A survey on educational data mining methods used for predicting students’ performance

Wen Xiao¹² | Ping Ji³ | Juan Hu⁴

¹School of Educational Science, Anhui Normal University, Wuhu, China
²Institute of Intelligent Education, Anhui Normal University, Wuhu, China
³Key Laboratory of Unmanned Aerial Vehicle Development & Data Application of Anhui Higher Education Institutes, Wanjiang University of Technology, Ma’anshan, China
⁴School of Computer and Information, Anhui Normal University, Wuhu, China

Correspondence
Wen Xiao, School of Educational Science, Anhui Normal University, Wuhu 241002, China.
Email: cyees@163.com

Funding information
Top discipline talents of the Anhui Provincial Department of Education, Grant/Award Numbers: gxbjzd2020101

Abstract
Predicting students’ performance is one of the most important issues in educational data mining (EDM), which has received more and more attention. By predicting students’ performance, we can identify students’ risk of academic failure and help instructors to take some actions such as guidance or interventions to help learners as early as possible, or carry out continual evaluation of learners as to optimize learning path or personalized learning resources recommendation. In this survey, we reviewed the 80 important studies on predicting students’ performance using EDM methods in 2016–2021, synthesized the procedure of building prediction model of students’ performance which contains four phases and 10 key steps, and compared and discussed the latest EDM methods used in all steps. We analyzed the challenges faced by previous studies in three aspects and put forward future suggestions on data collection, EDM methods used, and interpretation of prediction model. This survey provides a comprehensive understanding and practical guide for researchers in this field, and also provides direction for further research.

KEYWORDS
educational data mining, machine learning, predicting students’ performance, survey

JEL CLASSIFICATION
Computer and software engineering

1 | INTRODUCTION

With the rapid development of computer, communication, sensors, and other information technologies, people or devices in various fields are producing massive data every day, and human beings have entered the era of big data. Big data can be defined and described by the “5V” of volume, variety, value, velocity, and veracity.¹ In the field of education, the digital learning environments such as student information system (SIS), learning management system (LMS), massive open online course (MOOC), and virtual learning system are constantly producing educational big data. It is a great challenge for researchers and educational institutions to analyze the educational big data thoroughly and discover the hidden information so that learners, instructors, or managers can benefit from it.²
At present, educational data mining (EDM) is one of the most important ways to analyze educational big data effectively. As an interdisciplinary research field, EDM applies machine learning (ML), statistics, data mining (DM), educational psychology, cognitive psychology, and other theories and methods to analyze educational data, helping people to solve various problems in education effectively. The disciplines involved in EDM are shown in Figure 1. EDM methods are a combination of methods which come from statistics, ML, DM, and other fields. It can be roughly divided into six categories: data extraction, prediction, association mining, structure discovery, model-based recognition, and hybrid methods.

Predicting students’ performance is one of the most important issues in the field of EDM. By predicting students’ performance, we can identify the risk of students’ academic failure in the learning process as early as possible, so as to intervene and guide in advance. It can also provide recommendation basis for personalized learning, and provide support for educational administrators to make decisions by analyzing the factors affecting students’ performance. The basic process of predicting students’ performance using EDM is to collect students’ historical academic records and label them with performance level or GPA, use classification or regression algorithms from ML to establish a prediction model, which is trained by the labeled data, and apply the trained model to predict students’ performance in various applications after evaluation.

Although EDM is a new research field, it has developed rapidly. Some researchers have reviewed important studies in the field of EDM and listed problems in education that researchers try to solve by using EDM methods. Romero et al. reviewed the most cited papers and remarkable books in the field of EDM, discussed some important issues such as sources of educational data, the development of EDM, the tools and datasets used, and prospected the future development of EDM. Abdul et al. classified the EDM methods and listed the related studies according to the problems in education that researchers try to solve by using EDM methods. Said et al. listed the application of DM methods in education, and evaluated the possibility of various ML methods used in the field of Education. Ashish et al. provided a systematic review of literatures on clustering algorithm and its applicability in the context of EDM. These literatures had covered studies of more than three decades (1983–2016). Some researchers used the method of literature review to list the algorithms used in predicting students’ performance and the latent factors affecting students’ performance. Contributors of existed surveys used the method of systematic literature review (SLR) methodology to analyze keywords or abstracts to classify the studies in the field of EDM, focusing on the research direction and development trend in this field. However, researchers did not make a comprehensive analysis of the procedures used to establish the students’ performance prediction model, nor did they provide an overview of latest EDM methods used in other steps except establishing the prediction model.

To investigate the latest EDM methods used by researchers to predict student performance, this survey systematically reviews import studies about predicting students’ performance using EDM in 2016–2021, and synthesizes the procedure of establishing prediction model of students’ performance. According to the proposed procedure, we compare and analyze the EDM methods used in some key steps such as data collection and labeling, data preprocessing, establishing prediction model, evaluation and interpretation. The main contributions of this survey are as follows.
1. We synthesize the procedure of building prediction model of students’ performance using EDM method. This procedure contains four major phases and 10 key specific steps.

2. According to the proposed procedure, the EDM methods used in all steps of previous studies are investigated, merits and defects of different EDM methods are discussed and compared.

3. We summarize the previous studies on the predicting students’ performance using EDM, and put forward some suggestions for future work.

The remainder of this article is organized as follows: Section 2 describes the problem definition of students’ performance prediction, and briefly introduces and discusses related works. Section 3 outlines the survey methodology that we adopt in this research, as well as the research questions and objectives that we intend to address. We synthesize the procedure of building prediction model of students’ performance by using EDM method and discuss the EDM methods used in the key steps of previous studies in Section 4. Section 5 responds to the three research questions in this article and presents challenges of previous studies. We summarize this article and give some suggestions for future work in Section 5.

2 | RELATED WORKS

Although EDM is a new research field, there are many excellent studies in recent years. Some researchers have summarized previous studies in the field of EDM. Rambola et al. classified EDM methods into three categories: prediction, clustering, and relationship mining, and listed the ML or DM algorithms used in each category of methods. But this work did not sort out the use of EDM methods to solve problems in the field of education. 

Abdul et al. further divided EDM methods into six categories: data extraction, prediction, relationship mining, structure discovery, model discovery, and hybrid methods, introduced the ML or DM algorithms used in each category, classified some previous studies according to the EDM methods used, and pointed out that method of prediction which takes advantages of classification, regression, and density estimation algorithm of ML is the most commonly used EDM method. 

Said et al. reviewed more than 10 studies in EDM fields from 2016 to 2019, introduced the EDM methods used in these studies, and the educational problems they tried to solve. Said et al. also introduced the specific algorithms used in these studies to predict students’ performance, including decision tree (DT), support vector machine (SVM), naive Bayes, logistic regression, artificial neural network (ANN), and so on. 

Hanan et al. used the method of literature survey to summarize nearly 400 studies in the field of computer supported learning analysis (CSLA), computer supported predictive analysis (CSPA), computer supported behavioral analysis (CSBA), and computer supported visualization analysis (CSV A) from 2000 to 2017, and listed 10 EDM methods to solve the problems in the field of education, such as classification, clustering, data visualization, statistics, association rules, regression, sequential pattern mining, text mining, correlation analysis, and outlier detection. 

They also pointed out that the classification algorithm of ML can effectively predict students’ performance. Ashish et al. used the method of literature review to summarize nearly 160 studies using clustering algorithm in the field of EDM in recent 30 years (1983–2016), and listed the educational problems that researchers tried to solve in these studies, the specific algorithms used, the datasets used and their sources. They divided the problems in the field of education solved by clustering algorithm into five categories: analyzing students’ motivation and behavior, understanding students’ learning style, digital learning, and collaborative learning. 

Angelos et al. reviewed the application of soft computing methods such as ANN, DT, Bayesian, random forest (RF), and SVM in the field of EDM, covering nearly 300 studies in recent years. They enumerated the studies of soft computing methods in students’ learning process evaluation, learning result prediction, user system interaction quality evaluation, recommendation system to assist students’ learning, management decision support, and other aspects. 

From the statistical results of this study, it can be seen that predicting students’ performance based on classification algorithm is one of the most important problems in EDM, and more than 60% of studies in the field of EDM are related to this issue. Romero et al. comprehensively reviewed the studies in the field of EDM, introduced the most important international conferences, main publications, highly cited papers, and other information in the field of EDM. He also summarized the framework of EDM, the data generated by different educational environments, and the main tools used.

There are also some surveys on the predicting students’ performance using EDM. Shahiri et al. used the method of literature analysis to review studies on using DM methods to predict students’ performance from 2002 to 2015, 

and pointed out that the influencing factors of students’ performance mainly included CGPA, homework performance, quiz, classroom attendance, and other internal evaluation, as well as demographic information such as gender, age, family background, and so on. Algorithms that come from ML such as DT, ANN, naive Bayes, SVM, and so on are commonly used to establish the prediction model of students’ performance. Abdallah et al. used SLR method to analyze 62 papers from
IEEE Xplore, web of science, ACM, and other databases from 2010 to 2020 with a focus on three perspectives: the forms in which the learning outcomes are predicted, the predictive analytics models developed to forecast student learning, and the dominant factors impacting student outcomes. Saa et al. critically reviewed 36 research articles out of a total of 420 from 2009 to 2018 and analyzed by applying an SLR approach. Their results showed that the most common factors are grouped under four main categories, namely, students’ previous grades and class performance, students’ e-Learning activity, students’ demographics, and students’ social information. The most common DM techniques used to predict and classify students’ factors are DTs, naïve Bayes classifiers, and ANNs. Guerrero et al. reviewed 64 studies in recent 6 years and listed the algorithms and main objectives of each work. The results show that most of the ML methods used to predict students’ performance are supervised learning, and a few researchers use recommendation systems. The objectives of prediction mainly include students’ dropout, students’ performance, activities recommended and students’ knowledge. Anoopkumar et al. reviewed different DM methods especially the mostly used and trendy algorithms applied to EDM context. Bonde et al. listed the methods used and the main findings in 56 relevant literatures. Kamakshamma et al. explored the introduction to predictive analytics, applications of predictive analytics, and analysis of some predictive analytics tools in mining of educational data. Khan et al. presented a systematic review of EDM studies on students’ performance in classroom learning. It focused on identifying the predictors, methods used for such identification, time and aim of prediction. Kumar et al. identify different student attributes that are mainly used for predicting the student performance and identify and understand the five classifiers that include DT, NB, RB, KNN, and ANN, which are mainly used for predicting the student performance. Shingari et al. collects and consigns writing, distinguishes considerable work, and intervene it to processing instructors and expert bodies.

The comparison of existing surveys is shown in Table 1.

| Author          | Publication year | Covered year | Paper reviewed | Key research questions                                                                 |
|-----------------|------------------|--------------|----------------|----------------------------------------------------------------------------------------|
| Shahiri et al.  | 2015             | 2002–2015    | Not indicated  | 1. Important attributes used in predicting students’ performance                          |
|                 |                  |              |                | 2. Classifiers used for predicting students’ performance                                   |
| Abdallah et al. | 2021             | 2010–2020    | 62             | 1. The forms in which the learning outcomes are predicted                                 |
|                 |                  |              |                | 2. The predictive analytics models developed to forecast student learning                 |
|                 |                  |              |                | 3. The dominant factors impacting student outcomes.                                       |
| Saa et al.      | 2019             | 2009–2018    | 36             | 1. The factors affecting students’ performance in higher education                        |
|                 |                  |              |                | 2. Classifiers used to analyze and predict the students’ performance                      |
| Guerrero et al. | 2020             | 2014–2020    | 64             | 1. The machine learning method used in the proposed prediction model                     |
|                 |                  |              |                | 2. The objective of prediction                                                           |
| Anoopkumar et al.| 2016             | Not indicated| Not indicated  | 1. Data mining methods especially the mostly used and trendy algorithms applied to EDM context |
| Bonde et al.    | 2020             | Not indicated| 56             | 1. EDM method used in findings literatures                                                |
| Kamakshamma et al.| 2020          | 2010–2018    | Not indicated  | 1. Introduction to predictive analytics                                                  |
|                 |                  |              |                | 2. Applications of predictive analytics                                                   |
|                 |                  |              |                | 3. Analysis of some predictive analytics tools                                            |
| Khan et al.     | 2021             | 2000–2018    | 140            | 1. The factors influencing student performance in classroom learning                     |
|                 |                  |              |                | 2. The methods are used for finding these factors.                                       |
|                 |                  |              |                | 3. Possibility to predict before course commencement and the actual grade or score.     |
| Kumar et al.    | 2017             | 2007–2016    | 24             | 1. Important factors of students used for predicting students’ performance                |
|                 |                  |              |                | 2. Classifiers commonly used by researchers                                               |
|                 |                  |              |                | 3. Accuracy of different prediction models                                                |
Most contributors to existed surveys adopted the method of SLR to analyze different aspects of previous studies, such as publication year, main contributors, source of experimental dataset, and so on. Many studies are classified according to the EDM methods used or the problems solved in the field of education. Unfortunately, researchers did not make a comprehensive analysis of the procedures used to establish the students’ performance prediction model, nor did they provide an overview of latest EDM methods used in other steps except establishing the prediction model. Due to the defects of the existing surveys mentioned above, researchers cannot identify the essential process of establishing prediction models of students’ performance in different studies, and cannot comprehend commonly used EDM methods and their performance in different steps of building prediction models that are useful to further optimize the prediction model of students’ performance. At the same time, it is hard to investigate and summarize the main challenges and future directions in this research field.

3 | METHODOLOGY

In this article, we employed a simplified SLR methodology to perform a systematic review where the relevant academic works predicting students’ performance using EDM were identified, selected, and critically evaluated using several criteria, as presented in the results section. To streamline our contributions, we formulated three key research questions as follows:

- RQ1—Procedure of establishing prediction model. What procedure do researchers follow to establish the students’ performance prediction model? What are the main steps?
- RQ2—The EDM methods used in different steps. What EDM methods do researchers use in different steps in the procedure of building students’ performance prediction model?
- RQ3—Main challenges of previous studies. What are the challenges for previous studies in this field?

The main purpose of this survey is to comprehensively analyze the procedure and the latest EDM methods used by researchers in the past 5 years according to relevant literatures. In this survey, we performed search queries in online databases including Web of Science, Engineering Village, IEEE Xplore, and Science Direct. Figure 2 summarizes the general steps of our survey methodology.

The search terms included (“predict” OR “forecasting”) AND (“student performance”) AND (“Machine Learning” OR “Data Mining”). We focused on ML or DM-based methods used for predicting students’ performance. We considered studies published in English between 2016 and 2021, and retrieved 546 results by searching the databases. The inclusion criteria that were applied to select studies are shown in Table 2.

We then removed duplicates and screened the remaining studies. And finally, we identified 80 studies that met the eligibility criteria. Figure 3 describes the process of selecting studies.

After selecting the literatures, the data we extracted from the literature include:
TABLE 2 The inclusion criteria that were applied to select studies

| Inclusion criteria          | Description of criteria                                                                 |
|-----------------------------|----------------------------------------------------------------------------------------|
| Core research issues        | Predicting students’ performance using machine learning or data mining methods         |
| Empirical research          | The researchers conducted an experimental evaluation of the EDM methods used           |
| Language of publication     | English                                                                                 |
| Year of publication         | From Jan 2015 to June 2021                                                             |
| Availability of text        | Full text is accessible for analysis                                                    |

FIGURE 3 Flowchart of the process of selecting studies

1. What procedure do the researchers follow to establish the prediction model?
2. The data source and capacity of training data collected by researchers.
3. What EDM methods do researchers use to preprocess the raw data?
4. What EDM methods do researchers use to select features from the raw data?
5. What classifier do the researchers use to establish the prediction model and whether the ensemble method is used?
6. What indicators do researchers use to evaluate the performance of the prediction model?
7. How do researchers interpret the predicted models and results?

4 | RESULTS

In this section, we report the results of our survey, where we will answer our research questions, and elaborate on the interesting results we came up with from the extracted data.
4.1 | RQ1—Procedure of establishing prediction model

Some researchers have clearly stated that they used CRISP-DM\textsuperscript{21} to establish a prediction model for students’ performance.\textsuperscript{22–26} The execution process of CRISP-DM is shown in Figure 4, which mainly includes the following six stages:

1. Understanding the business: Understand the task, vision, and goal of the business system and what role the DM project will play, make the project plan, and identify the relevant constraints of the project. The most important thing at this stage is to clearly identify the requirements of the DM task.
2. Understanding data: The data tables and fields to be mined are identified, and the associations between data and basic characteristics of data are analyzed.
3. Preparing data: Integrate all the required data into a dataset, and clean and transform the data if necessary.
4. Creating models: Design and create models for analysis.
5. Evaluation: The model is evaluated by experiments.
6. Deployment: The prediction model is deployed and optimized according to the usage.

Although we can use CRISP-DM to describe the basic framework of establishing prediction model of students’ performance, it is obviously insufficient because some necessary procedures and steps are absent. For example, after data preprocessing, we need to select special features from dataset to find the features with high correlation with students’ performance through experts or automatic methods to avoid the interference of noise features. In addition, the preprocessed data need to be divided into training data and test data for training and evaluating prediction model, respectively. Therefore, we synthesized the procedure of building prediction model of students’ performance using EDM method from the remaining literatures. This procedure contains four major phases and 10 key specific steps as shown in Figure 5, and the steps indicated by dashed lines are optional.

4.1.1 | Data collection

The main purpose of the data collection is to collect raw data from multiple data sources to establish the prediction model. Data sources include SIS, course LMS used by educational institutions, and various kinds of data generated in virtual or mixed learning environment such as MOOC, Educational games, and so on. In addition to collecting data from various digital systems or environments, it is also possible to collect data about learners by using questionnaires. Since most of the ML classification algorithms used to build the prediction model of students’ performance are supervised learning algorithms, the collected data need to be labeled manually or automatically.
4.1.2 Data preprocessing

There may be many defects in the raw data collected in the previous phase, which need preprocessing to further improve the quality of data. This phase mainly includes four steps: data cleaning, data discretization, data normalization, and data balancing. The main goal of data cleaning is to deal with missing data, error, duplication, and noise in the original data, and to solve the inconsistency problem in the raw data. The goal of data discretization is to transform continuous values such as age or score into discrete values to meet the requirements of many ML algorithms. The purpose of normalization is to scale the values of each feature to a specific range, generally \([-1, 1]\) or \([0, 1]\), so as to avoid the influence of different orders of magnitude between features on the classification algorithm. If the number of tuples belonging to different categories in the raw data is greatly different, unbalanced data will lead to low accuracy of the trained prediction model. It is necessary to make the data which belong to different categories in the data almost equal through data balancing, so as to improve the quality of the prediction model.

4.1.3 Establish prediction model

In order to establish a high-quality prediction model of students’ performance, researchers should select the most relevant or significant features from the preprocessed data to reduce the impact of “high-dimensional curse” on the prediction
model. Feature selection can be done manually by experts according to their experience, or automatically by some feature selection algorithms. The purpose of data split is to divide the data into training data and test data; one part is used to train the prediction model and the other part is used to evaluate the prediction model. If researchers use unsupervised ML method to establish prediction model, the step of data split is not necessary. In order to further improve the accuracy of prediction, researchers can use some ensemble methods to integrate multiple prediction models which are built by different classification algorithms to improve the accuracy of entire model.

4.1.4 Evaluation and application

In order to evaluate the performance of the prediction model established in the previous phase, it is necessary to evaluate the prediction model. Through the evaluation results, researchers can find the strategies to further optimize the feature selection and other strategies. The prediction model that meets the evaluation requirements can be used alone or integrated into other systems, such as LMS, graduation qualification examination, and improve the interpretability of the prediction results.

4.2 RQ2—The EDM method used in different steps

In previous studies, most researchers established prediction models according to the procedure described in Section 4.1, but researchers used different EDM methods in the key steps of procedure. In this section, we investigated the EDM methods used in the key steps of existing studies, compared and discussed the performance of different methods.

4.2.1 Data collection

In order to establish prediction model of students' performance, it is necessary to collect students' historical academic records and label them before they can be used to train the prediction model established by supervised classification algorithm.

Source of data

The sources of data collected by researchers can be divided into three categories: digital systems used by educational institutions such student information management system (SIS) or course LMS, and so on, public dataset repository of ML such as UCI or Kaggle, and so on, and manual collection through questionnaires or surveys. In the selected studies, the proportion of these three sources of data is shown in Figure 6.

We can see from Figure 6 that 78% of the studies collect datasets from digital systems, which also reflects that various digital systems have been widely used in educational institutions. Most of the data sources are SIS or LMS used by specific educational institutions, but the features of the data in different studies are very different, which makes the prediction
models established in different studies highly dependent on the data collected and unable to compare their performance. Another eight studies \cite{23,27-33} collected data from Kaggle, \cite{34} and seven of them used dataset named xAPI-Edu-Data. \cite{35} Two studies \cite{36,37} used StudentPerformance \cite{38} dataset from UCI Machine Learning Repository. \cite{39} Researchers who collect data manually through the questionnaire usually design the questions in the questionnaire according to the predicted specific objectives. \cite{40-47} Especially, Hoang et al. conducted an online survey with a participation of students in various courses to analyze and show the effects of different learning styles on students’ performance. \cite{43} That test is based on the Felder-Soloman questionnaire \cite{48} with 44 questions which are divided into four dimensions. Sivasakthi et al. used a test which contains 60 questions to describe the performance of students in introductory programming. \cite{49} Obviously, only the performance of prediction models that are built by datasets from public data warehouse can be compared.

### Size of data

The size of data collected by different researchers varies greatly. For example, Mengash et al. collected 2039 students’ records from a college in a university, \cite{50} while Yousafzai et al. collected nearly 80,000 students’ records from the federal board of intermediate and secondary. \cite{51} In selected studies, the statistical values of instances included in the data collected by the researchers are shown in Table 3.

It can be seen from Table 3 that the size of data in 63.5% of the selected studies is less than 1000, which is very small. The small training datasets make the prediction model easy to overfitting and be affected by noise data, and may bring the problem of data imbalance. The size of data used for training prediction model is very small, which is a common and significant problem in the previous researches. In particular, Chui et al. proposed an improved conditional generative adversarial network based on deep SVM (ICGAN-DSVM) algorithm. In this algorithm, ICGAN aims at addressing the issue of low data volume by mimicking new training dataset. \cite{74} We also see that the data used in 16.5% of the selected studies exceeds 5000 instances. The large size of training data can improve the generalization ability of prediction model.

### Labeling data

Dataset with labeled can be used to train the prediction model established by supervised learning classification algorithm. Since researchers mostly collect students’ historical academic records from SIS, LMS, and other management information systems, most of the data have been labeled by educators, which can be directly used to train the prediction models. If the size of collected data is large and has not been labeled, it is very difficult to label manually, so automatic labeling methods are used to label the data necessarily. Predicting students’ performance with learning difficulties through prediction behavior (called PPP), \cite{56} which was established by Hooshyar et al., used K-means clustering algorithm \cite{106} to automatically divide students into three categories according to the characteristic vector of students’ homework submission behavior, and labeled them as procrastination/approximate-procrastination/ non-procrastination, respectively. It is an effective method to label data automatically before classification by using appropriate clustering algorithm.

Active learning \cite{107} and semi-supervised learning \cite{108} are effective methods to deal with a large number of unlabeled data. These methods only need a small amount of data labeled by experts manually, which can greatly reduce the amount of data need to be labeled manually and improve the accuracy of automatic labeling. We discussed this issue in section “Establish prediction model” later.

### 4.2.2 Data preprocessing

The phase of data preprocessing includes four key steps which are cleaning, discretization, normalization, and balancing multi-class. The purpose of this phase is to eliminate the missing values in collected data, discretize the continuous values
of features, unify the order of magnitude between different features, and solve the problem of multi-class imbalance that may exist in the collected data. After preprocessing, the quality of data used for training prediction model is further improved.

Conventional preprocessing

Conventional preprocessing mainly includes data cleaning, discretization, and normalization. In 81.5% of the selected studies, the researchers have carried out conventional preprocessing on the data.

The methods to deal with missing values include ignoring, discarding, filling manually, filling with attribute mean or median, filling with the same sample mean, filling with values determined by regression, and other speculative means. In particular, Júnior et al. introduced eight methods used to deal with missing value: ignoring, discarding, mean import, medium import, last observation carried forward (LOCF), linear interpolation, spline interpolation, and piecewise cubic Hermite interpolating polynomial (PCHIP). They analyzed the performance of these eight methods in predicting students’ performance through experiments. The experimental results in their research show that the ignoring and discarding methods have the best effect, while the medium imputation and spline interpolation methods have the worst performance.

Since DT, naive Bayes, and other classification algorithms used to establish prediction models have better performance for nominal features, it is necessary to discretize age, score, GPA, and other numeric features in data. The most widely used discretization methods in predicting students’ performance are unsupervised discretization methods under the guidance of expert experience.

The goal of normalization is to limit the values of different features in the same range, which is generally \([-1,1]\) or \([0,1]\), so as to avoid the impact of different dimensions of features on the performance of prediction model. The most commonly used normalization method in the selected studies is min-max normalization method that changes the original data value linearly. Let an original value of attribute \(A\) be \(v\), and the maximum and minimum values of this attribute are \(\text{Max}_A\) and \(\text{Min}_A\), respectively. After normalizing, the new value is \(v'\), and the maximum and minimum values of the new value are \(\text{newMax}_A\) and \(\text{newMin}_A\), respectively. Then the value of \(v'\) is calculated by Equation (1).

\[
v' = \frac{v - \text{Min}_A}{\text{Max}_A - \text{Min}_A} \times (\text{newMax}_A - \text{newMin}_A) + \text{newMin}_A
\]  

Balancing multi-class

If the number of instances belonging to different categories in the data varies greatly, it is called multi class imbalance which has a serious impact on the accuracy of the prediction model. However, the collected students’ historical academic records are multiclass unbalanced generally. For example, in a dataset of students’ course grades, numbers of records representing medium or good are much more than those of failing or excellent; in a dataset of students’ completing undergraduate programs, the number of students labeled as failing is far less than success. Therefore, in order to improve the accuracy of the prediction model, it is necessary to deal with the multi-class imbalance in the phase of data preprocessing. Unfortunately, among the studies we selected, only five studies clearly described the multi class balance methods used by the researchers (Table 4).

We can see from Table 4 that the three multi-class balance methods resampling, oversampling, and undersampling, which are commonly used in DM, have been applied in the research of predicting students’ performance.

| Study | Multi-class balancing methods                      |
|-------|---------------------------------------------------|
| 100   | Resampling                                        |
| 92    | Resampling                                        |
| 76    | Undersampling, oversampling                       |
| 93    | Undersampling, oversampling, hybrid sampling      |
| 36    | Undersampling, oversampling                       |
Especially, Hassan et al. compared the oversampling, undersampling, and hybrid methods for solving multiclass imbalance through experiments. The experimental results showed that random oversampling (ROS) method\textsuperscript{115} had the best performance in predicting students’ performance. Pristyanto et al. proposed a method to deal with multiclass imbalance by combining synthetic minority oversampling technology (SMOTE)\textsuperscript{116} and one-sided selection (OSS).\textsuperscript{117} The experimental results showed that the processed data further improved the prediction accuracy of SVM in predicting students’ performance.\textsuperscript{36}

4.2.3 Feature selection

Selecting the features that have a significant impact on students’ performance is the key step to establish the prediction model of students’ performance. Careful feature selection can significantly improve the performance of the prediction model. Saa et al. reviewed 36 studies from 2009 to 2018,\textsuperscript{14} and identified nine types of factors influencing students’ performance, among which the most commonly used four types are students’ previous grades and class performance, students’ e-learning activity, students’ demographics, and students’ social information, and the most frequently used single factor is cumulative grade point average (CGPA). Francis et al. divided the features that affect students’ performance into four categories: demographic features, academic features, behavioral features, and extra features.\textsuperscript{23} Demographic features include student number, name, gender, age, and so on; academic features include school grade, class level, semester, and so on; behavioral features include discussion, access to resources, comments, and other learning behaviors in online learning environment. The number of features used by researchers in the selected studies is shown in Table 5.

From Table 5, we can see that 70% of the selected studies contain more than 10 characteristics in the data collected by researchers, so it is necessary to select the features that are more relevant to the establishment of prediction model from these features to avoid the impact of irrelevant features on the accuracy of prediction results. The feature selection methods used in 35 of 80 selected studies can be divided into three categories: manual selection by experts, filter-based, and wrapper feature selection methods (Table 6). Instead of feature selection, researchers directly use all the features in the collected data in 45 studies of 80 selected studies.

As can be seen from Table 6, the feature selection methods most used by researchers are manual selection and correlation-based filtering methods. Manual selection requires researchers to select the features that have a significant impact on students’ performance from the data according to the theories of pedagogy, psychology, or expert experience. The correlation-based filtering method calculates the correlation degree between each feature and student performance

| Features | Ratio | Studies |
|----------|-------|---------|
| 1–10     | 30%   | 28,36,49,52,54,55,58–60,62,63,73,77,78,86,88–90,94,95,98,102,103,105 |
| 11–50    | 62.5% | 22–25,27,29,30,32,33,37,40–45,47,50,53,56,57,61,64–72,74–76,79–82,84,85,87,91–93,96,97,99,101,104 |
| 51–100   | 2.5%  | 46,83   |
| 101–     | 5%    | 26,31,51,100 |

| Feature selection method | Ratio | Study |
|--------------------------|-------|-------|
| Manual selection         | 16%   | 23–25,29,47,57,62,77,78,81,91,101,105 |
| Filter-based method      |       |       |
| Correlation              | 16%   | 22,31,50,67,68,71,79,82,92,93,97,104 |
| InformationGain          | 2.5%  | 30,99 |
| Classifier               | 5%    | 26,46,80,98 |
| Wrapper-based method     |       |       |
| Genetic algorithm        | 3.8%  | 27,37,51 |
| PCA                      | 1.2%  | 63    |
(Equation 2), and selects some features with the highest correlation degree. CFS\(^{118}\) is the most commonly used algorithm.

\[
R(f_i, c) = \frac{\text{Cov}(f_i, c)}{\sqrt{\text{Var}(f_i) \cdot \text{Var}(c)}}
\]

where \(R(f_i, c)\) denotes correlation between feature \(f_i\) and class \(c\), \(\text{Cov}(f_i, c)\) denotes covariance of features and class, and \(\text{Var}(f_i, c)\) denotes variance of features and class.

Similar to the correlation-based filtering method, the Information Gain-based filtering method calculates the information gain of each feature and student performance (Equation 3), and selects the best feature from the data.

\[
\text{Information Gain}(A) = \text{Info}(D) - \sum_{j=1}^{v} \left| \frac{|D_j|}{|D|} \right| \times \text{info}(D_j)
\]

Since gene-based algorithms are commonly used to solve global optimization problems, many feature selection methods based on gene algorithms have been proposed. Yousafzai et al. used genetic-based algorithm (GA)\(^ {119}\) to imitate the natural selection process of biological evolution for feature selection, and continuously adjusted the feature selection method through the evaluation of prediction results, so as to achieve the best feature selection in the process of approaching the best prediction results.\(^ {31}\) Turabieh et al. used a special method called binary GA, used ANN as the feedback model of prediction results, and continuously optimized feature selection by performing three operations similar to natural evolution of selection, crossover and mutation on feature selection scheme.\(^ {37}\) The method of feature selection based on genetic algorithm proposed by Farissi et al. is called GAFS, which is similar to the two methods mentioned above. The difference is that the results of KNN, DT, RF, and naive Bayes are used to optimize feature selection.\(^ {120}\) In particular, Khasanah et al. used five different methods, correlation-based, gain-ratio, information-gain, relief, symmetrical uncertainty, and took the common part of the selection results of different methods as the result of feature selection.\(^ {68}\)

### 4.2.4 Data split

After preprocessing and feature selection, the data need to be divided into two parts: one is used to train the prediction model, and the other is used to evaluate the performance of the prediction model. The researchers used the k-fold cross validation\(^ {121}\) to divide the data in 55 of 80 selected studies to make full use of the collected data, parameter \(k\) is generally set to 10. This method is suitable for the case where the size of the collected data is small. The researchers scale data manually in 25 of 80 selected studies, the division proportion is generally 70%–30% or 80%–20%.

### 4.2.5 Establish prediction model

In the selected 80 studies, almost all researchers (77 of 80) use supervised classification algorithm that comes from ML to establish prediction model of students’ performance. In order to investigate the performance of different classifiers, researchers generally use more than one classifier to establish prediction models in a study. Figures 7 and 8 show the statistics of supervised classifiers used in 77 of 80 selected studies.

As can be seen from Figures 7 and 8, 69% of the selected studies used more than one classifier, and the classifiers used by researchers more than 10 times were DT, NB, MLP, RF, SVM, LR, and KNN. The classification algorithms used by researchers are basically from Weka\(^ {122}\) or RapidMiner\(^ {123}\) and other open source DM tools. Saa et al. summarized the classification algorithms used in 36 studies from 2009 to 2018.\(^ {124}\) A total of 74 different classification algorithms were used in these studies, among which the most used eight algorithms were Naive Bayes, SVM, logistic regression, KNN, DT (ID3, C4.5, C5.0), and ANN. We discuss the application of these mostly used classifiers in predicting students’ performance as follows

1. DT is based on information entropy, while naive Bayes (NB) is based on probability. Both models are white box so researchers can interpret the models intuitively and discover the significant factors affecting students’ performance.
In addition, the training and prediction speed of these two models are very fast so they are generally welcomed by researchers.

2. According to the universal approximation theorem, an ANN with hidden layer and nonlinear mapping has the ability to approximate any function. Multilayer perceptron (MLP) is one of the most commonly used ANNs, which has the significant advantages of high fitting performance. MLP can also ensure good generalization performance by means of penalty coefficient and dropout etc.

3. RF is an ensemble learning model based on DT. It has the advantages of simple implementation, noise insensitive, and strong anti-overfitting ability, generally has higher prediction accuracy than a single DT.

4. Logistic regression is a classical regression algorithm, which is mainly used to solve the problem of binary classification. Logistic regression has the advantages of fast training speed, easy to understand, and easy to use, but it is difficult to deal with the problem of data imbalance and nonlinear division.

5. SVM is used for binary classification like LR, but it does not require the probability distribution of training data. SVM is suitable for the problem of binary classification students’ performance prediction with a large number of features.

6. K-nearest neighbor (KNN) is one of the simplest ML algorithms and is a very mature classification algorithm. The merits of KNN such as easy to use, insensitive to noise points, and so on, which make it one of the most effective algorithms to predict students’ performance. However, this algorithm is slow and easy to be interfered by irrelevant features, so it is only suitable for students’ academic performance dataset with small size and small number of features.

The improved ANNs used by researchers mainly include convolutional neural networks (CNN), recurrent neural network (RNN), and so on. In addition to the classical ANN algorithm such as MLP, and so on, some researchers have used more complex deep neural network to establish the prediction model of students’ performance, such as Tao et al. using graph convolution network, Chui et al. using counter neural network, and so on.

Some researchers try to use a variety of ML algorithms to establish prediction model of students’ performance. Almasri et al. proposed a unified framework to establish prediction model. The framework used clustering technique to group historical records of students into a set of homogeneous clusters. Then classifier model for each cluster is built and the final
unified classifiers along with the centroids at each cluster are used to establish prediction model of students’ performance. Francis et al. proposed the method of building prediction model. This method first uses the classical clustering algorithm k-means to group students, and then uses different classification algorithms to build prediction model for different groups, which also takes into account the heterogeneity between students.56

Active learning129 and semi-supervised learning108 are effective method to deal with a large number of unlabeled data. These methods only need a small amount of data labeled by experts manually, which can greatly reduce the amount of data need to be labeled manually and improve the accuracy of automatic labeling. In addition to using supervised learning methods in the selected studies, researchers also began to use semi-supervised learning, active learning and other methods to predict students’ performance. Kostopoulos et al. used specific algorithms to query the most useful unlabeled data and mark them by experts manually to reduce the number of data need to be marked manually.64 Tsiakmaki et al. introduce a fuzzy-based active learning method for predicting students’ academic performance which combines, in a modular way, autoML practices. Their experimental results revealing that the proposed method for the accurate prediction of students at risk of failure has better performance compared with the classical classifier.130 Hussain et al. proposed a method to predict students’ performance by combining unsupervised clustering and association rule mining methods with supervised classification methods.77

4.2.6 Ensemble method

In order to make full use of the advantages of different models, there are many researches using the ensemble method to integrate multiple prediction models. For the many prediction results generated by multiple prediction models, the ensemble method uses voting or weighting to reselect or combine these results, which produces better prediction results than single models. Research uses ensemble method in 19 of the selected 80 studies, and the number of ensemble method used is as shown in Figure 9.

From Figure 9, we can see that the classic boosting, bagging, and voting methods are the most used ensemble methods by researchers. The experimental results of many studies show that the prediction accuracy of the model using the ensemble method is higher than that of the model using a single algorithm. Bhat et al. reviewed the commonly used ensemble methods in the establishment of students’ performance prediction model, and used a dataset to carry out empirical research on different ensemble methods. The experimental results show that the boosting method has the best effect in its experimental environment.76 Hassan et al. empirically analyze five types of ensemble classifiers and seven sampling techniques. The experimental results show a hybrid technique ROS with AdaBoost produces the most excellent performance compared to the other benchmark techniques.93

4.2.7 Model evaluation

After the prediction model is established, it is necessary to compare the performance of different prediction models through evaluation, so as to select a more appropriate model. In EDM, predicting student performance is considered as a classification or regression problem. Therefore, researchers mainly use performance indicators of classification and

**FIGURE 9** Occurrences of different ensemble method in selected studies
regression in ML to evaluate the performance of prediction model. The performance indicators used in the selected 80 studies are shown in Figure 10.

From Figure 10, we can see that accuracy, precision, recall, F-measure, and sensitivity\(^{131}\) are the most commonly used performance indicators to evaluate the prediction model. The data collected by researchers are often multi-class unbalanced. Although accuracy is most used, it is not the best performance indicator to evaluate the prediction model. Precision and F-measure reflect the prediction accuracy of different class rather than all samples, and are more suitable for evaluating the performance of student performance prediction model.

It should be mentioned that although almost all the previous studies have used the above indicators to evaluate the established prediction models, and compared the different prediction models established in the same study, the prediction models established in different studies cannot be compared, because the training data and features used in previous studies are very different. In different studies, the conclusion about which algorithm is the best is very different. It can be seen that the prediction models established in the previous studies are extremely dependent on datasets, which are collected in specific studies, and it is very difficult to use this prediction models in other environments or systems.

4.2.8  |  Model interpretation

The main goal of predicting students’ performance is to identify the factors that have a significant impact on students’ performance and help instructor to intervene as early as possible. Therefore, this is an ML or DM task with requirement of high interpretability. However, very few studies have paid attention to the interpretability of the prediction model, that is, using pedagogy, cognitive psychology, and other learning theories to explain the principle of the prediction model, the prediction process and why different features can affect the prediction results. Xing et al. used activity theory\(^{132}\) to comprehensively quantify the activities of students participating in computer supported collaborative learning (CSCL) course. After collecting quantitative data, they used algorithm based on gene programming to establish prediction model.\(^{133}\) Sorour et al. proposed a method to extract “IF-THEN” rules from the prediction model established by DT, which improved the interpretability of prediction results.\(^{88}\) Arnedo et al. proposed a black-box technique to take advantage of the power and versatility of these methods, while making some decisions about the input data and design of the classifier that provide a rich output data set. A set of graphical tools is also proposed to exploit the output information and provide a meaningful guide to teachers and students.\(^{59}\)

4.3  |  RQ3—Main challenges of previous studies

Although previous studies have achieved considerable achievements, researchers in this field still face a series of challenges as follows:

1. Predicting students’ performance is one of the most important issues of EDM which requires high interpretability. The main purpose of predicting students’ performance is to find out the important factors that affect students’ learning.
According to the prediction results, instructor can intervene students’ learning process as soon as possible to achieve the goal of improving students’ learning performance. Although EDM is an interdisciplinary subject integrating ML, DM, pedagogy, and cognitive psychology, the prediction models currently established in previous studies generally lack the support of theories such as ML interpretability theory, pedagogy, and cognitive psychology, and cannot well identify and analyze the important factors affecting students’ performance. Few researchers focus on how to improve interpretability of prediction model such as choosing interpretable prediction model, using partial dependence, feature interaction, and other methods to explain the black box prediction model such as ANN, interpreting the prediction process of the sample, and so on, the interpretability of prediction process and results is not high.

2. The size and quality of historical academic data used in training prediction model need to be further improved. Most of the datasets used to train students’ performance prediction model are collected by different researchers themselves. The data source is single, the size of dataset is small, and the standards of datasets are not unified. There is a big gap between these datasets and the standard datasets used in other ML fields, such as UCI machine learning repository and Image-Net. Because the standards of datasets used in training prediction models are not unified, the prediction models established in different studies cannot be evaluated and compared in a unified way, and these models cannot be transplanted in different systems and environments. However, different from the classical classification and prediction problems such as image classification or speech recognition, predicting students’ performance is a very open problem, and the goals of prediction are varying greatly in different educational institutions and information systems.

3. The key technologies used to establish prediction model need to be further optimized. Most of the algorithms used in previous studies to establish the students’ performance prediction model directly come from the field of ML or DM. These algorithms are not adjusted and optimized according to the characteristics of the educational dataset and the task of predicting student performance. As a result, the performance and accuracy of the prediction model built by these algorithms need to be further improved. Most of the existing researches are ML algorithms based on supervised learning, which need a lot of labeled data for training model. The workload of data annotation is large, and the application of semi-supervised and unsupervised learning methods is lack.

5 | CONCLUSION AND FUTURE DIRECTIONS

5.1 | Key findings

EDM is one of the most advanced and effective ways to analyze educational big data. Predicting students’ performance is the most concerned research field of EDM. Almost half of the researches in the field of EDM are related to predicting students’ performance. Our survey makes a detailed investigation of nearly 80 studies using EDM to predict students’ performance in the past 5 years, providing a comprehensive understanding and practical guide for researchers in this field.

We synthesize the procedure of building prediction model of students’ performance by using EDM method; according to the procedure, the EDM methods used in the key steps of previous studies are discussed and compared. Below, we revisit each research question separately and highlight the main findings.

1. RQ1—Procedure of establishing prediction model. What procedure do researchers follow to establish the student’ performance prediction model? What are the key steps?

   The procedure of establishing prediction model includes four phases as follows: data collection, data preprocessing, establishing prediction model, evaluation, and application. It includes 10 key steps: collecting raw data, labeling data, handling missing value, discretization, normalization, balancing, feature selection, data split, establish prediction model, evaluation, and interpretation.

2. RQ2—The EDM method used. What EDM methods do researchers use in different steps in the procedure of building students’ performance prediction model?

Researchers have used many EDM methods in the 10 key steps of building prediction models. We summarize the most frequently used methods as follows: (1) Researchers mainly collect raw data from LMS, SIS, and other information management systems, and volume of data is generally small (less than 1000 instances); (2) Researchers mainly label the collected raw data manually, meanwhile some researchers use the automatic labeling method based on clustering
algorithm; According to the objective of prediction, researchers choose the labeled categories including binary labels such as success or fail, multi-class labels such as fail, pass, normal, excellent, and so on, or numeric label such as GPA. (3) Researchers use the methods of ignoring, discarding, filling manually to handle missing values in the data. Some experimental results show that the discarding method is the best; (4) Researchers use the unsupervised discretization method of equidistant or equifrequency automatic division to discretize the numerical values in the data; (5) The most commonly used normalization method is the min-max method; (6) Researchers use oversampling, undersampling, and hybrid methods to solve the problem of multi-class imbalance in data; (7) N-fold cross validation is a method that most researchers use to divide the data into training data and test data; (8) Most researchers use expert experience to manually select features, while some researchers use correlation-based or gene-based feature selection methods, which can significantly improve the accuracy of prediction; (9) Almost all researchers use supervised classification algorithms from ML to establish prediction models. The most commonly used algorithms are DT, NB, MLP, RF, SVM, KNN, and so on. Bagging, boosting, and other ensemble methods are used to integrate multiple prediction models by many researchers; (10) Accuracy, precision, recall, F-measure, and sensitivity are the most commonly used performance indicators to evaluate the prediction model.

3. RQ3—Main challenges of previous studies. What are the deficiencies in this field that can be further improved?

Although researchers have made obvious achievements in this field, there are still three deficiencies as follows. The EDM methods used by researchers are directly from the field of DM or ML, and they do not modify or optimize these methods according to characteristics of educational data and students’ performance prediction. Few researchers pay attention to the interpretability of prediction models, the weak interpretability of the prediction process, and the results will make it difficult for educators to identify the factors that have a significant impact on students’ academic performance, and also make prediction results highly questionable. The quality of data used for training prediction model needs to be further improved; in detail, researchers need to collect more students’ historical academic data and unify the standard of dataset etc.

5.2  Survey limitations

The limitations of this survey may come from the following aspects:

1. Although we try our best to collect important literature in this field between 2016 and 2021, we may miss some high-quality studies in this field due to the search keywords we choose. In addition, our organization does not have permission to access the full text of literature in databases such as world scientific and IGI global; the restriction of obtaining the full text of literature also leads us to miss the information extraction and evaluation of some literature.
2. Because most researchers use private datasets collected by them to establish prediction models, we cannot make empirical comparison between different prediction models.
3. In some studies, researchers did not fully report the details of establishing prediction models and experiments, which posed a challenge for us to extract information and evaluate the study.

5.3  Future directions

According to the above summary and discussion, we put forward the following suggestions for the future work of predicting students’ performance using EDM as follows:

1. Researchers should pay more attention to the interpretability of prediction process and results. We should make full use of modern teaching and learning theories and ML interpretability theories and methods to improve the interpretability of the prediction process and results, so that the prediction model can play a greater role in helping instructors and students.
2. Researchers should constantly improve the quality of educational data used to train students’ performance prediction models. Due to the small scale and non-uniform of educational data used by researchers in previous studies, the performance of prediction models established in different studies cannot be evaluated and compared, and these models cannot be transplanted between different environments and systems. We need to develop data standards and increase
data sources, establish a unified and optimized dataset for predicting students’ performance, and improve the portability and evaluability of the prediction model. At present, kaggle has become the most active digital community of DM and ML in the world. Every day, a large number of researchers and technical experts constantly improve the technology and effect of ML tasks of different types and goals through competitions. Researchers can also further optimize EDM methods used for predicting students’ performance by organizing competitions similar to those in the kaggle community.

3. Researchers should continue to optimize the procedure of establishing students’ performance prediction model and the EDM methods used to build prediction models. At present, researchers are still based on the general DM process to establish student performance prediction model, without considering the characteristics of EDM. Most of the EDM methods used by researchers in the previous studies to build students’ performance prediction model are directly from the field of DM and ML, and they are not optimized and integrated according to the characteristics of educational data. Under the guidance of modern teaching and learning theories, we should further optimize the EDM methods used to establish prediction models, such as data collection, feature selection, model establishment and evaluation, and constantly improve the accuracy and ease of use of prediction models. We should try to use semi-supervised or active learning algorithms to build students’ performance prediction model for reducing the workload of labeling data and improving performance of prediction model.

ACKNOWLEDGMENT
This work was supported by a funding project for top discipline talents of the Anhui Provincial Department of Education in 2020 under Grant gxbjzd2020101.

PEER REVIEW
The peer review history for this article is available at https://publons.com/publon/10.1002/eng2.12482.

DATA AVAILABILITY STATEMENT
Data sharing not applicable to this article as no datasets were generated or analysed during the current study

CONFLICT OF INTEREST
The authors have no conflict of interest relevant to this article.

ORCID
Wen Xiao https://orcid.org/0000-0003-1444-908X

REFERENCES
1. Srinivasan S. Data mining with big data. Comput Rev. 2015;56(9):556-557.
2. Milevski Z, Golova E, Zdravev Z. Big data for education: data mining, data analytics, and web dashboards. Europsko građansko procesno pravo – izabrane teme; 2012.
3. Romero C, Ventura S. Educational data mining: a review of the state of the art. IEEE Trans Syst Man Cybern C. 2010;40(6):601-618.
4. Algarni A. Data Mining in Education. Int J Adv Comput Sci Appl. 2016;7(6):456–461.
5. Aleem A, Gore MM. Educational data mining methods: a survey. Proceedings of the 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT); 2020.
6. Charitopoulos A, Rangoussi M, Koulouriotis D. On the use of soft computing methods in educational data mining and learning analytics research: a review of years 2010–2018. Int J Artif Intell Educ. 2020;30(3):371–430.
7. Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Mining in educational data: review and future directions. Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020); 2020.
8. Dutt A, Ismail MA, Herawan T. A systematic review on educational data mining. IEEE Access. 2017;5:15991-16005.
9. Aldowah H, Al-Samarraei H, Fauzy WM. Educational data mining and learning analytics for 21st century higher education: a review and synthesis. Telematics Inform. 2019;37(April):13-49.
10. Shahiri AM, Husain W, Rashid NAA. A review on predicting Student’s performance using data mining techniques. Proc Comput Sci. 2015;72:414-422.
11. Rastrollo-Guerrero JL, Gomez-Pulido JA, Durán-Dominguez A. Analyzing and predicting students’ performance by means of machine learning: a review. Appl Sci. 2020;10(3):1042.
12. Wohlin C, Prikладники R. Systematic literature reviews in software engineering. Inf Softw Technol. 2013;55(6):919-920.
13. Namoun A, Alshanqiti A. Predicting student performance using data mining and learning analytics techniques: a systematic literature review. Appl Sci. 2020;11(237):1-28.
14. Saa AA, Shaalan MAEK. Factors affecting students’ performance in higher education: a systematic review of predictive data mining techniques. Technol Knowl Learn. 2019;2019(24):567-598.
15. Anoopkumar M, Rahman A. A review on data mining techniques and factors used in educational data mining to predict student amelioration. Proceedings of the International Conference on Data Mining & Advanced Computing; 2016.
16. Bonde SN, Kirange DK. Educational data mining survey for predicting Student's academic performance. Proceeding of the International Conference on Computer Networks, Big Data and IoT (ICCCI - 2018); 2020.
17. Kamakshamma V, Bharati KF. Predictive Analytics in Mining of Educational Data: A Research Travelogue. Springer; 2020.
18. Khan A, Ghosh SK. Student Performance Analysis and Prediction in Classroom Learning: A Review of Educational Data Mining Studies. Vol 2. Education and Information Technologies; 2020.
19. Kumar M, Singh AJ, Handa D. Literature survey on student’s performance prediction in education using data mining techniques. Int J Educ Manag Eng. 2017;6(6):40-49.
20. Shingari I, Kumar D. A survey on various aspects of education data mining in predicting student performance. J. appl. sci. comput. 2018;5(6):38-42.
21. Wirth R, Hipp J. CRISP-DM: towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining; 2000; Springer-Verlag: London, UK.
22. Qazdar A, Er-Raha B, Cherkauel C, Mammass D. A machine learning algorithm framework for predicting students performance: a case study of baccalaureate students in Morocco. Educ Inf Technol. 2019;1:3577-3589.
23. Francis BK, Babu SS. Predicting academic performance of students using a hybrid data mining approach. J Med Syst. 2019;2019(24.6):3577-3589.
24. Tarmizi SS, Mutalib S, Hamid NH, Abdul-Rahman S, Ab Malik AM. A case study on student attrition prediction in higher education using data mining technique. Proceedings of the International Conference on Soft Computing in Data Science; 2019.
25. Moscoso-Zea O, Saa P, Luján-Mora S. Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. Australas J Eng Educ. 2019;24:1-10.
26. do Nascimento RL, das Neves Junior RB, de Almeida Neto MA, de Araújo Fagundes RA. Educational data mining: an application of regressors in predicting school dropout. Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition; 2018; Cham.
27. Wafi M, Faruq U, Supianto AA. Automatic feature selection for modified K-nearest neighbor to predict student's academic performance. Proceedings of the 2019 International Conference on Sustainable Information Engineering and Technology (SIET); 2019.
28. Hmiedi I, Najadat H, Halloush Z, Jalabneh I. Semi supervised prediction model in educational data mining. Proceedings of the 2019 International Arab Conference on Information Technology (ACIT); Vol. 3, December 2019:27-31; IEEE.
29. Brohi SN, Pillai TR, Kaur S, Kaur H, Sukumaran S, Asirvatham D. Accuracy comparison of machine learning algorithms for predictive analytics in higher education. Proceedings of the International Conference for Emerging Technologies in Computing; 2019.
30. Amrieh EA, Hamtini T, Aljarah I. Mining educational data to predict student's academic performance using ensemble methods. Int J Database Theory Appl. 2016;9(8):119-136.
31. Kumar AD, Selvam RP, Palanisamy V. Hybrid classification algorithms for predicting student performance. Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS); 2021; IEEE.
32. Saleem F, Ullah Z, Fakieh B, Kateb F. Intelligent decision support system for predicting student’s E-learning performance using ensemble machine learning. Mathematics. 2021;9(17):2078.
33. Rahman MH, Islam MR. Predict student's academic performance and evaluate the impact of different attributes on the performance using data mining techniques. Proceedings of the 2017 2nd International Conference on Electrical & Electronic Engineering (ICEEEE); 2017.
34. community, K. datasets; 2021. https://www.kaggle.com/datasets
35. Amrieh EA, Hamtini T, Aljarah I. Students’ academic performance dataset (xAPI-Edu-Data); 2016.
36. Pristyanto Y, Pratama I, Nugraha AF. Data level approach for imbalanced class handling on educational data mining multiclass classification. Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT); 2018.
37. Turabieh H. Hybrid machine learning classifiers to predict student performance. Proceedings of the 2019 2nd International Conference on New Trends in Computing Sciences (ICTCS); 2019.
38. Cortez F. Student performance data set; 2008.
39. Dua D, Graff C. UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science; 2019.
40. Widiyaningsih Y, Fitriani N, Sarwinda D. A semi-supervised learning approach for predicting student's performance: first-year students case study. Proceedings of the 2019 12th International Conference on Information & Communication Technology and System (ICTS); 2019.
41. Lekan AJ. Student academic performance prediction using support vector machine; 2019.
42. Jove E, Vázquez JA, Fernández-Ibáñez I, Casteleiro-Roca JL, Rolle JL. Hybrid intelligent system to predict the individual academic performance of engineering students. Int J Eng Educ. 2018;34(3):895-904.
43. Binh HT, Duy BT. Predicting students’ performance based on learning style by using artificial neural networks. Proceedings of the 2017 9th International Conference on Knowledge and Systems Engineering (KSE); 2017.
44. Sikder MF, Uddin MJ, Halder S. Predicting students yearly performance using neural network: a case study of BSMRSTU. Proceedings of the International Conference on Informatics; 2016.
45. Kasthuriarachchi K, Liyanage S. Predicting students’ academic performance using utility based educational data mining. Proceedings of the International Conference on Frontier Computing; 2018; Springer.

46. Hegde V, Prageeth PP. Higher education student dropout prediction and analysis through educational data mining. Proceedings of the 2018 2nd International Conference on Inventive Systems and Control (ICISC); 2018.

47. Nahar K, Shova BI, Ria T, Rashid HB, Islam AS. Mining educational data to predict students performance. Educ Inf Technol. 2021;1-17.

48. Felkel B, Gosky RM. A study of reliability and validity of the felder-soloman index of learning styles for business students; 2013.

49. Sivasakthi M. Classification and prediction based data mining algorithms to predict students' introductory programming performance. Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI); 2017.

50. Mengash HA. Using data mining techniques to predict student performance to support decision making in university admission systems. IEEE Access. 2020;8:55462-55470.

51. Yousafzai BK, Hayat M, Afzal S. Application of Machine Learning and Data Mining in Predicting the Performance of Intermediate and Secondary Education Level Student. Vol 25. Education and Information Technologies; 2020:3.

52. Widyahastuti F, Tjhin VU. Predicting students performance in final examination using linear regression and multilayer perceptron. Proceedings of the International Conference on Human System Interactions; 2017.

53. Uskov VL, Bakken JP, Byerly A, Shah A. Machine learning-based predictive analytics of student academic performance in STEM education. Proceedings of the 2019 IEEE Global Engineering Education Conference (EDUCON); 2019.

54. Ramanathan L, Geetha A, Khalid M, Swarnalatha P. Apply of sum of difference method to predict placement of students’ using educational data mining. Information Systems Design and Intelligent Applications. Springer; 2016:367–377.

55. Yusof MH, Abd Khalid I. Precision education reviews: a case study on predicting student's performance using feed forward neural network. Proceedings of the 2021 International Conference of Technology, Science and Administration (ICTSA); 2021.

56. Hooshyari D, Pedaste M, Yang Y. Mining educational data to predict students’ performance through procrastination behavior. Entropy. 2019;22(1):12.

57. Harvey JL, Kumar SA. A practical model for educators to predict student performance in K-12 education using machine learning. Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI); 2019.

58. Breikli BA, Zaki N, Mohamed EA. Using educational data mining techniques to predict student performance. Proceedings of the 2019 International Conference on Electrical and Computing Technologies and Applications (ICECTA); 2019.

59. Villagrá-Arnedo CJ, Gallego-Durán FJ, Llorens-Largo F, Compañ-Rosique P, Satorre-Cuerda R, Molina-Carmona R. Improving the expressiveness of black-box models for predicting student performance. Comput Hum Behav. 2017;72:621-631.

60. Asif R, Hina S, Haque SI. Predicting student academic performance using data mining methods. Int J Comput Sci Netw Sec. 2017;17(5):187-191.

61. Roy S, Garg A. Predicting academic performance of student using classification techniques. Proceedings of the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON); 2017.

62. Chihbeh F, Boumahdi F, Bouarfa H, Boukraa D. Predicting students performance using decision trees: case of an Algerian university. Proceedings of the International Conference on Mathematics and Information Technology; 2017.

63. Xu M, Yu L, Wu W. Predicting honors student performance using RBFNN and PCA method. Proceedings of the International Conference on Database Systems for Advanced Applications. 2017.

64. Kostopoulos G, Liptakis AD, Kotsiantis S & Gravvanis G Predicting student performance in distance higher education using active learning. Proceedings of the International Conference on Engineering Applications of Neural Networks; 2017;75-86; Springer, Cham.

65. Rao BM, Murthy BR. Prediction of student’s educational performance using machine learning techniques. Data Engineering and Communication Technology. Springer; 2020.

66. Jalota C, Agrawal R. Analysis of educational data mining using classification. Proceedings of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon); 2019; IEEE.

67. Ma X, Yang Y, Zhou Z. Using machine learning algorithm to predict student pass rates in online education. Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing; 2018.

68. Khasanah AU, Harwati A. Comparative study to predict student’s performance using educational data mining techniques. IOP Conf Ser Mater Sci Eng. 2017;215:012036.

69. Costa EB, Fonseca B, Santana MA, De Araújo FF, Rego J. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Comput Hum Behav. 2017;73(August):247-256.

70. Ahmed, DM, Abdulazeex AM, Zeebaree DQ, Ahmed FY. Predicting University’s students performance based on machine learning techniques. Proceedings of the 2021 IEEE International Conference on Automatic Control & Intelligent Systems (I2CACSIS); 2021; IEEE.

71. Aggarwal D, Mittal S, Bali V. Significance of non-academic parameters for predicting student performance using ensemble learning techniques. Int J Syst Dyn Appl. 2021;10(3):38–49.

72. Rosyid HA, Pujianto U, Invarian BG. Performance comparison of Naïve Bayes and neural network in predicting student violation. Proceedings of the 2021 3rd East Indonesia Conference on Computer and Information Technology (EICoCIT); 2021; IEEE.

73. Tsiakmaki M, Kostopoulos G, Kotsiantis S, Ragos O. Transfer learning from deep neural networks for predicting student performance. Appl Sci. 2020;10(6):2145.

74. Chui KT, Liu RW, Zhao M, De Pablos PO. Predicting students' performance with school and family tutoring using generative adversarial network-based deep support vector machine. IEEE Access. 2020;8:86745-86752.

75. Alsalman YS, Halemah NK, AlNagi ES, Salameh W. Using decision tree and artificial neural network to predict students academic performance. Proceedings of the 2019 10th International Conference on Information and Communication Systems (ICICS); 2019.
76. Bhat M, Zaman M, Butt M. An intelligent prediction system for educational data mining based on ensemble and filtering approaches. *Proc Comput Sci.* 2020;2(167):1471-1483.

77. Hussain S, Atallah R, Kamsin A, Hazarika J. Classification, clustering and association rule mining in educational datasets using data mining tools: a case study. *Proceedings of the Computer Science on-Line Conference*; 2018.

78. Altaf S, Soomro W, Rawi MI. Student performance prediction using multi-layers artificial neural networks: a case study on educational data mining. *Proceedings of the 2019 3rd International Conference*; 2019.

79. Hasan M, Aly M. Get more from less: a hybrid machine learning framework for improving early predictions in STEM education. *Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence*; 2019.

80. Devasia T, Vinushree TP, Hegde V. Prediction of students performance using educational data mining. *Proceedings of the International Conference on Data Mining & Advanced Computing*; 2016.

81. Lehr S, Liu H, Kingslestone S, Konyha A, Robaszewska N, Medinilla J. Use educational data mining to predict undergraduate retention. *Proceedings of the IEEE International Conference on Advanced Learning Technologies*; 2016.

82. Almasri A, Alkhawaldeh RS, Elebi E. Clustering-based EMT model for predicting student performance. *Arabian J Sci Eng.* 2020;3:10067-10078.

83. M Tsai, G.K., S Kotsiantis, O Ragos. Predicting student performance and its influential factors using hybrid regression and multi-label classification. *IEEE Access.* 2020; 8: p. 203827–203844.

84. López-Zambrano J. Towards portability of models for predicting students' final performance in university courses starting from moodle logs. *Appl Sci.* 2020;10(1):354.

85. Qu S, Li K, Wu B, Zhang X, Zhu K. Predicting student performance and deficiency in mastering knowledge points in MOOCs using multi-task learning. *Entropy.* 2019;21(12):1216.

86. Aydodu E. Predicting student final performance using artificial neural networks in online learning environments. *Educ Inf Technol.* 2020;25(1):1-15.

87. Alharbi Z, Cornford J, Dolder L, De La Iglesia B. Using data mining techniques to predict students at risk of poor performance. *Proceedings of the 2016 SAI Computing Conference (SAI)*; 2016.

88. Sorour SE, Mine T. Building an interpretable model of predicting student performance using comment data mining. *Proceedings of the 2016 5th IIAI International Congress on Advanced Applied Informatics (IIIAI-AAI)*; 2016.

89. Adekiman AI, Noma-Asoghae E. Data mining approach to predicting the performance of first year student in a university using the admission requirements. *Educ Inf Technol.* 2019;24(2):1527-1543.

90. Canagareddy DS, Hurbungs K. A Machine Learning Model to Predict the Performance of University Students. *Department of Software and Information Systems Faculty of Information Communication and Digital Technologies University of Mauritius Moka Mauritius; 2019.*

91. Verhun V, Batyuk V, Voityslyn V. Learning analysis as a tool for predicting student performance. *Proceedings of the 2018 IEEE 13th International and Technical Conference on Computer Sciences and Information Technologies (CSIT); 2018.*

92. Dhar J, Jodder AK. An effective recommendation system to forecast the best educational program using machine learning classification algorithms. *Ingénierie Des Syst D Inf.* 2020;25(5):559-568.

93. Hassan H, Ahmad NB, Anuar S. Improved students’ performance prediction for multi-class imbalanced problems using hybrid and ensemble approach in educational data mining. *J Phys Conf Ser.* 2020;1529:052041.

94. Rosado JT, Payne AP, Reborg CB. eMineProve: educational data mining for predicting performance improvement using classification method. *IOP Conf Ser Mater Sci Eng.* 2019;649:012018.

95. Agaoglu M. Predicting instructor performance using data mining techniques in higher education. *IEEE Access.* 2016;4:2379-2387.

96. Cagliero L, Canale L, Farinetti L, Baralis E, Venuto E. Predicting student academic performance by means of associative classification. *Appl Sci.* 2021;11(4):1420.

97. Buenaño-Fernández D, Gil D, Luján-Mora S. Application of machine learning in predicting performance for computer engineering students: a case study. *Sustainability.* 2019;11(10):2833.

98. Wu B, Qu S, Ni Y, Zhou Y, Wang P, Li Q. Predicting student performance using weblogs. *Proceedings of the 2019 14th International Conference on Computer Science & Education (ICCSE); 2019.* IEEE.

99. García-González JD, Skrita A. Predicting academic performance based on Students’ family environment: evidence for Colombia using classification trees. *Psychol Soc Educ.* 2019;11(3):299-311.

100. Angiani G, Ferrari A, Fornacciari P, Mordonini M, Tomaiaulo M. Real marks analysis for predicting students' performance. *Proceedings of the International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning; 2019.* Springer.

101. Helal S, Li J, Liu L, et al. Predicting academic performance by considering student heterogeneity. *Knowl-Based Syst.* 2018;161(December 1):134-146.

102. Azizah EN, Pujianto U, Nugraha E. Comparative performance between C4.5 and naive Bayes classifiers in predicting student academic performance in a virtual learning environment. *Proceedings of the International Conference on Education and Technology; 2018.*

103. Wang GH, Zhang J, Fu GS. Predicting student behaviors and performance in online learning using decision tree. *Proceedings of the 2018 7th International Conference of Educational Innovation through Technology (EITT); 2018.* IEEE.

104. Sobnath D, Kaduk T, Rehman IU, Isiaq O. Feature selection for UK disabled students’ engagement post higher education: a machine learning approach for a predictive employment model. *IEEE Access.* 2020;8:159530–159541.

105. Sukhbaatar O, Ogata K, Usagawa T. Mining educational data to predict academic dropouts: a case study in blended learning course. *IEEE Tencon; 2018.*
106. Macqueen J. Some methods for classification and analysis of MultiVariate observations. Proceedings of Berkeley Symposium on Mathematical Statistics & Probability; 1965.

107. Settles B, Craven M, Friedland L. Active learning with real annotation costs; 2008.

108. Ginzburg I, Roustant B, Salagnac P. Interactive semi-supervised machine learning for classification; 2012.

109. Han J, Pei J, Kamber M. Data Mining : Concepts and Techniques. Elsevier, 2011.

110. de Melo Júnior G, Alcalá SG, Furriel GP, Vieira SL. Missing data analysis using machine learning methods to predict the performance of technical students. Revista Brasileira de Computação Aplicada. 2020;12(2):134-143.

111. Pochon B, Scotton P. Method and apparatus for data normalization. US, US7356599 B2. 2008.

112. Wang S, Yao X. Multiclass imbalance problems: analysis and potential solutions. IEEE Trans Syst Man Cybern B. 2012;42(4):1119-1130.

113. Lahiri SN. Resampling methods for dependent data: Resampling methods for dependent data; 2003.

114. Miao J, Kirz J, Sayre D. The oversampling phasing method. Acta Crystallogr. 2000;56(Pt 10):1312-1315.

115. Liu XY, Wu J, Zhou ZH. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B. 2009;39(2):539-550.

116. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE synthetic minority over-sampling technique; 2011.

117. Kubat M. Adressing the curse of imbalanced training sets: one-sided selection. Proceedings of the International Conference on Machine Learning; 1997.

118. Hall MA. Correlation-Based Feature Selection for Machine Learning. Phd thesis. Waikato University; 2000.

119. Yang J, Honavar V. Feature subset selection using a genetic algorithm. IEEE Intell Syst Appl. 2002;13(2):44-49.

120. Farissi A, Dahlan HM. Genetic algorithm based feature selection for predicting student's academic performance. International Conference of Reliable Information and Communication Technology. Springer; 2020.

121. Battula K. Research of machine learning algorithms using K-fold cross validation. Int J Eng Adv Technol. 2021;8(6S):215-218.

122. Witten IH, Frank E, Trigg LE, Hall MA, Holmes G, Cunningham SJ. Weka: practical machine learning tools and techniques with Java implementations. ACM SIGMOD Rec. 1999;31(1):76-77.

123. Robert M, Lynch R. Data mining use cases and business analytics applications. Comput Rev. 2014;55(5):278-279.

124. Saa AA, Al-Emran M, Shaalan K. Factors affecting students' performance in higher education: a systematic review of predictive data mining techniques. Technol Knowl Learn. 2019;24(4):567–598.

125. Tikk D, Koczty LT, Gedeon TD. A survey on universal approximation and its limits in soft computing techniques. Int J Approx Reason. 2003;33(2):185-202.

126. Zeiler M, Fergus R. Visualizing and understanding convolutional neural networks. Proceedings of the European Conference on Computer Vision; 2013.

127. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. Eprint Arxiv; 2014.

128. Hai-tao P, Ming-qi F, Hong-bin Z, et al. Predicting academic performance of students in Chinese-foreign cooperation in running schools with graph convolutional network. Neural Comput Appl. 2021;33(2):637-645.

129. Settles B. Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers; 2012.

130. Tsiakmaki M, Kostopoulos G, Kotsiantis S, Ragos O. Fuzzy-based active learning for predicting student academic performance using autoML: a step-wise approach. J Comput High Educ. 2021;33:1-33.

131. Alpaydin E. . Introduction to Machine Learning. MIT press; 2004.

132. Barab SA, Barnett M, Yamagata-Lynch L, Squire K, Keating T. Using activity theory to understand the systemic tensions characterizing a technology-rich introductory astronomy course. Mind Cult Act. 2002;9:76-107.

133. Xing W, Guo R, Petakovic E, Goggins S. Participation-based student final performance prediction model through interpretable genetic programming: integrating learning analytics, educational data mining and theory. Comput Hum Behav. 2015;47(June):168-181.

134. Deng J, Dong W, Socher R, Li JJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009.

**How to cite this article:** Xiao W, Ji P, Hu J. A survey on educational data mining methods used for predicting students’ performance. Engineering Reports. 2022;4(5):e12482. doi: 10.1002/eng2.12482