frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 02 October 2025
DOI 10.3389/feduc.2025.1501796

OPEN ACCESS

EDITED BY
Mona Hmoud AlSheikh,
Imam Abdulrahman Bin Faisal University,
Saudi Arabia

REVIEWED BY
Najah Al-Shanableh,
Al al-Bayt University, Jordan
Mahdi-Reza Borna,
Tarbiat Modares University, Iran

*CORRESPONDENCE
Neill Smit
✉ Neill.Smit@nwu.ac.za

RECEIVED 25 September 2024
ACCEPTED 19 September 2025
PUBLISHED 02 October 2025

CITATION
Smit N, Osler Z and van der Merwe L (2025) At-risk student identification and interventions for data science programs at a South African university.
Front. Educ. 10:1501796.
doi: 10.3389/feduc.2025.1501796

At-risk student identification and interventions for data science programs at a South African university

Neill Smit*, Zonia Osler and Leandra van der Merwe

Centre for Business Mathematics and Informatics & Unit for Data Science and Computing, North-West University, Potchefstroom, South Africa

In this paper, thresholds are established to identify at-risk data science students at a South African university and an intervention process is proposed for handling identified at-risk students. An evaluation of student performance in the core program modules is conducted, focusing on the differences between the standard and extended data science programs offered by this university. Through this evaluation, mark thresholds are specified for core mathematics and statistics modules that can be used to detect at-risk students. A statistical analysis is conducted to determine the suitability of using the thresholds for identifying at-risk students. A fitted logistic regression model, using the number of threshold breaches as the predictor, yields significant predictor coefficients and odds ratios for both programs (p = 0.0014 and OR = 4.0367 for the standard program; p = 0.0405 and OR = 2.1174 for the extended program). For both programs, the Mann–Whitney test confirms a statistically significant difference in the number of threshold breaches between graduates and dropouts ($p < 0.0001$; p = 0.0273) and Fisher's exact test indicates an association between the number of breaches and dropout status (p = 0.0002; p = 0.0312). Lastly, sensitivity/specificity analysis using the number of breaches to classify students yields estimated AUC values of 0.7811 and 0.7074, respectively. An intervention process is also suggested for the data science programs to provide struggling students with advice throughout their academic life cycles. This study shows how a simple threshold approach can be used to design an understandable and program-specific at-risk identification strategy. Literature on extended programs is less common than literature on bridging programs, where the differences between these transition programs are also highlighted in this paper.

KEYWORDS

at-risk students, data science education, extended programs, intervention process, program evaluation

1 Introduction

1.1 Data science education

Data science is one of the fastest-growing career sectors in the world. The United States Bureau of Labor Statistics projects an employment growth rate for data scientists in the United States of 35% from 2022 to 2032, compared to an average employment growth rate over all occupations of 3% for the same period (United States Bureau of Labor Statistics, 2023). LinkedIn lists data scientists, as well as three other jobs related to data science, namely financial technology engineer, machine learning specialist, and big data specialist, under its list of the

top ten fastest growing jobs for 2022 (LinkedIn, 2022). Glassdoor placed data scientist as the third best job in its list of 50 best jobs in the United States for 2022, where several other data science related jobs also made this list (Glassdoor, 2023). Due to the high demand for data scientists, the demand for education and training in data science has skyrocketed, with more and more universities offering data science programs (see, for example, De Veaux et al., 2017; Voulo et al., 2024).

A host of literature exists on the design of data science programs and the development of this field at university level. Only some of the relevant literature is discussed here. The early data science education research focuses on the technical skills required by data scientists, although the multidisciplinary nature of data science is also considered. Cleveland (2001) discusses an action plan to expand six technical areas of statistics and data science for a university department, where the author provides guidelines for resource allocation in data science degrees.

In later research, there is a clear shift towards the growing incorporation of computers in data science education, as well as the integration of the fields of mathematics, statistics, and computer science as the building blocks for data science education. Challenges in statistics education, as well as new teaching innovations and the reform of statistics education, are considered in Tishkovskaya and Lancaster (2012). The authors recommend the incorporation of information technology and real-world practical problems in statistics education, as well as the use of web-based learning materials to supplement teaching. A comprehensive overview of the development of data science is given in Cao (2017). The author discusses various topics, such as the evolution of data science, major challenges and innovations in data science, data competency and education, industrialization and new career opportunities, and the future of data science. De Veaux et al. (2017) provide curriculum guidelines for undergraduate data science programs based on inputs from various mathematics, statistics, and computer science university departments across the United States. The integration of courses in these three fields, together with a capstone project, is seen as crucial in data science degrees.

The most recent research emphasizes the need for practical experience and applications to real-world problems as part of data science education. Furthermore, the needs and involvement of government and industry must be considered during the design of data science programs at universities. Zakari (2020) addresses the development of statistics and data science at university levels in Niger, focusing on the importance of collaboration with government and industry, as well as collaboration between university departments that need statistics courses to develop academics and researchers with the necessary skills. The design and development of a four-year undergraduate data science program at a university in Bhutan, with inputs from external stakeholders in terms of their needs, is discussed in Namgay et al. (2022). The authors focus not only on the course content and the multidisciplinary nature thereof, but also on the process followed to develop the program and obtain approval for its implementation from the university board. De Veaux et al. (2022) expand on the work of De Veaux et al. (2017), by emphasizing some key aspects of data science education that may be overlooked in many data science degrees. The authors recommend that education in data science should focus not only on the theoretical background and the straightforward application to data, but also on training future data scientists to solve real-world problems. Special attention should be given to defining the purpose for which the data were collected, assessing the quality and integrity of the data, thinking about ethical considerations, and effectively communicating the key findings that address the original problem (De Veaux et al., 2022).

Due to the mathematical and statistical nature of most data science programs, the admission requirements are typically very high, and few students are allowed into these programs. Keeping in mind the global need for data science graduates, some universities offer bridging and extended data science programs to provide an opportunity for admission to these programs for more students.

1.2 Bridging programs in higher education

The National Plan for Higher Education (NPHE) emphasizes that raising participation rates in higher education relies on enhancing the system's efficiency by increasing the number of graduates it produces (Council on Higher Education, 2014). To adhere to the national plan, bridging programs for prospective university students act as an intervention to increase the graduation rates and the performance of students entering universities. Typically for science, technology, engineering, and mathematics (STEM) degrees, the transition from school to university can be challenging, and bridging programs can ease this transition by introducing prospective students to some core concepts before the start of the academic year. These bridging programs also increase the number of students enrolled for certain degrees. Not only is the increase in participation adhering to the national plan, but it is also of great relevance for STEM professions, as there are major skills shortages in these professions and students often do not meet the admission requirements for STEM degrees. In this case, bridging programs are typically aimed at those students who missed the admission requirements by a small margin. After participating in the bridging program, students are tested to determine if they have gained the necessary skills to qualify for the degree they applied for.

Most of the research on the efficacy of bridging programs shows positive results in terms of student retention and performance. Some relevant literature, mostly focusing on bridging programs for STEM degrees, is discussed next.

Murphy et al. (2010) determine the effect of a bridging program on the graduation rates of minority groups in scientific and technical disciplines at an American university. The authors report that participation in a bridging program contributes to student retention and significantly increases the likelihood of graduation of the participants. Ssempebwa et al. (2012) investigate the effectiveness of university bridging programs at a Ugandan university, where the bridging program is aimed at attracting international students who would otherwise not qualify for admission. The authors find that the bridging program is effective and that there is not a significant difference in the performance of students who were admitted through the bridging program against that of students who were admitted through conventional routes.

Raines (2012) investigates the efficacy of a bridging program, which is aimed at addressing mathematics deficiencies in STEM majors, at an American university. The author states that the bridging program positively impacted the performance and retention rates of the participants. Bradford et al. (2021) consider the effectiveness of university bridging programs, with a focus on STEM students. The

authors analyze STEM bridging programs from 16 universities across the United States and find that participation in the programs had a significant effect on first-year performance and retention. Brady and Gallant (2021) report on a qualitative assessment of a bridging program for minority groups enrolling in STEM programs at an American university. The authors state that the participants felt that the program not only increased their knowledge of relevant mathematics and science but also facilitated their transition to university.

Besides the challenge of students not meeting program admission requirements, universities also face the challenge of high dropout rates in data science programs. This creates a need for interventions and the early identification of students who may potentially drop out.

1.3 Literature on at-risk student identification

The rise in dropout rates among university students remains a major concern for higher education administrators, with some institutions in South Africa experiencing dropout rates as high as 80% (Moodley and Singh, 2015). This issue is particularly evident in more challenging programs, such as statistics and data science. Babalola et al. (2022) discuss the challenges that lead to high dropout rates among undergraduate students enrolled in statistics programs and states that dropout rates in developing countries are always high among undergraduate students studying statistics. Major stress factors that contribute to students not advancing to the next year, include financial difficulties, accommodation issues, academic pressures, and incorrect field of study due to limited information regarding their career choice (Pillay and Ngcobo, 2010). It is crucial for institutions to identify at-risk students early to retain them through intervention strategies, especially for emerging and growing fields such as data science (Babalola et al., 2022).

The identification of at-risk students has been investigated in many schools and institutions. One approach to early identification is to implement performance thresholds from the start of the students' academic program. A study on such thresholds is Gordanier et al. (2019), where the effectiveness of early academic intervention in economics courses at a large public university is investigated. Students who fell below a 70% threshold on a performance measure or had an attendance rate below 75% were referred to the university's student success center for additional academic support. The authors state that the interventions improved student scores on common questions on the final exam by 6.5 to 7.5 percentage points for students at or near the performance threshold. The gains were particularly large for students who entered college with below-average mathematics placement scores. In another study by Beitelmal et al. (2022), the identification of threshold concepts in higher education, particularly in introductory statistics courses, is discussed. The authors argue that identifying and focusing on threshold concepts (key ideas that are crucial to understanding a subject) can help instructors address areas where students often struggle, leading to improved comprehension and performance.

Most of the recent research papers on at-risk student identification focus on predictive modeling, using machine learning and deep learning models. Given the extensive literature available on this topic, we discuss only a selection of recent publications in this area of research. For the interested reader, these selected papers also refer to many other studies on educational data mining, student performance prediction, and identification of at-risk students.

Cummings and Smolkowski (2015) discuss the use of receiver operating characteristic (ROC) curves and the area under the curve (AUC) to determine appropriate thresholds/cut-offs for identifying at-risk students via predictive models or screeners. Ortiz-Lozano et al. (2020) use classification trees (CT) built on academic and socio-demographic data to identify at-risk university students. Their findings support the need for early identification and interventions, and indicate that academic data is the main contributor to making accurate predictions. A Bayesian profile regression approach based on data from undergraduate students at an Italian university, including students' performance, motivation, and resilience, is investigated in Sarra et al. (2019). The authors were able to group students into nine profiles, each characterized by different dropout rates and combinations of covariates. Al-Shabandar et al. (2019) consider several machine learning models, including random forest (RF), logistic regression (LR), gradient boosting machine (GBM), and neural network (NN), to identify students who are at risk of dropping out of large open online courses. Their study indicated that all the classifiers performed well in terms of accuracy, with GBM achieving the highest accuracy.

Veerasamy et al. (2020) consider CT and RF, based on data from early course work, for predicting student performance in an introductory programming course. Jamjoom et al. (2021) accurately predict whether students would pass a course based on preliminary performance in the course. The authors use CT, k-nearest neighbors (kNN), naïve Bayes (NB) classifier, and support vector machines (SVM). All models performed very well in terms of accuracy, with CT and SVM achieving the highest accuracy. Various machine learning models, such as NB, RF, CT, kNN, SVM, AdaBoost, and LR, are investigated by Pek et al. (2022) for identifying at-risk students. Again, all models achieve high accuracy, with an ensemble model using SVM as the meta learner identified as the best model after optimizing hyperparameters. Jang et al. (2022) use several machine learning models to identify at-risk students in seven courses at a Korean university. Mostly online behavioral features are used as variables and LR was found to be the best model based on performance metrics such as AUC and accuracy.

Carneiro et al. (2022) consider kNN, CT, RF, NB, NN, and pruning-based rule induction for at-risk student identification. Their feature set includes socio-demographic and geographical variables in addition to academic performance variables. All the machine learning models had high accuracy, with the pruning-based rule induction being the best performing model. Köhler et al. (2022) use a wide range of machine learning models to predict which students are at risk of failing an introductory course at a Chilean university. The study involves engineering students who can choose between a 4-year program and a 6-year program. The authors identify SVM as the best performing model in terms of accuracy. Borna et al. (2024) analyze data from the Open University Learning Analytics Dataset to identify students who are at risk of withdrawing. The authors explored several classification models and found that RF had the highest accuracy. Atindama et al. (2025) discuss the impact of targeted interventions on the retention of at-risk engineering students at a private research university. The study focuses on historically underrepresented students and three different

intervention strategies. The authors use LR to predict on-time graduation, before and after interventions, and find that the tailored intervention strategies are effective. Kalita et al. (2025) consider a bidirectional long short-term memory (bi-LSTM) network to predict student performance and identify at-risk students at an American university. The authors find that the bi-LSTM model outperforms several other machine learning models, achieving an accuracy of 88%.

Given the discussed literature, most machine learning models can be used to accurately identify at-risk students. However, the best performing model varies across studies. This observation is supported by Jang et al. (2022), where several studies are listed that identify different best performing machine learning models for at-risk student identification.

Once the at-risk students are identified, action should be taken to intervene. The intervention strategies employed include transitional and orientation classes, motivation, and building positive relationships to improve the literacy and learning skills of students (Lowder et al., 2022). Sarra et al. (2019) suggest that all intervention programs should improve students' resilience by enhancing their ability to plan and set goals to manage their studies. Pérez (1998) describes a strategy that will first divide the at-risk students into meaningful subsets and then offer support to assist with the everyday problems, connection opportunities to allow networking between students, and transformation strategies to overcome barriers preventing students from reaching their full potential.

1.4 Motivation and layout of the paper

The North-West University (NWU) is one of the largest universities in South Africa, with over 50,000 students across its three campuses in Potchefstroom, Mahikeng, and Vanderbijlpark. The NWU is a balanced teaching-learning and research university that offers a broad spectrum of programs across eight faculties, with unique strengths and demographics on each campus. The NWU typically ranks among the top 1,000 universities in the world, according to several ranking systems (see, for example, North-West University, 2025b). The Vanderbijlpark campus (VC), located next to the Vaal River, is the smallest and fastest growing campus of the three. The vast majority of the students on the VC are African, where many of them come from schools with limited resources and poor households. Most of these students also make use of government subsidized student loans, which includes a small monthly stipend that is often their only means for covering basic living expenses. Due to under-resourced schooling, students often do not meet the requirements for the standard three-year degree programs. Therefore, extended programs for several degrees are mainly offered on the VC.

Lecturers in the data science programs at the VC of the NWU have recently become concerned with the dropout rates for these programs. The differences in graduation rates of the standard and extended data science programs are of specific importance, since the extended programs were introduced to provide an opportunity for students from under-resourced schools who do not meet the requirements for the standard programs to study a data science program. The secondary aim of introducing the extended programs was to increase the number of data science graduates, due to the need for more qualified data scientists in South Africa.

These are very demanding degrees, resulting in many students taking exceptionally long to graduate or dropping out of the programs after several years. The lecturers have recognized the importance of providing students with guidance regarding their future studies, as many dropouts leave the university without a formal qualification. Some factors contributing to student dropouts for the VC data science programs have been identified, which include the following:

- Financial and personal challenges
- Adaptability to a new environment
- Difficulty and intensity of the programs
- Lack/gaps in mathematics foundation
- Move to another program
- COVID-19 pandemic

In this paper, a simple threshold approach based on academic performance data is formulated to identify students who are at risk of dropping out of the data science programs at the VC. The aim of our research study is to determine the following:

- Can this simple threshold approach be used to effectively identify at-risk students in these programs?
- Is there a significant difference between the graduation rates for the standard and extended programs?
- Can an intervention process for at-risk students in these programs, aimed at guiding students through their academic life cycles, be proposed?

Identifying at-risk students early allows for discussions and interventions with students whose performances indicate that they are unlikely to graduate within the maximum time allowed by the university. We establish threshold marks for core modules in the programs through an evaluation of student performance. The main reason for opting for this benchmarking approach is for easier communication and application, compared to the use of machine learning models. Easily understandable thresholds are straightforward to communicate to students early in their academic life cycle and can serve as motivation for them to meet these thresholds rather than just trying to pass modules. The goal of the study is not to build the most accurate predictive model, for which machine learning models would be more appropriate. The use of understandable thresholds also makes it easier to develop a structured and practically applicable intervention process. While the thresholds are program-specific, a similar process can be followed to develop a tailored framework for other degree programs. We perform a statistical analysis on the use of the thresholds, including LR, formal tests, and sensitivity/specificity analysis, which further supports our approach.

There is limited literature that focus on extended degree programs, since these programs are much less common than bridging and other transitional programs. This paper may address this gap in literature by highlighting the differences between extended programs and bridging programs. The aim of these discussions is to stimulate conversations among educators regarding the role and viability of extended programs at their own institutions, particularly for degrees relating to STEM fields and professional areas experiencing skill shortages.

The contribution of this work and its future application is of great importance for various reasons. First, we believe that program-specific guidelines that can assist in the early identification of students at risk

of dropping out, together with an intervention process, could possibly increase graduation rates. Furthermore, students referred to other programs through the intervention process could at least leave the university with alternative degrees, rather than simply dropping out of university after several years with no formal qualifications. Second, improving graduation rates specifically for the extended programs could motivate the introduction of extended data science programs at other South African universities. The structural differences between the standard and extended programs at the VC could then provide a foundation for developing such programs at other universities. Third, several statistics and data science related professions are classified by the South African government as critical skills. This classification means that there are major shortages in these professions, ranging from corporate jobs to university lecturers and teachers. Furthermore, many qualified South African data scientists readily find work overseas and emigrate, mainly due to socio-political concerns in South Africa, which further exacerbates local skill shortages. Should the intervention process prove successful, more data science graduates could enter the job market to alleviate these skill shortages.

The remainder of the paper is structured as follows. In Section 2, the data science programs offered at the VC are briefly discussed. The key differences between the extended and standard programs are also highlighted. In Section 3, the performances of students participating in both extended and standard programs are evaluated. Mark thresholds for core modules are also identified in this section, with the aim of identifying at-risk students. The efficacy of the thresholds is evaluated against a more recent cohort of students. Lastly, statistical methods for evaluating the performance of the simple threshold approach are discussed. In Section 4, the results and interpretations of the statistical analyses are presented, and an intervention process, which has recently been employed for these programs, is suggested. The paper is concluded in Section 5 with some closing remarks.

2 The extended data science programs

The VC presents three data science related Bachelor of Science degree programs with different specializations. The standard programs take a minimum of 3 years to complete, where students have very busy schedules throughout the durations of the programs. The core modules of these degrees are centered around mathematics, statistics, and programming, with a focus on applications to business and finance. Modules on economics, accounting and business ethics are also included in all three programs. The specializations include degrees in financial mathematics (FM), quantitative risk management (QRM), and business analytics (BA).

The FM program includes additional modules on more advanced mathematics, covering topics such as multivariate calculus and real analysis, as well as some modules on risk management. The QRM program focuses more on risk management courses, including topics such as investment management, bank risk management, financial markets, and financial risk management. The BA program incorporates several additional modules on programming, covering topics such as object-oriented programming, data structures and algorithms, databases, and decision support systems.

Since these data science programs are mathematically demanding, the most important admission requirement is that applicants should have a mark of at least 70% for mathematics in Grade 12 (their final year of high school). However, in 2014, extended programs for these degrees were introduced to attract more students to the programs on the VC of the NWU. In addition to a lower admission point score, applicants need a mark of at least 50% for mathematics in grade 12 to qualify for the extended programs. The reader is referred to North-West University (2025a) for the complete admission requirements, detailed descriptions of each data science program, and the content of the modules included in each program. A discussion on the academic preparedness of students and the performance of the first two cohorts of students in all extended programs presented at the VC is provided in Du Plessis and Gerber (2012).

The extended data science programs at the VC were developed using certain principles of traditional bridging programs. The extension of the three-year data science programs to four-year programs could be seen as degrees for which there are prolonged bridging programs. However, there are some key differences and features.

Firstly, admission to the extended programs is not contingent on the student's performance in a test after completion of a short bridging program. The purpose is to allow students who do not qualify for the standard programs to enroll for these degrees via the extended programs, where they will build up the necessary skills to participate in the remainder of the programs.

Secondly, the degrees are extended by 1 year, where the first-year mathematics and statistics modules of the standard programs are split over the first 2 years in the extended programs. Additional basic mathematics and statistics modules are used in the extended programs to bring students who did not qualify for the standard programs up to speed. Thus, the programs are designed such that the core modules are aligned when extended program students enter their third year and standard program students enter their second year.

Lastly, the extended programs are in no way seen as inferior to the standard programs. Since all work covered in the standard programs is also covered in the extended programs, students graduating from the different programs should be equipped with the same skillset. Furthermore, students from the different programs are treated equally in terms of applications for postgraduate studies in these data science programs.

There have been many success stories from these extended programs. However, an increased number of student dropouts has been observed in recent years, which warrants a thorough investigation into the success of the programs. The recent implementation of proper procedures to terminate the studies of repeatedly underperforming students in both the extended and standard programs further motivates the need for such an investigation.

3 Methodology and establishment of performance thresholds

3.1 Investigation into student performance

In this section, we present an analysis of the performance of students enrolled in the standard and extended data science programs at VC. The first part of the analysis focuses on key indicators such as graduation and dropout rates for both the standard and extended

TABLE 1 Categorization of students for the standard programs.

Standard programs	Graduates	Dropouts	Ongoing	Total
Financial mathematics	24	11	4	39 (53.4%)
Quantitative risk management	12	13	0	25 (34.2%)
Business analytics	5	4	0	9 (12.3%)
Total	41 (56.2%)	28 (38.4%)	4 (5.5%)	73 (100%)

TABLE 2 Categorization of students for the extended programs.

Extended programs	Graduates	Dropouts	Ongoing	Total
Financial mathematics	7	16	9	32 (61.5%)
Quantitative risk management	7	4	1	12 (23.1%)
Business analytics	3	2	3	8 (15.4%)
Total	17 (32.7%)	22 (42.3%)	13 (25.0%)	52 (100%)

programs. This is followed by an evaluation of the number of years it took students to obtain their degrees. Using box-and-whisker plots, we examine the performance in core mathematics and statistics modules to serve as the foundation for establishing performance thresholds, which are motivated and discussed in Section 3.2.

The dataset for this analysis consists of the registration information and academic performance of all students who registered for the data science programs between 2014 and 2018. The performance of these students was considered up to the end of 2022 to allow at least the minimum qualification time for the 2018 registrants. The data was obtained from the Integrated Planning & Strategic Intelligence department at the NWU. It should be noted that the data represents complete student records, where the module marks for all completed modules are recorded. The reflected marks represent finalized module marks and have thus already undergone validation and approval from the respective educators for each module. After the extraction of the data, simple data cleaning steps are performed. These data cleaning steps include merging or elimination of duplicate records and, where necessary, correcting administrative inaccuracies from alternative student records. The extracted cohort consists of 125 students in total, where the performance of the entire cohort is considered in this section. Of the 125 students, 73 were registered for the standard programs and 52 were registered for the extended programs.

Students were categorized into three groups, based on their progression in the programs. The first group consisted of the students who graduated from the programs. The second group consisted of students who dropped out of the programs, whether they discontinued their studies or switched to another program presented at the VC. Dropouts are thus defined as students who did not complete any of the data science programs and are also not still busy with any of these programs. The last group consisted of ongoing students, which are those students who were still enrolled in the programs at the end of 2022.

Tables 1, 2 display the categorization of the students for the standard and extended programs, respectively. The FM program is by far the most popular for both the standard and extended programs, followed by the QRM program. Overall, more students register for the standard programs than for the extended programs. Clear differences in the graduation and dropout rates can be observed between the programs. The dropout rate for the extended FM program is almost double that of the standard FM program. The opposite holds for the QRM programs, where the dropout rate for the extended QRM program is much lower than that of the standard QRM program. The overall dropout rates for the standard and extended programs are very similar, but this could be attributed in part to the higher proportion of ongoing students in the extended programs. The higher proportion of ongoing students in the extended program also explains the much lower overall graduation rate for these programs.

The general findings suggest the need for support and intervention, particularly in addressing the high dropout rates, to improve program efficiency. We proceed with the analysis by examining how long it takes for students to graduate from the programs. Table 3 displays the time taken for students to graduate, measured in terms of the minimum number of years required to graduate from the programs (i.e., 3 years for the standard programs and 4 years for the extended programs). Note that most students in the extended programs do not complete their degrees in the minimum required time. Excluding the ongoing students in the dataset, only 25 of the 108 (23.1%) students from both the standard and extended programs graduate in the minimum required time. This extension of study years often leads to even more financial difficulty, since most of the students are subsidized through government bursaries with performance requirements. The need for a more in-depth look at student performance is motivated, where possible reasons why students are struggling in their study progression should be investigated in future research.

We continue our analysis by assessing the performance of students in the core mathematics and statistics modules within these programs. These modules were identified as core modules for two reasons. The first is that these modules consist of the most important technical work that data science students need to master. The second is that these modules serve as foundations for further modules that students often struggle with (motivated by discussions with various stakeholders in the programs). The module code and module name of the identified core modules are provided in Table 4. The reader is referred to the NWU yearbooks (North-West University, 2025a) for more details on these modules, where the general module outcomes for each module are described. The yearbooks also contain detailed descriptions of the full curriculums, student progression, and prerequisites for the standard and extended programs.

TABLE 3 Time taken to graduate.

Program	Minimum required time	1 additional year	2 + additional years	Total
Standard programs	21	13	7	41 (70.7%)
Extended programs	4	12	1	17 (29.3%)
Total	25 (43.1%)	25 (43.1%)	8 (13.8%)	58 (100%)

TABLE 4 Description of core modules.

Module code	Module description
WISS121	Introduction to Mathematics II (Extended)
STTF125	Introductory Statistical Inference (Extended)
MTHS111	Introductory Algebra and Calculus I
MTHS121	Introductory Algebra and Calculus II
STTN215	Probability and Sampling Theory
STTN225	Statistical Inference and Data Analysis
MTHS211	Multivariable Calculus I
MTHS222	Linear Algebra II

3.2 Establishing mark thresholds for at-risk identification

The marks obtained for the core mathematics and core statistics modules are presented as box-and-whisker plots in Figures 1, 2, respectively. Note that the marks for each core module are split between graduates and dropouts, where the marks for ongoing students are not considered. These plots provide insight into the central tendency, spread, and presence of outliers in the dataset. The upper and lower whiskers of the box-and-whisker show the range of marks, with outliers indicated by dots, while the box itself illustrates the interquartile range, with the median indicated by the middle bar. As expected, a clear difference in the distribution of marks for graduates and dropouts can be observed. We used key points in the distributions, such as the quartiles, to inform the establishment of thresholds. The goals of these thresholds are to monitor student progression through the program and to assist lecturers in advising students based on their academic performance in these critical modules.

To inform a starting point for setting a threshold for each module, we considered the first quartile (Q1) of the graduates and the third quartile (Q3) of the dropouts. The reasoning behind this approach was to capture the majority of graduates with a type of lower bound, defined by Q1 of the graduates' marks, while trying to exclude the majority of dropouts with a type of upper bound, defined by Q3 of the dropouts' marks. Consequently, the logic was that the bottom 25% of students in the graduate group may have been at risk of dropping out, and the top 25% of students in the dropout group may have had a chance of improving their performance to graduate from the programs. As an example, consider the module MTHS111. From Figure 1, the value of Q1 for graduates is 65% and the value of Q3 for dropouts is 62.5%. Using these values as a basis for discussions with stakeholders, the threshold for MTHS111 was set at 65%. Additionally, groupings of the core modules were created to define more holistic thresholds, with small adjustments in some cases. For all the holistic thresholds, the average mark of the grouped modules is considered, and a student is allowed a maximum downward deviation of 5% for one module in each grouping.

Using this approach, together with small adjustments made in consultation with various lecturers involved in these modules, the thresholds given in Table 5 were recommended. The small adjustments involve discretionary rounding to align the thresholds with either the lower or upper 5% band based on the expert opinion of the lecturers involved. This was done in cases where there is a slight difference between Q1 for graduates and Q3 for dropouts in order to ease communication and applicability of the thresholds. For example, first-year standard program and second-year extended program students should achieve at least 65% for MTHS111 and 60% for MTHS121. Rather using the holistic thresholds, an average of 60% for the two modules should be achieved, where a maximum downward deviation of 5% is allowed for one of the two modules. That is, if a student achieves a mark of 55% for MTHS111, the student will have to achieve a mark of 65% in MTHS121 to have an average for this group of 60%.

To test the validity of the established thresholds, we applied them to the students in our dataset, excluding those students who dropped out of the programs before the end of their first-year studies. This helped us evaluate how effective the thresholds are at identifying at-risk students. In Figure 3, we observe distinct differences in the adherence to the thresholds between dropouts and graduates, in both the standard and extended programs. We observe in the standard program that a higher percentage of graduates met all the criteria compared to dropouts, indicating the efficacy of the thresholds in predicting possible graduates. In the extended program, there is a similar trend, while the percentage of graduates meeting all criteria is somewhat lower than that of the standard program graduates. Many students who miss one of the thresholds still graduate, but it is clear that students missing two or more thresholds are much more likely to drop out of the programs. These insights highlight the practical value of the established thresholds in distinguishing between graduates and dropouts in both the standard and extended programs.

3.3 Further evaluation of the thresholds on a more recent cohort

In this section, the efficacy and validity of the thresholds are further explored by applying them to the next cohort of students. The aim of this extension is to determine whether there has recently been an increase in the number of dropouts, which would further motivate the need for interventions. The validation dataset consists of the 17 ongoing students from the previous dataset, as well as students who registered for the programs between 2019 and 2023. The performance of the students up to the end of 2023 was considered. The validation dataset consisted of 116 students, of which 47 were enrolled in the standard programs and 69 were enrolled in the extended programs.

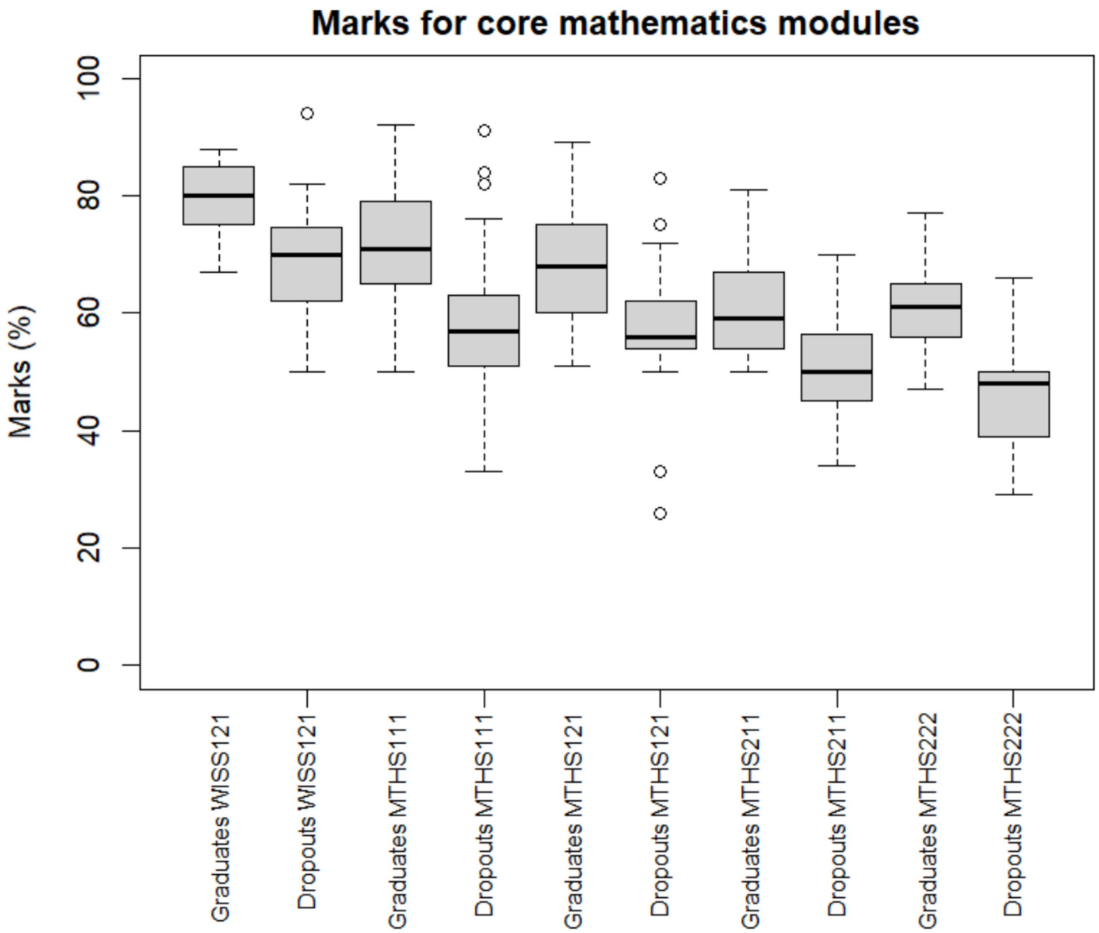

FIGURE 1
Boxplots of the distribution of marks for the core mathematics modules.

The extended programs are now attracting more students than the standard programs. This highlights the importance of the extended programs as a tool for increased student intake, since these students would not have met the requirements for the standard programs.

The same data cleaning adjustments as described in Section 3.1 were made and the students were again categorized as graduates, dropouts, or ongoing students. The categorizations of the validation dataset for the standard and extended programs are given in Tables 6, 7, respectively. The student numbers in this dataset are more evenly distributed between the three programs, compared to the dataset used in Section 3.1. Note that the numbers of ongoing students in the validation dataset are much higher, due to the performance being measured up to the end of 2023 while all students registered up to 2023 are considered. Thus, many students have only been enrolled for one, two, or three (for extended program students) years and cannot complete their degrees in this timeframe.

Although the ratio of graduates to dropouts will significantly improve as ongoing students complete their degrees, it is still of great concern. It is important to note that proper procedures were implemented from 2023 to terminate the studies of students who had been busy with their degrees for too long and did not show sufficient progress throughout 2023. Discussions were also held with students who performed very poorly, where some of them were advised to convert to a more suitable program at the NWU, so that they could still obtain a degree before leaving the university. These factors also contributed to the higher dropout rates.

For the validation dataset, the time taken for students to graduate is displayed in Table 8. There is a clear deterioration with respect to the time it takes students to graduate in this more recent dataset. The significant increase in the percentage of students who take 2 or more additional years to complete their degrees is of particular concern. Although the guidance accompanying the implementation of the derived thresholds might have a positive impact on the time taken to graduate, further investigation into the admission requirements for the programs might also be considered.

Next, the performance of students against the thresholds set in Section 3.2 is investigated. Recall that there are three thresholds set for the standard program students and four for the extended program students. Figure 4 displays the performance of graduates and dropouts for both programs against the thresholds.

It is interesting to see such large differences in threshold performance between the original dataset and the validation dataset. There is a higher percentage of graduates who miss more than one of the thresholds, supporting the concerns about declining student performance over the past few years. For the dropouts, however, there is a clear shift towards missing fewer thresholds before dropping out of the programs. This

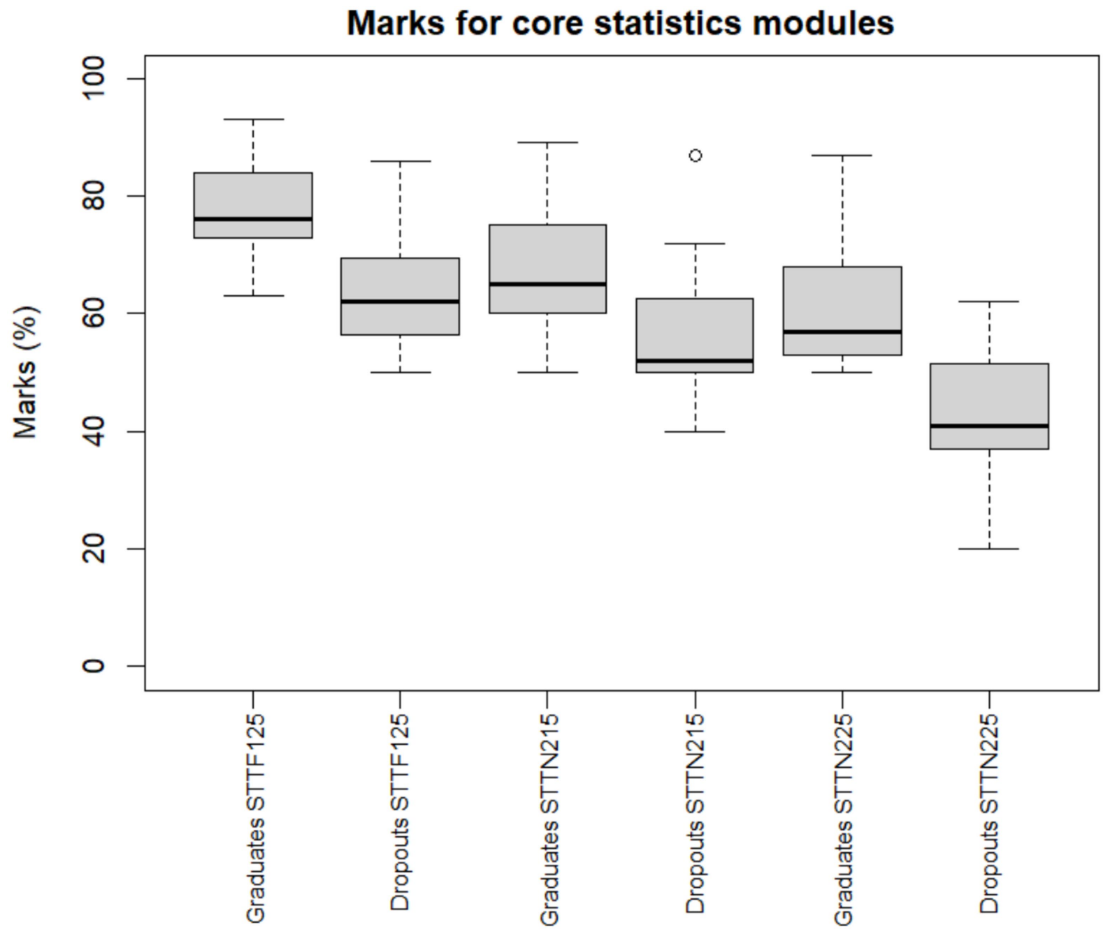

FIGURE 2
Boxplots of the distribution of marks for the core statistics modules.

TABLE 5 Recommended thresholds for core modules.

Year level	Module	Average
Extended first year	WISS121 (70%) and STTF125 (70%)	70%
Standard first year Extended second year	MTHS111 (65%) and MTHS121 (60%)	60%
Standard second year Extended third year	STTN215 (60%) and STTN225 (55%)	55%
Standard second year Extended third year	MTHS211 (55%) and MTHS222 (55%)	55%

could be indicative that the interventions and procedures recently implemented are effective to some degree. Dropouts who miss no thresholds might also be attributed to financial or other factors.

3.4 Assessing the validity of the threshold approach

To establish the validity of our simple threshold approach, several statistical methods and tests can be considered. For this analysis, missing a threshold is defined as a *breach,* where the number of breaches and dropout status for each student in the standard and extended programs are considered. Similar to the previous evaluations, the students of the standard and extended programs are considered separately. The methods considered can be used to determine whether there is a relationship between the number of breaches and dropout status and whether the number of threshold breaches can be used as an indicator of at-risk students.

First, we will consider two important formal tests. The tie-corrected Mann–Whitney test (see, for example, Lehmann and D'Abrera, 2006) can be used to determine whether two groups differ in their distributions. In the context of our study, this test can be used to assess whether the number of breaches differs between the graduates and dropouts. The tie-corrected test is used since the number of breaches variable can only take on a few distinct values, resulting in many ties in the data. Fisher's exact test (see, for example, Mehta and Patel, 1983) can be used to determine whether there is an association between two categorical variables. This test is more appropriate than a chi-square test, since exact p-values can be calculated for uneven group distributions and small sample sizes. For our study, this test can be used to establish if there is an association between the number of threshold breaches and dropout status.

Next, we will consider a logistic regression model that assesses the probability of being a dropout as a function of the number of breaches, which is given by

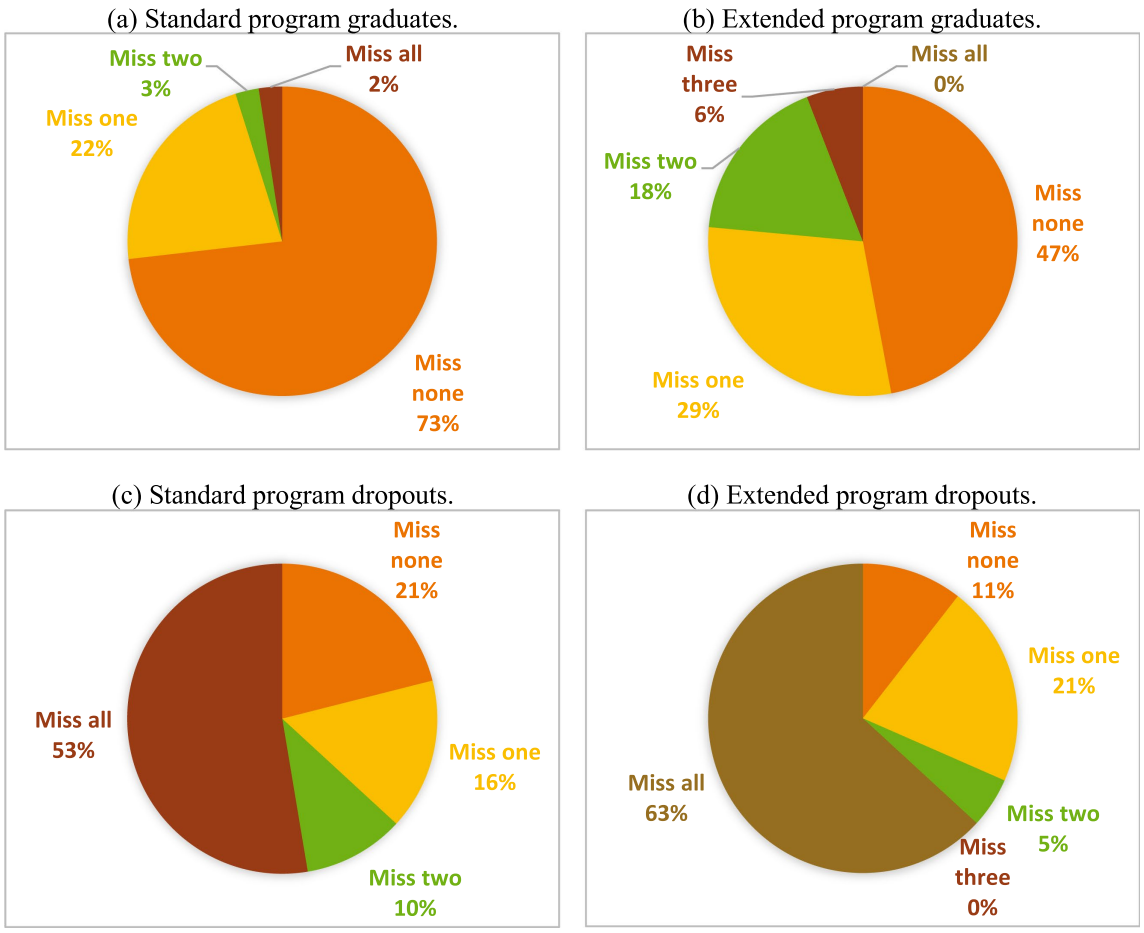

FIGURE 3
The proportion of students who missed a certain number of thresholds for the standard and extended programs. (a) standard program graduates, (b) extended program graduates, (c) standard program dropouts, (d) extended program dropouts.

TABLE 6 Categorization of the more recent cohort of students for the standard programs.

Standard programs	Graduates	Dropouts	Ongoing	Total
Financial mathematics	5	6	7	18 (38.3%)
Quantitative risk management	1	2	13	16 (34.0%)
Business analytics	3	3	7	13 (27.7%)
Total	9 (19.1%)	11 (23.4%)	27 (57.4%)	47 (100%)

TABLE 7 Categorization of the more recent cohort of students for the extended programs.

Extended programs	Graduates	Dropouts	Ongoing	Total
Financial mathematics	4	7	15	26 (37.7%)
Quantitative risk management	0	5	15	20 (29.0%)
Business analytics	2	7	14	23 (33.3%)
Total	6 (8.7%)	19 (27.5%)	44 (63.8%)	69 (100%)

$$\text{logit}\left(\text{P}\left(\text{Dropout}\right)\right) = \beta_0 + \beta_1 \cdot \text{Breaches}.$$

If the predictor coefficient β_1 is statistically significant, it would mean that the number of breaches is a suitable variable for identifying at-risk students. Furthermore, β_1 should be positive such that more breaches relates to a higher likelihood of dropping out.

Lastly, we will perform a sensitivity/specificity analysis, where the number of breaches is considered to classify students as dropouts or graduates. We consider all possible classification cases due to a limited number of possible classification cut-offs, where a student is classified as a dropout if they had a certain number of breaches or more. That is, sensitivity represents the proportion of dropouts correctly classified as dropouts under the specified classification cut-off, and specificity represents the proportion of correctly classified graduates for the same

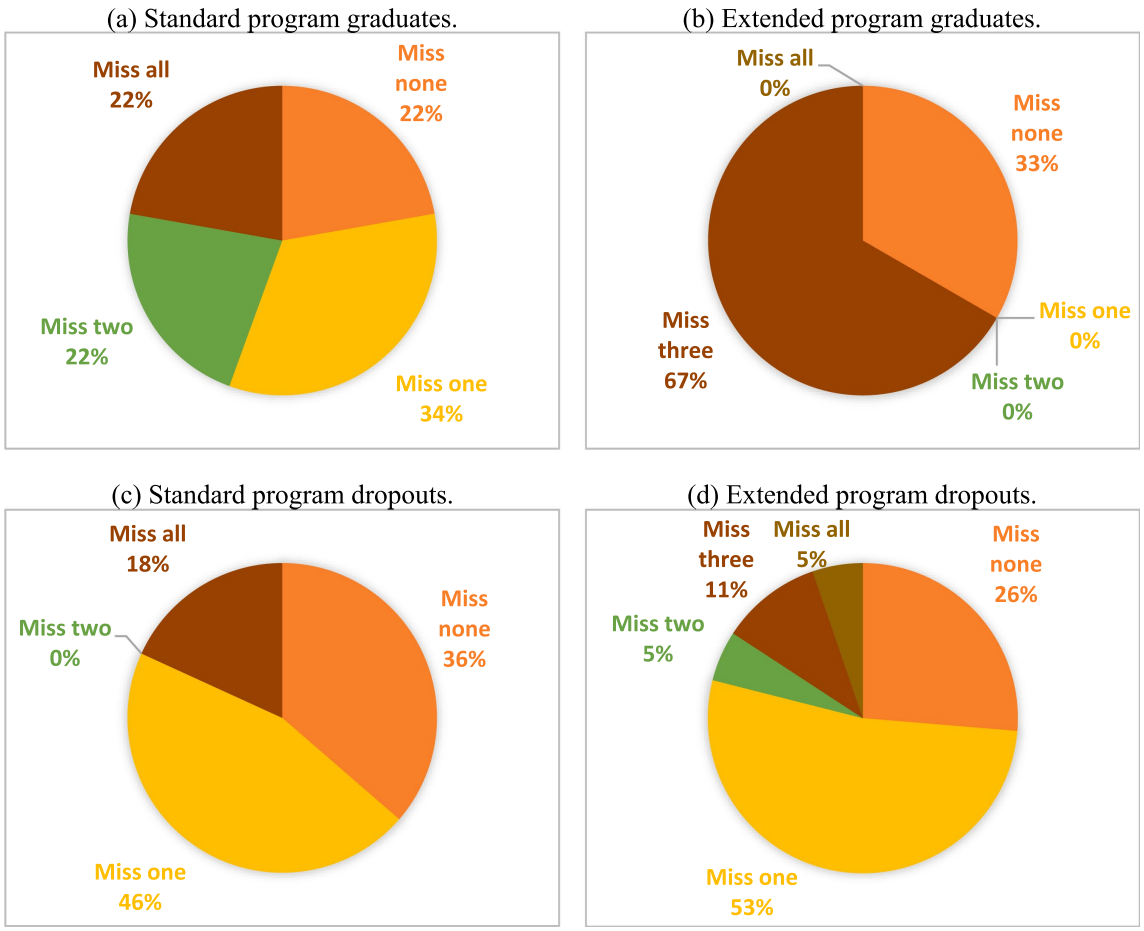

FIGURE 4
The performance of the more recent cohort of students against the thresholds in terms of the number of threshold breaches. (a) standard program graduates, (b) extended program graduates, (c) standard program dropouts, (d) extended program dropouts.

TABLE 8 Time taken to graduate for the more recent cohort of students.

Program	Minimum required time	1 additional year	2 + additional years	Total
Standard programs	2	4	3	9 (60.0%)
Extended programs	1	2	3	6 (40.0%)
Total	3 (20.0%)	6 (40.0%)	6 (40.0%)	15 (100%)

classification cut-off. The accuracy is also considered, which represents the overall proportion of correctly classified students under the specified classification cut-off. The values of these classification metrics can then be considered to determine an appropriate cut-off (number of breaches) to classify a student as an at-risk student. From the sensitivity/specificity analysis, we can also estimate the AUC. This metric indicates whether a model is effective at differentiating between two classes, which are the dropout and graduate classes in our case.

4 Results and discussion

4.1 Statistical analysis of the thresholds

In this section, we provide the results of the statistical analysis of the thresholds established in Section 3.2 and provide insightful discussions. The analysis is based on the data of the first cohort of students, which was used to establish the thresholds, i.e., all first-time entries into the programs between 2014 and 2018, monitored up until the end of 2022 (see Section 3.1. for further information).

First, let us consider the two formal tests. The p-values of the Mann–Whitney test for the standard program and extended programs are, respectively, <0.0001 and 0.0273. This indicates a statistically significant difference in the number of breaches for the graduates and dropouts, at a 5% significance level, for both programs. That is, dropouts typically have a higher number of breaches than graduates, suggesting that the number of breaches is a meaningful measure in terms of identifying at-risk students. Fisher's exact test is applied to examine the association between the number of breaches and dropout status. For the standard program, the p-value is 0.0002 and for the extended program, the p-value is 0.0312. Both are significant at a 5% significance level, indicating that the number of breaches and dropout

TABLE 9 Logistic regression results.

Model details	Standard program	Extended program
$\hat{\beta}_0$	−1.7228	−0.8020
$(p-\text{value})$	(<0.0001)	(0.1386)
$\hat{\beta}_1$	1.3954	0.7502
$(p-\text{value})$	(0.0014)	(0.0405)
Odds ratio	4.0367	2.1174
P(Dropout|Breaches = 0)	0.1515	0.3096
P(Dropout|Breaches = 1)	0.4189	0.4871
P(Dropout|Breaches = 2)	0.7442	0.6678
P(Dropout|Breaches = 3)	0.9215	0.8098
P(Dropout|Breaches = 4)	NA	0.9001

TABLE 10 Sensitivity/specificity analysis results.

Classification cut-off	Sensitivity	Specificity	Accuracy
Standard program			
≥ 0 Breaches	1.0000	0.0000	0.3167
≥ 1 Breaches	0.7895	0.7317	0.7500
≥ 2 Breaches	0.3158	0.9512	0.7500
≥ 3 Breaches	0.0526	0.9756	0.6833
Extended program			
≥ 0 Breaches	1.0000	0.0000	0.5278
≥ 1 Breaches	0.8947	0.4706	0.6944
≥ 2 Breaches	0.4210	0.7647	0.5833
≥ 3 Breaches	0.2105	0.9412	0.5556
≥ 4 Breaches	0.2105	1.0000	0.5833

status are not independent in either program. This suggests that the number of breaches has a significant association with the dropout status, further supporting the use of the number of breaches for identifying at-risk students.

Next, we consider a logistic regression model using only the number of breaches as a predictor. Table 9 provides details on the fitted logistic regression models for the standard and extended programs. For both models, the coefficient of the predictor (number of breaches) is positive and statistically significant at a 5% significance level. This supports our previous findings in that the number of breaches can be used as a simple and effective indicator for identifying at-risk students. Furthermore, the odds ratios indicate that each additional breach increases the odds of dropping out by factors of approximately 4 and 2, respectively, for the standard and extended programs. Both fitted models show a considerable probability of dropping out of the programs even with a single breach and notably higher probabilities for two or more breaches. This indicates that identifying a student as at-risk when two breaches has occurred may be too late, highlighting the need for early interventions. Lastly, the probability of dropping out of the extended program while having zero breaches is double that of the standard program. This may suggest that the extended program does not adequately prepare students for the alignment with the standard program (third year of the extended program and second year of the standard program) or that the admission requirements need to be reconsidered.

Lastly, we consider the sensitivity/specificity analysis for both programs under the possible classification cut-offs when using the number of breaches to classify students as graduates or dropouts. These results are presented in Table 10. Considering that the problem at hand is identifying students at risk of dropping out of the programs, it can be argued to some extent that identifying dropouts correctly is more important than identifying graduates correctly. This is due to the consequences of identifying a potential graduate as an at-risk student being less severe than the other way around. For the standard and extended programs, the sensitivity drops from 0.7895 to 0.3158 and from 0.8947 to 0.4210, respectively, between the classification cut-offs of "≥ 1 Breaches" and "≥ 2 Breaches." These decreases in sensitivity further support that identifying students as at-risk after two or more breaches might be too late and that early interventions are required. Based on these findings, it is recommended that students should be flagged as at-risk as soon as they have one breach, so that early interventions can be made. A satisfactory specificity is also achieved for the classification cut-off of one or more breaches, while taking into account that correctly classifying dropouts is more important. The accuracy metric also supports early at-risk identification, where the highest accuracy is achieved when using a classification cut-off of one or more breaches, for both the standard and extended programs. The AUC for the standard and extended programs are 0.7811 and 0.7074, respectively. This indicates satisfactory discrimination between dropouts and graduates when using only the number of breaches to identify at-risk students.

In summary, the findings of the statistical analysis support the need for the early identification of at-risk students. Most of the methods used indicate that students should be classified as at-risk after their first breach. This also means that early interventions are necessary. Considering the changes in the sensitivity between classification cut-offs, it may be important to consider light interventions after one breach and more intensive interventions after two or more breaches.

In the next section, an intervention process for at-risk students in the data science programs is presented. This intervention process has recently been employed at the university in an attempt to improve student retention and provide students with relevant advice throughout their academic life cycles.

4.2 A recently introduced intervention process

To be able to address the possible issues students have in terms of factors contributing to dropping out, a method is needed to flag the students who are struggling. As shown in Section 4.1, the number of threshold breaches can be effectively used to identify at-risk students. The statistical analysis supports classifying students as at-risk if one breach occurs, leading to the need for interventions early in the students' academic life cycles. These thresholds can be used to guide students better so that they can complete a qualification in the shortest possible time. The following intervention process is suggested and has

recently been employed for students enrolled in the data science programs at the VC:

1 First threshold breach: Group discussions are held with students after not meeting a threshold for the first time. During these discussions, the students are advised on the required personal commitment and effective time management expected to obtain one of the data science degrees. The students are also given an opportunity to raise any collective concerns and issues, such as problems with certain modules, lecturers, and the general academic environment. Lastly, students are made aware of student counseling services offered by the university and are encouraged to attend an organized group awareness session.
2 Additional threshold breaches: A one-on-one discussion is held with a student who misses multiple thresholds. During this discussion, causes of poor academic performance are discussed with the student and guidance/assistance is provided by the lecturers, where possible. In the case of more personal issues, the student is referred to the university's student counseling services.
3 Threshold breach and failing core modules: After a student fails a core module, a one-on-one discussion is held with the student. During this discussion, the student is made aware of the consequences of repeated underperformance, and a first formal warning is issued to the student. The lecturers also facilitate a conversation on alternative study options to consider, where it may be in the best interest of the student to change to a less demanding degree rather than not obtain a degree at all. Should the student again fail core modules, the exercise is repeated, where a second and final formal warning is issued to the student. At this stage, it is made clear that failing any more modules would result in the termination of the student's studies at the university.
4 Failing modules after the final warning: Should a student not meet the specific conditions set out in the final formal warning, or fail any more modules, the student's studies within the faculty are terminated in accordance with the university rules.

The above interventions would of course be tailored to the specific circumstances of the students, to achieve the best possible solution for them. The aim of the intervention process is to prevent any students from reaching the point of their studies being terminated (as set out in the university rules), as this would result in the student leaving the university with no formal qualification. The current cohort of ongoing students will be monitored to investigate the efficacy of the proposed intervention process.

5 Conclusion

In this paper, an evaluation of student performance and graduation rates was conducted for data science related programs at the VC. The comparison focused on the differences in performance between standard and extended data science programs. It is evident that the dropout rates of the extended programs are higher than those of the corresponding standard programs. From an initial cohort of students, thresholds were established to identify at-risk students in the programs. These thresholds were then applied to a more recent cohort of students, and concerns regarding declining student graduation rates were confirmed.

We can now consider whether our aims for the research study have been met. A statistical analysis was performed, which showed that a simple threshold approach, which is tailored to these specific programs, is effective in identifying at-risk students. It was also established that there is a significant difference in the graduation rates between the standard and extended programs, which indicates that a revision of the extended program framework for these programs might be required. Furthermore, the analysis showed that early identification of at-risk students is necessary, where we recommend classifying students as at-risk as soon as they have one threshold breach. The need for early interventions is also highlighted in this analysis. An intervention process was suggested that aims to improve student retention and provide students with appropriate advice throughout their academic life cycles. We hope that the discussions in this paper will encourage other educators to consider the role and viability of extended programs, as an alternative to short transitional programs, at their own institutions.

Although the study presents promising results for using the thresholds to identify at-risk data science students, several limitations should be acknowledged. First, the analysis is based on data from a single South African university, which may limit the generalizability of the findings to other institutions or contexts. The thresholds are tailored to our specific program and institution, but a similar process can be followed to develop a corresponding framework for other programs and institutions. Due to the limited literature and studies on extended programs, it is also not possible to thoroughly compare our findings to those of other international studies. Most of the existing literature focuses on the evaluation of bridging programs or short transitional programs, which are inherently different from an extended degree program (as highlighted in Section 2).

Additionally, although the threshold approach offers a simplistic at-risk identification method, it might not capture complex relationships between performance indicators and academic risk. The study also primarily considers academic performance in core modules, potentially overlooking other factors such as socio-economic background, mental health, and support systems, which could influence student success. Another limitation is the static nature of the thresholds, which may be fixed for a long period of time before considering a re-evaluation. Changes in the program structure, especially in the core modules, could necessitate a complete redesign of the approach without sufficient data or knowledge to motivate new thresholds.

Future research could consider expanding the dataset to include multiple institutions to enhance the robustness and generalizability of the thresholds. This would require an evaluation of other institutions' program structures to identify equivalent core modules to use for the benchmarking approach. Such a study would also be limited to South African universities to ensure consistency in terms of the educational framework and background of participants. Generalizability of the threshold approach may, however, be limited due to module content and difficulty level differences across institutions.

The primary direction of future research on at-risk identification for our programs is to consider machine learning models. The use of more complex models could provide deeper insights into the multifactorial nature of academic risk. From the literature review in Section 1, it is clear that at least LR, CT, RF, and SVM should be considered. Furthermore, many studies include socio-demographic and online behavioral features to complement academic performance features. The NWU has a learner management system (LMS) for module administration, content

distribution, and reporting, which can provide valuable insights on student engagement with module content and resources. The university also has certain socio-demographic information on students, which can be requested for research purposes, subject to ethics committee approval. Lastly, the intervention process is structured to gain insights into poor academic performance, where additional socio-economic and behavioral features can be created and tracked via group and individual discussions with students. The predictive performance of the machine learning models can then be compared to that of our simple threshold approach.

Lastly, a longitudinal study that track the long-term outcomes of students who receive interventions would help to assess the true impact of the proposed intervention process. This would involve monitoring dropout rates for the standard and extended programs over several years, while consistently adhering to the proposed interventions. Exploring non-academic indicators and student feedback could assist in refining the intervention process to be more holistic and responsive to student needs. However, such a study would require an extensive timeframe to properly evaluate the impact and implement changes over different groups of students. The study by Atindama et al. (2025) could serve as a valuable source for such a longitudinal study, where the efficacy of our intervention process and possible improvements to the intervention process can be considered.

As a closing note, consideration should be given to the fact that there is a significant proportion of graduates from the extended programs who would not otherwise have been given the opportunity to study towards obtaining a data science degree. Since there are many success stories, discontinuing the extended programs without proactive efforts to improve the graduation rates would demonstrate a lack of foresight, particularly considering the scarcity of STEM graduates in South Africa. However, it may be necessary to revise the admission requirements for the programs if the suggested intervention process is not effective.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: confidential student data from a specific institution are used and only summarized results on the data are discussed in this paper. Requests to access these datasets should be directed to Neill.Smit@nwu.ac.za.

Ethics statement

The studies involving humans were approved by Faculty of Natural and Agricultural Sciences Ethics Committee of the North-West University. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and institutional requirements.

Author contributions

NS: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. ZO: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. LM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Al-Shabandar, R., Hussain, A. J., Liatsis, P., and Keight, R. (2019). Detecting at-risk students with early interventions using machine learning techniques. *IEEE Access* 7, 149464–149478. doi: 10.1109/ACCESS.2019.2943351

Atindama, E., Ramsdell, M., Wick, D. P., Mondal, S., and Athavale, P. (2025). Impact of targeted interventions on success of high-risk engineering students: a focus on historically underrepresented students in STEM. *Front. Educ.* 10:1435279. doi: 10.3389/feduc.2025.1435279

Babalola, B. T., Awe, O. O., and Adarabioyo, M. I. (2022). "Challenges of statistics education that leads to high dropout rate among undergraduate statistics students in developing countries" in Promoting statistical practice and collaboration in developing countries, ed. O. O. Awe, K. Love and E. A. Vance (New York, NY: Chapman and Hall/CRC), 75–83.

Beitelmal, W. H., Littlejohn, R., Okonkwo, P. C., Hassan, I. U., Barhoumi, E. M., Khozaei, F., et al. (2022). Threshold concepts theory in higher education – introductory statistics courses as an example. *Educ. Sci.* 12:748. doi: 10.3390/educsci12110748

Borna, M. R., Saadat, H., Hojjati, A. T., and Akbari, E. (2024). Analyzing click data with AI: implications for student performance prediction and learning assessment. *Front. Educ.* 9:1421479. doi: 10.3389/feduc.2024.1421479

Bradford, B. C., Beier, M. E., and Oswald, F. L. (2021). A meta-analysis of university STEM summer bridge program effectiveness. *CBE Life Sci. Educ.* 20:ar21. doi: 10.1187/cbe.20-03-0046

Brady, A., and Gallant, D. (2021). Stem bridge program: underrepresented minority students' perceptions of Louis stokes Alliance for minority participation program impact. *J. Coll. Sci. Teach.* 50, 57–62. doi: 10.1080/0047231X.2021.12290534

Cao, L. (2017). Data science: a comprehensive overview. *ACM Comput. Surv.* 50:43. doi: 10.1145/3076253

Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., and Araújo, R. D. (2022). Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. *Rev. Bras. Inform. Educ.* 30, 379–395. doi: 10.5753/rbie.2022.2518

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *Int. Stat. Rev.* 69, 21–26. doi: 10.1111/j.1751-5823.2001.tb00477.x

Council on Higher Education. (2014) Framework for institutional quality enhancement in the second period of quality assurance. Available online at: https://www.che.ac.za/sites/default/files/publications/QEP%20Framework%20Feb%202014.pdf (Accessed June 18, 2025).

Cummings, K. D., and Smolkowski, K. (2015). Selecting students at risk of academic difficulties. *Assess. Eff. Interv.* 41, 55–61. doi: 10.1177/1534508415590396

De Veaux, R. D., Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., et al. (2017). Curriculum guidelines for undergraduate programs in data science. *Ann. Rev. Stat. Appl.* 4, 15–30. doi: 10.1146/annurev-statistics-060116-053930

De Veaux, R. D., Hoerl, R., Snee, R., and Velleman, P. (2022). Towards holistic data science education. *Stat. Educ. Res. J.* 21:2. doi: 10.52041/serj.v21i2.40

Du Plessis, L., and Gerber, D. (2012). Academic preparedness of students – an exploratory study. *J. Transdiscipl. Res. South. Afr.* 8, 81–94. doi: 10.4102/td.v8i1.7

Glassdoor (2023). 50 best jobs in America for 2022. Available online at: https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm (Accessed August 20, 2024).

Gordanier, J., Hauk, W., and Sankaran, C. (2019). Early intervention in college classes and improved student outcomes. *Econ. Educ. Rev.* 72, 23–29. doi: 10.1016/j.econedurev.2019.05.003

Jamjoom, M. M., Alabdulkreem, E. A., Hadjouni, M., Karim, F. K., and Qarh, M. A. (2021). Early prediction for at-risk students in an introductory programming course based on student self-efficacy. *Informatica* 45, 1–9. doi: 10.31449/inf.v45i6.3528

Jang, Y., Choi, S., Jung, H., and Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Educ. Inf. Technol.* 27, 12855–12889. doi: 10.1007/s10639-022-11120-6

Kalita, E., Alfarwan, A. M., El Aouifi, H., Kukkar, A., Hussain, S., Ali, T., et al. (2025). Predicting student academic performance using bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. *Front. Educ.* 10:1581247. doi: 10.3389/feduc.2025.1581247

Köhler, J., Hidalgo, L., and Jara, J. L. (2022). Using machine learning techniques to predict academic success in an introductory programming course in 2022 41st international conference of the Chilean computer science society (SCCC) (Chile: IEEE), 1–8.

Lehmann, E. L., and D'Abrera, H. J. (2006). Nonparametrics: Statistical methods based on ranks. New York: Springer.

LinkedIn (2022) Top 10 emerging and declining jobs in 2022. Available online at: https://www.linkedin.com/pulse/top-10-emerging-declining-jobs-2022-teamleasedigital/ (Accessed August 18, 2024).

Lowder, C., O'Brien, C., Hancock, D., Hachen, J., and Wang, C. (2022). High school success: a learning strategies intervention to reduce drop-out rates. *Urban Rev.* 54, 509–530. doi: 10.1007/s11256-021-00624-z

Mehta, C. R., and Patel, N. R. (1983). A network algorithm for performing fisher's exact test in r x c contingency tables. *J. Am. Stat. Assoc.* 78, 427–434. doi: 10.1080/01621459.1983.10477989

Moodley, P., and Singh, R. J. (2015). Addressing student dropout rates at South African universities. *Alternation* 17, 91–115. Available online at: https://alternation.ukzn.ac.za/Files/docs/22%20SpEd17/06%20Moodley%20F.pdf

Murphy, T. E., Gaughan, M., Hume, R., and Moore, S. G. (2010). College graduation rates for minority students in a selective technical university: will participation in a summer bridge program contribute to success? *Educ. Eval. Policy Anal.* 32, 70–83. doi: 10.3102/0162373709360064

Namgay, P., Wangdi, P., and Thinley, S. (2022). Designing and developing a data science programme in Bhutan in 2022 IEEE Frontiers in education conference (FIE). Sweden: IEEE.

North-West University (2025a) Faculty of Natural and Agricultural Sciences undergraduate yearbook. Availabel online at: https://studies.nwu.ac.za/studies/yearbooks (Accessed July 22, 2025).

North-West University. (2025b). The NWU rankings. Available online at: https://www.nwu.ac.za/rankings (Accessed July 22, 2025).

Ortiz-Lozano, J. M., Rua-Vieites, A., Bilbao-Calabuig, P., and Casadesús-Fa, M. (2020). University student retention: best time and data to identify undergraduate students at-risk of dropout. *Innov. Educ. Teach. Int.* 57, 74–85. doi: 10.1080/14703297.2018.1502090

Pek, R. Z., Özyer, S. T., Elhage, T., Özyer, T., and Alhajj, R. (2022). The role of machine learning in identifying students at-risk and minimizing failure. *IEEE Access* 11, 1224–1243. doi: 10.1109/ACCESS.2022.3232984

Pérez, L. X. (1998). Sorting, supporting, connecting, and transforming: intervention strategies for students at risk. *Community Coll. Rev.* 26, 63–78. doi: 10.1177/009155219802600105

Pillay, A. L., and Ngcobo, H. S. (2010). Sources of stress and support among rural-based first-year university students: an exploratory study. *S. Afr. J. Psychol.* 40, 234–240. doi: 10.1177/008124631004000302

Raines, J. M. (2012). FirstSTEP: a preliminary review of the effects of a summer bridge program on pre-college STEM majors. *J. STEM Educ.* 13, 22–29. Availabel online at: https://www.jstem.org/jstem/index.php/JSTEM/article/view/1682

Sarra, A., Fontanella, L., and Di Zio, S. (2019). Identifying students at-risk of academic failure within the educational data mining framework. *Soc. Indic. Res.* 146, 41–60. doi: 10.1007/s11205-018-1901-8

Ssempebwa, J., Eduan, W., and Mulumba, F. N. (2012). Effectiveness of university bridging programs in preparing students for university education: a case from East Africa. *J. Stud. Int. Educ.* 16, 140–156. doi: 10.1177/1028315311405062

Tishkovskaya, S., and Lancaster, G. A. (2012). Statistics education in the 21st century: a review of challenges, teaching innovations and strategies for reform. *J. Stat. Educ.* 20:641. doi: 10.1080/10691898.2012.11889641

United States Bureau of Labor Statistics. (2023). Occupational outlook handbook – data scientists. Available online at: https://www.bls.gov/ooh/math/data-scientists.htm#tab-1 (Accessed August 20, 2024).

Veerasamy, A. K., D'Souza, D., Apiola, M. V., Laakso, M. J., and Salakoski, T. (2020) Using early assessment performance as early warning signs to identify at-risk students in programming courses. 2020 IEEE Frontiers in education conference (FIE) (Sweden: IEEE).

Voulo, M., Evans, B., Hannon, G., Longenbach, S., and Spaen, B. (2024). Data science degree programs. Available online at: https://www.datascienceprograms.org/ (Accessed August 20, 2024).

Zakari, I. S. (2020). Promoting statistics in the era of data science and data-driven innovations. *Stat. Educ. Res. J.* 19, 226–237. doi: 10.52041/serj.v19i1.132