Mehran University Research Journal of Engineering and Technology

https://doi.org/10.22581/muet1982.2301.12

2023, 42(1) 120-136

Quality enhancement at higher education institutions by early identifying students at risk using data mining

Khalid Mahboob [a,*], Raheela Asif [b], Najmi Ghani Haider [c]

[a] *Department of Computer Science and Information Technology, N.E.D University of Engineering and Technology, Karachi Pakistan*

[b] *Department of Software Engineering, N.E.D University of Engineering and Technology, Karachi Pakistan*

[c] *Department of Computing Sciences, U.I.T University, Karachi Pakistan*

[*] Corresponding author: Khalid Mahboob, Email: nedian07@cloud.neduet.edu.pk

Received: 31 March 2022, Accepted: 15 December 2022, Published: 01 January 2023

K E Y W O R D S

Students

Courses

Accuracy

Performance

At-risk

Data

A B S T R A C T

Accurate prediction of students' academic performance is one of the challenges in maintaining quality standards in any Higher Education Institution (H.E.I.). To ensure the quality of teaching and learning, H.E.I.s often employ Self-Assessment Reports (S.A.R.s) in which identifying a student drop-out ratio is important. Hence, it is essential to identify at-risk students in a given academic program. This article aims to identify at-risk students early by proposing a data mining-based predictive framework to improve the student's learning experience and minimize the dropped-out ratio. The academic sub-attributes or indicators in each course that may affect the performance of students in higher education institutions used in this study to examine students' academic achievement and predict students' performance to distinguish at-risk students are the marks of assignments, mid-term, lab exams, semester marks, total, grade, grade point (G.P.), quality point (Q.P.), grade point average (G.P.A.), and credit hours data of multiple courses categorized according to three knowledge areas defined by Higher Education Commission (H.E.C.), Pakistan using data mining predictive techniques. The results indicate that the proposed methods can achieve maximum accuracy in predicting and identifying at-risk students in different courses.

1. Introduction

Higher education institutions (HEIs) play a crucial role in any region's economic and social development. The drop and failure rate of students is one of the serious concerns which HEIs must continuously monitor to ensure the standards of the offered courses and overall program. Students' learning and prediction of its associated parameters, for example, drop rate, and students-at-risk, should be a critical subject that any quality assurance method must consider. It means the quality enhancement methods at HEIs must be student-learning central. Educational data on student grades contain a significant amount of hidden and latent information, the analysis of which can be instrumental in optimizing students' learning experience [1].

In this regard, a university maintains a massive volume of academic data for their students using offline (paper-based) and online electronic methods. Recent

advances in data mining allow universities to deal with educational data and the information generated to support teachers and students in intelligent decision-making so that student performance can be tracked, and hence the quality of education be ensured [2-4].

According to [5], [6], Education Data Mining (EDM) is referred to as the development of methods and models that can be used to extract useful information from data derived from the educational environment. Student academic performance in university should be a big concern not only for the students and parents but also for the university administration and faculty. In general, EDM searches for new data models and creates new algorithms and patterns. EDM involves collecting data on the student's academic background to understand how students interact with their university resources [6], [7].

Student academic performance is an integral part of HEIs, and students' evaluation is essential in maintaining student performance and the efficacy of the learning process. From the analysis of student performance, a better strategic program can be well-organized during their training in an institution [8]. Therefore, every higher education organization should maintain a database with all necessary relevant data. Storing student academic records is a part of the educational database [9].

This research aims to predict the student's academic performance during semesters using data mining techniques to identify valuable predictors of students at risk in the early years of their education. In this way, the quality of education can be improved at HEIs. The primary purpose of this study is to classify courses that affect the student's success in tackling low academic performance during different semesters, then use these courses as an initial estimate for the expected success rate and to deal with their weaknesses [10].

Thus, the study intends to show early indicators of students' poor performance that can be used to target corrective actions for troubled students. Besides, the research also aims to explore features that can be used for predictions and the type of classifiers that can give the best results [21]. The paper is organized as follows: Section 2 describes the related works. Section 3 covers the problem statement of the study. Section 4 presents the data collection and description. Section 5 explains the experimental techniques/methodology. Section 6 encompasses the results and discussion, and Section 7 concludes the paper.

2. Related Works

Much of the work regarding analyzing and predicting students' academic performance has been done using data mining. In this section, we present a brief overview of research work related to quality enhancement in education by evaluating at-risk students' academic performance by employing data mining techniques:

The NBTree classification approach has been used in [1] for predicting student performance. They used students' data, educational data, and admission data for mining data to determine the performance prediction model. The experiments are implemented with two-level classifications at the university and faculty levels. The results from the model indicated that some attributes, such as gender, credit, test score, and GPA, significantly impacted students' academic performance.

A multi-dimensional procedural approach employs in [2] to identify the factors of students' backgrounds. Ten classification tree techniques and a multilayer perceptron algorithm were used to experiment with students' enrolment records. A Random tree, amongst other algorithms, was outperformed and assumed as an optimal algorithm for the study.

Data mining classification algorithms used in [3] on the university data to disclose data high potency using data mining applications instead of university management commencing a data mining project UNWE. The results of the different classifications were later compared concerning accuracy. The students' admission data at the university and the number of failures in first-year university exams were identified as the most influencing factors for the classification process.

An analytical model in [7] was developed for predicting and understanding the causes behind fresh students' attrition using five years of academic data with data mining techniques such as support vector machines, decision trees, neural networks, and logistic regression. The comparative analysis results showed that the ensemble's performance was better than individual models, while the balanced dataset generated well-predictive results over the unbalanced dataset. A sensitivity analysis of the models shows that the educational and financial variables are among the most significant predictors of the event. Among the four techniques, support vector machines generated the best prediction results.

A systematic literature review was proposed in [8] on student performance prediction utilizing data mining

techniques to improve student's achievements in their education. Various data mining methods and student attributes were proposed for predicting students' performance. Neural networks and decision trees are two highly used methods by researchers to predict students' performance.

A pattern of student records in [10] was suggested, and courses available to predict students' performance. The study consists of two parts: First, to understand the factors related to the success of the course, and second, to define predictors centered on students' performance. Classification and clustering methods analyze different aspects that may affect student performance in the course(s).

Social and demographics feature at the school level in [11] are used as influencing factors on students' academic performance. They practice three data mining classification techniques, i.e., Naïve Bayes, J48 decision tree, and Three different classification techniques used in [12], namely Decision Trees, Naïve Bayes, and K-NN of multiple engineering disciplines students taking pre-examination marks only with a single course to track and analyze engineering students pedagogical progression in their studies.

A web-based system in [15] proposed utilizing the Naive Bayesian data mining technique for practical information extraction. The 700 student records have been used in an experiment taking students' academic history and demographics to prevent drop-out at early stages, reduce failure rates, and take acceptable actions for forthcoming semester exams.

The history of students in [16] was examined to access the Learning Management System (LMS) data. Classification techniques are used to build a learning model based on Knowledge Discovery in Databases (KDD) for predicting learning behaviors. They conclude that the J48 decision tree algorithm and multiple linear regression can be used to generate LMS predictions. A model can provide a powerful learning tool that systematically analyzes and predicts student performance in DE learners.

The ID3 decision tree algorithm in [17] uses for building prediction models to investigate academic performance, particularly for female students in a Bachelor's program in Information Technology. They identify the required courses that may affect student performance in the said program.

The EDM model in [20] predicts student performance in a programming course. The study includes factors such as students' mathematics background, programming, problem-solving skills, gender, mathematics high school grade, region, prior experience in computer programming, and e-learning practice using rule extraction.

A recommender system was proposed in [21] to predict at-risk students with poor results based on admission data and the module results of first year using different demographic characteristics, including gender, age, disability, nationality, etc. The goal was achieved with reasonable accuracy using classification models to specify students with low-performance achievement by high probability.

A reduced training vector-based support vector machine (RTV-SVM) was proposed in [22] to predict at-risk and marginal students. The technique was adopted because it eliminates unnecessary training vectors to decrease training time and support vectors. The process was applied to seven courses. The results show that the proposed method can achieve an overall accuracy of 92.2-93.8% and 91.3-93.5% in predicting at-risk and marginalized students, respectively.

Predictive methods have been compared in [23] for identifying at-risk students in standard-based grading courses. The prediction methods use only the semester-performing data available for the course instructors. The analysis shows that the Naive Bayes and Ensemble models using the order of models (e.g., K-Nearest Neighbours, Support Vector Machine, and Naive Bayes Classifiers) had the best results among the seven model methods tested.

3. Problem Statement

Students' academic performance evaluation is essential for student retention at HEI and to prevent educational dropout at early stages. Students' academic performance can be judged through their marks in different courses they studied during their degree or educational program [25]. For this purpose, in this case study, the courses are categorized into three main categories according to the knowledge areas defined by the Higher Education Commission (HEC), Pakistan, in the curriculum of the Software Engineering degree program. Therefore, the objective of this article is to answer the following issues quantitatively.

1) Can we identify the at-risk students in the technical course(s) during semesters to enhance quality education with reasonable accuracy?

2) Can we identify the students at risk in a non-technical course(s) during semesters to enhance quality education with reasonable accuracy?

3) Can we identify the at-risk students in the mathematical course(s) during semesters to enhance quality education with reasonable accuracy?

The study will be compelling in that it can support the practice of a performance management system for the students and the university. Suppose the students at risk can be identified at the initial stages in different courses based on their score results. In that case, necessary action can be taken based on the student's academic performance, implementing a performance management system that is more accessible and more supportive of dealing with students at risk [13], [14]. The flow of steps involved in a proposed data mining-based predictive framework is shown below in Fig. 1.

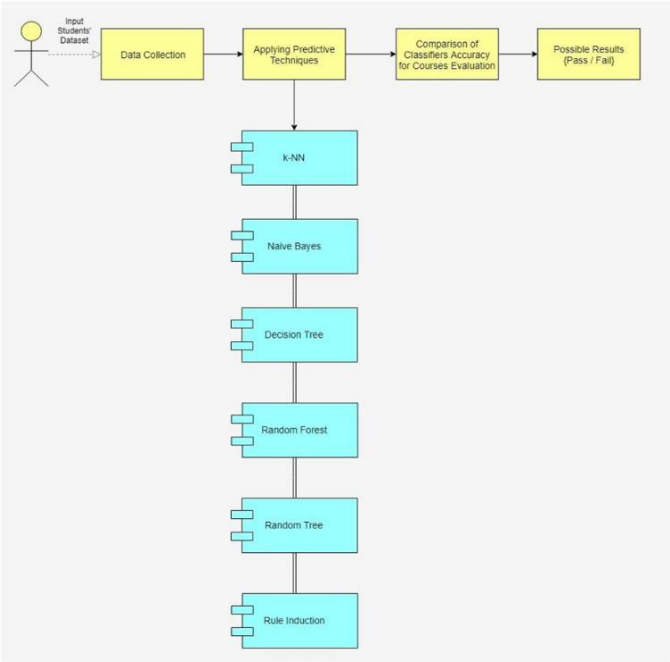

Fig. 1. Proposed data mining based predictive framework

4. Data Collection and Description

In this section, the analyses of two consecutive academic cohorts' data of the Software Engineering discipline at private sector H.E.I., Pakistan, which entailed overall 597 undergraduates enrolled in the academic batches with 255 students in batch 2016 and 342 students in batch 2017, have been collected for analyzing significant performances of students in academic courses that may affect academic performance. Semester sessional and semester examination marks have been considered for this purpose. Different courses are categorized as technical, non-technical, and mathematical category of courses with arbitrating sub-attributes for the evaluation of each course, including the marks of assignments, mid-term, lab exam, semester marks, total, grade, grade point (G.P.), quality point (Q.P.), grade point average (G.P.A.). Credit hours have been selected for assessing undergraduates' academic performance in a dataset [30]. The list of variables used in this study for the evaluation of technical courses is described in Table 1, the non-technical courses are described in Table 2, the mathematical courses are described in Table 3, and the list of each variable sub-attributes is described in Table 4.

Table 1

List of variables for Technical Courses

Variables	Course Description	Value
SWE-101	Introduction to Computing	P or F {Pass, Fail}
SWE-102	Programming Fundamentals	
SWE-103	Object Oriented Programming	
SWE-201	Introduction to Software Engineering	
SWE-203	Data Structure and Algorithm	
SWE-206	Digital Logic and Design	
SWE-202	Automata Theory and Formal Languages	
SWE-204	Operating Systems	
SWE-205	Software Requirement Engineering	
EE-110	Basic Electronics	
CE-207	Computer Organization and Architecture	

Table 2

List of variables for Non-Technical Courses

Variables	Course Description	Value
HS-101	Islamic Studies / Ethical Behavior	P or F {Pass, Fail}
HS-102	Technical English	
HS-103	Pakistan Studies and Aligarh Movement	
HS-201	Communication Skills	

Table 3

List of variables for Mathematical Courses

Variables	Course Description	Value
MS-103	Calculus and Analytical Geometry	P or F {Pass, Fail}
MS-105	Linear Algebra and Differential Equations	
MS-110	Applied Physics	
MS-204	Discrete Math	
MS-301	Probability and Statistics	

Table 4

List of Sub-Attributes of each Variable

Sub-Attributes	Values	Courses
Credit_Hours	{4}	SWE-101, SWE-102, SWE-103, MS-110, SWE-201, SWE-202, SWE-206, SWE-204, CE-207
	{3}	MS-103, HS-102, MS-105, HS-103, EE-110, HS-201, MS-204, MS-301, SWE-205
	{2}	HS-101, SWE-202
Assignments	{0-10}	SWE-101, SWE-102, SWE-103, EE-110, MS-110, SWE-201, SWE-203, SWE-206, SWE-204, CE-207, SWE-205
	{0-20}	MS-103, HS-102, HS-101, MS-105, HS-103, HS-201, MS-204, MS-301, SWE-202
Mid_Term	{0-20}	SWE-101, SWE-102, SWE-103, EE-110, MS-110, SWE-201, SWE-203, SWE-206, SWE-204, CE-207, SWE-205
	{0-30}	MS-103, HS-102, HS-101, MS-105, HS-103, HS-201, MS-204, MS-301, SWE-202
Lab_Exam	{0-20 | <10=F}	SWE-101, SWE-102, SWE-103, EE-110, MS-110, SWE-201, SWE-203, SWE-206, SWE-204, CE-207, SWE-205
Semester	{0-50 | <25=F}	All Courses
Total	{0-100 | <50=F}	All Courses
Grade	Grade_Point	% Marks Remarks
A+	4.00	90-100 Extra Ordinary
A	3.7-3.9	85-89 Excellent
A-	3.5-3.6	80-84 Very Good
B+	3.2-3.4	75-79 Good
B	3.0-3.1	70-74 Above Average
C+	2.5-2.9	65-69 Average
C	2.0-2.4	60-64 Satisfactory
D	1.0-1.9	50-59 Pass
F	0.00	0-49 Fail
Quality_Point	Grade_Point * Credit_Hours	
G.P.A	Quality_Point / Credit_Hours	

5. Experimental Techniques/Methodology

The following are the standard data mining predictive techniques employed to perform analyses of technical, non-technical, and mathematical academic courses to identify at-risk students during semesters. The RapidMiner, a data mining tool, is used to fulfill students' performance analyses of academic courses [18].

5.1 k-NN

k-NN is a lazy learning algorithm. Its objective is to use a database where data points are divided into multiple classes to predict the new sample point [24]. In this research, k-NN is used to classify an object (i.e., a student's w.r.t courses) with most K neighbors. The 11 closest neighbors (i.e., students' performances in technical courses), four most immediate neighbors (i.e., students' performances in 4 non-technical courses), and five closest neighbors (i.e., students' performances in 5 mathematical courses) have been applied to classify at-risk students. Euclidean distance between two points (x and y) has been calculated to obtain the nearest neighbor.

5.2 Naïve Bayes

It's easy and fast to forecast the test data levels. Regarding independence, the Naive Bayes classifier is better than others, such as logistic regression, and it requires fewer training data. It works well if variable input variables are equated to numerical variables [6, 26]. In this study, a Naïve Bayes classifier was used to contribute independently to the probability that the result will be either passed or failed in each course category.

5.3 Decision Tree

The decision tree provides a practical decision-making approach because it identifies the issue so that all

options are challenged. It can thoroughly analyze the possible consequences of the solution graphically [24]. In this analysis, each step separates data based on variables (i.e., course total marks) in each course category. In contrast, all data in each node has a single category label (e.g., P for pass or F for fail) or all variables employed. In this context, the Gain Ratio acronym GR and Gini Index acronym GI results are presented as a function of the estimation tree model training [26].

5.4 Random Forest

It creates numerous decision trees for multiple solutions and unites them to get a stable and accurate forecast [24]. The goal of Random Forest in this study is to construct a classification model that predicts target attribute values (e.g., P for pass or F for fail) for each course category based on the multiple input sub-attributes like marks of assignment, lab, mid-term examination, etc. The Gain Ratio GR and Gini Index GI results as measuring criteria are presented [26].

5.5 Random Tree

It behaves like a decision tree with one exception: only a random subcategory of attributes is accessible for each segment. The goal of Random Tree in this analysis is to create a classification model that estimates the label values (e.g., P for pass or F for fail) for each course category based on multiple input sub-attributes like marks of assignment, lab, mid-term examination, etc. The Gain Ratio GR and Gini Index GI results as measuring criteria are presented.

5.6 Rule Induction

It is a part of data mining, in which the formal rules are taken from the set of observations. Extracted rules can represent a complete scientific data model or describe the local model in the data. The rules have been removed for this study using the Information Gain acronym IG as a criterion to identify at-risk students' marks in various courses regarding whether they can pass or fail in a particular course.

The essential splitting parameters used as metrics during the process of decision-making are presented below.

1. Gain Ratio

It adjusts the reception of information for each attribute to allow the width and uniformity of the values.

2. Gini Index

This is a measure of the impurity of the dataset. The distribution of selected attributes provides a decrease in the average Gini index of the subsets obtained.

3. Information Gain

It calculates the entropy of all attributes, and the minimum entropy attribute is selected for the split. This method is biased in selecting attributes with many values.

6. Results and Discussion

The datasets comprising two Software Engineering cohorts have been analyzed consisting of 11 technical courses, four non-technical courses, and five mathematical courses with sub-attributes (as previously mentioned in Tables 1, 2, 3, and 4 in detail) for the academic session 2016-17 and 2017-18 taught during semesters. In this study, a student's academic performance will be a class or label P or F [19], representing a pass or fail in three categories: technical courses, non-technical courses, and mathematical courses encompassing semester records. Class or label permits division between strong and weak performances in different courses. The datasets of two cohorts are compared regarding range (minimum to maximum marks obtained) for all three categories, i.e., technical courses, non-technical courses, and mathematical courses are presented in Tables 5, 6, and 7.

Table 5

Comparison of datasets for Technical Courses

Predictors	Range of Dataset 1	Range of Dataset 2
SWE-101	[0; 97]	[0;93]
SWE-102	[0; 91]	[0;91]
SWE-103	[0; 97]	[0;97]
SWE-201	[0; 89]	[0;85]
SWE-203	[0; 96]	[0;95]
SWE-206	[0; 99]	[0;95]
SWE-202	[0; 100]	[0;100]
SWE-204	[0; 89]	[0;95]
SWE-205	[0; 97]	[0;91]
EE-110	[0; 88]	[0;99]
CE-207	[0; 97]	[0;99]

Table 6

Comparison of datasets for Non-Technical Courses

Predictors	Range of Dataset 1	Range of Dataset 2
HS-101	[0; 100]	[0;99]
HS-102	[0; 85]	[0;88]
HS-103	[0; 85]	[0;88]
HS-201	[0; 84]	[0;83]

Table 7

Comparison of datasets for Mathematical Courses

Predictors	Range of Dataset 1	Range of Dataset 2
MS-103	[0; 94]	[0;100]
MS-105	[0; 95]	[0;100]
MS-110	[0; 88]	[0;85]
MS-204	[0; 91]	[0;97]
MS-301	[0; 94]	[0;96]

As discussed in the previous section, the two datasets have been designed for the gathered data, i.e., Dataset 1 and 2. 70% of the academic data in both datasets are used for training, and 30% is used for testing. RapidMiner is used for research, statistical analysis, and data mining [18]. Each course with sub-attributes in each category is analyzed using different predictive techniques. The overall results for all three-course categories are obtained concerning the accuracy, precision, and recall using different classifiers [9]. The accuracy is the proportion of the total number of correct predictions, as shown in Eq. 1.

$$Accuracy = \frac{(Correct\ predictions)}{(Number\ of\ instances)} = \frac{(TP+TN)}{(TP+FP+F\ N+TN\ )} \quad (1)$$

The precision is the proportion of positive instances that were correctly identified as shown in Eq. 2.

$$Precision = \frac{(True\ positive\ predictions)}{(Total\ positive\ predictions)} = \frac{TP}{(TP+FP)} \quad (2)$$

The recall is the proportion of actual positive instances which were correctly identified as shown in Eq. 3.

$$Recall = \frac{(True\ positive\ predictions)}{(Number\ of\ positive\ instances)} = \frac{TP}{(TP+FN)} \quad (3)$$

The summary of comparison results of accuracy, precision, and recall of both datasets using different predictive techniques for technical courses, non-technical courses, and mathematical courses are mentioned in Tables 8, 9, and 10 correspondingly.

Table 8

Comparison of Analysis for Technical Courses

Techniques	Dataset 1			Dataset 2		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
k-NN	90.98%	97.32%	88.41%	98.54%	94.90%	100.00%
Naïve Bayes	83.92%	98.43%	76.22%	99.71%	100.00%	98.92%
Decision Tree with GR	95.29%	96.91%	95.73%	97.08%	94.62%	94.62%
Decision Tree with GI	94.51%	96.88%	94.51%	96.49%	93.55%	93.55%
Random Forest with GR	85.10%	89.87%	86.59%	93.86%	89.13%	88.17%
Random Forest with GI	85.88%	90.00%	87.80%	94.74%	92.13%	88.17%
Random Tree with GR	80.00%	91.85%	75.61%	83.33%	66.36%	78.49%
Random Tree with GI	78.43%	85.16%	80.49%	82.46%	67.37%	68.82%
Rule Induction with IG	92.16%	93.37%	94.51%	94.44%	87.00%	93.55%

Table 9

Comparison of Analysis for Non-Technical Courses

Techniques	Dataset 1			Dataset 2		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
k-NN	95.29%	96.88%	86.11%	99.71%	99.07%	100.00%
Naïve Bayes	97.65%	97.14%	94.44%	98.83%	96.40%	100.00%
Decision Tree with GR	97.65%	95.83%	95.83%	97.95%	100.00%	93.46%
Decision Tree with GI	95.29%	95.45%	87.50%	98.25%	99.03%	95.33%
Random Forest with GR	97.65%	97.14%	94.44%	99.12%	99.06%	98.13%
Random Forest with GI	96.08%	95.59%	90.28%	98.83%	99.05%	97.20%
Random Tree with GR	96.08%	96.97%	88.89%	95.91%	96.04%	90.65%
Random Tree with GI	92.94%	89.71%	84.72%	95.32%	93.33%	91.59%
Rule Induction with IG	90.98%	98.04%	69.44%	93.57%	96.70%	82.24%

Table 10

Comparison of Analysis for Mathematical Courses

Techniques	Dataset 1			Dataset 2		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
k-NN	95.29 %	96.88 %	86.11 %	99.71 %	99.07%	100.00 %
Naïve Bayes	97.65 %	97.14 %	94.44 %	98.83 %	96.40%	100.00 %
Decision Tree with GR	97.65 %	95.83 %	95.83 %	97.95 %	100.00 %	93.46%
Decision Tree with GI	95.29 %	95.45 %	87.50 %	98.25 %	99.03%	95.33%
Random Forest with GR	97.65 %	97.14 %	94.44 %	99.12 %	99.06%	98.13%
Random Forest with GI	96.08 %	95.59 %	90.28 %	98.83 %	99.05%	97.20%
Random Tree with GR	96.08 %	96.97 %	88.89 %	95.91 %	96.04%	90.65%
Random Tree with GI	92.94 %	89.71 %	84.72 %	95.32 %	93.33%	91.59%
Rule Induction with IG	90.98 %	98.04 %	69.44 %	93.57 %	96.70%	82.24%

The comparison results of accuracy with nine different classifiers for all three-course categories are depicted in Figs. 2, 3, and 4 on both datasets. From Fig. 2, it is clearly shown that each classifier attained the highest results regarding accuracy on dataset 2 as compared to dataset 1 for technical courses. The Naïve Bayes obtained the highest accuracy in comparing two datasets among nine techniques. In the same way, Random Tree with a Gini index shows the least accuracy for dataset 1 compared to dataset 2, among other methods. Fig. 3 shows that, again, classifiers' accuracy results are high on dataset 2 compared to dataset 1 for non-technical courses. The k-NN achieved the highest accuracy among overall comparisons between the two datasets. Likewise, the rule induction with information gain has achieved the least accuracy in general comparisons. Similarly, for mathematical courses, as shown in Fig. 4, the classifiers attained maximum accuracy on dataset 2 in contrast to dataset 1. The Naïve Bayes once more performed best in accuracy than the rest of the classifiers, whereas; Random Tree with gain ratio performed less for overall comparison.

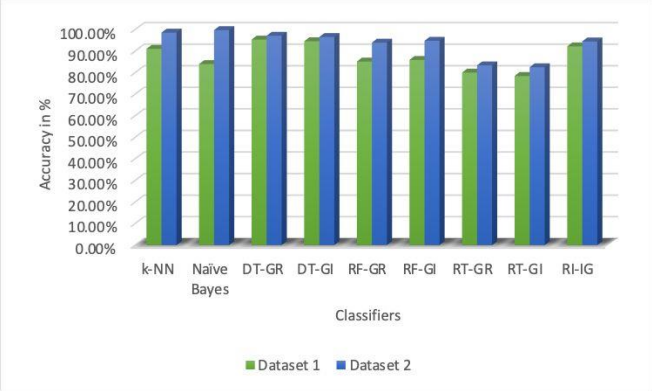

Fig. 2. Comparison in terms of accuracy of technical courses

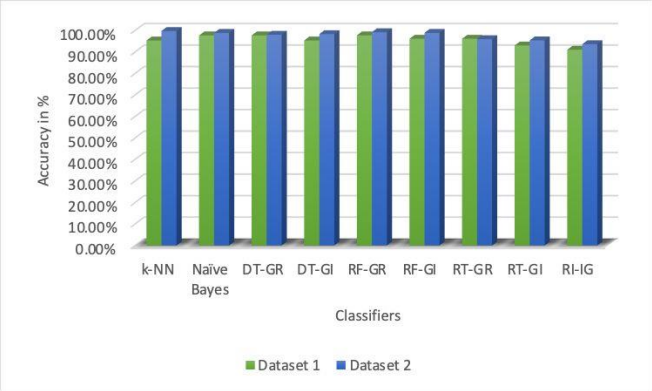

Fig. 3. Comparison in terms of accuracy of non-technical courses

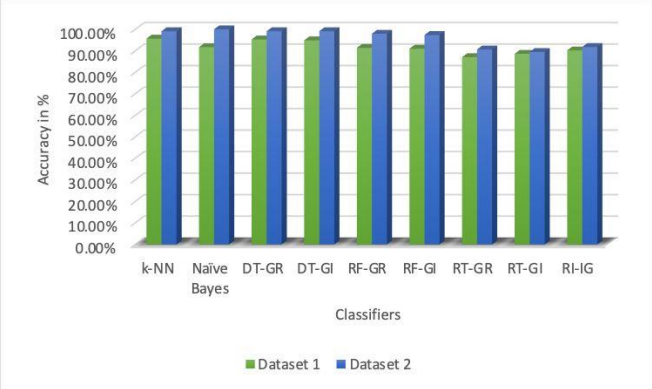

Fig. 4. Comparison in terms of accuracy of mathematical courses

After determining the best data mining predictive techniques, the robustness of the selected modelling techniques to the training scale of the dataset was investigated. Data volume can lead to the resulting low accuracy of the models [27]. However, an extensive training data set results in over-fitting the model to the training dataset and decreasing predicting accuracy. The classifiers' accuracy on the datasets shows how helpful these models are in identifying students at risk in forthcoming semesters [23]. Classifiers generally have given better results on dataset 2, probably because of the

more extensive dataset. There could be more instances in a dataset of better model training. The resultant confusion matrices of technical, non-technical, and mathematical courses comprising both datasets used in this case study are shown in Tables 11, 12, and 13, respectively. To understand the results of confusion matrices, let's consider the example of the classifier "k-NN" of a technical course category. There are 91 (81+10) of the actual class 'P' students: the classifier predicted 81 correctly as 'P' and ten wrongly as 'F' instances. Similarly, there are 164 (16+148) of the actual class 'F' students: the classifier predicted 148 correctly as 'F' and 16 wrongly as 'P' instances, and so on. All correct predictions are highlighted along the diagonals of the table [26].

Table 11

Confusion Matrices of Technical Courses

		Dataset I			Dataset II		
k-NN		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	81	16	83.51%	248	1	99.60%
	F	10	148	93.67%	1	92	98.92%
Class Recall		89.01%	90.24%		99.60%	98.92%	
Naïve Bayes		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	89	31	74.17%	249	1	99.60%
	F	2	133	98.52%	0	92	100.00%
Class Recall		97.80%	81.10%		100.00%	98.92%	
Decision Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	87	5	94.57%	244	7	97.21%
	F	4	159	97.55%	5	86	94.51%
Class Recall		95.60%	96.95%		97.99%	92.47%	
Decision Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	84	9	90.32%	243	7	97.20%
	F	7	155	95.68%	6	86	93.48%
Class Recall		92.31%	94.51%		97.59%	92.47%	
Random Forest with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	81	18	81.82%	235	13	94.76%
	F	10	146	93.59%	14	80	85.11%
Class Recall		89.01%	89.02%		94.38%	86.02%	
Random Forest with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	83	22	79.05%	239	14	94.47%
	F	8	142	94.67%	10	79	88.76%
Class Recall		91.21%	86.59%		95.98%	84.95%	
Random Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	84	30	73.68%	217	20	91.56%
	F	7	134	95.04%	32	73	69.52%
Class Recall		92.31%	81.71%		87.15%	78.49%	
Random Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	70	20	77.78%	226	16	93.39%

	F	21	144	87.27%	23	77	77.00%
	Class Recall	76.92%	87.80%		90.76%	82.80%	

Rule Induction with Information Gain		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	79	16	83.16%	237	7	97.13%
	F	12	148	92.50%	12	86	87.76%
Class Recall		86.81%	90.24%		95.18%	92.47%	

Table 12

Confusion Matrices of Non-Technical Courses

		Dataset I			Dataset II		
k-NN		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	181	7	96.28%	234	1	99.57%
	F	2	65	97.01%	1	106	99.07%
Class Recall		98.91%	90.28%		99.57%	99.07%	
Naïve Bayes		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	181	4	97.84%	231	0	100.00%
	F	2	68	97.14%	4	107	96.40%
Class Recall		98.91%	94.44%		98.30%	100.00%	
Decision Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	179	3	98.35%	234	2	99.15%
	F	4	69	94.52%	1	105	99.06%
Class Recall		97.81%	95.83%		99.57%	98.13%	
Decision Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	180	6	96.77%	233	2	99.15%
	F	3	66	95.65%	2	105	98.13%
Class Recall		98.36%	91.67%		99.15%	98.13%	
Random Forest with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	181	7	96.28%	233	10	95.88%
	F	2	65	97.01%	2	97	97.98%
Class Recall		98.91%	90.28%		99.15%	90.65%	
Random Forest with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	180	5	97.30%	233	7	97.08%
	F	3	67	95.71%	2	100	98.04%
Class Recall		98.36%	93.06%		99.15%	93.46%	
Random Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	181	8	95.77%	230	8	96.64%
	F	2	64	96.97%	5	99	95.19%
Class Recall		98.91%	88.89%		97.87%	92.52%	
Random Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	178	11	94.18%	228	6	97.44%
	F	5	61	92.42%	7	101	93.52%
Class Recall		97.27%	84.72%		97.02%	94.39%	
Rule Induction with Information Gain		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	180	17	91.37%	233	22	91.37%
	F	3	55	94.83%	2	85	97.70%
Class Recall		98.36%	76.39%		99.15%	79.44%	

Table 13

Confusion Matrices of Mathematical Courses

		Dataset I			Dataset II		
k-NN		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	98	7	93.33%	208	0	100.00%
	F	7	143	95.33%	0	134	100.00%
Class Recall		93.33%	95.33%		100.00%	100.00%	
Naïve Bayes		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	102	18	85.00%	208	0	100.00%
	F	3	132	97.78%	0	134	100.00%
Class Recall		97.14%	88.00%		100.00%	100.00%	
Decision Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	101	9	91.82%	208	0	100.00%
	F	4	141	97.24%	0	134	100.00%
Class Recall		96.19%	94.00%		100.00%	100.00%	
Decision Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	101	10	90.99%	208	0	100.00%
	F	4	140	97.22%	0	134	100.00%
Class Recall		96.19%	93.33%		100.00%	100.00%	
Random Forest with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	98	9	91.59%	189	0	100.00%
	F	7	141	95.27%	19	134	87.58%
Class Recall		93.33%	94.00%		90.87%	100.00%	
Random Forest with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	97	11	89.81%	199	10	95.22%
	F	8	139	94.56%	9	124	93.23%
Class Recall		92.38%	92.67%		95.67%	92.54%	
Random Tree with Gain Ratio		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	89	21	80.91%	180	2	98.90%
	F	16	129	88.97%	28	132	82.50%
Class Recall		84.76%	86.00%		86.54%	98.51%	
Random Tree with Gini Index		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	91	26	77.78%	186	21	89.86%
	F	14	124	89.86%	22	113	83.70%
Class Recall		86.67%	82.67%		89.42%	84.33%	
Rule Induction with Information Gain		Actual			Actual		
		P	F	Class Precision	P	F	Class Precision
Predicted	P	97	16	85.84%	191	5	97.45%
	F	8	134	94.37%	17	129	88.36%
Class Recall		92.38%	89.33%		91.83%	96.27%	

Further, the representations of decision trees with gain ratio and Gini index for K=11, K=4, and K=5 for both datasets are presented below. The successful predictors' results from different classifiers show that the students are less at-risk in these courses. The tree in Fig. 5 indicates that SWE-205, Software Requirement Engineering, is a successful predictor of students' academic performance during semesters in a technical course category for dataset 1. This suggests that the students who scored marks greater than at least 52% are more likely to pass a semester examination. The tree in Fig. 6 indicates that the course SWE-204, Operating Systems is a successful predictor of students' academic

performance during semesters in a technical course category for dataset 1. This suggests that the students who scored marks greater than 49.5% are more likely to pass a semester examination.

The tree in Fig. 7 specifies that the course HS-201, Communication Skills, is a successful predictor of students' academic performance during semesters in a non-technical course category for dataset 1. This proposes that the students who scored marks more significant than 41% are more likely to pass a semester examination. The tree in Fig. 8 specifies that the course HS-102, Technical English, is a successful predictor of students' academic performance during semesters in a non-technical course category for dataset 1. This proposes that the students who scored marks more significant than 46% are more likely to pass a semester examination.

The tree in Fig. 9 shows that the course MS-105, Linear Algebra, and Differential Equations, is a successful predictor of students' academic performance during semesters in a mathematical course category for dataset 1. This recommends that the students who scored marks greater than 48.5% are more likely to pass a semester examination. The tree in Fig. 10 shows again the course MS-105, which is Linear Algebra and Differential Equations, is a successful predictor of students' academic performance during semesters in a mathematical course category for dataset 1. This recommends that the students who scored marks greater than 48.5% are more likely to pass a semester examination [28].

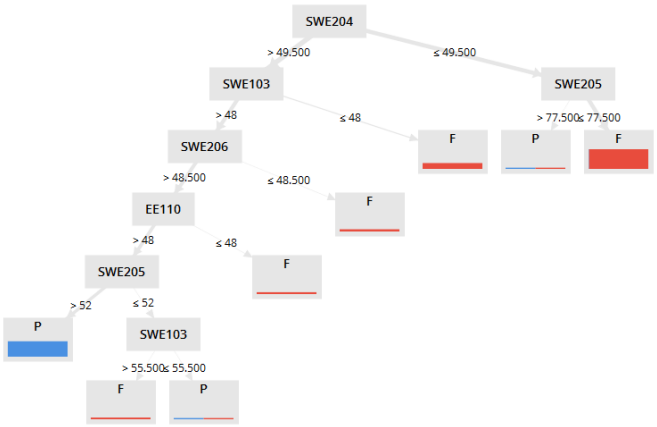

Fig. 5. Decision tree with gain ratio for dataset 1.

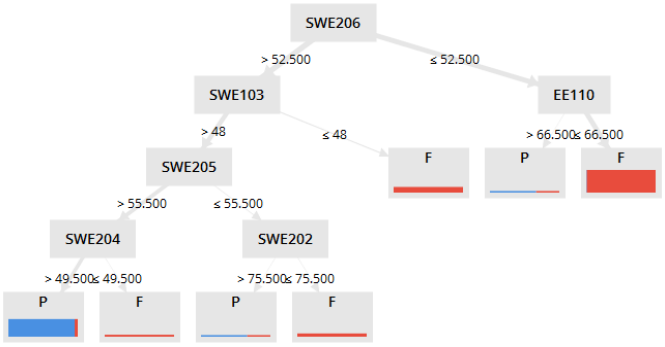

Fig. 6. Decision tree with Gini index for dataset 1

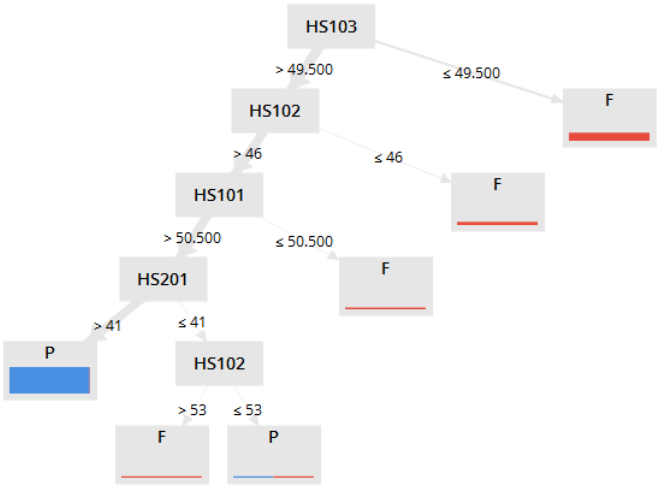

Fig. 7. Decision tree with gain ratio for dataset 1

Fig. 8. Decision tree with Gini index for dataset 1

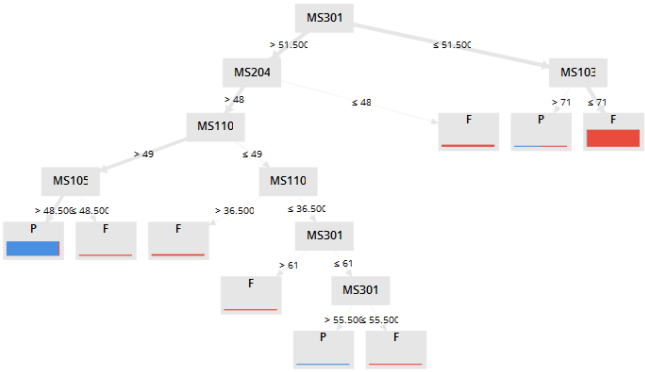

Fig. 9. Decision tree with gain ratio for dataset 1

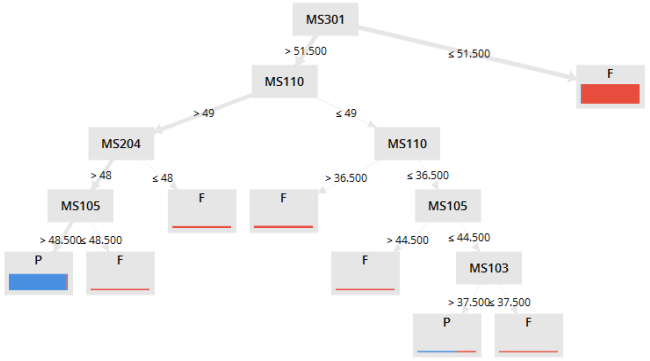

Fig. 10. Decision tree with Gini index for dataset 1

The tree in Fig. 11 indicates that the SWE-203, Data Structure, and Algorithm, is a successful predictor of students' academic performance during semesters in a technical course category for dataset 2. This suggests that the students who scored marks greater than 25.5% are more likely to pass a semester examination. The tree in Fig. 12 indicates that the course SWE-202, Automata Theory and Formal Languages, is a successful predictor of students' academic performance during semesters in a technical course category for dataset 2. This suggests that the students who scored marks greater than 27.5% are more likely to pass a semester examination.

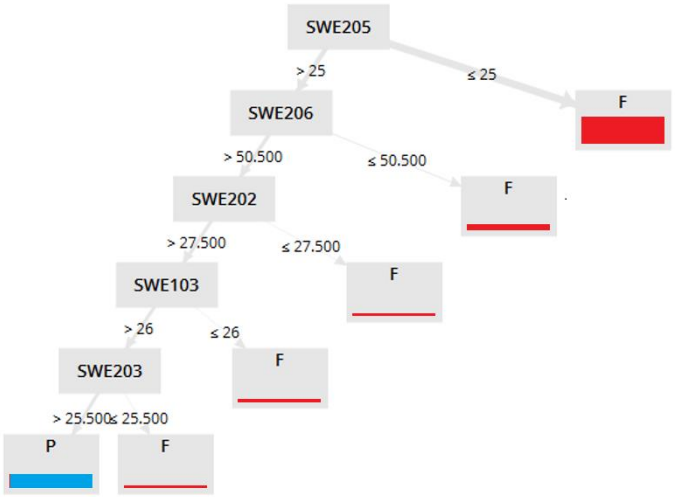

Fig. 11. Decision tree with gain ratio for dataset 2

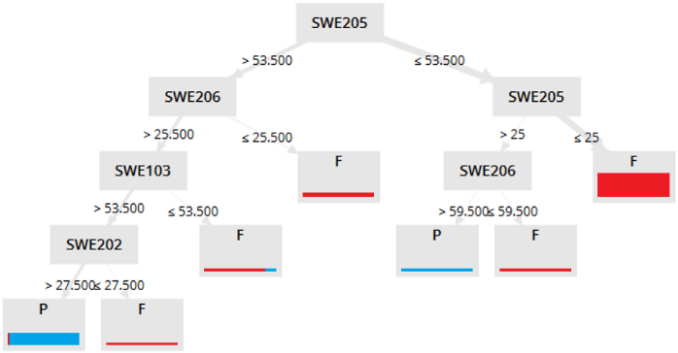

Fig. 12. Decision tree with Gini index for dataset 2

The tree in Fig. 13 specifies that the course HS-102, Technical English, is a successful predictor of students' academic performance during semesters in a non-technical course category for dataset 2. This proposes that the students who scored marks greater than 0.5% are more likely to pass a semester examination. The tree in Fig. 14 specifies the same result as the decision tree in Fig. 13.

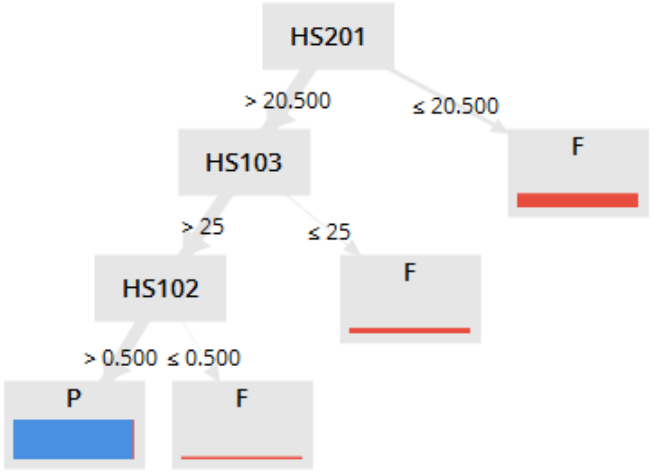

Fig. 13. Decision tree with Gain ratio for dataset 2

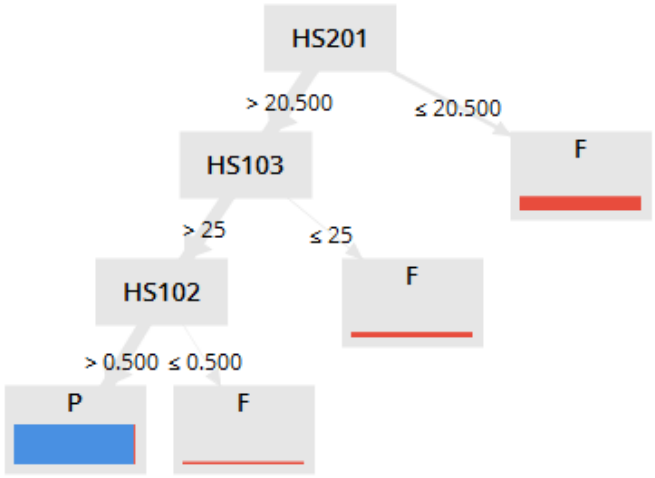

Fig. 14. Decision tree with Gini index for dataset 2

The tree in Fig. 15 shows that the course MS-204, Discrete Math, is a successful predictor of students' academic performance during semesters in a mathematical course category for dataset 2. This recommends that the students who scored marks more significant than 25% are more likely to pass a semester examination. The tree in Fig. 16 shows the same result as the decision tree in Fig. 15.

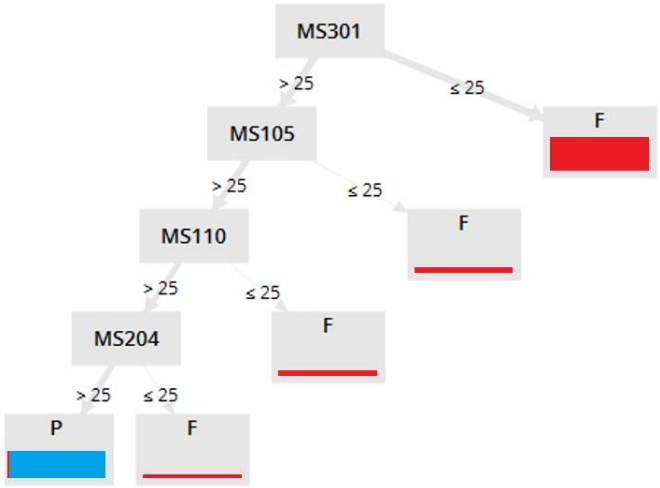

Fig. 15. Decision tree with gain ratio for dataset 2

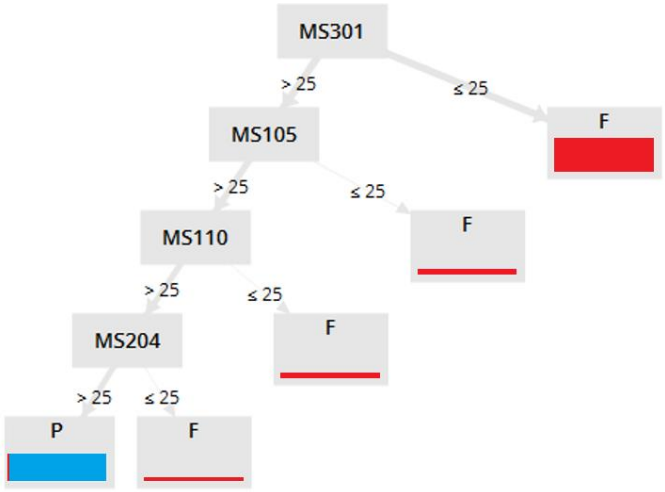

Fig. 16. Decision Tree with Gini Index for Dataset 2

The best rules extracted from rule induction with information gain for both datasets are shown in Tables 14 and 15.

Table 14

Results of Rule Induction for Dataset 1

Technical Courses Rules	if SWE206 ≤ 53.500 then F (5 / 119) if SWE103 > 48 and SWE202 > 62.500 then P (75 / 6) if SWE103 ≤ 48 then F (0 / 25) if SWE205 ≤ 62 then F (0 / 9) if SWE206 > 57.500 then P (10 / 1) if SWE204 ≤ 61 then F (0 / 4) else P (0 / 0)
Non-Technical Courses Rules	if HS201 > 51.500 then P (173 / 13) if HS103 ≤ 54 then F (0 / 44) if HS201 ≤ 41 then F (1 / 14) else P (8 / 0)
Mathematical Courses Rules	if MS301 ≤ 51.500 then F (2 / 123) if MS103 > 59.500 then P (83 / 6) if MS110 ≤ 49 then F (2 / 13) if MS105 > 48.500 and MS204 > 45.500 then P (18 / 1) else F (0 / 6)

Table 15

Results of Rule Induction for Dataset 2

Technical Courses Rules	if SWE205 ≤ 25 then F (207 / 0) if SWE206 > 54.500 then P (9 / 91) else F (29 / 1)
Non-Technical Courses Rules	if HS201 > 57.500 then P (214 / 13) if HS201 ≤ 20.500 then F (0 / 77) if HS103 > 50.500 and HS102 > 52.500 then P (19 / 1) else F (1 / 14)
Mathematical Courses Rules	if MS301 ≤ 25 then F (169 / 0) if MS110 > 56.500 then P (13 / 112) if MS204 ≤ 60.500 and MS301 > 57.500 then F (11 / 2) else P (13 / 18)

From the above analysis, all three research questions are answered positively. Different classifiers have shown different results based on identifying strong and weak academic performances in technical, non-technical, and mathematical courses with reasonable accuracy for the Software Engineering students that play a vital role in achieving and improving grades or G.P.A [29]. The core technical courses SWE-101 and SWE-102 taught in the first semester as baseline courses have been identified as low-performance predictors in the above analysis, which means the students are likely at-risk in these courses. There may be a chance of semester drop-out in a too early stage. Similarly, the core technical course SWE-103 taught in a second semester is considered an average performance predictor. Again, the base core technical course SWE-201 taught in the third semester is imparted as a low-performance

predictor indicating the students may be at-risk in a core course.

However, the results obtained from other classifiers such as Random Forest and Random Tree with gain ratio and Gini index should be presented in this study due to the numerous huge trees. Still, for the courses SWE-101 and SWE-103, Random Forest with gain ratio and Gini index has been a good performance predictor. In comparison, Random Forest with Gini index and Random Tree with Gini index has discovered the course SWE-102 as a good performance predictor. The course SWE-201 with Random Forest with gain ratio and Gini index and Random Tree with Gini index is a good performance predictor.

Although university passing criteria for the course is a minimum of 50 marks in total and a minimum of 25 marks in-sessional with at-least ten marks in a lab in a course with a lab, here, the assumption in a data list is that if a student fails in any one of the technical, non-technical, or mathematical courses, they will be treated as a fail. It is noteworthy that some results of the classifiers need to agree with the course passing criteria. So, no course in any of the course categories is essentially characterized as a successful academic predictor.

The present study is different from [10-12], [15], [17], [19], [29] in a view that many courses consider both sessional and semester final examination marks as well as other assessment parameters like course credit hours, grade point, quality point, and G.P.A with some more classifiers, are used to determine students' academic predictors that may affect their academic results to minimize drop-out ratio to enhance the quality of education in distinguished courses. Further, the academic courses are categorized into technical (considered core) courses, non-technical courses, and mathematical courses to find out how well students have performed and determine their strengths and difficulties in different courses based on their academic results.

7. Conclusion and Future Recommendation

The goal of this study is to identify at-risk students early to reduce the chances of academic drop-out and attrition by analyzing and predicting their academic performances in various courses used to teach Software Engineering technology at private sector H.E.I using predictive data mining techniques. Different academic courses taught during semesters have been classified into three categories: technical courses, non-technical courses, and mathematical courses with pre-examination and final semester examination marks belonging to two consecutive cohorts have been analyzed to improve students' learning experience in different courses. The most influential sub-attributes as variables in each course were selected for the analysis. Note that this study's sub-attributes results are not presented due to large data computations. So, only the results based on the total marks are presented in this study. Further, the results show credit hours do not directly affect a course's performance.

In both datasets, 30% of the academic data is utilized for testing, while 70% is used for training. Different classifiers predicted students' strong and weak academic performances in multiple courses in the three-course categories, positively answering the research questions. The results of the study show that the Naïve Bayes classifier achieved the maximum accuracy on both technical and mathematical courses categories respectively. While the k-NN classifier performed the best in the non-technical courses category. However, some results of the classifiers fail to meet the university passing criterion of the courses. Through these analyses, the students at risk can be identified in the early stages to not only prevent students from the semester dropping out but also to improve the efficacy of academic courses. Implementing the proposed data mining-based predictive framework sounds important in H.E.Is to properly abreast students by identifying their weaknesses and solving their problems and thus updating and improving academic curriculum systematically with the help of the educators and learners. In general, the techniques or methodology used in this study can be applied to any engineering and non-engineering disciplines at H.E.Is to identify the students at risk in the early stages.

In the future, faculty or instructor performances can also be analyzed by correlating the students' academic performance and faculty/instructor performance to evaluate students' insight in different courses and to improve students overall learning experience at H.E.Is and hence enhance the quality of education using data mining techniques.

8. References

[1] T. M. Christian and M. Ayub, "Exploration of classification using NBTree for predicting students' performance", International Conference on Data and Software Engineering, Bandung, Indonesia, pp. 1–6, 2014.

[2] M. Goga, S. Kuyoro, and N. Goga, "A

recommender for improving the student academic performance", Procedia - Social and Behavioral Sciences, vol. 180, no. November 2014, pp. 1481–1488, 2015.

[3] D. Kabakchieva, "Predicting student performance by using data mining methods for classification", Cybernetics and Information Technologies, vol. 13, no. 1, pp. 61–72, 2013.

[4] P. Guleria, M. Arora, and M. Sood, "Increasing quality of education using educational data mining", 2013 2nd International Conference on Information Management in the Knowledge Economy, pp. 118–122, 2013.

[5] P. Akulwar, S. Pardeshi, and A. Kamble, "Survey on different data mining techniques for prediction", 2nd International Conference on IoT in Social, Mobile, Analytics and Cloud, Palladam, India, pp. 513–519, 2019.

[6] S. S. Athani, S. A. Kodli, M. N. Banavasi, and P. G. S. Hiremath, "Student academic performance and social behavior predictor using data mining techniques", 2017 International Conference on Computing, Communication and Automation (ICCCA), pp. 170–174, 2017.

[7] D. Delen, "A comparative analysis of machine learning techniques for student retention management", Decision Support Systems, vol. 49, no. 4, pp. 498–506, 2010.

[8] 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.

[9] C. E. Lopez Guarin, E. L. Guzman, and F. A. Gonzalez, "A model to predict low academic performance at a specific enrollment using data mining", Revista Iberoamericana de Tecnologias del Aprendizaje, vol. 10, no. 3, pp. 119–125, 2015.

[10] N. A. Yassein, R. G. M Helali, and S. B. Mohomad, "Predicting student academic performance in KSA using data mining techniques", Journal of Information Technology and Software Engineering, vol. 07, no. 05, 2017.

[11] S. Roy and A. Garg, "Predicting academic performance of student using classification techniques", 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON 2017, vol. 2018-Janua, pp. 568–572, 2017.

[12] K. Mahboob, S. A. Ali, D. U. R. Khan, and F. Ali, "A comparative study of engineering students pedagogical progress", International Journal of Advanced Computer Science and Applications, vol. 9, no. 6, pp. 323–331, 2018.

[13] Z. J. Kovacic, "Predicting student success by mining enrolment data", Research in Higher Education Journal, vol. 15, no. iii, p. 1, 2012.

[14] E. Aguiar, N. V. Chawla, J. Brockman, G. A. Ambrose, and V. Goodrich, "Engagement vs performance: using electronic portfolios to predict first semester engineering student retention", ACM International Conference Proceeding Series, pp. 103–112, 2014.

[15] T. Devasia, T. P. Vinushree, and V. Hegde, "Prediction of students performance using educational data mining", Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, pp. 91–95, 2016.

[16] B. E. V. Comendador, L. W. Rabago, and B. T. Tanguilig, "An educational model based on knowledge discovery in databases (KDD) to predict learner's behavior using classification techniques", IEEE International Conference on Signal Processing, Communications and Computing, Conference Proceedings, pp. 1–6, 2016.

[17] Y. Altujjar, W. Altamimi, I. Al-Turaiki, and M. Al-Razgan, "Predicting critical courses affecting students performance: a case study", Procedia Computer Science, vol. 82, no. March, pp. 65–71, 2016.

[18] RapidMiner, 19-Oct-2021. [Online]. Available: https://rapidminer.com/.

[19] K. Deepika and N. Sathyanarayana, "Analyze and predicting the student academic performance using data mining tools", Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS 2018), pp. 76–81, 2018.

[20] A. F. Gamal, "An educational data mining model for predicting student performance in programming course", International Journal of Computer Applications, vol. 70, no. 17, pp. 22-28, 2013.

[21] Z. Alharbi, J. Cornford, L. Dolder, and B. De La Iglesia, “Using data mining techniques to predict students at risk of poor performance”, Proceedings of 2016 SAI Computing Conference, SAI 2016, pp. 523–531, 2016.

[22] K. T. Chui, D. C. L. Fung, M. D. Lytras, and T. M. Lam, “Predicting at-risk university students in a virtual learning environment via a machine learning algorithm”, Computers in Human Behavior, vol. 107, 2018.

[23] 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 and Education, vol. 103, pp. 1–15, 2016.

[24] E. Wakelam, A. Jefferies, N. Davey, and Y. Sun, “The potential for student performance prediction in small cohorts with minimal available attributes,” British Journal of Educational Technology, vol. 51, no. 2, pp. 347–370, 2019.

[25] A. Behr, M. Giese, H. D. Teguim Kamdjou, and K. Theune, “Dropping out of university: a literature review”, Review of Education, vol. 8, no. 2, pp. 614–652, 2020.

[26] R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, “Analyzing undergraduate students’ performance using educational data mining”, Computers and Education, vol. 113, pp. 177–194, 2017.

[27] J. Vinas-Forcade, C. Mels, M. Van Houtte, M. Valcke, and I. Derluyn, “Can failure be prevented? Using longitudinal data to identify at-risk students upon entering secondary school”, British Educational Research Journal, 2020.

[28] M. A. Hebert, S. R. Powell, J. Bohaty, and J. Roehling, “Piloting a mathematics-writing intervention with late elementary students at-risk for learning difficulties”, Learning Disabilities Research and Practice, vol. 34, no. 3, pp. 144–157, 2019.

[29] R. Dorta-Guerra, I. Marrero, B. Abdul-Jalbar, R. Trujillo-González, and N. V. Torres, “A new academic performance indicator for the first term of first-year science degrees students at La Laguna University: a predictive model”, FEBS Open Bio, vol. 9, no. 9, pp. 1493–1502, 2019.

[30] P. Crowther and S. Briant, “Predicting academic success: a longitudinal study of university design students”, International Journal of Art and Design Education, vol. 40, no. 1, pp. 20–34, 2021.

