ROLE OF DATA MINING IN EDUCATION FOR IMPROVING STUDENTS PERFORMANCE FOR SOCIAL CHANGE

Myla M. Arcinas¹, Guna Sekhar Sajja², Shazia Asif³, Sanjeev Gour⁴, Ethelbert Okoronkwo⁵, Mohd Naved⁶

¹Associate Professor, Behavioral Science Department
De La Salle University, Philippines
²Research Scholar
University of the Cumberlands, USA
³Sasif1@hct.ac.ae
⁴Career College, Bhopal
sunj129@gmail.com
⁵ethelbert.okoronkwo@gmail.com
⁶Assistant Professor, Department of Business Analytics
Alex Ekwaume Federal University Ndufu-Alike, Ikwo, Ebonyi State Nigeria
Jagannath University, Delhi-NCR, India.

ABSTRACT

Higher education institutions increasingly realize that they are in the service sector, with students as the primary clients. Higher education institutions place a high priority on improving student performance. Before establishing a performance development program, it is necessary to map out the student's current situation. Predicting a student's success is a major challenge for higher education administrators. This research aims to discover the characteristics that impact students' choice of a field of study in higher education. Predictive tools and procedures will be developed to forecast students’ behavior, attitudes, and performance to decide the students' choice of higher education. Early prediction of student performance aids in taking steps to improve student accomplishment. Several attempts have been made to anticipate student performance to reach a quality education standard; however, the prediction accuracy is not satisfactory. This paper attempts to cover every area of educational data mining. A paradigm for predicting student success is also provided and assessed.

Keywords: Data Mining, Education, Machine Learning, Prediction, Student Performance, Social Change

I. OVERVIEW OF DATA MINING EDUCATION SYSTEM

Students and alumni in higher education encounter significant hurdles. Institutions want to determine, for example, which students will enroll in specific course programs and which students will require support to graduate. Is it true that certain students are more likely to transfer than others? Which alumni groups are most likely to make pledges? Aside from these problems, conventional difficulties like enrollment management and time-to-degree continue to inspire higher education institutions to seek better solutions. One efficient approach to solving these student and alumni concerns is data analysis and presentation, often known as data mining [1].

Data mining [2] helps firms leverage their existing reporting skills to discover and comprehend hidden patterns in massive collections. These patterns are subsequently incorporated into data mining models and utilized to forecast individual behavior accurately. As a result of this understanding, institutions are better equipped to distribute resources and personnel. Data mining, for example, can provide an institution with the knowledge it needs to take action before a student drops out, or it may help an institution allocate resources more efficiently by providing an accurate prediction of how many students will attend a given course.
This research looks at the capabilities of data mining and how it may be used in higher education. Data mining employs a mixture of an explicit knowledge base, strong analytical capabilities, and domain expertise to find hidden trends and patterns. These trends and patterns serve as the foundation for prediction models, which allow analysts to generate new observations from current data. To generate new observations from existing data, a method of uncovering significant new correlations, patterns, and trends by sifting through enormous volumes of data housed in repositories and employing pattern recognition technology, as well as statistical and mathematical methodologies, is required. Data mining on very large or raw datasets should be conducted using supervised or unsupervised data mining methods. It should be noted that data mining cannot take place without direct interaction with unitary data.

EDM [3], as shown below in figure 1, is a study field that analyzes educational data to uncover useful information about various types of learners and their learning behavior and the impact of educational policies applied in diverse learning contexts. EDM uses raw data as input for various educational settings and converts it into relevant facts and knowledge. This data may help educational policymakers, school administrators, instructors, and students make educated decisions on managing and engaging with educational resources. It allows for data-driven decision-making in order to improve present educational procedures and learning resources.

A huge amount of student data has become available as technology use in educational information systems has advanced. This emphasizes the need to use EDM to analyze students' learning habits. EDM aids in the successful assessment of educational institutions in order to maximize the use of learning resources.

Figure 1 shows information mining from an educational information system, which gathers and uses educational data in learning objects, event logs, and student profiles. EDM [4] assignments are created using educational data in student profiles and knowledge modeling. This educational data assists students in making efficient use of learning resources and educators in making use of this found information, i.e., learning patterns, for performance prediction. EDM [5] is concerned with the numerous obstacles and problems linked with distinct parts of the learning phenomena [6].

These challenges are as follow:

- Finding relevant course materials for students.
- Organizing and delivering course material to end-users.
- Identifying the factors that affect the performance of students.
- Getting feedback from end-users on the delivered course material.
• Analyzing the significance of teaching aid provided, delivered course material, and feedback received from end-user.

Various practices, in contrast to the different educational datasets, leads to the unveiling of several problems. So, it is essential to select the correct problem formulation technique corresponding to the desired research objectives to get the desired results.

II. DATA MINING TECHNIQUES AND THEIR APPLICATIONS IN EDM

EDM approaches study data in order to find solutions to educational issues. It attempts to investigate the undiscovered patterns following an examination of curriculum, learning behavior, and student family information gathered from diverse educational institutions. The intended outcomes of the EDM include greater knowledge of the existing causes and consequences in the education system.

To demonstrate the link between data objects, association rules are utilized. Mining association rules provide for discovering rules of the following type: If antecedent, then (most likely) consequent, where antecedent and consequent are item sets consisting of one or more items. Association rule creation is often divided into two steps: To begin, minimal support is used to locate all frequent item sets in a database. Second, these frequent item sets and minimal confidence restrictions are utilized to create rules.

Machine learning (ML) [7] is a new area of data mining that allows a computer program to grow increasingly accurate in predicting outcomes without explicitly programming it. These ML techniques are often divided into two types: supervised and unsupervised learning techniques employ labeled training data for inference (classification, regression), whereas unsupervised learning techniques employ unlabeled data to identify existing hidden patterns (clustering).

The act of transforming an input collection of instances P into a unique collection of characteristics Q, also known as target attributes or labels, is known as classification. Classification techniques such as decision tree classifiers, Bayesian classifiers, artificial neural networks, nearest neighbor classifiers, random forest, and support vector machines are used in various applications [8]. We will go through each of them briefly. Each strategy is based on the learning algorithm that it utilizes.

A decision tree is one of the most basic and simple classifiers used to solve classification issues. A decision tree is a graph in which occurrences are classified by sorting them depending on their feature values. The decision tree is composed of nodes and branches, with each node representing a classification instance and each branch representing a value that the node might take on. In the decision, instant categorization starts at the root node, and instance sorting is done based on feature values.

Predicting the class label for a given collection of input qualities can be challenging in some situations. Furthermore, even when utilizing the specified input attribute set values to match some of the attributes in the training data set, class variables are non-deterministic. This is possible due to some noisy data and perplexing features that are not considered during analysis—for example, projecting the possibility of heart disease in a given person based on that individual's daily activity.

In this instance, most people who consume nutritious foods and exercise regularly are at risk of acquiring heart disease due to other variables such as smoking, alcohol intake, and potentially inheritance. In such circumstances, the categorization model is created based on well-recognized heart disease characteristics, which cannot offer correct information. In such situations, there is a requirement to describe probabilistic correlations between the attribute collection and the class label, and the Bayesian classifier is all about justifying such duties [9].

An artificial neural network (ANN) is based on biological neural networks used to build animal brains. ANN is also known as a connectionist system since it comprises linked nodes and directed linkages. Each linked connection is assigned a weight and is responsible for sending a signal from one node to another. When a node gets a signal, it processes it before passing it through to another node.

In ordinary ANN implementations, the signal at the link between artificial neurons is an actual number, and the output of each neuron is determined by a non-linear function of the sum of all its inputs. Because of the weights of artificial neurons and the connections between them, the signal intensity rises or decreases as learning continues [10].

www.turkjphysiotherrehabil.org 6521
There are two approaches to building a learning model in ML classification. One of them is that the model begins learning as soon as the training set is available; such models are known as eager learners. Another model observes all training examples but only achieves classification if the test instance's properties perfectly match any of the training instances. Such pupils are known as lazy pupils [11].

The Nearest Neighbour (NN) classifier treats each sample as a data point in a d-dimensional space, where d is the number of characteristics. It is determined the distance between the provided test example and all data points in the training set. The data point X's k-Nearest Neighbors are the k points nearest to the X.

The data point is then categorized based on the class labels of its neighbors. If a data point has more than one class labeled neighbor, the class label with the most class labels is applied to the data point. The precise value of k's nearest neighbors should be established. If the value of k is too low, it may misclassify owing to noise in the training data. On the other hand, if the value of k is too big, there is a risk of misclassification since the collection of nearest neighbors may contain data points situated far away from the test attribute's neighborhood.

Random forest is a supervised machine learning technique composed of a forest of judgments produced by many decision trees produced using random vectors. This method may be used to solve classification problems as well as regression operations. The random forest's outcome is connected to the number of trees it combines in the forest so that as the number of trees in the forest rises, so does the chance of attaining more accuracy. It is critical to understand that establishing the forest is not the same as generating decision trees [12].

Random forest classifiers are used in medical, finance, e-commerce, and the stock market. In banking, random classifiers are used to distinguish between loyal and fraudulent clients. Random Forest is used in medicine to discover the optimal mix of medications and diagnose illness based on a patient's past medical information. Random Forest classifier is used to watch a stock's activity in the stock market and then detect the loss and profit. Random Forest may be used to estimate user product suggestions in the context of e-commerce.

The Support Vector Machine is the supervised learning model used for categorization (SVM). It has received a great deal of interest in the categorization sector. A noticeable gap separates instances of the various categories in vector space in the SVM model. When a new sample arrives, it is mapped into the specified vector space, and its label is assigned to a category based on which side of the gap it falls [13]. An SVM can do non-linear classification effectively by employing the kernel technique.

Clustering, also known as cluster analysis, is the job of arranging a collection of things so that elements in one group are more similar to those in another. As the similarities between things in one group and the dissimilarities between things in other groups rose, so did the clustering. Clustering is the cornerstone of data mining and may be employed in various applications like image processing, data compression, computer graphics, machine learning, and many more.

Clustering can be used in combination with other object-categorization techniques such as classification, segmentation, and partitioning. When comparing Cluster analysis to classification, we may state that clustering is an example of unsupervised learning. Cluster analysis varies from classification in that classification retains knowledge of classes, but clustering does not maintain knowledge of classes. Furthermore, fresh samples are grouped into established classes in the case of classification, whereas groups are recommended in cluster analysis based on data patterns [14].

The author of [15] provides a method for categorizing students to estimate their final grade using attributes retrieved from recorded data in an educational web-based system. They create, develop, and test a collection of pattern classifiers, comparing their performance on an online course dataset. To separate the kids, four classifiers were utilized. The integration of many classifiers improves classification performance significantly.

They applied the genetic algorithm (GA) to increase prediction accuracy, and when compared to non-GA, the accuracy of combined classifier performance is around 10 to 12 percent higher when using the genetic algorithm.
This strategy is quite helpful in identifying students in danger early on, especially in big classrooms, and allows the instructor to offer relevant advice in a timely way. Researchers examined how association rules may be used in educational data mining to assess learning data in the paper.

They emphasized that the cosine and added value (or equivalently lift) functions are ideally suited to educational data and that teachers can readily grasp their results. They give data from the LMS for the case study (Learning Management System). Researchers showed how data mining might be used in higher education, specifically to increase student performance, in the article. They used the Database course to do this and collect all accessible data, including their use of the Model e-learning facility. They used an association rule, a classification rule using a decision tree, classified the students using EM-clustering, and found an outlier in the data using outlier analysis.

They put their newfound knowledge to good use by improving their performance. The article authors used cluster analysis and K-means algorithm approaches to investigate the association between student university entrance examination results and success. The university students were organized into clusters based on their characteristics, and the clustering procedure was carried out using the K-means clustering method.

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>• Detection of student behavior.</td>
</tr>
<tr>
<td></td>
<td>• Development of domain models.</td>
</tr>
<tr>
<td></td>
<td>• Discovering students learning styles and preferences.</td>
</tr>
<tr>
<td></td>
<td>• Understanding the educational outcomes of the students.</td>
</tr>
<tr>
<td>Clustering</td>
<td>• Grouping similar students based on learning behavior and their performance.</td>
</tr>
<tr>
<td>Predictive Modeling</td>
<td>• Prediction of either a student qualifies a course or not.</td>
</tr>
<tr>
<td>Relationship Mining</td>
<td>• Discovering the curriculum associations in a course.</td>
</tr>
<tr>
<td></td>
<td>• Finding bottlenecks in specific study programs.</td>
</tr>
<tr>
<td>Visual Analytics</td>
<td>• Analyzing of educational processes or erudition outcome by visualizing the model.</td>
</tr>
<tr>
<td>Discovery with Models</td>
<td>• Student characteristics or contextual variables.</td>
</tr>
<tr>
<td></td>
<td>• Determining the relations among different student behaviors.</td>
</tr>
<tr>
<td>Refinement of Data for Individual decision</td>
<td>• Labeling the data that helps in the improvement of the prediction model.</td>
</tr>
<tr>
<td></td>
<td>• Identification of the students’ learning patterns.</td>
</tr>
</tbody>
</table>

Figure 2: Various Data Mining Techniques and Their Role in Education Data Mining

III. FRAMEWORK FOR STUDENTS PERFORMANCE PREDICTION

Student performance data set [17] is used as input in the framework as in figure 3. This KDD data set is preprocessed to remove noise from the data and make data consistent. After preprocessing, a clean and consistent input data is available. Now, this input data is provided to machine learning algorithms like- SVM, Neural Network and Random Forest. These algorithms perform the classification of input data. Then this classification
data works as training data for the prediction task. When new student data is entered into the framework then based on learning data available in the classes, the framework predicts the performance.

![Figure 3: Student Performance Prediction Framework](image)

The machine learning algorithm's accuracy and error rate are shown below in figure 4 and figure 5.

![Figure 4 Accuracy Results of Classification Algorithms](image)
Higher education institutions are increasingly aware that they are in the service industry, with students serving as the key customers. Improving student performance is a top concern for higher education institutions. It is vital to map out the existing state of the students before designing a performance improvement program. For higher education administrators, predicting a student's performance is a huge task. Early prediction of student performance assists in the implementation of measures to increase student achievement. Several attempts have been made to forecast student performance to achieve a quality education standard; however, the prediction accuracy is inadequate. This website tries to cover every aspect of educational data mining. A model for predicting student achievement is also presented and evaluated.

REFERENCES


