Application of Data Mining Classification in Employee Performance Prediction

John M. Kirimi
School of Computing and Informatics
University of Nairobi
P. O. Box 30197 – 00100
Nairobi, Kenya

Christopher A. Moturi
School of Computing and Informatics
University of Nairobi
P. O. Box 30197 – 00100
Nairobi, Kenya

ABSTRACT
In emerging knowledge economies such as Kenya, organizations rely heavily on their human capital to build value. Consequently, performance management at the individual employee level is essential and the business case for implementing a system to measure and improve employee performance is strong. Data Mining can be used for knowledge discovery of interest in Human Resources Management (HRM). We used the Data Mining classification technique for the extraction of knowledge significant for predicting employee performance using previous appraisal records a public management development institute in Kenya. The Cross Industry Standard Process for Data Mining (CRISP-DM) was adopted for predictive analysis. Decision tree was the main Data Mining tool used to build the classification model, where several classification rules were generated. To validate the developed model, a prototype was constructed and the data collected from the institute's Human Resource Department was used. Results show that employee performance was highly affected by experience, age, academic qualification, professional training, gender, marital status and previous performance appraisal scores. This paper proposes a prediction model for employee performance forecasting that enables the human resource professionals to refocus on human capability criteria and thereby enhance the performance appraisal process of its human capital.

Keywords
Employee Performance Prediction, Data Mining, Data Mining Classification, C4.5 (J4.8) Algorithm.

1. INTRODUCTION
The explosive growth of available data as a result of computerization of almost every aspect of the operations of organizations has instinctive contributions to the development of intelligent decision making technologies. A young yet promising of these kind technologies is Data Mining which is the process of analyzing data from different perspectives and summarizing it into useful information. Data mining techniques are aimed at discovering knowledge from the available data and could be used for improving the processes. A historical overview Data Mining and its future directions in terms of standard for a Knowledge Discovery and Data Mining process model is given in [11].

There are many classification techniques in Data Mining such as Decision Tree, Neural Network, Rough Set Theory, Bayesian theory and Fuzzy logic [13]. Decision tree is among the popular classification techniques, which can produce the interpretable rules or logic statement [7]. The generated rules from the selected technique can be used for future prediction.

Data Mining stands out due to its wide-array of techniques from the different domains such as statistics, artificial intelligence, machine learning, algorithms, database systems and visualization. These influences serve as groundwork for its applications to business for which the human resource management is unexceptionally classified. Data Mining has gained popularity due to its tools with potential to identify trends within data and turn them into knowledge mostly with predictive attributes that could significantly lead to better and strong bases for decision making [18], [11].

The Kenya School of Government (the School) is a Public Management Development Institute established under an Act of Parliament. The mandate of the School is to provide learning, consultancy and research services designed to inform public policy, promote national development and standards of competence and integrity in a result-oriented public service (http://www.ksg.ac.ke/). In this knowledge economy, the School relies heavily on the human capital to build value. Consequently, performance management at the individual employee level is essential and the business case for implementing a system to measure and improve employee performance at the School is strong. Business organizations are interested to settle plans for correctly selecting proper employees. After recruiting employees, the management becomes concerned about the performance of these employees where they build evaluation systems in an attempt to preserve the good performance of employees [3].

The School introduced performance monitoring and evaluation system to measure the performance of its human resource in a fair, objective and comprehensive manner in order to create a result- oriented institution. Employee performance evaluation is systematically carried out, according to a definite plan, typically by a corresponding manager or supervisor. Employee appraisal results are useful in compensation decisions, promotions, training and development programs, feedback, and personal development. Employee performance evaluation at the School therefore forms a basis for many HR decisions. To maintain consistency and disdain from partiality, this research sought to propose a classifier model for employee performance prediction. With Data Mining functionalities such as classification, this model can be used for the extraction of knowledge significant for predicting employee performance from previous appraisal records. The knowledge discovered could support human resource managers in deciding apt vital enhancement trainings for employees to efficiently respond to performance evaluations and expectation.

2. RELATED WORK
Several studies have used Data Mining for extracting rules and predicting certain behaviors in several areas of science, information technology, management, education, biology and

medicine. Data mining tools [18] are necessary in order to analyze vast amount of data generated by large organizations and drawing fruitful conclusions and inferences. An overview of the Data Mining systems and some of its applications in the different fields is given in [12].

The abundance of data has attracted Data Mining research towards the domain of Human Resource Management. The review by [16] shows that HRM constitutes a noteworthy new domain of Data Mining research that is dominated by method- and technology-oriented work. However, there is need for specific domain requirements, such as performance evaluation, or compliance with legal standards. [8] Reviewed HR applications and talent management, the use of Data Mining technique in HR and proposed the potential HR system architecture for talent forecasting.[14] Showed the ability of data mining in improving the quality of the decision-making process in HRM Systems by showing how to discover and extract useful patterns from large data sets in order to find observable patterns in HR.

[17]Used a Naive Bayes classifier to predict job performance in a call center with the aim of knowing what levels of the attributes are indicative of individuals who perform well. By using operational records, they predicted future performance of sales agents, achieving satisfactory results. [3]Developed a Data Mining framework based on decision tree and association rules to generate useful rules for personnel selection. This framework can be used to develop an effective personnel selection mechanism to find the talents who are the most suitable to their own organizations. Their results show that specific recruitment and human resource management strategies were created.

[2]Used rule-based classification Data Mining technique to extract knowledge significant for predicting training needs of newly-hired faculty members in order to devise the necessary development programs. They used the Cross Industry Standard Process for Data Mining (CRISP-DM) in discovering significant models needed for predictive analysis and demonstrated the required professional trainings to prepare faculty members to perform their tasks effectively.

[1] Applied Data Mining techniques to build a classification model for predicting employee's performance. Their model that was based on CRISP and use of decision tree as the main Data Mining tool, was validated by experiments with real data collected from several companies. [15] Built models that used classification algorithm like decision trees and Naïve Bayes to rank the applicants for a job profile based on their resume and social media presence. Theirs is a match making system where the companies will be given a list of ranked candidates using information retrieval technique like two way relevance ranking.

[6]Proposed a potential Data Mining technique for talent forecasting by identifying potential talent using past experience knowledge. [7] Show how the potential human talent can be predicted using a decision tree classifier with the pattern of talent performance identified through the classification process. They used decision tree C4.5 classification algorithm to generate the classification rules for human talent performance records.

[5]Suggested a model for talent management that can be used as a decision support tool and performs several different type of analysis. This model uses supervised as well as unsupervised techniques. The authors have further tested the accuracy of prediction using four different classification algorithms and have achieved significant accuracy in the classification results. [19]Also built similar model but with data warehousing perspective of the data. [4]Applied Data Mining in the context of higher educational system to discover useful knowledge.

Each of the work cited above has primarily used classification supervised learning approach [10] or clustering – unsupervised learning for constructing Data Mining models in HRM. While cited studies substantiated the applications of Data Mining in the HRM domain, none has applied Data Mining to predict the employees performance based on their inherent characteristics which could be initially mined from previous employee appraisal records. Instead, predictions of performance and talents have been stressed out. However, in harmony with these applications, this paper strives to build a model for predicting employee performance from previous appraisal records that is parallel to the criteria used for performance evaluation.

3. METHODOLOGY

3.1 Research Design

The development of this research followed the Cross Industry Standard Process for Data Mining (CRISP-DM) model. The CRISP-DM model was best suited for this research because it provided a generic guide to develop Data Mining project lifecycle. The employee performance data was collected from database of the Human Resource Department at the Kenya School of Government. A series of experiments were conducted to test the model using attributes extracted from the employee appraisal form. Based on the complex and multi-level structure of the data, the generic process of Knowledge Discovery in Databases (KDD) was reformed for effective results. The classification process was carried out with three different Data Mining algorithms, ID3, C4.5 and Naïve Bayes to identify the best and most suitable classification algorithm. The chosen algorithm was then improvised to obtain the best classification rate. The sequences of steps followed by this research are illustrated in the figure 1.

The predictive attributes were extracted from the employee appraisal form and were used as performance attributes. They include:

i. Age
ii. Gender
iii. Marital status
iv. Qualification

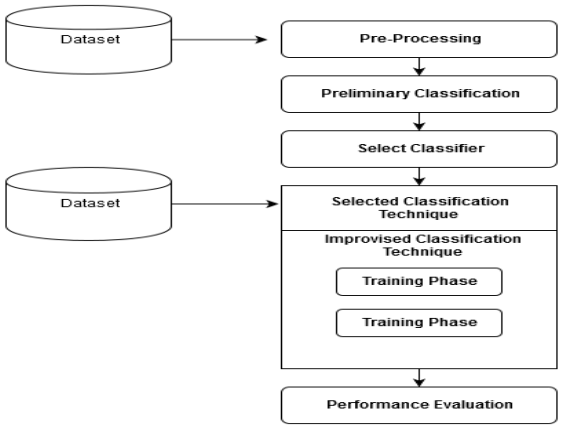

Figure 1: Steps of building the classifier

v. Specialization
vi. Professional training
vii. Job Group
viii. Experience
ix. Salary
x. Designation
xi. Performance Appraisal Score

The attributes were grouped into three categories,

i. **D**: Age, Gender and Marital status
ii. **E**: Qualification, Specialization, Professional Training
iii. **P**: JG, Experience, Salary, Designation, PAS

The attributes were input values and the Performance attribute serving as the target class. The target class attribute was discrete in nature with Outstanding, Exceed Expectation, Meet Expectation, Needs Improvement and Does not Meet Minimum Standards as the values

3.2 Data Understanding

The data for training the model was collected through the extraction and organization of the employee performance data from the Kenya School of Government database in a new flat file. The data collected was for five years and included 206 employee's performance appraisal records, described by 14 parameters. These parameters basically represented the skills set required to be possessed by an employee. This data was divided into two datasets. The first dataset of 110 records was used for training and testing, while the other dataset of 96 records was used for validating the model.

3.3 Data Cleaning

The raw data contained instances that were not applicable. This was due to errors and anomalies that had to be discarded. The data was transferred to Excel sheets. The types of data were then reviewed and modified. Data cleaning and filling-in of missing values in the data was performed before feature selection was applied to the dataset to identify the key attributes and obtain a reduced relevant subset of key attributes to be used in the classification exercise. These files were then prepared and converted to arff format that is compatible with the WEKA Data Mining toolkit which was used in building the model. The data was divided into three data sets. The first dataset was used for training the model, the second dataset was used for testing the model and the third data set was for validating the model.

3.4 Modeling and Experiments

First level classification considering all the parameters was performed. "Raw" classification without applying any dimensionality reduction technique was carried out on each of the below-mentioned datasets and also on the combination of the datasets to understand the significance of each individual dataset to the performance category of an employee. The following different combinations of preprocessed datasets were used for this classification step:

i. D, E, P, independently
ii. D &E
iii. D & P
iv. E & P
v. D & E & P

For these basic classifications, decision tree algorithm was the most suitable choice because of its interpretability. The tree structure generated as a result also provided important insights about significant attributes of the concerned dataset. A preliminary analysis was also carried out on dataset D & E & P with different classification techniques using Performance as target class and abovementioned continuous attributes as input attributes. The comparison of accuracy rate obtained for the various classification techniques is given in table 1 below

Table 1 : Classification Algorithms Accuracy Rate

NO:	Technique	Classification Accuracy
1	ID3	64.52%
2	Naïve Bayes	80.33%
3	C4.5 (J4.8)	92.60%

As seen from the table 1, C4.5 algorithm had the highest accuracy of 92.69% and was therefore best suited for the training and development of the classification model.

3.5 Classification Model

Important insights about significant attributes of the concerned dataset revealed that the dataset **P** was predominant and controlled the entire process of evaluation of an employee performance. The classifier thus did not consider attributes from this set as predictors, instead they were considered as class label attributes. C4.5 algorithm was applied to the training data with known result to obtain the rule set during the training phase. The classification rules obtained were then applied to the whole pre-processed data in testing phase and the results obtained analyzed. Figure 2 shows the prediction model used in the modelling process.

4. RESULTS AND DISCUSSION

4.1 Classification Model Results

The research found out that several factors had a great effect on employee performance. One of the most effective factors was the experience which had the maximum gain ratio. Other attributes that participated in the decision tree were Age, Qualification, Gender, Marital Status, Training and Performance Appraisal score for the year 2012 and 2013.

The experience attribute had the maximum gain ratio, which made it the starting node and most effective attribute. The attribute had positively affected the performance with employees who had more years of experience showing better performance than those with less years of experience. The figure 4 below illustrates this finding.

Age attribute showed a positive effect on performance. This effects on performance showed that younger employees portrayed poor performance. This could be due to newly working employees who did not have experience working in other companies. On the other hand, older employees may have had much experience that would influence their performance. It was observed that employees between the age of 35 years and 60 years showed better performance.

The qualification attribute, had also positively affected the performance. Employees with higher academic qualification performed better than ones with lower qualification. Figures from the experiments concluded that most employees with PhD and Master's degree had outstanding performance and exceeded expectations respectively.

The Gender attribute had shown an interesting influence on performance. It was observed that female employees exceeded expectation after undertaking a professional training in their area of expertise while male employees exceeded expectation after having 3 or more years of experience. But generally male employees had better performance compared to female employees.

Marital status, on the other hand, was clearer in its effect, since single employees had shown better performance than married employees. From the experiments, although there were more married employees than single employees, majority of single employees had exceeded expectations while majority of the married employees met expectations.

The Training attribute had a slight effect on performance. Employees who undergone professional training in their area of expertise showed better performance compared to those who did not. Generally, the number of employees who had outstanding performance after the training was higher than those who did not attend any training within the period of review.

Finally, Performance Appraisal Score for year 2012 and Performance Appraisal Score for year 2013 had some effect on performance. Employees with a Performance Appraisal Score of above 80 % marks for the two years had Exceeded Performance.

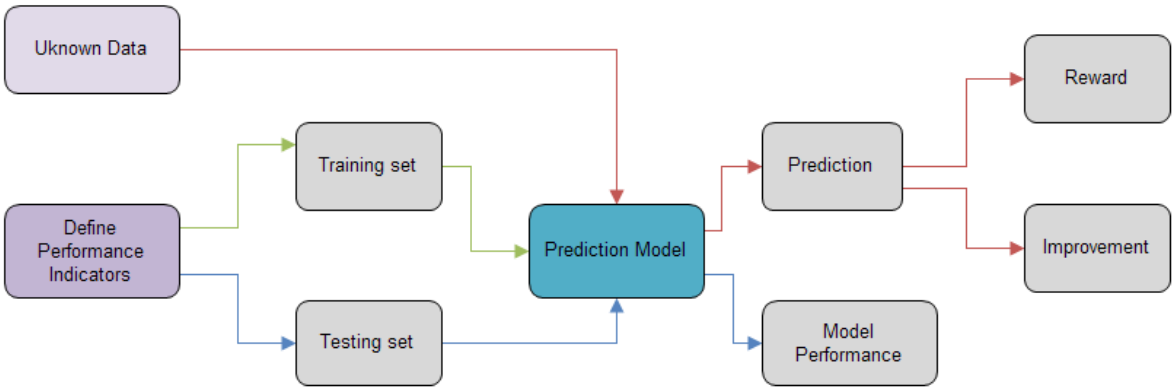

Figure 2: Prediction Model

4.1.1 C4.5 Decision Tree.

Figure 3 below shows the tree generated using the C4.5 algorithm. As indicated, experience attribute was the most effective factor.

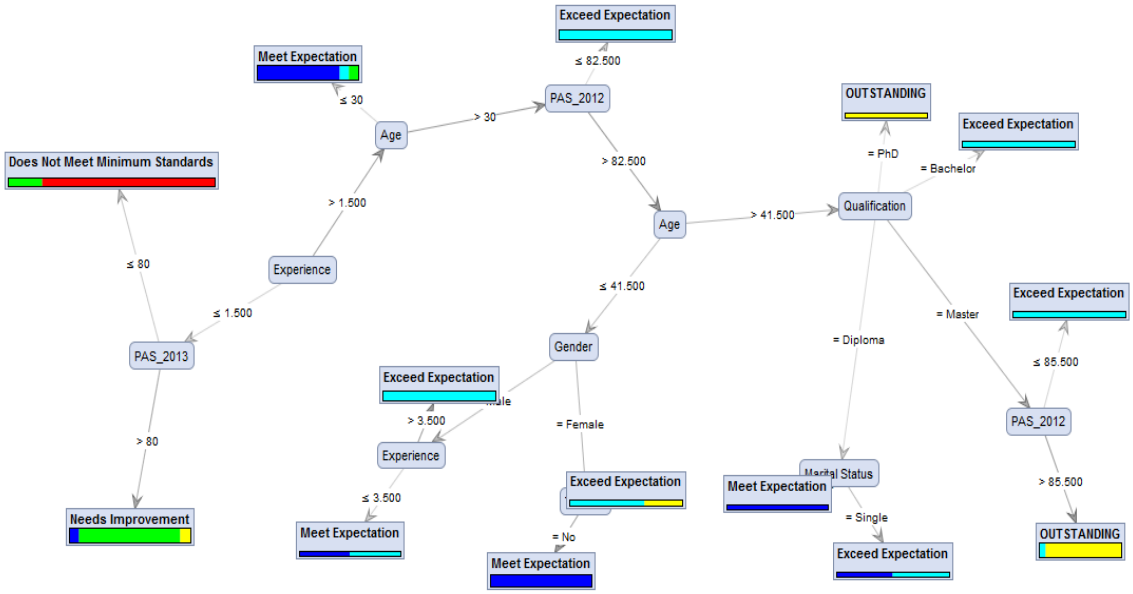

Figure 3: Decision Tree

Table 2: Classification Rules Generated By C4.5 Algorithm for Predicting Performance

#	Rule Antecedent	Performance Decision	# Of Instances
1	IF experience <= 1.5 & PAS_2013 > 80 THEN	Need Improvement	10
2	IF experience <= 1.5 & PAS_2013 <= 80 THEN	Does not Meet Minimum Standards	5
3	IF experience > 1.5 & age <= 30 then	Meet Expectation	9
4	IF experience > 1.5 & age > 30 & PAS_2012 <= 82.5 THEN	Exceed Expectation	7
5	IF experience > 1.5 & age > 30 and PAS_2012 > 82.5 & age <= 41.5 & gender ="Male" & experience > 3.5 THEN	Exceed Expectation	7
6	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age <= 41.5 & gender ="Male" & experience <= 3.5 THEN	Meet Expectation	1
7	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age <= 41.5 & gender ="Female" & training = Yes THEN	Exceed Expectation	2
8	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age <= 41.5 & gender ="Female" & training = No THEN	Meet Expectation	9
9	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "PhD" THEN	Outstanding	2
10	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "Master" & PAS_2012 > 85.5 THEN	Outstanding	11
11	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "Master" & PAS_2012 <= 85.5 THEN	Exceed Expectation	3
12	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "Diploma" and marital_status = "married" THEN	Meet Expectation	2
13	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "Diploma" & marital_status = "single" THEN	Exceed Expectation	1
14	IF experience > 1.5 & age > 30 & PAS_2012 > 82.5 & age > 41.5 & qualification = "Bachelor" THEN	Exceed Expectation	3

Figures 4 below illustrate the effects of experience attribute on performance. The experience attribute had the maximum gain ratio, which made it the starting node and most effective attribute.

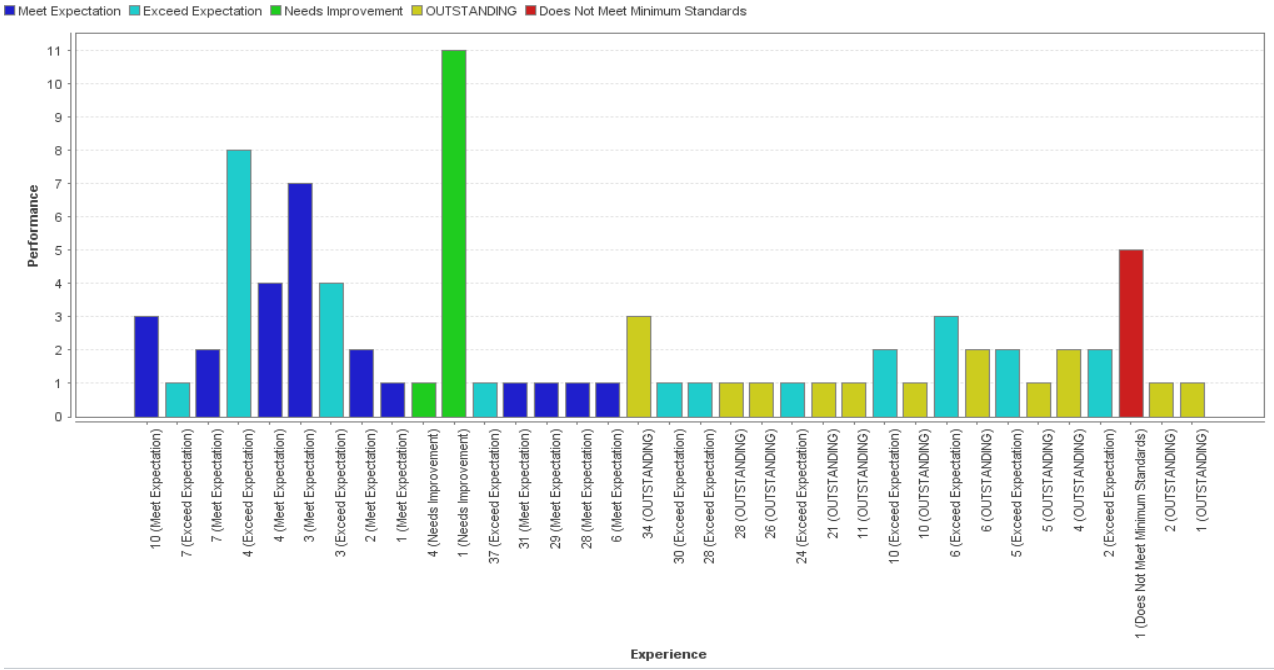

Figure 4: Effect of Experience attribute on Employee Performance

4.2 Classifier Model Evaluation

The classifier model generated after the classification process was evaluated using cross fold validation and full training evaluation techniques. Table 3 below shows the percentage accuracy results for each of the evaluation technique.

4.2.1 Cross Fold Validation

A 10 - fold cross validation with 70 % hold out was performed on the J48 algorithm. The percentage accuracy was 70.83%. A detailed accuracy by class is shown in the table 3 below.

Table 3: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.833	0.111	0.714	0.833	0.769	0.944	Meet Expectation
0.5	0.167	0.5	0.5	0.5	0.736	Exceed Expectation
1	0.045	0.667	1	0.8	0.977	Needs Improvement
0.714	0.059	0.833	0.714	0.769	0.87	Outstanding
0.667	0	1	0.667	0.8	0.833	Does Not Meet Minimum Standards
0.708	0.09	0.727	0.708	0.708	0.859	**Weighted Avg.**

Table 4: Confusion Matrix

a	b	c	d	e	Classified as
5	1	0	0	0	a = Meet Expectation
2	3	0	1	0	b = Exceed Expectation
0	0	2	0	0	c = Needs Improvement
0	2	0	5	0	d = Outstanding
0	0	1	0	2	e = Does Not Meet Minimum Standards

4.2.2 Evaluation using Full Training Dataset

A full training data set was used for the evaluation. The percentage accuracy for the J48 (C4.5) algorithm was 92.60%. A detailed accuracy by class is shown in the table below.

Table 5: Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.826	0.017	0.95	0.826	0.884	0.971	Meet Expectation
0.923	0.055	0.889	0.923	0.906	0.967	Exceed Expectation
1	0.014	0.923	1	0.96	0.999	Needs Improvement
1	0.015	0.938	1	0.968	0.994	Outstanding
1	0	1	1	1	1	Does Not Meet Minimum Standards
0.926	0.027	0.927	0.926	0.925	0.98	**Weighted Avg.**

Table 6: Confusion Matrix

a	b	c	d	e	Classified as
19	3	1	0	0	a = Meet Expectation
1	24	0	1	0	b = Exceed Expectation
0	0	12	0	0	c = Needs Improvement
0	0	0	15	0	d = Outstanding
0	0	0	0	5	e = Does Not Meet Minimum Standards

4.3 Deployment

The generated model was implemented in a web application. The essence of the web application was to map the results achieved after modeling phase to code. This was achievement by use of class methods in PHP. The result of the improved C4.5 algorithm was in the form of trees and this was translated to code in the form of if-else statements. The statements were then placed into PHP class method that accepted only the splitting attributes i.e. Experience, Age, Qualification, Gender, Marital Status, Training and Performance Appraisal score for the year 2012 and 2013 as method parameters The class method then returned the final result of that particular evaluation, indicating the performance category of an employee. Figure 5show Web page for Singular Evaluation.

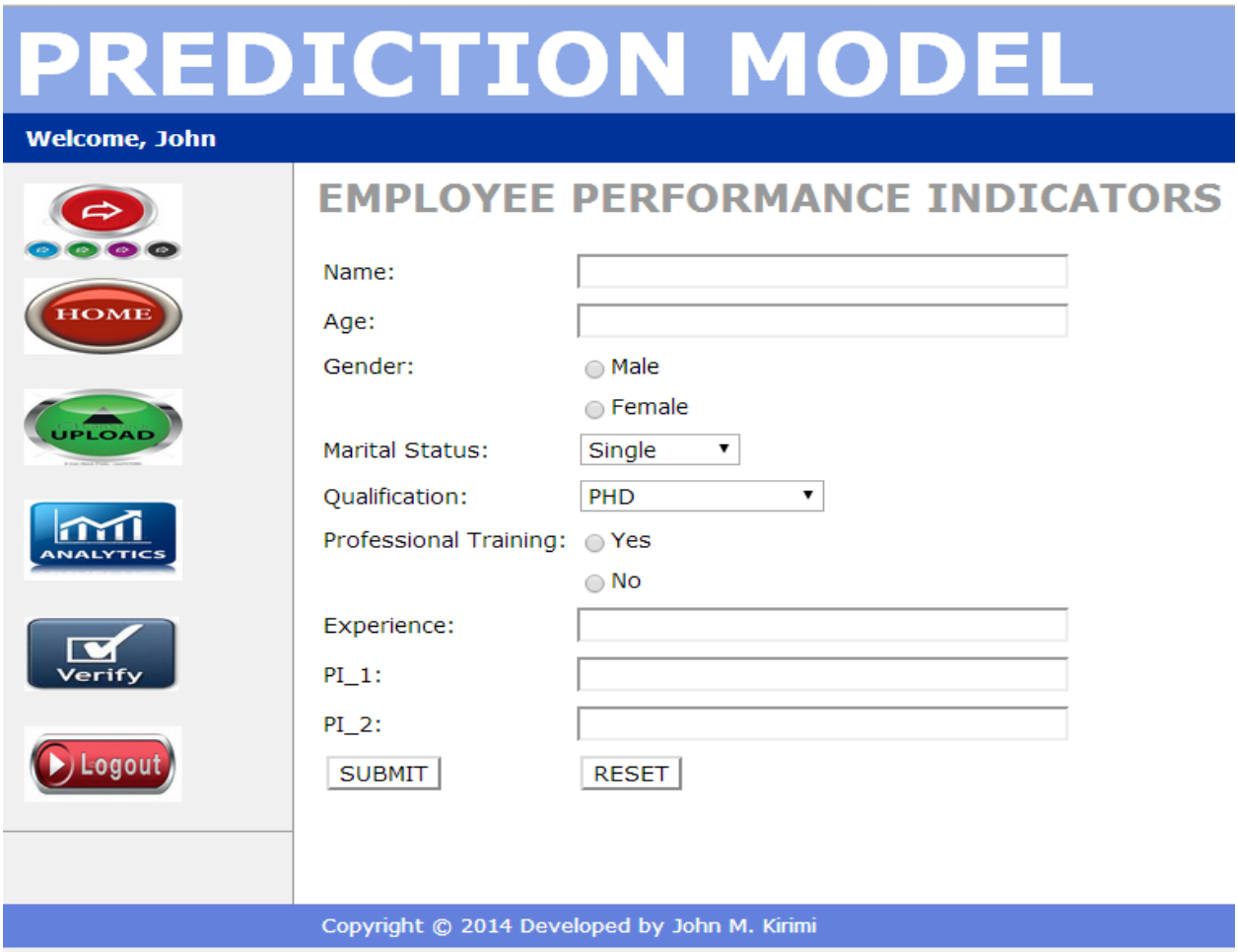

Figure 5: Web page for Singular Evaluation

5. CONCLUSION AND FUTURE WORK

This paper focused on the possibility of building a classification model for predicting employee performance. Many performance attributes were tested using performance appraisal score for the year 2012 and 2013. Some of the attributes were found effective on the performance prediction. The Experience attribute had the maximum gain ratio, which made it the starting node and most effective attribute. Other attributes that appeared on the decision tree include Age, Qualification, Gender, Marital Status, Training and Performance Appraisal Score.

The Age attribute did not show any clear effect while the Marital Status and Gender have shown some effect in predicting the performance. Educational factors like Academic Qualification and Professional Training have slightly affected the performance but not with clear trend. Finally, the effect of Performance Appraisal Score on employee performance was clear where employees with a score of above 80 % mark for the two years showing better performance.

For management of the School and HR Department, this model, and its subsequent enhancements, can be used in predicting the employee performance. Several actions can be taken in this case to avoid any risk related to hiring poorly performed employee.

In future work, it is recommended to extend the prediction of employee as a continuous value instead of predicting performance category of the employee. A comparative analysis of the category prediction (Classification) model and a value prediction would help to choose a more robust model. When the appropriate model is generated, software could be developed to be used by the HR including the rules generated for predicting performance of employees.

6. ACKNOWLEDGEMENT

The authors would like to thank the Management of Kenya School of Government for allowing access and use of the employee appraisal data. The authors also appreciate the assistance given by Joyce Maingi, the Principal Human Resource Officer at the School.

7. REFERENCES

[1]. Al-Radaideh, Q.A., Al-Nagi, E., (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, International Journal of Advanced Computer Science and Applications, 3(2), pp 144 – 151

[2]. Ancheta, R.A, Cabauatan, R.J.M., Lorena, B.T.T., Rabago, W., (2012). Predicting faculty development trainings and performance using rule-based classification algorithm, Asian Journal of Computer Science and Information Technology 2: 7, pp 203 – 209.

[3]. hein, C.F., Chen, L.F., (2008). Data Mining to improve personnel selection and enhance human capital: A case study in high technology industry, Expert Systems with Applications, 34(1), pp 280–290

[4]. Delavari, N., Phon-Amnuaisuk S., (2008). Data Mining Application in Higher Learning Institutions, Informatics in Education, 7(1), pp. 31–54

[5]. Hamidah J., Abdul R.H., Zulaiha A.O., (2009). Knowledge Discovery Techniques for Talent Forecasting in Human Resource Applications, World Academy of Science, Engineering and Technology, 3

[6]. Jantan, H., Hamdan, A. R., Othman, Z. A. (2011). Towards applying Data Mining techniques for talent management. International Conference on Computer Engineering and Applications IPCSIT Vol. 2.

[7]. Jantan, H., Hamdan, A. R., & Othman, Z. A. (2010). Human talent prediction in HRM using C4. 5 classification algorithm, International Journal on Computer Science and Engineering, 2(08-2010), pp 2526-2534

[8]. Jantan, H., Hamdan, A. R., Othman, Z. A. (2009). Knowledge discovery techniques for talent forecasting in human resource application. World Academy of Science, Engineering and Technology, Penang, Malaysia, pp 803-811

[9]. Jayanthi R., D.P. Goyal, S.I Ahson, (2008). Data Mining techniques for better decisions in human resource management systems, International Journal of Business Information Systems, 3(5), pp 464 – 481

[10]. Kotsiantis, S.B., (2007). Supervised machine learning: a review of classification techniques, Informatica, 31, pp 249-268.

[11]. Kurgan, L.A., Musilek, P. (2006). A survey of knowledge discovery and Data Mining Models, The Knowledge Engineering Review, 21(1), pp 1 - 24

[12]. Mishra, P., Padhy, N., Panigrahi, R. (2013). The survey of Data Mining applications and feature scope. Asian Journal of Computer Science & Information Technology, 2(4), pp 67 – 77

[13]. Phyu, T.N., (2009). Survey of classification techniques in data mining, Proceedings of the International Multi Conference Of Engineers And Computer Scientists, IMECS 2009, Vol 1

[14]. Ranjan, J., Goyal,D.P., S I Ahson, S.I., (2008). Data mining techniques for better decisions in human resource management systems, International Journal of Business Information Systems 3(5) pp 464 – 481

[15]. Sarda, V., Sakaria, P., Sindhu Nair, S., (2014). Relevance Ranking Algorithm for Job Portals, International Journal of Current Engineering and Technology, 4(5), pp 3157 – 3160

[16]. Strohmeier, S., Piazza, F., (2013). Domain driven Data Mining in human resource management: A review of current research, Expert Systems with Applications, 40(7), pp 2410–2420

[17]. Valle, M.A., Varas, S., Ruz, G.A., (2012). Job performance prediction in a call center using a Naive Bayes classifier, Expert Systems with Applications, 39(11), pp 9939–9945

[18]. Witten, I.,H., Frank, E., (2005). Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, Elsevier Inc.

[19]. Zhao, X.(2008). A Study of Performance Evaluation of HRM: Based on Data Mining, FITME 2008, International Seminar on Future Information Technology and Management Engineering

[20]. http://www.ksg.ac.ke/