Article Number: A82C94

A paper presented at the 10th Egerton University International Conference and Research Week, Egerton University, Njoro, Kenya.
30th March - 1st April 2016

ACADEMIC JOURNALS
expand your knowledge

Conference Proceedings

Full Length Research Paper

Prediction modelling of academic performance with logistic regression: A case of rural primary school students in Kenya

Mvurya Mgala* and Audrey Mbogho

Institute of Computing and Informatics, Technical University of Mombasa, Kenya.

Every year, when the Kenya Certificate of Primary Education (KCPE) examination results are released, the same story of mass failure in rural schools is repeated. Academic performance prediction modelling could provide an opportunity for learners' outcomes to be known early, before they sit for final examinations. This would be particularly useful for education stakeholders to initiate intervention measures to help students who require high intervention to pass final examinations. This study proposed that an academic performance prediction model could be built using Logistic Regression to classify students into two categories: those that will pass and those that will need intervention to pass. A six-step Cross-Industry Standard Process for Data Mining (CRISP-DM) theoretical framework was used to support the modelling process. Modelling was conducted using two datasets collected in Kwale County and Mombasa County. The first dataset had 2426 records having 22 features, collected from 54 rural primary schools. The second dataset had 1105 records with 19 features, collected from 11 peri-urban primary schools. Evaluation was conducted to investigate: (i) the prediction performance of Logistic Regression on the two datasets with all the features and; (ii) the prediction performance with an optimal subset of features. Two common performance measures (ROC area and F-Measure) were adopted. It was found that the model achieved a ROC area measure of 88.7% with all features and 88.5% with the optimal feature dataset. Similarly the F-Measure rate was 89.7% for all the features and 89.6% for the optimal feature subset. Further, a mobile application was implemented to facilitate the model use in rural areas where desktops cannot be used. Teachers in 15 schools used the model for two weeks to classify their Class Six and Class Seven students. Results show that nearly 80% of the students requiring high intervention could be determined. This high prediction performance means that the students who need high intervention could be determined early enough before the final examination. Further, this accuracy of prediction is good enough to motivate stakeholders to initiate strategic intervention measures.

Key words: Prediction modelling, academic performance, rural schools, prediction performance.

INTRODUCTION

Education is very important, as it prepares the human resource necessary for economic development of a country (Munyi and Orodho, 2015). Kenya, like any developing country has faced challenges to ensure quality education, especially in rural public schools. Most students in these schools perform below the national

average marks required for progression to secondary schools. Thus, they drop out of the school system and end up as unskilled labours. The problem of mass failure in rural public schools may have been catalysed by the free primary education policy. The initiative caused increased enrolment with the same limited resources (Somerset, 2009). Somarset argues that this increased enrolment is what affected the quality of education. Increased access is a good thing as long as the required resources are made available. Munyi and Orodho (2015) associates the problem with: overstretched facilities, overcrowding in schools, high teacher- pupil ratio, over-age children, insufficient textbooks, poverty, culture that impedes education, and limited support from the community.

Kwale is one of the Counties in Kenya that has many rural schools. Like many Counties with rural schools, they face the problem of mass failure; one that is likely to remain or worsen unless intervention measures are put in place. The strategies through which mass failure in primary schools can be reduced have received much attention from researchers. Studies have attempted to identify the causes for mass failure in rural public schools and proposed recommendations to the government to act on (Mweki, 2016). However, many such studies have not suggested ways that could be used in solving the problem. The government of Kenya has attempted to enforce policies for managing education in order to improve quality (Lucas et al., 2014). A drawback is that this is possible only at macro-level. The government approach lacks effectiveness because the policies are general (Achola and Pillai, 2016). The problem is caused by many and varied reasons, hence the need for specific strategies to assist individual students.

This study proposes to use Machine Learning, a technique that has been used successfully in developed counties to predict students at-risk of failing long before they sit for final examinations (Tamhane et al., 2014; Thai-Nghe et al., 2011). Techniques for determining the attributes that are most indicative of the target have also been used in a number of areas to reduce the size of the dataset used (Tang et al., 2014). This approach was used in this study, aimed at identifying the causes of poor performance in academic work, and to use these causes or attributes, to classify students into two categories, high-intervention and low-intervention. While high-intervention stands for the students that need strategic measures for them to get above average marks, low-intervention stands for the students that are above average, or those that may not need any strategic measure to achieve above average marks. The model predicts the students that require high intervention as early as one or two years before they sit for Kenya Certificate of Primary Education (KCPE). The high accuracy of prediction by the model is what motivates strategic intervention.

The paper is extracted from a bigger study as it is guided by the following questions:

i) What is the prediction performance of Logistic Regression on the two datasets when the complete datasets are used?
ii) What is the prediction performance of the two datasets using the optimal subset of features?

The rest of this paper is organised as follows: First is a description of the adopted methodological approach, followed by a presentation of the prediction performance results and discussion. The paper concludes with a summary of achievements of further research.

METHODS

This study used the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Kurgan and Musilek, 2006). CRISP-DM is a six-step process that include domain understanding, data understanding, data preparation, data mining, evaluation, and using discovered knowledge. These steps are illustrated in Figure 1.

Problem domain understanding

Problem domain understanding focuses on the area of interest in the problem solving process (Asamoah and Sharda, 2015). In this study, a preliminary survey with 7 education officers, 14 head teachers and 124 teachers in Kwale County was used to achieve this. Findings reveal that poor academic performance exists in the County, and that nearly 70% of the students who sit for KCPE do not achieve above the national of 250 marks out of 500 total marks.

Data understanding

In data understanding, data is examined closely to determine its quality and usefulness for the mining process (Asamoah and Sharda, 2015). This was achieved by scrutinising the collected 2426 student records consisting of 22 fields from rural schools, and the 1105 student records of 19 fields from peri-urban schools. The examination entailed: a consideration for the tools to be used for the type of data, the fields that need conversion to make them usable in WEKA, and the records that needed data to be either filled or deleted.

Data preparation

Data preparation entails combining tables from different sources

*Corresponding author. Email: mmgala@tum.ac.ke.

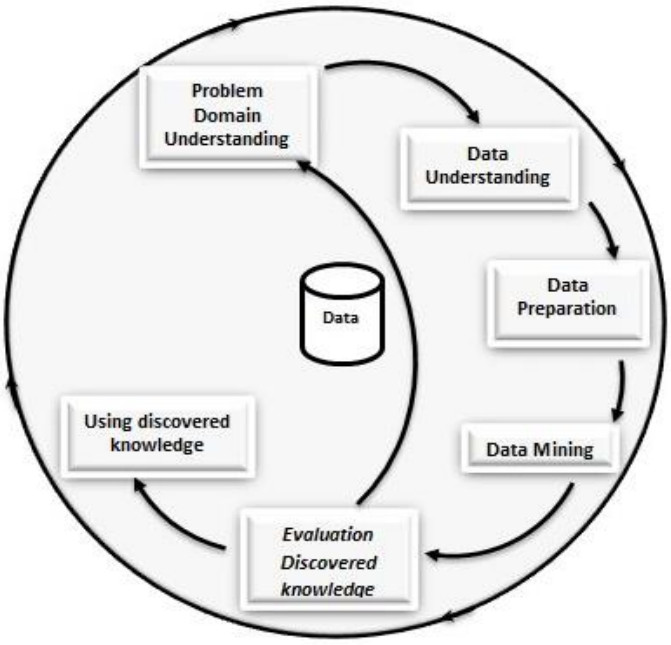

Figure 1. CRISP-DM Processes.
Source: Mgala (2016) and Kurgan and Musilek (2006).

and pre-processing the data. It entails solving data problems such as missing data that may hinder effective analysis (Sattler and Schallehn, 2001). Data preparation was achieved by: typing the manual records into excel, determining the validity of the typed data, cleaning the data by replacing missing values and deleting the records that did not have the target, discretising some of the data, and selecting the most predictive features.

Data mining

Data mining is the process of analysing data in order to extract meaningful patterns from it (Shafique and Qaiser, 2014). Logistic regression modelling technique was used to build the models using WEKA machine learning environment (Hall et al., 2009). The rural dataset was divided into 70% training data and 30% test data. The peri-urban data was divided into 60% training data and 40% test data; this was done to increase the number of test data for the peri-urban data. Using a 10-fold cross validation method (Refaeilzadeh et al., 2009), the modelling and evaluation was achieved as shown in the results.

Evaluation and discovered knowledge

Evaluation focuses on interpreting the model performance to determine whether it achieves a reasonable performance or not (Shafique and Qaiser, 2014). A comparison of the prediction performance for logistic regression on the different datasets was carried out; thereafter, two metric measures (F-Measure and ROC area) were used for the evaluation process.

Using the discovered knowledge

The stage explains how the discovered knowledge or results are to be used. The prediction results obtained using logistic regression was compiled into a report that is presented in conferences or presented to education stakeholders. The aim is to motivate initiation of strategic intervention among the stakeholders.

RESULTS AND DISCUSSION

The results of modelling and testing using the 10-fold cross validation method are presented in Table 1. Two metrics were selected to compare the model prediction performance with the different datasets, Receiver Operation Characteristic (ROC) Area and F-Measure. ROC Area is a metric curve generated by plotting sensitivity against specificity; it is a preferred measure because of its stability even in imbalanced classes (Jiménez-Valverde, 2012). F-Measure is the other preferred metric because it combines precision and recall to obtain an average value (Shaikh et al., 2015).

DISCUSSION

The aim of this research was to develop a model that could predict the students that require high intervention. The high prediction performance of the model is what would motivate strategic intervention among education stakeholders. Detailed results on feature selection and implementation of the model in a mobile form together with evaluation process results are presented in Mgala

Table 1. Results obtained using 10-fold cross validation with logistic regression.

Metric	Model Prediction Performance			
	Full-rural dataset (22 features)	Optimal rural dataset (7 features)	Peri-urban dataset (19 features)	Peri-urban dataset (7 features)
ROC Area (%)	88.7	88.5	89.7	90.2
F-Measure (%)	89.7	89.6	78.4	79.9

(2016). The results presented here show that the model prediction performance for the rural dataset was nearly the same for both the complete dataset and the optimal 7 feature dataset. This means, the seven features could be the most indicative of the problem of poor performance in Kwale County. These features are: test-marks, gender, family-income, student-age, teacher-shortage, student motivation, and study-time.

Importantly, the high metric values indicate that students requiring high intervention can be classified early before they sit for KCPE, and that the stakeholders will be motivated to initiate strategic intervention. On the other hand, the peri-urban data show slightly higher performance with the ROC Area metric. It also shows a noticeably lower performance for the F-Measure value. This shows that the model is sensitive to the type of data used. In the work by Mgala (2016), the optimal features picked were: test-marks, parent-education-level, student-age, teacher-absenteeism, student-discipline, gender, and family income. Four features are similar in both datasets while three are different, which explains the slight difference in performance.

Conclusion

This paper presents the CRISP-DM process for data mining and extends it to be applied in educational data from rural schools in Kwale County, Kenya. The process was used to model logistic regression. Results show that a high accuracy was obtained in predicting the students that require high intervention before they sit for KCPE. Using the cross validation method in the two datasets, and using preferred metrics, ROC Area and F-Measure, high values of over 88% were obtained with the rural dataset. This is the focus of the study. The peri-urban dataset shows a noticeable variation but also attained nearly 80% with optimal dataset. These results are high enough to motivate initiation of strategic measures for the students that are classified as requiring high intervention.

Future work will be necessary that will extend the study to other Counties which have rural schools to validate the model. It is also intended to further refine the model by using more data collected from the other Counties. The mobile tool will also be refined so that it could be adopted in the Kenyan education system and beyond.

CONFLICT OF INTERESTS

The authors have not declared any conflict of interests.

ACKNOWLEDGEMENTS

The authors are grateful to the Hasso Plattner Institute for funding this work, jointly with Technical University of Mombasa, and to the Kenyan NACOSTI for allowing us to collect data from Kwale and Mombasa Counties.

REFERENCES

Achola PP, Pillai VK (2016). Challenges of primary education in developing countries: Insights from Kenya. Routledge.

Asamoah D, Sharda R (2015). Adapting Crisp-Dm Process for Social Network Analytics: Application to Healthcare. AMCIS 2015 Proceedings. Retrieved from http://aisel.aisnet.org/amcis2015/BizAnalytics/GeneralPresentations/33.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009). The WEKA Data Mining Software: An Update. Sigkdd Explor. Newsl 11(1):10-18. https://doi.org/10.1145/1656274.1656278

Jiménez-Valverde A (2012). Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. Global Ecology. Biogeography 21(4):498–507. https://doi.org/10.1111/j.1466-8238.2011.00683.x.

Kurgan LA, Musilek P (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review 21(1):1–24. https://doi.org/10.1017/S0269888906000737.

Lucas AM, McEwan PJ, Ngware M, Oketch M (2014). Improving Early-Grade Literacy in East Africa: Experimental Evidence from Kenya and Uganda. Journal of Policy Analysis and Management 33(4):950-976. https://doi.org/10.1002/pam.21782

Mgala M (2016). Investigating prediction modelling of academic performance for students in rural schools in Kenya (Thesis). University of Cape Town. Retrieved from https://open.uct.ac.za/handle/11427/23463.

Munyi CM, Orodho JA (2015.). Wastage in Schools: What Are The Emerging Internal Efficiency Concerns in Public Primary Schools in Kyeni Division, Embu County, Kenya? Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1032.7890andrep=rep1andtype=pdf

Mweki PM (2016). Institutional Factors Influencing Pupils' Performance In Mathematics At Kenya Certificate Of Primary Education In Kathonzweni Sub County, Makueni County, Kenya (Thesis). University of Nairobi. Retrieved from http://erepository.uonbi.ac.ke:8080/xmlui/handle/11295/98422.

Refaeilzadeh P, Tang L, Liu H (2009). Cross-Validation. In L. LIU and

M. T. ÖZSU (Eds.), Encyclopedia of Database Systems Springer US. pp. 532-538. https://doi.org/10.1007/978- 0-387-39940-9_565.

Sattler KU, Schallehn E (2001). A data preparation framework based on a multidatabase language. In Proceedings 2001 International Database Engineering and Applications Symposium pp. 219-228). https://doi.org/10.1109/IDEAS.2001.938088.

Shafique U, Qaiser H (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). International Journal of Innovation and Scientific Research 12(1):217-22.

Shaikh A, Mahoto N, Khuhawar F, Memon M (2015). Performance Evaluation Of Classification Methods For Heart Disease Dataset. Sindh University Research Journal - SURJ (Science Series), 47(3). Retrieved from http://sujo.usindh.edu.pk/index.php/SURJ/article/view/1190.

Somerset A (2009). Universalising primary education in Kenya: the elusive goal. Comparative Education 45(2):233-250. https://doi.org/10.1080/03050060902920807.

Tamhane A, Ikbal S, Sengupta B, Duggirala M, Appleton J (2014). Predicting Student Risks Through Longitudinal Analysis. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining New York, NY, USA: ACM. pp. 1544-1552. https://doi.org/10.1145/2623330.2623355.

Tang J, Alelyani S, Liu H (2014). Feature selection for classification: A review. In C. Aggarwal (Ed.), Data Classification: Algorithms and Applications P37. CRC Press. Retrieved from https://www.crcpress.com/Data-Classification-Algorithms-and-Applications/Aggarwal/9781466586741.

Thai-Nghe N, Drumond L, Horváth T, Krohn-Grimberghe A, Nanopoulos A, Schmidt-Thieme L (2011). Factorization techniques for predicting student performance. Educational Recommender Systems and Technologies: Practices and Challenges (In Press). IGI Global. Retrieved from http://147.172.223.251/pub/pdfs/Nguyen_et_al_ERSAT_2011.pdf.