Predictive Identification Of At-Risk Students: Using Student Information System Data

Abdulmohsen Algarni*[1] and Hadeel Allahiq[2]

[1]King Khalid University, College of Computer Science, KSA
[2]School of Electronic and Computer Science, University of Southampton, UK

Abstract

As electronic systems are increasingly utilized in education, vast amounts of data have been accumulated in educational databases. Educational Data Mining (EDM) is an emerging research field that aims to explore data from educational institutions. Extracting knowledge from educational data can facilitate a better understanding and improvement of educational processes. Predicting student performance has become a hot topic, as it can help decision-makers identify the factors contributing to student success or failure. In this study, we utilized EDM techniques to create a framework for predicting student academic performance, using pre-admission information and first-year subject marks. In the first step, the most important attributes of the data were identified and then used an EDM algorithm to extract knowledge from the selected attributes. Our experimental results revealed that the random forest algorithm produced the most accurate predictions, achieving an accuracy rate of more than 78%.

1. Introduction

Some of the challenges currently facing universities are how to reduce student failure rates and how to best advise students regarding the successful completion of their academic programs. However, it can be hard to tell which students need help during the early stages. In the current era, the capabilities of Data Mining (DM) and Machine Learning (ML) have been able to contribute to the prediction of which students will struggle and the identification of their needs to reduce student failure rates in universities. With the development of technology, educational institutions worldwide have relied on storing student data electronically. As a result, countless valuable data are generated daily, though they are commonly underutilized. DM, which aims to extract valuable information from big data, could play an

important role in resolving this situation.

Recently, researchers have shown great interest in taking advantage of the possibilities of educational data mining (EDM) to understand the phenomena of teaching and learning behaviors. One exciting trend in EDM is the prediction of student performance, as this can help educational institutions and decision-makers to improve student success rates. For example, EDM could help in improving attendance rates [1] measuring the effectiveness of instructional methods, and detecting unacceptable grades obtained by students [2].

Although educational institutions are making efforts to maintain high levels of student enrollment and prevent students from dropping out, their activities have subtle side effects that may lower student outcomes and lead to higher dropout rates. Many researchers have sought to uncover and study the differences between certain aspects, such as student exam scores [3], the educational background of parents [4], the influence of ethnicity [5], and gender differences [6]. In addition, some researchers have investigated how pre-admission scores affect student performance [7]. However, no published reviews have explained how pre-admission scores, such as General Aptitude Test (GAT) scores, Standard Achievement Admission Test (SAAT) scores, high school achievement scores, and first-year performance, can predict the ultimate performance of students.

In this paper, we aimed to fill this research gap by predicting student academic performance based on pre-admission scores and performance during the first year of their studies in university. Early awareness of student performance enables universities to take proactive measures, improves student status, and aids management and faculty in developing clear and personalized learning plans for each student. Hence, this awareness could improve the use of university resources, speed up the process of efficient student graduation, and reduce the number of students who fail to complete their studies. It also serves as an additional indicator that can assist academic advisers in identifying students who need more help in a timely fashion. Thus, these students can be psychologically and educationally supported in improving their academic performance.

The remainder of this paper is structured in four sections as follows: the second section presents a literature review related to this field; the third section describes the proposed model, detailing the meaning of student academic performance and data modeling; the fourth section discusses the evaluation

process and contains descriptions of the evaluation methods and datasets used; the fifth section presents the results and discussion; and, finally, the last section provides our conclusions.

2. Literature Review

Around the world, universities, companies, and even individuals are generating endless amounts of electronic data. The huge amount of data generated every day leads to an increased use of DM and ML. DM constitutes techniques and steps that can be performed to extract knowledge from different types of data. The extracted knowledge can be used to improve designs [8]. Some researchers have defined DM as the ability to conduct deep analysis to find unexpected patterns and relationships within data that can help decision-makers to solve problems. To understand these patterns and relationships, historical data are used and analyzed to create understandable structures that can solve current or future problems. Recently, DM and data analysis have been used in many sciences and fields (e.g., marketing, industry, business, and banking). Each time these operations have been used, hidden patterns in complex big data have been revealed [9] and information of unrecognized value has been extracted.

In education, the emerging field of EDM is concerned with the study of unique patterns and valuable information in large databases from educational environments [10]. In recent years, EDM has emerged as a stand-alone field, as it has extended beyond the purpose for which the data in these databases were generated [11]. Researchers are using EDM to gain insights into student behavior patterns and their interactions with educational environments. In addition, EDM seeks to harness the power of data to discover ways to improve existing teaching methods. Therefore, educational institutions need to understand the importance of the field and its ability to help teachers in adapting learning styles to the individual needs of learners [12]. The better teachers understand the behaviors within educational data, the greater their ability to improve educational outputs [13]. However, some educational data have different characteristics that require different, adapted methods of data mining. This is because educational data are diverse and numerous.

DM is a broad field in which several different methods are used to obtain appropriate results for solving problems. Some of the most common of these methods are categorization, grouping, and correlation [14]. Recent studies have focused on the correlation between pre-university admission requirements and academic performance [15]. This research has shown that

using pre-university admission requirements can help in predicting student performance with high accuracy. Different studies used students' progress during the semester to predict final exam marks [16]. The data that were used included quizzes, assignments, and mid-semester exam results to predict the final exam mark. Similar studies used facets of student behavior, such as attendance and quiz results, to predict student exam marks in small student cohorts [17].

Several algorithms, such as decision trees (DTs), neural networks, discriminant analysis, and neural networks, were used to predict the performance of these students. However, the results were inappropriate due to the large differences between the numbers of samples [18].

In addition, several other researchers have used the Naïve Bayes (NB) classifier to solve the same problem [19,20]. NB produced good results, but the random forest (RF) algorithm outperformed many other algorithms, such as DTs, classification and regression algorithms, and NB, in a study that aimed to predict the academic performance of students in higher education [21].

The use of selected features has also varied among researchers, as some have tended to focus on educational features, while others have tried to identify non-educational features. For example, one study directed at educational features [7] used high school, SAAT, and GAT scores to predict student performance and showed that, overall, SAAT scores were critical factors in predicting student academic performance. Most previous research has focused on student test scores, achievements, and university stages to predict student performance. The differences between these studies reside in the samples, as some researchers have studied first-year students [3], while others have focused on fourth- year students, for whom historical data were available [20].

In previous research, there was great diversity among the non-educational features studied. For example, the discrepancy in scores between males and females has also intrigued researchers. Some have found that male students very slightly outperform female students in some courses [6], while others have found that females outperform males in many other cases [22]. Some have found no significant differences between the sexes. In addition to examining gender differences [23], some research has been carried out on the effects of race and personality on academic achievement [5] Other researchers have also found evidence that supports the suggestion that personality type influences student academic performance [24]

3. Proposed Model

3.1 Student Academic Performance

To achieve the aim of this research, it was important to clearly define student academic performance. This concept can be defined in several ways, for example, by using cumulative average scores or the time required for each student to complete their graduation requirements. In this paper, the students were classified into two classes as follows:

- Normal-level students (normal performance)
- At-risk students (low performance)

The normal-level students were those who completed their studies within the normal study plan period (i.e., less than or equal to 10 semesters) and had GPAs that were more than good. The at-risk students were those who exceeded the normal study plan period (over 10 semesters) or had GPAs of less than or equal to 2.75 out of 5. Therefore, students can be categorized into one of these two classes as follows:

$$\begin{cases} GPA > 2.75 \;\&\; |S| \leq 10 \;\&\; Graduated & \text{, Normal} \\ GPA \leq 2.75 \; Or \; |S| > 10 & \text{, At-risk} \end{cases} \quad (1)$$

Where |S| is the number of semesters during which a student has studied at university.

Using student attributes would help in extracting the knowledge necessary for predicting student performance in the early stages. However, student records contain large amounts of information; however, not all of the information is useful for predicting student performance.

Therefore, one of the necessary challenging tasks is the selection of the right features for use in predicting student performance. Then, the appropriate algorithm needs to be selected based on the characteristics of the dataset used to train the system. The main steps of the proposed model are shown in Figure 1.

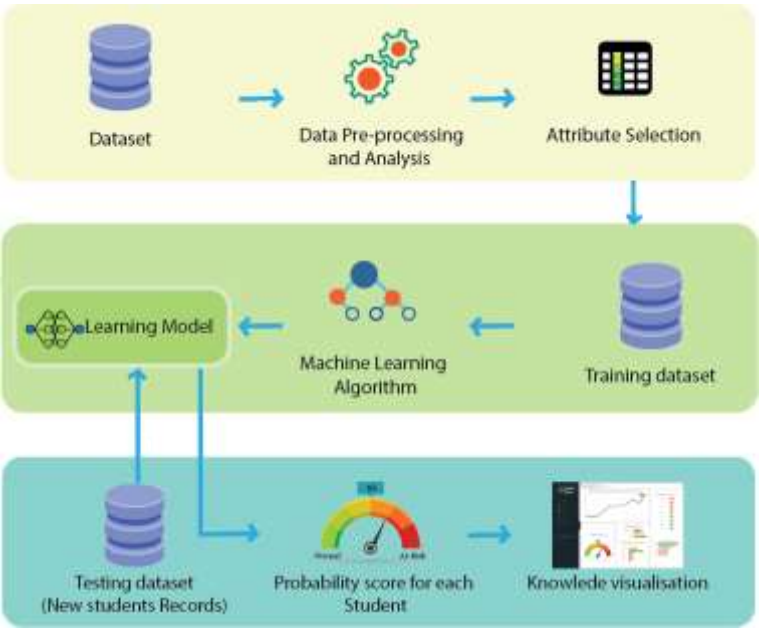

Figure 1: The proposed model.

3.2 Understanding the Data

It is critical to clearly understand the dataset we will be using in this study, including the amount of data, number of records, and attributes. Moreover, it is also essential to know the data type for each attribute and to have a certain amount of statistical information. This stage aims to ensure that the dataset is suitable for the problem and can provide beneficial results. Therefore, in this section, we will discuss the data used in this research in detail to ensure that the data are sufficient for achieving this project's desired results.

3.3 Dataset

The dataset used in this research is sourced from open data from King Khalid University (KKU). It is one of the largest public universities in Saudi Arabia and is spread over several cities in the southwest of the country. The data consists of 827 records for bachelor's degrees across four subject areas: Computer Science, Computer Engineering, Information Systems, and Network and Communications Engineering, as shown in Figure 2. Each student record consists of 32 attributes, such as gender, pre-admission tests, specialization, and first- and second-semester grades. All the features are shown in Table 1. However, not all features are useful for prediction. Some of the most important features are the SAAT mark, GAT mark, and the subjects studied in the student's first year. For the four majors studied, the first year of the study plan is the same. In the first semester, the student is assigned to one of four subjects: English

Language Course (011ENG-6), Introduction to Computers (011CSM-6), Mathematics (001MATH-3), and Introduction to Islamic Culture (111IC1-2). In the second semester, the student is asked to study the second stage of the same subjects.

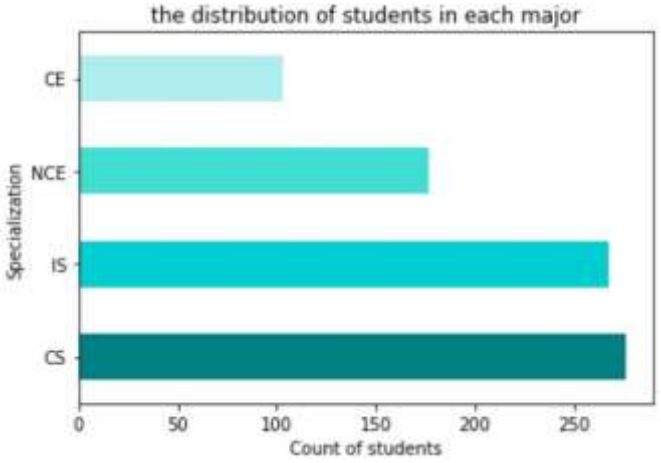

Figure 2: Number of students recorded in each department.

Figure 3 shows the distribution of the data in terms of the most important attributes. It is clear to us that the distribution of High School Accumulative Average (HSAA) leans to the right. The main reason for this is that, to attend university, the student must obtain high grades in high school. In contrast, we find that the distribution of GAT and SAAT results is normal. Thus, most students enrolled in the majors in this research have a relative GAT and SAAT average between 70 and 80. Regarding the distribution of subject grades, the grading system used ranges from 0 to 100, and the student must pass with a mark of at least 60 to succeed in the course. Therefore, we find that most subjects have a normal distribution but also find that some have a somewhat normal distribution, but with a long tail to the left. This is a normal condition due to some students failing to attend the final exam or withdrawing from the subject.

Table 1: Dataset attributes.

Attribute Name	Description	Attribute Name	Description
Sex	The gender of each student	HSAA	High School Accumulative Average
SAAT	Pre-Admission test	GAT	Pre-Admission test
Current headquarters	The branch of the university	Study Duration Semester	The duration of the student's study
Last semester	The student's last semester	Current Status	Student's academic status
GPA for the last semester	Last GPA of student	Appreciation	Student's appreciation
011ENG marks	English score for 1st semester	011ENG grade	English grade for 1st semester
011CSM marks	Introduction to Computer score for 1st semester	011CSM grade	Introduction to Computer grade for 1st semester
001MAT marks	Mathematics score for 1st	001MAT grade	Mathematics grade for 1st semester

	semester		
111IC1 marks	Introduction to Islamic Culture score- 1st semester	111IC1 grade	Introduction to Islamic Culture grade-1st semester
012MAT marks	Mathematics score for 2nd semester	012MAT grade	Mathematics score for 2nd semester
012ENG marks	English score for 2nd semester	012ENG grade	English grade for 1nd semester
112IC1 marks	Introduction to Islamic Culture score- 2nd semester	112IC1 grade	Introduction to Islamic Culture grade-2nd semester
012CSM marks	Introduction to Computer score for 2nd semester	012CSM grade	Introduction to Computer grade for 2nd semester
Current specialization	Students current specialization	Class type	Student's performance

3.4 Data Preparation

The existence of a dataset is one of the requirements for any data mining study. Therefore, collecting data and understanding the source of the dataset is very important. However, every dataset suffers from problems, including missing values, the duplication of the records and outliers, etc. Therefore, it is important to prepare data before using them to achieve accurate results. The dataset preparation process is carried out by applying data mining pre-processing steps to the data set; this helps in cleaning and normalizing the dataset. The expected output of this stage is a clean, balanced, and ready-to-use dataset.

In the pre-processing stage, certain attributes that do not improve the prediction value for our model were omitted (for example, the name of the college or secondary school). The second step was to deal with all the missing data. It should be noted that there were very few missing data with regard to grades (these missing data related to three students, for which there was also missing information regarding other features). Due to the difficulty of obtaining these results from the university, the data for these students were deleted from the dataset. All duplicate student records and irrelevant features, for example, regarding the type of high school attended, were also removed, as these would not achieve any improvement in prediction results. Table 2 shows the number of attributes and records before and after data processing.

Table 2: Comparing dataset before and after pre-processing.

	Attributes	Records
Before pre-processing	44	827
After pre-processing	28	823

Regarding the format of the data, information was provided in

Arabic; therefore, all words were translated into English to facilitate data mining. Moreover, a new column has been created representing the classes that represent the students' performance rating: 0 was assigned to low- performing students and 1 to students performing normally. Finally, any data with wrong values, or in a different format and that caused noise, has been corrected.

3.5 Data Modeling

Classification techniques are required to solve the student performance prediction problem. Based on the literature review, we noticed that no one particular algorithm consistently outperformed the others, although it was interesting that some algorithms were repeatedly used in research. Therefore, we chose to test the following algorithms, in addition to certain others that fit with the dataset: a DT, NB, logistic regression (LR), support vector machine (SVM), and RF.

- **Decision Trees**

DTs are considered to be among the most popular algorithms in the scientific community due to their ability to solve both classification and regression problems. DTs are based on hierarchical analysis [25]. Therefore, the shape of their external structure resembles a tree. Unlike other algorithms, DTs are characterized by their ease of use, in addition to their ability to deal with categorical and numerical data. The goal of using DTs is to reduce data uncertainty in order to achieve the best possible segmentation; therefore, in this study, we used an entropy meter to measure the uncertainty in our data. Entropy meters measure the randomness of data and the impurity of nodes. In DTs, the final division must be either "yes" or "no", without any impurities. Entropy meters measure this impurity according to the following equation [26]:

$$(S) = -(+)\log\ p(+) - p(-)\log\ p(-)$$

where p (−) is the negative class probability, p(+) is the positive class probability, and S is a subset of the training data.

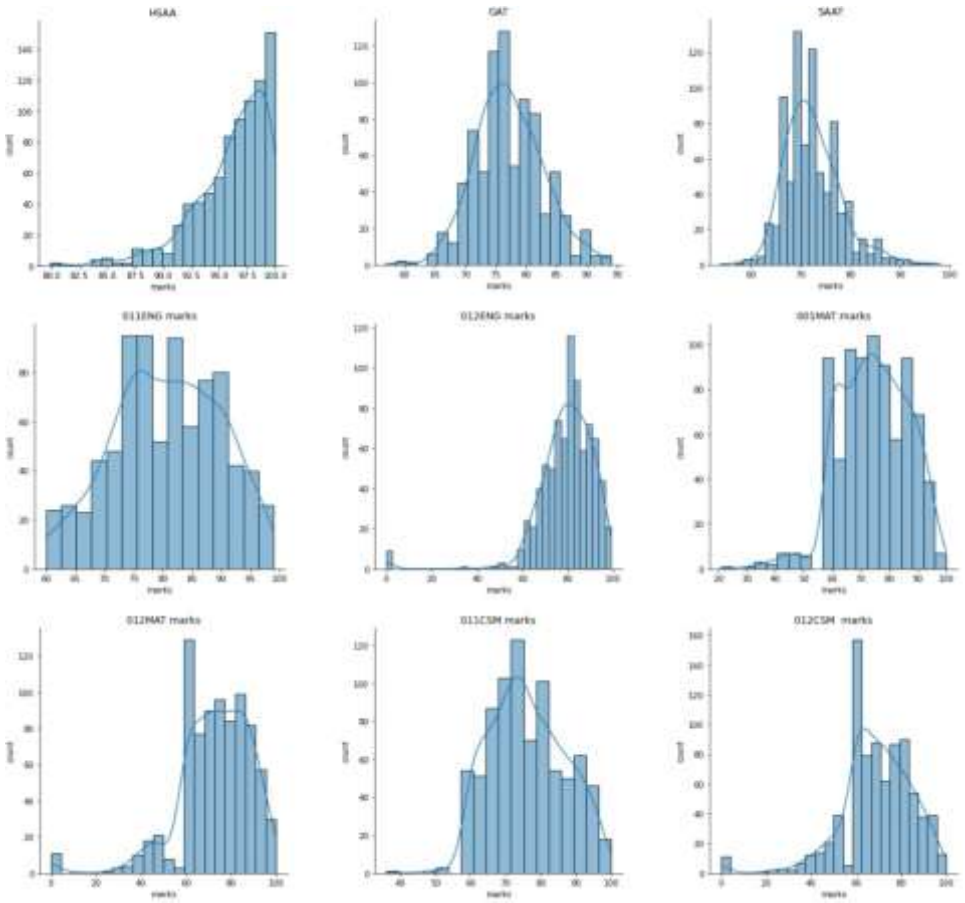

Figure 3: Data distribution of subjects.

- **Naïve Bayes (NB)**

NB is an ML paradigm and it produces good and acceptable results relatively quickly. NB is best known for its application in Bayes' theorem as follows:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(\mathrm{B})} \qquad (2)$$

The idea of Bayes' theorem revolves around calculating the conditional probabilities of independent variables, meaning that the presence of a feature in a class is independent of and is not affected by the presence of other features [27]. Using Bayes' theorem, it is possible to predict the occurrence of variable A given the occurrence of B. In this study, A represented the hypothesis and B was the evidence for the prediction.

- **Logistic Regression (LR)**

Logistic Regression LR is an algorithm that produces binary classification using linear regression equations. It is characterized by its simplicity and its ability to deal well with various features. In this paper, we used LR to align the task of

classifying students as either at-risk (represented by class 0) or normal-level (represented by class 1) as follows:

$$0 \leq (X) \leq 1$$

- **Support Vector Machine (SVM)**

SVMs are supervised algorithms that divide data into categories using maximum margins. As shown in Figure 4, the greater the distance from the data to the maximum margin, the more effective and accurate the classification of the SVM algorithm [28,29].

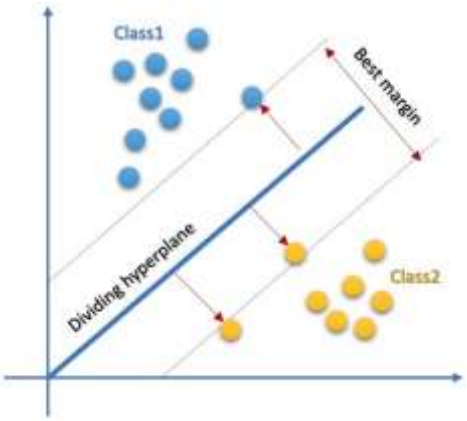

Figure 4: A representation of a support vector machine.

- Random Forest (RF)

RF algorithms are supervised learning algorithms that combine several DTs that operate individually and form strong predictors by working together [27]. Each DT uses the bagging concept, which allows it to take random samples from datasets that are equivalent to the size of the original dataset. Thus, each tree provides a different prediction. The prediction that occurs the most often is taken as the closest prediction for the target task [30].

4 Evaluation

4.1 Splitting the Dataset

After the pre-processing steps, the dataset contained 823 student records from between 2012 and 2015. Students accepted between 2012 and 2015 were chosen because we could easily classify each student. To implement the classifiers, we divided the dataset into two: 80% of the data were used for training and 20% were used for testing. In order to achieve the best distribution between the two groups, we ensured that all classes were represented in both training and testing datasets.

During the training phase, we imported the training data into the models for training. The testing dataset then provided an estimate for the predictive power of these models. To achieve the best distribution between the two groups, we ensured that all classes had the same distribution so that the success of the models was not affected by generalization and the results were more accurate.

As shown in Table 3, the training dataset contained 658 records. The at-risk students comprised 44.9%=296/658, while the normal-level students comprised 55.1%=362/658. In contrast, the total number of records in the testing dataset was 165, of which 44.8% = 74/165 were at-risk students and 55.1= 91/165 were normal-level students.

Table 3. The distribution of student records in the training and testing datasets.

Dataset	Total	At-Risk Students	Normal-Level Students
Training Dataset	658	296	362
Testing Dataset	165	74	91

4.2 Evaluation Metrics

The evaluation was an important stage in the prospecting process, aiming to identify the algorithm that best represented the training data and would enable long-term success. In this paper, we were interested in one category of students, namely the at-risk students, because they urgently needed help in order to complete their studies successfully. To perform the assessment, we used metrics that are well-known in the educational community, i.e., confusion scales, to determine the adequacy of the proposed model, as shown in Table 4.

Table 4. A confusion matrix of the classifiers.

		Actual Value	
		At-Risk	**Normal-Level**
Prediction Outcome	**At-Risk**	TP	FP
	Normal-Level	FN	TN

One of the most common metrics is accuracy. This measure represents the percentage of correct outcomes out of the total cases examined, as shown in the following equation:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}$$

However, accuracy is not completely reliable, especially when

the available data are unbalanced, because it can produce misleading results [31]. For this reason, we used accuracy in conjunction with other metrics to produce the most accurate overall rating. Other metrics that are widely used by researchers are precision, recall, and F1 − score. Precision is used to indicate the validity of model predictions by calculating the percentage of true positive outcomes that the classifier categorizes as being in a positive category as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

On the other hand, recall shows the proportion of true positive outcomes that the model rates correctly as follows:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Recall is commonly used when the purpose of the task is to detect the largest number of positive outcomes. However, it is impractical to compare the recall and precision results and take the best value. Therefore, recall and precision are often traded off to find a balance, i.e., the F1−score. Whenever these two scales decrease, the F1−score decreases accordingly and vice versa. The F1−score is calculated as follows:

$$F1 - score = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

5. Results and Discussion

5.1 Results

Table 5 shows the results of all four metrics for each model. We focused on the precision scores, which represented the number of positive outcomes that were classified as positive. LR produced the best result in terms of precision for classifying at-risk students (82%), which indicated that most of the students who were classified as at-risk were categorized correctly by the LR algorithm. Conversely, the DT produced the worst result. Figure 5 shows the precision results of all the tested models. Regarding recall, which represented the percentage of true outcomes that were classified as positive by the models (Table 5), the SVM algorithm outperformed all of the other algorithms with 80% . The higher the precision and recall, the better the prediction performance. The F1 − score was then used to compare all the models by creating a balance between the precision and recall results. Figure 6 shows a comparison between the F1−scores of all the models. Unsurprisingly, the RF and SVM algorithms yielded excellent results, outperforming the other classifiers by 0.74% and 0.72%, respectively. In terms of accuracy, Table 5 shows that the

RF algorithm achieved a high result of 78%, while the accuracy of the DT algorithm was lower than those of the other classifiers at 68%.

Table 5. A comparison of all tested algorithms, based on the selected evaluation

Metric	RF	DT	SVM	NB	LR
Precision	0.79	0.62	0.66	0.73	0.82
Recall	0.70	0.72	0.80	0.61	0.62
F1 – Score	0.74	0.67	0.72	0.66	0.71
Accuracy	0.78	0.68	0.73	0.72	0.77

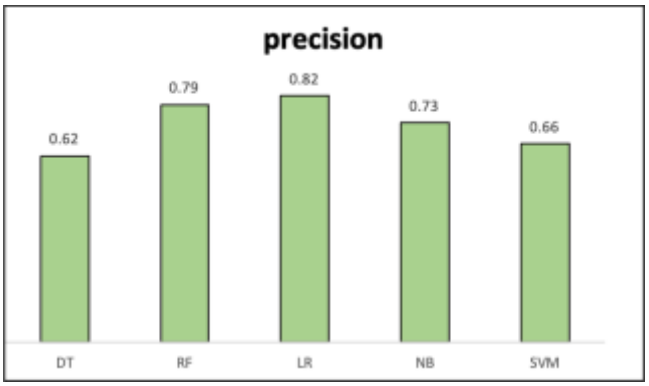

Figure 5: A comparison of the precision results of all models.

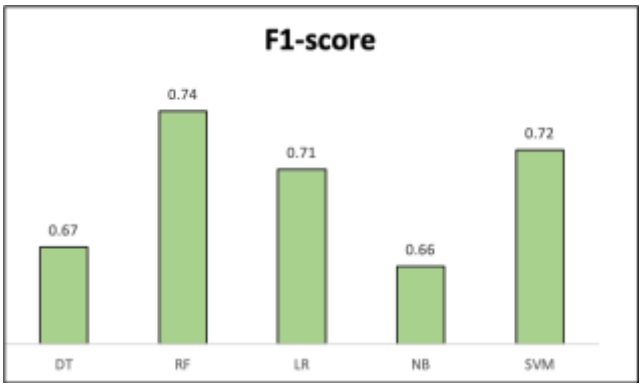

Figure 6: A comparison of the F1− scores of all models.

5.2 Discussion

The discussion is divided into two main sections. The first part discusses the common features and their importance in the best model results. Table 5 shows that some models produced better predictive results than others. In fact, the RF and SVM algorithms produced the best results in terms of F1 − score. However, it is generally better to predict false positives than misclassify true positives. it has been found that the RF algorithm performed better than the other classifiers within our framework. Because its results in the other metrics reflected its ability to predict which students were at risk in the early stages of their courses, we chose to use the RF algorithm to define the basic model within this framework.

5.2.1 Feature Importance

To predict student academic performance, especially that of at-risk students, we analyzed various attributes (i.e., high school, SAAT, GAT, MATH1, MATH 2, ENG1, ENG2, CSM1, and CSM2 scores). To find out which attributes were important in the prediction of student performance, we divided them into three sections, as shown in Table 6: pre-admission scores only (i.e., high school, SAAT, and GAT scores); pre-admission scores and first-semester scores (i.e., MATH1, ENG1, and CSM1 scores); and all attributes.

The main idea behind these classification categories was to determine the best time to start the prediction process for new students and the most important features on which to base decisions.

As shown in Table 6, using only pre-admission grades and first- or second-semester scores was insufficient for the prediction of student performance. Additionally, Figure 7 shows that using all attributes resulted in the best prediction results.

Table 6. The most important features affecting the prediction of student performance for RF Classifier:

Feature	Precision	Recall	*F*1 – *Score*	Accuracy
Pre-Admission Scores	0.66	0.55	0.60	0.67
Pre-Admission and First-Semester Scores	0.71	0.59	0.65	0.71
All Attributes	0.79	0.70	0.74	0.78

Despite our preference for combining pre-admission and first-year data, a recent literature review on the topic of predicting student academic performance found that only using pre-admission scores produced good results [7]. In particular, the study recommended allocating the highest importance weight within admission systems to the SAAT score. In fact, the results of that study, as well as our own results (Table 6), confirm that only using pre-admission scores could be a good predictor of student academic performance. However, as shown in Table 6, the use of more attributes, such as first- and second-semester grades, in addition to pre-admission grades, produced noticeable improvements in the results.

The precision results improved by 19.6% and the recall results improved by 27.27%. Finally, in terms of *F1–score* and accuracy, the results improved by 16.41% and 13.84%, respectively.

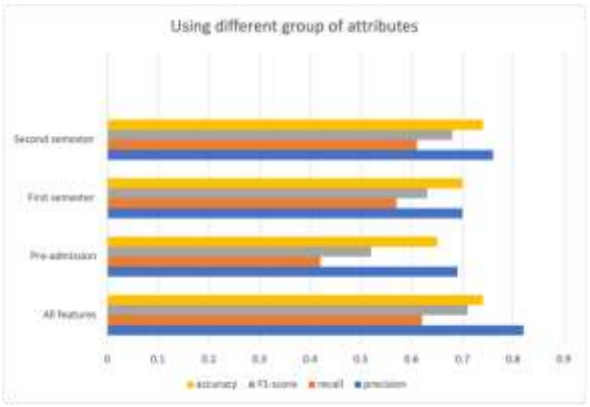

Figure 7: A comparison of the results obtained using different groups of attributes.

Based on the proposed model, we determined the best time to start predicting the academic performance of new students was after the first year, using first-semester and second-semester results in addition to pre-admission test scores. In order to determine the effect of each attribute on the results, we calculated the importance of each feature in the RF model. Feature importance plays an important role in any predictive model, including reducing data dimensionality and highlighting the best features for improving the efficiency and effectiveness. They can also be used to remove features that are less relevant to the desired goal of achieving a shorter training time.

We analyzed the different features in the RF model to establish the order of the features in terms of their importance. Figure 8 clearly shows the importance scores of each feature. The top five most important features for identifying at-risk students were high school GPA and CSM2, ENG2, ENG1, and MATH2 scores. To emphasize the importance of these features, we applied each of them in the RF model separately and the results are presented in Table 7.

Table 7. A comparison between the metrics of the most important features and the less important features in RF Classifier.

Feature	Precision	Recall	*F*1 – *Score*	Accuracy
Most Important Features	0.74	0.68	0.70	0.75
Less Important Features	0.62	0.50	0.55	0.64

By determining the most important features for improving teaching and learning processes, it becomes possible to highlight these in teaching processes in order to help improve student performance.

Figure 8: The importance of the selected attributes.

6. Conclusions

In this paper, we presented a framework that could enable universities to benefit from the large volumes of data that are stored in their electronic systems daily. Data concerning pre-admission information and first-year grades could be employed to predict the academic performance of students in the early stages of their educational journeys. In the first stage of this work, we chose the most effective attribute for the proposed model through data pre-processing, applying statistical analysis, and studying the various correlations between important features. In the second stage, we modeled the data using different algorithms and applied those algorithms to the training and testing data. Then, we evaluated the results using evaluation metrics, such as accuracy, recall, and F1 − score. Our evaluation results showed that the RF algorithm produced the best prediction results, with an accuracy of more than 78%.

References

Kasthuriarachchi, K.; Liyanage, S. Predicting Students' Academic Performance Using Utility Based Educational Data Mining. In Proceedings of the International Conference on Frontier Computing. Springer, 2018, pp. 29–39.

Hung, J.L.; Hsu, Y.C.; Rice, K. Integrating data mining in program evaluation of K-12 online education. Journal of Educational Technology & Society 2012, 15, 27–41.

Vandamme, J.P.; Meskens, N.; Superby, J.F.; et al. Predicting academic performance by data mining methods. Education Economics 2007, 15, 405.

Hasan, M.; et al. Predicting Student Performance to Reduce Dropout Using J48 Decision Tree Algorithm. PhD thesis, Daffodil International University, 2019.

Borg, M.O.; Stranahan, H. The effect of gender and race on student performance in principles of economics: The importance of personality type. Applied Economics 2002, 34, 589–598.

Monem, R.M. Does access to tutorial solutions enhance student performance? Evidence from an accounting course. Accounting &Finance 2007, 47, 123–142.

Mengash, H.A. Using data mining techniques to predict student performance to support decision making in university admission

systems. IEEE Access 2020, 8, 55462–55470.
Manjarres, A.V.; Sandoval, L.G.M.; Suárez, M.S. Data mining techniques applied in educational environments: Literature review. Digital Education Review 2018, pp. 235–266.
Han, J.; Pei, J.; Kamber, M. Data mining: concepts and techniques; Elsevier, 2011.
Educationaldatamining. Educational Data Mining. https://educationaldatamining.org/, 2021. [Online; last accessed 10 Jul. 2021].
Baker, R.; et al. Data mining for education. International encyclopedia of education 2010, 7, 112–118.
Greller, W.; Drachsler, H. Translating learning into numbers: A generic framework for learning analytics. Journal of Educational Technology & Society 2012, 15, 42–57.
Vahdat, M.; Ghio, A.; Oneto, L.; Anguita, D.; Funk, M.; Rauterberg, M. Advances in learning analytics and educational data mining. Proc. of ESANN2015 2015, pp. 297–306.
Romero, C.; Ventura, S. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 2010, 40, 601–618.
Qahmash, A.; Ahmad, N.; Algarni, A. Investigating Students'; Pre-University Admission Requirements and Their Correlation with Academic Performance for Medical Students: An Educational Data Mining Approach. Brain Sciences 2023, 13. https://doi.org/10.3390/brainsci13030456.
Hasan, H.R.; Rabby, A.S.A.; Islam, M.T.; Hossain, S.A. Machine Learning Algorithm for Student's Performance Prediction. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT),2019, pp. 1–7. https://doi.org/10.1109/ICCCNT45670.2019.8944629.
Wakelam, E.; Jefferies, A.; Davey, N.; Sun, Y. The potential for student performance prediction in small cohorts with minimal available attributes. British Journal of Educational Technology 2020, 51, 347–370. https://doi.org/https://doi.org/10.1111/bjet.12836.
Kovacic, Z. Early prediction of student success: Mining students' enrolment data. 2010.
Kotsiantis, S.; Pierrakeas, C.; Pintelas, P. PREDICTING STUDENTS'PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES. Applied Artificial Intelligence 2004, 18, 411–426.
Asif, R.; Merceron, A.; Pathan, M.K. Predicting student academic performance at degree level: a case study. International Journal of Intelligent Systems and Applications 2014, 7, 49–61.
Kumar, M.; Singh, A. Evaluation of Data Mining Techniques for Predicting Student's Performance. International Journal of Modern Education & Computer Science 2017, 9.
Tho, L.M. Some evidence on the determinants of student performance in the University of Malaya introductory accounting course. Accounting Education 1994, 3, 331–340.
Brookshire, R.G.; Palocsay, S.W. Factors contributing to the success of undergraduate business students in management science courses. Decision Sciences Journal of Innovative Education 2005, 3, 99–108.
Borg, M.O.; Stranahan, H.A. Personality type and student performance in upper-level economics courses: The importance of race and gender. The Journal of Economic Education 2002, 33, 3–14.
Aggarwal, C.C.; Zhai, C. A survey of text classification algorithms. In

Mining text data; Springer, 2012;
pp. 163–222.
Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends 2021, 2, 20–28.
Manda, K.R. Sentiment Analysis of Twitter Data Using Machine Learning and Deep Learning Methods, 2019.
Mahesh, B. Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet] 2020, 9, 381–386.
Mohammadi, M.; Rashid, T.A.; Karim, S.H.T.; Aldalwie, A.H.M.; Tho, Q.T.; Bidaki, M.; Rahmani, A.M.; Hosseinzadeh, M. A comprehensive survey and taxonomy of the SVM-based intrusion detection systems. Journal of Network and Computer Applications 2021, 178, 102983.
Géron, A. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow; " O'Reilly Media, Inc.", 2022.
Liu, B. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications); Springer, 2007.