International Journal of Future Generation Communication and Networking
Vol. 14, No. 1, (2021), pp. 5061-5070

An Exploration of Teaching-Learning Using Data Mining Classification Algorithms in Higher Education

Prashant Chintal[1], Haftom Gebregziabher[2] and Vikrant Shaga[3]

[1]Marathwada Institute of Technology, Dr. B.A.M University, Aurangabad, Maharashtra,India. prashant.chintal@gmail.com
[2]Ethiopian Technical University , Addis Ababa, Ethiopia
haftom.gebregziabher@ethernet.edu.et
[3]Department of Information Communication Technology, Ethiopian Technical University Addis Ababa, Ethiopia, shagavik@gmail.com

Abstract

Many technological advances have been made in recent years for enhancing teaching-learning. Students' engagement using activity-based learning is a major concern for the teachers. This study looked at student feedback in order to assist educational stakeholders in taking remedial measures to improve their students' performance. The amount of data saved in educational institutions is significantly expanding during this pandemic period. These data contain hidden information for the improvement of students' performance, teaching, planning, and so on. To achieve the goal of this study, the data mining techniques and the WEKA tool are used to implement the classification algorithms. We used J48, Naïve Bayes, REPTree, PART, and JRip classifiers for the experiments. Therefore, this study will also help the teachers to enhance their teaching mechanism to measure and improve students' understanding by considering overall factors of learning.

Keywords: activity-based learning, data mining, J48, Naïve Bayes, REPTree, PART, JRip

Introduction:

Education plays such a role as it increases and strengthens the creative and productive capacity of students. Educational institutes play an important role in the social and economic transformation of society [1]. Online education and training nowadays are one of the essential driving forces and a necessary condition for engaging students for online activity-based learning. After the engagement, the next concern is the understanding level of the students and measuring their learning skills. Activity-Based Learning (ABL) is a method enabling students to assess and learn from a practical experience (Learning by doing). It could be online or offline classes but learning by doing makes the learning easy for the students [2]. More than 8000 learning activities have been created to promote activity-based learning for more than 100 courses of one of the universities for enhancing the validity of knowledge assessment [3].

In this technological era, many Higher Education Institutes (HEI's) from different countries introduced ICT in education for improving the education process. The use of ICT in the education system has increased the efficiency of the educational processes. The students are learning through LMS (Learning Management Systems), Online learning portals, different online certification courses, and many other innovative approaches. But the important fact is whether they understand the concepts and the pedagogical activities. The role of technology is to provide useful information, general development, and increase creative thinking [4]. Predicting the performance of student help in identifying the risk of failure and accordingly

ISSN: 2233-7857 IJFGCN

help the teachers to take some essential steps for improving the performance of individual or group of students [5]. The regression algorithm of machine learning was used to make some empirical predictions about student performance. With the help of this research and its findings, a prototype version of a software application tool for tutors has been created to aid in proper decision-making [6].

For this research study, different classification algorithms were applied to develop the model. The reason for selecting these algorithms is they are easy to implement and use, robust to isolated noise points, handles missing values by ignoring the instance during probability estimate conclusions, robust to irrelevant attributes, is also easy to understand, handles all types of attributes and they are also computationally cheap and also various literature and articles show that these algorithms perform at high accuracy when applied in building model for student performance [7]

Data mining in education:

The study of collecting, cleaning, processing, analyzing, and extracting meaningful insights from data is known as data mining [8]. The study was carried out by educational researchers. The purpose of this study was to identify knowledge in order to conduct an analysis of student motivation on e-Learning systems using data mining techniques [9].

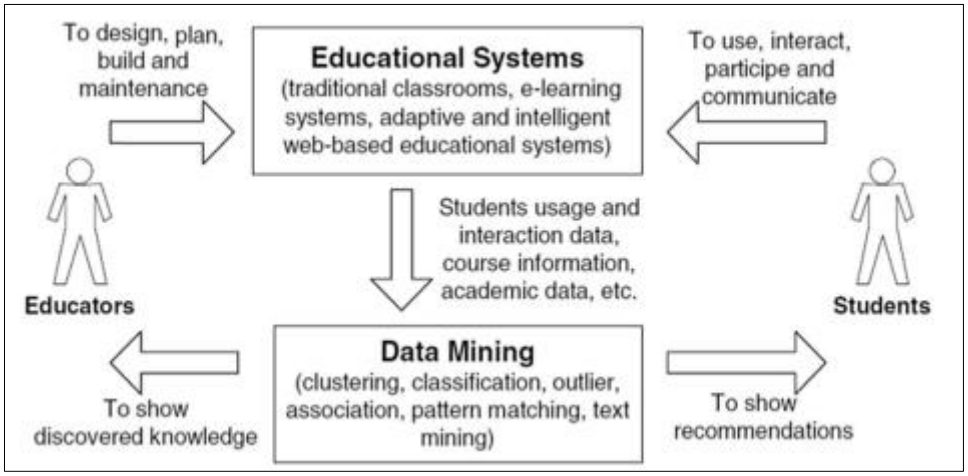

Figure 6. The cycle of applying data mining in educational systems [10]

The flow of educational systems using data mining techniques, also known as educational data mining, is depicted in Figure 6. (EDM). The entire participant in the educational system is involved in the iterative cycle of the educational process to fulfil their specific needs. Clustering, classification, outlier association, pattern matching, and text mining are some of the data mining techniques used by EDM [10].

Naive Bayes Classifier

The simplest probabilistic classifier is Naive Bayes, which is based on the Bayes theorem. In machine learning, Nave Bayes classifiers are used for any classification based on the conditional possibility or probabilities of the features attributed to a class, where the features are selected using feature selection methods. [11][12]

J48 Classifier

ISSN: 2233-7857 IJFGCN

C4.5, also known as J48 in the WEKA machine learning environment, is a prevalent decision tree classifier. A decision tree performs classification by repeatedly separating attributes into branches in the shape of a tree. Mathematical algorithms such as the Gain ratio and Information gain are frequently used to determine one attribute and its threshold in order to split the attributes into two subgroups. The first node is known as the root node, and the subsequent nodes are known as leaf nodes. The preceding procedure is repeated at each leaf node until the tree is complete. The final node is the end node. The J48 classifier is described as a versatile and widely used classification tool [12]. Decision trees have the well-known advantage of representing guidelines that users can easily interpret and understand, and they do not necessitate complex data preparation.

PART Classifier

PART (Projective Adaptive Resonance Theory) is simply a rule learner that divides and conquers, producing sets of decision list rules. Each new piece of data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule. In each iteration, PART constructs a decision tree and turns the best leaf into a rule [12].

JRip Classifier

The most popular algorithm is JRip, which uses sequential covering algorithms to create ordered rule lists. Classes are examined as they grow in size, and an initial rule for the class is generated with incrementally reduced error. JRip begins by treating all examples of a specific decision on the training data as a class and determining a set of rules that covers all members of that class. It then moves on to the next class and repeats the process until all classes have been covered [12].

REPTree Classifier

Reduces Error Pruning (REP) Tree, where Classifier is a fast type of decision tree learner that is built for the decision tree or regression tree using information gain and entropy, and which, like the C4.5 Algorithm, deals with missing values by breaking the respective instances into pieces. [13]

Dataset and Attribute Selection:

Participants were 182 post-graduate students of HEI's. All students belong to Computer Science, Information Technology, and Computer Application background with knowledge of using modern LMS. The feedback survey was conducted for the post-graduate courses to know the understanding and pedagogy are useful or not. The survey is divided into six special sections to collect information or feedback about the teaching-learning mechanism, course contents, and instructor interaction. These sections' questions are classified according to relevance, reflective thinking, interactivity, tutor support, peer support, and interpretation. [14] [15]

Table 1. Original attributes with their description and data type.

S.No	Attribute Name	Data Type	Description
1	Gender	Nominal	Gender of the candidate
2	Course	Nominal	Course’s name

ISSN: 2233-7857 IJFGCN

3	Subject	Nominal	Subjects in the courses
4	Relevance	Numeric	Self-assessment points based on students interest and important for improving professional practice
5	Reflective Thinking	Numeric	Self-assessment points based on critical thinking about self-learning, own ideas and readings
6	Interactivity	Numeric	Self-assessment points based on sharing ideas with other peers
7	Tutor Support	Numeric	Self-assessment points based on tutors encouragement and learning models and participation
8	Peer support	Numeric	Self-assessment points based on peers encouragement, contribution and participation in learning
9	Interpretation	Numeric	Self-assessment points based on understanding other students messages as well as tutors messages
10	Course Year	Numeric	Year on which course was conducted
11	Time taken for feedback (in minutes)	Numeric	Time taken by student to fill feedback
12	Performance	Numeric	Total score from online exam
13	Result	Nominal	Either pass, fail or not appeared
14	Institute	Nominal	The institute where student is trained
15	Place	Nominal	City from state
16	State	Nominal	State from country
17	Country	Nominal	Students' country

As shown in Table 1, there are 17 attributes. There were 17 attributes in total, including the target.

The entire target dataset was not used for the data mining task. Because the primary goal of this research was to use data mining techniques to identify major factors influencing students' performance based on their feedback. Because the entire attribute was not found useful for this study, the researcher performs attribute selection. For this study, the variables, or features, that describe the data were chosen to obtain a more essential and compact representation of the available information. A total of 13 attributes were chosen. Additionally, attribute selection was performed using the WEKA DM tool using WEKA attribute selection via *InfoGainAttributeEval* and search method ranker. Gain Ratio is a metric for evaluating attribute value by determining how informative it is about the class. The attributes with a score of less than 0.01 should be removed from the data set being analysed [16]. The ranked attribute is shown below in figure 2.

ISSN: 2233-7857 IJFGCN

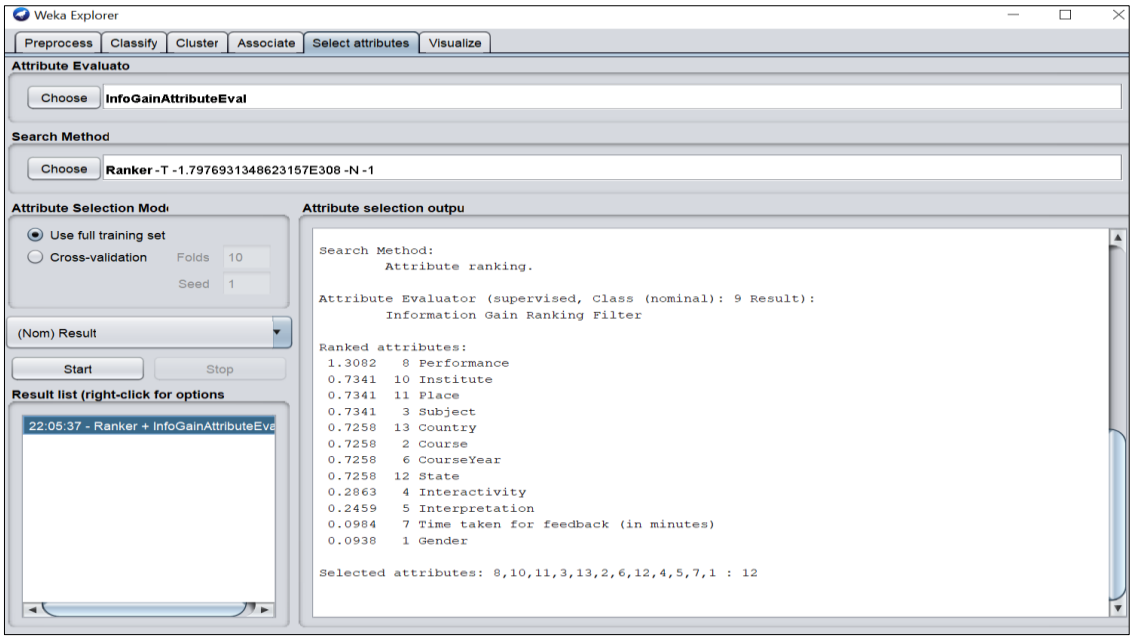

Figure 2. Attribute Ranking based on Gain Ratio

Table 2. Final list of selected attributes with their descriptions

S.No	Attribute Name	Data Type	Description
1	Gender	Nominal	Gender of the candidate
2	Course	Nominal	Course's name
3	Subject	Nominal	Subjects in the courses
4	Interactivity	Numeric	Self-assessment points based on sharing ideas with other peers
5	Interpretation	Numeric	Self-assessment points based on understanding other students messages as well as tutors messages
6	Course Year	Numeric	Year on which course was conducted
7	Time taken for feedback (in minutes)	Numeric	Time taken by student to fill feedback
8	Performance	Numeric	Total score from online exam
9	Result	Nominal	Either pass, fail or not appeared
10	Institute	Nominal	The institute where student is trained
11	Place	Nominal	City from state
12	State	Nominal	State from country
13	Country	Nominal	Students' country

Experiment Analysis:

To experiment with this research, we used classification algorithms J48, Naïve Bayes, REPTree, PART, and JRip. The experiments were done on 182 records and 13 attributes using 10-fold cross-validation. 10-fold Cross-validation divides data into ten sets of size n/10, trains and tests them, and then averages the results.

ISSN: 2233-7857 IJFGCN

Although there are several methods for evaluating models, in this study, the models were evaluated by comparing their accuracy in terms of confusion matrices such as True Positive Rate (TPR), False Positive Rate (FPR), True Negative Rate (TNR), False Negative Rate (FNR), Matthews' Correlation Coefficient (MCC), F-Measure, and the number of correctly classified instances. The number of correct and incorrect predictions made by the model is represented by a confusion matrix.[17]

a) Naive Bayes Classifier

```
Correctly Classified Instances         174               95.6044 %
Incorrectly Classified Instances         8                4.3956 %
Kappa statistic                          0.9242
Mean absolute error                      0.0312
Root mean squared error                  0.1645
Relative absolute error                  8.3307 %
Root relative squared error             38.0874 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 0.918     0.000     1.000       0.918   0.957       0.916    1.000     1.000     Pass
                 1.000     0.048     0.652       1.000   0.789       0.788    0.996     0.962     Fail
                 1.000     0.000     1.000       1.000   1.000       1.000    1.000     1.000     NA
Weighted Avg.    0.956     0.004     0.971       0.956   0.960       0.937    0.999     0.997

=== Confusion Matrix ===

  a  b  c   <-- classified as
 90  8  0 |  a = Pass
  0 15  0 |  b = Fail
  0  0 69 |  c = NA
```

Figure 3. Summary of outputs for Naïve Bayes using 10-fold cross-validation test mode

The experiment done with Naïve Bayes shows an accuracy of 95.60%. Among the three classes of Pass, Fail and Not Appeared students, the model correctly classified 91.8% of Pass student instances are correctly classified as passed students. From the total 182 cases, the model made 174 correct predictions and 8 incorrect predictions and the overall accuracy rate is 0.956.

b) J48 Classifier

```
Correctly Classified Instances         181               99.4505 %
Incorrectly Classified Instances         1                0.5495 %
Kappa statistic                          0.9901
Mean absolute error                      0.0037
Root mean squared error                  0.0605
Relative absolute error                  0.9783 %
Root relative squared error             14.0094 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area
                 1.000     0.012     0.990       1.000   0.995       0.989    0.994     0.990
                 0.933     0.000     1.000       0.933   0.966       0.963    0.967     0.939
                 1.000     0.000     1.000       1.000   1.000       1.000    1.000     1.000
Weighted Avg.    0.995     0.006     0.995       0.995   0.994       0.991    0.994     0.990

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  0  0 |  a = Pass
  1 14  0 |  b = Fail
  0  0 69 |  c = NA
```

Figure 4. Summary of outputs for J48 classifier using 10-fold cross-validation test mode

ISSN: 2233-7857 IJFGCN

The experiment done with the J48 classifier shows an accuracy of 99.45%. Among the three classes of Pass, Fail, and Not Appeared students, the model correctly classified 100% of Pass student instances are correctly classified as passed students. From the total 182 cases, the model made 181 correct predictions and 1 incorrect prediction and the overall accuracy rate is 0.995 which is a promising result.

c) PART Classifier

```
Correctly Classified Instances         181               99.4505 %
Incorrectly Classified Instances         1                0.5495 %
Kappa statistic                          0.9901
Mean absolute error                      0.0037
Root mean squared error                  0.0605
Relative absolute error                  0.9783 %
Root relative squared error             14.0094 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area
                 1.000     0.012     0.990       1.000   0.995       0.989    0.994     0.990
                 0.933     0.000     1.000       0.933   0.966       0.963    0.967     0.939
                 1.000     0.000     1.000       1.000   1.000       1.000    1.000     1.000
Weighted Avg.    0.995     0.006     0.995       0.995   0.994       0.991    0.994     0.990

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  0  0 |  a = Pass
  1 14  0 |  b = Fail
  0  0 69 |  c = NA
```

Figure 5. Summary of outputs for PART classifier using 10-fold cross-validation test mode

The experiment done with the PART classifier shows an accuracy of 99.45%. Among the three classes of Pass, Fail, and Not Appeared students, the model correctly classified 100% of Pass student instances are correctly classified as passed students. From the total 182 cases, the model made 181 correct predictions and 1 incorrect prediction and the overall accuracy rate is 0.995 which is again a promising result.

d) JRip Classifier

```
Correctly Classified Instances         180               98.9011 %
Incorrectly Classified Instances         2                1.0989 %
Kappa statistic                          0.9802
Mean absolute error                      0.0077
Root mean squared error                  0.0856
Relative absolute error                  2.0666 %
Root relative squared error             19.8185 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area
                 1.000     0.024     0.980       1.000   0.990       0.978    0.987     0.978
                 0.867     0.000     1.000       0.867   0.929       0.925    0.929     0.878
                 1.000     0.000     1.000       1.000   1.000       1.000    1.000     1.000
Weighted Avg.    0.989     0.013     0.989       0.989   0.989       0.982    0.987     0.978

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  0  0 |  a = Pass
  2 13  0 |  b = Fail
  0  0 69 |  c = NA
```

Figure 6. Summary of outputs for JRip classifier using 10-fold cross-validation test mode

ISSN: 2233-7857 IJFGCN

The experiment done with the JRip classifier shows an accuracy of 98.90%. Among the three classes of Pass, Fail, and Not Appeared students, the model correctly classified 100% of Pass student instances are correctly classified as passed students. From the total 182 cases, the model made 180 correct predictions and 2 incorrect predictions and the overall accuracy rate is 0.989.

a) REPTree Classifier

```
Correctly Classified Instances         180               98.9011 %
Incorrectly Classified Instances         2                1.0989 %
Kappa statistic                          0.9802
Mean absolute error                      0.0085
Root mean squared error                  0.0857
Relative absolute error                  2.2744 %
Root relative squared error             19.8303 %
Total Number of Instances              182

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area
                 1.000    0.012    0.990      1.000    0.995      0.989    0.992     0.987
                 0.867    0.000    1.000      0.867    0.929      0.925    0.922     0.878
                 1.000    0.009    0.986      1.000    0.993      0.988    0.996     0.986
Weighted Avg.    0.989    0.010    0.989      0.989    0.989      0.984    0.988     0.977

=== Confusion Matrix ===

  a  b  c   <-- classified as
 98  0  0 |  a = Pass
  1 13  1 |  b = Fail
  0  0 69 |  c = NA
```

Figure 7. Summary of outputs for REPTree classifier using 10-fold cross-validation test mode

The experiment done with the REPTree classifier shows an accuracy of 98.90%. Among the three classes of Pass, Fail, and Not Appeared students, the model correctly classified 100% of Pass student instances are correctly classified as passed students. From the total 182 cases, the model made 180 correct predictions and 2 incorrect predictions and the overall accuracy rate is 0.989.

Results and discussion:

After performing the experiments, the next step was comparing the model's performance and selecting the best model. To select the best model for predicting the students' performance J48, Naïve Bayes, REPTree, PART, and JRip classifiers approaches using cross-validation (10-folds) were used. Accuracy, TP Rate, FP Rate, Precision, F-measure, and MCC were used to compare the models.

This study's goals include selecting a better classification technique for building a model that performs best in handling prediction and identifying the students' performance. As a result, five classification models were chosen.

Table 3. Accuracies of Classification Models

Algorithm	**Accuracy (in %)**
Naïve Bayes	95.60
J48	99.45

ISSN: 2233-7857 IJFGCN

REPTree	98.90
PART	99.45
Jrip	98.90

Conclusions:

Early prediction of the students' performance can help in making different and timely managerial decisions at each level based on the necessary feedback to improve the academic performance of students.

The goal of this research was to look into the potential applications of data mining technology in Ethiopian and Indian higher education, specifically on post-graduate student feedback data sets, in order to develop a predictive model that could help students improve their performance during online and activity-based learning.

The original dataset which was 182 instances with 17 numbers of attributes is changed to 13 relevant attributes using the attribute selection method.

The dataset is pre-processed using MS excel and WEKA made a suitable experiment using J48, REPTree, Naïve Bayes, JRip, and PART classifier algorithms for extracting hidden knowledge.

Hence, data mining classifiers, J48, and PART are observed to be best among the selected classifiers with an accuracy of 99.45% which means that the models are successfully predicting the PASS students based on their own feedback.

References:

[1] C. Grant, "The contribution of education to economic growth," 2017.
[2] S. Lijanporn and J. Khlaisang, "ScienceDirect The development of an activity-based learning model using educational mobile application to enhance discipline of elementary school students," *Procedia-Social Behav. Sci.*, vol. 174, pp. 1707–1712, 2015, doi: 10.1016/j.sbspro.2015.01.825.
[3] M. Andergassen *et al.*, "The evolution of e-learning platforms from content to activity based learning: The case of Learn@WU," *Proc. 2015 Int. Conf. Interact. Collab. Learn. ICL 2015*, pp. 779–784, Nov. 2015, doi: 10.1109/ICL.2015.7318127.
[4] M. Lavrenova, N. V. Lalak, and T. I. Molnar, "Preparation of Future Teachers for Use of ICT in Primary School," *Rev. Rom. pentru Educ. Multidimens.*, vol. 12, no. 1Sup1, pp. 185–195, Apr. 2020, doi: 10.18662/RREM/12.1SUP1/230.
[5] T. Mishra, D. Kumar, and S. Gupta, "Mining students' data for prediction performance," *Int. Conf. Adv. Comput. Commun. Technol. ACCT*, pp. 255–262, 2014, doi: 10.1109/ACCT.2014.105.
[6] S. B. Kotsiantis, "Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades," *Artif. Intell. Rev. 2011 374*, vol. 37, no. 4, pp. 331–344, May 2011, doi: 10.1007/S10462-011-9234-X.
[7] M. Tiwari, M. B. Jha, and O. Yadav, "'Performance analysis of Data Mining algorithms in Weka,'" *IOSR J. Comput. Eng.*, vol. 6, no. 3, pp. 32–41, Accessed: Jul. 07, 2021. [Online]. Available: www.iosrjournals.org.

ISSN: 2233-7857 IJFGCN

[8] C. C. Aggarwal, "An Introduction to Data Mining," *Data Min.*, pp. 1–26, 2015, doi: 10.1007/978-3-319-14142-8_1.

[9] N. Hidayat, R. Wardoyo, and S. N. Azhari, "Educational Data Mining (EDM) as a model for students' evaluation in learning environment," *Proc. 3rd Int. Conf. Informatics Comput. ICIC 2018*, Oct. 2018, doi: 10.1109/IAC.2018.8780459.

[10] S. Alturki, I. Hulpuș, · Heiner Stuckenschmidt, H. Stuckenschmidt, and S. Alturki, "Predicting Academic Outcomes: A Survey from 2007 Till 2018," *Technol. Knowl. Learn.*, doi: 10.1007/s10758-020-09476-0.

[11] A. Al-Malaise, A. Malibari, and M. Alkhozae, "STUDENTS' PERFORMANCE PREDICTION SYSTEM USING MULTI AGENT DATA MINING TECHNIQUE," *Int. J. Data Min. Knowl. Manag. Process*, vol. 4, no. 5, 2014, doi: 10.5121/ijdkp.2014.4501.

[12] V. S. Parsania, N. N. Jani, and N. H. Bhalodiya, "Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis," Accessed: Jul. 08, 2021. [Online]. Available: www.darshan.ac.in.

[13] M. Belouch, S. El Hadaj, and M. I. Labsiv, "A Two-Stage Classifier Approach using RepTree Algorithm for Network Intrusion Detection," *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 6, 2017, Accessed: Jul. 07, 2021. [Online]. Available: www.ijacsa.thesai.org.

[14] V. Shaga, S. Sayyad, K. Vengatesan, and A. Kumar, "Fact findings of exploring ICT model in teaching learning," *Int. J. Sci. Technol. Res.*, vol. 8, no. 12, pp. 2051–2054, 2019.

[15] V. Shaga, H. Gebregziabher, and P. Chintal, "Predicting Performance of Students Considering Individual Feedback at Online Learning Using Logistic Regression Model," *Lect. Notes Networks Syst.*, vol. 191, pp. 111–120, 2022, doi: 10.1007/978-981-16-0739-4_11.

[16] E. Osmanbegović, M. Suljić, and H. Agić, "DETERMINING DOMINANT FACTOR FOR STUDENTS PERFORMANCE PREDICTION BY USING DATA MINING CLASSIFICATION ALGORITHMS," *Broj*, vol. 34, no. 34, 2014.

[17] R. Kumari Dash, "Selection Of The Best Classifier From Different Datasets Using WEKA," Accessed: Jul. 07, 2021. [Online]. Available: www.ijert.org.

ISSN: 2233-7857 IJFGCN