli, S., 2016. Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Information Systems, 25(4), pp.289-302.

Murumba, J. and Micheni, E., 2017. Big data analytics in higher education: a review. The International Journal of Engineering and Science, 6(06), pp.14-21.

Nagesh, A.S. and Satyamurty, C.V.S., 2018. Application of clustering algorithm for analysis of student academic performance. Int. J. Comput. Sci. Eng, 6(1), pp.381-384.

Najibi, A., 2020. Racial discrimination in face recognition technology. Harvard Online: Science Policy and Social Justice, 24.

Ng'ambi, D., Brown, C., Bozalek, V., Gachago, D., & Wood, D. 2016. Technology enhanced teaching and learning in South African higher education–A rearview of a 20 year journey. British Journal of Educational Technology, 47(5): 843-858.

Ngalo-Morrison, L. 2017. Factors influencing the academic attainment of undergraduate sponsored students at the University of the Western Cape: a strength-based approach (Doctoral dissertation, University of the Western Cape).

Nichols, J.A., Herbert Chan, H.W. and Baker, M.A., 2019. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophysical reviews, 11(1), pp.111-118.

Ningrum, R. K. and N. W. D. Ekayani. 2019. "Predictive Value of Entrance Test with the Academic Achievement of Medical Students." Journal of Physics: Conference Series, 1402: 022068. IOP Publishing.

Nyamupangedengu, E., 2017. Investigating factors that impact the success of students in a Higher Education classroom: a case study. Journal of Education (University of KwaZulu-Natal), (68), pp.113-130.

Ogunleye, J.O., 2022. Predictive Data Analysis Using Linear Regression and Random Forest.

Omuya, E.O., Okeyo, G. and Kimwele, M., 2022. Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. Engineering Reports, p.e12579.

Oraison, H., Konjarski, L. and Howe, S., 2019. Does university prepare students for employment?: Alignment between graduate attributes, accreditation requirements and industry employability criteria. Journal of Teaching and Learning for Graduate Employability, 10(1), pp.173-194.

Orlikowski, W.J., 1992. The duality of technology: Rethinking the concept of technology in organizations. Organization science, 3(3), pp.398-427.

Orlikowski, W. J. 2000. Using technology and constituting structures: A practice lens for studying technology in organizations. Organization science, 11(4): 404-428.

Osho, O., Musa, F.A., Misra, S., Uduimoh, A.A., Adewunmi, A. and Ahuja, R., 2019, October. AbsoluteSecure: a tri-layered data security system. In International Conference on Information and Software Technologies (pp. 243-255). Springer, Cham.

Oza, A., 2018. Fraud detection using machine learning. TRANSFER, 528812(4097), p.532909.

Pan, L., Patterson, N., McKenzie, S., Rajasegarar, S., Wood-Bradley, G., Rough, J., Luo, W., Lanham, E. and Coldwell-Neilson, J., 2020. Gathering Intelligence on Student Information Behavior Using Data Mining. Library Trends, 68(4), pp.636-658.

Papamitsiou, Z., & Economides, A. A. 2014. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4).

Pedroso, John Erwin & Celestial, Joemari & Cataluña, Jonel. (2022). Social Media Platforms as Enhancement Tools of Academic Performance in the Midst of COVID-19 Pandemic. 3. 2785-2793.

Petrova-Dimitrova, V., 2022, September. Reinforcement learning algorithms using for agent behaviour modelling and researching. In AIP Conference Proceedings (Vol. 2449, No. 1, p. 040009). AIP Publishing LLC.

Picciano, A. G. 2012. The evolution of big data and learning analytics in American higher education. Journal of Asynchronous Learning Networks, 16(3): 9-20.

Pillay, R. 2017. Crafting a meso practice course using elements of authentic learning for undergraduate social work students in South Africa (Doctoral dissertation, University of the Western Cape).

Porter, S., 2015. What Are MOOCs. To MOOC or Not to MOOC, pp.1-7.

Portugal, I., Alencar, P. and Cowan, D., 2018. The use of machine learning algorithms in recommender systems: A systematic review. Expert Systems with Applications, 97, pp.205-227.

Prusty, S., Patnaik, S. and Dash, S.K., 2022. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. Frontiers in Nanotechnology, 4, p.972421.

Pugliese, R., Regondi, S. and Marini, R., 2021. Machine learning-based approach: Global trends, research directions, and regulatory standpoints. Data Science and Management, 4, pp.19-29.

Prinsloo, P. and Roberts, J., 2022. Analysis of Higher Education (HE) Systems' Approach in South Africa. (Open) Educational Resources around the World.

Rabella, M.F. 2016. How does big data impact education? Accessed online at: http://oecdinsights.org/2016/11/07/how-does-big-data-impact-education/ [21st of January 2019].

Raghupathi, W., & Raghupathi, V. 2014. Big data analytics in healthcare: promise and potential. Health information science and systems, 2(1): 3.

Rahi, S. 2017. Research Design and Methods: A Systematic Review of Research Paradigms, Sampling Issues and Instruments Development. International Journal of Economics & Management Sciences, 6(2): 1-5.

Rahmani, A.M., Azhir, E., Ali, S., Mohammadi, M., Ahmed, O.H., Ghafour, M.Y., Ahmed, S.H. and Hosseinzadeh, M., 2021. Artificial intelligence approaches and mechanisms for big data analytics: a systematic study. PeerJ Computer Science, 7, p.e488.

Rajak, A., Shrivastava, A.K. and Vidushi, 2020. Applying and comparing machine learning classification algorithms for predicting the results of students. Journal of Discrete Mathematical Sciences and Cryptography, 23(2), pp.419-427.

Ramachandra, M.N., Srinivasa Rao, M., Lai, W.C., Parameshachari, B.D., Ananda Babu, J. and Hemalatha, K.L., 2022. An Efficient and Secure Big Data Storage in Cloud Environment by Using Triple Data Encryption Standard. Big Data and Cognitive Computing, 6(4), p.101.

Rains, S.A. and Bonito, J.A., 2017. Adaptive structuration theory. The international encyclopedia of organizational communication, pp.1-9.

Ratih, I.D., Retnaningsih, S.M., Islahulhaq, I. and Dewi, V.M., 2022, October. Synthetic minority over-sampling technique nominal continous logistic regression for imbalanced data. In AIP Conference Proceedings (Vol. 2668, No. 1, p. 070021). AIP Publishing LLC.

Rao, C.R. and Gudivada, V.N., 2018. Computational analysis and understanding of natural languages: principles, methods and applications. Elsevier.

Rawat, S., Kumar, D., Khattri, C. and Kumar, P., 2021. Machine Learning Classification Algorithms for Systematic Analysis to Understand Learners Drop out of MOOCs courses.

Refaeilzadeh, P., Tang, L. and Liu, H., 2009. Cross-validation. Encyclopedia of database systems, 5, pp.532-538.

Reschly, A. L., & Christenson, S. L. 2006. Prediction of dropout among students with mild disabilities: A case for the inclusion of student engagement variables. Remedial and Special Education, 27(5): 276-292.

Reyes, J. A. 2015. The skinny on big data in education: Learning analytics simplified. TechTrends, 59(2): 75-80.

Rodriguez, M.Z., Comin, C.H., Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D.F. and Rodrigues, F.A., 2019. Clustering algorithms: A comparative approach. PloS one, 14(1), p.e0210236.

Rohit, W. (2017, April 24). Handling imbalanced dataset in supervised learning using family of SMOTE algorithm. Data Science Central. https://www.datasciencecentral.com/handling-imbalanced-data-sets-in-supervised-learning-using-family/

Rokach, L. and Maimon, O., 2005. Decision trees. In Data mining and knowledge discovery handbook (pp. 165-192). Springer, Boston, MA.

Rolf, B., Jackson, I., Müller, M., Lang, S., Reggelin, T. and Ivanov, D., 2022. A review on reinforcement learning algorithms and applications in supply chain management. International Journal of Production Research, pp.1-29.

Romero, C., & Ventura, S. 2010. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6): 601-618.

Romero, C. and Ventura, S., 2020. Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), p.e1355.

Rose, J., & Scheepers, R. (2001). Structuration theory and information system development-frameworks for practice. ECIS 2001 Proceedings, 80.

Sarker, I.H., 2021. Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. SN Computer Science, 2(5), p.377.

Rustagi, M. and Goel, N., 2022. Predictive Analytics: A study of its Advantages and Applications. IARS'International Research Journal, 12(01), pp.60-63.
Sah, S., 2020. Machine learning: a review of learning types.

Salami, H.O., Ibrahim, R.S. and Yahaya, M.O., 2016. Detecting Anomalies in Students' Results Using Decision Trees. International Journal of Modern Education & Computer Science, 8(7).

Sankara Subbu, R., 2017. Brief Study of Classification Algorithms in Machine Learning.

Sauber-Cole, R. and Khoshgoftaar, T.M., 2022. The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey. Journal of Big Data, 9(1), p.98.

Savov, V., 2011. Visualized: a zettabyte, https://www.engadget.com/2011-06-29-visualized-a-zettabyte.html Accessed [11th of August 2022]

Sarica, A., Cerasa, A. and Quattrone, A., 2017. Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in aging neuroscience, 9, p.329.

Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

Sarker, I.H., Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN Comput Sci 2021; 2: 420.

Seify, M., Sepehri, M., Hosseini-far, A. and Darvish, A., 2022. Fraud Detection in Supply Chain with Machine Learning. IFAC-PapersOnLine, 55(10), pp.406-411.

Sestino, A., Prete, M.I., Piper, L. and Guido, G., 2020. Internet of Things and Big Data as enablers for business digitalization strategies. Technovation, 98, p.102173.

Seyedan, M. and Mafakheri, F., 2020. Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. Journal of Big Data, 7(1), pp.1-22.

Sharma, M.2017. Implementation of Big data analytics in Education Industry. Journal of Electrical and Computer Engineering, 19: 36-39. 10.9790/0661-1906033639.

Shmilovici, A., 2009. Support vector machines. In Data mining and knowledge discovery handbook (pp. 231-247). Springer, Boston, MA.

Siemens, G., & Long, P. 2011. Penetrating the fog: Analytics in learning and education. EDUCAUSE review, 46(5): 30.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. American Behavioral Scientist, 57(10): 1380-1400.

Silhavy, R., Silhavy, P. and Prokopova, Z., 2017. Analysis and selection of a regression model for the use case points method using a stepwise approach. Journal of Systems and Software, 125, pp.1-14.

Singh, V., Chen, S.S., Singhania, M., Nanavati, B. and Gupta, A., 2022. How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries–A review and research agenda. International Journal of Information Management Data Insights, 2(2), p.100094.

Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. Journal of business research, 70, pp.263-286.

Smaya, H., 2022. The Influence of Big Data Analytics in the Industry. Open Access Library Journal, 9(2), pp.1-12.

Schmidt, J., Marques, M.R., Botti, S. and Marques, M.A., 2019. Recent advances and applications of machine learning in solid-state materials science. npj Computational Materials, 5(1), pp.1-36.

Soltanzadeh, P. and Hashemzadeh, M., 2021. RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. Information Sciences, 542, pp.92-111.

Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.

Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. and Abreu, R., 2015. A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. International Educational Data Mining Society.

Strauss, A., & Corbin, J. 1994. Grounded theory methodology. Handbook of qualitative research, 17: 273-85.

Strydom, F. & Loots, S. 2018. Data analytics as key to student success: Improving student success and completion in higher-education institutions through gathering, analysing and organising data. Available Online at: https://www.sowetanlive.co.za/news/south-africa/2018-08-17-data-analytics-as-key-to-student-success/

Sun, Y., Ren, Z. and Zheng, W., 2022. Research on Face Recognition Algorithm Based on Image Processing. Computational Intelligence and Neuroscience, 2022.

Sun, A.Y. and Scanlon, B.R., 2019. How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7), p.073001.

Swartz, R., Ivancheva, M., Czerniewicz, L. et al. Between a rock and a hard place: dilemmas regarding the purpose of public universities in South Africa. High Educ 77, 567–583 (2019). https://doi.org/10.1007/s10734-018-0291-9

Tan, H., 2021, August. Machine Learning Algorithm for Classification. In Journal of Physics: Conference Series (Vol. 1994, No. 1, p. 012016). IOP Publishing.

Tan, Y., 2022. Application Research on Face Image Evaluation Algorithm of Deep Learning Mobile Terminal for Student Check-In Management. Computational Intelligence and Neuroscience, 2022.

Tao, H., Niu, X., Fu, L., Yuan, S., Wang, X., Zhang, J. and Hu, Y., 2022. DeepRS: A Library of Recommendation Algorithms Based on Deep Learning. International Journal of Computational Intelligence Systems, 15(1), pp.1-12.

Tasmin, R., Muhammad, R.N. and Aziati, A.N., 2020, September. Big Data Analytics Applicability in Higher Learning Educational System. In IOP Conference Series: Materials Science and Engineering (Vol. 917, No. 1, p. 012064). IOP Publishing.

Temraz, M. and Keane, M.T., 2022. Solving the class imbalance problem using a counterfactual method for data augmentation. Machine Learning with Applications, 9, p.100375.

Tennenholtz, G., Zahavy, T. and Mannor, S., 2018. Train on validation: squeezing the data lemon. arXiv preprint arXiv:1802.05846.

Tercan, H. and Meisen, T., 2022. Machine learning and deep learning based predictive quality in manufacturing: a systematic review. Journal of Intelligent Manufacturing, pp.1-27.

Tewari, D.D. and Ilesanmi, K.D., 2020. Teaching and learning interaction in South Africa’s higher education: Some weak links. Cogent Social Sciences, 6(1), p.1740519.

Thudumu, S., Branch, P., Jin, J. and Singh, J.J., 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. Journal of Big Data, 7(1), pp.1-30.

Togelius, J., 2019. Playing smart: On games, intelligence, and artificial intelligence. MIT Press.

Trunfio, T.A., Scala, A., Giglio, C., Rossi, G., Borrelli, A., Romano, M. and Improta, G., 2022. Multiple regression model to analyze the total LOS for patients undergoing laparoscopic appendectomy. BMC Medical Informatics and Decision Making, 22(1), pp.1-8.

Tseng, H.H., Wei, L., Cui, S., Luo, Y., Ten Haken, R.K. and El Naqa, I., 2020. Machine learning and imaging informatics in oncology. Oncology, 98(6), pp.344-362.

Twum-Darko, M., 2014. Sustainable local economic development: the role of informatics in determining municipal revenue management. Journal of Economics and Behavioral Studies, 6(6), pp.466-476.

Vashisht, V., Pandey, A.K. and Yadav, S.P., 2021. Speech recognition using machine learning. IEIE Transactions on Smart Processing & Computing, 10(3), pp.233-239.

van den Berg, C. L. 2017. A framework to teach digital innovation skills to South African Information Systems students.

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.

Venkatasubramaniam, A., Wolfson, J., Mitchell, N., Barnes, T., JaKa, M. and French, S., 2017. Decision trees in epidemiological research. Emerging themes in epidemiology, 14(1), pp.1-12.

Wach, M. and Chomiak-Orsa, I., 2021. The application of predictive analysis in decision-making processes on the example of mining company's investment projects. Procedia Computer Science, 192, pp.5058-5066.

Wang, S., Dai, Y., Shen, J. and Xuan, J., 2021. Research on expansion and classification of imbalanced data based on SMOTE algorithm. Scientific Reports, 11(1), pp.1-11.

Walker, S., 2021. Machine Learning and Corporate Fraud Detection. University of California, Berkeley.

Webb, H. W., & LeRouge, C. 2009. Modeling ERP Academic Deployment via Adaptive Structuration Theory. In Encyclopedia of Information Science and Technology, Second Edition (pp. 2638-2645). IGI Global.

Wells, C., 2016. Maryland Universities to Use Data to Predict Student Success-or Failure. The Baltimore Sun, 11.

Wells, L. and Bednarz, T., 2021. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. Frontiers in artificial intelligence, 4, p.550030.

West, D.M. and Allen, J.R., 2018. How artificial intelligence is transforming the world. Report. April, 24, p.2018.

Westphal, M. and Brannath, W., 2020. Evaluation of multiple prediction models: A novel view on model selection and performance assessment. Statistical Methods in Medical Research, 29(6), pp.1728-1745.

White, M., Becker, J. and du Plessis, M., 2021. Unintended positive consequences of development centres in university graduates.
Frontiers in Psychology, 12, p.5421.

Winham, S.J., Slater, A.J. and Motsinger-Reif, A.A., 2010. A comparison of internal validation techniques for multifactor dimensionality reduction. Bmc Bioinformatics, 11(1), pp.1-16.

Wong, W. and C. Hinnant, C., 2022. Competing perspectives on the Big Data revolution: a typology of applications in public policy. Journal of Economic Policy Reform, pp.1-15.

Xie, Y., Zhang, K., Kou, H. and Mokarram, M.J., 2022. Private anomaly detection of student health conditions based on wearable sensors in mobile cloud computing. Journal of Cloud Computing, 11(1), pp.1-12.

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.W. and Qiu, J., 2021. Artificial intelligence: A powerful paradigm for scientific research. The Innovation, 2(4), p.100179.

Yayla, R., Yayla, H.N., Ortaç, G. and Bilgin, T.T., 2021. A classification approach with machine learning methods for technical problems of distance education: Turkey example. Open Praxis, 13(3), pp.312-322.

Yi, X., Xu, Y., Hu, Q., Krishnamoorthy, S., Li, W. and Tang, Z., 2022. ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection. Complex & Intelligent Systems, pp.1-26.

Yoshihara, K. and Takahashi, K., 2022. A simple method for unsupervised anomaly detection: An application to Web time series data. PloS one, 17(1), p.e0262463.

Xu, D. and Tian, Y., 2015. A comprehensive survey of clustering algorithms. Annals of Data Science, 2(2), pp.165-193.

Xu, Y. and Goodacre, R., 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of analysis and testing, 2(3), pp.249-262.

Zhang, C., 2022. Research on Literature Clustering Algorithm for Massive Scientific and Technical Literature Query Service. Computational Intelligence and Neuroscience, 2022.

Zhang, Y., 2012, September. Support vector machine classification algorithm and its application. In International conference on information computing and applications (pp. 179-186). Springer, Berlin, Heidelberg.

Zhang, Z., 2020. Predictive analytics in the era of big data: opportunities and challenges. Annals of translational medicine, 8(4).

Zheng, L., 2022. Innovative Elements Gathering Scheme For Students With Communication Adaptation Barriers Under The Guidance Of Entrepreneurial Innovation Spirit. Psychiatria Danubina, 34(suppl 4), pp.650-650.

Zimmer, M., Viappiani, P. and Weng, P., 2014, May. Teacher-student framework: a reinforcement learning approach. In AAMAS Workshop Autonomous Robots and Multirobot Systems.

Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V., 2017. Machine learning on big data: Opportunities and challenges. Neurocomputing, 237, pp.350-361.

Zinshteyn, M., 2016. The colleges are watching. The Atlantic.