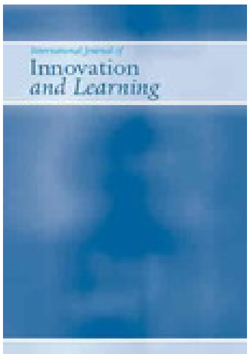


International Journal of Innovation and Learning

ISSN online: 1741-8089 - ISSN print: 1471-8197
https://www.inderscience.com/ijil

A survey on predicting at-risk students through learning analytics

Kam Cheong Li, Billy Tak-Ming Wong, Maggie Liu

DOI: 10.1504/IJIL.2024.10065503

Article History:

Received:	31 March 2024
Last revised:	03 June 2024
Accepted:	04 June 2024
Published online:	26 July 2024

A survey on predicting at-risk students through learning analytics

Kam Cheong Li, Billy Tak-Ming Wong* and Maggie Liu

Hong Kong Metropolitan University,
Homantin, Hong Kong, China
Email: kcli@hkmu.edu.hk
Email: tamiwong@hkmu.edu.hk
Email: mjliu@hkmu.edu.hk
*Corresponding author

Abstract: This paper analyses the adoption of learning analytics to predict at-risk students. A total of 233 research articles between 2004 and 2023 were collected from Scopus for this study. They were analysed in terms of the relevant types and sources of data, targets of prediction, learning analytics methods, and performance metrics. The results show that data related to students' academic performance, socio-demographics, and learning behaviours have been commonly collected. Most studies have addressed the identification of students who have a higher chance of poor academic performance or dropping out of their courses. Decision trees, random forests, and artificial neural networks are the most frequently used techniques for prediction, with ensemble methods gaining popularity in recent years. Classification accuracy, recall, sensitivity, and true positive rate are commonly used as performance metrics for evaluation. The results reveal the potential of learning analytics for informing timely and evidence-based support for at-risk students.

Keywords: students at-risk; prediction; learning analytics; educational data mining; student support.

Reference to this paper should be made as follows: Li, K.C., Wong, B.T-M. and Liu, M. (2024) 'A survey on predicting at-risk students through learning analytics', *Int. J. Innovation and Learning*, Vol. 36, No. 5, pp.1–15.

Biographical notes: Kam Cheong Li is the Dean of the School of Open Learning and Director of the Institute for Research in Open and Innovative Education at Hong Kong Metropolitan University. He has devoted to higher education and teaching development for 30 years. He has also been appointed Visiting Professor of Middlesex University in the UK, and a number of institutions in China, such as Hebei Academy of Social Sciences and Hebei University of Economics and Business. His research interests lie in technology in education, pedagogical approaches, and educational innovation.

Billy Tak-Ming Wong is the Deputy Director of the Institute for Research in Open and Innovative Education and a senior research coordinator at Hong Kong Metropolitan University. He has been involved in various research projects related to technology-enhanced education. His research areas include mobile learning, computer-enhanced learning and learning analytics.

Maggie Liu is a Research Associate at Hong Kong Metropolitan University. Her research interests lie in the areas of internationalisation of higher education and ICT-supported teaching and learning.

This paper was originally accepted for a special issue on 'Technology-Mediated Learning: Innovative Practices for the New Normal' guest edited by Dr. Naraphorn Paoprasert, Dr. Simon K.S. Cheung, Dr. Kam Cheong Li and Dr. Peerayuth Charnsethikul.

This paper is a revised and expanded version of a paper entitled 'Prediction of at-risk students using learning analytics: a literature review' presented at the 6th International Conference on Technology in Education at Bangkok, Thailand, 19–21 December 2023.

1 Introduction

Identifying at-risk students is a crucial responsibility for educators and educational institutions. At-risk students include those who are likely to drop out or fail to meet academic standards due to various factors such as socio-economic characteristics, learning disabilities, or lack of engagement (Laskey and Hetzel, 2011). Identifying these at-risk students in order to provide timely support will significantly and positively impact their academic and personal development. To address this issue, learning analytics has emerged as an effective tool for predicting at-risk students in the past decade. Learning analytics refers to the process of collecting, analysing, and interpreting data from various sources such as student performance, behaviours, and engagement to understand and optimise learning (Long and Siemens, 2011). With systematic management and data analysis related to student learning, learning analytics facilitates the identification of patterns and trends that reveal the students who are at risk of falling behind. This proactive approach plays an important role in helping instructors to offer personalised interventions for students who may require additional learning support.

There has been a broad range of scholarly interest in utilising learning analytics to predict at-risk students. For example, Akçapınar et al. (2019) developed a model based on the k-nearest neighbours algorithm to forecast students' academic performance at the end of a term, using interaction data in an online learning environment. Choi et al. (2018) applied logistic regression and linear regression to identify students at risk of failing a course by processing in-class clicker data. Bayazit et al. (2022) presented a predictive model for identifying at-risk students in a blended learning setting, and compared the performance of different learning analytics techniques for such purposes. Russell et al. (2020) investigated students' use of a learning analytics platform that provides weekly performance feedback in a chemistry course. Their findings indicate that the use of the learning analytics platform by at-risk students is associated with achieving a passing grade at the end of the course.

Despite the abundance of research on learning analytics for predicting at-risk students, relevant review studies on this topic have limitations, such as narrow scopes and limited data sources (Na and Tasir, 2017; Tamada et al., 2019; Li et al., 2022; Shafiq et al., 2022; Nurmalitasari et al., 2023). To address these limitations, this paper aims to provide a comprehensive survey on the adoption of learning analytics for predicting at-risk students. It seeks to answer the following research questions:

1 What are the types and sources of data commonly collected to predict at-risk students using learning analytics?

2 What are the targets of prediction to identify at-risk students through learning analytics?

3 What are the learning analytics methods frequently employed for prediction?

4 What are the performance metrics commonly used for evaluation of the prediction?

2 Related studies

The adoption of learning analytics has been increasingly popular in the educational context, and a wide range of review studies have been conducted. Some of these studies have provided a holistic overview of the learning analytics development in the education sector, covering the features and trends, research approaches, analytical techniques, benefits, and challenges (Avella et al., 2016; Wong, 2017; Wong et al., 2018; Viberg et al., 2018). There have also been reviews on particular areas. For example, Li and Wong (2020) analysed the features and patterns of learning analytics in STEM education. Omer et al. (2023) summarised the steps and benefits of integrating learning analytics in programming courses. Wong et al. (2023a) examined the application of learning analytics for personalised learning.

Reviews on learning analytics have also addressed the specific purpose of making predictions. Sghir et al. (2023) identified five main areas of prediction, namely enrolment, performance, at-risk students, engagement, and satisfaction. Based on a summary of case studies on learning analytics dashboards which incorporate predictive analytics, Ramaswami et al. (2023) investigated how learning analytics have been used to meet the diverse needs and interests of different stakeholders. Li et al. (2024) reported an analysis on predictive analytics with respect to university student admission.

There have been reviews related to learning analytics techniques for supporting at-risk students. For example, Li et al. (2022) conducted a review of empirical studies that adopt learning analytics to predict the likelihood of students persisting in STEM education delivered through massive open online courses (MOOCs) and online learning. They identified a range of student-related factors that contribute to retention, including individual features, enrolment properties, academic performance, and learning engagement.

Tamada et al. (2019) surveyed the solutions used to predict student dropout in virtual learning environments. Their study revealed that supervised machine learning techniques, such as logistic regression and support vector machine, have been more commonly utilised for dropout prediction compared to unsupervised machine learning techniques. The data used in related studies covers diverse aspects, including clickstream data, forum participation, and event logs.

In their study, Shafiq et al. (2022) investigated the factors that can aid in identifying high-risk students and those that fail to recognise such students in three learning environments, namely traditional learning, blended learning, and online learning. Their findings indicate that academic results are suitable data for predicting dropout, while time-related factors such as time spent in learning activities, and socio-demographic factors such as age and gender are less significant for predicting at-risk students. They also discovered that supervised machine learning and deep learning techniques are widely used in student retention.

The study conducted by Nurmalitasari et al. (2023) presented a systematic review of studies that employed predictive learning analytics to predict student dropout rates. The review summarised the variables used to predict student dropouts and identified the characteristics of these variables, such as low frequency, nominal, and ordinal measurement levels. They also illustrated the data processing tasks such as data cleaning, data integration, data reduction, and data transformation, as well as learning techniques such as logistic regression, decision trees, support vector machine, and Naive Bayes.

Na and Tasir (2017) investigated the identification of at-risk students by analysing their online learning behaviours. They discovered that most of the related work focused on the attributes or indicators for identifying at-risk students. The common ones include learning level, network data, and learning emotion. Logistic regression and decision trees were the most frequently used analytic techniques. Approaches to assist at-risk students mostly involve interventions in aspects such as course designs, teaching methods, pedagogical recommendations, and instructional materials.

However, the existing reviews on the use of learning analytics for predicting at-risk students have limitations in terms of scope, such as focusing specifically on online learning in Li et al. (2022) and Tamada et al. (2019), and sources of data, such as covering only 39 studies as in Nurmalitasari et al. (2023), or only including publications up to 2017 as in Na and Tasir (2017). The present study aims to address these limitations by surveying the most recent related studies and analysing how learning analytics methods have been utilised to identify at-risk students.

3 Research method

3.1 Search strategies and selection procedures

For this study, research articles addressing the use of learning analytics to predict at-risk students were searched in Scopus, which is a widely accepted database for literature reviews (Li and Wong, 2021; Wong et al., 2023a, 2023b). The keywords [('learning analytics' OR 'educational data mining') AND ('attrition' OR 'dropout' OR 'academic performance' OR 'final grade' OR 'final score') AND 'predict*'] were used to search relevant articles. The time range was limited to the past two decades, from 2004 to 2023, and the document type was set as 'Article'. An initial search returned 412 articles, which were then screened according to the following inclusion criteria:

1 The article presents the design and implementation of a learning analytics-based approach for predicting at-risk students.

2 The article is written in English.

3 The article is available in full text.

Based on the above criteria, a total of 233 articles were finally included for further analysis. Figure 1 illustrates the flow chart for article search strategies and selection procedures.

Figure 2 presents the annual count of research publications on the adoption of learning analytics to predict at-risk students. There was an absence of relevant publications on this topic from 2004 to 2008, and no more than three publications in each year between 2009 and 2014. However, there has been a noticeable increase in the number of publications in the recent decade, particularly since 2018, and the trend peaked in 2022 and 2023 with 48 publications for each year. These findings imply a mounting interest in leveraging learning analytics to predict students who are at risk of academic failure.

Figure 1 Procedures for search and selection of articles

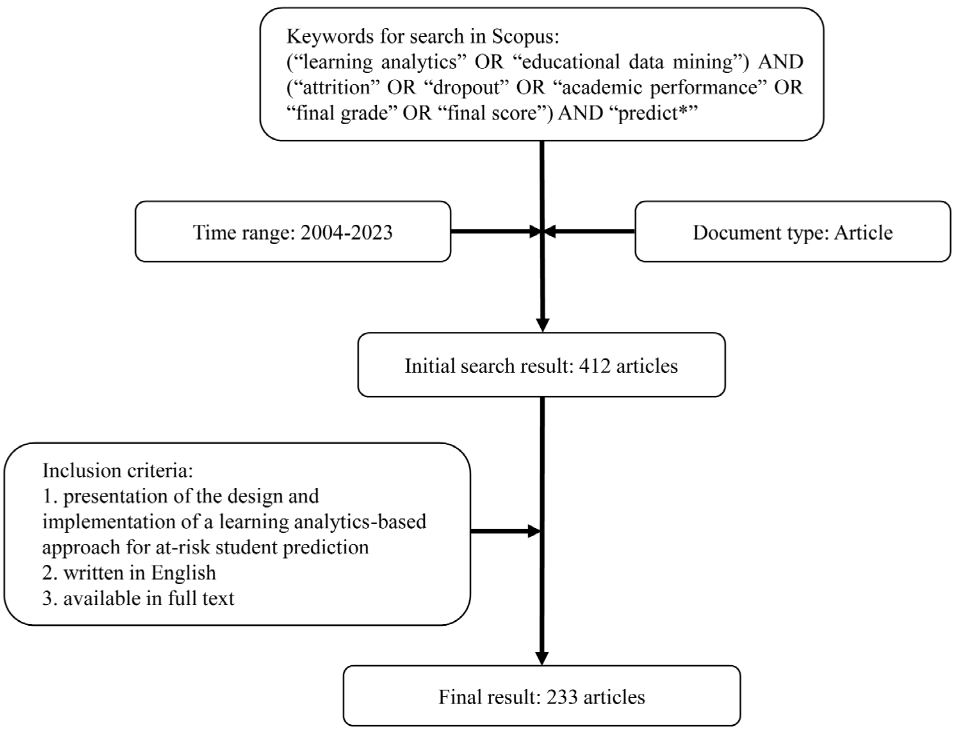
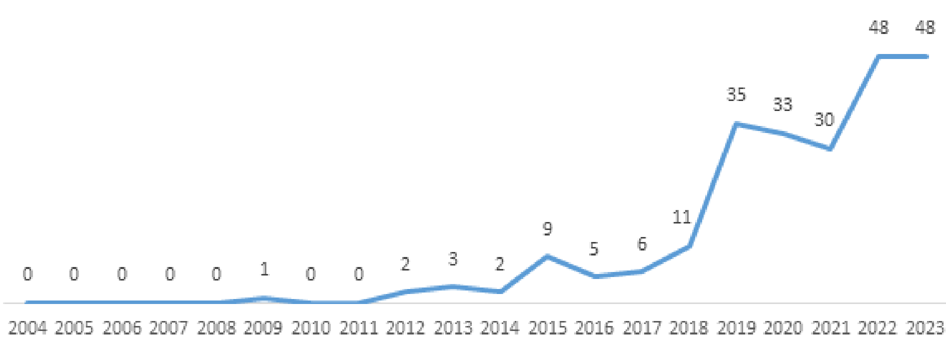


Figure 2 Publication years of the articles (see online version for colours)



3.2 Data extraction and coding

Each selected article was thoroughly examined, and relevant information was extracted and categorised to address the research questions. The information covers the following aspects:

1 types and sources of data used in prediction

2 targets of prediction

3 learning analytics techniques adopted for the prediction

4 evaluation measures of the prediction.

The data extraction and coding were conducted by two researchers. Any disagreements during the data examination process were discussed until a consensus was reached.

4 Results

4.1 Types and sources of data

Table 1 reports the types and sources of data used to support the predictions in the studies. Students' academic performance data was the most frequently utilised, followed by data on their socio-demographic information, online learning behaviour, and academic history. Offline learning behaviour data was least used for predicting at-risk students. These data were acquired through various sources.

Academic performance data was collected from assessments, including scores of assignments, quizzes, exams, lab work, and projects, as well as course grades, and GPA.

Socio-demographic information was acquired mainly through student surveys. This type of data includes

1 demographic information such as gender, age, ethnicity, marital status, and health status

2 socio-economic information such as employment status, financial support, residence type, family income, number of siblings, as well as parents' educational level and occupations

3 personal and social lifestyle such as alcohol consumption habits, time spent with friends, and romantic relationships

4 psychological attributes such as learning motivation and anxiety.

Online learning behaviour data was captured through learning management systems (LMS) and learning platforms. It involves clickstream data of student interactions with the LMS and learning platforms, including

1 participation in events such as assignments, exercises, quizzes and forum discussion

2 access to course materials such as course pages, e-books, lecture notes and videos

3 frequency and time spent on involvement in such learning activities.

Data on academic history was gathered from administrative records. Examples include

1 pre-enrolment background such as high school information, high school grades, and scores on standardised tests
2 enrolment information such as admission test scores, reasons for selecting the school/programme, and mode of study
3 data related to number of course failures and attempts, as well as attendance and absence records.

Offline learning behaviour data involved learning behaviours inside and outside of the classroom, which could be obtained through student surveys and classroom observations. Instances include participation in after-school tutoring courses, time spent on studying, frequency of library visits, and data observed in face-to-face classes such as raising hands and taking notes.

Table 1 Types and sources of data involved in the studies

Type of data	*Source*	*Frequency*
Academic performance	Assessment	159
Socio-demographic information	Student surveys	135
Online learning behaviour	Learning management systems and learning platforms	111
Academic history	Administrative records	95
Offline learning behaviour	Student surveys and classroom observations	25

4.2 Targets of prediction

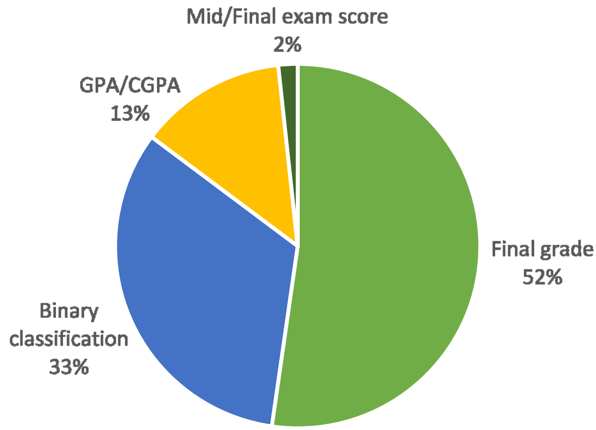
Table 2 presents the targets of prediction addressed in related studies to identify at-risk students. Four types of targets were identified, with academic performance being the most frequently predicted, followed by students who may drop out from their courses or programmes. Additionally, one study focused on predicting student assignment submission, and another predicting student engagement.

Table 2 Targets of prediction made in the studies

Target of prediction	*Frequency*
Academic performance	176
Course/programme dropout	62
Assignment submission	1
Engagement	1

As shown in Figure 3, the majority of academic performance predictions were based on final grades (52%). Some studies used binary classifications such as successful/unsuccessful, at risk/not at risk, and pass/fail to denote academic performance (33%). GPA/CGPA were predicted relatively infrequently (13%), while mid/final exam scores were only predicted in 2% of the related studies.

Figure 3 Types of prediction under the academic performance category (see online version for colours)



4.3 Learning analytics methods

In the reviewed studies, three types of prediction tasks were identified, namely classification, regression, and clustering. Table 3 shows the top 10 techniques used to predict at-risk students based on their frequencies. Eight of these techniques were utilised for the classification task, with decision tree, random forest, and artificial neural network being the most popular. Multiple linear regression and linear regression, the 9th and 10th ranked techniques, were used for the regression task.

Table 3 Top 10 most frequently used techniques to predict at-risk students

Task	*Technique*	*Frequency*
Classification	Decision tree	129
	Random forest	102
	Artificial neural network	100
	Support vector machine	94
	Naïve Bayes	84
	Logistic regression	75
	k-Nearest neighbours	58
	Rule Induction/rule-based classification	22
Regression	Multiple linear regression	19
	Linear regression	15

Table 4 Types of ensemble learning methods in the studies

Ensemble approach	*Frequency*
Boosting	20
Bagging	16
Stacking	12
Voting	5

About 9% of the studies reviewed employed ensemble learning to improve the accuracy of predictions by merging predictions from multiple models. Bagging, boosting, stacking, and voting are the four major types of ensemble learning methods. Some of the studies used at least two of these methods. Table 4 presents the distribution of the ensemble learning methods identified in the studies in terms of frequency. Boosting is the most commonly adopted method, followed by bagging and stacking, while voting is the least used method.

4.4 Performance metrics

A variety of performance metrics are available for measuring the prediction performance of machine learning models. Table 5 depicts the top 10 performance metrics that were most frequently adopted in the reviewed studies. The most frequently used performance metrics are accuracy/classification accuracy, which were identified in nearly two-thirds of the studies. Recall/sensitivity/true positive rate, precision, and F-measure/F-score/ F1-measure /F1-score were also commonly used, with a usage rate of about half of the total, respectively. These metrics, together with Area under the receiver operating characteristic curve, specificity/true negative rate, and receiver operating characteristic curve, were all used for the classification tasks. The remaining relatively less used metrics, such as root mean square error, mean absolute error, and coefficient of determination/R squared and Adjusted R squared, were used for the regression tasks.

Table 5 Top 10 most frequently used performance metrics

Task	*Performance metric*	*Frequency*
Classification	Accuracy/classification accuracy	155
	Recall/sensitivity/true positive rate	123
	Precision	107
	F-measure/F-score/F1-measure /F1-score	101
	Area under the receiver operating characteristic curve	47
	Specificity/true negative rate	20
	Receiver operating characteristic curve	16
Regression	Root mean square error	28
	Mean absolute error	22
	Coefficient of determination/R squared and Adjusted R squared	21

5 Discussion

The findings address the four research questions in this study. The first research question concerns the types and sources of data used to predict at-risk students. The findings show that academic performance data collected from assessments is the most commonly used. Conventionally, students' learning progress has been monitored based on academic results. Academic success has been regarded as a significant factor influencing student dropout (Sultana, et al., 2017; Nuanmeesri et al., 2022). As noted by Buenaño-Fernández et al. (2019), students' previous academic performance has become a valuable source of information for identifying students who may be at risk, and it is one of the most widely used applications of learning analytics.

Socio-demographic information acquired through student surveys is another type of data commonly used in predictions of at-risk students. Previous studies have demonstrated that socio-demographic characteristics of students are key factors in predicting their academic performance and probability of dropout. Gender and age are among the most common demographic variables to predict academic success and student dropout. Notably, however, while some studies have found gender (e.g., Krishnan et al., 2019) and age (e.g., Yasmin, 2013) to be major features affecting students' academic performance and dropout decision, other studies have found these variables to not have a significant impact on the prediction (e.g., Alturki et al., 2022; Sithole et al., 2023; Tamada et al., 2022). This indicates that while socio-demographic information can be useful in predicting at-risk students, its predictive power may vary depending on the specific contexts and other variables included in the analysis.

Online learning behaviour data, which can be recorded automatically in the databases of LMS and learning platforms, is also a frequently used source for identifying at-risk students. This would be relevant to the increasing popularity of online learning and the development of personalised learning (Wong et al., 2023a) and smart education (Li and Wong, 2022). Analysis of LMS activity data keeps teachers informed about the dynamics of students' learning patterns and progress, helps discover potential at-risk students in a timely manner, and enables effective intervention to improve learning outcomes (Herodotou et al., 2019). For example, dropout prediction can be performed weekly by comparing the historical patterns of interaction with the data of current students, and alert messages can be sent to teachers so that they may take proactive and personalised actions to mitigate students' learning problems (Cambruzzi et al., 2015).

The second research question relates to the targets of prediction addressed in the previous studies. Most of the studies aimed to identify at-risk students by forecasting their academic performance. Poor academic performance, such as low final grades, course failure, and low scores in mid/final exams were commonly selected as target variables for the prediction. These variables have shown a significant correlation with student dropout, and their prediction is crucial in identifying the learning challenges that students encounter and informing teachers to provide timely support to them (Bedregal-Alpaca et al., 2020). For example, Cogliano et al. (2022) built a prediction model to identify students who were likely to perform poorly in a course, and provided a digital self-regulated learning skill training programme to some of these students as the treatment group. Their results showed that the treatment group students performed better than those struggling students who had not received the training in examinations. Moreover their performance did not even differ significantly from those who were predicted to perform well.

The findings demonstrate that predicting course/programme dropout is another common target, which directly addresses the risk of dropout. Analysis of factors that contribute to academic failure and dropout helps teachers improve their teaching methods and provide a better learning experience for students. This can also aid educational institutions in developing effective strategies to reduce dropout rates. For example, Shiao et al. (2023) developed a learning platform that regularly updates data on students' learning progress and dropout predictions. The platform alerts teachers about students who are struggling with their studies and provides students with personalised learning recommendations on learning materials and methods. After three years of implementation, the dropout risk of students was found to have decreased.

The third research question pertains to learning analytics methods used to identify at-risk students. The findings reveal that classification and regression are the two primary tasks that learning algorithms aim to accomplish. Consistent with previous studies, classification was found to be the most common task for identifying at-risk students (Huang et al., 2020; Yang et al., 2020). Among the machine learning algorithms, decision tree and random forest were the most commonly applied, and they have been evidenced to be effective in predicting academic performance (Huynh-Cam et al., 2021). The third frequently used machine learning method is Artificial Neural Network, which was utilised in 100 studies with 94 of them published between 2019 and 2023, indicating the growing popularity of this technique in recent years. This suggests the superior performance of this method and the useful insights it provides into the factors that impact the educational process (Sandoval-Palis et al., 2020).

As claimed by Sghir et al. (2023), the selection of appropriate algorithms for predictive analysis depends on various factors, including the purpose and settings, dataset size, data characteristics, and prediction targets. Instead of relying on a single method for prediction, about 73% of the reviewed studies were found to involve more than one machine learning technique or algorithm. Some studies conducted experiments to compare the performance of multiple techniques and selected the best one (e.g., Orrego Granados et al., 2022; Ramaswami et al., 2019; Queiroga at el., 2022). Several studies used popular algorithms such as decision tree, linear regression, and support vector machine as baseline models to evaluate the prediction performance of the proposed model (e.g., Alhassan et al., 2020; Nayak et al., 2023).

It has been observed that the use of ensemble methods has rapidly increased in recent years. They feature a combination of several algorithms or baseline models into one optimal model to complement the limitations of different models so as to achieve a higher prediction accuracy (Siddique et al., 2021). By leveraging the power of various machine learning techniques, ensemble models have often attained solid accuracy, which is indeed superior to those single-based models (e.g., Nahar et al., 2021; Predić et al., 2018; Verma et al., 2022). Major types of ensemble methods employed in the reviewed studies include boosting, bagging, stacking, and voting. Boosting generates and combines several weak models into a strong one, which can help reduce bias and prediction errors. Bagging is mostly used to reduce variance in noisy data. Stacking combines the predictions of different based models with a single meta model to minimise error and enhance prediction accuracy, which is 'useful when different techniques are all good for tackling the same problem, but in different ways' (Talamás-Carvajal and Ceballos, 2023, p. 12171). Voting is an algorithm that is used to aggregate multiple decisions either by majority vote or plurality to improve the overall prediction performance. In the reviewed studies, these ensemble methods were used not only individually, but also in combination (e.g., Balcioğlu and Artar, 2023; Memon et al., 2022).

The fourth research question concerns the performance metrics used for evaluating the predictions. The findings reveal that the performance metrics are determined by the learning analytics tasks. For classification tasks, accuracy, recall, precision, and F-measure have been frequently used to assess performance in related studies. For regression tasks, common metrics include root mean square error, mean absolute error, and coefficient of determination. Notably, although accuracy is the most widely used performance metric, this metric may generate misleading evaluation results about the prediction performance of models when the data is not distributed in a balanced way (e.g., the number of students who drop out is significantly smaller than the number of those who persist) (Barros et al., 2019).

6 Conclusions

This study analysed the literature on the prediction of at-risk students through learning analytics. It provides a comprehensive overview of how the research and practice is developing in this domain. The findings revealed the main types of data with various sources used for predictions, including academic performance data, socio-demographic information, online learning behaviour data, academic history, and offline learning behaviour data. Using multiple sources and types of data were recommended for learning analytics to improve precision and accuracy (Carter et al., 2017), highlighting the need to combine various data types (Omer et al., 2023). Most studies addressed the identification of students who have a higher chance of poor academic performance or dropping out of their courses, indicating that the primary goal of learning analytics in predicting at-risk students is to provide early interventions to prevent academic failure and dropout. The findings also showed that a broad range of learning analytics methods and performance metrics were used in the studies to serve specific tasks, and most studies adopted multiple techniques and metrics. These results suggest that the use of multiple techniques and metrics is necessary to improve the overall effectiveness of learning analytics in predicting at-risk students.

These findings provide insights for future research in the field. With the rapid advancement of artificial intelligence, it would be worthwhile exploring the potential of using relevant techniques for predicting at-risk students. Furthermore, the findings reveal the need to investigate the differences in at-risk student patterns across multiple subject disciplines. Furthermore, future studies need to examine the relationships between the targets of prediction, the data used for prediction, the analytics techniques, and the performance metrics for evaluation. This would help to identify the most effective combination of these factors in different contexts for predicting at-risk students.

Acknowledgements

The work described in this paper was partially supported by grants from Hong Kong Metropolitan University (2021/011 and CP/2022/04).

References

Akçapınar, G., Altun, A. and Aşkar, P. (2019) 'Using learning analytics to develop early-warning system for at-risk students', *International Journal of Educational Technology in Higher Education*, Vol. 16, No. 1, pp.1–20.

Alhassan, A., Zafar, B. and Mueen, A. (2020) 'Predict students' academic performance based on their assessment grades and online activity data', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 4, pp.185–194.

Alturki, S., Cohausz, L. and Stuckenschmidt, H. (2022) 'Predicting Master's students' academic performance: an empirical study in Germany', *Smart Learning Environments*, Vol. 9, No. 1, p.38.

Avella, J.T., Kebritchi, M., Nunn, S.G. and Kanai, T. (2016) 'Learning analytics methods, benefits, and challenges in higher education: a systematic literature review', *Online Learning*, Vol. 20, No. 2, pp.13–29.

Balcioğlu, Y.S. and Artar, M. (2023) 'Predicting academic performance of students with machine learning', *Information Development*, DOI: 10.1177/02666669231213023.

Barros, T.M., Neto, P.A.S., Silva, I. and Guedes, L.A. (2019) 'Predictive models for imbalanced data: a school dropout perspective', *Education Sciences*, Vol. 9, No. 4, p.275.

Bayazit, A., Apaydin, N. and Gonullu, I. (2022) 'Predicting at-risk students in an online flipped anatomy course using learning analytics', *Education Sciences*, Vol. 12, No. 9, p.581.

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zarate-Valderrama, J. and Yanque-Churo, P. (2020) 'Classification models for determining types of academic risk and predicting dropout in university students', *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 1, pp.266–272.

Buenaño-Fernández, D., Gil, D. and Luján-Mora, S. (2019) 'Application of machine learning in predicting performance for computer engineering students: a case study', *Sustainability*, Vol. 11, No. 10, p.2833.

Cambruzzi, W., Rigo, S.J. and Barbosa, J.L.V. (2015) 'Dropout prediction and reduction in distance education courses with the learning analytics multitrail approach', *Journal of Universal Computer Science*, Vol. 21, No. 1, pp.23–47.

Carter, A.S., Hundhausen, C.D. and Adesope, O. (2017) 'Blending measures of programming and social behavior into predictive models of student achievement in early computing courses', *ACM Transactions on Computing Education (TOCE)*, Vol. 17, No. 3, p.12.

Choi, S.P., Lam, S.S., Li, K.C. and Wong, B.T. (2018) 'Learning analytics at low cost: at-risk student prediction with clicker data and systematic proactive interventions', *Journal of Educational Technology and Society*, Vol. 21, No. 2, pp.273–290.

Cogliano, M.C., Bernacki, M.L., Hilpert, J.C. and Strong, C.L. (2022) 'A self-regulated learning analytics prediction-and-intervention design: detecting and supporting struggling biology students', *Journal of Educational Psychology*, Vol. 114, No. 8, pp.1801–1816.

Herodotou, C., Rienties, B., Boroowa, A., Zdrahal, Z. and Hlosta, M. (2019) 'A large-scale implementation of predictive learning analytics in higher education: the teachers' role and perspective', *Educational Technology Research and Development*, Vol. 67, pp.1273–1306.

Huang, A.Y.Q., Lu, O.H.T., Huang, J.C.H., Yin, C.J. and Yang, S.J.H. (2020) 'Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs', *Interactive Learning Environments*, Vol. 28, No. 2, pp.206–230.

Huynh-Cam, T.T., Chen, L.S. and Le, H. (2021) 'Using decision trees and random forest algorithms to predict and determine factors contributing to first-year university students' learning performance', *Algorithms*, Vol. 14, No. 11, p.318.

Krishnan, R., Beegum, H. and Sherimon, P.C. (2019) 'Academic performance based on gender using filter ranker algorithms - an experimental analysis in Sultanate of Oman', *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8, No. 11, pp.3502–3506.

Laskey, M.L. and Hetzel, C.J. (2011) 'Investigating factors related to retention of at-risk college students', *Learning Assistance Review*, Vol. 16, No. 1, pp.31–43.

Li, C., Herbert, N., Yeom, S. and Montgomery, J. (2022) 'Retention factors in STEM education identified using learning analytics: a systematic review', *Education Sciences*, Vol. 12, No. 11, p.781.

Li, K.C. and Wong, B.T.M. (2020) 'Trends of learning analytics in STE(A)M education: a review of case studies', *Interactive Technology and Smart Education*, Vol. 17, No. 3, pp.323–335.

Li, K.C. and Wong, B.T.M. (2021) 'A review of the use of open educational resources: the benefits, challenges and good practices in higher education', International Journal of Innovation and Learning, Vol. 30, No. 3, pp.279–298.

Li, K.C. and Wong, B.T.M. (2022) 'How smart learning has been achieved: a review of the literature (2011–2020)', *International Journal of Mobile Learning and Organisation*, Vol. 16, No. 3, pp.310–322.

Li, K.C., Wong, B.T.M. and Chan, H.T. (2024) 'Prediction of at-risk students using learning analytics: a literature review', in Cheung et al. (Eds.): *Technology in Education. Innovative Practices for the New Normal. Communications in Computer and Information Science*, Vol. 1974, pp.119–128, Springer, Singapore.

Long, P. and Siemens, G. (2011) 'Penetrating the fog: analytics in learning and education', *Educause Review*, Vol. 46, No. 5, pp.30–40.

Memon, M.Q., Lu, Y., Yu, S., Memon, A. and Memon, A.R. (2022) 'The critical feature selection approach using ensemble meta based models to predict academic performances', *International Arab Journal of Information Technology*, Vol. 19, No. Special Issue 3A, pp.523–529.

Na, K.S. and Tasir, Z. (2017) 'Identifying at-risk students in online learning by analysing learning behaviour: a systematic review', in *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, November, pp.118–123, IEEE.

Nahar, K., Shova, B.I., Ria, T., Rashid, H.B. and Islam, A.H.M.S. (2021) 'Mining educational data to predict students performance: a comparative study of data mining techniques', Education and Information Technologies, Vol. 26, No. 5, pp.6051–6067.

Nayak, P., Vaheed, S., Gupta, S. and Mohan, N. (2023) 'Predicting students' academic performance by mining the educational data through machine learning-based classification model', *Education and Information Technologies*, Vol. 28, No. 11, pp.14611–14637.

Nuanmeesri, S., Poomhiran, L., Chopvitayakun, S. and Kadmateekarun, P. (2022) 'Improving dropout forecasting during the COVID-19 pandemic through feature selection and multilayer perceptron neural network', *International Journal of Information and Education Technology*, Vol. 12, No. 9, pp.851–857.

Nurmalitasari, A., Long, Z. and Mohd Noor, M.F. (2023) 'The predictive learning analytics for student dropout using data mining technique: a systematic literature review', in Ismail et al. (Eds.): *Advances in Technology Transfer Through IoT and IT Solutions*, pp.9–17, Springer, Cham.

Omer, U., Tehseen, R., Farooq, M.S. and Abid, A. (2023) 'Learning analytics in programming courses: Review and implications', *Education and Information Technologies*, Vol. 28, No. 9, pp.11221–11268.

Orrego Granados, D., Ugalde, J., Salas, R., Torres, R. and López-Gonzales, J.L. (2022) 'Visual-predictive data analysis approach for the academic performance of students from a Peruvian University', *Applied Sciences*, Vol. 12, No. 21, p.11251.

Predić, B., Dimić, G., Rančić, D., Štrbac, P., Maček, N. and Spalević, P. (2018) 'Improving final grade prediction accuracy in blended learning environment using voting ensembles', *Computer Applications in Engineering Education*, Vol. 26, No. 6, pp.2294–2306.

Queiroga, E.M., Batista Machado, M.F., Paragarino, V.R., Primo, T.T. and Cechinel, C. (2022) 'Early prediction of at-risk students in secondary education: a countrywide K-12 learning analytics initiative in Uruguay', *Information*, Vol. 13, No. 9, p.401.

Ramaswami, G., Susnjak, T., Mathrani, A. and Umer, R. (2023) 'Use of predictive analytics within learning analytics dashboards: a review of case studies', *Technology, Knowledge and Learning*, Vol. 28, No. 3, pp.959–980.

Russell, J.E., Smith, A. and Larsen, R. (2020) 'Elements of success: supporting at-risk student resilience through learning analytics', *Computers and Education*, Vol. 152, p.103890.

Sandoval-Palis, I., Naranjo, D., Gilar-Corbi, R. and Pozo-Rico, T. (2020) 'Neural network model for predicting student failure in the academic leveling course of Escuela Politécnica Nacional', *Frontiers in Psychology*, Vol. 11, p.515531.

Sghir, N., Adadi, A. and Lahmer, M. (2023) 'Recent advances in predictive learning analytics: a decade systematic review (2012–2022)', *Education and Information Technologies*, Vol. 28, No. 7, pp.8299–8333.

Shafiq, D.A., Marjani, M., Habeeb, R.A.A. and Asirvatham, D. (2022) 'Student retention using educational data mining and predictive analytics: a systematic literature review', *IEEE Access*, Vol. 10, pp.72480–72503.

Shiao, Y.T., Chen, C.H., Wu, K.F., Chen, B.L., Chou, Y.H. and Wu, T.N. (2023) 'Reducing dropout rate through a deep learning model for sustainable education: long-term tracking of learning outcomes of an undergraduate cohort from 2018 to 2021', *Smart Learning Environments*, Vol. 10, No. 1, p.55.

Siddique, A., Jan, A., Majeed, F., Qahmash, A.I., Quadri, N.N. and Wahab, M.O.A. (2021) 'Predicting academic performance using an efficient model based on fusion of classifiers', *Applied Sciences*, Vol. 11, No. 24, p.11845.

Sithole, S.T.M., Ran, G., de Lange, P., Tharapos, M., O'Connell, B. and Beatson, N. (2023) 'Data mining: will first-year results predict the likelihood of completing subsequent units in accounting programs?', *Accounting Education*, Vol. 32, No. 4, pp.409–444.

Sultana, S., Khan, S. and Abbas, M.A. (2017) 'Predicting performance of electrical engineering students using cognitive and non-cognitive features for identification of potential dropouts', *International Journal of Electrical Engineering and Education*, Vol. 54, No. 2, pp.105–118.

Talamás-Carvajal, J.A. and Ceballos, H.G. (2023) 'A stacking ensemble machine learning method for early identification of students at risk of dropout', *Education and Information Technologies*, Vol. 28, No. 9, pp.12169–12189.

Tamada, M.M., de Magalhães Netto, J.F. and de Lima, D.P.R. (2019) 'Predicting and reducing dropout in virtual learning using machine learning techniques: a systematic review', in *2019 IEEE Frontiers in Education Conference (FIE)*, October, pp.1–9, IEEE.

Tamada, M.M., Giusti, R. and de Magalhães Netto, J.F. (2022) 'Predicting students at risk of dropout in technical course', *Electronics*, Vol. 11, No. 3, p.468.

Verma, S., Yadav, R.K. and Kholiya, K. (2022) 'A scalable machine learning-based ensemble approach to enhance the prediction accuracy for identifying students at-risk', *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 8, pp.185–192.

Viberg, O., Hatakka, M., Bälter, O. and Mavroudi, A. (2018) 'The current landscape of learning analytics in higher education', *Computers in Human Behavior*, Vol. 89, pp.98–110.

Wong, B.T.M. (2017) 'Learning analytics in higher education: an analysis of case studies', *Asian Association of Open Universities Journal*, Vol. 12, No. 1, pp.21–40.

Wong, B.T.M., Li, K.C. and Cheung, S.K. (2023a) 'An analysis of learning analytics in personalised learning', *Journal of Computing in Higher Education*, Vol. 35, No. 3, pp.371–390.

Wong, B.T., Li, K.C., Chan, H.T. and Cheung, S.K. (2023b) 'HyFlex learning research and practice: a longitudinal analysis', *Sustainability*, Vol. 15, No. 12, p.9699.

Wong, B.T.M., Li, K.C. and Choi, S.P.M. (2018) 'Trends in learning analytics practices: a review of higher education institutions', *Interactive Technology and Smart Education*, Vol. 15, No. 2, pp.132–154.

Yang, J., DeVore, S., Hewagallage, D., Miller, P., Ryan, Q.X. and Stewart, J. (2020) 'Using machine learning to identify the most at-risk students in physics classes', *Physical Review Physics Education Research*, Vol. 16, No. 2, p.20130.

Yasmin (2013) 'Application of the classification tree model in predicting learner dropout behaviour in open and distance learning', *Distance Education*, Vol. 34, No. 2, pp.218–231.