International Journal of Computer Applications Technology and Research
Volume 4– Issue 9, 693 - 697, 2015, ISSN: 2319–8656

A Model for Predicting Students' Academic Performance using a Hybrid of K-means and Decision tree Algorithms

Thaddeus Matundura Ogwoka
School of Open, Distance and eLearning (SoDeL)
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

Wilson Cheruiyot
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

George Okeyo
School of Computer Science and Information Technology
Jomo Kenyatta University of Agriculture and Technology (JKUAT)
Nairobi, Kenya

Abstract: Higher learning institutions nowadays operate in a more complex and competitive due to a high demand from prospective students and an emerging increase of universities both public and private. Management of Universities face challenges and concerns of predicting students' academic performance in to put mechanisms in place prior enough for their improvement. This research aims at employing Decision tree and K-means data mining algorithms to model an approach to predict the performance of students in advance so as to devise mechanisms of alleviating student dropout rates and improve on performance. In Kenya for example, there has been witnessed an increase student enrolling in universities since the Government started free primary education. Therefore the Government expects an increased workforce of professionals from these institutions without compromising quality so as to achieve its millennium development and vision 2030. Backlog of students not finishing their studies in stipulated time due to poor performance is another issue that can be addressed from the results of this research since predicting student performance in advance will enable University management to devise ways of assisting weak students and even make more decisions on how to select students for particular courses. Previous studies have been done Educational Data Mining mostly focusing on factors affecting students' performance and also used different algorithms in predicting students' performance. In all these researches, accuracy of prediction is key and what researchers look forward to try and improve.

Keywords: Data Mining; Decision tree; K-means; Educational Data Mining

1. INTRODUCTION

To predict how students may perform during their learning process is a complex task despite continuous increase of data in the databases relating to students academics in institutions of higher learning. According to (Marquez et al., 2013), the academic management systems are not designed properly to support educational managers to investigate which students are at risk of dropping out of university. Data Mining is the process of discovering interesting patterns and knowledge from large amounts of data (Han and Kamber, 2003). From educational Data Mining (EDM) website , "Educational Data Mining is an emerging discipline , concerned with developing methods for exploring the unique types of data that come from educational settings, and use using them to better understand students and the setting in which they learn" . To analyze students' learning process is complex, but thanks to EDM in which its methods and approaches can be used to predict students' performance like the model being proposed in this research paper. Hence university managers will have options of to come with strategies of improving student academic performance (Borka and Rajeswari, 2013). This research proposes the use of Decision tree classification and K-means clustering algorithms to develop a model for predicting the academic performance of students in higher level institutions like universities. Prediction is a method of carrying out Educational Data Mining (EDM) using clustering algorithms like K-means and classification algorithms like decision trees to predict student performance (Ramesh et al., 2013). In university, the performance a student at the end of every semester determines whether a student is to progress to the nest academic year which leads to the completion of his/her studies. Passing in these semester examinations is crucial since it will determine whether a student is to get to final year which later realizes a student graduating and ushered into the economic development of a country. This is one of the reasons institutions of higher learning are established for (Patel et al., 2013). In this research paper, WEKA knowledge analysis software tool is used to for the analysis the algorithms used and the model performance.

2. RELATED WORK

Great work has been done and is always being done by this area of Educational Data Mining. From (Shovon and Haque, 2012)'s research, where they used k-means algorithm to predict the student learning activities by clustering them into: "Good", "Medium", and "Low" based on their GPA. They used 50 students as the training samples and concluded that their prediction accuracy was low and needed improvement in future. Our proposed model has realized improved accuracy by using 173 students as the algorithm training samples. (Yedav and Pal, 2012) using decision trees' ID3, CART, and C4.5 classifiers, conducted a study to predict student academic performance and realized an accuracy of 62.22%, 62.27%, and 67.77% respectively. In our research, using decision tree's J48, we realized a prediction accuracy of 98.8439% on the student training instances. (Kalpesh and Pal, 2013) in their model of predicting students' performance using decision tree's ID3 and C4.5 on 173 training datasets, achieved a prediction accuracy of 75.145% at 47.6 milliseconds execution time.

2.1 Data Clustering

Data clustering is unsupervised statistical analysis technique, which is used to segment large data into homogeneous groups called clusters, in order to discover hidden patterns and relationships to help in quick decision making (Shovon and

Haque, 2012). K-means is simple algorithm that partitions "n" observations into k clusters in which each member belongs to a cluster of nearest mean (Mustafa et al., 2010).

Algorithm 1 Basic K-means Algorithm.

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

Figure 1. Basic K-means algorithm (Mustafa et al., 2010)

2.2 Decision tree classification algorithm

Decision tree is a data mining technique which can be applied in prediction tasks, it is a tree-like structure in which the root and each internal node are labelled with a question (Shovon and Haque, 2012). It is a classifier with the following structure: **Decision node** specifies a test on a single attribute, **Leaf node** indicates the value of the target attribute, **Arc/edge** split of an attribute, and **Path** is a junction test to make final decision.

3. PROPOSED MODEL APPROACH

In The proposed model is to take into account the educational data mining requirements as cited by other authors (Baker and Yasef, 2009) of keeping continuous record of student progress. The goal of this study is to predict students' academic performance. The approach will contain among others: Data tier which is the collection of student semester records, Application layer that will contain data extraction, loading, and knowledge repository. Then finally the presentation layer for displaying results analyzed to the user.

3.1 Data processing and analysis

In data mining, before running tests on the collected data instances, it is necessary to clean and prepare the data for use. In our research, sample data from the Technical University of Mombasa-Kenya student management system was cleaned to look at the relevance of the data attributes to be able to remove any redundancy, or irrelevant features, and analyzed using WEKA software tool. In our study, we have considered dataset of undergraduate students pursuing Bachelor of Science in Technology (BTIT), Bachelor science in Information Technology (BSIT) both government sponsored and self-sponsored, part-time and full-time students at the department of computer science and information technology (CSIT) in the Technical university of Mombasa which was our case study. According to (Suchita and Rajeswari, 2013), on the basis of the data collected some variable attributes are considered to predict student academic performance: Attendance%, Assignment%, Unit tests% and University result%. But for the purpose of this research, only useful fields from the single combined table were selected for our study. Some of the selected variables from the database and recommended attributes for academic prediction are shown below.

Table 1. Selected variables from student records

VARIABLE	DESCRIPTION	POSSIBLE VALUE
Dept.	Student's department	{CSIT}
AdmNo	Student's admission number	{Char}
Gender	Student's sex	{M,F}
Course	Student's course	{BSIT,BTIT}
EntryMode	Student's entry mode	{JAB, Self-sponsored(SS)}
CourseMode	Student's mode of studying	{Full-time(FT), Part-time(PT)}
semesterMarks	Student's total semester marks	{Numeric}
CAT	Whether Student sat for Continous Assessment Tests	{Y, N}
Attendance	Student's Average Semester Attendance	{Good, Poor}
MeanMark	Student's Average Semester Mark	{Numeric}

3.2 Analysis of data

Here we are evaluating the Decision tree algorithm using J48 classifier and K-means algorithm using Simple-Kmeans option from WEKA. From our framework architecture, it consisted of Data Tier where data from students is collected and interesting variables selected. Application tier where data cleaning and mining using algorithms is done producing a prediction model. We used First semester results to train the algorithm (this was the training Dataset), the second semester final results was predicted by the algorithms (this was our Test dataset). After successful learning, the algorithms were tested for prediction of second semester results in which the final grade column was left blank for the algorithm to predict the students' performance (predicted results).

AdmNo	Gender	Course	EntryModde	CourseMode	SemesterMarks	CAT	Attendance	MeanMark	FinalGrade
BSIT/0018/2012	M	BSIT	JAB	FT	468	Y	Poor	59	
BSIT/0029/2012	M	BSIT	JAB	FT	508	Y	Good	64	
BSIT/0048/2012	M	BSIT	JAB	FT	441	N	Poor	55	
BSIT/0501/2012	M	BSIT	JAB	FT	460	Y	Good	58	to be predicted
BSIT/0521/2012	M	BSIT	JAB	FT	516	N	Poor	65	
BSIT/0541/2012	M	BSIT	JAB	FT	455	Y	Good	57	
BSIT/0551/2012	M	BSIT	JAB	FT	532	Y	Good	67	
BSIT/0581/2012	M	BSIT	JAB	FT	545	Y	Good	68	
BSIT/0591/2012	M	BSIT	JAB	FT	531	Y	Good	66	

Figure 2. Part of test dataset used for prediction

Part of the training data used for training the algorithms (first semester student records-cleaned) is shown below:

AdmNo	Gender	Course	EntryModde	CourseMode	SemesterMarks	CAT	Attendance	MeanMark	FinalGrade
BSIT/0018/2012	M	BSIT	JAB	FT	427	Y	Good	53	P
BSIT/0029/2012	M	BSIT	JAB	FT	498	Y	Good	61	P
BSIT/0048/2012	M	BSIT	JAB	FT	301	N	Poor	38	F
BSIT/0501/2012	M	BSIT	JAB	FT	404	Y	Good	51	P
BSIT/0521/2012	M	BSIT	JAB	FT	353	N	Poor	44	F
BSIT/0541/2012	M	BSIT	JAB	FT	470	Y	Good	59	P
BSIT/0551/2012	M	BSIT	JAB	FT	448	Y	Good	56	P
BSIT/0581/2012	M	BSIT	JAB	FT	450	Y	Good	56	P

Figure 3. Part of training data set

This are part of the student datasets which is used to train the algorithms in our model. It represents the first semester student results. The test dataset of records represents second semester students' records in which the model is supposed to predict how the student will perform. And use the model to

cluster students in groups according to their predicted result relationships.

4. RESULTS & DISCUSIONS

Here we evaluated the experiments conducted using the algorithms (Decision tree using J48 and K-means) the analysis and results was done in WEKA. The decision tree algorithm was used to do the prediction and show in tree-like structure the results which will help in making decisions from the predictions made. The K-means algorithms run on the same dataset on the same WEKA tool, was used to group the predicted students into several groups of our choice in relation to the attributes of the records as shown below. Several classification metrics are used to evaluate the results (Manahaes et al., 2013):

- **Accuracy :** the measure of correctly classified instances
- **True positive (TP):** the proportion of positive cases (P) correctly classified as such and
- **True Negative (TN):** the proportion of negative cases (F) correctly classified as such.

After loading the training dataset into WEKA, the preprocessing yielded the following results as shown in Figure 4 below.

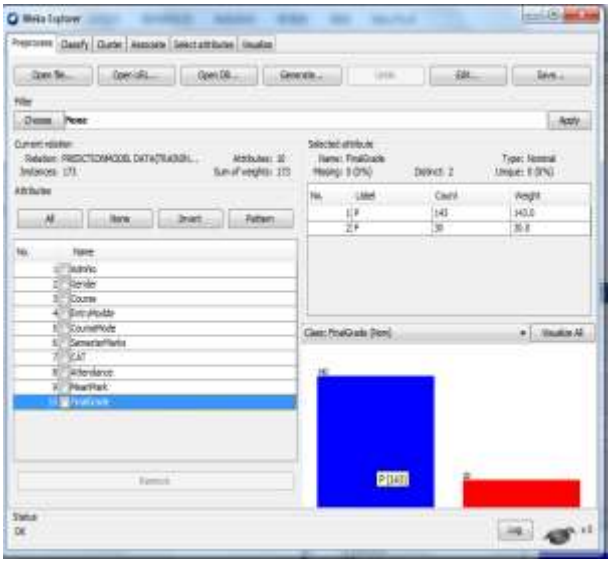
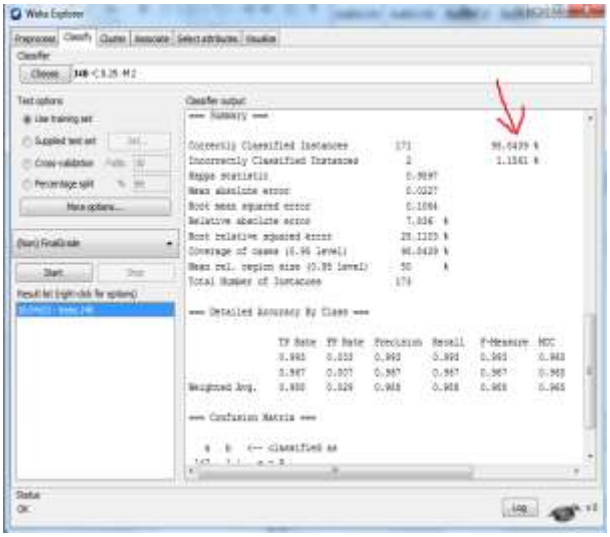


Figure 4. preprocessed training datasset results

From preprocessing results of training dataset, 143 students are identified as passed-P (in blue color as per the bar graph) and 30 students as failed-F (in red color as per the bar graph) in the fig 1above. This is true as per the original dataset, we had a total of 173 student samples. After preprocessing, by using Decision tree's J48 classifier, the algorithm was trained using the same dataset. The results are as shown in fig 5 below.



From the training result, 171 intances were classified correctly with an **accuracy of 98.8439%,** while 2 instances were incorectly classified with an **error** of **1.1561%.** The kappa statistic which takes account of similarities between classsses was 0.95978% which is better. :A kappa value greater than zero indicates the classifier is doing better than chance" (Manhaes et al, 2011). **TP Rate:** Rate of true positives (these indicated students who were correctly clasified in relation to final grade as the class)-the model realised 0.988 out of 1. **FP Rate:** Rate of false positives (these indicated students who were wrongly classified as belonging to a given final grade as aclass)-the model generated 0.029 out of 1. **Precision**: these represents a proportion of students that are truly of aclass(given final grade) divided by the total students classified as that class-the model generated 0.988 out of 1. **Recall**: these represents the proportion of students from the experiment who were classified as a given class (final grade) diveided by the actual total in that class (eqivalent to TP Rate)-the model generated 0.988 out of 1. **F-measure**: this is acombined measure of Predcision and Recall calculated as 2*Precision*Recall/(Precision+Recall) and it was 0.988. From all these acuracy measures, they all approached 1, hence showing our algorithm was learning well from the training instances.

The figure 6 below shows decision tree view generated by the model

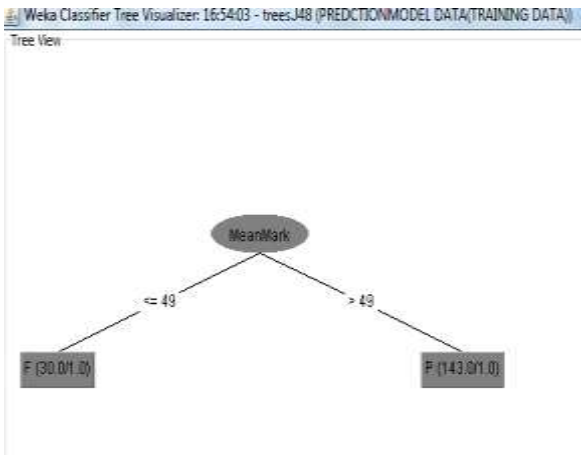


Figure 6 decision tree view of the model

The view above shows in "tree"-structure one of the views that can be used in decision making. From the above results, 30 students whose mean marks were <=49 failed (F) while those with >49 mean marks in their first semester results passed (P).

4.1 Prediction results

After learning using first semester students results, we tested the algorithm with same students but on different results (second semester records) and the Class attribute "FinalGrade" was left blank for the algorithm to predict as was shown in figure 2. Using decision tree algorithm, on the test data, the following output were realized as predicted results as shown in figure 7 below.

```
@attribute AdmNo {BSIT/0018/2012,BSIT/029J/2012,I
@attribute Gender {M,F}
@attribute Course {BSIT,BTIT}
@attribute EntryModde {JAB,SSP}
@attribute CourseMode {FT,PT}
@attribute SemesterMarks numeric
@attribute CAT {Y,N}
@attribute Attendance {Good,Poor}
@attribute MeanMark numeric                    predicted
@attribute 'prediction margin' numeric
@attribute 'predicted FinalGrade' {P,F}
@attribute FinalGrade {P,F}

@data
BSIT/0018/2012,M,BSIT,JAB,FT,468,Y,Poor,59,1,P,?
BSIT/029J/2012,M,BSIT,JAB,FT,508,Y,Good,64,1,P,?
BSIT/048J/2012,M,BSIT,JAB,FT,441,N,Poor,55,1,P,?
BSIT/050J/2012,M,BSIT,JAB,FT,460,Y,Good,58,1,P,?
BSIT/052J/2012,M,BSIT,JAB,FT,516,N,Poor,65,1,P,?
BSIT/054J/2012,M,BSIT,JAB,FT,455,Y,Good,57,1,P,?
```

Figure 7 part of predicted results from the model

As shown from figure 7 above, the model was able to predict the second semester's final grades (as shown by arrow), and as shown from the decision tree view of the predicted results in figure 8 below, 30 students who failed (F) had their end of semester two mean mark <=49 while 143 students passed with a mean mark of >49.

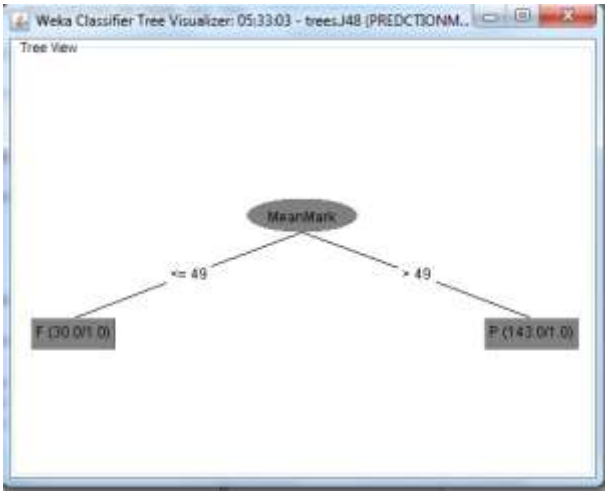


Figure 8 predicted results decision tree view

4.2 K-means clustering predicted results

From the same model, using WEKA, the following results were output.

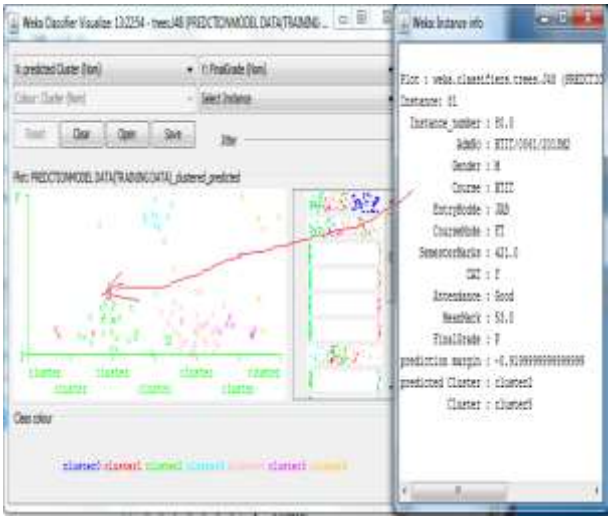


Figure 9 predicted results clusters

"Clustering is the process of grouping a set of elements in such a way that the elements in the same group or cluster are more similar to each other than to those in other groups or cluster" (Kalpesh et al., 2013). As shown from the above fig 4.8, a student of admission number "BTIT/0061/2013M2" who is a male, entry mode by JAB, learning full-time, belonging to cluster0 passed with 53 mean mark, his attendance was good and had sat for the continuous assessment test.

4.3 Results summary

According to (Kalpesh et al., 2013) who predicted students' academic performance using ID3 and C4.5 algorithms, obtained the following accuracy as compared to what we obtained using J48 and K-Means algorithms as shown in table 2 and table 3 below.

Table 2 Algorithm accuracy. Source: (Ogwoka et al., 2015)

Algorithm	Total students	Correctly predicted students	Accuracy (%)	Execution time in milliseconds
J45	173	171	98.8439	20

Table 3 Algorithm accuracy. Source: (Kalpesh et al.; 2013)

Algorithm	Total students	Correctly predicted students	Accuracy (%)	Execution time in milliseconds
ID3	173	130	75.145	47.6
C4.5		130	75.145	39.1

5. SUMMARY OF FINDINGS

The main objective of our research was to apply decision tree and k-means algorithms to create a model for predicting students' performance. To achieve this, our model was realized through stepwise specific objectives:

1. We evaluated decision tree and k-means algorithms in terms of their operations using WEKA free software tool and other written literature.
2. We successfully applied decision tree and k-means algorithms and created a model of predicting students' academic performance where we analyzed 173 undergraduate students of Technical University of Mombasa's Computing and information technology department using first semester results to predict second semester results.
3. We tested the model of predicting students' academic performance and realized an accuracy of 98.8439% at an execution time of 20 milliseconds.

From the most previous researchers we looked into, their biggest challenge was to have an increased accuracy. All the their models as was seen in the previous literature, their accuracies were less than 90% with an execution time as big as 47.6 milliseconds (Kalpesh et al., 2013). Our results of 0.05 milliseconds execution time and accuracy of 98.8439% has reduced this gap.

6. CONCLUSIONS

A model for predicting students' academic performance using Decision tree and k-means algorithms has an improved accuracy and easily be implemented in institutions of higher to do prediction of students' performance and also mine interesting features pertaining academics of students.

7. RECOMMENDATIONS AND FUTURE WORK

From the results and findings of the experiments done in this study, the researcher recommends the adoption of student performance prediction models as Education Data Mining is an emerging data Mining discipline. In our research, WEKA does not update automatically on test dataset predicted as is the case on training dataset, hence to view the results you have to save in a file. In future, we will explore if WEKA has improved on this feature to use in our model or research more on more other open source data mining and analysis tools on this recommendation.

8. REFERENCES

[1] Marquec-vera, C. Cano, A. Romero, C., and Ventura, S. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. Applied Intelligence, 2013, 1:1-16.

[2] Han, J., Kamber, M., and Pei, J. (2013). Data Mining Concepts and Techniques, third edition. Morgan Kaufmann.

[3] JEDM-Journal of Educational Data Mining, (JEDM. ISSN2157-2100). Meaning of Educational Data Mining. Retrieved from www.educationaldatamining.org/JEDM/index.php/JEDM.

[4] Burka, S., and Rajeswari, K. Predicting Student Academic Performance using Data Mining, 2013, 2(7), 213-219.

[5] Ramesh, P., Parkavi, P., and Ramar, K Predicting Student Performance. A statistical and Data Mining approach, 2013, 63(8), 0975-8887.

[6] Patel, M., Abdul, K., and Pappal, P. Educational Data Mining and its Role in Education Field, 2013, 5(2), 2458-2461.

[7] Ajay, P., and Saurabh, P., Data Mining Techniques in EDM for Predicting the Performance of Students, 2013, 2(6), 2279-0764.

[8] Shovon, M., and Haque, M., An approach of Improving Student Academic Performance by using K-means clustering Algorithm and Decision tree, 2012, 3:8

[9] Ahmed, A., and Elaraby, I., Data Mining: A prediction of Student performance using classification method, 2014, 2:43-47.

[10] Mustafa, T., Ayesha, S., and Khan, M. 2010. Data Mining Model for Higher Education System. European Journal of Scientific, 1450-216X.