International Journal of Computer Applications Technology and Research
Volume 7–Issue 03, 122-128, 2018, ISSN:-2319–8656

Improving Student Enrollment Prediction Using Ensemble Classifiers

Stephen Kahara Wanjau
Directorate of ICT
Murang’a University of Technology
Murang’a, Kenya

Geoffrey Muchiri Muketha
School of Computing and IT
Murang’a University of Technology
Murang’a, Kenya

Abstract: In the recent years, data mining has been utilized in education settings for extracting and manipulating data, and for establishing patterns in order to produce useful information for decision making. There is a growing need for higher education institutions to be more informed and knowledgeable about their students, and for them to understand some of the reasons behind students’ choice to enroll and pursue careers. One of the ways in which this can be done is for such institutions to obtain information and knowledge about their students by mining, processing and analyzing the data they accumulate about them. In this paper, we propose a general framework for mining student data enrolled in Science, Technology, Engineering and Mathematics (STEM) using performance weighted ensemble classifiers. We train an ensemble of classification models from enrollment data streams to improve the quality of student data by eliminating noisy instances, and hence improving predictive accuracy. We empirically compare our technique with single model based techniques and show that using ensemble models not only gives better predictive accuracies on student enrollment in STEM, but also provides better rules for understanding the factors that influence student enrollment in STEM disciplines.

Keywords: Ensemble classification, STEM, predictive modeling, machine learning, WEKA.

1. INTRODUCTION

Strengthening the scientific workforce has been and continues to be of importance for every country in the world. Preparing an educated workforce to enter Science, Technology, Engineering and Mathematics (STEM) careers is important for scientific innovations and technological advancements, as well as economic development and competitiveness [1]. In addition to expanding the nation’s workforce capacity in STEM, broadening participation and success in STEM is also imperative for women given their historical underrepresentation and the occupational opportunities associated with these fields.

Higher Education Institutions (HEIs) in Kenya offer a variety of academic programs with admission of new student held every year. Student applications are selected based exclusively on one criterion, their performance in the Secondary School Final Examination (KCSE), an academic exam that largely evaluates four components: Mathematics, Sciences, Social sciences, and Languages. Every academic program has a previously defined number of places that are occupied by the students with higher marks, ensuring a high academic quality of the students. As HEIs increasingly compete to attract and retain students in their institutions, they can take advantage of data mining, particularly in predicting enrollment. These institutions can collect data about students from the admission process including the test scores results, the decision for enrollment, and some socio-demographic attributes. This data can be used to predict future student enrollment using data mining techniques.

Machine learning has in the recent years found larger and wider applications in Higher Education Institutions and is showing am increasing trend in scientific research, an area of inquiry, termed as Educational Data Mining (EDM) [1]. EDM aims towards discovering useful information from large amounts of electronic data collected by educational systems. EDM typically consists of research to take educational data and apply data mining techniques such as prediction (including classification), discovery of latent structure (such as clustering and q-matrix discovery), relationship mining (such as association rule mining and sequential pattern mining), and discovery with models to understand learning and learner individual differences and choices better [2], [3].

Researchers in educational data mining have used many data mining techniques such as Decision Trees, Support Vector Machines, Neural Networks, Naïve Bayes, K-Nearest neighbor, among others to discover many kinds of knowledge such as association rules, classifications and clustering [4]. The discovered knowledge has been used for prediction regarding enrolment of students in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students, prediction about students’ performance among others [5].

Prediction modeling lies at the core of many EDM applications whose success depends critically on the quality of the classifier [6]. There has been substantial research in developing sophisticated prediction models and algorithms with the goal of improving classification accuracy, and currently there is a rich body of such classifiers. However, although the topic of explanation and prediction of enrollment is widely researched, prediction of student enrollment in higher education institutions is still the most topical debate in higher learning institutions. These institutions would like to know, for example which student will enroll in which particular course, and which students will need assistance in order to graduate [7]. One approach to effectively address these student challenges is through the analysis and presentation of data or data mining.

Predicting student enrollment in higher education institutions is a complex decision making process that is more than merely relying on test scores. Previous research indicate that student enrollment, particularly STEM courses depends on diverse factors such as personal, socio-economic, family and other environmental variables [8], [9]. The scope of this paper is to predict enrollment in STEM disciplines and to determine the factors that influence the enrollment of students, using data mining techniques.

Ensemble classification has received much attention in the machine learning community and has demonstrated promising capabilities in improving classification accuracy. Ensemble methods combine multiple models into one usually more accurate than the best of its components. In this paper, we suggest an ensemble classifier framework for assessing and predicting student enrollment in STEM courses in Higher Education Institutions. The study focuses on improving the quality of student enrollment training data by identifying and eliminating mislabeled instances by using multiple classification algorithms.

The rest of the paper is organized as follows: Section II describes the related works including ensemble methods in machine learning and related empirical studies on educational data mining using ensemble methods. Section III describes the methodology used in this study and the experiment conducted. Section IV presents results and discussion. Finally, section V presents the conclusions of the study.

2. ENSEMBLE CLASSIFICATION

Ensemble modeling has been the most influential development in Data Mining and Machine Learning in the past decade. The approach includes combining multiple analytical models and then synthesizing the results into one usually more accurate than the best of its components [9]. An ensemble of classifiers blends predictions from multiple models with two goals: The first goal is to boost the overall prediction accuracy compared to a single classifier and the second one is to achieve a better generalizability owing to different specialized classifiers. Consequently, an ensemble can find solutions where a single prediction model would have difficulties. The main underlying principle is that an ensemble can select a set of hypotheses out of a much larger hypothesis space and combine their predictions into one [10]. The philosophy of the ensemble classifier is that another base classifier compensates the errors made by one base classifier. The following sub sections details different base classifiers and the ensemble classifiers.

2.1 Base Classifiers

Rahman and Tasnim [11] describe base classifiers as individual classifiers used to construct the ensemble classifiers. The following are the common base classifiers: (1) Decision Tree Induction – Classification via a divide and conquer approach that creates structured nodes and leafs from the dataset. (2) Logistics Regression – Classification via extension of the idea of linear regression to situations where outcome variables are categorical. (3) Nearest Neighbor – Classification of objects via a majority vote of its neighbors, with the object being assigned to the class most common. (4) Neural Networks – Classification by use of artificial neural networks. (5) Naïve Bayes Methods – Probabilistic methods of classification based on Bayes Theorem, and (6) Support Vector Machines – Use of hyper-planes to separate different instances into their respective classes.

2.2 Ensemble Classifiers

Many methods for constructing ensembles have been developed. Rahman and Verma [12] argued that ensemble classifier generation methods can be broadly classified into six groups that that are based on (i) manipulation of the training parameters, (ii) manipulation of the error function, (iii) manipulation of the feature space, (iv) manipulation of the output labels, (v) clustering, and (vi) manipulation of the training patterns.

2.2.1 Manipulation of the Training Parameters

The first method for constructing ensembles manipulates the training data set to generate multiple hypotheses. The learning algorithm is run several times, each time with a different subset of the training data set [13]. This technique works especially well for unstable learning algorithms whose output classifier undergoes major changes in response to small changes in the training data: Decision tree, neural network, and rule learning algorithms are all unstable, linear regression, nearest neighbor, and linear threshold algorithms are generally very stable. Different network weights are used to train the base neural network learning process [11]. These methods achieve better generalization.

2.2.2 Manipulation of the Error Function

The second method for constructing ensembles is by augmenting the error function of the base classifiers. In this case, an error is imposed if base classifiers make identical errors on similar patterns [11]. An example of such an ensemble is the Negative correlation learning. The idea behind negative correlation learning is to encourage different individual networks in an ensemble to learn different parts or aspects of a training data so that the ensemble can learn the whole training data better [14].

2.2.3 Manipulation of the Feature Space

The third general technique for generating multiple classifiers is to manipulate the set of input features (feature subsets) available to the learning algorithm. According to Dietterich [13] this technique only works when the input features are highly redundant.

2.2.4 Manipulation of the Output Labels

A fourth general technique for constructing an ensemble of classifiers is to manipulate the output targets. Each base classifier is generated by switching the class labels of a fraction of training patterns that are selected at random from the original training data set [12]. Each member of each class receives a vote and the class with the most votes is the prediction of the ensemble.

2.2.5 Ensemble Classifier Generation by Clustering

Another method of generating ensemble classifiers is by partitioning the training data set into non-overlapping clusters and training base classifiers on them [12] and the patterns that tend to stay close in Euclidean space naturally are identified by this process [13]. A pattern can belong to one cluster only therefore; a selection approach is followed for obtaining the ensemble class decision. These methods aim to reduce the learning complexity of large data sets

2.2.6 Manipulation of the Training Patterns

The last method for constructing ensembles is by manipulating the training patterns whereby the base classifiers are trained on different subsets of the training patterns [12]. The largest set of ensembles are built with different learning parameters, such as number of neighbors in a k Nearest Neighbor rule, and initial weights in a Multi Layer Perceptron.

3. RELATED EMPIRICAL STUDIES

Stapel, Zheng, and Pinkwart [15] study investigated an approach that decomposes the math content structure underlying an online math learning platform, trains specialized classifiers on the resulting activity scopes and uses those classifiers in an ensemble to predict student performance on learning objectives. The study results suggested that the approach yields a robust performance prediction setup that can correctly classify 73.5% of the students in the dataset. This was an improvement over every other classification approach that they tested in their study. Further examinations revealed that the ensemble also outperforms the best single-scope classifier in an early prediction or early warning setting.

In their study, Satyanarayana and Nuckowski [16] used multiple classifiers (Decision Trees-J48, Naïve Bayes and Random Forest) to improve the quality of student data by eliminating noisy instances, and hence improving predictive accuracy. The results showed that student data when filtered can show a huge improvement in predictive accuracy. The study also compared single filters with ensemble filters and showed that using ensemble filters works better for identifying and eliminating noisy instances.

Pardos, Gowda, Baker, and Heffernan [17] study investigated the effectiveness of ensemble methods to improve prediction of post-test scores for students using a Cognitive Tutor for Genetics. Nine algorithms for predicting latent student knowledge in the post-test were used. The study found that ensembling at the level of the post-test rather than at the level of performance within the tutor software resulted to poor prediction of the post-test, based on past successes of combined algorithms at predicting the post-test. The study gave a few possible reasons for this. First of all, the data set used in this study was relatively small, with only 76 students. Ensembling methods can be expected to be more effective for larger data sets, as more complex models can only achieve optimal performance for large data sets. This is a general problem for analyses of post-test prediction.

In their study, Shradha and Gayathri [18] used educational data mining to analyze why the post-graduate students' performance was going down and overcome the problem of low grades at AIMIT College, Mangalore, India for the academic year 2014-2015. In their study, they compared base classifiers with an ensemble model. The study used J48, Decision Table and Naïve Bayes as base classifers and bagging ensemble model. The study concluded that J48 algorithm was doing better than the Naïve Bayesian. Also, bagging ensemble technique provided accuracy which was comparable to J48. Hence, this approach could aid the institution to find out means to enhance their students' performance.

4. METHODOLOGY

4.1 Study Design

This study adapted the Cross Industry Standard Process for Data Mining (CRISP-DM) process model suggested by Nisbet, Elder and Miner [19] as a guiding framework. The framework breaks down a data mining project in phases which allow the building and implementation of a data mining model to be used in a real environment, helping to support business decisions. Figure I give an overview of the key stages in the adapted methodology.

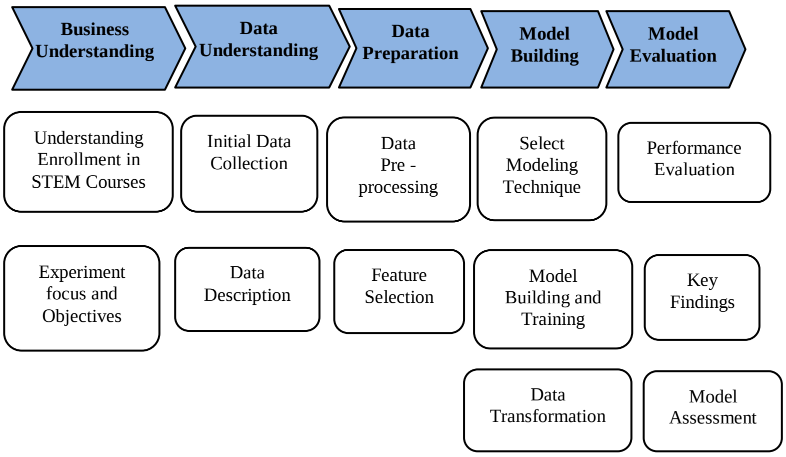

Figure I: Adapted Methodology for Research Implementation

4.1.1 Business Understanding

This phase begins with the setting up of goals for the data mining project. The goal of this stage of the process is to uncover important factors that could influence the outcome of the project [19]. Some of the activities in this stage include identifying the target variable, listing all the important predictor variables, acquiring the suitable institutional dataset for analyses and modeling, and generating descriptive statistics for some variables.

4.1.2 Data Understanding

Data understanding phase starts with data collection and getting used to the data to identify potential patterns in the data. This stage involves activities including data acquirement, data integration, initial data description, and data quality assessment activities. Data has to be acquired before it can be used. The data set used in this study was collected through the questionnaire survey at Murang'a University of Technology, a Public University in Kenya.

4.1.3 Data preparation

Data preparation is the phase of the data mining project that covers all activities needed to construct the final dataset. Initially the dataset was collected in Ms Excel sheet and preprocessing done. Feature selection was used as a method to select relevant attributes (or features) from the full set of attributes as a measure of dimensionality reduction. Two statistical methods were adopted to determine the importance of each independent variable. These methods include Chi-Square Attribute evaluation and Information Gain Attribute evaluation.

4.1.4 Modeling

This phase in data mining project involves building and selecting models. The usual practice is to create a series of models using different statistical algorithms or data mining techniques. The open source software WEKA, offering a wide range of machine learning algorithms for Data Mining tasks, was used as a data mining tool for the research implementation. The selected attributes were transformed into a form acceptable to WEKA.

4.1.5 Evaluation

This stage involves considering various models and choosing the best one based on their predictive performance. The resultant models, namely J48, Naïve Bayes, and CART were evaluated alongside bagging. Classification accuracy of the models was calculated based on the percentage of total prediction that was correct.

4.2 Experiment

4.2.1 Data Collection

Data was collected from sampled students through a personally administered structured questionnaire at Murang'a University of Technology, Kenya for the academic year 2016-2017. The target population was grouped into two mutually exclusive groups namely; STEM (Science, Technology, Engineering and Mathematics) and non-STEM Majors. Aside from the demographic data, data about their interests and motivations to enroll in the courses of their choice, academic qualification and educational contexts was collected. Table I shows the identified attributes and possible values that were taken as an input for our analysis.

Table I: Factors affecting Students Enrollment in STEM

S/No	Attribute	Possible Values
1	Career Flexibility	{Yes, No}
2	High School Final Grade	{A,A-,B+,B,B-,C+}
3	Math Grade	{A,A-,B+,B,B-,C+}
4	Pre - University awareness	{Yes, No}
5	Teacher Inspiration	{Yes, No}
6	Financial Aid	{Yes, No}
7	Extracurricular	{Yes, No}
8	Societal Expectation	{Yes, No}
9	Parent Career	{STEM , Non-STEM}
10	Self Efficacy	{Yes, No}
11	Career Earning	{Yes, No}
12	Gender	{Male, Female}
13	Age	Below 20 Years 20 – 25Years 26 – 30 Years 31 and above

14	Family Income	Less than 10,000;
		10,001 – 20,000;
		20,001 – 30,000;
		30,001 – 40,000;
		40,001 – 50,000;
		50,001 and above

4.2.2 Data Transformation

The collected data attributes were transformed into numerical values, where we assigned different numerical values to each of the attribute values. This data was then transformed into forms acceptable to WEKA data mining software. The data file was saved in Comma Separated Value (CSV) file format in Microsoft excel and later was converted to Attribute Relation File Format (ARFF) file inside WEKA software for easy use.

4.2.3 Data Modeling

To find the main reasons that affects the students' choice to enroll in STEM courses the study used three base classification algorithms together with an ensemble model method, so that we can find accurate or exact factors affecting students' enrollment in STEM. Using algorithms in ensemble model, we will find the actual factors that effects students' choice to enroll in STEM. The following are the methods that we were using for classification:-

4.2.3.1 J48 Algorithm

J48 is a decision tree algorithm and an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. In order to classify a new item, the algorithm first needs to create a decision tree based on the attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly.

4.2.3.2 Naïve Bayes Algorithm

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given dataset [20]. The Naive Bayesian classifier is based on the Bayes' theorem with independence assumptions between predictors.

4.2.3.3 CART

Classification and Regression Tree (CART) is one of the commonly used Decision Tree algorithms. It is a recursive algorithm, which partitions the training dataset by doing binary splits. At each level of the decision tree, the algorithm identify a condition - which variable and level to be used for splitting input node (data sample) into two child nodes.

4.2.3.4 Bagging

Bagging is the technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model. Bagging algorithm uses bootstrap samples to build the base predictors. Each bootstrap sample of *m* instances is formed by uniformly sampling m instances from the training dataset with replacement.

5. RESULTS AND DISCUSSION

We collected students' information by distributing structured questionnaire among 220 students and 209 responses were collected. This data was preprocessed and recorded into Microsoft Excel file and then through online conversion tool, the Excel file was converted into .arff file which is supported by the WEKA software tool [21]. We used Weka 3.6 software for our analysis. Table II shows the results obtained from the experiment.

Table II: Comparison of Algorithms

S/No	Algorithm	Correctly Classified instances (%)	Incorrectly Classified instances (%)
1	J48	84	16
2	CART	77	23
3	Naïve Bayes	72	28
4	Bagging	82	18

The information on Table II shows comparison details of the algorithms that were used in our analysis. When we compared the models, we found that the J48 Algorithm correctly classified 84% of the instances and 16% of the instances incorrectly classified. The classification error is less compared to the other two baseline classification algorithm, that is, CART (23% Incorrectly Classified Instances) and Naïve Bayes (28% Incorrectly Classified Instances). From these results we can conclude that among the three base classification algorithms that we used J48 algorithm was best suited for predicting enrollment of students in STEM courses.

We observed in the experiments with the baseline classifiers, that their classification accuracy can vary a lot based on random sampling of the training and test data. One of the reasons for this instability is because the base classifiers are highly susceptible to noisy training data and have a tendency to overfit.

To reduce chances of over-fitting, the most popular and simple techniques is called ensemble learning where multiple models are trained and their results are combined together in some way. One of the most popular methods is called bagging. In bagging, samples of the training data are bootstrapped. In other words, the samples are selected with replacement from the original training set.

The models are trained on each sample. Bagging makes each training set different with an emphasis on different training instances. In this study, bagging ensemble model was developed that gave 82% of Correctly Classified Instances.

Table III shows the attributes and the values obtained by applying the Karl Pearson Co-efficient Technique.

Table III: Values obtained by Karl Pearson Co-efficient Technique

S/No	Attribute	Coefficient of Determination (R^2) Value
1	High School Final Grade	0.981
2	Career Flexibility	0.842
3	Math Grade	0.763
4	Self Efficacy	0.714
5	Teacher Inspiration	0.692

The results from Table III show the five most significant attributes that highly affects the students choice to enroll in STEM courses in the University. These are the attributes that we can consider as factors which the institutions must focus on while considering enrollment of students in STEM related courses.

6. CONCLUSIONS

There are many factors that may affect students' choice to enroll and pursue a career in STEM in higher education institutions. These factors can be used during the admission process to ensure that students are admitted in the courses that best fit them. To categorize the students' based on the association between choice to enroll in a STEM major and attributes, a good classification is needed. In addition, rather than depending on the outcome of a single technique, ensemble model could do better. In our analysis, we found that J48 algorithm is doing better than Naïve Bayesian and the CART algorithms.

Also, the study results demonstrated that bagging technique provides accuracy which is comparable to J48. Moreover, the correlation between the attributes and the choice to enroll in STEM courses was computed and found that five significant attributes were highly affecting the students' choice to enroll in STEM courses. These attributes include the score obtained from the high school final exam, student score in Mathematics subject, expected career flexibility, belief in the ability to succeed in a STEM related career, and the inspiration from the high school teacher. Therefore, this approach could help institutions of higher learning to find out means to enhance student enrollment in STEM disciplines.]

In future work, the effects of using different base classifiers alongside other ensemble algorithms on classification accuracy and execution time as parameters can be investigated.

7. REFERENCES

[1] Lichtenberger, E. and George-Jackson, C. “Predicting High School Students’ Interest in Majoring in a STEM Field: Insight into High School Students’ Postsecondary Plans,” *Journal of Career and Technical Education, 28*(1), 19-38, 2013.

[2] Kulkarni, S., Rampure, G., and Yadav, B. “Understanding Educational Data Mining (EDM),” *International Journal of Electronics and Computer Science Engineering*, 2(2), 773-777, 2013.

[3] Baker, R. and Yacef, K. “The State of Educational Data mining in 2009: A Review and Future Visions, “ *Journal of Educational Data Mining, 1*(1), 3-17, October, 2009.

[4] Romero, C. and Ventura, S. “Educational Data Mining: A Review of the State of the Art,” *Systems, Man, and Cybernetics,Part C: Applications and Reviews, IEEE Transactions, 40*(6), 601-618, 2010.

[5] Sarala, V. and Krishnaiah, J. “Empirical Study of Data Mining Techniques in Education System,” *International Journal of Advances in Computer Science and Technology (IJACST)*, 15-21, 2015.

[6] Baradwaj, B. and Pal, S. “Mining Educational Data to Analyze Students’ Performance,” *International Journal of Advanced Computer Science and Applications, 2*(6), 63-69, 2011.

[7] Namdeo,V., Singh, A., Singh, D. and Jain, R. “RESULT ANALYSIS USING CLASSIFICATION,” *International Journal of Computer Applications, 1*(22), 22-26, 2010.

[8] Nandeshwar, A. andChaudhari, S. *Enrollment Prediction Models Using Data Mining.* [Unpublished], April 22, 2009.

[9] Wang, X. “Modeling Entrance into STEM Fields of Study among Students Beginning at Beginning at Community Colleges and Four-Year Institutions,” *Research in Higher Education*, 54 (6), 664-669, September, 2013.

[10] Rokach, L. “Ensemble-based classifiers,” *Artificial Intelligence Review*, 33, 1-39, 2010.

[11] Rahman, A. and Tasnim, S. “Ensemble classifiers and their applications: A review,” *International Journal of Computer Trends and Technology*, 10(1), 31-35, 2014.

[12] Rahman, A. and Verma, B. “Ensemble Classifier Generation using Non–uniform Layered Clustering and Genetic Algorithm,” *Elsevier Knowledge Based Systems*, 43, 30-42, May, 2013.

[13] Dietterich, G. T. (n.d.), *Ensemble Methods in Machine Learning.* Retrieved November 2016, from web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

[14] Liua, Y. and Yao, X. “Ensemble learning via negative correlation,” *Neural Networks, 12*, 1399-1404, 1999.

[15] Stapel, M., Zheng, Z. and Pinkwart, N. “An Ensemble Method to Predict Student Performance in an Online Math Learning Environment,” *Proceedings of the 9th International Conference on Educational Data Mining*, pp. 231-238, 2016.

[16] Satyanarayana, A. and Nuckowski, M. “Data Mining using Ensemble Classifiers for Improved Prediction of Student Academic Performance,” *ASEE Mid-Atlantic Section Spring 2016 Conference.* Washington D.C: George Washington University, April 8-9, 2016.

[17] Pardos, Z., Gowda, S., Baker, R., and Heffernan, N. “Ensembling Predictions of Student Post-Test Scores for an Intelligent Tutoring System,” *Educational Data Mining*, 2011.

[18] Shradha, S. and Gayathri, “Approach for Predicting Student Performance Using Ensemble Model Method,” *International Journal of Innovative Research in Computer and Communication Engineering, 2*(Special Issue 5), 161-169, October, 2014.

[19] Nisbet, R., Elder, J., and Miner, G. *Handbook of statistical analysis and data mining applications.* Amsterdam: Elsevier, 2009.

[20] Sage, S. and Langley, P. “Induction of Selective Bayesian Clasifiers,” *ARXIV*, pp. 399- 406, 2013.

[21] School, W. (2015). *Introduction to Weka - A Toolkit for Machine Learning.* Retrieved April 22, 2015, from http://www.iasri.res.in: http://www.iasri.res.in/ebook/win_school_aa/notes/WEKA.pdf

Stephen Kahara Wanjau currently serves as the Director of ICT at Murang'a University of Technology, Kenya. He received his BSc. degree in Information Sciences from Moi University, Kenya in 2006 and a Master of Science degree in Organizational Development from the United States International University – Africa in 2010. He is a master student in the Department of Computing, School of Computing and IT at Jomo Kenyatta University of Agriculture and Technology, Kenya. His research interests are machine learning, artificial intelligence, Knowledge management, and cloud computing.

Geoffrey Muchiri Muketha is Associate Professor of Computer Science & Dean of School of Computing and Information Technology at Murang'a University of Technology. He received his BSc. degree in Information Sciences from Moi University, Kenya, his MSc. degree in Computer Science from Periyar University, India, and his PhD degree in Software Engineering from Universiti Putra Malaysia. His research interests are software metrics, software quality and intelligent systems.