IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11242

Data Mining Methods and Techniques in Higher Education

Shilpa Kulkarni[1], Dr. Sasikala P[2]

Asst. Professor, Dept. of Computer Science, Nrupatunga University,

Nruptatunga Road, Bangalore – 560001, Karnataka[1,2]

Abstract: Data analysis is crucial for decision support in every firm, whether it is a manufacturing unit or an educational institution. Data mining techniques are used in a wide range of fields. This paper proposes the use of data mining techniques to improve the efficiency of higher education institutions. When data mining techniques such as clustering, decision trees, and association are used to higher education processes, they can aid in improving student performance, life cycle management, course selection, retention rate monitoring, and grant fund management. This is a strategy for determining how data mining tools affect higher education. Educational Data Mining (EDM) is an interdisciplinary research area focused on data mining's application in the field of education. It uses a number of tools and techniques from machine learning, statistics, data mining, and data analysis to analyses data created during teaching and learning. Educational Data Mining is the process of converting raw data from large educational databases into useful and meaningful information that can be used to better understand students and their learning environments, improve teacher assistance, and make educational system decisions. The goal of this research is to give a broad overview of educational data mining, including its uses and benefits.

Keywords: Educational data mining (EDM), learning analytics (LA), knowledge discovery in databases (KDD), data mining techniques, data mining methods. EDM tools, visualizations tools.

1. INTRODUCTION

In recent years, educational data mining (EDM) and learning analytics (LA) have gained appeal as alternate techniques to working with educational data to frequentist and Bayesian approaches. Data mining, also known as knowledge discovery in databases (KDD), comprises methodologies that explore for innovative and generalizable patterns and findings rather than attempting to validate prior beliefs. Today's universities operate in a complex and competitive environment. Due to rapid technical improvements and lower-cost IT equipment, the amount of data kept in educational databases is continuously expanding, but if this data is not further assessed, it remains merely vast amounts of data. Using data mining tools, methodologies, and strategies, we may study this data and uncover hidden patterns and information. Data mining is a method of discovering patterns and relationships in data in order to make better decisions. Statistics, artificial intelligence, neural networks, database systems, machine learning, pattern recognition, data visualization, knowledge acquisition, and information theory approaches are all used in this multidisciplinary discipline. "The process of sifting through huge volumes of data stored in repositories and applying pattern recognition technologies as well as statistical and mathematical approaches to find crucial new connections, patterns, and trends" [1].

Data mining can be used in a variety of ways. It's a term used in finance to describe the process of analyzing customer behavior data in order to increase client loyalty. It also assists in the discovery of illicit activity by revealing hidden links between various financial indicators. By gathering historical data and translating it into useable and legitimate information, it can detect both fraudulent and non-fraudulent behaviors. Data mining in healthcare can aid in the discovery of links between diseases and treatment outcomes. It also assists in the detection of healthcare insurance fraud. Law enforcement agencies utilize it to spot money laundering, narcotics trafficking, and other criminal patterns. In the telecommunications industry, data mining is frequently used to boost profitability by delivering personalized services and to reduce customer churn by researching demographic traits and forecasting user behavior. The results of the data mining technique can be used to develop successful marketing campaigns and pricing strategies. In marketing and sales, data mining techniques are used to find hidden trends in prior purchasing data. In market basket analysis, data mining results are utilized to discover customer behavior buying patterns and provide information on combinations of products purchased together. It can also be used to predict future trends and customer buying habits. In the banking industry, data mining tools are routinely used to predict client attrition, as well as to detect fraud and insolvency [2].

Data mining has a number of disadvantages, most notably in terms of user privacy and security. It must be clear how and with whom the information will be used and shared. The implementation stage is highly costly because data mining tools and processes work with massive amounts of data. Competent IT skills is essential for data preparation and identifying

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11242

the best model and technique for analysis. Data mining techniques may have significant ramifications and expenses because they are not 100 percent accurate. The method of applying educational data mining to analyses the effects of educational techniques is depicted in Figure 1. Lecturers are in charge of planning and teaching lessons using a variety of methods. Students use and participate in these activities alone and in groups. Data mining is a technique for classifying, detecting trends, and associating distinct concepts with data [3].

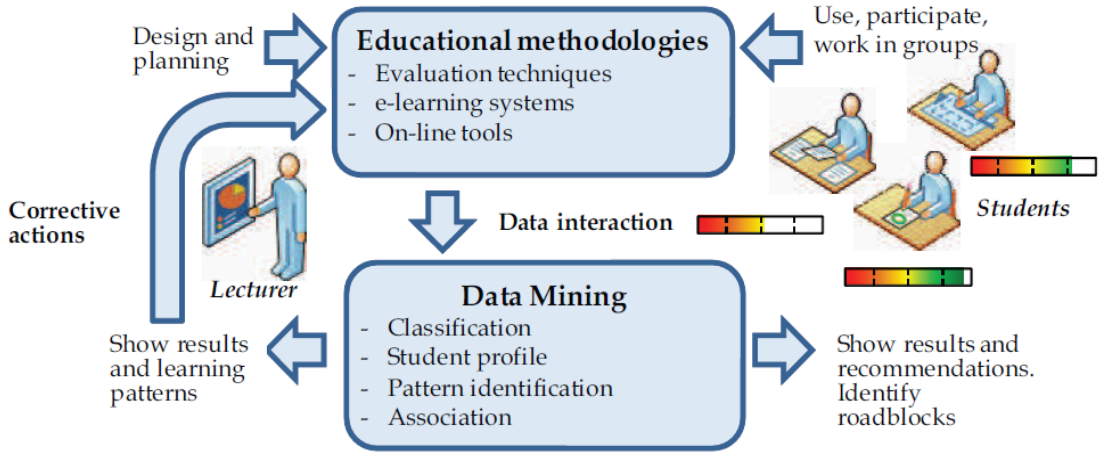

FIGURE 1. Data mining for identification of the student' learning pattern [3]

2. LITERATURE REVIEW

Several academics have been successful in their studies of data mining applications in education, as evidenced by prior reviews on the subject. Different data mining approaches are used in the studies to support various educational activities. Predicting student dropout, predicting students' cognitive skills in an educational setting, predicting at-risk students and slow learners, predicting student course selection and career choices, student retention management, and predicting student performance are some of the features supported. Predicting students' performance, on the other hand, is one of the most essential and helpful notions in educational data mining. Predicting a student's performance entails guessing an unknown value, such as the student's grade or mark [4].

3. EDUCATIONAL DATA MINING (EDM)

Many educational institutions' data collection and storage became too large, and educational data analysis could no longer be done manually. EDM is a new discipline that arose from the use of data mining tools to analyse educational data. There are various definitions of EDM, but they all agree that it is an interdisciplinary research area that analyses data collected during teaching and learning in order to discover previously unknown information, relationships, and patterns in large data repositories using methods and techniques from machine learning, statistics, data mining, and data analysis [5].

3.1 Educational Data Mining Process

The EDM process is divided into four stages. The initial process of data mining is problem definition, which involves translating a specific problem into a data mining problem. The project purpose and objectives, as well as the primary research topics, are developed during this phase. The second phase, Data preparation and gathering, is the most time-consuming. In data mining, data quality is a crucial issue. Source data must be found, cleansed, and formatted in a predetermined manner at this phase. Following that, there is a Modeling and evaluating phase, in which the parameters are set to optimal values and various modelling techniques are chosen and applied. The deployment phase is the final step in the data mining process, during which the results are organized and displayed in graphs and reports. It is vital to note that data mining is an iterative process, which means that the process does not end when a certain solution is implemented [6]. It could just be a new source of data for a new data mining technique.

3.2 Methods and Techniques

Educational data mining employs a variety of tools, algorithms, and techniques. Classification, clustering, prediction, and association are the most common applications. Neural networks, decision trees, regression analysis, and cluster analysis are the most often used data mining techniques.

Classification is a data mining process that divides data into desired categories or classes in a collection. It aids in the analysis of data and the prediction of outcomes. Classification's purpose is to correctly anticipate the target class for each

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11242

case in the data. The classifier training algorithm determines the set of parameters required for classification using pre-classified instances [7]. This technique is frequently used in the educational industry to classify pupils based on factors such as age, gender, grades, knowledge, academic achievements, motivation, behavior, demographic or regional characteristics, and so on.

A. Clustering Techniques

It's a way of categorizing a group of abstract objects into groups of parallel elements. Data analysis, pattern recognition, and image processing are among the most common applications of clustering algorithms.

B. Neural Networks

One of the key advantages of neural networks is that they may discover tasks depending on the data provided for training or preliminary experience. The neural network is capable of identifying all conceivable interactions between variables and predictors.

C. Decision trees

Decision tree strategies are more reliable for understanding and identifying the most promising factors in the shortest amount of time. It can also be used to determine the relativity of two or more variables. ID3, CART, C4.5, Random tree, and CHIAD are the most common algorithms used in decision trees.

D. Bayesian Classifier

To determine the parameters, this is a very easy and uncomplicated procedure that takes very little data preparation. The class conditional independencies between subsets of variables are clear in a Bayesian classifier. A graphical representation of causal links will enhance the learning process.

E. Neural networks.

Neural networks are one of the most commonly utilized algorithms in the field of education for predicting student performance. The most common rationale for employing ANN is that it can classify patterns without the need for any prior training. Because it is natively parallel and thus able to speed up the computation process, ANN is highly suited for prediction operations in the educational data mining arena.

F. Naive Bayes.

The naive Bayes technique is a supervised classifier that employs two simplifications: one that uses the conditional independence assumption and the other that ignores the denominator. It is based on using the Bayes' theorem with strong naïve independence assumptions between the explanatory variables.

G. K- nearest neighbor.

K-nearest-neighbor classifiers are analogy-based algorithms that learn by comparing themselves to test training tuples with similar characteristics. The algorithms can be used to return a real-valued forecast for an unidentified tuple in numeric predictions. As a result, the technique restores the reasonable value associated with the unidentified tuple's k-nearest neighbors. KNN performed well when used to predict student performance.

Examining the tasks performed and the technologies utilized can help institutional researchers better understand how data mining can help them. The following are the several types of data mining tasks: classification, estimation, segmentation, and description. Table 2.1 summarizes the tasks and the tools that go with them.

Table 1. Classification of Data Mining Tasks and Tools [7]

Tasks	Supervised	Unsupervised
Classification	Memory based reasoning, genetic algorithm, C&RT, link analysis, C5.0, ANN	Kohonen nets
Estimation	ANN, C&RT	--
Segmentation	Market basket analysis, memory based reasoning, link analysis, rule induction	Cluster detection, K-means, generalized rule induction, APRIORI
Description	Rule induction, market basket analysis	Spatial visualization

ANN, SVM, NB, CLU, RF, LR, DT, EM, KNN, NN, ANN, C45, CART are some of the other strategies that have been published in the recent literature for predicting student retention/dropout at various HE schools around the world.

2.3 Powerful EDM Tools

Data mining has a wide range of applications, including commodity, service, or product marketing and promotion, artificial intelligence research, biological sciences, crime investigations, and high-level government intelligence. As a result of their widespread use and the complexity involved in designing data mining applications, a large number of data mining tools have been produced throughout the years. Each instrument has its own set of advantages and disadvantages [8].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11242

A. Rapid Miner (YALE)

Rapid Miner is a software platform that combines machine learning, data mining, text mining, predictive analytics, and business analytics in one place. It is utilized for corporate and industrial applications, as well as research, education, training, rapid prototyping, and application development, and it supports all parts of the data mining process. Rapid Miner is a client/server program that can be operated on cloud infrastructures or as SaaS.

The graphical programming language of Rapid Miner is more powerful than most other data mining programs, and it includes many user-defined capabilities. Rapid Miner's Batch Cross Validation operator, for example, may do cross-validation at several layers. This capability is very useful for generalizability assessments, and it is a substantial advantage over the graphical languages of most other data mining applications. Rapid Miner also includes a number of metrics for model evaluations, as well as graphics such as Receiver-Operating Curves, to help with model fit evaluation. Models can be exported as mathematical models or xml files, which can then be used to run Rapid Miner code to run the model on new data. Rapid Miner's Application Program Interface (API), which can be integrated into Java or Python programs, can handle a range of tasks that the graphical programming language couldn't handle [9].

All of Weka's algorithms are included in Rapid Miner, which are listed below. Newer versions of Rapid Miner now feature crowd-sourced algorithm and parameter suggestions. Rapid Miner includes a plethora of tutorials to assist you in learning how to use the graphical programming language. Rapid Miner is free for academic use, and commercial licenses are available through Rapid-I.

B. WEKA

Weka stands for Waikato Environment for Knowledge Analysis. It's a set of machine learning methods for data mining. This tool is incredibly useful, and data mining professionals use it primarily for predictive modelling and fact analysis. When compared to rapid miner, it offers a number of benefits and supports popular educational data mining tasks such clustering, classification, visualization, data pretreatment, regression, and selection. These algorithms can be applied on a data set directly or called from Java code. The Weka workbench is a collection of data analytics and predictive modelling visualization tools and algorithms, as well as graphical user interfaces providing easy access to these capabilities [10].

Weka comes with a vast range of classification, clustering, and association mining algorithms that can be used individually or in combination, with techniques like bagging, boosting, and stacking. Data mining techniques can be accessed via the command line, a graphical user interface (GUI), or a Java API. The command line interface and APIs are more powerful than the graphical user interface, which does not allow users to use all of the advanced features. Weka can generate mathematical models or PMML (Predictive Modeling Markup Language) files, which may then be used to run the model on new data with the Weka scoring plugin [11].

C. R-Programming

R-Programming is used to write the modules, and R-Programming is written in C and Fortran. Data analysis and statistical software are built using the R programming language, and this software is utilized in the middle of the data miners. In educational data mining, R- Programming employs graphical, statistical, time series analysis, classification, and clustering methodologies to improve student performance.

D. Orange

For educational data mining researchers, The Orange is a Python-based open source tool. The machine learning method and the few bioinformatics and text mining features included to the Orange tool are its main advantages. Orange is a component-based data mining and machine learning software suite that includes a visual programming front-end for exploratory data analysis and visualization, as well as Python bindings and libraries. The software includes data preparation, feature scoring and filtering, modelling, model evaluation, and exploration methodologies [12].

E. KNIME

This KNIME tool is created in Java and is based on Eclipse. The major components of data preprocessing technique, transformation, extraction, and loading, were done well in this tool. Through a graphical user interface idea, the KNIME tool will allow the building of nodes for data processing. The open source scheme included business intelligence, financial data analysis, reporting, integration platform, and data analytics. It also includes a number of specialized algorithms for sentiment analysis and social network analysis, among other things.

F. Spark MLlib

Spark is a distributed data processing platform for large-scale data processing over several computer processors. Spark has an API that allows it to connect to a variety of computer languages, including Java, Python, and SQL, allowing these languages to be utilized for distributed processing. Several standard machine learning and data mining methods are implemented in Spark's MLlib machine learning framework. MLlib's capabilities is still restricted, and it is strictly a programmed tool, but its distributed nature makes it a quick and efficient option.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 2, February 2022

DOI: 10.17148/IJARCCE.2022.11242

3.3 VISUALIZATIONS TOOLS [9]

A. Tableau

Tableau is a set of interactive data visualization and analysis tools. Tableau is frequently used in educational settings to analyses student data, generate actionable insights, improve teaching practices, and expedite educational reporting, despite its primary focus on business intelligence. Tableau's main advantage is that it doesn't require any programming knowledge to analyses large amounts of data from many sources, making a wide range of visualizations available to a broader audience. Tableau allows you to connect or import data from a wide range of recognized formats (e.g., databases, data warehouses, and log data). Tableau also allows users to develop sophisticated and interactive dashboards that offer dynamic real-time representations to end users. Tableau, on the other hand, has some limitations, such as the inability to perform predictive analytics or relational data mining. Tableau is also not extendable, and as a commercial solution, it does not allow for interface with other software platforms.

B. D3js

D3.js is a JavaScript framework that enables academics and practitioners to alter data-driven documents and generate complex, dynamic data visualizations that require data handling and are optimized for modern web browsers. The ability to build a wide range of data visualizations without the need for installation, the flexibility to reuse code, and the fact that it is free and open source are just a few of the benefits of D3.js. Increased educational research implementation, on the other hand, is riddled with challenges. D3.js is a technology that requires extensive programming knowledge and has a number of compatibility issues as well as speed limitations for large data sets. Finally, there is no method to hide data from visualization users, therefore data pre-processing is required to maintain privacy and security [13].

3.4 Specialized EDM and LA Applications

In the preceding part, we looked at some general-purpose EDM modelling and analysis tools. Specific data and analysis goals, on the other hand, may need the use of additional specialized algorithms not found in these general-purpose tools.

A. Tools for Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) is a popular method for estimating latent knowledge that entails evaluating a student's knowledge while they learn online. This is different from the type of educational measurement often employed in tests in that knowledge changes while being measured while online learning is taking place. This forecasts whether a student has achieved or not mastered a specific ability in an intelligent tutoring system or similar application. BKT models are often fitted using one of two methods: brute force grid search or Expectation Maximization (EM).

B. Text Mining

Text mining is a fast expanding field of data mining, with a large number of programmes, apps, and APIs accessible for tagging, processing, and identifying textual data. Parts of speech, sentence structure, and semantic word meaning can all be processed using text analysis software.

4. DISCUSSION

From 2010 to 2020, the outcomes of the commonly used student performance prediction systems were examined and plotted on a graph to show how they differ in overall forecast accuracy. Figure 1 depicts the diagram.

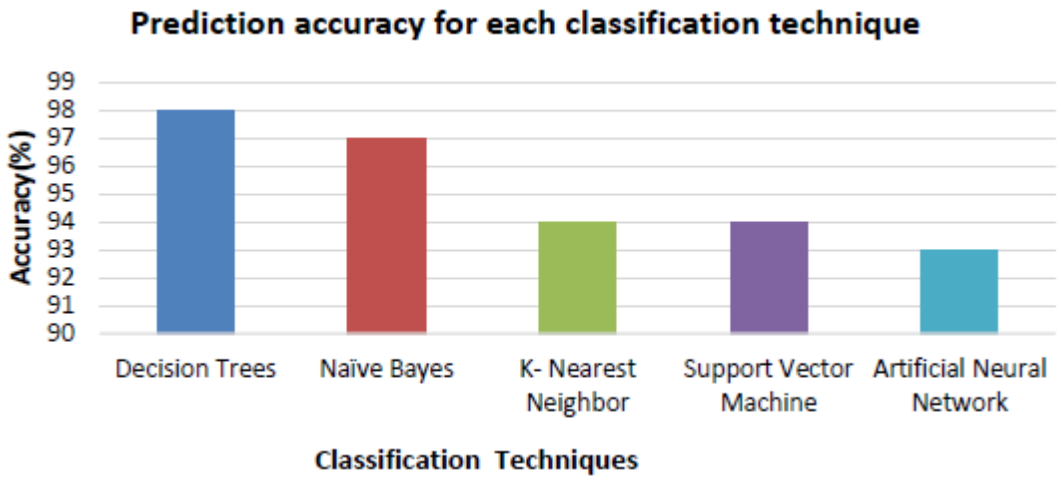

FIGURE 1. Performance prediction accuracy (2010-2020) [14]

Figure 1 was created by analyzing the most generally used methods for predicting student performance in prior research and comparing their results in order to determine which approach had the highest accuracy, and the results were plotted on the graph in figure 1. The overall accuracy of the conventional prediction methods used to estimate student

DOI: 10.17148/IJARCCE.2022.11242

performance that the authors investigated in this study, all grouped by their algorithms from 2012 to 2020, is shown in the Figure. In comparison to all other methods, the decision trees method has a high prediction accuracy of (98%) as shown in Figure 1. The second most accurate method was naive bayes (97 percent). Then came the k-nearest neighbor and support vector machine methods, both of which had a similar level of accuracy (94 percent). The artificial neural network has the lowest accuracy in predicting student achievement (93 percent) [14].

CONCLUSION

Educational data mining is a relatively young topic that holds a lot of promise for all parties engaged in the educational process. Data mining techniques were developed to automatically detect hidden knowledge and patterns in data. Educational data mining can be used to classify and predict student performance, dropouts, and instructor performance. It can help teachers track academic success in order to improve the teaching process, as well as students choose courses and manage their education. In order for a university to be profitable, educational data mining can be utilized to attract, maintain, and retain students. Analyzing student data is necessary for discovering, recognizing, and comprehending if educational practices are effective. In this study, we looked at the benefits and applications of data mining techniques in a range of educational settings. The main goal of the study is to demonstrate how useful educational data mining tools can be and to urge others to use them as well.

REFERENCES

1. Dr. P. Nithya, B. Umamaheswari, A. Umadevi – "A Survey on Educational Data Mining in Field of Education" – International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5 Issue 1, January 2016.
2. Aykroyd, R.G.; Leiva, V.; Ruggeri, F. Recent developments of control charts, identification of big data sources and future trends of current research. Technol. Forecast. Soc. Chang., 144, 221–232, 2019.
3. Hooshyar, D.; Pedaste, M.; Yang, Y. Mining educational data to predict students' performance through procrastination behavior, Entropy, Volume 22, Issue 12, 2020.
4. Bakhshinategh, B.; Zaiane, O.R.; Elatia, S.; Ipperciel, D. Educational data mining applications and tasks: A survey of the last 10 years. Educ. Inf. Technol., pp. 537–553, Volume 23, 2018.
5. Del Bonifro, F.; Gabbrielli, M.; Lisanti, G.; Zingaro, S.P. Student Dropout Prediction. In Artificial Intelligence in Education; Bittencourt, I., Cukurova, M., Muldner, K., Luckin, R., Millán E., Eds.; Springer: Cham, Switzerland, 2020.
6. Lázaro, N.; Callejas, Z.; Griol, D. Predicting computer engineering students dropout in cuban higher education with pre-enrollment and early performance data. J. Technol. Sci. Educ. 2020, 10, 241–258.
7. Mduma, N.; Kalegele, K.; Machuve, D. Machine learning approach for reducing student's dropout rates. Int. J. Adv. Comput. Res., Volume 9, 156–169, 2019.
8. P.Sinha, M. Arora, N. Mishra, "Framework for a Knowledge Management Platform in Higher Education Institutions", Volume 2, Issue 4, September 2012
9. Goyal, R. Vohra, "Applications of Data Mining in Higher Education", International Journal of Computer Science Issues, Vol. 9, Issue 2, No1, March 2012.
10. Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. (2014). Computer-Assisted Reading and Discovery for Student Generated Text in Massive Open Online Courses. Journal of Learning Analytics, 2(1), 156–184.
11. A. U. Khasanah et al., "A comparative study to predict student's performance using educational data mining techniques," in IOP Conference Series: Materials Science and Engineering, vol. 215, p. 012036, IOP Publishing, 2017.
12. M. Makhtar, H. Nawang, and S. N. Wan Shamsuddin, "Analysis on students' performance using naive bayes classifier." Journal of Theoretical & Applied Information Technology, vol. 95, no. 16, 2017.
13. S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of students' academic performance using weka," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447–459, 2018.
14. N D Lynn and A W R Emanuel, Using Data Mining Techniques to Predict Students' Performance. A Review, IOP Conf. Series: Materials Science and Engineering 1096 (2021). doi:10.1088/1757-899X/1096/1/012083.

