

Received 25 April 2023, accepted 13 May 2023, date of publication 17 May 2023, date of current version 24 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3277225

RESEARCH ARTICLE

Student Performance Patterns in Engineering at the University of Johannesburg: An Exploratory Data Analysis

MFWABO MAPHOSA¹, WESLEY DOORSAMY², (Senior Member, IEEE), AND BABU S. PAUL¹

¹Institute for Intelligent Systems, University of Johannesburg, Johannesburg 2006, South Africa

²School of Electronic and Electrical Engineering, University of Leeds, LS2 9JT Leeds, U.K.


Corresponding author: Mfowabo Maphosa (mfowabo@gmail.com)

ABSTRACT Globally, the increased demand for engineers is not matched by an increase in graduates. This is further exacerbated by the fact that student dropout rates in engineering are higher than in other disciplines. Understanding engineering students' performance patterns and potential influences can lead to developing interventions to improve engineering students' success. Recent advances in data science and educational data mining have made it possible to extract valuable information from historical data, which can supplement interventions. This study sought to extract insights and information from real-world data, analyse correlations in the dataset's variables and better understand the influences of student performance. Exploratory data analysis was applied to the dataset to visualise the dataset and infer the correlations between variables provided in the dataset on student performance patterns. We used Python for data analysis and visualising the correlation between variables. The results show gender disparity in engineering enrollments, with only a quarter of female students enrolled. The study also indicates that the completion rates could be much higher. Another finding is that most students who drop out do so because of choosing the wrong qualifications. Furthermore, when comparing the percentages, female students performed slightly better than their male counterparts. The correlation analysis shows no relationships between gender, race, admission point score, mathematics marks and science marks with student performance in engineering. Understanding student performance patterns can reduce dropout rates by correctly advising students to enrol on the most suitable programmes, and aid support interventions are needed to improve student success in engineering.

INDEX TERMS Engineering education, exploratory data analysis, extended programmes, mainstream programmes, student performance.

I. INTRODUCTION

Research has shown an increasing need for more science, technology, engineering, and mathematics (STEM) graduates to satisfy the demand for these skills in the global economy [1]. Worldwide, concerns have been over declining STEM qualifications enrolment and economic growth's adverse effects [2]. Africa has been trying to overcome a significant engineering skills shortage [3]. Engineering is uniquely positioned to contribute to some of the United Nations Sustainable Development Goals (SDGs), such as

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu .

water, energy infrastructure, and agricultural technologies that support food security and biomedical technologies to reduce the disease burden [4].

The educational sector in Africa has suffered many setbacks due to underdevelopment, financial hardship, insufficient budget and corruption [5]. Despite increased funding, the education sector in South Africa still needs to produce the necessary skills required by the labour market [6]. In South Africa, the higher education sector consists of public and private institutions. Recently, there has been considerable growth in enrolments at public higher education institutions (HEI) due to increased funding by the government. Student academic performance in higher education is researched

extensively to tackle academic underachievement, increased university dropout rates, and graduation delays, among other tenacious challenges [7]. The ability to accurately predict students' academic performance is crucial because it affects admission decisions that HEIs make towards providing better educational services [8]. HEIs can identify and detect patterns based on statistical analysis [9].

Data mining tools can be used for multiple purposes, such as analysing student characteristics and predicting several outcomes, such as transferability, choosing elective courses, choosing the right career path, reducing student dropout rates and succeeding based on the student's performance in high school [10]. Once students' needs have been identified, instructors can implement intervention strategies to arrest the situation [9]. Student dropout is a critical issue that requires global analysis. Student dropout wastes resources and even affects the evaluation processes of educational institutions. Studies have shown that engineering students' dropout rate is higher than in other disciplines [11]; [12]; [13]. Thus, there is a clear need to understand student performance patterns and to identify the factors that influence student performance, especially in engineering.

South Africa faces several problems in engineering education, such as equality of access and success for black African students, inclusion, studying costs and dropout of academically eligible students [14]. One study analysed data from the 2005 student cohort registered for a bachelor's degree in engineering provided by the South Africa Council on Higher Education (CHE). This analysis showed that only a quarter of the students completed the degree in the required time of four years, almost a fifth (19%) finished in five years, and more than half (55%) completed the degree after six years [15]. A similar study focused on the results of the 2009-2013 cohorts of engineering students at the University of Johannesburg, precisely the mining engineering degree. The study found that common factors were identified as to why some students failed or took longer to complete a degree; these were grouped into socioeconomic and essential education/student readiness factors. The survey revealed that the average graduation rate of the institution was 40% [16].

Several measures have been implemented to increase student success, such as bridging, foundation and extended curriculum programmes. The background for these programmes recognised the need for more high-quality graduate output in South African higher education [17]. In 2006, foundation programmes were combined with the mainstream first-year courses into the extended curriculum model, which meant that the first year was spread over two years [18]. The bridging programmes aim to improve inadequate secondary education preparation for university students [19]. In contrast, the foundation programme seeks to lay an academic foundation for subsequent study. Several courses make up a foundation programme, which ensures accreditation for specific modules added to the first year of study [15].

An extended programme is one in which a qualification's minimum completion time is increased with a slightly different curriculum. Such a curriculum gives students more time while including developmental courses [20]. The extended programme offers an alternative curriculum framework and closure of articulation gaps [21]. The powerful feature of giving students more time for tuition is central to the bridging, foundation, and extended curriculum programmes. This is because underprepared students need more time and should be given more tuition if they want to succeed at university [19].

The South African government has been proactive in developing the engineering profession in the country. In 2001, the Department of Education unveiled the National Plan that restructured higher education by re-grouping 36 universities and technikons into 23 HEIs [22]. The National Plan, among others, introduced a Bachelor of Technology (B Tech) degree. This requires a year of academic work after receiving the diploma, and such graduates are called "engineering technicians". The B Tech is more pragmatically oriented than the Bachelor of Science because of the experiential education component of the National Diploma (N Dip) [23].

Exploratory data analysis (EDA) is a methodology that employs various techniques to maximise understanding of the data, uncover underlying structure, extract important variables, detect outliers and anomalies, test underlying assumptions and develop suitable models. It is a powerful framework for having a big picture of a phenomenon [24]. EDA is used to summarise, visualise and understand data through graphical representation and serves as the foundation for further statistical analysis. Its purpose is to examine the data for distribution, outliers, and anomalies. EDA is also used to explore interrelationships between variables and identify interesting subsets of observations [25]. EDA is intended to support the analyst's natural pattern recognition. EDA has gained considerable traction as the benchmark for analysing datasets. EDA visualises, plots, and manipulates data without assumptions that help assess data quality and build [26].

Research has been conducted on student academic performance, with much of this research being based on surveys of students' cohorts as they progress through their studies [27]. Although this work has been valuable in formulating and validating theories about student academic performance, the practical utility of survey-based approaches has been questioned based on the need for more accuracy, generalisability of results, and the high cost of conducting such studies [28]. Maphosa et al. noted that using actual data to investigate student performance issues represents a growing research area on which researchers and practitioners can focus. This study uses real student data from the UJ to explore student performance patterns in engineering [29]. There needs to be a greater understanding of the factors that influence student success in engineering and a need to use exploratory data analysis to extract

insights from educational data, particularly in developing countries.

The proposed approach is based mainly on EDA, an elementary technique that lets the data speak for itself. It uses Python's powerful and versatile data analysis ecosystem to enrich EDA with appropriate and convenient graphical ways to visualise results. The study's main contribution is using real-world data to understand student performance patterns and the potential influences thereof. The research objectives for this study are to determine the student performance patterns in engineering and identify any correlations between the different variables provided in the dataset. Based on these objectives, below are the research questions for the study:

- What are the student performance patterns in main-stream and extended engineering programmes?
- Are there any correlations between the different variables provided in the dataset and student performance in engineering?

The remainder of the paper is structured as follows. Section II provides the context for the study case, the UJ. Section III presents the data and methods employed in the study. Section IV presents the results of this study, and Section V discusses the implications of these findings. Section VI wraps up the paper's discussion and suggests additional research.

II. CONTEXT

There are 26 public universities in South Africa (Universities South Africa, 2022). This study is based on data about engineering students from the UJ, a university located in the most populated city within the most densely populated province of the country, Gauteng. Gauteng has the country's most prominent university student population [30]. UJ was officially launched on 1 January 2005, resulting from a merger of the Rand Afrikaans University, Witwatersrand Technikon and the two Vista campuses of Soweto and the East Rand [31]. UJ was chosen as the case because it is a medium-sized university with about 50 000 students, making it one of the largest contact universities in South Africa.

In 2001, the University of Johannesburg (UJ) implemented a bridging programme that accepted students rejected from any N Dip in Engineering programme because they needed to meet the entry requirements. However, one of the requirements for continuing engineering studies at the UJ after completing the bridging programme was that students could not repeat the programme or fail any of the six courses offered [18]. In 2007, the UJ instituted the Engineering extended curriculum programme and phased out the foundation programme [18]. As in the bridging and foundation programmes, the Senate ruling that required all subjects to be passed had to be removed as required by the Department of Education. If they failed their first year, students admitted to the extended curriculum N Dip in Engineering were now permitted to repeat [18].

The Faculty of Engineering and the Built Environment (FEBE) at the UJ offers engineering technology and engineer-

ing science undergraduate and postgraduate programmes. FEBE has five centres, 12 departments and a Postgraduate School of Engineering Management [32]. Engineering Council of South Africa (ECSA) is a signatory of the Dublin, Sydney and Washington Accords, an international agreement among bodies responsible for accrediting engineering degree programmes. The (Bachelor of Engineering) B Eng and Bachelor of Engineering Technology (B EngTech) degrees offered by FEBE are registered with the ECSA.

The graduates for the N. Dip become technicians, the B EngTech graduates become technologists, and the B Eng graduates become engineers. BEngTech aims to become more 'knowledge professionalised' to enable technical specialisation following on-the-job training, experience, and clarification in postgraduate research programs. The N Dip programme is characterised by contextual relevance to emphasise more task-specific knowledge. The B EngTech's insistence on continuous professional development, as opposed to N Dip graduates' emphasis on job-based skills and knowledge of situations and procedures [33].

III. DATA AND METHODS

This study used a descriptive, exploratory data analysis on quantitative data obtained from UJ's Institutional Planning, Evaluation and Monitoring department. Ethical clearance to use the data was obtained from FEBE (ethical clearance number - UJ_FEBE_FEPC_00685) and the Research and Innovation Team within the university.

A. DATASET OVERVIEW

The dataset is about students registered for the 2016 and 2017 academic years studying towards N Dip and B Eng Tech qualifications in civil, electrical, industrial, and mechanical engineering. The dataset included registration data such as date of birth, race, gender, marital status, ethnicity and the student's home location (urban or rural). It also contained registration data such as the academic year and qualification, whether the student was on the extended programme, whether the student had cancelled their registration, the student's final result, and the study period at the start of the academic year. Also, the dataset contained student entry data – admission point scores (APS) and the student's previous activity. In South Africa, APS determines if a student qualifies for a particular programme. APS is calculated using matric (high school leaving grade) subject marks. The dataset also contained performance data that included graduation year and the years it took to complete the qualification for students who had completed their studies for the B Eng Tech: Engineering: Electrical degree. Each student's record was analysed individually for this degree to determine whether they graduated within the minimum time required.

B. DATASET CLEANING AND CODING

The initial dataset contained sheets with the 2016 and 2017 student registration data with 4 242 records and the matric data with 8 662 records (mathematics marks and

TABLE 1. Variables retained in the dataset.

Variable	Description
Student Number	Student's number of the student
Academic year	Year of registration
Extended	Enrolled for the extended programme
Qualification	The qualification the student is enrolled for
Cancelled	If the student cancelled studies
Result	Student's results for the year
Period of study	Student's year of study
Gender	Student's gender
Age in 2017	Student's age in 2017
Marital status	Student's marital status
Race	Student's ethnicity
APS	The student's admission point score
Location	Student's home location
Previous activity	Student's activity the previous year
Mathematics mark	Student's mark for mathematics in high school
Science mark	Student's mark for science in high school

science marks from high school). Non-numerical data in the dataset was coded numerically to allow analysis. There were no duplicates in the registration dataset when analysed per academic year.

The matric dataset was cleaned by removing duplicates in the dataset. Using student numbers, mathematics and physical science marks were merged into the student registration data using Python's merge functionality to create a combined dataset. The left merge was used to retain all the information for students in the registration data with the mathematics and physical science marks. Where students' mathematics and science marks were available, these were linked to the student's registration data. Table 1 shows a description of the variables retained in the final dataset. Fig. 1 shows the summary of the final dataset that was analysed.

C. MISSING DATA ANALYSIS

The quality of data analysis depends on the data used. Analysing the data quality to check missing values is a crucial and fundamental stage in EDA. Situational awareness can be developed from the overall quality of the data, which can be instructive and help improve data analysis by recognising and accounting for these data gaps. Knowing a data set's patterns and missing data landscape can help determine how to handle missing data in subsequent analysis steps. The Python package missingno is designed to help users visualise data set completeness and understand missing data [34]. Fig. 2 provides a summary of the missing data. As shown, values are missing for three variables – marital status, APS, and location. The left axis shows the ratio of missing values. The top values indicate each variable's total number of instances at the bottom.

A correlogram in Fig. 3 presents nullity correlations when analysing missing data. This correlogram, also created with the missingno package, aims to illustrate how one variable's presence or absence influences another's presence or absence. Nullity correlation in the map ranges from -1 to 1, with values

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 4242 entries, 0 to 4241
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Student Number        4242 non-null   int64
1   Academic Year          4242 non-null   int64
2   Extended               4242 non-null   int64
3   Qualification          4242 non-null   int64
4   Cancelled              4242 non-null   int64
5   Result                 4242 non-null   int64
6   Period Of Study        4242 non-null   int64
7   Gender                 4242 non-null   int64
8   Age in 2017            4242 non-null   int64
9   Marital status         4041 non-null   float64
10  Race                   4242 non-null   int64
11  APS                    3653 non-null   float64
12  Location               2819 non-null   float64
13  Previous Activity       4242 non-null   int64
14  Mathematics Mark       3750 non-null   float64
15  Science Mark           3724 non-null   float64
dtypes: float64(5), int64(11)
memory usage: 563.4 KB
```

FIGURE 1. Final dataset summary showing the variables extracted from the initial dataset and the values for each variable.

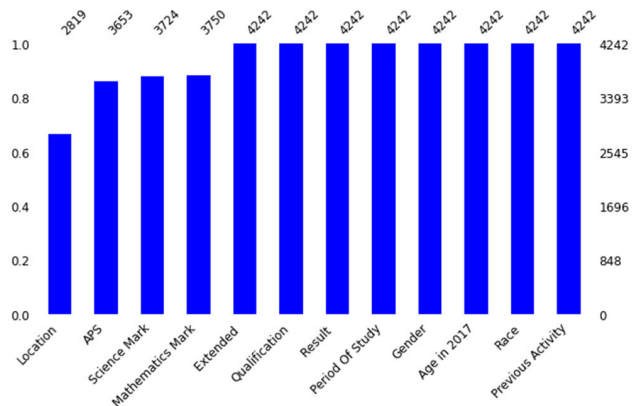


FIGURE 2. Bar chart depicting missing data.

closest to -1 denoting a negative correlation, values closest to 1 indicating a positive correlation, and 0 representing no nullity correlation. The graph excludes variables that are either always empty or have no missing data [34]. Variables with no missing data or always open are omitted from the chart. As shown in Fig. 3, a positive relationship exists between APS scores, mathematics marks, and science marks. This is expected, as mathematics and science marks are used to calculate APS scores. APS and location have a moderate positive correlation. The map shows no nullity correlation between mathematics marks and science marks.

Analysing the available data as if there are no missing data can produce biased results when there are missing data [35]. Although standard imputation methods perform worse than simple listwise deletion (if one value is missing in that row, the whole row is excluded), handling missing categorical data with imputation still requires a careful

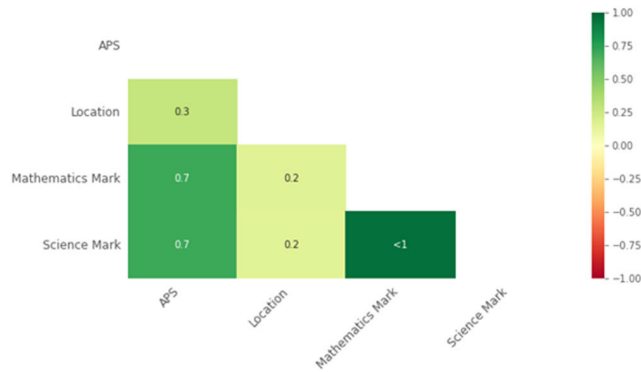


FIGURE 3. Nullity correlation heatmap.

approach because little is known about the mechanism of underlying missing data [36]. This analysis showed that the target data are generally of good quality. Because imputation methods for categorical missing data are pretty complex and are separate from EDA, exploratory research can proceed without requiring a rigorous strategy for handling missing data.

IV. RESULTS

In this section, an exploratory analysis of the dataset is performed. We used various EDA techniques to analyse our datasets and visualise the results to evaluate engineering students' performance. We set up several Python packages, including Pandas, Seaborn, Matplotlib, Missingno and Matplotlib.

A. EXPLORATORY DATA ANALYSIS

The purpose of EDA is to provide an overview of the dataset's structure and students' performance. Fig. 4 shows the distribution of the dataset in terms of gender, the academic year of registration, whether students enrolled for the extended programme, and if students cancelled their qualifications. As shown, male represents three-quarters of the dataset. This aligns with prior findings that show that African women are underrepresented in engineering programmes across the African continent [37].

Regarding the academic year of registration, 51% registered in the 2016 academic year and 49% in the 2017 academic year. 42% of students were registered for the extended programmes, and 58% were on the mainstream programmes. 5% discontinued their studies in the 2016 and 2017 academic years, with the remainder continuing their studies. Analysis of the reasons for cancelling shows that the highest reason was that students made a wrong choice in the qualification registered for, accounting for 30%, followed by personal problems accounting for almost 18%. Change of qualifications accounts for 16% and administrative for 14%. Adaptation problems, deceased students, financial and health problems, working circumstances, transfer and completed qualifications combined accounted for almost 23% of the reasons for cancelling.

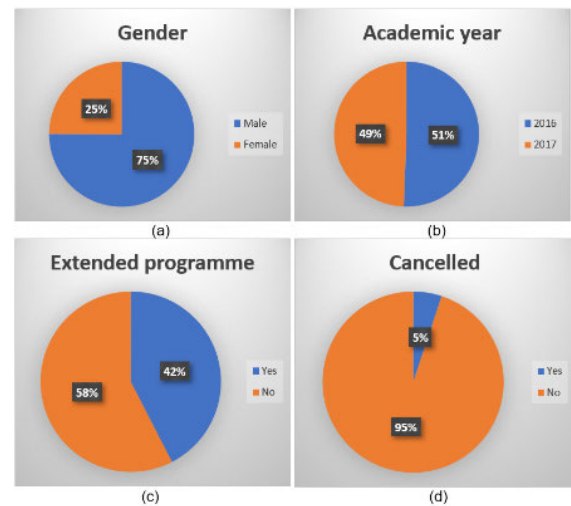


FIGURE 4. Dataset summary showing students' gender, the academic year of registration, whether students enrolled for the extended programme, and whether students cancelled their studies.

B. ANALYSIS OF STUDENT PERFORMANCE IN MAINSTREAM VS EXTENDED PROGRAMMES

Table 2 compares the mainstream programmes with the extended programmes in terms of qualification, level of study and final result for the year. The N Dip qualifications dominate, accounting for just over 87%, and the B Eng Tech accounting for about 13%. When comparing student registrations between the mainstream and the extended programmes, the data show that more students enrolled for the mainstream in all programmes except the industrial engineering for the B Eng Tech and the N Dip. Regarding the level of study, there were more registrations for the mainstream programmes compared to the extended programmes. The mainstream programmes have no fourth level because these are three-year qualifications.

The registration numbers are almost the same for the 2016 and 2017 academic years for both the mainstream and the extended programme, as shown in Table 2. The gender distribution also indicates that males dominate, with almost 79% for the mainstream programmes and just over 70% for the extended programme. Analysis of the results shows that nearly 22% of students in the mainstream had completed their studies compared to 15% in the extended programme. Just over 64% of students in the mainstream were continuing with their studies compared to 54% in an extended programme. There is a similar trend for 'no re-admission', 'no result' and 'no/slow progress' for both the mainstream and the extended programmes.

As shown in Table 2, in terms of the study period, students enrolled for the third year of their studies dominate, accounting for 32%, followed by second years with over 30% and first years with over 28%. Fourth years are the lowest representing 9.5%. This is understandable, as only extended programmes have the fourth year of study. Analysis of the result grouping shows that by August 2022, 19%

TABLE 2. Comparison of the mainstream and extended programme stats.

	Characteristic	Mainstream	Extended	Total	Percentage
Qualification	B Eng Tech: Engineering: Civil	79	53	132	3.1
	B Eng Tech: Engineering: Electrical	92	49	141	3.3
	B Eng Tech: Engineering: Industrial	44	79	123	2.9
	B Eng Tech: Engineering: Mechanical	102	51	153	3.6
	N Dip: Engineering: Civil	450	236	686	16.2
	N Dip: Engineering: Electrical	745	305	1 050	24.7
	N Dip: Engineering: Industrial	284	631	915	21.6
	N Dip: Engineering: Mechanical	645	397	1 042	24.6
	Total	2441	1801	4 242	100.0
Level of study	First-year	742	464	1 206	28.4
	Second year	820	455	1 275	30.1
	Third year	879	477	1 356	32.0
	Fourth-year	-	405	405	9.5
	Total	2441	1801	4 242	100.0
Result	No re-admission	127	95	222	5.2
	No result	129	77	206	4.9
	No/slow progress	326	202	528	12.4
	Continuing with studies	1 326	1 157	2 483	58.5
	Obtained qualification	533	270	803	19
	Total	2441	1801	4 242	100.0

of students had completed their qualifications, and 58.5% were continuing their studies. Similarly, 5.2% of students were not readmitted due to poor performance, 4.9% did not have results, and 12.4 had no/slow progress. Most students in the 'no/slow progress' category needed experiential training, followed by students awaiting special exam marks to continue.

When comparing the no-re-admission and the continuing studies categories students, there is an almost equal distribution between the mainstream and the extended programmes. There is a similar trend in the no result and the no/slow progress categories, with the mainstream accounting for over 60% and the extended for almost 40%. As expected, the mainstream dominates the obtained qualification category because the extended programme is a year longer than the mainstream programme. The most common reason for the no/slow progress is students with incomplete or outstanding experiential learning, accounting for 93.1%, followed by students having failed a module twice, accounting for 4.2% and then students' progress depending on results for a supplementary or special assessment. At UJ, although the university undertakes to assist students in obtaining suitable experiential learning placements at approved companies, it is the student's responsibility to get appropriate experiential learning placements at approved companies. After completing each level of experiential learning, the students must ensure that they hand in all the documented evidence of their having finished their experiential learning; this should be done according to the submission dates stipulated by FEBE [38]. The high percentage could be due to

students needing help finding or completing experiential learning.

A boxplot graphically portrays a continuous variable's distribution and shows how a continuous variable varies between groups. Boxplots show a dataset's first, second, and third quartiles, interquartile range, and outliers. The median is indicated by a line dividing the box. The position of the box's dividing line serves as a visual representation of the distribution's symmetry and skewness; symmetry happens when the line splits the box into two halves, and skewness occurs when one half is larger than the other [39]. The minimum, the 25th percentile, the median, the 75th percentile, and the maximum are displayed in a box plot.

Fig. 5 compares the APS showing the APS for mainstream programmes (N) and extended programmes (Y) in a boxplot. The tails are of the same length, indicating that there is no imbalance in the dataset. A sample from a normal population should have whiskers roughly the same length as the box or slightly longer. As can be seen, the whiskers are longer than the box. Students on the mainstream programme have higher APS values, with outlier APS in the lower adjacent values.

The outliers on the higher adjacent values for the APS values for students on the extended programme indicate that students with higher APS values also choose to do the extended programmes. The outliers on the lower adjacent values for the APS values for students on the extended programme indicate the acceptance of students who have gone through alternative pathways. These include students with low APS scores and enrolled for the National and Technical Education Department's Nated (N3, N4, N5 and N6)

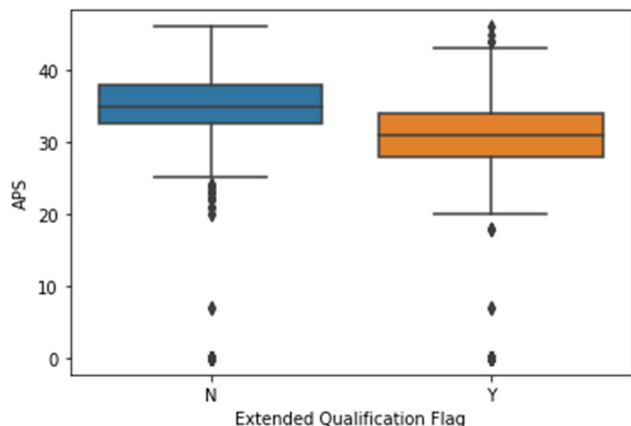


FIGURE 5. Comparison of the APS for students enrolled for the mainstream (N) and extended (Y) programmes.

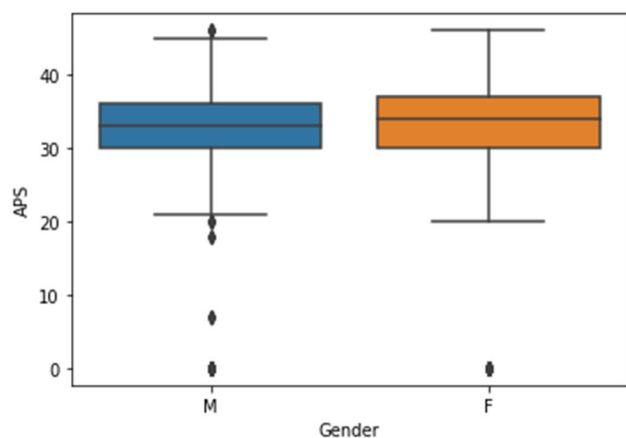


FIGURE 6. Comparison of the APS obtained by male (M) and female (F) students using gender.

qualifications at Vocational Education and Training (VET) institutions. Although in a limited capacity, universities and VET institutions collaborate to create alternatives that colleges or universities can offer to act as entry points into university programmes [40].

Using a boxplot, fig. 6 compares the APS values for male (M) and female (F) students. As can be seen, for male students, the potential outlier is picked out in the boxplot, with outlier APS values in both the lower and upper adjacent values. The fact that the upper outliers are not all that much higher than the highest APS value and the lower outliers are not all that much lower than the lowest APS values suggests that these values have no significant issues. These APS values are for students who got admitted to the extended programmes. Male students’ median APS scores are more significant than female students. The interquartile and overall ranges are reasonably similar, as depicted by the lengths of the boxes.

Table 3 compares the academic performance of male and female students. The results show that over 21% of female students had completed their studies compared to 18% of

TABLE 3. Academic performance of male and female students.

Variable	Male		Female	
	Count	Percentage	Count	Percentage
No re-admission	185	5.8	37	3.5
No result	162	5.1	44	4.1
No/slow progress	400	12.6	128	12.1
Continuing with studies	1 858	58.4	625	59.1
Obtained qualification	579	18.2	224	21.2
Total	3 184	100	1 058	100

male students. The statistics for those continuing their studies are almost the same, with males having 58% and females having 59%. There is a similar trend for ‘no re-admissions’, ‘no result’ and ‘no/slow progress’ for both the mainstream and the extended programmes.

C. ANALYSIS OF GRADUATION RATES

The dataset contained graduation information for students registered for the B Eng Tech in Electrical Engineering. Table 4 analyses the statistics for students who completed their qualifications. Of the 141 students enrolled in the 2017 academic year, only 70 completed their qualifications by the end of the first semester in 2022. After three years, 13.5% of students had completed their qualifications, and regarding gender, 13.4% of males and 13% of females had completed their qualifications. After four years, 30.5% of students had completed their studies – 28.6% males and 39% females. After five years, 46.8% of students had completed their qualifications – males (42.9%) and females (65%). By the end of the first semester in 2022, 49.6% of the students had completed their qualifications, with females accounting for 22.9% and males 77.1%. When comparing the mainstream and the extended programme graduation rate, it is clear that the mainstream students account for just over 71% of all students who graduated and 35.5% of all students registered for the qualification.

Fig. 7 compares the APS scores with the total years to complete the qualification. One would have expected students with lower APS to take longer to complete their qualifications than those with higher APS. As can be seen, there is no correlation between APS and the time taken to complete qualifications. There is an equal representation for the students with low, medium or high APS scores in the students who took three, four, five or six years to complete their qualifications.

D. CORRELATION ANALYSIS

A stable or positive relationship occurs when higher values of the first variable correlate with higher values of the second variable, and lower values of the first variable correlate with lower values of the second variable, indicating a stable or positive relationship. Higher values of the first variable correlate with lower values of the second variable giving a positive correlation. In comparison, lower values of the first

TABLE 4. B Eng Tech Engineering: Electrical Completion Statistics.

Enrollment Year	Graduation Year	Mainstream	Extended	Male	Female
2017	2019	19	-	16	3
2017	2020	16	8	18	6
2017	2021	13	10	17	6
2017	2022*	2	2	3	1
Total		50	20	54	16

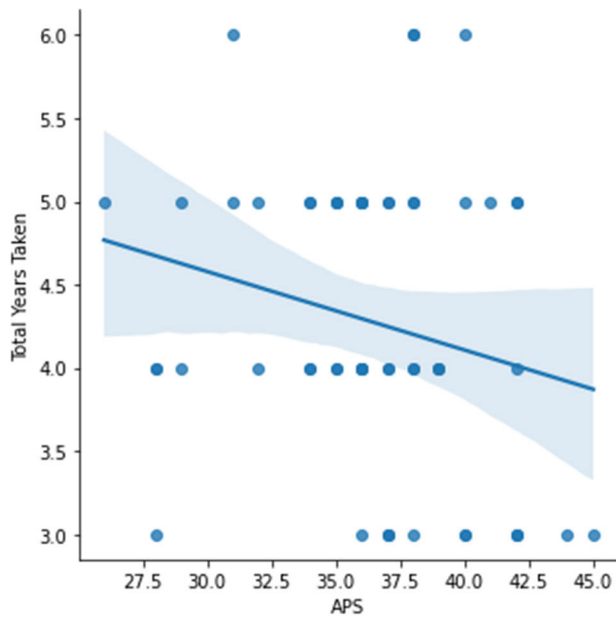


FIGURE 7. A comparison of APS obtained by students versus the years taken to complete qualifications.

variable correlate with higher values of the second variable (scores decrease from left to right). Often the linearity or non-linearity of the relationship is essential. A linear relationship exists when the second variable changes proportionately to changes in the first variable.

We visualised the correlation grid created using a heat map using the Python Pandas corr() function to obtain correlations. Fig. 8 shows the correlation matrix with the heat map. The blue shades indicate a negative correlation, while the red shades indicate a positive correlation.

The correlation heatmap shows a strong positive relationship between mathematics and high school science marks. This finding suggests that students who do well in one subject also do well in the other subject, as shown by the score of 0.65. A moderate positive relationship exists between APS and the science mark (0.58) and a low positive relationship between APS and the mathematics mark (0.46). When comparing the various variables against student performance (result), the correlation map shows a negligible relationship, as shown by a correlation of less than 0.30. The correlation between the period of study and the result was 0.28. The rest of the correlations were very low, less than 0.10. Overall, the data suggest no relationships exist between gender, race,

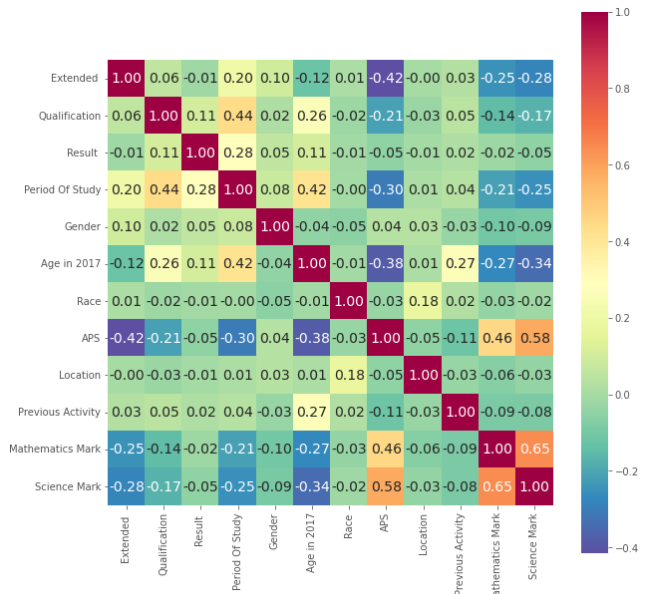


FIGURE 8. Heat map showing the correlations among the dataset variables.

APS, mathematics mark, science mark and previous activity with performance in engineering (result).

V. DISCUSSION

Globally, the increased demand for engineers is not matched by an increase in graduates. This is further exacerbated by the fact that student dropout rates in engineering are higher than in other disciplines. Understanding engineering students’ performance patterns and potential influences can lead to developing interventions to improve engineering students’ success. While noble, the use of survey data can lead to biases. There is a need to use actual data to understand issues in engineering education [29]. There are many new opportunities to gain meaningful empirical evidence from data, and commercial and social scientific research is analysing enormous data sets more frequently. This study focuses on understanding the engineering students’ performance patterns at a medium-sized university, the UJ, with an average student record size, which has never been explored in the educational or learning analytics domain.

The study’s findings show that there is still a wide gender disparity in enrolment in engineering, with three-quarters of male students. Exploring the results in more detail reveals some interesting gender differences. While females’ admissions rates in engineering are less than that of males, females performed better than males when comparing the completion rates of one qualification. The issue of gender disparity in STEM has been a focus area in research. Research has shown that the weight of evidence no longer supports that gender disparities result from differences in innate ability. Instead, gender differences in STEM appear to result partly from differences in perceived values and opportunities in environments and from general implicit and

explicit biases that shape perceptions of those values and environments [41]. Knowing the root causes of women's underrepresentation in admissions would enable targeted corrective measures to be implemented. Similar measures can be used for reasons to increase female participation in engineering.

Analysis of the reasons why students cancelled (dropped out) shows that almost a third of them dropped out because of choosing the wrong qualifications. Another study showed that the wrong choice of qualification, workload issues or academic difficulties contributes most strongly to the decision to drop out [42]. Inadequate guidance and education are cited as factors leading to or dropping out of engineering and emphasise the need to improve risk management through guidance in design program selection and identification [43]. In the South African context, a suggestion has been made to create opportunities for career guidance and screen students before the final acceptance to prevent students from choosing the wrong qualifications [16]. The availability of data and advances in educational data mining initiatives makes it opportune to develop a recommender system that can reduce the problem of selecting the wrong qualifications by recommending the most suitable qualifications to students.

Analysis of the students with no/slow progress shows that 93% had incomplete or outstanding experiential learning documentation, leading to their inability to complete their studies or to graduate. Experiential learning, or learning through experience, is often defined as the opposite of traditional or didactic learning, where students are relatively passive in listening or reading about the experiences of others. In design, a similar concept is often expressed by active learning, learning by doing, or hands-on learning, which refers to the student's active role [44].

The study found that regarding the students who completed their qualifications, those on the mainstream performed slightly better than those on the extended programme. This could be because, by nature, the mainstream programme attracts stronger students than the extended programme. An earlier study found no marked difference in the student's performance in mainstream and extended programmes [45]. This indicates the success of extended programmes. Although the programmes have lower entrance requirements than mainstream programmes, they are still considered inferior, and students are frequently stigmatised [46].

The study's findings regarding graduation rates show that only about 14% of the 141 students enrolled for the B Eng Tech in Electrical Engineering had completed their qualifications in the minimum time of three years. After six years, only about half of the enrolled students had completed their qualifications. This finding is similar to an earlier study that showed that South Africa has one of the lowest graduation rates in the world at approximately 15% [47]. A survey of the graduation trends in South African engineering students showed that students who graduated on

time obtained many credits in their first year. Such a result suggests that universities should consider providing academic and financial support [48].

There is an assumption that past student performance determines future performance. This is only sometimes the case, as social factors could have influenced past and future performance [49]. The relationship between entry requirements and students' performance has been a subject of interest even in engineering education. Past research has shown that core-course academic variables are critical in determining engineering education success (retention and graduation). Ghani and Mohamed found a moderate correlation between entry requirements and students' performance [50]. Odukoya et al. found a weak correlation between entry requirements and student performance [51]. The data analysis revealed that no relationships exist between gender, race, APS, mathematics mark, science mark and previous activity with performance in engineering (result). This finding is consistent with the previous research that found no significant relationships between entry requirements and student performance.

The results from this study should be considered in the context of the study's weaknesses. The study identifies correlations but not causal relationships between academic variables and student performance in engineering and does not consider non-academic factors. Further research is required to demonstrate causation and consider these other aspects. The study also used EDA instead of machine learning algorithms because it aims to analyse and understand datasets to find patterns, relationships, and insights. EDA is often done before applying machine learning algorithms to a dataset. More variables are necessary to increase the reliability of the findings and identify the causal mechanisms underlying the relationship between different variables regarding student data. The other limitation of this study is the use of the 2016 and 2017 cohorts and, thus, needs to reflect more recent trends in engineering education. Although a restriction, using the 2017 cohort allowed us to track graduation rates until 2022 for one specialisation. The dataset did not contain information about application choices. Therefore, it was impossible to track if the student were placed within their first choice. Future research could focus on a more recent cohort. Furthermore, there is merit in investigating the influence of first, second or third-choice admissions on performance in engineering.

Future research may also examine non-traditional routes into engineering, such as vocational training, and the effects of extracurricular activities and other non-academic activities on student performance. Studies may consider motivation, personality, and study habits as non-cognitive factors affecting students' success in engineering. It might also look into the use of interventions like peer support or mentoring programmes as a way to help students succeed. To make the results more generalisable, the study might also be broadened to cover a more extensive and varied sample of students.

VI. CONCLUSION

An EDA method for identifying patterns in engineering education student performance has been presented in this paper. The study has achieved its research objectives of determining student performance patterns and identifying any correlations between the different variables provided in the dataset. We presented an exploratory data analysis of a dataset of B Eng Tech and N Dip programmes in engineering for the 2016 and 2017 cohorts at UJ. Because the visual representations are light and easy to understand, the results or outputs produced in the form of graphs can make it easy for anyone to understand information about the current situation. We also showed that our findings agree with the previous methods presented in the literature.

The research results confirm the general trend that there is a gender disparity in enrollment figures in engineering. Looking at the results in more detail reveals some interesting gender differences. Although women's enrollment rates in engineering are lower than men's, women perform better than men. Analysis data shows that almost a third of students drop out due to choosing the wrong qualifications. There is, therefore, a need to offer better career guidance and counselling to reduce the number of students dropping out because of choosing the wrong qualifications. A recommender system can reduce the challenge of choosing the wrong qualifications by recommending the most appropriate qualifications to students.

The study noted that students needed help completing experiential learning, a qualification requirement. This could be due to the impact of COVID-19, which restricted access to workplaces, meaning students needed help finding companies to complete the experiential learning component. Therefore, universities need to be proactive and rethink the experiential learning requirements during and after the impact of pandemics to ensure that students are not affected by factors beyond their control.

Students on the mainstream programmes performed marginally better than those on the extended programmes. This might be the case since mainstream programmes attract more capable students than the extended programme. HEIs must address issues around the stigmatisation students face on the extended programme to ensure that the programme is still attractive to students. Maintaining high engineering enrolment means extended programmes are critical in this endeavour. The study confirms that South Africa's graduation rates are still low, and interventions are needed to increase this. The availability of extensive data and advances in machine learning and educational data mining presents an opportunity. The study's findings are consistent with the previous findings that there is an insignificant correlation between the entry requirements and academic performance. Understanding student performance patterns on its own may not be necessarily helpful for sustainable and impactful impacts in improving student performance in engineering. Student performance patterns must be used to design and better understand student support interventions. The fact that

30% of students that cancelled their studies did so because they had chosen the wrong qualification shows a need to advise students better.

ACKNOWLEDGMENT

The authors would like to thank to the Institutional Planning, Evaluation and Monitoring Department, UJ, for compiling the dataset used for the study.

REFERENCES

- [1] National Academies of Sciences, Engineering and Medicine, *Promising Practices for Strengthening the Regional Stem Workforce Development Ecosystem*. Washington, DC, USA: National Academies, 2016.
- [2] M. Cole, *Literature Review Update: Student Identity About Science, Technology, Engineering and Mathematics Subject Choices and Career Aspirations*. Melbourne, VIC, Australia: Australian Council of Learned Academies, 2013.
- [3] *Engineering: Issues, Challenges and Opportunities for Development*, UNESCO Publishing, Paris, France, 2010.
- [4] M. Klassen and M. Wallace, "Engineering ecosystems: A conceptual framework for research and training in sub-Saharan Africa," in *Proc. 7th Afr. Eng. Educ. Ass. Int. Conf.*, Lagos, Nigeria, 2019, pp. 1–12.
- [5] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, pp. 1–21, 2019.
- [6] *Education Series Volume V: Higher Education and Skills in South Africa*, Statistics South Africa, Pretoria, South Africa, 2017.
- [7] A. Namoun and A. Alshantiri, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," *Appl. Sci.*, vol. 11, no. 1, p. 237, Dec. 2020.
- [8] J.-F. Chen and Q. H. Do, "Training neural networks to predict student academic performance: A comparison of cuckoo search and gravitational search algorithms," *Int. J. Comput. Intell. Appl.*, vol. 13, no. 1, Mar. 2014, Art. no. 1450005.
- [9] S. Akgun and C. Greenhow, "Artificial intelligence in education: Addressing ethical challenges in K-12 settings," *AI Ethics*, vol. 2, pp. 1–10, Sep. 2021.
- [10] A. Kunjir, P. Pardeshi, S. Doshi, and K. Naik, "Recommendation of data mining technique in higher education," *Int. J. Comput. Eng. Res.*, vol. 5, no. 3, pp. 29–34, 2015.
- [11] L. Paura and I. Arhipova, "Student dropout rate in engineering education study program," *Eng. Rural Dev.*, vol. 2016, pp. 641–646, May 2016.
- [12] S. Sultana, S. Khan, and M. A. Abbas, "Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts," *Int. J. Electr. Eng. Educ.*, vol. 54, no. 2, pp. 105–118, Apr. 2017.
- [13] D. Buenaño-Fernández, D. Gil, and S. Luján-Mora, "Application of machine learning in predicting performance for computer engineering students: A case study," *Sustainability*, vol. 11, no. 10, pp. 1–18, 2019.
- [14] N. Ahmed, B. Kloot, and B. I. Collier-Reed, "Why students leave engineering and built environment programmes when they are academically eligible to continue," *Eur. J. Eng. Educ.*, vol. 40, no. 2, pp. 128–144, Mar. 2015.
- [15] J. P. Grayson, "The consequences of early adjustment to university," *High. Educ.*, vol. 46, pp. 411–429, Dec. 2003.
- [16] M. Mpanza, *The Throughput of Mining Engineering Students in the University of Johannesburg (2009 to 2013 Cohorts)*. Accessed: Jan. 16, 2023. [Online]. Available: <https://ujcontent.uj.ac.za/esploro/outputs/conferencePaper/The-throughput-of-mining-engineering-students/9913574007691#file-0>
- [17] I. Scott, "Designing the South African higher education system for student success," *J. Student Affairs Afr.*, vol. 6, no. 1, pp. 1–17, Jul. 2018.
- [18] P. Machika, "Redefining access for success in engineering extended programmes," *South Afr. J. Higher Educ.*, vol. 26, no. 5, pp. 987–1000, Jan. 2016.
- [19] B. Kloot, J. Case, and D. Marshall, "A critical review of the educational philosophies underpinning science and engineering foundation programmes," *South Afr. J. Higher Educ.*, vol. 22, no. 4, pp. 799–816, Feb. 2009.

- [20] C. Boughey, "Understanding teaching and learning at foundation level: A 'critical' imperative," in *Beyond the University Gates: Provision of Extended Curriculum Programmes in South Africa*, C. Hutchings and J. Garraway, Eds., Cape Town, South Africa: Cape Peninsula Univ. Technology Fundani, 2010, pp. 4–7.
- [21] J. Case, D. Marshall, and D. Grayson, "Mind the gap: Science and engineering education at the secondary-tertiary interface," *S. Afr. J. Sci.*, vol. 109, no. 7, pp. 1–5, 2013.
- [22] I. Christiansen and N. Baijnath, "Perceptions of the 'University of Technology' notion at higher education institutions," *South Afr. J. Higher Educ.*, vol. 21, no. 2, pp. 207–217, Sep. 2007.
- [23] B. Kloot and S. Rouvrais, "The south African engineering education model with a European perspective: History, analogies, transformations and challenges," *Eur. J. Eng. Educ.*, vol. 42, no. 2, pp. 188–202, Mar. 2017.
- [24] NIST/SEMATECH. (Apr. 2012). *e-Handbook of Statistical Methods*. Accessed: Apr. 4, 2022. [Online]. Available: <http://www.itl.nist.gov/div898/handbook/>
- [25] D. T. Larose and C. D. Larose, "Exploratory data analysis," in *Discovering Knowledge in Data: An Introduction to Data Mining*, D. T. Larose and C. D. Larose, Eds. Hoboken, NJ, USA: Wiley, 2014, pp. 51–90.
- [26] M. Komorowski, D. C. Marshall, J. D. Saliccioli, and Y. Crutain, "Exploratory data analysis," in *Secondary Analysis of Electronic Health Records*. Cham, Switzerland: Springer, 2016, pp. 185–203.
- [27] S. Palmer, "Modelling engineering student academic performance using academic analytics," *Int. J. Eng. Educ.*, vol. 29, no. 1, pp. 132–138, 2013.
- [28] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decis. Support Syst.*, vol. 49, no. 4, pp. 498–506, Nov. 2010.
- [29] M. Maphosa, W. Doorsamy, and B. S. Paul, "Factors influencing students' choice of and success in STEM: A bibliometric analysis and topic modeling approach," *IEEE Trans. Educ.*, vol. 65, no. 4, pp. 1–13, Nov. 2022.
- [30] S. Vandeyar and A. Mohale, "Shifting perceptions of black students in a South African University residence," *South Afr. J. Higher Educ.*, vol. 31, no. 5, pp. 263–276, Sep. 2017.
- [31] South African History Online. (2011). *Rand Afrikaans University is Established*. Accessed: Aug. 24, 2022. [Online]. Available: <https://www.sahistory.org.za/dated-event/rand-afrikaans-university-established>
- [32] University of Johannesburg. (2021). *Faculty of Engineering & the Built Environment*. Accessed: Apr. 6, 2022. [Online]. Available: <https://www.uj.ac.za/faculties/engineering-the-built-environment/>
- [33] W. Doorsamy and K. Padayachee, "Conceptualising the knower for a new engineering technology curriculum," *J. Eng., Design Technol.*, vol. 17, no. 4, pp. 808–818, Aug. 2019.
- [34] A. Bilogur, "Missingno: A missing data visualization suite," *J. Open Source Softw.*, vol. 3, no. 22, p. 547, Feb. 2018.
- [35] A. Agresti, *An Introduction to Categorical Data Analysis*, 2nd ed. Hoboken, NJ, USA: Wiley, 2018.
- [36] P. D. Allison, "Imputation of categorical variables with PROC MI," *SUGI 30 Proc.*, vol. 113, no. 30, pp. 1–14, 2005.
- [37] E. Fredua-Kwarteng and C. Effah, "Gender inequity in African University engineering programs," *Int. Higher Educ.*, vol. 89, pp. 18–19, Apr. 2017.
- [38] University of Johannesburg. (2022). *FEBE 2022 Undergraduate Yearbook*. Accessed: Dec. 29, 2022. [Online]. Available: <https://www.uj.ac.za/wp-content/uploads/2022/01/febe-2022-undergraduate-yearbook.pdf>
- [39] J. Tukey, *Exploratory Data Analysis*, 1st ed. London, U.K.: Pearson, 1977.
- [40] J. Papier and S. Needham, "Higher level vocational education in South Africa: Dilemmas of a differentiated system," in *Equity and Access to High Skills through Higher Vocational Education*. Cham, Switzerland: Palgrave Macmillan, 2022, pp. 81–101.
- [41] T. E. S. Charlesworth and M. R. Banaji, "Gender in science, technology, engineering, and mathematics: Issues, causes, solutions," *J. Neurosci.*, vol. 39, no. 37, pp. 7228–7243, Sep. 2019.
- [42] E. Welp-Park, S. Preymann, D. Nömeier, and V. Rammer, "Academic difficulties, wrong choice of study programme or a lacking sense of belonging?—Reinvestigating the reasons for early dropout," in *Proc. Cross-Cultural Business Conf.*, 2022, p. 216.
- [43] B. N. Geisinger and D. R. Raman, "Why they leave: Understanding student attrition from engineering majors," *Int. J. Eng. Educ.*, vol. 29, no. 4, pp. 914–925, 2013.
- [44] P. C. Wankat and F. S. Oreovicz, *Teaching Engineering*, 2nd ed., West Lafayette, Indiana: Purdue Univ. Press, 2015.
- [45] R. G. Lekhehle, "Comparing academic performance of students in mainstream and extended programmes at a higher education institution in South Africa," M.S. dissertation, UKZN, Durban, 2020.
- [46] D. Hlalele and G. Alexander, "University access, inclusion and social justice," *South Afr. J. Higher Educ.*, vol. 26, no. 3, pp. 487–502, 2012.
- [47] P. N. Mafenya, "Increasing undergraduate throughput and success rate through mobile technologies: A South African distance learning case study," *Medit. J. Social Sci.*, pp. 428–434, Jul. 2014.
- [48] A. V. Bengesai and V. Paideya, "An analysis of academic and institutional factors affecting graduation among engineering students at a South African University," *Afr. J. Res. Math., Sci. Technol. Educ.*, vol. 22, no. 2, pp. 137–148, May 2018.
- [49] M. Maphosa, W. Doorsamy, and B. Paul, "A review of recommender systems for choosing elective courses," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 9, pp. 287–295, 2020.
- [50] A. A. Ghani and R. Mohamed, "The effect of entry requirement for civil engineering student performance," *J. Sci. Technol.*, vol. 9, no. 4, pp. 44–49, 2017.
- [51] J. A. Odukoya, S. I. Popoola, A. A. Atayero, D. O. Omole, J. A. Badejo, T. M. John, and O. O. Olowo, "Learning analytics: Dataset for empirical evaluation of entry requirements into engineering undergraduate programs in a Nigerian University," *Data Brief*, vol. 17, pp. 998–1014, Apr. 2018.

MFWABO MAPHOSA received the B.Sc. degree in information technology and computer science, the B.Sc. degree (cum laude) in information systems, and the M.Sc. degree in computing from the University of South Africa. He is currently pursuing the Ph.D. degree in electronic and electrical engineering with the University of Johannesburg, South Africa. He is also a Lecturer with the Faculty of Engineering, Built Environment and Information Technology, University of Pretoria, South Africa. He has published several journal articles and local and international conference proceedings. He is a peer reviewer of several journals and conferences. His research interests include artificial intelligence, machine learning, educational data mining, e-waste, engineering education, and ICT4D.

WESLEY DOORSAMY (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electrical engineering and the master's Diploma degree in higher education from the University of Witwatersrand, South Africa, the Graduate Diploma degree in data science from the University of London, and the M.Sc. degree in computer science (with artificial intelligence) from the University of York. Currently, he is a Lecturer and a Curriculum Redefined Lead with the School of Electronic and Electrical Engineering, University of Leeds. Prior to joining the University of Leeds, he was an Associate Professor with the Institute for Intelligent Systems, University of Johannesburg. He is a Senior Member of the South African Institute of Electrical Engineers (SAIEE), an affiliate of the African Academy of Sciences (AAS), and a National Research Foundation (NRF) Rated Researcher. His research interests include condition monitoring, applied machine learning, and data analytics.

BABU S. PAUL received the B.Tech. and M.Tech. degrees in radiophysics and electronics from the University of Calcutta, West Bengal, India, in 1999 and 2003, respectively, and the Ph.D. degree from the Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, India. He was with Philips India Ltd., from 1999 to 2000. From 2000 to 2002, he was a Lecturer with the Electronics and Communication Engineering Department, SMIT, Sikkim, India. He joined the University of Johannesburg, South Africa, in 2010, and he was the Head of the Department with the Department of Electrical and Electronic Engineering Technology, from April 2015 to March 2018. He is currently an Associate Professor and the Director of the Institute for Intelligent Systems, University of Johannesburg. His research interests include artificial intelligence and machine learning.

...