applied sciences

MDPI

Article

Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average?

Ahmet Emin Tatar [1,†] and Dilek Düştegör [2,*,†]

[1] Department of Mathematics and Statistics, KFUPM, Dhahran 31261, Saudi Arabia; atatar@kfupm.edu.sa
[2] Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, Dammam 31441, Saudi Arabia
* Correspondence: ddustegor@iau.edu.sa
† These authors contributed equally to this work.

Received: 22 June 2020; Accepted: 16 July 2020; Published: 19 July 2020

Abstract: Predicting the academic standing of a student at the graduation time can be very useful, for example, in helping institutions select among candidates, or in helping potentially weak students in overcoming educational challenges. Most studies use individual course grades to represent college performance, with a recent trend towards using grade point average (GPA) per semester. It is unknown however which of these representations can yield the best predictive power, due to the lack of a comparative study. To answer this question, a case study is conducted that generates two sets of classification models, using respectively individual course grades and GPAs. Comprehensive sets of experiments are conducted, spanning different student data, using several well-known machine learning algorithms, and trying various prediction window sizes. Results show that using course grades yields better accuracy if the prediction is done before the third term, whereas using GPAs achieves better accuracy otherwise. Most importantly, variance analysis on the experiment results reveals interesting insights easily generalizable: individual course grades with short prediction window induces noise, and using GPAs with long prediction window causes over-simplification. The demonstrated analytical approach can be applied to any dataset to determine when to use which college performance representation for enhanced prediction.

Keywords: academic performance; course grades; data mining; grade point average; machine learning; prediction; undergraduate

1. Introduction and Motivation

Educational Data Mining (EDM) is a fast-growing scientific field offering the potential to analyze a variety of student features to harness valuable knowledge from them. To this end, a plethora of predictive algorithms were effectively applied in educational contexts for numerous purposes using a variety of data and student records. As compiled in the review paper [1], two main application purposes can be identified in the college contexts: predictors and early warning systems (EWS). A predictor, "given a specific set of input data, aims to anticipate the outcome of a course or degree" [1], and a EWS "performs the same tasks as a predictor, and reports its findings to a teacher and/or to students at an early enough stage so that measures can be taken to avoid or mitigate potentially negative outcomes" [1]. Common prediction goals are listed as risk of failing a course, dropout risk, grade prediction, and graduation rate.

Among the various prediction goals, prediction of academic performance at graduation time especially, is of tremendous importance, as this information can be useful for:

- Enabling the educational institution to identify students (not) likely to complete the program and help in their admission decisions,
- Identifying students at risk and provide adequate advising and tailored help towards reduced failure rates,
- Detecting high achiever students and help them enhance their career paths,
- Analyzing factors of key importance and mobilize educational efforts for continuous quality improvements.

Looking at the literature on prediction of academic performance at the graduation time, we can observe that all studies rely mainly on four types of information on students, namely: (1) demographics and socio-economic, (2) high-school related, (3) college enrollment, and (4) college performance (up to the time of prediction).

Commonly used demographics and socio-economic information are sex/race [2], household income [3], age, first generation student [4], marital status, parents' jobs and educational levels [2]. Among the high-school related information, high-school GPA [2], pre-college marks [5,6], college admission test scores [3], public or private high-school [2], are frequently observed. As college related, in terms of enrollment information, the major and campus [2], a student's full-time vs. part-time status as well as whether s/he has a scholarship [3], enrolled hours and earned credit hours [4], year of entry and program [2,7] are often used. Finally, we observe that college performance has mostly been represented with grades from courses taken earlier [2,4,6,8], unless the prediction model is meant to be used at admission time [3,8,9].

Based on the above background, we observe that the bulk of previous studies used datasets with relatively large dimensionality of observations, some of them being expensive to measure (when not already available in the records). This, often combined with small samples, caused the curse of dimensionality, potentially yielding models with sub-optimal predictive power.

Recently, there is a trend to use only college performance [8,9], or using courses average per semester instead of individual courses grades (i.e., GPA per semester, or CGPA for cumulative average at time of prediction) [2,4,7]. However, there is no study comparing the performance of EDM models using individual course grades vs. grade point averages. It is unknown whether these two college performance representations are equivalent and can be used interchangeably, or if one is superior to the other in yielding better predictive power.

The main purpose of the present study, therefore, is to elucidate this matter by answering the following research question:

> *Is the individual course grade or grade point average more relevant for predicting student graduation academic performance?*

To answer this question, recent student data compiled at the College of Computer Science and Information Technology (CCSIT) from Imam Abdulrahman bin Faisal University (IAU) are used to generate two sets of predictive models, one using individual course grades, the other using the grade point averages. Thus, predictive power of respective models can be compared. However, it is well known that the performance of such models can also be affected by (1) student data used (besides the academic performance), (2) the data mining technique applied, as well as (3) how far from graduation the prediction is performed. Therefore, a comprehensive set of experiments is designed for spanning the whole search space made of student information besides the college performance, several machine learning methods commonly used in the literature, and prediction window of various sizes.

In the following sections, we first describe the research methodology, including, the dataset description, its preprocessing, the methods used, the experimental setup, and the evaluation criteria. Then, each conducted experiment and its results are reported, followed by the discussion and the concluding remarks.

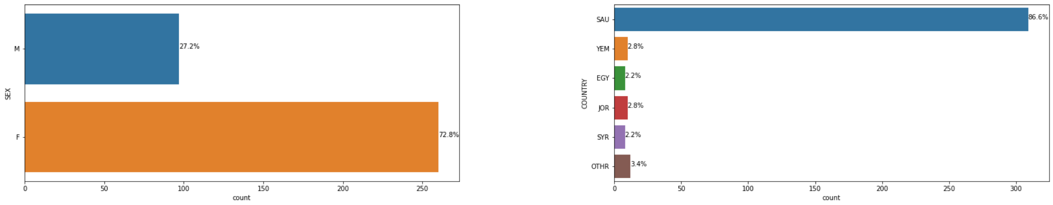
2. Research Methodology

2.1. The Dataset and its Preprocessing

Our dataset contains records of 357 students who were admitted to the CCSIT at IAU from Fall 2011 to Fall 2013 (included), and thus includes three batches of students. The institutional review board at IAU reviewed and approved using the data anonymously (application approved on 19 December 2018; IRB Number: 2018-09-304). Two programs of CCSIT are included in this study, namely Computer Science (CS) and Computer Information Systems (CIS). During the first three years, all the students of the College follow the same plan. In their first year, they attend the Preparatory program where they take mainly intensive English Language courses. In their second and third years, called General Years, the students take courses fundamental to computer and information sciences. At the end of their third year, students select either CS or CIS program based on their interests.

Student records populated from IAU learning management system contain features of three different nature: the demographic features, the pre-college features, and the college records including enrollment information and college performance.

The demographic features consist of gender and nationality (see Figure 1). The female gender dominates the CCSIT as the College is one of the top ranking colleges in the Eastern Region of Saudi Arabia for females. As expected, the dominant nationality is Saudi Arabian with over 85%. The other significant countries represented are Yemen (YEM), Egypt (EGY), Jordan (JOR), and Syria (SYR). There are nine other countries represented — Morocco, Pakistan, Palestine, Ethiopia, India, Iraq, USA, Bahrain, and Sudan —each with less than 0.5% grouped in the OTHR class.



(**a**) Gender Ratio (**b**) Country Ratio

Figure 1. Bar charts showing the demographic information of the dataset.

The demographic features being nominal, we need to convert them into numerical features to use them in machine learning models. We used three different approaches for this purpose.

1. The first approach is "*dummification*". For a nominal feature X with n categories, n new features $\{X_1, \ldots, X_n\}$, called *dummy* features, are created so that if a sample belongs to the i^{th} category, then $X_i = 1$ and $X_j = 0$ for all $1 \leq j \leq n$ and $i \neq j$.
2. The second approach is a derivative of the *dummification* approach. There is a redundancy between the new features $\{X_1, \ldots, X_n\}$. If we know the values of any $n-1$ features, then we can find out the value the missing one. This redundancy may reduce the performance of a machine learning model. Therefore, in this variant of the dummification method, we drop one of the new features, the first one to be exact. We can summarize the dummification approach and its variant as substituting a nominal variable by a binary vector of size n and $n-1$, respectively.
3. In case the dataset has too many nominal features with too many categories each and limited number of samples, the dummification of all the nominal features can reduce the performance of the machine learning algorithm as with the new features the number of total features can increase drastically. To avoid such complications, one can label a category of the nominal feature by its probability. This way, the total number of features do not change. For instance, in our dataset, this approach would have replaced the Female class by 0.728 and the Male class by 0.272.

We experimented with all three approaches. We did not see a major difference in the performance metrics when Logistic Regression or Random Forest machine learning methods are used. However we observed a significant decline in performance when Naive Bayes method is used with either the redundant or the non-redundant dummification which can be explained by the introduction of the new features that are not probabilistically independent. Because of this performance drop, we adopted the third approach in all our experiments.

The pre-college features consists of scores obtained from three national exams. These are numeric scores over 100. The only preprocessing we applied to these features were standardization. (i.e., shifted the mean to 0 and scaled the standard deviation to 1).

The academic records are the third group of features, which contain all transcript information, including admission term, graduation term, and letter grades for all the courses taken per term, for all terms including preparatory year until graduation. We only use the numerical values of the letter grades as described in Table 1. The irregular students, as they are very rare, are not included in this study. Thus, per semester, students take the courses as shown in Table 2 that is prepared based on the degree plan (The actual degree plans can be found at the links [10] for CS program and [11] for CIS program).

Table 1. Conversion table from ordinal to numerical value for letter grades (defined by the university).

Letter Grade	A+	A	B+	B	C+	C	D+	D	F
Numeric Grade	5	4.75	4.5	4	3.5	3	2.5	2	1

Table 2. Courses taken by term by regular CCSIT students during the academic years 2011/2012, 2012/2013, and 2013/2014.

Terms	**Course List**
1st Term	MATH 111, COMP 131, LRSK 141, PHEDU 162
2nd Term	ENGL 101, MATH 112, LRSK 142, COMP 122
3rd Term	CIS 211, CS 211, MATH 211, PHYS 212, ISLM 271
4th Term	CS 221, CS 222, STAT 207, BIOL 222, ISLM 272
5th Term	CIS 313, MATH 301, CS 311, CS 314, CIS 315, ISLM 273
6th Term	CS 310, CS 321, CIS 321, CIS 325, MGMT 290, CIS 413

The target variable in all the models is the graduation GPA, the weighted mean of the numeric scores of all the courses taken by a student. To draw more meaningful results, we use the graduation GPA not as a numerical feature but as an ordinal feature with three categories determined by the university. A student whose graduation GPA out of 5 is greater than or equal to 4.5 belongs to the class *"High GPA"*, between 4.5 and 3.75 (included) to the *"Average GPA"* class, and less than 3.75 to the *"Low GPA"* class. Figure 2 shows the distribution of three classes in the dataset.

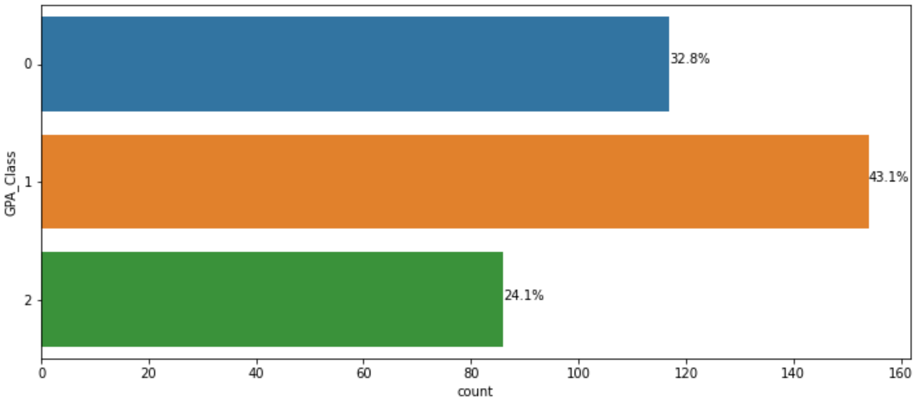


Figure 2. The bar chart showing the three classes in the target variable. The GPA Classes 0, 1, and 2 represent the classes Low GPA, Average GPA, and High GPA, respectively.

2.2. Experimental Set-Up

To answer the research question, we develop several classification models (and not regression, as the target variable has been transformed into an ordinal variable in Section 2.1) that differ with respect to (1) the way college performance is defined, (2) the type of student's data included, (3) the machine learning algorithm applied, (4) the size of the performance window, and (5) the size of the observation window (historical data).

In this study, we define the "term performance" of a student in two different ways. In the first representation, *by courses*, term performance is represented by a vector of size equal to the number of courses that should be taken in that term according to Table 2 with components being the numeric scores of the courses. In the second representation, *by GPA*, we represent term performance by the numeric weighted average of the courses with weights being the credit hours. Comparing the results obtained from the models *by courses* vs the models *by GPA* allows identifying which of the individual course grades and grade averages is more relevant for predicting student's graduation academic performance, thus answer the main research question of this study.

Then, the college performance data for students is modeled using two observation window size. In the first approach, only the immediate past term performance is included in the model (either last term *by courses*, or last term *by GPA*). In the second approach, a cumulative view is adopted where all the past terms' performance is included in the analysis (either cumulative *by courses*, or cumulative *by GPA*). The first approach corresponds to using only one term as history window, while the second approach corresponds to using all past terms data since the student joined the college. The reason to consider models that include last term performance only, is to isolate the term the most impactful to the student's success.

Figure 3 shows a sample student transcript data for the first 6 terms. For instance, let us consider predicting the graduation GPA class at the end of the second year, which is the end of the term 4, if we want to use one term observation history, then we use the term 4 performance alone, either as the term courses which is the vector $[2.5, 2, 4, 2.5, 4]$, or the GPA that is calculated as $(3 \times 2.5 + 4 \times 2 + 3 \times 4 + 4 \times 2.5 + 2 \times 4)/16 = 2.84375$. On the other hand, if we want to use all past observations cumulatively, then we use all past terms performance, either as all past courses, which is to say the vector $[4.5; 4.5; 4; 4; 4; 3.5; 4.75; 4.5; 3.5; 3; 2.5; 2.5; 4.75; 2.5; 2; 4; 2.5; 4; 2; 3; 2; 2.5; 3; 5; 2; 3; 2; 3.5; 3; 4.5]$, or the accumulated GPAs as $[4.3125; 4.1; 3.1; 2.84375]$.

For investigating the impact of the prediction window, we develop six models at different times of the curriculum, (1) as early as by the end of the first semester of the preparatory year, *term 1*, (2) by the end of preparatory year, *term 2*, (3) after the first semester of the general year, *term 3*, (4) by the end of the first general year, *term 4*, (5) after completing the first term of the second common year, *term 5*,

(6) by the end of the common general years, *term 6*. With reference to the student in Figure 3, the above described models correspond respectively to using only the term 1 data, adding one term at a time until all six terms data are used.

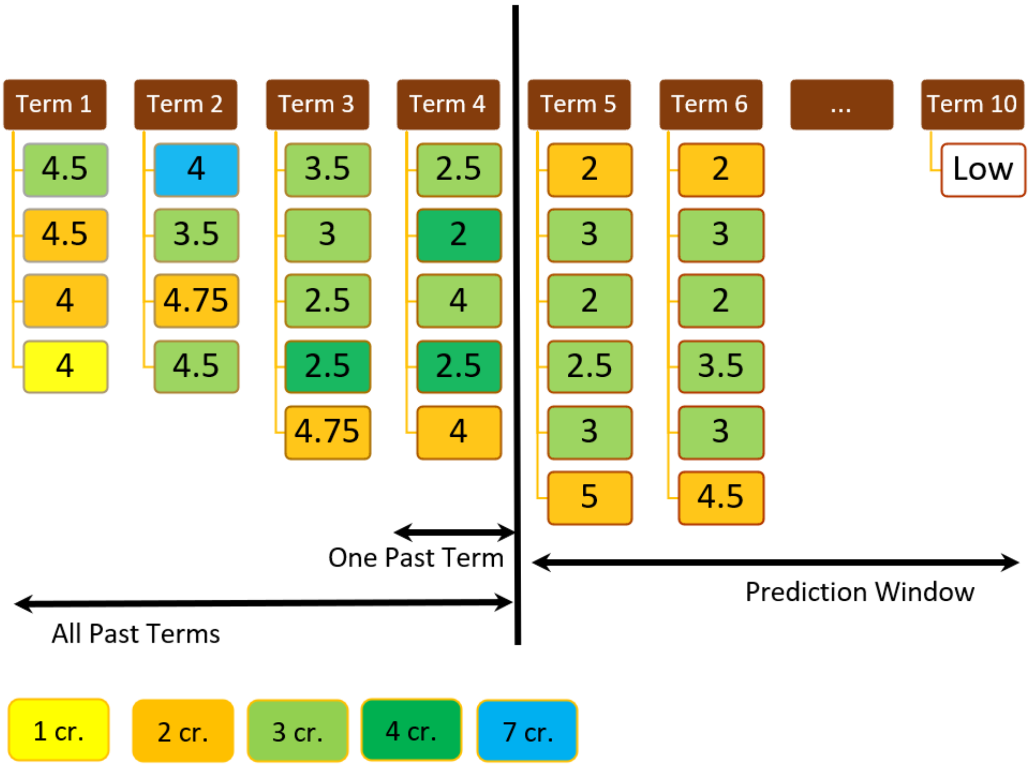


Figure 3. Sampe student transcript data (color code is, yellow: 1, orange: 2, light green: 3, dark green: 4, blue: 7 credits each).

We develop machine learning (ML) models working with the algorithm commonly used in EDM, namely Logistic Regression (LR), Random Forest (RF), and Naive Bayes (NB) with the accuracy as the performance metric. *Logistic Regression* is a linear model used for classification. It is often the first model considered due to its simplicity and interpretability. *Random Forest* is an ensemble method that fits number of decision trees. It makes a prediction based on the average of the predictions from the decision trees which is the method most used to predict graduation performance in the literature as identified in the review paper [12]. *Naive Bayes*, runner up method in the literature [12], is a statistical method based on the Bayes' Theorem. Its performance depends on the statistical independence of the features.

Finally, in order to investigate the impact of set of features on the model performance, we designed four experiments that exclude some features as seen in Table 3. Please note that we excluded all the four experiments which do not include academic records as they are not relevant to the goal of this study.

Table 3. Experiment scenarios (+ indicates inclusion of the feature set).

Terms	Demographics	Pre-College	Academic Records
Scenario 1 (S1)			+
Scenario 2 (S2)	+		+
Scenario 3 (S3)		+	+
Scenario 4 (S4)	+	+	+

Thus, a total of 288 experiments are conducted. Figure 4 recapitulates them.

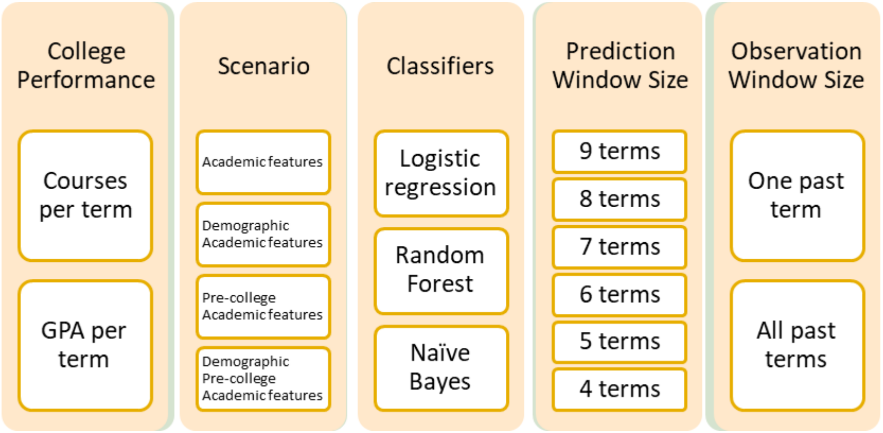


Figure 4. Experiments Conducted.

2.3. Performance Evaluation

As a base case model, instead of a simple random guess, we develop naive models based on the statistical facts that only uses term performance features and not any demographic or pre-college features. The idea behind these models is the following: for every term, we calculate the average term performance across all the students. No matter how these terms' performance is calculated, *Single-Course*, *Single-GPA* , *Cumulative-Course* , and *Cumulative-GPA*, if for a student they are always equal or above the mean of the term performance calculated across all the students, then that student is classified as *High GPA* student (i.e., a student always better than the average). Conversely, if they are always below, then the student is classified as *Low GPA* student (i.e., a student always lower than the average). All the other cases are classified as *Average GPA*. Table 4 illustrates the naive models with three sample students using the term performance by current GPA's.

Table 4. The table shows how Students A, B, and C are classified by the naive models by the end of Term 4, assuming the means of the first 4 term GPA's are 4, 3.25, 3.5, and 3.75, respectively.

	Term 1	Term 2	Term 3	Term 4	Class (*Cumulative-GPA*)	Class (*Single-GPA*)
Student A	4.1	3.5	3.5	4	High GPA	High GPA
Student B	3	3	3	3	Low GPA	Low GPA
Student C	4	3.2	3.75	4	Average GPA	High GPA

While developing the naive models, with a total of 357 samples, the size of the dataset can be problematic. If we use all the samples to develop our model, then we do not have any samples to estimate the true performance (performance of the model on an unseen data) of our models. Therefore, we split our dataset into training and testing datasets. We develop our model on the training dataset and evaluate its performance on the testing dataset. Performance indicators obtained using this approach, called *hold-out technique*, are more realistic. Yet, there are still some concerns. First of all, due to the hold-out samples, the learning is not 100%. To improve learning, we use split ratios with high training percentage. This creates yet another problem. Due to the small size of testing sets, the performance results may vary significantly. To reduce this variance, we can use repeated training and testing phases or use subsampling methods such as *k*-fold cross validation, or even repeated subsampling methods. We decide to use the repeated subsampling to minimize the variance in the accuracy scores. For this, the dataset is divided randomly into training and testing at the ratio of 4:1. Then, we calculate the statistics on the training dataset, do the classification of the samples on the testing dataset based on that statistics, and record the accuracy of the classification. Finally, we repeat this experiment 500 times and report the arithmetic mean as the result. Table 5 reports the performance of naives models.

Table 5. Results of the Naive Models.

		Term 1	Term 2	Term 3	Term 4	Term 5	Term 6
Single	**Course**	0.539	0.534	0.621	0.698	0.630	0.672
	GPA	0.460	0.498	0.556	0.553	0.561	0.558
Cumulative	**Course**	0.531	0.549	0.533	0.526	0.514	0.503
	GPA	0.470	0.618	0.671	0.736	0.734	0.742

3. Experiments and Results

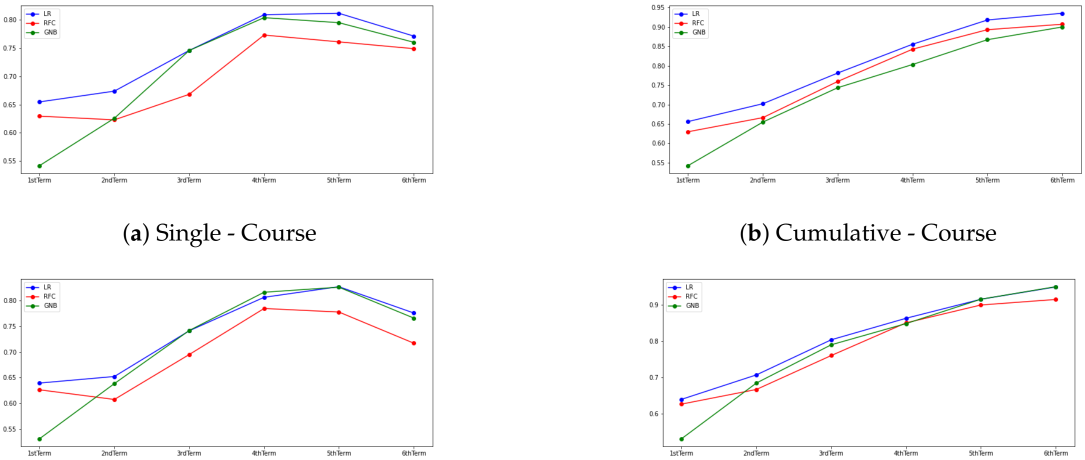
All experiments are done on Python 3. We design the ML models on Python's scikit-learn library version 0.22.2 with default hyper-parameters. For the ML models, we also use the repeated 5-fold stratified cross-validation with 100 repetitions. Since our goal is to observe the change of performance when the academic features are used either *by course* or *by GPA*, hyper-parameter search is not relevant. Nevertheless, when we tested random hyper-parameters, we observed the same trends as explained in the Discussion Section 4. We record the performance of the model both on the training and the testing datasets. All results are reported in following sub-sections per scenario.

3.1. Scenario 1: Academic Features Only

The first scenario only includes academic features, and the performance of the obtained prediction models are reported in Table 6. The best performance by the end of term 1, is the LR method with term performance represented *by courses*, whether single or cumulative, with 65.6%. When the prediction is performed later, the best performance is systematically obtained again with the LR method, but with the term performance represented *by cumulative GPA*. Please note that GNB shows the same best performance for the terms 5 and 6, and second best for the terms 3 and 4. Looking at Figure 5b,d, we observe that performance of the cumulative models are improving from 65.6% (term 1) to 94.9% (term 6) with decreasing prediction window size. Finally, Figure 5a,c show that among the models using only one past term, performance reaches a pick mostly when the prediction is performed by the end of term 4 or term 5.

Table 6. Accuracy results of the ML Models only with the academic features.

		ML Algoth	Term 1	Term 2	Term 3	Term 4	Term 5	Term 6
Single	**Course**	LR	**0.656**	0.673	0.746	0.809	0.812	0.774
		RFC	0.630	0.623	0.670	0.774	0.760	0.749
		GNB	0.542	0.625	0.745	0.804	0.796	0.760
	GPA	LR	0.639	0.652	0.741	0.8027	0.778	0.717
		RFC	0.626	0.607	0.695	0.785	0.778	0.717
		GNB	0.531	0.638	0.741	0.816	0.826	0.766
Cumulative	**Course**	LR	**0.656**	0.702	0.781	0.856	**0.918**	0.936
		RFC	0.630	0.666	0.760	0.843	0.893	0.907
		GNB	0.543	0.655	0.744	0.803	0.867	0.900
	GPA	LR	0.639	**0.707**	**0.803**	**0.863**	**0.915**	**0.949**
		RFC	0.626	0.667	0.760	0.850	0.899	0.915
		GNB	0.531	0.684	0.790	0.848	**0.915**	**0.949**



(**a**) Single - Course

(**b**) Cumulative - Course

(**c**) Single - GPA

(**d**) Cumulative - GPA

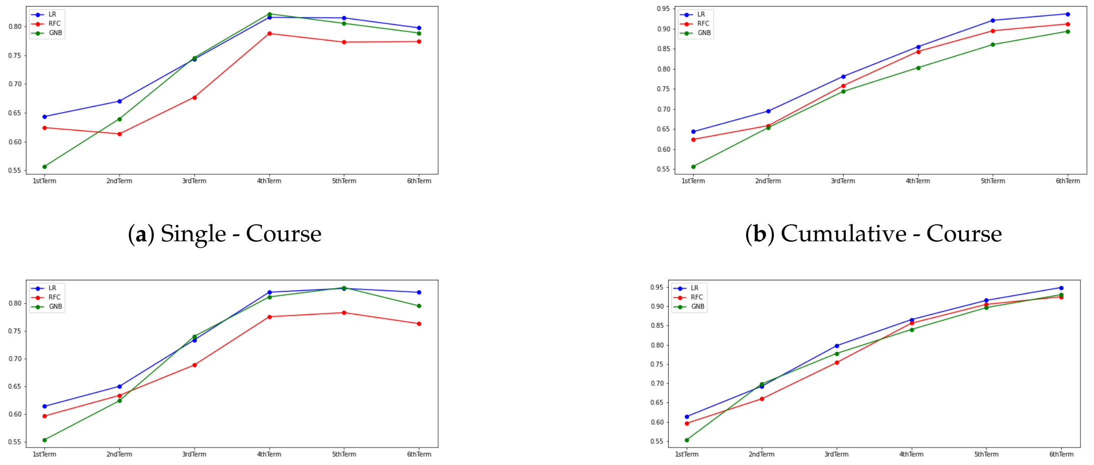
Figure 5. Accuracy plots along the terms of the ML Models only with the academic features.

3.2. Scenario 2: Demographics and Academic Features

The second scenario includes demographics and academic features. Performance of the obtained prediction models is reported in Table 7. The best performance by the end of term 1, is the LR method with the term performance represented *by courses*, whether single or cumulative, with 64.4%. When the prediction is performed later, we observe similar results with the scenario 1, i.e., the best performance is mainly obtained with the LR method, with the term performance represented *by cumulative GPA*. Please note that GNB shows a slightly superior performance for the term 4, and the second best for the terms 5 and 6. Looking at Figure 6b,d, we observe that the performance of cumulative models are improving from 64.4% (term 1) to 94.9% (term 6) with decreasing prediction window size. Again, same as for the scenario 1, the models using only one past term reach a performance pick mostly when the prediction is performed by the end of term 4 or term 5 (see Figure 6a,c).

Table 7. Accuracy results of the ML Models with the academic and demographic features.

		ML Algoth	1st Term	2nd Term	3rd Term	4th Term	5th Term	6th Term
Single	**Course**	LR	**0.644**	0.670	0.743	0.816	0.815	0.798
		RFC	0.624	0.614	0.677	0.788	0.773	0.774
		GNB	0.556	0.640	0.746	0.822	0.805	0.789
	GPA	LR	0.614	0.652	0.734	0.820	0.827	0.820
		RFC	0.597	0.634	0.689	0.776	0.783	0.764
		GNB	0.554	0.624	0.741	0.812	0.829	0.796
Cumulative	**Course**	LR	**0.644**	0.695	0.781	0.855	**0.921**	0.937
		RFC	0.624	0.659	0.758	0.843	0.895	0.912
		GNB	0.557	0.654	0.744	0.803	0.860	0.894
	GPA	LR	0.614	0.693	**0.798**	**0.866**	**0.916**	**0.949**
		RFC	0.597	0.660	0.754	0.857	0.905	0.925
		GNB	0.554	**0.699**	0.778	0.840	0.897	0.931



(**a**) Single - Course (**b**) Cumulative - Course

(**c**) Single - GPA (**d**) Cumulative - GPA

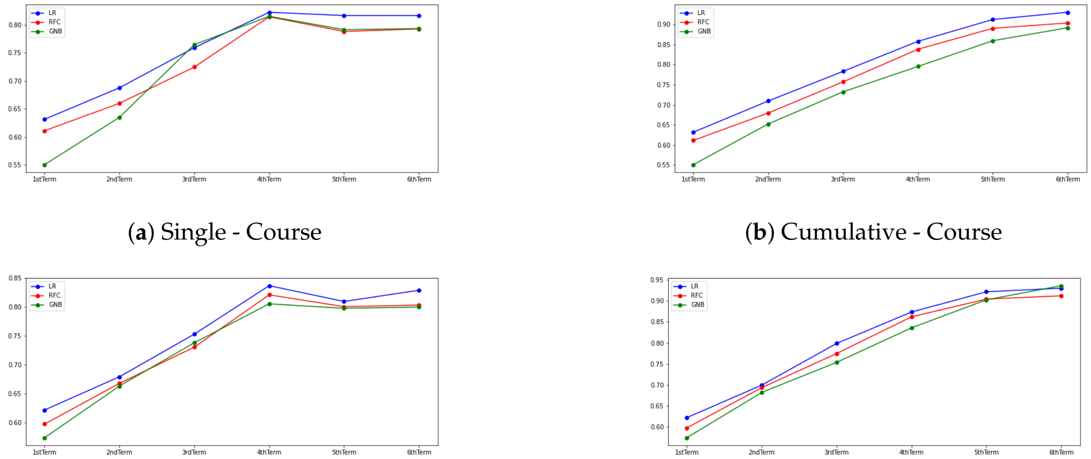
Figure 6. Accuracy plots along the terms of the ML Models with the academic and demographic features.

3.3. Scenario 3: Pre-College and Academic Features

The third scenario includes pre-college and academic features. Performance of the obtained prediction models is reported in Table 8. The best performance by the end of term 1, is again the LR method with term performance represented *by courses*, whether single or cumulative, with 63.2%. When the prediction is performed later, we observe similar results with the previous two scenarios, in other words the best performance is mainly obtained with the LR method, with term performance represented *by cumulative GPA*. The GNB shows a slightly superior performance for the term 6. Figure 7b,d shows that the performance of the cumulative models are improving from 63.2% (term 1) to 93.7% (term 6) with the decreased prediction window size. Again, same as for previous two scenarios, the models using only one past term reach a performance pick for prediction performed by the end of term 4 (see Figure 7a,c).

Table 8. Accuracy results of the the ML Models with the academic and pre-college features.

		ML Algoth	1st Term	2nd Term	3rd Term	4th Term	5th Term	6th Term
Single	**Course**	LR	**0.632**	0.688	0.759	0.822	0.817	0.817
		RFC	0.611	0.660	0.725	0.814	0.788	0.793
		GNB	0.551	0.635	0.765	0.815	0.791	0.793
	GPA	LR	0.622	0.679	0.753	0.837	0.810	0.829
		RFC	0.597	0.668	0.731	0.821	0.801	0.804
		GNB	0.574	0.663	0.738	0.806	0.798	0.800
Cumulative	**Course**	LR	**0.632**	**0.709**	0.783	0.858	0.912	0.931
		RFC	0.611	0.679	0.757	0.838	0.891	0.904
		GNB	0.551	0.652	0.732	0.795	0.860	0.892
	GPA	LR	0.622	0.699	**0.799**	**0.874**	**0.922**	0.931
		RFC	0.597	0.693	0.775	0.862	0.905	0.912
		GNB	0.574	0.682	0.754	0.836	0.903	**0.937**



(**a**) Single - Course

(**b**) Cumulative - Course

(**c**) Single - GPA

(**d**) Cumulative - GPA

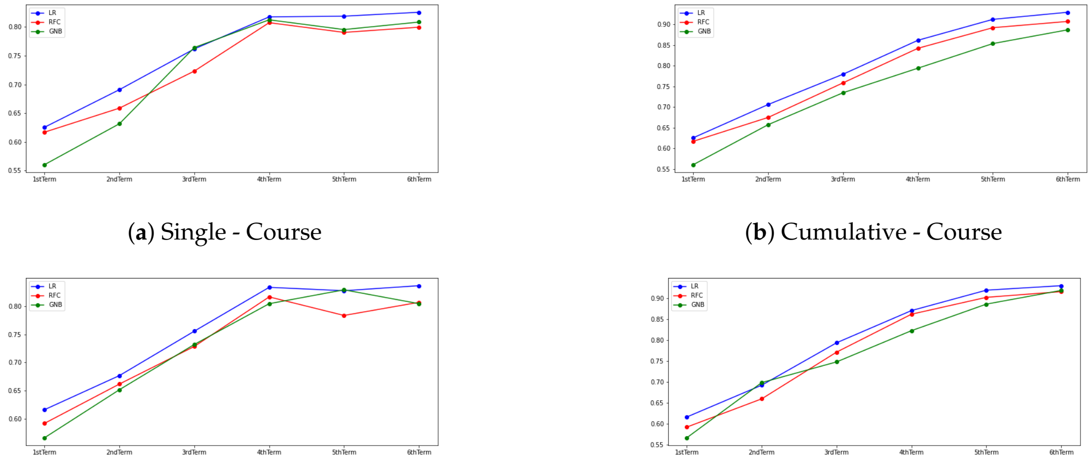
Figure 7. Accuracy plots along the terms of the ML Models with the academic and pre-college features.

3.4. Scenario 4: Demographics, Pre-College, and Academic Features

The last scenario includes all students data, that is to say, demographics, pre-college, and academic features. Performance of the obtained prediction models is reported in Table 9. The best performance by the end of term 1, is the LR method with the term performance represented *by courses*, whether single or cumulative, with 62.5%. When the prediction is performed later by the end of the terms 3, 4, 5, and 6, we observe similar results with all the past scenarios, that is, the best performance is mainly obtained with the LR method, with term the performance represented *by cumulative GPA*. Looking at Figure 8b,d, we observe that the performance of the cumulative models are improving from 62.5% (term 1) to 93.5% (term 6) with decreasing prediction window size. Models using only one past term reach a performance pick mostly when the prediction is performed by the end of term 4 or term 6 (see Figure 8a,c).

Table 9. Accuracy results of the ML Models with the academic, demographic, and pre-college features.

		ML Algoth	1st Term	2nd Term	3rd Term	4th Term	5th Term	6th Term
Single	**Course**	LR	**0.626**	0.691	0.762	0.818	0.819	0.826
		RFC	0.617	0.659	0.723	0.808	0.791	0.800
		GNB	0.560	0.632	0.764	0.813	0.796	0.809
	GPA	LR	0.616	0.679	0.757	0.830	0.821	0.836
		RFC	0.591	0.665	0.725	0.819	0.801	0.807
		GNB	0.565	0.651	0.732	0.804	0.798	0.802
Cumulative	**Course**	LR	**0.625**	**0.706**	0.779	0.861	0.912	**0.929**
		RFC	0.617	0.675	0.758	0.842	0.892	0.907
		GNB	0.560	0.657	0.735	0.794	0.853	0.887
	GPA	LR	0.616	0.697	**0.794**	**0.873**	**0.921**	**0.930**
		RFC	0.596	0.693	0.773	0.862	0.903	0.918
		GNB	0.565	0.687	0.751	0.825	0.887	0.919



(**a**) Single - Course

(**b**) Cumulative - Course

(**c**) Single - GPA

(**d**) Cumulative - GPA

Figure 8. Accuracy plots along the terms of the ML Models with the academic, demographic, and pre-college features.

4. Discussion

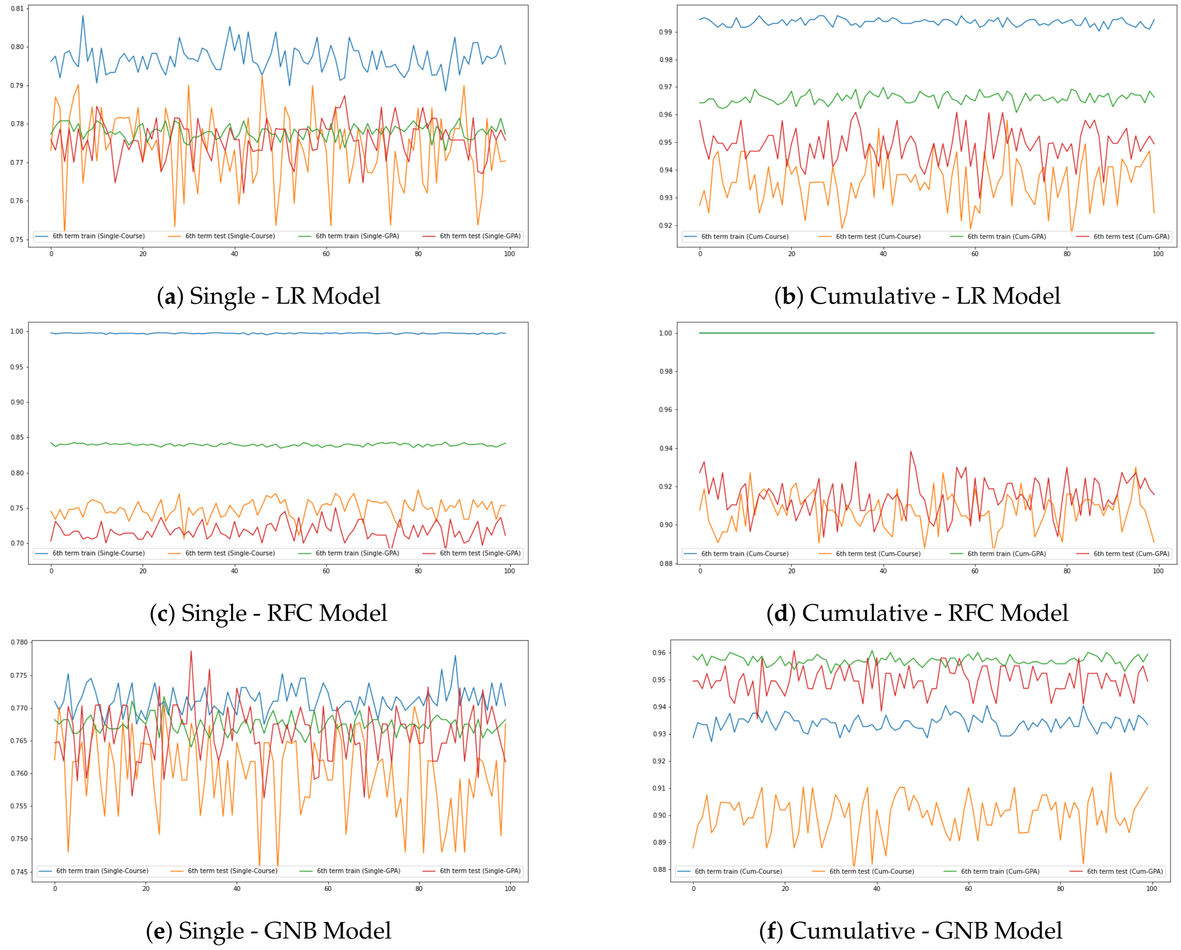
In this section, we are going to discuss our findings from two perspectives: (1) findings that can be generalized to any dataset and (2) findings specific to our dataset.

Looking at Table 10, which summarizes the best ML models per term, one can see the answer to the research question, namely, which representation of college academic performance is more relevant for predicting student graduation academic performance. When the models are compared based on how the academic performance is defined, that is whether *by courses* or *by GPA* as explained in Section 2.2, we notice, in general, that the models where the academic records are used *by GPA* achieve higher accuracy scores at the later terms whereas the models where the academic records are used *by courses* achieve higher accuracy scores at the earlier terms. To be more precise, we can see in Table 10 that course grades yields better results until the end of term 2, and GPA gives better results afterwards.

Table 10. Summary of the best models per term.

	Term 1	**Term 2**	**Term 3**	**Term 4**	**Term 5**	**Term 6**
Random Guess	0.333	0.333	0.333	0.333	0.333	0.333
Naive Model	0.539	0.618	0.671	0.736	0.734	0.742
Best cumulative ML Performance	0.656	0.709	0.803	0.874	0.922	0.949
Best cumulative ML Model	LR	LR	LR	LR	LR	LR
Academic performance	by course	by course	by GPA	by GPA	by GPA	by GPA
Scenario	S1	S3	S1	S4	S3	S1

However, in an attempt to gain more insight, we analyzed the variance of the training and testing performance of ML models. We observed that after term 2, the course models all had a higher variance. For example, Figure 9 illustrates the train and test accuracy scores by the number of experiments of all ML models on term 6. In all the plots, it is clearly visible that the difference between the green and the red curves (train and test of the GPA models) is less than the difference between the blue and the orange curves (train and test of the Course models). This indicates that if we introduce academic performance into the models *by courses*, in later terms, the models learn the noise (i.e., overfitting), which results in performance loss. This is not the case for GPA models in later terms, since one GPA per term would replace several (up to six at IAU-CCSIT) individual course grades, thus representing an equivalent information with reduced size of academic performance features. On the other hand, if we introduce the academic performance into the model as GPA at earlier terms, then we are over-simplifying the model by reducing the size of the academic features to a single number per term. We expect this observation to hold in any academic dataset and thus conclude that for earlier predictions academic performance should be used *by courses* and for later predictions *by GPA*.



(**a**) Single - LR Model (**b**) Cumulative - LR Model

(**c**) Single - RFC Model (**d**) Cumulative - RFC Model

(**e**) Single - GNB Model (**f**) Cumulative - GNB Model

Figure 9. Variance plots of ML models for term 6.

The conducted experiments also allow answering several common EDM questions with respect to the specific IAU-CCSIT dataset. Our experiment results show that all the models perform better than a random guess (which can be considered to be roughly 33.3%). Moreover, ML models perform significantly better than the naive models that we defined in Section 2.3 as our base line result. We can conclude that the prediction models can be used as early as by the end of term 1, knowing that delaying the prediction thus gathering more academic data about the student will improve the performance of the classifier. Within ML models, LR shows the best overall performance with GNB being the runner up. As for the highest performance, both LR and GNB records 94.9% accuracy using all six current GPAs cumulatively. The poor performance of RFC can be explained by the overfitting which is very clear from Figure 9c,d where training performance reaches 100%. Looking at the four scenarios and the corresponding results, we observe that the demographics is not a significant group of features as scenario 2 never yields the best performance. Academic records alone give the best performance models when used by the end of term 1, term 3, and term 6. For the terms 2 and 5, knowing about pre-college exam results slightly improves the performance. Certainly, these results should be interpreted as specific to the IAU-CCSIT dataset and congruent with the many case studies in the field ([13] and in their references).

Finally, we can draw conclusions from the models where the academic records are used alone or cumulatively. As expected, when the academic records, again alone or with the other features, are used cumulatively we get the best accuracy scores. Yet, we can extract some valuable information from the models where the academic performance is used alone. For instance, from the accuracy graphs we see that the accuracy scores peaked at term 4 and stabilized afterwards. Hence, term 4, which is the end of the general year 1, has the maximum impact on the graduation GPA class. We can thus conclude that the term 4 is a good moment to start predicting the graduation performance. This information can also be shared with students explaining them that an extra effort they will put in their studies on term 4 will have higher impact on their graduation GPA.

5. Conclusions

With a plethora of studies in EDM for predicting a student's academic success at graduation time, this study investigated which of the individual course grades or grade averages is more relevant for predicting student graduation academic performance. Although both types of data are interchangeably used in the literature, there is no study comparing the performance of EDM models using grade averages vs. individual course grades. It is unknown when and how to use these two college performance representations to attain best predictive power. To elucidate this matter, a comprehensive set of experiments were conducted on the recent student data compiled from the second author's college.

The experiment results show that for earlier predictions, individual course grades should be used to represent academic performance, while it is preferable to use GPAs for prediction after a few terms. We explain based on variance analysis that this will help avoiding oversimplification and noise, as both can lower the performance of a predictive model. This is a novel contribution to the field of EDM, that will enable scientists and educators to decide which representation to adopt depending on the time of prediction.

The second main contribution of this study is to investigate the individual impact of each semester on the graduation academic performance. The results of such an analysis can help identifying when is the best time to do the prediction (e.g., in order not to miss the most impactful term), or can help in advising and motivating the students about when to put extra efforts in their studies.

Author Contributions: Conceptualization, D.D.; Methodology, A.E.T.; Software, A.E.T.; Validation, A.E.T.; Formal Analysis, A.E.T.; Investigation, D.D.; Resources, D.D.; Writing—Original Draft Preparation, A.E.T.; Writing—Reviewing—Editing, D.D.; Visualization, A.E.T.; Supervision, D.D.; Project Administration, D.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research did not receive any grant.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liz-Domínguez, M.; Caeiro-Rodríguez, M.; Llamas-Nistal, M.; Mikic-Fonte, F.A. Systematic literature review of predictive analysis tools in higher education. *Appl. Sci.* **2019**, *9*, 5569. [CrossRef]
2. Miguéis, V.; Freitas, A.; Garcia, P.J.; Silva, A. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decis. Support Syst.* **2018**, *115*, 36–51. [CrossRef]
3. Trussel, J.M.; Burke-Smalley, L. Demography and student success: Early warning tools to drive intervention. *J. Educ. Bus.* **2018**, *93*, 363–372. [CrossRef]
4. Mason, C.; Twomey, J.; Wright, D.; Whitman, L. Predicting Engineering Student Attrition Risk Using a Probabilistic Neural Network and Comparing Results with a Backpropagation Neural Network and Logistic Regression. *Res. High. Educ.* **2018**, *59*, 382–400. [CrossRef]
5. Asif, R.; Hina, S.; Haque, S.I. Predicting student academic performance using data mining methods. *Int. J. Comput. Sci. Netw. Secur.* **2017**, *17*, 187–191.
6. Asif, R.; Merceron, A.; Pathan, M. Predicting student academic performance at degree level: A case study. *Int. J. Intell. Syst. Appl.* **2014**, *7*, 49–61. [CrossRef]

7. Adekitan, A.I.; Salau, O. The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon* **2019**, *5*, e01250. [CrossRef] [PubMed]
8. Jiménez, F.; Paoletti, A.; Sánchez, G.; Sciavicco, G. Predicting the risk of academic dropout with temporal multi-objective optimization. *IEEE Trans. Learn. Technol.* **2019**, *12*, 225–236. [CrossRef]
9. Aluko, O.; Adenuga, O.; Kukoyi, P.; Aliu, S.; Oyedeji, J. Predicting the academic success of architecture students by pre-enrolment requirement: Using machine-learning techniques. *Constr. Econ. Build.* **2016**, *16*, 86. [CrossRef]
10. Computer Science Program. Available online: https://www.iau.edu.sa/en/colleges/college-of-computer-science-and-information-technology/programs/bachelor-of-science-in-computer-science-cs (accessed on 15 December 2019).
11. Computer Information Systems Program. Available online: https://www.iau.edu.sa/en/colleges/college-of-computer-science-and-information-technology/programs/bachelor-of-science-in-computer-information-systems-cis (accessed on 15 December 2019).
12. Alyahyan, E.; Düştegör, D. Predicting academic success in higher education: Literature review and best practices. *Int. J. Educ. Technol. High. Educ.* **2020**, *17*, 3. [CrossRef]
13. Alyahyan, E. Predicting undergraduate students' Success in a Saudi University Using Data Mining Techniques. Master's Thesis, Imam Abdulrahman bin Faisal University, Dammam, Saudi Arabia, 2019.