Academic Journal of Nawroz University (AJNU), Vol.10, No.3, 2021

https://doi.org/10.25007/ajnu.v10n3a875

Predicting Student Performance Using Knowledge Discovery Database and Data Mining for Duhok Polytechnic University Students Records as a Sample

Bareen Haval, Duhok Polytechnic University, Technical Institute of Administration, Dept. of Management Information System, Iraq - Kurdistan
Karwan Jameel Abdulrahman, Duhok Polytechnic University, Technical College of Administration, Dept. of Information Technology Management, Iraq - Kurdistan
Aras Rajab Abrahim, Duhok Polytechnic University, Technical College of Administration, Dept. of Information Technology Management, Iraq - Kurdistan

ABSTRACT

This article presents the results of connecting educational data mining techniques to the academic performance of students. Three classification models (Decision Tree, Random Forest and Deep Learning) have been developed to analyze data sets and to predict the performance of students. The projected submission of the three classificatory was calculated and matched. The academic history and data of the students from the Office of the Registrar were used to train the models. Our analysis aims to evaluate the results of students using various variables including student's grade. Data from (221) students with (10) different attributes were used, the attributes were provided by the university registration office which are the basic information of each student. The results of this study are very important since they provide a better understanding of student success assessments and stress the importance of data mining in education. The main purpose of this study is to show the student successful forecast using data mining techniques to improve academic programs. The main purpose of this study is to solve the problem of student's low academic performance by successfully predicting their academic and social information by using data mining techniques to improve future academic programs. The results of this research indicate that the Decision Tree classifier overtakes the other two classifiers as it achieves a total prediction accuracy of 97%.

KEYWORDS: Mining Educational Data, Models of Classification, Prediction, Data Set, Estimation.

1. Introduction

Currently, education is not confined to the classrooms, as many students are learning by using different learning instruments like the online education system, Massive Open Online Courses (MOOC), Intelligent tutorial, Web-based education structure, Project-based Learning, seminars, workshops, etc. Educational Data Mining (EDM), the latest trend in data mining and Knowledge Discovery in Database (KDD) finds essential knowledge from education information systems, such as admission structures, registrations, curriculum control systems (moodle and blackboard, etc.), and any new systems that deal with students at different levels.

Scientists in this area seek to identify useful analysis to help enhance the learning experience in education systems and promote students to achieve their educational objectives. Data mining is the process of deep information retrieval of vast numbers of original information. It can be described as the untrivial filtration of previously unrecognized and potentially valuable data. It is used to acknowledge data that could potentially and to show it in a way that is widely identified [1]. Data mining is a process of acquiring large sets of data using algorithms and techniques demonstrated in statistics, machine learning and data base management [2]. Standard data analysis methodologies typically involve old-fashioned research and delayed, expensive, and genuinely open to interpretation data studies [3].

Data mining techniques are used to forecast the academic performance of the students and the reason of using these techniques is to have knowledge of the factors that have impact on the performance of the students wither academic or social problems and try

to solve them to have better performance by the students. The methodology adopted for this research is based on Database as well as Data Knowledge Discovery Method and Mining. Mining.

2. Related work

The literature review demonstrates that many researchers have remained concerned with EDM issues over the last few years. Data Mining Techniques (DMT) utilized in various study areas, mainly for obtaining and examining a large number of data sets. The performance of DMT in learning is defined as Educational Data Mining (EDM), the guideline embraced in higher education institutions (IHLs) to measure the progress of students [4]. The prediction is among the significant elements in EDM. It is crucial to identify and predict the academic performance and achievement of pupils, which discussed in [5].

In obtaining data, which impacts the accuracy of prediction, different components, and pointless data could be seen. Appropriate dedication strategies were used in order to maximize the relevance of features and to eliminate unnecessary repetition [6]. In [7], the aspects reported, and their positive and negative findings were acknowledged in the Mathematics Learning student, and the KDD methodology was applied using the data mining method. A combined literature assessment on the clustering algorithms and their appropriateness nor usability in the EDM background was conducted in [8]. A review was carried out on the importance of simplification using various means [9]. Methods for data mining to achieve student performance were implemented. Two phases of student achievement are highlighted [10].

In 2000, the outcomes of a study [11] were analyses about discovering weak students by taking classification models out of data, and involving these in additional courses for advanced support. The retention of students' data is also a concern Luan has discussed using clustering, neural networking, and decision-making patterns to estimate the students at risk of failure. [12][13].

[14] Also works with educational outcomes estimation, the identification of dropout students relative to population features (e.g., sex, age, marital status) and success (e.g., labeling of a specific job). [15] provides data through a virtual arithmetic learning program and utilizes a predictive model to forecast the mathematical test results based on individual skills. [16] predict students who at risk of dropout, which determine the performance factors of first-year undergraduates, which categorize students into the three groups – low risk, medium risk and high risk, the use of decision trees, random methodology for forestry, neural networks and linear discrimination analyses.

3. Classifiers Used in Data Mining

There are different methods for building a model or classifier. Random Forest, Deep Learning, and Decision Tree were being used in this study.

3.1 Random Forest

Random Forest is monitored for machine education Classification approach, regression and other tasks works by building a number of decision-making trees Coaching and the production of the class mode Specific trees classes. In contrast to the decision tree where every node is divided into the main strengths, in random forest, each node is divided into one subset with the best randomly selected predictors at the node. The strategy of random forest is very good when compared to many other algorithms of classification, including the neural network. Support for vector machine and discriminatory analysis Others and is robust to overfit. It is a structured method of classification that consists of many decision trees randomly generated. The forest developed by the group of decision-makers trained by the packaging method, that's one of the neural networks. Random forest creates and combines multiple decision trees to achieve greater accuracy

and stable forecasting [17] [18].

3.2 Decision Tree

Each node of the tree implies an arrangement, and the branch node implies an option between several alternative methods. Decision trees are often established up for the sole purpose of decision-making to obtain information based on its impact on the result, and the root node is put at the top. The greatest impact is set first. Each module is recursively divided until a leaf node is approached. The overall result is a tree-like system that decides on a condition at all levels, and the choice of the previous level decides on the next course. This approach has been used in many studies. Since this approach uses a tree such as a structural system wherein the internal nodes represent the test performed on a specific attribute, and the end/leaf node contains a correct or incorrect response to the test given input from either the testing procedure and not an autonomous document [19].

3.3 Deep Learning in Neural Network

Another common method in educational data mining is deep learning in the neural network. It can perceive likely interfaces between variables of predictors. Even in complicated ways, deep learning cannot be identified in the relationship between the variables (dependent/independent). Therefore, it is one of the best prediction methods in the neural network method [20]. Another involvement of neural networks is learning, and preparation are used to get the best achievable reliability of data inputs user input was obtained from the input nodes, and the output was sent to the user's system from the output layer. The input and output layers are concealed in the midpoint. Unseen detecting system never collaborate directly with the user only with neurons. The process is monitored by the evaluation of patterns and results. Therefore, the representation of knowledge preparation is guaranteed, [21].

4. METHODOLOGY

The technique planned for describing the students' academic performance in this research paper is consistent with the representation of data mining knowledge and databases. Documenting or analyzing knowledge usually consists of the following stages:

- Data set cleaning: it is the removal of noise and conflicting information.
- Data integration: Multiple data sets integrated into this procedure.
- Data selection: at this stage, the data relevant to our required job assignment managed to recover.
- Transformation of data: Information transformed into systems that are adequate for information retrieval.
- Data mining: Different original data patterns retrieved to be useful in this phase.

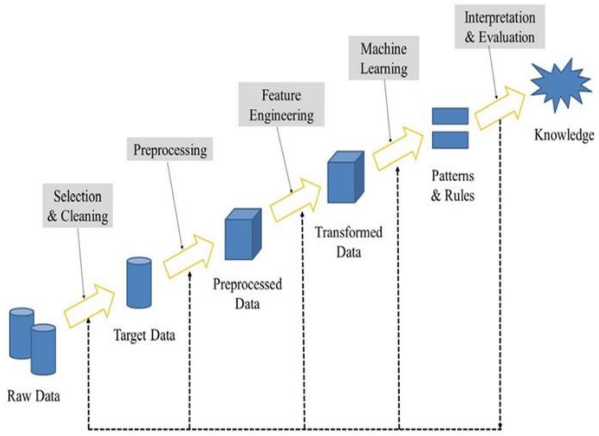

Figure 1. Knowledge discovery in databases

The primary steps in this method are:

4.1 Data Gathering

This study used data of students of Technical College of Administration and Technical Institute of Administration at Duhok Polytechnic University. It focuses on 221 students in the four academic levels (1, 2, 3 and 4). The data gathered are from the college's and the institute's registration office.

The data of 221 students were analyzed by focusing on 10 attributes related to their academic achievement and to their backgrounds.

Table. 1: Data Attributes in the Research

no	attributes	Possible values

1	St.ID	Digits
2	Dept.	Alphabetic
3	Year	1, 2, 3, 4
4	Address	Alphabetic
5	Age	Digits
6	Gender	1: Male or 0: Female
7	Civil Status	1: Single or 2: Married
8	Economic Status	1: Work with study or 0: Don't work with study
9	Grade	Digits
10	Remarks	Excellent, Very Good, Good, Fair, Pass, and Weak

4.2 Data Pre-processing:

To link data mining algorithms, sets of data were organized. Large-scale pre-processing strategies such as data cleaning and data conversion were implemented.

The process includes the following steps:

- **Cleaning Data:** fill in null values, noisy smooth data, identify or extract information, and correct inconsistencies.
- **Data reduction:** numerous datasets, databases or documents used.
- **Transformation of data:** normalization and accumulation.

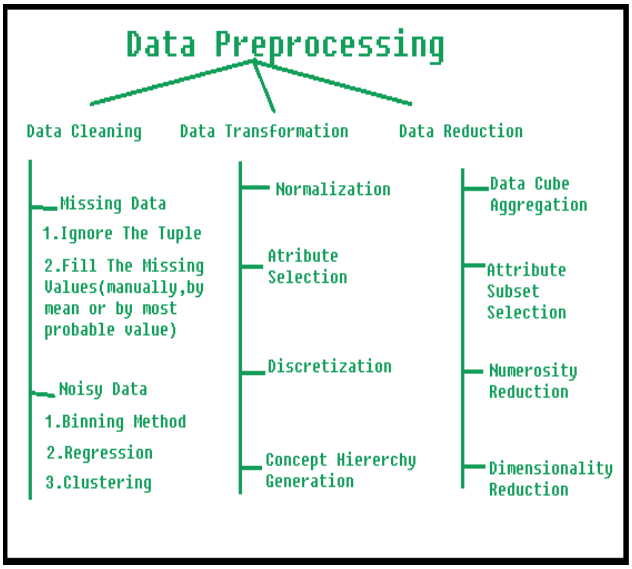

Figure 2. Data pre-processing procedure

In this paper, Rapid miner was used for data cleaning and normalizing student's data, as shown below:

Step 1:

Opening the Data Set in (Turbo Prep) Case by pressing (Turbo Prep) command.

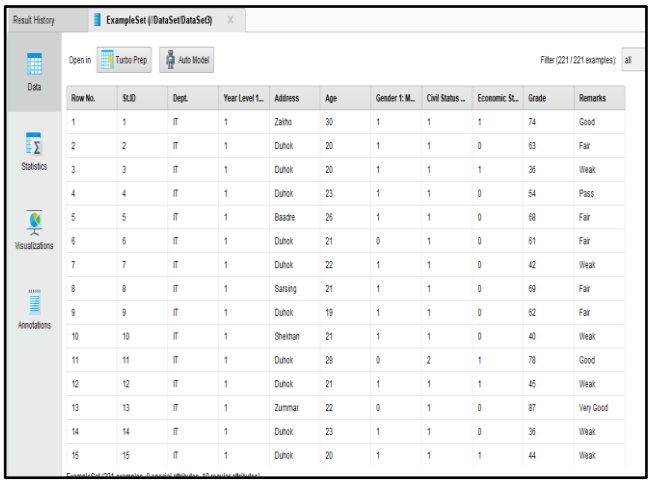

Figure 3. Opening Data Set in Turbo Prep State

Step 2:

Selecting all columns of attributes of the data set, then clicking (CLEANSE) option.

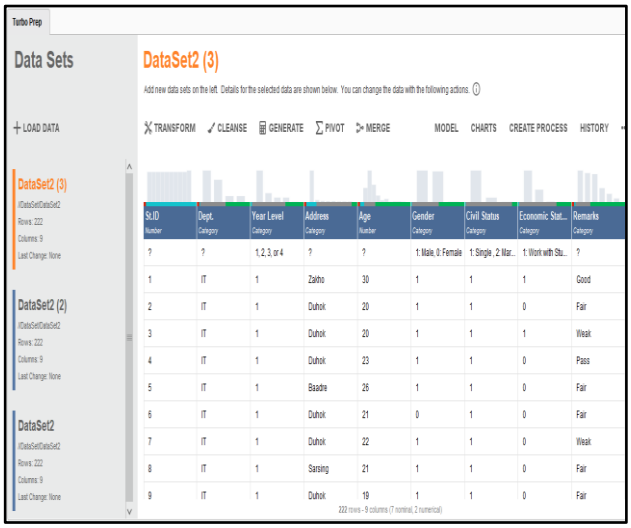

Figure 4. CLEANSE Window

Step 3:

Clicking (AUTO CLEANSING) option to perform data cleaning.

Figure 5. AUTO CLEANSING Window

Step 4:

Selecting a column which is dependent on the data set, then clicking (Next) option.

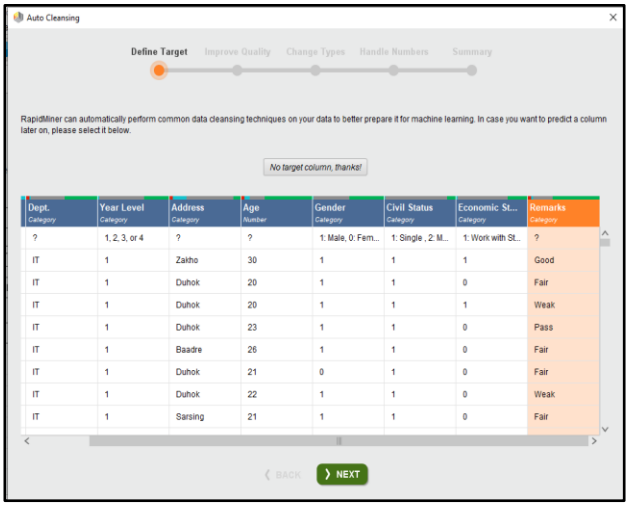

Figure 6. Defining target window

Step 5:

Selecting all options that you want to do such as (Perform PCA, and Perform Normalization), then clicking (Next) option.

Figure 7. Handling numbers step window

Step 6:

The result of data cleaning appear as follows:

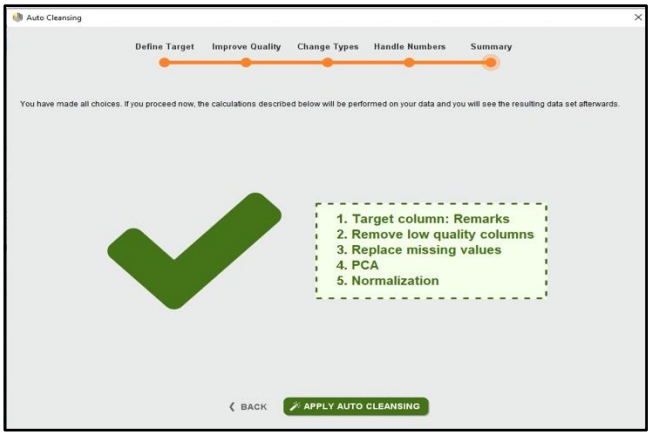

Figure 8. Summary step window

3. Model Development

The three strategies that perfectly matched to our crucial objective were selected, namely, Random Forest, Deep Learning and Decision-making to triangulate the findings. The result of each technique will be assessed to detect accuracy and error rate.

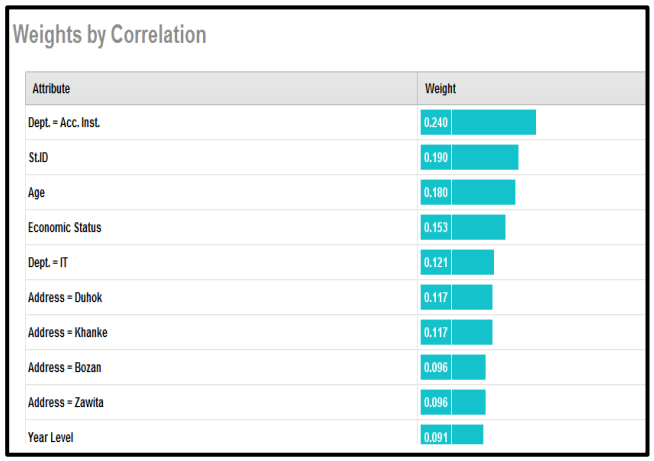

Figure 9. Weights by correlation

5. RESULTS AND DISCUSSION

The findings vary with different data mining quantities by using RapidMiner as a Data Mining Technique and implementing classification methods on the data set.

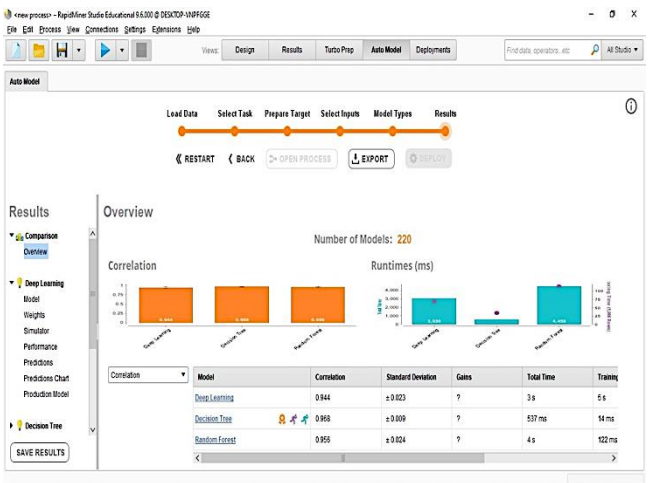

Figure 10. Correlation of models

Figure 10 shows the correlation within the data set. The weight by correlation provider computes the weight of attributes utilizing correlation with regard to the label attribute. The greater the weight of an attribute, the more accurate it is. Polynominal attributes cannot be implemented because the polynomial classes do not deliver information regarding their ordering. Thus, the weights are more or less alleged according to the internal numeric

values of the classes. Binominal labels operate as numerical items due to their recognition as 0 and 1. Finally, decision tree seems to have the highest precision of (97 %) between classifiers. Table 2 compares the results of each classifier.

Table2. Comparative results of classifiers about correlation

Model	Correlation	Standard Deviation	Total Time	Training Time (1,000 Rows)	Scoring Time (1,000 Rows)
Random forest	96%	2%	4839.0	117.6	90.9
Deep learning	94%	2%	2580.0	4511.3	113.6
Decision tree	97%	1%	488.0	13.6	34.1

Figure. 11 shows the relative error of the data set in each classifier. Decision tree has the least significant relative error (5.8%).

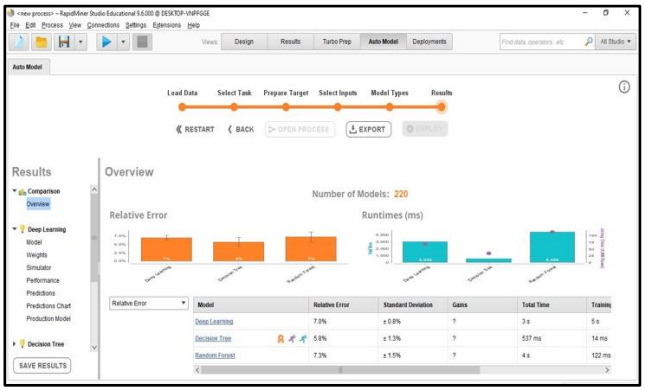

Figure 11. Relative error results of models

Table3. Comparative results of Classifiers about relative errors

Model	Relative Error	Standard Deviation	Total Time	Training Time (1,000 Rows)
Deep learning	7.0%	0.8%	2580.0	4511.3
Decision tree	5.8%	1.3%	488.0	13.6
Random forest	7.3%	1.5%	4839.0	117.6

6. CONCLUSION

This research aims at predicting and evaluating the academic performance of students. Three techniques have been applied, namely random forest, decision tree, and deep learning. These methods were applied on the information collected from Technical College of Administration and Technical Institute of Administration at Duhok Polytechnic University. Three class labels were studied to predict the academic achievement of students. The results show that the classifier of Decision Tree beats the other two classifiers by achieving (97%) of the total projection accuracy. Random Forest has (96%) accuracy and Deep Learning has (94%) accuracy rate. Additional tests with more massive datasets plus different courses and different education levels can be carried out for further study and a program may be designed to explore all factors/attributes automatically.

7. References

1. D. J. Prajapati and J. H. Prajapati, "Handling Missing Values: Application to University Data Set," Int. J. Emerg. trends Eng. Dev., vol. 1, 2011.
2. A. Feelders, H. Daniels, and M. Holsheimer, "Methodological and practical aspects of data mining," Inf. Manag., vol. 37, no. 5, pp. 271–281, 2000.
3. L. K. Long and M. D. Troutt, "Data mining for human resource information systems," in Data mining: Opportunities and challenges, IGI Global, 2003, pp. 366–381.
4. J. Chamizo-Gonzalez, E. I. Cano-Montero, E. Urquia-Grande, and C. I. Muñoz-Colomina, "Educational data mining for improving learning outcomes in teaching accounting within higher education," Int. J. Inf. Learn. Technol., 2015.
5. A. Mueen, B. Zafar, and U. Manzoor, "Modeling and predicting students' academic performance using data mining techniques," Int. J. Mod. Educ. Comput. Sci., vol. 8, no. 11, p. 36, 2016.
6. W. Punlumjeak and N. Rachburee, "A comparative study of feature selection techniques for classify student performance," in 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE), 2015, pp. 425–429.
7. S. O. da Fonseca and A. A. Namen, "Data mining on inep databases: An initial analysis aiming to improve brazilian educational system," Educ. em Rev., vol. 32, no. 1, pp. 133–157, 2016.
8. A. Dutt, M. A. Ismail, and T. Herawan, "A systematic review on educational data mining," Ieee Access, vol. 5, pp. 15991–16005, 2017.
9. S. Slater, S. Joksimović, V. Kovanovic, R. S. Baker, and D. Gasevic, "Tools for educational data mining: A review," J. Educ. Behav. Stat., vol. 42, no. 1, pp. 85–106, 2017.
10. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, "Analyzing undergraduate students' performance using educational data mining," Comput. Educ., vol. 113, pp. 177–194, 2017.

11. Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee, "Targeting the right students using data mining," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 2000, pp. 457–464.
12. J. Luan, "Data mining applications in higher education," SPSS Exec., vol. 7, 2004.
13. D. Kabakchieva, "Predicting student performance by using data mining methods for classification," Cybern. Inf. Technol., vol. 13, no. 1, pp. 61–72, 2013.
14. S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "PREDICTING STUDENTS'PERFORMANCE IN DISTANCE LEARNING USING MACHINE LEARNING TECHNIQUES," Appl. Artif. Intell., vol. 18, no. 5, pp. 411–426, 2004.
15. Z. A. Pardos, N. T. Heffernan, B. Anderson, C. L. Heffernan, and W. P. Schools, "Using fine-grained skill models to fit student performance with Bayesian networks," Handb. Educ. data Min., vol. 417, 2010.
16. J.-F. Superby, J. P. Vandamme, and N. Meskens, "Determination of factors influencing the achievement of the first-year university students using data mining methods," in Workshop on educational data mining, 2006, vol. 32, p. 234.
17. Romero, C., Lopez, M.-I., Luna, J.-M. and Ventura, S. (2013). Predicting students´´ final performance from participation in online discussion forums. Computers & Education. 68, 458–472.
18. Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. Expert systems with applications. 33(1), 135–146
19. ygthB. Kapur, N. Ahluwalia, and R. Sathyaraj, "Comparative study on marks prediction using data mining and classification algorithms," Int. J. Adv. Res. Comput. Sci., vol. 8, no. 3, 2017.
20. A. M. Shahiri and W. Husain, "A review on predicting student's performance using data mining techniques," Procedia Comput. Sci., vol. 72, pp. 414–422, 2015.
21. S. S. Abu-Naser, I. S. Zaqout, M. Abu Ghosh, R. R. Atallah, and E. Alajrami, "Predicting student performance using artificial neural network: In the faculty of engineering and information technology," 2015.