Journal of Information Technology Education: Innovations in Practice

An Official Publication of the Informing Science Institute
InformingScience.org

JITEiip.org

Volume 20, 2021

USING EDUCATIONAL DATA MINING TO PREDICT STUDENTS' ACADEMIC PERFORMANCE FOR APPLYING EARLY INTERVENTIONS

Sarah Alturki*	Data and Web Science Group, Faculty of Business Informatics and Mathematics, University of Mannheim, Mannheim, Germany	alturki@informatik.uni-mannheim.de
Nazik Alturki	Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia	namalturki@pnu.edu.sa

* Corresponding author

ABSTRACT

Aim/Purpose	One of the main objectives of higher education institutions is to provide a high-quality education to their students and reduce dropout rates. This can be achieved by predicting students' academic achievement early using Educational Data Mining (EDM). This study aims to predict students' final grades and identify honorary students at an early stage.
Background	EDM research has emerged as an exciting research area, which can unfold valuable knowledge from educational databases for many purposes, such as identifying the dropouts and students who need special attention and discovering honorary students for allocating scholarships.
Methodology	In this work, we have collected 300 undergraduate students' records from three departments of a Computer and Information Science College at a university located in Saudi Arabia. We compared the performance of six data mining methods in predicting academic achievement. Those methods are C4.5, Simple CART, LADTree, Naïve Bayes, Bayes Net with ADTree, and Random Forest.

Accepting Editor Dennis Kira | Received: May 9, 2021 | Revised: June 24, July 11, July 13, July 16, 2021 | Accepted: July 17, 2021.
Cite as: Alturki, S., & Alturki, N. (2021). Using educational data mining to predict students' academic performance for applying early interventions. *Journal of Information Technology Education: Innovations in Practice, 20,* 121-137. https://doi.org/10.28945/4835

Contribution	We tested the significance of correlation attribute predictors using four different methods. We found 9 out of 18 proposed features with a significant correlation for predicting students' academic achievement after their 4th semester. Those features are student GPA during the first four semesters, the number of failed courses during the first four semesters, and the grades of three core courses, i.e., database fundamentals, programming language (1), and computer network fundamentals.
Findings	The empirical results show the following: (i) the main features that can predict students' academic achievement are the student GPA during the first four semesters, the number of failed courses during the first four semesters, and the grades of three core courses; (ii) Naïve Bayes classifier performed better than Tree-based Models in predicting students' academic achievement in general, however, Random Forest outperformed Naïve Bayes in predicting honorary students; (iii) English language skills do not play an essential role in students' success at the college of Computer and Information Sciences; and (iv) studying an orientation year does not contribute to students' success.
Recommendations for Practitioners	We would recommend instructors to consider using EDM in predicting students' academic achievement and benefit from that in customizing students' learning experience based on their different needs.
Recommendations for Researchers	We would highly endorse that researchers apply more EDM studies across various universities and compare between them. For example, future research could investigate the effects of offering tutoring sessions for students who fail core courses in their first semesters, examine the role of language skills in social science programs, and examine the role of the orientation year in other programs.
Impact on Society	The prediction of academic performance can help both teachers and students in many ways. It also enables the early discovery of honorary students. Thus, well-deserved opportunities can be offered; for example, scholarships, internships, and workshops. It can also help identify students who require special attention to take an appropriate intervention at the earliest stage possible. Moreover, instructors can be aware of each student's capability and customize the teaching tasks based on students' needs.
Future Research	For future work, the experiment can be repeated with a larger dataset. It could also be extended with more distinctive attributes to reach more accurate results that are useful for improving the students' learning outcomes. Moreover, experiments could be done using other data mining algorithms to get a broader approach and more valuable and accurate outputs.
Keywords	Educational Data Mining (EDM), prediction of academic achievement, higher education

INTRODUCTION

Since one of the benchmarks of a high-quality university is based on the students' excellent record of academic performance, students' failure or low achievement is a big concern for higher education institutions. Moreover, low academic achievement is a problem that negatively affects the individual and the community, e.g., it may lead to unemployment. Therefore, it is essential to address such issues by investigating the factors associated with students' success and finding ways for early intervention to help low-performing students (Jayaprakash & Jaiganesh, 2019).

Data Mining (DM) is widely applied in the Education field (Romero & Ventura, 2010) and is one of the most popular techniques to analyze students' performance. DM refers to extracting or "mining" knowledge from large amounts of data to discover hidden patterns and relationships that are helpful in decision making. Currently, there is an increase of research interest in using DM in education due to its high potential in improving educational institutes (Baradwaj & Pal, 2011). The practice of DM methods applied to educational data is known as Educational Data Mining (EDM) (Baker & Yacef, 2009). It concerns developing methods that discover knowledge from educational environment data (Han et al., 2011) that are drawn from a variety of domains, including DM and machine learning, psychometrics and other areas of statistics, information visualization, and computational modeling (Romero & Ventura, 2007). EDM provides educators and students with useful insights into the education process, resulting in suitable actions and decisions that improve academic success (Kotsiantis, 2009).

Predicting students' academic performance is one of EDM's main focuses (Fan et al., 2019). Such prediction brings many advantages, for instance, providing valuable feedback, offering recommendations, supporting personalized learning, and strategically planning educational programs (Hellas et al., 2018). There are three types of predictions in higher education: (i) predicting students' academic performance or GPA at a degree level; (ii) predicting students' failure or drop out of a degree; and (iii) predicting students' results on particular courses (Alturki et al., 2020). In this study, we address the first type. We provide a conceptual framework that may be used to create a recommender system that uses classification algorithms to predict students' academic achievement at an early stage. Such a system will enable higher education institutions to create quality graduates and reduce student attrition. Furthermore, it will assist educational institutions in early intervention to improve students' performance.

The rest of the paper is organized as follows. The next section presents the study's overview and the research questions, the background of the study, and the related work. Following that, the research methodology is explained. Then we provide details of the experimental results and discussion. Finally, we conclude with a summary of the study's primary outcomes, outline our research limitations, and suggest future lines of research.

STUDY OVERVIEW

In this section, the research questions are outlined, the empirical study background is presented, and related previous work is discussed.

RESEARCH QUESTIONS

The four research questions proposed for this study are as follows:

RQ.1: What are the significant attributes for predicting student academic achievement in the College of Computer and Information Sciences?

RQ.2: Is it possible to discover honorary students at an early stage of their studies?

RQ.3: Do academic language skills play a role in students' success?

RQ.4: Does studying an Orientation year (or Preparatory year) affect students' success?

THE COLLEGE OF COMPUTER AND INFORMATION SCIENCES

At the studied College of Computer and Information Sciences, a 5.00-grade point average (GPA) system is used for evaluating students' achievement each semester. The final GPA is calculated by dividing the total quality points earned by the total number of credit hours for which grades were assigned. The college offers three majors: Computer Science (CS), Information Technology (IT), and Information Systems (IS). These majors are within the same realm of study. However, each major

focuses on specific aspects of the field of computer and information sciences. The CS major focuses on the theory of computational applications, i.e., understanding the "why" behind computer programs. The major is based on algorithms and mathematics – the language of computers. CS students also learn the fundamentals of programming languages, linear and discrete mathematics, and software design and development. Students study the machine itself and understand how and why various computer processes operate the way they do.

The IT major focuses more on network models and their protocols and the types of traffic generated, and their quality of service requirements. Students learn how to deal with performance issues in networks and competence in the use of techniques to analyze and optimize performance. IT students also focus on internet design principles, internet routing design, internet application protocols, and cryptography and security.

Students enrolled in the IS major focus on meeting the needs of users in an organizational context through the selection, creation, practical application, integration, and administration of computing technologies. Students also learn how to use and apply current technical concepts and practices and how to analyze, identify and define the requirements that must be satisfied to address IT problems or opportunities faced by organizations or individuals. Moreover, students learn the fundamentals of effectively designing IT-based solutions and integrating them into the user environment, along with identifying and evaluating current and emerging technologies and discuss their applicability to solve the users' needs.

Previously, all three majors could be completed within five years, i.e., ten semesters (two semesters as part of the orientation [Preparatory] year, eight semesters as part of the major). An orientation year is a program designed to prepare students for their higher education. During the orientation year, students are given courses of the English language and foundational undergraduate-level courses such as basics in physics and mathematics. The aim of this year is to allow students to adjust to the new academic environment and teaching system. However, the orientation year has been recently eliminated from the college programs, making it possible to graduate within four academic years.

There are twelve core college courses shared between the three majors taught in the English language. Programming Language (1), Programming Language (2), Database Fundamentals, and Computer Networks Fundamentals, are among the introductory courses taught during the first two years of all three majors.

Related Work on Predictions in Higher Education

Substantial research on the effectiveness of teaching methods indicates that the quality of teaching is often reflected by the achievements of learners (Ganyaupfu, 2013). Educators should apply appropriate teaching methods that best suit specific objectives (Ganyaupfu, 2013). Adjusting the teaching according to specific students' needs can be achieved by using methods of EDM. Various EDM applications can be applied in educational institutes (Romero & Ventura, 2013). Predicting student academic performance is one of the leading applications. This section of the paper provides a literature review on related previous studies and draws particular attention to the attributes that researchers have widely been using to predict academic achievement.

Based on Alturki et al. (2020), the features that are used for predicting academic achievement can be classified into three categories. They are: (i) demographics, (ii) pre-enrollment features, and (iii) post-enrollment features. Although the demographical features are heavily used for predicting academic achievement, the extent to which they are useful is unclear (Alturki et al., 2020). One of the top used features in this category is gender (Aulck et al., 2016; Daud et al., 2017; Garg, 2018; Kovačić, 2010; Osmanbegović & Suljic, 2012; Shakeel & Butt, 2015). However, some researchers found that gender does not significantly impact the overall prediction (Kovačić, 2010; Osmanbegović & Suljic, 2012). Age is also one of the common features used to predict academic achievement (Aulck et al., 2016; Kemper, 2020; Kovačić, 2010; Yehuala, 2015). However, Kovačić (2010) reported that age does not

significantly impact predicting academic success. Country of origin is also used widely for predicting academic achievement (Abu Saa, 2016; Aulck et al., 2016; Kemper, 2020).

Using pre-enrollment features to predict students' academic achievement is significant, especially if the prediction is to be performed at an early stage (Alturki et al., 2020). Previous GPA is one of the most popular used features for predicting academic success (Abu Saa, 2016; Aluko et al., 2018; Garg, 2018; Huang & Fang, 2013; Kabakchieva, 2013; Kovačić, 2010; Osmanbegović & Suljic, 2012; Pal & Pal, 2013; Thai-Nghe et al., 2007). Academic language skills have also been widely used for predicting academic success (Abu Saa, 2016; Asif et al., 2017; Badr et al., 2016; Bani-Salameh, 2018; Thai-Nghe et al., 2007). Although some researchers (e.g. Arsad et al., 2014) claim that academic language skills do not affect students' success in "non-linguistic courses," others (e.g. Wait & Gressel, 2009) found a significant relationship. Bani-Salameh (2018) performed a study on medical students' performance and its relation to the teaching language. Students were given the same test in two languages, namely Arabic (native language), and English (teaching language), and compared their performance. The results indicate a weak correlation between students' performance and the teaching language.

Using post-enrollment features for predicting students' academic achievement can maximize the prediction accuracy as such features represent students' current situation in the program (Alturki et al., 2020). The achievement of the previous semester, which is part of the total GPA, has been used in many studies to predict students' success (e.g. Abu Saa, 2016; Al luhaybi et al., 2018; Asif et al., 2017; Kabakchieva, 2013; Thai-Nghe et al., 2007; Yehuala, 2015); that is, since students' success is highly dependent on previously acquired knowledge. Asif et al. (2017) found that the results of a four-year program's first and second-year courses play a significant role in predicting graduation performance. Using students' grades that are earned in quizzes and examinations have also been widely used in various studies for predicting students' academic success (e.g. Al luhaybi et al., 2018; Aulck et al., 2016; Badr et al., 2016; Huang & Fang, 2013; Kemper et al., 2020; Pradeep & Thomas, 2015; Shakeel & Butt, 2015; Villwock et al., 2015; Yadav et al., 2011; Yassein et al., 2017). Failure in examinations has also been used as a predictor; for instance, Kabakchieva (2013) used a dataset of 10,330 students to predict their performance using five classes (bad, average, good, very good, and excellent) and found that the number of failures at the first-year exams is among the most influencing features in the classification. It is relatively reasonable to use such a feature as a predictor as a "high rate of academic failure is often observed in the first year of studies" (Gilar-Corbi et al., 2020, p. 1). Academic load, measured in terms of credit hours and course difficulty per a single semester (Szafran & Austin, 2002), is also used to predict students' success (e.g. Abu Saa, 2016; Yehuala, 2015). In fact, Yehuala (2015) found that the number of credit hours is one of the main significant attributes for predicting academic achievement.

In terms of the performance of different DM methods, there is no single method that works best in all settings. Therefore, researchers often explore two or more DM methods to reveal which generates the best accuracy in their particular case (Alturki et al., 2020). Yohannes and Ahmed (2018) designed an application to assist higher education institutions in predicting their students' academic performance upon graduation (eighth semester). They used students' scores for core and non-core courses from the first to the sixth semester and found that Support Vector Regression and Linear Regression performed better than Neural Networks. A study by Khasanah and Harwati (2017) conducted Feature Selection to find high influence attributes with students' performance. They used students' demographics, pre-enrolment information, and post-enrolment information. They found that Bayesian Network outperforms Decision Trees and that students' attendance and students' GPA in the first semester are the two most important features for predicting academic achievement. Shakeel and Butt (2015) used demographics, pre-enrollment features, and post-enrollment features to predict students who are likely to drop out. They found that Naïve Bayes performed best, followed by Random Forest, then J48, then Logistic Regression.

One identified research gap is that none of the reviewed studies explored the effect of the orientation year on predicting students' academic achievement. McMullen (2014) emphasizes the

importance of the orientation year on students' success as they provided statistical proof that students in Saudi Arabia believe that they leave secondary school without the skills necessary to enter their academic majors. Since the orientation year has been shown to improve students' attrition (Davig & Spain, 2003), it is, therefore, essential to examine its impact on predicting students' achievement.

RESEARCH METHODOLOGY

DATA COLLECTION

Participants

The data set of 300 female students aged between 20 and 22 used in this study was obtained randomly from a College of Computer and Information Sciences at a Saudi university. One hundred of the collected records are from the IS major, one hundred from the IT major, and one hundred records from the CS major. While 117 of the collected students have studied an orientation year, the remaining 183 have not. It is important to note that the GPA of the orientation year is not accounted for those who have studied it, i.e., it has no effect on the final GPA. In the collected data, the majority of the students had an "Accepted" GPA, and the minority had a "Poor" GPA. Figure 1 shows the distribution of the students' final academic achievement.

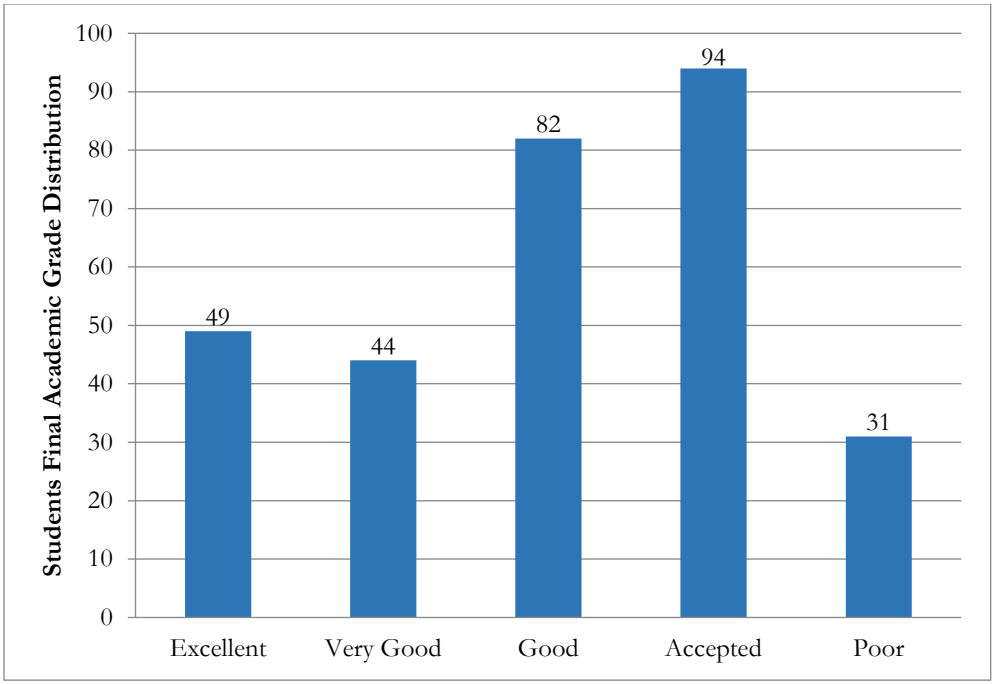

Figure 1. Distribution of the final academic grade of the collected data

There are three types of features for predicting students' academic achievement: demographics such as gender and age, pre-enrollment features such as previous GPA, and post-enrolment features such as grades achieved in each course (Alturki et al., 2020). In our study, we applied one pre-enrollment feature, which is the 'secondary school graduation percentage', and the rest are post-enrollment features, which are: two 'English course grades', 'the GPA of the first four semesters', 'academic load in first four semesters', 'number of failed courses in the first four semesters, and grades of the college's core courses. Although there are multiple shared courses between all three majors, we only considered the courses taken in the first four semesters as we aim to perform the prediction at an early stage, i.e., before the third academic year. Table 1 below shows the dataset characteristics.

Table 1. Main characteristics of the collected dataset

Feature	Description	Type	Value
GradGPA	Graduation grade	Nominal	Excellent, Very good, Good, Accepted, and Poor
SecPer	Secondary school achievement	Numeric	0 – 100%
Major	The major of the student	Nominal	CS, IS, and IT
PY	Whether the student studied an orientation year or not	Boolean	Yes or no
GPA1	Student's GPA in 1st academic semester	Numeric	0 - 5
GPA2	Student's GPA in 2nd academic semester	Numeric	0 - 5
GPA3	Student's GPA in 3rd academic semester	Numeric	0 - 5
GPA4	Student's GPA in 4th academic semester	Numeric	0 - 5
Hrs/sem1	Student's academic load per 1st semester	Numeric	12 - 24 hours
Hrs/sem2	Student's academic load per 2nd semester	Numeric	12 - 24 hours
Hrs/sem3	Student's academic load per 3rd semester	Numeric	12 - 24 hours
Hrs/sem4	Student's academic load per 4th semester	Numeric	12 - 24 hours
F/year1	Number of failed courses in the 1st academic year	Numeric	≥ 1
F/year2	Number of failed courses in the 2nd academic year	Numeric	≥ 1
Prog1	Student's grade in programming (1)	Nominal	A+, A, B+, B, C+, C, D+, D, and F
Prog2	Student's grade in programming (2)	Nominal	A+, A, B+, B, C+, C, D+, D, and F
DB	Student's grade in Database's fundamentals	Nominal	A+, A, B+, B, C+, C, D+, D, and F
NW	Student's grade in Computer Networks fundamentals	Nominal	A+, A, B+, B, C+, C, D+, D, and F

Data Preprocessing

In the raw dataset, the final GPA is within the range of 0–5.0, where 5.0 is the best possible GPA score. However, since the final GPA is in the form of an integer, and the predicted class should be categorical (nominal) values, we transformed the GPA into five categories according to the grading system. Table 2 shows the grading classification that is used in this study.

Table 2. Classification of academic grading

GPA	Grade	Symbol
4.5 - 5	Excellent	A
4.00 – 4.5	Very good	B
3.25 – 4.00	Good	C
2.5 – 3.25	Accepted	D
Less than 2.5	Poor	E

Data Mining Methods

Data Mining (DM) methods can be classified into two categories: supervised and unsupervised methods. The unsupervised methods uncover hidden patterns in unlabeled data to find patterns in a dataset, and there are no output variables to predict. On the other hand, supervised methods (also known as predictive or directive) predict the value of the output variables based on the inputs. In this study, we address supervised DM methods to predict the value of the output variables based on the inputs. To achieve this, a model is developed from training data where the values of inputs and outputs are previously labeled. 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).

Several classification algorithms can be used to predict the students' graduation performance at the end of the degree. However, the literature review suggests that, in general, there is no single classifier that works best in all contexts to provide a good prediction. Following are the six DM methods applied in this study:

C4.5: The C4.5 (J48 in Weka) is an extension of the ID3 algorithm (Interactive Dichotomize 3). This algorithm is used to generate a tree-shaped structure that represents sets of decisions. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting measure is the normalized information gain. The attribute with the highest normalized information gain is selected to make the decision.

Simple CART: The Classification and Regression Tree (CART) analysis uses a learning sample which is a set of historical data with preassigned classes for all observations for building decision trees. It is a learning technique that gives the results as either classification or regression trees, depending on the categorical or numeric data set. It uses cross-validation or a large independent test sample to select the best tree from the sequence of trees considered in the pruning process. During the implementation phase of CART, the dataset is split into the two subgroups that are the most different concerning the outcome. This procedure is continued on each subgroup until some minimum subgroup size is reached (Kalmegh, 2015).

LADTree: The Logical Analysis of Data (LAD) is a classification method built based on learning a logical expression. Since LAD is a binary classifier, it can differentiate between positive and negative samples (Amudha et al., 2011). For a dataset processed by LAD, a large set of patterns are produced, and a subset of them is selected to satisfy the above assumption such that each pattern in the model satisfies specific requirements in terms of prevalence and homogeneity (Buhmann, 2003).

Naïve Bayes: The Naive Bayes classifier, which is based on the work of Thomas Bayes, simplifies learning by assuming that features are independent given class. The Naïve Bayes follows the hypothesis that the data belongs to a particular class, then the probability for the hypothesis is calculated to be true. Thus, only one scan of the data is required. In cases of training data, each training example can incrementally increase/decrease the probability that a hypothesis is correct. Thus, Naïve Bayes perfectly fits domains containing uncertainty (Nielsen & Jensen, 2007).

Bayes Net with ADTree: A Bayesian network is a probabilistic graphical model representing a set of variables and their conditional dependencies via a directed acyclic graph. They are ideal for taking an event that occurred and predicting the possibility that any of several possible known causes was the contributing factor. An Alternating Decision Tree (ADTree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. It consists of an alternation of decision nodes that specify a predicate condition, and prediction nodes that contain a single number. An ADTree classifies an instance by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed. Bayes Net associated with an ADTree combines both methods to improve the classification accuracy.

Random Forest: As the name implies, a Random Forest is a tree-based classifier that functions as an ensemble depending on a collection of random variables (Cutler et al., 2012). Random Forest is a mixture of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. In the training stage, random forests apply the general technique known as "bagging" to individual trees in the ensemble (Caie et al., 2021). Bagging repetitively chooses a random sample with replacement from the training set and fits trees to these samples where each tree is grown without any pruning.

Data Mining Tool

For this study, we used the WEKA software package developed at the University of Waikato in New Zealand. The name stands for Waikato Environment for Knowledge Analysis. This package has been implemented in the Java software language and stands out as one of the most competent and comprehensive packages with machine learning algorithms. WEKA supports several standard data mining tasks, specifically data preprocessing, clustering, classification, regression, visualization, and feature selection (Kalmegh, 2015).

Experimental Results

Different options are available for evaluating performance measures. A few of them are cross-validation using a different number of folds and percentage split, which is used to split the dataset into two partitions – one is used for training the dataset, and the other is used for testing. In this study, a percentage split is applied, i.e., 70% of the dataset has been used for training the model, and the remaining 30% is used for testing purposes. Data sets were randomly partitioned into training and testing datasets via 10-fold CV. In 10-fold CV, the data set is divided into 10-subsets, and the holdout method is repeated ten times. Each time, one of the ten subsets is used as the test set, and the other nine subsets are put together to form the training set.

Individual Feature Analysis

As selecting the right features improves classifiers' performance, we have explored different feature selection techniques to understand the influence of different features on the target "final grade". They are Search-Based, Correlation Based, Information Gain Based, and Wrapper with Naïve Bayes. The search method is either BestFirst, which searches the space of attribute subsets by greedy hill-climbing augmented with a backtracking facility, or Ranker, which ranks attributes by their evaluations. Table 3 describes the selected methods and their results.

Table 3. Comparison between different attribute selection methods

Attribute Evaluator	Description	Search Method	Selected Attributes
Search Based (CfsSubsetEval)	Evaluates the effect of a subset of features by considering the predictive ability of each one along with the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred	BestFirst	GPA1 (100%) GPA2 (100%) GPA3 (100%) GPA4 (100%) DB (100%) F/year1 (80%) NW (30%)
Correlation Based (CorrelationAttributeEval)	Evaluates the influence of an attribute by measuring the correlation (by Pearson formula) between attribute and the target attribute.	Ranker	1. GPA3 2. GPA4 3. F/year2 4. GPA2 5. GPA1 6. F/year1 7. SecGPA 8. DB 9. Hrs/sem4
Information Gain Based (InfoGainAttributeEval)	Evaluates the influence of an attribute by measuring the information gain concerning the target attribute.	Ranker	1. GPA3 2. GPA4 3. DB 4. GPA2 5. GPA1 6. prog1 7. prog2 8. F/year2 9. F/year1
Wrapper with Naïve Bayes (WrapperSubsetEval)	Evaluates attribute sets by using a learning scheme. Cross validation is used to estimate the accuracy of the learning scheme for a set of attributes.	BestFirst	GPA3 (100 %) GPA4 (100 %) GPA1 (90%) NW (90%) F/year1 (70%) DB (60%) Prog1 (40%)

Based on the results provided above, we conclude that the attributes that have the most influence on academic achievement are the student's GPA for each semester, the number of failed courses in 1st and 2nd year, their grade in the 'Database fundamentals', and 'Programming 1' core courses. On the other hand, 'English skills', 'secondary school GPA', 'academic load', and 'programming 2' are found surprisingly to have no effect on the prediction of students' academic achievement.

Performance Evaluation of Different DM Methods

In this section, two case studies are provided for predicting students' academic achievement. In the first case, the prediction is performed after the 3rd academic semester whereas, in the second, it is performed after the 4th semester. In both cases, we compare the performance of six classifiers in terms of the following: (1) classifier accuracy, which is the total number of correct predictions divided by the total number of predictions made for a dataset; (2) Receiver Operating Characteristic (ROC) to examine the performance of a binary classifier by creating a graph of the True Positives vs False Positives for every classification threshold (the higher the ROC, the better the results); (3) F Measure, which provides a way to combine both precisions and recall into a single measure that captures both properties (a poor F-Measure score is 0.0 and a best F-Measure score is 1.0); and finally (4) the accuracy of allocating excellent students.

Predicting final academic grade after the 3rd semester

In this academic prediction case, six attributes are selected: the GPA from the 1st, 2nd, and 3rd semester; the number of failed courses during the 1st year; and the grades of the two core courses that are taken during the first three semesters, i.e., Programming Language (1), and Database Fundamentals. Table 4 below compares the different classifiers' results in predicting students' academic achievement after their 3rd semester. It can be observed that Naïve Bayes and Random Forest perform the best with an accuracy of 63.33% and 63%, respectively. Those two classifiers also perform the best in predicting honorary students with an equal accuracy of 77.6%.

Table 4. Accuracy of predicting final academic grade after the 3rd semester

Classifier	Accuracy	ROC	F Measure	Honorary students (Accuracy)
J48	55.67%	.783	.549	71.4%
SimpleCart	59.67%	.824	.576	75.5%
LADTree	58.33%	.850	.585	75.5%
Naïve Bayes	63.33%	.892	.637	77.6%
Bayes Net with ADTree	60.33%	.877	.603	77.6 %
Random Forest	63%	.872	.615	85.8%

Predicting final academic grade after the 4th semester

In this academic prediction case, nine attributes are selected to predict students' graduation grades: the GPA from semesters 1, 2, 3, and 4, number of failed courses during the 1st year, number of failed courses during the 2nd year, Programming Language (1) grade, Database Fundamentals grade, and Computer Networks Fundamentals grade. Table 5 compares the different classifiers' results in predicting students' academic achievement after their 4th semester. We can notice that Random Forest performs the best compared to the rest of the Tree-based Models with 67.6% accuracy. However, Naïve Bayes outperformed all five tree-based classifiers with an accuracy of 69.67%. In terms of which classifier performs the best in predicting honorary students, Random Forest significantly outperformed the rest of the classifiers with 92.6% accuracy.

Table 5. Accuracy of predicting final academic grade after the 4th semester

Classifier	Accuracy	ROC	F Measure	Honorary students (Accuracy)
J48	61.67%	.815	.611	77.6%
SimpleCart	63.3%	.833	.633	79.6%
LADTree	63%	.875	.630	75.5%
Naïve Bayes	69.67%	.917	.697	77.6%
Bayes Net with ADTree	66.33%	.911	.666	79.6%
Random Forest	67.6%	.909	.660	92.6%

Discussion of the Results

In this section, we answer our research questions and discuss the results obtained from the study.

As a general observation, it is clear that, with the increase of attributes, the model's accuracy increases as all six classifiers performed better in the second case than they did in the first; that is, they acquired a higher accuracy, ROC, and F measure. In terms of which classifier performed better in predicting academic achievement, the results are similar to the results of Shakeel and Butt (2015). We found that Random Forest performs the best compared to the rest of the Tree-based Models with 63% and 67.6% accuracy in the first and second case respectively. Random Forests generally yield better results as they are much more robust than a single decision tree; that is, they aggregate many decision trees to limit error and overfitting due to bias. While single decision trees search for the most important feature when splitting a node, Random Forests search for the best feature among a random subset of features. Such wide diversity usually results in a better model. The second-best Tree-based classifier was Bayes Net with ADTree with an accuracy of 60.33% and 66.33% in the first and second case respectively. In binary classification trees, such as J48, CART, an instance follows only one path through the tree, ADTree follows all paths for which all decision nodes are true and sums any prediction nodes that are traversed. However, Naïve Bayes outperformed all five tree-based classifiers with an accuracy of 63.33% and 69.67% in the first case and second case respectively. The reason behind the good performance of Naïve Bayes is described by Domingos & Pazzani (1996) as follows: "Naïve Bayes is commonly thought to be optimal, in the sense of achieving the best possible accuracy, only when the independence assumption holds, and perhaps close to optimal when the attributes are only slightly dependent. However, this very restrictive condition seems to be inconsistent with the Naïve Bayes' surprisingly good performance in a wide variety of domains, including many where there are clear dependencies between the attributes."

Referring to our first research question ("what are the significant attributes for predicting student academic achievement in the College of Computer and Information Sciences?"), we found that students' earned GPA in all four semesters plays a significant role in predicting students' academic achievement. This supports the findings of Asif et al. (2017). However, the students' earned GPAs in the 3rd and 4th semesters are found to have more influence on the prediction than the 1st and 2nd semesters. That is relatively reasonable as the courses become more challenging as students escalate in semesters and the variation in students' skills start to show more. We have also found that the number of failed courses in the first two years of the program plays a significant role in predicting students' academic achievement. This is in line with the results of Kabakchieva (2013). In terms of which fundamental course has the largest effect on students' success, we found that the Database Fundamentals course had the highest impact. On the other hand, we have unexpectedly found that the academic load does not play a significant role in predicting students' success. This contradicts the

findings of Yehuala (2015). Although previous GPA from secondary school is one of the most popular used features for predicting academic success (Abu Saa, 2016; Aluko et al., 2018; Garg, 2018; Huang & Fang, 2013; Kabakchieva, 2013; Kovačić, 2010; Osmanbegović & Suljic, 2012; Pal & Pal, 2013; Thai-Nghe et al., 2007), we have found it to have no significant effect on the overall prediction.

When it comes to our second research question regarding the ability to predict honorary students, we can conclude that it is possible to identify honorary students at an early stage of their bachelor's studies at Computer and Information Sciences colleges. In terms of which classifier performed best in such prediction, Random Forest outperformed the other Tree-based classifiers and outperformed Naïve Bayes with an accuracy of 85.8% in the first case and 92.6% in the second.

To answer our third research question regarding whether academic language skills play a role in students' success, we found that English language skills did not play a role in students' success. This is in accordance with the findings of Arsad et al. (2014) and Bani-Salameh (2018) and might be due to the fact that the nature of courses under the College of Computer and Information Science is not linguistics and rather more scientific.

The fourth research question in this paper was related to investigating whether studying an Orientation year affects students' success. Although the orientation year has shown to improve students' academic achievement (Davig & Spain, 2003; McMullen, 2014), we found that it has no effect on students' success in the College of Computer and Information Science and can therefore conclude that removing the orientation year from the study programs was a reasonable action that did not carry drawbacks on the students.

CONCLUSION

The capabilities of the DM techniques can provide useful insights for predicting the final academic performance of students. This type of prediction can help both teachers and students in many ways. It can also enable the early discovery of honorary students. Thus, well-deserved opportunities can be offered, e.g., scholarships, internships, and workshops. It can also help identify students who require special attention to take an appropriate intervention at the earliest stage possible. Moreover, instructors can be aware of each student's capability and thus can customize the teaching tasks based on students' needs, e.g., offering extracurricular learning material to students facing difficulties, using different teaching strategies, and providing online tutoring videos for those who need.

The main objectives of this study were to predict students' final grades at a degree level and identify honorary students at an early stage of their studying journey. We tested the significance of correlation attributes predictors using four different methods. We have found 9 out of 18 proposed features with a significant correlation for predicting students' academic achievement after their 4th semester. Those features are student GPA during the first four semesters, the number of failed courses during the first four semesters, and the grades of three core courses, i.e., Database Fundamentals, Programming Language (1), and Computer Network Fundamentals. We performed the prediction of students' academic achievement by investigated six supervised DM algorithms. As the classifiers' performances are evaluated based on their predictive accuracy, we found that Naïve Bayes performed better than the five selected Tree-based Models for predicting students' final graduation grades. However, Random Forest performed significantly better than Naïve Bayes in predicting honorary students with an accuracy of 85.8%, 92.6% after the third and fourth semesters, respectively.

Evaluating bachelor's programs should be an ongoing cycle. Therefore, it was essential to make sure that eliminating the orientation year from the College of Computer and Information Systems was an ideal change. The study results reveal existing evidence that the orientation year does not have an impact on students' success.

One of the limitations of this present study is that we did not investigate the influence of gender and age on the prediction of academic achievement, since all of the participants are females and

belong to the same age group. Also, due to time limitations, this research has been carried out including only students from one college and one university.

For future work, the results of this study can be used to design a recommender system that enables timely interventions at Computer and Information Systems Colleges. The experiment can be repeated with a larger dataset. It could also be extended with more distinctive attributes to reach more accurate results helpful in improving the students' learning outcomes, e.g., gender, students' attendance, students' e-learning activity information, and students' satisfaction. Future research could also investigate the effects of offering tutoring sessions for students who fail core courses in their first semesters, examine the role of language skills in social science programs, and examine the role of orientation year in other programs.

We would recommend instructors to consider using EDM in predicting students' academic achievement and benefit from that in customizing students' learning experience based on their different needs. We would also highly endorse that researchers apply more EDM studies across different universities and compare between them.

ACKNOWLEDGMENT

This research was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program.

REFERENCES

Abu Saa, A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5): 212-200. https://doi.org/10.14569/IJACSA.2016.070531

Al luhaybi, M., Tucker, A., & Yousefi, L. (2018). The prediction of student failure using classification methods: A case study. *Computer Science & Information Technology*, 79-90. https://doi.org/10.5121/csit.2018.80506

Alturki, S., Hulpus, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 1-33. https://doi.org/10.1007/s10758-020-09476-0

Aluko, R. O., Daniel, E. I., Shamsideen Oshodi, O.., Aigbavboa, C. O., & Abisuga, A. O. (2018). Towards reliable prediction of academic performance of architecture students using data mining techniques. *Journal of Engineering, Design and Technology, 16*(3), 385-397. https://doi.org/10.1108/JEDT-08-2017-0081

Amudha, J., Soman, K., & Kiran, Y. (2011). Feature selection in top-down visual attention model using WEKA. *International Journal of Computer Applications*, *24*(4). https://doi.org/10.5120/2955-3895

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. *WSEAS Transactions on Advances in Engineering Education, 11*, 44-53. https://www.wseas.org/multimedia/journals/education/2014/a105710-111.pdf

Asif, R., Merceron, A., Ali, S. A., & Haider, N. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, *113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. *Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York*. 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. https://doi.org/10.1016/j.procs.2016.04.012

Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, *1*(1), 3-17. http://doi.org/10.5281/zenodo.3554658

Bani-Salameh, H. N. (2018). Teaching language effects on students' performance. *Health Professions Education, 4*(1), 27-30. https://doi.org/10.1016/j.hpe.2017.01.005

Baradwaj, B. K., & Pal, S. (2011). Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications, 2*(6), 63-69. https://doi.org/10.14569/IJACSA.2011.020609

Buhmann, M. D. (2003). *Radial basis functions: Theory and implementations* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511543241

Caie, P. D., Dimitriou, N., & Arandjelović, O. (2021). Precision medicine in digital pathology via image analysis and machine learning. In S. Cohen (Ed.), *Artificial intelligence and deep learning in pathology* (pp. 149-173). Elsevier. https://doi.org/10.1016/b978-0-323-67538-3.00008-7

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Ma (Eds.), *Ensemble machine learning* (pp. 157-175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Daud, A., Aljohani, N. R., Abbasi, R., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting student performance using advanced learning analytics. *Proceedings of the 26th International Conference on World Wide Web Companion,* 415-421. https://doi.org/10.1145/3041021.3054164

Davig, W. B., & Spain, J. W. (2003). Impact on freshmen retention of orientation course content: Proposed persistence model. *Journal of College Student Retention: Research, Theory & Practice*, *5*(3), 305-323. https://doi.org/10.2190/V6B4-PQAW-TTV0-CJCU

Domingos, P. M. & Pazzani, M. (1996). Beyond independence: Conditions for the optimality of the simple bayesian classifier *Proceedings of the 13th ICML Conference on on Machine Learning,* 105–112. http://www.ics.uci.edu/~pazzani/Publications/mlc96-pedro.pdf

Fan, Y., Liu, Y., Chen, H., & Ma, J. (2019). Data mining-based design and implementation of college physical education performance management and analysis system. *International Journal of Emerging Technologies in Learning*, *14*(6), 87-97. https://doi.org/10.3991/ijet.v14i06.10159

Ganyaupfu, E. M. (2013). Teaching methods and students' academic performance. *International Journal of Humanities and Social Science Invention, 2*(9), 29-35.

Garg, R. (2018). Predicting student performance of different regions of Punjab using classification techniques. *International Journal of Advanced Research inComputer Science, 9*(1), 236-241. https://doi.org/10.26483/ijarcs.v9i1.5234

Gilar-Corbi, R., Pozo-Rico, T., Castejón, J. L., Sánchez, T., Sandoval-Palis, I., & Vidal, J. (2020). Academic achievement and failure in university studies: Motivational and emotional factors. *Sustainability, 12*(23), 1-14. https://doi.org/10.3390/su12239798

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V., utica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. (2018). Predicting academic performance: A systematic literature review. In B. Scharlau & G. Rossling (Eds.), *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*. Association for Computing Machinery. https://doi.org/10.1145/3293881.3295783

Huang, S., & Fang, N. (2013). Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models. *Computers and Education*, *61*, 133-145 https://doi.org/10.1016/j.compedu.2012.08.015

Jayaprakash, S., & Jaiganesh, V. (2019) A conceptual framework to predict academic performance of students using classification algorithm. *International Research Journal of Engineering and Technology, 6*(9), 721-733.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and Information Technologies*, *13*(1), 61-72. https://doi.org/10.2478/cait-2013-0006

Kalmegh, S. (2015). Analysis of WEKA data mining algorithm REPTree, simple cart and randomtree for classification of Indian news. *International Journal of Innovative Science, Engineering & Technology*, *2*(2), 438-446. http://ijiset.com/vol2/v2s2/IJISET_V2_I2_63.pdf

Kemper, L., Vorhoff, G., & Wigger, B. U. (2020). Predicting student dropout: A machine learning approach. *European Journal of Higher Education*, *10*(1), 28-47. https://doi.org/10.1080/21568235.2020.1718520

Khasanah, A. U., & Harwati, H. (2017). A comparative study to predict student's performance using educational data mining techniques. *IOP Conference Series: Materials Science and Engineering, 215*(1). https://doi.org/10.1088/1757-899X/215/1/012036

Kotsiantis, S. (2009). Educational data mining: A case study for predicting dropout-prone students. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *1*(2), 101. https://doi.org/10.1504/IJKESDP.2009.022718

Kovačić, Z. J. (2010, June). Early prediction of student success: Mining students enrolment data. *Proceedings of the 2010 Informing Science and Information Technology Education Conference*, *Cassino, Italy,* 647-665. https://proceedings.informingscience.org/InSITE2010/InSITE10p647-665Kovacic873.pdf

McMullen, M. (2014). The value and attributes of an effective preparatory English program: Perceptions of Saudi University students. *English Language Teaching*, *7*(7). https://doi.org/10.5539/elt.v7n7p131

Nielsen, T. D., & Jensen, F. V. (2007). *Bayesian networks and decision graphs* (2nd ed.). Springer.

Osmanbegović, E., & Suljic, M. (2012). Data mining approach for predicting student performance. *Journal of Economics and Business*, *10*(1), 3-12.

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), 1560-1565.

Pradeep, A., & Thomas, J. (2015). Predicting college students dropout using EDM techniques. *International Journal of Computer Applications*, *123*(5), 26-34. https://doi.org/10.5120/ijca2015905328

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *ScienceDirect*, *33*(1), 135-146. https://doi.org/10.1016/j.eswa.2006.04.005

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, *40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., & Ventura, S. (2013). Data mining in education. *WIREs Data Mining and Knowledge Discovery*, *3*(1), 12-27. https://doi.org/10.1002/widm.1075

Shakeel, K., & Butt, N. (2015, May). Educational data mining to reduce student dropout rate by using classification. *Proceedings of the 253rd OMICS International Conference on Big Data Analysis & Data Mining*, *Lexington, Kentucky.*

Szafran, R. F., & Austin, S. F. (2002). The effect of academic load on success for new college students: Is lighter better? *NACADA Journal*, *22*(2), 26-38. https://doi.org/10.12930/0271-9517-22.2.26

Thai-Nghe, N., Janecek, P., & Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. *37th Annual Frontiers in Education Conference - Global Engineering: Knowledge without Borders, Opportunities without Passports*, T2G-7-T2G-12. https://doi.org/10.1109/FIE.2007.4417993

Villwock, R., Appio, A., & Andreta, A. A. (2015). Educational data mining with focus on dropout rates. *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. https://doi.org/10.1002/j.2168-9830.2009.tb01035.x

Witten, I. H., Frank, E., Hall, M. A., & Pal., C. J. (2017). Part II: More advanced machine learning schemes, *Data mining: Practical machine learning tools and techniques* (4th ed, pp. 205-208). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-804291-5.00021-0

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.

Yassein, N. A., Gaffer, R., Helali, M., & Mohomad, S. B. (2017). Predicting student academic performance in KSA using data mining techniques. *Journal of Information Technology & Software Engineering, 7*(5). https://doi.org/10.4172/2165-7866.1000213

Yehuala, M. A. (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*(4), 91-94.

Yohannes, E., & Ahmed, S. (2018). Prediction of student academic performance using neural network, linear regression and support vector regression: A case study. *International Journal of Computer Applications*, *180*(40), 39-47. https://doi.org/10.5120/ijca2018917057

AUTHORS

Sarah Alturki is currently a Ph.D. student working with the Data and Web Science group at the School of Business Informatics and Mathematics at the University of Mannheim, Germany. She earned her Master's degree in Applied Computer Science from SRH University of Applied Sciences in Heidelberg, Germany. She is currently focusing on projects and research related to empowering technology in higher education.

Nazik Alturki is an Assistant Professor at the Information Systems Department, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Saudi Arabia. She earned her Ph.D. degree in Information Systems from the University of Melbourne, Australia. Her research interests include Health Informatics, Big Data, Data Analytics, and Data Mining.