International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

Utilizing Decision Tree Technique to Predict and Analyze Student Performance

Muskan Chakrawarti[1], Prof. Jayesh Jain[2], Prof. Saurabh Sharma[3], Prof. Vishal Paranjape[4]
[1,2,3,4]*Global Nature Care Sangathan Group of Institutions, Jabalpur (M.P), India*

Abstract: **analyses student data using data mining techniques to build a predictive model for academic success. In big relational databases, data mining technically refers to the act of identifying correlations or patterns among numerous fields. By applying analysis approaches to evaluate student and instructor performance, data mining is also used to sort educational problems. In this study, classification techniques like decision trees are used to assess student performance. The job can be processed based on a variety of characteristics to forecast how well the students will accomplish each activity. The advancement of prediction/classification approaches, which are used to evaluate a person's skill expertise based on their academic performance by the breadth of their knowledge, has been the main emphasis of this study article. Providing information on the findings and the particular demands for study improvement, such as supporting students through their learning process and making quick decisions to reduce academic risk and desertion. Finally, some suggestions and ideas for the performance'sfuture evolution are presented. Analyzing slow learners who are probably studying in unfavourable conditions helps them to develop their skills as early as possible to reach the objective.**

Keywords: **Data mining, Classification algorithms, decision trees.**

I. INTRODUCTION

The analysis phase of "knowledge discovery in databases" is data mining. Data mining is the process of analysing data from several angles and synthesising the results into meaningful knowledge. Data mining software is one of the analytical techniques used for data analysis. It enables users to analyse data from several perspectives, classify it, and describe established relationships. In recent years, there has been a growing interest in utilising data mining to investigate scientific questions within educational research, a field of study known as educational data mining [8]. An ability of student performance is essential in the education environment, which is influenced by many qualitative attributes, such as Student Identity, gender, age, Specialty, Lower class Grade, higher Class Grade, Extra knowledge or skill, Resource, Attendance, Time spent studying, Class Test Grade (Internal), Seminar Performance, Lab Work, Quiz, E-Exercise, E-Homework, Over all Semester exam Percentage.

Numerous data mining techniques, including K-nearestneighbor, decision tree, Nave Bayes, neural network, fuzzy, and genetic, are used in the educational context [13].

The rapid expansion of educational data is a significant element in the institution. The primary objective of any educational institution is to enhance the quality of education. Prediction of student performance in educational institutions is one means of achieving a high standard of education. The staff of an educational institution should identify students who are likely to fail tests. Due to the vast number of registered students, it is difficult to recognise them early. Many elements, such as personal, social, and demographic information, might affect a student's academic achievement. It is difficult to extract meaningful information from a vast database [13]. EDM refers to educational data mining, or data mining in education. Educational data mining (EDM) is a study field that combines data mining, machine learning, and education.

The use of learning and statistics to data created by educational institutions, such as universities and intelligent tutoring systems. In addition, these methods are incapable of uncovering useful concealed information. Online Exam is being introduced since a demand exists.

Utilizing classification algorithms for education data modelling. The number of applications has increased during the past six years [14]. Researchers prefer to adopt a single technique in their studies on student ratings like those listed above. In addition, these methods are incapable of uncovering useful concealed information. Exam structure is being developed that will benefit both colleges and students. With this paradigm, institutions can register and administer examinations. Students are able to take examinations and check their findings with help, while instructors can assess student performance. This approach is an attempt to eliminate the existing problems in the manual exam administration system. Examination System meets the needs of institutes for exam administration. Thus, the objective of the model is to develop a system that saves the institutes and the students time and effort. Institutes enter the exam questions they desire. These questions are presented as a test to pupils who qualify. Students' responses are subsequently assessed, and their scores are generated and kept.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

The institutes can then use this score to evaluate their performance. In this study analyses the student performance by utilising data mining technique like classification, decision tree algorithm using to create the classifier model on base on dataset consisted of replies of students to courses evaluation questions.

II. Related Work

Mustafa Agaoglu [1] study in educational mining focuses on modelling student's performance instead of instructors' performance. One of the common ways to evaluate instructors' work is the course evaluation questionnaire to evaluate based on students' impression. In this study, classifier models are constructed using four distinct classification techniques: decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis. Their performances are compared across a dataset comprising of replies of students to a genuine course evaluation questionnaire utilising accuracy, precision, recall, and specificity performance criteria. Although all the classifier models show comparably high classification capabilities, C5.0 classifier is the best with respect to accuracy, precision, and specificity. In addition, a study of the variable relevance for each classifier model is done. Accordingly, it is shown that many of the items in the course evaluation form appear to be irrelevant. Furthermore, the data demonstrates that the instructors' success relies on the students' impression.

Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta [2] employ multiple classification strategies to construct performance prediction model based on students' social integration, academic integration, and numerous emotional skills which have not been included so far. Two algorithms J48 (Implementation of C4.5) and Random Tree have been applied to the data of MCA students of colleges affiliated to Guru Gobind Singh Indraprastha University to predict third semester performance. Random Tree is proven to be more accurate in forecasting performance than J48 algorithm.

Keno C. Piad,Menchita Dumlao,Melvin A. Ballera,Shaneth C. Ambat [3] forecasts the employability of IT graduates using nine variables. First, numerous classification algorithms in data mining were evaluated creating logistic regression with accuracy of 78.4 is implemented. Based on logistic regression analysis, three academic characteristics directly effect; IT Core, IT Professional and Gender indicated as important predictors for employability.

The results were collected based on the five year profiles of 515 students randomly selected at the placement office tracer study.

Bipin Bihari Jayasingh [4] commences a sample study that is taken for a particular institution, in the particular atmosphere, for the particular batch and particular set of students. The sample data are acquired from a classroom by giving a questionnaire with questions about inquiry-based and deductive learning to two distinct groups of students. The system is designed and tested twice after teaching the topic using inductive technique and implemented utilising attribute relevance, discriminant rules of class discrimination mining. Using bar graphs, the data demonstrate that the two cohorts of learners from different years have distinct learning characteristics.

S. M. Merchán [5] describes and evaluates the experience of using specific data mining methods and procedures to the data of 932 Systems Engineering students from El Bosque University in Bogotá, Colombia; an endeavour that has been undertaken.

Pursued in order to develop a predictive model for students' academic success. As part of an iterative discovery and learning process, the experience is examined based on the outcomes of each iteration. Each generated result is evaluated based on the predicted results, the input and output characteristics of the data, what the theory mandates, and the relevance of the obtained model in terms of prediction accuracy. Said pertinence is evaluated taking into consideration unique data regarding the population researched, and the specific demands exhibited by the institution.

Konstantina Chrysafiadi and Maria Virvou [6] provide an innovative approach to web-based education that provides personalised teaching in the field of programming languages. This strategy is fully developed and assessed in a Fuzzy Knowledge State Definer educational application module (FuzKSD). Specifically, FuzKSD provides user modelling by dynamically recognising and updating the student's knowledge level for all domain knowledge topics. Fuzzy Cognitive Maps (FCMs) are utilised to depict the dependencies between domain ideas in FuzKSD's functioning. FuzKSD use fuzzy sets to describe the knowledge level of students as a subset of domain knowledge.

Based on the study of M. Mayilvaganan and D. Kalpanadevi [7], this paper focuses on the enhancement of Prediction/Classification approaches that are used to examine the skill mastery of students based on their academic achievement and the extent of their knowledge.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

In addition, the study compares the performance of C4.5 algorithm, AODE, Nave Bayesian classifier algorithm, Multi Label K-Nearest Neighbor method, and decision tree algorithm to determine the optimal accuracy of classification algorithm and performance analysis of students using the Weka tool.

Crist'obal Romero [8] Educational data mining (EDM) is an emerging interdisciplinary field of study that focuses on the development of techniques to analyse data originating from an educational context. EDM analyses educational data with computational methods in order to investigate educational topics. This study examines the most significant studies conducted in this topic to yet. It begins by introducing EDM and describing the various user groups, educational environment kinds, and data they give. It then lists the most typical/common challenges in the educational environment that have been resolved by data-mining approaches, and it concludes with a discussion of some of the most interesting future research directions.

R. V. Mane and Priyanka Patil [9] offer the Generalized Sequential Pattern mining technique for discovering frequent patterns in student databases and the Frequent Pattern tree approach for constructing a tree based on frequent patterns. This tree can be used to predict if a student will pass or fail. Once it is determined that a student is at risk of failing, he or she can get performance enhancement advice.

III. Proposed System

We will provide a system that allows the user to administer tests on specified educational or subject categories. When a student completes a test, the system will calculate the user's performance using a decision tree algorithm. The system will advise to the instructor the topics the student needs to review or improve upon. There is a need for a computerised system to manage all examination writing tasks in order to resolve the issues associated with manual examination writing. We suggest an application that will give a flexible working environment, facilitate work, and save the time required for report preparation and other paper operations.

Today, many companies conduct online tests successfully over the world and publish results online, but they do not measure student performance and teachers are unaware of students' weak points; we are working on this issue. The primary benefit is that the examination of replies can be totally automated for all questions, while other essay-type questions can be evaluated manually or by an automated system, depending on the nature of the question and the criteria. In order to increase efficiency, transparency, and dependability, institutions should implement this new examination management technology.

Student modelling can be defined as the act of collecting pertinent information to infer the current cognitive state of the student and representing it in a way that makes it accessible and useful for evaluating their performance. The proposed student model includes two user types. The first user consists of students who can administer tests, manage their profiles, and check their performance, results, and expert recommendations, among other things. The second user is a teacher or instructor who can access all student data, their results, and each student's performance. Models aid in analysing student performance, their weaknesses, and the need for score improvement. Each time a teacher interacts with the system, a test is administered to measure the student's level of knowledge.

In the proposed model, users must first register before receiving an authorised username and password. When a student logs in, he or she can take a test based on the subcategory or category selected. After a test has been completed, the system automatically computes the result using data mining and provides a score along with recommendations for increasing the score and valuable feedback.

This record is also viewed by teachers so they may evaluate student performance and take appropriate steps to enhance it before the student enters a key area. This technique will monitor and assess the academic performance of students at various year levels prior to the final exam in order to predict their deficiencies. Teachers can play the role of administrators with the authority to add test subjects, themes, and questions. Fig. 1 depicts a flowchart that outlines the entirety of these operations.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

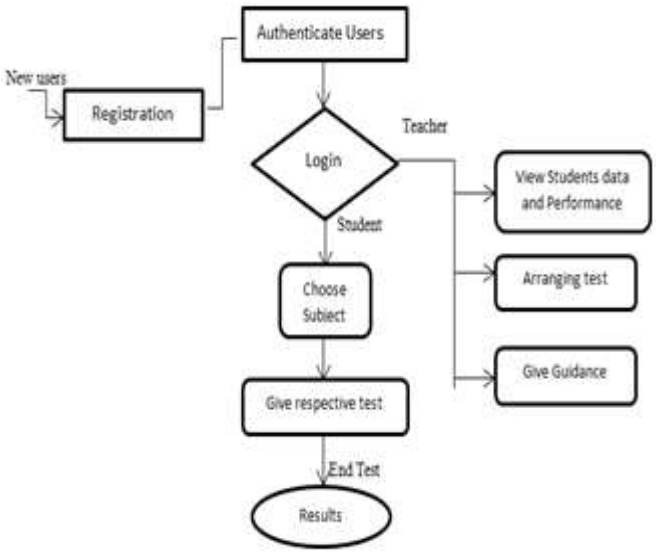

Fig. 1. Data Flow Diagram

IV. Proposed Methodology

The major objective of the proposed methodology is to build the classification model that classifies a students' performance. The classifiers, has been built by combining the Standard for Data Mining that includes student data and finally application of data mining techniques which is classification in present study. In other words, using this Decision tree algorithm, we wanted to be able to guide student towards achievement of good score that we felt they would enjoy doing.Tree-based methods classify instances by sorting the instances down the tree from the root to some leaf node, which provides the classification of a particular instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute [15].The benefits of having a decision tree are as follows –

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

Decision Tree Induction Algorithm

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner.

Decision Trees

Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. A decision tree is a flow-chart-like tree structure, where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Arcs from an internal node to its children are labelled with distinct outcomes of the test. Each leaf node has a class label associated with it.

V. Data Mining Technique And Classification

Data mining is very promising as a new effective technique for decision making processes. Through Educational data mining is an analysis of discipline to developing the methods for exploring the unique types of data from educational settings and it is used for improvement of students in better way [8] .Data mining techniques are applied in higher education more and more to give insights to educational and administrative problems in order to increase the managerial effectiveness. However, most of the educational mining research focuses on modelling student's performance.Data mining technique can give the input for the teachers and students about the student academic results. This technique can analysis the database patterns to forecast student performance, so this allows the teachers to prepare like a remedial program (needing extra help for learning) or more additional assignments for the students.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

Several DM algorithms (naıve Bayes, Bayes net, support vector machines, logistic regression, and decision trees) have been compared to detect student mental models in ITSs [16]. Unsupervised (clustering) and supervised (classification) machine learning have been proposed to reduce development costs in building user models and to facilitate transferability in intelligent learning environments. Clustering and classification of learning variables have been used to measure the online learner's motivation.

Classification is one of the most studied problems in machine learning and data mining .It consists in predicting the value of a categorical attribute (the class) based on the values of other attributes (predicting attributes). A search algorithm is used to induce a classifier from a set of correctly classified data instances called the training set. Another set of correctly classified data instances, known as the testing set, is used to measure the quality of the classifier obtained. Different kinds of models, such as decision trees or rules, can be used to represent classifiers.In Classification process, the derive model is to predict the class of objects whose class label is unknown. Generally, the classification of data has two step process are learning and a classification step which is used to predict class labels for training data. In classification step, test data are used to estimate the accuracy of classification rules. There are many techniques that can be used for classification techniques such as decision tree, Bayesian methods, Bayesian network, rule based algorithms, neural network, support vector machine, association rule mining, k-nearest- neighbor, case- based reasoning, genetic algorithms, rough sets and fuzzy logic. In this study, we focus on classification techniques such as decision tree [7].

A. Data Preparations

The data set used in this study was obtained from a student's database which we are created for our application.

In this step data stored in different tables was joined in a single table after joining process errors were removed.

B. Data selection and transformation

In this step only those fields were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables was extracted from the database.

C. Decision Tree Algorithm

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification. At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists.

All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

Algorithm: Generate_decision_tree

Step 1- Start
Step 2-Take input which is given by User
I_n= {I_1, ...,I_n}
Step 3-Dataset preparation
D_n = { {I_1,…I_n}D}
Step 4-Dataset elaboration
D_I ={S_1,….S_n, C_1,…..C_n, I_1,……I_n, a_1, ... a_n}
Step 5- Processing
While(D_n!=0)
{
If (a_n==I_n)
Check C_n, S_n;
}
Step 6- Result Generation

R={ S_c, S_n, C_n};
Where,
I_n = Input given by users
D_n = Dataset
D = Database
D_I = Dataset contents
S_c=Score
a_1,….a_n= Answer
S_1,….S_n= Subject
C_1,….C_n= Category

We also using Generalized Sequential Pattern mining algorithmfor predicting the student's performance as pass or fail. Once the student is found at the risk of failure he/she can be provided guidance for performance improvement.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

Generation Sequential pattern Mining Algorithm

```
Step 1-Start
Step 2- Take input from Dataset
Step 3- Processing
 While ( Cn!=0 )
    {
       Qn={{Q1,....Qn} Cn}
PData= {count(Qn), So(Qn), Cr(Qn)}
 Rc= PData{( count(Qn)-Cr(Qn)) Cn}
    }
Step 4- Result Generation
         R=Rc;
 R will show the weak category of student.
Where,
 Qn=Questions(Total)
 {(Qn)Cn}= Questions regarding to category.
 So= Solved Questions
 Cr= Correct questions
Rc= Result Category
```

It provides weak categories or subjects of students from his/her performance.

VI. Simulation And Results

The simulation studies involve comparison of ID3 and C4.5 accuracy with different data set size, this comparison is presented graphically in Fig.2.

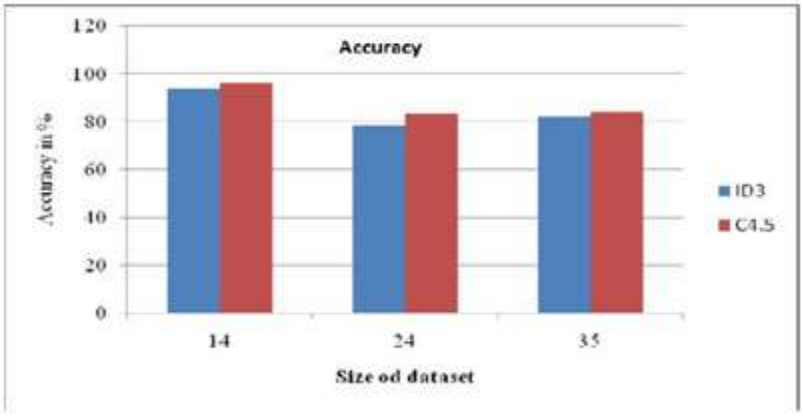

Fig.2. Comparison of Accuracy for ID3 & C4.5 Algorithm

The 2nd parameter compared between ID3 and C4.5 is the execution time which is show in Fig.3.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

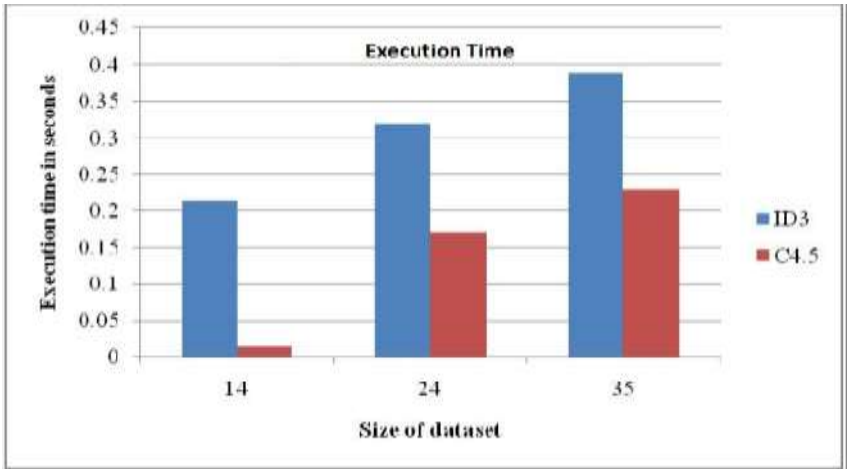

Fig.3. Comparison of Execution Time for ID3 & C4.5 Algorithm

In result student getting the all records of given test is show in Fig. 4. Figure shows the test given by student marks obtains, total marks, date oaf test and provide suggestion to study for improving performance.

Fig.4. Student Record

In Fig.5 shows the student performance record of test given by that particular student which involve subject, total marks, marks obtain, date of conducting test, weak concept shoes the category in which student is weak.

Fig.5. Student performancereport

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 11, Issue 08, August 2022)

In Fig.6 indicate performance of student in graphical to teacher.

Fig.6. Performance result

VII. CONCLUSION

Academic success of students of any professional Institution has become the major issue for the management. An early analysis of students at risk of poor performance helps the management take timely action to improve their performance through extra coaching and counselling. The result of this study indicates that data mining techniques capabilities provided effective improving tools for analysis student performance. In this paper, data mining is utilized to analyse course evaluation questionnaires. Here, the most important variables that separate "satisfactory" and "not satisfactory" student performances and there weakness' in particular subject or field. Hopefully, these can help instructors to improve their performances. Tree-based methods classify instances by sorting the instances down the tree from the root to some leaf node, which provides the classification of a particular instance. Each node in the tree specifies a test of some attribute of the instance and each branch descending from that node corresponds to one of the possible values for this attribute. This paper focuses on analysis student academic performance by using advantage of data mining techniques model.

REFERENCES

[1] Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access , Volume: 4 ,2016.

[2] Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta,"Mining Students' Data for Performance Prediction," in fourth International Conference on Advanced Computing & Communication Technologies,2014.

[3] Keno C. Piad, Menchita Dumlao, Melvin A. Ballera, Shaneth C. Ambat," Predicting IT Employability Using Data Mining Techniques," in third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016.

[4] Bipin Bihari Jayasingh,"A Data Mining Approach to Inquiry Based Inductive Learning Practice In Engineering Education," in IEEE 6th International Conference on Advanced Computing,2016.

[5] S. M. Merchán,"Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic,"IEEE Latin America Transactions, vol. 14, no. 6, June 2016.

[6] Konstantina Chrysafiadi and Maria Virvou," Fuzzy Logic for adaptive instruction in an e-learning environment for computer programming," IEEE Transactions on Fuzzy Systems ,Volume: 23, Issue: 1, Feb. 2015.

[7] M. Mayilvaganan,D. Kalpanadevi ," Comparison of Classification Techniques for predicting the performance of Students Academic Environment," in International Conference on Communication and Network Technologies (ICCNT), 2014.

[8] Crist´obal Romero," Educational Data Mining: A Review of the State of the Art," IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 40, No. 6, November 2010.

[9] Priyanka Anandrao Patil, R. V. Mane," Prediction of Students Performance Using Frequent Pattern Tree," Sixth International Conference on Computational Intelligence and Communication Networks,2014.

[10] Behrouz Minaei-Bidgoli, Deborah A. Kashy , Gerd Kortemeyer, William F. Punch, "Predicting Student Performance: An Application Of Data Mining Methods With An Educational Web-Based System," in 33'd ASEE/IEEE Frontiers in Education Conference, T2A-13,November 5- 4,2003.

[11] Peter Brusilovsky, Sibel Somyürek , Julio Guerra , Roya Hosseini , Vladimir Zadorozhny , Paula J. Durlach,"Open Social Student Modeling for Personalized Learning," IEEE Transactions on Emerging Topics in Computing, Volume: 4, Issue: 3, July-Sept. 2016.

[12] Pedro G. Espejo, Sebasti´an Ventura, and Francisco Herrera," A Survey on the Application of Genetic Programming to Classification," IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 40, No. 2, March 2010.

[13] Carlos Márquez Vera, Cristóbal Romero Morales and Sebastián Ventura Soto,"Predicting of school failure and dropout by using data mining techniques", The IEEE Journal of Latin-American Learning Technologies (IEEE-RITA) , Vol. 8, No. 1, pp 7-14, Feb 2013.

[14] A. Peña-Ayala, "Review: Educational data mining: A survey and a data mining-based analysis of recent works," Expert Systems with Applications, vol. 41, no. 4, pp. 1432-1462, 2014.

[15] R.S.J.D Baker and K.Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions" , Journal of Educational Data Mining, 1, Vol 1, No 1, 2009.

[16] V. Rus, M. Lintean, and R. Azevedo, "Automatic detection of student mental models during prior knowledge activation in MetaTutor," in Proc. Int. Conf. Educ. Data Mining, Cordoba, Spain, 2009, pp. 161–170.