Comparative Analysis of Cluster and Random Tree Algorithm Based on Students Performance Outcome

Baba Saleh Ahmed[1], Ali Baba Dauda[2,], Mohammed Shettima[3]
[1]Mai Idris Alooma Polytechnic, Geidam
[2]University of Maiduguri, Nigeria
[3]Mai Idris Alooma Polytechnic, Geidam

ali.dauda@nmimaid.edu.ng[2]
ms.albe89@miapoly.edu.ng[3]

Corresponding Author: fusam20@gmail.com +2348020514430

Abstract
Student performance is the major tool use in determining the status of students at all levels of education. Data mining tools are nowadays used in determining the students' performance and it is greatly helps in making analysis and decision based on the performance. This paper aims at comparing Cluster algorithm and Random/Decision Tree algorithm. Rapid miner studio is used to determine the best algorithm to determining student performance. We conducted this work in Mai Idris Alooma Polytechnic, Geidam Yobe State, where student score sheets that contains six (6) attributes and 669 tuples was used as the dataset for this work. One attribute was selected as a label attribute that determine the Student performance in case of supervised learning, while on the other hand average within centroid distance of all the clusters is measured to see closeness within performance of students. Student grade was used for determining performance of students as label attribute. The findings show that Random Tree algorithm has a higher class precision with an accuracy of about 73.73% compared to that of measures of average within centroid distance. The findings will equally help in marking a sound academic planning in future. Finally, the analysis of the results obtained will go a long way in making recommendations for future work.
Key words: *Cluster, Random Tree, Decision Tree, Algorithm, Data Mining, MIAP.*

Introduction

The population of our students in Nigerian polytechnics is growing higher beyond expectations. Students from different background will come together under the same umbrella to learn various concepts of knowledge. Their performance varies on different fields of knowledge, therefore, the biggest challenge faced is the analysis of their performance. The institutions will be desperate to know the overall performance of their students so that they can make appropriate academic planning ahead of future times to come. Data mining is the process of extracting patterns, thereby evaluating these patterns to obtain knowledge from a huge data. The need for Data mining has arisen since the size of the data is widening and there is a need for transforming such data into a useful information, as such Data mining tools with embedded statistical algorithms will be used to determine the student performance in the Mai Idris Alooma Polytechnic (MIAP), Geidam (Prasada & Reo M. C. 2016).

Amudha, (2016), in his work, Data mining is very good in making decision because data has to undergo Cleaning and Integration, then Data Selection and retrieval of the relevant data, then the data will be transform for Mining, then the Mining will give a Pattern for evaluations, then the last stage is knowledge discovery with the help of the patterns obtained earlier.

This research will be conducted in school of science MIAP, data will be collected from the various department of the school of science which will be gathered and prepared for the comparison of different algorithm in the data mining tool which will be used in the analysis of data collected. Furthermore, it will be used for the prediction of the best algorithm that determining the best performance of the students in school of Science and as well the entire polytechnic.

Statement of the problem

Many researches have conducted on the issues of student performance in order to know the actual performance of the students of a particular institution, which are conducted on the concepts of student CGPA or the aggregates marks obtained by each student. These mention tools CGPA and Marks aggregate can't give the actual performance students based on the following reasons:

A student with CGPA 3.5 will be on the same class with other student with CGPA 3.99 or 4.0.

A student with the highest mark aggregate in one institution will return the lowest class position in the same type of sister institution.

Therefore, based on the above reasons this research work will used Student Grade as the means of determining student performance.

Dataset for this work

The dataset for this work will be collected from all the departments of school of science MIAP, which will be selected and gathered in the same file and data mining algorithm will be design for such data to extract the actual performance of the student. 669 datasets will be used for this work.

Related Literature

Tools for Data Mining

The research work will have used *Rapid Miner* as a tool for extraction of data from the data set given. Different researches have been conducted on the comparison of algorithms in rapid miner using supervised learning technics, where the dataset has been trained and a label is assign on trained datasets to see the performance of the model and the accuracy of the results obtained (Richard et al., 2007). While on the other hand unsupervised learning technics is also applied where the trained datasets will not be assign any label, a Deep learning model is adopted without an explicit instruction on what the model should do on the training datasets, the model is allowed to find features and pattern for prediction (Bradly, 2002).

Data mining tools like **rapid miner** is a very useful tool for knowledge discovery in database. Knowledge discovery and data mining are synonymously inter-related but most time people are confusing about the concepts. Data mining is also a step to knowledge discovery in database (Gurmeet, & Williamjit, 2016), uses various data Mining Techniques both classification and non-classifiers to extract data from student's percentage (%) marks obtained by students and compared to obtained the best data mining tool to determined student performance. Decision tree J48 yields a better result compared to Naïve Bayesian technique in terms of accuracy (Gurmeet, & Williamjit, 2016).

Classification Techniques

Decision Tree

Richard et al., (2007). In some Random forest classification applications with the presence and missing of some values in the dataset, a high classification accuracy is recorded with cross validation and in case of lichen data by using an independent test set.

Random Forest is a supervised learning algorithm that is used for either classification or regression analysis.

But it is largely used for classification problems (Tutorialspoints, 2020).

Nunung & Yudho (2014), one of the popular data mining classifier known today is a *Decision Tree*. Decision tree is popular because of its simplicity in the interpretation of results of datasets. It is tree-like in structure having one root node and branches best on the interpretation of the trained datasets. Each branch is an interpretation of a particular test on attribute. And finally the leaf node represents a classes of distribution.

Obadiah et al (2019), mentioned that decision tree algorithm requires large number of dataset to be train and predict patterns for evaluation and to obtain knowledge from it.

Clustering Techniques

Cluster algorithm can be categorized by the type of data to be analyzed. The similarity of the dataset is used to group the data, while on the other hand the theory is defining the number of cluster in a particular dataset. Most common cluster algorithm used are found in various data mining tools such as Rapid miner and Weka tools (Anuradha, et al, 2015). Subsequently, clustering adopts various techniques like:

Portioning Method: this results in a set of clusters where each object belongs to one cluster and each cluster must contain at least one object (Anuradha, et al, 2015).

Hierarchical methods: this is a type of clustering technique that group data objects into a hierarchy of tree-like structure (Berking, 2006).

Others are: Density-Based Method, Grid-Based Method, Model-Based Method and Constraint-Based Method (Amudha, 2016).

Research Design

This research will be conducted using Rapid Miner studio. Student score sheet is the major source of dataset for this analysis of the student performance in school of science MIAP. There are three different departments in the faculty. All the three departments will provide us with the necessary score sheets for two consecutive years. All the score sheets gathered will be treated on the same file in order to obtain a single set of data to make analysis of the performance of their results.

Classification algorithm will be tested first using *random forest/decision tree algorithm*. Then later the cluster algorithm will be applied on the same set of data using *K-mean algorithm* where results will be collected and compared for the actual analysis (Nunung, & Yudho, 2014). Below are the steps followed to achieved the experiments:

1 *Preparing Data:* The dataset for this work will be prepared and keep in base where it will be fetch for analysis (Nunung, & Yudho, 2014).

2 *Data Cleaning:* This stage is checking for corrupt data, inaccurate data and irrelevant data. The data has to be in a consistent state, inconsistencies should be removed (Amudha, 2016).

3 *Data Integration:* The data collected for this work will be put in excel format, which comprises the six (6) attributes with grade as the label attribute (Nunung, & Yudho, 2014).

4 *Data Selection and Transformation:* The data relevant to tihs work will be selected by the data mining tool, and the data will be transform into normal form appropriate for as required by the procedure (Usama, et-al 1996).

5 *Data Mining:* The process of discovering knowledge from the patterns obtained which is extracted from the given dataset with the help of machine learning and some statistical algorithms, which gives an information for further analysis (wikipedia.org, n.d.).

6 *Evaluation and Presentation:* After the extraction of the data some valuable patterns are obtained for evaluation. This will help in obtaining the knowledge for evaluations and further decision making analysis (Usama, et-al 1996). The figure below shows the structure of the knowledge discovery:

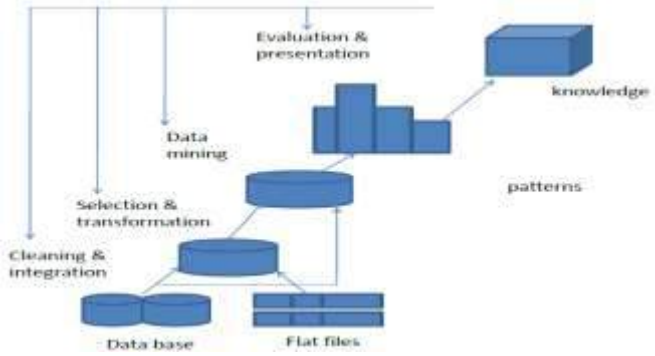

Fig 1: *Knowledge discovery in database (KDD)* (Amudha, 2016).

Methodology

The methodology for this work follows the above steps in the research design. Two different algorithm will be design one for each decision tree algorithm under a supervised learning, and on the other hand k-mean algorithm will be design to reflect the unsupervised learning, and in both the performance of the algorithm

will be tested to see the effectiveness of each algorithm. Dataset will be loaded in either .xls format or in .cvs format. The figure below shows the research methodology design:

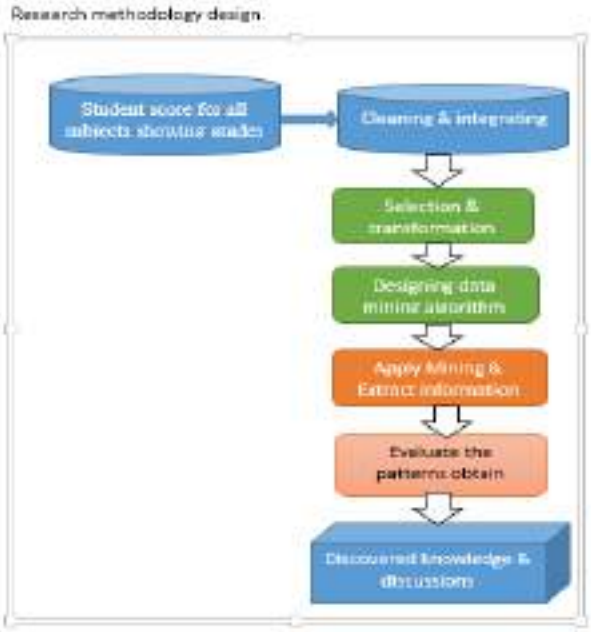

Fig 2: research methodology design.

Data Mining Process

The student score sheets for two years is sampled from all the three departments in the schools, in which 669 student's results were collected, trained and tested using a Random Tree algorithm to obtain the following results:

Experimental design for classification algorithm

The student score sheet comprises different attributes in which some of the attributes were excluded like the student name which cannot be easily classified. Read excel operator is used to load the data in Microsoft Excel format. The filter operator is used to filter the information. A Decision tree algorithm is used as a classifier for this experiment and a model is trained to obtain the performance of the algorithm.

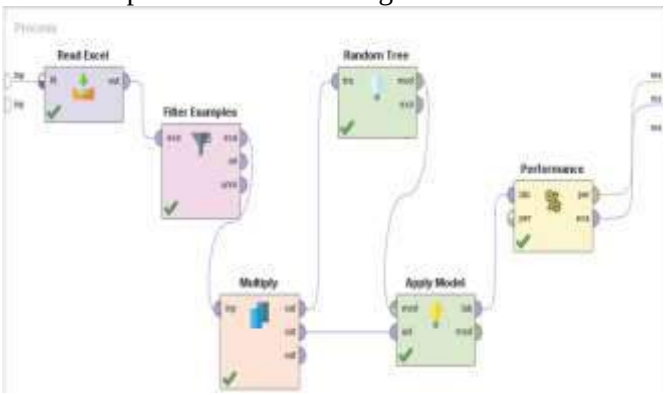

Fig 3: *Decision Tree model and the Performance of the model*

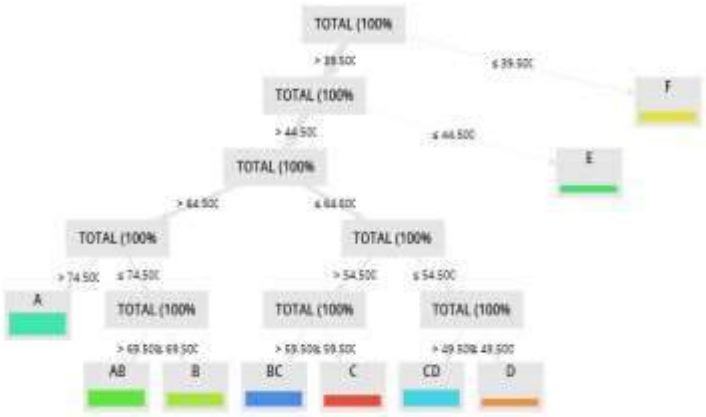

Fig 4: *Tree obtained from the decision tree algorithm model*

The above decision tree is showing the example data set classified, the leaves of the tree are showing the grades and the vertexes are showing the magnitude of the marks obtained by the students in various courses within two years. The performance of this classifier is shown below where the accuracy of the algorithm is 73.73%. The class precisions of each grade and their corresponding class recalls. The precision of each can be obtain by the following formula:

Precision = (True Positive value)/(True Positive Value + False Positive value)*100. Or

Precision = TP/(TP + FP)*100.

On the other hand, the recalls for each class can be obtain by:

Recall = (True Positive value)/(True Positive Value + False Negative value)*100. Or

Recall = TP/(TP + FN)*100.

The Accuracy of the algorithm can also be obtain by the following formula:

Accuracy = (True Positive + True Negative)/(True Positive + True Negative + False Positive + False Negative)*100. Or

Accuracy = (PT + TN)/(TP +TN + FP + FN)*100 (Ashmeet, & Sathyaraj, 2016)..

The figure below shows the accuracy of the classification model. Various class attributes have been shown together with their various *class precisions* and their *class recalls.* The class precision is the actual and truly classified variable to be true and those that are wrongly classified to be true, and the class precision is obtain based on the above formula. Also, the class recall is actual predicted class variable to be true and the wrongly predicted class variable to

be true, which is equally calculated based on the above formula.

Fig 5: *The performance vector of the model*

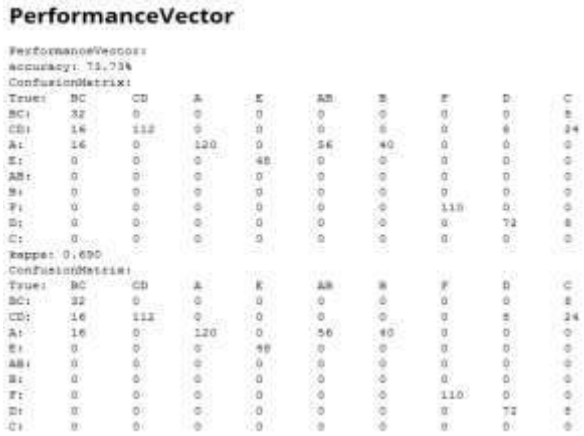

Fig 6: *Performance vector showing Confusion Matrix*

The experimental design for k-means algorithm

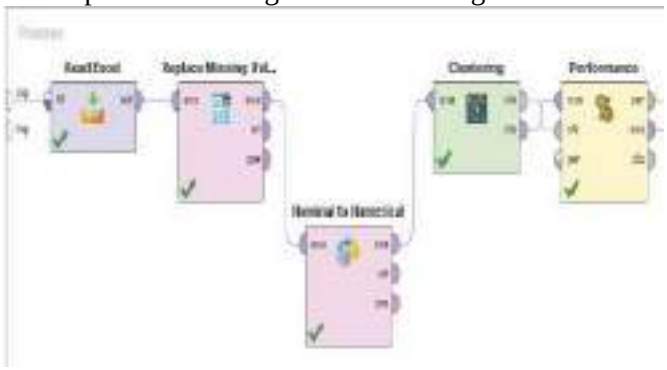

Fig 7: *The k-mean algorithm and the performance of the algorithm*

On the other hand a cluster algorithm above is design by loading the dataset from the *Read Excel* operator through the *Replace Missing value* operator because it takes care of any missing value in the dataset. The *Nominal to Numerical* operator is equally use because of Nominal datasets is not group by *k-means cluster algorithm* which is design for numerical values. Lastly, the *performance* operator determines the average within centroid distance of each cluster (tutorialspoints.com, 2020).

Again, we can study the average centroid distance obtain for all the clusters in all the variables of the distribution. The ***Centroid distance*** simply means a vector that contains single number for each variables of the distribution, where each number stands for the mean of a variable for the observation in that cluster (Jesse, & Mark, 2006).

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -70.759
Avg. within centroid distance_cluster_0: -52.310
Avg. within centroid distance_cluster_1: -188.499
Avg. within centroid distance_cluster_2: -51.398
Avg. within centroid distance_cluster_3: -130.648
Avg. within centroid distance_cluster_4: -79.279
Davies Bouldin: -0.737
```

Fig 8: *performance vector showing average centroid distance for each cluster*

Discussion of Results

Decision tree algorithm and cluster algorithm

The above result can be viewed and compared in order to come up with the best algorithm for predicting student result. The result obtained from the two algorithm are not synonymous in term of comparative analysis, but we can view them separately and draw a conclusion on the best algorithm of two results.

First, the decision tree algorithm can be seen in terms of its accuracy, precision and recall.

True Positive	True Negative	Accuracy	Class Precision	Class Recall
BC = 32	Nil	73.73%	80%	50%
CD = 112	48	73.73%	70%	100%
A = 120	112	73.73%	51.72%	100%
E = 48	Nil	73.73%	100%	100%
AB = Nil	Nil	73.73%	0%	0%
B = Nil	Nil	73.73%	0%	0%
F = 110	Nil	73.73%	100%	100%
D = 72	8	73.73%	90%	90%
C = 0	Nil	73.73%	0%	0%

Table 1: Showing TP, TN, Accuracy, Class precision and Class recalls

Secondly, the k-mean algorithm performance vector generates five (5) different clusters that gives the

Average Within Centroid Distance: 70.759. And the other parameter is Davies Bouldin: 0.737 as it can be seen in figure 8:. The average within centroid distance is calculated by averaging all distance within all clusters. While on the other hand Davies Bouldin is returning cluster with low intra cluster distance or the shortest distance within the centroid and (it is higher in intra cluster similarities). The higher the intra cluster distance the lower the similarities between different clusters (RapidMiner, n.d,).

The subsequent chart is a graph of bar chart showing the general performance of the students obtained in various grades in different courses within the period of that two years. This graph along can served as a tool for decision making for academic future planning

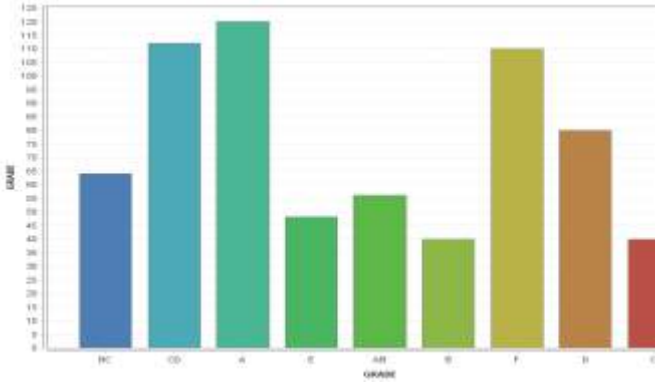

Fig 9**:** *Bar chart showing the grades obtained by all the students*

Conclusion

In this study two data mining techniques were used to determine the best algorithm for predicting the best performance of students, i.e *classification algorithm* and *cluster algorithm.* The results obtain for the ***random tree*** algorithm is precisely better to the cluster algorithm using ***K-mean***. This is because the random tree is supervised learning where a label is assign and predictions is based on the label attribute and is clearly shown, while on the other hand k-mean algorithm predict with the help of *Nominal to numerical* operator as it is not meant for non-numeric values.

Therefore, random tree algorithm can be used in different data mining tools to predict the best performance of students in the entire polytechnic.

References

[1] K. Prasada Reo, M. C. (2016). Predicting Learning Behavior of Student Using Classification Techniques. *International journal of computing Application (0975-8887) volume 139-No 7, 15-19.*

[2] Bradly C. Love (2002). Comparing Supervised and Supervised Category Learning. Psychonomic Bulletin & Review 2002, 9(4), 829-835.

[3] Richard D. Cutler et. Al, (2007). Random Forest for Classification in Ecology. *Ecological Society of America, 88(11), 2007, pp. 2783-279*

[4] http://tutorialspoints.com (2020). *Classification Algorithm-Random Forest.*

[5] Nunung N. Q. and Yudho G. S. (2014), Employees' Attendance Patterns Prediction Using Classification Algorithm. *International Journal of Computing, communications and Instrumentation Engg. (IJCCIE) vol. 1, Issue 1 (2014) ISSN 2349-1469 EISSN 2349-1477.*

[6] Anuradha C, Velmurungan T & Anandavally R, (2015). International journal of Singapore. ISSN: 0976-268X

[7] Berking, P.P.(2006). "A Survey of Clustering Data Mining Techniques", Grouping Multi dimentional data, Springer Berlin Heidelberg, pp.25-71.

[8] Amudha S. (2016). "An overview of Clustering Algorithm in Data Mining". *International Research Journal of Engineering and Technology (IRJET), e- ISSN: 2395-0056, p- ISSN 1395-0075.*

[9] Usama, et al (1996). From Data Mining to Knowledge Discovery in Databases. *Al magazine Volume 17 Number 3.*

[10] Ashmeet S. and Sathyaraj R. (2016). Comparison between Classification Algorithms on Different Datasets Methodologies Using Rapidminer. *International Jounal of Advanced Research in Computer and Communication Engineering. ISSN (online) 2278-1021. ISSN (Print) 2319-5940*

[11] https://en.wikipedia.org/wiki/Data_mining

[12] Jesse D. and Mark G. (2006). The Relationship Between Precision-Recall and ROC Curves. *Appearing in the preceedings of the ss23rd International Conference on Machince Learning, Pittsburgh. PA*

[13] Gurmeet K. & Williamjit S. (2016). Prediction of Student Performance Using Weka Tool. *An International Journal of Engineering Sciences.* Vol 17 ISSN 2229-6913 (Print), web presence: http://www.ijoes.vidyapublications.com

[14] Obadiah M. M., Kelvin O., Raphael A. (2019). Toward Prediction of Students' Academic Performance in Secondary School Using Decision Trees. *International Journal of*

Research and Innovation in Applied Science. Volume IV, ISSUE X, ISSN 2454-6194.
[15] Jeromie R. E. et al (2019). Application of Decision Tree Algorithm for Prediction of Student's Academic Performance. *International Journal of Innovative Technology and Exploring Engineering.* ISSN: 2278-3075, volume-8 Issue-6S.
[16] RapidMiner Studio "n.d.". Fundamental Terms in RapidMiner Studio. In R. Studio, *RapidMiner Studio Manual. Boston, London, Dortmund, Budapest: Global Leader in Predictive Analytics Software.*