International Journal of Information Sciences and Application.
ISSN 0974-2255 Volume 4, Number 1 (2012), pp. 57-76

http://www.irphouse.com

Machine Learning Approach to Determining the Influence of Family Background Factors on Students' Academic Performance

[1]Kuyoro 'Shade O., [2]Prof. Nicolae Goga Ph.D., [1]Dr. Oludele Awodele Ph.D and [1]Dr. Samuel Okolie Ph.D

[1]Department of Computer Science, Babcock University, Nigeria
afolashadeng@gmail.com
[2]University of Groningen, The Netherlands or Politenica Bucharest
n.goga@rug.nl

Abstract

Machine learning has been successfully applied to many domains such as fraud detection, medicine, banking, bioinformatics, and so on. The application of this paradigm to enhancing and evaluating the higher education tasks is a new research area. There have been various works bordering on students' performance, and related problems but the focus of this work is on applying machine learning algorithms to students' data for predictive purposes in an educational environment. We propose to develop novel approach based on machine learning algorithms to be delivered to educational institutions for guiding their planning of educational activities within the scope of increasing the academic performance of the students by taking into account their family background factors. One thousand five hundred (1500) records of students in three Nigerian tertiary institutions will be used. The students' academic performance will be measured by Cumulative Grade Point Average (CGPA) at the end of first year. Waikato Environment for Knowledge Analysis (WEKA) and See5 will be used to generate three decision tree models, Artificial Neural Networks and two rulesets. These algorithms will be compared based on their accuracy level and confusion matrices to determine the optimal model. The rules generated from the optimal model will be disseminated to educational administrators to guide in their planning activities.

Keywords: Decision trees, Neural Networks, Family Background, educational planning activities, machine learning.

AREA: Computer Science

RESEARCH FIELDS: Artificial Intelligence, Machine Learning and Data Mining

INTRODUCTION

BACKGROUNG TO THE STUDY

Machine learning has proven to be of great value in various application domains. It is especially useful in data mining problems where large databases may contain valuable implicit regularities that can be discovered automatically; poorly understood domains where humans might not have the knowledge needed to develop effective algorithms such as face recognition from images; and domains where the program must dynamically adapt to changing conditions.

Machine learning involves searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner. It draws on ideas from a diverse set of disciplines including artificial intelligence, probability and statistics, computational complexity, information theory, psychology and neurobiology, control theory, evolutionary models and philosophy.

Designing a machine learning approach involves a number of design choices such as choosing the type of training experience, the target function to be learned, a representation for this target function, and an algorithm for learning the target function from training examples.

The most commonly used machine learning algorithms are: 1. Artificial Neural Network -nonlinear predictive model that learns through training which resembles biological neural networks in structure. 2. Decision trees -tree-shaped structures that represent sets of decisions which generate rules for the classification of a dataset e.g. CART, ID3, C4.5, C5 etc. 3. Genetic Algorithms -optimization techniques that use process such as genetics combination, mutation, and natural selection in a design based on concepts of evolution. It tries to mimic the way nature works. It is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics. 4. Rule Induction -extraction of useful if-then rules from data based on statistical significance. 5. Regression Methods -identify the best linear pattern in order to predict the value of one characteristic under study in relation to another.

The application of machine learning approaches to educational data is a recent trend in research. The differential students' performance in tertiary institutions has been and is still a source of great concern and research interest to the higher education managements, government and parents because of the importance education has on the national development. Academic institutions are increasingly required to monitor their performance and the performance of their students. This gives rise to a need to collate, analyze and interpret data, in order to have evidence to inform academic policies that are aimed at improving student retention rates, allocating teaching and support resources, or creating intervention strategies to mitigate factors that may affect student performance adversely. There are a number of reasons for this: 1. Tertiary education should aim to maximize the potential of each student. Therefore, a careful examination of student outcomes against some benchmark or expected outcome may provide evidence as to whether student potential is being realized. Such insights may also help the institutions to prioritize scarce resources, to focus them on specific problem areas, 2. Institutions have an obligation to deliver value for money to the bodies that fund them, 3. Institutions are often judged by the quality of the awards

they provide; for instance, the more honours level graduates a course provides, the better the course is perceived to be. This provides additional incentive for institutions to take proactive steps to investigate students' data with a view to finding useful information.

The literature is replete with various works in machine learning area bordering on university admission, student performance, and related problems. Many of the studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data. Some of these factors seemed to be stronger than others; but there is no consistent agreement among different studies. However, all studies show that academic success is dependent on many factors; grades and achievements, personality and expectations, as well as academic environments.

From various studies that have been carried out in this area, the observed poor performance of students in tertiary institutions has been partly traced to poor academic background. However, the focus of this work is on applying machine learning algorithms to students' data for predictive purposes in educational environment to determine the influence of family background factors on the academic performance of students. One thousand five hundred (1500) records of students in three Nigerian Universities will be used. The students' academic performance will be measured by cumulative grade point average (CGPA) at the end of first year.

PROBLEM STATEMENT

The application of machine learning algorithms to family background data for educational planning purposes is a new research area. There have been research reported in the literature bordering on predicting students academic performance using machine learning algorithms but till now, there is no good procedure for educational administrators to guide their planning activities for improving the academic performance of their students taking into account the family background factors of students, neither is there research directly related to this. Our problem statement is the following: "How machine learning algorithms can be used to generate a procedure for educational administrators in guiding their planning activities for improving the academic performance of their students by taking into account the family background factors of their students".

AIM AND OBJECTIVES

The aim of this study is to use machine learning algorithms to determine the influence of family background factors on students' academic performance. The general objectives are:

- To use machine learning algorithms to determine the influence of family background factors on students' academic performance.
- To recognize patterns mapping students' academic performance to their family background.
- To develop novel methodology based on machine learning algorithms to be

delivered to educational institutions for guiding their planning of educational activities within the scope of increasing the academic performance of the students by taking into account their family background factors.

The specific objectives are:

- To demonstrate how machine learning algorithms can be applied to educational datasets
- To create database representing the students' family background.
- To model the students' academic performance based on data collected
- To compare models generated from the machine learning algorithms (CART, C4.5 and C5, ANN) using the accuracy level and confusion matrices
- To identify which model is most suitable for prediction of students' academic performance based on family background factor.
- To generate if-then rules from the optimal model.
- To examine the rules to see what can be learned from them.
- To design a predictive system framework based on the rules generated

RESEARCH QUESTIONS

- What specific machine learning algorithm best model the student performance?
- To what extent do family background factors affect students' current academic performance?
- What specific "if-then" rules can be generated to predict student performance?
- Can a framework for predictive system be developed from these rules?

SIGNIFICANCE OF THE STUDY

This research is significant for several reasons. From a scientific perspective it will apply and compare the fitness of different machine learning algorithms to a new domain, this will lead to the gain of new knowledge. From a practical perspective, the main objective of any higher educational institution is to improve the quality of managerial decisions and to impart quality education. Good prediction of student's success in higher learning institution is one way to reach the highest level of quality in higher education system. This study will provide an improved method of achieving this focusing on students' family background. This will be of great benefit to educational administration in guiding their planning activities. It is will also be beneficiary to colleges, universities, parents or guardian and the Nigerian society in the long run.

SCOPE OF STUDY

Datasets for the purpose of the study will be obtained from three tertiary institutions in Nigeria. The institutions will be selected at random. One thousand five hundred

records of students will be used.

EXPECTED CONTRIBUTIONS

The main contribution of this research is to explore the possible application of machine learning algorithms to educational datasets; thus providing better insight into the process of applying machine learning algorithms to real world large data sets. It will also provide a deeper knowledge of individual student background enabling the higher education management to take necessary actions to aid each student in improving or maintaining his/her academic performance. The results of this work will provide a useful snapshot of family background factors affecting students' academic performance as well as demonstrate a trend by which other students with similar data can be identified and assisted. It will also provide a framework of a predictive system for higher education management to aid their decision-making.

LITERATURE REWIEW

There has been significant progress in the field of machine learning bordering on its application to the areas of medicine, natural language processing, software development and inspection, financial investing, biometrics and so on.

Padberg et al, 2004 uses neural networks to estimate how many defects are hidden in a software document. Input for the models are metrics that were collected when applying a standard quality assurance technique on the software inspection document. Two key ingredients for a successful application of neural networks to small data sets identified are: Adapting the size, complexity, and input dimension of the networks to the amount of information available for training; and using Bayesian techniques instead of cross-validation for determining model parameters and selecting the final model. For inspections, the machine learning approach is highly successful and outperforms the previously existing defect estimation methods in software engineering by a factor of 4 in accuracy on the standard benchmark.

Gaweda et al, 2005 presents application of reinforcement learning to drug dosing personalization in treatment of chronic conditions. Reinforcement learning is a machine learning paradigm that mimics the trial-and-error skill acquisition typical for humans and animals. In treatment of chronic illnesses, finding the optimal dose amount for an individual is a process that is usually based on trial-and-error. The challenge of personalized anemia treatment with recombinant human erythropoietin was focused on; the application of a standard reinforcement learning method, called Q-learning, to guide the physician in selecting the optimal erythropoietin dose was demonstrated. Random exploration in Q-learning from the drug dosing perspective was addressed, and smart exploration method was proposed. Computer simulations were used to compare the outcomes from reinforcement learning-based anemia treatment to those achieved by a standard dosing protocol used at a dialysis unit.

Salah et al, 2007 survey the use of machine learning methods for biometrics applications and relevant research issues. The article focused on three areas of interest: offline methods for biometric template construction and recognition,

information fusion methods for integrating multiple biometrics to obtain robust results, and methods for dealing with temporal information. Exemplary and influential machine learning approaches in the context of specific biometrics applications were introduced to create novel machine learning solutions to challenging biometrics problems.

Rada et al, 2008 reviewed the case for knowledge-based machine learning in financial investing. While machine learning exploits knowledge, it also relies heavily on the evolutionary computation paradigm of learning, namely reproduction with change and selection of the fit. A model for financial investing was presented; review of what has been reported in the literature as regards knowledge-based and machine-learning-based methods for financial investing and a design of a financial investing system is described which incorporates the key features identified through the review of related literature. The emerging trend of incorporating knowledge-based methods into evolutionary methods for financial investing suggests opportunities for future researchers.

Sokolova et al, 2008 present applications of machine learning techniques to problems in natural language processing that require work with very large amounts of text. Such problems came into focus after the Internet and other computer-based environments acquired the status of the prime medium for text delivery and exchange. In all cases discussed, an algorithm ensured a meaningful result, be it the knowledge of consumer opinions, the protection of personal information or the selection of news reports. The elements of opinion mining, news monitoring and privacy protection as well as text representation, feature selection, and word category and text classification problems were discussed. The applications presented combine scientific interest and significant economic potential.

The literature is full of works relating to university admission, student performance, and related problems. Recently the field of machine learning is experiencing a new trend of its application to educational data especially in relation to predicting students' academic performance.

Bakare (1975) summarized the factors and variables affecting students' academic performance into the intellective and non-intellective factors, emphasizing that the intellectual abilities were the best measure. He categorized causes of poor academic performance into four major classes namely: Causes resident in society, Causes resident in school, Causes resident in the family and Causes resident in the student.

Anderson et al., (1994) studied the effect of factors such as gender, student age, and students' high school scores in mathematics, English, and economics, on the level of university attainment. According to the study, students who received better scores in high school also performed better in university. Also men had better grades than women and choose to drop from school less often.

McKenzie and Schweitzer (2001) investigated academic, psychosocial, cognitive and demographic predictors of academic performance to improve interventions and support services for student at risk of academic problems. They recommended implementing stringent record keeping procedures at the university level to enable researchers to fully examine the relationship between age, previous academic performance and university achievement.

Golding and Donaldson (2006) stated that the use of performance in first year computer science course is a possible factor, which may determine academic performance. They also showed that gender and age have no significant correlation as predictive factors.

Delavari and Beikzadeh (2004) proposed a model for using data mining in a higher educational system to improve the efficiency and effectiveness of the traditional processes. Kalles and Pierrakeas (2004) in an effort to analyze students' academic performance through the academic years, as measured by the students home work assignments, attempted to derive short rules that explain and predict success or failure in the final exams using different machine learning techniques (decision trees, neural networks, Naive Bayes, instance-based learning, logistic regression and support vector machines) and compared them with genetic algorithm based induction of decision trees.

Delavari et al (2005) proposed an analysis model and used it as a roadmap for the application of data mining in higher educational system. The model allows the decision makers to better predict which students are less likely to perform well in that specific course, or those who are less likely to be successful in it.

Adeyemo and Kuye (2006) presented an evaluation of the factors that contributed to the academic performance of students admitted into the university. The variables of interest were the entry qualification and admission mode and how these factors affect the academic performance of the students. The study indicated that the observed performance of student whose admission into tertiary institution is through the University Matriculation Examinations (UME) depends more on their respective Senior School Certificate Examination (SSCE) performance than their entry scores in the UME examination used as the basis for their admission.

Superby et al. (2006) and Vandamme et al. (2007) studied correlations of various parameters such as attendance, estimated chance of success, previous academic experience and study skills. They found out that changing process factors during a student's stay at the university plays a large part in academic performance. In addition they experimented on predicting students' performance using decision tree, neural networks and linear discriminant analysis. The rates of prediction obtained were not particularly good due to the difficulty to classify students into 3 groups, namely, high risk, medium risk and low risk, before the first university examinations.

Ogor 2007 developed a methodology by the derivation of performance prediction indicators in deploying a simple student performance assessment and monitoring system within a teaching and learning environment by mainly focusing on performance monitoring of students' continuous assessment (tests) and examination scores in order to predict their final achievement status upon graduation. Based on various data mining techniques (DMT) and the application of machine learning processes, rules are derived that enable the classification of students in their predicted classes. The deployment of the prototyped solution, integrates measuring, 'recycling' and reporting procedures in the new system to optimize prediction accuracy.

Osofisan and Olamiti (2009) investigated the academic background in relationship with the performance of students in a computer science programme in a Nigerian university. The study indicated that the grade obtained from SSCE in mathematics is

the highest determinant used by the C4.5 learning algorithm in building the model of the students' performance. The study showed that if a student does not finish his programme in the normal number of (four) academic sessions for whatever reasons he would still graduate with minimum of second class lower if he took further mathematics at SSCE examination. Students who spend more than four academic sessions in the programme and did not take further mathematics at SSCE examination are more likely to graduate with class below second class lower.

Affendey et al (2010) used attribute importance analysis to rank influencing factors that contribute to the prediction of students' academic performance. The study predict whether, a first year student will graduate higher or lower than a second class upper and compare the prediction accuracy of several classification methods. It also ascertains the major courses that contribute to the overall academic performance within the bachelor of computer science program and displayed a trend suggesting that these courses were a determining factor in predicting performance.

Kovačić (2010) used feature selection and classification trees to show the most important factors for student success. The study shows that the most important factors separating successful from unsuccessful students are: ethnicity, course programme and course block. The risk estimated by the cross-validation and the gain diagram suggests that all trees, based only on enrolment data are not quite good in separating successful from unsuccessful students.

Yu et al (2010) explore the issue of student retention in tertiary institutions from a new perspective by using three data mining algorithms -classification trees, multivariate adaptive regression splines (MARS), and neural networks. They identified transferred hours, residency, and ethnicity as crucial factors to students' retention.

Quadri and Kalyankari (2010) used decision trees to make important design decisions about the interdependencies among the properties of drop out students. The study provides examples of how data mining technique can be used to improve the effectiveness and efficiency of the modeling process. It also addressed the capabilities and strengths of data mining technology in identifying drop out students and to guide the teachers to concentrate on appropriate features associated and counsel the students or arrange for financial aid to them.

Bhardwaj and Pal (2011) constructed Bayes classification prediction model for students' performance in their study and identified the difference between high learners and slow learners student. In another study, these authors constructed classification task –decision tree to evaluate student's performance and extracted knowledge that describes students' performance thus, helps in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising/counseling earlier.

Pandey and Pal (2011) used Bayes classification to identify students who consistently perform well and the underformer, thus reduced the drop out ratio to a significant level and improve the performance level of the institution.

Marquez-Vera et al (2011) attempted to improve accuracy in the prediction of final student performance and of which students might fail. They compared ten classification algorithms and ten-fold crossvalidation, identifying white-box

classification algorithm as the optimal approach. The problem of classifying unbalanced data by rebalancing data and using cost sensitive classification was also resolved.

Other studies tried to identify the significant factors that can influence tertiary students' academic performance in a more detailed way. Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others; but there is no consistent agreement among different studies. However, all studies show that academic success is dependent on many factors, where grades and achievements, personality and expectations, as well as sociological background all play significant roles.

This study will serve as a furtherance of other studies relating to students' academic performance showing an in-depth application of machine learning algorithms on students' data as well as demonstrate the application of machine learning algorithms to large educational datasets relating to students family background which to the best of our knowledge has not been done somewhere else.

METHODOLOGY

DESIGN OF EXPERIMENT

In this research, we will focus on comparing the performance of machine learning algorithms that are trained with data relating to students family background factors with the aim of obtaining a robust predictive system for the administrations to be used in their educational planning. For collection data to train the machine learning algorithms, quantitative approach will be used. A tool will be built based on database that interfaces with See5 and WEKA for training the algorithms. For finding the optimal predictive model we will make a comparative research on the machine learning algorithms using cross-validation and misclassification errors benchmarks. For the training, 40% of the data will be used while the remaining 60% will be used to validate.

After determining the optimal algorithm, we will design the framework of a predictive system that can be used by educational administration based on if-then rules generated from the optimal algorithm. However, it is out of scope of this current research to validate the predictive system with real world data because of the time constraint. The methodology is detailed below:

DATA COLLECTION AND PREPARATION

The data set to be used in this study will be obtained from three Nigerian Universities' Students Record Systems. Both the institutions and the students' records to be used will be selected at random. The size of the dataset will be 1,500. The real-world data set from the Students Record may not store sufficient students' family background information therefore there may be need to extract some of the background information from the entrance questionnaires that are given to students to fill as part of entrance registration requirements. Other variables that may be extracted from

these questionnaires include mother's educational qualification, father's educational qualification, Sponsor, Family size, student's position in the family, mother's occupation, father's occupation, marital status of parents, and Average family income.

During data collection, the relevant data will be gathered. Once the data has been collected, it quality will be verified. Incomplete data will be eliminated and the data will be cleaned by filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Finally, the cleaned data will be stored in different tables and later joined in a single table to remove errors. A database of record will be created for the data collected.

VARIABLES SELECTION AND TRANSFORMATION

Only the fields required for the model will be selected. The information for the variables selected will be extracted from the database that will be created for the purpose of this study. Assumed predictor and response variables which will be derived from the database are given in Table 1.

Table 1: Data Format

S/N	Variable Name	Variable format	Variable Type
1.	Gender	Male, Female	Categorical
2.	Average Family Income		Continuous
3.	Mother's educational qualification	No formal education, Primary, SSCE, 1st degree, 2nd degree, PhD	Categorical
4.	Father's educational qualification	No formal education, Primary, SSCE, 1st degree, 2nd degree, PhD	Categorical
5.	Marital status of parents	Married, Divorced, Separated, Widowed	Categorical
6.	Mother's occupation	Unemployed, Government worker, Private, Self employed	Categorical
7.	Father's occupation	Unemployed, Government worker, Private, Self employed	Categorical
8.	Family size		Continuous
9.	Student's position in the family	1st born, last born, only child, others	Categorical
10.	Sponsor	Parents, Scholarship, Self, Others	Categorical
13.	Current CGPA	A: 4.5-5.0, B+:4.0-4.49, B: 3.5-3.99, C+: 3.0-3.49, C: 2.5-2.99, D:2.0-2.49, E: 1.0-1.99, F:<1.0	Categorical

MODEL BUILDING

Waikato Environment for Knowldege Analysis (WEKA) and See5 will be used to build software tool for all experiments. See5 for Windows can only generate trees based on C5 algorithm while WEKA allows many algorithms giving room for comparison to determine the optimal model in this study. See5 for Windows is a

sophisticated machine learning tool for discovering patterns that delineate categories, assemble them into classifiers, and use them to make predictions. It is an improvement on Quinlan's ID3 and C4.5 algorithms. See5 classifiers are expressed as decision trees/Seetrees or sets of if-then rules forms that are generally easy to understand.

WEKA is a collection of machine learning algorithms tools for data pre-processing, classification, regression, clustering, association rules and visualization. There are many machine learning algorithms implemented in WEKA including Bayesian classifiers, Trees, Rules, Functions, Lazy classifiers and miscellaneous classifiers.

See5

The first step in building a Seetree in See5 is to collect a set of data values that See5 can analyze. This dataset will have instances for which the actual value of the class variable is known and the associated predictor variables. This will be done at data preparation stage. Each entry in the dataset provides values for the class and predictor variables for a specific instance. Each entry is known as a case, row, record, observation or vector. Each student record represents a case. The attributes description names, labels, classes, and discrete values will be represented by arbitrary strings of characters and saved in a file with .name extension. The data file provides information on the training cases from which See5 will extract patterns. Data values will be separated by comma and saved in data file with .data extension. Figure 1 and 2 below are instances of data file and attribute description file respectively. Figure 3 shows the typical See5 Window with the data file selected.

```
22,Female,19,46,2,Middle,1stdegree,1stdegree,Married,Selfemployed,Private,7,others,Federa
l,Urban,Rural,Scholarship,D
21,Female,25,55,2,Middle,1stdegree,1stdegree,Married,Selfemployed,Government,7,onlychil
d,Private,Urban,Semiurban,Parents,C
16,Female,20,48,1,Upper,2nddegree,2nddegree,Married,Private,Government,7,others,Private,
Urban,Semiurban,Parents,B+
16,Female,20,47,1,Upper,1stdegree,1stdegree,Married,Government,Government,7,others,Priv
ate,Urban,Semiurban,Parents,B+
17,Male,21,45,1,Upper,1stdegree,1stdegree,Married,Government,Government,4,1stborn,Priv
ate,Urban,Semiurban,Parents,E
16,Female,20,40,1,Upper,1stdegree,SSCE,Married,Government,Government,3,1stborn,Privat
e,Urban,Semiurban,Parents,B+
16,Male,18,43,1,Upper,1stdegree,1stdegree,Married,Government,Government,3,1stborn,Fede
ral,Semiurban,Urban,Parents,B
18,Male,19,48,1,Upper,1stdegree,SSCE,Married,Government,Government,3,others,Private,U
rban,Urban,Parents,B+
16,Male,25,49,1,Upper,1stdegree,1stdegree,Married,Government,Government,7,others,Privat
e,Semiurban,Urban,Parents,D
17,Male,30,61,1,Upper,1stdegree,1stdegree,Married,Government,Government,6,others,Privat
e,Semiurban,Urban,Parents,C
```

Figure 1 Example See5 data format

100L CGPA. |the target attribute
Gender: Male, Female.
total SSCE score: continuous.
post UME score: continuous
Year lapse before admission: continuous.
Social Class:Upper, Middle, Lower.
Mother\'s education: Primary, SSCE, 1stdegree, 2nddegree, PhD.
Father\'s education: Primary, SSCE, 1stdegree, 2nddegree, PhD.
Parents marital status: Married, Divorced, Separated, Widowed.
Mother\'s occupation: Government, Private, Selfemployed.
Father\'s occupation: Government, Private, Selfemployed.
Family size: continuous.
Student\'s Position in the family: 1stborn, lastborn, onlychild, others.
Secondary School Type: Private, State, Federal.
Secondary School Location: Rural, Semiurban, Urban.
Residence Location: Rural, Semiurban, Urban.
Sponsor: Parents, Scholarship, Self, Others.
100L CGPA: A, B+, B, C, D, E, F.

Figure 2 Example See5 attributes' description

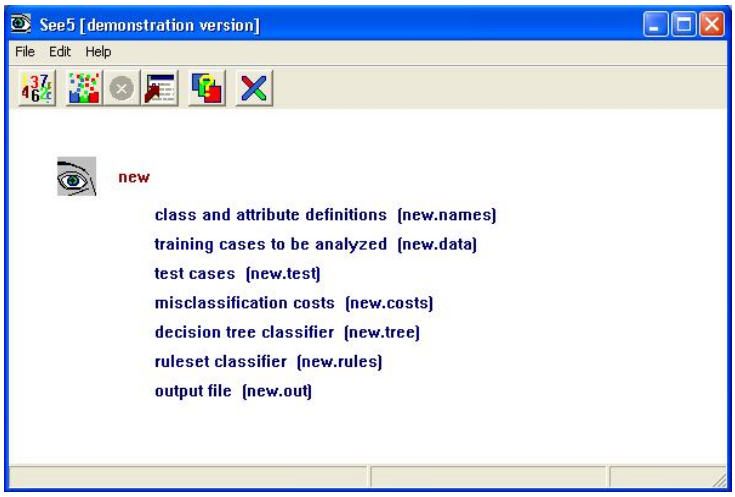

Figure 3 See5 Window

As soon as the names, data, and optional files had been set up in See5 the classifiers can be constructed using the relevant options that See5 provides which affect the type of classifier that See5 will produce and the way it will be constructed (Figure 3).

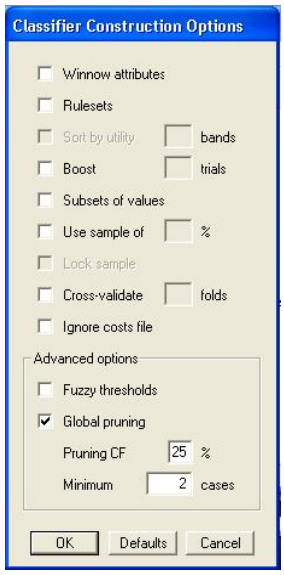

Figure 4: See5 Classifier Construction Options Window

When See5 is invoked with the default values of all options, as shown in figure 4, it will construct a single decision tree. Classifiers called rulesets that consist of unordered collections of (relatively) simple if-then rules are generated when the ruleset option is checked. These rulesets are easier to interpret than the decision trees and a ruleset generated from a tree usually has fewer rules than the tree has leaves. The Boost option with x trials instructs See5 to construct up to x classifiers. In this study Single Seetree, Rulesets and Boost SeeTree will be generated using See5.

WEKA

WEKA is data mining software developed in Java. It has a GUI Chooser from which any one of the four major WEKA applications can be selected. For the purpose of this study the Explorer application will be used as shown in Figure 5.

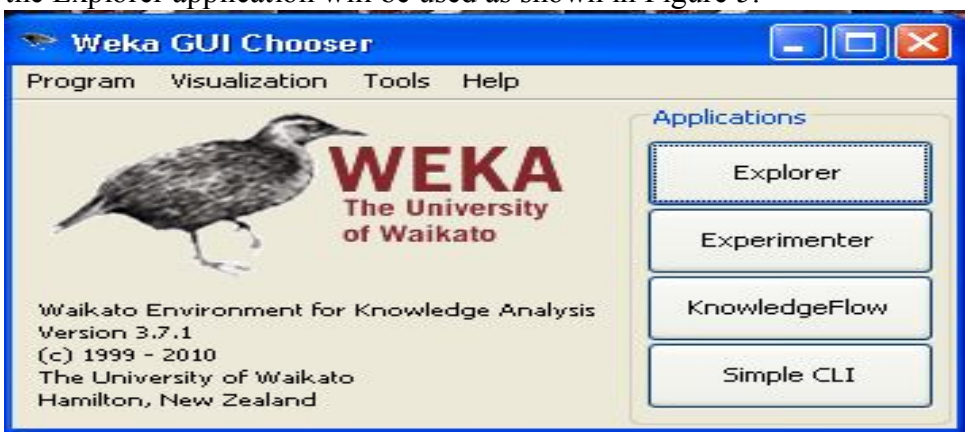

Figure 5: Weka GUI Chooser

The data file to be used for WEKA is similar to that of See5, the only difference is that the attribute and the data are put together in one file and the data is saved in .arff format. The file begins with @RELATION to indicate the datafile name, each attribute starts with @ATTRIBUTE and the dataset begins with @DATA. A typical WEKA file is shown in Figure 6.

```
Students data from Babcock University
@RELATION data

@ATTRIBUTE ageonentry NUMERIC
@ATTRIBUTE Gender {Male, Female}
@ATTRIBUTE SocialClass {Upper, Middle, Lower}
@ATTRIBUTE Mothereducation {Primary, SSCE, 1stdegree, 2nddegree, PhD}
@ATTRIBUTE Fathereducation {Primary, SSCE, 1stdegree, 2nddegree, PhD}
@ATTRIBUTE ParentsMaritalstatus {Married, Divorced, Separated, Widowed}
@ATTRIBUTE Motheroccupation {Government, Private, Selfemployed}
@ATTRIBUTE Fatheroccupation {Government, Private, Selfemployed}
@ATTRIBUTE Familysize NUMERIC
@ATTRIBUTE StudentPositioninthefamily {1stborn, lastborn, onlychild, others}
@ATTRIBUTE SecondarySchoolType {Private, State, Federal}
@ATTRIBUTE SecondarySchoolLocation {Rural, Semiurban, Urban}
@ATTRIBUTE ResidenceLocation {Rural, Semiurban, Urban}
@ATTRIBUTE Sponsor {Parents, Scholarship, Self, Others}
@ATTRIBUTE 100LCGPA {A, B+, B, C, D, E, F}

@DATA
16,Male,26,43,1,Upper,1stdegree,1stdegree,Married,Private,Private,3,1stborn,Private,Urban,S
emiurban,Parents,B+
17,Male,18,49,1,Upper,1stdegree,2nddegree,Married,Government,Government,4,onlychild,Pr
ivate,Urban,Semiurban,Parents,B
16,Male,21,53,1,Upper,1stdegree,1stdegree,Married,Private,Government,4,1stborn,Private,Ur
ban,Semiurban,Parents,C
18,Male,19,58,2,Upper,SSCE,1stdegree,Separated,Government,Government,5,lastborn,Privat
e,Urban,Semiurban,Parents,C
16,Female,30,56,1,Upper,1stdegree,1stdegree,Married,Government,Government,5,1stborn,Pri
vate,Urban,Semiurban,Self,B
18,Male,32,61,2,Upper,2nddegree,1stdegree,Divorced,Government,Government,6,1stborn,Pri
vate,Urban,Urban,Parents,D
19,Male,22,40,2,Upper,1stdegree,PhD,Divorced,Government,Government,4,1stborn,Private,
Urban,Urban,Parents,C
16,Male,18,43,1,Upper,1stdegree,1stdegree,Married,Government,Government,7,onlychild,Pri
vate,Urban,Urban,Parents,D
17,Male,30,62,1,Upper,Primary,SSCE,Married,Selfemployed,Selfemployed,4,1stborn,Federal
,Semiurban,Semiurban,Parents,B+
18,Male,24,53,2,Upper,1stdegree,1stdegree,Divorced,Government,Government,8,others,Fede
ral,Semiurban,Semiurban,Parents,C
```

Figure 6 WEKA file format

The Explorer window of WEKA has six tabs. The first tab is Preprocess that enable the formatted data to be loaded into WEKA environment. Once the data has been loaded, the Preprocess panel will show a variety of information as shown in Figure 7

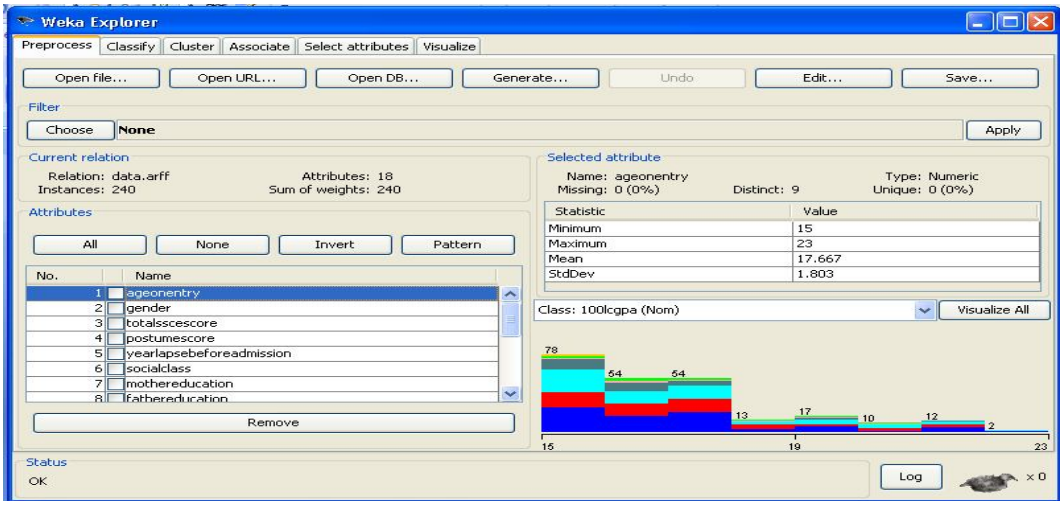

Figure 7 WEKA Explorer

There are several classifiers available in WEKA but CART (SimpleCart), C4.5 (J48) and ANN will be used in this study. RIpple DOwn Rule Learner (Ridor) rule based learner will be generated using WEKA. Attribute importance analysis will be carried out to rank the attributes by significance using Information gain. This is similar to attribute usage in See5. Finally, Correlation-based Feature Subset Selection (CfS) and Consistency Subset Selection (CoE) filter algorithm will be used to rank and select the attributes that are most useful. The F-measure and the AUC which are well known measures of probability tree learning will be used as evaluation metrics for models generated by WEKA classifiers.

ALGORITHMS

The two major algorithms to be used in this study are Decision trees and Neural Networks.

Decision trees

A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision-making. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome. The three widely used decision tree learning algorithms are:

ID3, CART and C4.5. In this study, WEKA and See5 will be used to construct CART, C4.5 and C5 decision tree algorithms.

Neural Networks

Neural networks try to mimic interconnected neurons in animal brains in order to make the algorithm capable of complex learning for extracting patterns and detecting trends. It is built upon the premise that real world data structures are complex, and thus it necessitates complex learning systems. A trained neural network can be viewed as an "expert" in the category of information it has been given to analyze. This expert system can provide projections given new solutions to a problem and answer "what if" questions. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

A typical neural network is composed of three types of layers, namely, the input layer, hidden layer, and output layer. It is important to note that there are three types of layers, not three layers, in the network. There may be more than one hidden layer and it depends on how complex the researcher wants the model to be.

The input layer contains the input data; the output layer is the result whereas the hidden layer performs data transformation and manipulation. Because the input and the output are mediated by the hidden layer, neural networks are commonly seen as a "black box."

The network is completely connected in the sense that each node in the layer is connected to each node in the next layer. Each connection has a weight and at the initial stage and these weights are just randomly assigned. A common technique in neural networks to fit a model is called back propagation. During the process of back propagation, the residuals between the predicated and the actual errors in the initial model are fed back to the network. In this sense, back propagation is in a similar vein to residual analysis in EDA (Behrens and Yu, 2003). Since the network performs problem-solving through learning by examples, its operation can be unpredictable. Thus, this iterative loop continues one layer at a time until the errors are minimized. Neural networks address the problem of under-determination of theory by evidence with use of multiple paths for model construction. Each path-searching process is called a "tour" and the desired result is that only one best model emerges out of many tours. Like other machine learning algorithms, neural networks also incorporate cross-validation to avoid capitalization on chance alone in one single sample.

In this study, WEKA will be employed to construct neural nets. Different combinations of hidden layers (1-3), tours (3-20), and k-fold cross-validation (2-5) will be explored. However, the results may not be repeatable due to numerous possibilities of path-searching during touring and random spilts of subsets during cross-validation. The major objective of running neural nets is to select a subset of best predictors as well as examine the non-linear relationship between the probability of retention and the variables.

The models built from the 2 classes of algorithms will be trained with 40% of the data and 10-fold cross-validation will be used to compute confusion matrices and accuracy level to measure the validity of the models. If-then rules that can be used for predictive purposes will be generated from the optimal algorithms after a lot of comparison. Finally, a framework for a predictive system will be designed to serve as a recommender for the higher education management decision-making.

EXPECTED OUTPUT

The expected output from this study include an optimal method for recommending students based on machine learning algorithms, a database of students with their various family background information, if-then rules mapping students performance to their family background and a framework of a recommender system for decision making for administrator in higher education institutions. The beneficiaries of the outcome of this study include but not limited to higher education management, parents or guardians and government.

CONCLUSION

This report presents a proposal for research in the area of machine learning for determining the influence of family background factor on student academic performance. The main purpose of this research is to show that this new approach will have several benefits to strengthen the educational system planning and monitoring activities. To achieve this objective, an optimal machine learning algorithm will be derived, rules based on this algorithm will be generated and a framework for predictive system based on these rules will be designed. Although the application of this method to the area of educational evaluation is quite new, a series of experiments will be carried out using small student academic performance dataset before launching out into the larger dataset. The proposed predictive system based on machine learning algorithm rule sets is expected to offer simplicity that will be quite useful in the area of educational evaluation, which needs a system that is easily understood by many people such as educators, policy planners, parents and students.

This project proposal describes the previous attempts in using machine learning approach in educational evaluation and describes the newly proposed method. This is particularly useful in giving direction on what this research should focus on. The work to be carried out in the rest of this project, if successful, will lead to the establishment of a systematic approach to predicting students' academic performance considering their family background which will help reinforce decision made in monitoring or offering assistant to students. This study to the best of our knowledge has not been done somewhere else. Other works from literature has focused on grades of students from high school, gender, personality, classroom and peer pressure, but none of this research has been on family background factors. Therefore this research is new and will be a novelty if successfully carried out.

WORK PLAN

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Jan
Proposal	x	x	x	x									
System Design				x	x	x							
Database Creation					x	x	x						
Model training and building							x	x	x				
Publication									x	x	x		
Final defense												x	x

BUDGET

	$
Computer systems	1,000
Literature	1,000
System Design	1,000
Database and Coding	1,000
Software (See5)	1,400
Publications	600
Conferences	2,000
Dissertation writing and defense	1,000
Total	9,000

REFERENCES

[1] Adeyemo, A. B and Kuye G., 2006. Mining Students' Academic Performance using Decision Tree Algorithms. *Journal of Information Technology Impact* 6(3): 161-170.

[2] Affendey, L.S., I.H.M. Paris, N. Mustapha, M.N. Sulaiman and Z. Muda, 2010. Ranking of influencing factors in predicting students academic performance. *Inform. Technol. J.*, 9: 832-837.

[3] Anderson, G., Benjamin, D., and Fuss, M. 1994. The Determinant of Success in University, Introductory Economics Courses. *Journal of Economic Education.* 25:99-120.

[4] Asikhia O.A. 2010 Students and Teachers' Perception of the Causes of Poor Academic Performance in Ogun State Secondary Schools [Nigeria]: Implication for Counseling for National Development. *European Journal of Social Sciences*. 13(2):229-242.

[5] Bakare, C.C. 1975. Some Psychological Correlates of Academic Success and Failure. *African Journal of Educational Research*.

[6] Bhardwaj B. K. and Pal S. Data Mining: A prediction for performance improvement using classification, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011

[7] Bhardwaj B. K. and Pal S. Mining Educational Data to Analyze Students‟ Performance (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011

[8] Delavari N, Beikzadeh M. R, Amnuaisuk S. 2005. Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System. *6th Annual International Conference*: ITEHT Juan Dolio, Dominican Republic.

[9] Delavari N, Beikzadeh M. R. 2004 A New Model for Using Data Mining in Higher Educational System*, 5th International Conference on Information Technology based Higher Education and Training*: ITEHT ’04, Istanbul, Turkey.

[10] Gaweda, A. E., Muezzinoglu, M. K., Aronoff, G. R., Jacobs, A. A., Zurada, J. M., & Brier, M. E. 2005. Individualization of pharmacological anemia management using reinforcement learning. *Neural Networks*, *18*, 826–834. doi:10.1016/j.neunet.2005.06.020

[11] Golding, P. and O. Donaldson, 2006. Predicting academic performance. *Proceedings of the 36th ASEE/IEEE Frontiers in Education Conference T1D-21*, San Diego, CA.,1-6.

[12] Han J, Kamber M. 2003. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, New Delhi.

[13] Kalles D., Pierrakeas C. 2004, *Analyzing student performance in distance learning with genetic algorithms and decision trees*, Hellenic Open University, Patras, Greece.

[14] Kovačić Z. J. 2010 Early Prediction of Student Success: Mining Students Enrolment Data *Proceedings of Informing Science & IT Education Conference (InSITE)* pp 647:655

[15] Kuyoro S. O. 2010, Investigating the Effect of Students Socio-Economic/Family Background on Students Academic Performance in Tertiary Institutions Using Decision Tree Algorithms. Department of Computer Science, University of Ibadan. Unpublished MSc Thesis

[16] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten 2009. *The WEKA Data Mining Software: An Update*; SIGKDD Explorations, Volume 11, Issue 1.

[17] Marquez-Vera C., Romero C. and Ventura S. Predicting School Failure Using Data Mining. Proceedings of the 4th International Conference on Educational Data Mining, pp 271-276, 6 Jul 2011

[18] McKenzie, K. and R. Schweitzer, 2001. Who succeeds at University? Factors predicting academic performance in first year Australian university students. *Higher Educ. Res. Dev.*, 20: 21-33.

[19] Ogor E. N. Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques in proceeding Fourth Congress of Electronics, Robotics and Automotive Mechanics 2007 pp352-359

[20] Osofisan, A. O. and Olamiti, A. O. 2009. Academic Background of Students and Performance in a Computer Science Programme in a Nigerian University. *European Journal of Social Sciences*. 9(4): 564-572.

[21] Padberg, F., Ragg, T., & Schoknecht, R. 2004. Using machine learning for estimating the defect content after an inspection. *IEEE Transactions on Software Engineering*, *30*(1), 17–28. doi:10.1109/TSE.2004.1265733

[22] Pandey U. K. and Pal S. Data Mining : A prediction of performer or underperformer using classification. International Journal of Computer Science and Information Technologies, Vol. 2 (2) , 2011, 686-690

[23] Quadri M. N. and Kalyankari N.V. 2010 Drop Out Feature of Student Data for Academic Performance Using Decision Tree Techniques, *Global Journal of Computer Science and Technology* Vol. 10 Issue 2, pp2-5

[24] Quinlan J.R. 1993 C4.5: *Programs for Machine Learning*. Morgan Kaufmann, San Mateo California.

[25] Quinlan J.R. 1996. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77-90.

[26] Rada, R., & Ha, L. 2008. Intelligent technologies for investing: A review of engineering literature. *Intelligent Decision Technologies*, *2*(3), 167–178.

[27] Salah A. A., Çınar H., Akarun L., & Sankur B. 2007. Robust facial landmarking for registration. *Annales des Télécommunications*, *62*(1-2), 1608–1633

[28] Sembiring S., Zarlis M., Hartama D., Ramliana S, and Elvi W. Prediction of student academic performance by an application of data mining techniques in the proceeding *2011 International Conference on Management and Artificial Intelligence IPEDR vol.6 (2011) © (2011) IACSIT Press, Bali, Indonesia*

[29] Sokolova M., Nastase V., & Szpakowicz S. 2008. *The telling tail: Signals of success in electronic negotiation texts.* Paper presented at the Third International Joint Conference on Natural Language Processing (IJCNLP 2008).

[30] Superby, J.F., J.P. Vandamme and N. Meskens, 2006. Determination of factors influencing the achievement of the first-year university students using data mining methods. *Proceedings of the 8th international conference on intelligent tutoring systems, Educational Data Mining Workshop, (ITS'06),* Jhongali, Taiwan, 37-44.

[31] Vandamme, J.P., N. Meskens and J.F. Superby, 2007. Predicting academic performance by data mining methods. *Educ. Econ*., 15: 405-419.

[32] Yu C. H., DiGangi S., Jannasch-Pennell A. and Kaprolet C. A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year. Journal of Data Science 8(2010), 307-325