Technology, Knowledge and Learning
https://doi.org/10.1007/s10758-020-09476-0

ORIGINAL RESEARCH

Predicting Academic Outcomes: A Survey from 2007 Till 2018

Sarah Alturki[1] · **Ioana Hulpuș**[1] · **Heiner Stuckenschmidt**[1]

Accepted: 19 September 2020

Abstract
The tremendous growth of educational institutions' electronic data provides the opportunity to extract information that can be used to predict students' overall success, predict students' dropout rate, evaluate the performance of teachers and instructors, improve the learning material according to students' needs, and much more. This paper aims to review the latest trends in predicting students' performance in higher education. We provide a comprehensive background for understanding Educational Data Mining (EDM). We also explain the measures of determining academic success and highlight the strengths and weaknesses of the most common data mining (DM) tools and methods used nowadays. Moreover, we provide a rich literature review of the EDM work that has been published during the past 12 years (2007–2018) with focus on the prediction of academic performance in higher education. We analyze the most commonly used features and methods in predicting academic achievement, and highlight the benefits of the mostly used DM tools in EDM. The results of this paper could assist researchers and educational planners who are attempting to carry out EDM solutions in the domain of higher education as we highlight the type of features that the previous researches found to have significant impact on the prediction, as well as the benefits and drawbacks of the DM methods and tools used for predicting academic outcomes.

Keywords Prediction · Higher education · Educational data mining · Academic achievement

✉ Sarah Alturki
alturki@informatik.uni-mannheim.de

Ioana Hulpuș
ioana@informatik.uni-mannheim.de

Heiner Stuckenschmidt
heiner@informatik.uni-mannheim.de

[1] Data and Web Science Group, Faculty of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany

Published online: 28 September 2020

1 Introduction

Since higher education plays an essential role in the development of a society (Pinheiro et al. 2015), increasing student success is a long-term goal for academic institutions. In order to increase students' success rate, it is vital to understand and define academic success. The definition of academic success is rather complex and wide-ranging; therefore, it is frequently misused within educational research. However, the study of York et al. (2015) suggests a theoretically grounded definition of academic success that is made up of six components: (1) academic achievement, which is nearly entirely measured with course grades and grade point average (GPA), (2) satisfaction, which is often captured either by course evaluation or institutional surveys, (3) persistence, which is measured by retention between particular years of college and degree attainment rates, (4) acquisition of skills and competencies, which can be measured by assignments and course evaluations, (5) attainment of learning objectives, which can also be measured by assignments and course evaluations, and finally (6) career success, which can be determined by job attainment rates, promotion histories, career satisfaction and professional goal attainment.

A second crucial requirement for maximizing students' success is the identification of the factors that have an effect over academic performance. Awareness of students' success factors could assist in achieving the highest level of quality education (Yassein et al. 2017). It can potentially help in providing a clear and strong description of the types of knowledge and behaviour that are associated with adequate performance. Such awareness can be gained by using methods of data mining (DM) over educational records. The practice of DM methods applied to educational data is known as Educational Data Mining (EDM) (Baker and Yacef 2009). It is drawn from a variety of domains, including DM and machine learning, psychometrics and other areas of statistics, information visualization, and computational modelling (Romero and Ventura 2007). Generally, EDM refers to techniques and tools designed for automatically extracting useful information and patterns from huge data repositories related to people learning activities in an educational environment (Nithya et al. 2016). Those tools employ machine learning algorithms, database systems, statistical analysis, and artificial intelligence. The DM techniques include regression, clustering, classification, association, and prediction.

Figure 1 by Romero and Ventura (2007) shows the application of DM in educational systems, and we use it in order to position the approaches analyzed in this survey within the EDM landscape. The use of DM in educational systems is represented as an iterative cycle of hypothesis formation, testing, and refinement where the systems can be directed to support the specific needs of every participant in the educational process. EDM carries an array of DM techniques, to (1) support relationship analysis, classification, and clustering, (2) elaborate educational hypotheses, and (3) provide learning assistant (Baker and Yacef 2009).

The scope of this paper lies within the DM phase in the EDM cycle. There is an extensive variety of methods within DM. However, as this survey covers predictions in EDM, we address only supervised DM methods (also known as predictive or directive). Supervised techniques such as classification and regression predict the value of the output variables based on the inputs. To do this, a model is developed from training data where the values of input and output are previously labelled. The model generalizes the relationship between the inputs and outputs and uses it to predict other datasets where only inputs are known (Witten et al. 2017).

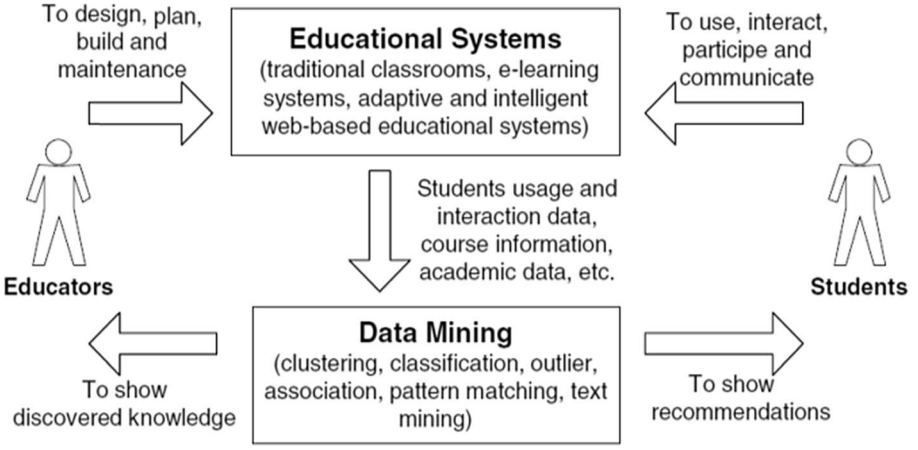

Fig. 1 The cycle of applying data mining in educational systems (Romero and Ventura 2007)

After having introduced the EDM concept and the supervised techniques, the rest of the paper is organized as follows: We start by giving an overview of existing EDM surveys, along with the research question and research methododogy. Afterwards, we provide an outline of the task of prediction in EDM. This task involves essential decision making regarding the DM tool to be used in performing the predictions, the prediction methods and techniques, and the selection of features (or predictors). We then continue with a rich literature review regarding predicting students' academic achievement in the past 12 years. Lastly, we present a summary of the main results and the reached conclusions, as well as future lines of research.

2 Overview

The EDM literature is rapidly growing and needs to be brought up-to-date frequently to take new studies into account. Romero and Ventura (2007) conducted a literature survey on EDM covering published work between 1995 and 2005. They reviewed different types of educational systems (traditional classrooms and distance learning) and how EDM can be applied on them. They also described the DM techniques that have been applied in educational systems. Another EDM survey was conducted by Peña-Ayala (2014) and analyzes EDM studies published between 2010 and 2013. It provides an analysis of the EDM weakness, strengths, threats, and opportunities. Dutt et al. (2017) performed a systematic literature review covering research between 1983 and 2016 on clustering algorithms only, which are unsupervised techniques. Unsupervised techniques uncover hidden patterns in unlabeled data with the aim of finding patterns in a dataset. In unsupervised DM, there are no output variables to predict. In their study, they viewed the applicability and usability of clustering algorithms in the context of EDM. They concluded that the key benefit of the clustering algorithms is that it provides a fairly explicit schema of students' learning style when using a number of attributes (e.g., time spent to complete tasks, students' behavior in class, and students' motivation towards learning) to cluster. Kumar et al. (2017) conducted a literature survey covering publications between 2007 and 2016 on students' performance prediction. They report the used DM methods and their accuracies. However, they do not

report the used DM tools. They found that GPA and students' internal marks are important attributes for predicting academic achievement. Our literature survey covers different and more recently published papers (2007–2018) and focuses on prediction tasks in EDM using supervised methods. In this section of the paper, we outline our research questions and the research methodology used to collect relevant literature.

2.1 Research Questions

The research questions proposed in this literature survey are as follows:

- What are the measurable aspects of the prediction of student academic achievement in higher education?
- Which are the best DM methods to predict students' academic achievement in higher education?

2.2 Research Methodology

To answer our research questions, we followed a quantitative approach by collecting information from 22 individual published studies regarding predictions performed in higher education institutes and universities. We displayed the collected information using tables to allow random access and to simplify comparison between the different data.

2.2.1 Search Strategy

Three databases were used to search and filter out the papers that were relevant to our investigation. They are as follows: SpingerLink,[1] ScienceDirect,[2] and ACM Digital Library.[3] We searched different types of publications, including Journal articles as well as conference and workshop proceedings. Our search strings were written using Boolean operators like AND, and OR to further produce more relevant results, e.g., (student success) AND (factors OR features OR attributes OR characteristics OR aspects) AND (educational data mining) AND (prediction OR estimation). We have also hand-searched journals to identify papers that might have not yet been included in electronic databases, or have not been indexed. Hand-searching involves the page-by-page inspection of relevant conference proceedings or journal issues to find studies relevant to our research.

2.2.2 Inclusion Criteria

The inclusion criteria to determine relevant literature are as follows:

- Studies that have been conducted between 2007 and 2018.
- Studies that reported the used data mining method for performing the prediction.
- Studies that reported the features used for performing the prediction.

[1] https://link.springer.com/.
[2] https://www.sciencedirect.com/.
[3] https://www.dl.acm.org/.

2.2.3 Exclusion Criteria

The following principles were used to exclude the literature that was not relevant for this research:

- Studies that focused on unsupervised analysis.
- Studies that did not include analysis or prediction of academic success.
- Studies which are not performed on a higher education level, e.g., elementary or secondary school.

3 Overview of Features, Algorithms and Software Used for Prediction in EDM

Being able to predict a student's academic success serves as an essential research topic in many academic disciplines due to its benefits to both teaching and learning activities (Holland and Nichols 1964). Performing predictions in educational environments generally includes three steps: (1) selecting the right features (predictors), (2) selecting the right method or technique, and (3) selecting the proper data mining software. As is the case with most supervised machine learning tasks, these choices are essential for the reliability of the findings, for reducing the likelihood of errors, and ultimately for the overall performance of the solution. In the following, we give a short account of the features, methods and tools that we have frequently encountered in our survey, in the DM phase of the EDM process.

3.1 Overview of Features Used in EDM

Many features have been researched with the scope of determining their ability to predict students' academic performance. Based on the studies that have been reviewed in our research investigation, these features can be classified into three general categories:

- *Demographical features*, which as the title suggests, include gender, age, marital status, background, income, occupation, mobility (transportation), disability, parents' education level, etc.
- *Pre-enrollment features*, which are related to students' achievements before their enrollment such as their GPA from previous studies, previous major (field of study), previous institute, language proficiency, grades earned in prerequisite courses, as well as pre-enrollment exams, e.g., Scholastic Aptitude Test (SAT) and Graduate Record Examinations (GRE), etc.
- *Post-enrollment features*, which are related to students' attitude after enrollment and during the course such as attendance, assignments, scores earned in quizzes and final exams, lab work, writing notes in class, boredom level during lectures, the number of credit hours per semester, etc.

We provide details on the commonality as well as performance of these features as revealed by our literature survey in Sect. 4.1.

3.2 Overview of Data Mining Methods Used in EDM

In EDM, multiple prediction methods have been researched. Since there is no definite answer to the question of which is the best DM method as every method has its advantages and limitations, most researchers often explore two or more techniques to reveal which method generates the best accuracy in their specific case and adopt it. Table 1 summarizes the advantages and weaknesses of the most common DM methods used today in predicting students' academic achievement. We provide details on the usage of these algorithms in EDM in Sect. 4.2.

3.3 Overview of Data Mining Software Used in EDM

In recent years, a number of tools have been developed with the purpose of conducting DM research. According to Slater et al. (2017), the 7 tools represented in Table 2 are the ones that offer algorithms that can be used to model and predict processes and relationships in educational data. They are all well documented and they are all cross-platform applications as they may run on Microsoft Windows, Linux, and mac OS. We get back to analyzing the use of these tools in the context of EDM, in Sect. 4.3.

4 Comprehensive Review of Academic Achievement Prediction Literature

In this section, we survey the different types of predictions that have been performed in higher education institutions. We summarize 22 recent studies regarding the prediction of academic achievement in higher education. These studies have been conducted in different countries around the world between the year 2007 and 2018. We also demonstrate the most significant findings out of the literature study by discussing the outcomes of the previous works.

4.1 Features Used in Predicting Students' Academic Achievement

Figure 2 shows the prevalence of the most commonly used features for predicting academic achievement in higher education, as encountered in the studied literature. As can be seen, gender and the GPA are used in more than half of the studies: 14 (63%) and 12 (54%) respectively. Their frequencies are followed by those of age (40%) and language proficiency (31%). The other features such as income, nationality, marital status, employment status and attendance are each used in less than 30% of the publications.

In the following, we analyze these features in more detail.

4.1.1 Demographics

4.1.1.1 Gender As Fig. 2 illustrates, it can be concluded that gender has been used the most compared to other demographics in predicting academic achievement. This should come as no surprise since the relationship between gender and academic achievement of students has been discussed for decades (Eitle 2005), resulting in a substantial body of literature.

Table 1 Pros and cons of the most common data mining methods

DM method	Definition	Pros.	Cons.
Decision trees (Clark 2013; Kaushal and Shukla 2014; Yu-Wei and David. 2015)	A classification method in which each internal node serves as a "test" on a feature, each branch serve as the outcome of the test, and each leaf node correspond as a class label (decision taken after computing all features)	Easy to understand Can handle missing values	Could suffer from overfitting Less accuracy with continuous variables
Support vector machine (SVM) (Clark 2013; Harrington 2011; Tomar and Agarwal 2013)	A supervised learning model with associated learning algorithms which analyze data being used for classification and regression analysis by applying the statistics of support vectors to categorize unlabeled data	High accuracy Can handle different data types Effective in high dimensional space	Black box Training processes can take time High algorithmic complexity
Naïve Bayes (Harrington 2011; Yu-Wei and David 2015)	A classification method that assumes that the prognostic features are conditionally independent and that there are no hidden features that could affect the process of prediction	Simple to use Can deal with missing and noisy data	Black box Assumes that all features are independent and equally important
Neural Networks (Clark 2013; Kaushal and Shukla 2014; Tomar and Agarwal 2013)	A method that learns to perform tasks by considering examples, generally without being programmed with any task-specific rules	High accuracy Can handle missing and noisy data	Black box Difficult in dealing with big data High complexity
Logistic regression (Geng 2006; Yu-Wei and David 2015)	A statistical model that is often taken to apply to a binary dependent variable. More formally, a logistic model is one where the log-odds of the probability of an event is a linear combination of independent or predictor variables	Easy to understand Provides probability outcome	Does not handle missing values well Could suffer from over-fitting

Table 1 (continued)

DM method	Definition	Pros.	Cons.
K-nearest neighbour (Clark 2013; Yu-Wei and David 2015)	A classification and regression method that stores available cases and classifies new cases based on a similarity measure. A case is classified by a majority vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function	Nonparametric Easy to understand the output Robustness to noisy training data	Black box Difficult in handling mixed data type Assumes that all features are equally important Sensitive to outliers
Rule induction (An et al. 1997; Domingos 1995)	An area of machine learning where formal rules (If–then) are extracted based on statistical significance from a set of data observations	Low computational-space cost Can make effective use of statistical measures to combat noise	Slow training time Has trouble recognizing exceptions or small, low-frequency sections of the space

"Black box" indicates that the results of the method are not easy to explain and therefore do not support straightforward human interpretability

Table 2 Pros and cons of the different data mining tools and packages

DM tool and source	Pro-graming language	Pros.	Cons.
RapidMiner[1] (commercial)	Java	Supports command line and Graphical User Interface (GUI) Displays visualizations Performs cross-validation at multiple levels Provides wide range of metrics for model assessments	Limited functionality for engineering new features out of existing features
SPSS[2] (Commercial)	Java	Supports command line and GUI Easily visualize the process Functionality for creating new features out of existing ones	Minimum support for modelling Less flexible than other tools Difficult to customize Slow in handling large data sets
WEKA[3] Open source	Java	Supports command line and GUI Displays visualizations	Does not support the creation of new features
KNIME[4] (Open source)	Java	Supports GUI Displays visualizations Has the ability to integrate data from multiple sources Has extensions allowing interface with R, Python, Java, and SQL	Does not support interactive execution Not all nodes can be streamed
Orange[5] (Open source)	Python Cython C++ C	Supports command line and GUI Customizable visualizations Easy to understand interface	Limited in the scale of data that it can work with, comparable to Excel Less suitable for big projects
Spark MLLib[6] (Open source)	Scale SQL Java R Python	Displays visualizations Can connect with several programming languages through API	Purely programmatic tool (less usability for non-programmers)

Table 2 (continued)

DM tool and source	Pro-graming language	Pros.	Cons.
KEEL[7] (Open source)	Java	Supports command line and GUI Displays visualizations Supports discretization algorithms Feature selection support with a broad range of algorithms Extensive support for missing data	Limited functionality for engineering new features out of existing ones Limited support for clustering and factor analysis Limited support for association rule mining

[1] rapid-i.com/content/view/181/190/

[2] https://www.ibm.com/analytics/dk/da/technology/spss/

[3] https://www.cs.waikato.ac.nz/ml/weka/

[4] https://www.knime.com

[5] https://www.orange.biolab.si

[6] https://www.spark.apache.org/mllib/

[7] https://www.sci2s.ugr.es/keel/

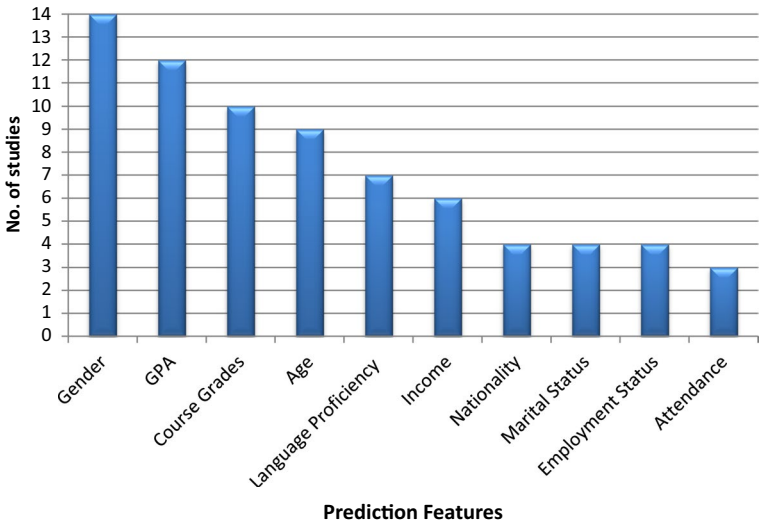

Fig. 2 Mostly used features in predicting students' academic achievement between 2007 and 2018

In the literature that this study targets, while some of the researchers found no significant gender difference between students (Goni et al. 2015), others found a significant difference with either the male (Chang 2008) or the female (Simsek and Balaban 2010) performing better, based on the specific subject. Unfortunately, none of these aforementioned studies also attempted the prediction task. Regarding the studies that actually used gender for predicting academic outcomes, we identify 14 studies. However, only 2 studies (Kovačić 2010; Osmanbegović et al. 2012) reported its impact on the overall prediction. Both concluded that gender does not have a significant impact on this task.

4.1.1.2 Age The second most frequently used demographical feature for predicting academic success is age. A potential explanation to the prevalence of this feature is the fact that many researchers in the past found a positive relationship between age and performance (Sturman 2003; Watkins and Hattie 1985). These previous studies try to explain the positive correlation between age and academic achievement by suspecting that older students were more highly motivated, more experienced, and that they possess effective study habits. Regarding the studies that we target, unfortunately most of the studies do not report the individual impact of this feature. Exception is a study by Kovačić (2010) who found that age actually does not have a significant impact on predicting academic success.

4.1.1.3 Marital Status The relationship between marital status and the academic achievement of students has also been widely discussed in the literature, specifically in 18% of the studies we surveyed. Yess (2009) investigated the influence of marital status on the scholastic achievement of 240 Community Colleges students in the US. The result revealed that it was a significant predictor of achievement. Another study by Ma and Wooster (2009) investigated how the marital status of College students can affect their academic performance using a study sample of 374 students. Their investigation revealed that married students had higher grades than unmarried students.

4.1.1.4 Other Demographic Features There are also other demographics, such as income, that have been used often as predictors (Daud et al. 2017; Nghe et al. 2007; Pal and Pal 2013; Ali et al. 2013; Villwock et al. 2015; Yadav and Pal 2012). Among them, Ali et al. (2013) examined the individual features affecting academic performance of graduate students, including student's social economic status. Using a sample of 100 randomly selected students, they found that income significantly contributes to students' success. Students' employment status has also been used several times as a predictor of academic achievement (Daud et al. 2017; Kovačić 2010; Nghe et al. 2007, Mohamadian et al. 2015). Among them, the study of Mohamadian et al. (2015) investigates the relationship between employment status and academic achievement using a sample of 235 students. Their results showed that unemployed students had significantly higher academic achievement than employed students. They believe that working students devote less time to study and as a consequence, achieve less success.

We conclude that although the demographical features are heavily used, the extent to which they are useful in the academic achievement prediction task is not yet clear, with multiple studies either not reporting the individual contribution of the features, or with studies reaching opposing conclusions, particularly with respect to gender and age. Since previous researchers claim that gender, age, marital status, etc. have an effect on students' success then, the latest EDM research tends to use them as features for predicting academic success, but yet with unclear success. In our analysis, it has become apparent that the use of demographical features, as well as their choice might be strongly influenced by the cultural background of the countries where the study is held. For instance, when the study is performed in a collectivistic country (e.g., India, and Malaysia), we witness features related to the family of the student, such as family support (Sembiring et al. 2011), family income (Yadav and Pal 2012; Pal and Pal 2013; Villwock et al. 2015; Abu Saa 2016; Daud et al. 2017), family size (Yadav and Pal 2012), and parent's qualifications (Abu Saa 2016). This was not the case in studies performed in individualistic countries. (e.g., United States and Europe). While individualistic cultures tend to focus on personal achievement, collectivist cultures prioritize family and team goals over individual requirements (Kim 1995). This might mean that students from individualistic cultures could be more competitive than those from collectivistic cultures. We therefore observe that this finding certainly deserves further investigation in order to improve our understanding of the cultural impact over academic performance.

4.1.2 Pre-enrollment Features

4.1.2.1 GPA With respect to using students' previous qualifications for predicting academic achievement, the mostly used feature is GPA (Nghe et al. 2007; Kovačić 2010; Osmanbegović et al. 2012; Huang and Fang 2013; Pal and Pal 2013; Kabakchieva 2013; Abu Saa 2016). In fact, Ibrahim and Rusli (2007) found that GPA is the most significant feature (with an 87% correlation) to predict students' success compared to some demographical and pre-enrolment features.

4.1.2.2 Academic Language Skills Also, academic language skills have been used frequently in the list of features to predict student achievement (Nghe et al. 2007; Abu Saa 2016; Badr et al. 2016; Asif et al. 2017). Academic Language is the language being used in textbooks, spoken in classrooms, and presented on tests and examinations. While some

researchers (Arsad et al. 2014;) found out that the academic language skills do not affect students' success in "knowledge courses" or "non-linguistic courses" other researchers (Wait and Gressel 2009) found a significant relationship and concluded that students who are proficient in the teaching language will be much better equipped to acquire new knowledge through reading and listening, and will also be better in expressing their ideas through oral discussions and oral exams. As to the significance of using language proficiency in prediction, most of the viewed literature did not report the significance of using academic language skills as a predictor. However, Badr et al. (2016) reported that they acquired better accuracy (67.33%) when their prediction model depended on only language skills and no other feature.

To conclude, we believe that using pre-enrollment features to predict students' academic achievement is significant, especially if the prediction is to be performed at a very early stage (i.e., before the start of the program) as there are no other measurable features available at that point of time. As per choosing the predictor features, although the literature evaluating the impact of the individual features is scarce, the few studies that do it, do agree that both GPA and academic language skills have a positive impact on the prediction.

4.1.3 Post-enrollment Features

4.1.3.1 Grades When it comes to using students' post-enrollment features in predicting academic achievement, the grades earned in quizzes and examinations have been mostly used (Al luhaybi et al. 2018; Aulck et al. 2017; Badr et al. 2016; Huang and Fang 2013; Kemper 2018; Pradeep and Thomas 2015; Shakeel and Anwer Butt 2015; Villwock et al. 2015; Yadav et al. 2011; Yassein et al. 2017). In the study of Huang and Fang (2013), the earned grade in a mid-term exam was found to be the most important feature affecting prediction accuracy.

4.1.3.2 Results in Previous Semester The success rate of the previous semester, which was usually measured by GPA, has also been used often (Nghe et al. 2007; Kabakchieva 2013; Alemu Yehuala 2015; Abu Saa 2016; Asif et al. 2017; Al luhaybi et al. 2018; Kemper 2018) in the studies we have reviewed. That is due to the fact that students' success is highly dependent on previously acquired knowledge or skills. Asif et al. (2017) found that the marks of the first and second year courses of a four-year program play a role in predicting the graduation performance in a program. Likewise, Al luhaybi et al. (2018) found that the results of the core modules of the first year of the academic program have a high impact on the prediction of the high risk of failure students.

4.1.3.3 Attendance A number of studies have used attendance in predicting academic achievement (Al luhaybi et al. 2018; Pradeep and Thomas 2015; Yadav et al. 2011; Yassein et al. 2017) as increased attendance could be seen as a direct indicator of students' success. Lukkarinen et al. (2016) investigated the relationship between university students' class attendance and learning performance by using data from a course in a university in which attendance to classes was not mandatory. They found that attendance is positively and significantly related to students' performance. Another study by Alija (2013) used binary logistic regression to study the relationship between attendance and students' achievement. They found that students who regularly attend the lectures have more chances to receive passing grades.

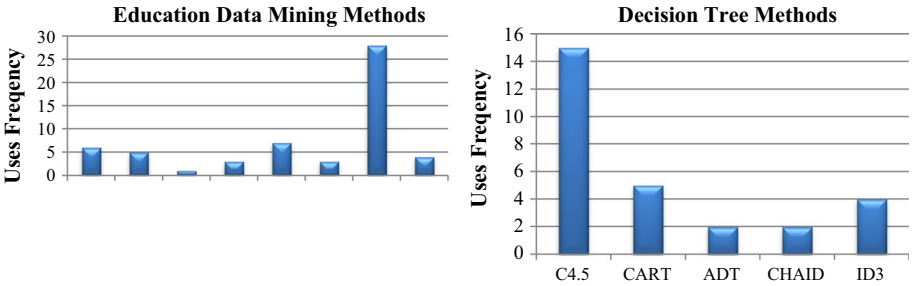

Fig. 3 Mostly used data mining methods in educational predictions between 2007 and 2018

4.1.3.4 Other Post-enrollment Features Balancing the academic load is vital to academic achievement. It is measured in terms of credit hours and course difficulty (Szafran and Austin 2002). Therefore, the choice of registered courses (Alemu Yehuala 2015; Aulck et al. 2017) and the total credit hours per semester (Alemu Yehuala 2015; Abu Saa 2016) have been used as predictors for academic success. In fact, Alemu Yehuala (2015) found that that the number of credit hours is one of the main significant attributes for predicting academic achievement.

We conclude that using post-enrollment features for predicting students' academic achievement can contribute to maximizing the accuracy of the prediction. This is due to the fact that such features represent students' current situation in the program rather than depending on their previous condition only.

4.2 Mostly Used Data Mining Methods in Predicting Students' Achievement

By observing the viewed studies, it can be noticed that most researches explored several methods to predict students' success and did not rely on the results of just one method. They often compared the results of each method to determine the best-fit technique for the specific dataset and consequently ensure the highest accuracy rates when deploying the model.

As seen in Fig. 3, the mostly used DM methods in the covered studies are decision trees. Due to their usability and efficiency, they have become one of the most effective and popular methods in machine learning since their introduction in the 1960 s (Song and Lu 2015). Knowledge models under this paradigm can be directly transformed into a set of IF–THEN rules. CHAID, CART, C4.5, and ID3 (Jain et al. 2017) are all decision tree algorithms. However, C4.5 (J48 in Weka) appears to be more popular than the rest of the decision tree algorithms. It has been used in 15 studies leading to a range of accuracies between 0.364 and 0.945. It was assessed to be the best scoring method in five studies (Alemu Yehuala 2015; Kabakchieva 2013; Nghe et al. 2007; Yadav and Pal 2012) and second best scoring method in two studies (Osmanbegović et al. 2012; Abu Saa 2016). CART has also been used in 5 of the reviewed studies leading to a range of accuracies between 0.40 and 0.622. It was found to be the best scoring method in three cases (Kovačić 2010; Yadav et al. 2011; Abu Saa 2016). ID3 was applied in 4 studies and was assessed to be the best method in the study of (Pal and Pal 2013) with 0.78 accuracy, and the worst method in the study of (Abu Saa 2016) with 0.333

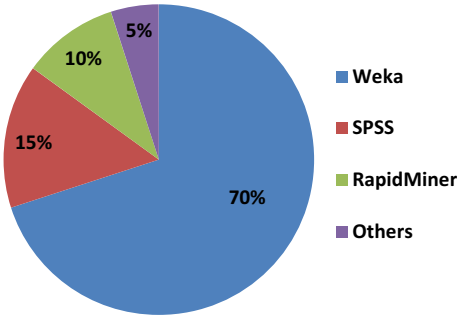

Fig. 4 Mostly used data mining tools between 2007 and 2018

accuracy. ADT was used in 2 studies only. In the first study (Pal and Pal 2013), it produced 0.6950 accuracy, while in the second (Pradeep and Thomas 2015), it obtained an accuracy of 0.995 and was assessed as the best scoring method. CHAID was also used in two studies only. It produced an accuracy of 0.594 in the first study (Kovačić 2010) and an accuracy of 0.341 in the second (Abu Saa 2016).

Rule-based classifiers such as JRip, NNge, OneR, and Ridor (Lakshmi 2012) have also been used several times by researchers. They often produced good results as they gave an accuracy of 0.545 (Kabakchieva 2013) in its worst cases and 0.982 (Pradeep and Thomas 2015) in its best. Also, Naïve Bayes produced outstanding results in most cases with accuracy above 0.75.

Even though sophisticated techniques, like neural networks or support vector machines, may outperform logistic regression and decision trees regarding prediction accuracy (HoYu et al. 2010), they are deemed to be less suitable for EDM purposes. Knowledge models obtained under these paradigms are considered to be black-box mechanisms, i.e., they can achieve high accuracy rates but can be difficult for people to comprehend.

4.3 Mostly Used Data Mining Tools in Predicting Students' Achievement

Based on the studies we have viewed in this paper, the open-source Weka tool appears to be the most widely used DM tool for predicting academic results (see Fig. 4). It is intended for machine learning and DM and was developed at the University of Waikato in New Zealand. Weka supports several standard DM tasks like data clustering, classification, regression, pre-processing, visualization and feature selection. Weka has become popular with academic researchers in recent years due to its highly active community.

SPSS and Rapid miner tools have also been used by the EDM researchers quite often in comparison to the rest of the DM tools. The advantage of the IBM SPSS tool is that it offers the user much control and enables to develop the predictive models quickly using business expertise (Brahmeswara Kadaru and Umamaheswararao 2017). Likewise, RapidMiner, formerly called as "Yale", has many benefits including the multiple deployment options based on the user's preferences.

After overviewing the features, methods and tools used in the target literature, we now analyze in detail the same literature, by shifting the focus to the sub-tasks that comprise the academic achievement task.

4.4 Per-Task Analysis

From the viewed literature, the prediction of students' performance in higher education can be broadly classified into three areas based on the prediction of (1) academic performance or GPA at a degree level, (2) failure or drop out of a degree, and (3) academic performance at a course level. In this section of the paper, we present the viewed literature using bullet points and tables. The bullet points show only the studies which have reported the significance of certain features on the prediction, whereas the tables show all the viewed studies, including the prediction task, where the study was held, the features used for the prediction, the DM tool, the DM method, and the accuracy of the prediction.

4.4.1 Prediction of Students' Academic Performance or GPA at a Degree Level

One of the most known standards for assessing the quality of universities is based on the excellence records of their students' academic outcome. A primary application area of prediction in EDM is predicting students' GPA or overall academic performance, e.g., excellent, very good, good, etc. This type of prediction is useful in different contexts in universities, like for instance, identifying excellent students for allocating scholarships. Following are the studies that have reported the impact of the success features on the prediction of students' academic performance or GPA at a degree level:

- Sembiring et al. (2011) sampled 300 students to predict the final grade of students from the faculty of computer systems and software engineering. They used innovative features that were not visible in the rest of the studies. The significance of each feature was tested using multi-variant analysis methods. They found that family support had the most impact (52.6% contribution) on the prediction, followed by engaging time, then study behaviour, and finally study interest. On the other hand, students' personal beliefs did not have any impact.
- Kabakchieva (2013) used a dataset of 10,330 students to predict their performance using 5 classes (Bad, average, good, very good and excellent). They found out that the classifiers perform differently for the five classes. Another finding is that the post-enrollment features related to students' university admission score, and numbers of failures at the first year exams are among the most influencing features in the classification.
- Abu Saa (2016) collected data from 270 students using a survey distributed in daily classes and online with the aim of predicting students' performance in an IT Department. They found that the students' performance is not totally dependent on post-enrolment features, such as their academic efforts, but that on the contrary, there are many other features that have equal to more significant influences as well. This includes demographical features, such as gender, and mother occupation, as well as pre-enrolment features, such as high school grade, and University fees discount.
- Asif et al. (2017) predicted students' performance using a sample of 210 undergraduate students. The features they used to perform the prediction are marks only. The results of their study showed that it is possible to predict the graduation performance in a four-year university program using only pre-university marks and marks of first and second-year courses with a reasonable accuracy.

Table 3 Prediction of students' academic performance or GPA at a degree level

Author (s)	Prediction	Predictor features	DM method (s)	DM tool	Result
Nghe et al. (2007)	Predict students' GPA at the end of the first year of their Master program using three models at the Asian Institute of Technology (AIT), Thailand	Demographics: gender, marital status, income, and age Pre-enrolment features: academic institute, previous GPA, English proficiency, and TOEFL score	Decision-trees (J48) Bayesian-tree	Weka	J48 produced better accuracy (91.98%) for 2 classes (pass/fail), (67.74%) for 3 classes (Fail/Good/Very Good) and (63.25%) for 4 classes (Fail/Fair/Good/Very Good)
Nghe et al. (2007)	Predict students' GPA at the end of the third year using three models in Can Tho University (CTU), Vietnam	Demographics: gender, age, family, job, and religion Pre-enrolment features: English skill, entry marks range, the field of study, and faculty Post-enrollment features: second-year GPA	Decision-trees (J48) Bayesian-tree	Weka	J48 produced better accuracy (92.86%) for 2 classes (pass/fail), (84.18%) for 3 classes (Fail/Good/Very Good) and (66.69%) for 4 classes (Fail/Fair/Good/Very Good)
Yadav et al. (2011)	Predict computer master students' performance at VBS Purvanchal University in Jaunpur, India	Post-enrollment features: attendance, test grade, seminar grade, assignment grade, and lab work	Decision trees (ID3, CART, and C4.5)	Weka	CART produced the best accuracy (56.25%) followed by ID3 (52.08%) then C4.5 (45.83%)
Sembiring et al. (2011)	Predict final grade of students from the faculty of computer systems and software engineering at the University of Malaysia Pahang (UMP) in Malaysia	Demographics: personal beliefs, and family support Post-enrollment features: interest, study behaviour, and engaging time	Support Vector Machine (SVM)	Rapid-Miner	SVM produced high accuracy (83%)

Table 3 (continued)

Author (s)	Prediction	Predictor features	DM method (s)	DM tool	Result
Yadav and Pal (2012)	Predict Engineering student academic performance at VBS Purvanchal University in Jaunpur, India	Demographics: gender, the medium of instruction, location, accommodation type, parents' qualification, parents' occupation, annual family income and etc. Pre-enrolment features: previous grades, and admission type	Decision trees (ID3, CART, and C4.5.)	Weka	C4.5 produced the best accuracy (67.77%) followed by ID3 and CART with the same accuracy (62.22%)
Pal and Pal (2013)	Predict student performance at VBS Purvanchal University in Jaunpur, India	Demographics: gender, medium of instruction, college location, student accommodation, family size, family income, parents' qualification, parents' occupation, and category Pre-enrolment features: High School grade, senior secondary grade, admission type, and BCA result	Decision tree (ID3, ADT) Bagging	Weka	ID3 produced the best accuracy (78%) followed by bagging (73%) then ADT (69.50%)

Table 3 (continued)

Author (s)	Prediction	Predictor features	DM method (s)	DM tool	Result
Kabakchieva (2013)	Predict students' performance (Bad, average, good, very good and excellent) at the University of National and World Economy (UNWE) in Bulgaria	Demographics: gender, age, student speciality, and current semester Pre-enrolment features: place, and profile of the secondary school, secondary school GPA, admission exam score, and admission year Post-enrolment features: GPA achieved during the first year of university	Decision tree (J48) K-nearest neighbour Bayesian-rule induction (JRip, OneR)	Weka	J48 produced the best accuracy (66.5%) followed by JRip (63%) then KNN (60%) and Bayesian (≈60%) then finally OneR (54.5%)
Abu Saa (2016)	Predict students' performance in the IT Department at Ajman University of Science and Technology in Ajman, United Arab Emirates	Demographics: gender, nationality, first language, teaching language, living location, sponsorship, parent working in the university, student discounts, transportation method, family size, family income, parent's marital status, parent's qualifications, parent's occupation, and number of friends Pre-enrolment features: high school percentage Post-enrollment features: previous semester GPA, number of credit hours, the average number of hours spent with friends per week	Decision trees (CART, C4.5., ID3, CHAID) Naïve Bayes	Rapid-Miner and Weka	CART produced best accuracy (40%) followed by C4.5 (36.40%) then Naïve Bayes (35.19%) then CHAID (34.07%) then finally ID3 (33.33%)

Table 3 (continued)

Author (s)	Prediction	Predictor features	DM method (s)	DM tool	Result
Asif et al. (2017)	Predict students' performance using 2 classes (low/high) at the end of the third year of their IT bachelor degree in Pakistan	Pre-enrolment features: total marks in High School Certificate, marks in mathematics, and sum of the marks in mathematics, physics and chemistry in an entrance examination Post-enrollment features: examination marks of all the courses taught in different academic years	Decision trees Random Forest Rule induction Naïve Bayes Neural NW K-nearest neighbour	Rapid-Miner	Naive Bayes produced best accuracy (83.65%) followed by 1-nearest neighbor (74.04%) then random forests (71.15%) then decision tree (69.23%) then neural NW (62.50%) then finally rule induction (55.77%)
Yassein et al. (2017)	Predict students' academic Performance in Saudi Arabia	Post-enrollment features: assignments, attendance, lab work attendance, final exam mark, mid-exam grades, education type, and success rate	Decision tree (C4.5) Two-step clusterin	SPSS & clementine	n/a

Table 3 provides a summary of the studies we analyzed, with respect to the used features, tools, methods and accuracy.

4.4.2 Prediction of Students' Failure or Drop Out of a Degree

Student failure or dropout is a significant concern in the education and policy-making communities (Demetriou and Schmitz-Sciborski 2011). High dropout rates and poor academic performance among students are examples of the most common issues that affect the reputation of an educational institution. The negative consequences of dropping out of the educational system are considerable, both for the individuals as well as the teaching institutions. Therefore, preventing educational dropout poses a significant challenge to institutions of higher education. This prevention can be made by predicting students at risk at an early stage. Following are the studies that have been performed to predict students' failure or drop out of a degree:

- Pradeep and Thomas (2015) Predicted bachelor student dropout using the records of 150 students who have been enrolled in a Technology program. Interestingly, the number of used features was reduced from 67 features to the best 13 using Attribute Selection Algorithms provided in WEKA tool. The selected features were mostly post-enrollment features such as attendance, taking notes from class, and some courses scores. Features such as age, gender and religion were neglected as they did not have an effect on the overall prediction.
- Alemu Yehuala (2015) used 11,873 student records to predict university students who are at risk of failure. They found out that the 6 main features determining the failure or success of students are: number of students in a class, number of courses given in a semester, higher education, entrance certificate, examination result of a student, and gender.
- Villwock et al. (2015) investigated the factors that may influence the students' decision to drop out from a Mathematics Major. It was possible to observe that the courses that contributed to dropouts in the Major differ in different years. Considering only the subjects taken in the first year, the course that most contributed to dropouts was "Differential and Integral Calculus I", and considering the first 2 years, it was "Finite Mathematics". It was also concluded that the work factor is the feature that most contributed to the decision of dropping out. They believe that this is due to the fact that the working student has little time to devote to extracurricular study. They also found that marital status and age contributed to the decision of dropping out as well.
- Daud et al. (2017) used 776 student instances to predict the completion or dropout of students from multiple universities in Pakistan. 23 features (selected by the feature extraction process) were chosen for the experiment. They concluded that the features that are most influential for predicting students' performance are students' natural gas expenditure, electricity expenditure, self-employment and location.
- Aulck et al. (2017) used a dataset of over 32,500 students to predict student drop out in an Electrical Engineering department. Examining individual features revealed that the strongest features for the prediction of students' drop out are GPA in math, English, chemistry, and psychology courses, year of enrollment, and birth year.

Table 4 Prediction of students' failure or drop out of a degree

Author (s)	Prediction	Predictor features	DM Method (s)	DM tool	Result
Pradeep and Thomas (2015)	Predict bachelor student dropout in Technology programme at Mahatma Gandhi University (MG University) in Kerala, India	Pre-enrolment features: score obtained in Mathematics in 12th grade Post-enrollment features: level of boredom in classes, attendance, taking notes from class, scores in Basic Electrical Engineering, Basic Electronics Information Technology, Basic Mechanical Engineering, Engineering Graphics, Engineering Mechanics, Engineering Mechanics1, EC, and difficulty level in Engineering Mechanics1	-Decision trees (ADT, J48, Random tree, REP tree) Rule induction (JRip, NNge, OneR, Ridor)	Weka	ADT obtained best accuracy (99.5%) followed by JRip (98.02%) then NNge and random tree with same accuracy (97.02%) then Ridor (96.53%) then REP Tree (95.05%) then J48 (94.55%) then finally OneR (89.60%).
Alemu Yehuala (2015)	Predict university students at risk of failure at Debre Markos University in Ethiopia	Demographics: gender, age, region, identity, and socio-family past Pre-enrolment features: Higher Education, Entrance examination result Post-enrollment features: study field, college, semester GPA, no. of students in class, participation in optional activities, meeting with lecturers, views on academic context, teaching professors, registered courses, number of courses per semester, and total credit hours per semester	Decision tree (J48) Naïve Bayes	Weka	J48 produced better accuracy (91.62%-92.33%) than Naïve Bayes (86.3%-87.4%)

Table 4 (continued)

Author (s)	Prediction	Predictor features	DM Method (s)	DM tool	Result
Shakeel and Anwer Butt (2015)	Predict students who are likely to drop out and student needing further help in University of Gujrat (UOG) in Pakistan	Demographics: gender Pre-enrolment features: enrollment status Post-enrollment features: registered courses, matriculation marks, intermediate marks, sessional marks, midterm marks, final term objective marks, final term subjective marks, compulsory subject marks, and general subject marks	Decision tree (J48) Naive Bayes Random Forest Bayesian Logistic Regression	Weka	Naive Bayes produced the best accuracy (91.93%) then Random Forest (88.71%), then J48 (87.09%) then Bayesian Logistic Regression (66.13%)
Villwock et al. (2015)	Identify courses contributing to student's decision of dropping out the Mathematics Major at Universidade Estadual do Oeste do Paraná – UNIOESTE in Brazil	Demographics: family income, domestic budget, expenses with the university, residence, housing, and more Pre-enrolment features: result in the course of Differential and Integral Calculus I Post-enrollment features: information on the courses taken in the first two years of the Major	Decision tree (J48)	Weka	J48 produced high accuracy (91.84%)

Table 4 (continued)

Author (s)	Prediction	Predictor features	DM Method (s)	DM tool	Result
Daud et al. (2017)	Predict the completion or drop-out of students from different universities in Pakistan	Demographics: gender, marital Status, house ownership, Scholarship, Self-employed, electricity bill, natural gas bill, telephone bill, water bill, food expenses, miscellaneous expenditure, medical, family expenditure on education, accommodation expenses, studying family members, dependent family member, family Income, and family Assets Pre-enrolment features: previous institution type, previous program	Support vector machine Bayes network Naïve Bayes Decision trees (C4.5, CART)	Weka	SVM produced best accuracy (86.7%) followed by Bayes network & Naïve Bayes with the same accuracy (84.8%) then C4.5 (76.6%) then finally CART (71%)
Aulck et al. (2017)	Predict student drop out using the first semester's grades in the Electrical Engineering department at the Eindhoven University of Technology in the USA	Demographics: gender, race, residency status, and birthdate Pre-enrolment features: previous schooling, (SAT and ACT scores, if available), transcript records Post-enrolment features: received grades, taken classes, and time at which they were taken	Logistic regression Random forests K-nearest neighbor	n/a	Logistic regression produced the best accuracy (66.59%) then k-nearest neighbor (64.60%) then random forests (62.24%)

Table 4 (continued)

Author (s)	Prediction	Predictor features	DM Method (s)	DM tool	Result
Kemper (2018)	Predict student dropout at KIT university in Karlsruhe, Germany	Demographics: gender, age, origin, and date of matriculation Post-enrollment features: study status, exam ID, exam grade, exam result, average grade in all exams, average grade in passed exams, average grade in failed exams, count of all exams, count of passed exams, and count of failed exams	Logistic regressions Decision trees	n/a	Decision trees to produce slightly better accuracy (91.3%) than Logistic regressions (90.08%)

Table 4 provides a summary of the studies we analyzed regarding the prediction of students' failure or drop out of a degree, with respect to the used features, tools, methods and accuracy of the prediction.

4.4.3 Prediction of Students' Results on Particular Courses

The prediction of a student's achievement at a course level can help instructors develop a good understanding of how well the students in their classes perform and as a result, take proactive measures to improve students' learning experience. For instance, if the prediction shows that some of the students in the class are "at risk of failing the course", educators may consider taking specific proactive measures to help those students achieve better in the given course. This can be done by adopting a variety of active and cooperative learning strategies. Following is a brief presentation of some studies that have been performed to predict students' results on particular courses:

- Kovačić (2010) collected data from 453 students to predict their performance in an "Information Systems" course. They tried to find out whether the successful vs unsuccessful student can be distinguished in terms of demographic features (such as gender, age, ethnicity, and disability) or by study environment (such as course program, faculty or course block). Their results suggest that the information gathered during the enrolment process (demographics, secondary school, working status, and early enrolment) are not sufficient for an accurate distinction between successful and unsuccessful students.
- Osmanbegović et al. (2012) used a dataset of 257 student records to predict their performance in a "Business Informatics" course. They performed an analysis to determine the importance of each feature individually. The results of their analysis revealed that the GPA impacts output the most, followed by entrance exam, then the study material, then the average weekly hours dedicated to studying. On the other hand, the number of household members, distance of residence from the faculty, and gender had the smallest output impact.
- Huang and Fang (2013) used the data of 323 undergraduate students to predict their performance in a "Dynamics" course. One of their interesting findings is that the grades that students earn in pre-requisite courses might not truly reflect the knowledge of the students in those topics. This is due to the fact that they may have taken pre-requisite courses years ago, and by the time they take the dependent course, their knowledge in the pre-requisite courses may have improved.
- Badr et al. (2016) used 203 students' records to predict their performance in a "Programming" course. They analyzed the relationship between the programming course and the other courses and found out that only the English courses have a direct effect on the prediction.
- Al luhaybi et al. (2018) collected data from 129 students to predict the students at high risk of failure in four computer science core modules. The predicted class feature is the "Overall Grade", which is the final grade obtained by the student in the targeted module. The overall grade has five possible values A: Excellent, B: Very Good, C: Good, D: Acceptable, and F: Fail, which have been merged on to Low risk, Medium risk and High risk of failure to improve the classification results. A significant finding in their study was that student qualifications on the program

entry have a high impact on their academic performance. They also found out that some of the final grades in previous modules are influencing the students' academic results in the current modules.

Table 5 provides a summary of the studies we analyzed regarding the prediction of students' results on particular courses, with respect to the used features, tools, methods and accuracy of the prediction.

As it can be seen from Tables 4 and 5, there are no particular features, tools or methods used for particular tasks, but rather the same methodologies are used across the three tasks.

5 Current Trends and Future Work

While this survey covers academic prediction studies performed between 2007 and 2018, there are also more recent studies that have been published in 2019 and 2020. Many of these approaches (Adekitan and Salau 2019; Berens et al. 2019; Bhutto et al. 2020) still rely on traditional machine learning methods, such as SVM, decision trees, logistic regression, and Naïve Bayes. However, there are also some new data mining methods that have been explored, such as Structural Equation Modeling (Nabizadeh et al. 2019) and probabilistic neural networks (Adekitan and Salau 2019). Although deep neural nets have seen a growing popularity in the machine learning community, particularly with applications to natural language processing, they are still not adopted in the EDM literature. This is probably due to their need of very large training data, whose sourcing is problematic in educational contexts.

With respect to the type of features used to perform the predictions, demographical features (Berens et al. 2019; Bhutto et al. 2020; Nabizadeh et al. 2019), previous GPA (Adekitan and Salau 2019; Berens et al. 2019; Bhutto et al. 2020; Nabizadeh et al. 2019), as well as students' satisfaction and interaction with system (Bhutto et al. 2020) are still commonly used. However, some new features are also being investigated such as cognitive and metacognitive learning strategies (Nabizadeh et al. 2019). These strategies include students' effort in rehearsal, elaboration, organization, critical thinking, and metacognitive self-regulation. Intrestengly, using such features shows promising results. This comes to confirm that prediction of students' performance is still a very actively researched problem, whose current solutions can still be improved, and that the factors that mostly influence academic outcomes and hence can be used to predict future performances are still not widely understood.

With respect to opportunities for further research in the domain of EDM, we identify two critical gaps in the previous literature. First relates to the investigation of the relation between personality and academic achievement. In light of the recent personality measures such as IPIP-NEU (Goldberg et al. 2006), we identify the opportunity of using such measures for academic performance achievement. However, the complexity of collecting such data is an important challenge to overcome. The second gap we identify relates to student self-assessment. None of the studies reports on analyzing the relationship between self-assessment and actual performance. We expect that the inclusion of such measures has the potential to further improve the accuracy of academic performance prediction.

Table 5 Prediction of students' results on particular courses

Author (s)	Prediction	Predictor features	DM method (s)	DM Tool	Result
Kovačić (2010)	Predict successful and unsuccessful student in Information Systems course in New Zealand	Demographics: gender, age, ethnicity, disability, work status, early enrollment, Course program, and course block Pre-enrolment features: GPA of secondary school	Decision trees (CHAID, CART)	SPSS	CART produced better accuracy (60.5%) than CHAID (59.4%)
Osmanbegović et al. (2012)	Predict students' success in business informatics course in the Faculty of Economics, in Tuzla, Bosnia and Herzegovina	Demographics: gender, scholarship, marital status, income, distance, and no. of household members Pre-enrolment features: GPA, high school, entrance exam score Post-enrollment features: grade importance, studying time, and internet reach	Decision tree (J48) Naïve Bayes multilayer prediction	Weka	Naïve Bayes produced the best accuracy (76.65%) followed by J48 (73.93%) then multilayer prediction (71.2%)
Huang and Fang (2013)	Predict student academic performance in Engineering Dynamics in Utah University in the USA	Pre-enrollment features: GPA, grades earned in four prerequisite courses: Engineering Statics, Calculus I, Calculus II, and Physics Post-enrollment features: scores earned in three Dynamics mid-exams	Multivariate linear regression Neural Networks Support Vector Machine	SPSS	The developed predictive models have an average prediction accuracy of 86.8–90.7%
Badr et al. (2016)	Predicting students' grades in a programming course for KSU mathematics department in Riyadh, Saudi Arabia	Post-enrollment features: Model 1: grades in 2 English courses and 2 math courses Model 2: grades in two English courses only	CBA rule-generation algorithm	LUCS-KDD	Model 2 produced better accuracy (67.33%) than model 1 (62.75%)

Table 5 (continued)

Author (s)	Prediction	Predictor features	DM method (s)	DM Tool	Result
Al luhaybi et al. (2018)	Predict 2nd-year computer science student academic performance in 4 computer science core courses at Brunel University in London, UK	Demographics: gender, age, and country Pre-enrolment features: previous institute, qualifications, enrollment status Post-enrollment features: program name, chosen route, study mode, first-year final grades, fees status, and model related data such as attendance and tutor	Decision tree (J48) Naïve Bayes	Weka and Java API	Naïve Bayes produced slightly better accuracy (78.79%) than J48 (77.3%)

6 Summary and Conclusion

Educational data mining is an area that holds exciting opportunities for researchers and practitioners all around the world. This field assists in improving institutional effectiveness by supporting decision making and enhancing student learning to reach visible and measurable targets. This paper provides a rich literature review on the prediction of academic achievement in higher education for the past 12 years with the final aim of providing researchers and educational planners with information to assist them when attempting to carry out an EDM solution.

This paper revealed that a considerable amount of work has been performed in analyzing and predicting academic performance. It showed that classification and regression algorithms can be used successfully to predict students' academic achievement in both course and degree level. It can be seen that most of the reviewed EDM research in the past decade has been completed using the open source Weka machine learning software. We found that the most used methods for predicting academic achievement are decision tree algorithms, with C4.5 being most popular among them, most likely because such white box classification algorithms obtain models that can explain predictions by IF–THEN rules. These rules can be simply interpreted by non-expert DM users, such as teachers, and can be directly applied in decision making. On the other hand, neural networks, support vector machines and K-nearest neighbour were not frequently used as compared to the rest of the DM methods. Such methods are not suitable for EDM purpose due to their black-box mechanisms.

We also found that the used features broadly differ based on the specific settings of the institute, culture, and country. However, gender, age, previous GPA and the proficiency level of the academic language are the features that most researchers agreed on when predicting students' academic achievement in higher education regardless of their environment, i.e., where they come from and what they believe in. Nevertheless, an essential limitation of the surveyed literature is the fact that only a few studies investigate and report the significance of each predictor. Rather, the vast majority of studies report only the final results, making it challenging to judge the value of each feature, even for the very widely used ones. We therefore conclude that more research is needed first to deepen our understanding of the contribution of each traditionally used feature, and second, to extend the set of features and methodologies for further improving the current prediction accuracies.

Funding Open Access funding enabled and organized by Projekt DEAL. Funding was provided by Universität Mannheim.

References

Abu Saa, A. (2016). Educational data mining and students' performance prediction. *International Journal of Advanced Computer Science and Applications*. https://doi.org/10.1007/s10462-018-9620-8.

Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon, 5*(2), e01250.

Al luhaybi, M., Tucker, A., & Yousefi, L. (2018). The prediction of student failure using classification methods: A case study. In *Computer science and information technology* (pp. 79–90). Academy & Industry Research Collaboration Center (AIRCC).

Alemu Yehuala, M. (2015). Application of data mining techniques for student success and failure prediction (The Case Of Debre_Markos University). *International Journal of Scientific & Technology Research*, *4*(04).

Ali, S., Haider, Z., Munir, F., Khan, H., & Ahmed, A. (2013). Factors contributing to the students academic performance: A case study of Islamia University Sub-Campus. *American Journal of Educational Research, 1*(8), 283–289.

Alija, S. (2013). How attendance affects the general success of the student. *International Journal of Academic Research in Business and Social Sciences, 3*(1), 168.

An, A., Cercone, N., & Chan, C. (1997). Integrating rule induction and case-based reasoning to inhance problem solving. In J. Carnegie, M. U. Carbonell, & University of S. Siekmann (Eds.), *Second international conference on case-based reasoning* (pp. 499–458). Springer, Berlin.

Arsad, P. M., Buniyamin, N., & Manan, J. A. (2014). *Students' English language proficiency and its impact on the overall student's academic performance: An analysis and prediction using Neural Network Model*.

Asif, R., Merceron, A., Abbas Ali, S., & Ghani Haider, N. E. D. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113,* 177–194.

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). *Predicting student dropout in higher education*. Seattle. Retrieved 22 August, 2018 from https://arxiv.org/pdf/1606.06364.pdf.

Badr, G., Algobail, A., Almutairi, H., & Almutery, M. (2016). Predicting students' performance in university courses: A Case Study And Tool in KSU mathematics department. *Procedia Computer Science, 82,* 80–89.

Berens, J., Oster, S., Schneider, K., & Burghoff, J. (2019). Early detection of students at risk: Predicting student dropouts using administrative student data and machine learning methods. *Schumpeter School of Business and Economics, 11*(3), 1–32.

Bhutto, S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020). Predicting students' academic performance through supervised machine learning. In *ICISCT 2020: 2nd international conference on information science and communication technology*. Institute of Electrical and Electronics Engineers Inc.

Brahmeswara Kadaru, B., & Umamaheswararao, M. (2017). An overview of general data mining tools. *International Research Journal of Engineering and Technology, 04*(09), 930–936.

Chang, Y. (2008). Gender differences in science achievement, science self-concept, and science values. In *The proceedings of international association for the evaluation of educational achievement (IRC)*.

Clark, M. (2013). *An introduction to machine learning with applications in R*. Retrieved 22 August 2018 from http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/ML_inR.pdf.

Daud, A., Aljohani, N. R., Ayaz Abbasi, R., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. In *WWW'17 companion: Proceedings of the 26th international conference on world wide web companion*. Perth, Australia: ACM.

Demetriou, C. P., & Schmitz-Sciborski, A. (2011). Integration, motivation, strengths and optimism : Retention theories past, present and future. In R. Hayes (Ed.), *Proceedings of the 7th national symposium on student retention* (pp. 300–312). Charleston: The University of Oklahoma.

Domingos, P. (1995). *Rule induction and instance-based learning a unified approach*. Retrieved 28 August 2018 from https://pdfs.semanticscholar.org/3f9b/a769643cdc530e93f80cea49889415792099.pdf.

Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *IEEE Access*. Institute of Electrical and Electronics Engineers Inc.

Eitle, T. M. (2005). Do gender and race matter? Explaining the relationship between sports participation and achievement. *Sociological Spectrum, 25*(2), 177–195.

Geng, M. (2006). *A comparison of logistic regression to random forests for exploring differences in risk factors associated with stage at diagnosis between black and white colon cancer patients*. Master Degree Thesis. Department of Biostatistics, University of Pittsburg.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84.

Goni, U., YaganaWali, S., Kaltum, A. H., & Waziri, B. M. (2015). Gender difference in students' academic performance in colleges of education in Borno State, Nigeria: Implications for counselling. *Journal of Education and Practice, 6*(32), 107–114.

Harrington, P. (2011). Machine learning in space: Extending our reach. *Machine Learning, 84*(3), 335–340.

Holland, J. L., & Nichols, R. C. (1964). Prediction of academic and extra-curricular achievement in college. *Journal of Educational Psychology, 55*(1), 55–65.

HoYu, C., DiGangi, S., Jannasch-Pennell, A., & Kaprolet, C. (2010). a data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science, 8,* 307–325.

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models.

Ian, H. W., Frank, E., Hall, M. A., & Christopher, J. P. (2017). *Data mining: Practical machine learning tools and techniques—Part II: More advanced machine learning schemes* (4th ed.). Burlington: Morgan Kaufmann.

Ibrahim, Z., & Rusli, D. (2007). Predicting students' academic performance: Comparing artificial neural network, decision tree and linear regression. In *21st Annual SAS Malaysia forum*.

Jain, A., Rawat, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications, 163*(8), 975–8887.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies, 13*(1), 61–72.

Kaushal, A., & Shukla, M. (2014). Comparative analysis to highlight pros and cons of data mining techniques-clustering, neural network and decision tree. *International Journal of Computer Science and Information Technologies, 5*(1), 651–656.

Kemper, L. (2018). *Predicting student dropout: A machine learning approach*. Karlsruhe Institute of Technology, Internal Report.

Kim, U. (1995). *Individualism and collectivism a psychological, cultural and ecological analysis*. Nordic Institute of Asian Studies Report Series No. 21, ISBN: 8787062194.

Kovačić, Z. J. (2010). Early prediction of student success: Mining students enrolment data. In *Proceedings of informing science and IT education conference* (pp. 647–665).

Kumar, M., Singh, A. J., & Handa, D. (2017). Literature survey on student's performance prediction in education using data mining techniques. *International Journal of Education and Management Engineering, 7*(6), 40–49.

Lakshmi, D. (2012). Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface, 1*(1), 05–09.

Lukkarinen, A., Koivukangas, P., & Seppälä, T. (2016). Relationship between class attendance and student performance. *Procedia: Social and Behavioral Sciences, 228*(July), 341–347.

Ma, L.-C., & Wooster, A. R. (2009). Marital status and academic performance in college. *College Student Journal, 13*(2), 106–111.

Mohamadian, Z., Fallah, S., Safdarian, A., & Jalali, Z. (2015). An open access. *Indian Journal of Fundamental and Applied Life Sciences, 5*(S1), 1262–1270.

Nabizadeh, S., Hajian, S., Sheikhan, Z., & Rafiei, F. (2019). Prediction of academic achievement based on learning strategies and outcome expectations among medical students. *BMC Medical Education, 19*(1), 99.

Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *2007 37th Annual frontiers in education conference—Global engineering: knowledge without borders, opportunities without passports* (pp. T2G-7–T2G-12), IEEE.

Nithya, P., Umamaheswari, B., & Umadevi, A. (2016). A survey on educational data mining in field of education. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 5*(1), 1–16.

Osmanbegović, E., Suljic, M., & Suljić, M. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business*, *X*(1).

Pal, A. K., & Pal, S. (2013). Analysis and mining of educational data for predicting the performance of students. *International Journal of Electronics Communication and Computer Engineering, 4*(5), 2278–4209.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*.

Pinheiro, R., Wangenge-Ouma, G., Balbachevsky, E., & Cai, Y. (2015). The role of higher education in society and the changing institutionalized features in higher education. In *The Palgrave international handbook of higher education policy and governance* (pp. 225–242). London: Palgrave Macmillan UK.

Pradeep, A., & Thomas, J. (2015). Predicting college students dropout using EDM techniques. *International Journal of Computer Applications, 123*(5), 26–34.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *ScienceDirect, 33*(33), 134–146.

Sembiring, S., Zarlis, M., Hartama, D., Wani, E., & Magister, P. (2011). Prediction of student academic performance by an application of data mining techniques. In *International conference on management and artificial intelligence* (Vol. 6, pp. 110–114).

Shakeel, K., & Anwer Butt, N. (2015). Educational data mining to reduce student dropout rate by using classification. In *253rd OMICS international conference on big data analysis & data mining*. Lexington.

Simsek, A., & Balaban, J. (2010). Learning strategies of successful and unsuccessful university students. *Contemporary Educational Technology, 1*(1), 36–45.

Slater, S., Joksimović, S., Kovanovic, V., Baker, R. S., & Gasevic, D. (2017). Tools for educational data mining: A review. *Journal of Educational and Behavioral Statistics, 42*(1), 85–106.

Song, Y.-Y., & Lu, Y. (2015). Decision tree methods: Applications for classification and prediction. *Shanghai Archives of Psychiatry, 27*(2), 130–135.

Sturman, M. C. (2003). Searching for the inverted U-shaped relationship between time and performance: Meta-analyses of the experience/performance, tenure/performance, and age/performance relationships.

Szafran, R. F., & Austin, S. F. (2002). The effect of academic load on success for new college students: Is lighter better? *NACADA Journal, 22,* 26–38.

Tomar, D., & Agarwal, S. (2013). A survey on data mining approaches for healthcare. *International Journal of Bio-Science and Bio-Technology, 5*(5), 241–266.

Villwock, R., Appio, A., & Andreta, A. A. (2015). Educational data mining with focus on dropout rates. *IJCSNS International Journal of Computer Science and Network Security, 15*(3), 17–23.

Wait, I. W., & Gressel, J. W. (2009). Relationship between TOEFL score and academic success for international engineering students. *Journal of Engineering Education, 98*(4), 389–398.

Watkins, D., & Hattie, J. (1985). A longitudinal study of the approaches to learning of Australian tertiary students: Human learning. *Journal of Practical Research & Applications, 4*(2), 127–141.

Baker, R. S. J. d., & Yacef, K. (2009). *JEDM Journal of Educational Data Mining*, *1*(1), 3–17.

Yadav, S. K., Bharadwaj, B., & Pal, S. (2011). Data mining applications: A comparative study for predicting student's performance. *International Journal of Innovative Technology and Creative Engineering, 1*(12), 13–19.

Yadav, S. K., & Pal, S. (2012). Data mining: A prediction for performance improvement of engineering students using classification. *World of Computer Science and Information Technology Journal (WCSIT), 2*(2), 51–56.

Yassein, N. A., Gaffer, R., Helali, M., & Mohomad, S. B. (2017). Citation: Predicting student academic performance in KSA using data mining techniques. *Journal of Information Technology and Software Engineering, 7*(5), 213.

Yess, J. P. (2009). Influence of marriage on the scholastic achievement of community college students: Humanities, social sciences and law. *American Journal of Educational Research., 4*(2), 103–118.

York, T. T., Gibson, C., & Rankin, S. (2015). Defining and measuring academic success: Practical Assessment. *Research & Evaluation, 20*(5), 462.

Yu-Wei, C., & David, C. (2015). *Machine learning with R cookbook: Explore over 110 recipes to analyze data and build predictive models with the simple and easy-to-use R code*. Birmingham: Packt Publishing.