Early Detection of Student Degree-level Academic performance using educational data mining

Areej Fatemah Meghji [1], **Naeem Ahmed Mahoto** [1], **Yousef Asiri** [Corresp., 2], **Hani Alshahrani** [2], **Adel Sulaiman** [2], **Asadullah Shaikh** [3]

[1] Department of Software Engineering, Mehran University of Engineering and Technology Jamshoro, Hyderabad, Jamshoro, Pakistan

[2] Department of Computer Science, College of Computer Science and Information Systems, Najran University, Najran, Najran, Saudi Arabia

[3] Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran, Najran, Saudi Arabia

Corresponding Author: Yousef Asiri
Email address: yasiri@nu.edu.sa

Higher educational institutes generate massive amounts of student data. This data needs to be explored in-depth to understand various facets of student learning behavior better. The educational data mining approach has given provisions to extract useful and non-trivial knowledge from large collections of student data. Using the educational data mining method of classification, this research analyzes data of 291 university students in an attempt to predict student performance at the end of a 4-year degree program. A student segmentation framework has also been proposed to identify students at various levels of academic performance. Coupled with the prediction model, the proposed segmentation framework provides a useful mechanism for devising pedagogical policies to increase the quality of education by mitigating academic failure and encouraging higher performance. The experimental results indicate the effectiveness of the proposed framework and the applicability of classifying students into multiple performance levels using a small subset of courses being taught in the initial two years of the 4-year degree program.

Early Detection of Student Degree-Level Academic Performance Using Educational Data Mining

Areej Fatemah Meghji[1], Naeem Ahmed Mahoto[1], Yousef Asiri[2,*], Hani Alshahrani[2], Adel Sulaiman[2], and Asadullah Shaikh[3]

[1]**Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro 76062, Pakistan**
[2]**Department of Computer Science, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia**
[3]**Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia**

Corresponding author:
Yousef Asiri[2]

Email address: yasiri@nu.edu.sa

ABSTRACT

Higher educational institutes generate massive amounts of student data. This data needs to be explored in-depth to better understand various facets of student learning behavior. The approach of educational data mining has given provisions to extract useful and non-trivial knowledge from large collections of student data. Using the educational data mining method of classification, this research analyzes data of 291 university students in an attempt to predict student performance at the end of a 4-year degree program. A student segmentation framework has also been proposed to identify students at various levels of academic performance. Coupled with the prediction model, the proposed segmentation framework provides a useful mechanism for devising pedagogical policies to increase the quality of education by mitigating academic failure and encouraging higher performance. The experimental results indicate the effectiveness of the proposed framework and the applicability of classifying students into multiple performance levels using a small subset of courses being taught in the initial two years of the 4-year degree program.

INTRODUCTION

For centuries, the method of educating a large set of students has revolved around instructions being passed to them in a classroom setting (Romero and Ventura, 2013). An instructor delivers lectures and gives tasks; a student attempts to solve these tasks to the best of his/her ability. By monitoring student class behavior, observing their engagement patterns, and checking their task solutions, the instructor can better assess how well a student has grasped concepts. These observations or feedback help instructors revise and modify course contents and the method of lecture delivery. This feedback is an essential component of higher education systems (Bransford et al., 1999). Sadly, an increase in the number of students in a class makes it difficult for the instructor to obtain and record this feedback from each student. The absence of this traditional feedback channel necessitates the exploration of other sources of available data that may aid higher educational institutes create additional feedback loops.

Higher educational institutes collect and store vast amounts of student data (Baek and Doleck, 2022; Khan and Ghosh, 2021). This data includes student demographics, test scores, course assessments, and so on. In recent years, instead of simply storing this data in filing cabinets, an immense amount of research has been conducted on exploring this data to better understand various facets of student learning and behavior. The field of Educational Data Mining (EDM) is an evolving area of research that gained momentum in 2008 (Khan and Ghosh, 2021; Baker, 2014). To find meaningful patterns and hidden insights in the data emerging from the sector of education, EDM builds on techniques from data mining,

machine learning, and statistics to analyze this data (Viberg et al., 2018). EDM aims to extract knowledge from educational data and use it for improved feedback and decision-making (Berland et al., 2014). A unique feature of educational data is the internal hierarchy and correlation amongst data at different levels. Taking this into consideration, EDM approaches explicitly exploit the non-independence and multi-level hierarchy in educational data to predict an overall pattern (Romero and Ventura, 2020). There are five key approaches or research areas in EDM: prediction, relationship mining, clustering, discovery within models, and distillation of data for human judgment (Peterson et al., 2010).

Classification is a popular approach within prediction (Khan and Ghosh, 2021; Viberg et al., 2018). In classification, educational data is fed to an algorithm specifically designed to infer or predict the value of an attribute (class label) based on the patterns or relationships discovered within certain other attributes (predictor variables). Classification has been applied at various levels of granularity to address an ever-increasing set of problems within the educational domain such as inferring a student's emotional state (D'mello et al., 2008), predicting student drop-outs (Agrusti et al., 2019; Márquez-Vera et al., 2016; Delen, 2010), developing recommender systems (Mimis et al., 2019; Erdt et al., 2015), predicting student retention (Shafiq et al., 2022), examining the use of learning materials uploaded in an e-learning platform (Valsamidis et al., 2011), and to identify patterns associated with student success in e-learning platforms (Sánchez et al., 2023). A key application area has been predicting student academic outcomes (Xiao et al., 2022; Nahar et al., 2021; Viberg et al., 2018; Romero and Ventura, 2020; Fernandes et al., 2019). Research in this area has been carried out to predict student success in a course, their grades in a semester, and, to a smaller extent, their success in terms of exam verdict or grades at the end of a degree (Asad et al., 2022; Romero and Ventura, 2013; Berland et al., 2014; Nghe et al., 2007; Asif et al., 2017).

Goal of the Research

By analyzing the most basic student data collected by a higher educational institute, this research aims to devise a classification model that predicts student end-of-degree performance at an early stage during the course of the degree. The goal is to not only predict student performance in terms of academic achievement but also discover courses that impact this performance. This has been done to provide instructors and policy-makers the feedback needed to meet their objective of creating a student-centric learning environment. The predictions made by the model have also been used to devise a segmentation framework that can effectively classify students into learner categories and further help in designing a pragmatic pedagogical policy.

Research Questions

In light of the goal of the research, the work presented in this paper attempts to answer the following questions:

- Is the generation of a predictive model for early detection of student end-of-degree performance possible using the most basic and readily available learning data collected by higher educational institutes?
- Can courses that strongly influence the final prediction of student end-of-degree performance be ascertained to provide intervention?
- Can a segmentation framework be devised to help design a pragmatic pedagogical policy?

The rest of this paper is organized as follows: A review of the related literature has been presented in the section Related Work followed by the section Classification which outlines the process of classification, the working mechanism of some popular classifiers used in this paper, and the metrices used to evaluate the performance of a classifier. The section Research Methodology explores the experimental setup of this research followed by the Experimental Results and Discussion. Lastly, a conclusion and suggestions for future work have been presented in the section Conclusion and Future Work.

RELATED WORK

Higher educational institutes constantly strive to provide an environment that fosters student-centric learning (Romero and Ventura, 2020). The proper analysis of data emerging from the sector of higher education has the potential to manifest results that can not only help enhance student performance but

also elevate teaching effectiveness. EDM is being increasingly used to improve educational outcomes. In particular, researchers have focused on developing classification models to predict student performance (Baek and Doleck, 2022; Xiao et al., 2022).

Nghe et al. (2007) investigated students' undergraduate and postgraduate academic performance at two universities. For Can Tho university in Vietnam, a total of 20,492 undergraduate student records between the years 1995 to 2002 have been explored. The attributes of gender, English language skill, age, family job, and CGPA in the second year of study have been used to predict GPA at the end of the third year of education. Decision trees and Bayesian classifiers have been used to classify student performance. Experiments have been conducted to classify students into four GPA-based classes: fail, fair, good, and very good; three classes: fail, good, and very good; and two classes: fail and pass. The decision tree outperformed in all the experiments. It was observed that accuracy of the classification model increased when the number of class labels was decreased; classifier performance for four classes was 72.95% which improved to 86.47% when made to predict three classes. The performance further improved to to 94.03% for prediction of two class labels. For the Asian Institute of Technology in Thailand, 936 student records were explored between 2003 and 2005. The attributes of university entrance GPA, proficiency in English, and gender have been used to predict GPA at the end of the first year of the master's program. Here, too, the decision tree outperformed with an accuracy of 70.62% for four classes, 74.36% for three classes, and 92.74% for two classes.

Miguéis et al. (2018) explored data of 2459 students belonging to an engineering and technology school of a European public research university between the years 2003 to 2015. Student data available after the first year of a degree program has been used to predict degree-level student academic performance. The data for this research included socio-demographic features, social-economic features, high school background, and data of the first year of the degree. Several classification algorithms have been explored, including Naïve Bayes (NB), Sequential Minimal Optimization (SMO), decision trees, and Random Forests (RF). The classification model based on RF exhibited an accuracy of 96.1%.

Kabakchieva (2013) analyzed data of 10330 students across 20 attributes between the years 2007 to 2009 in a Bulgarian educational institute. After an initial exploration of data, 6 attributes have been removed, and the study has been conducted using student attributes that, among others, included gender, previous education, score in the university entrance exam, and current semester score. Student performance has been classified into five classes (excellent, very good, good, average, or bad) using the decision tree, NB, K-nearest neighbor, and rule-based classifiers. Although the decision tree-based J48 outperformed, all the classifiers achieved an accuracy of less than 70%. The university admission score was discovered as the most influencing attribute towards the final class prediction.

Aman et al. (2019) analyzed data of 1021 students pertaining to academic, demographic, and socio-economic attributes between the years 2014 to 2017. To ascertain the relevance of the considered category of attributes, experiments were performed using only academic and combinations of academic, demographic, and socio-economic attributes. Some attributes considered in this research include gender, division obtained in previous studies, literacy rate, study mode, and the index of poverty of student residential areas. The best results were found using all attributes.

In contrast to the studies discussed thus far, a significant decrease in the dataset size can be observed in the remaining studies. Nahar et al. (2021) have predicted student performance by experimenting on data of 80 students from the department of CSE, Notre Dame University Bangladesh. Student performance has been classified into three categories of good, bad, and medium, based on data from student mark sheets and a behavior survey. Experiments have been conducted using decision trees, NB, RF, and techniques of bagging and boosting. The accuracy of their experiments ranged between 64% to 75% on the test data.

Zimmermann et al. (2015) explored data of 171 students belonging to the Bachelor and Master programs in Computer Science at ETH Zurich, Switzerland. The research analyzed how efficiently student undergraduate performance could indicate student graduate-level performance. Using linear regression in conjunction with variable selection strategies, this research showed that 54% of the variance in graduate-level performance could be explained by undergraduate-level performance. The grade point average of the third year was highlighted as the most significant indicator of overall student performance.

Asad et al. (2022) used the attributes sessional marks and internal marks obtained by different sets of students undertaking five different courses to predict if a student will be safe or at risk of failure in the course. Combining the data across the groups, a total of 176 student records in a bachelor degree program have been analyzed in the research. Experiments have been conducted using decision trees, NB,

RF, SMO, and Linear Regression. The accuracy of their experiments ranged between 88% to 95%. One concern in the used dataset is the imbalance in the class labels which may have caused biased results.

Nieto et al. (2019) explored data from students belonging to a public sector engineering university in Colombia. A total of 19 attributes comprising of student academic and certain derived variables (1st, 2nd, and 3rd quartile of grades) have been explored. Classification approaches of RF, decision trees, and SMO have been used with a varying set of feature-selected attributes. The classification model based on SMO achieved the highest accuracy of 84.43% using all 19 attributes.

Asif et al. (2017) explored the data of 210 students of a public sector university in Pakistan to predict their degree-level performance. This research analyzed the marks obtained by students in various subjects during the first two years of the university degree. Each subject has been treated as an indicator of the final performance prediction. The findings of the research indicate that student performance at the degree level could be successfully predicted by solely using academic marks. Although the NB classifier exhibited the best results with an accuracy of 83.65%, it was established that all classification models are not human interpretable; a model based on the NB classifier could not be used to visualize the generated model. The model based on the decision tree was used to derive subjects that influenced the degree level performance. The decision tree classifier exhibited an accuracy of 69.23%.

In light of the discussed papers, it is evident that various sets of student learning and descriptive attributes have been used to predict student end-of-degree performance with varying degrees of success. Researchers have explored personal features such as age, gender, marital status, parents' education level and job, as well as student learning data such as marks/grade obtained in high school, marks obtained in university entrance exam, and academic marks across various subjects. Some studies have made use of only academic marks, while others have used either a combination of academic and derived or academic and demographic attributes. Based on the reviewed studies it was observed that classifier accuracy is strongly influenced by the number of class labels being predicted; a greater aggregation of academic performance leads to a higher classifier accuracy (Asif et al., 2017; Nghe et al., 2007). Another observation was that most studies have focused solely on predicting student performance and not on finding the factors/features that increase or decrease this performance (Nahar et al., 2021; Aman et al., 2019; Nieto et al., 2019; Kabakchieva, 2013; Nghe et al., 2007). The resultant model needs to be interpreted in order to provide feedback necessary for academic improvement (Xiao et al., 2022). Based on the explored literature, it is also apparent that there is no 'best' classification algorithm; different classifiers have outperformed each other in the discussed papers based on the nature of the examined data. A trend that can be seen is that experiments have mostly been conducted using decision trees, NB, RF, and SMO. It remains to be seen, however, if the performance of the final classification model is significantly influenced by varying the number of class labels, using feature selection, and using academic attributes in conjunction with derived and demographic attributes.

CLASSIFICATION

Classification is a popular approach of prediction which, after learning from a set of data, constructs a model that can be used to predict a designated class label for new and, as yet, unseen data (Mohammed et al., 2016). This process can be broken down into two stages of operation. The first stage is termed the training or learning phase, where labeled educational data are fed to a classification algorithm (classifier) (Romero and Ventura, 2013). The classifier examines and analyzes this data and generates a classification model Quinlan (1993). The generated classification model represents the pattern or logic of how the provided data is categorized into one or more class labels. Thus, classification can be regarded as the task of approximating a mapping function f from certain input variables x to discrete output variables y or $y = f(x)$. An important consideration during this stage is ensuring that the dataset used to train the model has a balanced representation of the class labels (Miguéis et al., 2018). If the sample used to train the model has a biased or skewed distribution towards the classes, the resultant model might have poor predictive performance, especially towards the minority class (Hassan et al., 2021).

Classifiers can be broadly categorized based on their internal mechanism of generating a classification model (Han et al., 2011). Some popular classifiers include decision trees, rule-induction, probabilistic, support vector machines, and memory-based classifiers (Khan and Ghosh, 2021).

Decision tree classifiers are so named as the model generated by them resembles a flow chart or tree structure (Baker and Inventado, 2014). Every internal node in the tree represents a conditional test. Each branch represents the outcome of the test. Starting from a root node, the tree branches out into internal

nodes and branches that finally conclude at some leaf node. The leaf node represents the class label. The root node represents the most significant attribute of the dataset and can be determined using various approaches, including entropy, information gain, and GINI index (Mohammed et al., 2016). J48 is a popular classifier in this category. The RF classifier builds on the concept of decision trees. Instead of generating a single decision tree, the RF generates a forest of decision trees. A class label is established by taking into consideration the output of all the generated trees (Asif et al., 2017). A key attraction of decision tree-based classifiers is their simplicity and the fact that the resultant model can be deciphered. The visual representation of the tree can be used to identify attributes that most strongly influence the final prediction of a class label as well as to understand the exact combination of the attributes and their precise configurations that lead to a particular class label (Viberg et al., 2018; Quinlan, 1993).

Another set of classifiers that generate understandable models is rule-based. If-Then conditions are utilized to generate the target function based on the training data (Han et al., 2011). JRip is a popular rule-based classifier that specifically handles overfitting while learning through reduced error pruning. The NB is a probabilistic classifier that works on the Bayes Theorem. This classifier is quick and resistant to overfitting (Mohammed et al., 2016). SMO is another popular classifier that iteratively trains a support vector machine. It is used to solve optimization problems by incrementally dividing problems into smaller sub-problems (Han et al., 2011).

Unlike classifiers that learn from the training set and then discard it once a mapping function or model of their understanding has been generated, memory-based classifiers store the entire training set. To classify new data items, these classifiers compare the test data with the entire stored training set at run-time (Mohammed et al., 2016) For this reason, these classifiers termed instance-based or lazy. These classifiers are computationally expensive, requiring considerable storage space, especially if the training set is large. However, these classifiers do not make assumptions on the training data and thus are adaptable to problems where the learned assumptions may fail. KStar is a popular memory-based classifier.

The second stage of the classification process is the test phase (Khan and Ghosh, 2021). Once the mapping function has been approximated, it is used to predict the class label of new data that the model has not been trained on. Labels for the test data are known yet kept hidden to evaluate the performance of the model. An important consideration while building a classification model is how well the model learns the target function from the training data and how accurately the model generalizes to new data (Xiao et al., 2022; Romero and Ventura, 2020).

The output of a classification model may be one of the four possibilities: true positive (TP), false positive (FP), true negative (TN), or false negative (FN) (Zeng, 2020). Consider a scenario where the data has been categorized into two classes: P and N. A TP is a correct prediction made by the model for class P. Similarly, a TN is a correct prediction made for class N. FP and FN are incorrect predictions. An FP means an incorrect prediction for class P; data that should have been classified into class N has been incorrectly labeled as belonging to class P.

A confusion matrix is often employed for evaluating the performance of a classification model (Bucos and Drăgulescu, 2018). Table 1 provides the structure of a confusion matrix for a binary classifier.

Table 1. Binary confusion matrix

	Predicted: P	**Predicted: N**
Actual: P	TP	FN
Actual: N	FP	TN

Accuracy, precision, recall, and F_1 score are some evaluation measures computed using the confusion matrix. Accuracy is a measure of correctness. It is used to evaluate how often the predictions made by a classifier are correct (Nieto et al., 2019; Farsi, 2021). Accuracy can be measured using the formula:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{1}$$

The recall is the ability of a classification model to find all relevant cases (points of interest) in the provided data. It measures how many instances of interest were predicted correctly out of all the instances of interest (Farsi, 2021).

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

Precision is used to measure the fraction of instances the classification model considers relevant that actually are relevant (Aman et al., 2019; Hassan et al., 2021). This metric is used to quantify the correct positive predictions. In other words, it is the ratio of correct positive predictions to all positive predictions made by the model.

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

The F_1 score or F-value is the harmonic mean of precision and recall (Khan and Ghosh, 2021; Farsi, 2021). As it takes into account both FP and FN, it performs well on balanced and imbalanced datasets (Hassan et al., 2021). F_1 score is measured as:

$$F_1 Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

The weighted F_1 score is an average of F-values across all class labels, weighted based on the class distribution (class size). Apart from these evaluation metrics, Kappa is also commonly used to evaluate the performance of a classification model (Peterson et al., 2010; Fleiss, 1971). Kappa works under the assumption that a correct prediction could have been made simply by chance. This assumption makes Kappa a useful measure for evaluating the performance of classifiers trained on balanced as well as imbalanced data. Kappa can have a value between 0-1. Similar to accuracy, a higher Kappa value is better; a value above 0.3 signifies that the output of the classifier is not based on chance (Asif et al., 2017).

RESEARCH METHODOLOGY

Fig.1 outlines the research methodology followed in this paper. An explanation of each step is provided in the subsequent sections.

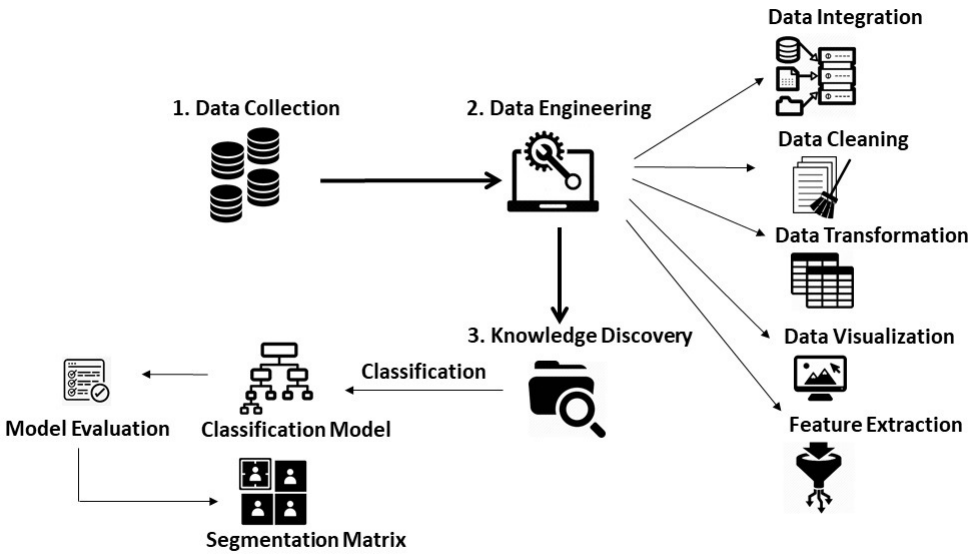

Figure 1. Research Methodology

Data Collection

The first step of this research was the collection of student data. The current research explores data of students enrolled in the Bachelor of Engineering degree program at the Department of Software Engineering, Mehran University of Engineering and Technology, Pakistan. Data from 291 students belonging to three consecutive batches (13SW: 2013-2016, 14SW: 2014-2017, and 15SW: 2015-2019) has been collected. The data was collected from two main sources: institutional records and individual student files. The institutional records comprised of marks obtained by the student in each subject during the course of the bachelors degree. A total of 28 subjects (theory and practical treated as separate heads) are taught by the end of the 2nd year of education in the Department of Software Engineering. As this research attempts to predict student end-of-degree performance based on student academic achievement in the initial two years of the degree program, every subject has been treated as a feature for the prediction of student end-of-degree performance.

Studies by Miguéis et al. (2018) and Asif et al. (2017) suggest academic background prior to the enrolment into the university may influence student performance at the university level, thus, the marks obtained in the university admission test, Higher Secondary Certificate (HSC) exams, and Secondary School Certificate (SSC) exams have also been collected through student files maintained by the department. This data was maintained manually and had to be extracted and computerized so that it could be used in this research. The examined literature presents some conflict over the significance and level of influence of the demographic attribute of gender on overall student academic performance (Khan and Ghosh, 2021). To examine the influence of gender on the prediction of student performance, this attribute has also been considered in this research.

Building on the premise that derived attributes can play a significant role towards student performance prediction (Nieto et al., 2019), two derived attributes: 1st year accumulative score and 2nd year accumulative score have been computed; bringing the total number of attributes to 34. A description of some attributes used in this research has been presented in Table 2. The complete list of attributes used in this research, along with their description, has been provided in the Appendix. The data used in this study has also been attached as an additional file named DegreeData_Classification.csv. Although the student identities have been anonymized by substituting student ids with unique identifiers, this data is not meant for publication as it may be considered sensitive for the university and the students. The Advanced Studies and Research Board (ASRB), in its 136^{th} meeting at Mehran University of Engineering and Technology, approved this study with reference resolution number 136.43.

Table 2. Attributes in the Dataset

Attribute	Description	Type	Value
SSC	SSC Exam Marks	Academic	0-850
HSC	HSC Exam Marks	Academic	0 1100
Ad_Test	University Admission Test Marks	Academic	0-100
ENG11	Functional English	Academic	0-100
MTH108	Applied Calculus	Academic	0-100
SW111	Computer Programming	Academic	0-100
SW111_Pr	Computer Programming Practical	Academic	0-50
Score_First	1st Year Accumulated Score	Derived	0-10
Score_Second	2nd Year Accumulated Score	Derived	0-20
Gender	Student Gender	Demographic	M-F

Data Engineering

Data Integration, Cleaning and Transformation After obtaining data of all the attributes considered in this study, the data of all the three batches was integrated into a single dataset. The data was then analyzed to ensure it did not contain missing or erroneous entries. As no missing or null values were uncovered, the data did not require further scrutiny.

As per the policy of Mehran University of Engineering and Technology (set in accordance to the Higher Education Commission of Pakistan), the final percentage of a student in a bachelor degree is computed by the following formula:

$$Final\% = 0.1 \times 1^{st}\,year\% + 0.2 \times 2^{nd}\,year\% + 0.3 \times 3^{rd}\,year\% + 0.4 \times 4^{th}\,year\% \quad (5)$$

The final percentage at the end of the degree is calculated by summing 10% of the percentage obtained in the 1st year, 20% of the percentage obtained in the 2nd year, 30% of the percentage obtained in the 3rd year, and 40% of the percentage obtained in the final year of the degree. Per the marks obtained in the 1st year of the degree program, the percentage of each student at the end of the 1st year has been computed. A similar practice was followed for the 2nd year percentage. Using these values, the accumulated scores or 10% of the 1st year percentage and 20% of the 2nd year percentage have been computed. The computed values of accumulated scores have been treated as derived attributes in this research.

As this research measures academic success in terms of the total percentage obtained at the end of the degree, experiments have been conducted under two settings. For the first set of experiments, four classes have been established based on student academic success:

1. Class A: High-Achievers (>=85%)
2. Class B: Above-Average (75% - 84%)
3. Class C: Average-Achievers (65% - 74%)
4. Class D: Under-Achievers (<65%)

For the second set of experiments, two classes have been established based on student academic success:

1. Class SP: Satisfactory Performance (>=75%)
2. Class NI: Need Improvement (<=74%)

Data Visualization Classifiers are vulnerable when trained on imbalanced class labels (Hassan et al., 2021); with imbalanced labels resulting in classification models that provide unreliable and biased predictions. Before proceeding with the experiment, it is important to ensure that each label is well-balanced. The class distribution details provided in Table 3 help analyze the class labels' distribution and ensure the results' authenticity for the next step.

Table 3. Class distribution for the considered batches

Student Details	15SW	14SW	13SW	Total
Students in Class A	10	16	22	**48**
Students in Class B	40	27	31	**98**
Students in Class C	37	28	22	**87**
Students in Class D	24	15	19	**58**
Students in Class SP	50	43	53	**146**
Students in Class NI	61	43	41	**145**
Total Students	**111**	**86**	**94**	**291**

Taking a closer look at the figures provided in Table 3, the class labels for both sets of experiments have a balanced distribution. Classes A and D have slightly lesser representation than classes B and C but the values are within the acceptable percentage (Khan and Ghosh, 2021; Asif et al., 2017). The classes SP and NI for the second experiment are equally represented.

Feature Selection Although the amount of data used to train a classifier has great influence on the effectiveness of the generated model, the size of the data alone does not ensure the accuracy and quality of the generated model (Asif et al., 2017). The number of attributes (dimensions / features) being explored, the level of influence these attributes have on the prediction of the class label, and the removal of attributes that inversely affect the prediction of the class label can greatly improve the quality of the generated

model (Matharaarachchi et al., 2022). Thus, an important step before knowledge discovery is ensuring the use of optimal attributes for the classifier (Farsi, 2021).

CfsSubsetEval is a correlation-based feature evaluator in Weka (Witten and Frank, 2002; Eibe et al., 2016) which utilizes Pearson's correlation (r) to determine attributes that strongly influence the prediction of the class label (Hall, 1998). As this research uses a large number of attributes, feature selection using CfsSubsetEval has been explored to find the most significant attributes (see Table 4).

Another unique aspect of this research is that feature selection has been applied on the collected data in three stages. First, feature selection has been applied on all the academic attributes only; the demographic attribute of gender and the derived attributes have not been used. Second, derived attributes have been added to the academic attributes, and feature selection has been applied to the combination. Finally, all the academic, derived, and demographic attributes have been used. This has been done to better analyze the relevance of the features on the final prediction.

Knowledge Discovery

Following the reviewed literature (Nghe et al., 2007; Miguéis et al., 2018; Kabakchieva, 2013; Zimmermann et al., 2015; Nieto et al., 2019; Asif et al., 2017; Aman et al., 2019; Asad et al., 2022), this research makes use of widely popular classification algorithms to predict student performance at the end of a 4-year degree program. Experiments have been conducted using different combinations of the collected data attributes. The first set of experiments has been conducted using all the academic attributes provided in the Appendix. The second set uses a combination of all the academic and derived attributes. The demographic attribute of gender has been added to the existing attributes for the third set of experiments. The last set of experiments focused on attributes discovered during feature selection (see Table.4). Experiments have been conducted on the feature-selected subset of i) academic, ii) academic and derived, and iii) academic, derived, and demographic attributes.

For the discovery of the most optimal classification model, the generated models have been evaluated using the metrices of Accuracy, F-Score (weighted average), and Kappa. The statistical difference in classifier performance has also been examined by means of p-value, computed using the Friedman test (k-1 degrees of freedom), to establish the significance of the results Settouti et al., 2016.

EXPERIMENTAL RESULTS AND DISCUSSION

Table. 4 presents the resulting feature selected attributes obtained using CfsSubsetEval.

As explained in the section Knowledge Discovery, experiments have been conducted on different combinations of the collected data using widely popular classification algorithms. Table 5 presents the results of the various conducted experiments. The attribute set used with each of the classifiers has been presented in the first column. Since the same classifier has been used with various sets of attributes, please note that α has been used to indicate the attribute set under which the result with the highest accuracy has been generated by a classifier when predicting four class labels, and $*$ has been used to indicate the attribute set under which the result with the highest accuracy has been generated by a classifier when predicting two class labels.

The p-value for the observed classifier performance has been computed to monitor the statistical significance of the results. The statistical difference (p-value) for classifier performance has been presented in Table 6.

Several classifiers have exhibited good performance. For experiment#1, the most optimal performance has been exhibited by the model generated by the NB classifier with an accuracy of 84.87%, weighted average F_1 score of 0.848, and a Kappa score of 0.7942 on a feature selected subset of academic, derived, and demographic attributes. The model generated by the RF classifier has the second highest accuracy of 83.50%, followed by the SMO classifier with an accuracy of 82.13% and the J48 classifier with an accuracy of 81.44%. Unlike the NB classifier, RF, SMO, and J48 showed better performance while working with a feature selected subset of academic and derived attributes. Interestingly, apart from the model generated by NB, demographics have not been featured in the optimal model generated by any other classifier. Another interesting observation is that the models based on the decision tree and rule-based classifiers have exhibited better performance when working with a combination of the academic and derived attributes; the performance of these classifiers has decreased when working with the demographic attribute of gender. The performance of all classifiers improved in terms of accuracy and Kappa when working with a feature-selected subset of attributes.

Table 4. Feature selected attributes

	Experiment#1 (4 classes)			Experiment#2 (2 classes)		
SNo	**AC**	**AC+DR**	**AC+DR+DM**	**AC**	**AC+DR**	**AC+DR+DM**
1	ES121	SW125	Gender	SSC	SSC	Gender
2	MTH112	SW215	SW125	SW111_Pr	EL101_Pr	SSC
3	SW121	SW214	SW215	ES121	SW121_Pr	EL101_Pr
4	SW122	SW224	SW214	SW122	SW125	SW121_Pr
5	SW125	SW223	SW224	SW121	MTH212	SW125
6	MTH212	SW221	SW223	SW121_Pr	SW215	MTH212
7	SW211	SW221_Pr	SW221	SW125	SW214	SW215
8	SW214	SW212	SW221_Pr	MTH212	SW211	SW214
9	SW215	SW222	SW212	SW215	SW223	SW211
10	SW224	SW222_Pr	SW222	SW214	SW221	SW223
11	MTH217	Score_First	SW222_Pr	SW224	SW221_Pr	SW221
12	SW223	Score_Second	Score_First	SW211	SW212	SW221_Pr
13	SW221		Score_Second	SW223	SW222	SW212
14	SW221_Pr			SW221	Score_Second	SW222
15	SW212			SW221_Pr		Score_Second
16	SW222			SW212		
17	SW222_Pr			SW221_Pr		
18				SW222		
19				SW222_Pr		

AC=Academic; DR=Derived; DM=Demographic

For experiment#2, the model generated by the SMO classifier exhibited the highest accuracy of 93.13%, weighted average F_1 score of 0.935, and a Kappa score of 0.8694 followed by the NB classifier with an accuracy of 92.78%. Looking at the results presented in Table 5, it can be seen that the accuracy, F_1 score and Kappa scores have greatly improved when working with two class labels. The accuracy of models generated by all six classifiers is approximately equal to or above 90%. Like experiment#1, the model with the highest accuracy has been built using academic, derived, and demographic attributes. Most classifiers have shown an improvement when working with a feature-selected subset of attributes.

The results are presented in Table.5 demonstrate that it is possible to generate a model for the early detection of student end-of-degree performance using the most basic and readily available learning data collected by higher educational institutes. Thus the first research question has been answered in the affirmative.

Classification Model

Even though experiments have been conducted with several classifiers, as previously established by Asif et al. (2017), and discussed in the section Classification, the target courses cannot be identified with all classifiers. Keeping in mind that a goal of this research has been not only the early prediction student academic performance, but also the identification of courses that play a significant role in the final academic performance of the student, a trade-off is being made between classifier accuracy in favour of the interpretability of the model.

The results of the decision-tree classifier J48 have been considered here to identify courses that can help educators provide the necessary intervention, at an early stage, to at-risk students. Due to the extensive size of the generated model (tree) for classification of students into 4-classes, it has been split in two parts. The left-side of the J48 tree for the classification of students into 4-classes has been presented in Fig.2 and the right-side of the J48 tree has been presented in Fig.3.

As explained in the Classification section, the root node of a decision tree identifies the attribute which most strongly influences the final prediction of the class label. Similarly, nodes at a higher level in the tree (closer to the root node) play a stronger role in influencing the final class label. From the model in Fig.2 and Fig.3, it can be observed that the derived attribute of the accumulated score at the end of the 2nd year is the most important feature towards the final prediction of student performance. Subjects SW214, SW221_Pr, SW222, SW212, and SW223 have been identified as the main subjects that affect student

Table 5. Performance Evaluation

	Experiment1 (4 classes)			Experiment2 (2 classes)		
Classifier	**Accuracy (%)**	**F_1 Score**	**Kappa Score**	**Accuracy (%)**	**F_1 Score**	**Kappa Score**
NB $_{(AC)}$	78.01	0.776	0.7015	89.35	0.893	0.7869
NB $_{(AC+DR)}$	80.06	0.800	0.7288	90.72	0.907	0.8144
NB $_{(AC+DR+DM)}$	79.72	0.796	0.7243	90.38	0.904	0.8075
NB $_{(FS\ AC)}$	83.50	0.834	0.7761	91.41	0.914	0.8282
NB $_{(FS\ AC+DR)}{}^{*}$	84.53	0.844	0.7897	92.78	0.928	0.8557
NB $_{(FS\ AC+DR+DM)}{}^{\alpha *}$	84.87	0.848	0.7942	92.78	0.928	0.8557
J48 $_{(AC)}$	68.72	0.684	0.5744	89.00	0.890	0.7801
J48 $_{(AC+DR)}$	73.19	0.732	0.6333	89.00	0.890	0.7801
J48 $_{(AC+DR+DM)}$	72.16	0.722	0.6183	89.35	0.893	0.7870
J48 $_{(FS\ AC)}$	70.44	0.703	0.5958	89.00	0.890	0.7801
J48 $_{(FS\ AC+DR)}{}^{\alpha *}$	81.44	0.814	0.7480	90.72	0.907	0.8145
J48 $_{(FS\ AC+DR+DM)}{}^{*}$	79.72	0.796	0.7244	90.72	0.907	0.8145
JRip $_{(AC)}$	70.44	0.702	0.5973	87.97	0.880	0.7595
JRip $_{(AC+DR)}{}^{*}$	77.66	0.773	0.6977	90.38	0.904	0.8075
JRip $_{(AC+DR+DM)}$	72.16	0.719	0.6214	89.35	0.893	0.7870
JRip $_{(FS\ AC)}$	71.13	0.707	0.6069	90.03	0.900	0.8007
JRip $_{(FS\ AC+DR)}{}^{\alpha}$	74.91	0.749	0.6561	88.32	0.883	0.7663
JRip $_{(FS\ AC+DR+DM)}$	73.53	0.731	0.6396	86.94	0.869	0.7388
RF $_{(AC)}$	81.09	0.812	0.7397	91.07	0.911	0.8213
RF $_{(AC+DR)}{}^{\alpha}$	83.50	0.835	0.7737	90.72	0.907	0.8144
RF $_{(AC+DR+DM)}$	80.75	0.808	0.7360	91.07	0.911	0.8213
RF $_{(FS\ AC)}{}^{*}$	71.13	0.707	0.6069	90.03	0.900	0.8007
RF $_{(FS\ AC+DR)}{}^{\alpha}$	83.50	0.836	0.7731	91.07	0.911	0.8213
RF $_{(FS\ AC+DR+DM)}{}^{\alpha}$	83.50	0.835	0.7731	90.72	0.907	0.8144
SMO $_{(AC)}$	79.72	0.798	0.7211	92.09	0.921	0.8419
SMO $_{(AC+DR)}$	81.78	0.819	0.7494	92.44	0.924	0.8488
SMO $_{(AC+DR+DM)}{}^{*}$	81.78	0.818	0.7498	93.47	0.935	0.8694
SMO $_{(FS\ AC)}$	81.44	0.816	0.7443	91.75	0.918	0.8351
SMO $_{(FS\ AC+DR)}{}^{\alpha}$	82.13	0.822	0.7537	92.78	0.928	0.8557
SMO $_{(FS\ AC+DR+DM)}$	81.78	0.819	0.7496	93.13	0.931	0.8625
KStar $_{(AC)}$	72.51	0.726	0.6227	87.29	0.873	0.7457
KStar $_{(AC+DR)}$	75.94	0.760	0.6700	88.32	0.883	0.7663
KStar $_{(AC+DR+DM)}$	76.28	0.764	0.6745	87.97	0.880	0.7594
KStar $_{(FS\ AC)}{}^{\alpha *}$	76.97	0.771	0.6823	89.35	0.893	0.7869
KStar $_{(FS\ AC+DR)}$	75.60	0.757	0.6656	87.29	0.873	0.7457
KStar $_{(FS\ AC+DR+DM)}$	75.26	0.754	0.6605	89.00	0.890	0.7800

Note: $^{\alpha}$ indicates the result with highest accuracy generated by a classifier when predicting four class labels and * indicates the result with the highest accuracy generated by a classifier when predicting two class labels

end-of-degree performance. Following the path from the root to the class label, some interpretations that can be made from the model are:

- Having a 2nd year accumulated score of less than 11 will result in graduating under Class-D.
- Having a 2nd year accumulated score between 11 and 12, obtaining more than 34 marks in SW222_Pr, greater than 42 marks in SW223, and obtaining more than 64 marks in the subject SW212 will result in graduating under Class-C.
- Having a 2nd year accumulated score between 14-16 and greater than 43 marks in SW221_Pr will result in graduating under Class-B.

Observing the J48 tree for the classification of students into 2-classes (see Fig.4), the derived attribute

Table 6. Statistical difference in classifier performance

Experiment1 (4 classes)		Experiment2 (2 classes)	
Classifier	***p*-Value**	**Classifier**	***p*-Value**
NB vs J48*	0.01431	NB vs J48*	0.01431
NB vs JRip*	0.01431	NB vs JRip*	0.01431
NB vs RF	1	NB vs RF	0.68309
NB vs SMO	1	NB vs SMO*	0.04123
NB vs KStar*	0.01431	NB vs KStar*	0.01431
J48 vs JRip	0.68309	J48 vs JRip	0.68309
J48 vs RF*	0.01431	J48 vs RF*	0.04123
J48 vs SMO*	0.01431	J48 vs SMO*	0.01431
J48 vs KStar	1	J48 vs KStar	0.10247
RF vs JRip*	0.04123	RF vs JRip*	0.04123
RF vs KStar	0.10247	RF vs KStar*	0.01431
JRip vs Kstar	0.10247	JRip vs Kstar	0.10247
SMO vs JRip*	0.01431	SMO vs JRip*	0.01431
SMO vs RF	0.41422	SMO vs RF*	0.01431
SMO vs KStar*	0.01431	SMO vs KStar*	0.01431

* p-value significant $p \leq 0.05$

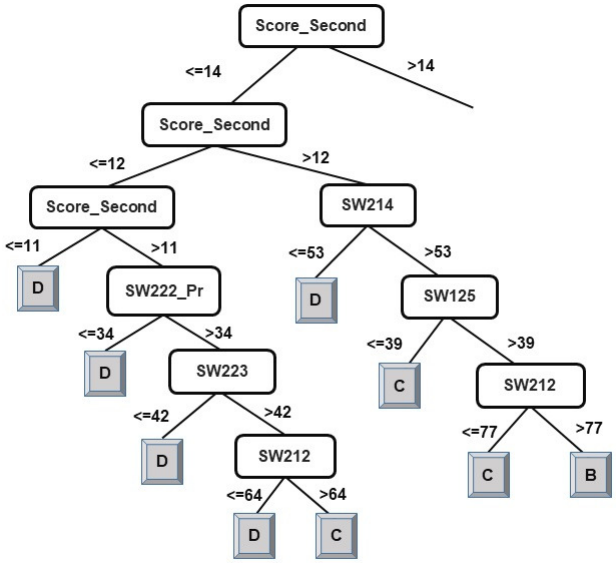

Figure 2. J48 model for predicting the class of learners at the end-of-degree (I)

of the accumulated score at the end of the 2nd year is the most important feature towards the final prediction of student performance into two classes. The subjects SW125, SW221_Pr, EL101_Pr, SW214, SW223, and SW211 have been identified as playing a key role in the final prediction of student end-of-degree performance. Some interpretations that can be made from the model presented in Fig.4 are:

- Having a 2nd year accumulated score of less than or equal to 14 and a score of less than or equal to 39 in SW125 will result in graduating under Class NI.
- Having a 2nd year accumulated score of between 14-15 and a score of greater than 43 in EL101_Pr will result in graduating under Class SP.

An examination of the model presented in Fig.2, Fig.3, and Fig.4 answers research question two. It is

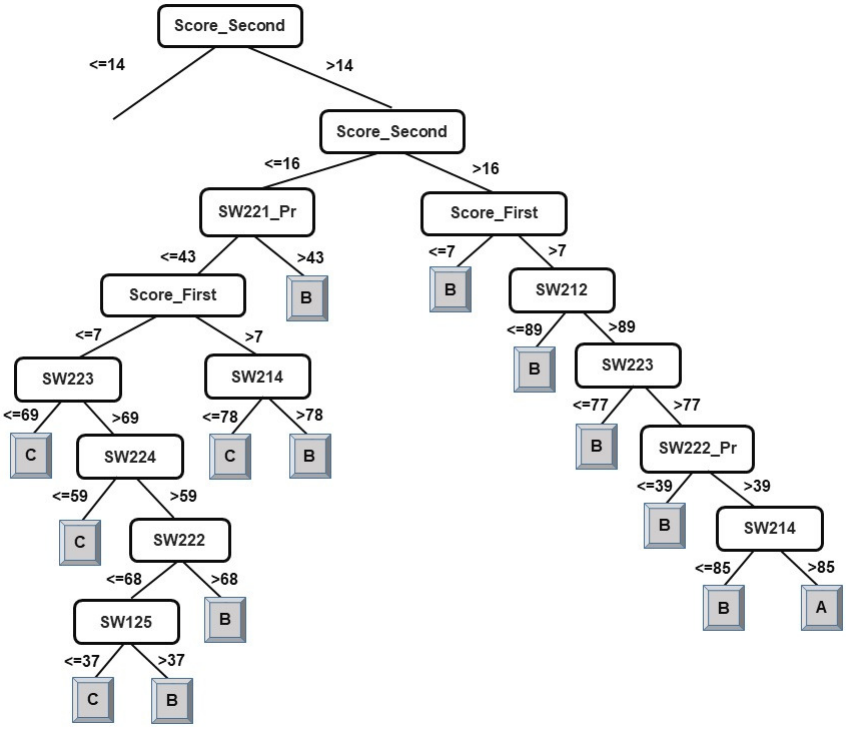

Figure 3. J48 model for predicting the class of learners at the end-of-degree (II)

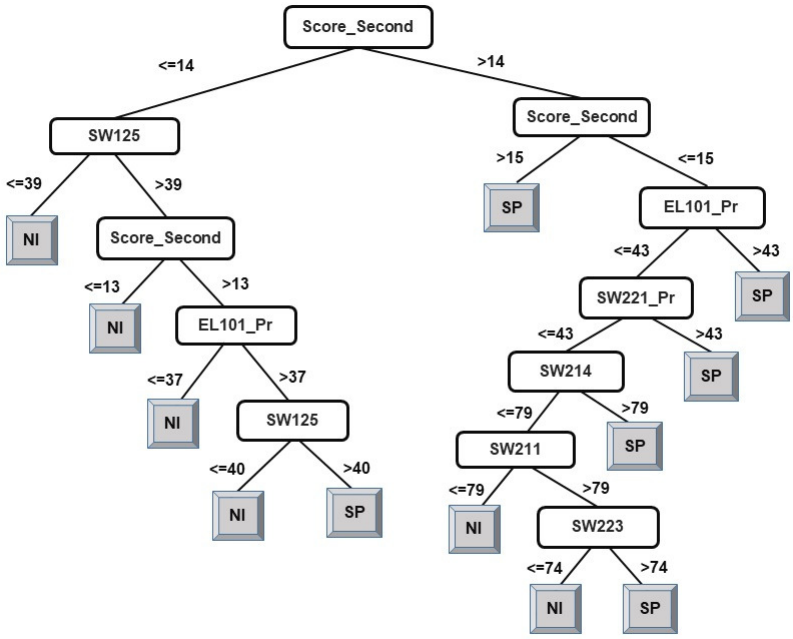

Figure 4. J48 model for predicting the class of learners at the end-of-degree (2-classes)

now safe to conclude that courses which strongly influence the final prediction of student end-of-degree performance can be ascertained.

Segmentation Framework

To identify students for intervention and necessary pedagogical actions, a segmentation framework in the form of a cross-tabular matrix has been proposed. To generate the segmentation matrix, student academic performance at the end of the 2nd year of their university education has been computed. Using the classification model, student performance at the end of the degree has been predicted. The segmentation matrix confronts the observed student performance at the end of the 2nd year against the final performance predicted by the model. As this research uses two approaches to classify students, two segmentation matrixes have been generated. Fig.5 presents the segmentation matrix where students have been classified into two classes based on the percentage obtained at the end of the degree: SP (satisfactory performance: >=75%) or NI (needs improvement: <=74%).

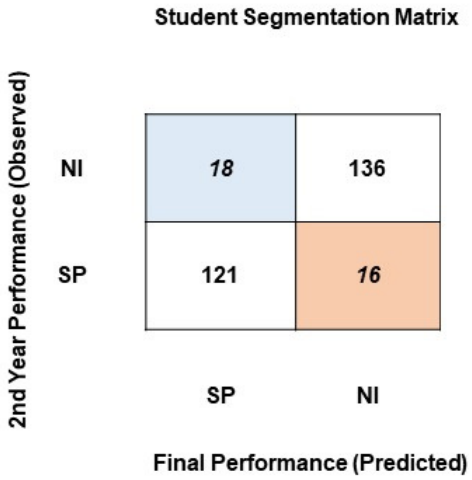

Figure 5. Student segmentation matrix (2-classes)

Evident from Fig.5, a majority of students stay in the same segment at the end of the degree as they did at the end of the 2nd year of the degree program: 121 students reside in the satisfactory performance segment, and 136 students reside in the needs improvement segment. 16 students have moved from the satisfactory segment to the needs improvement segment. The segmentation matrix raises two main concerns. First, a very large proportion of students (136) is persistently performing below a satisfactory performance level. Second, 16 students that resided in the satisfactory segment up until their 2nd year fall into the needs improvement segment by the completion of their degree. As evident from their prior results, these students have the potential to perform better. The students in these two segments are being neglected by the educational institute. A system of feedback, intervention, mediation, and active involvement of the instructors and policy-makers can help students move from these segments.

Using two classes allows us to understand student performance to a small extent. However, bifurcating these classes into further subdivisions will help pinpoint students across various performance levels. Fig.6 presents the segmentation matrix where students have been classified using the second approach. Here, students have been segregated into 4 classes: A (high achievers: >=85%), B (above-average achievers: 75% - 84%), C (average achievers: 65%-74%), and D (underachievers: <65%).

Observing the diagonal of the segmentation matrix in Fig.6, most students graduate in the same performance segment they belonged to at the end of their 2nd year: 43 students in Class-A, 63 in Class-B, 50 in Class-C, and 55 in Class-D. The cells adjacent to the diagonal identify students whose performance changes after the 2nd year. Observing the last row of the segmentation matrix, 13 students that were in Class-A at the end of the 2nd year are predicted to finish their education in Class-B, and 4 students that reside in Class-A are predicted to finish their education under Class-C. Observing the second row from the top, 21 students who reside in Class-C at the end of the 2nd year have been predicted to complete the degree in the Class-B performance segment. These students have potential, and perhaps having the right pedagogical strategies may help them jump up to the high-achiever segment. A major concern in

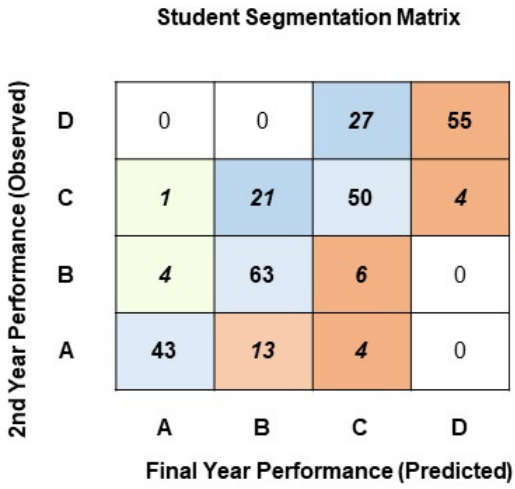

Figure 6. Student segmentation matrix (4-classes)

the segmentation matrix is the top-right cell: 55 students that are predicted to complete their degree as underachievers in Class-D.

The suggested approach identifies 16 student segments allowing the institute to design a pedagogical policy to specifically target each segment. A robust, pragmatic policy can be devised to mitigate factors that lead to poor performance levels and identify academically motivated students. Using the approach proposed in this research, it can be concluded that a segmentation framework based on student performance can be devised to help design a pragmatic pedagogical policy.

Discussion

Consistent with the research conducted in (Nghe et al., 2007; Miguéis et al., 2018; Nieto et al., 2019; Asif et al., 2017; Aman et al., 2019), the current research validates that it is possible to successfully predict student performance at the end of the degree using student data at some earlier point during the course of the degree.

As the reviewed studies differed in the attributes used, efforts were made in the current research to conduct experiments that would build upon concepts provided in the mentioned studies. Thus, the first set of experiments focused only on academic attributes (marks in SSC, HSC, university admission test, and the marks in subjects studied in the first two years of the degree program). The results outlined in Table 5 clearly indicate that student performance can be predicted using only academic attributes. However, it needs to be noted that using a feature-selected subset of academic attributes greatly improved the performance of the classifiers. In the case of the NB classifier, the performance in terms of accuracy increased from 78.01% to 83.50% when working with 4-classes and from 89.35% to 91.41% when predicting 2-classes.

The current research computed the attributes of accumulated scores at the end of the 1st and 2nd years of the degree program. The addition of these derived attributes significantly improved the classifiers' performance. Considering the results of experiment#1, the J48 classifier exhibited an accuracy of 68.72% using only academic attributes, which improved to 73.19% with the addition of the derived attributes. This accuracy further increased to 81.44% when a feature-selected subset of academic and derived attributes was used. Similarly, for the NB classifier, an accuracy of 84.53% was observed on a feature-selected subset of academic and derived attributes.

The addition of the attribute of gender did not play a significant role in the final prediction for all classifiers. Although the addition of this attribute improved the performance of the NB and SMO classifiers, it had the opposite effect on the performance of the J48, KStar, and JRip classifiers. Thus the conflict of using the attribute of gender still stands (Khan and Ghosh, 2021). It was also observed that classifier performance is inversely proportional to the number of class labels. The lesser the number of class labels, the better the performance of the classifier.

At the end of the experiments, it can be concluded that a classification model to predict the class a student will graduate in can successfully be generated with a subset of academic and derived attributes. Using feature selection greatly improves the classifiers' overall performance and can aid in reducing the complexity of the final model. A segmentation matrix can then be generated using the classification model. The proposed segmentation framework can be useful for proactively devising a pedagogical policy that targets each performance segment.

CONCLUSION AND FUTURE WORK

This research explores and analyzes the most basic student data available in a 4-year degree program. Three research questions have been investigated in this paper. The first question focused on the generation of a classification model for early identification of student end-of-degree performance using the most basic and readily available learning data collected by higher educational institutes. It was observed that student performance at the end of a degree program could successfully be predicted using a feature-selected combination of academic and derived attributes. The second question focused on deriving courses that strongly influence the final prediction of student performance. The model generated using the J48 classifier indicates that certain courses do influence the final prediction of student performance. Furthermore, the marks obtained in these courses can be used to classify students into various performance levels and thus be used to provide intervention to students at risk of obtaining poor grades. The third question involved the generation of a segmentation framework. A cross-tabular segmentation matrix has been used to confront the computed student performance at the end of the 2^{nd} year against the final performance as predicted by the generated model. The resultant segmentation matrix identifies students in various performance segments. The early identification of these students provides the opportunity to robustly devise a pragmatic policy to specifically target each performance level.

This research aims to provide instructors and policymakers with the much-needed feedback to truly create a student-centric learning environment. Several courses have been identified as indicators of student performance in this research. An important future direction can be to explore student performance in these courses. This will provide the educational institute an added opportunity to improve educational outcomes. Also, using the approach outlined in this paper, predictive models can be built for the early identification of student performance across the other degree programs offered by the university. The early prediction of student performance will help in designing a pedagogical policy that can increase the quality of education by not only mitigating academic failure but also by encouraging higher performance.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors are thankful to the Deanship of Scientific Research at Najran University for funding this work under the Research Collaboration Funding program grant code (NU/RC/SERC/11/7).

Grant Disclosure

The following grant information was disclosed by the authors:

Deanship of Scientific Research, Ministry of Higher Education Saudi Arabia: NU/RC/SERC/11/7

Competing Interests

The authors declare that they have no competing interests

Author Contributions

Areej Fatemah Meghji conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and tables, authored or reviewed drafts of the paper, and approved the final draft.

Naeem Ahmed Mahoto conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Yousef Asiri conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Hani Alshahrani performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Adel Sulaiman performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Asadullah Shaikh performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The used dataset has been anonymized and attached as an additional file named DegreeData_Classification.csv. However, this data is not meant for publication as it is considered sensitive for the university and the students. The data described in the research is available in the Appendix. Further data analysis and experimental results are available in Table 2, Table 3, Table 4, Table 5, and Table 6.

REFERENCES

Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of E-Learning and Knowledge Society*, 15(3):161–182.

Aman, F., Rauf, A., Ali, R., Iqbal, F., and Khattak, A. M. (2019). A predictive model for predicting students academic performance. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–4. IEEE.

Asad, R., Arooj, S., and Rehman, S. U. (2022). Study of educational data mining approaches for student performance analysis. *Technical Journal*, 27(01):68–81.

Asif, R., Merceron, A., Ali, S. A., and Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113:177–194.

Baek, C. and Doleck, T. (2022). Educational data mining: A bibliometric analysis of an emerging field. *IEEE Access*, 10:31289–31296.

Baker, R. S. (2014). Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, 29(3):78–82.

Baker, R. S. and Inventado, P. S. (2014). Educational data mining and learning analytics. In *Learning analytics*, pages 61–75. Springer.

Berland, M., Baker, R. S., and Blikstein, P. (2014). Educational data mining and learning analytics: Applications to constructionist research. *Technology, Knowledge and Learning*, 19(1-2):205–220.

Bransford, J. D., Brown, A. L., and Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* National Academy Press.

Bucos, M. and Drăgulescu, B. (2018). Predicting student success using data generated in traditional educational environments. *TEM Journal*, 7(3):617.

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, 49(4):498–506.

D'mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., and Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, 18(1):45–80.

Eibe, F., Hall, M. A., and Witten, I. H. (2016). The weka workbench. online appendix for data mining: practical machine learning tools and techniques. In *Morgan Kaufmann*. Morgan Kaufmann Publishers.

Erdt, M., Fernandez, A., and Rensing, C. (2015). Evaluating recommender systems for technology enhanced learning: a quantitative survey. *IEEE Transactions on Learning Technologies*, 8(4):326–344.

Farsi, M. (2021). Filter-based feature selection and machine-learning classification of cancer data. *INTELLIGENT AUTOMATION AND SOFT COMPUTING*, 28(1):83–92.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., and Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, 94:335–343.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hassan, M. M., Eesa, A. S., Mohammed, A. J., and Arabo, W. K. (2021). Oversampling method based on gaussian distribution and k-means clustering. *Computers, Materials and Continua*, 69(1):451–469.

Kabakchieva, D. (2013). Predicting student performance by using data mining methods for classification. *Cybernetics and information technologies*, 13(1):61–72.

Khan, A. and Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies*, 26(1):205–240.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124.

Matharaarachchi, S., Domaratzki, M., and Muthukumarana, S. (2022). Minimizing features while maintaining performance in data classification problems. *PeerJ Computer Science*, 8:e1081.

Miguéis, V. L., Freitas, A., Garcia, P. J., and Silva, A. (2018). Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, 115:36–51.

Mimis, M., El Hajji, M., Es-Saady, Y., Guejdi, A. O., Douzi, H., and Mammass, D. (2019). A framework for smart academic guidance using educational data mining. *Education and Information Technologies*, 24(2):1379–1393.

Mohammed, M., Khan, M. B., and Bashier, E. B. M. (2016). *Machine learning: algorithms and applications*. Crc Press.

Nahar, K., Shova, B. I., Ria, T., Rashid, H. B., and Islam, A. (2021). Mining educational data to predict students performance. *Education and Information Technologies*, 26(5):6051–6067.

Nghe, N. T., Janecek, P., and Haddawy, P. (2007). A comparative analysis of techniques for predicting academic performance. In *2007 37th annual frontiers in education conference-global engineering: knowledge without borders, opportunities without passports*, pages T2G–7. IEEE.

Nieto, Y., García-Díaz, V., and Montenegro, C. (2019). Decision-making model at higher educational institutions based on machine learning. *JUCS-Journal of Universal Computer Science*, 25:1301.

Peterson, P. L., Baker, E., and McGaw, B. (2010). *International encyclopedia of education*. Elsevier Ltd.

Quinlan, J. R. (1993). Program for machine learning. *C4. 5*.

Romero, C. and Ventura, S. (2013). Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 3(1):12–27.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1355.

Sánchez, A., Vidal-Silva, C., Mancilla, G., Tupac-Yupanqui, M., and Rubio, J. M. (2023). Sustainable e-learning by data mining—successful results in a chilean university. *Sustainability*, 15(2):895.

Settouti, N., El Amine Bechar, M., and Amine Chikh, M. (2016). Statistical comparisons of the top 10 algorithms in data mining for classification task. *International Journal of Interactive Multimedia and Artificial Intelligence*, 4:46–51.

Shafiq, D. A., Marjani, M., Habeeb, R. A. A., and Asirvatham, D. (2022). Student retention using educational data mining and predictive analytics: A systematic literature review. *IEEE Access*.

Valsamidis, S., Kontogiannis, S., Kazanidis, I., and Karakos, A. (2011). E-learning platform usage analysis. *Interdisciplinary Journal of E-Learning and Learning Objects*, 7(1):185–204.

Viberg, O., Hatakka, M., Bälter, O., and Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Computers in Human Behavior*, 89:98–110.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Xiao, W., Ji, P., and Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. *Engineering Reports*, 4(5):e12482.

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. *Communications in Statistics-Theory and Methods*, 49(9):2080–2093.

Zimmermann, J., Brodersen, K. H., Heinimann, H. R., and Buhmann, J. M. (2015). A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance. *Journal of Educational Data Mining*, 7(3):151–176.