International Journal of Artificial Intelligence
p-ISSN: 2407-7275, e-ISSN: 2686-3251

Original Research Paper

PSAP: Improving Accuracy of Students' Final Grade Prediction using ID3 and C4.5

Ismail Yusuf Panessai[1*], Muhammad Modi Lakulu[1], Mohd Hishamuddin bin Abdul Rahman[1], Noor Anida Zaria binti Mohd Noor[1], Nor Syazwani binti Mat Salleh[1], Aldrin Aran Bilong[1].

[1] *Faculty of Faculty of Art, Computing and Creative Industry, Universiti Pendidikan Sultan Idris, Malaysia*

Article History
Received:
17.10.2019

Revised:
18.11.2019

Accepted:
30.11.2019

***Corresponding Author:**
Ismail Yusuf Panessai
Email:
ismailyusuf@fskik.upsi.edu.my

PSAP: Improving Accuracy of Students' Final Grade Prediction using ID3 and C4.5

Abstract: This study was aimed to increase the performance of the Predicting Student Academic Performance (PSAP) system, and the outcome is to develop a web application that can be used to analyze student performance during present semester. Development of the web-based application was based on the evolutionary prototyping model. The study also analyses the accuracy of the classifier that is constructed for the prediction features in the web application. Qualitative approaches by user evaluation questionnaire were used for this study. A number of few personnel expert users which are lecturers from Universiti Pendidikan Sultan Idris were chosen as respondents. Each respondent is instructed to answer a total of 27 questions regarding respondent's background and web application design. The accuracy of the classifier for the prediction features is tested by using the confusion matrix by using the test set of 24 rows. The findings showed the views of respondents on the aspects of interface design, functionality, navigation, and reliability of the web-based application that is developed. The result also showed that accuracy for the classifier constructed by using ID3 classification model (C4.5) is 79.18% and the highest compared to Naïve Bayes and Generalized Linear classification model.

Keyword: Predicting, Student Academic Performance, Educational Data Mining; C4.5, ID3.

DOI: 10.36079/lamintang.ijai-0602.42

1. Introduction

This Nowadays, it is very common there would be one or two students will fail on a course. Failing to achieve good grade may result of student to extend the semester or dropping out university. Increase in rate of the student dropout may reflect badly at a university's reputation. According to Cherif et al. [1], the causes of the student failing can be categorized in a major area such as student-related factors, life, and socioeconomic issues and failure in the educational system.

Another cause of student failure is that lecturer is unable to keep up with all students that are taking the course. This is because there is a lot of students is taking the course and lecturers often don't pay much attention to on each individual performance in the class. Besides, the lecturer does not have the appropriate tools to predict students' academic performance and unable detect an early sign of student is going to fail the course. There have been various student retention programs that are being developed as efforts to reduce the rates of university dropout. In UPSI, there was a program called "Program Jejak Cemerlang" held by Faculty of Art, Computing and Creative Industry involving students that acquire less than 3.0 CGPS. This is one of many the attempts by the university to improve the student's motivation in achieving better CGPS for the upcoming semester. Besides that, another strategy in the attempt to detect potential student dropout is to develop a web-based application that would able to analyze student academic performance, predict their mark in the upcoming final exam and generate a report containing a list of possible candidates that have a high chance of failing the course. This early prediction allows lecturers to choose on which student they should pay more attention and be well prepared during the lecture.

In our education system, assessments are methods or tools used by educators to assess, measure, and document the academic readiness, learning progress, skill acquisition or student education needs. This includes mid-term exam, group project, final exam, quizzes and other assessments that can be utilized to measure the academic performance of the student. These assessments also known as coursework will be recorded by lecturer and uploaded to the university's database. These data will be kept and will be reviewed to determine whether the students have all the requirements to graduate. The coursework will also determine the student's grade on the course. With the current technology of data mining, we can utilize the student's coursework data to predict the student's grade. This can be made possible by using the data mining process by compiling large data to identify patterns and establish relationships to solve problem through data analysis. Data mining tools allow us to predict future trends. We also can identify the patterns that eventually will lead students to fail a course.

Therefore, in this project, we will develop a system that allows lecturer to monitor their students' academic performance as well as to predict students' grade for the course before the student is undertaking the final exam. Besides that, the system will allow lecturer to manage the coursework information that is constructed for the course. In addition, lecturers will able to import students' coursework mark in order to construct the model for the prediction process.

2. Related Studies

The data mining has been applied to many fields, which include the educational field. Nghe et al. [2], Amirah et al [3] have provided an excellent review on how other studies have been applying data mining in the educational field especially on predicting student academic performance.

Osmanbegović and Suljic [4] have differentiate three supervised data mining algorithms that are used on preliminary estimation data to predict success in the course (pass or failure) and the performance of learning methods measured by their predicted accuracy, learning facilities and user-friendly features. The results show that Naïve Bayes producers outperform predictive performance and neural network methods. It has also been shown that a good classifier model should be accurate and understandable by the professor. Adhatrao et al. [5] developed a system that can predict student performance from the past result by using the concept of data mining techniques under Classification. They have analyzed data sets containing information about students, such as gender, X and XII grade test scores, scores and ranks in entrance examinations and results in the first year of the previous student group. Using the ID3 (Dichotomizer 3) and C4.5 algorithms for this data, they have foreseen the general achievement of new students and students who are accepted into future exams.

Nagy et al. [6] show how data mining can be used in education by developing a "Student Advisory Framework" that uses classifications and clusters to build smart systems. This system can be used to provide a variety of consultations to first-year university students to pursue a particular educational track in which they may succeed, aimed at reducing the high academic failures among these students.

DOI: 10.36079/lamintang.ijai-0602.42

Chen et al. [7] recommend a model that predicts student performance, based on the results of standardized examinations, including university entrance examinations, high school graduation examinations, and other influential factors. In this study, the approach to problems based on artificial neural networks (ANNs) with two nail-inspired meta-heuristic algorithms, Cuckoo Search (CS) and Cuckoo Optimization Algorithms (COA) is recommended. Specifically, they use previous exam results and other factors, such as the location of the secondary school students and the gender of the students as input variables, and predict student achievement.

Natek and Zwilling [8] in their study, ask whether it is possible to predict the success rate of students enrolling in their course? Are there any specific student characteristics, which can be attributed to the student success rate? Are the relevant student data for institutions of higher learning (HEI) where they can predict the success rate of the student? Unfortunately, data mining algorithms work well with large data sets, while student data, available for IPT, related courses are limited and included in the small dataset category. Their research results support the conclusion that data mining is generally not limited to large data sets as a majority but not all previous authors and investigations expect this discovery. The use of specialized data mining techniques for structured small data sets can also meet the usable results. Smaller data mining data from research is an example of the use of relevant data mining technology to develop HEI's knowledge management systems in the education domain.

A prediction model for student academic performance, for undergraduate and postgraduate students in the field of Computer and Electronics and Communication using the two selected classification methods developed by Hamsa et al. [9] by using Decision Tree and Fuzzy Genetic Algorithm. Parameters such as internal marks, session marks and entry scores are selected to do this work. Internal value is a combination of attendance value, the average value obtained from two session exams, and assignment scores. The entry value for the student level is a weighted score obtained from the test scores 10 and 12 and the entry value. In the case of a master's degree student, that includes the degree exam marks and entry marks. The resulting prediction model can be used to identify student performance for each subject. A systematic approach can be taken to improve performance over time.

Another research by Sweeney et al. [10] applied for data mining in the field of education. In their experiments, the Factorization Machine (FM), Random Forest (RF), and the Personalization Multi-Linear Regression model produced the lowest prediction error. By comparing feature interests across populations and across models, they reveal strong relationships between instructor characteristics and student performance. Research conduct by Sweeney et al. [10] found the main differences between transfer and non-transfer students. In the end, they found that the hybrid FM-RF method can be used to accurately predict values for new students and who re-take new and existing courses. The application of these techniques promises student-level planning, instructor intervention, and personal advice, all of which can improve retention and academic performance.

Adejo and Connolly [11] presented research on the use of basic and aggregate classifiers in predicting students' academic performance using UWS student data. Furthermore, the important variables needed for the prediction were identified using PCA. They showed that there is a correlation between the identified variables and that the improvements in the prediction model can be obtained using the generalization of the stack (stacking set) with the resulting effect of the increase in accuracy, the error rate reduced and the predictive value. A very precise model suggested by Chiheb et al. [12] propose a system whereby undergraduate students and postgraduate students can be classified according to their decisions and their performance can be predicted for years to come based on current results and their historical data. The system can also be used as an early warning tool for high school students and helps graduates to choose the appropriate master discipline for their studies. Livieris et al. [13] proposed new decision support software to predict students' performance in the final exams. The proposed software is based on a novel 2-step classification technique that performs better than any single learning algorithm studied. The main advantages of the presented tool are the simple and user-friendly interface and the possibility to be implemented on any platform under any operating system.

3. Methodology

This research contains three phases which are analysis phase, development phase, and evaluation phase. In the analysis phase, researcher doing observations of the early research and analyse the user

DOI: 10.36079/lamintang.ijai-0602.42

target. The Second phase is development phase where RDB (requirement-design-build) model will be use in this research. The last phase is evaluation phase.

Evaluations are consist of two, i.e. evaluation to know how the performance of the system and to find out if the predicting can be used to increase the understanding level of students academic performance.

3.1. Dataset Used for Prediction

Transcripts data for students who are undertaking the Software Engineering Program at Universiti Pendidikan Sultan Idris in the year 2015 were collected from the university's database management system and the totaling 123 students. The collected data was organized in a Microsoft Excel sheet. Each of the student records had the following attributes: Matric No, pre-University academic qualification type, (Cumulative Grade Point Average) CGPA, and the coursework mark excluding the final exam mark and finally the students' grade for the course. Table 1 shows the list of attributes of the data set.

Table 1. List of Features in This Study

Attributes	Type	Values
Student ID	Alphanumeric strings	
pre-University academic qualification type	Varchar	
CGPA	Floating-point	[0.00-4.00]
Total of coursework mark	Floating-point	[0.00-100.00]
Final grade	Varchar	[A, B, C, D, E, F]

3.2. RDBR Model

The Requirement, Design, Build and Refine (RDBR) model is best applied when detailed information related to input and output requirements of the system is not available. Normally, the RDBR model is used when a similar system does not yet exist or in case of a large and complex system where there is no manual process to determine the requirements. Since that the input output requirements of the PSAP application is not very clear and the application have very limited number of similar systems that have been developed, the most appropriate software development model that have been decided for the development of the PSAP web application is the RDBR model.

Besides, RDBR model has benefit when used in developing a software project. Firstly, it increases user involvement in the software development which will lead the user to better understand what the users really want as final product. Other than that, defect can be detected on earlier phase which can reduces time taken and cost in order to complete the project. Any missing, confusing and difficult functionality would be identified and these functionalities can be modified, added or removed in order to meet the user requirement.

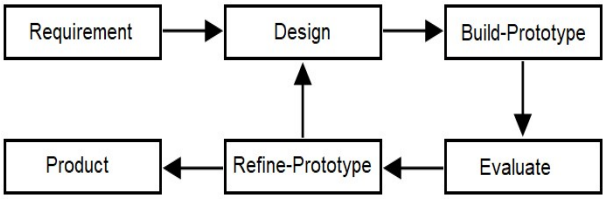

Figure 1. RDBR Model

3.2.1. Requirement

In requirement phase, the developer will identify the basic requirement of the software and give more accentuation on the user interface particularly. The other detail of requirements which are least important can be ignored in requirement phase such as system performance, security and other

DOI: 10.36079/lamintang.ijai-0602.42

complicated detail. The aim of early construction of a prototype is to demonstrate the actual feeling of when the user is interacting with the web application.

3.2.2. Design

After the basic requirement is collected and analysed, a preliminary design of the prototype is generated as guideline in the construction of the prototype. The design is not very detail and would not be the final design. The design includes only the important aspect of the system which gives an idea of how the system operates to the user.

3.2.3. Build-Prototype

In build-Prototype phase, the first prototype of the system is constructed based on the preliminary design. Normally, the prototype built is a scaled-down system and demonstrates the resemblance of the functionality of the final product.

3.2.4. Evaluate

In the evaluate phase, the user will be presented with the prototype. The user will evaluate the prototype and identify the strength and the weakness, what should be added and removed. The feedbacks are collected and analysed. In the evaluation phase, any missing requirements that are not yet noticed will be identified.

3.2.5. Refine-Prototype

The feedbacks from the user are discussed at refining prototype phase and if the users are yet not satisfied with the current prototype, the requirement of the prototype will be refined based on the feedbacks. Thus, the phase will repeat the process of building a new prototype. The developer will do a quick design of the new prototype based on the refined requirements according to the feedback from the user evaluation.

Then, a new prototype is constructed based on the quick design. The new prototype will be represented to the user for evaluation and user feedback will be collected and analysed. The prototype will be refined until the prototype meets the user expectations. Whenever the user satisfied with the final prototype, then the stage will be proceeding to engineer the product.

3.2.6. Product

Once the user expectations are completely met, the final system will be constructed based on the final prototype. The final system is tested and evaluated thoroughly followed by the routine maintenance on regular basis for preventing large-scale failures and minimizing downtime.

4. PSAP Web-Based Application

For the purpose of this research, we developed a user-friendly web-based application, which is called PSAP (Predicting Student Academic Performance) that used for predicting student performance at the final examinations based on their CGPA on the pre-university academic qualification and total of coursework's mark.

The following are the main features of the PSAP web-based application.

Home Page : This module is optionally used by user to manage the course and coursework structures.

Prediction : This module is used to accept student's information and then show the prediction result.

Data Training : This module is used to import the data set of previous student's academic performance information including their coursework mark.

Firstly, the user/educator can use the uploaded data in the database of the web or they can upload their own data collected from their past courses in CSV file format. This function can be accessed on the "Data Training" section. The user can download the sample file as a reference of what the CSV file should contain.

Next, by scrolling down the pages, PSAP will display the decision tree constructed based on the data that are in the database of the web application. Last, user will be able to predict their student

DOI: 10.36079/lamintang.ijai-0602.42

performance by navigating to the "Prediction" section. User will have to select a course as a base for prediction. The course and coursework information can be edited in home page.

Figure 2 – Figure 6 shows the PSAP's user interface. In Figure 5, the user then have to enter their student's academic information in order to view the prediction result in Figure 6.

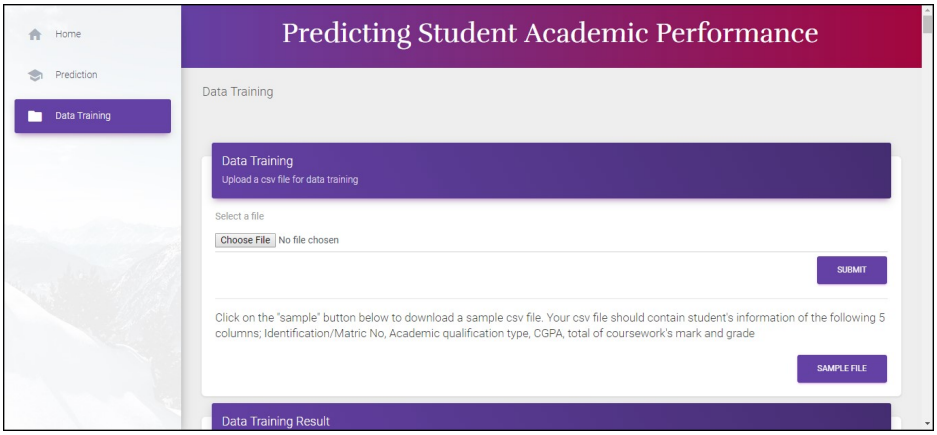

Figure 2. Data Training Interface

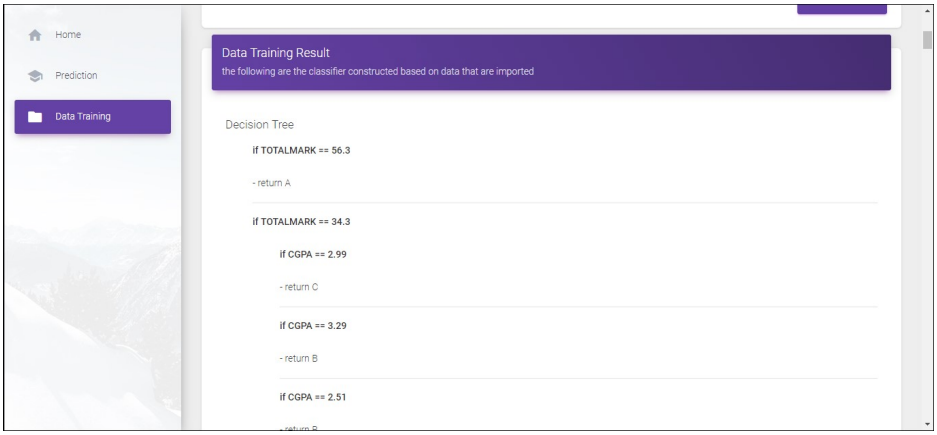

Figure 3. PSAP: Data Training Result

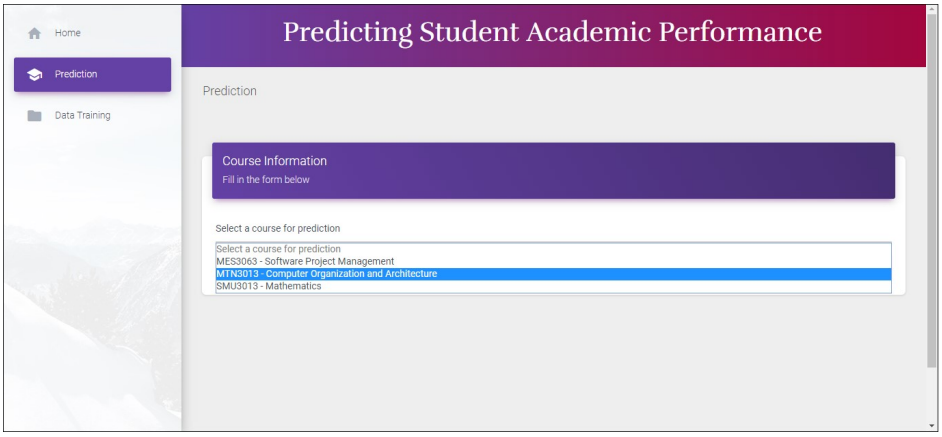

Figure 4. PSAP: Select a Course

DOI: 10.36079/lamintang.ijai-0602.42

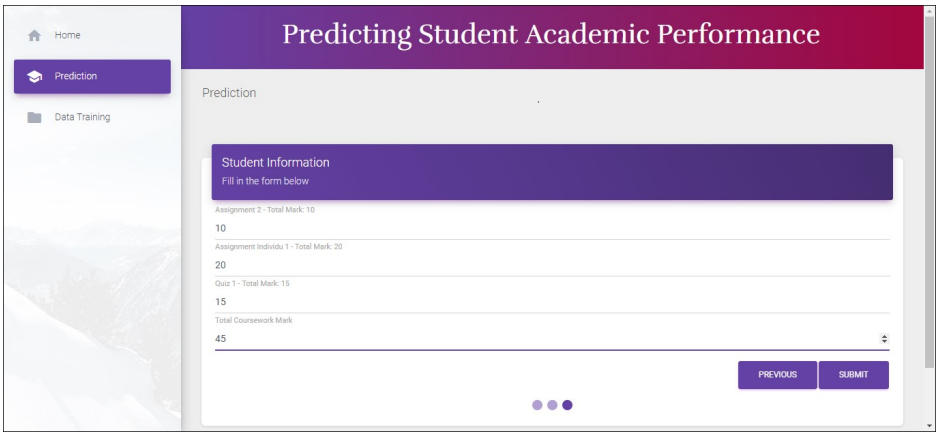

Figure 5. PSAP: Student Coursework Mark Form

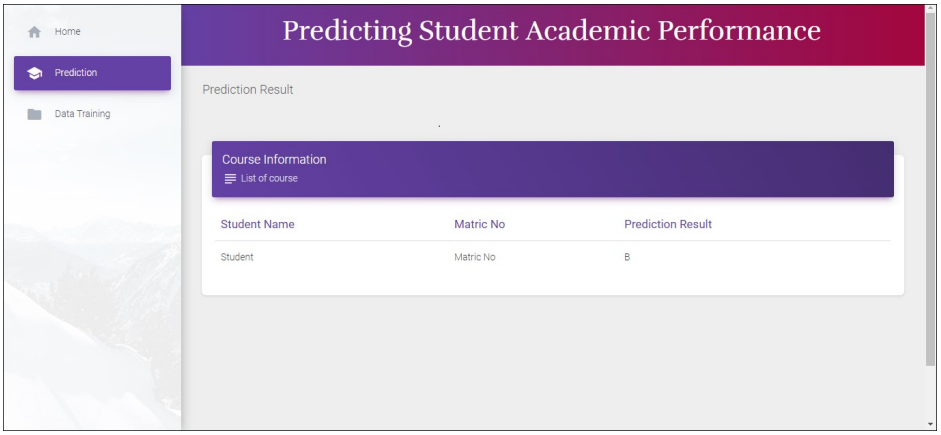

Figure 6. PSAP: Prediction Result

5. Analysis

There are three classification model applied to the data set. Each model is evaluated by using confusion matrix and the result is shown in Figure 7. Figure 7 shown that C4.5 is highest accuracy compared to the other two model. While Figure 8 shows the decision tree model visualization.

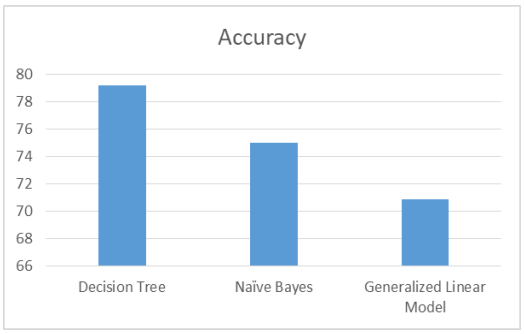

Figure 7. Classifier Model's Accuracy Comparison

DOI: 10.36079/lamintang.ijai-0602.42

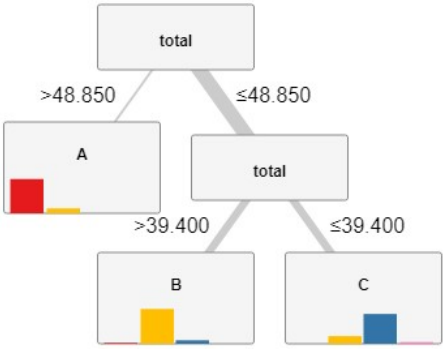

Figure 8. Decision Tree Visualization

Table 3 shows the result of confusion matrix evaluation of decision tree classifier. And since the decision tree model has the highest accuracy, it will be implemented to the system that is going to be developed.

Table 2. Confusion Matrix for Decision Tree C4.5 Algorithm

	Actually A	Actually B	Actually C	Actually D+	Class Precision
Predicted A	1	0	0	0	100.00%
Predicted B	1	10	2	0	76.92%
Predicted C	0	2	8	0	80.00%
Predicted D+	0	0	0	0	0.00%
Class Recall	50.00%	83.33%	80.00%	0.00%	

In attempt to increase the quality of the PSAP is to increase the number of rows of dataset in the databases. This will eventually decrease the number of cases of missing data thus increase the accuracy of the classifier model.

Other suggestions that are made are to add another column at the prediction result which is reasoning of the prediction result are made by the classifier. For instance, if the prediction result is "A" then the system will display a description of why the student will achieve that grade such as "Student have good past academic qualification and score excellent in coursework – Assignment 1".

6. Conclusion

As a conclusion, PSAP web-based application has benefit of allowing lecturer to monitor and predicting student grade. This helps lecturer to identify promising students and also provides them an opportunity to pay attention to and also improve those who would probably get lower grades. However, there are some improvement must be made in order to increase the level of user-friendliness of this application. Besides, improvement should be made to increase the level of user satisfaction of this application based on the user requirements.

Acknowledgement

This research worked under grant GPU-Khas Education 2017-0311-107-01 issued by Universiti Pendidikan Sultan Idris. Please do not hesitate to contact ismail.lamintang@ yahoo.com for any questions related to the data, coding, commercialization and others.

References

[1] Cherif, A. H., Adams, G. E., Movahedzadeh, F., Martyn, M. A., & Jeremy, D. (2014). Why Do Students Fail? Faculty's Perspective. 2014 Collection of Papers - Creating & Supporting Learning Environments.

[2] Nghe, N. T., Janecek, P., & Haddawy, P. (2007). A Comparative Analysis of Techniques for Predicting Academic Performance. 2007 37th Annual Frontiers in Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports (pp. T2G-7-T2G-12). Milwaukee, WI: IEEE.

DOI: 10.36079/lamintang.ijai-0602.42

[3] Amirah, M., Wahidah, H., & Nur'aini, A. (2015). A Review on Predicting Student's Performance using Data Mining Techniques. Procedia Computer Science, 72, 414-422.

[4] Osmanbegović, E., & Suljic, M. (2012). Data Mining Approach for Predicting Student Performance. Journal of Economics and Business/Economic Review, 10, 3-12.

[5] Adhatrao, K., Gaykar, A., Dhawan, A., Jha, R., & Honrao, V. (2013). Predicting Students' Performance Using ID3 And C4.5 Classification Algorithms. International Journal of Data Mining & Knowledge Management Process (IJDKP), 3(5).

[6] Nagy, H. M., Aly, W. M., & Hegazy, O. F. (2013). An Educational Data Mining System for Advising Higher Education Students. International Journal of Computer, Control, Quantum and Information Engineering, 7(10), 1266-1270.

[7] Chen, J.-F., Hsieh, H.-N., & Do, Q. H. (2014). Predicting Student Academic Performance: A Comparison of Two Meta-Heuristic Algorithms Inspired by Cuckoo Birds for Training Neural Networks. Algorithms, 7, 538-553.

[8] Natek, S., & Zwilling, M. (2014). Student data mining solution–knowledge management system related to higher education institutions. Expert Systems with Applications, 41, 6400-6407.

[9] Hamsa, H., Indiradevi, S., & Kizhakkethottam, J. J. (2016). Student Academic Performance Prediction Model Using Decision Tree and Fuzzy Genetic Algorithm. Procedia Technology, 25, 326-332.

[10] Sweeney, M., Rangwala, H., Lester, J., & Johri, A. (2016). Next-Term Student Performance Prediction: A Recommender Systems Approach. JEDM | Journal of Educational Data Mining, 8(1), 22-51.

[11] Adejo, O. W., & Connolly, T. (2017). Predicting student academic performance using multi-model heterogeneous ensemble approach. Journal of Applied Research in Higher Education, 10(1), 61-75.

[12] Chiheb, F., Boumahdi, F., Bouarfa, H., & Boukraa, D. (2017). Predicting Students Performance Using Decision Trees: Case of an Algerian University. 2017 International Conference on Mathematics and Information Technology (ICMIT) (pp. 113-121). Adrar: IEEE.

[13] Livieris, I., Drakopoulou, K., Kotsilieris, T., Tampakas, V., & Pintelas, P. (2017). DSS-PSP - A Decision Support Software for Evaluating Students' Performance. In G. Boracchi, L. Iliadis, C. Jayne, & A. Likas, Engineering Applications of Neural Networks. EANN 2017 (Vol. 744, pp. 63-74). Athen: Springer, Cham.

No		Strongly disagree	Disagree	Not sure	Agree	Strongly Agree
(A) Interface						
8	Website screen design is appropriate for media learning					
9	The fonts that are used are easy to read					
10	The web site does not contain spelling errors					
11	The graphics used in this software are interesting and helpful					
12	The color selection used in this website is appropriate					
(B) Functionality						
13	This website allows you to view, add, update and delete existing courses information					
14	This website allows you to view, add, update and delete existing coursework's information					
15	Files can be uploaded easily					
16	You are able to get and view results when using the "prediction" function					
17	All the buttons are functioning well					
(C) Navigation						
18	There is a clear indication of the current location					
19	There is a clearly-identified link to the Home page					
20	All major parts of the site are accessible from the Home page					
21	The navigation menu helps you use the web site easily					
22	Scrolling within a website page is easy					
(D) Reliabilty						
23	This website is suitable for use by lecturers who are teaching courses in university					
24	This website provided useful information for the lecturer about their student academic performance					
25	This website is suitable for use by students in university					
26	The website is user-friendly					