Africa Journal of Technical & Vocational Education and Training

Enrolment Management in Higher Learning Institutions: Student Retention Prediction

Joseph Ngemu
Wote Technical Training Institute, Makueni, Kenya

Abstract

Student retention has become one of the most important priorities for decision makers in higher learning institutions (HLI). The increasing competition for students among tertiary institutions has resulted in a greater emphasis on student retention. Improving student retention starts with a thorough understanding of the reasons behind the attrition. In an effort to address this issue, the study used student demographic and institutional data along with several business intelligences (BI) techniques and analytical tools, to develop prototype to predict likelihood of student persistence or dropout with the goal to identify factors that can be used to identify students who are at risk of dropping out of tertiary institution program. 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 attribute selection 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 business intelligence system that can serve as valuable tool in predicting student withdrawal or persistence as well as recommend the necessary intervention strategies to adopt, leading to better education efficiency and graduation rate.

Key words: ***Business intelligence, retention, attrition, WEKA, classifiers***

ISSN 2518-2722

Introduction

Recently, management information systems, ERPs and online systems in education have increased, and student digital data has come to big data size. This makes possible to draw rules and predictions about the students by processing educational data with data mining techniques. All kinds of information about the student's socioeconomic environment, learning environment, demographic or course enrolment data can be used for prediction, which affect the persistence or withdrawal of a student.

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, 2004). 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.

Purpose of the Study

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.

Specific Research Objectives

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

ii. Apply business intelligence concepts in the modelling process for the prediction of likelihood of dropping out.

iii. Construct a BI prototype for predicting likelihood student withdrawal

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

Literature Review

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 (Olszak & Ziemba, 2004). 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 (Watson & Wixsom, 2007).

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).

Despite the fact that BI can play an important role in student data analysis for decision making and strategic planning and in addressing 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 & 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 dropout issue.

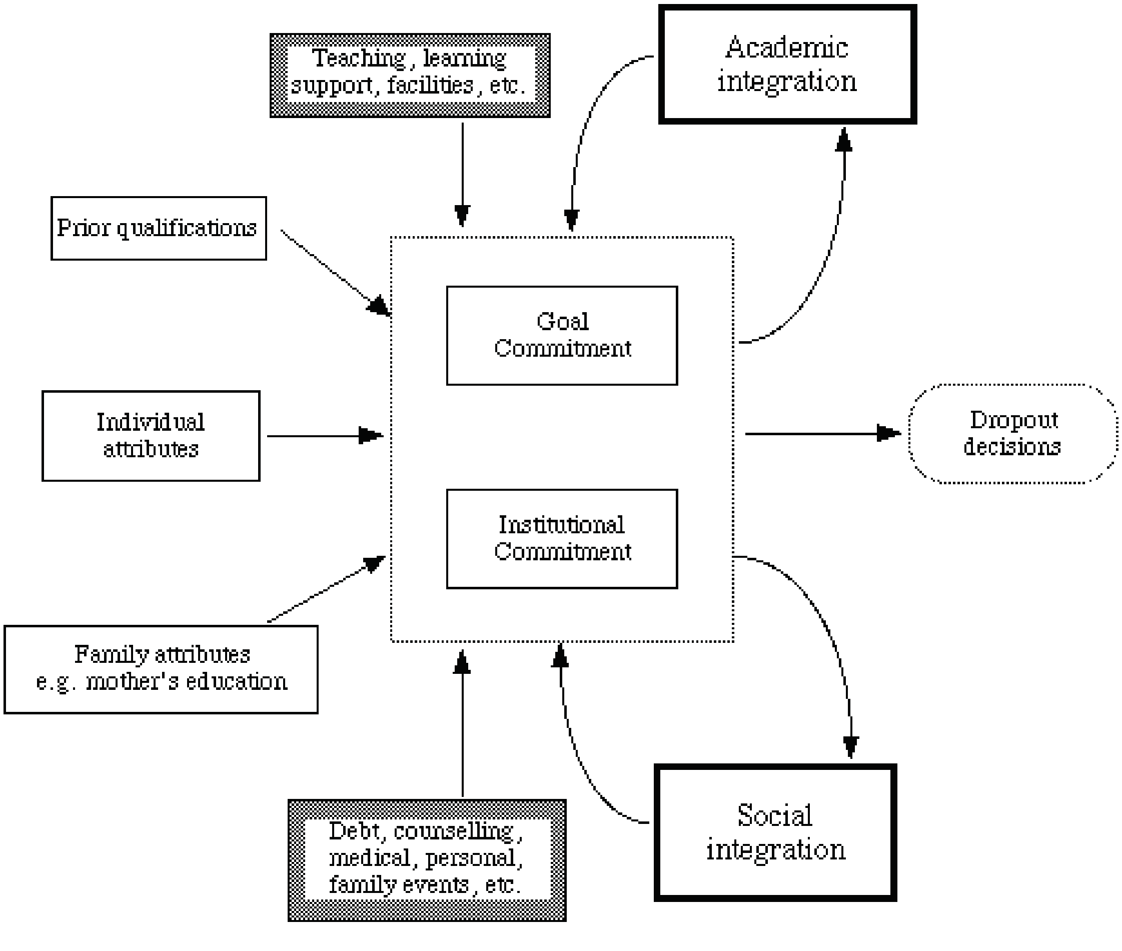

Teaching, learning support, facilities, etc.
Academic integration
Prior qualifications
Goal Commitment
Individual attributes
Dropout decisions
Institutional Commitment
Family attributes e.g. mother's education
Debt, counselling, medical, personal, family events, etc.
Social integration

Figure 1: Student Retention Model (Tinto, 1975) Dropout from Higher Education

Methodology

In this study, the retention of the students at the end of academic year are estimated by using the student data obtained from higher learning institution. The aim of this study is to find ways of improving the efficiency of higher learning institution systems by applying business intelligent techniques to predict the students' withdrawal to support educators to take precautions for the children at risk. A number of data pre-processing were applied to increase the accuracy rate of the predictor model. A wrapper method for feature subset selection was applied to find the optimal subset of features. After that, four popular business intelligence techniques/algorithms (Decision tree, Naïve bayes, multi-layer perceptron and support vector machine) were used and compared in terms of classification accuracy rate.

In this study, four well-known classification algorithms (Decision tree, Naïve bayes, multilayer perceptron and support vector machine) were employed on the educational datasets to predict the likelihood of withdrawal of students.

Naive Bayes

Naive Bayes classifiers are a family of algorithms. These classifiers are based on Bayes' Theorem, which finds the possibility of a new event based on previously occurring events. Each classification is independent of one another but has a common principle.

Decision Tree

A decision tree uses a tree like graph. Decision trees are like flowchart but not noncyclic. The tree consists of nodes and branches. Nodes and branches are arranged in a row. Root node is on the top of a tree and represents the entire dataset. Entropy is calculated when determining nodes in a tree. It models decisions with efficacy, results, and resource costs. In this study, decision tree technique is preferred because it is easy to understand and interpret.

Multilayer Perceptron

Multilayer perceptron (MLPs) is flexible machine learning techniques that can fit complex nonlinear mappings. MLPs are the most popular neural network type, consisting on a feedforward network of processing neurons that are grouped into layers and connected by weighted links.

Support Vector Machines

Support vector machines is flexible machine learning techniques that can fit complex nonlinear mappings. Support vector machines transforms the input variables into a high dimensional feature space and then finds the best hyperplane that models the data in the feature space

Figure 1 illustrates the workflow of data mining model for classification. In the first step, feature selection algorithms are applied on the educational data. Next, classification algorithms are used to build a good model which can accurately map inputs to desired outputs. The model evaluation phase provides feedback to the feature selection and learning phases for adjustment to improve classification performance. Once a model is built, then, in the second phase, it is used to predict label of new student data.

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 2.

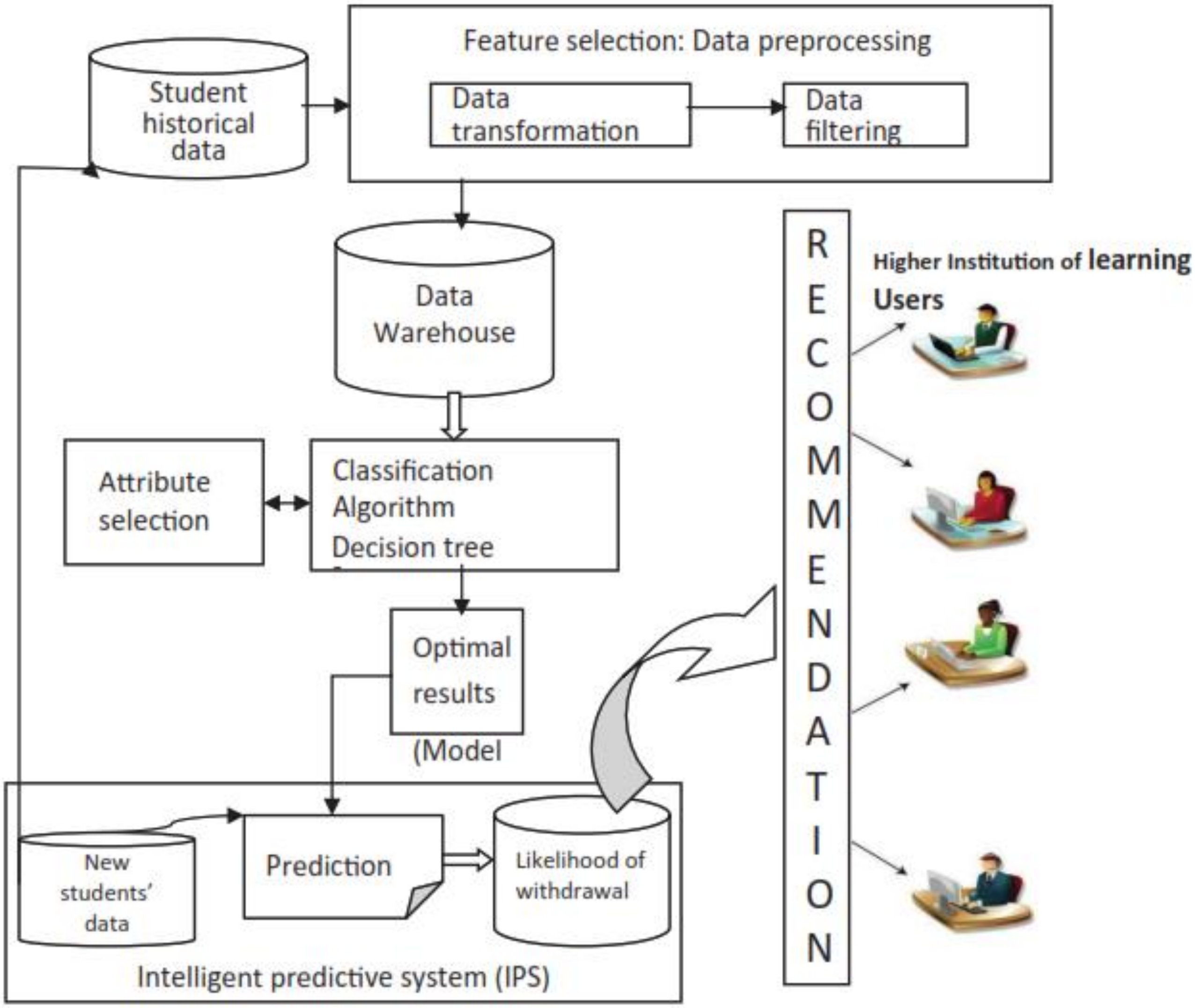


Figure 2. Block Diagram of the Proposed Student Retention Predictive Model

Experimental Studies

In this study, the feature subset selection and classification operation were conducted by using WEKA open-source data mining software (Hall, Frank, Holmes, Pfahringer & Reutemann, 2009). In each experiment, 10-fold cross-validation was performed to evaluate the classification models. The classification accuracy of the algorithms for the test data was measured as given in Eq.1:

$$Accuracy\ (T) = \frac{\sum_{l=1}^{|T|} eval(t1)}{|T|} \qquad where\ \mathrm{eval}(t) = \begin{cases} 1, if\ classify\ (t) = C \\ 0, if\ classify(t) \neq C \end{cases} \quad \ldots \text{Eq 1.}$$

Where *T* is a test set that consists of a set of data items to be classified; *c* is the actual class of the item *t, where t € T;* and *classify (t)* returns the classification output of *t* by the algorithm.

Dataset description.

In this study, higher learning dataset (Hämäläinen & Vinni, 2010) was used to predict student withdrawal. The data set was sourced from the University database systems and files for this study. Dataset attributes are Student demographic data and course enrollment data extracted from the student records system as well as from other primary sources. The dataset have 14 attributes as shown in Table 1.

Table 1
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

Data Partition

The input data was randomly divided into three datasets: a training data set, test dataset and validation set. The training dataset was used to build the model. The model was then tested using test data to compute a realistic estimate of the performance of the model on unobserved data. A ratio of 60% of the data was used for training, and 30% for testing, and 10% for validation following standard data mining practice (Frank & Hall, 2011).

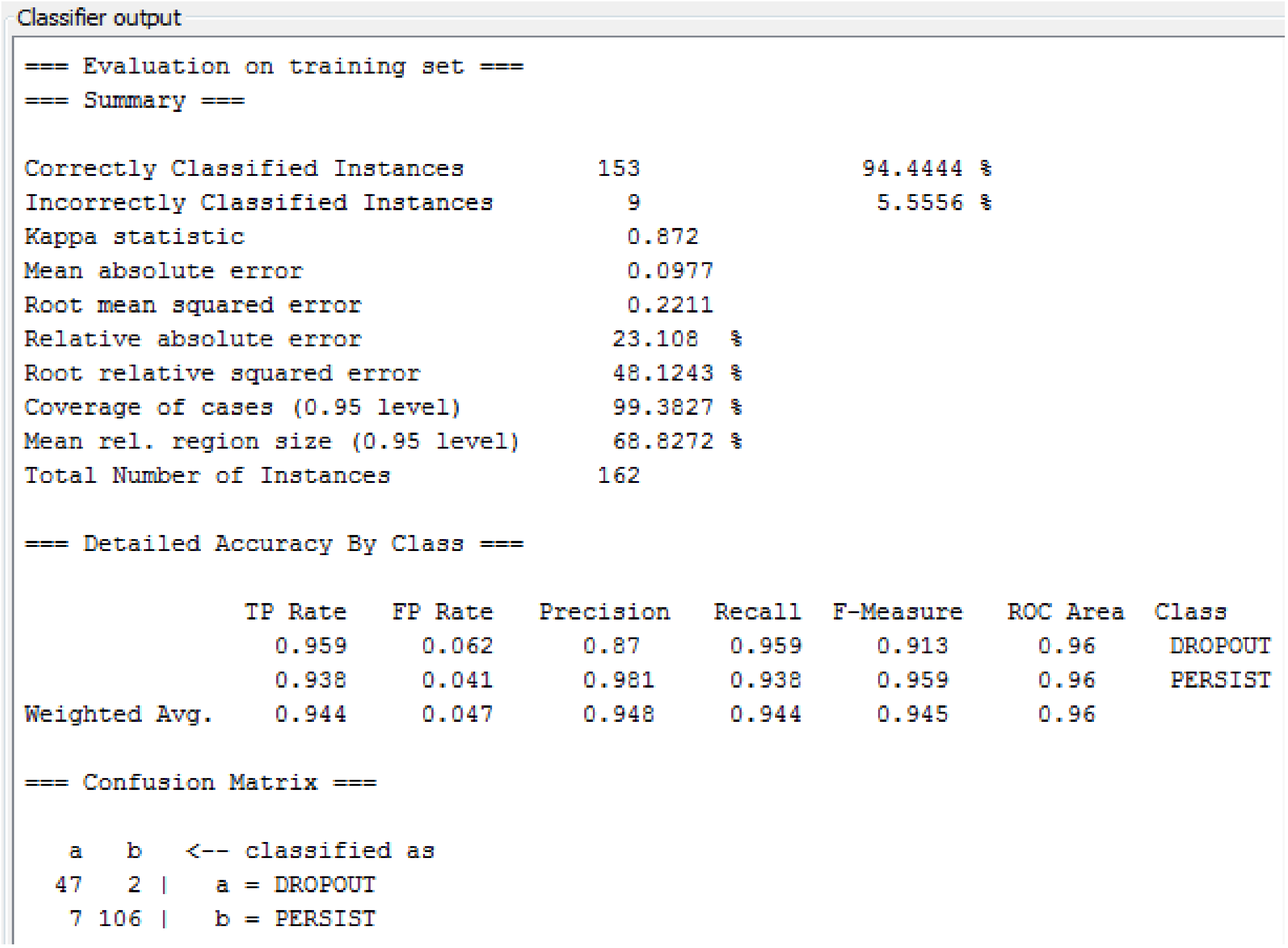
Data Analysis and Results

In the classification J48, Naïve bayes, Multilayer perceptron and SVM were used. 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 Table 2. 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 were the best four attributes.

Table 2

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

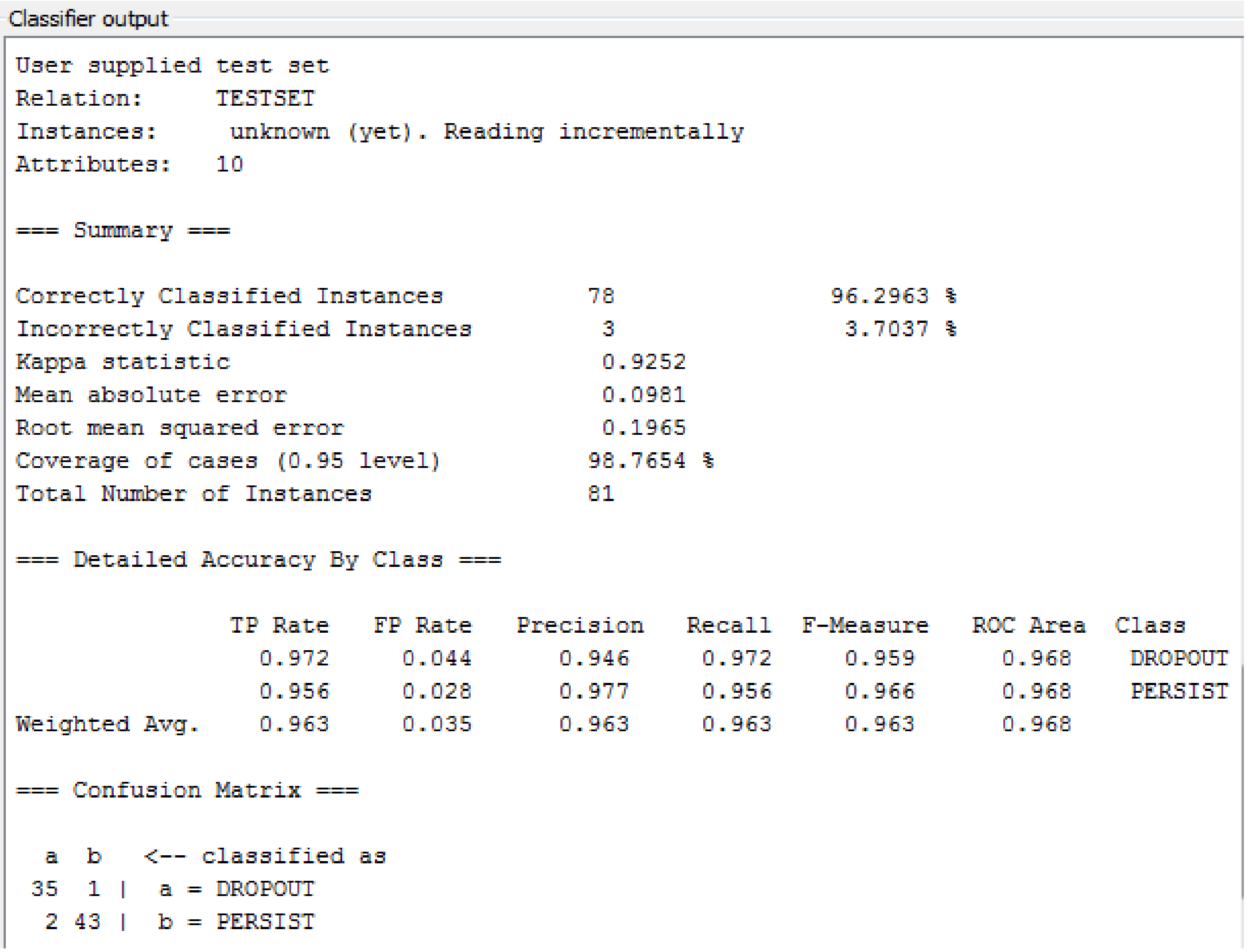


```
=== Evaluation on training set ===
=== Summary ===

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 3. Screenshot of Model building using training data set

Interpretation of Results of the Training Dataset

The model classifies 153 instances correctly with an accurate rate of 94.4%, as depicted in Figure 4, this indicates that the results obtained from training data are optimistic and can be relied on for future or new predictions



```
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 4: Screenshot of Prediction for Test Data to Test Model Accuracy

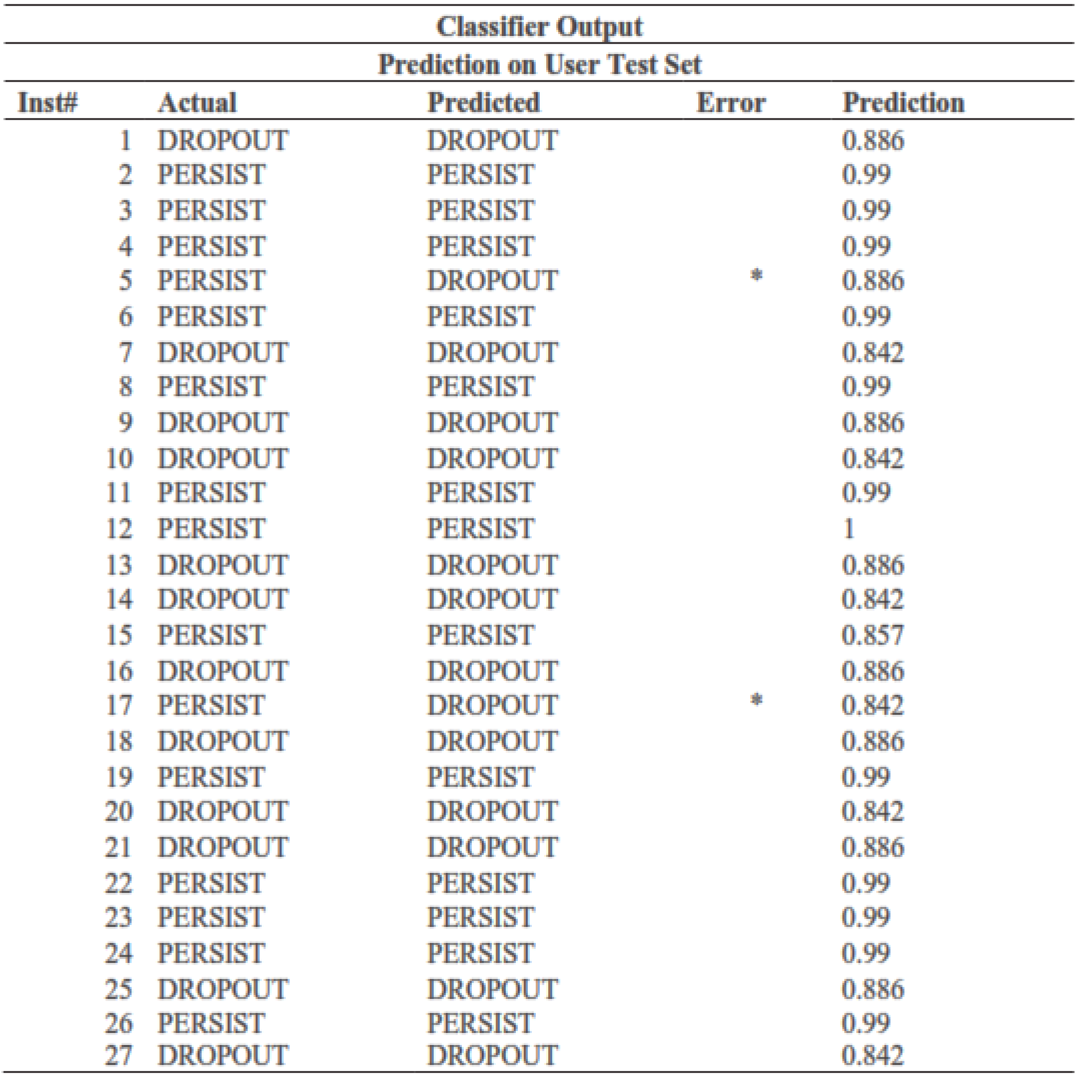
Interpretation of Results of the Test Dataset

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

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 modelling 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.

Table 3
Prediction Results



Classifier Output				
Prediction on User Test Set				
Inst#	**Actual**	**Predicted**	**Error**	**Prediction**
1	DROPOUT	DROPOUT		0.886
2	PERSIST	PERSIST		0.99
3	PERSIST	PERSIST		0.99
4	PERSIST	PERSIST		0.99
5	PERSIST	DROPOUT	*	0.886
6	PERSIST	PERSIST		0.99
7	DROPOUT	DROPOUT		0.842
8	PERSIST	PERSIST		0.99
9	DROPOUT	DROPOUT		0.886
10	DROPOUT	DROPOUT		0.842
11	PERSIST	PERSIST		0.99
12	PERSIST	PERSIST		1
13	DROPOUT	DROPOUT		0.886
14	DROPOUT	DROPOUT		0.842
15	PERSIST	PERSIST		0.857
16	DROPOUT	DROPOUT		0.886
17	PERSIST	DROPOUT	*	0.842
18	DROPOUT	DROPOUT		0.886
19	PERSIST	PERSIST		0.99
20	DROPOUT	DROPOUT		0.842
21	DROPOUT	DROPOUT		0.886
22	PERSIST	PERSIST		0.99
23	PERSIST	PERSIST		0.99
24	PERSIST	PERSIST		0.99
25	DROPOUT	DROPOUT		0.886
26	PERSIST	PERSIST		0.99
27	DROPOUT	DROPOUT		0.842

Tree Visualization

This is the graphical representation of the classification tree.

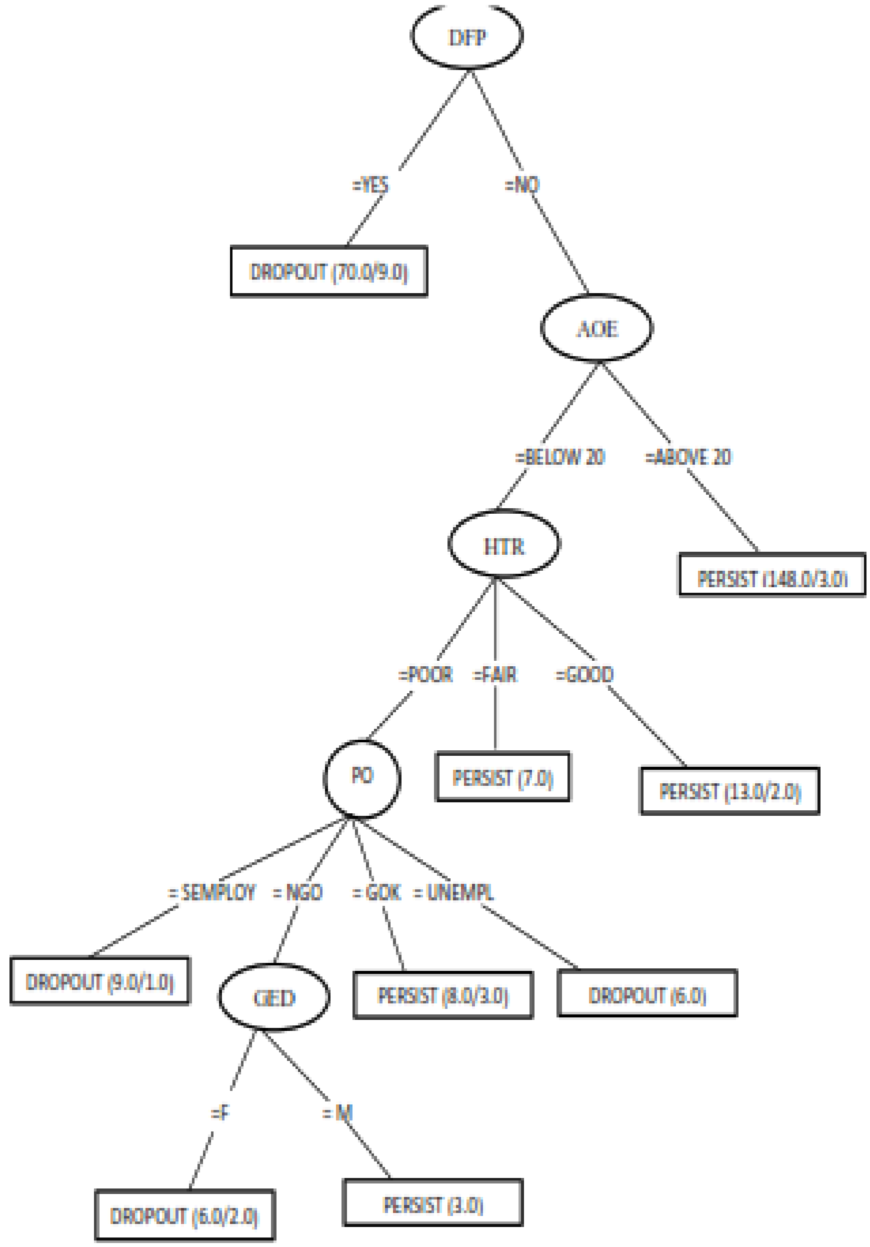


Figure 7: Visualization of decision tree of Dropout

Conclusions

In our research study, we were able to build a model that has the ability 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 pre-processed dataset from a HLI. The results obtained from experiments with the classifier show that the classifier is capable of performing classification with an accuracy of 94.4% for dataset obtained from the HLI. Finally, techniques and methods developed were integrated into a Java based application for use in predicting the likelihood of a student withdrawing in future.

Few patterns, which came across during the course of the study, are:

1. The drop out features seems to be high due to the difficulties in paying fees.
2. If the age of entry of the student is below twenty years, parent is self-employed or unemployed and has poor health, then the student is likely to drop out.
3. If a student has poor health, age of entry of the student is below twenty years and parent employed, then the student is likely to drop out if is a female.

Recommendations

By gaining a deep understanding of student retention patterns and tendencies, we are able 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 makers can utilize necessary action and directly or indirectly intervene by providing extra academic counselling, and financial aid. Therefore, the Higher Learning institution management system is able 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.

References

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

Carey, K. (2004). A matter of degrees. Improving graduation rates in four-year colleges and universities. New York: The Education Trust

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.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. ACM SIGKDD explorations newsletter. 2009

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 Systemic.

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

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

Watson, H. J., & Wixom B. H., (2007). The current state of business intelligence, *Computer*, 40(9) 96-99, September 2007. doi:10.1109/MC.2007.331