ISSN No:-2456-2165

Application of Data Mining to Predict Students Learning Behavior: A Case Study of Kebbi State Polytechnic Dakin-gari

Suleiman Sahabi[1*], Anas Shehu[2*], Shamsu Sani[2], Abubakar Sani[3],
[1] Department of Sciences Kebbi State Polytechnic Dakin-gari, Dakin-gari, 862106, Nigeria.
[2] Department of Computer Science Kebbi State Polytechnic Dakin-gari, Dakin-gari, 862106, Nigeria.
[3] Department of Computer Science Yusuf Maitama Sule University, Kano, 700214, Nigeria.

Abstract:- This research was conducted in data mining. To our knowledge, no research covered seven programs of higher institutions for data mining purposes. This work used real dataset of students from seven department of Kebbi State Polytechnic Dakin-gari. Classification, association and clustering were used to discover hidden patterns in the dataset. WEKA workbench was used to run the experiment and evaluate the results. Classification was done with optimum accuracy where four classes were identified i.e. weaker, weak, good and better students. After association rule was done, the authors found out strong correlation between (ATT and CA) with (GPA), (EF and CA) with (GPA) and (EF and CA) with (CO). And that affect the students in their GPA results. Same test data was used for hierarchical clustering where two clusters were returned and 162 tuples was distributed between the clusters in 81% and 19% fashion, Conclusively, the authors strongly feel that the management of Kebbi State Polytechnic Dakin-gari with a matter of urgency need to tackle these discovered pattern to minimize rate of failure and drop out within the students.

Keywords:- *WEKA, Data Mining Tool, Dakin-Gari, Clustering, Association, Classification, Metric.*

I. INTRODUCTION

Modern technology allows students and staff of the universities, polytechnics and colleges to have a cohesive collaboration than ever before. This allows more data collection in the long run that also involve students submitting their academic and non-academic—personal data to school involved. This provides the foundations of data mining for pattern and rule discovery in education domain. Data Mining (DM) is a technique where extraction is done on data in order to discover hidden, relevant and useful information to users [1]. Application of data mining cut across almost all industry and professional areas, educational sector inclusive. Applying DM concepts and techniques on educational data is called educational data mining (EDM). EDM as emerging discipline [2] is imperative to allow school managements to predict student's performance of success or failure. However, understanding and applying it in school allow the management to better manage resources [3]. Important it is for schools to assist in providing good education to students which will make them employable as well as earning reputation for the concern institution [4]. On the Students part, they may wish to know how best they can perform in each particular subject based on prediction [1], [5] which by extension, reduces insecurity in state as these students will be counseled properly. Thus, educational data mining is ultimately aimed at solving educational issues as these data contain useful information that need to be mine for good decision-making. DM has its own tasks ranging from descriptive such as association, clustering, classification etc., or predictive that uses rule set such as decision tree, neural networks, and support vector. The rules to be working on are: Association, Classification, Clustering, and outlier

➢ *Research Contributions*

The contributions of our research work are as follows:

- To determine and present the total number of all National Diploma I students of Kebbi State Polytechnic Dakin-gari.
- To apply students data set for data mining techniques.
- To obtain the mined results, make evaluation and predict student's academic performance

II. METHODOLOGY

This section present the methodology used in the process of research study. Brief explanation of data mining techniques, dataset collection and used, WEKA toolbox used, experimental setup and the results discussion are part of this section. Figure 1 below depicts the whole research activity diagram.

ISSN No:-2456-2165

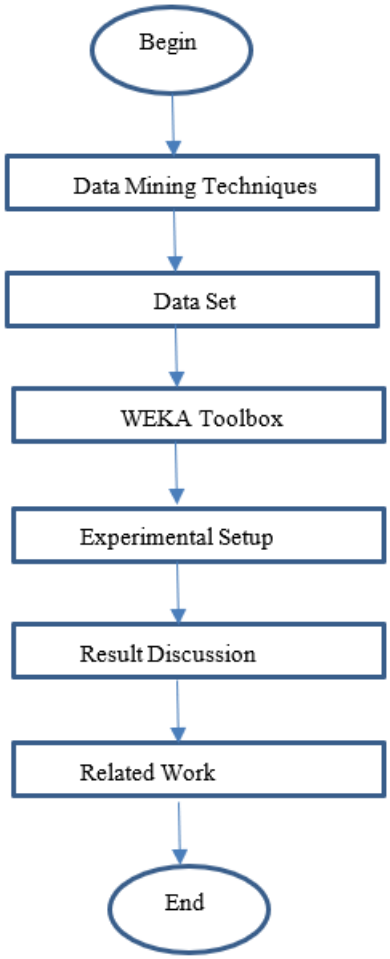

Fig 1. Activity Diagram

III. DATA MINING TECHNIQUES

In educational data mining, association rule, classification and clustering are the most commonplace technique [6]. Thus, as fa as our work is concern, we choose to give an overview of them only.

- Association: this technique is used in data mining where the need arises where hidden relationship between data items need to be exposed. The steps in association rules is twofold. Minimum support is used to get frequent item sets in data and lastly minimum confidence constraint is used on these frequent item sets to develop rules. There are data mining algorithms that supports association rule (see [7])
- Classification: called supervised learning. Is used to form groups and classes with the help of classifier or model by constructing the model and implement the model usage. In the model build up, each sample represents one particular class in the training data set. For example, providing students score in a particular course, one may predict his final results within say three classes (Excellent, Good, Fail). As association, classification also supports various algorithms refer [8], [6], [9] and [10] for details. Depending on the choice of algorithm used, way of finding out relationship differs.
- Clustering: called cluster analysis or unsupervised learning. The technique identify group of objects that have similar attributes, trend or pattern which all are different from other group. For detail about clustering refer [3].

➢ *Dataset*

The dataset to be used in our work is students' dataset collected from seven department and other units from Kebbi State Polytechnic Dakingari. It contained 172 tuples which, after data cleanup became 162 records only. Tables 1 below shows an overview of the datasets.

Table 1. Dataset

Dataset	Attributes	Records	Size (KB)
Students	7	162	2

➢ *WEKA Toolbox*

WEKA is a popular open source software used for data mining purposes with JAVA implemented algorithms. It was programmed to assist users to speed up testing machine learning in a dynamic fashion [11]. The workbench contains a bunch of visualization tools which is portable and can work on many computer platforms. It supports disparate data mining task ranging from data preprocessing, clustering, classification, association, visualization, DataSinks and feature selection. This workbench supports different data file formats such as .arff, .csv among others. It was developed in New Zealand at University of Waikato. For in depth WEKA understanding refer [12]

➢ *Experimental Setup*

In this work, the experiments were conducted on a laptop computer running 64-bit Windows 8 (6.2, Build 9200) with AMD E-300 CPU with Radeon (TM) processor at 1.3GHZ clock speed with 4 GB RAM. As for the data mining tasks, this study use WEKA toolbox, explained above. However, all the tasks understudy and evaluation metrics are implemented in the toolbox.

IV. RESULTS AND DISCUSSION

This section will analyze the results obtained in running the mining tasks. Also, some evaluation and error measuring techniques will be used in the process of analyses, i.e. confusion matrix, True Positive or Recall, Precision, F-score, Accuracy, Mean Absolute Error, Root Mean Square Error, Relative Absolute Error, Root Relative Squared Error.

➢ *Confusion Matrix. I*

t is also called contingency table. Performance measures are deduce from the matrix which produce the output of four classification. The outputs are 1. True Positive (Right positive prediction) 2. True Negative (Right negative prediction) 3. False Positive (Wrong positive prediction) 4. False Negative (Wrong negative prediction).

ISSN No:-2456-2165

- *True Positive or Recall.*

Here recall takes the number of right classification and divide it by the total number of positive.

Thus, R = TP/ (TP + FN) = TP/P

- *Precision.*

Takes the number of right positive classification and divide it by total number of positive classifications.

Thus, P = TP/ (TP + FP)

- *F-score.*

Is the harmonic mean of precision and true positive.

Thus, F = 2PR / (P + R)

- *Accuracy.*

Takes the number of all right classification and divide it by all classes.

Thus, Accuracy = (TP + TN)/ (TP + TN + FN + FP)= (TP+TN)/ (P+N)

Mean Absolute Error. Takes the estimation on how far the prediction is different from the actual values.
Root Mean Square Error. It evaluates the differences between the predictor values and the actual observed values.
Relative Absolute Error. Is the ratio of absolute error by the magnitude of the actual value?
Root Relative Squared Error. Takes the mean absolute error and divide it by the classification model error.

Table 2. Dataset description

Attribute	Description	Values
SE	Sex	(Male, Female)
SH	Study Hours	(Good, Average, Poor)
ATT	Class Attendance Percentage	If percentage>=75 then Good, if percentage >= 51 but < 75 then Average, if percentage <50 then poor
GPA	Grade Point Average	If GPA >= 0.00 but < 1.00 then Weaker, if GPA >= 1.00 but < 1.5 then Weak, if GPA >=1.5 but <2.4 then Average, if GPA >= 2.4 but < 3.50 then Good, if GPA >= 3.50 but < 4.50 then Better, if GPA >=4.50 then Best
CA	Continuous Assessment	(Best, Better, Good, Average, Poor)
CO	Carry Over	(Yes, No)
EF	English Fluency	(Yes, No)

The WEKA feature selection found attributes that have more influence with the help of correlation-based attributes evaluation, gain-ratio attribute evaluation, information-gain attribute evaluation, and relief attribute evaluation, symmetrical uncertainty attribute evaluation. These attributes were used by WEKA for our classification. J48 classifier was used for generating decision tree with 10 fold cross validation and the accuracy of the classification is 73%. Table 3 below shows the statistical accuracy of the classification.

Table 3. Statistical Accuracy

TP Rate	FP Rate	Precision	Recall	F-score	ROC Area	Class
0.919	0.563	0.527	0.919	0.67	0.698	Weaker
0	0	0	0	0	0.657	Weak
0.943	0.178	0.485	0.943	0.641	0.91	Good
0	0	0	0	0	0.758	Better

From the table above, we can observe that ROC Area for all classes are greater than 0.6. That means classification process succeeded in training the set and the results returned by the algorithm were relevant and needed.

To find an association using our dataset to uncover the frequent If/Then patterns, Apriori Algorithm was used to find the correlation between the attributes. Minimum support was 0.6 (with 90 instances) and minimum confidence was 0.9 and the number of cycles were 4

Table 4. Sample of Typical Rules

		Support	Confidence	Lift
(ATT = poor, CA = poor)	(GPA >=0.00 < 1.00)	0.16	0.800	3.333
(ATT = Good, CA = Best)	(GPA >= 4.50)	0.15	0.750	3.125
(EF = No, CA = poor)	(GPA >=0.00 < 1.00)	0.14	0.700	3.333
(EF = Yes, CA = Good)	(GPA >+ 2.40 < 3.50)	0.14	0.682	3.099
(CA = poor, EF = No)	(CO = Yes)	0.13	0.66	2.778

From the above table, we can understand the existence of strong correlation between (ATT and CA) with (GPA), (EF and CA) with (GPA). There is also another correlation between (EF and CA) with (CO).

ISSN No:-2456-2165

Table 5. Sample of cluster statistics

Algorithm	No. of Clusters	Cluster instances
Hierarchical	2	132 (81%)
		30 (19)

From the table above, it is observed that hierarchical clustering algorithm was the candidate for clustering purposes. Two clusters were formed Out of 162 students. Cluster 0 contains highest percentage and cluster 1 the lowest share.

V. RELATED WORK

Hadzagic M. et al (2013) different pattern discovery techniques were explained in detail. The techniques are rule pattern identification, data clustering, categorization and singularity. Though the bedrock of their work was the used of R to apply data mining tasks in maritime traffic analysis with automatic identification system (AIS) data sets. In [7] mining was done on 151 students data of Islamic University of Gaza where the main four techniques—descriptive tools were used. The outcome of the work identified those that provide accurate prediction regarding students' performance. The work is more of comparative analyses of the tasks. On the other hand, [4] used classification algorithms in their prediction and made complex comparison where rule based algorithm outperformed Decision Tree and N Bayes. However, the work was conducted on 8 years level 100 student's data. Comparative analyses between two algorithms, J48 and Random tree were conducted to ascertain better performance in prediction where J48 came out superior in [7]. Data sets of 90 students were also used in [4] where classifier algorithms such as ID3, C4.5 & CART were used to ascertained students' performance. Eventually, ID3 came out victorious in the area of accuracy. In the work of [2], concept of data mining was provided plus knowledge discovery in data processes. In their work also step by step explanation of techniques were presented. Interestingly, the work used WEKA tool to experiment the data set of undergraduate students using four courses. ZeroR algorithm and DBSCAN were used for classification and grouping the students. In the work of [13] compare six data mining tools were compared in order to have global idea of their behaviors, functions and working environment. The authors reported that only two among the six have full functionalities to carrying out data mining tasks. But the research study presented by [3] used machine learning algorithms such as Support Vector Machine, ANN and Random Forest to know what student's behavior, performance and others impact his choice of profession Out of these three, SVM has 94% accuracy, followed by ANN with 84% and lastly random forest with the score of 83%. On the other hand the authors reported the error rate to be 17%, 16% and 6% for random forest, ANN and SVM respectively. In the work of [13] LDA and pLSA were used extensively to predict students results for each lecture that was conducted using student's comment data sets. In the work. F-accuracy said a lot of differences between the classes of the students. Also, the research showed that class C comments provided with highest prediction.

Four clustering algorithms such as k-Means, k-Medoids, Fuzzy C Means and Expectation Maximization using students data sets were compared based on three factors—purity, normalized mutual information & execution time in a study conducted by [14]. The data sets is for 1531 students and came from two schools. Execution time of k-Medoids is lesser than all and, EM has highest. For purity value, again EM has largest then followed by FCM. While k-means has largest NMI value. Another research work conducted by [8], data mining and machine learning algorithms were practically compared with regards to functionality. The work used inbuilt weather data present in WEKA tool. Their experiment indicated that PRISM had higher result accuracy then, followed by SVM and lastly IBK than other classifiers. It was also reported that Naïve Bayes had higher FP than all which the authors tagged it as "not good" classifier. Apriori algorithm was proposed in [7] using R to mine the association rules of thirty four subjects of one hundred students of computer science. Strong correlations were found between English related courses. Again, strong correlation between technical courses such as Digit logic and database principle, interface and compiling principle among others were also found. Application of logistic model trees in WEKA environment to select the right course of study and predict the future outcome of the examinations results of the student was proposed in [15]. Five experiments were ran with accuracy 0f 79%, 52%, 81.5%, 60.4% and 83.2% respectively. Again, the authors used evaluation metrics and compared the proposed work with random forest and j48 algorithms. In another study presented by [6], related work in the EDM was prepared and presented. The work highlighted three most prevalent EDM techniques such as association, classification and clustering. Few examples of algorithms they supported were explained. Three most prominent EDM techniques were discussed in the work of [9] with some examples of supported algorithms. However, the bedrock of the study was to classify and predict student's results and job placement using J48 classification algorithm. The authors experiment were conducted in a WEKA environment. Seven algorithms were thoroughly compared with three different student data set for best prediction performance. In the end, results indicated that Random forest and C5.0 outperformed J48, CART, NG, KNN, and SVM. In a similar work by [4], seven algorithms were compared with three different data sets in predicting student's performance. In the same work, Pearson correlation was used and the conclusion was C5.0 and J48 outclassed remaining five algorithms in terms of prediction accuracy. Two algorithms namely, J48 and Random Tree were applied on MCA student's dataset for performance comparison using in the work of [1]. Random Tree was found to be more accurate in the study. The work concluded that second semester result have influence on third semester student's results. In another similar study conducted by [10], J48, PART, Random Forest and Bayes Network classifiers were measured against each other using level 300 student's data sets and WEKA toolset. The researchers concluded that Random Forest was the best in terms of accuracy and classification error.

Based on these literatures, we can confirm that none of the work have explored more than 3 program in the process of mining the students' data. Therefore, there is need to conduct mining application on different level of students data sets for better performance and prediction. As such, we intend to cover seven programs (computer science, business administration, public administration, electrical engineering, library and information science, laboratory science and fishery) with seven different data sets from their departments in Kebbi State Polytechnic Dakin-gari, Nigeria.

VI. CONCLUSION

In this research, it could be concluded that the test data provide desired results. Classification was done with optimum accuracy where four classes were identified i.e. weaker, weak, good and better students with the help of WEKA. Association was also done and strong correlation that exist between (ATT and CA) with (GPA), (EF and CA) with (GPA) and (EF and CA) with (CO). That affect the students in their GPA results. Additionally, hierarchical clustering algorithm was used in the work and two clusters were returned and 162 was distributed between the clusters in 81% and 19%, Conclusively, the authors feel that the management of Kebbi State Polytechnic Dakin-gari need to with a matter of urgency need to tackle these discovered pattern to minimize rate of failure and drop out within the students

REFERENCES

[1]. Mishra T., Kumar D. and Gupta S., "Mining Students' Data For Performance Prediction," in Fourth International Conference on Advanced Computing & Communication Techniques, 2014.
[2]. Aher S.B. and Lobo L.M.R.J., "Data Mining in Educational System Using WEKA," in International Conference on Emerging Technology Trends, 2011.
[3]. Arcinas M.M, Sajja G.S., Asif S., Gour S., Okoronkwo E. and Naved M., "Role of Data Mining in Education For Improving Students Performance For Social Change," Turkish Journal of Physiotherapy and Rehabilitation, vol. 32, no. 3, pp. 6519-6526, 2021.
[4]. Sathe M.T. and Adamuthe A.C., "Comparative Study of Supervised Algorithms For Prediction of Student's Performance," I.J. Modern Education and Computer Science, pp. 1-21, 08 February 2021.
[5]. Ma Y., Liu C., Wong C.K., Yu P.S. Lee S.M., "Targeting The Right Students Using Data Mining," in KDD, Boston, MA USA, 2000.
[6]. Saleh M.A., Palaniapan S. and Abdalla N.A., "Education is an Overview of Data Mining and The Ability to Predict The Performance of Students," UNNES, vol. 15, no. 1, pp. 19-28, 2021.
[7]. Wu X. and Zeng Y., "Using Apriori Algorithm on Student's Performance Data For Association Rules Mining," in 2nd International Seminar on Education Research on Social Science, 2019.
[8]. Kumar D. and Suman, "Performancae Analysis of Various Data Mining Algorithms: A Review," International Journal of Computer Applications, vol. 32, no. 6, pp. 9-15, October 2011.
[9]. Sumathi K., Kannan S. and Nagarajan K., "Data Mining: Analysis of Student Database Using Classification Techniques," International Journal of Computer Applications, vol. 141, no. 8, pp. 22-27, May 2016.
[10]. Hussain S., Dahan N.A., Ba-Alwib F.M. and Ribata N., "Educational Data Mining and Analysis of Students' Academic Performance Using WEKA," Indonesian Journal of Electrical Engineering and Computer Science, vol. 9, no. 2, pp. 447-459, February 2018.
[11]. Ali F.M.N. and Hamed A.A.M., "Usage Apriori and Clustering Algorithms in WEKA Tools to Mining Dataset of Traffic Accidents," Journal of Information and Telecommunication, vol. 2, no. 3, pp. 231-245, 2018.
[12]. Bouckaert R.R., Frank E., Hall M., Kirkby R., Reutemann P., Seewald A. and Scuse D., "WEKA Manual For Version 3-7-8," 2013.
[13]. Govindasamy K. and Velmurugan T., "Analysis of Student Acafemic Performance Using Clustering Techniques," International JoApplied Mathematicsurnal of Pure and, vol. 119, no. 15, pp. 309-323, 2018.
[14]. Sorout S.E., Goda K. and Mine T., "Comment Data Mining to Estimate Student Performance Considering Consecutive Lessons," Journal of Educational Technology & Society, vol. 20, no. 1, pp. 73-86, 17 May 2016.
[15]. R. A. A. R. a. I. F. Aman F., "A Predictive Model for Predicting Students Academic Performance," IEEE, pp. 1-4, 2019.