South African Journal of Information Management
ISSN: (Online) 1560-683X, (Print) 2078-1865

Exploring factors influencing academic literacy – A data-driven perspective

Authors:
Janus Roestenburg[1]
Cornelius J. Kruger[1]
Mariska Nel[2]
Zander Janse van Rensburg[3]

Affiliations:
[1]Department of Computer Science and Information Systems, Faculty of Natural and Agricultural Sciences, North-West University, Potchefstroom, South Africa

[2]School of Languages: Academic Literacy, Faculty of Humanities, North-West University, Mahikeng, South Africa

[3]Writing Centre, Faculty of Humanities, North-West University, Potchefstroom, South Africa

Corresponding author:
Zander Janse van Rensburg,
zander.jansevanrensburg@nwu.ac.za

Dates:
Received: 19 June 2023
Accepted: 16 Feb. 2024
Published: 29 Mar. 2024

How to cite this article:
Roestenburg, J., Kruger, C.J., Nel, M. & Janse van Rensburg, Z., 2024, 'Exploring factors influencing academic literacy – A data-driven perspective', *South African Journal of Information Management* 26(1), a1729. https://doi.org/10.4102/sajim.v26i1.1729

Background: Data science and machine learning have shown their usefulness in business and are gaining prevalence in the educational sector. In illustrating the potential of educational data mining (EDM) and learning analytics (LA), this article illustrates how such methods can be applied to the South African higher education institution (HEI) environment to enhance the teaching and learning of academic literacy modules.

Objectives: The objective of this study is to determine if data science and machine learning methods can be effectively applied to the context of academic literacy teaching and learning and provide stakeholders with valuable decision support.

Method: The method applied in this study is a variation of the knowledge discovery and data mining process specifically adapted for discovery in the educational environment.

Results: This study illustrates that utilising educational data can support the educational environment by measuring pedagogical support, examining the learning process, supporting strategic decision-making, and predicting student performance.

Conclusion: Educators can improve module offerings and students' academic acculturation by applying EDM and LA to data collected from academic literacy modules.

Contribution: This manuscript contributes to the field of EDM and LA by illustrating that methods from these research fields can be applied to the South African educational context and produce valuable insights using local data, providing practical proof of its feasibility and usefulness. This is aligned with the scope of this journal as it pertains to innovations in information management and competitive intelligence.

Keywords: educational data mining; learning analytics; academic literacy; machine learning; applied linguistics; student support; student success; academic acculturation.

Introduction

Modern education generates increasing amounts of learner data, which are underutilised to a larger extent. The increasing use of e-Learning methods in conjunction with learning management systems (LMS) creates large quantities of data (Romero & Ventura 2020). This is especially true in the academic literacy (AL) context – a service module through which the academic literacy skills of approximately 12 000 first-year students annually (six modules, two languages, offered as both distance and contact modules) are developed and improved. However, to put this data to use and improve educational environments is challenging. Previous studies attempting to gauge the adoption of learning analytics (LA) in South Africa, that is the application of data analytics methods to educational data, have shown that the adoption of such methods is still in its infancy (Lemmens & Henn 2016). Furthermore, it is shown that to utilise educational data effectively, solutions often need to be tailored to specific use cases and environments, even more so in the developing South African context (Molokeng & Van Belle 2021; Ngqulu 2018).

This study aims to evaluate the feasibility and usefulness of exploiting educational data, specifically in the context of academic literacy teaching and learning at a South African higher education institution. To achieve this aim, methods from the fields of LA and educational data mining (EDM) are applied to data collected from an academic literacy department in a South African university in 2022. These methods include descriptive analytics to gain a better understanding of the educational context, and predictive modelling to attempt to predict educational outcomes.

This article reviews the application of data analytics to higher education and the context of academic literacy in the first two chapters. Hereafter, the motivation for and goals of the study are given. The method used to analyse the data is described. A detailed analysis is conducted, and finally, the results are discussed.

Higher education and data analytics approaches

Information technology (IT) has become an integral part of education acting as a network connecting nodes of information between various key role players in HEIs (Singh 2021). The incorporation of IT and IT-related practices into the day-to-day teaching and learning practices has improved the overall accessibility of education and enhanced the transfer of knowledge between teachers and learners (Raja & Nagasubramani 2018). However, there are some costs associated with the invention of communication technologies, which need to be carefully analysed to optimise the embedded promises and possibilities (Nayar & Kumar 2018). To wit, IT educational systems generate a large amount of diverse data that often need analysis, data that subsequently can be exploited through EDM and LA (Romero & Ventura 2020).

Although both communities use data-intensive approaches to enhance educational research, they differ slightly in their methods and goals. Educational data mining is concerned with applying data mining (DM) techniques to educational datasets (Bakhshinategh et al. 2018). In contrast, LA collects, measures and analyses educational data based on data analytics principles using data analytics (DA) methods (Romero & Ventura 2020). Data mining is the process of discovering insightful, interesting and novel patterns, as well as descriptive, understandable and predictive models from large-scale data (Zaki, Meira & Meira 2014). Data analytics is defined as the application of computer systems to the analysis of large data sets and uses methods from statistics, machine learning, system theory and other fields to support decision-making (Runkler 2020). The distinction between DM and DA, and by extension EDM and LA, is that DM focuses on the discovery of patterns and knowledge and DA focussing on the support of decisions. The end goal of both EDM and LA is to generate statistics to support the educational process. Considering that EDM and LA share many of the same methods and techniques and combine computer science, statistics and education to form an interdisciplinary field, the application of both is often complementary and depends on the specific research problem (Baker & Inventado 2014; Romero & Ventura 2020).

Educational data mining/LA has been applied to several areas of the educational environment illustrating an ability to aid the task of teaching and learning. Learning analytics is especially efficient at describing students' behaviour and how they engage with module material. Zhang et al. (2018) conducted a study in which statistical analysis, visualisation and correlation analysis were used to analyse how students interact with module content. They discovered which time of day students prefer to access module content and which day of the week students access content frequently. Further, Zhang et al. (2018) measured which resources are most often accessed by the students and which content topics they preferred. This knowledge enables instructors to determine student engagement with module content and, combined with statistics on module grades, could provide more information on the effectiveness of study content. Providing decision-makers with accurate statistics regarding the modules they manage can give a better understanding of underlying student behavioural trends and enable them to improve decision-making capabilities.

Another popular research topic in EDM/LA is predictive modelling. The ability to predict students' performance proves to be a highly desirable measure for instructors (Hung et al. 2019). By predicting the marks achieved or outcomes of modules taken by students, potentially at-risk students may be identified early on, and timely intervention – based on the predictive modelling – may prevent such students from failing, thereby enhancing student success and the throughput rates for the module and HEI. To this end, EDM/LA has been implemented in previous studies to identify at-risk students (Akçapınar, Altun & Aşkar 2019; Choi et al. 2018).

Identifying relevant features in student enrolment data, acceptance test outcomes or demographic data may all serve as valuable parameters for prediction models. Furthermore, it has been demonstrated that using EDM and/or LA to identify at-risk students is feasible and that output from such research can support instructors better in assisting students (Foster & Siddle 2020). Besides supporting instructors or decision-makers in the academic environment, EDM/LA may also have a positive impact on the students. Providing high-quality feedback to students is fundamental to improving academic performance (Wisniewski, Zierer & Hattie 2020). Educational data mining has a student-oriented approach to providing feedback to students. Using knowledge discovery models, feedback may be customised to fit the student's needs and maximise the effectiveness of this feedback. Intelligent tutoring systems (ITSs), massive open online courses (MOOCs) and LMSs may all use EDM to customise the content provided to students based on their peers or academic performance. This may be in the form of providing additional or adapted content to struggling students or recommending new modules to study in the case of MOOCs (Romero & Ventura 2020). The feedback may also aid instructors, especially since LA approaches feedback from the perspective of instructors delivering insights on student performance. By applying analytics to student grades, instructors are presented with more descriptive information that they may use to better support students (Lang et al. 2017).

Educational data mining/LA as a field provides diverse options to analyse and determine how students interact with university resources, study content, as well as the teaching and learning (pedagogical) approaches followed in the specific module. However, most problems experienced in HEIs are unique to the module and therefore require tailored methods and research topics to address the specific problem (Ray & Saeed 2018). Apart from studies examining academic background like that of Fernandes et al. (2019) which created prediction models for the academic performance of public school learners in the capital of Brazil, or case studies presented in EDM/LA handbooks (Baker & Inventado 2014; Lang et al. 2017; Romero et al. 2010) with examples of the application of methods in other educational contexts, no studies could be found producing findings having direct relation to applying EDM/LA to explore academic literacy data. Previous studies apply methods to other fields such as measuring the originality of neuropsychology students in their assignments (Baker & Inventado 2014) or providing dashboards to students of their content interactions for a specific course (Lang et al. 2017).

Academic literacy

Academic literacy can be seen as having a transdisciplinary nature, drawing on anthropology, the New Literacy Studies movement, applied sociolinguistics, as well as systemic functional linguistics (SFL) (writing in particular), literary theory, rhetorical studies, critical discourse studies, communication studies, language and learning, sociology and socio-cultural theories of learning, psychology, and multimodality (Van Dyk & Van de Poel 2013). According to Nel and Janse van Rensburg (2022), academic literacy refers to a student's ability to 'transfer knowledge and to move between different discourse communities' (Nel & Janse van Rensburg 2022). Academic literacy can also be seen as 'the knowledge and skills required to communicate and function effectively and efficiently in different academic communities and achieve well-defined academic goals' (Van Dyk & Van De Poel 2013). This refers to a process of academic acculturation where students acculturate to the new academic (and linguistic) environment, and learn how to communicate within their respective fields of study. From these definitions, the relationship between academic literacy skills and a successful academic career should be evident, since academic acculturation will result in a higher throughput rate.

In its most basic form, academic literacy can be viewed as the ability of learners to read and write academic content, which includes comprehension, application and synthesis of theoretical knowledge (Defazio et al. 2010; Weideman 2013). From a pedagogical perspective, academic literacy abilities are influenced by a student's reading proficiencies and development, academic discourse abilities, and the various skills gained from secondary education (Andrianatos 2019; Walt & Mostert 2018). This refers to the notion that it is not a single language skill, but rather a wide range of diverse factors, influencing individual academic literacy ability.

With most HEI programme offerings being presented in English, the use of English as a second language (L2) has a definite impact on the academic literacy levels (and success rate) of students, and even more so in the diverse South African context, where students often have a background of speaking an African language as a first language (Weideman 2013). Studies have explored the factors influencing academic literacy, specifically secondary education language performance (Fleisch, Schoer & Cliff 2015).

From the discussion above, it should be evident why academic literacy development programmes are incorporated into the programme offerings of most HEIs in South Africa (SA). In most instances, the academic literacy offering refers to a mandatory first-year module(s) that aim to assist the student with their academic acculturation process. To support students with this 'acculturation process' a strong focus is placed on skills that will enable students to access, process, and produce information in their field of study (discourse community). In these modules, but particularly at the institution, we find ourselves, students' reading and writing abilities are assessed before the commencement of the module to measure their literacy ability's part and parcel of a placement protocol. Depending on the results of the placement protocol, students complete the required academic literacy modules.

Given the sheer number of students completing academic literacy modules (not only at the institution we find ourselves) and therefore completing tests and assessments completed by students, as well as the pedagogical resources created and published to LMSs, it is apparent that academic literacy should provide valuable data to analyse. However, the absence of machine learning and data mining, and the usage of extensive data with a variety of features more akin to Big Data, indicates that new knowledge may be gained from a study incorporating methods from EDM/LA.

By analysing academic and educational data, valuable insights can be gained, potentially aiding academic literacy departments in enhancing teaching and learning. Most data are obtained from students from diverse backgrounds representing different faculties. Although data mining has been applied to education in general, not many studies focus on academic literacy, especially where data is captured from highly diverse sources. This article argues that exploratory analysis is necessary to determine if statistically significant trends can be identified in the academic literacy offering.

Motivation

The previous section illustrates how the effective application of EDM/LA can improve the educational environment and support its goals. Furthermore, AL is shown to be of importance not only for the language ability of students but also for their ability to produce and interpret academic content. When considering these points, the following problem statement and research questions are formulated.

Problem statement

Academic literacy stands to benefit from the application of EDM/LA methods, but until now few studies have investigated the feasibility of applying these methods in the

South African context. This study aims to apply such methods and illustrate their feasibility.

Research question 1

Which methods from EDM and LA can prove valuable for academic literacy stakeholders, and why?

Research question 2

When applying such methods to a relevant data set, what insights could be gained?

Research question 3

Which improvements could be made to support the application of such methods to the educational environment, specifically regarding data collection and quality?

By investigating this research problem and addressing the research questions, it is hoped that this study could provide proof that EDM/LA could be effectively applied to the AL educational environment in the South African context and provide some recommendations towards future application.

Method

A systematic process, such as knowledge discovery and data (KDD) mining, can be followed to produce insights from educational data by applying either EDM or LA methods. Knowledge discovery and data is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro & Smyth 1996). Although the KDD process was defined and developed in 1996 by Fayyad et al. (1996), it remains foundational for knowledge discovery through data. The following sections briefly describe how the EDM/LA knowledge discovery cycle, based on the KDD process as proposed by Romero and Ventura (2020), enhances statistical analysis and understanding.

Defining the educational environment

The emergence of information and communications technology (ICT) and its impact quickly extended to the educational sector, with many applications gaining prevalence. Understanding the IT systems that generate data and how they contribute to the educational environment is essential to discovering knowledge in the academic context. Student administration has long been enhanced by leveraging IT and often contains student demographic data, enrolment information, and student grades which can all be used for analysis (Cunha & Miller 2014). Besides student administration systems, ICT has enabled instructors to change how module content is delivered. Learning management systems are digitising the entire classroom environment (Bradley 2021) and generate data that aid instructors in understanding their student's learning and learning patterns. To conduct this study, the systems that generate data relevant to academic literacy and to which the researchers may gain access were identified and analysed.

Gather educational data

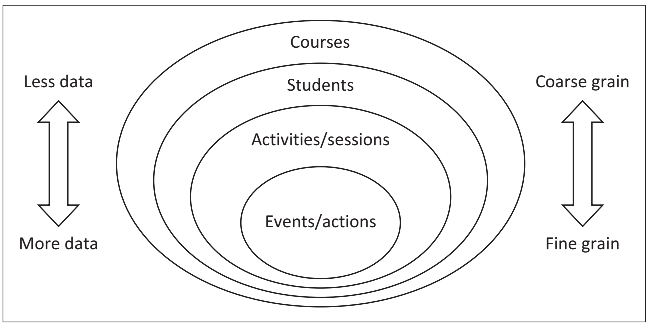
For EDM/LA methods to be used, data must be gathered from source systems. In education, the grain of data may vary with the granularity describing different aspects of the environment (Romero & Ventura 2020). Fine-grain educational data thus represent single events in the educational environment. An example of this would be students accessing a resource on an LMS or events like submitting assignments. In contrast, coarse grain data are student demographic or study module data. Generally, the quantity of data increases with a finer grain and decreases at a coarse level as shown in Figure 1.

The data captured for this study contains both fine-grain information about student events/actions, such as accessing study content on an LMS, and higher granularity data, such as student demographic and enrolment data, as well as data describing the structure of the academic literacy module.

Data pre-processing

In data mining, the data pre-processing phase is an essential but difficult and complicated task. Data pre-processing consists of all the tasks and actions before data analysis starts. It aims to produce a data set from an original data set that is more useful by removing problematic aspects of the data while preserving valuable information (Famili et al. 1997). Data pre-processing may therefore be divided into two distinct categories (data preparation and data reduction), each with distinct methods (García, Luengo & Herrera 2015), and addressing specific challenges to improve the usability of data for analysis.

Implementing techniques such as data cleaning, transforming, integration, normalisation, missing data imputation and noise identification, data preparation converts prior useless data into data that can be used by data mining techniques to produce accurate results. Data reduction obtains a reduced representation of the original data, which may increase the accuracy of certain DM techniques or focus on specific aspects of a more extensive data set. Typically, data reduction is made by implementing techniques such as feature selection, instance selection and discretisation (García et al. 2015).



Source: Romero, C. & Ventura, S., 2020, 'Educational data mining and learning analytics: An updated survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(3), e1355. https://doi.org/10.1002/widm.1355

FIGURE 1: Various levels of granularity and their relationship to the amount of data.

Apply methods and techniques

Once data is pre-processed, researchers apply DM methods that best fit their needs. The DM method is usually based on application and the problems they address as well as the nature of the available data (Bakhshinategh et al. 2018).

Interpretation and application of new knowledge

Educational data mining/LA methods in education can provide stakeholders with insights regarding assessment, pedagogy and epistemology (Lang et al. 2017). However, for studies to yield positive results, they must be presented to stakeholders so they can act on the findings (Baker & Inventado 2014). As EDM/LA research output comes in different forms, stakeholders ultimately determine the decision-making value and the application of knowledge gained based on the output produced. As such, studies can focus on complex artefacts that send real-time alerts to educators, show live data visualisations (Bakhshinategh et al. 2018), or present reports and visualisations (Siemens 2013). The data gathered for this study is historic and will be presented using reports and visualisations rather than live feedback.

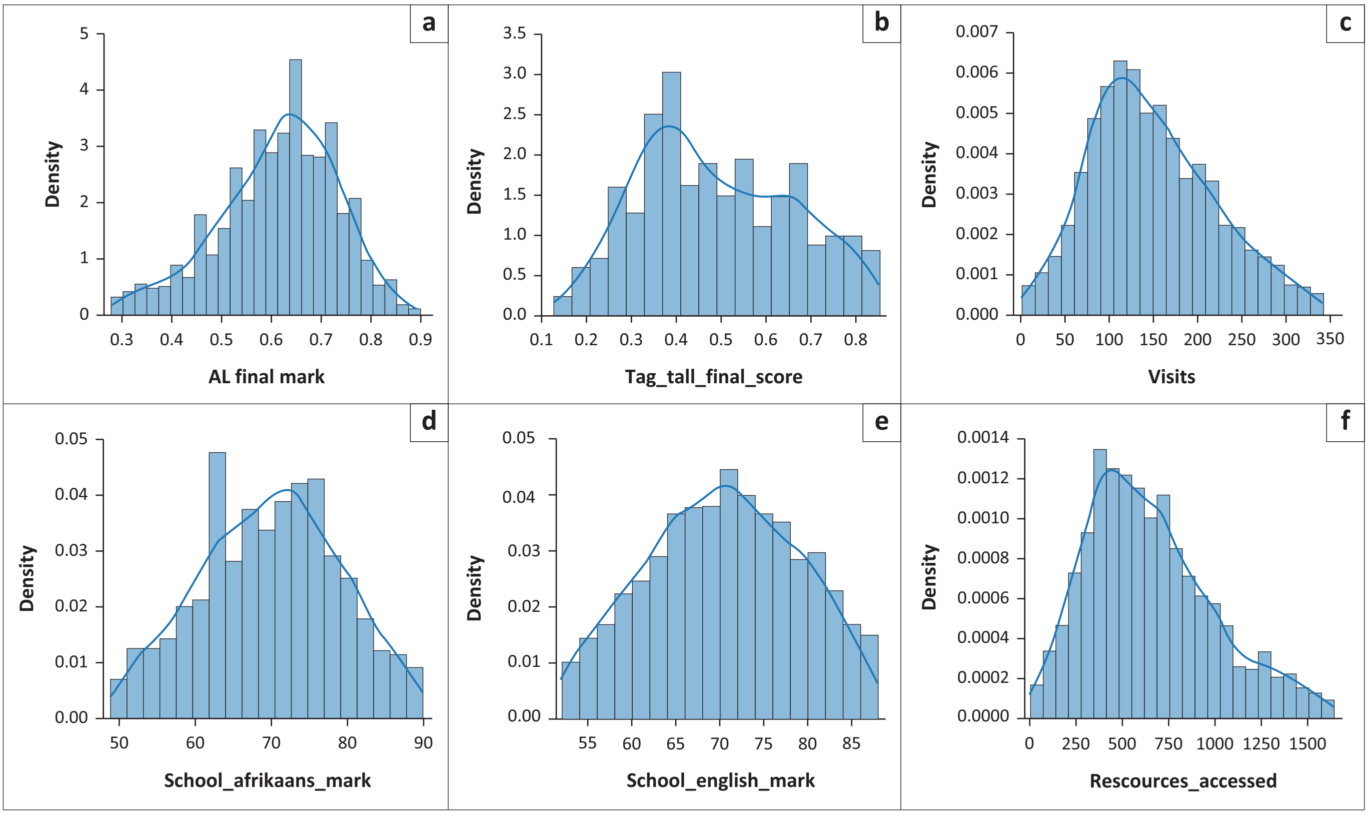
Analysis

After gathering data from various source systems, two datasets were compiled to be used in the analysis. The first data set represents students with features for student demographics, enrolment data, secondary education Afrikaans and English marks, preliminary language test marks, LMS usage statistics, and the final mark achieved for the academic literacy module. The second data set, being of a finer grain (as depicted in Figure 1), comprises individual visit statistics to the LMS with features for the date accessed and the number of visits. The first dataset, representing individual students, had a population size of 2288 students. The 'site visits' data set of a finer grain had 145122 records representing LMS visits per date.

Descriptive analysis

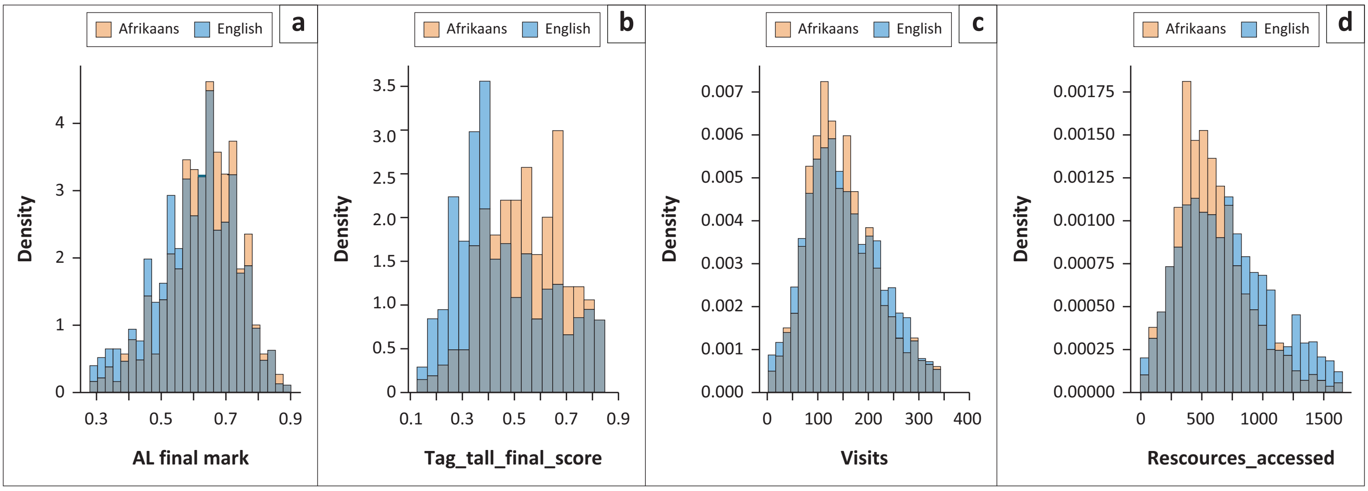
Using the student profile data set, feature selection was implemented by performing samples of individual variables in the data set. The continuous features selected were the final mark achieved, preliminary language test outcome, visits to LMS, resources accessed on LMS, Afrikaans secondary education mark, and English secondary education mark. Some samples contained outliers, which were removed by calculating the z-scores of records and removing those within two standard deviations of the mean.

Using the cleaned samples, probability distributions were plotted to represent the distributions of the variables. The plots in Figure 2 show that the final marks achieved for the module and secondary education language marks were



AL, academic literacy.

FIGURE 2: Probability distributions of continuous variables. (a) Probability distribution of final marks achieved for academic literacy; (b) Probability distribution of preliminarily language test outcomes; (c) Probability distribution of visits to a learning management system; (d) Probability distributor of Afrikaans marks achieved during secondary education; (e) Probability distribution of English marks achieved during secondary education; (f) Probability distribution of learning recourses accessed on a learning management system.



AL, academic literacy.

FIGURE 3: Probability distributions by language. (a) Language comparative probability distributions of academic literacy final marks achieved; (b) Language comparative probability distributions of preliminary language test outcomes; (c) Language comparative probability distributions of visits to a learning management system; (d) Language comparative probability distributions of learning recourses accessed on a learning management system.

normally distributed, and LMS usage statistics were slightly skewed to the right. The mean of preliminary language test outcomes (top middle graph) is below the passing mark of 50%, indicating that many students failed the initial language test and thus had poor language proficiency at the start of their first academic year.

Plotting the distributions by the language in which the academic literacy module was taken enables the reader to distinguish between students' behaviour according to their pedagogical language as seen in Figure 3. These distributions show that Afrikaans students performed better than English students for the module final marks and preliminary language test scores.

However, further inspecting this trend by calculating the proportion of students who received academic instruction in English but studied a different first language in secondary education, 41.62% of students in the population studied a different first language from English in secondary education, compared to 0.87% for Afrikaans students. This may explain why Afrikaans students outperformed English students and correlate with the observation of Fleisch et al. (2015) that the performance of L1 English students differs from L2 speakers. This finding can enable stakeholders to develop additional resources and interventions to further support L2 speakers in their academic literacy development process, which in turn could yield better student success ratios, while simultaneously improving multilingualism.

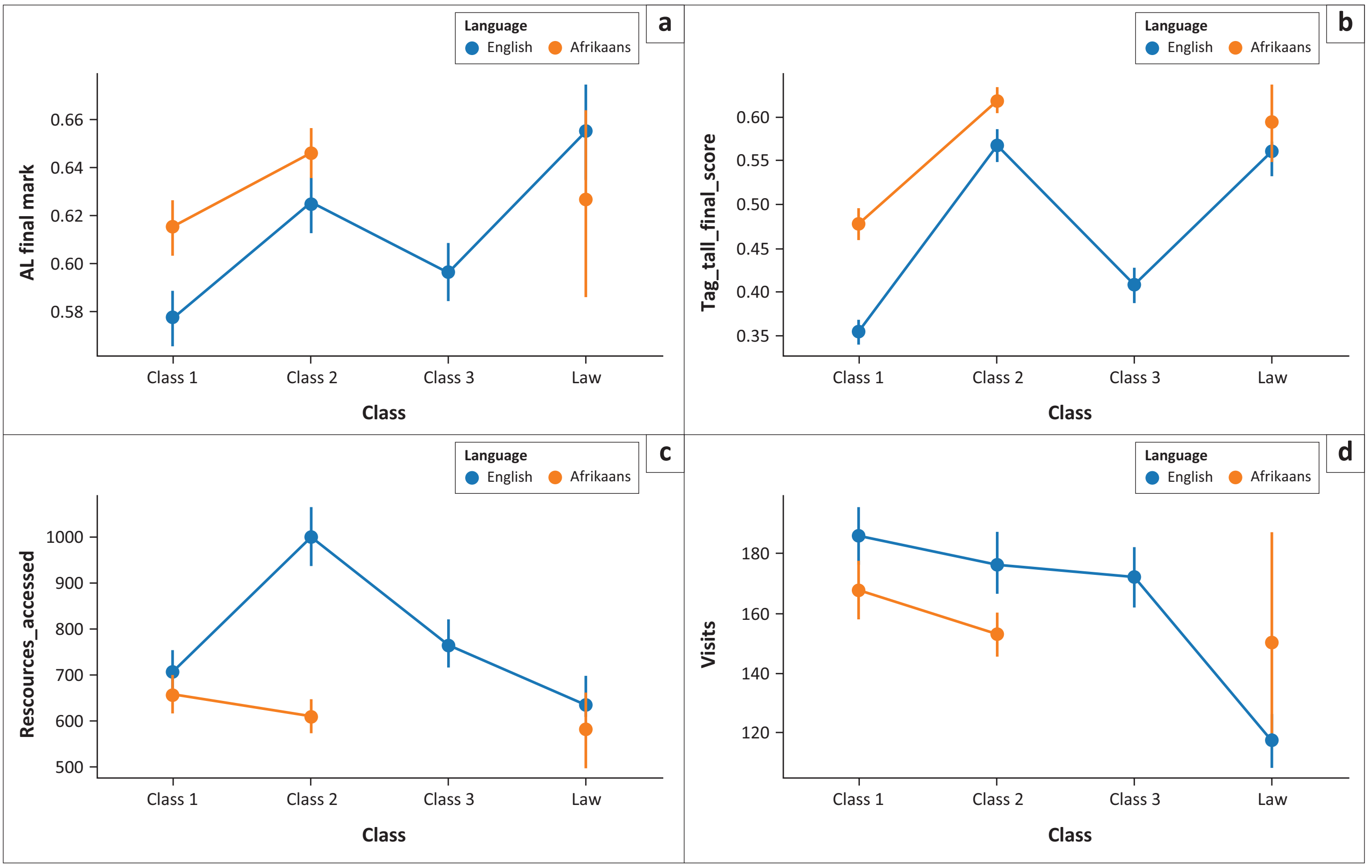
Plotting features on point plots and distinguishing by the different physical classes into which students were grouped at the commencement of the module allows stakeholders to differentiate between the performance of classes (Figure 4).

From the point plots, the same trend in better Afrikaans performance could be seen. Furthermore, classes one and three of the English classes had lower scores for both module outcomes and language test marks. In this case, point plots proved to be very well suited not only to distinguish between the performance of classes but this finding may aid decision-makers in providing better support for students on a class-by-class basis.

Using the visits to the LMS data set, time series plots were created for both English and Afrikaans students' visits to the LMS as shown in Figure 5. This indicated a higher variance between the English classes in the number of visits. For both the Afrikaans and English classes, spikes in visits were identified throughout the first half of the semester, followed by a flattening in the middle of the semester, and finally, a spike before the commencement of the examination period. The result from this analysis confirms the assumptions that students tend to: (1) 'lose interest' in the academic literacy module, (2) not consider the module as important or relevant, and (3) do not pay continuous attention to the module and tend to care only when final assessments need to be submitted. This type of plot may aid decision-makers in identifying classes which lack proper LMS usage and identify periods in the semester when more emphasis should be placed on LMS usage. This will enable the stakeholders to strategise and reconsider the module content and the planning of the delivery of content to ensure continuous participation by most students registered for the course and thereby counter the flattening seen in LMS visits in the middle of the semester.

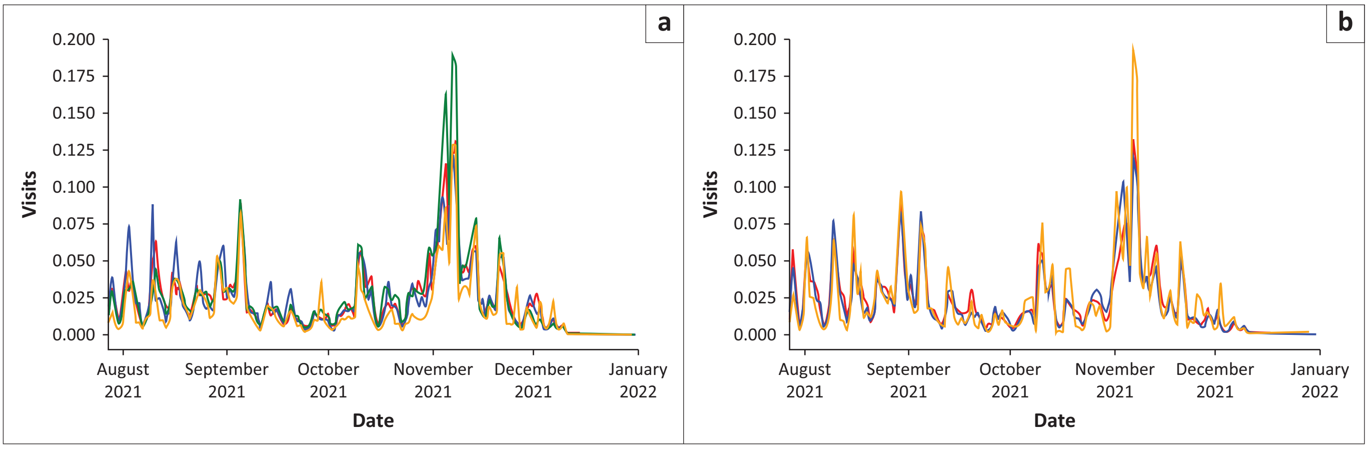
Using scatterplots combined with linear regression, the relationship between visits to LMS and resources accessed on LMS was compared to the module mark achieved (Figure 6).

Two correlation coefficient methods were used to determine the strength of the relationship between the variables namely Pearson's correlation coefficient with



AL, academic literacy.

FIGURE 4: Point plots distinguishing between class and language. (a) Point plot illustrating academic literacy final mark achieved per pedagogical class; (b) Point plot illustrating preliminary language test outcomes per pedagogical class; (c) Point plot illustrating learning recourses accessed using a learning management system per pedagogical class; (d) Point plot illustrating visits to a learning management system per pedagogical class.



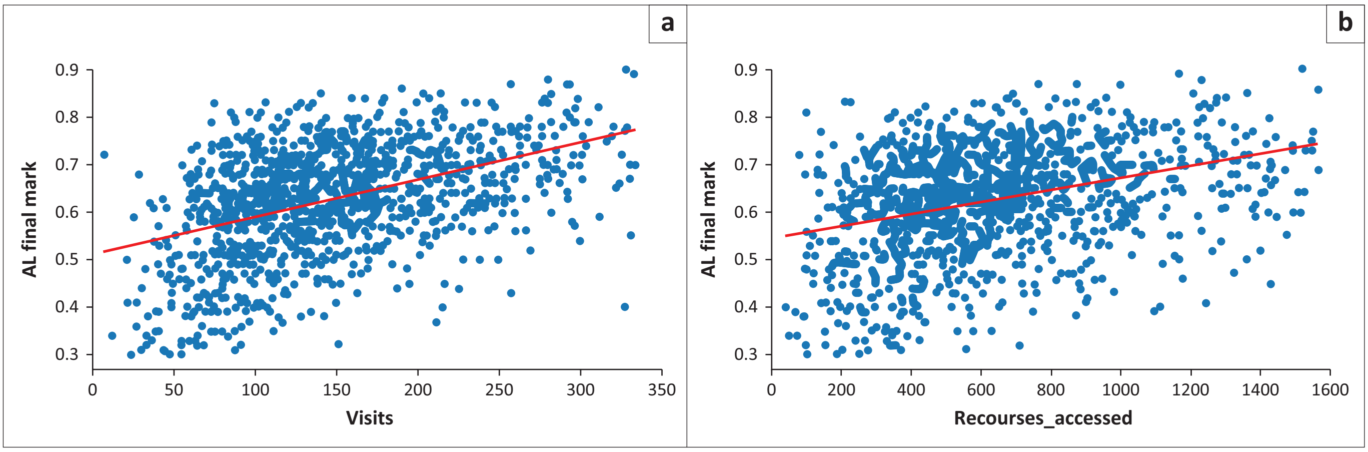
AL, academic literacy.

FIGURE 5: Time plots showing visits to learning management systems. (a) Time series plot of visits to a learning management system over the course of one semester for English academic literacy students; (b) Time series plot of visits to a learning management system over the course of one semester for Afrikaans academic literacy students.

correlation values of 0.44 for Visits and 0.34 for Resources Accessed, and Spearman's rank correlation coefficient with values of 0.42 and 0.34, respectively. For both visits and resources accessed, a weak positive correlation existed with module outcome, meaning that encouraging LMS usage could benefit students. This correlates with the point plots where class two for English students had higher academic performance and resource usage.

Predictive analysis

A predictive analysis was conducted to prove the feasibility of using models to categorically predict whether a student will pass and to predict a continuous variable representing the final mark achieved using features available to stakeholders at the commencement of the academic literacy module. Two predictive methods were used in the predictive analysis, the first being logistic regression used to classify students according to module outcome, and the



AL, academic literacy.

FIGURE 6: Scatterplots of learning management systems usage and resources accessed compared to academic literacy mark. (a) Scatter plot of visits to a learning management system and academic literacy final mark achieved; (b) Scatter plot of academic resources accessed on a learning management system and academic literacy final mark achieved.

TABLE 1: Performance measures of the logistic regression model.

Predicted class	Precision	Recall	F-measure
Pass	0.85	0.60	0.71
Fail	0.16	0.43	0.24

second k-nearest neighbours to predict the final mark achieved.

For the logistic regression, a dataset was created after applying feature selection on variables that would be available to instructors at the commencement of the module. Features that had slight variance between passing and failing students and variables with a *p*-value of less than 0.05 were removed from the model to increase its accuracy. The model was able to obtain an f1 score of 0.71 when predicting passing students and 0.24 when predicting failing students. Table 1 gives a more detailed summary of the performance measures used to evaluate the logistic regression model.

The f1 score, being the average of the prediction precision and recall value, serves as a good indicator of the effectiveness of the model in predicting module outcome. The respective f1 scores, however, were meagre and show that using only study programmes, school language marks, academic literacy language and class is inadequate to build an accurate logistic regression model.

K-nearest neighbours, being an unsupervised learning algorithm, are not dependent on the precision or recall values. Using the same features for k-nearest neighbours, a model was built to predict the final mark achieved for the module. Splitting training and test sets by 70/30, providing the lowest root mean squared error (RMSE), and using six for the value of k, the model could predict module marks with an RMSE of 9.2% for training data and 12.21% for test data. This means that the model was able to predict academic literacy marks achieved with an average error of 12.21%. Although the RMSE is high, k-nearest neighbours may still be used by stakeholders to some extent to predict expected academic literacy performance for students.

The predictive analysis performed in this study illustrates the possibility of using machine learning models in the context of academic literacy. Although the logistic regression did not provide enough accuracy to be used operationally, it may become relevant if more diverse features were to be introduced to the model. K-nearest neighbours can be implemented operationally to predict module outcomes and support decision-making, while also possibly benefiting from more added features.

Discussion

The methods applied in this article show that educational data can be leveraged to support teaching and learning academic literacy. Using descriptive statistics, many meaningful insights were gained into the nature of the data. The descriptive analysis provided insights into student behaviour by plotting probability distributions and may enable decision-makers to utilise the point plots regarding class performance factors better. Time series plots proved to be very efficient in determining LMS usage trends and distinguishing between classes. When exploring the effects of LMS usage and resources accessed, it was determined that both these variables influenced the module's outcome.

The predictive analysis proved the feasibility of applying predictive models to academic data in the context of the specific HEI. It showed that using features from data sets available to decision-makers at the commencement of the academic literacy module, predictions may be made with a certain level of confidence about student performance, a strong indication that these features influence the outcome of the module and by extension, academic literacy.

The data that was obtained for this study represents only a fraction of the data generated by HEIs in South Africa. Demographic data such as gender, ethnicity, secondary education language institution, parental income, and many more variables could be introduced in similar future research. The contribution of additional data will be unknown until the data has been used and analysed in similar studies.

Therefore, to increase the potential of future research, the democratisation of educational data would need to be increased allowing for more diverse data and more extensive research.

When considering the validity and reliability of this study, it is important to note that the diversity of educational environments could produce different outcomes when applying the same methods. However, the core concept that this article aims to prove, that methods from EDM and LA could be effectively applied in the relevant context remains valid. This is supported by the variety of generalisable methods both these approaches offer, and the variety of studies in which such methods are effectively applied. This sentiment further underlines the need to tailor similar solutions to the environment at hand.

Conclusion

With a proven track record of supporting pedagogy, EDM and LA methods can provide effective solutions to educational problems. This study serves as proof that such methods could be of significant use to the academic literacy offering of HEIs. This study made use of a fraction of the possible data that may be obtained in the educational environment, and being able to obtain more data could further increase the benefits of applying EDM/LA. In addition to more variety in data, there exists more methods and implementations for EDM/LA research that are not used in this study that can increase the benefits of such research. To exploit the full potential of Big Data in education in the South African context, further studies should be done to measure its feasibility and illustrate its usefulness.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

Authors' contributions

J.R. was at the time a honours student at the Computer Science department conducting this study under supervision of C.J.K. Z.J.v.R. and M.N. were approached by J.R. to perform this research on data collected at the NWU Subject-Group: Academic Literacy. M.N. provided the data and assisted J.R. in utilising the data within this research project. Z.J.v.R. and M.N. helped finalise the article.

Ethical considerations

This article followed all ethical standards for research without direct contact with human or animal subjects.

Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

Due to the nature of the research, and/or because of ethical, legal, or commercial reasons, supporting data is not available. The data used in this study is not public data and is owned by the educational institution where the study was conducted.

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

References

Akçapınar, G., Altun, A. & Aşkar, P., 2019, 'Using learning analytics to develop an early-warning system for at-risk students', *International Journal of Educational Technology in Higher Education* 16(1), 1–20. https://doi.org/10.1186/s41239-019-0172-z

Andrianatos, K., 2019, 'Barriers to reading in higher education: Rethinking reading support', *Reading & Writing – Journal of the Reading Association of South Africa* 10(1), 1–9. https://doi.org/10.4102/rw.v10i1.241

Baker, R.S. & Inventado, P.S., 2014, 'Educational data mining and learning analytics', in J.A. Larusson & B. White (eds.), *Learning analytics: From research to practice*, pp. 61–75, Springer, New York, NY.

Bakhshinategh, B., Zaiane, O.R., ElAtia, S. & Ipperciel, D., 2018, 'Educational data mining applications and tasks: A survey of the last 10 years', *Education and Information Technologies* 23(1), 537–553. https://doi.org/10.1007/s10639-017-9616-z

Bradley, V.M., 2021, 'Learning management systems (lms) use with online instruction', *International Journal of Technology in Education (IJTE)* 4(1), 68–92. https://doi.org/10.46328/ijte.36

Choi, S.P.M., Lam, S.S., Li, K.C. & Wong, B.T.M., 2018, 'Learning analytics at low-cost at-risk student prediction with clicker data and systematic proactive interventions', *Journal of Educational Technology & Society* 21(2), 273–290, viewed 03 July 2022, from http://www.jstor.org.nwulib.nwu.ac.za/stable/26388407.

Cunha, J.M. & Miller, T., 2014, 'Measuring value-added in higher education: Possibilities and limitations in the use of administrative data', *Economics of Education Review* 42, 64–77. https://doi.org/10.1016/j.econedurev.2014.06.001

Defazio, J., Jones, J., Tennant, F. & Hook, S.A., 2010, 'Academic literacy: The importance and impact of writing across the curriculum – A case study', *Journal of the Scholarship of Teaching and Learning* 10(2), 34–47.

Famili, A., Shen, W.-M., Weber, R. & Simoudis, E., 1997, 'Data preprocessing and intelligent data analysis', *Intelligent Data Analysis* 1(1), 3–23. https://doi.org/10.3233/IDA-1997-1102

Fayyad, U.M., Piatetsky-Shapiro, G. & Smyth, P., 1996, 'Knowledge discovery and data mining: Towards a unifying framework', in E. Simoudis, J. Han & U.M. Fayyad (eds.), *KDD'96 conference proceedings,* pp. 82–88, AAAI Press, Portland, OR.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R. & Van Erven, G., 2019, 'Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil', *Journal of Business Research* 94, 335–343.

Fleisch, B., Schoer, V. & Cliff, A., 2015, 'When signals are lost in aggregation: A comparison of language marks and competencies of first-year university students', *South African Journal of Higher Education* 29(5), 156–178. https://doi.org/10.10520/EJC182512

Foster, E. & Siddle, R., 2020, 'The effectiveness of learning analytics for identifying at-risk students in higher education', *Assessment & Evaluation in Higher Education* 45(6), 842–854. https://doi.org/10.1080/02602938.2019.1682118

García, S., Luengo, J. & Herrera, F., 2015, *Data preprocessing in data mining*, in J. Kacprzyk & J.C. Jain (eds.), vol. 72, pp. 59–139, Springer International Publishing, Cham.

Hung, J.-L., Shelton, B.E., Yang, J. & Du, X., 2019, 'Improving predictive modeling for at-risk student identification: A multistage approach', *IEEE Transactions on Learning Technologies* 12(2), 148–157. https://doi.org/10.1109/TLT.2019.2911072

Lang, C., Siemens, G., Wise, A. & Gasevic, D., 2017, *Handbook of learning analytics*, SOLAR, Society for Learning Analytics and Research, New York, NY.

Lemmens, J.-C. & Henn, M., 2016, *Learning analytics: A south african higher education perspective*, pp. 231–253, Institutional Research in South African Higher Education, Stellenbosch.

Molokeng, P.M. & Van Belle, J.P., 2021, 'Investigating the use of learning analytics at South Africa's higher education institutions', in E. Toledo Gómez (ed.), *The 2018 International Conference on Digital Science,* pp. 59–70, Springer International Publishing, Cham.

Nayar, K.B. & Kumar, V., June 2018, 'Cost benefit analysis of cloud computing in education', *International Journal of Business Information Systems* 27(2), 205–221. https://doi.org/10.1504/IJBIS.2018.089112

Nel, M. & Janse van Rensburg, Z., 2022, 'A holistic, continuous approach to nwu students' academic acculturation: The role of academic literacy and the writing centre', in M.M. Fernandes-Martins, M. Fourie, (ed.), *Transformative pedagogies*, pp. 299–335, Axiom Academic Publishers, Potchefstroom.

Ngqulu, N., 2018, 'Investigating the adoption and the application of learning analytics in south african higher education institutions (heis)', in E. Ivala (ed.), *International Conference on e-Learning*, pp. 545–XVI, Academic Conferences International Limited, June 2018, Cape Town.

Raja, R. & Nagasubramani, P., 2018, 'Impact of modern technology in education', *Journal of Applied and Advanced Research* 3(1), 33–35. https://doi.org/10.21839/jaar.2018.v3iS1.165

Ray, S. & Saeed, M., 2018, 'Applications of educational data mining and learning analytics tools in handling big data in higher education', in M.M. Alani, H. Tawfik, M. Saeed & O. Anya (eds.), *Applications of big data analytics*, pp. 135–160, Springer, Cham.

Romero, C. & Ventura, S., 2020, 'Educational data mining and learning analytics: An updated survey', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10(3), e1355. https://doi.org/10.1002/widm.1355

Romero, C., Ventura, S., Pechenizkiy, M. & Baker, R.S., 2010, *Handbook of educational data mining*, CRC press, Boca Raton, FL.

Runkler, T.A., 2020, *Data analytics*, Springer, Wiesbaden.

Siemens, G., 2013, 'Learning analytics: The emergence of a discipline', *American Behavioral Scientist* 57(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Singh, M.N., 2021, 'Inroad of digital technology in education: Age of digital classroom', *Higher Education for the Future* 8(1), 20–30. https://doi.org/10.1177/2347631120980272

Van Dyk, T. & Van de Poel, K., 2013, 'Towards a responsible agenda for academic literacy development: Considerations that will benefit students and society', *Journal for Language Teaching= Ijenali Yekufundzisa Lulwimi= Tydskrif vir Taalonderrig* 47(2), 43–69. https://doi.org/10.4314/jlt.v47i2.3

Walt, J.L.V.d. & Mostert, A., 2018, 'Academic literacy and the development of inference skills at secondary school level', *Journal for Language Teaching = Ijenali Yekufundzisa Lulwimi = Tydskrif vir Taalonderrig* 52(1), 62–80. https://doi.org/10.4314/jlt.v52i1.4

Weideman, A., 2013, 'Academic literacy interventions: What are we not yet doing, or not yet doing right?', *Journal for Language Teaching= Ijenali Yekufundzisa Lulwimi= Tydskrif vir Taalonderrig* 47(2), 11–23. https://doi.org/10.4314/jlt.v47i2.1

Wisniewski, B., Zierer, K. & Hattie, J., 2020, 'The power of feedback revisited: A meta-analysis of educational feedback research', *Frontiers in Psychology* 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Zaki, M.J., Meira Jr, W. & Meira, W., 2014, *Data mining and analysis: Fundamental concepts and algorithms*, Cambridge University Press, New York.

Zhang, J.H., Zhang, Y.X., Zou, Q. & Huang, S., 2018, 'What learning analytics tells us: Group behavior analysis and individual learning diagnosis based on long-term and large-scale data', *Journal of Educational Technology & Society* 21(2), 245–258.