Canadian Journal of Pure and Applied Sciences
Vol. 16, No. 1, pp. 5393-5398, Feb 2022
Online ISSN: 1920-3853; Print ISSN: 1715-9997
Available online at www.cjpas.net

PREDICTION OF STUDENT PERFORMANCE IN GRADUATION PROJECT IN INFORMATION SYSTEMS PROGRAM USING MACHINE LEARNING ALGORITHMS IN WEKA

Badr Mohammed Almezaini and *Muhammad Asif Khan
Department of Information Systems, College of Computer Science and Engineering
Taibah University, Madina al Munawwara, Saudi Arabia

ABSTRACT

Educational institutions strive to monitor and develop student academic performance by difference means in order to prepare quality graduates. Educational data mining is increasingly becoming a latest trend which aids educational institutions to predict academic performance of students using machine learning techniques. Researchers have conducted research to predict student academic performance, but did not consider the prerequisite courses for final graduation project which is major culmination activity of an undergraduate degree program. As result students could not develop quality projects. In this current study we have used student data in three prerequisite courses required to begin graduation project. We have used data of three prerequisite courses and applied Naïve Bayes. J48 and Neural Network algorithms in WEKA to predict student performance in final graduation project. The accuracy and confusion matrix have been discussed and the results obtained in both the classifiers also elaborated. The results help students to focus and complete graduation projects with high quality

Keywords: WEKA, education data mining, machine learning, data mining, graduation project.

INTRODUCTION

Higher education institutions are always concerned about student academic performance and their marketability after graduation. Therefore, it is important to keep the curriculum updated in order to align students' knowledge and skills with industry requirements. In view of rapidly changing technologies and business models the department of Information Systems has introduced database management systems, web applications development and information systems project management courses as prerequisites to begin final year graduation project. Students are expected to complete the three courses successfully before they could take graduation project. Students need to gain good, extensive knowledge and skills in the above three courses in order to complete graduation project. The required three courses are offered in level 6,7 and 8 whereas final graduation project starts in level 9 and ends in level 10.

Currently we are facing a problem of weak graduation projects due to lack of database, web development and project management skills among students. Students do not know whether they would be able to make a final graduation project with an acceptable quality. This research will help students and supervisors to know expected performance in projects based on grades in the above stated courses. During the final year students focus on their projects and their assigned supervisors guide them throughout the project duration. This research will facilitate supervisors to guide students more effectively in view of expected performance. The graduation projects reflect student knowledge and capabilities in the latest tools and technologies. Therefore, students must show competency and knowledge in the above stated three required courses in order to build good projects.

There is an emerging trend of educational data mining (EDM) to extract important information for predicting student academic performance (Algarni, 2016). There are many advantages of EDM and predicting academic performance of student is one of its most significant benefit (Fan *et al.*, 2019). It can be used to predict student overall academic performance in a program or performance in specific courses (Alturki *et al.*, 2020). There are various methods and techniques in EDM used to discover knowledge and predictions such as Neural Networks, Classification, Naïve Bayes, Clustering, J48, K-nearest neighbours, Decision Tree etc. We briefly define the different techniques below to develop our understanding.

Neural Networks

Neural networks techniques is used mainly on continuous data values which show a pattern for prediction. It is a learning system that is comprised of a network of

*Corresponding author e-mail: asifkhan2k@yahoo.com functions which is provided with a data to translate into desired results. This technique extracts information from complex data and predicts trends.

Classification

This approach learns from the data given to it and makes classifications to the new data. It is used for both structure and unstructured data and predicts classes which are also known as labels or categories.

Naïve Bayes

This classification classifier is simple and easy to build for large dataset. This requires a small amount of training data to produce good results. It shows excellent accuracy and fast processing.

J48

This algorithm is an statistical classifier which generates a decision tree. It can create decision tree for both continuous and categorial data attributes.

K-Nearest Neighbour

This technique determines the classification based on k nearest neighbour. A new data point is classified based on the nearest available labelled points to the new data point.

Decision Tree

A classification model is built by this decision tree algorithm analogous to a tree structure. In tree structure, a rectangle represents internal node whereas oval shows a leaf node.

In this study we use Naïve Bayes, Neural Network and J48 techniques to predict graduation project performance of students based on their marks obtained in the prerequisite three courses. The reason of selecting the above stated three techniques is their reliable and consistent efficiency in predicting The high accuracy of prediction will help instructors to warn students in advance to keep the right track in their projects and complete successfully. This study uses student data from the last three years and based on the grades obtained in the required three courses performance in graduation project is predicted.

RERELATED WORK

In educational institutions data mining i.e. educational data mining (EDM) is widely used in order to obtain valuable information to develop our understanding of academic and learning process (Romero and Ventura, 2010). The variables related to student academic performance are identified and extracted to predict academic performance.

For predicting academic achievement many researchers have conducted various studies. A study (Aulck *et al.*, 2016) found that demographic features such as age, location are important to predict student academic achievement (Aulck *et al.*, 2016; Kemper *et al.*, 2020) In a separate study (Milos, 2016) researchers used WEKA tool and used two different data sets of different students to predict final marks. In another study (Al-Radaideh, 2006) three different classification methods were used and it was concluded a decision tree was best to predict student grade in C++ course. In a study Classification technique was used to predict the number of enrolled students in view of academic data (Shannaq *et al*., 2010). In another study (Asif *et al*., 2017) described that in a 4-year program the first- and second-year academic performances are important to predict academic achievement. In some studies it was found that credit hours load is one of the important features used to predict student academic performance (Saa, 2016 ; Yehuala, 2015). In a research conducted by (Hilal, 2017), using WEKA tool a comparison of five classifiers to predict academic performance of students is presented. It is found that Bayesian Network classifier is the best in prediction of student performance. In a study (Sayali and Mohini, 2014), using classification technique and data from student management system weak students in terms of academic performance were identified in order prevent from failure. In an interesting study (Maryam *et al*., 2017) researchers evaluated the performance of different feature selection algorithms using different classification models during educational data mining. It was stated to predict student performance selection of relevant features is significant in building a model. In another study final students grades of an online forum were predicted using Clustering techniques in WEKA (Lopez *et al*., 2012). There is a study (Miranda *et al*., 2013) in which a genetic algorithm is used in order to identify attributes that impact student performance. The researchers in a study (Baradwaj and Pal, 2011) used a classification algorithm named ID3 to predict student grade in master of computer application course with aim to improve performance. In another study (Mashael and Muna, 2016) researchers used classification technique in WEKA to predict student final grade based on mandatory courses and identified important course that impacts final grade. In a study (Ertie, 2019) a decision tree classifier is used to predict grades of students in research projects using WEKA. In another study (ElGamal, 2013) algorithm is applied to extract rules which predict student performance in a programming course.

MATERIALS AND METHODS

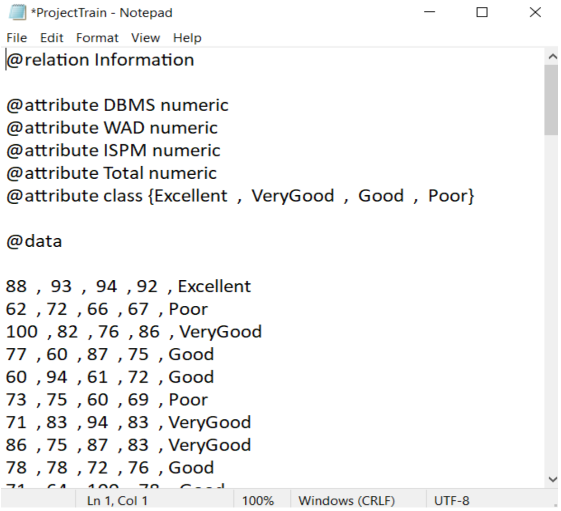
The three prerequisite courses i.e. database management systems, web applications development and information system project management for graduation project are offered in information system undergraduate program at level 6, level 7 and level 8 respectively. The data was comprised of 143 instances and each instance consists of five attributes including a class. The data was preprocessed in excel sheet in order to prepare it to be able to be processed in the data mining tool. Table 1 shows the main properties of the dataset.

Table 1. Data features and characteristics.

Feature	Description	Data type	Value
DBMS	Marks in database management systems	Numeric	60 - 100
WAD	Marks in the web application development	Numeric	60-100
ISPM	Information systems project management	Numeric	60-100
Total	Total marks in the three required courses	Numeric	60-100
Class	Performance in graduation project	Nominal	Excellent, Very good, Good, Poor

During the data preparation process, the attribute total is calculated as the mean marks obtained in the three prerequisite courses whereas, instances with missing values were removed. The types of the three courses and total attributes are numeric type and the class is nominal type.

In order to process data in WEKA we prepared the data file in the required format so that classifiers could be applied on the data. Figure 1 shows sample of data.



```
@relation Information

@attribute DBMS numeric
@attribute WAD numeric
@attribute ISPM numeric
@attribute Total numeric
@attribute class {Excellent , VeryGood , Good , Poor}

@data

88 , 93 , 94 ,92 , Excellent
62 ,72 ,66 ,67 ,Poor
100 ,82 ,76 ,86 ,VeryGood
77 ,60 ,87 ,75 ,Good
60 ,94 ,61 ,72 ,Good
73 ,75 ,60 ,69 ,Poor
71 ,83 ,94 ,83 ,VeryGood
86 ,75 ,87 ,83 ,VeryGood
78 ,78 ,72 ,76 ,Good
```

Fig. 1. A sample of training data file.

In order to create a model, we separated training data from testing data i.e. testing data is unseen. The training data file consists of 99 instances whereas testing data file comprised of 43 instances. In any classifier accuracy of prediction is important and although 100% accuracy is unattainable maximum accuracy is the goal of classification.

A classification model makes various correct and incorrect predictions which can be compared with the actual values. A confusion matrix depicts such possibilities a classification model can make. Performance of an algorithm can be visualized by using confusion matrix (Alqahtani, 2021). Table 2 shows a confusion matrix.

Table 2. Confusion matrix.

Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)
	Positive	**Negative**

It is clear from the confusion matrix that when a model classifier predicts positive instances that correspond to the target instances, the rate is True positive i.e.

TP rate = TP / (TP+FN)

Similarly, when a model predicts Negative instances but actually they are positive it is known False positive i.e.

FP rate = FP / {TN+FP)

A precision in a classifier shows the proportion of correct classification from the instances predicted as positive i.e.

Precision = TP / (TP + FP)

A recall in a classifier is the proportion of correct classification from actual correct instances i.e.

Recall = TP / (TP + FN)

In order to evaluate a model, accuracy is a measure that is defined as below

Accuracy = (TP + TN) / (TP + FP + FN + TN)

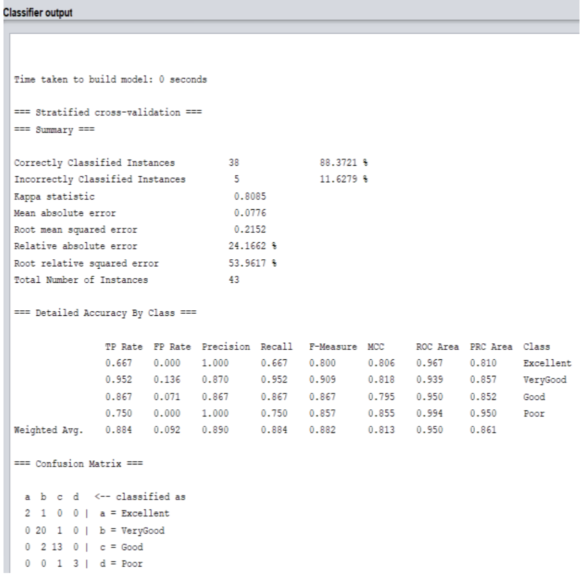
In the current study we have selected two classifiers i.e. Naïve Bayes and J48 to predict student performance in graduation project based on the three prerequisite courses.

RESULTS AND DISCUSSION

We selected the tree classifiers to predict performance in graduation project i.e. Naïve Bayes, J48 and Neural Network classifiers. The reason is their good output for the problem considered.

Naïve Bayes Model

After collecting, cleaning and formatting data into the required format we selected a classifier Naïve Bayes in WEKA. In order to train the model training data with 99 instances provided and the model predicted instances with 91.91%. Figure 2 depicts the result.



Classifier output

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          38               88.3721 %
Incorrectly Classified Instances         5               11.6279 %
Kappa statistic                          0.8085
Mean absolute error                      0.0776
Root mean squared error                  0.2152
Relative absolute error                 24.1662 %
Root relative squared error             53.9617 %
Total Number of Instances               43

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.667    0.000    1.000      0.667    0.800      0.806    0.967     0.810     Excellent
                 0.952    0.136    0.870      0.952    0.909      0.818    0.939     0.857     VeryGood
                 0.867    0.071    0.867      0.867    0.867      0.795    0.950     0.852     Good
                 0.750    0.000    1.000      0.750    0.857      0.855    0.994     0.950     Poor
Weighted Avg.    0.884    0.092    0.890      0.884    0.882      0.813    0.950     0.861

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  2  1  0  0 |  a = Excellent
  0 20  1  0 |  b = VeryGood
  0  2 13  0 |  c = Good
  0  0  1  3 |  d = Poor
```

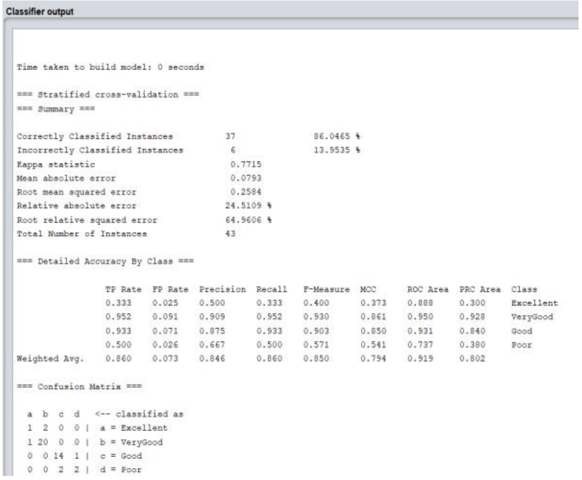
Fig. 2. Naïve Bayes Model testing data results.

The confusion matrix shows the instances in each class were classified more than 90% correctly. Table 3 shows the results of precision, recall and F-values.

Table 3. Accuracy data by class.

	Excellent	Very Good	Good	Poor
True Positive	0.667	0.952	0.867	0.750
False Positive	0.000	0.136	0.071	0.000
Precision	1.000	0.870	0.867	1.000
Recall	0.667	0.952	0.867	0.750
F-Value	0.800	0.909	0.857	0.818

The data shows the model provides reasonable accuracy rate 88.37% with the 10-fold cross validation of testing data with 43 instances. It is to be noted that the model was trained with training data of 99 instances and obtained 91.91% accuracy with 10-fold cross validation.



Classifier output

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          37               86.0465 %
Incorrectly Classified Instances         6               13.9535 %
Kappa statistic                          0.7715
Mean absolute error                      0.0793
Root mean squared error                  0.2584
Relative absolute error                 24.5109 %
Root relative squared error             64.9606 %
Total Number of Instances               43

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.333    0.025    0.500      0.333    0.400      0.373    0.888     0.300     Excellent
                 0.952    0.091    0.909      0.952    0.930      0.861    0.950     0.928     VeryGood
                 0.933    0.071    0.875      0.933    0.903      0.850    0.931     0.840     Good
                 0.500    0.026    0.667      0.500    0.571      0.541    0.737     0.380     Poor
Weighted Avg.    0.860    0.073    0.846      0.860    0.850      0.794    0.919     0.802

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  1  2  0  0 |  a = Excellent
  1 20  0  0 |  b = VeryGood
  0  0 14  1 |  c = Good
  0  0  2  2 |  d = Poor
```

Fig. 3. J48 Model with testing data.

J48 Model

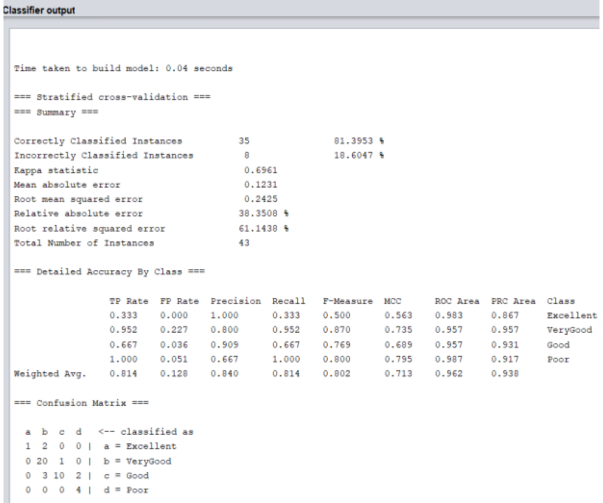
This model is an statistical algorithm and an extension of ID3 (Iterative Dichotomiser 3) model which is developed by Ross Quinlan. This model is used to generate decision tree from a given dataset. We used the training data by applying J48 algorithm with 10-fold cross validation and the model predicted 90.90% accuracy from 99 instances. However, with the same test data of 43 instances the accuracy rate 86.04% was achieved. Figure 3 show the tree created by the algorithm.

The confusion matrix depicts the classifier successfully predicted the classes 86.04% where low accuracy was found in predicting Excellent class. Table 4 show the accuracy data by class.

Table 4. Accuracy data by class.

	Excellent	Very Good	Good	Poor
True Positive	0.333	0.952	0.933	0.500
False Positive	0.025	0.091	0.071	0.026
Precision	0.500	0.909	0.875	0.667
Recall	0.333	0.952	0.933	0.500
F-Value	0.400	0.930	0.903	0.571

The confusion matrix shows the model could not identify Excellent class properly as the accuracy is 33% whereas Poor class was reasonable identified. However, remaining classes were classified with more than 93%.



Classifier output

```
Time taken to build model: 0.04 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          35               81.3953 %
Incorrectly Classified Instances         8               18.6047 %
Kappa statistic                          0.6961
Mean absolute error                      0.1231
Root mean squared error                  0.2425
Relative absolute error                 38.3508 %
Root relative squared error             61.1438 %
Total Number of Instances               43

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.333    0.000    1.000      0.333    0.500      0.563    0.983     0.867     Excellent
                 0.952    0.227    0.800      0.952    0.870      0.735    0.957     0.957     VeryGood
                 0.667    0.036    0.909      0.667    0.769      0.689    0.957     0.931     Good
                 1.000    0.051    0.667      1.000    0.800      0.795    0.987     0.917     Poor
Weighted Avg.    0.814    0.128    0.840      0.814    0.802      0.713    0.962     0.938

=== Confusion Matrix ===

  a  b  c  d   <-- classified as
  1  2  0  0 |  a = Excellent
  0 20  1  0 |  b = VeryGood
  0  3 10  2 |  c = Good
  0  0  0  4 |  d = Poor
```

Fig. 4. Neural Network Model with testing data.

Nerual Network Model

This model is known as multilayer perception in which nodes are connected with each other and work in parallel to produce output. We provided the training data to train the model and the accuracy rate was achieved 93.93% under 10-fold cross validation from the 99 instances. However, on testing data the model produced classification of 81.39% It is to be noted that in the training data the model identified 70% class of Excellent, but with the testing data the same class achieved only 33% which is very low as compare to other classes. Figure 4 shows the results of the model from 43 instances.

It is clear from the confusion matrix that the model did not classify Excellent class was not it is evident from the F-value and recall values. The Table 5 depicts the accuracy data by class.

Table 5. Accuracy data by class.

	Excellent	Very Good	Good	Poor
True Positive	0.333	0.952	0.667	1.000
False Positive	0.000	0.227	0.036	0.051
Precision	1.000	0.800	0.909	0.667
Recall	0.333	0.952	0.667	1.000
F-Value	0.500	0.870	0.769	0.800

This model predicts all the classes with high accuracy except the excellent class where the accuracy is 33%. Overall the model has shown the good performance in predicting the performance based on the required courses.

CONCLUSION

This study shows that student performance based on three required courses for graduation project can be predicted by using different classifiers in WEKA. However, it is found the accuracy of prediction varies from one classifier to another. In this study we found Naïve Bayes classifier gives high accuracy in prediction than J48 and Neural Networks classifiers under 10-fold cross validation. By using Naïve Bayes model student performance in graduation projects can be improved as graduation projects are important in student career. In future we would like to increase the data and number of classifiers using other data mining tools in order to compare results and predict results with higher accuracy.

REFERENCES

Alqahtani, A. 2021. Product Sentiment Analysis for Amazon Reviews. International Journal of Computer Science and Information Technology. 13(3).

Abu Saa, A. 2016. Educational data mining and students' performance prediction. International Journal of Advanced Computer Science and Applications. 7(5): 212-200.

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

Alturki, S., Hulpus, I. and H. Stuckenschmidt. 2020. Predicting academic outcomes: A survey from 2007 till 2018. Technology, Knowledge and Learning. 1-33.

Asif, R., Merceron, A., Ali, A. and N. Haider, N. 2017. Analyzing undergraduate students' performance using educational data mining. Computers and Education. 113: 177-194.

Aulck, L., Velagapudi, N., Blumenstock, J. and West, J. 2016. Predicting student dropout in higher education. Proceedings of the ICML Workshop on #Data4Good: Machine Learning in Social Good Applications, New York, USA.

Baradwaj, BK. and Pal, S. 2011. Mining educational data to analyze students' performance. International Journal of Advanced Computer Science and Applications. 2:63-69.

ElGamal, AF. 2013. An educational data mining model for predicting student performance in programming course. International Journal of Computer Applications. 70(17):22-28.

Ertie, A. 2019. A Decision Tree Approach for Predicting Student Grades in Research Project using Weka. International Journal of Advanced Computer Science and Applications. 10(7):285-289.

Fan, Y., Liu, Y., Chen, H. and Ma, J. 2019. Data mining-based design and implementation of college physical education performance management and analysis system. International Journal of Emerging Technologies in Learning. 14(6):87-97.

Hilal, A. 2017. Analysis of Students' Performance by Using Different Data Mining Classifiers. International Journal of Modern Education and Computer Science. 8:9-15.

Kemper, L., Vorhoff, G. and U. Wigger, U. 2020. Predicting student dropout: A machine learning approach. European Journal of Higher Education. 10(1):28-47.

López, M., Luna, J., Romero, C. and Ventura, S. 2012. Classification via clustering for predicting final marks based on student participation in forums. In 5th International Conference on Educational Data Mining, ERIC. 148-151.

Maryam, Z., Manzoor, H. and Savia, K. 2017. Performance analysis of feature selection algorithm for educational data mining. In IEEE Conference on Big Data and Analytics. 7-12.

Mashael, A. and Muna, A. 2016. Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology. 6(7):528-533.

Milos, I., Petar, S., Mladen, V. and Wejdan, A. 2016. Students' success prediction using Weka tool. Infoteh-Jahorina. 15:684-689.

Miranda, T., Martin, A. and Prasanna, V. 2013. An Analysis of Students Performance Using Genetic Algorithm. Journal of Computer Sciences and Applications. 1(4):75-79.

Q. AI-Radaideh, Q., AI-Shawakfa, E. and M. AI-Najjar, M. 2006. Mining student data using decision trees. International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan.

Romero, C. and S. Ventura, S. 2010. Educational data mining: A review of the state of the art. IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews. 40(6):601-618.

Sayali, R. and Mohini, M. 2014. Quality improvisation of student performance using data mining techniques. International Journal of Scientific and Research Publications. 4(4):1-4.

Shannaq, B., Rafael, Y. and Alexandro, V. 2010. Student Relationship in Higher Education Using Data Mining Techniques. Global Journal of Computer Science and Technology. 10(11):54-59.

Yehuala, M. 2015. Application of data mining techniques for student success and failure prediction (the case of Debre_Markos University). International Journal of Scientific & Technology Research. 4(4):91-94.

Received: Dec 22, 2021; Accepted: Jan 28, 2022