ISSN No. 0976-5697

Volume 4, No. 3, March 2013 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Data Mining Based Predictions for Employees Skill Enhancement Using Pro-Skill-Improvement Program and Performance Using Classifier Scheme Algorithm

Akhilesh k. Sharma
Asst. Prof.
Sir Padampat Singhania University, Udaipur
akhileshshm@gmail.com

Dr. Kamaljit Lakhtaria
Asst. Prof.
Sir Padampat Singhania University, Udaipur
kamaljit.lakhtaria@spsu.ac.in

Santosh Vishwakarma
Asst. Prof.
Sir Padampat Singhania University, Udaipur
santosh.vishwakarma@spsu.ac.in

Abstract: In the research paper we proposed extraction of knowledge significant for predicting the skill development program needs of newly-hired employees, in order to enhance employees inherent potentials using classifier scheme for data mining technique. In discovering significant models needed for predictive analysis, the Cross Industry Standard Process for Data Mining (CRISP-DM) was employed. Results show that professional skill development programs are needed in order to prepare employees to perform their tasks more efficiently.The knowledge flow model of the Opensource tool is also used to frame the elements. The tool from Waikato university is used for the framing of the model adapted.

Keywords: Data mining, skill development, Classifier, CRISP-DM,KDD, Bayes Net, JRIP

I. INTRODUCTION

In today's modern era of technology there is an explosive growth of available data as a result of computerization [1] of almost every aspect of the day-to-day operations of organizations has instinctive contributions to the development of intelligent decision making technologies for any kind of enterprise oriented decision. A young yet promising of this kind is data mining. It standouts due to its wide-array of techniques from the different domains such as statistics, artificial intelligence, machine learning, algorithms, database systems and visualization[1][18]. These influences served as groundwork for its applications to business for which the academic institution is unexceptionally classified. Generally, regardless of discipline, data mining has gained popularity due to its tools with potentials to identify trends within data [18] and turn them out into knowledge [1] mostly with predictive attributes that could significantly lead to better and strong bases for decision making with a wide range of Open source tools availability.

In todays scenario there is a Widely-held academic data mining researches are student-centric with much emphasis on predicting student learning outcomes/ performance like in [3, 19, 20] and [21]. Subsequently, this was extended to its applications to human resource management. In [24] for instance, data mining is used to measure the quality of a teacher. In other institutions as in [2],[4],[5],and [6], employee performance have been studied using data mining techniques. From the educational context, once hired, a teacher takes part of the academic institution's goal to increase productivity and to provide excellent customer service. While these roles are expected to be executed in the entire pedagogical process for the performance. Training and productivity development programs are administered to help augment teacher's inherent capabilities to perform.

The Organization Of The Paper: The paper is organized as follows: First is the Introduction which is entailing the overall description of the paper. the second consolidates the related studies in educational data mining; third provides the research problem and objectives; fourth are detailed discussions of the methodology and process of knowledge discovery; section 5 exhibits the simulation, analysis and results; and finally, the paper ends with a conclusion and an outlook for future directions.

II. RELATED WORKS

There have been quite a lot of studies of data mining in the educational domain. These concerned about students and employees' performances.

Related to this is the study of Bhardwaj and Pal [19] in which a data model was used to predict student's performance with emphasis on identifying the difference of high learners and slow learners using byes classification. Decision tree as a classification algorithm has been utilized in [21] to predict the final grade of a student in a particular course. The same algorithm has been applied in [23] on past student performance data to generate a model to predict student performance with highlights on identifying dropouts and students who need special attention and allow teachers to provide appropriate advising or counselling. Conversely, Pandey and Pal [24] have considered the qualities the

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

teacher must possess in order to determine how to tackle the problems arising in teaching, key points to be remembered while teaching and the amount of knowledge of the teaching process. In the course of identifying significant recommendations, John Dewey's principle of bipolar and Reybem's tri-polar education systems have been used to establish a model to evaluate the teachership on the basis of student feedback using data mining. While [2] used classification technique to build models to predict new applicant's performance, [5], [8], [9] and [11] used the same to forecast employee's talents. Another technique called fuzzy has been applied in [10] to build a practical model for improving the efficiency and effectiveness of human resource management while [15] has improved and employed it to evaluate the performance of employees of commercial banks. Other noteworthy researches that have added significant contributions to this study may be referred from [16] to [24]. While cited studies substantiated the applications of data mining in the educational domain, there is none that has applied data mining to predict the training and development needs of employees based on their inherent characteristics which could be initially mined from entry credentials such as resume and other supporting documents. Also, classifications as to what training needs are required to individual and a group of employees were not included. 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 training and development needs that is parallel to the criteria used for performance evaluation.

III. RESEARCH PROBLEM AND OBJECTIVES

Human capital is of a high concern for companies' management where their most interest is in hiring the highly qualified personnel which are expected to perform highly as well [2]. Foremost, the Human Resource (HR). Management plays the role of ensuring this by closely adhering to the standards set by the higher management or by some heuristic needs of applicants with distinctive qualifications and potentials.

However, oft-quoted factors that may affect employee performance are attributed to educational backgrounds, working experiences, as well as personal qualities. These when converged provide a picture of how well an employee performs his tasks. Assessment of human resource performance is a sensitive task. To avoid partiality, an efficient tool to deal with various data and assist managers to make decisions and plans is of great help. In data mining, historical data such as those attributes that influence performance could be exploited as learning experiences. These can be used to predict future circumstances and rich resource of knowledge and decision supports.

This is commonly referred to as supervised, classification or inductive learning [25]. For instance, in academic institutions like Technological Institute of the Philippines, training and developments are administered to newly-hired faculty members. Some are of institutional purposes such as policy and practices orientation, and behavioural and functional competence related trainings. However, to make training and development needs parallel to areas concerning evaluation, and with data mining's predictive capabilities, this study strives to propose a data mining approach to help HR forecast these needs. This helps managers and decision makers to identify development programs resembled to the needs to enhance intrinsic qualifications of teachers and areas pertaining to performance evaluation.

IV. METHODOLOGY AND KNOWLEDGE DISCOVERY

Generally, to carry out the process of discovering pieces of information needed to draw out scientific predictions, a framework serves as a guide to drill-down the details and roll-up the entirety of a procedure. This paper combines the CRISP-DM methodology (Cross Industry Standard Process for Data Mining)[26] and Process of Knowledge Discovery in [1] in which data mining is a significant step. The iterative and sequence of steps are shown in Figure 1. These include business understanding, data understanding, data preparation, modeling, evaluation and deployment along with the data discovery processes such as data cleaning, data integration, data selection, data transformation, data mining, evaluation and presentation.

A. Business Understanding:

In the section of Business understanding with proper endorsement and approval of some academic administrators, questions as to how the (DM) data mining functionalities are best applied in any Technological Institute has been identified. Recent studies had However, oft-quoted factors that may affect employees performance are attributed to educational backgrounds, working experiences, as well as personal qualities. These when converged provide a picture of how well an employee performs his tasks. Assessment of human resource performance is a sensitive task. To avoid partiality, an efficient tool to deal with various data and assist managers to make decisions and plans is of great help. In data mining, historical data such as those attributes that influence performance could be exploited as learning experiences. These can be used to predict future circumstances and rich resource of knowledge and decision supports. This is commonly referred to as supervised, classification or inductive learning [25]. For instance, in academic institutions like Technological Institute of the Philippines, training and developments are administered to newly-hired faculty members. Some are of institutional purposes such as policy and practices orientation, and behavioural and functional competence related trainings. However, to make training and development needs parallel to areas concerning evaluation, and with data mining's predictive capabilities, this study strives to propose a data mining approach to help HR forecast these needs. This helps managers and decision makers to identify development programs resembled to the needs to enhance intrinsic qualifications of teachers and areas pertaining to performance evaluation.

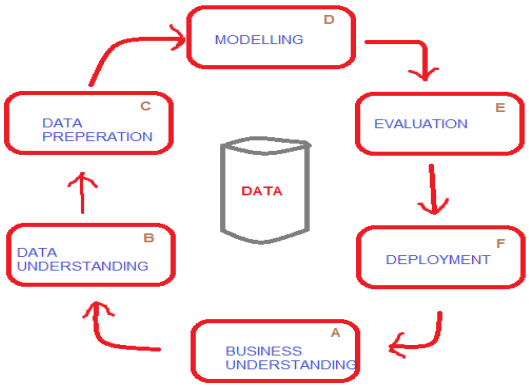

Figure1: Knowledge Data Discovery Framework for KDD

For this reason, the Human Resource Department was considered as the paper's area of interest. Along with the responsibilities of an academic institution to provide quality education to students is they examine on how its human resources particularly the teachers perform in the realization of its goals as clearly defined by the so-called graduate attributes. Commissioned to carry out these duties is the Human Resource Department. With respect to the needs of each department, qualified applicants are selected. Once hired, employees potential are re-examined to see how well they can execute their duties. To evaluate their performance, standard evaluation tool of the institution is used. Apart from the development trainings and seminars administered for newly hired teachers such as start-of-semester orientation, integrity of the workplace, on-the-job training program, and ICT Skills Enhancement, there is no such tool used to link these trainings to entry profile of teachers as well as a device to make these parallel to the criteria of performance evaluation. To digest then, the current state of administering these HR roles and policies, processes and evaluation tool were identified and reviewed. To substantiate these are actual data such as faculty profile, and performance evaluation criteria. These were used to predict the training needs of teachers as well as to forecast their performance to the significant areas of evaluation. To figure out the interrelationships among the attributes affecting teacher performance, Table 1 and 2 are exhibited

B. Data Preparation:

Tuples structured using the template presented in Table 3 were discretized to fit in with the requirements of data classification. Table 4 shows the possible values of each attribute of the data set. For instance, original values for graduate and undergraduate courses were changed to vertically-aligned, aligned, and not-aligned. The process of discretization was performed to most attributes in order to avoid numerical values.

V. SIMULATION RESULTS & ANALYSIS:

As shown in table with the priorities highlighted in Table 1 and the possible values of discretized data set in Table 2, class tuples of the data set were labeled based on their achievable heuristic importance to the main goals or objectives. And the class labels here include professional training and personal training. The Waikato tool Environment for Knowledge Analysis (WEKA) 3.60 (knowledge Flow) serves as an intelligent tool for data analysis and predictive modeling (as it is an open source tool). It is in this regard that aggregated data was transformed into Attribute-Relation File Format (ARFF), a WEKA readable format of data set. Figure 2 shows the WEKA Explorer with the loaded arff-formatted data set ready for analysis.

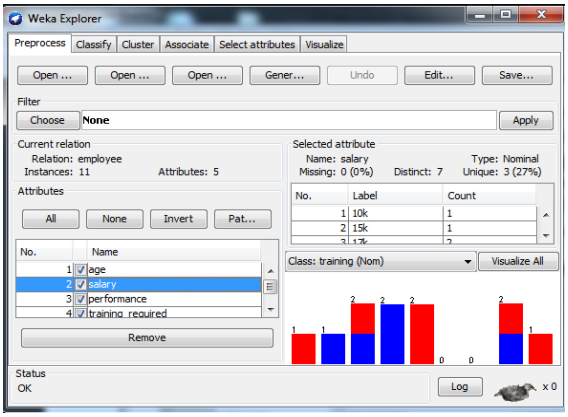

Figure 2. WEKA Explorer and Data Set

With WEKA's open source wide collection of free analytical tools, and also data mining algorithms, it was chosen among others to be the primary tool for classifying and associating employee attributes with training and performance level. Generally, data mining classification technique is a two-step process consisting of learning step where classification model is constructed and classification (see in figure of knowledge flow) step where the model is used to predict class labels for a given data. In the first step, a classifier (commonly referred to as model) is built describing a predetermined set of data classes. It is in this stage where a classification algorithm builds the classifier by analyzing or learning from a training set made up of data set tuples and their associated class labels too. Because of the class label for which each of the training tuple is provided, this step is known as supervised learning [1]. In the second step, test data are used to estimate the accuracy of the classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data tuples [1].

As we know the Supervised learning is simply a formalization/specialization of the idea of learning from the examples. In supervised learning, generally the learner (typically, a computer program) is provided with two sets of data, *a training set* and *a test set*. The idea is for the learner to "learn" from a set of labeled examples in the training set so that it can identify unlabeled examples in the test set with the highest possible accuracy for the prediction. For example, a training set might consist of images of different types of fruit (say, grapes and nectarines), where the identity of the fruit in each image is given to the learner for examination. The test set would then consist of more unidentified pieces of fruit, but from the same classes. The goal is for the learner to identify the elements in the test set. There are so many different approaches which attempt to build the best possible method of classifying examples of the test set by using the data given in the training set.

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

Table 1 : Result summarization for the Bayes Network classification:

```
=================================================
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
 0.4       0.833      0.286      0.4       0.333      0.067 professional_training
 0.167     0.6        0.25       0.167     0.2        0.067  personal_training

=================================================
Weighted Avg.    0.273    0.706    0.266    0.273    0.261    0.067
=================================================
```

Various evaluation metrics for predictive accuracy of a classifier include hold-out and random sub-sampling, cross-validation and bootstrap methods for example. These methods share similar characteristics since they are all based on randomly sampled partitions of a given data but differ in terms of processes and techniques [1]. In hold-out method, the given data are randomly partitioned into two independent sets, *a training set* and a *test set*. Typically, two-thirds of the data are allocated to the *training set*, and the remaining one-third is allocated to the *test set* [1]. In cross-validation, the initial data are randomly partitioned into mutually exclusive subsets or folds, each of approximately equal size. Training and testing is performed several times depending on set folds. In the first iteration, a partition is reserved as test set and the remaining partitions are collectively used to train a model [1].

The bootstrap method is for samples the given training tuples uniformly with replacement. That is, each time a tuple is selected, it is equally likely to be selected again and re-added to the training set [1]. For simplicity, clarity and dependability, the hold-out method is used to selecting a training set to derive a model and a test set to assess the accuracy of the generated model.we used the technique of the bayes net for the further classifier for having simulated our desired content for the development And also we used the JRIP for the classifier on the data set employee ARFF data set which dealt for the descretization on data for preprocesss & then we used the classifier in weka for the JRIP to give & prove our results as in figure3. In Weka, this is performed using percentage split [28]. The process is simplified for illustrative purposes in Figure 3.

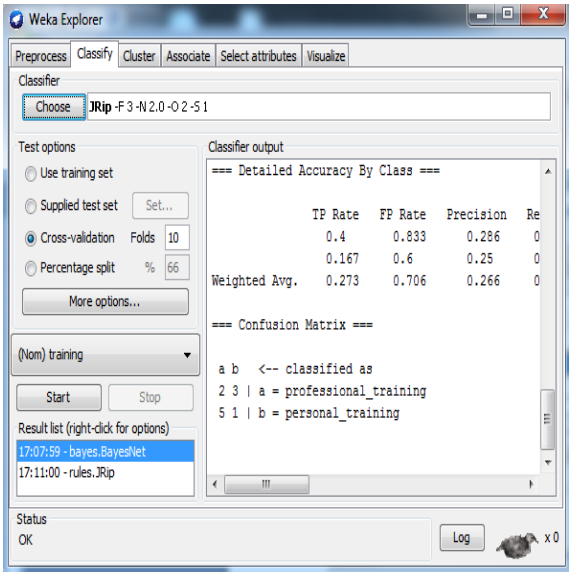

Figure 3: Result Data set description using Simulation on Bayes Net

A. *Modelling:*

Classification and association have been chosen as the most appropriate data mining functionalities for training and performance predictions. The former is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorically discrete, unordered class labels [1]. The basic classification techniques include decision tree, bayesian, and rule-based. In contrast to the first two techniques, rule-based classification learned model is represented as a set of IF-THEN rules. These rules are generated either from a decision tree or directly from the training data using sequential covering algorithm (SCA). An if-then rule is an expression of the form IF condition THEN conclusion in which the "If" part (left side) is the rule antecedent or precondition and "Then" part (right side) is the rule consequent[1].

Table 2: Pridiction results:

```
=================================================
=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0         0.167     0         0         0          0.417  professional_training
  0.833     1         0.5       0.833     0.625      0.417  personal_training
-------------------------------------------------------------------------------------
Weighted Avg.    0.455    0.621    0.273    0.455    0.341    0.417
=================================================
```

Rule-based classification is also attributed to accuracy and coverage. These measures whether or not a rule antecedent are satisfied and the rule covers the data set tuples. From a class-labeled data set, coverage refers to the number of tuples covered by the rule while accuracy determines the tuples correctly classified by the rule. Significant in rule-quality measures that considers both accuracy and coverage is the First-Order Inductive Learner (FOIL), a sequential covering algorithm that learns the first order logic rules [1]. Definitions of these measures are provided in Figure 4.

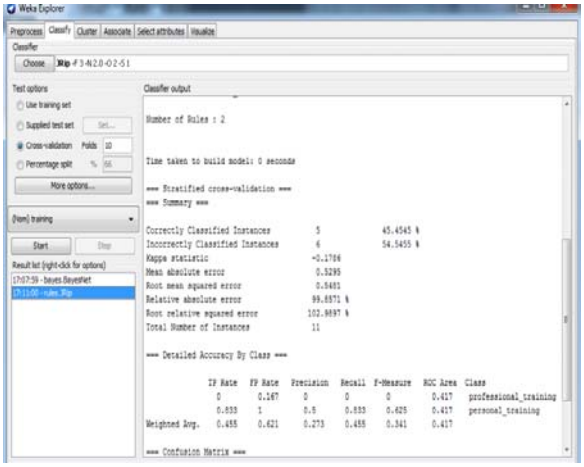

Figure 4. Coverage and Accuracy Classification Measures

At this point, to aggregate the classification tool's specific supports, this study used a rule-based classification technique using sequential covering algorithm (SCA) and hold-out method of partitioning sets of data. For rules quality, FOIL was chosen as it considers both coverage and accuracy. In WEKA, a cloned Ripper algorithm called Jrip[29] is designed to execute classification of data sets while simulating the process of sequential covering algorithm. Selected due

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

to its capability to directly classify data set without having to base rules on a decision tree, Figure 5 provides evidence of its simplicity.

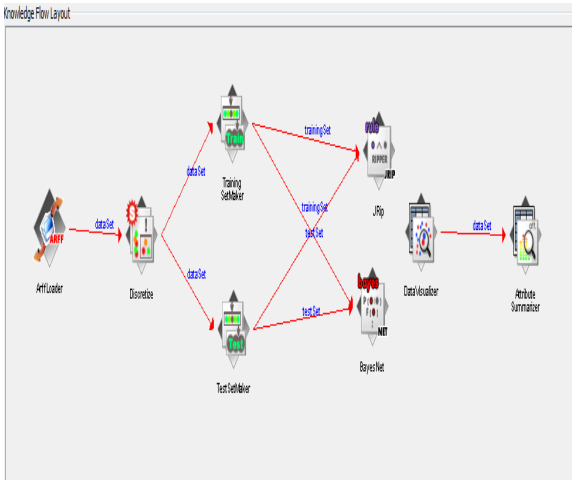

Figure 5: Knowledge Flow Layout for our Strategy

=== Confusion Matrix ===
a b <-- classified as
2 3 | a = professional_training
5 1 | b = personal_training
==========

VI. CONCLUSION AND FUTURE WORK

The buzzword now a days DM, is the Discovering knowledge based from previous learning makes data mining an efficient and effective tool to predict/forcast future occurrences which are responsible that may affect decisions for the current situations. In this paper, the predictions were made on the performance of newly-hired employees based on their current training needs. This give us a concluding remark that sensitive issues like these would bring out implications as to how and what programs are needed to enhance the inherent potentials of employees. Moreover, this study introduces potential applications of data mining in the educational domain not just for human resource management and student-related concerns but for other academic and non-academic areas also. This study further outlooks for applications of results to analyze enhancement programs for senior employees of any organization and to identify patterns affecting both teacher and student performance using other data mining techniques such as association rules.

VII. REFERENCES

[1]. Jiawei Han, Michlelline Kamber, & Jian Pei (2008). Data mining: concepts and techniques. Morgan Kaufmann publishers

[2]. Qasem A. Al-Radaideh, Eman Al Nagi (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012

[3]. Mohammed M. Abu Tair, Alaa M. El-Halees. (2012) Mining Educational Data to Improve Students' Performance: A Case Study. Volume 2 No. 2, February 2012. ISSN 2223-4985

[4]. Stavrou-Costea, E. (2005). The challenges of human resource management towards organizational effectiveness A comparative study in Southern EU. Journal of European Industrial, 2005. **29**(2): p. 112-134.

[5]. Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman (2010). Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application. 2010 International Journal of Human and Social Sciences 5:11

[6]. Qasem A. Al-Radaideh, Eman Al Nagi (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012 .

[7]. Data Mining Techniques. http://www.dataminingtechniques.net/data-mining-tutorial/data-mining-processes/

[8]. Hamidah Jantan, Abdul Razak Hamdan, and Zulaiha Ali Othman Human (2010). Talent Prediction in HRM using C4.5 Classification Algorithm. (IJCSE) International Journal on Computer Science and Engineering. Vol. 02, No. 08, 2010, 2526-2534

[9]. Hamidah Jantan, Abdul Razak Hamdan, Zulaiha Ali Othman. Towards applying Data Mining Techniques for Talent Management

[10]. Han Jing (2009). Application of Fuzzy Data mining Algorithm in Performance Evaluation of Human Resource. 2009 International Forum on Computer Science-Technology and Applications

[11]. Hamidah Jantan , Abdul Razak Hamdan and Zulaiha Ali Othman (2011). Talent Knowledge Acquisition using Data Mining Classification Techniques. 3rd Conference on Data Mining and Optimization (DMO)

[12]. ZHAO Xin (2008). A Study of Performance Evaluation of HRM: Based on Data mining. International Seminar on Future Information Technology and Management Engineering

[13]. Yan Huang (2009). Study of College Human Resources Data Mining Based on the SOM Algorithm. 2009 Asia Pacific Conference on Information Processing

[14]. Chen Xiaofan, and Wang Fengbin (2010). Application of Data Mining on Enterprise Human Resource Performance Management. 3rd International Conference on Information Management, Innovation Management and Industrial Engineering

[15]. Honglei Zhang (2009). Fuzzy Evaluation on the Performance of Human Resources Management of Commercial Banks Based on Improved Algorithm. 2009 2nd International Conference on Power Electronics and Intelligent Transportation System.

[16]. Qiangwei Wang, Boyang Li and Jinglu Hu. (2009). Feature Selection for Human Resource Selection Based on Affinity Propagation and SVM Sensitivity Analysis.

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council ,
IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India

[17]. Yuan Qi and Rosalind W. Picard (2002). Context-sensitive Bayesian Classifiers and Application to Mouse Pressure Pattern Clasification
[18]. The Value in Mining Data. http://www.information-management.com/ news/4618-1.html
[19]. Brijesh Kumar Bhardwaj, Saurabh Pal (2012). Data Mining: A prediction for performance improvement using classification
[20]. Qasem A. Al-Radaideh, Emad M. Al, et. al. Mining Student Data Using Decision Trees. [22] Singh, C. (2011). Extraction and analysis of faculty performance of management discipline from student feedback using clustering and association rule mining techniques. Electronics Computer Technology (ICECT).
[21]. Surjeet Kumar Yadav, Brijesh Bharadwaj (2012). Data Mining Applications: A comparative Study for Predicting Student's performance
[22]. Umesh Kumar Pandey (2012). Mining Data to Find Adept Teachers in Dealing with Students. I.J. Intelligent Systems and Applications
[23]. Supervised-learning. http://www.cs.uic.edu
[24]. Data Mining Processes. http://www.dataminingtechniques.net/data-mining-tutorial/data-mining-processes/
[25]. Data Mining - Classification II. http://www-staff.lboro.ac.uk/~comds2/coc131/COC131_tutorial_w7. pdf
[26]. WEKA Primer. http://weka.wikispaces.com/Primer
[27]. Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems.
[28]. http://www.dataminingtechniques.net/data-mining-tutorial/data-mining-processes/
[29]. Supervised Learning and Bayesian Classification, Erik G. Learned-Miller, September 12, 2011.
[30]. A. Sharma, J.malik, P. goyal, Top ten Algorithms in Data mining, INDIACOM-2010, New Delhi.
[31]. Akhilesh K Sharma, Kamaljit I Lakhtaria, The Feature Extraction methods for the Musical archives to classify the Music using Open Source Feature Extraction Library,Institution of Engeeniring (IEI)-LACE-13,MNIT Jaipur, 4th-Feb-2013.

CONFERENCE PAPER
II International Conference on
"Advance Computing and Creating Entrepreneurs (ACCE2013)"
On 19-20 Feb 2013
Organized by
2nd SIG-WNs, Div IV & Udaipur Chapter , CSI , IEEE Computer Society Chapter India Council , IEEE Student Chapter Geetanjali Institute of Technical Studies, Udaipur, Rajasthan, India