EDUCATIONAL DATA MINING: TOOLS AND TECHNIQUES STUDY

Smita J. Ghorpade[1], Seema S. Patil[2], Ratna S. Chaudhari[3]

Assistant Professor[1], Assistant Professor[2], Assistant Professor[3]
Department of Computer Science
K.R.T. Art's, B.H. Commerce, A.M. Science (KTHM) College, Nashik, India

Abstract— The potential influence of data mining analytics in higher education is a novel emerging field of research. Data from various educational organizations is explored and made operational, for various facets concerned with students. Educational data mining is an interdisciplinary research domain that enables to understand student's learning behavior, student's academic performance and analyses of the learning pattern. Educational data mining provides intrinsic knowledge about education system. This knowledge is useful to administrators for decision making, enhance the teaching learning process and develop student focused strategy. This paper presents various components of educational data mining, literature survey, objectives of EDM, tools and techniques used in EDM.

IndexTerms – **Data Mining, Educational Data Mining, Educational System, EDM Techniques, EDM Tools**

I. INTRODUCTION

Data Mining can be described as systematic processing of Big Data sets and finding hidden patterns and facts. Data mining techniques can be applied in various fields like business, medical, marketing, fraud detection etc. In recent years, applying data mining in education field is a developing interdisciplinary research area known as Educational Data Mining (EDM). In EDM process, unstructured or semi-structured data is collected from various educational systems and converted into inherent useful information that could have a great impact on educational research and practice. This process is very similar to other application areas of data mining like business, market basket, genetics, and medicine. EDM refers to tools and techniques used to automatically extract meaningful data from large repositories of data generated by learning activities in education stream.

The main goal of EDM is to understand student behavior, learning and improve educational outcomes. EDM discovers domain content knowledge, assessment outcomes, educational functionalities and applications, effect of instructional strategies surrounded by various learning environment. Learning Management System track information such as how many times each student has accessed each object, how many minutes the learning object was displayed on user's computer screen. For intelligent tutoring systems, learner submits a solution to a problem; these solutions are matched with expected solution and does analysis (B.M. Monjurul Alom et al. 2018).

In Higher Educational Institutes (HEIs) there are many important facets related to learning and teaching process. Major important aspects are Educational Data Mining (EDM), Academic Analytics and Learning Analytics (LA). EDM helps to determine the hidden learner's data in the learning environment. The learner's data is collected and reported by using learning analytics. Both LA and EDM aimed at enlightening quality education by improving interventions based on the nature of the analysis of big data in the learning environment. Academic analytics uses analytics methods to achieve the wants of institutional, functional, administration, and accounting decision-making practices. LA plays a vital role to improve HEIs performance and help them to face challenges in academia. Learning Analytics is a powerful resource that worked as a mirror for Higher Educational Institutes. Recently, distance education and online courses is growing need. The most focused areas are the adoption of "Massive Open Online Courses (MOOCs)" and "Learning Management Systems (LMS)" platforms in the educational environment. HEIs needs to turn its attention towards online education and analyses its challenges and improve quality. In online courses major challenges are network issue, unable to attend the course, engage learner efficiently, bored, puzzled, learners progress. EDM and LA technologies effectively recognize and support both students and teachers in the learning process.

This paper aims to illustrate key features of EDM, objectives of EDM, different data mining techniques used in higher education, and important tools used in EDM.

Following figure1 shows the different areas involved in Educational Data Mining.

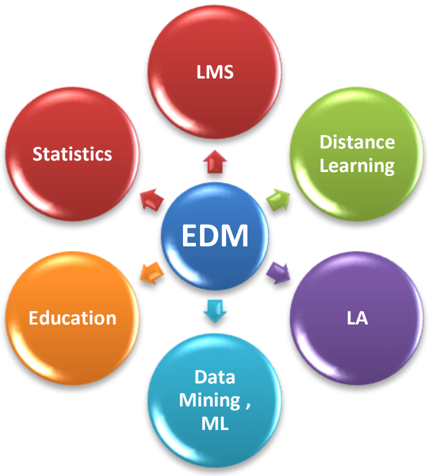


Figure. 1 EDM in Higher Education

II. GOALS OF EDM

1. Predicting Learner’s behaviors: Identify the characteristics of learners like knowledge, cognition, motivations, and attitudes. Construct a student model based on these characteristics and categorize students.

2. Instructional Design: Design and construct a teaching material in such way that learner should enjoy it and get fruitful knowledge domain. It is necessary to use different conceptual models, pictures to simplify the concepts.

3. Effective Pedagogy: It is necessary to use effective learner centric pedagogy. The five major approaches for effective pedagogy are Constructivist, Collaborative, Integrative, Reflective and Inquiry Based Learning. Effective pedagogies reflect on longer term learning outcomes as well as short-term goals. A student-centered teaching and learning approach, results into positive impacts in the learner’s behavior.

III. STAKEHOLDERS OF EDM

1. Learners: Optimizing individual learning styles, learning materials, and learning experiences, or recommending them.

2. Educators: Analyzing students’ learning behaviors, gaining the most supportive instruction, and predicting student learning pattern to increase effectiveness in teaching-learning process.

3. Researchers/Developers: Use data mining techniques for effectiveness, rectify learning materials, improving learning systems, in teaching learning process use effective pedagogy also analyses learning theories through data mining. Bridging data mining and learning science.

4. Organizations: EDM improves decision- making process in higher education learning institutions in terms of quality, efficiency and cost, such as admission processes and financial resources distribution (source - https://towardsdatascience.com/why-iseducational-data-mining-importantin-the-researche78ed1a17908)

IV. RELATED WORK

In last few decades, data mining techniques are used in educational environments actively and gaining much popularity in recent times. About EDM different surveys have been published. This section summarizes these works in this field. First EDM survey was identified in the literature in 2007 by Romero and Ventura. Later on lots of innovations happened in this area some of these are mentioned here. (Eduardo Fernandes et al., 2018) In this research, author has prepared a methodology for analyzing the predictive performance of students. In this paper author has used descriptive statistical analysis techniques to find new patterns of knowledge for students of public high school in the Federal District of Brazil. This discovered knowledge provides information to teachers, counsellors, city officials. This knowledge assists them in the social work initiatives, educational material and development of public policies in order to support students in the schools. Author has selected third year high school students as a target group.

The main goal is to improve their grades to at least a passing level so that their academic trajectory follows uninterrupted, guaranteeing their graduation, and a possibility of higher education. Author has used Gradient Boosting Machine (GBM) at the beginning of 2015 and 2016. Author has calculated achievement scores and become available bimonthly grades. These grades verify an impact on the precision of the model. This helps in to identify weaker students which were prone to failure. The proposed method is based on CRISP-DM and Student database is used. In experiment certain attributes are analyzed such as student’s attendance, student's ability to access a school or not, and the student's housing situation are relevant to the student's performance, social and personal attributes. These attributes are important and relevant factors to for students result, whether student passes or fails at the end of the school year.

(B.M. Monjurul Alom et al., 2018) In this paper author explores an attractive interdisciplinary research domain that deals with educational context – EDM. This paper analyses and makes an assessment according to gender. Author has calculated successive rates of completion rate of higher education in Australia. Author has posited four research questions. For analysis, data was collected for year 2004 to 2015 from Australia University. Author has used Wilson Calculator, Orange, T, Rapid miner, Weka, KNIME, Tangra, Orange and tools for analysis purpose. Finally, author concluded that gender played an important role in completion of higher education. Some states have socio-economic effects. (Hanan Aldowah et al.,2019) This paper has focuses on four main aspects: learning analytics, predictive analytics, behavioral analytics and visualization analytics. Author has done the survey from 2000 till 2017. For developing EDM system, most of the data mining techniques are suitable. This review has found that some of the data mining are not normally used due to the complexity in nature. In higher education institutes, EDM/LA provide significant benefits. Most of the organizations has adopted this EDM system. This study aims to select right data mining technique for respective application. Additionally, this study also shows that, the application of EDM and LA in higher education are helpful for developing student-centered provision, decision making and real-time application base tools and techniques used.

(Brijesh Kumar Baradwaj et al., 2011) In this research, student's performance is evaluated based on classification task. There are many classification approaches. This research study has used decision tree method. For classification of data, ID3 algorithm is used. Data used for analysis are seminar and assignment marks, attendance, and class test. This data is fetched from student's previous database. This system also predicts the performance of students at the end of semester. (Insha Majedd et al.,2018) Author addressed work of major contributors in Educational Data Mining field. This paper elaborates the use of educational data mining in education system. This paper shows work done is Educational Data Mining and used in academic mining which is understandable to researchers. Author has shown comparison of different data mining techniques in education field. This study shows analysis of data using KNN, Decision Tree, Neural Network, SVM, Naïve Bayes techniques.

V. EDM TECHNIQUES

Educational Data Mining uses various techniques for analysis, data processing, to identify patterns, model construction and predicting results. Some of the techniques are:

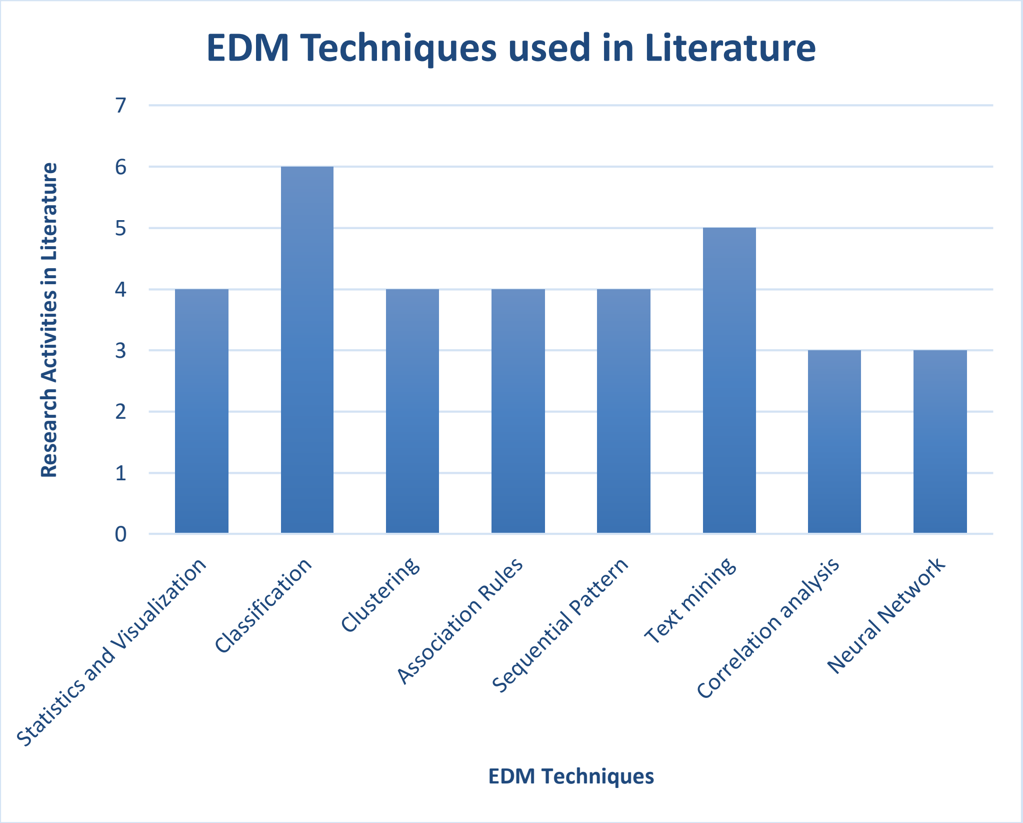
1. **Statistics and Visualization:** Statistical analysis plays a vital role in many data mining algorithms. There are many popular statistical techniques used for data analysis, particularly for numeric data. These techniques have been applied extensively to scientific data as well as to all sectors. Visual data mining shows implied and useful knowledge from enormous amount of data sets using data and knowledge visualization techniques. One picture is worth than one thousand word. The human visual system in controlled by the eyes and brain, the latter of which can be thought of as a powerful, highly parallel processing and reasoning engine containing a large knowledge base. Visualization is an attractive and effective tool for comprehension of data distributions, patterns, clusters, outliers.
2. **Classification:** is the form of data analysis that extracts relevant data through classification models. Such models, called classifiers which predicts categorical class labels. In machine learning, many classification methods have been proposed. Most of the algorithms are memory resident typically assume small data size. Data mining classification technique has strongest feature-scalability. These algorithms are capable of handling large amount of disk-resident data. In EDM also it plays vital role for data analysis and classification. It is supervised learning.
3. **Clustering**: Clustering is the process of grouping a set of data objects into multiple groups called as clusters. Objects within a cluster have high similarity. But objects are very dissimilar with other clusters. For the objects involve in distance measure, similarities and dissimilarities are assessed. This assessment is based on the attribute values of the objects. Scalability and Incremental clusters are the striking feature of clustering. It is an example of unsupervised learning.
4. **Association Rules and Sequential Pattern Mining**: Frequent patterns are patterns that appear frequently in a data set. Finding frequent patterns plays an essential role in mining associations, correlations and many other interesting relationships among data. After finding association among data sets, certain rules are generated. It works on massive amount of data in retail community and industries. Association rule mining finds strong associations and relationships among large data sets. This rule shows association among frequent items in itemset. Sequence mining discovers interesting patterns in data with respect to some subjective or objective measure of interest. To find correlations, frequent patterns lot of tools are available.
5. **Text Mining**: Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics and computational linguistics. It derives high quality information from text. This is typically done through the discovery of patterns and trends by means such as statistical pattern learning, topic modelling and statistical language modelling. Extract meaningful numeric indices of the text from unstructured information is the main purpose of text mining.
6. **Correlation Analysis**: Correlation Analysis is statistical method that is used to discover if there is a relationship between two variables; and how strong that variables are related to each other. For nominal data, we use chi-square test. Commonly we use correlation coefficient and covariance for numeric data, both methods access how one attribute's values vary from those of another.
7. **Neural Network**: It is a set of connected units in which each connection has a weight associated with it. To predict the correct output, learner adjust the weights during learning phase. The main advantage of neural network is that it provides high tolerance of noisy data. It is a technique to improve the interpretability of the trained network by using extracted rules for learning networks. To predict academic performance based on residency, ethnicity.

VI. DATA MINING TECHNIQUES USED IN HIGHER EDUCATION INSTITUTIONS

In above section, various data mining techniques used in EDM are introduced. These techniques are used to enrich the quality in higher education institutions. Visualization, Prediction, Monitoring, Association among elements, similarities among items, are the activities carry out by above techniques. The summary of usage of these techniques in EDM are shown in following table 1.

Table 1: EDM Techniques

Techniques	**Its use in Literature during period 2015-2020**
Statistics and Visualization	[1],[17], [22], [25]
Classification	[4],[20],[21],[27],[34],[40]
Clustering	[4],[26], [28],[31],[33]
Association Rules	[24],[29],[30]
Sequential Pattern	[19], [23],[36],[42]
Text mining	[32],[35],[37],[39],[40]
Correlation analysis	[18],[34],[44]
Neural Network	[38],[41],[43]



VII. IMPORTANT TOOLS FOR EDM

There is a wide range of algorithms and modelling frameworks that can be used to model and predict processes and relationships in educational data. In section V some of EDM techniques are discussed. To work on that algorithms and techniques following EDM tools are used by researcher. Table 2 shows effective tools used in EDM.

Table 2: EDM Tools

Tools	Description
Rapid Miner	A software platform used for data preparation, text mining and predictive analysis and create models.
WEKA (Waikato Environment for Knowledge Analysis)	It is a software tool which is used to implement data mining algorithms and build a model for algorithm
SPSS (Statistical Package for the Social Sciences)	SPSS is primarily a statistical package, and offers a range of statistical tests, regression frameworks, correlations, and factor analyses
KNIME (Konstanz Information Miner)	A tool used for data cleaning and analysis
Orange	It is a data analysis and visualization tool
KEEL (Knowledge Extraction Based on Evolutionary Learning)	It assesses evolutionary algorithms for data mining problems including regression, classification, clustering, pattern mining etc
Tangra	It offers various data mining methods from statistical learning, data analysis, and machine learning.
H2O	It is used to perform data analysis on the data held in cloud computing application systems
Rattle	It is GUI based data mining tool that uses R stats programming language. Rattle exposes the statistical power of R by providing considerable data mining functionality.
Spark MLLib	It well suited for distributed data. It is framework for data which is distributed across various nodes.
D3js	Data visualization tool used for complex data visualizations that require data handling

VIII. CONCLUSION AND FUTURE INSIGHTS

Data Mining is one of important analytical tool. It is used to enhance decision making, for finding new patterns and relationships for organizations. Educational Data Mining is an emerging discipline concerned with education system. EDM is extremely beneficial for predicting student performance, behavior. It also provides services to improve teaching- learning process, develop good instructional design, apply different pedagogy in teaching learning. This paper describes the key features of EDM, goals of EDM, objectives of EDM. This study also covers data mining techniques used in education, important tools of EDM. It presents a survey of tools used in EDM and presents review of current trends in EDM.

Apart from this there are still some challenges in Educational Data Mining. Educational data is growing rapidly and from this distilling massive amount of data requires set of algorithms. Traditional data mining algorithms have specific objectives and functionality; they cannot be applied directly to educational problems.

REFERENCES

[1] B.M. Monjurul Alom et al. 2018.Educational Data Mining Perspective from Primary to University Education in Australia. Information Technology and Computer Science.

[2] Hanan Aldowah et al. 2019. Educational data mining and learning analytics for 21st century higher education: A review and Synthesis. Telematics and Informatics.

[3] https://towardsdatascience.com/why-iseducational-data-mining-importantin-the-researche78ed1a17908

[4] Eduardo Fernandes et al. 2018. Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil. Journal of Business Research.

[5] Slater, S. et al. 2017. Tools for educational data mining: A Review.

[6] C. Romero et al. 2007. Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications Vol. 33,135-146.

[7] Brijesh Kumar Baradwaj et al. 2011.Mining Educational Data to Analyze Student's Performance. International Journal of Advanced Computer Science and Applications Vol-2, No. 6.

[8] Insha Majeed et al. 2018. Current State of Art of Academic data Mining and Future Vision. Indian Journal of Computer Science and Engineering (IJCSE).

[9] https://www.softwaretestinghelp.com/datamining-tools/

[10] M. N. Quadri. 2019. Methods and Usage of Educational Data Mining Tools. International Journal of Information Technology (IIJIT) Vol 7, Issue 3.

[11] Jiechao Cheng. 2017. Data Mining research in Education.

[12] Ben Kei Daniel. 2019. Big Data and data science: A critical review of issues for educational research. British Journal of Educational Technology, Vol 50.

[13] Siti Khadijah Mohamad et al. 2013. Educational Data Mining: A review. Procedia-Social and Behavioral Sciences.

[14] Ginika Mahajan et al. 2020. Educational Data Mining: A state-of-the-art survey on tools and techniques used in EDM. International Journal of Computer Applications & Information Technology Vol. 12, Issue No. 1.

[15] Hui-Chun Hung et al. 2020. Applying Educational Data Mining to Explore Students' Learning Patterns in the Flipped Learning Approach for Coding Education. Symmetry MDPI.

[16] S. Lonn, S. J. Aguilar et al.2015. Investigating student motivation in the context of a learning analytics intervention during a summer bridge program. Computers in Human Behavior, Vol. 47, pp. 90-97.

[17] Peter Sorenson et. al. Learner Interaction Monitoring System (LiMS): Capturing the Behaviors of Online Learners and Evaluating Online Training Courses.

[18] Abeer AlJarrah et al. 2018 Investigating temporal access in a flipped classroom :procrastination persists. International Journal of Educational Technology in Higher Education.

[19] Donia Malekian et al. 2020. Prediction of Students' Assessment Readiness in Online Learning Environments: The Sequence Matters. LAK20, March 23-27.

[20] Babandi Usman et al. 2020. Data Mining: Predicting of Student Performance Using Classification Technique. in International Journal of Information Processing and Communication (IJIPC) Vol. 8 No. 1 pp. 92-101.

[21] Maricel A. Timbal. 2019 Analysis of Student-at-Risk of Dropping out (SARDO) Using Decision Tree: An Intelligent Predictive Model for Reduction. International Journal of Machine Learning and Computing, Vol. 9, No. 3.

[22] Diana M. Naranjo et al. 2019. A Visual Dashboard to Track Learning Analytics for Educational Cloud Computing. Sensors 2019, 19, 2952; doi:10.3390/s19132952.

[23] Jacqueline Wonga et al. Exploring sequences of learner activities in relation to self-regulated learning in a massive open online course", in Computers & Education journal. homepage: www.elsevier.com/locate/compedu

[24] Sadiq Husssain et al. 2018. Classification, clustering and Association Rule Mining in Educational Datasets using Data Mining Tools : A case Study. Computer Science On-line Conference CSOC pp 196-211.

[25] Alenjandro pena-Ayala et al. Educational Data Mining: A survey and a data mining based analysis of recent works" in Expert Systems with Applications journal homepage: www.elsevier.com/locate/eswa.

[26] Natthakan Iam-On et al. 2017. Generating descriptive model for student dropout: a review of clustering approach. Iam-On and Boongoen Hum. Cent. Comput. Inf. Sci. (2017) 7:1 DOI 10.1186/s13673-016-0083-0.

[27] Bogdan Drăgulescu et al. Predicting Assignment Submissions in a Multiclass Classification Problem. TEM Journal 4(3):244-254.

[28] D. J. Salas et al. 2016. Supporting the Acquisition of Scientific Skills by the Use of Learning Analytics. International Conference on Web-Based Learning, Springer, pp. 281-293.

[29] D. Suganya et al. 2018. STUDENT PERFORMANCE DASHBOARD USING MINING APPROACH. International Journal of Pure and Applied Mathematics Volume 119 No. 12, 409-421.

[30] Pornthep Rojanavasu et al. 2019. Educational Data Analytics using Association Rule Mining and Classification International ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI-NCON)

[31] Yu Li et al. 2019.Educational data mining for students' performance based on fuzzy C-means clustering eISSN 2051-3305 doi: 10.1049/joe.2019.0938 www.ietdl.org.

[32] Swapna Gottipati et al. 2018. Text analytics approach to extract course improvement suggestions from students' feedback. Research and Practice in Technology Enhanced Learning.

[33] Husain Salilul Akareema et al. 2016. Determinants of education quality: what makes students' perception different. OPEN REVIEW OF EDUCATIONAL RESEARCH, VOL. 3, NO. 1, 52–67 http://dx.doi.org/10.1080/23265507.2016.1155167

[34] Nikola Tomasevic et al. 2019. An overview and comparison of supervised data mining techniques for student exam performance prediction. Computers & Education journal homepage: www.elsevier.com

[35] M. Erkens et al. 2016. Improving collaborative learning in the classroom: Text mining based grouping and representing. International Journal of Computer-Supported Collaborative Learning, Vol. 11, no. 4, pp. 387-415.

[36] J. Lamsa, R. et al. 2020. The potential of temporal analysis: Combining log data and lag sequential analysis to investigate temporal differences between scaffolded and non-scaffolded group inquiry-based learning processes. Computers & Education, Vol. 143, p. 103674.

[37] R. Ferreira Mello et al. 2019. Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 9, no. 6, pp. e1332.

[38] Maria Tsiakmaki et al. 2020. Transfer Learning from Deep Neural Networks for Predicting Student Performance. Appl. Sci. 10, 2145; doi:10.3390/app10062145

[39] H. Hind et al. 2017. Applying text mining to predict learners' cognitive engagement, in Proc. of the Mediterranean Symposium on Smart City Application, ACM.

[40] S Hussain et al. 2018. Classification, clustering and association rule mining in educational datasets using data mining tools: A case study CSOC2018: Computer Science On-line Conference

[41] Hamid Karimi. 2020. Online Academic Course Performance Prediction using Relational Graph Convolutional Neural Network. Proceedings of The 13th International Conference on Educational Data Mining.

[42] Nakamura et al. 2015. Sequential Pattern Mining System for Analysis of Programming Learning History. IEEE International Conference on Data Science and Data Intensive Systems IEEE pp. 69-74

[43] Syed Atif Ali Shah et al. 2020. An Enhanced Deep Neural Network for Predicting Workplace Absenteeism. Research Article Open Access , Article ID 5843932 | https://doi.org/10.1155/2020/5843932.

[44] A. AlJarrah et al. 2018. Investigating temporal access in a flipped classroom: procrastination persists. International Journal of Educational Technology in Higher Education, Vol. 15, no. 1, pp. 1, 201.