Investigating mobile based prediction modelling of academic performance for primary school pupils: a data mining approach.

PhD Research Proposal
AUGUST 2013

ICT for Development Research Laboratory
University of Cape Town
Department of Computer Science.

Mvurya Mgala
Sign ……………………….. Date …3/6/2013…………………………

Supervisors

Dr Audrey Mbogho
Sign ……………………….. Date …4/6/2013…………………………

Professor Gary Marsden
Sign ……………………….. Date …3rd June 2013………………………
ABSTRACT

Low academic performance of primary school pupils in some regions has presented a worrying trend. Research has shown this to be a widespread problem among developing nations. The problem has been attributed to many factors, ranging from students’ personal factors, teacher factors, school factors, to family background factors. Studies have been conducted on the various causes using different approaches in a bid to tackle the challenge of this trend. Albeit the many worthy attempts to temper the problem, it has persisted.

It is proposed that the first step towards intervention is to appreciate that there are many predictors of academic performance that affect different clusters of pupils. This calls for a thorough analysis of factors that affect academic performance among pupils. Knowledge discovery from databases (KDD) is a modern technique for turning data into knowledge. This study will apply methods and techniques associated with KDD for making sense of data. It is now common knowledge that data mining methods exist for pattern discovery and extraction. The patterns discovered can be used to build prediction models for estimating the values of future cases.

This study aims at understanding how the methods and techniques of KDD can be tapped to build a prediction model for predicting grades of the pupils. The study will focus on revealing the impact of the factors surrounding pupils’ academic performance. Many of the studies found in literature have focused on determining the causes of low academic performance in primary schools. This research will go further to determine the level of impact of these causes. The focus will be to discover the causes with the highest impact which can be used to build a prediction model. The model will be used by education stakeholders to acquire advance knowledge of pupil’s performance with a view to come up with intervention measures and informed decision making. In the course of carrying out the study, the contribution and challenges when designing the prediction model will be highlighted.

The study will focus on two key subjects; English and Mathematics, it will be carried out with primary schools in Kenya. The process of KDD will be followed. A number of other methodologies will be used at each stage of the process, including semi-structured interviews, focus groups, questionnaires and secondary data extraction. Testing of the prediction model will be conducted using part of the data collected. Evaluation of the model will be done using the confusion matrix. A mobile prototype artefact will be designed and evaluated. The study will be conducted following ethical practices as spelled out by UCT rules and regulations.

The original contribution of this research will be the design and creation of a prediction model and a mobile artefact for predicting primary school pupil academic grades. Through review of literature from most major journals, no other such work has been done especially in the developing world.
1. Background of the study

In the developing world, most educational structures are modelled after the British education system of 7-4-3-2-3 or after the US educational system of 8-4-4 [15]. In Kenya, the 8-4-4 system is used. A pupil learns for eight years in primary school; from class one to class eight. They then sit for the Kenya certificate of primary education (KCPE). For a student to ascend to the next level of four years in secondary, they have to pass this examination. This explains the emphasis on high academic performance. Low academic performance among school going children has long been a disturbing trend. This is evidenced by research on the reasons of low academic achievement [12, 14, 28, 26, 42, 43, 46].

Research in Ghana indicates that poor academic performance of pupils has been a concern for many years [26]. The study reported that schools in one district performed consistently poorly in the basic education certificate examinations. The same study identified factors that cause low performance to be attributed to teacher factors, pupil factors, parent factors and school factors. Research carried out in Kenya indicates that low academic achievement is an important topic for developing countries where new policies such as free primary education (FPE) have been implemented and curricula reviewed to promote greater educational access and attainment [11].

Many attempts have been made to determine the causes of low academic achievement. Such causes originate from the students, the school and the family [28], pupil gender, primary school type and household socio-economic status [9], and poverty [35]. The causes are indeed very complex, have many aspects and are interrelated, which aggravates the problem of low performance.

The reasons for low performance are complex [46], hence the attempts to study the problem using data mining techniques [5]. These techniques are part of the process of knowledge discovery from databases. The concept of knowledge discovery from databases was first coined at the first KDD workshop in 1989[52]. Early papers on data mining to Knowledge discovery in databases [52, 53] paved the way for further research in the application of the concept.

Data mining techniques have been applied in some cases in the field of education [1, 23, 31, 39, 51]. However, the literature shows this has mainly been done in the developed world, mostly focusing on online learning. This study focuses on the traditional classroom learning, being the method in use in the developing world.

Romero and Ventura [5] classified the education system as traditional classroom and distance learning. Traditionally, decision making involved observing the student’s behaviour, analysing historical data and estimating the effectiveness of pedagogical strategies. Because the volumes of data have increased and there is a lack of powerful and easy to use data analysis tools in the developing world, important decisions are made based not on the information rich data but rather on the decision maker’s intuition [32].

This study will focus on facilitating decision making by providing stakeholders with a tool to extract and evaluate knowledge hidden in the large amounts of data [32]. An education institution has many diverse and varied sources of information [55]. This abundant data available in primary schools can provide an opportunity to use data analysis tools for in depth analysis such as classification and clustering. A survey of decision making by data mining tools in education [2, 20, 25], shows that the majority of users were in higher education. Little has been done in primary schools.

The motivation behind this research is the exploration of a data mining tool that would predict performance for primary school pupils, in an attempt to help decision makers to make interventions and reverse the trend of poor academic performance. The benefits of data mining to gain deeper understanding of patterns previously unseen using current available reporting capacities would help turn around the trend of low academic performance in the developing world primary schools.
Finally, it is not the intention of this research to attempt to suggest changes in the education structures or the examinations systems but to strengthen the current education systems and complement the current examination systems by aiding the stakeholders to make informed and accurate decisions. The research aims to give widespread access to the prediction tool in a bid to foster better decision making and proposing of education policies.

2. Literature review
   2.1. Existing models for predicting students’ academic performance.

   With the general emphasis on record keeping in traditional classroom education, it is not surprising that knowledge discovery from databases in education has attracted considerable attention in recent years. In an attempt to understand the problem of discrepancy in academic performance, studies have focused on anything from finding the causes of low academic performance to how to turn around low performing schools. While some research has focused on understanding the causes [11, 26, 36, 42, 44, 46], other work has sought to design prediction models for academic performance [24, 30, 40, 49, 50].

   Many of the factors that cause low academic achievement can be grouped into four groups [26];
   - **School factors**
     These include: limited teaching materials, inadequate textbooks and insufficiently trained teachers;
   - **Teacher factors**
     These include: lateness to school, absenteeism, use of local language in teaching, inability to complete the syllabi, low interest in children’s understanding of the lesson and not hardworking;
   - **Pupil characteristics**
     Which include: absenteeism and truancy, use of local language in the classroom, lack of interest in the lesson and little help with studies in their homes.
   - **Parent factors**
     These include: inability to provide meals, textbooks and basic school needs, lack of interaction with children’s teachers and low involvement in parent teachers association.

   The focus of this study is to identify the differences between the high achieving and the low achieving schools in order to extract the causes of low academic achievement.

   Much of the latest design oriented work on prediction models is biased towards higher education and online learning [4, 6, 18, 31, 33]. However, there is some work based on traditional classroom education. An early study by Hostetler [50] predicted students’ success in an introductory programming course in a University. The study sought to find out the extent of a student’s aptitude in computer programming by measuring certain cognitive skills, personal traits and past academic achievement. Results from the prediction model confirmed past skills and the student’s GPA to be the factors most closely associated with success. The study used multiple regression developed from five predictors. The emphasis in this study [50] was to predict success so as to effectively counsel the students. The study did not focus on causes such as parent and teacher factors.

   Another early study predicted academic performance using neural networks [30]. Two neural networks for general mapping problems, back propagation and counter propagation, were trained to predict students’ grades in Calculus I from placement test responses. Results showed that back propagation networks with a sufficient number of hidden units (25 to 31) achieved very high levels of accuracy in predicting outcomes for either the training data or testing data that is reasonably similar to the training data. In [49] the researcher similarly trained a back propagation, feed-forward neural network to predict student performance in a large undergraduate computer science subject at the
University of New South Wales. In this work, the prediction used partial grades obtained during teaching session to predict the final grade. In both [30] and [49] test scores were used as the only predictor. This study will use multiple predictors.

A recent study sought to apply several data mining methods, including, discriminant analysis, neural networks, random forests and decision trees, to predict students’ academic success [24]. The study classified university students into three groups: the ‘low risk’ students, the ‘medium-risk’ students and the ‘high-risk’ students. The best prediction was made by discriminant analysis followed by the neural network. Closely related to this is a study done on machine learning techniques [45]. The work presented a case study of predicting students’ marks. Students’ key demographic characteristics and their marks in a small number of written assignments constituted the training set for a regression method in order to predict the student’s performance. The study used model trees, neural networks, linear regression, and locally weighted linear regression and support vector machines for the purpose of comparison. A prototype version of a software support tool was built for use by tutors. The focus of both studies was to get the most significant variables correlated to academic success among first year university students.

In Learning Management Systems (LMS), prediction of students’ grades has also been studied [3]. In this work, a grammar guided genetic programming algorithm, G3P-M1 was applied to predict if the student will pass or fail a module considering the time dedicated, the number and type of activities done by the student during the course including; quizzes, assignments and forums. Results of the experiments show that G3P-M1 achieves better performance with more accurate models and a better trade-off between such contradictory metrics as sensitivity and specificity. In another LMS, research has been conducted to identify successful learners from interactive behaviour [21]. The study argues that the interaction behaviours of successful, high-achieving learners when using LMS are different compared to the behaviour of learners who are having more difficulty mastering the course materials. The study was carried out among university students. The results of the study indicate that the results of formal assessments are unnecessary to accurately predict which learners will be successful with the course content and which ones will struggle with the course. The study applied decision trees constructed using the study data.

In another study of a traditional classroom, in the school of Computing and Information Technology in Jamaica [40], prediction of academic performance of students was conducted using the relationship between students’ demographic attributes (age and gender), qualification on entry, aptitude test scores, and performance in first year courses. The researcher was able to examine their overall performance in the program using stepwise multiple regression analysis. This study used admission criteria to the university and prediction of future success so as to come up with an admission policy.

Finally, studies on prediction modelling in Ethiopia used a Bayesian network to predict the performance of high school students [45]. The emphasis was to accurately estimate the student’s future performance so as to provide them with adequate assistance in the learning process. The study proved that Bayesian networks can successfully be used in education as a classification approach. The attributes used included; gender, group work attitude, interest in Mathematics, achievement motivation, self-confidence, shyness, English performance and Mathematics performance. The study was able to automatically predict performance of students in secondary school. Another work in Nigeria used artificial neural networks to predict performance in an engineering course in the university [54]. In these studies, the focus was to predict the performance so as to provide the students with adequate assistance in the learning process.

Much of the work found in literature and surveyed here has focused on secondary and tertiary education. In addition, most of the methods applied have included neural networks, regression analysis, discriminant analysis, random forest, decision trees and grammar guided genetic
programming algorithm. This proposal differs in that the focus is on primary school performance, which forms the foundation on which all of the student’s future learning is based. In addition, the method that will be applied is Bayesian networks which do not appear much in the literature on academic performance prediction, yet is has many benefits. Ranging from its graphic nature, making it easy for exploration and understanding to their solid theoretical base [13]. This provides an opportunity for research into what kind of a prediction model will be suitable for primary school pupils. The research will not only focus on student attributes but will look at other factors such as school factors, teacher factors and parent factors to predict pupil grades.

2.2. Classification algorithms for predicting academic performance

Studies have been conducted on the application of data mining techniques in traditional education. One of the earliest studies was on discovering enrolment knowledge in university databases [2]. Knowledge discovery was applied in the form of statements such as; “Pattern P holds for data in Range R”. A recent study [25] analysed the impact of curriculum revisions in a Brazilian university. A number of techniques were applied such as summarization, association and classification. In a related study, a scoring function based on association was used [55]. Another study done for higher education on a comprehensive analysis of students characteristics, proposed to use different unsupervised and supervised data mining algorithms (C5.0 and Genetic Algorithm) to do clustering and prediction [5, 20]. Bayesian network classifiers have also been used to predict academic performance [41].

Classification tasks are the most widespread usage of supervised data mining. In a large dataset, the aim is to generate a classification for the target variable for new records not currently in the database. The algorithm learns which combinations of variables are associated with which classes of the target categorical variables. The algorithm would then look at new records, in the test and validation sets. Based on the classification in the training set, the algorithm would assign classifications to the new records [10]. The relationship discovered is represented in a structure known as a model. Models describe and explain phenomena which are hidden in the dataset and can be used for predicting the value of the target attribute knowing the values of the input attributes.

Classification problems have been widely studied in statistics and artificial intelligence. A variety of different common classification approaches used in artificial intelligence include decision trees, neural networks, genetic algorithms and Bayesian networks [41]. This study will involve classifying pupils into categories of grades; therefore it will seek to apply one of these approaches. A brief concept of each of them is given below.

Decision trees [14], also called classification and regression trees, are a flexible and robust analytical method which deal with nonlinear relationships, high-order interactions and missing values. The method is also simple to understand and gives easily interpretable results. Trees explain variation of a single response variable by repeatedly splitting the data into more homogeneous groups using combinations and explanatory variables. Trees are represented graphically, which aids in exploration and understanding. Decision tree construction performed on very large datasets leads to bushy or meaningless results [34]. This study will involve a large dataset; which will result in many branches because of the large number of attributes. Such a large number of branches is computationally expensive, making this approach a poor choice for the proposed study.

Neural Networks [54] are inspired by the structure of the brain. A neural network consists of a set of highly interconnected entries, called processing elements or units. Each unit is designed to mimic its biological counterpart, the neuron. Each accepts a weighted set of inputs and responds with an output. There is no need to specify a function to which the data are to be fitted. The function is an outcome of the process of creating a network. The network is able to capture almost arbitrarily
nonlinear relationships. Neural networks are also readily updated as more historical data becomes available. Although the prediction accuracy is high and the networks are robust, they may be unsuitable for this study because they require long training time, the learned functions are difficult to understand, and it is difficult to incorporate domain knowledge.

Genetic Algorithms [8] are heuristic learning models based on principles drawn from natural evolution and selective breeding. Features that distinguish genetic algorithms from other search methods include: (a) population of structures that can be interpreted as candidate solutions to the given problem (b) the competitive selection of structures for reproduction, based on each structure fitness as a solution to the given problem (c) idealised genetic operators that alter the selected structure in order to create new structures for further testing. In many applications these features enable genetic algorithms to rapidly improve the average fitness of the population and quickly identify high performance regions of very complex search spaces. Genetic algorithms however are a method of search, often applied to optimisation or learning [48].

Bayesian Networks [40], also known as belief networks or causal probabilistic networks, draw their roots from a branch of probability and statistics known as decision theory. Some attractive aspects of the Bayesian include; (a) each training vector can be used to update probability distributions which in turn affect the probability that a given hypothesis is true (b) provide more flexibility in that a hypothesis does not get completely ruled out from few examples (c) prior knowledge can be easily implemented in the form of prior probability distributions [13]. The structure of Bayesian networks is a graphical illustration among the set of variables that it models. It consists of a directed acyclic graph and conditional probability distributions associated with the vertices of the graph. The directed acyclic graph represents the structure of the application domain. Nodes, which are usually drawn as circles or ovals, represent random variables and arcs represent directed probability dependencies among them [19, 54]. Every vertex is associated a table of conditional probabilities of the vertex given each state of its parents. Bayesian Classification calculates explicit probabilities for hypotheses making it the most practical approach to the proposed study. In addition they predict multiple hypotheses weighted by their probabilities and can provide standard optimal decision making against which other methods can be measured.

A classification approach will be used in this study because the problem of prediction will require classification of observed instances into predominant categories [36]. Pupils will be classified into categories of their performance. The study proposes to use Bayesian networks because of their solid theoretical base, and their graphical nature which aids exploration and understanding [13]

3. Outcomes

3.1 Statement of the problem

There is an apparent lack of design and testing of a prediction model artefact in the developing world geared towards predicting academic performance for primary school pupils. This research will be carried out to fill this gap.

The literature suggests that tittle has been done to tap into the power of prediction classification approaches and artificial intelligence algorithms in addressing the problem of low academic performance in primary schools. Specifically, little research has been conducted with the view to constructing a prediction model intervention that:

- Provides an alternative to dependence on final examinations to determine students’ abilities
- Provide means by which decision makers can make accurate decisions and come up with working policies.

It is hoped that this research will offer a step towards solving the low performance problem.
The experimental work that will be carried out in this research will involve primary school pupils who are in their final year. This research will be undertaken to provide additional insight into carrying out experimental research in developing nations, especially one that involves designing of a mobile phone prediction artefact.

3.2 Research Questions
In order to address the purpose of this research, the study will seek to answer the following questions:

a) How can the Bayesian classifier be modelled from the primary schools data?

b) How can a Bayesian model be used for prediction of primary school pupils’ academic performance?

c) What mobile phone application artefact can be designed to automatically predict academic performance?

3.3 Research Objectives
The main objective of this research is to design and create a prediction artefact that predicts pupils’ academic performance early before the final examination. In doing so and in order to address the research questions above, the study will seek:

a) To develop the Bayesian classifier model from the data sets aside for training

b) To predict academic performance of the pupils using the test data set aside.

c) To develop a mobile phone application software artefact to automate the prediction process.

3.4 Significance of the study
Findings of the study will contribute to the field of computer science in the following way:

- Provide a process to design and create a prediction model artefact that predicts academic performance for primary school pupils

Findings of the research will contribute to the field of Knowledge discovery from databases in the following ways:

- Expose the social and technological issues that influence the successful design, implementation and adoption of an academic performance prediction model.

- Support and enrich the classification approaches in implementation and adoption of prediction systems.

4. Methodology

4.1 Target Population
The research will be targeted towards primary school pupils. In this regard the major group involved in the experiment will be final year pupils in 10 primary schools in Kwale County in Kenya. All schools will be approached upon obtaining ethical clearance from the respective institutions.

4.2 Sampling Design
Stratified sampling will be used since the target group is known, being final year pupils of the selected 10 primary schools in Kwale County. Identification and measurement of the attributes will be done by first collecting a list of the attributes through literature and having semi-structured interviews with 18 education officers in the county. Questionnaires will be given to 50 teachers and 200 pupils in the 10 primary schools. Based on the collected information, the best attributes will be selected.
Data on the pupils’ previous performance will be collected from the schools records office (e.g. report forms from class six). The questionnaires will be used to measure teacher and student attributes. Likert scale will be adopted to register the extent of agreement or disagreement. Questionnaires will be filled out in the presence of the researcher.

4.3 Design Tools
MySQL database will be used to prepare and store the data. A belief network modelling software such as GeNiE 2.0 will be used to design the prediction model. A mobile phone application will be created using Android for the artefact. All these tools are open source and readily available.

4.4 Data Collection
A qualitative and experimental approach will be used, drawing upon a combination of structured interviews, focus groups, questionnaires and secondary data search.

Structured Interviews: These will be conducted as a preliminary survey to gather the attributes that contribute to low performance.

Focus Groups: This will consist of education officers and head teachers of primary schools in Kwale County. They will be detailed on the concept of predicting academic performance and the prediction artefact. The head teachers will be used regularly throughout the research to give feedback on user satisfaction.

Interviews and Questionnaires: Questionnaires will be given to the pupils to gather their personal and family background attributes and to the teachers to gather teacher attributes and school attributes. Both semi-structured interviews and questionnaires will be conducted to collect data to be used in the design of the prediction model and the mobile artefact in predicting pupil performance. The questionnaires will be fed into the Lime Survey for analysis [38].

Secondary data search: Secondary data will be obtained from the schools’ records offices. The target data is marks for Mathematics and English from previous report forms for end of term and end of year examinations.

5. Testing and Evaluation
Testing: In testing the algorithm will be fed with new records from the test set, for which no information about the classes is available. Based on the classification in the training set, where information is available about the predictor variables and the (already classified) target variable, the algorithm will assign classification on the new records, as described in [10].

Evaluation: Evaluation will be done to determine how well the classification algorithm is performing. A confusion matrix of correct and incorrect classification will be drawn. The rows will show the actual values and the columns will show the predicted values [10]. Percentages of correct classification will be obtained as was done in [41]. The mobile base prediction artefact will be evaluated using a field-based usability evaluation methodology as was done in [7]. The head teachers from the 10 primary schools will be given the mobile base prediction artefact. At the end of the term, questionnaires will be given to them to collect their feedback on usability and accuracy of prediction.

Impact evaluation: An impact evaluation assesses the extent to which and intervention has caused desired changes in the intended group, it assesses the outcome of change resulting from an intervention [27]. This study will randomly select five pupils per school who will be in their final year. The prediction artefact will be used to determine their likely outcome in the final examination. Some intervention measures will be put in place and the pupils’ final results compared with the
predicted grades. The study will therefore be able to propose possible interventions to the stakeholders.

6. Work Plan

<table>
<thead>
<tr>
<th>Date</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013 July</td>
<td>Semi-structured interviews</td>
</tr>
<tr>
<td>2013 July</td>
<td>Questionnaires to teachers and students</td>
</tr>
<tr>
<td>2013 October</td>
<td>Data preprocessing</td>
</tr>
<tr>
<td>2013 December</td>
<td>Bayesian classifier Model developed using training set.</td>
</tr>
<tr>
<td>2014 March</td>
<td>Prediction model developed using test set.</td>
</tr>
<tr>
<td>2014 July</td>
<td>User centered design of Mobile application.</td>
</tr>
<tr>
<td>2014 September</td>
<td>Mobile application artefact developed</td>
</tr>
<tr>
<td>2014 December</td>
<td>Field-based evaluation of mobile artefact.</td>
</tr>
<tr>
<td>2015 April</td>
<td>Impact evaluation</td>
</tr>
<tr>
<td>2015 August</td>
<td>Thesis write-up completed</td>
</tr>
</tbody>
</table>

7. References


