Materials Today: Proceedings xxx (xxxx) xxx

Contents lists available at ScienceDirect

Materials Today: Proceedings

journal homepage: www.elsevier.com/locate/matpr

Classification and prediction of student performance data using various machine learning algorithms

Harikumar Pallathadka [a,*], Alex Wenda [b], Edwin Ramirez-Asís [c], Maximiliano Asís-López [d], Judith Flores-Albornoz [e], Khongdet Phasinam [f]

[a] Manipur International University, Imphal, Manipur, India
[b] Department of Electrical Engineering, Faculty of Science and Technology, State Islamic University of Sultan Syarif Kasim Riau, Indonesia
[c] Management and turism Faculty, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Perú
[d] PhD in Computer Science and Engineering, Science Faculty, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Perú
[e] PhD in Environmental Engineering, Environmental Sciences Faculty, Universidad Nacional Santiago Antúnez de Mayolo, Huaraz, Perú
[f] Faculty of Food and Agricultural Technology, Pibulsongkram Rajabhat University, Phitsanulok, Thailand

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Educational Data Mining
Machine Learning
Student Performance
Classification
Prediction

ABSTRACT

In today's competitive world, it is critical for an institute to forecast student performance, classify individuals based on their talents, and attempt to enhance their performance in future tests. Students should be advised well in advance to concentrate their efforts in a specific area in order to improve their academic achievement. This type of analysis assists an institute in lowering its failure rates. Based on their prior performance in comparable courses, this study predicts students' performance in a course. Data mining is a collection of techniques used to uncover hidden patterns in massive amounts of existing data. These patterns may be valuable for analysis and prediction. Education data mining refers to the collection of data mining applications in the field of education. These applications are concerned with the analysis of data from students and teachers. The analysis might be used for categorization or prediction. Machine learning such as Nave Bayes, ID3, C4.5, and SVM are investigated. UCI machinery student performance data set is used in experimental study. Algorithms are analysed on certain parameters like- accuracy, error rate.

Selection and peer-review under responsibility of the scientific committee of the International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology.

1. Introduction

Educational data mining [1] refers to data mining techniques used to analyze educational data. Educational institutions store a vast quantity of data in order to keep track of students, faculty, and courses. This data contains personal and academic information about students, personal and academic information about faculty, syllabus, question papers, circulars, and so forth. Various universities and independent organizations have begun to use educational data mining to improve the lives of their students and faculty. These strategies are included into their application programs in order for them to be compatible with their databases. A few instances of educational data mining are shown below.

One of the most significant requirements for every institute is the performance of its students. The performance of students [2] can be anticipated based on their prior academic performance. According to the findings, students' abilities and interests might be linked to their performance. This type of analysis allows teachers to focus more attention on the pupils who need it the most. The success of a teacher is frequently measured by the performance of his students. Every institute must assess its faculty strength. Teachers can be assessed depending on their students' performance, comments, and so on. This type of analysis assists an institute in improving the quality of its instruction. Question papers can be evaluated to determine the level of difficulty. Such information assists an institute in normalizing the marks of all students in multi-session examinations.

* Corresponding author.
E-mail addresses: harikumar@miu.edu.in (H. Pallathadka), alexwenda@uin-suska.ac.id (A. Wenda), ehramireza@unasam.edu.pe (E. Ramirez-Asís), masisl@unasam.edu.pe (M. Asís-López), jfloresa@unasam.edu.pe (J. Flores-Albornoz), phasinam@psru.ac.th (K. Phasinam).

https://doi.org/10.1016/j.matpr.2021.07.382
2214-7853/
Selection and peer-review under responsibility of the scientific committee of the International Conference on Nanoelectronics, Nanophotonics, Nanomaterials, Nanobioscience & Nanotechnology.

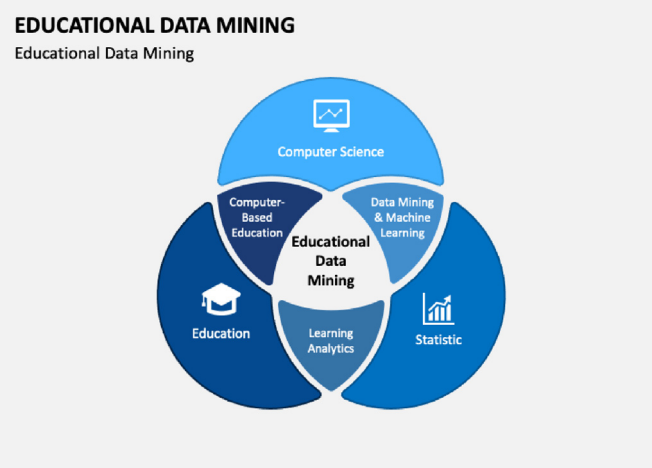

Fig. 1. Machine learning and Educational Data Mining.

As seen in Fig. 1, machine learning [3] is critical in educational data mining. It gives the capacity to forecast in the educational sector. One advantage of this approach is that it can identify recurring queries. Question papers from competitive examinations can be evaluated to determine the average weightage of each topic. Every course provides a collection of subjects on a semester or yearly basis. Some issues are connected to one another, while others are not. It is frequently seen that if a student does not study fundamental courses, he does poorly in advanced courses. Education data mining [4 5] assists in identifying such a group of disciplines that are interdependent. This information assists pupils in determining which disciplines are crucial in the future.

2. Literature survey

A literature review is a methodical approach to reviewing and comprehending current material that has been offered thus far. The primary goal of doing a literature study is to determine the extent to which an issue has been solved and what future alternatives for improvement exist. Every real-life problem is always handled incrementally. A researcher's solution is utilized as a foundation to improve it by decreasing or eliminating its constraints. Similarly, in the field of data mining-based research, it is necessary to apply contemporary algorithms as well as modern data schemes to improve the efficiency of our system. Sometimes limits are caused by the limits of the algorithms we utilized.

Constraints can also be attributed to the limitations of the data that we have utilized. In both circumstances, a good survey is essential before fixing any or both of these types of constraints to determine if such improvements have previously been done by someone else or not. If the same research effort has already been done someplace, there is no use in repeating it. It is necessary to consider all past solutions that have been suggested in order to evaluate their benefits and drawbacks. Later on, the knowledge gathered from the evaluation of literature assists a researcher in identifying a research need. Following that, a set of objectives may be identified, and a suggested system might be created to meet those objectives.

Many types and techniques of assessments investigated accessible in teaching and learning to assist educators in discovering excellent assessment techniques. Technology support was also deemed unavoidable and essential for assessment. For standard evaluation, a rubric was proposed as the most effective method [6].

A method[7] suggested based on a thorough analytic hierarchy method for evaluating teacher performance, which did quantitative and qualitative analysis. It is examined to see if the degree comparison design is reasonable by projecting itself towards a consistent aim, ensuring impartiality in subjective evaluation by various institutions and colleges. The research gap is that it has to be filled by correlating assessments with actual performance.

A qualitative and quantitative study is performed to get a better understanding of the value of partnership between higher education institutions and industry. Several impediments for this collaboration have been found, including reluctance to change, an aging academician, a cultural gap, an attitude toward innovation, a tendency to isolate, and insufficient facility. Among the benefits noted were efficiency, efficacy, and the quality of graduates in terms of employability [8]. Research work in [9] concentrated on students' learning processes. The topic of students as subjects in terms of their learning advantages has been explored. The use of experiments was thought to be a method of motivating pupils in their topics.

A student satisfaction index evaluation method based on factor analysis is provided [10], then empirically tested and updated the model. Student happiness was shown to be closely connected to instructional equipment, materials (particularly E-learning), network resources, and instructor supervision, rather than academic standards of instructors or teaching techniques.

A comprehensive teaching learning process management system is build and implemented that includes process tracking and control in addition to teaching resource management. The system also includes modules for autonomous resource allocation, job flow, and rule engine [11].

A survey on data mining techniques used in higher education is conducted [12]. They discovered that data mining is a vital tool in the education business since it aids in the discovery of several trends common in educational data relating to various areas such as the teaching learning process. The research gap is that no single instrument can meet the needs of all educational systems. Furthermore, the technologies are too sophisticated for educators to use and must be integrated into an e-learning environment.

Research [13] investigated the predictive model quality provided by machine learning algorithms for student retention management. When compared to other models, decision trees offer categorization rules that are easier to understand. The experimental findings shown that predictive models create a concise and accurate prediction list for student retention in addition to identi-

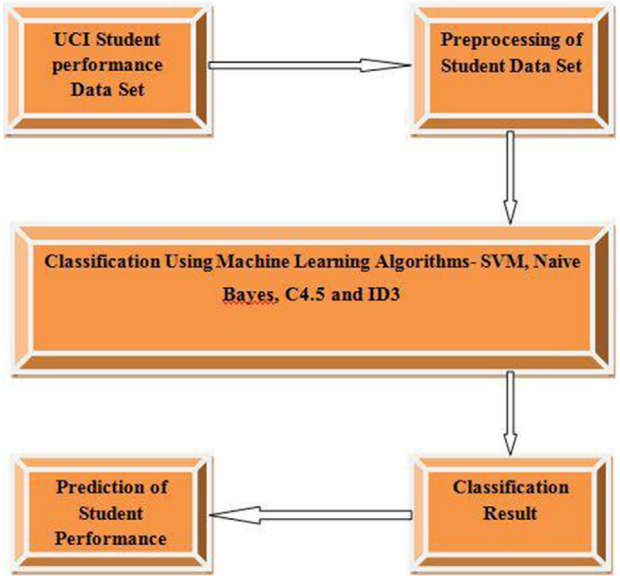

Fig. 2. Framework for Student Performance Prediction.

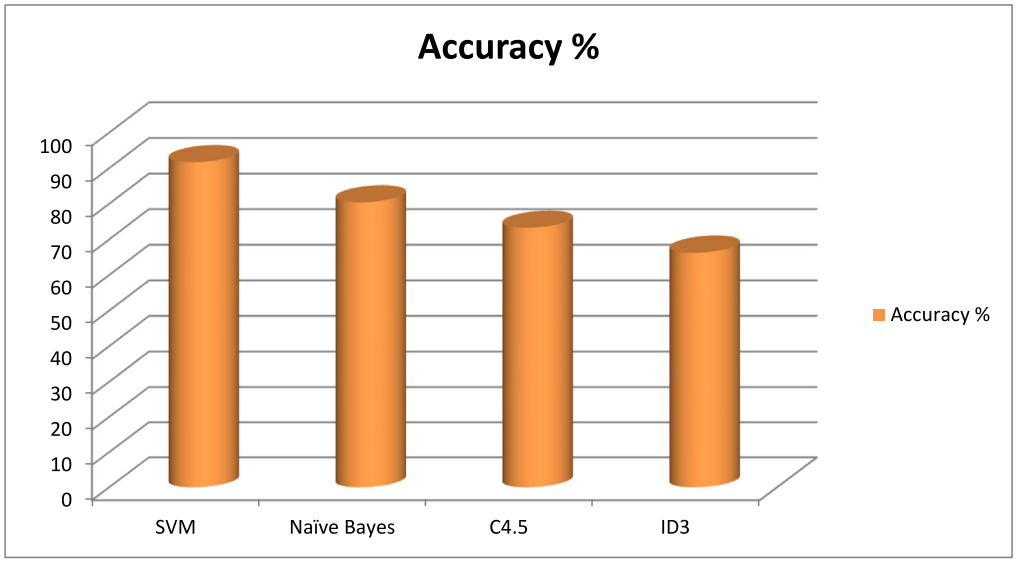

Fig. 3. Student Data Classification Results.

fying students that require particular attention, hence lowering the dropout rate. The research gap is that the anticipated dropout students' shortcomings are not visible enough to permit prompt correction interventions.

Method [14] investigated and examined the usefulness of several classification algorithms in predicting student academic achievement based on a variety of parameters. In contrast to Reptree, SimpleCart, Decision table, and J48 machine learning algorithms, it was discovered that neural network-based categorization had the highest precision, followed by Naive Bayes and ID3 algorithms. However, a multi-strategy machine learning technique can overcome the shortcomings of any of these models.

Ttwo classification methods, decision tree and fuzzy genetic algorithm used [15], to predict students' academic performance in both bachelors and masters degrees, which served as a feed forward mechanism for teaching faculty to pay more attention to these students before it was too late. It also aided in the placement of skilled pupils in reputable organizations. The decision tree revealed more students in the danger class, but the genetic algorithm revealed more passed students as a result of students being classified as safe between the danger and safe states. The research gap is that the decision tree takes a pessimistic approach, but the genetic algorithm takes a completely optimistic one, which may result in uncertainty entering into the outcomes. Faculty lacks unique variables that contribute to low performance in each student.

Framework [16] used ID3 and C4.5 independently to assess course assessment surveys and classify good and poor student performances as well as their topic deficiencies. Instructors have an edge in identifying students who are at danger of low performance and taking timely steps to enhance their performance. It was also discovered that for small data sets, both methods were equally accurate, however for big data sets, C4.5 was somewhat more accurate than ID3. The experiment identifies prospective poor performance but not the underlying causes of such deficiencies.

A review of decision tree, C4.5, Naive Bayesian, RIPPER, and SVM prediction algorithms and compared their top results in several aspects. In terms of FP rate, Precision, F-M, Recall, and MCC, the Naive Bayes method outperforms others. The algorithms' usefulness in quality management of teaching and learning in higher education institutions has yet to be investigated [17].

Researchers [18] employed a decision tree classification methodology on students' assessment results to identify pupils at risk of poor performance in order to enhance the quality management of the teaching learning process. The research gap is that this methodology does not identify pupils' individual limitations.

A model was developed [19] that successfully divided students into one of two groups based on their performance at the conclusion of their first academic year, as well as identifying relevant variables influencing their success. The model is based on information about student accomplishment in high school and their courses after finishing their first year of study, as well as the rank of preferences provided to the observed faculty, and it attempts to categorize students into one of two groups depending on their academic accomplishment.

3. Proposed methodology & result analysis

A framework for student performance prediction is shown below in Fig. 2. This framework has student performance data set as input. This student data set is preprocessed to remove noise from the data , to make input data set consistent. Then various machine learning algorithms- Nave Bayes, ID3, C4.5, and SVM are applied on input data set. Classification of data is performed. Results of classification of various algorithms are compared.

In experimental analysis UCI machinery student performance data set [20] is used. This data set has 33 attributes and 649 instances. This data set was donated by University of Minho, Portugal. The accuracy of various machine learning algorithms is shown below in Fig. 3. Machine learning algorithms were applied on student data set. The results obtained in terms of accuracy of classification are shown in graph.

4. Conclusion

The performance of students is one of the most significant criterion for every college. The performance of students can be anticipated based on their prior academic results. According to the findings, students' talents and interests might be linked to their performance. This type of analysis allows teachers to focus more on the pupils who need it the most. A teacher's success is frequently measured by the performance of his students. Every institute should assess its faculty strength. Teachers can be assessed depending on their students' results, comments, and so on. This type of analysis aids an institute in improving the quality of its

instruction. Question papers can be evaluated to determine difficulty levels. Education data mining is a set of data mining applications used in the field of education. These programs deal with data analysis from students and teachers. The analysis might be used to categorize or forecast. Machine learning algorithms such as Nave Bayes, ID3, C4.5, and SVM are being researched. In the experimental investigation, a data set of UCI machinery student performance is employed. Algorithms are evaluated based on characteristics such as accuracy and error rate. SVM is the most accurate technique for classifying a data set of student performance.

CRediT authorship contribution statement

Harikumar Pallathadka: Visualization. **Alex Wenda:** Data curation, Conceptualization, Methodology. **Edwin Ramirez-Asís:** . **Maximiliano Asís-López:** . **Judith Flores-Albornoz:** Writing – original draft. **Khongdet Phasinam:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] L. Ji, X. Zhang, L. Zhang, Research on the Algorithm of Education Data Mining Based on Big Data, in: 2020 IEEE 2nd International Conference on Computer Science and Educational Informatization (CSEI), 2020, pp. 344–350, https://doi.org/10.1109/CSEI50228.2020.9142529.

[2] A. Aleem, M.M. Gore, Educational Data Mining Methods: A Survey, in: 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT), 2020, pp. 182–188, https://doi.org/10.1109/CSNT48778.2020.9115734.

[3] R. Manne, S.C. Kantheti, Application of Artificial Intelligence in Healthcare: Chances and Challenges, Current Journal of Applied Science and Technology 40 (6) (2021) 78–89, https://doi.org/10.9734/cjast/2021/v40i631320.

[4] A. Hicham, A. Jeghal, A. Sabri, H. Tairi, A Survey on Educational Data Mining [2014-2019], International Conference on Intelligent Systems and Computer Vision (ISCV) 2020 (2020) 1–6, https://doi.org/10.1109/ISCV49265.2020.9204013.

[5] S. Kovalev, A. Kolodenkova, E. Muntyan, Educational Data Mining: Current Problems and Solutions, V International Conference on Information Technologies in Engineering Education (Inforino) 2020 (2020) 1–5, https://doi.org/10.1109/Inforino48376.2020.9111699.

[6] S.R. Hamidi, Z.A. Shaffiei, S.M. Sarif, N. Ashar, Exploratory study of assessment in teaching and learning, International Conference on Research and Innovation in Information Systems (ICRIIS) (2013) 398–403.

[7] J. Li, J. Zhao, G. Xue, Design of the index system of the college teachers' performance evaluation based on AHP approach, International Conference On Machine Learning And Cybernetics, Guilin: IEEE 2018 (2011) 995–1000.

[8] K. Ramakrishnan, N.M. Yasin, Higher learning institution — Industry collaboration: A necessity to improve teaching and learning process, in: 6th International Conference on Computer Science & Education (ICCSE), 2011, pp. 1445–1449.

[9] Staron, M., (2007), Using Experiments in Software Engineering asan Auxiliary Tool for Teaching – A Qualitative Evaluation from the Perspective of Students' Learning Process', 29th International Conference on Software Engineering, ICSE 2007, pp. 673 – 676.

[10] X. Feng, G. Hui, Study on the Evaluation Model of Student Satisfaction Based on Factor Analysis, International Conference on Computational Intelligence and Software Engineering (CiSE) (2010) 1–4.

[11] Hua, Z., Xue-qing, L., Jie-cai, Z., & Jiang-man, X. (2009), 'Research and implementation of Course Teaching learning Process Management System', IEEE International Symposium on IT in Medicine & Education, ITIME '09., Vol. 1, pp. 865 – 871.

[12] D.K. Gautam et al., 'Accreditation of engineers for effective implementation of the Washington accord', Achieving excellence through Accreditation, First world summit on accreditation WOSA-2012, NBA, New Delhi, 2012, pp. 1–15.

[13] S. Yadav, B. Bharadwaj, S. Pal, Mining Education Data to Predict Student's Retention: A comparative Study Retrieved from Retrieved from International Journal Of Computer Science And Information Security 10 (2) (2012) 113–117. http://sites.google.com/site/ijcsis/.

[14] K. David Kolo, S. A. Adepoju, J. Kolo Alhassan, A Decision Tree Approach for Predicting Students Academic Performance, International Journal Of Education And Management Engineering 5 (5) (2015) 12–19, https://doi.org/10.5815/ijeme10.5815/ijeme.2015.0510.5815/ijeme.2015.05.02.

[15] V. Dhanalakshmi, D. Bino, A. Saravanan, Opinion mining from student feedback data using supervised learning algorithms, in: 3rd MEC International Conference on Big Data and Smart City, Piscataway, New Jersey, IEEE, 2016, pp. 1–5.

[16] A.B. Raut, A.A. Nichat, Students Performance Prediction Using Decision Tree Technique, International Journal of Computational Intelligence Research 13 (7) (2017) 1735–1741.

[17] P.V.V.S. Eswara Rao, S.K. Sankar, Survey on Educational Data Mining Techniques, International Journal Of Engineering And Computer Science (2017), https://doi.org/10.18535/ijecs10.18535/ijecs/v6i4.41.

[18] G. Kavitha, L. Raj, Educational Data Mining and Learning Analytics - Educational Assistance for Teaching and Learning, International Journal Of Computer & Organization Trends 41 (1) (2017) 21–25.

[19] J. Mesarić, D. Šebalj, Decision trees for predicting the academic success of students, Croatian Operational Research Review 7 (2) (2016) 367–388, https://doi.org/10.17535/crorr10.17535/crorr.201610.17535/crorr.2016.0025.

[20] https://archive.ics.uci.edu/ml/datasets/Student+Performance