Data-Driven Intervention-Level Prediction Modeling for Academic Performance

Mvurya Mgala
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
mmgala@cs.uct.ac.za

Audrey Mbogho
Dept of Computer Science
University of Cape Town
HPI School in ICT4D, 7701 Cape Town
audrey.mbogho@uct.ac.za

ABSTRACT

Poor academic performance in final exams at primary school level in Kenya is a strong indicator that the student will not attain the desired career in future. It is therefore important to be able to predict the students who are likely to achieve below average marks and need high intervention early enough for them to improve their marks. This paper reports on a study to classify primary school students into two categories, those that need high intervention and the rest. The prediction can be initiated as early as two years before the final exam. An important highlight of this study is its focus on rural schools in a developing country. A total of 2426 records of students are used to build intervention prediction models. In the first set of experiments all the features are used. An optimal subset of features is then determined and a second set of experiments carried out. Results demonstrate that it is possible to attain reasonably accurate intervention prediction models even with the reduced dataset. The insights obtained will be used to build a mobile prediction tool that can be utilized by education stakeholders in rural regions where there is lack of electricity.

Categories and Subject Descriptors

D.2.8 [**Machine learning**]: Metrics—*classification algorithms*

General Terms

Algorithms, Classification, Model.

Keywords

Predicting academic performance; prediction model; machine learning, technology in education; data mining

1. INTRODUCTION

This study is being conducted using primary school data collected from one of the rural regions in Kenya. Over the years nearly 70% of the students in this region have been scoring below average mark in the national examination.

Among the primary goals of an education system is to ensure students who finish a given stage have acquired sufficient skills and knowledge to transition to a promising career. The effectiveness of any education system to meet this goal is a major determinant of the socioeconomic status of the people going through that education system. In particular, this is the stage at which the opportunities for a bright future are nurtured.

Given that the trend in education is to achieve universal primary education where children are able to complete a full course of primary schooling [21], it is disturbing to note that in many developing countries, thousands of children complete primary schools with low grades and are forced to drop out of the school system with no skills for meaningful employment [14, 22]. Since the final examinations are given great importance as a measure of a student's intelligence [27], and hence, the ability to contribute to the economic development of any country, it is important to identify students at risk of poor performance early enough so that personalized intervention plans can be started to improve the final examination marks. These interventions will come from the education stakeholders such as education officers, parents, head teachers, teachers and the community.

Parents and other stakeholders of education in rural regions of Kenya rely heavily on the teachers' input for good academic performance of the students. However, free primary education has seen a huge increase in enrollment [30], which overwhelms the teachers with work. Teachers rely mainly on continuous assessments to determine if the student might be in the category requiring high intervention or low intervention. This makes the process mainly reliant on the teacher and there is no objective way of determining the level of intervention the student may require. The dependence on test marks, especially for the final year of study may not allow enough time to initiate pro-active intervention measures to improve academic performance. It is with this understanding that this study has taken the initiative to use machine learning techniques to detect patterns which can be associated with poor academic performance. The study utilizes factors gathered from surveys carried out with education officers, teachers and literature search.

This study has been motivated by the need to develop a computer-based prediction tool, reliable enough for classifying a student in order to determine whether they will require high intervention for them to achieve a final exam score high enough for high school entrance.

ICTD '15, May 15 - 18, 2015, Singapore, Singapore
Copyright 2015 ACM 978-1-4503-3163-0/15/05 ...$15.00
http://dx.doi.org/10.1145/2737856.2738012.

The research is driven by the following questions:

- Which of the selected set of features that affect academic performance form a subset of the most predictive features for the final exam mark and hence will streamline the mobile intervention level prediction tool's interface?

- Which among the common supervised algorithms when used with the most predictive set of features performs best in classifying the students into the categories of high intervention and low intervention?

- To what extent is the class teachers' categorization of the students into high and low intervention comparable with the developed mobile intervention level prediction tool classification accuracy?

- What is the teachers' perception of the mobile intervention level prediction tool?

The first question is important because determining an optimal subset reduces the hypothesis search space and the storage requirements, as too much data may be irrelevant or redundant to the learning task. For this study, most importantly there will be a streamlined mobile tool interface.

The second question is at the core of this study as it is important to determine the best supervised algorithm among the different classifiers including: regression, instance based, decision trees, Bayesian, Kernel methods, and Artificial neural network. This is what the study requires to implement, the most accurate model as a mobile-based intervention level prediction tool.

The importance of the third question is in the fact that the mobile intervention level prediction tool's accuracy of classifying the students will be meaningful and sensible if it does not contradict the common knowledge that the teachers have about their students' ability level. Finally, the fourth question gives insight into the impact level the tool has had on the teachers who use it.

The study adopts a theoretical perspective of predictivism [3]. It is in the category of predictive research that goes beyond explanation to the prediction of precise relationships between dimensions of characteristics of a phenomenon using data mining techniques. The methods used can be divided into three steps: 1) developing predictive models, 2)validating their performance and, 3) studying their impact of use [32].

In this study, the models are built based on data collected from students who were in Standard 8 (equivalent to Grade 8 in other countries such as the US) in 2013 and sat their Kenya Certificate of Primary Education (KCPE) exam at the end of the year. The dataset consists of only the students who sat the exam and had marks. Cheating and absent cases are deleted, reducing the total number of records to 2426. Mean imputation is used to address missing values for the test marks which consist of end year test marks for Standard 6 and 7 and end of first term examination marks for Standard 8. Missing values on the questionnaire records are imputed with the most frequently occurring values. Features are selected as a preprocessing step using a technique called filters [15], which filters out unwanted features independent of any learning algorithm using a filter algorithm before learning commences.

The rest of this paper is structured as follows: in section 2, related work is discussed. A description of the dataset used for the analysis, its preprocessing and feature selection are presented in sections 3. Section 4 discusses the methodological framework. Section 5 discusses the experiments and results. Section 6 gives a discussion of the insights, and future work while section 7 gives the conclusion.

2. RELATED WORK

There is a wealth of research available [29] on predicting academic performance using machine learning that has been carried out for the developed world. It does not appear that much has been done for the developing world. A few selected papers are reviewed to put this research in context.

Previous research has studied predicting academic performance using different techniques and data sets. Affendey et al. [1] classified university students into two classes: either first-second-upper or second-lower-third classes. They used features such as subjects taught and a dataset of 2427 records. Applying the dataset to different classifiers they were able to determine the subjects that influence performance and the best classifiers: Naive Bayes, AODE and RBFNetwork. Kotsiantis [17] compared some of the state of the art regression algorithms to find out which algorithm is more appropriate to predict students' marks. Anozie and Junker [2] predicted students' scores on end of year state accountability exams from dynamic testing metrics developed for intelligent tutoring system log data using linear regression models. Comparing the performance and usefulness of different data mining techniques, Romero et al. [25] classified students based on their Moodle usage data and final marks obtained in their respective courses.

An earlier work carried out by Chamillard [5] aimed at helping professors guide their students to focus on potential areas of difficulty and give them insight into the relationships between the courses so that they can implement changes on the curriculum. Vandamme et al. [31] carried out a study similar to ours, that classified students into three groups as early as possible: the 'low-risk' students with a high chance of good academic performance; the 'medium- risk' students who may succeed due to measures taken by the university; and the 'high-risk' students, who have a high chance of failing. Jovanovich et al. [13] proposed a method for selecting the best among a set of classification models. They used a class labeling, which we have adopted, that separates students with highest grade (label value 1), from the rest (label value 0).

Different predictors as indicated by Golding et al. [10] have been used in predicting students' academic performance. The study points out that the task of finding effective predictors for academic performance remains incomplete. Rumberger and Lim [26] summarized twenty five years of research on factors that cause students to drop out. They categorized the factors as; individual level factors and factors associated with institutional characteristics. At the student level, the student's background focused on demographics, health variables, previous school experience and performance. These findings have reinforced the list of attributes compiled in this study using surveys from education officers and head teachers.

Previous studies have also emphasized the importance of feature subset selection as a way of enhancing the performance of learning algorithms. It reduces the hypothesis search space and the storage requirements, as too much data may be irrelevant or redundant to the learning task [12]. Yu and Liu [33] define feature redundancy and propose to perform explicit redundancy analysis in feature selection by developing a correlation-based method for relevance and redundancy analysis. According to Kohavi et al. [15], features can be classified into strongly relevant, weakly relevant and irrelevant. Strong relevance indicates that the feature is always necessary for the optimal subset; it cannot be removed without affecting the original conditional class distribution. Weak relevance suggests that the feature is not necessary but may become necessary for the optimal subset at certain conditions. Irrelevance indicates that the feature is not necessary at all. An optimal subset therefore should include all strongly relevant features none of irrelevant and a subset of weakly relevant features. The filter approach used in [12] and [16] prove to be practical because the features are selected independent of the algorithm and this is much faster compared to the wrapper approach [4] which is based on a specific algorithm. Our study uses the insights from these previous studies in performing feature selection as a preprocessing step in developing the intervention prediction model. We determine strongly relevant, weakly relevant and irrelevant features by using the correlation-based method.

To conclude this section we review a recent work where Tamhane et al. [29] predicts students at risk of poor performance so that the right personalized intervention plans can be initiated. A large data set was used as obtained in a longitudinal analysis. Although the emphasis is on school dropouts in a developed world where digital data is available, they give our study very useful insights as they report on the preprocessing step and on the relative performance of various machine learning techniques used for building predictive models. Their data set contains demographic features such as ethnicity, disability, free meals and behavioral data.

Generally, all these studies provide insight to systematically follow the steps of preprocessing and algorithm selection in the process of the intervention prediction model development.

3. DATA DESCRIPTION

Kwale County is one of the administrative regions in the Coast Province of Kenya, which is further subdivided into three Districts: Matuga (Kwale), Msambweni and Kinango. There are a total of 328 primary schools with over 15,000 students doing the KCPE exam every year.

Primary school education starts at age 6 or 7 years. Students learn for eight class years as a requirement to complete primary school education. Each class year starts in January and ends in December, with a four-week break in April, August and December. The first year is called Standard One and the subsequent year is Standard Two, all the way until Standard 8. Standard One through Standard Four are known as lower primary and Standard Five through Eight are referred to as upper primary.

At present, term test marks, at least for the upper primary, are stored securely by the class teacher of each class and copies are stored in the head teacher's office. This is the data that was made available to this research. It consists of end of year marks for Standard Six and Standard Seven and end of first term marks for Standard Eight. The first term marks for Standard Eight are picked because this is the last school-based exam before the KCPE exam and may be a good indicator of the final marks that the student will attain.

The second category of data is obtained through questionnaires filled by the students. It contains students' personal factors, family and school factors. The third category of data is the KCPE marks which is made available to this study by the County director's office.

It is important to mention that the privacy of students whose data is captured is maintained by ensuring their names and personal numbers are not used in the dataset for experiments.

3.1 Background on the Education System in Kenya

It is important to understand the structure of the education system and the final examination, as it will help in clarifying the type of features used, and the definition of intervention levels.

Primary schools in Kenya are classified as private or public. Private schools are owned, funded and managed by individuals as a business. Public schools on the other hand are built using public funds, get teachers who are paid by the government and parents are involved in running the schools through school management committees (SMC). In this research our focus is on the public schools since they form the education system and any education policy made directly affects them. Free education also only applies to public schools.

Since 1985 public education in Kenya has been based on the 8-4-4 system. The examining body, Kenya National Examination Council (KNEC) is mandated to give exams to students after 8 years of primary level and 4 years of secondary level. These exams are very important as they determine whether or not a student will proceed to the next level.

Standard Eight students take the KCPE exam, a standardized national examination taken by all students at the same time. They are tested in five subjects:

- Language I: English grammar, reading, comprehension and composition writing
- Language II: Swahili grammar, reading, comprehension and composition writing
- Mathematics
- Science
- Social Studies (History, Civics, Geography, Religion)

Each subject carries a maximum of 100 marks, making a total of 500 possible marks. It is important to point out that all subjects carry equal weight at this level and secondary school admission is only determined by the total marks the student has achieved. No emphasis is given on maths or languages even though these two are considered important. The exams are also mainly objective apart from the composition writing component for both Languages I and II, which constitute 40% of the total marks in each of the two language subjects.

The marks obtained in KCPE determine which secondary school a student will attend. The excellent students are admitted to national schools which have the best facilities and resources. There are a total of 105 national schools in the country, with at least two national schools in each of the 47 counties recently set up. On average the two national schools admit 430 students each year, which makes them quite competitive and for only the best students. The good students are admitted to second tier provincial schools and the average students are admitted to district or community schools and some to private schools, the rest drop out of the school system. In the last three years nearly 25% of the total number of candidates every year in the country have missed secondary school admission. This is mainly because the students who score less than 250 marks, majority of whom are from rural areas, are considered failures and therefore dropout of the school system. This research focuses on those students who are likely going to fail in order to improve their marks to a level where they can get admission in the community schools or private secondary schools.

4. METHODOLOGICAL FRAMEWORK

The modeling techniques considered in this study are classification algorithms which are based on supervised learning. These techniques are used because our data for training and testing is labeled(the data sets contain the features and the target). The aim is to categorize between the students who require high intervention and those that require low intervention (a binary classification process).

The methodology has the following phases: data collection, data preparation, data partitioning, training models, evaluating models and feature reduction. Feature reduction is carried out to determine an optimum feature subset. A second set of experiments is carried out to determine the possibility of using an optimum data set that could save on computer resources and time. The best performing trained and tested model can then be used to predict final marks using features of new students.

4.1 Data collection

Students' demographic data, behavior and attitude data, parent and school factors are collected using questionnaires that are given to the students. Test marks for end of year exams when the students were in Standard Six, Standard Seven and Term One marks for Standard Eight are collected from the examination offices in the schools. Results for the final exam (KCPE) for the same cohort of students are collected from the County education office. Identifying student information is removed while preparing the data set.

During the actual experimentation, we use the average test marks for the three tests and a total of 21 student responses from the questionnaire. As mentioned earlier, the intervention is derived using the national threshold of 250 out of the possible 500 marks. We adopt this KCPE mark where below 250 marks requires high intervention and 250 and above requires low intervention. As stated earlier, our major interest is to identify the group that will fall into the category of below the 250 mark, which is the risk group that needs high intervention. The study addresses the problem of predicting intervention levels by transforming it into a binary classification problem, where high intervention students are labeled as high samples and low intervention students are labeled as low samples.

4.2 Data Preparation

The record for each student is completed by adding test marks, questionnaire responses and the KCPE marks. The data, however, is seen to be incomplete. Some students' test marks are available but they were absent during the filling in of the questionnaires. Others were absent during KCPE examinations. A number of students lack test marks for one or two tests because they transfered to new schools in Standard Seven or Eight.

For the sake of simplicity in data preprocessing, we impute missing data for the two categories: the test marks and the students' responses. For the test marks, we use a common imputation technique which simply replaces the missing values with the mean of the feature. For the second category, we replace the missing value with the most frequently occurring values. In the case where a student does not have KCPE marks, we delete the complete record since the KCPE marks are the target variable. This preprocessing step leaves 2426 records and 22 features for the study.

4.3 Data Partitioning

In the experiments, 10-fold cross validation is used to evaluate the accuracy of predictions. This divides the entire data set into 10 equal parts. Prediction is repeated 10 times, each time keeping one of the 10 parts as test data and the other 9 parts as training data to build the prediction model. Finally, test results on all the 10 parts are accumulated together to calculate the average prediction accuracy.

4.4 Building Prediction Models

The study uses a number of common classifiers as intervention level prediction models. These models are trained and tested using the 10-fold cross validation data set as explained in section 4.3. The implementation is carried out in WEKA [11] with its default settings. We choose eight classifiers:

- Logistic regression [23] is a statistical technique based on the logic function. It makes use of estimating the probability of a binary event occurring (e.g. whether a student will require high intervention or low intervention to get good marks in KCPE).

- Multilayer perceptron [9] is a type of artificial neural network made up of simple interconnected neurons, also called nodes. The nodes are connected by weights and output signals which are a function of the sum of inputs to the node modified by a simple nonlinear activation function. The superposition of many simple nonlinear activation functions enable the multilayer perceptron to approximate non linear functions.

- Sequential minimal optimization algorithm (SMO) is an algorithm for training support vector machines [24]. SMO is preferred because it breaks large quadratic programming (QP) optimization problems into a series of smallest possible QP problems. These small QP problems are solved analytically hence avoids using a time-consuming numerical QP optimization as inner loop. SMO is thus faster for linear SVMs and sparse real world data sets.

- Bayesian network classifiers are directed acyclic graphs that allow efficient representation of joint probability distribution on a set of random variables[7].

- Naïve Bayes classifier is a simple Bayesian classifier with strong assumptions of independence among features and it is competitive with state-of-the-art classifiers [7].

- Lazy learners are a strategy for building classifiers where models are not learnt from the whole data set. Selected patterns are made depending on the query received and a classification model is learnt with these selected patterns [8]. We choose the locally weighted learning (LWL) type.

- Random forest classifier, is an ensemble of supervised machine learning technique which uses a decision tree as the base classifier [18].

- J48 algorithm is an implementation of the C4.5 decision tree learner. The implementation produces a decision tree model. Greedy techniques are used to induce it for classification [28].

All the 22 features per student are used as input to predict a binary output as to whether a student needs high intervention or low intervention.

4.5 Model Evaluation

The data set contains 2426 samples having a distribution of 68% high samples and 32% low samples, a clear indication that there is a skew that exists as most students require high intervention. The skew has important implications on the evaluation criteria. To balance this skew, a second dataset collected from a city setup in well performing schools will also be used.

As explained earlier, student intervention level prediction is in effect a binary classification task with the goal of categorizing the students into two groups: 1) high intervention and 2) low intervention depending on whether the student is likely to score less than 250 marks or more. Since there is a skew in the data towards the high intervention, it is likely that a simple classifier will assign high interventions and give high classification accuracy even with very poor low intervention class prediction accuracy. For this reason, this study employs three different measures: the percentage accuracy, the receiver operating characteristics (ROC) curve and the F-measure. The ROC curve [6] is a useful graphical technique for organizing, visualizing and selecting classifiers based on their performance. The curves have an attractive property of being insensitive to changes in class distribution (see Figure 1 for a sample plot). The F-Measure [19] is a single measure for performance that deals with three types of errors simultaneously. It combines precision which deals with substitution and insertion errors, and recall which deals with substitution and deletion errors.

4.6 Feature Reduction

Feature reduction seeks to determine an optimal feature subset. As much as we use all the available student features to predict the intervention levels, education stakeholders are also keen on discovering the features that most affect academic performance. Their greatest interest is to know which feature subset is optimal or more indicative of high intervention among the rest of the features. For this reason, the study has employed a correlation-based method.

Correlation [33] is applied widely in machine learning to determine relevance. This study adopts the information gain measure to determine the features that correlate more to the final exam mark and are ranked as shown in Figure 2.

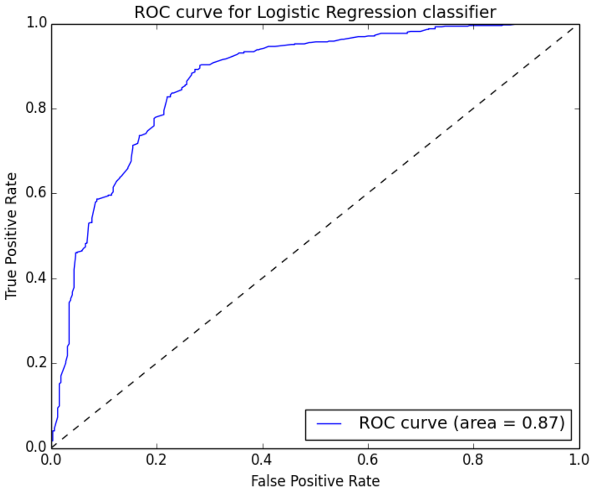


Figure 1: ROC curve for intervention prediction using logistic regression for KCPE marks

5. EXPERIMENTS AND RESULTS

This section discusses the results of experiments carried out to evaluate the students' intervention level prediction models. The experiments have two goals: 1) to measure intervention level prediction accuracy and 2) to analyze aspects of the students' data so as to derive insights that would be important to the education stakeholders. Specifically, of interest is identifying the features that are more important for the prediction task.

5.1 Intervention Prediction Accuracy

Table 1 shows the comparison of the three metric values obtained using eight classifiers for intervention predictions. We use various classifier implementations in WEKA [11] with its default settings. From the experiments performed, logistic regression is seen to have the highest percentage accuracy, ROC area value and F-measure (see Table 1).

Table 1: Comparison of classifier performance for intervention prediction in KCPE total marks using the complete data set of 22 features

Algorithm	Accuracy	ROC area	F-measure
Lazy(LWL)	83.4707	0.828	0.834
Multi. perceptron	79.761	0.828	0.794
Logistic	83.883	0.878	0.836
Tree (J48)	82.6051	0.779	0.819
Random Forest	81.3273	0.841	0.811
Bayes Net	83.2646	0.870	0.830
Naïve Bayes	73.9901	0.835	0.748
SMO	83.388	0.800	0.832

5.2 Determining a Reduced Feature Subset

Figure 2 shows a chart of all the features of the dataset that are fed into the feature selection algorithm with their corresponding information gains. The larger the value of information gain, the more strongly relevant the feature is to the training.

The features and their correlation values as given by the information gain algorithm are: test scores (0.21653), student's gender (0.02252), shortage of teachers (0.01807), student's motivation (0.01613), family income (0.0133), student age (0.01185), study time (0.0108), teacher attitude (0.00972), student absenteeism (0.00725), teacher commitment (0.00612), parent encouragement (0.00611), student education attitude (0.0045), school facilities (0.00895), command of English (0.008), distance to school (0.00584), student discipline (0.00584). As seen in Figure 2 test marks give the longest bar because it is overly co-related with the final exam mark. The research interest therefore is to determine which features for each student correlate with the test marks which in tern correlates with the final exam marks [20].

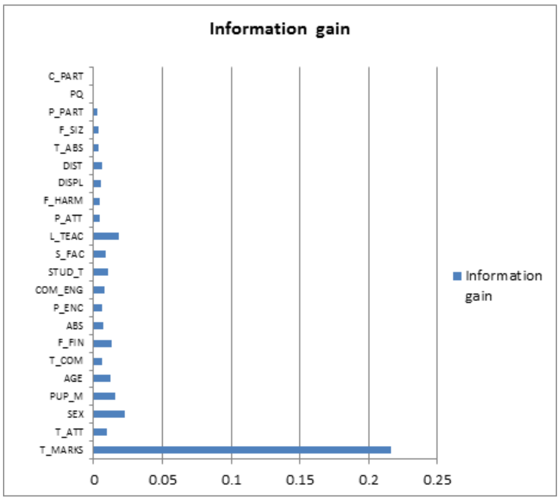


Figure 2: Information gain for the features

5.3 Intervention Prediction Models using the Optimum Feature Subset

The features obtained using the information gain approach are: test marks, gender, age, student motivation, study time, family income, teacher attitude and shortage of teachers.

Experiments are conducted to determine the best classifier using the eight-feature subset identified as potentially optimal. Table 2 shows the results of the experiments. Logistic regression still achieves the best prediction accuracy with performance similar to that of the entire feature set. This confirms that the 8 features are strongly indicative of future performance in KCPE examinations. It is also interesting to note that some classifiers (i.e. multilayer perceptron, SMO, naïve bayes, random forest) apparently have improved prediction accuracy, further confirming that the features eliminated are actually redundant and their presence may degrade the prediction accuracy.

To determine the impact of test marks, which is overly co-related with the target, a set of experiments are carried out without it. As seen in Table 3, the prediction accuracy reduces. For the highest performing, logistic regression, Multilayer perceptron and J48 the reduction is 12.9107, 12.7958 and 12.9843 respectively. This proves test marks is strongly relevant which means it is necessary, it cannot be removed without affecting the original conditional class distribution. The remaining features though still give 70.7749% accuracy for logistic regression which could give a fare classification.

5.4 Cost benefit analysis

This is a graphical analysis of the benefits associated with using the model summed up and the costs associated with it subtracted. Figure 3 shows the curves related to cost/benefit analysis. The one on the left is the threshold curve, while the one on the right is the cost benefit curve. The lowest point of the curve marked 'X' is the minimum cost/benefit point. This is also the point at which the classifier has the highest accuracy 83.8829%. At this point the cost incurred by the classifier for misclassifying 159 Low Intervention students and 232 High Intervention students (as shown in the confusion matrix) is 391 units. The cost that would be incurred if the students were randomly classified is 1028.48 units. The gain or profit obtained from using the classifier is therefore 637.48 units, which is a substantial profit.

Table 2: Comparison of classifier performance for intervention prediction in KCPE total marks using the optimal data set of 8 features

Algorithm	Accuracy	ROC area	F-measure
Lazy(LWL)	83.4707	0.835	0.834
Multi. perceptron	83.3059	0.863	0.829
Logistic	83.6356	0.874	0.834
Tree (J48)	83.2234	0.799	0.830
Random Forest	81.6573	0.851	0.813
Bayes Net	83.0173	0.871	0.826
Naïve Bayes	82.6876	0.866	0.827
SMO	83.4707	0.803	0.833

Table 3: Comparison of classifier performance without the most influencial feature (Test marks)

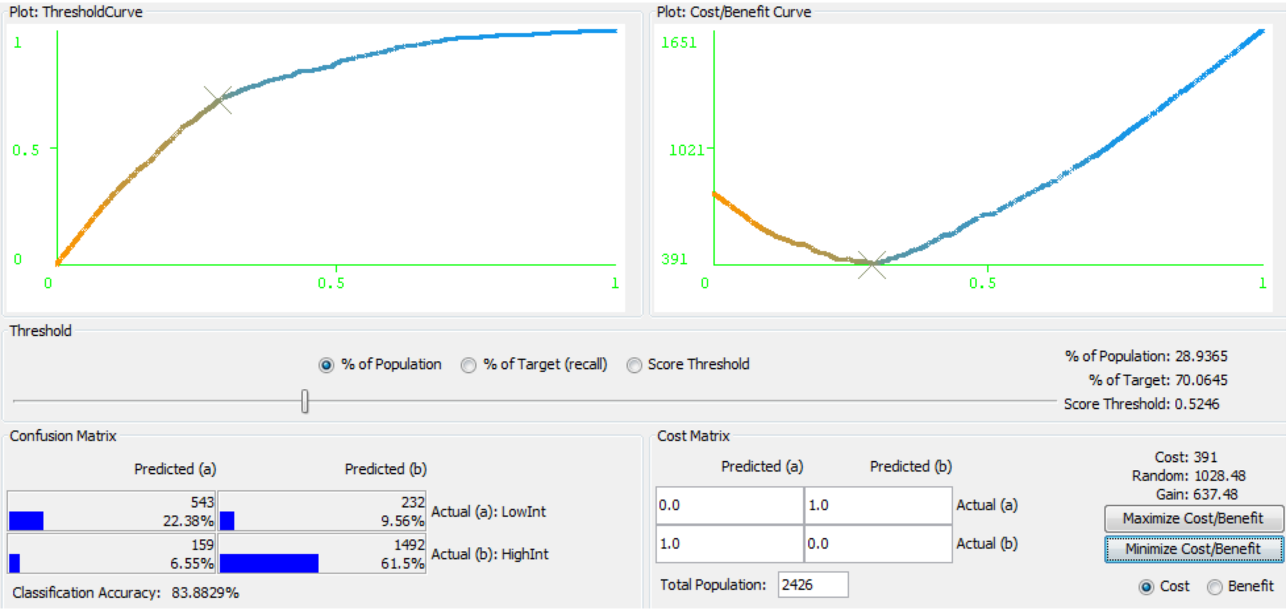
Algorithm	Accuracy	ROC area	F-measure
Lazy(LWL)	67.9308	0.660	0.559
Multi. perceptron	70.6101	0.694	0.674
Logistic	70.7749	0.691	0.672
Tree (J48)	70.2391	0.664	0.679
Random Forest	68.7964	0.698	0.663
Bayes Net	69.8681	0.684	0.638
Naïve Bayes	68.8376	0.697	0.691
SMO	67.601	0.560	0.556

6. DISCUSSION

In this section we discuss the experimental results that have potential value to education stakeholders.

First, the study results show that a student's level of intervention can be predicted with reasonable accuracy (see in Table 1. percentage accuracy: 83.88%, ROC area: 0.878 and F-Measure: 0.836) using a simple logistic regression classifier. A significant point to note is that the features are not time bound, hence predictions can be performed as early as for Standard Six students. Early prediction will be advantageous to allow enough time for initiating intervention measures.

Second, the study can claim the possibility of determining a reduced feature subset quickly using a correlation-based



Threshold

● % of Population ○ % of Target (recall) ○ Score Threshold

% of Population: 28.9365
% of Target: 70.0645
Score Threshold: 0.5246

Confusion Matrix

Predicted (a)	Predicted (b)	
543 22.38%	232 9.56%	Actual (a): LowInt
159 6.55%	1492 61.5%	Actual (b): HighInt

Classification Accuracy: 83.8829%

Cost Matrix

Predicted (a)	Predicted (b)	
0.0	1.0	Actual (a)
1.0	0.0	Actual (b)

Total Population: 2426

Cost: 391
Random: 1028.48
Gain: 637.48

● Cost ○ Benefit

Figure 3: Cost benefit analysis

method. The results from Tables 1 and 2 demonstrate that the eight features (test marks, gender, age, student motivation, study time, family income, teacher attitude and shortage of teachers) give reasonably accurate results when used on the same classifiers and can save computational time.

Third, the logistic regression model outperforms the other seven models. This is seen in the results when the full data set is used and when the reduced data set is used. This implies that logistic regression is the most suitable prediction model for the type of data set being used in this study.

This study plans to carry out further work on three fronts: 1) to make use of another data set from city schools in Kenya for the purpose of comparing and testing the methods used to see whether they are generalizable, 2) to improve the overall prediction accuracy and 3) to develop a mobile intervention prediction tool that will be used for practical purposes in rural regions.

7. CONCLUSION

This study is about predicting students who require high intervention as early as Standard Six. These are the students classified as the high intervention group or the group that will score below 250 marks in KCPE. The study is carried out using data collected from a rural region in Kenya with a history of low academic performance. Using a set of test marks, students' personal factors, school and family factors, the study constructs intervention level prediction models that are able to identify students who need high intervention with a reasonable degree of accuracy. The key observations from the experiments are: that predictions for KCPE exam performance may be made with a reasonably good accuracy, that it is possible to determine a greatly reduced feature subset which can achieve predictions similar to using all the features. Overall, this study has shown that data from a rural region can be used to build data-driven intervention prediction models for academic performance. This study hopes to motivate education stakeholders to initiate early intervention measures in order to assist as many students as possible to achieve above the national average exam mark.

8. ACKNOWLEDGMENTS

Our thanks to Hasso Plattner Institute for funding this work, to the university of Cape Town for providing a conducive environment for doing this research, and to the Kenyan NACOSTI for allowing us to collect data.

9. REFERENCES

[1] L. S. Affendey, I. Paris, N. Mustapha, M. N. Sulaiman, and Z. Muda. Ranking of influencing factors in predicting students' academic performance. *Information Technology Journal*, 9(4):832–837, 2010.

[2] N. Anozie and B. W. Junker. Predicting end-of-year accountability assessment scores from monthly student records in an online tutoring system. Educational Data Mining: Papers from the AAAI Workshop. Menlo Park, CA: AAAI Press, 2006.

[3] E. C. Barnes. *The paradox of predictivism.* Cambridge University Press, 2008.

[4] C. V. Bratu, T. Muresan, and R. Potolea. Improving classification accuracy through feature selection. In *Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on*, pages 25–32. IEEE, 2008.

[5] A. Chamillard. Using student performance predictions in a computer science curriculum. In *ACM SIGCSE Bulletin*, volume 38, pages 260–264. ACM, 2006.

[6] T. Fawcett. Roc graphs: Nootes and practical considerations for researchers. *Machine learning*, 31:1–38, 2004.

[7] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

[8] I. M. Galván, J. M. Valls, M. García, and P. Isasi. A lazy learning approach for building classification models. *International Journal of Intelligent Systems*, 26(8):773–786, 2011.

[9] M. Gardner and S. Dorling. Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.

[10] P. Golding and O. Donaldson. Predicting academic performance. In *Frontiers in education conference, 36th Annual*, pages 21–26. IEEE, 2006.

[11] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD exploration newsletter*, 11(1):10–18, 2009.

[12] M. A. Hall. Feature selection for discrete and numeric class machine learning. 1999.

[13] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic. Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study. *International Journal of Computational Intelligence Systems*, 5(3):597–610, 2012.

[14] B. Kaani. *Nature and prevalence of reading difficulties among school- dropouts: A case of selected school areas in Chipata District*. PhD thesis, 2014.

[15] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. Mlc++: A machine learning library in c++. In *Tools with Artificial Intelligence, 1994. Proceedings., Sixth International Conference on*, pages 740–743. IEEE, 1994.

[16] S. Kotsiantis, C. Pierrakeas, and P. Pintelas. Efficiency of machine learning techniques in predicting students' performance in distance learning systems. *Educational Software Development Laboratory Department of Mathematics, University of Patras, Greece*, 2002.

[17] S. B. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, 37(4):331–344, 2012.

[18] V. Y. Kulkarni, M. Petare, and P. Sinha. Analyzing random forest classifier with different split measures. In *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012*, pages 691–699. Springer, 2014.

[19] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. 1999.

[20] M. Mgala and A. Mbogho. Selecting relevant features for classifier optimization. In *Advanced Machine Learning Technologies and Applications*, pages 211–222. Springer, 2014.

[21] K. Mundy. *Education for all and the new development compact.* Springer, 2006.

[22] M. K. Mzuza, Y. Yudong, and F. Kapute. Analysis of factors causing poor passing rates and high dropouts rates among primary school girls in malawi. *World Journal of education*, 4(1):p48, 1999.

[23] F. T. Ngo, R. Govindu, and A. Agarwal. Assessing the predictive utility of logistic regression, classification and regression tree, chi-squared automatic interaction detection, and neural network models in predicting inmate misconduct. *American Journal of Crime Justice*, pages 1–28, 2014.

[24] J. Platt et al. Sequential minimal optimisation: A fast algorithm for training support vector machines. 1998.

[25] C. Romero, S. Ventura, P. G. Espejo, and C. Hervás. Data mining algorithms to classify students. In *EDM*, pages 8–17, 2008.

[26] R. Rumberger and S. A. Lin. Why students drop out of school: A review of 25 years of research, 2008.

[27] C. P. Smith. Relationships between achievement-related motives and intelligence, performance level, and persistence. *The Journal of abnormal and social Psychology*, 68(5):523, 1964.

[28] T. Soman and P. O. Bobbie. Classification of arrhythmia using machine learning techniques. *WSEAS Transactions on computers*, 4(6):548–552, 2005.

[29] A. Tamhane, S. Ikbal, B. Sengupta, M. Duggirala, and J. Appleton. Predicting student risk through longitudinal analysis. In *Proceeding of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1544–1552. ACM, 2014.

[30] J. Tooley, P. Dixon, and J. Stanfield. Impact of free primary education in kenya a case study of private schools in kenya, kibera. *Education management Administration & Leadership*, 36(4):449–469, 2008.

[31] J.-P. Vandamme, N. Meskens, and J.-F. SUperby. Predicting academic performance by data mining methods. *Education Economics*, 15(4):405–419, 2007.

[32] A. K. Waljee, P. D. Higgins, and A. G. Singal. A primer on predictive models. *Clinical and translational gastroenterology*, 5(1):e44, 2014.

[33] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research*, 5:1205–1224, 2004.