UNIVERSITY OF NAIROBI

SCHOOL OF COMPUTING AND INFORMATICS

MSC. COMPUTATIONAL INTELLIGENCE

BUSINESS INTELLIGENCE STUDENT RETENTION SYSTEM FOR HIGHER LEARNING INSTITUTIONS

(A CASE FOR MACHAKOS UNIVERSITY COLLEGE)

BY:

JOSEPH MUTUKU NGEMU

P52/65120/2013

SUPERVISOR

DR. ELISHA . T.O. OPIYO

A Research project submitted in partial fulfillment of the requirements of the Degree of Master of Science in Computational Intelligence at the University of Nairobi.

December, 2014

Declaration

This project is my original work and to the best of my knowledge it has not been presented anywhere for the purpose of an academic award.

Signature: _____ Date: _____

NGEMU, Joseph Mutuku
(P52/65120/2013)

Supervisor's approval

This project report has been submitted in partial fulfillment of the requirements for the Master of Science in Computational Intelligence of the University of Nairobi with my approval as the University Supervisor.

Signature: _____ Date: _____

Dr. E. T. O. Opiyo

School of Computing and Informatics

University of Nairobi.

Acknowledgement

To the Almighty God for the gift of life so as to get this far I have come. To my loved ones, for their great support, care and encouragement throughout my academic years.

To my supervisor Dr. E. T. O. Opiyo who has opened my eyes to the research world. His guidance, support, and positive criticism made this project a success. To the panellists, Dr. Agnes Wausi, Dr. Muchemi to whom I am grateful for positive criticism that has led to success of this project.

To my classmates and friends, who shared ideas and provided assistance during this project.

I say God Bless you all.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

BI – Business Intelligence

ICT – Information Communication Technology

RPS –Retention Predictive System

HLI – Higher Learning Institution

SRS – Software Requirements Specification

WEKA – Waikato Environment for Knowledge analysis

GUI - Graphical User Interface

DT – Decision Tree

CART - Classification and regression trees

IPS - Intelligent predictive system

JDK - Java Development Kit

JRE - Java Runtime Environment

KCSE- Kenya Certificate of Secondary education

ABSTRACT

Student retention is an essential part of many enrollment management systems. It affects university rankings, school reputation, and financial wellbeing. Student retention has become one of the most important priorities for decision makers in higher education institutions. Improving student retention starts with a thorough understanding of the reasons behind the attrition. Such an understanding is the basis for accurately predicting at-risk students and appropriately intervening to retain them. In this study, using student demographic and institutional data along with several business intelligence techniques (Decision tree, Naïve bayes, multilayer perceptron and support vector machine), we developed prototype to predict and to explain the reasons behind student attrition. This study used classification models generated using Waikato Environment for Knowledge Analysis (WEKA). The model was built using the 10-fold cross validation, and holdout method (60% of the data was used as training and the remaining as test and validation). Random sampling techniques were used in selecting the datasets. The sensitivity analysis of the models revealed that the student age on entry, parent occupation, health of student and financial variables are among the most important predictors of the phenomenon. Results of the classifiers were compared using accuracy level, confusion matrices and speed of model building benchmarks. The study shows that identifying the relevant student background factors can be incorporated to design a prototype that can serve as valuable tool in predicting student withdrawal as well as recommend the necessary intervention strategies to adopt.

Keywords: *Business Intelligence, Retention, Decision Trees, Naïve Bayes, Multilayer Perceptron, Support Vector Machine, WEKA* .

CHAPTER 1: INTRODUCTION

1.1 Background of the study

Business Intelligence or BI is a broad category of applications and technologies for gathering, storing, analyzing and providing access to data to help enterprise users make better business decisions. BI improves decisions by supplying timely, accurate, valuable, and actionable insights.

With the rapid advancement and development of Information and Communication Technologies (ICT), organizations are now able to generate, collect and distribute huge amount data from internal and external sources, this is also happening in higher learning institutions. As the concept of Business intelligence (BI) is steadily rising up the priority list within many institutions, it is necessary to explore the potential of BI in making better use of student data in support of student management and decision making. It is hoped that the application of BI systems will help managers and academic staff take a more proactive approach in student management and strategic planning through well informed and evidence-based decisions.

Organizations' requirements to improve quality of decision-making and quality of service should turn to the development of information technology infrastructure that will represent a holistic approach to business operations, customers etc.(Wells and Hess, 2004). Theory and practice from many studies show that organizations' requirements to improve quality of decision-making and quality of service are largerly met by BI systems (Liautaud and Hammond, 2002). Therefore, BI has become a strategic initiative, and many business leaders now regard BI as instrumental in driving business effectiveness and innovation (Watson and Wixsom, 2007). Moreover, BI has been used in many other sectors, for instance, in manufacturing companies, in retailing sector for user profiling, in financial services for claims analysis and fraud detection, in transportation for fleet management, in telecommunications for identifying reasons for churn and health care for outcomes analysis.

However, Bi technologies have not been widely used in higher learning institution; despite that BI can also play an important role in student data analysis for decision making and strategic planning. Most of the current student information systems in higher learning institution are just a collection of student data.

The key benefit of applying business intelligent to this problem of student withdrawal is that often there are multiple complex factors which influence a student's likelihood to withdraw. Business intelligent tools enables us to analyze historical data sets at an institution, identify the combination of factors which are most closely correlated with student withdrawal and build a model which allows us to predict the likelihood of individual student withdrawal in the future. This gives us a really powerful way to understand retention and a proactive way to manage retention issues

1.2 Problem Statement

An issue of concern in higher learning institutions across the world is the retention and success of students in their studies. This is a particularly pressing issue in the context of widening participation for under-represented student groups, easing student diversity and educational quality assurance and accountability processes. As well as the personal impact and loss of life chances for students, non-completion has financial implications for students in developing countries (and their families), and for society and the economy through the loss of potential skills and knowledge. Unfortunately, most institutions have not yet been able to translate what we know about student retention into forms of action that have led to substantial gains in student persistence and graduation. Though some have, many have not (Carey, 2005).

Lack of efficient educational system, lack of systems for predicting the likelihood of individual student withdrawal in the future and lack of information about the potential factors that may influence student withdrawal has been a challenge to many higher learning institutions when it comes to management of student retention issues.

Information is the new key enterprise asset as organizations across the globe not only leverage, but compete on information. But the pragmatic truth is that, while BI technologies continue to grow and mature, the promise of an efficient and effective BI environment that fits the real needs faced by higher learning institution users and decision makers day by day remains a challenge.

1.3 Objectives of the study

1.3.0 General research objectives

The goal of this research is to find ways of improving the efficiency of higher learning institution systems by applying business intelligent techniques on educational databases. This can potentially reduce the incidents of student retention.

1.3.1 Specific research objectives

i. Formulate specifications for a BI predictive model for higher learning institutions.

ii. Identification of different factors, which affects a student's retention rate and design a BI predictive model for higher learning institutions

iii. Apply business intelligence concepts in the modeling process for the prediction of likelihood of dropping out.

iv. Construct a BI prototype for predicting likelihood student withdrawal

v. Validation of the developed model for students studying in Higher Learning Institutions.

1.4 Research Questions

Based on the above research objectives, the research questions include:-

i. What are the variables, in the provided dataset that most determine the likelihood of drop out?

ii. Is it possible to integrate the application development, database and machine learning tool environments in the prototype modeling?

iii. What are the benefits of automating the BI predictive model for prediction of student withdrawal?

1.5 Significance of study

Through the process of business intelligent, educational institution can leverage hidden information in its data, uncovering associations, patterns, and trends that can lead to improving the educational processes.

The results and findings of the study is expected to assist and guide administrators and students in higher learning institutions, Quality Assurance & standards officers, and the

Ministry of Education to understand the factors that influence student withdrawal and come up with the best intervention programmes to address the issue so as to enhance and promote completion rate in higher learning institutions.

Supporting institutional decision making with more informed and evidence based information in all levels, i.e. academic staff and administrative staff at the operational level, faculty managers at the tactical level and registrar at the strategic level.

1.6 Scope of the study

The scope of the project shall be limited to activities that lead to data preprocessing and model building. The main idea of our project is to predict students' learning activities and withdrawal using business intelligence techniques. The information generated after implementing business intelligence can be helpful for both HLI administrators as well as students. The educational business intelligence allows analyzing and better understanding of the educational information and processes from a wide variety of different perspectives.

1.8 Definition of important terms

Student retention is defined by Seidman (2005), as the extent to which learner remain within a higher education institution, and complete a programme of study in a pre-determined time-period.

Business intelligence (BI) is a set of theories, methodologies, architectures, and technologies that transform raw data into meaningful and useful information for business analysis purposes. BI can handle enormous amounts of unstructured data to help identify, develop and otherwise create new strategic business opportunities. BI allows for the easy interpretation of volumes of data. Identifying new opportunities and implementing an effective strategy can provide a competitive market advantage and long-term stability.

Student refers to a learner, or someone who attends an educational institution.

Decision Tree classifier (DT) is a powerful and popular classification and prediction technique (Chaudhuri, 1998).

CHAPTER 2: LITERATURE REVIEW

2.0 Introduction

In this section, a review of existing literature on Business Intelligent and retention problem is done. A description is made of the Business Intelligent and machine learning classification techniques that are used in the prediction of the likelihood of student withdrawal and finally an outline of the factors that influences student withdrawal.

2.1 Higher Education Management and Efficiency

The efficiency of the study process can be measured by the student graduation rate, which is an important criterion in several national models of financing higher education institutions. This aspect of the efficiency of the study process ignores that the graduation rates are under the influence of external factors which are beyond the control of decision-makers at higher education institutions, which is taken into account in the study of Sav (2012).

2.2 Introduction to Student retention

Student retention is one of the most important issues facing higher education today. With one-third of college students dropping out of school each year, it's an issue universities across the country have noticed, but few have found a workable solution to the problem. Admissions offices are already stretched to the breaking point, budgets are tight across the board, and developing and instituting a feasible student retention program can feel like an insurmountable challenge (University Business, May 2008). At its core, the retention of college students is a complex issue, representing an inter play of personal, institutional, and societal factors, with likely associated detrimental costs and implications to all three audiences (Brunsden, 2000).

According to Seidman (2005), student retention refers to the extent to which learner remain within a higher education institution, and complete a programme of study in a pre-determined time-period. Student retention can be measured through completion rate and continuation rate. Scholars of higher education, especially retention experts, have variously defined retention amplifying certain elements based on their own theoretical perspective. According to a sample of definitions that may be found in the research literature, retention refers to: Successful completion of students' academic goals of degree attainment (Levitz, 2001). Students meeting clearly defined educational goals whether

they are course credits, career advancement, or achievement of new skills (Tinto, 1993). Students' successful academic and social integration into the college community, marked by the feeling that one fits at the institution and positive educational attitudes and experiences (Bean, 1980).The match between students' motivation and academic ability and their academic and social characteristics (Cabrera, Castaneda, Nora, Hengstler, 1992).The degree of direct involvement of students in the academic and social life of their institutions (Astin, 1984). The by-product of student success and satisfaction and ultimately an indicator of institutional success (Noel and Levitz, 1985).

Student retention has remained, over many decades, a strong area of concern in postsecondary education, for good reason. Retaining students is a key factor in an institution's maintenance of its stability and reputation, and it helps students and society to avoid the all-too-common consequences associated with a lack of education, including under employment and poverty.

2.3 Student retention models

2.3.1 Tinto's Student Retention Model

Of greatest note are Tinto's Student Retention Model (1975), Astin's Theory of Involvement (1984), Braxton's support of active learning (2000), Bean's Student Attrition Model (1982), and Chickering's Student Development Theory (1969). As the majority of research in the areas of student retention will sometimes link to remedial/developmental education, it is important to consider these interlocking theories and to provide some background.

The most commonly referred to model in the student retention/dropout literature is Tinto's. It was first offered in a literature review (Tinto, 1975)

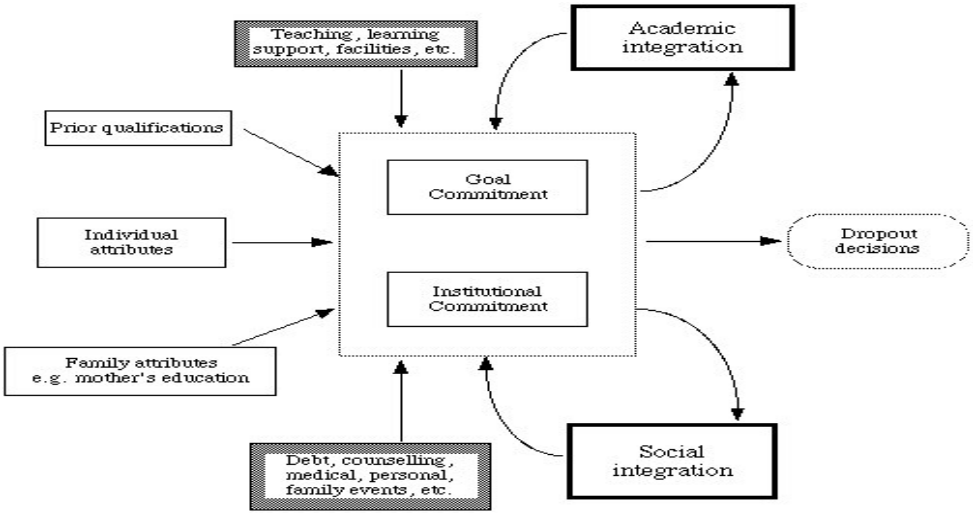


Figure1: student retention model (Tinto,V, 1975) "Dropout from Higher Education:

Its central idea is that of "integration": it claims that whether a student persists or drops out is quite strongly predicted by their degree of academic integration, and social integration. These evolve over time, as integration and commitment interact, with dropouts depending on commitment at the time of the decision.

2.3.2 Bean and Eaton Geometric Model of Student Persistence and Achievement

John Bean and Shevawn Eaton (2000) offer an integrated multi-level model of causes of dropping out. Their model combines individual characteristics and background variables. Examples include high school experiences; students' intentions or educational goals; family support; indicators of students' academic standing and social integration in college; how students interact with the institutional bureaucratic structures; external factors (i.e., financial situation or personal relationships outside of college); and ultimately students' attitudes toward themselves and the school, including feelings of fit and loyalty to the institution.

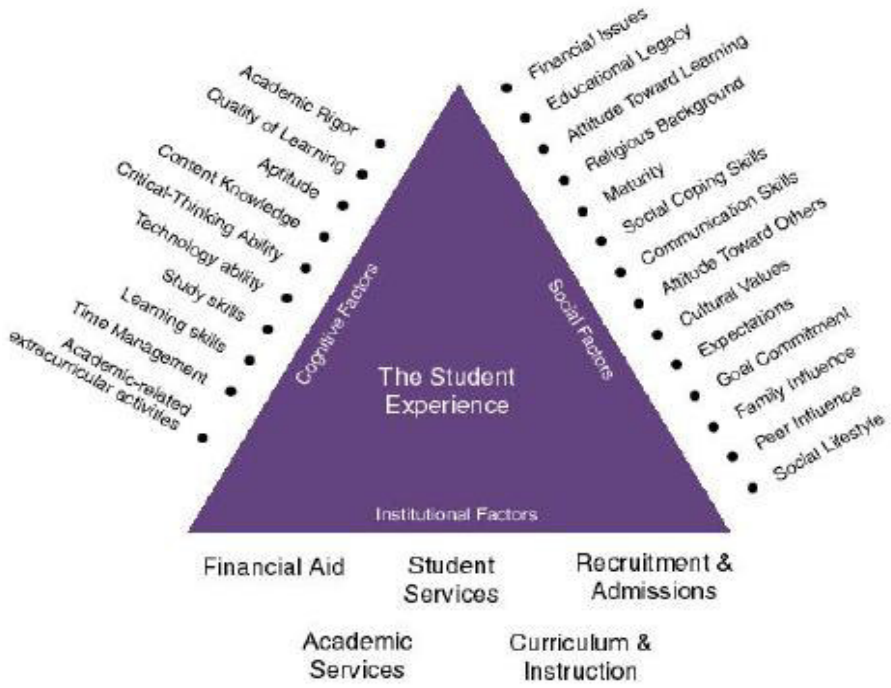


Figure 2: Model of Student Persistence and Achievement (bean, 2000)

The geometric model differs from others by placing the student at the center of the model, rather than an indifferent element to a flow chart or structural equation model.

2.3.2.1 Cognitive Factors

The cognitive factors relate to the intelligence, knowledge, and academic ability a student brings with him or her to the college environment. These factors may be measured by such variables as course selection and completion in high school, aptitude, or extracurricular involvement in academic-related areas. Cognitive factors are important because they directly relate to the student's ability to comprehend and complete the academic portion of the college curriculum.

The decision-making process is an important part of the models described Tinto (1975, 1993) describes the decision-making process regarding goal commitment and dropout, Bean (1982) describes an intent to leave, and Anderson (1985) identifies value conflicts and career indecision among the important variables that a student controls through the set of social and cultural values instilled in him or her.

2.3.2.2 Social Factors

The second factor related to student persistence and performance is the set of social factors impacting on students. Such factors include parental and peer support, the

development or existence of career goals, educational legacy, and the ability to cope in social situations A student who is brought up in a culturally and educationally rich

environment will develop skills critical to postsecondary, career, and personal success. Students hailing from less supportive environments may bring with them deficiencies in their self-esteem and efficacy, especially as they relate to academics when compared with students from more advantaged backgrounds.

2.3.2.3 Institutional Factors

College is undoubtedly the biggest social change a traditional-age student has ever undertaken. College presents stresses, at some level, to all students. Substantial research exists on the stresses of freshman year, especially on minority and low-income students. Regardless of one's subscription to either Gennep's social anthropology theory (Tinto, 1988) or to Valentine's biculturation theory (Rendon, Jalomo, and Nora, 2000; Valentine, 1971), how the institution reacts to students is of primary importance to retention, persistence, and completion. The geometric model places this set of factors at the base of the triangle because it is the college that forms the foundation for college success. It is here that the institution can identify and match the needs of individual students, a student cohort group, or the student body as a whole.

2.3.3 Swail student retention framework

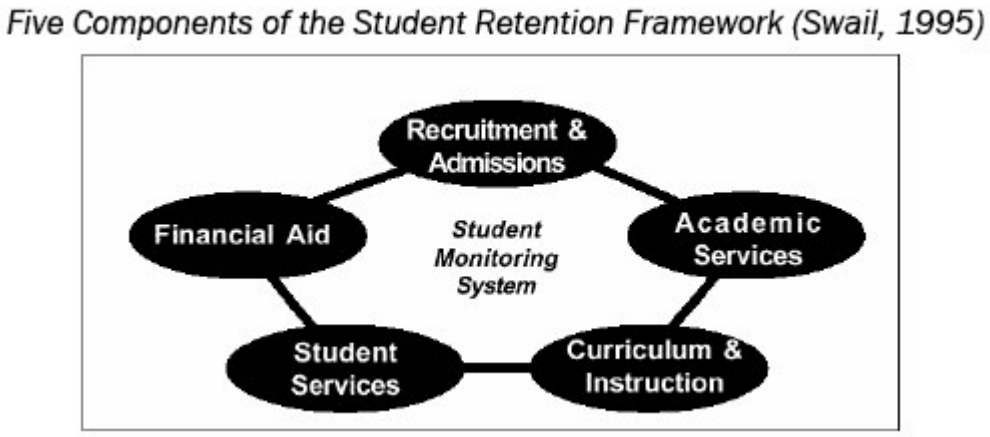


Figure 3: student retention framework (Swail, 1995)

The retention framework is classified into five components based on an extensive review of current literature.

2.3.3.1 Financial Aid

Financial aid is a critical part of the persistence puzzle. For students from low-income backgrounds, many of whom are students of color, finances are a make-it or break-it issue. A strong financial aid office is often the sign of a well-oiled campus, where latitude is given to students who have special financial needs.

Student financial issues have frequently been identified as a barrier to completion, especially by students from lower socio-economic groups (Ozga & Sukhnandan, 1997; Yorke et al., 1997; Yorke, 1999; Dodgson & Bolam, 2002). HEFCE-funded research in 1997 found that financial hardship exercised some impact on early withdrawal (Yorke et al 1997 and Ozga & Sukhnandan, 1997). In particular, students from the two lowest socio-economic groups were more likely to withdraw because of financial difficulties than students from the top two social groups. The House of Commons Select Committee Report (2001) on student retention found finance and part-time employment to be contributory factors to early withdrawal. Both lack of money and concern about debt adversely affect retention.

2.3.3.2 Recruitment and Admissions / Institutional and course match

Tinto (1993) and other researchers (Astin, 1975; Cope and Hannah, 1975) discuss the importance of matching students' goals and expectations to a college's mission. The role of the recruitment and admissions offices must be clarified, first, to identify students whose career and educational goals are closely matched to the institutional mission and, second, to admit only those students to college.

2.3.3.3 Academic Services

The academic services component is the most diversified and expansive component in the framework. The focus of academic services in terms of student retention and persistence is on providing supplementary support to students in addition to practice with classroom lectures. Effective academic advising is important to laying out an appropriate course map for students (Forrest, 1982; Beal and Noel, 1980). To be effective, it is important that students receive guidance that reflects their needs and incorporates the knowledge of campus programming and bureaucratic practices. Prospective advisers need to be trained accordingly to handle a variety of issues during advising sessions.

2.3.3.4 Curriculum and Instruction

The continued development of curricula and pedagogical practice is perhaps the most important and fundamental need that colleges must address in terms of student retention. The need to revise current practices, especially in gatekeeper courses, stems from what Tobias (1990) acknowledges as the practice of designing courses that are "unapologetically competitive, selective & intimidating, and designed to winnow out all but the 'top tier'"

2.3.3.5 Student Services / Personal circumstances

Neisler (1992) concluded that personal, emotional, and family problems, in addition to feelings of isolation and adjustment to college life, are strong barriers to retention for African American students. Therefore, the campus must focus on developing an atmosphere that is supportive, safe, and pluralistic. Personal circumstances can include mental and physical health problems, caring for a relative, childcare, bereavement etc.

2.4 Introduction to Business Intelligence

Business intelligence is a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes. (Carlo Vercellis, 2009).

Business Intelligence (BI) systems provide a proposal that faces needs of contemporary organisations. Main tasks that are to be faced by the BI systems include intelligent exploration, integration, aggregation and a multidimensional analysis of data originating from various information resources. Systems of a BI standard combine data from internal information systems of an organization and they integrate data coming from the particular environment e.g. statistics, financial and investment portals and miscellaneous databases. Such systems are meant to provide adequate and reliable up-to-date information on different aspects of enterprise activities. As the first research results show, the BI systems contribute to improvement and transparency of information flows and knowledge management and they also enable organizations to (Kalakota, & Robinson, 1999; Liautaud, & Hammond, 2002; Moss, & Alert, 2003), discover, monitor and do analysis.

Socio-economic reality of contemporary organisations has made organisations face some necessity to look for instruments that would facilitate effective acquiring, processing and analyzing vast amounts of data that come from different and dispersed sources and that

would serve as some basis for discovering new knowledge. For long time management information systems (MIS) have been supporting organisations in their different tasks. However, today many IT systems have undergone significant depreciation. Hitherto existing management information systems (i.e. MIS, DSS, ES, EIS) have not always met decision makers' expectations

It is assumed that BI may support decision making on all levels of management regardless of the level of their structuralisation (Olszak, & Ziemba, 2003). On the strategic level, BI makes it possible to set objectives precisely and to follow realisation of such established objectives. BI allows for performing different comparative reports, e.g. on historical results, profitability of particular offers, effectiveness of distribution channels along with carrying out simulations of development or forecasting future results on the basis of some assumptions.

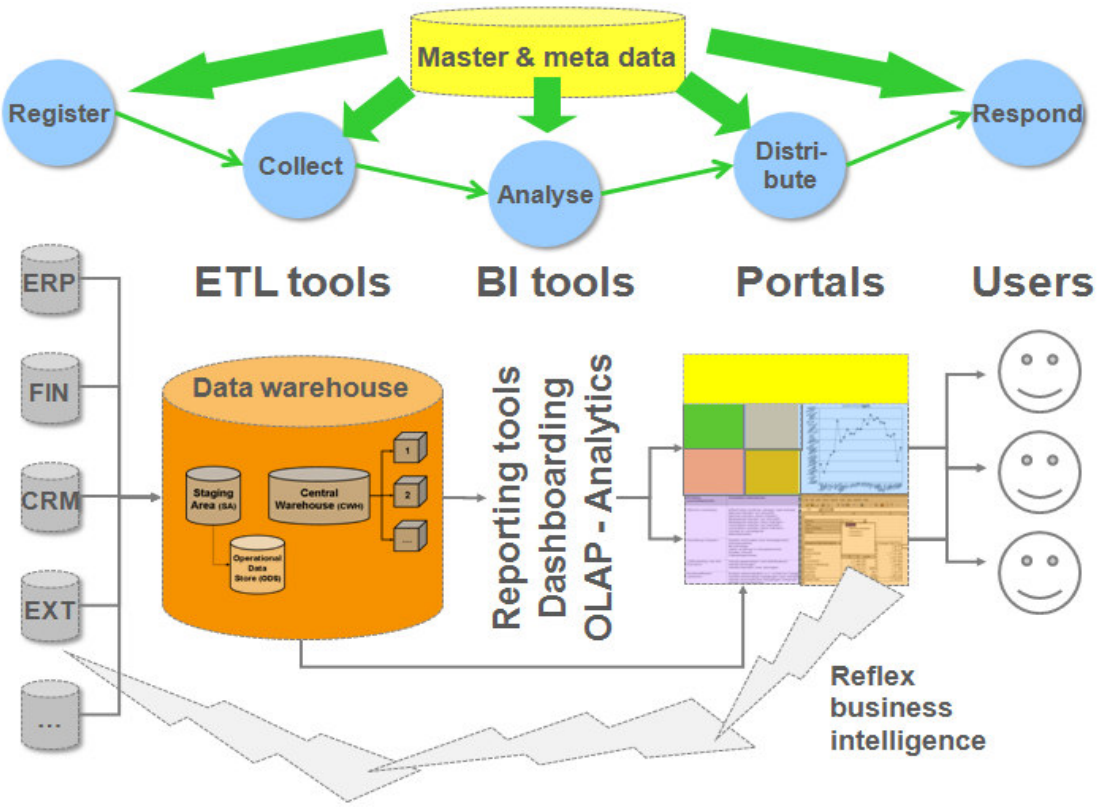


Figure 4: Conceptual Model (Ziemba, 2003). "For Understanding BI systems in decision making"

2.4.1 Application of Business Intelligence in student retention.

Higher learning institution data is massive therefore there is great need to use business intelligence to address several important and critical issues related to student retention. The patterns or trends that are discovered guide decision making such as forecasting retention and anticipating student's future fate. Business intelligence is an essential step in the process of knowledge discovery in database in which intelligent techniques are applied in order to extract patterns (Kharya, 2012).

In academic institutions, vast amount of data is produced on daily basis. Turning those data into information can give the academic institutions the necessary edge to not just stay in the game, but to stay ahead of it (Ballard, 2006).

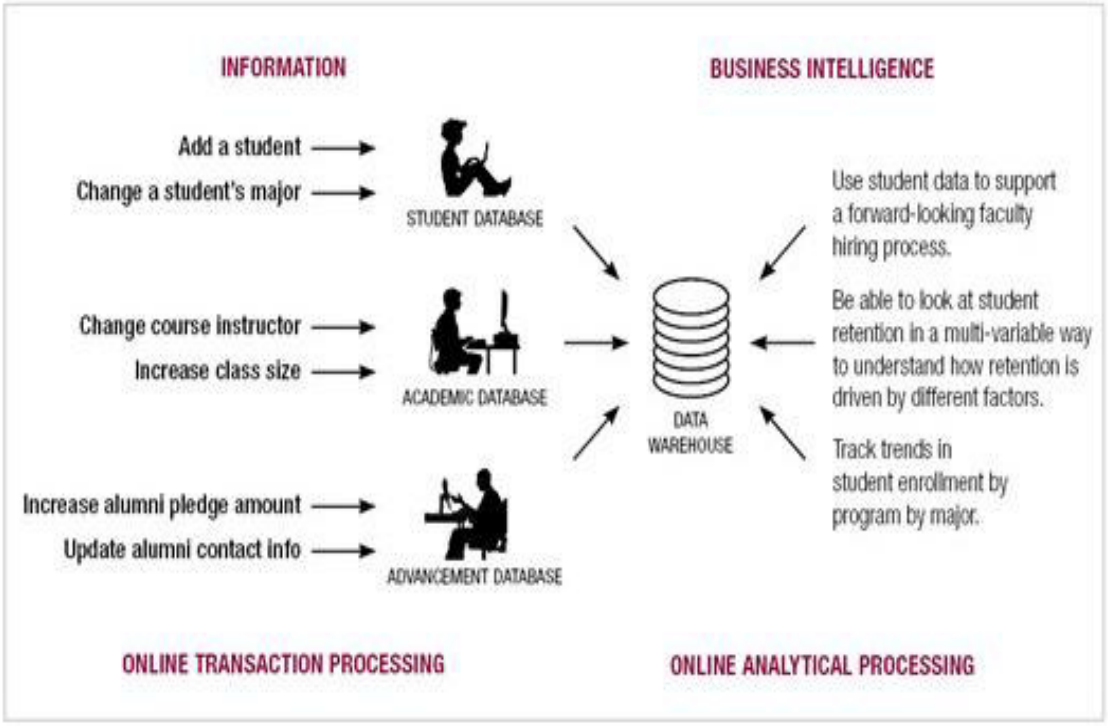


Figure 5: Business Intelligence System Architecture (Ballard, 2006), for Higher Learning Institutions

2.4.2 BI Predictive Models Used In student retention

Predictive modeling uses the data sets that one has collected so as to derive a mathematical model which will be used to predict outcomes of student persistent or dropout (Frank, E.,2011). The main goal of a predictive model is that it should be very accurate in its results as they are used in decision making for potential users. To discover

the hidden pattern from a huge data warehouse, BI software packages use a variety of algorithms to find out the relationship between different data and variables.

Predictive models can be built using different approaches which (Frank, E.,2011)defines some of them as stated below. A Classification Algorithm is a procedure for selecting a hypothesis from a set of alternatives that best fits a set of observations.

2.5 Review of prediction algorithms

2.5.1 Decision Tree classifier (DT)

DT is a powerful and popular classification and prediction technique (Chaudhuri, 1998). Hämäläinen and Vinni (2010) stress, that DT is the most common DM technique in the literature. There are several popular decision tree algorithms such as ID3, C4.5, and CART (classification and regression trees). DT is in the form of a tree structure, where each node is either a leaf node (indicating the value of the target class of examples) or a decision node (specifying a test to be carried out on a single attribute value, with one branch and sub-tree for each possible outcome of the test) (Berson, Smith & Thearling, 2000). DTs have many advantages such as very fast classification of unknown records, easy interpretation of small-sized trees, robust structure to the outliers' effects, and a clear indication of most important fields for prediction but DTs are very sensitive to over-fitting particularly in small data-sets (Hämäläinen & Vinni, 2010).

In this study, to generate a decision tree, the C4.5 (Quinlan, 1993) algorithm was used, which is an extension of Quinlan's earlier ID3 algorithm. To construct the tree, entropy measure was used in the determination of nodes. Since the attributes with higher the entropy cause more uncertainty in outcome, they were selected in order of entropy.

DT is well-known to be one other effective classification technique in several domains (Chau et al., 2001). It is a way of representing series of rules that lead to a class or value. DT models are commonly used in data mining to examine data and induce the tree and its rules that will be used to make predictions.

DT is developed through an iterative process of splitting data into discreet groups, where the objective is to maximize the distance between groups at each split (Two Crows, 1999). How the split is done depends on the algorithm used to implement it. In principle, it is possible to construct as many DTs as possible from a given dataset of attributes. While some of these trees are more accurate than others, finding the optimal tree is

computationally impossible when the search space is large. Efficient algorithms have, however, been developed to induce a reasonable accuracy within a reasonable amount of time. An example of such algorithms is the Hunt's algorithm that forms the bases of many existing decision tree induction algorithms, which include C4.5. These algorithms usually employ greedy strategy in searching the attributes space (Witten and Frank, 2005) and use for partitioning the data. This point is illustrated by how Hunt algorithm works from a high level point of view. A decision tree is grown in a recursive fashion by partitioning the training dataset into successive subsets (Perlich et al., 2004).

Decision Tree has been successfully applied to many areas in business intelligent to solve classification problems. Gama et al.(2003) used DT based algorithm to classify online data streams. Ding et al. (2002) also used DT to classify spatial data streams. Knab et al. (2006) applied DT technique to predict defect density in software source code files. These applications indicate how useful DT can be in solving classification problems in data mining and Business intelligence.

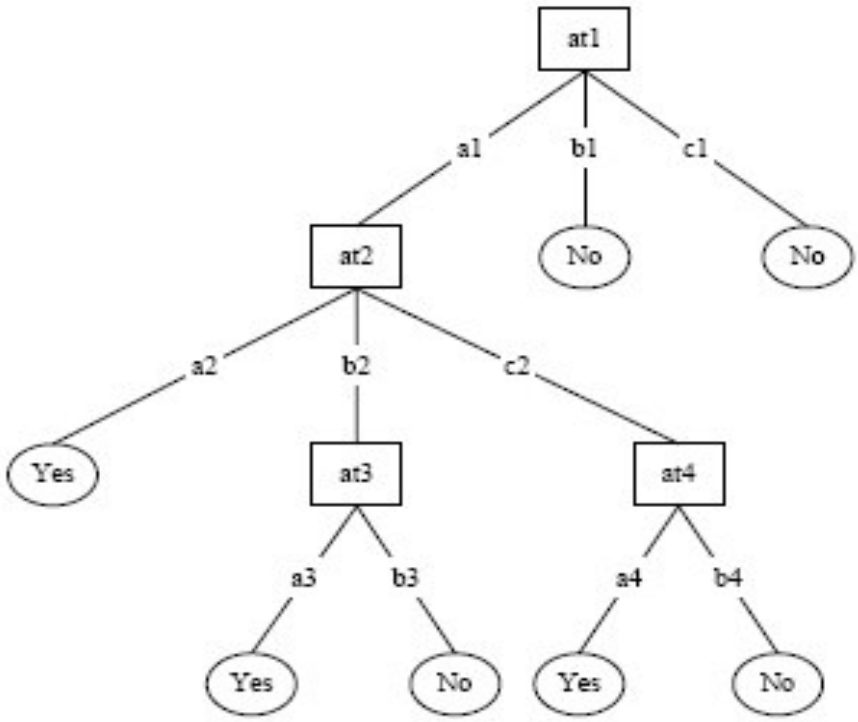


Figure 6: Decision tree, Source: Gama et al, 2003

2.5.2 Multilayered perceptrons

Perceptrons can only classify linearly separable sets of instances. If a straight line or plane can be drawn to seperate the input instances into their correct categories, input instances are linearly separable and the perceptron will find the solution. If the instances are not linearly separable learning will never reach a point where all instances are classified properly. Multilayered Perceptrons (Artificial Neural Networks) have been created to try to solve this problem (Rumelhart et al., 1986). Zhang (2000) provided an overview of existing work in Artificial Neural Networks (ANNs). Thus, in this

study, apart from a brief description of the ANNs we will mainly refer to some more recent articles. A multi-layer neural network consists of large number of units (neurons) joined together in a pattern of connections

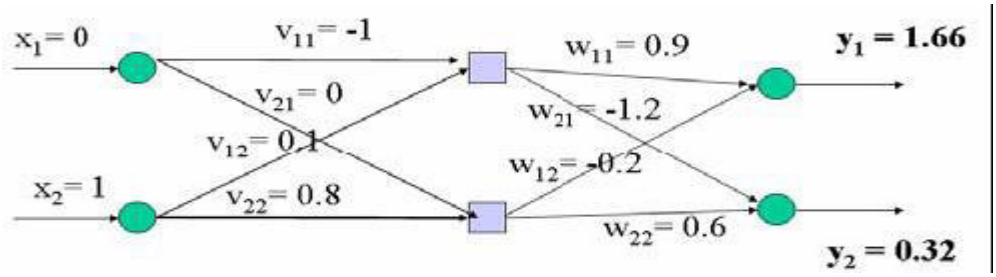


Figure7: Feed-forward ANN

2.5.3 Naive Bayes classifiers

Naive Bayesian networks (NB) are very simple Bayesian networks which are composed of directed acyclic graphs with only one parent (representing the unobserved node) and several children (corresponding to observed nodes) with a strong assumption of independence among child nodes in the context of their parent (Good, 1950).Thus, the independence model (Naive Bayes) is based on estimating (Nilsson, 1965):

$$R = \frac{P(i \mid X)}{P(j \mid X)} = \frac{P(i)P(X \mid i)}{P(j)P(X \mid j)} = \frac{P(i)\prod P(X_r \mid i)}{P(j)\prod P(X_r \mid j)}$$

Comparing these two probabilities, the larger probability indicates that the class label value that is more likely to be the actual label (if R>1: predict i elsepredict j). Cestnik et al (1987) first used the Naive Bayes in ML community. Since the Bayes classification

algorithm uses a product operation to compute the probabilities P(X, i), it is especially prone to being unduly impacted by probabilities of 0. This can be avoided by using Laplace estimator or m-esimate, by adding one to all numerators and adding the number of added ones to the denominator (Cestnik, 1990). The assumption of independence among child nodes is clearly almost always wrong and for this reason naïve Bayes classifiers are usually less accurate that other more sophisticated learning algorithms such as ANNs and SVMs.

2.5.3 Support Vector Machines

Support Vector Machines (SVMs) are the newest supervised machine learning technique (Vapnik, 1995). An excellent survey of SVMs can be found in (Burges, 1998), and a more recent book is by (Cristianini & Shawe-Taylor, 2000). In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel tricks, implicitly mapping their inputs into high- dimensional feature spaces.

2.5.4 Comparison of learning algorithms

No single learning algorithm can uniformly outperform other algorithms over all datasets. Features of learning techniques are compared in Table1 (from evidence of existing empirical and theoretical studies),(Dietterich, 2000).

Table1: Comparing learning algorithms (**** stars represent the best and * star the worst performance) , (Dietterich, 2000).

	Decision Trees	Neural Networks	Naïve Bayes	kNN	SVM	Rule-learners
Accuracy in general	**	***	*	**	****	**
Speed of learning with respect to number of attributes and the number of instances	***	*	****	****	*	**
Speed of classification	****	****	****	*	****	****
Tolerance to missing values	***	*	****	*	**	**
Tolerance to irrelevant attributes	***	*	**	**	****	**
Tolerance to redundant attributes	**	**	*	**	***	**
Tolerance to highly interdependent attributes (e.g. parity problems)	**	***	*	*	***	**
Dealing with discrete/binary/continuous attributes	****	***(not discrete)	***(not continuous)	***(not directly discrete)	**(not discrete)	***(not directly continuous)
Tolerance to noise	**	**	***	*	**	*
Dealing with danger of overfitting	**	*	***	***	**	**
Attempts for incremental learning	**	***	****	****	**	*
Explanation ability/transparency of knowledge/classifications	****	*	****	**	*	****
Model parameter handling	***	*	****	***	*	***

2.6 Strategy of the BI System Development

The development of a BI system is usually challenging because it is constrained by the reality of information. System developers must understand the business requirements, formats and deficiencies of the data sources, existing system and various needs of business users. According to Kimball and Caserta (2004), the whole flow for these processes can be viewed as shown in Figure 4.

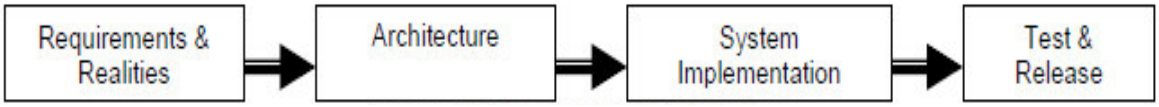


Figure 8: BI system design methodology, Source: Caserta, 2004

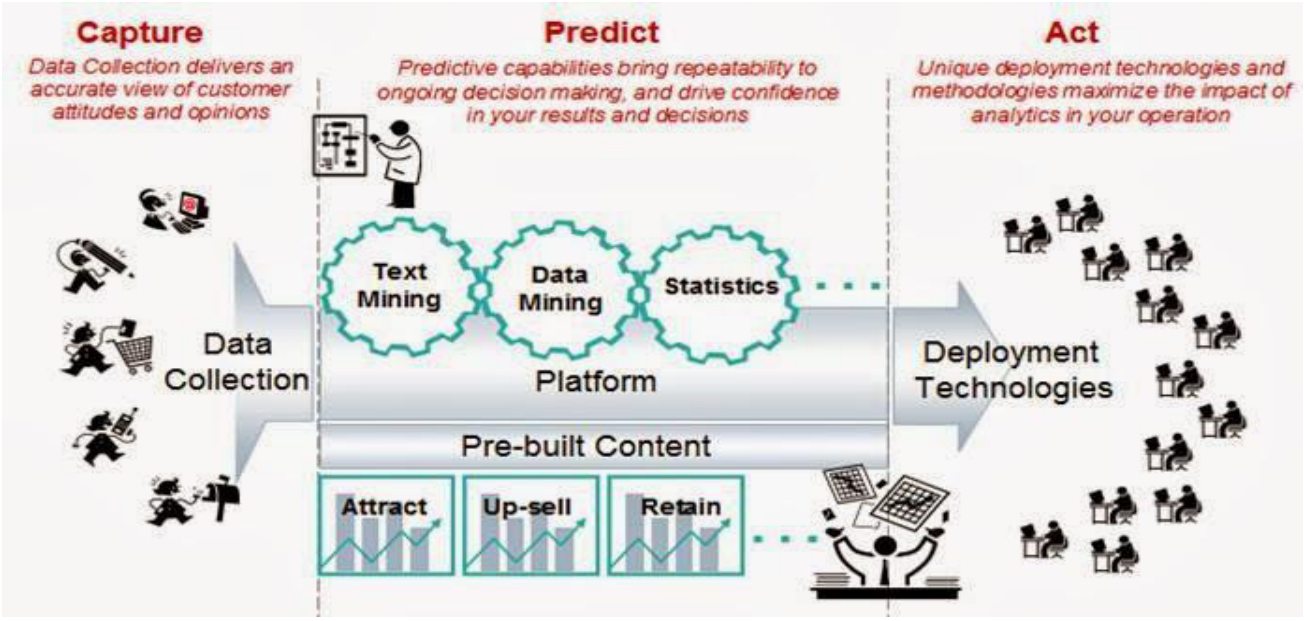


Figure 9: Business Intelligence system process flow (Kharya, 2012).

2.6.1 Success factors of BI implementation

According to (Kimball et al, 2008), there are three critical areas that organizations should assess before getting ready to do a BI project:

a) The level of commitment and sponsorship of the project from senior management

b) The level of business need for creating a BI implementation

c) The amount and quality of business data available.

2.7 Gaps to be filled

Despite the fact that BI can play an important role in student data analysis for decision making and strategic planning and address the issues of retention, most of the current student information systems in higher learning institution are just a collection of student data. BI technologies have not been widely used in higher learning institution (Watson and Wixsom, 2007). This study presents a BI project to generate predictive model for student retention management and construct a BI prototype for predicting the likelihood of student withdrawal .This will help decision makers to know what actions to be taken beforehand in case of drop-out issue.

2.8 Conceptual Framework of the proposed system architecture

The proposed BI Retention prediction System aims to address the challenges of student retention in higher learning institution. Thus, increasing retention has become a goal for many institutions, and a way of judging the quality of education. The proposed framework is presented in figure 10 below.

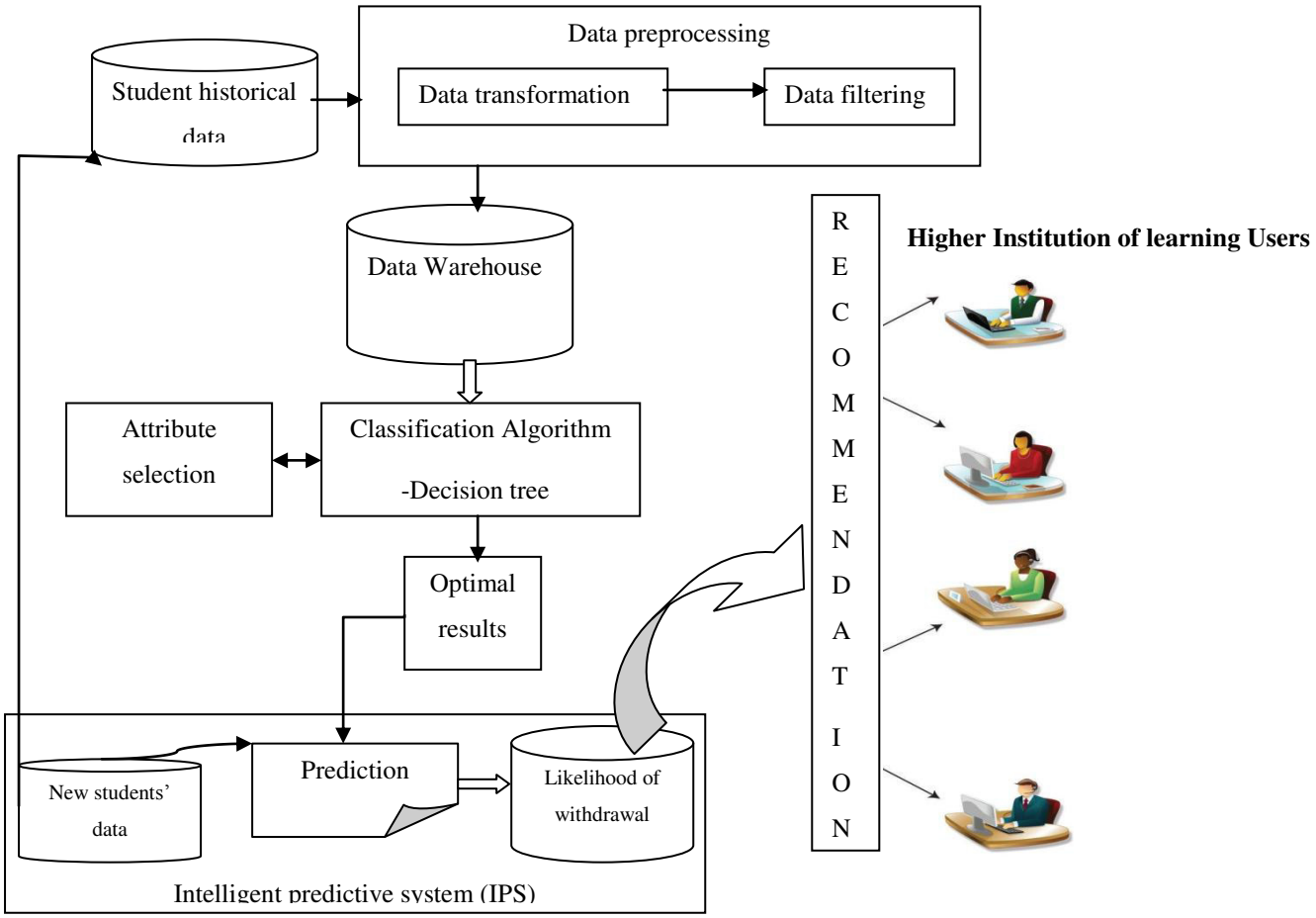


Figure 10: Block diagram of the proposed student retention predictive System.

The proposed retention predictive system aggregates three components:

Data Acquisition and Storage component

The Data Acquisition and Storage component responsible for storing the students' data, gathered from different data sources in a data warehouse.

Data warehousing (DW) is playing a major role in the integration process in BI (Turban et al.,2011). Zeng et al. (2006) suggest that data mining support BI including classification and prediction. The rapidly expanding volume of historical and real-time data contributes to the demand for and provision of data mining tools and it has become a critical role for advanced analytics in BI (Shim et al., 2002).

Model building component

The Model building component, responsible for obtaining knowledge about the students, through appropriates classification algorithms such as decision trees. Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and you move to the next node and the next until you reach a leaf that tells you the predicted output.

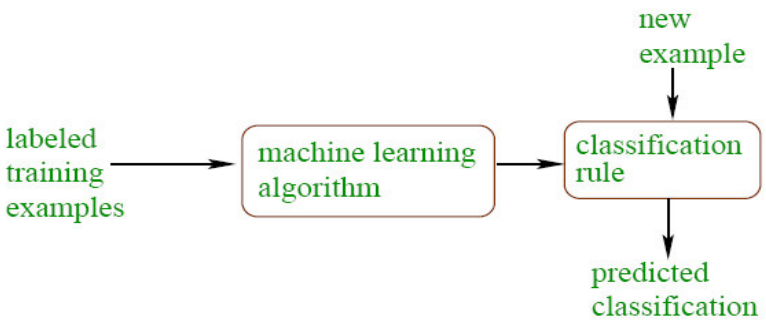


Figure 11: Block diagram of model building (Cestnik, 1990)

Intelligent Predictive System component

The Intelligent Predictive System component responsible for mapping the pattern in the rules generated with the new student data to predict likelihood of withdrawal or persistence.

CHAPTER 3: METHODOLOGY

3.1 Introduction

In this chapter, the research methodology is outlined. Data collection and techniques was also done. It presents the methodology and tools that were used to collect the data and methods of analyzing the data.

The initial requirements for the system was obtained from institutions of higher learning using questionnaires and carrying out interviews. This was done using Spiral model methodology.

Spiral model methodology was used in the system specification, system design and implementation. The spiral model methodology is a systems development lifecycle model which combines the features of the Prototyping Model and the Waterfall Model and has detailed process for specifying, designing, and implementing prototypes. The spiral model is favored for large, expensive and complicated projects.

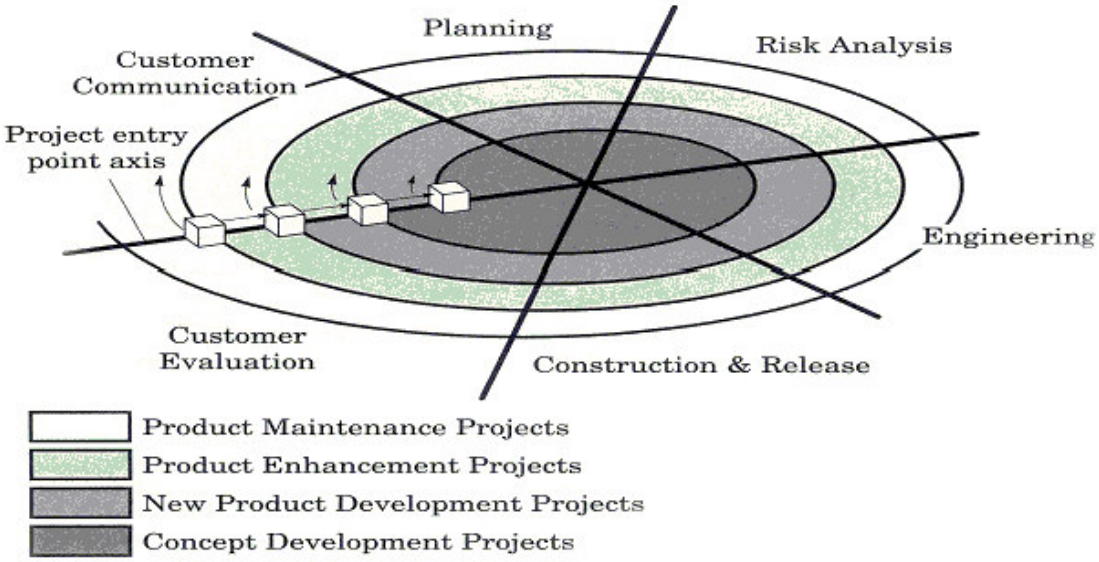


Figure 12: Spiral model methodology (Boehm, 2000).

3.2 Over view of Spiral Model:

Depending on the model it may have 3-6 task regions (framework activities). In this case, the researcher shall consider a '6-task region model'.

The regions are:

1) The customer communication task – to establish effective communication between developer and customer regarding the type of assessment i.e. student, lecturer and course, expected by the user.
2) The planning task – planning the next assessment to be implemented, designing attributes to be considered and dividing implementation work.
3) The risk analysis task – to assess both technical and management risks regarding implementation of the user interface, coding and management of the database.
4) The engineering task – to build one of more representations of the application/ GUI using swing class of Java, connecting WEKA and dataset with Java.
5) The construction and release task – to integrate the application, testing of all test cases, install the executable file and provide user support (e.g., documentation and training).
6) The customer evaluation task – to obtain user feedback based on the evaluation of the software representation created during the engineering stage and implemented during install stage.

Thus, the spiral model was used because of its advantages as follows:

a) This model gives the researcher the opportunity to build a model and then improving it in the next iteration based on the user's feedback.
b) Estimates (i.e. budget, schedule, etc) get more realistic as work progresses because important issues are discovered earlier.
c) It is more able to cope with the (nearly inevitable) changes that software development generally entails.
d) Software engineers (who can get restless with protracted design processes) can get their hands in and start working on a project earlier.

3.3 Requirement analysis: Feasibility and risk analysis

This involves preparation of the Software Requirements Specification (SRS) document. The SRS is a specification of the requirements for the software product we will produce in our project. The basic issues addressed are:

1) Functionality: The software is to be used for analyzing the performance of the students and lecturers by generating reports as well as maintaining records. The user will be provided with various options like student. The user has the option of providing single input or selecting an external file consisting of multiple tuples.

2) Interfaces: The software will have JAVA forms for client interaction. These forms use the JAVA swing class to accept input from authenticated user and to run specific data mining algorithm to service the user's request. The UI will retrieve data from the local host present at the back-end using Java DB connectivity. The data retrieved will then be processed and the appropriate BI techniques will be applied to produce results. The WEKA toolkit will be run using the JAVA to WEKA connectivity.

3) Quality Attributes: The quality attributes include correct authentication of the user which will be based on login-ID and password as well as the size of the dataset which will improve the accuracy of predictions.

4) Design constraints imposed on implementation: The design constraint would be that the server should work consistently without going down. The database would be developed using Microsoft SQL Server 2000 in order to maintain uniformity and consistency. The front-end of the software should be developed using JAVA as per specifications. The linking of the database with the front-end is to be done using the JAVA database connectivity.

3.3.1 Technical feasibility:

This discusses whether it is possible to provide a solution to the proposed problem, given the technical expertise, software systems and other technological resources.

The project is technically feasible due to following issues:-

- The application can work perfectly if there is an access to the JAVA, and WEKA Toolkit.

- It provides easy and fairly accurate results to the user who wants to know about the analysis of student retention issues.
- All the processing and analysis work is done at the back-end without user's knowledge. Thus user is kept hidden from the technical details by providing him only a GUI to interact with the system and giving him fastest possible, optimal results.

3.3.2 Economic feasibility:

The Economic feasibility study of the project was done to ensure that the development of the required software will be profitable for our application and the project can become profitable in the estimated period of time.

1) The project is economically feasible as all the softwares used to implement it (JAVA, and WEKA) are open-source and it provides an in-depth analysis of the available resources with just a click of a button.
2) The project can recover its cost as the current demand for such an application is very high.

3.3.3 Risk Analysis:

Following is the table for various risk factors and their possible impact, along with contingency plan in store to tackle such an event:

RISK ITEM	**CONTINGENCY PLAN**
Deletion of database	Back-up database at regular intervals
Inadequate software capabilities	Using latest back-end facilities
Developing the wrong interface	Prototyping, scenarios, task analysis.

3.4 Software requirement specification document

The purpose of the SRS was to give detailed view of the system, the requirements, the functions, and constraints to the users.

User Interface
The user interface contains a page for login for the authority. There is an interface to do prediction on unlabelled instances. These interfaces will be built using JAVA swing class.

Hardware Interface

- Windows XP or higher
- 256 MB RAM
- Pentium Duo core Processor.
- 1GB Hard Disk Space.

Software Interface
WEKA Data Mining Toolkit an open source toolkit is the main foundation of the project. Using this toolkit, we applied various classification techniques on the dataset. But, this toolkit will be accessed only through JAVA and should not be accessible to the user directly.

Memory Constraints: No specific memory constraints are applicable to the system.

3.5 Methodological framework

The business intelligence model considered in our study was based on supervised learning (classification) techniques given that labeled training data was available. Classification is the process of finding a model that describes and distinguishes data, classes or concepts for the class of objects whose class labels is known. Our methodology consists of data collection, data-preprocessing, building classification model using training data and evaluation of the generated models using test data. Trained and tested model was then used to score incoming data. In this study we used student data from Machakos university college database having 270 attributes and 14 instances. It consist of attributes like DFP,AOE, PO, HTH etc, these attributes predict the likelihood of a student withdrawal. Also different classifiers were applied in the classification such as decision tree, naïve bayes, support vector machine and multilayer perceptrons

3.6 Sources of Data and Target Population

This study used survey-based secondary data provided by Machakos University College. Machakos University College is reliable source since they are public higher learning institution that keeps data about students so as to analyze and deduce information from the data which later enables them to make sound decisions. The outcome of the patterns

are to help policy makers, educational administrators and the affected, to be able to make timely and rational decisions.

3.7 Sampling and Sample Definition

The prevalent data volumes was noted to be relatively low and thus non accommodative on learning from a sample of the data. The entire population of the data collected was thus used in the modeling process.

Sampling was however applied at the data preprocessing stage and to reduce any biases in the sampling process, resample technique was used.

The sampling methodology thus involved the generation of a random subsample of the dataset using sampling without replacement (to ensure an equal chance for every individual attribute to be selected).

3.8 Description of the Basic Dataset

The Basic Dataset in this study refers to the raw dataset that was sourced from the University database systems and files for this study. Student demographic data and course enrollment data was extracted from the student records system as well as from other primary sources.

Table 2: variables in the basic dataset

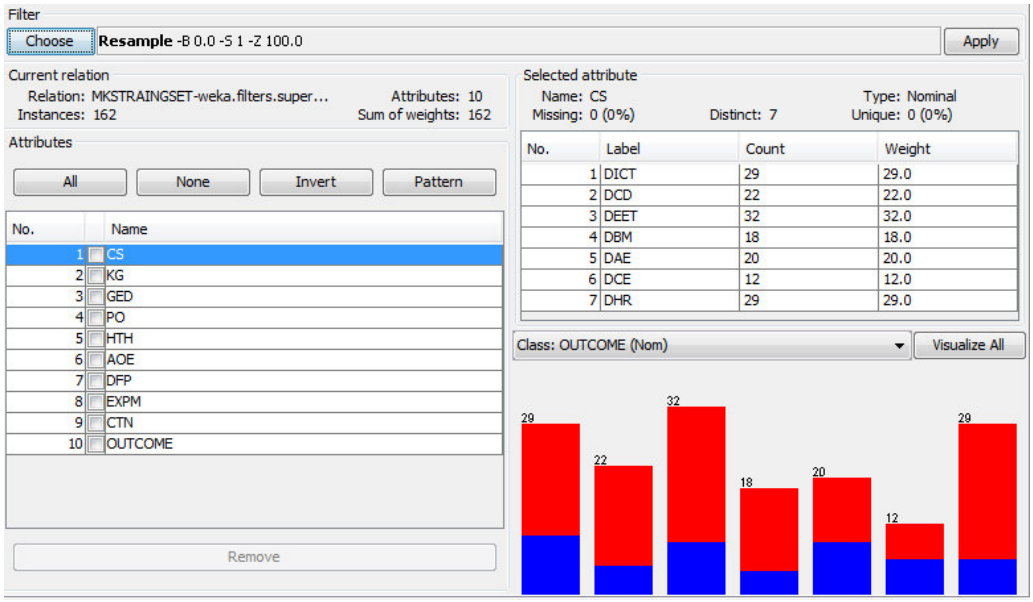
VARIABLES	DESCRIPTION	POSSIBLE VALUES
CS	Course taken	DICT,DCE,DAE,DCD,DBM,DEET,DHR
KG	KCSE grade	C-,C+,B-,B+
GED	Gender	F,M
FEQ	Fathers Education qualification	DEG, SEC.CERT, DIP, PRI.CERT,NONE, MSC, DR
MEQ	Mothers Education qualification	DEG, SEC.CERT, DIP, PRI.CERT,NONE, MSC, DR
DFP	Difficulties in fees payment	NO, YES
FO	Fathers occupation	GOK,UNEMPL,SEMPLOY,NGO
MSP	Marital status of parents	MARRIED, SEPARETED,SINGLE
SP	Sponsor/guardian	PARENT,SELF,,SCHOLARSHIP,ORG
AOE	Age on Entry	BELOW 20, ABOVE 20
EXM	Whether course expectations are met	YES, NO
HTH	Health	GOOD, FAIR, POOR
CTN	Course match	APPROPRIATE, NOT-APPROPRIATE
OUTCOME	Actual outcome	PERSIST, DROPOUT

3.9 Data Preprocessing: Transformation, Selection of attributes

The attributes from the original dataset are not necessarily of the most analytical relevance in the indication and revealing of pattern. Transformations are attribute filters that are done to realize new attributes that could be of increased predictive power. Other filters implemented in this study was remove, a preprocessing technique that omits a range of attributes from the dataset one at time that have lower ranks to improve the accuracy of the classification algorithm.

3.9.0 Data Partition

The input data was randomly divided into three datasets: a training data set, test data set and validation set. The training data set was used to build the model. Model was then tested using test data to compute a realistic estimate of the performance of the model on unobserved data. We used a ratio of 60% of the data used for training, and 30% for testing, and 10% for validation following standard data mining practice



3.9.1 Attribute selection

Attribute selection searches through all possible combinations of attributes in the data and finds which subset of attributes works best for prediction. The attributes relating to students' family background factors and previous academic achievement were considered. The attributes used in this study was ranked in order of importance using information gain and gain ratio measures. Information gain evaluates the worth of an attribute by measuring the information gain with respect to the class whereas gain ratio

evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

Data Collection and Analysis

The data obtained from Machakos University College was preprocessed and then passed through data mining and business intelligence tool namely WEKA where we were able to discover patterns that was helpful in decision making.

3.10 Model building and Validation

3.11.0 Model Building

The model building supported in this study is a classification in the search for the perfect model. The population for which a model is built is further divided into three sets: training, testing and validation. The ratio of the sample population is set at approximately 60%: 30%: 10% with the motivation to avoid occurrence of over-fitting and thus increase model accuracy and applicability in the performance dataset.

3.11.1 Model Validation

Model validation, in most cases relies on stakeholder and data based techniques. In this study, we investigate the usage and automation of the model validation process.

3.11.2 Prototype Development Methodology

The study involved the delivery of a software based prototype. Spiral model development methodology was chosen as the methodology for the development of the prototype. The choice of spiral methodology was based on the fact that a software prototype was to be delivered within a short time.

3.11 Modeling Techniques and Tools Used

The BI model considered in our study was based on supervised learning (classification) technique. The software tool used was WEKA an open-source and free software used for knowledge analysis and downloadable from the internet and used under the GNU license. WEKA implements different machine learning algorithms. The presentation of results and the development of the prototype were done using JAVA while the data will be stored in JavaDB.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Introduction

This chapter outlines the design process and implementation of the software prototype that was built for the purpose of experimentation in this study. The implementation of the system in terms of the data set used, the programming strategies selected and the testing process is outlined.

4.2 Data Analysis and Results

4.2.1 Predictive model/ Basic Classification Results using WEKA

In the classification we used J48, Naïve bayes, Multilayer perceptron and SVM. These classification algorithms were selected because they are considered as "white box" classification model, that is, they provide explanation for the classification and can be used directly for decision making. Each classifier belongs to a different family of classifiers implemented in WEKA. J48 relate to Decision trees, the multilayer perceptron belong to neural networks, Naïve bayes belongs to Bayesian network and SMO belong to support vector machine. Since they are from different classifiers family, they yielded different models that classify differently on some inputs. Attribute importance analysis was carried out to rank the attributes by significance using Information gain and gain ratio attribute evaluators. Ranker's Search method was used to achieve this. The outcome is presented in Table3 and Figure13. The ranking of both attribute evaluators was done using ranker search method. Among the attributes used in this study, it was discovered that DFP, AOE, PO and HTH are the best four attributes. The outcome of both evaluators is similar as shown in Figure 13.

Table 3: Attributes ranking using information gain and gain ratio

GAIN RATIO				**INFORMATION GAIN**			
s/n	Value	Attribute	Rank	s/n	Value	Attribute	Rank
7	0.42436	DFP	1	7	0.35036	DFN	1
6	0.15285	AOE	2	6	0.13401	AOE	2
4	0.06074	PO	3	4	0.11784	PO	3
5	0.03477	HTH	4	5	0.05483	HTH	4
9	0.02686	CTN	5	2	0.04203	KG	5
2	0.01728	KG	6	9	0.0232	CTN	6
8	0.01301	EXPM	7	8	0.01122	EXPM	7
3	0.00399	GED	8	1	0.00792	CS	8
1	0.00293	CS	9	3	0.00394	GED	9

Attribute ranking (with respect to the class attribute) according to information gain and gain ratio criteria show that DFP, AOE, PO and HTH are the best attributes. These attributes outperform other attributes in their contribution to the outcome of students' withdrawal or persistence in HLI as shown in Figure yy below.

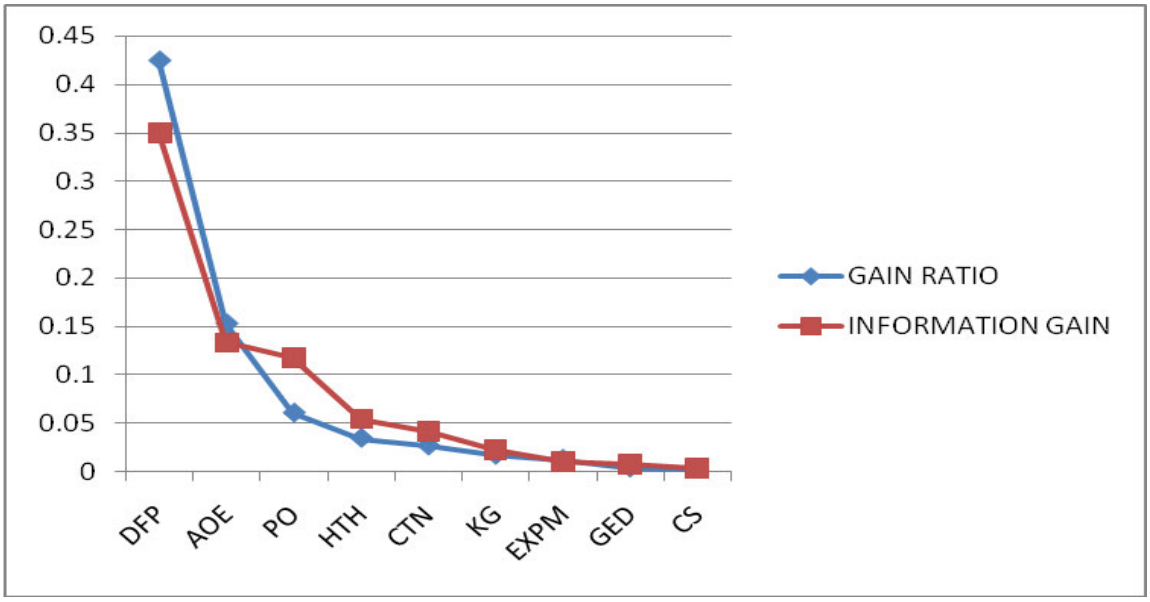


Figure 13: Information gain and gain ratio of the attributes for attribute selection

Comparison of learning algorithms

No single learning algorithm can uniformly outperform other algorithms over all datasets. Features of learning techniques are compared in Table below from the models built.

Sno.	Algorithm	Execution time on 10 – fold cross validation	Accuracy on 10 –fold cross validation	Recall On 10 –fold cross validation
1	J48	0.02sec.	94.8	94.4
2	Naïve Bayes	0.03sec	90.1	90.1
3	Multilayer perceptron	2.23sec	93.4	93.2
4	SVM	0.08 sec	90.3	90.1

Based on all the benchmarks used to measure the algorithms employed in this study, it is discovered that J48 performance is better than all other algorithms. We focus on designing our predictive system on the most suitable algorithm which is J48 in this domain.

4.2.2 Training data set

To produce the model a training data was used, we used a data set with known output values and use this data set to build our model. Then, whenever we have a new data point, with an unknown output value, we put it through the model and produce our expected output. However, this type of model takes an entire training set and divide it into two parts, i.e about 60-70% of the data is taken and put into our training set, which we use to create the model; then the remaining data set is put into a test data set, which we use immediately after creating the model to test the accuracy of our model.

The test data was created to control over fitting, after the model is created it is tested to ensure that the accuracy of the model built does not decrease with the test set. This ensures that our model will accurately predict future unknown values.

Classifier output

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         153               94.4444 %
Incorrectly Classified Instances         9                5.5556 %
Kappa statistic                          0.872
Mean absolute error                      0.0977
Root mean squared error                  0.2211
Relative absolute error                 23.108  %
Root relative squared error             48.1243 %
Coverage of cases (0.95 level)          99.3827 %
Mean rel. region size (0.95 level)      68.8272 %
Total Number of Instances              162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.959     0.062      0.87      0.959     0.913      0.96     DROPOUT
                 0.938     0.041      0.981     0.938     0.959      0.96     PERSIST
Weighted Avg.    0.944     0.047      0.948     0.944     0.945      0.96

=== Confusion Matrix ===

   a   b   <-- classified as
  47   2 |   a = DROPOUT
   7 106 |   b = PERSIST
```

Figure 14: model building using training data set.

Interpretation of results of the training data set

The model classifies 153 instances correctly with an accurate rate of 94.4%, this indicates that the results obtained from training data are optimistic and can be relied on for future or new predictions.

4.2.3 Test data set

Classifier output

```
User supplied test set
Relation:     TESTSET
Instances:     unknown (yet). Reading incrementally
Attributes:   10

=== Summary ===

Correctly Classified Instances          78               96.2963 %
Incorrectly Classified Instances         3                3.7037 %
Kappa statistic                          0.9252
Mean absolute error                      0.0981
Root mean squared error                  0.1965
Coverage of cases (0.95 level)          98.7654 %
Total Number of Instances               81

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.972     0.044      0.946     0.972     0.959      0.968    DROPOUT
                 0.956     0.028      0.977     0.956     0.966      0.968    PERSIST
Weighted Avg.    0.963     0.035      0.963     0.963     0.963      0.968

=== Confusion Matrix ===

  a  b   <-- classified as
 35  1 |  a = DROPOUT
  2 43 |  b = PERSIST
```

Figure 15: Prediction for test data to test model accuracy.

Interpretation of results of the test data set

The model classifies 78 instances correctly with an accurate rate of 96.3%, this indicates that our model will accurately predict future unknown values.

4.2.4 Visualization of results

Visualization is very useful in practice; it helps to determine difficulty of the learning problem by visualizing a 2-D plot of the current working relation

Model perfomance.

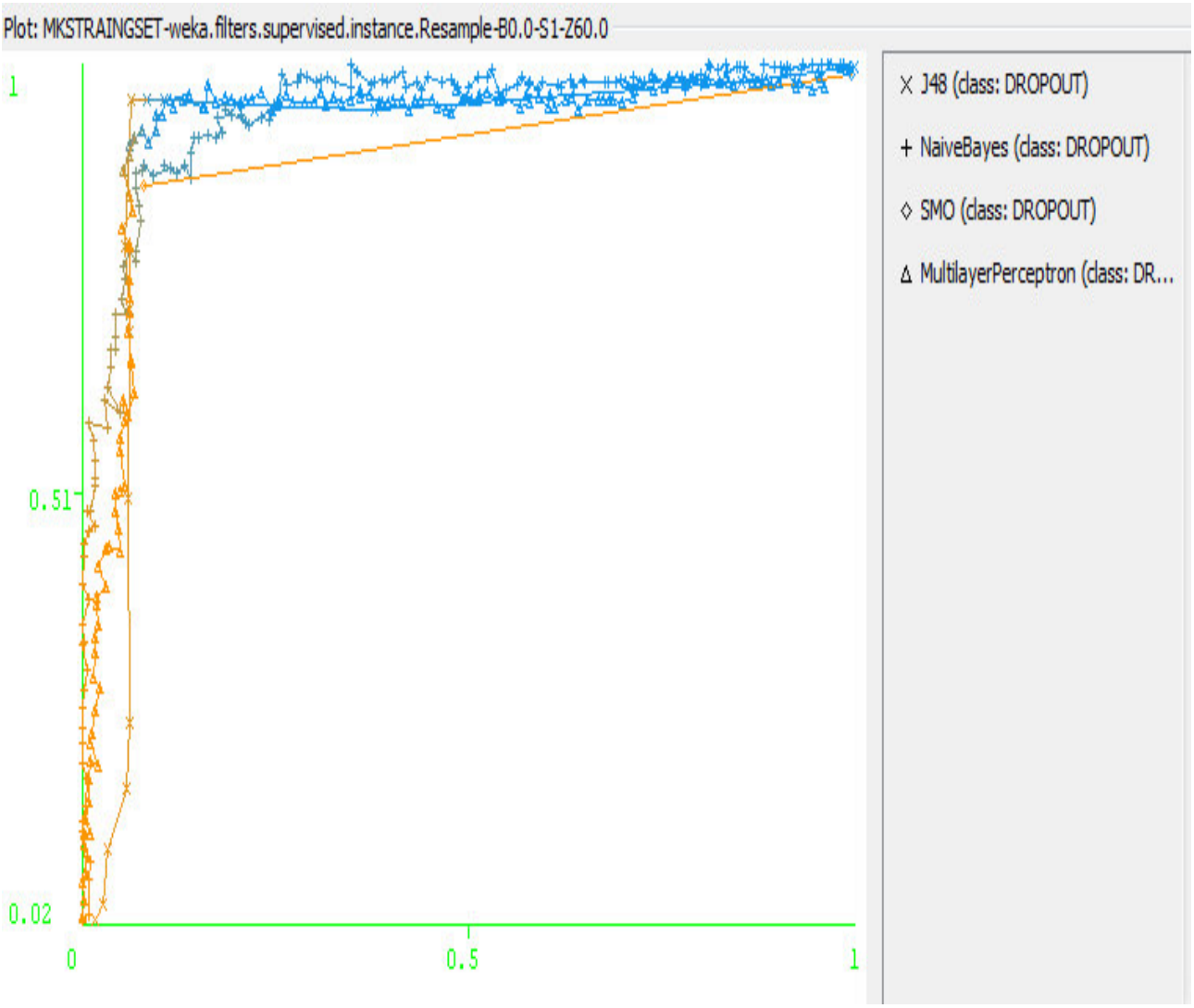


Figure 16: Visualization of models performance chart.

Tree visualization

This is the graphical representation of the classification tree.

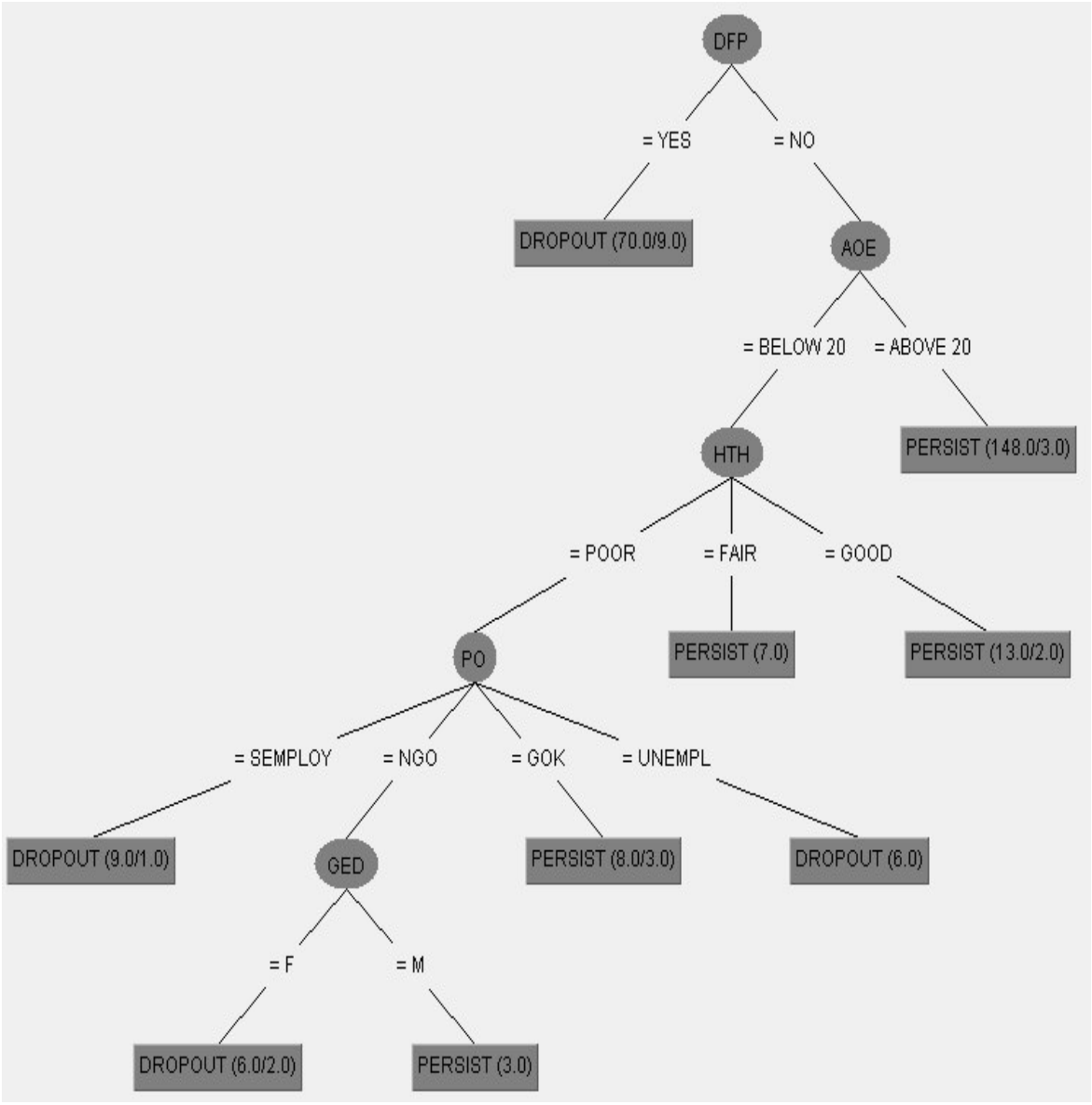


Figure 17: Visualization of decision tree of Dropout.

4.3 Using the classification Algorithm in our Dataset

Classification is used to find a model that segregates data into predefined classes. Classification is based on the features present in the data. The result is a description of the present data and a better understanding of each class in the database. Thus classification provides a model for describing future data. Prediction helps users make a decision. Predictive modeling for knowledge discovery in databases predicts unknown or future values of some attributes of interest based on the values of other attributes in a database.

Classifier output

```
=== Predictions on user test set ===

 inst#     actual  predicted error prediction
     1  1:DROPOUT  1:DROPOUT        0.886
     2  2:PERSIST  2:PERSIST        0.99
     3  2:PERSIST  2:PERSIST        0.99
     4  2:PERSIST  2:PERSIST        0.99
     5  2:PERSIST  1:DROPOUT   +    0.886
     6  2:PERSIST  2:PERSIST        0.99
     7  1:DROPOUT  1:DROPOUT        0.842
     8  2:PERSIST  2:PERSIST        0.99
     9  1:DROPOUT  1:DROPOUT        0.886
    10  1:DROPOUT  1:DROPOUT        0.842
    11  2:PERSIST  2:PERSIST        0.99
    12  2:PERSIST  2:PERSIST        1
    13  1:DROPOUT  1:DROPOUT        0.886
    14  1:DROPOUT  1:DROPOUT        0.842
    15  2:PERSIST  2:PERSIST        0.857
    16  1:DROPOUT  1:DROPOUT        0.886
    17  2:PERSIST  1:DROPOUT   +    0.842
    18  1:DROPOUT  1:DROPOUT        0.886
    19  2:PERSIST  2:PERSIST        0.99
    20  1:DROPOUT  1:DROPOUT        0.842
    21  1:DROPOUT  1:DROPOUT        0.886
    22  2:PERSIST  2:PERSIST        0.99
    23  2:PERSIST  2:PERSIST        0.99
    24  2:PERSIST  2:PERSIST        0.99
    25  1:DROPOUT  1:DROPOUT        0.886
    26  2:PERSIST  2:PERSIST        0.99
    27  1:DROPOUT  1:DROPOUT        0.842
```

Figure 18: prediction results.

4.4 Graphical User Interface

The retention predictor system for HLI was implemented in the following way: The front-end consist of a user-friendly interface implemented using Java.The back-end consist of a datasets. The actual BI classification was implemented using the WEKA Toolkit which assist in procuring insightful patterns in the educational data.

For the construction of the user interface, the Swing package of Java was used. Swing is the primary Java GUI widget toolkit. It provides a graphical user interface (GUI) for Java programs. Swing was developed to provide a more sophisticated set of GUI components than the earlier Abstract Window Toolkit. Swing provides a native look and feel that emulates the look and feel of several platforms, and also supports a pluggable look and feel that allows applications to have a look and feel unrelated to the underlying platform. It has more powerful and flexible components than AWT. The implementation of the interface is shown below:

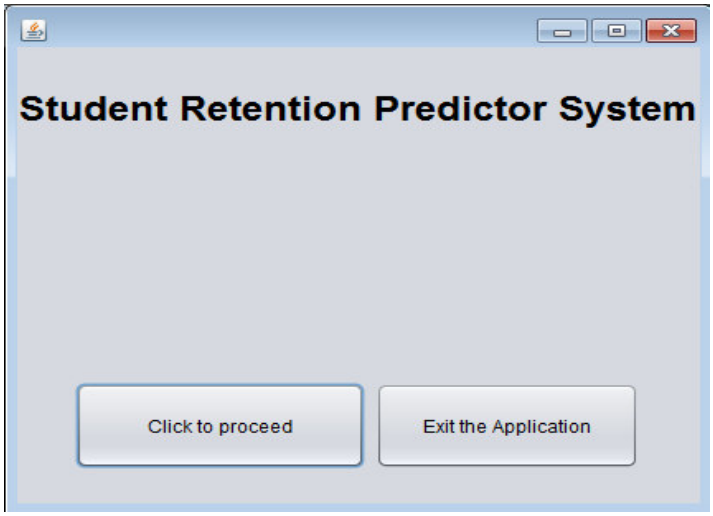


Figure 19: welcome and login window

Dropout Window

This is the next window which will be displayed only if the user gives correct login_ID and password. It presents the user with the main menu which contains various analyses to be performed. This window accepts the student's data and predicts whether the student is liable to drop out. This can be used to take action before-hand to prevent future students from dropping out.

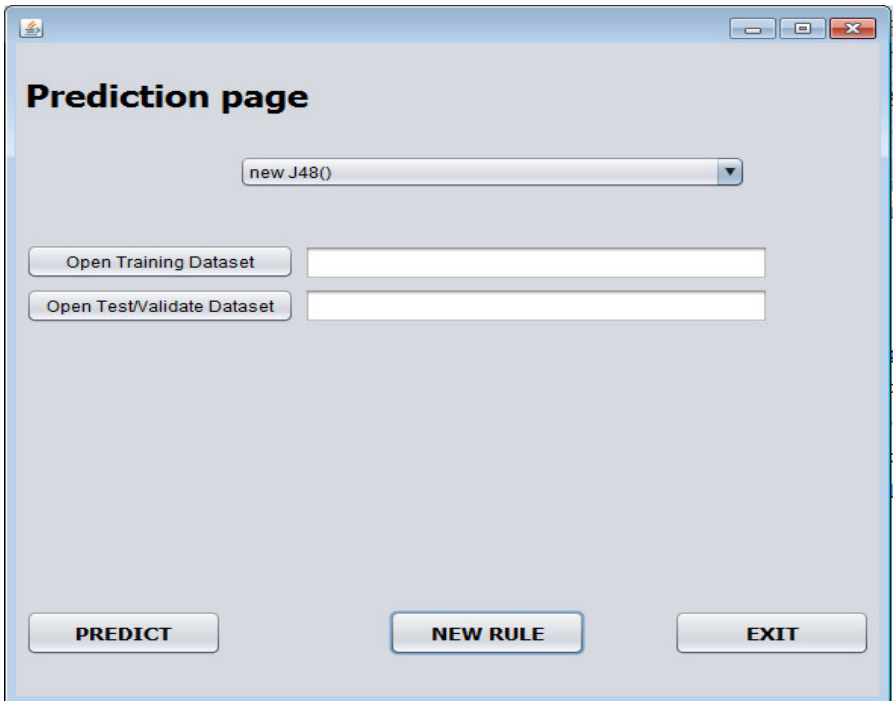


Figure 20: prediction window

4.5 Testing approach

All the components will initially be unit tested to check correct functioning and to check whether they are error free. Also, the error handling capacity of the code will be observed and rectified accordingly.

Table 4: test plan

TEST MODULE	FEATURE TO BE TESTED	TESTING ENVIRONMENT
Login attempt	Validation of forms and password	Manual
Computing parameters	Validation of forms	Manual
Selecting external file	File is present and being read	Manual
Registering WEKA editors	Presence of weka.jar	Manual
Error Handling	Checking the JAVA code	Manual

4.6 Proposed Prototype

The output of the analysis performed was patterns that will enable the Higher Learning institution administrators, student admission officers, students as well as policy makers on making sound decisions. These decisions may include best intervention programmes to address the retention issue and creating awareness. We propose a prototype that will assist HLI to manage retention issues. This system will be used to predict the likelihood of a student to persist or drop out school. The users of this application will be the higher learning institutions

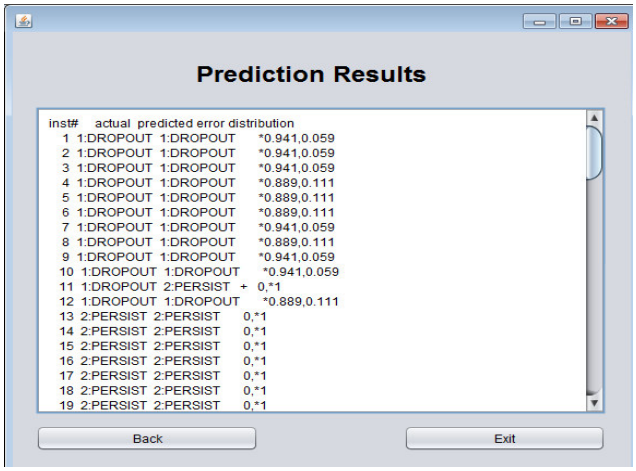


Figure 21: prototype prediction results window.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 CONCLUSION

In our research study we were able to build a prototype that has the ability to load a model and fetch data for prediction from the database of the Higher learning institution. A number of classification models were considered as specified in the literature review and compared in analysis stage out of which we chose to use the decision tree (J48) classifier model because of its performance in adapting it to the data collected. We developed a J48 classifier that integrates an information gain and gain ratio in Waikato Environment for Knowledge analysis (WEKA) Tool Kit and trained it on a preprocessed dataset from a HLI. The results obtained from experiments with the classifier (see Chapter 4 above) show that the classifier is capable of performing classification with an accuracy of 94.4% for dataset obtained from the HLI. Finally, we integrate the techniques and methods developed into a Java based application for use in predicting the likelihood of a student withdrawing in future.

Further this research has shown that it is possible to predict the dropout for different students. The study has revealed, some advantages of J48 model over naïve bayes, multilayer perceptron and support vector machine over other models. One of the advantages realized is that J48 could predict with more accuracy on small volumes of data with noise.

ANN and DT are the methods widely adopted mostly due to their prevalence in the field of BI and proven ability to form models across wide range of application area. More so with advancement in BI the two have proven to be the most versatile and accurate. Also compared with other techniques, they are well established for adoption in performance prediction (vddss et al, 2012).

By gaining a deep understanding of student retention patterns and tendencies, we are enabled to predict which students are most likely to dropout, or those who are most likely to persist. By identifying these students and future prediction of their further outcome, the faculty and managerial decision maker can utilize necessary action and directly or indirectly intervene by providing extra academic counseling, and financial aid. Therefore the Higher Learning institution management system is enabled to improve their policy

making, setting new strategies, and having more advanced decision making procedures. The final result of such model is improving the quality of higher educational system.

Few patterns which we came across during the course of the study are listed below:

1. If difficult in fees payment = YES, then outcome = DROPOUT
2. If difficult in fees payment = NO, student health = GOOD, then outcome = PERSIST
3. If difficult in fees payment = YES, age of entry of student < 20 years, and parent occupation= self employed OR unemployed, then outcome, = DROPOUT
4. If difficult in fees payment = NO, student health=poor, and parent occupation =GOK, then outcome= PERSIST
5. If student health = poor, age of student< 20, parent occupation =NGO, and gender= female then outcome= DROPOUT

5.1.0 To identify different factors that affects student's retention rate

In the literature review, a number of retention models used in prediction of student retention for example Tinto's Student Retention Model, Astin's Theory of Involvement, and Bean's Student Attrition Model have identified that student financial issues, personal, emotional, and family problems, in addition to feelings of isolation and adjustment to college life have frequently been identified as a barrier to completion, especially by students from lower socio-economic groups. Finally it was found that the student age on entry, parent occupation, health of student and financial variables are among the most important predictors of the phenomenon.

5.1.1 To explain business intelligence technologies used in student retention prediction.

In the literature review, a number of BI techniques used in prediction of student retention for example DT, ANN, SVM revealed that the BI systems contribute to improvement and transparency of information flows and knowledge management and discover hidden patterns in student data. Finally the study has revealed some advantages of J48 model over naïve bayes, multilayer perceptron and support vector machine over other models. One of the advantages realized is that J48 could predict with more accuracy on small volumes of data with noise.

5.1.2 To develop and validate BI predictive retention prototype.

A student predictor prototype was developed in Java, WEKA toolkit and My SQL SERVER. The prototype was able to predict the likelihood of student withdrawal.

5.2 CONTRIBUTIONS

This research contributes to the body of knowledge. Further a number of Business Intelligence models have been evaluated on their performance in retention prediction for HLI. The research has revealed that BI technologies can be used efficiently in HLI to enhance education efficiency. On the other hand, the researcher has proposed a system that can be adopted by HLI to perform student retention prediction for better education efficiency.

5.3 MANAGERIAL IMPLICATION

The finding of this research have important implication for HLI specifically registrar. Any HLI that needs to establish its policy upon future dropout prediction may use this finding. Big volumes of past student data are available to many HLI. This data can be a rich source of knowledge, if only properly used. This can be very beneficial for the HLI using BI to extract knowledge and useful information from this available source of data. Thus, one of the managerial implication of this research is to inform managers about the advantages and importance of BI in their strategic planning

5.4 LIMITATIONS FACED

We were not able to collect more information associated to the student social and cultural factors.

5.5 FUTURE WORK

The future scope of the system may provide facilities of generation of more reports to evaluate the retention issue. It can be implemented on a wide basis for all the Higher Learning institutions in Kenya, by associating students' personal information with test score and social-cultural factors in determining retention.

Functionalities for accommodating other classifiers other than the J48 classifier can be developed into the application. These classifiers include Naïve bayes, Support Vector Machines and Multilayer perceptron. Results from the various classifiers can be compared in a report interface for the best classification technique to be selected by the user.

REFERENCES

Abdul-Alim, J. (2011). *Higher education leaders must boost college completion efforts, U.S. Education Department's Kanter says*. Retrieved from http://www.diverse education.com/article/15498/

Aristovnik, A., & Obadić, A. (2011). The Funding and Efficiency of Higher Education in Croatia and Slovenia: A Non-Parametric Comparison with EU and OECD Countries. *SSRN Electronic Journal*. doi:10.2139/ssrn.1735654

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Banks, J. (1993). The canon debate, knowledge construction, and multicultural education. *Educational Researcher*, *22*(5), 4.

Bill Patrick (2001) "Students Matter: Student Retention: who stays and who leaves" *The University Newsletter* http://www.gla.ac.uk/newsletter/227/html/news15.html

Berson, A.; Smith, S. and Thearling, K. (2000). *Building Data Mining Applications for CRM.*New York: McGraw-Hill Professional Publishing.

Chaudhuri, S. (1998). Data Mining and Database Systems: Where is the Intersection? In*IEEE Bulletin of the Technical Committee on Data Engineering, 21(1)*, (pp. 4-8).

Chaudhary, S. (2004). Management factors for strategic BI success. In *Business intelligence in digital economy. Opportunities, limitations and risks*. IDEA Group Publishing.

Dresner, H. J., Buytendijk, F., Linden, A., Friedman, T., Strange, K. H., Knox, M., & Camn, M. (2002).*The business intelligence center: An essential business strategy*. Gartner Research.
Hsu, J. (2004). Data mining and business intelligence: Tools, technology and applications.

H.Frank, E., Hall, M. A., ―Data Mining: Practical Machine Learning Tools and Techniques‖, 3rd Ed. Morgan Kaufmann, 2011.

Hämäläinen, W. and Vinni, M. (2010). Classifiers for educational technology. In C. Romero, S. Ventura, M. Pechenizkiy, R.S.J.d. Baker (eds.), *Handbook of Educational Data Mining,*(pp. 54-74). CRC Press.

Liautaud, B., & Hammond, M. (2002). *E-business intelligence. Turning information into knowledge into profit*. New York: McGraw-Hill.

Olszak, C. M., & Ziemba, E. (2004). Business intelligence systems as a new generation of decision support systems. *Proceedings PISTA 2004, International Conference on Politics and Information Systems: Technologies and Applications*. Orlando: The International Institute of Informatics and Systemics.

Tinto,V. (1975) "Dropout from Higher Education: A Theoretical Synthesis of Recent Research" *Review of Educational Research* vol.45, pp.89-125.

Thomas, L 2002, 'Student Retention in Higher Education: *The role of institutional habitus'*, Journal of Education Policy vol. 17, no. 4, pp. 423-432.

Thomas, L, Quinn, J, Slack, K & Casey, L 2002, Student Services: *Effective approaches to retaining students in higher education, Institute for Access Studies:* Staffordshire University.

Thomas,E.A.M. (2002b) "Student retention in Higher Education: The role of institutional habitus" *Journal of Educational Policy* vol.17 no.4 pp.423-432

Ozga,J & Sukhnandan,L. (1998) "Undergraduate non-completion: Developing an explanatory model" *Higher Education Quarterly* vol.52 no.3 pp.316-333

Mitchell, T. (1997). *Machine Learning.* New York: McGraw Hill

Ryan, M.P. & Glenn, P.A. (2003). *IMachakos University Collegeeasing One-Year Retention Rates By Focusing On Academic Competence:* An Empirical Odyssey. Journal of College Student Retention, 4(3), 297-324. Retrieved June 20, 2007, from Research Library database. (Document ID: 567498231).

Seidman, Alan. *College student retention*: formula for student success. Westport, CT: Praeger Publishers, 2005.print.

Thanassoulis, E., Kortelainen, M., Johnes, G., & Johnes, J. (2010). Costs and efficiency of
higher education institutions in England: a DEA analysis⋆. *Journal of the Operational Research Society*, *62*(7), 1282–1297. doi:10.1057/jors.2010.68

Vygotsky, L. S. (1978). *Mind and society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Appendix One (Source code in Java)

```
package ngemu;

import java.text.DateFormat;
import java.text.SimpleDateFormat;
import java.util.Calendar;
import weka.classifiers.Classifier;
import weka.classifiers.Evaluation;
import weka.classifiers.trees.J48;
//import weka.classifiers.functions.MultilayerPerceptron;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.converters.ConverterUtils;

/**
 *
 * @author user
 */
public class Tumia_Weka {

        public final static void main(String[] args) throws Exception
        {
                test();
        }


        /**
         * Method to use Weka_Use.
         *
         * @throws Exception
         */
        public final static void test () throws Exception
        {
```

```
                DateFormat dateFormat = new SimpleDateFormat("yyyy/MM/dd
HH:mm:ss");
                Calendar cal = Calendar.getInstance();

          //out.setTextArea("\t["+ dateFormat.format(cal.getTime()) +"]");
          //out.setVisible(true);
                System.out.println("\t["+ dateFormat.format(cal.getTime()) +"]");
                String filename;
        filename = "C:\\Users\\bi\\Desktop\\Ngemu\\TRAININGDATASET.arff";
                Instances allData = new
ConverterUtils.DataSource(filename).getDataSet();

                        /* 2. Build Instances */
                double percent = 0.8;
                String attributes_filter = "1-15";//,28-60,63-68,62,70";

                Instances testInstances                 = buildInstancesP(allData,
attributes_filter, true, percent, 1);
                Instances learningInstances = buildInstancesP(allData, attributes_filter,
true, 0, percent);

                        /* 3. Evaluation */
                Classifier classifier     = learning (learningInstances);
                Evaluation eval                         = evaluation(learningInstances,
testInstances, classifier, true);

                System.out.println(eval.errorRate());

                dateFormat = new SimpleDateFormat("yyyy/MM/dd HH:mm:ss");
                cal = Calendar.getInstance();
           //out.setTextArea("\t["+ dateFormat.format(cal.getTime()) +"]");
                System.out.println("\t["+ dateFormat.format(cal.getTime()) +"]");
        }

        /**
         * Used to build an Instances from some Percents of an other one.
```

```
	* It's possible to add a filter on the attributes and choose if the new Instances got
or not
	* only different following rows.
	*
	* @param data
	*                              The Instances to extract the new Instances.
	* @param attributes_filter
	*                              String which represents the attributes to remove.
	*                              <i>"1-4"</i> | <i>"28"</i> | <i>"1-70,45,68-
72"</i> | <i>""</i> | <i>...</i>
	* @param differentNext
	*                              Only different following rows or not?
	* @param percent_start
	*                              Percent indicating the first line of the selection.
	* @param percent_end
	*                              Percent indicating the last line of the selection.
	*
	* @return The new extracted and filtered Instances.
	*
	* @throws Exception
	*/
	public static Instances buildInstancesP (Instances data, String attributes_filter,
boolean differentNext, double percent_start, double percent_end) throws Exception
	{
		/* Security (on percent_start and percent_end) */
		percent_end    = Math.max(percent_end, 0);
		percent_start  = Math.max(percent_start, 0);
		percent_end    = Math.min(percent_end, 1);
		percent_start  = Math.min(percent_start, 1);
		if (percent_end < percent_start){
			double temp = percent_start;
			percent_start = percent_end;
			percent_end = temp;
		}
		/* Rows selection */
```

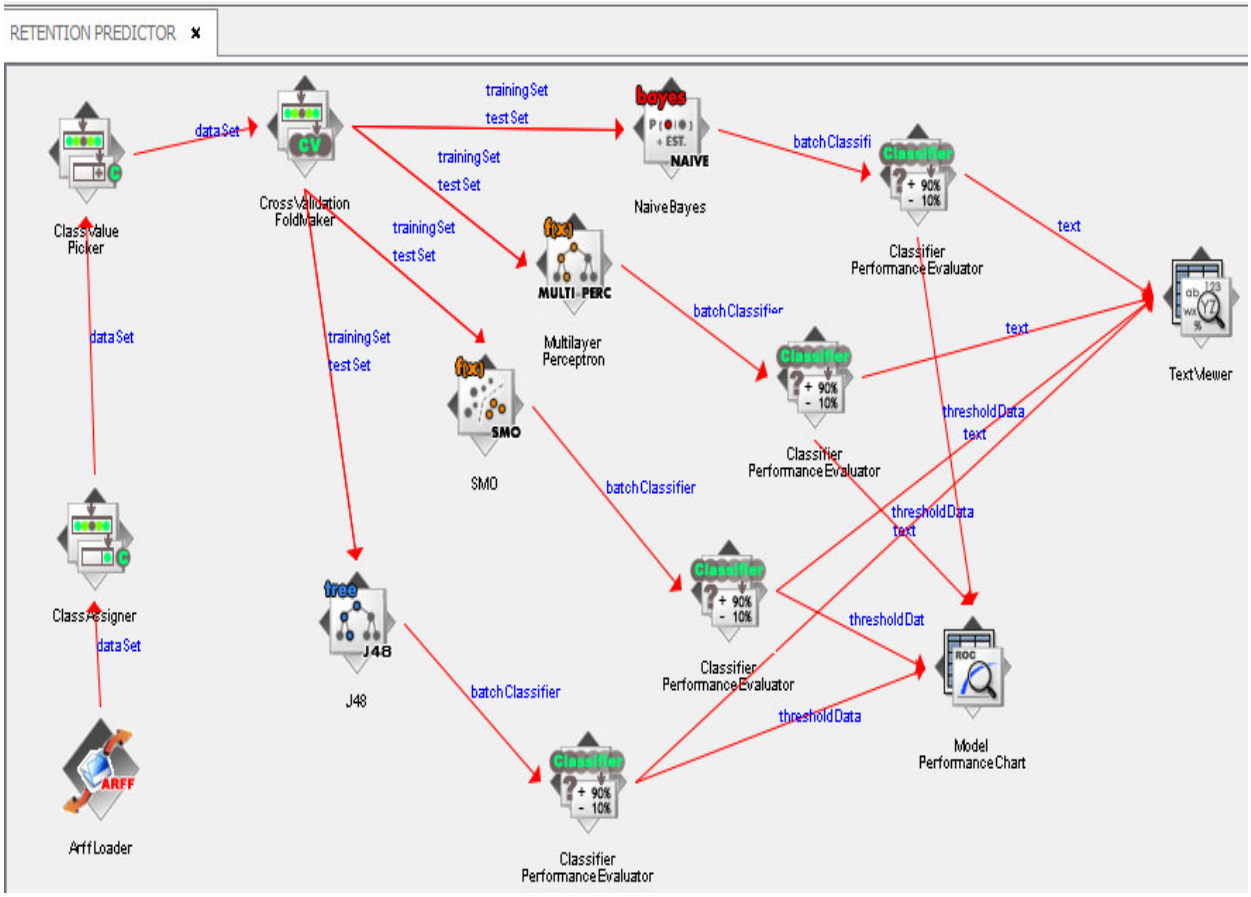
```
		Instances instances = Weka_ManageInstances.percentSelection(data, percent_start, percent_end);
			/* Attributes Filter */
		if (attributes_filter.length()>0) instances = Weka_ManageInstances.attributSelection(instances, attributes_filter);
			/* Set class attributes. */
		if (instances.classIndex() == -1)
			instances.setClassIndex(instances.numAttributes() - 1);
			/* Delete rows whose the next is same. */
		if(differentNext) instances = Weka_ManageInstances.differentNextSelection(instances);

		return(instances);
	}

	/**
	 * Used to build an Instances extracted between two row Numbers of an other one.
	 * It's possible to add a filter on the attributes and choose if the new Instances got or not
	 * only different following rows.
	 *
	 * @param data
	 *				The Instances to extract the new Instances.
	 * @param attributes_filter
	 *				String which represents the attributes to remove.
	 *				<i>"1-4"</i> | <i>"28"</i> | <i>"1-70,45,68-72"</i> | <i>""</i> | <i>...</i>
	 * @param differentNext
	 *				Only different following rows or not?
	 * @param num_start
	 *				Line number indicating the first line of the selection.
	 * @param num_end
	 *				Line number indicating the last line of the selection.
	 *
	 * @return The new extracted and filtered Instances.
	 *
```

```
	 * @throws Exception
	 */
	public static Instances buildInstancesN (Instances data, String attributes_filter, boolean differentNext, int num_start, int num_end) throws Exception
	{
			/* Security (on num_end and num_start) */
		num_end		= Math.max(num_end, 0);
		num_start	= Math.max(num_start, 0);
		num_end			= Math.min(num_end, data.numInstances());
		num_start	= Math.min(num_start, data.numInstances());
		if (num_end < num_start){
			int temp = num_start;
			num_start = num_end;
			num_end = temp;
		}
			/* Rows selection */
		Instances learningInstances = Weka_ManageInstances.rowNumberSelection(data, num_start, num_end);
			/* Attributes Filter */
		if (attributes_filter.length()>0) learningInstances = Weka_ManageInstances.attributSelection(learningInstances, attributes_filter);
			/* Set class attributes */
		if (learningInstances.classIndex() == -1)
			learningInstances.setClassIndex(learningInstances.numAttributes() - 1);
			/* Delete rows whose the next is same. */
		if(differentNext) learningInstances = Weka_ManageInstances.differentNextSelection(learningInstances);

		return(learningInstances);
	}

	/**
	 * Used to train a Classifier.
	 * (In comments, the some other methods to learn).
	 *
```

```
 * @param learningInstances
 *                                  Instances used to build the Classifier.
 * @return
 *                          The trained Classifier.
 *
 * @throws Exception
 */
public static Classifier learning (Instances learningInstances) throws Exception
{
                /* Methods to learn */
        //Classifier classifier = new MultilayerPerceptron();
        Classifier classifier = new J48();
        // Classifier classifier = new PaceRegression();
        // Classifier classifier = new GaussianProcesses();
        // Classifier classifier = new IsotonicRegression();
        // Classifier classifier = new RBFNetwork();
        // Classifier classifier = new SimpleLinearRegression();
        // Classifier classifier = new LinearRegression();
        // Classifier classifier = new SMOreg();
        // Classifier classifier = new LeastMedSq();

        classifier.buildClassifier(learningInstances);

        return (classifier);
}

/**
 * Used to evaluate an Instances of test and display it.
 *
 * @param learningInstances
 *                                  The Instances used to build the Classifier.
 * @param testInstances
 *                                  The Instances of test data.
 * @param classifier
 *                                  The Classifier trained by the learning Instances.
```

```
    * @param displayResults
         *
         * @return The Evaluation of the test data.
         *
         * @throws Exception
         */
        public static Evaluation evaluation (Instances learningInstances, Instances
testInstances, Classifier classifier, boolean displayResults) throws Exception
        {
        Outcome out=new Outcome();
                for (int i = 0; i < testInstances.numInstances() && displayResults; i++)
                {
                        Instance instance = testInstances.instance(i);
                out.setTextArea(instance + " => "
                        + classifier.classifyInstance(instance)+"\n");
                        //System.out.println(instance + " => "
                        //+ classifier.classifyInstance(instance));
               // out.setVisible(true);
                }

                Evaluation eval = new Evaluation(learningInstances);
                eval.evaluateModel(classifier, testInstances);
                if (displayResults)
out.setTextArea(eval.toSummaryString("\nResults\n======\n", true));
                if (displayResults)
System.out.println(eval.toSummaryString("\nResults\n======\n", true));
                out.setVisible(true);
                return(eval);

        }
}
```

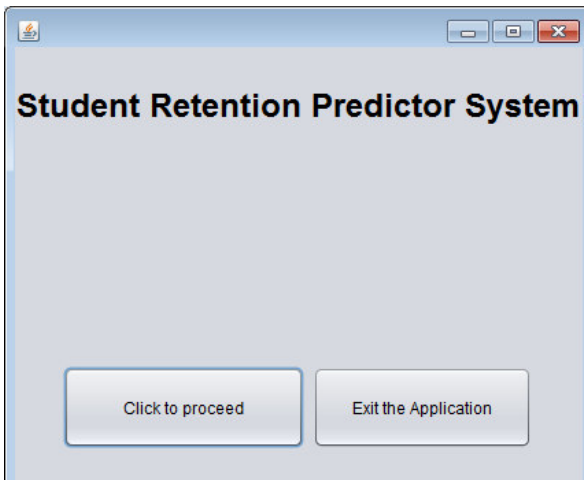
Appendix Two (Systems Documentation)



Appendix Three (User manual)

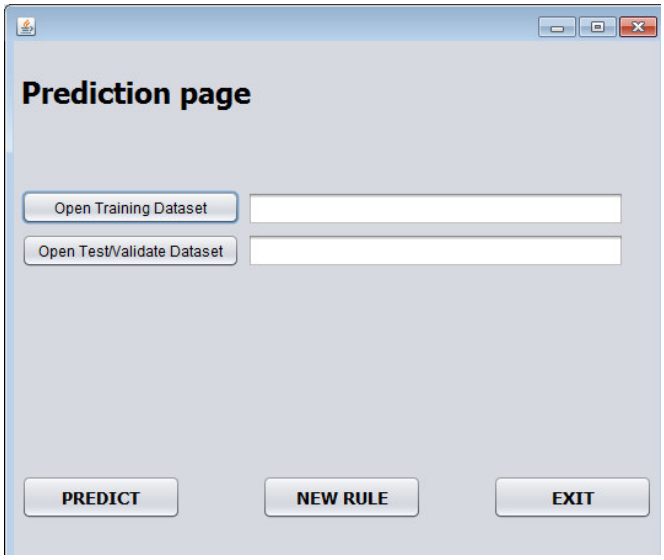
Login

To load the application, run welcome file then click proceed.

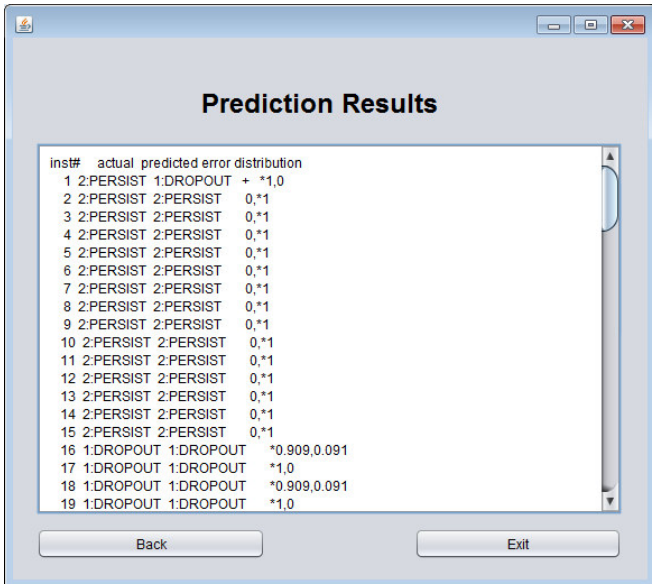


A login window appears as show below

After a successful login the main window for prediction appears



Choose open Training dataset to load training set and open Test dataset tab to load test set then click New Rule tab to choose the classifier and finally click predict Tab. The prediction is done and displayed as below.



Appendix Four (Raw data set)

Part of the Raw Data

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	RN	STN	CS	KG	GED	FEQ	MEQ	FO	HTH	SP	AOE	DFP	EXM	MSP	CTN	AOUTCOME
2	1023	Katiwa M Meshack	DCE	C+	F	DEG	PRI.CERT	GOK	GOOD	PARENT	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST
3	1033	Patrick Musyoka	DCE	C	M	SEC.CERT	NONE	UNEMPL	FAIR	SELF	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST
4	1047	Sylvester Musila Munyiva	DCE	C+	F	DIP	SEC.CERT	SEMPLOY	POOR	SELF	ABOVE 20	NO	NO	SEPARATED	APPRPRIATE	DROPOUT
5	1057	Jackson Ngao Peter	DCE	B-	M	PRI.CERT	SEC.CERT	SEMPLOY	GOOD	ORG	BELOW 20	YES	YES	SINGLE	APPRPRIATE	PERSIST
6	1068	Peter Mwedwa Mbuvi	DCE	C-	M	NONE	DIP	SEMPLOY	POOR	PARENT	BELOW 20	YES	YES	SINGLE	NOT-APPRPRIATE	DROPOUT
7	1074	Mwanzia M Jacob	DCE	C	M	MSC	DIP	UNEMPL	FAIR	SELF	ABOVE 20	YES	YES	SINGLE	APPRPRIATE	PERSIST
8	1083	Edward Mwinzi Nzioka	DCE	C	M	SEC.CERT	PRI.CERT	GOK	FAIR	GURDIAN	ABOVE 20	NO	YES	SINGLE	APPRPRIATE	PERSIST
9	844	Michael Wanjohi Ndungi	DCE	C-	M	DIP	NONE	GOK	GOOD	ORG	BELOW 20	YES	NO	SEPARATED	NOT-APPRPRIATE	DROPOUT
10	1007	Hoseah Mutua Musyimi	DCE	C	M	DIP	DEG	NGO		PARENT	ABOVE 20	NO	YES	SINGLE	APPRPRIATE	PERSIST
11	1129	Daniel Kitonga	DCE	B-	M	PRI.CERT	SEC.CERT	NGO	POOR	SELF	ABOVE 20	YES	YES	MARRIED	APPRPRIATE	PERSIST
12	1133	John Musavi	DCE	C	M	NONE	NONE	GOK	POOR	SCHOLARSHIP	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST
13	1139	John Musau	DCE	C	M	DR	DEG	GOK	GOOD	PARENT	ABOVE 20	NO	YES	SINGLE	NOT-APPRPRIATE	PERSIST
14	1142	Dickson Kasina	DCE	B	M	SEC.CERT	SEC.CERT	SEMPLOY	FAIR	PARENT	ABOVE 20	YES	YES	SINGLE	APPRPRIATE	PERSIST
15	1198	Nicholas Kyalo	DCE	C-	M	NONE	DIP	UNEMPL	GOOD	PARENT	BELOW 20	YES	NO	SINGLE	NOT-APPRPRIATE	DROPOUT
16	1203	Jane Musembi Ngila	DCE	C-	F	DEG	DIP	UNEMPL	FAIR	GURDIAN	ABOVE 20	NO	NO	SINGLE	APPRPRIATE	DROPOUT
17	1291	Robert Mutua Mbuvi	DCE	B-	M	SEC.CERT	PRI.CERT	SEMPLOY	GOOD	SELF	ABOVE 20	YES	YES	MARRIED	NOT-APPRPRIATE	PERSIST
18	1040	Antony Nzeko Kaloki	DCE	C+	M	DIP	NONE	NGO	FAIR	SELF	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST
19	902	Alphonce Mumo Nyamai	DCE	B	M	DIP	PRI.CERT	NGO	GOOD	SELF	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST
20	970	Festus Muema	DCE	C+	M	PRI.CERT	PRI.CERT	GOK	POOR	SCHOLARSHIP	ABOVE 20	NO	YES	SINGLE	APPRPRIATE	PERSIST
21	971	Peter Nthenge	DCE	C	M	NONE	NONE	SEMPLOY	POOR	GURDIAN	ABOVE 20	NO	YES	SINGLE	APPRPRIATE	PERSIST
22	976	Philip Kioko Ngayai	DCE	C-	M	PRI.CERT	DIP	SEMPLOY	POOR	PARENT	BELOW 20	YES	NO	SINGLE	APPRPRIATE	DROPOUT
23	1011	Felis Kitavi David	DCE	B	F	PRI.CERT	DIP	GOK	GOOD	GURDIAN	ABOVE 20	YES	NO	SINGLE	APPRPRIATE	PERSIST
24	1012	Mercy Mulei	DCE	C	F	NONE	DEG	NGO	FAIR	SCHOLARSHIP	ABOVE 20	NO	YES	SINGLE	NOT-APPRPRIATE	PERSIST
25	1017	Mwangi Jackson Mutungi	DCE	B-	M	DIP	DEG	SEMPLOY	POOR	PARENT	BELOW 20	YES	NO	MARRIED	NOT-APPRPRIATE	DROPOUT
26	1022	Fredrick Kyungu	DAE	C	M	DIP	SEC.CERT	SEMPLOY	GOOD	SELF	ABOVE 20	NO	YES	SINGLE	NOT-APPRPRIATE	PERSIST
27	1045	Christopher Mutua Mutei	DAE	B	M	DEG	DEG	SEMPLOY	GOOD	SCHOLARSHIP	ABOVE 20	NO	NO	SINGLE	NOT-APPRPRIATE	PERSIST
28	1093	Timothy Nyamai Musau	DAE	C+	M	DEG	DEG	GOK	POOR	ORG	BELOW 20	YES	NO	SINGLE	NOT-APPRPRIATE	DROPOUT
29	1096	Richard Mutwiwa Mutua	DAE	B-	M	SEC.CERT	DEG	GOK	FAIR	SELF	ABOVE 20	YES	YES	SINGLE	APPRPRIATE	PERSIST
30	1098	Kelvin Muthini Mutua	DAE	C+	M	DEG	DIP	GOK	GOOD	SCHOLARSHIP	ABOVE 20	NO	NO	SINGLE	APPRPRIATE	PERSIST
31	1099	Dennis Ivuti Muendo	DAE	B	M	DEG	DEG	GOK		GURDIAN	ABOVE 20	NO	YES	MARRIED	APPRPRIATE	PERSIST

Appendix Five (Test results)

Model classifier

Classifier output

```
=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         153               94.4444 %
Incorrectly Classified Instances         9                5.5556 %
Kappa statistic                          0.872
Mean absolute error                      0.0977
Root mean squared error                  0.2211
Relative absolute error                 23.108  %
Root relative squared error             48.1243 %
Coverage of cases (0.95 level)          99.3827 %
Mean rel. region size (0.95 level)      68.8272 %
Total Number of Instances              162

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.959     0.062      0.87      0.959     0.913      0.96     DROPOUT
                 0.938     0.041      0.981     0.938     0.959      0.96     PERSIST
Weighted Avg.    0.944     0.047      0.948     0.944     0.945      0.96

=== Confusion Matrix ===

   a   b   <-- classified as
  47   2 |   a = DROPOUT
   7 106 |   b = PERSIST
```

Prediction on validation data

```
=== Re-evaluation on test set ===

User supplied test set
Relation:     VALIDATIONSET
Instances:     unknown (yet). Reading incrementally
Attributes:   10

=== Predictions on user test set ===

 inst#     actual  predicted error prediction
     1  1:DROPOUT  1:DROPOUT        0.886
     2  2:PERSIST  2:PERSIST        0.99
     3  2:PERSIST  2:PERSIST        0.99
     4  2:PERSIST  2:PERSIST        0.99
     5  2:PERSIST  1:DROPOUT   +    0.886
     6  2:PERSIST  2:PERSIST        0.99
     7  1:DROPOUT  1:DROPOUT        0.842
     8  2:PERSIST  2:PERSIST        0.99
     9  1:DROPOUT  1:DROPOUT        0.886
    10  1:DROPOUT  1:DROPOUT        0.842
    11  2:PERSIST  2:PERSIST        0.99
    12  2:PERSIST  2:PERSIST        1
    13  1:DROPOUT  1:DROPOUT        0.886
    14  1:DROPOUT  1:DROPOUT        0.842
    15  2:PERSIST  2:PERSIST        0.857
    16  1:DROPOUT  1:DROPOUT        0.886
    17  2:PERSIST  1:DROPOUT   +    0.842
    18  1:DROPOUT  1:DROPOUT        0.886
    23  2:PERSIST  2:PERSIST        0.99
    24  2:PERSIST  2:PERSIST        0.99
    25  1:DROPOUT  1:DROPOUT        0.886
    26  2:PERSIST  2:PERSIST        0.99
    27  1:DROPOUT  1:DROPOUT        0.842

=== Summary ===

Correctly Classified Instances          25               92.5926 %
Incorrectly Classified Instances         2                7.4074 %
Kappa statistic                          0.8525
Mean absolute error                      0.1324
Root mean squared error                  0.2532
Coverage of cases (0.95 level)         100      %
Total Number of Instances               27

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.133      0.857     1         0.923      0.939    DROPOUT
                 0.867     0          1         0.867     0.929      0.939    PERSIST
Weighted Avg.    0.926     0.059      0.937     0.926     0.926      0.939

=== Confusion Matrix ===

  a  b   <-- classified as
 12  0 |   a = DROPOUT
  2 13 |   b = PERSIST
```