DATA MINING TECHNIQUE AS A TOOL FOR INSTRUCTORS' PERFORMANCE EVALUATION IN HIGHER EDUCATIONAL INSTITUTION

ASANBE M.O
Department of Computer Science
Villanova Polytechnic Imesi-Ile
Osun State, Nigeria
moasanbe@gmail.com

&

OLAGUNJU M.P
Department of Computer Science
Villanova Polytechnic Imesi-Ile
Osun State, Nigeria
moasanbe@gmail.com

A0BSTRACT

Educational Data Mining (EDM) is an evolving field exploring pedagogical data by applying different machine learning techniques/tools. It can be considered as interdisciplinary research field which provides intrinsic knowledge of teaching and learning process for effective education. The main objective of any educational institution is to provide quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge that predicts teachers' performance. This study presents an efficient system model for evaluation and prediction of teachers' performance in higher institutions of learning using data mining technologies. To achieve the objectives of this work, a two-layered classifier system was designed; it consists of an Artificial Neural Network (ANN) and Decision Tree. The classifier system was tested successfully using case study data from a Nigerian University in the South West of Nigeria. The data consists of academic qualifications for teachers as well as their experiences and grades of students in courses they taught among others. The attribute selected were evaluated using two feature selection methods in order to get a subset of the attributes that would make for a compact and accurate predictive model. The WEKA machine learning tool was used for the mining. The results show that, among the six attributes used, Working Experience, and Rank are rated the best two attributes that contributed mostly to the performance of teachers in this study. Also, considering the time taken to build the models and performance accuracy level, C4.5 decision tree outperformed the other two algorithms (ID3 and MLP) with good performance of 83.5% accuracy level and acceptable kappa statistics of 0.743. It does mean that C4.5 decision tree is best algorithm suitable for predicting teachers' performance in relation to the other two algorithms in this work.

Keywords: Educational Data Mining, Decision Tree, Artificial Neural Networks, Machine learning, WEKA

INTRODUCTION

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories [1].Data mining refers to extracting knowledge from large amount of data. The term data mining is variously named as "Knowledge mining from data", "knowledge mining" or "Knowledge Discovery from Database". However, Data Mining refers to a particular step in the Knowledge discovery process. It consists of particular algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (models) over the data [2]. Xingquan and Ian [3] defined data mining as the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. Data mining techniques have been in use to solve educational problems and to perform crucial analysis in the educational sector. This is to enhance educational standards and management such as investigating the areas of learning recommendation systems, learning material arrangement, continuous student assessments and evaluation of educational websites.

According to Romero *et al.*, [4], there are increasing research interests using data mining in education. This new emerging field, called Educational Data Mining is concerned with developing methods that discover knowledge from databases originating from educational environments. The data can be personal or academic which can be used to understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems, to improve curriculums and many other benefits. The desire is to extract hidden but useful knowledge from data through data mining tools. The recent decline in the standard of education in most developing countries has necessitated researches that will help proffer solutions to some of the problems. Performance evaluation has been described as a systematic process of evaluating an individual worker's job performance and effectiveness in relation to certain pre-established criteria and organizational objectives [5].

It has been observed that the continuous decline in average academic performance of students at higher educational institution can be traced back to the performance of their teachers. Most organizations use performance appraisal system to evaluate the teacher's performance while some institutions use student rating technique to measure teaching effectiveness of instructors. Efficiencies of these methods have been doubted over the years.

This study proposed to evaluate teacher's performance on the basis of different factors, using data mining techniques. Data mining is a powerful technology for analyzing important information from historical data, thus information needed for this study exist already within the system. This study presents an efficient system model with an algorithm for evaluation of teachers' performance in higher institutions of learning essentially to overcome the limitations of conventional

approaches. The focus of this work is on designing a framework based on academic background factors and socio-economic factors among other factors which can be used predict teacher's performance and recommend necessary actions for improvement.

REVIEW OF RELATED WORKS

Extensive literature reviews of the EDM research field are provided by Romero and Ventura [6], covering the research efforts in the area between 1995 and 2005. In their paper, they surveyed the application of data mining to traditional educational systems, particular web-based courses, well-known learning content management systems, and adaptive and intelligent web-based educational systems. Another extensive review was carried out by Baker and Yacef [7] covering the period between 2005 and 2009. They reviewed the history and current trends in the field of Educational Data Mining (EDM), considered the methodological profile of research in the early years of EDM, compared to 2008 and 2009, and discussed trends and shifts in the research conducted by this community. Another more recent survey was done by Rajni and Malaya [8], which focused on components, research trends (1998 to 2012) of EDM highlighting its related Tools, Techniques and educational Outcomes. They also highlighted the Challenges in EDM. A study by Varun and Arupama [9] examined the application of data mining techniques in higher educational institution to extract useful information from huge data sets and provided analytical tool to view and use this information for decision making processes by taking real life examples.

Chin-Chia Hsu and Tao Huang [10] conducted a study on the use of data mining technology to evaluate student's academic achievement via multiple channels of enrolment like joint recruitment enrolment, athletic enrolment and application enrolment. A similar study was carried out by Osofisan and Olamiti [11] where they investigated the academic background in relationship with the performance of students in a computer science programme in a Nigerian university. Their study showed that the grade obtained from senior secondary school examination (SSCE) in mathematics is the highest determinant of students' performance using the C4.5 learning algorithm in building the model of the student's performance.

Mardikyan and Badur [12] conducted a study to investigate the factors associated with the assessment of instructors teaching performance using two different data mining techniques: regression analysis and decision trees. For regression analysis the stepwise regression method was used and for decision trees CHAID and CART algorithms were applied.As a result of their study, they found that instructors, who have well prepared course outlines, use satisfactory materials, help the student outside the lectures, grade exams fairly and on time receive higher evaluations.

Hemaid and El-Halees [13]; Surjeet and Saurabh[14] used C4.5, ID3 and CART decision tree algorithms on engineering student's data to predict their performance in the final exam. Prediction models that included all personal, social, psychological and other environmental variables were necessary for the effective prediction of the performance of the students. C4.5 technique has highest accuracy of 67.7778% compared to other methods ID3 and CART algorithms.

A study by Pal and Pal [15] analyzed some numbers of parameters for the derivation of performance prediction indicators needed for teachers performance assessment, monitoring and evaluation. Four classification algorithms (Naïve Bayes, ID3, CART and LAD) based on Weka were used in their study. Their results showed that the best algorithm based on data is Naïve Bayes classification.

Surjeet *et al*, [16] carried out a research on mining educational data to predict student's retention. In the study machine learning algorithms (ID3, C4.5 and ADT) were applied to analyze and extract information from existing student data. Their study established predictive models and showed that machine learning algorithm such as Alternating Decision Tree (ADT) can learn predictive models from the student retention data accumulated from previous year.

Aranuwa and Sellapan [17] used directed modeling that is, an intelligent technique for evaluation of instructors' performance in higher institutions of learning, and proposed an optimal algorithm and designing a system framework suitable for predicting instructors' performance as well as recommended necessary action to be taken to aid school administrators in decision making considering the limitations of the classical methodologies. The proposed system, if fully implemented, will aid school administrators in decision making, provide basis for instructors' performance improvement that will optimize students' academic outcomes and improve standard of education. Consequently, this will contribute to successful achievement of the goals.

A similar study was carried out by Hemaid and El-Halees [13] to examine the factors associated with the assessment of teacher's performance. In this study, data was collected for teachers from the Ministry of Education and Higher Education in Gaza City. They proposed a model to evaluate their performance through the use of techniques of data mining like association, classification rules (Decision Tree, Rule Induction, K-NN, Naïve Bayesian (Kernel)) to determine ways that can help them to better serve the educational process and hopefully improve their performance and thus reflect it on the performance of teachers in the classroom. In each tasks, they presented the extracted knowledge and described its importance in teacher performance domain.

METHODOLOGY

The study aims at using the classification method of Data mining for the prediction of teachers' performance. The prediction model was developed using the Classification methods of the Data mining technique. The Neural Network data mining technique (the multilayer perceptron algorithm) and Decision Trees methods specifically the ID3 (Iterative Dichotomiser 3) and C4.5 algorithms (the C4.5 is implemented in WEKA by the classifier class: weka.classifiers.trees.J48) were used and their performances were compared to each other. The WEKA 3.6.13 Data mining software tool was also used to carry out the prediction processes.

Data Acquisition

The raw data used in this study was collected from an Academic Department of a University in Ondo State, South West Nigeria. The data included two basic categories of variables, the first group consists 350 records of the teachers' data variables as shown in table 1. The second group of variables includes the students' learning outcome (results) from 2010 to 2015 academics sessions as shown in table 2.

Table 1: Teachers' Data variable

VARIABLE NAME	VARIABLE FORMAT	VARIABLE TYPE
Teachers' ID	1,2,3…	Numerical
University	Name of university	Categorical
Name	Surname and other names	Categorical
Gender	Male, Female	Categorical
Appointment Status	Permanent, Temporal, Contract	Categorical
Employment Status	Old, New	Categorical
Rank	Prof, Reader, SL, L1, L2, AL, GA	Categorical
Present Age	20,30,…	Numerical
Working Experience	Year	Numerical
Highest Qualification	PhD, Master, Bachelor	Categorical
Year of last Qualification	1999, 2000,…	Numerical
Professional Qualification	Yes/No	Categorical

Preprocessing data and feature extraction

The data preprocessing was carefully done to avoid incomplete records. The fields selected for the model include: Appointment status, Rank, University working experience, Highest Qualification, Year of last Qualification, Professional Qualification and Result. The instrument and variable selected for this work were based on pedagogical surveys, facts from literatures and experts in the domain. The reliability and validity of the instruments used construe with the theories posited by the experts on survey design and guidelines for survey design according to Bradburn *et al* [18].

Table 2: Student's learning outcome

VARIABLE NAME	VARIABLE FORMAT	VARIABLE TYPE
Course ID	1,2,3…	Numerical
Course Name	Name of the course	Categorical
Course Lecturer	Name of Lecturer that took the course	Categorical
Course Performance	Satisfactory, Average, Poor	Categorically

System Design

The tasks involve machine learning and classification algorithms; hence in the research work a two-layered classifier system was designed to achieve the objective of the work as shown in figure 1. Layer 1 consists of an Artificial Neural Networks (ANN) and layer 2 is made up of Decision Trees classifiers. These classifiers have been selected because of their performances in various domains. They have both been successfully applied to a variety of real-world classification tasks in industry, business, science and education with good performances. The Neural Networks is known for its predictive accuracy, ability and aptitude to learn and remember [19]. Decision Tree classifiers are considered "white box" classification model as they can provide explanation for their models and can be used directly for decision making [20]. These abilities and aptitude are best suited and of good requirement for any effective and efficient intelligent system.

Analysis Proposed Model

The goal of the proposed system it to aid higher institution management in determining teachers' performance and recommend necessary actions to be taken on individual teacher based on the prediction from the intelligent evaluation system. The proposed system framework subsumes five components: The first and the second components take care of data acquisition and storage, responsible for storing teachers' data, gathered from different data sources proposed in a data warehouse.

The third component is model building, responsible for obtaining knowledge about the teachers, through appropriates classification models. Different classification algorithms are proposed in search for the best model with high predictive accuracy. The fourth component is for mapping pattern in the rules generated with the teacher data to predict performance and the fifth component is the recommendation, responsible for recommending necessary action to be carried out on individual teacher based on the prediction from the intelligent evaluation system as shown in figure 2.

Implementation

The data is analyzed and implemented in WEKA (Waikato Environment for Knowledge Analysis)Version 3.6.13.WEKA contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. The input data file is accepted in ARFF format. This file contains complete information regarding the set of all attributes and also the values for that attribute. There are 16 decision tree algorithms like ID3, J48, Simple CART etc. that are implemented in WEKA. An ARFF file consists of two distinct sections: the Header section defines attribute name, type and relations, start with a keyword @Relation <data-name> @attribute <attribute-name><type> or {range} and the Data section lists the data records, starts with @Data list of data instances. Any line start with % is the comments.

Dataset Training and Testing

The data set used for this study was stored in a Microsoft Excel spreadsheet named "teacherData.csv". For easy usage, the data was converted into ARFF format, that is, the file with (.arff extension) named as "teacherData.arff". The data was loaded into Weka and the attributes are recognized. The dataset includes 216 (61.89%) permanents, 72(20.34%) temporal and 61 (17.48%) contract staff ranging from professor to assistant lecturer. Out of the total 349 cases included in this study 200 (57.31%) were used as the training set while the remaining 149 (42.69%) were used as the test set.

RESULTS AND DISCUSSION

The proposed model was developed using WEKA. The model was built with three machine learning algorithms: C4.5 Decision Tree Classification Algorithm, ID3 Decision Tree Classification Algorithm and MLP Neural Network. A comparative analysis of the performance of the models was carried out.
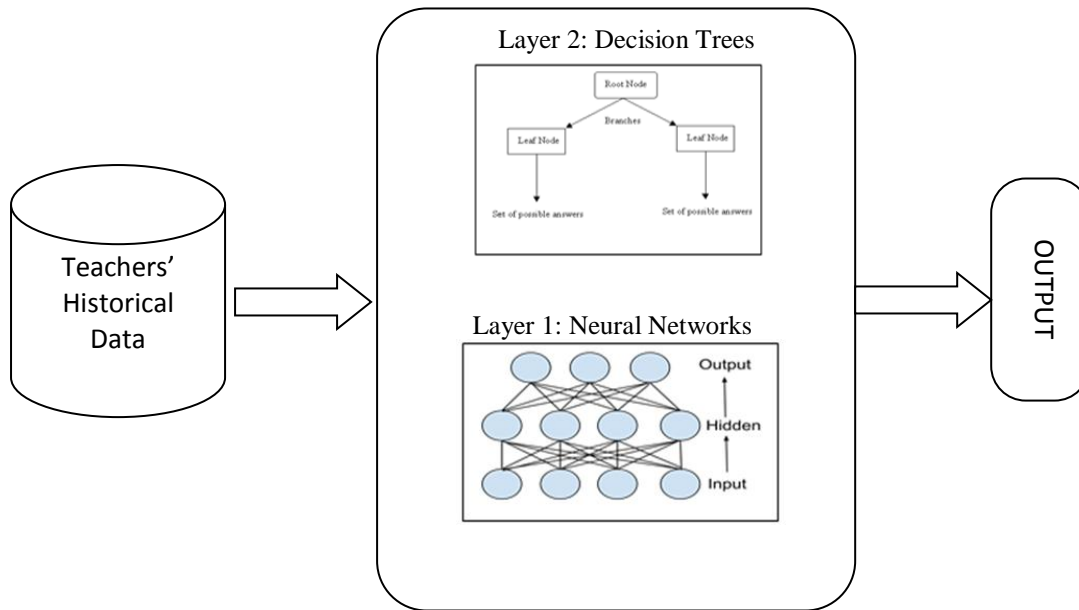
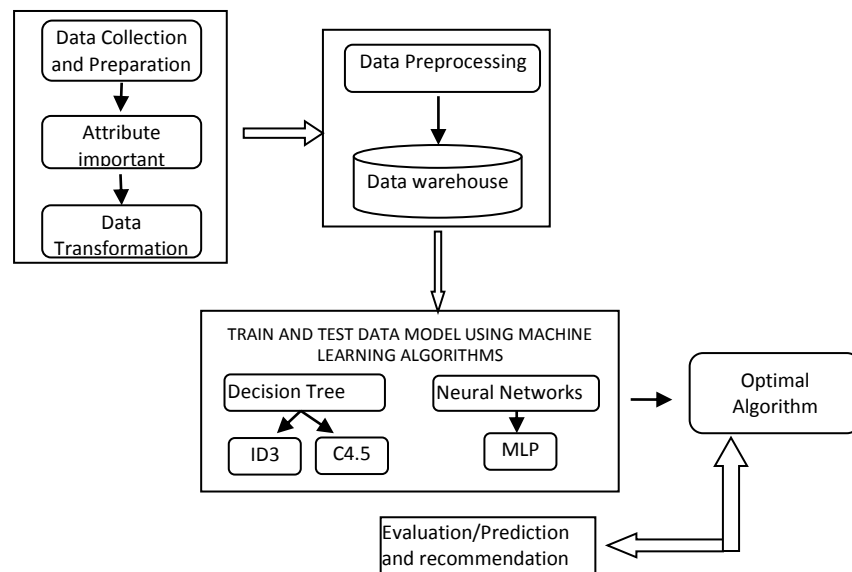**Fig 1: Two-layered classifier system**

Fig 2: Architecture of the proposed model

Attribute importance analysis was first carried out to rank the attributes used in this work by significance using wrapper method and filter method. The results of the two methods construe with each other are shown in table 3 and figure 3.

Table 3: Attributes Ranking using Information Gain Ratio

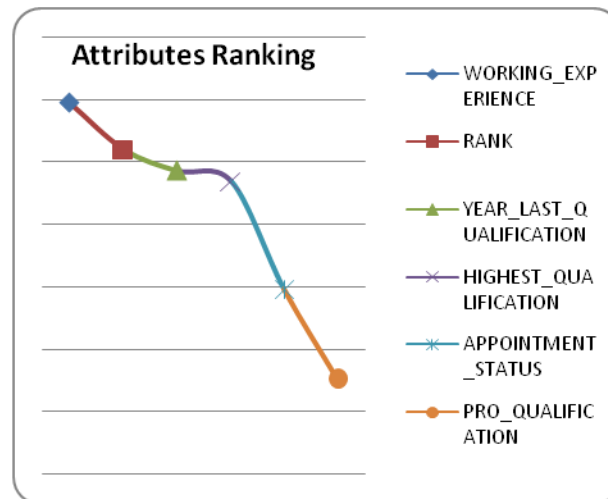| RANK ATTRIBUTES	VALUES	RANKS
WORKING_EXPERIENCE	0.298	1
RANK	0.2605	2
YEAR_LAST_QUALIFICATION	0.2435	3
HIGHEST_QUALIFICATION	0.2349	4
APPOINTMENT_STATUS	0.1483	5
PRO_QUALIFICATION	0.0763	6

Figure 3: Information gain ratio of the attributes

4.1   Model Comparison

The Weighted averages of the models were compared using different performance measures like:

- accuracy
- True Negative Rate
- True Positive Rate

- F-Measure and
- ROC.

The best model was then selected using Tables 4, Table 5 and Figure 4. The performances of these models were evaluated based on these criteria:

- Prediction accuracy
- Time taken to build the model and
- Error rate.

These are illustrated in table 5. C4.5 algorithm predicts better than the ID3 and MLP algorithms since its accuracy is the highest compared to others.

The results obtained from the analysis demonstrated a slight higher performance of model built with decision tree (C4.5 algorithm) over neural network (MLP algorithm). Both C4.5 and MLP algorithms results show great superiority over ID3 algorithm in terms of performance. C4.5 algorithm performed better than other algorithms not only in terms of the number of correctly classified instances also in terms of RMSE, MAE, RAE. Neural Network performed well in classification as well as in prediction but suffered from lack of speed. The ID3 Decision Tree was the fastest, but did not perform well at the classification. Also the rules generated makes C4.5 decision tree algorithm clearer and understandable.

Table 4: Performance Summary of the models.

Algorithms	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
ID3	0.71	0.171	0.711	0.71	0.708	0.87
C4.5	0.835	0.098	0.098	0.835	0.835	0.936
MLP	0.825	0.093	0.828	0.825	0.824	0.927

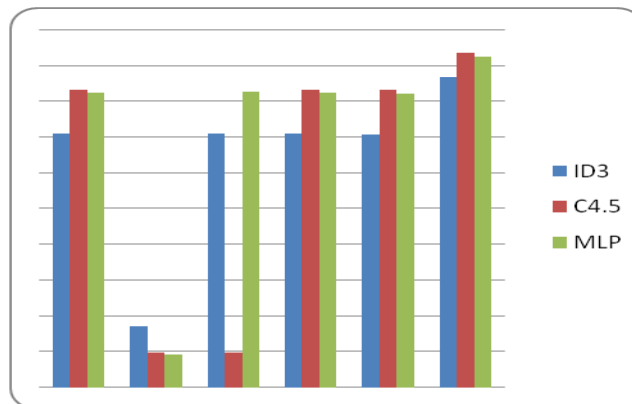**Figure 4: Comparison between performance measure parameters**

Table 5: Comparative analysis on the models

| METRIC	ID3	C4.5	MLP
Time take to build the model	0.02 sec	0.09 sec	6.24 sec
Correctly classified instances	71%	83.5%	82.5%
Incorrectly classified instances	29%	16.5%	17.5%
Kappa statistics	0.5461	0.743	0.7295
Mean Absolute Error	0.2513	0.167	0.1876
Root Mean Squared Error	0.3545	0.289	0.3075
Relative Absolute Error	58.8598%	39.1273%	43.9501%
Root Relative Squared Error	76.7465%	62.5708%	66.5732%
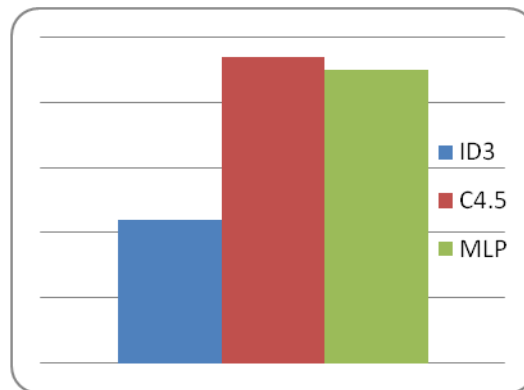
Figure 6: Prediction Accuracy

CONCLUSION

This study confirms that data mining techniques can be applied in the prediction of teachers' performance and the resulting models of this study are worthy of educational testing. To improve upon the classification accuracy of the models further researches should be conducted using different classification algorithms and other data mining techniques such as, Naïve Bayes classifier, genetic algorithm as well as data from other universities. Also, expanded data set with more distinctive attributes (such as subject mastery) to get more accurate results can also be used to carry out predictions to improve the classification accuracy.

ENDNOTES

[1] Ogunde A.O and Ajibade D.A (2014): A Data Mining System for Predicting University Students' Graduation Grades Using ID3 Decision Tree Algorithm. Journal of Computer Science and Information Technology. March 2014, Vol. 2, No 1, pp 21 – 46.

[2] Albashiri K. A. (2013). Data Partitioning and Association Rule Mining Using a Multi-Agent System. International Journal of Engineering Science and Innovative Technology (IJESIT), Volume 2, Issue 5, pp 161-169.

[3] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining: Challenges and Realities", ISBN 9781-59904-252, Hershey, New York, 2007.

[4] Romero C., Ventura S., Garcia E. (2008) Data mining in course management systems: Moodle case study and tutorial, Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008

[5] Abu-Doleh J, Weir D. Dimensions of Performance Appraisal Systems in Jordanian Private and Public Organizations. International Journal of Human Resource *Management*, 2007; 18(1): 75-84.

[6] C. Romero, S. Ventura (2007): "Educational data mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146,

[7] Baker, R. S. J. D., and K. Yacef. (2009): The State of Educational Data Mining in 2009: AReview and Future Visions.*Journal of Educational Data Mining* 1 (1): 3–17.

[8] Rajni J. and Malaya D.B (2013): A Survey on Educational Data Mining and Research Trends. International Journal of Database Management Systems (IJDMS) Vol.5, No.3

[9] Varun Kumar and AnupamaChadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011

[10] Chin Chia Hsu and Tao Huang (2006): The use of Data Mining Technology to Evaluate Student's Academic Achievement via multiple Channels of Enrolment. An empirical analysis of St. John's University of Technology.

[11] Osofisan A.O. and Olamiti A.O. (2009): Academic Background of Students and Performance in Computer Science Programme in a Nigerian University. *European Journal of Social Science. Vol. 33 Issues 4. 2009.*

[12] MardikyanS., and Badur B. (2011). Analyzing teaching Performance of Instructors Using Data Mining techniques. Informatics in Education, 2011, Vol. 10, No. 2, pp 245 – 257.

[13] Hemaid and El-Halees (2015): Improving Teacher Performance using DataMiningInternational Journal of Advanced Research in Computer and Communication EngineeringVol. 4, Issue 2, February 2015.

[14] Surjeet K.Y and Saurabh P (2012): Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56, 2012

[15] Pal K. and Pal S. (2013): Evaluation of Teacher's Performance: A Data Mining Approach. International Journal of Computer Science and Mobile Computing. IJCSMC, Vol 2, Issue. 12, Dec., 2013, pg. 359 – 369.

[16] Surjeet K.Y et al, (2012): Mining Educational Data to Predict Student's Retention: A comparative study, volume 10, No 2.

[17] Aranuwa F.O., and Sellapan P.,(2013): A data mining model for evaluation ofinstructors'performance in higher institutions of learning usingmachine learning algorithms, *International Journal of Conceptions on Computing and Information Technology Vol. 1, sue 2, Dec' 2013; ISSN: 2345 – 9808*

[18] Bradburn N, Sudman S, WansinkB.Asking Questions: The Definitive Guide to Questionnaire Design. Jossey-Bass, 2004

[19] Zurada JM. (2006): "Introduction to Artificial Neural System", *Jaico Publishing House*,.

[20] Romero C, Olmo JL, Ventura S. (2013): A meta-learning approach for recommending a subset of white-box classification algorithms for Moodle datasets. *Department of Computer Science, University of Cordoba, Spain*,