Predicting Students' Academic Performance Using Data Mining Method

Kenth C. Novo

Staff, Research and Development, Northeastern Mindanao State University

DOI : https://doi.org/10.51583/IJLTEMAS.2024.131016

Received: 28 October 2024; Accepted: 03 November 2024; Published: 14 November 2024

Abstract: Predicting students' academic performance is a critical task in educational institutions to identify students likely to succeed or need intervention. This study leverages data mining techniques to analyze and predict student outcomes, contributing to more informed decision-making in educational management. Using data from various academic sources, the research implements three popular data mining algorithms—Naive Bayes, Multilayer Perceptron, and C4.5 Decision Tree—to classify student success. These techniques allow the identification of patterns and dependencies in student data, providing insights that can assist educators in creating tailored interventions. The results indicate that the Naive Bayes algorithm outperforms other classifiers in terms of prediction accuracy, making it a viable tool for predicting academic performance. This research underscores the potential of data mining to enhance educational outcomes by allowing proactive responses to students' academic needs.

Keywords: Academic performance prediction, Student performance analysis, Data mining in education

I. Introduction

Predicting students' academic performance in students has long been a popular study issue. The major goal of the admissions process is to identify students who are likely to succeed after being admitted into the institution. The quality of admitted students has a significant impact on the institution's academic achievement, research, and training. Failure to make an accurate admission decision may result in the admission of an inappropriate student to the program. As a result, admissions staff seek to learn more about each student's academic potential. Accurate forecasts assist admissions staff in distinguishing between eligible and unsuitable candidates for a particular academic program, as well as identifying those who are likely to succeed at the institutions of higher education whose mission is to contribute to a higher standard of living.

The success of human capital building in higher education is the topic of ongoing investigation. As a result, higher education institutions must be able to forecast student performance since the quality of the teaching process is determined by the capacity to satisfy students' demands. In this way, critical data and information are collected on a regular basis, examined by the proper authorities, and quality criteria are established. Higher education institution quality entails offering services that most likely suit the demands of students, academic staff, and other education system participants. Participants in the educational process generate a massive quantity of data that must be gathered, integrated, and used once they have fulfilled their commitments through proper activities. By transforming this information into knowledge, and the enjoyment of all participants. Data mining on data from the higher education system might help all players in the educational process: students, professors, administration, supporting administration, and the social community.

II. Method

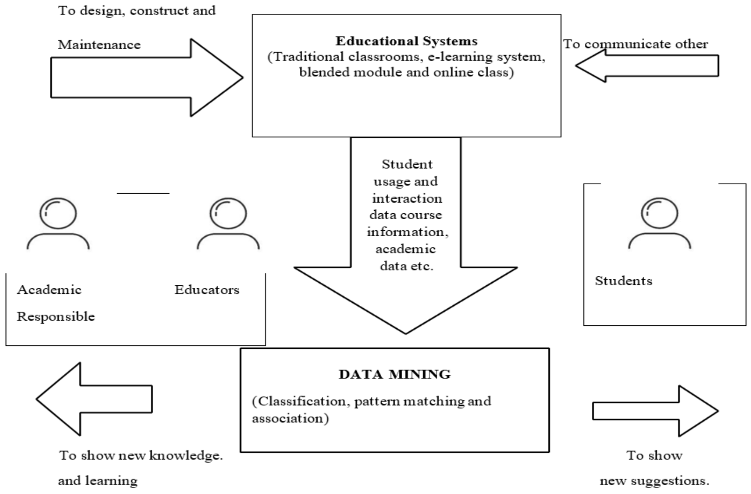
To design, construct and
Maintenance
Educational Systems
(Traditional classrooms, e-learning system, blended module and online class)
To communicate other
Student usage and interaction data course information, academic data etc.
Academic
Responsible
Educators
Students
DATA MINING
(Classification, pattern matching and association)
To show new knowledge. and learning
To show new suggestions.

Figure 1.1 cycle of applying data mining in educational system.

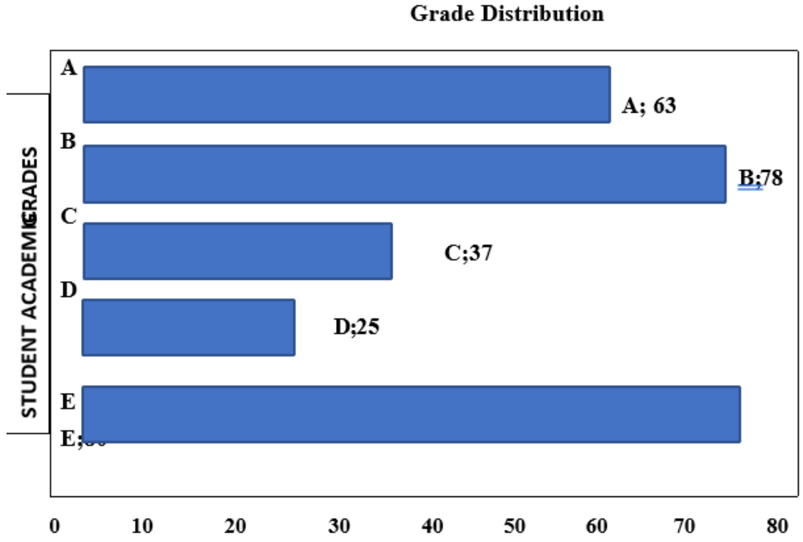

Figure 2.1 Student academic grades distribution of every student.

Number of Students

The output variable is the students' evaluations in the class.

There are various types of academic years, two of which are as follows:

Using the six classes that are coded in the way: pupils' labels are the same as theirs.

Table 2.2 shows the final grades. Through the use of two classes written in this language way: failure category A, success in category B. Table 2.3 shows that the test was passed.

Figure 2.2. There are six different classifications for the ultimate grade of the students

Class	Grade	Student	Percentage
1	**A**	63	25.12%
2	**B**	78	32.54%
3	**C**	37	13.07%
4	**D**	25	10.43%
5	**E**	80	35.76%

Table 2.3 Two class labels regarding as students' final grade

Class	Grade	Student	Percentage
1	A	80	35.76%
2	B	203	78.88%

Since the prediction error rate in the first example will be substantially larger due to the varied distribution of grades across classes, the advantage will be provided to the second scenario of this study.

Data Mining

The use of data mining in educational systems may be tailored to meet the unique demands of each student. Those who take part in the educational process. Additional exercises, educational materials, and assignments that would benefit and increase the student's learning are necessary. Professors would have access to feedback, as well as the ability to divide students into groups depending on their need for supervision and monitoring, as well as the ability to identify the most common errors. discover the most effective measures, and so forth. The factors that will increase system performance will be given to the administration and administrative employees. Data mining is a computer approach of data processing that has been effectively implemented in a variety of fields with the goal of extracting meaningful knowledge from data.

Data mining techniques are used to create a model in which unknown data attempts to identify new data information. Regardless of where they came from, all data mining approaches have one thing in common: they all use automation to uncover new correlations and dependencies among characteristics in the data. If the purpose of the analysis is to categorize data by class, then the new information about the classes to which data belongs is the new information. As a result, the algorithms are classified into two categories: supervised and unsupervised algorithms.

Algorithms that are supervised.

The output conditions are not explicitly represented in the data set when mining is "unsupervised" or "undirected": the task of an unsupervised algorithm is to discover automatically inherent patterns in data without prior information about which class the data could belong to, and it does not involve any supervision. The purpose of a model like this is to find data patterns in a large number of input variables. Even though an unsupervised learning algorithm was not developed for prediction tasks, the model it produces can sometimes be utilized for them. This category includes clustering methods and association rules.

Supervised algorithms are those that take data from a known class to develop models, and then predict the class to which new data will belong based on the model. This category includes categorization methods. The process of learning a function that maps data into one of many predetermined classes is referred to as data classification methods. An input data set including vectors of attribute values and their matching classes is provided to any classification method that is based on inductive learning. A classification technique's purpose is to create a model.

It allows for the automatic classification of future data points based on a set of predefined features. These systems use a set of cases as input, each of which belongs to one of a restricted number of classes and is specified by the values for a set of fixed characteristics. They provide a classifier that can reliably predict which class a new case belongs to as an output. Decision trees, induction rules or classification rules, probabilistic or Bayesian networks, neural networks, and hybrid techniques are the most frequent classification approaches. In this study, we investigated the impact of three algorithms for intelligent data analysis: C4.5, Multilayer Perceptron, and Naive Bayes. There are many different classifiers in the literature, and it is impossible to choose the best because they differ in many aspects such as learning rate, amount of data for training, classification speed, robustness, and so on.

(Wu and Kumar, 2009). These methods are used to create classification models, with the goal of predicting the class (student success) to which a fresh unlabeled sample will belong. The three categorization strategies chosen are utilized to determine the best method for predicting student performance.

The Naive Bayes algorithm (NB)

A basic classification approach based on probability theory, specifically the Bayesian theorem (Witten and Frank, 2000). It's named naïve because it simplifies difficulties by depending on two key assumptions: prognostic characteristics are conditionally independent with familiar categorization, and there are no hidden traits that might influence the prediction process. This classifier is a potential approach to probabilistic knowledge discovery, as well as an extremely efficient data categorization technique. One of the most extensively used and popular neural networks is the multilayer perceptron (MLP) algorithm. The input layer of the network is made up of sensory components, one or more hidden layers of processing elements, and the processing elements' output layer (Witten and Frank, 2000). MLP is particularly well suited to approximate a classification function that divides the example indicated by the vector attribute values into one or more classes (where we are unfamiliar with the connection between input and output attributes). C4.5 is the most prevalent and, in many ways, the most frequently utilized decision tree algorithm today. In 1993, Professor Ross Quinlan created the C4.5 decision tree method, which is the outcome of research that dates back to the ID3 algorithm (originally suggested by Ross Quinlan in 1986). Handling missing values, classification of continuous characteristics, decision tree pruning, rule formulation, and other features are included in C4.5. The divide and conquer strategy is used in the basic design of C4.5 algorithms to create an appropriate tree.

III. Result and Discussions

This package was written in the Java programming language, and it is now widely regarded as the most capable and complete bundle of machine learning algorithms available in academic and nonprofit settings. It is typical to study the influence of input factors during students' prediction success, in which the impact of specific input variables of the model on the output variable has been investigated, in order to gain a better understanding of the relevance of the input variables. For the evaluation of input variables, four tests were used: Chi-square test, One R-test, Info Gain test, and Gain Ratio test. The following metrics were used to track the results of each test: Attribute (name of the attribute), Merit (measure of goodness), Merit dev (deviation, i.e. measure of goodness deviation), Rank (average position occupied by attribute), Rank and dev (deviation, deviation takes attribute's position), Rank and dev (deviation, deviation takes attribute's position).

Diverse algorithms provide substantially different outcomes, i.e., they each account for attribute relevance in a different way. Instead of picking one method and trusting it, the end result of attribute ranking is the average value of all the algorithms. Three supervised data mining algorithms were used to preoperative assessment data to predict course outcome (passed or failed), and the learning techniques' performance was assessed based on their prediction accuracy, simplicity of learning, and user pleasant qualities. The results show that the Nave Bayes classifier beats decision tree and neural network approaches for prediction. It has also been said that an effective classifier model must be accurate as well as understandable to instructors. Because data mining techniques were performed after the data was obtained, this study was based on typical classroom situations. This approach, however, may be utilized

to assist students and Methods, Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop.

Acknowledgement

I would like to express my sincere gratitude to my advisor, Catherine R. Alimboyon, for their unwavering guidance and support throughout this research. My heartfelt thanks to the faculty and staff of the North Eastern Mindanao State University, especially the College of Information Technology Education, for providing the resources and assistance needed to complete this study.

I am also deeply grateful to my family and friends for their continuous encouragement. Lastly, my appreciation goes to all the students and educators who contributed data for this research, making this work possible.

Author's Biography

Kenth C. Novo born on February 20, 1992, in Brgy. Dagocdoc, Tandag City, Surigao del Sur, Philippines. Earned my Bachelor of Science in Computer Science from North Eastern Mindanao State University, where I developed a strong interest in data mining and educational technology.

References

1. IJCSMC. (2013). IJCSMC, 2(7).
2. Ibrahim, Z. (2007). Faculty of Information Technology and Quantitative Sciences. Universiti Teknologi MARA Malaysia.
3. Ackerman, P. L., Kanfer, R., & Beier, M. E. (2013).
4. Knowledge-Based Systems. (2008).
5. Educational Technology & Society. (2021). 24(1).
6. Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018).
7. Huang, A. Y. Q., Lu, O. H. T., Huang, J. C. H., & Yin, C. J. (2019).

Appendix

Appendix A: Data Mining Algorithms Used in the Study

This appendix provides details on the three data mining algorithms used in this study to predict students' academic performance.

1. **Naive Bayes Algorithm (NB)**
   - The Naive Bayes classifier is a probabilistic model based on the Bayes theorem. It assumes conditional independence between features, making it computationally efficient for large datasets. In this study, NB was used to classify students based on their performance data, predicting whether they would succeed or fail in their courses.
2. **Multilayer Perceptron (MLP)**
   - MLP is a type of artificial neural network that consists of an input layer, one or more hidden layers, and an output layer. It is used to model complex relationships between inputs and outputs by learning a function that maps input data to its corresponding output class. MLP was applied to the dataset to improve prediction accuracy by capturing nonlinear patterns in the data.
3. **C4.5 Decision Tree Algorithm**
   - C4.5 is a decision tree algorithm that splits the dataset into subsets based on the value of the most significant attribute, thus generating a tree where each node represents a decision. The algorithm handles both continuous and categorical data and includes pruning methods to prevent overfitting. It was used to classify students into various performance categories based on historical academic data.

Appendix B: Student Grade Distribution

The table below shows the distribution of student grades in the study sample, categorized into five grade ranges:

Grade	Number of Students	Percentage (%)
A	63	25.12%
B	78	32.54%
C	37	13.07%
D	25	10.43%
E	80	35.76%

Appendix C: Input Variables for Prediction Models

The following input variables were considered for predicting students’ academic performance:

- **Course information**: Includes data on the type of course and subject.
- **Student interaction data**: Reflects how students engaged with course materials and online resources.
- **Academic data**: Consists of grades from previous subjects, attendance records, and participation in school activities.

Appendix D: Evaluation Metrics

To assess the performance of the data mining models, the following metrics were used:

- **Accuracy**: The percentage of correct predictions made by the model.
- **Chi-square test**: Evaluates the significance of input variables.
- **Info Gain test**: Measures the importance of each feature in relation to the target output.
- **Gain Ratio test**: A refinement of the Info Gain test, which considers the intrinsic information of a split.