A Decision Tree Approach for Predicting Student Grades in Research Project using Weka

Ertie C. Abana
School of Engineering, Architecture, Interior Design and Information Technology Education
University of Saint Louis, Tuguegarao City, Cagayan, Philippines

***Abstract*—Data mining in education is an emerging multidiscipline research field especially with the upsurge of new technologies used in educational systems that led to the storage of massive student data. This study used classification, a data mining process, in evaluating computer engineering student's data to identify students who need academic counseling in the subject. There were five attributes considered in building the classification model. The decision tree was chosen as the classifier for the model. The accuracy of the decision tree algorithms, Random Tree, RepTree and J48, were compared using cross-validation wherein Random Tree returned the highest accuracy of 75.188%. Waikato Environment for Knowledge Analysis (WEKA) data mining tool was used in generating the classification model. The classification rules extracted from the decision tree was used in the algorithm of the Research Project Grade Predictor application which was developed using Visual C#. The application will help research instructors or advisers to easily identify students who need more attention because they are predicted to have low grades.**

***Keywords*—Data mining; classification rules; decision tree; educational data mining; WEKA**

I. Introduction

The developed world has experienced rapid increase in technology and information over the past few years. The Information Age has led to speedy flow and availability of information. This information comes from the massive data that are being extracted from different databases. When this data is analyzed using the statistical methods that are continuously being refined and perfected [1], valuable answers to business, social, environmental and educational problems are being discovered. In discovering valuable answers to many problems, new technology has emerged affecting human life in various spheres directly or indirectly [2]. This technology is called data mining or knowledge discovery in databases (KDD). Data mining is utilized to extract important information from complex databases [3][4][5]. The primary function of data mining is applying different methods and algorithms to preprocess, classify, cluster and associate the data to discover useful patterns [6][7] of stored data.

Education is one of the areas that benefited most in the emergence of data mining. This kind of data mining is called Educational Data Mining (EDM). It is used by educational institutions to provide better service to their students. Data mining also allows schools to use stored data to improve teaching and learning processes [4]. Educators will know so much more about the student's process which can improve students' performance in school. Moreover, it can be used to make data-informed decisions about what should people be doing for education. EDM can be used in many ways but perhaps the most common application of EDM is predicting a student's academic performance. Several studies along this area predict students' achievement in their subjects like mathematics [8], physics, chemistry, and biology [9]. All of these studies have the goal to identify at-risk students and identify priority learning needs for different groups of students [4] to create interventions and improve their performance.

Research is a subject in college that is embedded in the curriculum of any course. The activities in this subject are highly considered as a high-impact educational practice [10]. It is where lifelong learners' vital skills and attitude are being cultivated through inquiry [11][12]. The practice in most schools is that students are guided by an adviser when undertaking a research project during a specific period [13]. Students need to develop the skills necessary [13] for their research process especially in applied disciplines such as engineering, architecture and information technology. For example, computer engineering students are required to have a high level of proficiency in programming. Although skills are necessary to perform well in research, other factors like backlog and research method grade may serve as an indicator to student's performance. Backlogs are often considered as one of the factors in predicting students' academic performance [4][14] because this is considered as a burden to students. The grade in research method also serves as the basis on how the student will perform in a research project because all the basics of research are being taught in this subject. The gender of student doing the research is also important most especially in engineering disciplines because sometimes research projects being built are too heavy for female students to handle.

Although there are already studies that predict student's academic performance in subjects like math and science, none have studied about predicting a student's performance in an undergraduate research project course. This study addressed this gap. It proposed a classification model specifically decision tree algorithm in predicting the possible grade of a computer engineering student in Research Project.

The data mining software WEKA was utilized in the preprocessing of data, construction of classification model, and interpretation of the model. The decision tree generated was used to create a grade prediction software application. This software can be used in identifying students who needs academic counseling so that their performance in research will improve sufficiently and they will be able to produce a publishable or patentable research project output.

II. Methodology

A. Data Mining Software Utilized

The data mining software WEKA shown in Fig. 1 is programmed using Java. This software was developed at the University of Waikato in New Zealand [16]. It has many machine learning algorithms for different data mining tasks. It contains features that are used in data preparation and preprocessing, classification, clustering, association rules mining, regression, and visualization. WEKA is widely-used free software licensed under GNU General Public License (GPL). This software is not only recognized as a landmark system in data mining but also in machine learning [15]. Academia and business circles have been using this software for different purposes.

B. Collection and Preprocessing of Data

The researcher has been handling the Research subjects of the computer engineering program for the past four years. The grades of the students from the research subjects, particularly Research Method (RM) and Research Project (RP) were used as attributes for the project. RM is a pre-requisite subject of RP. It served as one of the attribute predictors in the model. On the other hand, RP served as the attribute class being predicted in the classification model. Three other attribute predictors were used which includes gender, backlog, and programming proficiency. The RM grades, RP grades, gender, and backlog data for the project has been collected from the University of Saint Louis Tuguegarao. The backlog was traced based on the year the student graduated. If the student graduated semester/s after completing RP, it means that the student still has backlogs. The programming proficiency was filled out manually by the researcher based on the student's programming proficiency level. The levels were Fundamental Awareness (basic knowledge), Novice (limited experience), Intermediate (practical application) and Advanced (applied theory). The RM and RP grades were converted into letter grades which includes A (92%–100%), B+ (87%–91%), B (83%–86%), C+ (79%–82%), C (75%–78%). This letter grade conversion was based on the letter grade equivalence of Ateneo de Manila University, except that it was only up to C since grades below this are considered failed.

The data were first collected in Microsoft Excel worksheet and initial preprocessing was done. The dataset contains 133 instances wherein each instance contains the five (5) attributes. The possible values of the different attributes are shown in Table I.

C. Classification Model Building

After the data collection and preprocessing, the classification models were finally built. The classifier used in the study was the decision tree. Decision tree has been used in numerous studies on prediction of student's academic performance [17][18][19] because classification rules can be derived in a single view. The Random Tree, RepTree and J48 decision tree were used for the model construction. Fig. 2, Fig. 3 and Fig. 4 shows the constructed decision tree for Random Tree, RepTree and J48, respectively. In the decision trees, the leaf node was represented by rectangle while the root node was represented by an oval [17].

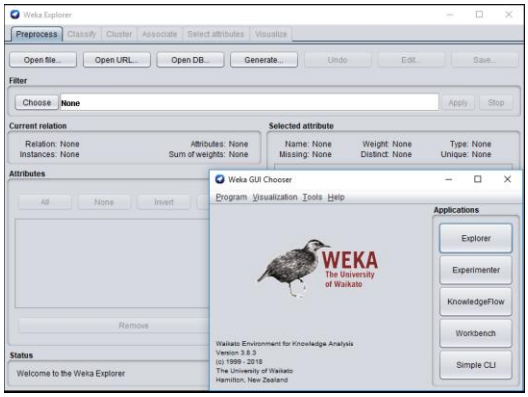

Fig. 1. The Graphical user Interface of WEKA.

TABLE I. The Attributes used in the Classification Model

Attributes Name	Possible Values
Gender	Male, Female
Backlog	Yes, No
Programming Proficiency	Fundamental Awareness, Novice, Intermediate, Advanced
RM Grade	A, B+, B, C+, C
RP Grade	A, B+, B, C+, C

III. Results and Discussion

A. Model Evaluation and Interpretation

Cross-validation was used to measure the predictive performance of the classification models. It was also used in previous studies [14][15][16] because it checks how a model performs when new data set or test data are used. Cross-validation is important because when a model is fit, it is usually fit only to the training dataset. With cross-validation, the prediction accuracy of the model can be seen when there is new data. In this study, the 10-fold cross-validation feature of WEKA was used to evaluate the classification models. The three different decision tree algorithms Random Tree, RepTree and J48 were compared. The result of compression is depicted in Table II for the cross-validation method. The decision tree with the highest accuracy was achieved by the Random Tree decision tree algorithm. The over-all accuracy of this classification model was 75.188% which means out of the 133 student grades in RP, 100 were correctly classified. This accuracy is better than that of previous studies [14][15][18] that also conducted prediction of student's academic performance but in general. The RepTree and the J48 were less accurate with both having 69.925% accuracy. From the results, it is noticeable that the accuracy of the classification models is acceptable but not very high. More samples should be collected and more attributes should also be added to have a very good classification model.

TABLE II. Acuracy of the Decision Tree Algorithms

Decision Tree	Accuracy (%)	Build Time	Correctly Classified Instances	Incorrectly Classified Instances
Random Tree	75.188	0.00	100	33
RepTree	69.925	0.02	90	43
J48	69.925	0.02	93	40

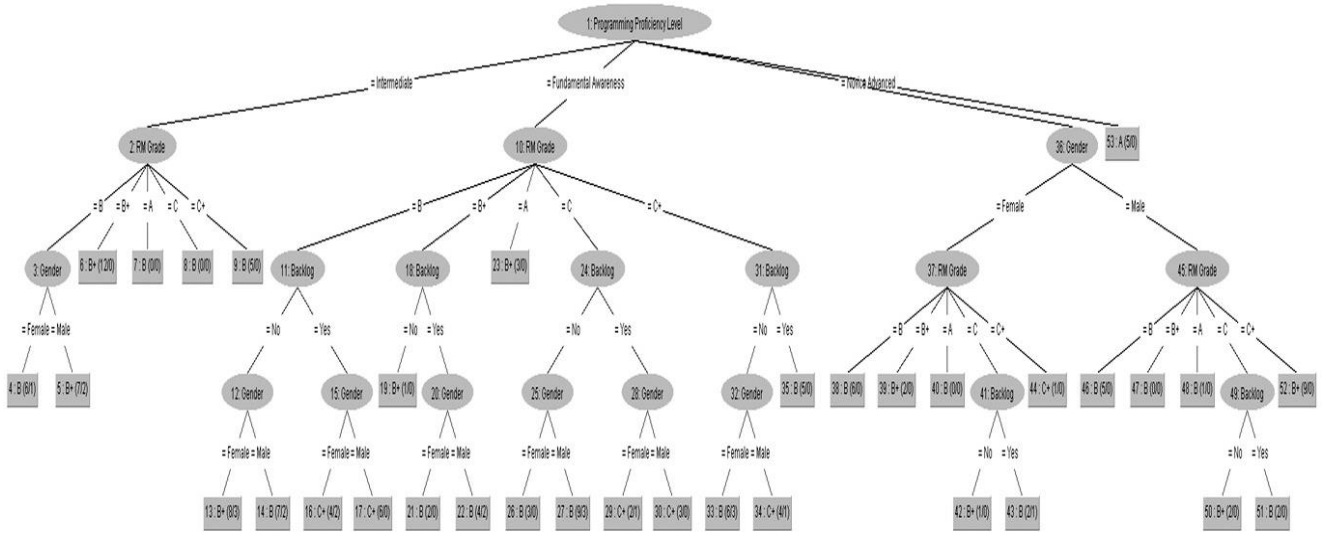

Fig. 2. The Constructed Random Tree Decision Tree using WEKA.

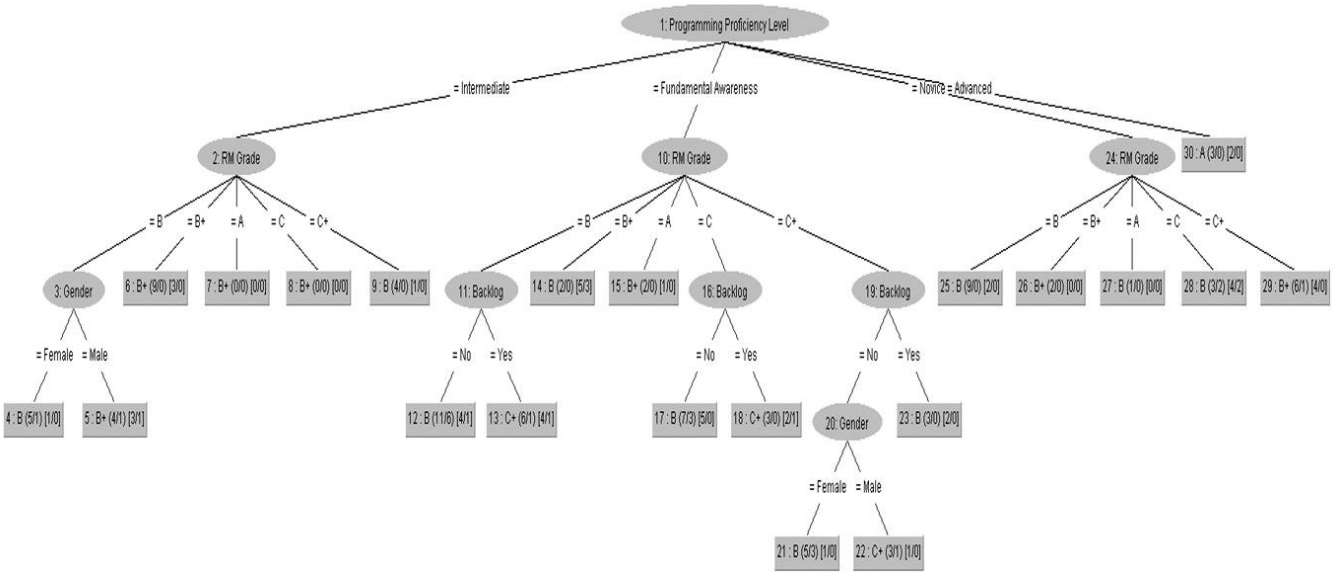

Fig. 3. The Constructed RepTree Decision Tree using WEKA.

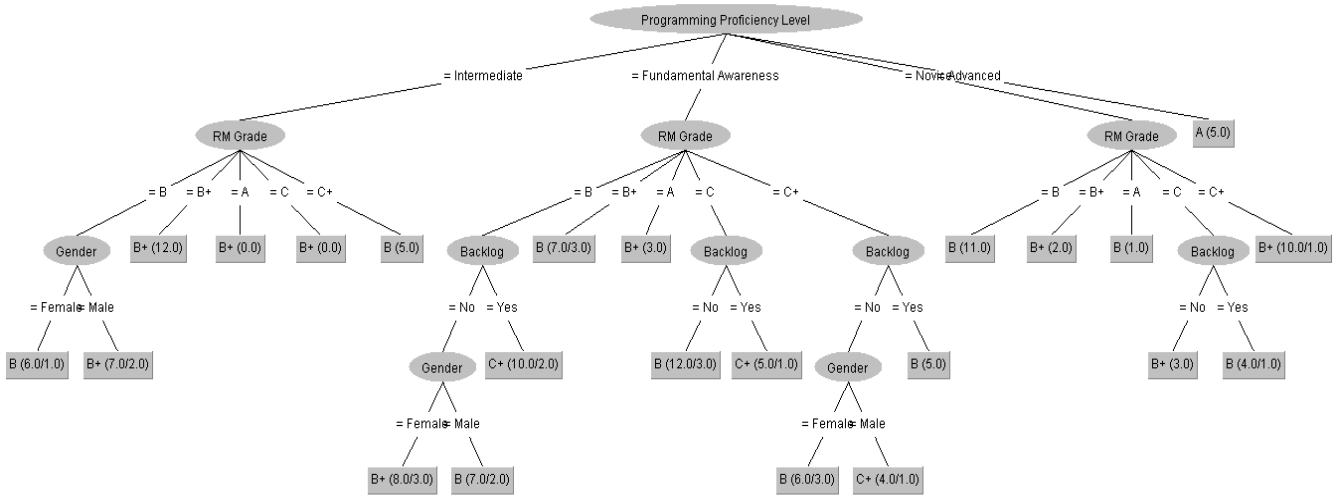

Fig. 4. The Constructed J48 Decision Tree using WEKA.

B. Classification Rules

Since Random Tree decision tree returned the highest accuracy after the 10-fold cross-validation, it was used to extract the classification rules. The rules were generated by getting the leaf nodes that were on the path of a root node in the decision tree. The logical conjunction of every leaf node from the path of a root node forms the rule while the root node represents the predicted grade.

A total of twenty-eight classification rules were extracted from the Random Tree decision tree as shown in Table III. The rules serve as a condition that when it is met, it would return an equivalent predicted grade. When the grade of a student is predicted, it can be used to determine if the student needs help in the research class.

C. RP Grade Predictor Software Application

Using the extracted classification rules from the generated decision tree, a Research Project Grade Predictor application shown in Fig. 5 was developed. The application was developed using the Visual C# programming language. Visual C# is one of the programming languages embedded in the Microsoft Visual Studio Express. The Microsoft Visual Studio Express is a collection free function-limited Integrated Development Environments (IDE) developed by Microsoft. The Visual C# IDE is a powerful and easy to use objected-oriented [20] programming language.

The software application was designed using a card base design to bridge the gap between interaction and usability [21] in a synchronized manner. It has a simple Graphical User Interface (GUI) that is user-centered wherein users are expected to run the application without training [22]. By using this application, the research instructor can now conduct proper counseling to students with low predicted grades.

TABLE III. THE SET OF CLASSIFICATION RULES

Rule No.	Rules	Predicted Grade	No. of Instances
1	If Programming Proficiency=Fundamental Awareness, RM Grade=A	B+	3/0
2	If Programming Proficiency=Fundamental Awareness, RM Grade=B+, Backlog=Yes, Gender=Male	B	4/2
3	If Programming Proficiency=Fundamental Awareness, RM Grade=B+, Backlog=Yes, Gender=Female	B	2/0
4	If Programming Proficiency=Fundamental Awareness, RM Grade=B+, Backlog=No	B+	1/0
5	If Programming Proficiency=Fundamental Awareness, RM Grade=B, Backlog=Yes, Gender=Male	C+	6/0
6	If Programming Proficiency=Fundamental Awareness, RM Grade=B, Backlog=Yes, Gender=Female	C+	4/2
7	If Programming Proficiency=Fundamental Awareness, RM Grade=B, Backlog=No, Gender=Male	B	7/2
8	If Programming Proficiency=Fundamental Awareness, RM Grade=B, Backlog=No, Gender=Female	B+	8/3
9	If Programming Proficiency=Fundamental Awareness, RM Grade=C+, Backlog=Yes	B	5/0
10	If Programming Proficiency=Fundamental Awareness, RM Grade=C+, Backlog=No, Gender=Male	C+	4/1
11	If Programming Proficiency=Fundamental Awareness, RM Grade=C+, Backlog=No, Gender=Female	B	6/3
12	If Programming Proficiency=Fundamental Awareness, RM Grade=C, Backlog=Yes, Gender=Male	C+	3/0
13	If Programming Proficiency=Fundamental Awareness, RM Grade=C, Backlog=Yes, Gender=Female	C+	2/1
14	If Programming Proficiency=Fundamental Awareness, RM Grade=C, Backlog=No, Gender=Male	B	9/3
15	If Programming Proficiency=Fundamental Awareness, RM Grade=C, Backlog=No, Gender=Female	B	3/0
16	If Programming Proficiency=Novice, Gender=Male, RM Grade=A	B	1/0
17	If Programming Proficiency=Novice, Gender=Male, RM Grade=B+	B	0/0
18	If Programming Proficiency=Novice, Gender=Male, RM Grade=B	B	5/0
19	If Programming Proficiency=Novice, Gender=Male, RM Grade=B	B	5/0
20	If Programming Proficiency=Novice, Gender=Male, RM Grade=C, Backlog=Yes	B	2/0
21	If Programming Proficiency=Novice, Gender=Male, RM Grade=C, Backlog=No	B+	2/0
22	If Programming Proficiency=Intermediate, RM Grade=A	B	0/0
23	If Programming Proficiency=Intermediate, RM Grade=B+	B+	12/0
24	If Programming Proficiency=Intermediate, RM Grade=B, Gender=Male	B+	7/2
25	If Programming Proficiency=Intermediate, RM Grade=B, Gender=Female	B	6/1
26	If Programming Proficiency=Intermediate, RM Grade=C+	B	5/0
27	If Programming Proficiency=Intermediate, RM Grade=C	B	0/0
28	If Programming Proficiency=Advanced	A	5/0

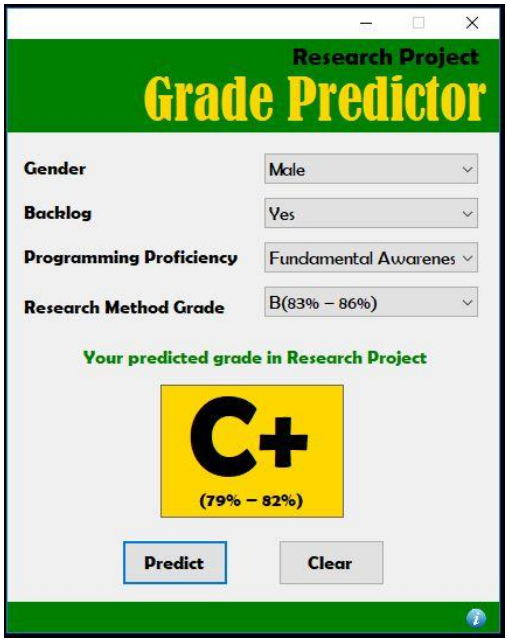

Fig. 5. The Grade Predictor Software.

IV. Conclusion

This study developed a classification model using a decision tree approach in predicting student grades in Research Project. It was limited to the use of only three decision algorithms which include Random Tree, RepTree and J48. The classification rules extracted from the Random Tree decision tree was used to create a software application that can be used by research instructors in identifying students who need academic counseling to improve their performance in research. The resulting accuracy of the classification model after the cross-validation means more samples and more attributes is still needed to arrive with a highly accurate prediction.

For future works, other decision tree algorithms should be used to analyze the data. The software application that can be developed with this kind of study can also be improved by adding a feature like allowing multiple student data to be analyzed at the same time.

Acknowledgment

I am thankful to God for giving me wisdom to finish this study. I am also thankful to the University of Saint Louis for all the help during the conduct of the research.

References

[1] B. Kitts, G. Melli, & K. Rexer, “Data Mining Case Studies,” In The First International Workshop on Data Mining Case Studies, IEEE International Conference on Data Mining, Huston, USA, 2005.

[2] D. Kumar & D. Bhardwaj, “Rise of data mining: current and future application areas,” International Journal of Computer Science Issues (IJCSI), 8(5), pp. 256, 2011.

[3] J. Han, J. Pei, & M. Kamber, “Data mining: concepts and techniques,” Elsevier , 2011.

[4] A. Algarni, “Data mining in education,” International Journal of Advanced Computer Science and Applications, 7(6), 456-461, 2016.

[5] J. Jacob, K. Jha, P. Kotak, & S. Puthran, “Educational Data Mining techniques and their applications,” In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT), pp. 1344-1348, 2015.

[6] Q. A. Al-Radaideh, E. M. Al-Shawakfa, & M. I. Al-Najjar, “Mining student data using decision trees,” In International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, 2006.

[7] M. Aggarwal, & A. Bhatia, "Pattern discovery techniques in online data mining," International Journal of Engineering and Technical Research, 3(7), pp. 28-31, 2015.

[8] C. T. Lye, L. N. Ng, M. D. Hassan, W. W. Goh, C. Y. Law, & N. Ismail, Predicting Pre-university student's Mathematics achievement,” Procedia-Social and Behavioral Sciences, 8, pp. 299-306, 2010.

[9] A. Yağcı, & C. Mustafa, "Predictions of academic achievements of vocational and technical high school students with artificial neural networks in science courses (physics, chemistry and biology) in Turkey and measures to be taken for their failures." In SHS Web of Conferences, 37, pp. 1-9, 2017.

[10] R. Imafuku, T. Saiki, C. Kawakami, & Y. Suzuki, “How do students' perceptions of research and approaches to learning change in undergraduate research?,” International journal of medical education, 6, pp. 47-55, 2015.

[11] A. B. Hunter, S. L. Laursen, & E. Seymour, “Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development,” Science education, 91(1), pp. 36-74, 2007.

[12] M. Healey, & A. Jenkins, “Developing undergraduate research and inquiry,” York Higher Education Academy, 2009.

[13] K. Zimbardi, & P. Myatt, “Embedding undergraduate research experiences within the curriculum: a cross-disciplinary study of the key characteristics guiding implementation,” Studies in Higher Education, 39(2), pp. 233-250, 2014.

[14] C. Lulla, Y. Agarwal, S. Kankariya, P. Sakaray, & P. Alappanavar, “Student academic performance prediction using machine learning and data mining techniques,” International Journal of Computer Science and Mobile Computing, 6(5), pp. 301-307, 2017.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, & I. H. Witten, “The WEKA data mining software: an update,” ACM SIGKDD explorations newsletter, 11(1), pp. 10-18,”

[16] A. H. Wahbeh, Q. A. Al-Radaideh, M. N. Al-Kabi, & E. M. Al-Shawakfa, “A comparison study between data mining tools over some classification methods,” International Journal of Advanced Computer Science and Applications, 8(2), pp. 18-26, 2011.

[17] M. Pandey, & V. K. Sharma, “A decision tree algorithm pertaining to the student performance analysis and prediction,” International Journal of Computer Applications, 13(1), pp. 61-72, 2013.

[18] K. D. Kolo, S. A. Adepoju, & J. K. Alhassan, “A decision tree approach for predicting students academic performance,” International Journal of Education and Management Engineering, 5(5), pp. 1-5, 2015.

[19] Q. A. Al-Radaideh, E. M. Al-Shawakfa, & M. I. Al-Najjar, "Mining student data using decision trees,” In International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan, pp. 1-5, 2006.

[20] E. Abana, K. H. Bulauitan, R. K. Vicente, M. Rafael, & J. B. Flores, “Electronic Glove: a Teaching Aid for the Hearing Impaired,” International Journal of Electrical and Computer Engineering, 8(4), pp. 2290-2298, 2018.

[21] J. Kiruthika, S. Khaddaj, D. Greenhill, & J. Francik, “User Experience design in web applications,” International Conference on Computational Science and Engineering, Paris, New York, pp. 642-646, 2016.

[22] E. Abana, M. Pacion, R. Sordilla R, D. Montaner, D. Agpaoa, R. M. Allam, "Rakebot: a robotic rake for mixing paddy in sun drying," Indonesian Journal of Electrical Engineering and Computer Science (IJEECS), 14(3), pp. 1165-1170, 2019.