International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Predicting At-Risk Students in Higher Education

Abdulmohsen Algarni*[1], Maha Abdullah*[2], Hadeel Allahiq[3], Ayman Qahmash[4]

Submitted: 26/04/2023 **Revised**: 27/06/2023 **Accepted**: 06/07/2023

Abstract: Student performance prediction is very important and can help an educational institute to increase the success rate among students. A common problem universities face is that of students failing to complete the academic program or taking a long time to do so. Identifying at-risk students in the early stages would help to provide them with the support they need. At-risk students can be described in different ways, depending on the educational system and its requirements. In this paper, at-risk students are defined as those with a low GPA of less than 2.75 out of 5 or who have failed to graduate. The focus is on the attributes that can help recognize at-risk students in advance. The results of this study proved that at-risk students can be predicted at an early stage based on their gender, and marks on pre-admission exams, in high school, and the first semesters of their academic programs.

Keywords*: At-Risk students' prediction; Classification; Data mining; Decision Tree; Educational Data Mining.*

1. Introduction

Education plays an important role in the development of societies, and educational institutions have recently been using electronic systems to store information about their students and employees. Such systems can store enormous amounts of data, which grow dramatically every year as existing students' progress through universities and new students enroll. The use of available data can help an understanding of many educational phenomena, thus enabling the provision of solutions to some of the problems faced by students, teachers, and universities.

Data mining (DM) has garnered considerable attention because of the large amounts of data that can be stored in several formats, such as texts, images, files, audio, and video. Converting this data into practical and meaningful knowledge is, therefore, an extremely important task, and the knowledge discovered through DM techniques plays a significant role in decision-making. Educational data mining (EDM) is concerned with developing strategies that can extract knowledge from educational data environments. The benefits of the use of EDM tools are not limited to analyzing student behavior, as they enable the use of available prior information to predict the performance of students in their future studies. Universities face many challenges related to student performance, which reflects on the quality of their educational outcomes. As poor student achievement impacts both universities and students, applying EDM techniques, which turn raw data stored in educational systems into useful knowledge, would help solve the problem, and have a significant impact on educational practice and research.

Educational data can be derived from diverse educational environments. Romero and Ventura [1] suggested e-learning systems, student information systems (SISs), intelligent tutoring systems (ITSs), and adaptive educational hypermedia systems (AEHSs) as examples. Some researchers used students' activity sequences in the Massive Open Online Courses (MOOC) platform to predict at-risk students [2]. Other studies used the discussion data in the online system to predict student performance [3]. Based on the problem, several methods can be applied to solve a specific task. The most used are classification, regression, association rule mining, clustering, outlier detection, discovery with models, and sequential pattern mining [4]. Most of the available research focuses on extracting knowledge from student learning management [5-8].

Various educational data mining techniques apply to educational issues. EDM can be used to predict exam results, warn students at risk before final exams [9], predict students' final GPAs from their grades in previous courses, evaluate the courses with the greatest impact on students' final GPAs [10], predict students' academic outcomes at the end of the school year [11], support admission planning and predict students' employment positions after graduation [12], predict low- and high-achieving students [13], predict student drop-out rates [14], predict at-risk students in a specific course [15], and predict slow learners [16].

This research focuses on applying classification methods to student data available in SISs to extract useful knowledge that would identify at-risk students at the early stages. Predicting at-risk students would help to provide them with the required support. To achieve that, the main attribute that affects student performance needs to be selected, after which feature extraction will be applied to the selected

[1] *King Khalid University, College of Computer Science, KSA*
[2] *King Khalid University, College of Computer Science, KSA*
E-mail: 441813258@kku.edu.sa
[3]*School of Electronic and Computer Science, University of Southampton*
E-mail: hhaa1n20@soton.ac.uk
[4] *King Khalid University, College of Computer Science, KSA*
E-mail: a.qahmash@kku.edu.sa
** Corresponding Author Email: a.algarni@kku.edu.sa*

attribute. Using the extracted features can help to predict at-risk students.

The remainder of this paper is organized as follows. Section 2 introduces a detailed overview of the related works. Section 3 reviews the concepts of the proposed method, including the dataset used. Section 4 introduces the algorithms most used for predicting at-risk students. Finally, Section 5 offers the concluding remarks.

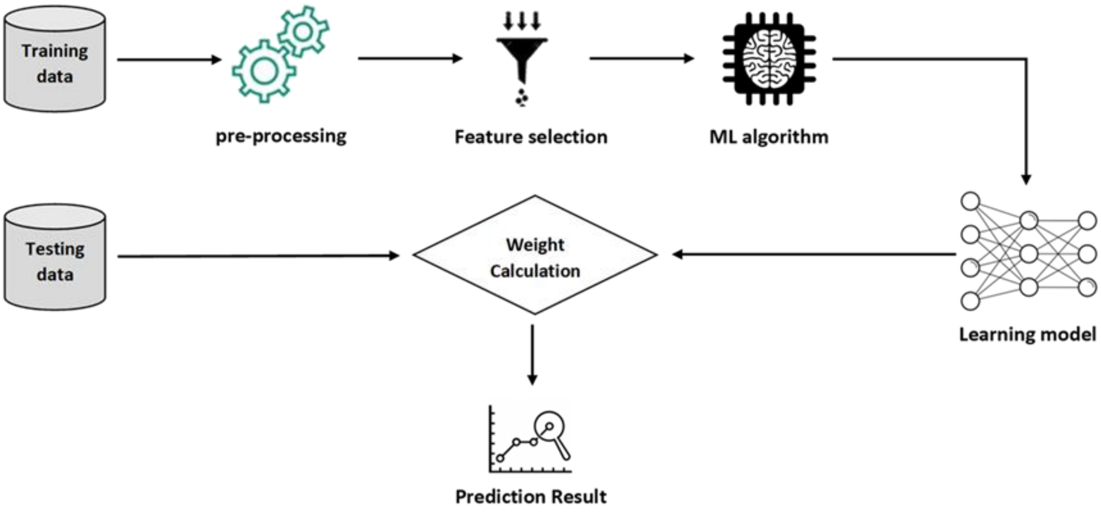

Fig 1: Proposed Model.

2. Related Work

The process of EDM turns raw data stored in educational systems into useful knowledge that could potentially have considerable influence on educational practice and research. Educational data can be derived from diverse educational environments such as e-Learning systems, SIS, intelligent tutoring systems (ITS), and adaptive educational hypermedia systems (AEHS) [1]. Based on the problem at hand, several methods can be applied to solve a specific task, of which those used most are classification, regression, association rule mining, clustering, outlier detection, discovery with models, and sequential pattern mining [4].

There are multiple educational issues to which various educational data mining techniques are applied. The goal of applying DM methods to educational data differs, depending on the problem under study [9,10,11]. It can be used in several way such as predicting students Mark, GPA, final result at the end of academic year. Moreover, EDM can support education organization in admission planning, employments rate, student's dropout rate [12,13,16]. In university datasets, numerous attributes can be found. Both selected attributes and prediction methods can lead to research directions [17].

EDM methods have been applied to the data of both public schools and universities. Numerous attributes, with academic, personal, economic, social, and institutional factors, can be found in the dataset, among which personal factors are the main cause of students dropping out of university [18]. Some researchers have proposed studying attributes like social networks and the internet as a learning resource to determine their influence on student performance. Absences and grades are the most relevant attributes for predicting students' end-of-year academic outcomes [11]. Demographic attributes such as neighborhood, age, and school are also potential indicators of a student's academic failure or success [11], [19]. However, the Standard Achievement Admission Test (SAAT) is one of the most common attributes used to predict student performance in universities [20].

To predict student performance, various DM classification techniques have been applied, of which the decision tree (DT) is the most commonly used. It is used to predict students' final GPAs based on grades achieved in previous courses [10]. Naive Bayes (NB) and different types of decision trees like CART, C4.5, and ID3 are applied to establish models that can be used to predict students' grades [21]. Some studies apply several classification methods to predict student performance, such as a rule learner (OneR), a nearest neighbor algorithm (IB1), a neural network (Multi-Layer Perceptron), and a decision tree [22]. Moreover, artificial neural network (ANN), Support Vector Machine (SVM), DT, and NB are applied to predict student performance to support decision-making [20]. Some attributes like preadmission exams are used to predict student performance in specific programs [20]. In this research, we believe that using both pre-admission and first-semester marks will provide better results in the prediction of student performance.

3. Proposed Model

In this paper, we propose a framework to predict at-risk students at early stages using student information available in the SIS systems. These systems contain numerous attributes, some important and others not. However, it is important to define who are at-risk students to select the most attributes that would affect them. At-risk students can be defined in different ways depending on many factors, including the educational settings they study in, or the curriculum of the program they enrol in. Based on studying the students' records, at-risk students can be defined as those with GPAs of less than 2.75.

The academic records in the dataset used concern 743 students. Student records reveal that 452 students graduated with GPAs over 2.75 (excellent, very good, and good), while the remaining (291) students have GPAs under 2.75 (pass, fail). Of the 291 with low GPAs, 171 had GPAs of less than 2 and cannot be awarded a degree. Therefore, students with low academic achievement are considered at-risk. The low-academic achievement students are those with a very low GPA (less than 2.75), who are vulnerable to academic warnings. Moreover, a low GPA may lead to dropping out or academic expulsion. It also reflects the poor knowledge of the students in their study field and in general. Indeed, the extended consequences of low GPAs may affect the future careers of students.

Selecting the right attributes and applying machine learning algorithms to them can be challenging. Figure 1 shows the model presented in this paper. The following section discusses the proposed model in more detail.

To achieve this, the following steps will be followed:

1. The pre-processing steps will be applied to the dataset to perform data cleansing and normalization.
2. To select the most important attributes that affect at-risk students, Pearson's product-moment correlation coefficient has been used. It determines which variables have a higher correlation with at-risk students. The variable with the highest correlation coefficient contributes more than the others to predict at-risk students. Moreover, the built-in feature importance function of the random forest can be used to rank the different attributes based on their importance.
3. Based on the characteristics of the dataset and selected attributes, four classification techniques will be applied, viz., DT, SVM, NB, and Random Forest (RF). Then the parameters that control the performance of the model will be set to achieve possible higher scores on the evaluation measures.
4. Then the evaluation measures will be calculated to assess the result.

The previous steps can be repeated as often as needed until an acceptable result is obtained.

3.1. DATA SET

The data set consists of 743 student records for both males and females from the departments of Computer Science (CS) and Information Systems (IS) at King Khalid University in Saudi Arabia, as shown in Figure 2. Each student's record consists of numerous attributes. However, this research focused on a few attributes, as shown in Table 1, including the student's preadmission information, first semester marks, students' status, and their cumulative grade point averages (CGPAs). The data are limited to students who enrolled during 2013–2015, including two semesters each year.

3.2. Data Description

It is critical to understand the domain of the dataset. The main attributes that can be used are the required pre-admission exam marks and first-semester subject marks and grades. The admission exams include the high school accumulative average (HSAA), a general aptitude test (GAT), and a standard achievement admission test (SAAT).

The study plan for both departments (CS and IS) is identical. Each program consists of 152 study hours distributed across five years with two semesters each. In the first semester, every student is required to take four subjects, the courses being Intensive English Program 1 (011ENG6), Mathematics 1 (001MATH-3), Introduction to Computers (011CSM-6), and Entrance to Islamic Culture (111IC1-2). The marks scored by students in these courses are categorized into nine grades: A+ (95–100), A (90–94), B+ (85– 89), B (80–84), C+ (75–79), C (70–74), D+ (65–69), D (60–64), and F (less than 60). The final total GPA is calculated out of 5. The general grades for graduation are as follows: Excellent (GPAs 4.50 –5.00), Very Good (3.75–4.49), Good (2.75–3.74), and Pass (2.00 –2.74). A student must attain a GPA of at least 2.00 to graduate. Table 1 provides a brief description of each attribute. The corelation between the most common attributes and at-risk students is presented in Figure 3.

3.3. Feature Selection

Numerous attributes were available for use. However, not all were useful to apply to prediction. Therefore, to reduce the dimension of the data, a basic statistical method was applied to select the most important attributes that can be used to predict at-risk students.

First, all attributes that have numerous NULL values, such as the name of the high school, were removed. Then all attributes containing nearly the same values, such as age, were ignored. As shown in Table 2, the correlation between the selected attributes has been calculated. Table 2 shows the correlation between the independent variables and the dependent variable. The heat map in Figure 4 shows the correlation between different attributes. Consequently, only the attribute (Gender, HSAA, GAT, SAAT, ENGLISH, Math, and CSM) was selected.

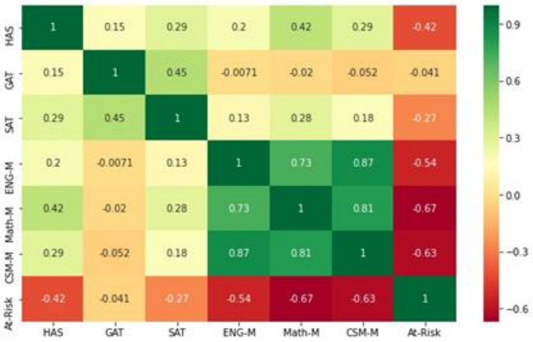

Fig 2: Correlation among the most common attributes and at-risk students.

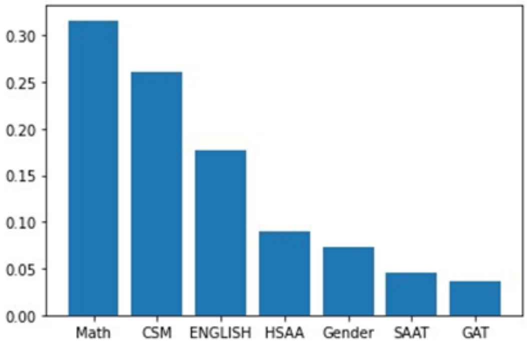

Fig 3: Feature Importance.

Table 1: Correlation between the independent variables and the dependent variable

	HSAA	GAT	SAAT	ENG	MATH	CSM
At-risk (1)	-0.42	-0.041	-0.27	0.54	-0.67	-0.63

3.3.1. Correlation Coefficient

Table 2 presents the correlation between the independent variables in the dataset, viz., HSAA, GAT, SAAT, ENGLISH, Math, and CSM with the target (dependent variable), At-risk. The first semester courses are highly correlated to the at-risk variable, with a correlation coefficient of more than (-0.5), while the pre-admission information is less correlated with the target (at-risk), with a correlation coefficient of less than (-0.5).

3.3.2. Feature Importance

To assign an importance score to each feature, the ensemble-based RF model was used to rank the different features based on their importance. Such techniques are applied to rank the features based on their contribution to predicting at-risk students. Features with higher importance scores play important roles in predicting at-risk students. In other words, the higher the importance score of the feature, the better it predicts at-risk students.

Figure 4 shows the importance of each attribute based on its contribution to predicting at-risk students—the most important attribute is Math, which has been given a higher score, followed by CSM, ENGLISH, and HSAA. In contrast, the other attributes of gender and pre-admission exams (GAT and SAAT) are considered less important attributes. The attributes assigned low importance scores are insufficient to predict at-risk students, compared to other attributes.

4. Machine Learning Algorithms

The problem of at-risk students was considered one of classification. Two classes have been defined: at-risk and not at-risk students, based on which multiple classification techniques were applied to predict at-risk students. Based on many EDM studies in the literature, the most frequently used classification methods are DT, NB, SVM, and RF.

4.1. Design Tree (Dt) Algorithm

DT is a non-parametric supervised learning technique, used for solving classification and regression problems [23]. A

DT structure resembles a tree, with the root node being the first node in the tree, no incoming edges, and at least one outgoing edge. The internal nodes were used to test the attributes, with each branch symbolizing an outcome of a test. Each terminal node, known as a leaf node, refers to a target or class label. A DT sorts an instance from the root outward, toward certain terminal nodes, representing the classification of the instance.

There are various types of DT, including iterative dichotomiser 3 (ID3), successor of ID3 (C4.5), classification and regression trees (CARTs), and conditional inference trees [23]. Two different impurity-based criteria are used to measure the quality of a split: entropy and Gini impurity [24]. The differences between the probability distributions of target variables can be measured using Gini impurity [23]. Sometimes, data are mislabeled due to random labeling. The frequency with which data are mislabeled when they are randomly labeled can be measured using Gini impurity. Gini values fall between 0 and 1. A value of 0 means that samples belong to the same target or class, whereas a value of 1 implies that the elements are randomly distributed across several classes.

Gini impurity can be calculated by the following formula:

where p_j is the probability of class j.

$$GiniIndex = 1 - \sum_{j=1}^{n} p_j^2$$

The amount of information that describes samples precisely is measured by entropy. As with Gini impurity, a value of 0 indicates that the samples are similar and belong to the same class, while a maximum value of 1 means that samples are divided equally.

Entropy is calculated by the following formula:

$$\text{Entropy} = -\sum_{j=0}^{n} p_j * \log(p_j)$$

where p_j is the probability of class j.

There are some advantages to using DT, one of which is its easiness of interpretation. Unlike other techniques can also deal with both categorical and numerical data. Nevertheless, DT has a few disadvantages, one of which is susceptibility to overfitting, meaning that a complex tree does not generalize well.

Table 2: Number of at-risk and not at-risk students in the training and testing datasets.

Dataset	Total number of students	Number of at-risk students	Number of not at-risk students
Training dataset	594	215	379
Testing Dataset	149	76	73

However, this can be handled by using parameters to control the size of a tree, such as pruning or setting the maximum depth to a suitable number [23]. This research used a CART model to predict at-risk students by learning the decision rules derived from the dataset's features. CART was introduced in 1984 by Breiman [25] to analyze binary recursive partitioning, meaning that each group of instances represented by DT nodes is split into two groups [26].

4.2. Random Forest (Rf) Algorithm

RF is a supervised learning technique developed by Breiman [27], where the features are selected randomly for each decision split [28]. RF is preferred for high-dimensional data due to its ability to handle missing values. It can also deal with categorical, continuous, and binary data, and is less sensitive to outliers [28]. Like DT, RF can also be used for both classification and regression problems. Further, RF can measure the importance of variables, as well as impute missing values [29]. RF trees depend on binary recursive partitioning trees [30]. The predictor space is split by a binary partition series (split for single variables). The whole predictor space is included in the root node of the tree. The final partition of the predictor space is composed of leaf nodes (terminal nodes), and the internal nodes are split into two successor nodes [30].

4.3. Support Vector Machine (SVM) Algorithm

SVM is a supervised learning technique introduced by Vladimir Vapnik [31]. Its basic idea is to ascertain which optimal hyperplane can classify unseen data according to the training data, which can be described as generalized linear classification. Further, SVM is known as a maximum margin classifier due to its capability of minimizing errors in empirical classification on the one hand and maximizing the geometric margin on the other [32]. SVM is sensitive to the type of kernel function. Kernel functions are mainly used when values cannot be separated linearly [31], but no single kernel is preferred in all cases; selecting the best kernel function depends on the nature of the problem being considered. The most common types of kernels are linear, polynomial, radial basis function (RBF), and sigmoid.

A linear kernel can be expressed as:

$$K\ (X_i, X_j) = X_i^T X_j$$

A polynomial kernel is not a constant one and is preferred in cases where all the training samples are normalized [31]. It can be defined as follows:

$$K\ (X_i, X_j) = (\gamma\ X_i^T X_j + r)^{\ d}.,\ \gamma > 0$$

An RBF kernel is a Gaussian kernel, whereby a measure of distance is smoothed by a radial function (an exponential function) [33]. Unlike a linear kernel, an RBF kernel can deal with non-linear relationships between a target and its attributes. Again, unlike a linear kernel, an RBF kernel maps non-linear data points onto a higher dimensional space. It acquires fewer hyperparameters than a polynomial kernel, and can be defined as follows:

$$K\ (X_i, X_j) = exp\ (-\gamma \|X_i - X_j\|^2).,\ \gamma > 0$$

A sigmoid kernel can be expressed as

$$K(X_i, X_j) = tanh(\gamma X_i^T X_j + r).$$

In the above kernel expressions, the γ, d, and r are the parameters of the kernel.

4.4. D. Naive Bayes (Nb) Algorithm

NB is a probabilistic supervised learning technique, the main idea behind which is to apply the Bayes theorem with a “naive" assumption of conditional independence between variables [34]. This assumption is not always met but can provide reasonable performance with low computation times. The Bayes theorem is used to calculate conditional probabilities. An NB classifier supposes that the presence or non-presence of a certain feature of a class is unassociated with the presence or non-presence of any other feature [34]. There are several types of NB classifiers, including Gaussian NB, multinomial NB, Complement NB, and Bernoulli NB.

5. Experimental Results and Discussion

This section narrates the overall result and discussion. To apply the model, the data set was split into two groups: a training dataset and a testing dataset (80% and 20%, respectively). The training dataset was used to train the model, and the testing dataset to test it. In this context, the data set was slitted sequentially according to years. Therefore, the data of student enrollees to the university in 2013 and 2014 were used as a training set, and the remaining data of student enrollees to the university in 2015 were used as a testing set. This was done for using senior students’ data to predict junior students’ future.

Then, based on the students enrolled at the university in 2013 or 2014, the performance of those enrolled in 2015 can be predicted.

The training dataset comprised 594 student records, the at-risk students constituting 36.2% of the data. Moreover, 149 student records (51.0%) were included in the testing dataset of the records of those considered at-risk students. Table 3 shows the distribution of students in the training and testing datasets.

5.1. Evaluation Measures

The most commonly used evaluation measures in the data mining and machine learning fields are accuracy, precision, recall, and F-measure. Accuracy is simply the number of observations predicted correctly, over the total observations. Accuracy is a good measure in the case of a balanced dataset, where the positive and negative observations are almost the same. However, for an unbalanced dataset, accuracy can be a misleading metric. In other words, evaluating the model using the accuracy measure alone is not reliable. For example, taking a sample with 99 positive samples and one negative sample, if all samples are classified as positive, the accuracy score will be 99% [35].

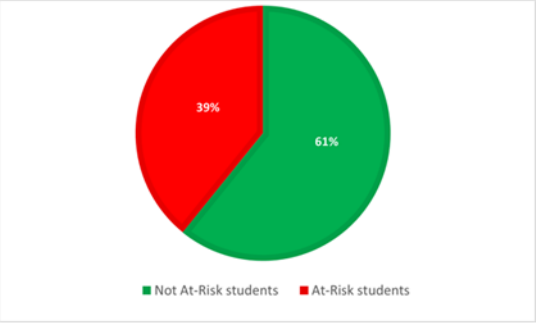

Fig 4: The ratio of students At-Risk and Not At-Risk

Precision and Recall can be used to evaluate the model in more accurate ways. Whereas Precision measures the accuracy of positive predictions, recall measures the completeness of positive predictions. In most problems, both high precision and recall offer the best result.

However. it is difficult to compare the two models’ effectiveness based on the two measures. Hence, another technique called F-measure, also known as F1-score is applied, which is a metric that considers both precision and recall.

Precision, recall, F-measure, and accuracy can be calculated using the following equations:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

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

$$F - measure = \frac{2 * \ precision * recall}{precision + recall}$$

where TP (true positive) is where a student is predicted as at-risk, and the actual class of the student is also at-risk; FP (false positive) is where the student is predicted as not at-risk and the actual class is at-risk; TN (true negative) is where the student is predicted as not at-risk and the actual class too is not at-risk; and FN (false negative) is where the student is predicted as at-risk and the actual class is not at-risk.

5.2. AT-RISK STUDENTS

At-risk students can be defined in different ways, depending on many factors, including the educational settings they study in, or the curriculum of the program they have enrolled in. In this research, students with GPAs under the threshold (GPA=2) will be classified as at-risk.

Figure 5 illustrates the ratio between students at risk and not at risk in the dataset. Figure 6 represents the final GPAs achieved by students, which reveals that several students have obtained a GPA under 2.75 (Pass, Fail).

Table 3: Result of predicting the at-risk students.

Classifier	Precision	Recall	F1-score	Accuracy
RF	0.86	0.87	0.86	0.86
DT	0.89	0.87	0.88	0.88
NB	0.89	0.78	0.83	0.84
SVM	0.86	0.87	0.86	0.86

5.3. Result

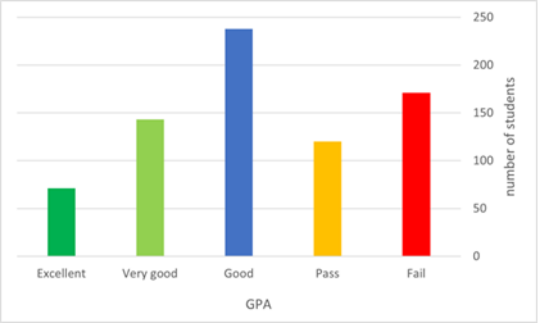

Fig 5: The distribution of student GPAs in the data set.

The result of applying different classification methods is presented in Table 4. It is clearly shown that at-risk students can be predicted based on first-semester courses, pre-admission exams, high school averages, and gender. Based on the results in Table 4, all classifiers achieved a very good result. However, the decision tree algorithm produced the best result in the F-measure.

As shown in Figure 7, DT offered the best result in predicting at-risk students in the F-Measure result. Both RF and SVM produced an equal result concerning all evaluation measures, while the decision tree outperformed other classifiers in terms of accuracy and F1 score. Further, NB performed worse than other techniques concerning accuracy, recall, and F1score.

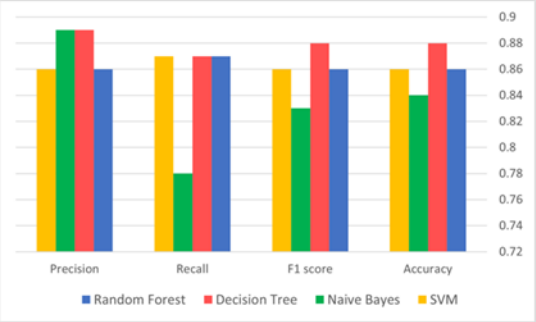

Fig 6: Comparison of the performance of all classifiers.

Table 4: Common features of the at-risk students.

1	IF (Math ≤6) AND (CSM ≤35) AND (Gender = Male) AND (HSAAs ≤90), Then student at-risk.
2	IF (Math [> 6 & ≤30]) AND (CSM ≤61) AND (Gender = Male), Then student at-risk.
3	IF (CSM ≤56) AND (Math ≤59) AND (Gender = Male), Then student at-risk.
4	IF (CSM ≤67) AND (Math ≤22) AND (English ≤80), Then student at-risk.
5	IF (CSM ≤24) AND (Math ≤58) AND (English ≤73), Then student at-risk.
6	IF (CSM ≤68) AND (Math ≤32), Then student at-risk.
7	IF (CSM ≤68) AND (Math ≤56) AND (English ≤83), Then student at-risk.

Different studies using different attributes came up with the same result [19], [20]. However, NB outperformed SVM and RF in terms of the precision score. Despite the slight variance in the results of all classifiers, the overall result achieved by each was good. Generally, all the predictive models developed achieved high scores in all evaluation measures, which reflects their power to predict at-risk students in the early stages of their studies.

5.4. Comparing All Attributes with Pre-Admission Attributes

Table 5: All attributes *vs* pre-admission attributes.

Attributes	Precision	Recall	F1-score	Accuracy
All Attributes	0.89	0.87	0.88	0.88
Pre-Admission	0.77	0.71	0.74	0.74
% chg	15.59 %	22.54 %	18.92 %	18.92 %

Table 6 shows the comparison result of using DT to predict at-risk students using all the attributes (Gender, HSAAs, GAT, SAAT, CP, English, Math, and CSM) and using preadmission attributes (HSAAs, GAT, and SAAT). The result shows that using all attributes produced a better result. In fact, first semester marks (for English, Math, and CSM) improved the prediction results remarkably.

Some researchers [20] have focused only on pre-admission scores (HSAAs, GAT, and SAAT) to predict students' performance,

Table 6: Most important attributes *vs* less important attributes.

Attributes	Precision	Recall	F1-score	Accuracy
Most Important Attributes	0.88	0.84	0.86	0.86
Less Important Attributes	0.84	0.76	0.80	0.81
% chg	4.77 %	10.53 %	7.5 %	6.18 %

applying different classifiers, such as DT, SVM, and NB. The results presented in Table 6 confirmed that predicting students' performance using pre-admission scores alone gave acceptable results to some extent, but using additional attributes, such as first-semester marks, considerably improved the results of the classifier. In terms of precision, it improved the results significantly (by 15.59%) and in terms of recall, by 22.54%; and in terms F1-score and accuracy, it improved the results by 18.92%.

Applying different feature selection techniques (see Table 8) ranked the attributes according to their importance for predicting at-risk students. To understand the effect of the most important

attribute compared with less important attributes, please see Table 7. The results showed that the performance of the classifier using the most important attributes was better than that performance using the less important attribute.

5.5. Common Features of At-Risk Students

To identify the common features of the at-risk students, DT generates a rule to classify them. The extracted rules provided a clear picture of the following features (Table 5) that all at-risk students have in common.

It can be observed that all marks for the CSM course scored by at-risk students were less than 68 for all the extracted features. In the case of the Math course, all the students' marks were below 59, meaning that they did not pass the course because their marks were below 60. Moreover, the marks scored by at-risk students for the English course were below 83.

Since males constituted the largest portion of the at-risk students, they dominated in all extracted features, compared to females. As can be seen from all the extracted features of the at-risk students, the CSM and Math courses were present notably, emphasizing the importance of these two subjects for identifying at-risk students. By contrast, the other attributes did not play a prominent role. In conclusion, out of all extracted features of the at-risk students, the following was the inclusive and general feature: regardless of gender and HSAAs, if the CSM was ≤ 68, Math was ≤ 59, and English was ≤ 83, the student is possibly at-risk. Consequently, the institute needs to pay more attention to those subjects and to the students who got low marks in them.

6. Conclusion

The main goal of this study was to provide a framework to predict at-risk students, using the attributes that could help to achieve this goal. Specifically, two groups of attributes were used. The first one is pre-admission information, including HSAA, pre-admission exams, and gender. The second group comprises academic results for the first semester courses. The study has been conducted on the academic records of undergraduate students in the College of Computer Science. To predict the at-risk students, four classification techniques were applied: DT, RF, NB, and SVM. The results show that DT achieved 88% in terms of accuracy and F1-score, higher than the other classifiers. To select features, a technique like feature importance methods was used to rank the attributes according to their importance in predicting at-risk students. Moreover, Pearson's product-moment correlation coefficient was used to determine the correlation

Table 7: Importance score for each attribute

	Attributes							
method	Math	CSM	English	HSAAs	Gender	SAAT	GAT	CP
FI	0.344287	0.260451	0.169005	0.088935	0.054436	0.042612	0.034718	0.005555
RFE	True [1]	True [1]	True [1]	True [1]	False [1]	False [2]	False [3]	False [4]

between the independent variables (pre-admission information, and first-semester courses) and the dependent variable (at-risk students). The results of both techniques indicate that first-semester courses play the most important role in predicting at-risk students. The results of this study demonstrate that students at risk can be predicted early using pre-admission and first-semester scores using DT. Further research could be conducted using more attributes from second-semester marks.

References

[1] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[2] M. Fei and D.-Y. Yeung, "Temporal models for predicting student dropout in massive open online courses," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, 2015, pp. 256–263.

[3] C. Romero, M.-I. López, J.-M. Luna, and S. Ventura, "Predicting students' final performance from participation in on-line discussion forums," *Computers Education*, vol. 68, pp. 458–472, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131513001607

[4] R. B. Sachin and M. S. Vijay, "A survey and future vision of data mining in educational field," in *2012 Second International Conference on Advanced Computing & Communication Technologies*. IEEE, 2012, pp. 96–100.

[5] Y.-C. Chang, W.-Y. Kao, C.-P. Chu, and C.-H. Chiu, "A learning style classification mechanism for e-learning," *Computers & Education*, vol. 53, no. 2, pp. 273–285, 2009.

[6] F. Castro, A. V. Alacena, A. N. Castells, and J. Minguillon, "Detecting atypical student behaviour on a e-learning system," in *Actas del I Simposio Nacional de Tecnologías de la Información y de las Comunicaciones en la Educación*. Thomson-Paraninfo, 2005, pp. 153–160.

[7] F. Castro, A. Vellido, A. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," in *Evolution of teaching and learning paradigms in intelligent environment*. Springer, 2007, pp. 183–221.

[8] I. Lykourentzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950–965, 2009.

[9] G. Ben-Zadok, A. Hershkovitz, E. Mintz, and R. Nachmias, "Examining online learning processes based on log files analysis: A case study," in *5th International Conference on Multimedia and ICT in Education (mICTE'09)*, 2009.

[10] M. A. Al-Barrak and M. Al-Razgan, "Predicting students' final gap using decision trees: A case study," *International Journal of Information and Education Technology*, vol. 6, no. 7, p. 528, 2016.

[11] E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, "Educational data mining: Predictive analysis of academic performance of public-school students in the capital of Brazil," *Journal of Business Research*, vol. 94, pp. 335–343, 2019.

[12] P. Rojanavasu, "Educational data analytics using association rule mining and classification," in *2019 Joint International Conference on Digital Arts, Media, and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT-NCON)*. IEEE, 2019, pp. 142–145.

[13] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, vol. 113, pp. 177–194, 2017.

[14] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study." *International Working Group on Educational Data Mining*, 2009.

[15] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Computers & Education*, vol. 103, pp. 1–15, 2016.

[16] P. Kaur, M. Singh, and G. S. Josan, "Classification and prediction-based data mining algorithms to predict slow learners in education sector," *Procedia Computer Science*, vol. 57, pp. 500–508, 2015.

[17] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Computer Science*, vol. 72, pp. 414–422, 2015, the Third Information Systems International Conference 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050915036182

[18] M. Alban and D. Mauricio, "Predicting university dropout through data mining: A systematic literature," *Indian Journal of Science and Technology*, vol. 12, no. 4, pp. 1–12, 2019.

[19] H. Altabrawee, O. A. J. Ali, and S. Q. Ajmi, "Predicting students' performance using machine learning techniques," *JOURNAL OF UNIVERSITY OF*

BABYLON for Pure and Applied Sciences, vol. 27, no. 1, pp. 194–205, 2019.

[20] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462–55470, 2020.

[21] A. A. Saa, "Educational data mining & students' performance prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, pp. 212–220, 2016.

[22] A. K. Pal and S. Pal, "Data mining techniques in edm for predicting the performance of students," *International Journal of Computer and Information Technology*, vol. 2, no. 06, 2013.

[23] L. Rokach and O. Maimon, "Decision trees," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2005, pp. 165–192.

[24] N. Naik and S. Purohit, "Comparative study of binary classification methods to analyze a massive dataset on virtual machine," *Procedia Computer Science*, vol. 112, pp. 1863–1870, 2017.

[25] H. Sharma and S. Kumar, "A survey on decision tree algorithms of classification in data mining," *International Journal of Science and Research (IJSR)*, vol. 5, no. 4, pp. 2094–2097, 2016.

[26] R. J. Lewis, "An introduction to classification and regression tree (cart) analysis," in *Annual meeting of the Society for Academic Emergency Medicine in San Francisco, California*, vol. 14, 2000.

[27] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[28] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, "Random forests and decision trees," *International Journal of Computer Science Issues (IJCSI)*, vol. 9, no. 5, p. 272, 2012.

[29] A. Cutler, D. R. Cutler, and J. R. Stevens, "Random forests," in *Ensemble Machine Learning*. Springer, 2012, pp. 157–175.

[30] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees (The Wadsworth statistics/probability series) Chapman and Hall," New York, NY, pp. 1–358, 1984.

[31] R. Amami, D. B. Ayed, and N. Ellouze, "Practical selection of svm supervised parameters with different feature representations for vowel recognition," *arXiv preprint arXiv:1507.06020*, 2015.

[32] K. S. Durgesh and B. Lekha, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, no. 1, pp. 1–7, 2010.

[33] H. Suo, M. Li, P. Lu, and Y. Yan, "Using svm as back-end classifier for language identification," EURASIP Journal on Audio, Speech, and Music Processing, vol. 2008, pp. 1–6, 2008.

[34] P. K. Singh and M. S. Husain, "Books reviews using naıve Bayes and clustering classifier," in Conference: Second International Conference on Emerging Research in Computing, Information, Communication and Applications'(ERCICA-14), 2014, pp. 886–891.

[35] B. Liu. Web data mining: exploring hyperlinks, contents, and usage data. Data-centric systems and applications. Springer, Berlin, 2007.

[36] Prof. Amruta Bijwar, Prof. Madhuri Zambre. (2018). Voltage Protection and Harmonics Cancellation in Low Voltage Distribution Network. International Journal of New Practices in Management and Engineering, 7(04), 01 - 07. https://doi.org/10.17762/ijnpme.v7i04.68

[37] Kumar, D. ., & Sonia, S. (2023). Resources Efficient Dynamic Clustering Algorithm for Flying Ad-Hoc Network. International Journal on Recent and Innovation Trends in Computing and Communication, 11(2s), 106–117. https://doi.org/10.17762/ijritcc.v11i2s.6034

[38] Sherje, N. P., Agrawal, S. A., Umbarkar, A. M., Dharme, A. M., & Dhabliya, D. (2021). Experimental evaluation of mechatronics based cushioning performance in hydraulic cylinder. Materials Today: Proceedings, doi:10.1016/j.matpr.2020.12.1021