International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1953158

Attributes of Low Performing Students In E-Learning System Using Clustering Technique

Ebiemi Allen Ekubo
North-West University, Potchefstroom, South Africa

ABSTRACT

Data mining in education is considered to be one of the relevant and fast growing areas in data mining, with free access to datasets available online, researchers have continued to analyze and produce knowledge which has improved the educational sector. With many research geared towards predicting student results, this paper offers a different approach of gaining knowledge of student data by presenting the attributes of low-performing students. The idea is to group students with low grades and discover the core attributes of these category of students, thereby providing stakeholders with these attributes which should be looked out for in current and prospective students. The dataset used in this research was collected from an e-Learning system called Kalboard 360. The k-means clustering technique embedded in the WEKA tool was used to group these category of students into two clusters. The knowledge gained from the mining process shows that lower-level absentee students with parents that do not actively participate in their learning process are most likely to perform poorly in their studies.

Keywords : Data Mining, E-learning, WEKA, Low-Performing Students.

I. INTRODUCTION

In modern times the need to understand data has become prevalent, with the advent of big data, almost every sector has embraced the idea of gaining knowledge from the set of huge available data to enable them to make better decisions. In Educational data mining, the concern is to develop methods which provides solutions to educational problems that occur within educational settings (Baker, 2010). These solutions when offered are meant to provide all stakeholders with the right information they need to improve the state of the education sector (Romero et al, 2010). A prevailing research area within the education sector is to discover the attributes and causes of low performing students. Many researchers have combined different attributes in order to predict the performance of students. In line with this, this paper aims to discover the attributes of low performing students using an e-learning online dataset, this research would focus on only the students with low grade marks and present the knowledge gained. The k-means algorithm is used in this research to group the low performing students into similar clusters.

With free dataset available online for researchers to use, mainly extracted from online educational sources. This research makes use of the data available at www.kaggle.com/aljarah/xAPI-Edu-Data (Amrieh et al, 2016). In their papers Amrieh et al, made use of this data to predict student's academic performance employing different data mining techniques (Amrieh et al, 2016) and to pre-process and analyse the dataset for improving student's performance (Amrieh et al, 2015). The full description of the dataset is given in the methodology section.

The organization of the remaining parts of this paper are as follows: Section 2 reviews related literature, section 3 presents the proposed methodology, section 4 gives results and discussions on the results obtained from the experiment, section 5 concludes and provides insights for future work.

II. LITERATURE REVIEW

The increase in educational data both in the traditional learning environment and web-based educational environment has generated massive data repositories about teaching and students' information which provides a good opportunity for data mining (Castro et al, 2007; Baker, 2010). Several data mining techniques such as clustering, classification and regression are used to build predictive models, the most widely used technique is classification (Loh, 2011). Classification task aims to predict the label of a class for a set of unlabelled items (Zaki et al, 2014). Classification algorithms build models by learning from a set of trained data and later tested on new sets of data (Han et al, 2012; Zaki et al, 2014). Examples of methods used in classification are Decision Tree Induction, Naïve Bayes classifiers, k-nearest neighbour, Rule-Based classifiers, Neural networks and Support vector machines (Han et al, 2012; Bramer, 2013). In cases where labelling objects cannot be performed easily probably due to the huge amount of data or unknown labels, grouping similar data would be necessary to gain an understanding of the data (Han et al, 2012). Clustering is the process of grouping similar objects that are different from objects in other clusters (Bramer, 2013; Tan et al, 2013). The k-means clustering is a type of clustering technique where individual object is apportioned to exactly one of a set of clusters, it is inherent to start by determining the value of k, that is how many clusters should be formed from the data (Bramer, 2013).

Researchers have used clustering technique to identify learning attributes and behaviours. Bouchet et al. (2012) used clustering technique to distinguish students with prior knowledge on topics, strategies and learning performance and characterized the students as having high or low learning gains. Chan and Bauer (2014) used clustering techniques to identify first-year students at-risk in chemistry by analysing affective variables with results showing that students in the high affective group perform better than those in the low affective group. Ayesha et al. (2010) made use of the k-means clustering technique to predict the learning activities of students form a database with features that includes quizzes and exams scores, the information gained from the research is sent to the class teacher to ensure the number of students that perform poorly are reduced.

Although a lot of research is concerned about dealing with student performance issues, researchers have focused more on predicting students at risk and less on investigating the attributes of students that are already in the low performing category. This research will focus on those attributes and the information gained could serve as a warning sign for administrators to look out for in their scholars.

III. METHODS AND MATERIAL

This paper makes use of the clustering technique to evaluate the attributes of low performing students. The methodology starts by collecting data from the online source, the data collected was in an excel file with a .csv format which is compatible with the Modelling tool used in the mining process. The next step which is the pre-processing step involved cleaning and selecting the dataset using features available in the excel package, this step selected only the students with low level performance for the mining process. The k-means clustering technique is used as the data mining technique to discover the

core attributes of low performing students from the dataset. The next step presents the results gotten from the mining process and the final step of the methodology presents the knowledge gained.

In figure 1 below, the major steps of the proposed methodology are depicted.

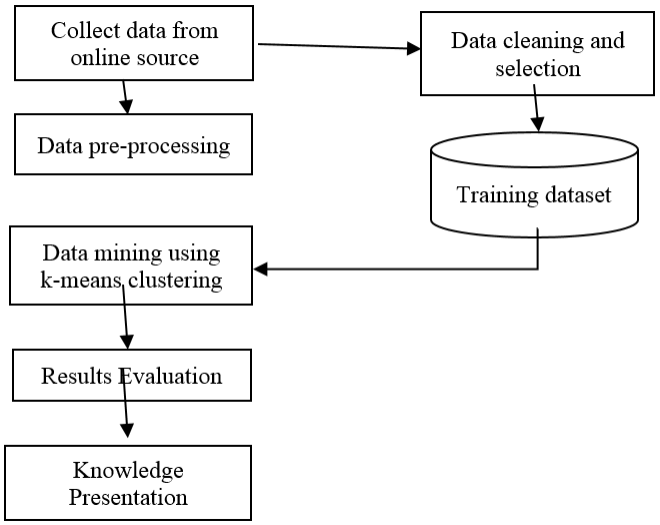

Figure 1. Low performing student model research steps

3.1 Data Selection

The dataset is an educational data which was collected from a learning management system called Kalboard 360, the data was collected using a tracker tool called experience API [10]. The dataset originally consisted of 480 records with 16 features and the class feature which is the total mark of a student, from this dataset 127 records of students classified as Low-Level in their total mark was selected in line with the aim of the research, which is to discover the attributes of low-performing students. The features in the dataset are majorly grouped into three; demographic features such as nationality and gender, academic background features such as grade level and educational stage, the behavioural features such as resources visited and participation in discussion groups, table 1 shows the description of all the features.

Feature Category	Feature	Description
Demographic Features	Nationality	Nationality of the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Gender	The gender of the student (female or male)
	Place of Birth	Place of birth for the student (Jordan, Kuwait, Lebanon, Saudi Arabia, Iran, USA)
	Parent responsible for student	Student's parent (father or mum)
Academic Background Features	Educational Stages	Stage student belongs such as (primary, middle and high school levels)
	Grade Levels	Grade student belongs as (G-01, G-02, G-03, G-04, G-05, G-06, G-07, G-08, G-09, G-10, G-11and G-12)
	Section ID	Classroom student belongs as (A, B, C)
	Semester	School year semester as (First or second)
	Topic	Course topic as (Math, English, IT, Arabic, Science, Quran)
	Student Absence Days	Student absence days (Above-7, Under-7)
Parent Participation in learning process	Parent Answering Survey	Parent is answering the surveys that provided from school or not (Yes, No).
	Parent School Satisfaction	This feature obtains the Degree of parent satisfaction from school as follow (Good, Bad)
Behavioral Features	Discussion Groups	How many times the student participated in discussion groups (numeric:0-100)
	Visited Resources	How many times the student visits a course content (numeric: 0-100)
	Raised hand in class	How many times the student raises his/her

		hand on classroom (numeric: 0-100)
	Viewing Announcements	How many times the student checks the new announcements(numeric:0-100)
Performance Level	Class	Total score (Low-Level: 0 to 69, Middle-Level: 70 to 89, High: interval 90 to 100)

3.2 Data Visualization

Data visualization which is an important part of the preprocessing task ensures the data used is easily understood. The graphical representation of some features are given in the figures below. In Figure 2 the distribution of student gender for the selected dataset is 103 males and 24 females.

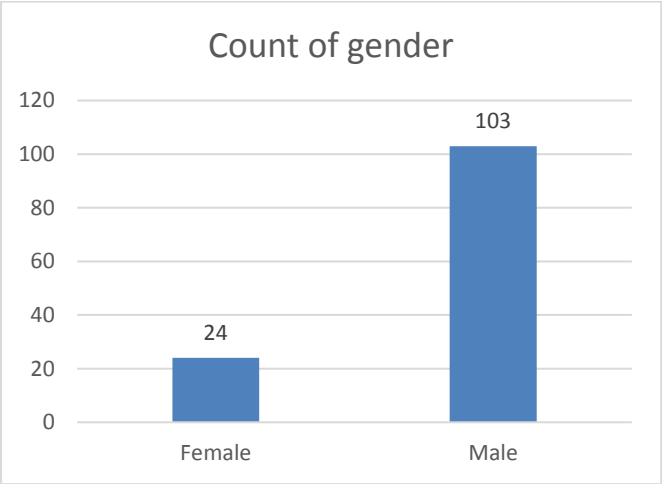

Figure 2: Count of gender

As shown in figure 3 the number of students absent above 7 days are 116 and 11 are absent below 7 days.

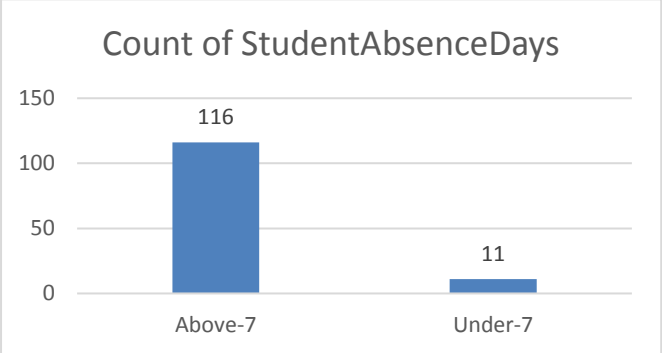

Figure 3: Count of student Absent

Figure 4 depicts the number of parents that answered surveys provided by the school, 99 parents did not answer surveys provided by the school while 28 parents did.

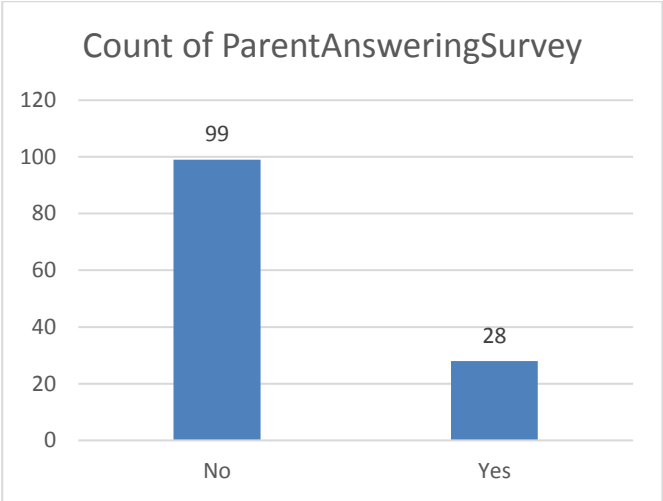

Figure 4: Count of Parents Answering Survey

The features shown graphically were selected to show their impact on students' performance.

IV. RESULTS AND DISCUSSION

WEKA (Waikato Environment for Knowledge Analysis) tool was used to run the experiment in this research. The k-means clustering technique was used to group these students to have a better understanding of their similar features. All the 127 records were used as the training set for this experiment. Figure 5 shows the number of iterations and initial starting point of the cluster.

```
kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 551.5533967761523

Initial starting points (random):

Cluster 0: M,Jordan,Jordan,MiddleSchool,G-07,B,Science,S,Father,52,10,13,6,No,Bad,Above-7
Cluster 1: M,Syria,Jordan,MiddleSchool,G-08,A,Spanish,S,Father,9,7,21,20,Yes,Good,Above-7
```

Figure 5: Number of iterations and initial starting points of cluster

From figure 6, the final cluster centroids are shown with the average value of the features contained in

them. Two clusters were formed with one cluster having 96 instances and the other 31 instances. Similar average value of features was found in both clusters and distinctions are only clear in the StageID, GradeID, ParentAnswering Survey and Parent School Satisfaction features.

Final cluster centroids:

Attribute	Full Data (127.0)	Cluster# 0 (96.0)	1 (31.0)
gender	M	M	M
NationalITy	KW	KW	KW
PlaceofBirth	KuwaIT	KuwaIT	KuwaIT
StageID	lowerlevel	lowerlevel	MiddleSchool
GradeID	G-02	G-02	G-08
SectionID	A	A	A
Topic	IT	IT	IT
Semester	F	F	F
Relation	Father	Father	Father
raisedhands	16.8898	16.875	16.9355
VisITedResources	18.3228	15.6458	26.6129
AnnouncementsView	15.5748	14.0521	20.2903
Discussion	30.8346	32.2396	26.4839
ParentAnsweringSurvey	No	No	Yes
ParentschoolSatisfaction	Bad	Bad	Good
StudentAbsenceDays	Above-7	Above-7	Above-7

Figure 6: Final cluster centroids

Figure 7 shows the percentage of instances gotten for both clusters found in the dataset. Cluster 0 with 96 instances has 76% of the total instances, while cluster 1 with 31 instances has 24%.

```
=== Model and evaluation on training set ===

Clustered Instances

0       96 ( 76%)
1       31 ( 24%)
```

Figure 7: Clustered instances

V. CONCLUSION

In this paper, we have been able to present the attributes of low-performing students from features collected from an e-Learning system called Kalboard 360. The dataset collected which originally had 480 records where reduced to 127 records containing students classified as low performers. The k-means clustering technique embedded in the WEKA tool was used to group these category of students. Two clusters of students where identified with many similarities such as behavioral and demographic features. Both clusters had average of student absent days to be above 7 days. The first cluster of students contains 96 instances (making up 76% of the dataset) with lower averages on behavioral features, it is important to note the parents' participation in the learning process of these group of students, their parents did not take part in surveys and their satisfaction with the school was rated bad, these students also fall into the lower level stage in their educational stage.

The second cluster of students contains 31 instances (making up 24% of the dataset) with slightly higher averages on behavioral features, these cluster of students have parents that take part in surveys and their satisfaction with the school was rated good, these students fall into the middle level stage in their educational stage.

From the information gained from the clusters, we can conclude that lower level absentee students with parents that do not actively participate in their learning process are most likely to perform poorly in their studies. Stakeholders should therefore look out for these attributes in new and prospective students. Since this assumption is based on the limited dataset used in this research, we recommend the use of a larger dataset in combination with other data mining techniques. Future research activity would include analyzing data with low performing students further partitioned to enable a predictive model to be built that would provide stakeholders with information for better decision making.

VI. REFERENCES

[1]. Amrieh, E.A. et al, 2016. www.kaggle.com/aljarah/xAPI-Edu-Data.

[2]. Amrieh, E.A. et al, 2016. Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, Vol. 9, No.8, pp. 119-136.

[3]. Amrieh, E.A. et al, 2015. Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), pp. 1-5.

[4]. Ayesha, S. et al, 2010. "Data Mining Model for Higher Education System", European Journal of Scientific Research, vol. 43, no. 1, pp. 24-29, 2010.

[5]. Baker, R.S., 2010. Data mining for education. International encyclopedia of education, Vol. 7, No. 3, pp.112-118.

[6]. Bouchet, F. et al, 2012. Identifying Students' Characteristic Learning Behaviors in an Intelligent Tutoring System Fostering Self-Regulated Learning. International Educational Data Mining Society, pp. 65-72.

[7]. Bramer, M., 2013. Principles of data mining (Vol. 180). London.: Springer.

[8]. Castro, F. et al., 2007. Applying data mining techniques to e-learning problems. In Evolution of teaching and learning paradigms in intelligent environment, Springer Berlin Heidelberg, pp. 183-221.

[9]. Chan, J. Y. and Bauer, C.F. 2014. Identifying at-risk students in general chemistry via cluster analysis of affective characteristics. Journal of chemical education, vol. 91 no. 9, pp.1417-1425.

[10]. Han, J. et al, 2012. Data mining: concepts and techniques. Elsevier.

[11]. Loh, W.Y., 2011. "Classification and regression trees," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 1, no. 1, pp.14-23.

[12]. Romero, C.B. et al. eds., 2010. Handbook of educational data mining. CRC Press.

[13]. Tan, P. N. et al, 2013. Introduction to Data Mining. Pearson Education Limited.

[14]. Zaki, M.J. et al, 2014. Data mining and analysis: fundamental concepts and algorithms. Cambridge University Press.

Cite this article as :

Ebiemi Allen Ekubo, "Attributes of Low Performing Students In E-Learning System Using Clustering Technique", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 3, pp. 480-485, May-June 2019.
Available at doi : https://doi.org/10.32628/CSEIT1953158
Journal URL : http://ijsrcseit.com/CSEIT1953158