



Predicting Student Performance to Improve Academic Advising Using the Random Forest Algorithm

Mirna Nachouki, Department of Information Technology, Ajman University, UAE*

 <https://orcid.org/0000-0003-3490-0878>

Mahmoud About Naaj, Department of Information Technology, Ajman University, UAE

 <https://orcid.org/0000-0001-8440-5889>

ABSTRACT

The COVID-19 pandemic constrained higher education institutions to switch to online teaching, which led to major changes in students' learning behavior, affecting their overall performance. Thus, students' academic performance needs to be meticulously monitored to help institutions identify students at risk of academic failure, preventing them from dropping out of the program or graduating late. This paper proposes a CGPA predicting model (CPM) that detects poor academic performance by predicting their graduation cumulative grade point average (CGPA). The proposed model uses a two-layer process that provides students with an estimated final CGPA, given their progress in second- and third-year courses. This work allows academic advisors to make suitable remedial arrangements to improve students' academic performance. Through extensive simulations on a data set related to students registered in an undergraduate information technology program gathered over the years, the authors demonstrate that the CPM attains accurate performance predictions compared to benchmark methods.

KEYWORDS

Academic Second-and-Third-Year Grades, CGPA Prediction, Educational Data Mining

INTRODUCTION

Student success is one of the main goals of educational institutions. It is measured by academic performance, or the extent to which students meet the standards defined by these institutions. According to Tuckman (1975), student performance refers to observing student knowledge, skills, concepts, understanding, and ideas. In the higher education environment, academic student performance depends on instructors' and program coordinators' standards and reflects the achievement of their short- and long-term educational goals (Sundar, 2013).

Prediction of accurate academic student performance in universities is considered an important tool that helps in various decisions related to student admission, retention, graduation, and adapted educational support based on student data observation. Student performance is a significant indicator in measuring institutions' effectiveness and a crucial factor in students' future success, particularly in

DOI: 10.4018/IJDET.296702

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

countries' prosperity. For this reason, higher education institutions focus today on improving student performance and enhancing the quality of their educational programs. An in-depth analysis of the learners' previous records can play a vital role in providing quality education to learners.

Early prediction of academic student performance helps institutions provide appropriate actions to improve students' retention and success rates. Educational data mining (EDM) involves analysis and improvement in the prediction methods of student performance. With EDM techniques, researchers can develop prediction models to detect, monitor, and improve student achievement (Alyahyan & Düşteğör, 2020).

Predicting academic student performance may also improve curriculum content and plan for adequate academic advising for students. Data mining techniques allow researchers to examine data sets and obtain conclusions that help improve the educational learning process. Various techniques have been applied for this purpose. Machine learning, collaborative filtering, Bayesian networks, artificial neural networks, random forest decision trees, rule-based systems, and correlation analysis have been applied to predict the risk of dropping out of the university, students' achievement, or grades. All these techniques classify the significant factors that affect and foresee overall student academic performance. However, they differ in precision/accuracy, complexity, and sample data size requirements.

This paper focuses on developing a prediction model of students' academic performance based on their high school average score and second and third-year grades in a four-year information technology program. It explores the performance of random forest (RF) machine learning in predicting student performance to achieve high predicting accuracy. The proposed methodology proved its worth by achieving accurate results. The result can be used by students, advisors, and program coordinators to reduce education difficulties, improve students' results, provide better quality education, and develop plans for education policy.

BACKGROUND

Every year, higher education institutions collect large amounts of student data that could be transformed into knowledge, which can help instructors, program coordinators, and policymakers analyze and make adequate decisions. It can also provide timely information to different stakeholders that enhance the quality of their educational processes. Early student performance prediction can help universities provide timely actions, like planning for appropriate training to improve students' learning experience, thus improving their success rates. In addition, detecting at-risk students early would provide more time for them to improve their performance (Riestra-González et al., 2021).

Academic performance analysis has gained popularity in the past 20 years. Researchers used various prediction and classification methods to provide clues to help students improve their performance and assist educational institutions in improving quality and making better administrative decisions.

Using Data Mining Techniques to Predict Academic Students' Success

Data mining (DM) is often used to predict academic students' performance. It is usually used by

- Program coordinators, to improve curriculum design.
- Academic advisors, to guide their students in choosing the right courses to register for and to improve their performance in the coming semesters (Han et al., 2010).
- Universities, to improve students' success rate, reduce retention, and develop future plans for education policy.

EDM techniques are currently used in various educational situations to measure students' grades, understanding, and performance based on their behavior and attitude. EDM uses DM techniques to

analyze and produce knowledge from an educational data framework (Romero & Ventura, 2007). EDM is used by different groups of users who can manipulate the extracted knowledge based on their vision and objectives. For example, instructors can utilize the produced knowledge to improve their teaching methodologies (Zorrilla et al., 2005). In addition, program coordinators can make better decisions to enhance learning outcomes (Kotsiantis et al., 2010).

The EDM approach has attracted many researchers to analyze related academic data sets in the context of high school and university students and then extract relevant information to assist academic advisors in understanding how students acquire their knowledge and skills and enhance their learning outcomes (Saa et al., 2019).

Classification is an important DM technique that extracts knowledge from related academic data sets to classify and predict students' performance (Chung & Lee, 2019). It consists of discovering a model categorizing data into classes to predict the class of not yet defined entities. It involves two phases. In the training phase, a classification algorithm constructs a model by learning from a data set (training data) and its associated class features—named class label attributes. Then, the previously built model is applied to classify test data independent of the training data in the test phase.

To predict students' graduation performance, researchers have applied various classification algorithms. They concluded that no single classifier technique might be applied in all situations and provide acceptable prediction results (Asif et al., 2017). Moreover, researchers applied different strategies, such as decision trees, neural networks, naïve Bayes, k-nearest neighbor, and various others, to explore the best classification techniques suitable for analyzing specific kinds of data.

Random forest (RF) was considered among the best DM techniques used to predict academic students' performance (Rao et al., 2016). This technique has also shown excellent results to predict high school students' attrition using several binary classification attributes (Chung & Lee, 2019). Hussain et al. (2018) presented a similar study using data mining tools and techniques to enhance students' academic performance and prevent dropout. Four classification algorithms—J48 decision tree, Bayes network, and RF—were used. This study used students' demographic, socioeconomic, and academic data to perform the prediction. The findings showed that RF outperformed the other classifiers based on accuracy and classifier errors. Abubakar & Ahmad (2017) predicted students' performance based on the following prediction features: (1) students' interaction in an e-learning environment, (2) their grades, and (3) their prerequisite knowledge. Again, the authors used the RF algorithm to predict student performance. They concluded that this technique outperformed other popular data mining techniques, such as decision trees and k-nearest neighbor.

Xu et al. (2017) developed a bilayered structure comprising multiple base predictors and a cascade of an ensemble predictor for making predictions based on students' evolving performance states. A data-driven approach was used in discovering course relevance based on latent factor models and probabilistic matrix factorization. They showed that the proposed method achieves performance superior to the benchmark approaches. Furthermore, among the implemented base predictors, they found that random forest performs the best and k-nearest neighbor performs the worst in most cases.

Hasan et al. (2020) concluded that RF provided a high classification accuracy rate and outperformed the other algorithms applied (with an accuracy of 88.3%) in predicting successful students. Rao et al. (2016) compared the performance of the following three algorithms: J48 decision tree, naïve Bayes, and RF. The findings showed that RF provided better accuracy, especially with large data set size, though it needed much time to construct the model and provide the results.

Adekitan & Salau (2019) applied a data mining approach to predict students' final graduation CGPA using the GPA students got every semester of their first three years of study. They have applied six different data mining algorithms on the same data set for comparative performance analysis. According to their results, the logistic regression algorithm produced the best accuracy (89.15%), followed by tree ensemble (87.884%), decision tree (87.85%), and random forest (87.70%), then naïve Bayes (86.438%), and finally probabilistic neural network (85.895%). Thus, they have concluded

that engineering students' final CGPA can be predicted based on their performance in the first six academic semesters.

Influential Factors in Predicting Students' Performance

Sixty-nine percent of the studies addressing the prediction of academic performance have cited two main features: students' demographics (such as age, gender, race/ethnicity, annual family income, disability, and parents' background) and student high school average (Shahiri & Husain 2015; Almarabeh, 2017; Hamoud et al., 2018; Mengash, 2020; Alturki et al., 2020; Fahd et al., 2021). A literature review by Alyahyan and Düşteğör (2020) revealed that prior academic achievement is the most crucial factor cited in more than 40% of research. Shahiri and Husain (2015) and Garg (2018) stated that the high school average is considered the most critical factor that impacts students' performance reliability. Moreover, this factor influences the prediction of many other different topics (Aluko et al., 2018). Other factors, such as students' marks obtained by various course assessment tools (Almarabeh, 2017; Hamoud et al., 2018), their GPA and CGPA (Ahmad et al., 2015; Hamoud et al., 2018) as well as their attendance (Almarabeh, 2017) have also been cited in as having a significant impact on the prediction of academic students' achievement.

Several researchers applied various machine learning methods to predict student academic performance at the end of the study program. However, they could not detect which students may need immediate attention to lower the chances of failing or getting low grades in courses.

The need for this study emerged during the COVID-19 pandemic, when students could not meet face-to-face with their advisors to discuss issues concerning their academic performance and possible delay in their graduation. This study contributes to the current body of literature by identifying low-performing and struggling students by predicting their CGPA at the end of the second year and third year, and applying various teaching and learning strategies and supplementary measures to improve their performance.

The high school average score was used in this work as a substitution for demographic and behavioral factors as many studies have suggested that there is a correlation between high school performance and non-academic factors (Aluko et al., 2018; Shahiri & Husain 2015; Garg, 2018).

PURPOSE OF THE STUDY

This work aims to investigate whether it is possible to obtain an adequate accuracy in predicting students' academic performance based on their high school score average and second- and third-years grades in a four-year information technology program. The objectives of this study are to:

1. Identify attributes that influence students' academic performance.
2. Develop a valid model for predicting students' academic performance.
3. Identify, at early stages, students with low performance.
4. Categorize pillar courses that have a significant impact on students' academic journey.

SIGNIFICANCE OF THE STUDY

This study is designed to provide academic institutions with a model that will provide decision-makers with pertinent data to enhance students' retention and graduation rates. More specifically, it will help:

- Department heads in identifying, at early stages, students at risk of dropping out from the program, specifically at the end of the second year.
- Advisors in providing dedicated guidance to students with poor academic performance; for example, recommending students to register elective courses to get them more engaged in the program.

- Program coordinators in designing pertinent syllabi for all courses that are considered as pillars in the program.
- Educators in developing adequate workshops for students to motivate them to improve their academic performance.

METHODOLOGY

Predicting students' performance is an essential factor in students' academic success. However, students enrolled in the same majors can have different backgrounds, resulting in different sequences in which they prefer to register for courses. Moreover, not all the courses taken by students are equally weighted to predict their future performance. For this reason, we will consider IT courses in the students' major to decrease the complexity of the problem and reduce the noise that may arise in the prediction process.

This work examines the use of a random forest algorithm to predict students' graduation CGPA. In addition, it uses grades of the second- and third-year courses to measure students' performance.

Random Forest Algorithm

The random forest algorithm is considered one of the classification learning algorithms that deploy collections of lower-level classifiers to train and predict autonomous data sets (Gollapudi, 2016). The algorithm finds a result based on ensemble methods that classify estimators' results for each predicted statement. The algorithm finds a result based on ensemble methods that classify estimators' results for each predicted statement and uses this result to provide high-level results for distinct fundamental classifiers. Combining all the results is expected to provide better performance.

A random forest algorithm consists of creating several decision trees. The more decision trees the algorithm produces, the more accurate the result is. This algorithm creates decision trees on randomly selected data samples, generates a prediction from each tree, and then uses the voting technique to select the best solution and the final class of the test data sets (Lee & Chung, 2019). The random forest algorithm uses a nonlinear technique to discover interrelationships between the various attributes involved in the problem. The random forest technique has been used to examine several interesting problems and produce applicable classification models. Applications using this technique are found in Rodriguez-Galliano et al. (2012). In addition, several researchers have compared the random forest technique's performance to other data mining techniques in different problems (Verikas et al., 2011).

Predicting students' academic performance consists of a typical classification task—the most commonly applied technique in the educational data mining field (Shahiri & Husain, 2015). Constructing decision trees for learning and prediction was found to be the most successful technique in classification-based problems. Random forest or random decision trees are suitable for solving such problems because a multitude of decision trees will be constructed. Then the selected output class will be based on the class that they more often predict. Moreover, the RF technique provides accurate predictions and is widely used in predicting students' performance in the education sector. This technique does not prune trees like other tree-based algorithms. Instead, at each tree node, splitting is considered for a random subset of features, resulting in features being split into more and smaller random subsets, increasing the diversity among the forest of trees, leading to its outperformance compared to other decisions trees-based algorithms (Fahd et al., 2021).

Technical Specification

The CGPA predicting model (CPM) is implemented using the Anaconda 3 distribution of Python, using the scikit-learn and pandas (<https://pandas.pydata.org/>) libraries, which include a vast array of functionality for all aspects of data mining and machine learning experimentation and research. Specifically, the *RandomForestClassifier()* comprised the core scikit-learn modules used for the classifier. More specifically, the following Python functions were used:

- *describe()* computes a summary of statistics relating to the features to identify anomalies. This function gives the mean, standard deviation, minimum, maximum, and lower and upper percentiles. The lower percentile is 25, and the upper percentile is 75. The 50th percentile is the same as the median.
- *train_test_split()* splits data. As a starting point, we split the data set into training and testing at a ratio of 70:30. Seventy percent of the data set is applied to train the model, study the relationship between the features and the target, and fit the model. Secondly, we split the training data into training and validation at a ratio of 40:30. The purpose of the validation set is to estimate the prediction error for the selected data. Finally, the test set is used to assess the generalization error of the model in order to output the corresponding GPA target.
- *RandomForestClassifier()* instantiates a model with the following parameters: `random_state = 5`, `n_estimators = 100`, `criterion: gini`, `min_sample_split = 2`, `min_sample_leaf = 1`, and `max_features = number of features`.
- *fit()* trains the model on training data.
- *predict()* uses the model on the test data.

CPM Applicability

The CPM was tested on the educational Kaggle data set (Samin, 2020). This data set contains 525 rows related to students' marks in seven different courses (English, botany, zoology, physics, chemistry, mathematics, and ICT) to predict students' GPA. First, 70% of the data set (367 records) was applied to train and validate the model using k-fold ($k = 5$) cross-validation and study the relationship between the features (courses) and the target (GPA). The model was then tested by making predictions on a testing set of 158 records (30%) to output the corresponding GPA target.

It was found that the root mean square error (RMSE) provided by the CPM for GPA predictions is 0.2723, the mean absolute error is 0.21, and the mean squared error is 7.41, giving 94.29% accuracy of predicting graduation GPA. These results show that the CPM achieves high prediction on student performance.

Data Set

Data related to students who have graduated with a four-year bachelor's degree in information technology (IT) were used in this study. Data were collected during the academic years 2020–21 and 2021–22 of a private university in the United Arab Emirates (UAE). Only data of students who have completed the graduation requirements were used to train, validate, and test the model. The population consists of 105 graduates, of which 73.3% male and 26.7% female, coming from 24 different nationalities, mainly from the Middle East region. All students admitted in the program have at least a 70% average score in a scientific high school certificate or equivalent. Seventy-four records were allocated from training using k-fold cross-validation ($k = 5$), and the remaining were held out for testing.

The first-year courses consist of general education requirements and introductory IT courses, which do not reflect students' actual performance in an IT program. The program's core courses start in the second year. Therefore, we have decided to consider students' grades related to the second- and third-year core IT courses to predict students' graduation CGPA in the suggested model, in addition to their high school average.

Predicting students' performance at the end of the second year is important to detect, at an early stage, those students who are at the borderline and need dedicated academic support to succeed in their educational journey. Third-year prediction is also important as, at this level, students are mature enough to assume their responsibility and take tangible steps towards enhancing their CGPA a year before graduation.

The complete list of features considered in the CPM is detailed in Table 1. Next, the CPM examines the targets (CGPA) to study the relationships between the features and these targets. Next, it learns how to predict the CGPA from the features by constructing 100 decision trees. It is then tested by making predictions on a testing set of data with access to the features (without the targets). Finally, the predictions are compared to the known answers (targets) to determine how far our CGPA predictions are from the actual values.

Table 1. Features used in the model

High School Grade Average
Second-year IT courses:
INT201 Object-Oriented Programming
INT202 Discrete Mathematics
INT203 Computer Organization
INT204 Data Structures
INT205 Fundamentals of Data Communications and Networking
INT206 Fundamentals of Web Systems
Third-year IT courses:
INT301 Operating Systems
INT302 Database Management Systems
INT303 Fundamentals of Information Security
INT304 Human–Computer Interaction
INT305 Software Engineering
INT306 Computer Ethics and Professional Development
INT307 IT Project Management
INT308 Enterprise Systems
INT311 Advanced Computer Networks
INT312 Network Security

This process has been executed twice: first with students' high school average and IT courses related to the second-year level and a second time with students' high school average and the second- and third-year IT courses. We can compare these predictions to the actual value to judge how accurate the model is.

RESULTS AND ANALYSIS

The CPM was used to predict the CGPA of 30% of the 105 already-graduated students. Table 2 summarizes statistics relating to the mean, standard deviation (std), min, max, and lower (25%) and upper (75%) percentiles. These numbers demonstrate that no data values seem anomalous, and no zeros appear in any measurement columns.

We have calculated Pearson correlation and significance one-tailed test of the outcome (CGPA) and all the variables listed in Table 1. It was found that r (Pearson correlation) is between .351 and .780. Furthermore, these results show a positive correlation between the outcome (CGPA) and all variables at a p -value < 0.05 (Table 3).

Table 2. Summary statistics related to the features

Features	Mean	Std	Min	25%	50% (Median)	75%	Max
High School Grade Average	81.62	9.11	56	76	82	88	99
Object-Oriented Programming	79.89	10.87	63	73	77	87	95
Discrete Mathematics	77.24	10.3	63	67	77	87	95
Computer Organization	77.6	11.04	63	67	77	87	95
Data Structures	76.4	10.39	63	67	73	83	95
Fundamentals of Data Communications and Networking	80.16	9.74	63	73	77	87	95
Fundamentals of Web Systems	82.73	9.19	63	77	83	95	95
Operating Systems	78.33	9.52	63	73	77	83	95
Database Management Systems	80.23	9.39	63	73	77	87	95
Fundamentals of Information Security	81.3	9.92	63	73	83	87	95
Human-Computer Interaction	81.19	8.39	63	77	83	87	95
Software Engineering	78.35	9.75	63	73	77	87	95
Computer Ethics and Professional Development	81.91	8.86	63	73	83	87	95
IT Project Management	78.8	10.23	63	73	77	87	95
Enterprise Systems	83.53	9.93	63	73	83	95	95
Advanced Computer Networks	77.89	10.27	63	73	77	87	95
Network Security	79.72	9.75	63	73	77	87	95

Predicting Students' Performance Using Second-Year Course Grades

The CPM was first applied to predict the CGPA of graduated students to test its performance. Figure 1 shows the actual and predicted CGPA for each student with regard to their second-year courses grades. In this figure, the x-axis represents the number of actual and predicted CGPA; the y-axis represents students' CGPA ranging between 2.0 and 4.0. In order to obtain these answers, our CPM uses seven features:

- Students' performance in high school
- Their grades in
 - Object-Oriented Programming
 - Discrete Mathematics
 - Computer Organization
 - Data Structures
 - Fundamentals of Networking & Data Communication
 - Fundamentals of Web Systems

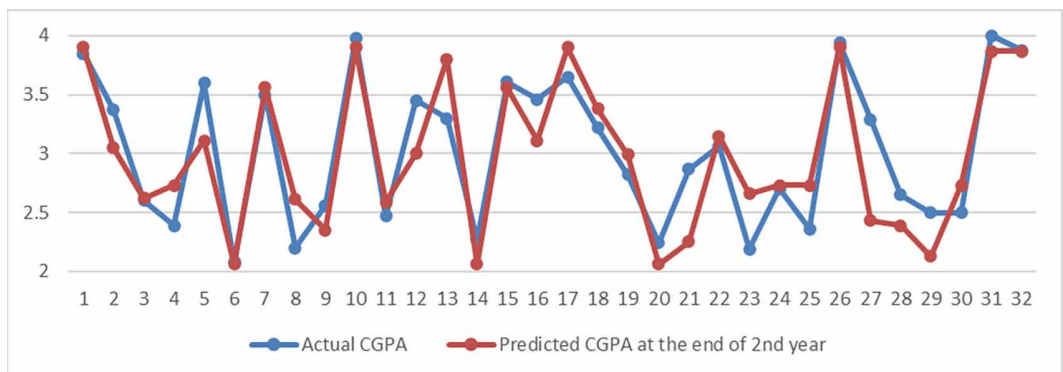
The RMSE of the CPM's predictions at the end of the second year is 0.3179, the mean absolute error is 0.247, and the mean squared error is 10.11, giving 91.32% accuracy of predicting graduation CGPA.

Figure 2 shows that students' high school average is the best predictor of their performance at the end of their second year. The second most important factor is the course Discrete Mathematics. As a prerequisite to Data Structures, the course Object-Oriented Programming does not highly contribute to

Table 3. Pearson Correlation and Significance One-Tailed Test of the CGPA and all variables

	Pearson Correlation	p
	CGPA	CGPA
CGPA	1.000	.
High School score	.351	.000
INT201	.621	.000
INT202	.706	.000
INT203	.752	.000
INT204	.664	.000
INT205	.617	.000
INT206	.622	.000
INT301	.612	.000
INT302	.716	.000
INT303	.673	.000
INT304	.630	.000
INT305	.754	.000
INT306	.592	.000
INT307	.702	.000
INT308	.668	.000
INT311	.739	.000
INT312	.780	.000

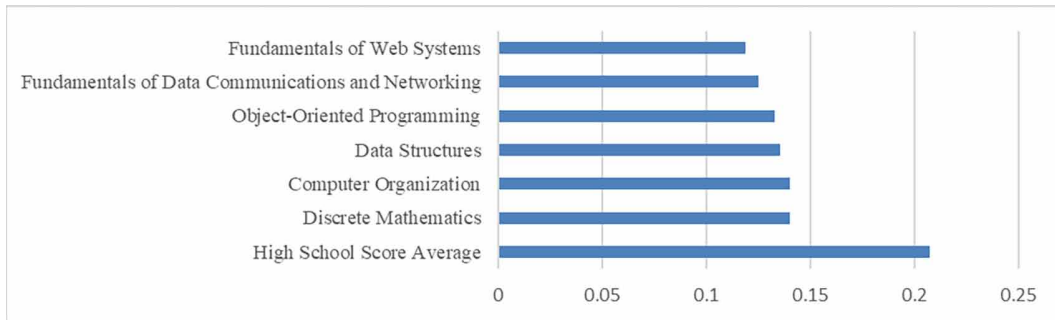
Figure 1. Actual and predicted CGPA based on selected features



the prediction of students' performance, and the accurate level of students' programming knowledge and skills is measured after the completion of Data Structures. Finally, this figure shows that the course Fundamentals of Web Systems is less critical in predicting students' final CGPA.

In the random forest algorithm, feature importance is calculated by measuring the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated

Figure 2. The predictive power of selected features in ascending order of importance

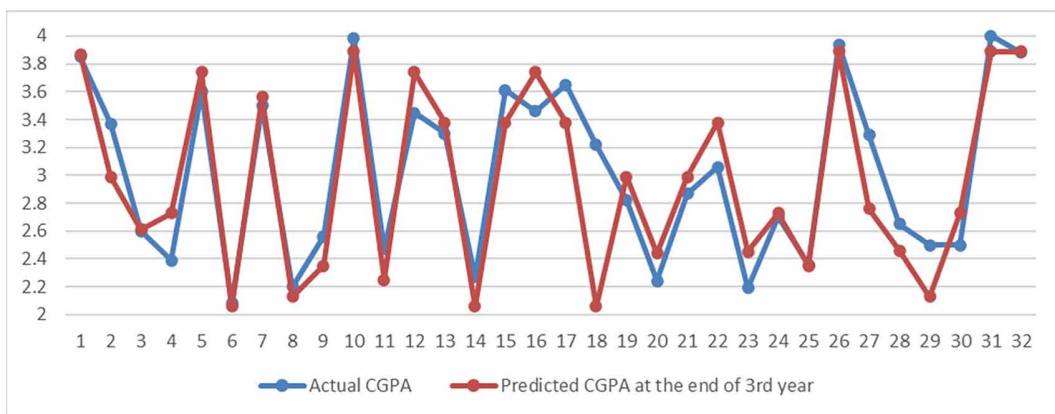


by the number of samples that reach the node, divided by the total number of samples. The higher the value, the more important the feature.

Predicting Students' Performance Using Third-Year Course Grades

To test our CPM's performance in another situation, we have applied it to predict the CGPA of graduated students to a broader set of IT core courses. This set of courses are offered in the second and the third year. Figure 3 shows the actual and predicted CGPA for each student at the end of the third year. In this current situation, the CPM uses 17 features: students' performance in high school, and their grades in the following courses: Object-Oriented Programming, Discrete Mathematics, Computer Organization, Data Structures, Fundamentals of Data Communication & Networking, Fundamentals of Web Systems, Database Management Systems, Software Engineering, Operating Systems, Fundamentals of Information Security, Human-Computer Interaction, IT Project Management, Enterprise Systems, Computer Ethics & Professional Development, Advanced Computer Networks, and Network Security.

Figure 3. Actual and predicted CGPA based on all 17 features at the end of the third year



Accuracy and Errors Measures

The RMSE of the CPM's second run predictions is 0.2984, the mean absolute error is 0.208, and the mean squared error is 8.9 giving 92.87% accuracy of predicting graduation CGPA.

Figure 4. The predictive power of all the features in ascending order of importance

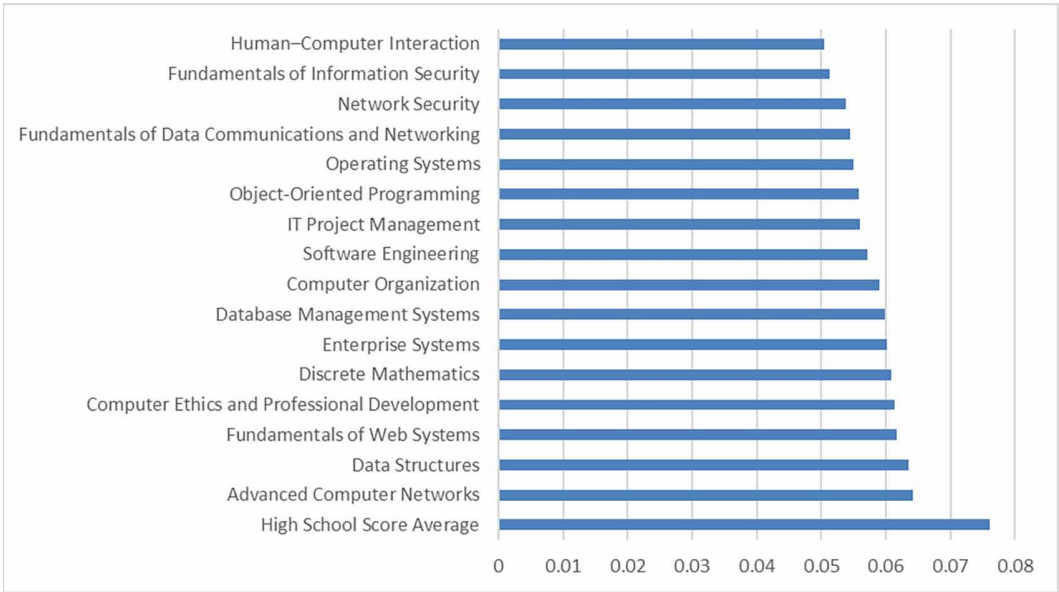
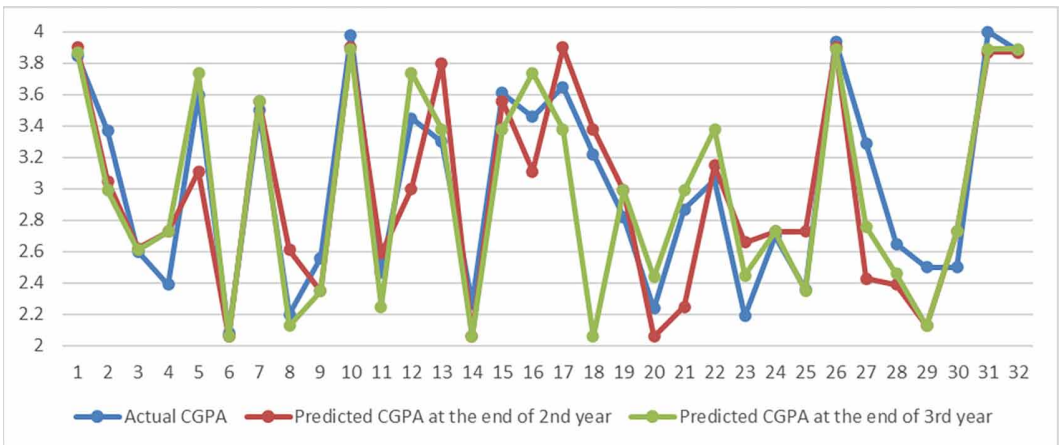


Figure 4 shows that the high school grade average maintains the best predictor of students' performance at the end of their third year.

These two experiments show that the model can indicate students' final performance early, allowing the advisor to take appropriate actions early.

Figure 5. Actual and predicted CGPA in both experiments



As per the IT program study plan, Discrete Mathematics, Computer Organization, and Object-Oriented Programming are offered in the third semester. Data Structures, Fundamentals of Data Communications, and Fundamentals of Web Systems are offered in the fourth semester. It is worth mentioning that based on both experiments, the order of importance of these courses in both semesters

is the same (Table 4 (a) and (b); Table 4 (c) and (d)). This finding indicates that the CPM is reliable. Moreover, in the second experiment, all performance values are improved, as shown in Table 5 and Figure 5.

Table 4. (a – d) Comparative features importance in the two experiments

Third-Semester Features' Importance in the First Experiment		Third-Semester Features' Importance in the Second Experiment	
Discrete Mathematics	0.140193	Discrete Mathematics	0.060794
Computer Organization	0.139934	Computer Organization	0.058912
Object-Oriented Programming	0.133045	Object-Oriented Programming	0.055773
(a)		(b)	
Fourth semester Features Importance in the First Experiment		Fourth semester Features Importance in the Second Experiment	
Data Structures	0.135728	Data Structures	0.063488
Fundamentals of Data Communications and Networking	0.125054	Fundamentals of Web Systems	0.061614
Fundamentals of Web Systems	0.118749	Fundamentals of Data Communications and Networking	0.054391
(c)		(d)	

Table 5. Comparative performance of the two experiments

	First Experiment Performance	Second Experiment Performance
Root Mean Square Error	0.3179	0.2984
Mean Absolute Error	0.247	0.208
Mean Squared Error	10.11	8.9
Accuracy	91.32%	92.87%

Paired Sample T-Test Between Actual and Predicted CGPA

We conducted a paired sample (dependent) t-test to compare actual students' CGPA with the predicted CGPA. Table 6 presents the results of the test.

Hypothesis: No significant difference between students' actual and predicted CGPA.

The test results showed that actual students' CGPA and predicted CGPA scores were strongly positively correlated ($r = 0.892$). Moreover, the difference was not statistically significant between actual CGPA ($M = 3.0175$, $SD = 0.62071$) and predicted CGPA scores ($M = 2.9668$, $SD = .65590$); ($t_{31} = .922$, $p > 0.05$). On average, actual CGPA scores were 0.0507 points higher than predicted CGPA scores (95% Confidence Interval is: $-.05910, .15660$) Thus, the hypothesis was accepted, meaning no significant difference exists between students' actual and predicted CGPA.

Table 6. Paired sample t-test

Paired Samples Statistics		Mean	N	Std. Deviation	Std. Error Mean	
Pair 1	Actual	3.0175	32	0.62071	.10973	
	Predicted	2.9688	32	.65590	.11595	

Paired Samples Correlations		N	Correlation	Sig.
Pair 1	Actual & Predicted	32	.892	.000

Paired Samples Tests		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
					Lower	Upper			
Pair 1	Actual - Predicted	.04875	.29913	.05288	-.05910	.15660	.922	31	.364

Paired Samples Effect Sizes		Standardizer*	Point Estimate	95% Confidence Interval of the Difference		
				Lower	Upper	
Pair 1	Actual - Predicted	Cohen's d	.29913	.163	-.187	.511
		Hedges' correction	.30281	.161	-.185	.504

DISCUSSION

This proposed CPM found a strong correlation between students' final CGPA and their high school average score, with a total importance of 0.207 and 0.076, as shown in Figures 2 and 4, respectively. Thus, students with higher high school average scores obtain higher final CGPA in their major. This theory is confirmed by Thiele et al. (2016), who found a strong relationship between students' contextual background characteristics (such as school grades, school type, and gender) and their academic performance at university. Therefore, high school average grade provides an adequate indication of students' academic performance once they get into university. The findings of this study can be explained by the fact that, in general, students tend to have the same academic performance throughout their academic journey. In other words, if a student has relatively good grades in high school, he/she would continue to have good grades at the university level. Similarly, students having bad grades in high school are likely to continue getting low performance at university level. Several research results show a robust relationship between high school grades and college performance and retention and graduate rates (Asif et al., 2017; Kotsiantis et al., 2010).

The 91.32% and 92.87% accuracy levels obtained in the previous section show that the CPM can be used as a useful prediction tool to identify students at risk and prevent them from dropping out. This conclusion is also strengthened by the 94.29% accuracy obtained by the CPM when applied to the validation data related to the educational Kaggle data set (Samir, 2020).

Adekitan & Salau (2019) have applied the random forest algorithm to predict engineering students' final CGPA based on their grades in the program's first three years with an accuracy of 87.7%. The results obtained by the CPM reflected that students' final academic performance is

highly related to their high school average. Thus, they represent the strongest predictor of students' final CGPA consistently throughout the undergraduate four-year information technology program. Shahiri & Husain (2015) have also applied the decision trees technique to evaluate and predict students' performance using course grades, student demographics, and high school background with an accuracy level of 73%. Amrieh et al. (2016) have applied random forest techniques to predict students' performance based on behavioral features. The accuracy level obtained is 75.6%.

The CPM results outperformed the findings obtained by the works mentioned above that use similar factors in predicting students' academic performance. Although other studies have considered various factors in predicting students' academic performance, the results found using the CPM are considered satisfactory to provide a significant contribution to department heads, program coordinators, educators, and advisors to guide their students towards improving their marks.

Furthermore, the CPM results reveal that the second-year courses Discrete Mathematics, Computer Organization, and Data Structures have a high impact on the students' final CGPA. This finding allows academic advisors to give personalized guidance to low-performing students and their instructors at the end of the first year, such as providing extra course activities to these students and/or assigning them dedicated assistant instructors to monitor their progress and encourage them to improve their performance. The same applies to Advanced Computer Networks and Computer Ethics, which are identified as the most important subjects among the third-year course offerings. These prompt, appropriate, and personalized actions taken by academic advisors towards low-performing students would improve students' final CGPA and prevent them from receiving an academic warning in their final year of study.

In addition, the CPM results allow the program coordinator to carefully design the content and assessment tools of the courses that are found to have a major effect on students' academic performance.

CONCLUSION

This study discusses factors related to students' academic success or failure in a four-year bachelor of science program in information technology. The findings are expected to provide insight to university stakeholders, and more specifically, to program coordinators, student advisors, and instructors, on the aspects that contribute to students' academic performance. Understanding the various factors that impact and delay students' academic progress is very important; it can help academic institutions design strategies that can improve students' academic achievements and enhance the efficiency of education.

This study suggests an educational model, CPM, to predict students' academic performance, which various decision-makers could use within higher education institutions. The results generated indicate that our CPM can be incorporated in an advising tool to identify students that might drop out of the program. It can also identify students with poor academic performance by predicting their graduation CGPA based on their grades in core courses in a four-year bachelor's degree program and their high school grade average. This prediction provides students estimated final CGPAs, given their progress using second-year courses grades, which will allow early identification of students at risk of dropping out from the program. This prediction is then enforced with third-year courses grades to help academic advisors identify low-performing students and recommend that they register in elective courses to engage them more in the program. Taking a right corrective step will result in improving students' performance, graduation, and retention rates.

The CPM also identifies courses that have an important weight in determining students' final CGPA, which helps the program coordinator design the delivery and assessment of these courses judiciously.

LIMITATIONS

This study revealed the eligibility of the CPM model to efficiently predict students' performance at different levels of their studies using their grades in core program courses and their high school grade average. However, more research is needed to improve the CPM in predicting students with

low performance by considering various factors affecting students' performance, specifically those related to a particular major or a specific institution.

By categorizing and examining these factors, students' academic performance could be improved by designing adequate remedial tutorials to enhance the overall educational process. Furthermore, this study was applied to a small set of data that constitutes the graduates of an IT program; it is worth mentioning that the CPM was first applied to 82 records and provided an accuracy of 88.68% and 92.36% in predicting the final CGPA at the end of the second and third years, respectively. This experiment was then repeated with 105 records; the accuracy of both experiments improved to 91.32% and 92.87%, respectively. With a larger data set, the RF classification performance would increase.

FUTURE WORK

In the future, the model could be easily and quickly adopted by other higher education institutions, with larger data sets based on their needs. In addition, other research directions could be explored, such as:

- Implementing different classification algorithms and comparing the results.
- Improving the classification accuracy by using a different hybrid classification algorithm.
- Using multiple data sets from various disciplines to strengthen the validity of the predictive model.
- Constructing an updated CPM considering demographic and behavioral factors to determine the impact of these factors on student performance.

Moreover, in this study, we have considered the high school average score only to predict students' academic performance. In the future, we may expand this research to include the type of high school certificate (UAE, American, British, French, Indian, etc.).

REFERENCES

- Abubakar, Y., & Ahmad, N. B. H. (2017). Prediction of students' performance in E-learning environment using Random Forest. *International Journal of Innovative Computing*, 7(2).
- 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. doi:10.1016/j.heliyon.2019.e01250 PMID:30886917
- Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The prediction of students' academic performance using classification data mining techniques. *Applied Mathematical Sciences*, 9(129), 6415–6426. doi:10.12988/ams.2015.53289
- Almarabeh, H. (2017). Analysis of students' performance by using different data mining classifiers. *International Journal of Modern Education and Computer Science*, 9(8), 9–15. doi:10.5815/ijmeecs.2017.08.02
- Alturki, S., Hulpuş, I., & Stuckenschmidt, H. (2020). Predicting academic outcomes: A survey from 2007 till 2018. *Technology, Knowledge and Learning*, 1-33.
- Aluko, R. O., Daniel, E. I., Oshodi, O. S., 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*.
- Alyahyan, E., & Düşteğör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17(1), 1–21. doi:10.1186/s41239-020-0177-7
- Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining educational data to predict student's academic performance using ensemble methods. *International Journal of Database Theory and Application*, 9(8), 119–136. doi:10.14257/ijta.2016.9.8.13
- Asif, R., Merceron, A., Ali, S. A., & Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177–194. doi:10.1016/j.compedu.2017.05.007
- Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, 96, 346–353. doi:10.1016/j.childyouth.2018.11.030
- Fahd, K., Miah, S. J., & Ahmed, K. (2021). *Predicting student performance in a blended learning environment using learning management system interaction data*. Applied Computing and Informatics. doi:10.1108/ACI-06-2021-0150
- Garg, R. (2018). Predicting student performance of different regions of Punjab using classification techniques. *International Journal of Advanced Research in Computer Science*, 9(1), 236–241. doi:10.26483/ijarcs.v9i1.5234
- Gollapudi, S. (2016). *Practical machine learning*. Packt Publishing.
- Hamoud, A., Hashim, A. S., & Awadh, W. A. (2018). Predicting student performance in higher education institutions using decision tree analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(2), 26–31. doi:10.9781/ijimai.2018.02.004
- Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufman.
- Hasan, R., Palaniappan, S., Mahmood, S., Abbas, A., Sarker, K. U., & Sattar, M. U. (2020). Predicting student performance in higher educational institutions using video learning analytics and data mining techniques. *Applied Sciences (Basel, Switzerland)*, 10(11), 3894. doi:10.3390/app10113894
- Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2), 447–459. doi:10.11591/ijeecs.v9.i2.pp447-459
- Kotsiantis, S., Patriarcheas, K., & Xenos, M. (2010). A combinational incremental ensemble of classifiers as a technique for predicting students' performance in distance education. *Knowledge-Based Systems*, 23(6), 529–535. doi:10.1016/j.knsys.2010.03.010
- Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. *Applied Sciences (Basel, Switzerland)*, 9(15), 3093. doi:10.3390/app9153093

- Mengash, H. A. (2020). Using data mining techniques to predict student performance to support decision making in university admission systems. *IEEE Access: Practical Innovations, Open Solutions*, 8, 55462–55470. doi:10.1109/ACCESS.2020.2981905
- Rao, K. P., Rao, M. C., & Ramesh, B. (2016). Predicting learning behavior of students using classification techniques. *International Journal of Computers and Applications*, 139(7), 15–19. doi:10.5120/ijca2016909188
- Riestra-González, M., del Puerto Paule-Ruíz, M., & Ortin, F. (2021). Massive LMS log data analysis for the early prediction of course-agnostic student performance. *Computers & Education*, 163, 104–108. doi:10.1016/j.compedu.2020.104108
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104. doi:10.1016/j.isprsjprs.2011.11.002
- Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert Systems with Applications*, 33(1), 135–146. doi:10.1016/j.eswa.2006.04.005
- Saa, A. A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, 24(4), 567–598. doi:10.1007/s10758-019-09408-7
- Samir, T. I. (2020). *HSC level students' GPA dataset*. Kaggle. <https://www.kaggle.com/tahamidulislamsamin/final-training>
- Shahiri, A. M., Husain, W., & Rashid, N. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 72, 414–422. doi:10.1016/j.procs.2015.12.157
- Sundar, P. P. (2013). A comparative study for predicting students academic performance using Bayesian network classifiers. *IOSR Journal of Engineering*.
- Thiele, T., Singleton, A., Pope, D., & Stanistreet, D. (2016). Predicting students' academic performance based on school and socio-demographic characteristics. *Studies in Higher Education*, 41(8), 1424–1446. doi:10.1080/03075709.2014.974528
- Tuckman, H. P. (1975). Teacher effectiveness and student performance. *The Journal of Economic Education*, 7(1), 34–39. doi:10.1080/00220485.1975.10845419
- Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2), 330–349. doi:10.1016/j.patcog.2010.08.011
- Xu, J., Moon, K. H., & Van Der Schaar, M. (2017). A machine learning approach for tracking and predicting student performance in degree programs. *IEEE Journal of Selected Topics in Signal Processing*, 11(5), 742–753. doi:10.1109/JSTSP.2017.2692560
- Zorrilla, M. E., Menasalvas, E., Marin, D., Mora, E., & Segovia, J. (2005, February). Web usage mining project for improving web-based learning sites. In *International conference on computer aided systems theory* (pp. 205–210). Springer. doi:10.1007/11556985_26

Mirna Nachouki, Ph.D., Member of Association for Computing Machinery (ACM), is the Head of the Information Technology Department, College of Engineering and Information Technology, at Ajman University. She has obtained her Ph.D. (1995) and MSc. (1991) in Computer Science from the University of Paul Sabatier, Toulouse III, France and BSc (1990) in Computer Science from the University of Pau et Pays de l'Adour, France. She served as acting dean of the former College of Information Technology. She has more than 22 years of teaching and research experience. Her current research interests include IT Education, Blended Learning, Assessments, Computer Ethics, Cloud Computing, and Educational Data Mining.

Mahmoud Abou Naaj (Member of IEEE, ACM, and ISI), Ph.D., is an Associate Professor at the Department of Information Technology, College of Engineering and Information Technology, at Ajman University. He has obtained his Ph.D. (1983) and MSc. (1979) in Computer Science from the University of Leeds (UK) and BSc (1975) in Mathematics from the University of Baghdad, Iraq. He served as the College of General Studies Dean, dean of the College of Information Technology, and Dean of Admission and Registration at Ajman University. He has more than 35 years of teaching and research experience. His current research interests include IT Education, Blended Learning, Assessments, Computer Ethics, Computer Algorithms, and Educational Data Mining.