# IMPROVING A STUDENT'S ACADEMIC PERFORMANCE: THE APPLICATION OF MACHINE LEARNING AND BIG DATA ANALYTICS


**Abstract**

The paper explores the utilisation of machine learning algorithms and big data analytics in enhancing academic and psycho-socio performance among students. It suggests that South African higher education can benefit from appropriate technology and data application to address issues like lack of personalised support, varied student readiness, and resource constraints affecting performance. Employing adaptive structuration theory (AST), derived from Giddens' structuration theory, the study examines how institutions incorporate new technologies to adaptively improve student outcomes. Through a thorough literature review encompassing student performance factors, teaching methodologies, technology integration, and big data analytics (BDA), the research proposes establishing an IT infrastructure capable of integrating diverse student data types for machine learning analysis. This data-driven approach aims to personalise curricula, identify at-risk students, and enhance pedagogy to bolster learning outcomes. By bridging technological capabilities with practical implementation, the study offers a framework for local universities to make informed, data-driven decisions tailored to their challenges. It underscores the potential for data analytics to create supportive, personalised learning environments conducive to student success within the South African higher education context.



**Keywords:**Machine learning algorithms; Big data analytics; Student performance; Personalised support; Adaptive structuration theory (AST)




## 1. Introduction

Education is a fundamental global building block for individual and societal development (Brende, 2015). Pedros et al. (2022) highlight that "education enables upward socioeconomic mobility" and is critical 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 completing their studies (Mlambo et al., 2021). Specifically, inefficiencies have been widely 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 (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 (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 (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 (Badaru&Adu, 2022). However, research signals that these technologies have yielded mixed outcomes, often failing to fulfil their primary objective of strengthening teaching and learning to boost student performance as anticipated (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). Chaurasia et al. (2018) position BDA as an accelerator of disruptive innovation in higher education. Reyes (2015) summarises that 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 needs to be explored and underutilised within higher education in South Africa (Gamede, 2022). Existing research signals that South African IHL generates expansive stores of student data from platforms like LMS. Still, more analytics is needed to extract value from these assets to strengthen teaching and propel performance optimisation efforts (Reyes, 2015). As such, an opportunity exists to address this gap by examining how BDA can strategically be leveraged using locally generated higher education datasets to model and predict student academic performance.
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 research objective was to examine the current use of software to improve students' performance, and the question was, "How is software currently used to improve students' performances?"

## 2. Literature review

Access to higher education has been made more accessible in the post-apartheid era in South Africa (Ng'ambi et al., 2016). However, student performance still needs to reflect high academic success. Indeed, the Higher Education Department reported that six per cent of South Africans have university degrees. This translates to just over 1.7 million people in the country holding degrees. South Africa's degree accomplishment is trailing several other countries categorised as middle-income, although the country is fairer 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 is a source of concern for policymakers and educators.

The issue with low performance could be attributed to a wide range of students needing personalised support, owing to different readiness levels and the under-preparedness of students entering IHL (Van Den Berg, 2017). Whether from disadvantaged or advantaged backgrounds, skills like information literacy, academic writing, critical thinking, and language skills were 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 is divided between internal and external factors. Internal factors refer to class size, facilities, technologies used in the class, teacher's role, and class environment. In South Africa, some of the following factors have been summarised as follows:

- **Inadequate funding**: More funding is needed for higher education in South Africa. This leads to insufficient resourcing and student support, impacting their performance (Prinsloo & Roberts, 2022).
- **Inadequate support for students**: Many students in South Africa need help to afford the costs of higher education, including tuition fees, accommodation, and living expenses. The inadequacy leads to financial strain and stress, negatively impacting their performance (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 rural areas. The lack of access impacts their ability to succeed in higher education (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 various factors, including differences in access to education and support and socioeconomic and cultural differences (Prinsloo & Roberts, 2022).

In recent years, several efforts have been made to improve the performance of higher education students in South Africa. 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 student support: The government and universities have also implemented initiatives to improve student support, including tutoring, counselling 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. However, drawing from Ayuk & Koma (2019), efforts are ongoing to address these disparities, including initiatives to increase access to education for historically disadvantaged groups, such as black students and students from rural areas increased efforts are required.
- **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 to improve students' performance. IHL has used several ways to improve students' performance, including using software to analyse big data on students. Institutions of higher learning worldwide have set objectives to improve students' performance by enhancing teaching and learning. Due to the proliferation of software in the marketplace, Information communication technology (ICT) has also been used in academic environments. Indeed, in an era of fast-paced developing technologies, IHL has also incorporated ICT to keep up with market changes. The results are varied.

Some Studies have shown a strong correlation between ICT use and students' motivation and academic performance. At the same time, other studies show little to no correlation between ICT tools and students' performance (Mlambo et al., 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 believe technology is disruptive, and universities still need to cope. Big Data Analytics is emerging, and IHL worldwide is reaping the benefits. IHL can improve students' performance by customising the curriculum for one student at a time. With regards to this research, ICT refers to Big Data analytics.

**3. Overview of Underpinning Theory and Adaptive Structuration Theory (AST)**

The theory of adaptive structuration that underpins this research is discussed. Rose & Scheepers (2001) explain that theory is used in Information communication technology (ICT) because it can assist in determining the interaction between technologies and organisations, 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.

DeSanctis and Poole reformed Structuration Theory into Adaptive Structuration Theory 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 "the production and reproduction of the social systems through members' use of rules and resources in interaction". AST is often used to study information and communication technologies (ICTs) in organisations and society. AST suggests that ICTs are not neutral tools; instead, they are shaped by and shape the social structures and human agency within which they are used. AST is applied to various social and organisational phenomena, including social media, the adoption and diffusion of innovations, and the implementation and management of information systems. It is a valuable 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 organisations. AST examines the transformation process from the perspectives of "types of structures provided by the advanced technologies and the structures that emerge in human action as people interact with these technologies". 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.

**4. Finding and Analysis**

One of the main goals of Institutions of higher learning in South Africa is to improve students' performances through enhanced teaching and learning. However, inadequate mining or analytics of Big

Data has led to insufficient information to provide teaching and learning structures that will empower and improve students' full potential (Reyes, 2015). Many South African institutions make use of various approaches and methods in an attempt to achieve this goal. Some of the approaches include using software and analysing data (Prinsloo, 2022). Despite these methods and approaches, including the use of software, challenges persist in the attempts to improve students' performances. The challenges persist because the software is incorrectly applied relative to the available data or the selected software is unsuitable, thereby giving results that do not help improve students' performances (Prinsloo, 2022). 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 rather than the factors that influence the results. Thus, it is critical to employ a framework encompassing aspects of influence, enabling forecast, and ensuring consistency towards performance improvement.

### 4.1 The application of ML algorithm to the problem

#### 4.1.1 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 5.1: 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 is both quantitative and qualitative. As machine learning does not understand text, this text will eventually be transformed into numbers before it is fed to an algorithm. The following provides the meaning of some of the features:

- Married: If a student is married or single.
- Home-to-school travel time: A student's time to reach home.
- Student's home address: If a student lives in a city or town.
- Living status: If a student lives with their 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 Final Mark (FM) is the target label, which is the final grade of the student that the algorithm is attempting to predict. The algorithm is fed data about students and predicts the final mark based on entered data. This will help the lecturer to predict and identify students at risk of failing.

### 4.1.2 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 are 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’’.
- Drop the columns that have one value only because these types of columns do not improve the model building. Removing any such columns will thus reduce the number of features (columns). The model shows that eight columns containing unique values were dropped: ("Student's guardian", "Married", "Qual", and" year").

#### (b) Null values in each column validation

After removing the columns containing unique values, missing values in each column was investigated. Figuring out how many null values exist in each column has some benefits.

- Any machine learning model will not accept null (nan) values. It will only accept numbers. Hence, any column with null value will raise an error.
- Also, if a column has too many null values, it is best practice to delete that column because no accurate prediction can be made to replace them accurately.
- 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, which occurs most of the time compared to other values.

#### (c) Checking the data type of each column

Checking each column's datatype will help identify non-numeric columns and transform them into numeric ones to be fed into the model. Again, machine learning models only accept numerical values, so non-numerical values must be transformed. Most columns appear to have an object (string) datatype; some columns appear 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 column's names were renamed into a more readable format. After cleaning the data, the final shape shows that all the columns have numbers.

### (d) Exploratory data analysis (EDA) and the distribution of the target column (FM)

In this step, visualisation was used on some columns to see their distribution and gain insights about the data.Exploratory Data Analysis (EDA) was a crucial step in the machine learning process, especially when predicting student performance. Understanding the distribution of the target variable (in this case, the FM column representing student scores) and other relevant features provided valuable insights for model building and selection.

Histograms was used in EDA to examine the distribution of numerical variables. By plotting histograms, insights into the frequency distribution of student scores and other relevant factors such as marital status was gained.In the Figure 5.1 below, after plotting the histogram of the (FM) column, most students' scores are greater than 40 and less than 80. Thisindicates a distribution with a peak around the mid-range scores.  This suggests that most students fall within this score range, providing a baseline understanding of the performance distribution.

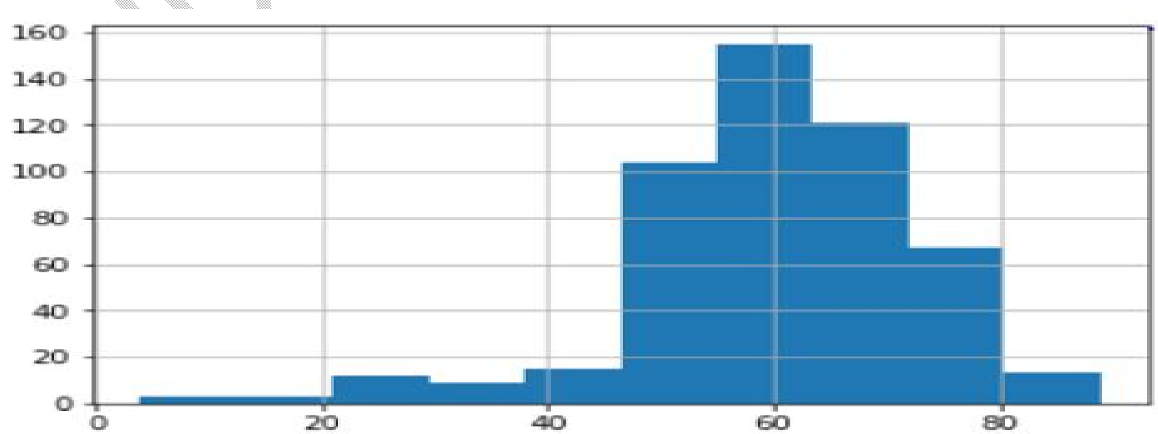

*Figure 5.1: Histogram of the Final Mark*

In Figure 5.2 below illustrates that most students are unmarried and as such married students are less in number than single ones.

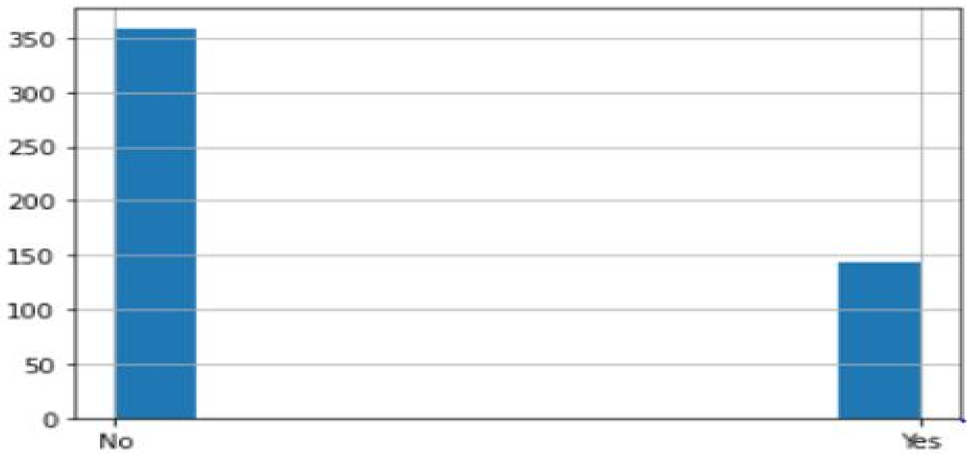


*Figure 5.2:  Married Vs. Single*

From the Figures 5.3 and Figure 5.4 below, student's scores lie between 50 to 75. The average score of married students is 60.69. The maximum score obtained by a married student is 88.0. The average score of Single students is 59.738. The maximum score obtained by a Single student is 89.0.Theseillustrate that unmarried students outnumber married students, indicating that marital status is a relevant factor to consider in predicting student performance. This insight highlights potential demographic influences on academic achievement.

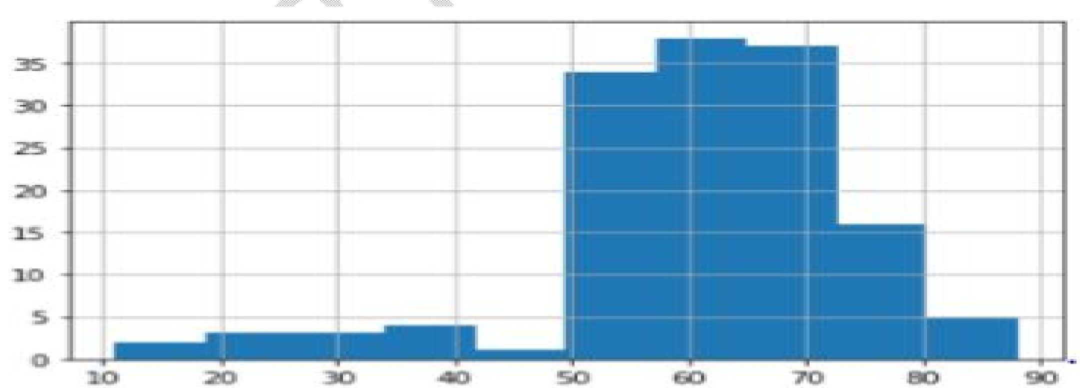


*Figure 5.3: Distribution of Marks (married)*

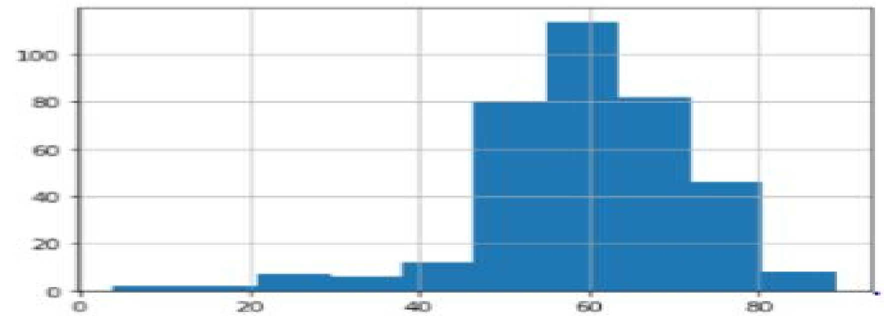


*Figure 5.4: Distribution of Marks (single)*

These insights obtained from exploratory data analysis, particularly through histogram visualisation and distribution analysis, provide valuable information for machine learning model development. Understanding the distribution of student performance relative to demographic factors like marital status can help in feature selection, model interpretation, and ultimately improve the accuracy of predictions regarding student outcomes.

Cross-validation is a valuable technique for assessing the performance of an ML model. It offers a more robust estimate of the model performance. Cross-validation helps the model be trained and evaluated on multiple different data splits. Finally, cross-validation is also helpful in preventing overfitting, as the model is not trained and assessed on the same data.Cross-validation is useful in cases where the designer does not desire to use a validation set or there needs to be more 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.

**6.5 Recommendations and Future Research**

To use BDA to predict students' performance,universities need to be educated on the different use cases of BDA in education. Also, it is imperative to identify potential barriers to IHL's adoption of BDA. Once that is determined, a strategy can be implemented to address data analytics.Rose & Scheepers (2001) explain that theory is used in Information communication technology (ICT) because it can assist in determining the interaction between technologies and organisations, 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. IHL also dealt with privacy issues regarding the data; further work could focus on ensuring the students' data's security, confidentiality, and privacy protection. This will thus assist in streamlining all the delays and difficulties encountered in collecting data. The scope of this study was focused on one cohort, MBIS. Further research could also take data from undergraduates across different qualifications and cohorts for a comparative study.Finally, additional research is also recommended to extend system abilities to predict individual students' grades, monitor students' learning patterns, and suggest intervention methods applicable to stakeholders.

## 7. 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 needs to achieve its objectives. This has caused IHL to look at other alternatives, such as LMS, which have produced mixed results. Across the world, other IHLs have now started looking at different technologies, such as BDA, to improve students' performance. Institutions of higher learning globally have set objectives to enhance students' performance by improving teaching and learning. Due to the proliferation of software in the marketplace, Information communication technology (ICT) has also been used in academic environments. Indeed, in an era of fast-paced developing technologies, IHL has also incorporated ICT to keep up with market changes. Across the world, ICT tools such as big data analytics have been adopted and used to improve students' performances through enhanced teaching and learning. Indeed, educational organisations explore using big data to match students' careers and goals with their studies, conduct evidence-based planning, and improve students' efforts. Also, Big data Analytics can assist in discovering students' learning patterns, which can be used to create teaching and learning programs specific to individuals. In South Africa, there is potential for enhancing IHL decision-making through Big Data analytics, yet various obstacles hinder this progress.



## References

[1] 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.

[2] Brenda, 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/

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

[4] 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.

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] Prinsloo, P. and Roberts, J., 2022. Analysis of Higher Education (HE) Systems' Approach in South Africa. (Open) Educational Resources around the World.

[13] Reyes, J. A. 2015. The skinny on big data in education: Learning analytics simplified. TechTrends, 59(2): 75-80.

[14] Rose, J., & Scheepers, R. (2001). Structuration theory and information system development frameworks for practice. ECIS 2001 Proceedings, 80.

[15] Prinsloo, P. and Roberts, J., 2022. Analysis of Higher Education (HE) Systems' Approach in South Africa. (Open) Educational Resources around the World.

[16] Pedroso, John Erwin & Celestial, Joemari& Cataluña, Jonel. 2022. Social Media Platforms as Enhancement Tools of Academic Performance amid COVID-19 Pandemic. 3. 2785-2793.

[17] 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.

[18] 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.

[19] van den Berg, C. L. 2017. A framework to teach digital innovation skills to South African Information Systems students.

[20] White, M., Becker, J. and du Plessis, M., 2021. Unintended positive consequences of development centres in university graduates. Frontiers in Psychology, 12, p.5421.