

Research Article

Fuzzy Logic and Machine Learning-Enabled Recommendation System to Predict Suitable Academic Program for Students

Tribhuwan Kumar ,¹ K. Sakthidasan Sankaran ,² Mahyudin Ritonga ,³ Shazia Asif ,⁴ C. Sathiya Kumar ,⁵ Shoaib Mohammad ,⁶ Sudhakar Sengan ,⁷ and Evans Asenso ,⁸

¹College of Science and Humanities at Sulail, Prince Sattam Bin Abdulaziz University, Al Kharj 11942, Saudi Arabia ²Department of ECE, Hindustan Institute of Technology and Science, Chennai, India

³Universitas Muhammadiyah Sumatera Barat, Padang, Indonesia

⁴*Higher College, Sharjah, UAE*

⁵Department of Computational Intelligence, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India

⁶School of Law, IMS Unison University, Dehradun, India

⁷Department of Computer Science and Engineering, PSN College of Engineering and Technology (Autonomous), Tirunelveli 627152, Tamil Nadu, India

⁸Department of Agricultural Engineering, University of Ghana, Accra, Ghana

Correspondence should be addressed to Evans Asenso; easenso@ug.edu.gh

Received 10 May 2022; Accepted 22 July 2022; Published 11 August 2022

Academic Editor: Mukesh Soni

Copyright © 2022 Tribhuwan Kumar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, educational data mining has gained a considerable lot of interest as a consequence of the large number of pedagogical content that can be gathered from a range of sources. This is because there is a lot of instructional information that can be obtained. The data mining tools collaborate with academics to improve students' learning strategies by analyzing, sifting through, and estimating components that are pertinent to students' characteristics or patterns of behavior. This is accomplished through the following steps: EDM is utilized in the vast majority of instances to develop the classification model, which then assigns a certain class to each student based on the known properties of the training dataset. Before putting the classification model into use, it is possible to utilize a test dataset to verify that the model is accurate. This article provides a description of a recommendation system that determines the most beneficial academic program for students by utilizing fuzzy logic and machine learning. The compilation of a student dataset has begun. It includes a total of 21 features and 1000 individual cases. The initial step is to employ the CFS attribute selection method. This methodology selects 15 of the initial set of 21 characteristics. Following the completion of the data gathering, it is put through various machine learning methods such fuzzy SVM, random forest, and C4.5. This methodology that has been offered makes predictions about the academic program that is best suitable for students.

1. Introduction

In recent years, EDM (educational data mining) has attracted a significant lot of interest as a consequence of the enormous amount of instructional material that can be obtained from a variety of sources. Primary purpose of EDM is to increase the efficacy of data mining models in order to secure and protect the large amount of educational data collected and to establish a safe learning environment for students. In this technique, many models for data management and analytics have been employed [1]. A number of prediction methods, including classification, regression, and latent component analysis, were also used to make the predictions.

The data mining tools work in collaboration with academics to enhance students' learning techniques by assessing, filtering, and estimating aspects that are relevant to students' traits or behavioral patterns [2]. Increase the number of students who are placed, and the number of graduates who are hired is one of the most difficult goals for any educational institution to accomplish. EDM techniques such as clustering and visualization are displayed in Figure 1 as examples of popular EDM approaches.

The building of an ML (machine learning) model, which is a computationally demanding process, makes use of resampling and iterative classification algorithms. Through the use of machine learning approaches that combine optimal subset selection, traditional classifier flaws such as over-fitting and distributional demands on parameters may be removed. The original machine learning (ML) approaches in computer science were based on statistics rather than predicting group characteristics; instead, they started with an arbitrary group separator and tweaked it repeatedly until it met the requirements of the classification groups in question [3]. When the ML functions become unstable, it is a good idea to review the tuning variables as well as the individual ML functions.

In addition, because of the nonstatistical character of these approaches, the data may be presented in a number of formats, such as nominal data, in order to achieve the maximum possible classification precision. It is defined as a process that can be quantified and measured. With the help of machine learning, we are able to solve the classification issue in the context of collective characteristics [4]. These applications have benefited from the advancements in pattern analysis and machine learning (ML) during the last several decades, which have broadened the range of features that may be employed in these applications.

To minimize the amount of redundant characteristics, which are referred to as "overload" and "difficult processes," several ways have been implemented in order to reduce complexity. It is critical to pick characteristics that are most relevant to the issue at hand in order to learn more about it while also reducing the amount of processing necessary. In this work, we look at models that employ specific values to define subsets of features in order to enhance overall prediction efficiency. Clusters are collections of data elements that have been grouped together. To put it another way, a cluster is a collection of items that are all the same inside the cluster but are not related to any other clusters at all.

Clustering is one of the most essential UL statistical processes, and it is described here [5]. This preprocessing approach reduces the quantity of the data for useful clusters, which may subsequently be used in further in-depth studies in the DM model when they have been identified. Because the data are saved in a cluster format, which is a lossy compression approach, the file size is reduced as a result of the compression. Clustering is a challenging idea to categories since they overlap so much. Traditionally, clustering models such as hierarchical and partitional clustering have been divided into two groups. Understanding the subtle distinctions between clustering and supervised classification, on the other hand, is crucial. In supervised classification, prelabeled data patterns are utilized to make decisions about the data. In this procedure, unlabeled datasets are identified and clustered, with the objective being to estimate the labeling of those datasets while the clustering is taking place.



FIGURE 1: Educational data mining.

When it comes to the design of clusters that are used to develop a clustering solution, it is common to distinguish between partitional and hierarchical techniques. In addition, there is a distinction between "hard" and "soft" models, which is related to the way the items in the dataset are mapped into clusters and how the models are trained.

Data classification frequently necessitates the use of two sophisticated methods. This step of the learning process begins with an examination of a collection of training dataset samples to establish the classes that will be used. NN, rulebased methods, and data-driven techniques are examples of methodology. The technique for data classification is illustrated. Every instance is considered to be located in the previous class, which is defined as follows: it is sampled in the second phase, and this is done with the aid of an additional dataset that has been created to test the classification accuracy of a method. Once a suitable level of accuracy has been attained, this approach is used to categorize future data instances by assigning them a class label. Classification is, without a doubt, a critical component of the decisionmaking process. One of the classification ideas that is employed is the Bayesian technique [6, 7].

In the majority of cases, EDM is employed in the following areas: in order to create the classification model, assign a specific class to each student based on the known attributes of the training dataset. A test dataset may be used to ensure that the classification model is valid before deploying it. It is possible to use the categorization model to assign a new student to a certain class. It is possible to anticipate a student's academic progress using data from either online or offline sources. Content recommendations, course registration, and academic program enrolment for students are all included in this category.

Literature survey contains survey of various methods for educational data mining. Methodology section presents a fuzzy logic and machine learning-enabled recommendation system to predict suitable academic program for students. A student dataset is prepared. It is having 1000 instances and 21 attributes. First, CFS attribute selection algorithm is applied. This algorithm selects 15 attributes out of original 21 attributes. Then, machine learning algorithms, namely, fuzzy SVM, random forest, and C4.5, are applied on the dataset. This proposed model predicts suitable academic program for students. Result section presents details related to dataset and accuracy achieved by the prediction model.

2. Literature Survey

Educational institutions may be able to solve these challenges through the use of EDM and learning abilities, which allow them to precisely anticipate the number of students who will graduate and be put in positions. DM, according to research conducted by Ji et al. [8], assists educators in the transformation of every model into data, such as learning objectives, learning actions, learning prioritization, participation and competition, functions, and accomplishment tendencies in a variety of learning events, according to the findings of their study. Recently, data mining and machine learning models in educational data have been employed to solve the issues raised above.

Two of the most widely used machine learning (ML) algorithms, known as predictive and descriptive analytics, are unsupervised learning (UL) and supervised learning (classification), which are two of the most popular ways to machine learning (ML) algorithms [9].

Both of these considerations are at the center of our study [10]. Students' performance in two disciplines, such as a Bachelor of Business Administration degree, should be evaluated in Finland and Spain. Then, examine what effects their performance in other areas, such as their motivation or priorities, should be evaluated in both countries. The primary objective of this study is to evaluate the effectiveness of a specific latent class model, Bayesian profile regression, in identifying students who are more likely to fail their classes. It is feasible to build student profiles that are associated with the highest levels of academic risk based on the performance, motivation, and resiliency of the kids. The data for this study are gathered via online questions submitted by undergraduate students at an Italian institution. The data are collected in real time and utilized to build a sample of data for the study.

Education at the next level is developed and implemented through the use of EDM and learning analytics [11]. For the goal of creating learning approaches, it provides a systematic paradigm for collecting, calculating, reporting, and operating on digitalized data. By incorporating EDM and LA into the educational setting, instructors may come up with innovative solutions to the problem of interaction. It is a virtual educational platform that serves to bridge the gap that exists between instructors and learners. LMS is an abbreviation for learning management system. With the aid of this technology, students and teachers may communicate more effectively with one another. It enables both instructors and students to share knowledge and answer issues and concerns in a safe environment.

DM and data analytics were the focus of Ray et al. [12] research in the educational business, where they examined how to handle the data collected. EDM and LA approaches

may be used to handle enormous amounts of data by both commercial and noncommercial organizations. EDM and LA also give a complete analysis of how the role of shareholders in PG-level educational institutions is influenced by their findings. A brief description of how these models may be implemented, as well as how students' learning processes can be analyzed, and how they can be utilized to provide complete feedback, is also given in this part. These models finally have an impact on the administrative ideas that are acceptable for all stakeholders in the educational process.

The NB classification has been used to develop a model for predicting student dropout, which has just been published by Hegde et al. [13] in the R programming language. The next stage will be to take a more in-depth look at why students fail or succeed in their first year, and whether or not they will be dropped. As previously noted, there are a variety of factors that might lead to a kid being dismissed from school without warning. The capacity to predict whether or not a kid would drop out of school is tremendously valuable to company owners.

When dealing with large student databases, a number of data preparation techniques have been employed in order to generate the students' marks in accordance with the assessment modules. Student grades have been fine-tuned in the data preparation stage before being utilized to extract the categorical component from the data. As a result, there are no unique grades for each of the courses that have been recorded.

Following that, an investigation of the preprocessing procedures of EDM data was conducted. It is generally acknowledged that educational information should not be saved in the same way as other types of data since there are so many variables, such as diverse data sources, applications, and human error. As a result, the coursework estimation ratio has been utilized to investigate alternative module assessment approaches while simultaneously producing transcription data for students enrolled in the course. The coursework assessment ratios have been demonstrated to have an effect on classifiers that use radio frequency technology (CARs).

The academic performance of a university's students is typically used to determine the excellence of the institution. When it comes to EDM applications, popular and successful one is the prediction of students' grade point averages and educational achievement, which is classified in the EDM forecast as "excellent," "very good," "good," "moderate," and so on and so forth. This form of prediction may be used to identify the most deserving candidates for scholarship grants at a number of institutions. For the undergraduate level, subsequent elements such as grade point average (GPA) and academic efficiency have been researched in the literature. According to the researchers, about 300 pupils were utilized to predict the final grades of students in the faculty of computer systems and software engineering [14]. Using multivariate analytic models, the significance of a feature has been sampled to determine its relative importance. According to experts, a student's ability to succeed may be predicted by their family's support. Furthermore, the outcome is unaffected by the pupils' level of interest.

4

Asshraf et al. [15] conducted an investigation of the performance of 210 undergraduate students by analyzing their data. The qualities are used to make predictions about students' grades. Finally, the research demonstrates that it is possible to predict a student's graduation performance in a final semester of university using their preuniversity grades as well as the marks from their first- and second-year courses with greater accuracy using their preuniversity grades and the marks from their first- and second-year courses.

Pradeep et al. [16] predicted the dropout rate for bachelor students using a sample of students enrolled in the technology program. When using the Weka tool, the attribute selection algorithms help to limit the influence of the characteristics that are employed. Postenrolment characteristics like attendance, paying attention in class, and grades received are usually stated as determining factors in choosing a university. For reliable prediction of academic accomplishment, it is not required to include other information such as age, gender, or religion in the questionnaire.

The goal of this study is to determine whether or not hyper-parameter optimization can be employed effectively in educational environments, which is the subject of this study [17]. When it comes to forecasting student, achievement based on the role performed by the online learning environment, automatic machine learning has shown to be particularly successful. In order to obtain visible and intelligible results in the search space at the same time, both rule-based and tree-based approaches are employed. The final point to mention is that an impressive number of findings indicate the ability of auto ML systems to achieve extraordinary results. It is demonstrated that the DM model may be utilized to analyze educational data in the field of education [18]. When forecasting the number of students who will drop out of college or university, a researcher uses a classifier that is specifically designed for this purpose. To assess whether or not a student falls into the dropout group, it is necessary to have detailed information on their curriculum. Students who have enrolled in university throughout the preceding few decades have been able to access authentic academic data about themselves. When dealing with irregular data, it is necessary to preprocess the raw data.

The usage of decision support systems (DSS) is a vital element in the EDM process (DSS). The goal of an application is to make the process of making a decision easier. For example, giving comments and sending alerts, planning events, making ideas, and improving course materials are all instances of this sort of work. Although DSS is primarily intended for use by teachers, it is not capable of being used successfully by students, managers, or researchers.

New models of decision-making have been proposed in recent studies, including the following. Participants, on the other hand, are the primary emphasis of these recommendations. Instructors, for example, could provide courses to students. RS methods such as associative/content-based filtering, association rules, and hybrid systems are common examples of RS techniques. These have been widely employed in the field of business data management (EDM). Suggestions are also provided via the use of discovery techniques. To give an example, Vialardi-Sacón et al. [19] employed a performance forecasting tool to generate ideas for their clients. This model is used to forecast a student's performance in each class and to provide course suggestions based on that performance estimate.

The fundamental purpose of this study's research is to install a classifier based on locally produced student characteristics as soon as possible [20]. After that, we employ preprocessed data to determine student characteristics, which are subsequently used in the FS as well as in future learning and evaluation via Weka, among other things. It has been possible to construct predictor tools by utilizing the NB classification model, which is extremely accurate. It is consequently possible to utilize the tool as a sampling tool since it may anticipate students' roles in following studies based on its qualities.

3. Methodology

This section presents a fuzzy logic and machine learningenabled recommendation system to predict suitable academic program for students as shown in Figure 2. A student dataset is prepared. It is having 1000 instances and 21 attributes. First, CFS attribute selection algorithm is applied. This algorithm selects 15 attributes out of original 21 attributes. Then, machine learning algorithms, namely, fuzzy SVM, random forest, and C4.5, are applied on the dataset. This proposed model predicts suitable academic program for students.

When determining the value of a subset of characteristics, CFS takes into account both the unique ability of each characteristic and the degree of duplication that occurs between the characteristics. Characteristics with a high relationship to the class, but a low correlation with the other criteria, are considered for selection [21].

The categorizing approach may be used in a supervised or especially unsupervised setting, depending on the situation. This is a well-known truth in the business community. This is the source of the fact that support vector networks are supervised learning standards. An SVM may be used to build feature points or attribute states by projecting them onto nonlinear hyperplanes and planar projections. The performance of SVMs is significantly influenced by the use of Gaussian kernels, data variance and standard deviation, and kernel selection procedures, among other factors. When using fuzzy SVM, each training point corresponds to a single class with pinpoint accuracy. When dealing with data that are stochastic or probabilistic in nature, it is necessary to collect prelearning data on the datasets themselves. In this section, we will take a look at a number of different stochastic connections [22].

Using random forests for classification and regression problems may prove to be beneficial in the near future. Regression methods are used in order to anticipate the outcomes of each decision tree that is generated during the process of training the decision trees. It has a low standard deviation when used to create predictions and is capable of quickly combining numerous bits of information. Because of the cryptic nature of random forest categorization, the general public was first skeptical of its efficacy. On the contrary, it has shown superior performance in a predictionbased challenge [23].



FIGURE 2: Fuzzy logic and machine learning-enabled recommendation system to predict suitable academic program for students.

A decision tree is a kind of classification algorithm that is often employed in machine learning models because of the speed and accuracy they provide. Tree trimming may be accomplished in a number of ways with the help of this method, which is quite adaptable. Following the process of pruning, only a small number of discoveries are produced that are easily understandable. Some scientists feel that overfitting might be done by the removal of trees. C4.5 method iterative classification continues until the data are classified as accurately as possible by creating pure leaf nodes, at which point the operation is completed. C4.5 method iterative classification allows for the most accurate possible results to be obtained from training data without the need for extraneous rules that merely identify a certain behavior [24].

4. Results and Discussion

A dataset of 1000 students has been created with 21 attributes like age, sex, university name, city, country, father's education, mother's education, family size, health status, travel time, family relationship, study hours, extra activities, hobbies, alcohol consumption, course name, course description, skills, difficulty level, results, and attendance.

First features are selected using correlation feature selection (CFS) algorithm. Then, these 15 attributes are selected, age, father's education, mother's education, health status, travel time, study hours, extra activities, hobbies, alcohol consumption, course name, course description, skills, difficulty level, results, and attendance.

For performance comparison, three parameters, accuracy, sensitivity, and specificity, are used.

Accuracy =
$$\frac{(TP + TN)}{(TP + TN + FP + FN)}$$
,
Sensitivity = $\frac{TP}{(TP + FN)}$, (1)

Specificity =
$$\frac{1N}{(TN + FP)}$$



FIGURE 3: Accuracy of machine learning techniques for recommendation system to predict suitable academic program for students.



FIGURE 4: Sensitivity of machine learning techniques for recommendation system to predict suitable academic program for students.



FIGURE 5: Specificity of machine learning techniques for recommendation system to predict suitable academic program for students.

where TP = true positive, TN = true negative, FP = false positive, and FN = false negative.

Results of different machine learning predictors are shown in Figures 3–5. Accuracy of fuzzy SVM is better than random forest and C4.5 algorithm.

5. Conclusion

As a result of the large quantity of instructional content that can be collected from many sources, educational data mining has garnered a great deal of attention over the last few years. Student learning approaches are improved via the use of data mining technologies, which work in partnership with academics to identify features that are significant to students' qualities or behavioral patterns and then analyze, filter, and estimate those aspects. It is most often used to develop the classification model, which then assigns a particular class to each student on the basis of known qualities from a training dataset. Before implementing a classification model, it is possible to test it on a test dataset to check that it is valid. It is described in this study how a recommendation system that employs fuzzy logic and machine learning may be used to find the optimum academic program to be followed by students. A student data collection is currently being compiled. It is made up of 1000 occurrences and 21 attributes, in total. After that, the CFS attribute selection process is used. This approach selects 15 characteristics from the original 21 characteristics. Following data gathering, machine learning algorithms such as fuzzy SVM, random forest, and C4.5 are used to the data. This proposed model predicts the best suited academic program for students based on their individual needs.

Data Availability

The data shall be made available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

- R. S. Baker and P. S. Inventado, *Educational Data Mining and Learning Analytics*Learning analytics Springer, New York, NY, 2014.
- [2] B. K. Baradwaj and S. Pal, "Mining educational data to analyze students' performance," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 6, pp. 63–69, 2012.
- [3] V. Hemamalini, S. Rajarajeswari, S. Nachiyappan et al., "Food quality inspection and grading using efficient image segmentation and machine learning-based system," *Journal of Food Quality*, vol. 2022, pp. 1–6, 2022.
- [4] A. Gupta and L. K. Awasthi, "Peer-to-peer networks and computation: current trends and future perspectives," *Computing and Informatics*, vol. 30, no. 3, pp. 559–594, 2011, http://www.cai2.sk/ojs/index.php/cai/article/view/184.
- [5] M. Shabaz and U. Garg, "Predicting future diseases based on existing health status using link prediction," *World Journal of*

Engineering, vol. 19, no. 1, pp. 29–32, ahead-of-print (Issue ahead-of-print). Emerald, 2021.

- [6] M. M. Ezz, "Advisory system for student enrollment in university based on variety of machine learning algorithms," *International Journal of Computing Academic Research* (*IJCAR*), vol. 4, no. 2, pp. 34–45, 2015.
- [7] M. Shabaz and U. Garg, "Shabaz-urvashi link prediction (sulp): a novel approach to predict future friends in a social network," in *Journal of Creative Communications*vol. 16, no. Issue 1, pp. 27–44, SAGE Publications, 2020.
- [8] H. Ji, K. Park, J. Jo, and H. Lim, "Mining students activities from a computer supported collaborative learning system based on peer to peer network," *Peer-to-Peer Networking and Applications*, vol. 9, no. 3, pp. 465–476, 2016.
- [9] C. Wu, P. Lu, F. Xu, J. Duan, X. Hua, and M. Shabaz, "The prediction models of anaphylactic disease," *Informatics in Medicine Unlocked*, vol. 24, p. 100535, 2021.
- [10] M. d. M. Camacho-Miñano, C. del Campo, E. Urquía-Grande, D. Pascual-Ezama, M. Akpinar, and C. Rivero, "Solving the mystery about the factors conditioning higher education students' assessment: Finland versus Spain," *Education* + *Training*, vol. 62, no. 6, pp. 617–630, 2020.
- [11] A. Van Barneveld, K. E. Arnold, and J. P. Campbell, "Analytics in higher education: establishing a common language," *EDUCAUSE learning initiative*, vol. 1, no. 1, 2012.
- [12] S. Ray and M. Saeed, Applications of educational data mining and learning analytics tools in handling big data in higher education. International Conference on Applications of Big Data Analytics, , pp. 135–160, geeksforgeek, 2018.
- [13] V. Hegde and P. P. Prageeth, "Higher education student dropout prediction and analysis through educational data mining," *International Conference on Inventive Systems and Control (ICISC)*, pp. 694–699, 2018.
- [14] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "Prediction of Student Academic Performance by an Application of Data Mining Techniques," *International Conference on Management and Artificial Intelligence*, pp. 110–114, 2011.
- [15] M. Ashraf, M. Zaman, and M. Ahmed, "An intelligent prediction system forEducational data mining based on ensemble and filtering approaches," *Procedia Computer Science*, vol. 167, pp. 1471–1483, 2020.
- [16] A. Pradeep and J. Thomas, "Predicting college students dropout using EDM techniques," *International Journal of Computer Application*, vol. 123, no. 5, pp. 26–34, 2015.
- [17] M. Tsiakmaki, G. Kostopoulos, S. Kotsiantis, and O. Ragos, "Implementing AutoML in educational data mining for prediction tasks," *Applied Sciences*, vol. 10, no. 1, pp. 90–117, 2019.
- [18] M. Utari, B. Warsito, and R. Kusumaningrum, "Implementation of Data Mining for Drop-Out Prediction Using Random Forest Method," in *Proceedings of the International Conferenceon Information and Communication Technology*, pp. 1–5, Yogyakarta, Indonesia, June2020.
- [19] C. Vialardi-Sacín, L. Shaftr, J. Braver, and A. Ortigosa, "Recommendation in higher education using data mining techniques," *Journal of Educational Data Mining*, pp. 190–199, 2009.
- [20] A. A. Rimi, A. A. Ibrahim, and O. Bayat, "Developing classifier forthe prediction of students' performance using data mining ClassificationTechniques," AURUM Mühendislik Sistemleri ve Mimarlik Dergisi, vol. 4, no. 1, pp. 73–91, 2020.

- [21] M. Aggarwal, "Performance analysis of different feature selection methods in intrusion detection," *International Journal* of Scientific & Technology Research, vol. 2, pp. 225–231, 2013.
- [22] M. Nilashi, H. Ahmadi, A. A. Manaf et al., "Coronary heart disease diagnosis through self-organizing map and fuzzy support vector machine with incremental updates," *International Journal of Fuzzy Systems*, vol. 22, no. 4, pp. 1376–1388, 2020.
- [23] S. Jayaprakash, S. Krishnan, and V. Jaiganesh, "Predicting students academic performance using an improved random forest classifier," in *Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics*, pp. 238–243, Pune, India, March 2020.
- [24] L. I. P. Aji and A. Sunyoto, "An implementation of C4.5 classification algorithm to analyze student's performance," in *Proceedings of the 2020 3rd International Conference on Information and Communications Technology*, pp. 126–130, Yogyakarta, Indonesia, November 2020.