# A Novel Approach for Faculty Appraisal in Educational Data Mining using CLEMENTINE TOOL

**Ramakrishna Gandi, Prathimarani Palla, Madhuri Thimmapuram, Daniel Prasanth T**

***Abstract— Data mining, the concept of unseen predictive information from big databases is a powerful novel technology with great potential used in various commercial uses including banking, retail industry, e-commerce, telecommunication industry, DNA analysis remote sensing, bioinformatics etc. Education is a required element for the progress of nation. Mining in educational environment is called Educational Data Mining. Educational data mining is concerned with developing new methods to discover knowledge from educational database. In order to analyze opinion of students about their teachers in Professor Appraisal System, this paper surveys an application of data mining in Professor Appraisal System & also present result analysis using CLEMENTINE 12.0 tool. There are varieties of popular data mining task within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. How each of data mining tasks can be applied to education system is explained. In this paper we analyze the performance of final Faculty Appraisal of a semester of a computer engineering department, Vignan Institute of Information Technology College of engineering & is presented the result which it is achieved using CLEMENTINE 12.0 tool. We have verified hidden patterns of Faculty Appraisal by students and is predicted that which Faculty will be invited to faculty classes and which Faculty will be refusing and department heads due to Appraisal reasons will ask explanations with them.***

***Index Terms— Classification, Clustering, Association rule, Data mining, Appraisal, CLEMENTINE 12.0.***

## I. INTRODUCTION

Data mining has involved a great deal of responsiveness in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the forthcoming need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration [1]. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. The transition won't occur automatically; in this case, there is a need for data mining [2]. Mining applied in education was published in 1995 by Sanjeev and Zytkow. Researchers gathered the knowledge discovery as terms like “P pattern for data in the range R” from university database [3]. Vranić and Skoćır was examined how to improve some aspects of educational quality with data mining algorithms and techniques by taking a specific course students as target audience in academic environments [4].In this paper we have collected information and results of a appraisal about 30 professors in Vignan Institute of Information Technology College of Engineering, Department of Computer Engineering on professor's performances in classroom then with data mining algorithms such Association Rule and decision trees (C&RT) , it is proceeded to analyze and predict acceptation of a professor for continuing the teaching in that subject .There are new rules and relations between selected parameters such as Teaching, Professor Degree, Preparation, Communication, Class Control, Teaching experience, Approved Staff to next semesters on professor appraisal system that is interested for Heads of Departments of Institution.

**Revised Version Manuscript Received on April 27, 2016.**

**Ramakrishna Gandi**, CSE Department,Vignan’s Institute of Information Technology, Visakhapatnam(A.P),India.

**Prathimarani Palla**, CSE Department, Vignan’s Institute of Information Technology, Visakhapatnam (A.P),India.

**Madhuri Thimmapuram**, CSE Department, Vignan’s Institute of Information Technology, Visakhapatnam (A.P), India.

**Daniel Prasanth T**,CSE Department, Vignan’s Institute of Information Technology, Visakhapatnam (A.P),India.

## II. METHODOLOGY

In this research study, We have followed a popular data mining methodology called Cross Industry Standard Process for Data Mining (CRISP-DM), which is a six-step process [5]:

- **Problem explanation**: Comprises understanding development goals with business perspective.
- **Understanding the data**: Includes identifying the sources of data.
- **Formulating the data**: Includes pre-processing, cleaning, and transforming the relevant data into a form that can be used by data mining algorithms.
- **Creating the models**: Includes developing a wide range of models using comparable analytical techniques.
- **Assessing the models**: Includes evaluating and assessing the validity and the utility of the models against each other and against the goals of the study.
- **Using the model**: Includes in such activities as deploying the models for use in decision making processes.

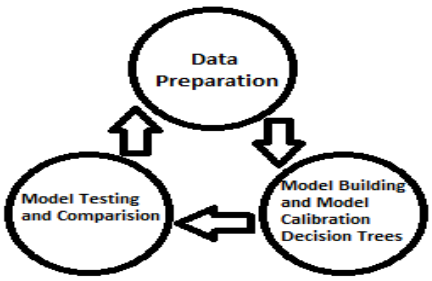


**Fig.1.A graphical illustration of the methodology employed in this study**



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

### A. DATA

In this study 34 records were used which is taken from feedback_2014_15_sem_1 of Department of computer engineering, Vignan Institute of Information Technology College of Engineering. Dataset have professors' information such as Teaching, Preparation, Communication, Class Control along with this we have included Professor Degree, Professor Experience, Approved Staff.

**Table 1. The list of independent variables used in this study**

| Variable Name | Data Type | Description |
|---|---|---|
| Teaching | Text | Teaching Score |
| Professor Degree | Text | Professors Degree |
| Preparation | Text | Preparation Score |
| Communication | Text | Communication Score |
| Class_ Control | Text | Class Control Score |
| Teaching_ experience | Text | Teaching experience of Professor |
| Approved_ Staff | Text | Approved Professor or not |

**Table 2. The list of independent variables and values used in this study**

| Variable Name | Data Type | Values |
|---|---|---|
| Teaching | Text | {Excellent, Good, Satisfactory ,Poor} |
| Professor_ Degree | Text | {BE,ME,PHD} |
| Preparation | Text | {Excellent, Good, Satisfactory, Poor} |
| Communication | Text | {Excellent, Good, Satisfactory, Poor} |
| Class_ Control | Text | {Excellent, Good, Satisfactory, Poor} |
| Teaching_ experience | Text | {TRUE,FALSE} |
| Approved_ Staff | Text | {Yes, No} |

Teaching score of professors which are studying in Vignan Institute of Information Technology College of Engineering, Computer Engineering Department Faculty are represented by the word system. Score ranges of these words are shown in Table 3.

**Table 3. The output variable (Evaluation score) used in the study**

| Raw Score | Nominal Representation |
|---|---|
| Score < 60 | Poor |
| 60<=Score < 75 | Satisfactory |
| 75<=Score< 85 | Good |
| 85<=Score<= 100 | Excellent |

**Table 4. The output variable (Teaching experience) used in the study**

| Raw –Years of Teaching | Nominal Representation |
|---|---|
| Years < 3 | False |
| Years >= 3 | TRUE |

### B. Background

In this research we have used **CLEMENTINE 12.0** and Data mining. The following subsections contain a summary of these topics.

#### a. CLEMENTINE 12.0

Clementine is a mature data mining toolkit which aims to allow domain experts (normal users) to do their own data mining. Clementine has a visual programming or data flow interface, which simplifies the data mining process. Clementine applications include customer segmentation for marketing companies, fraud detection, credit scoring, load forecasting for utility companies, and profit prediction for retailers. SPSS Clementine was one of the very first general purpose data mining tools, and one of the most popular data mining packages[6].

#### b. Data Mining

Data mining is the method of defining interesting knowledge from big amount of data stored in database, data warehouse or other information sources. It includes various tasks such as classification, clustering, association rule etc.

#### 1. Association Rule

Association rules are used to show the relationship between data items. Association rule generation consists of two separate steps: First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules [6]. Support & confidence are the normal method used to measure the quality of association rule. Association rule can be used in educational data mining and professor˝s appraisal system for analyzing the learning data.

#### 2. Classification

Classification is a data mining task that maps the data into predefined groups & classes. It is also called as supervised learning .It consists of two steps:

2.1 Model construction: It consists of set of predetermined classes. Each sample is assumed to belong to a predefined class. The set of sample used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae.

2.2 Model usage: This model is used for classifying upcoming or unidentified objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur. [6]

#### c. Clustering

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group. Clustering can be considered the most important unsupervised learning technique. In educational data mining and professor's appraisal system, clustering has been used to group the professors according to their behavior e.g. clustering can be used to distinguish active professor from non-active professor according to their performance in activities

## III. ARCHITECTURE OF PROPOSED SYSTEM

In this paper, it is done a feedback of 2014_15 academic year and semester-1 survey from 319 students then it is prepared results of this survey for 34 professors.

#### a. The Explorer Interface of CLEMENTINE

Initially "preprocess" [7] will have been selected. This is the tab you select when you want to tell CLEMENTINE where to find the data set that you want to use. At the start of a session, you see the Clementine User Interface



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

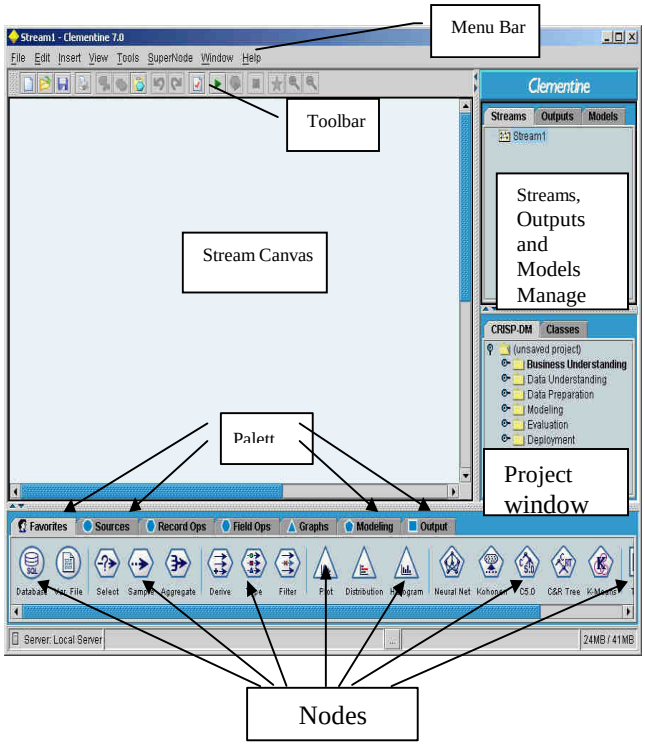


**Fig 3.1 Clementine User Interface**

Clementine enables you to mine data by visual programming techniques using the Stream Canvas. This is the main work area in Clementine and can be thought of as a surface on which to place icons. These icons represent operations to be carried out on the data and are often referred to as nodes.

The nodes are contained in palettes, located across the bottom of the Clementine window. Each palette contains a related group of nodes that are available to add to the data stream. For example, the Sources palette contains nodes that you can use to read data into your model and the Graphs palette contains nodes that you can use to explore your data visually. Which icons are shown depends on the active, selected palette.

The Favorites palette is a customizable collection of nodes that the analyst uses most frequently. It contains a default collection of nodes, but these can be easily modified within the Palette Manager (reached by clicking Tools. Favorites). (either text or chart) or a model. At the upper right of the Clementine window (shown above), there are three types of manager tabs. Each tab (Streams, Outputs, and Models) is used to view and manage the corresponding type of object. You can use the Streams tab to open, rename, save, and delete streams created in a session. Clementine output, such as graphs and tables, are stored in the Outputs tab. You can save output objects directly from this manager.

The Models tab is the most important of the manager tabs as it contains the results of the machine learning and modeling conducted in Clementine. These models can be browsed directly from the Models tab or added to the current stream displayed in the canvas. At the lower right of the Clementine window we have the Projects window. This window offers you a best- practice way to organize your data mining work. The CRISP-DM tab helps you to organize streams, output, and annotations according to the phases of the CRISP-DM process model (mentioned in Chapter 1). Even though some items do not typically involve work in Clementine, the CRISP-DM tab includes all six phases of the CRISP-DM process model so that you have a central location for storing and tracking all materials associated with the project. For example, the Business Understanding phase typically involves gathering requirements and meeting with colleagues to determine goals rather than working with data in Clementine. The CRISP-DM tab allows you to store your notes from such meetings in the Business Understanding folder of a project file for future reference and inclusion in reports.

The Classes tab in the Project window organizes your work in Clementine categorically by the type of objects created. Objects can be added to any of the following categories:

- Streams
- Nodes
- Models
- Tables, graphs, reports
- Other (non-Clementine files, such as slide shows or white papers relevant to your data mining work)

If we turn our attention to the Clementine menu bar there are eight menu options:

- File allows the user to create, open and save Clementine streams and projects. Streams can also be printed from this menu.
- Edit allows the user to perform editing operations: for example copy/paste objects; clear manager tabs; edit individual nodes.
- Insert allows the user to insert a particular node, as alternative to dragging a node from the palette.
- View allows the user to toggle between hiding and displaying items (for example: the toolbar or the Project window).
- Tools allows the user to manipulate the environment in which Clementine works and provides facilities for working with Scripts.

Supermodel allows the user to create, edit and save a condensed stream. Super nodes are discussed in the Data Manipulation with Clementine training course.

- Window allows the user to close related windows (for example, all open output windows).
- Help allows the user to access help on a variety of topics or view a tutorial. [7]

### *b. Reading Data Files into Clementine*

Clementine reads a variety of different file types, including data stored in spreadsheets and databases, using the nodes within the Sources palette. Here, data can be read in from text files, in either free-field or fixed-field format, using the Var. File and Fixed File source nodes.

### *c. Reading Data from Free-Field Text Files*

The Var. File node reads data from a free-field(delimited)



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

text file .We demonstrate this by reading a comma-separated data file with field names in this first record. The figure below shows the beginning of the file(using Notepad).

```
Teaching,Professor_Degree,Preparation,Communication,Class_Contro
l,Teaching_Experience,Approved_Staff
Excellent, ME, Excellent, Excellent, Excellent, TRUE, Yes Excellent,
ME, Good, Good, Excellent, TRUE, No Excellent, ME, Excellent,
Excellent, Excellent, TRUE, No Excellent, ME, Good, Excellent, Good,
TRUE, Yes Excellent, ME, Excellent, Excellent, Excellent, TRUE, Yes
Good, ME, Good, Good, Good, TRUE, No
Satisfactory, ME, Satisfactory, Good, Satisfactory, TRUE, Yes
Satisfactory, ME, Satisfactory, Good, Satisfactory, TRUE, Yes Good,
ME, Excellent, Good, Excellent, FALSE, No
Excellent, ME, Excellent, Excellent, Excellent, TRUE, No Excellent,
PHD, Excellent, Excellent, Excellent, FALSE, No Good, ME, Good,
Good, Satisfactory, TRUE, Yes
Good, ME, Good, Good, Good, TRUE, Yes
Good, ME, Good, Excellent, Excellent, FALSE, No Good, BE, Good,
Satisfactory, Satisfactory, FALSE, No Good, ME, Excellent, Excellent,
Excellent, FALSE, No Good, ME, Good, Good, Good, TRUE, No
Excellent, BE, Good, Excellent, Excellent, TRUE, No Good, ME, Good,
Good, Good, TRUE, No
Excellent, ME, Excellent, Excellent, Excellent, TRUE, Yes Excellent,
BE, Excellent, Excellent, Good, TRUE, No Good, ME, Good,
Satisfactory, Good, TRUE, Yes Excellent, ME, Excellent, Excellent,
Excellent, FALSE, No Excellent, ME, Excellent, Excellent, Excellent,
TRUE, Yes Good, BE, Good, Good, Good, FALSE, No
Good, BE, Good, Good, Good, TRUE, No
Excellent, ME, Excellent, Excellent, Excellent, TRUE, No Excellent,
ME, Excellent, Excellent, Excellent, TRUE, Yes Good, ME, Good,
Good, Poor, FALSE, No
Excellent, ME, Excellent, Good, Excellent, TRUE, Yes Excellent, ME,
Good, Satisfactory, Good, TRUE, Yes Good, BE, Good, Excellent,
Good, FALSE, No
Good, ME, Good, Good, Good, TRUE, No
Good, ME, Good, Satisfactory, Satisfactory, FALSE, No
```

**Figure3.2 Free-field Text File (Sampledata.txt)**

We will read this file into Clementine using Var. File source node. It is better to empty the stream canvas and to start from scratch.

Click the Var. File node in the Sources palette

Position the cursor on the left side of the Stream Canvas and click once. A copy of icon should appear in the Stream Canvas .This source node represents the process of reading a text data file into Clementine. To link this node to a specific file, it needs to be edited.

Right-click on the Var. File node, then click Edit (alternatively, double-click the Var. File node).

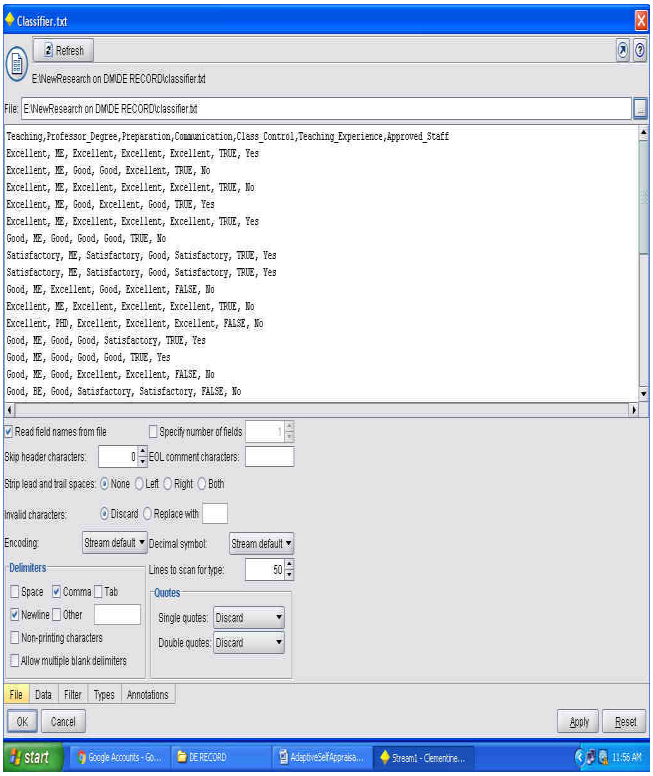


**Fig 3.3 Variable File Node Dialog Box**

The first thing to do is to specify the file name. The file list button is used to browse through directories and to identify the data file.

Click the file list button, and then move to the E:\NewResearchonDM\DE directory. Click Sampledata.txt ,and then click Open.

The Var. File dialog gives a preview of the first lines of data. Note, that the first line of data contains field names. These names can be read directly into Clementine by checking the Read field names from file check box. The characters used to separate the individual fields with in the file are specified under Delimiters. To connect the nodes: Right-click the Var. File node (Sampledata.txt) and select the Connect from the conext pop-up menu, and then click the Table node in the Stream Canvas. An arrow in fig 3.4 appears connecting the source node to the Table node. The direction of the connecting arrow indicates the direction flow, passing from source nodes into Table output node.

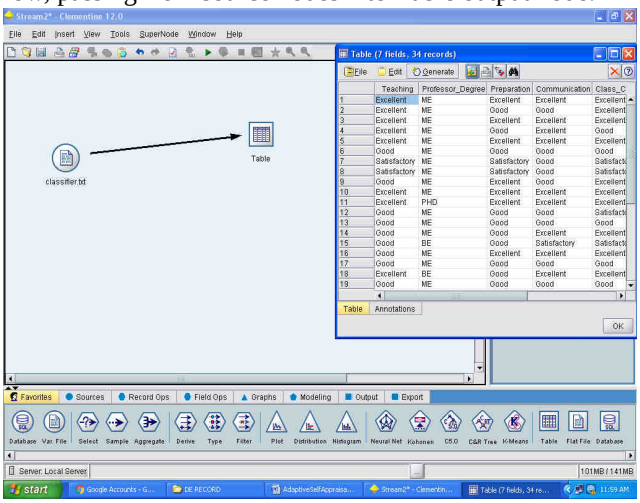


**Fig3.4 File Node Connected to Table Node**



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

The output style can be chosen by editing the Table Node:

Double-click on the Table node

Click the Output tab.

### *d. Choosing a classifier*

The Classification and Regression (C&R) Tree model [8].[6] generates a decision tree to predict or classify future observations. CART builds a binary tree by splitting the records at each node according to a function of a single input field. The measure used to evaluate a potential splitter is diversity. The best splitter is the one that decreases the diversity of the record sets by the greatest. This method uses recursive partitioning to split the training records into segments with similar output field values. The CART tree node starts by examining the input fields to find the best split, measured by the reduction in an impurity index that results from the split. The initial split produces two nodes, each of which is attempted to split in the same manner as the root node. In this way all the input fields are examined to find candidate splitters. If the field only takes on one value, it is eliminated from consideration. The best field for each of the remaining fields is determined. When no split can be found that significantly decreases the diversity of a given node, it is labeled as a leaf node. CART trees gives the option to first grow the tree and then prune based on a cost-complexity algorithm that adjusts the risk estimate based on the number of terminal nodes. This method, which allows the tree to grow large before pruning based on more complex criteria, may result in smaller trees with better cross- validation properties. Increasing the number of terminal nodes generally reduces the risk for the current (training) data, but the actual risk may be higher when the model is generalized to unseen data. To train CART model [9] there should be one or more In fields and exactly one Out field. Target and predictor fields can be range or categorical. Fields set to both or none are ignored. Fields used in the model must have their types fully instantiated, and any ordinal fields used in the model must have numeric storage (not string). If necessary, the Reclassify node can be used to convert them.

### *e. Choosing the experimental procedures*

**STEPS:**

1) Select Text File (Sampledata.txt) from Var. File
2) Select type icon from field option and make any field as Output in given file (here select Approved Staff ) in fig 3.5.
3) Select C&RT and connect it from type (fig 3.5).
4) Then right click on C&RT and select Execute Instruction. (Fig 3.5)

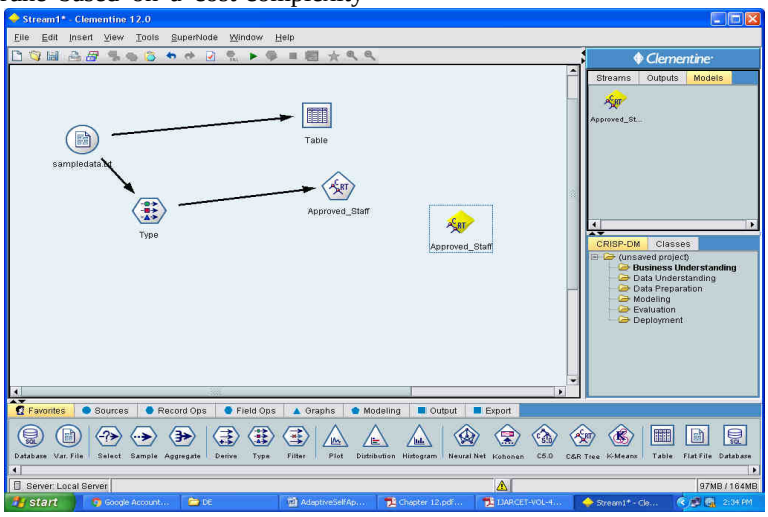


**Fig 3.5 File Node Connected to C&RT Model Node**

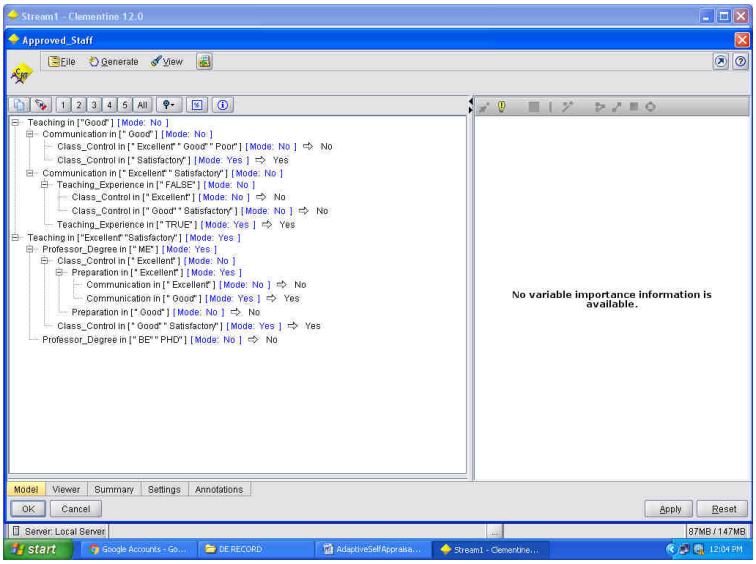


**Fig 3.6 Fully Un Folded TREE**



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

5) It will generate a model.(fig 3.5)
6) Select C&RT model from models pallet.(fig 3.5)
7) Click on C&RT model and click browse, it will display output.(fig 3.6-fig 3.7)

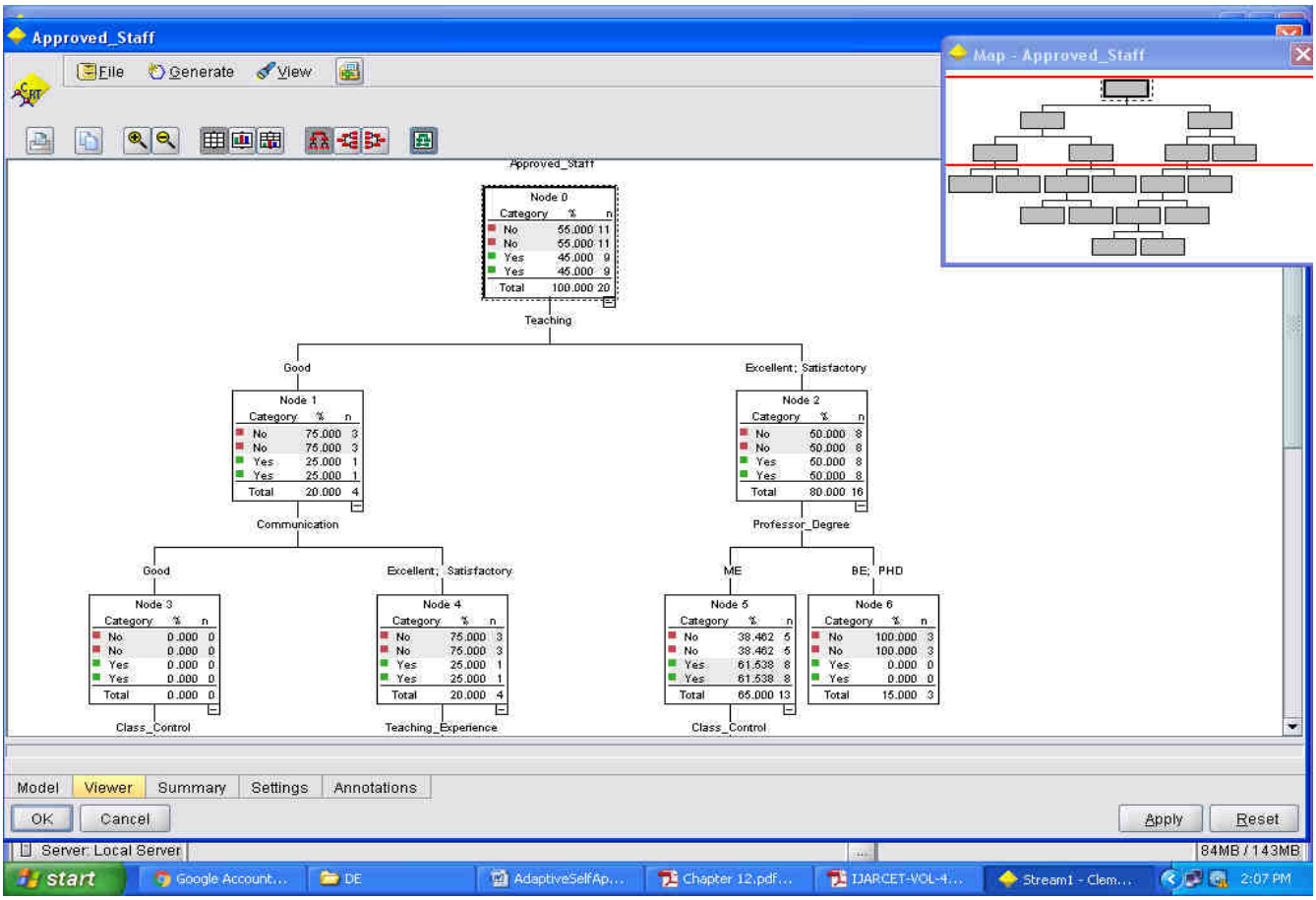


**Fig 3.7a Top Portion of Decision Tree in Viewer Tab**

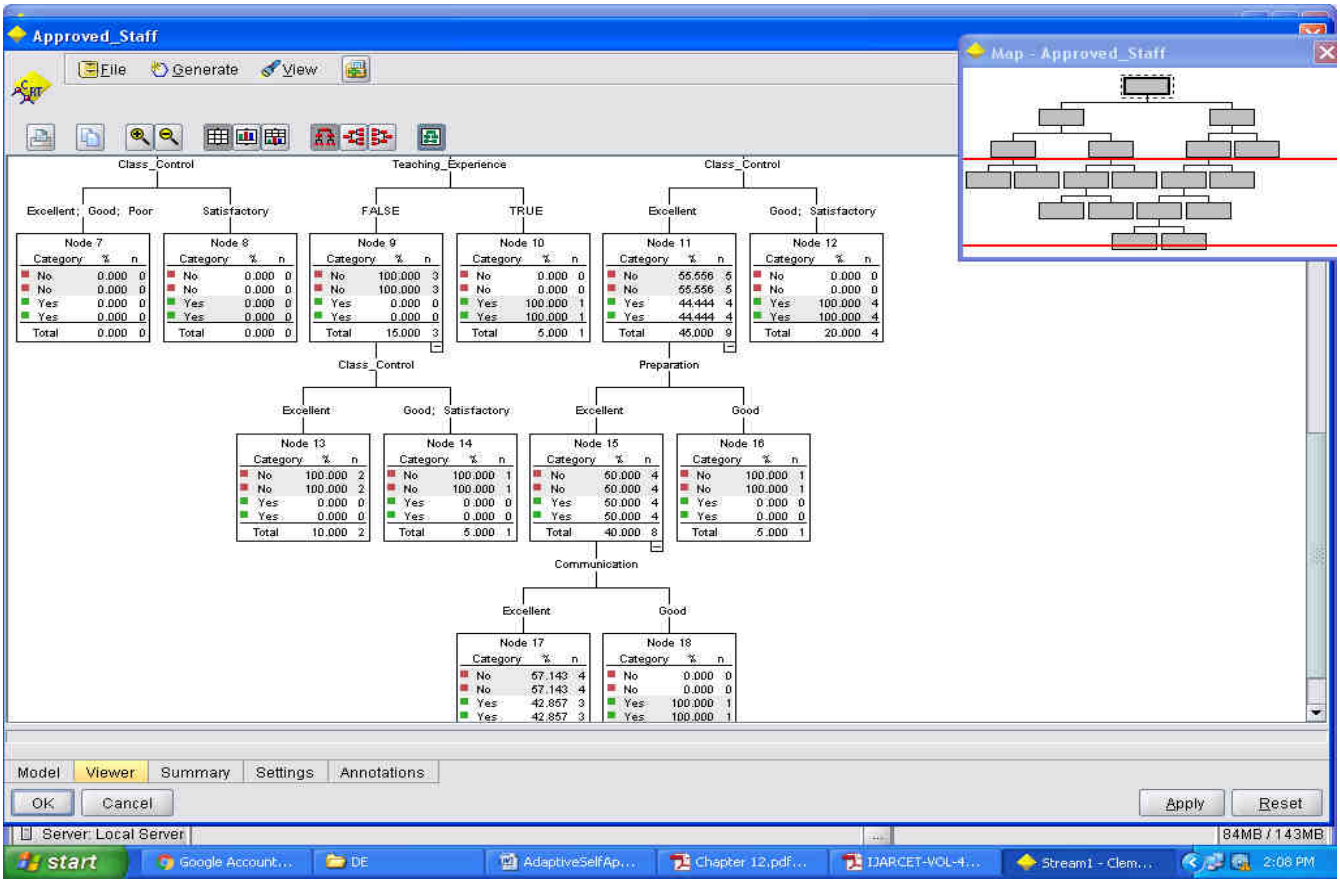


**Fig 3.7b Bottom Portion of Decision Tree in Viewer Tab**



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*

## IV. CONCLUSION

At Professor's appraisal, teaching score of students is very important factor that many colleges/universities gather this information on performance of professor. New rules by using data mining and CRT tree as a decision tree in this paper are results that heads of institutes could use these rules in future decisions to submit new professors and continue with elected old professors. For example is discovered these rules as you see in Fig.3.7

1- IF(Teaching=GOOD)THEN (Approved_Staff is Yes means next semester who continue his/her teaching)

2-IF(Teaching=Excellent)AND(Teaching_experience=FALSE means is low)THEN(Approved_Staff is Yes means next semester who continue his/her teaching)

and etc. Correctness of this rules depending variety of datasets and statistical instances can vary. But data mining tools such as CLEMENTINE TOOL as is showed in this paper can conclude variety results that help head of the departments in Engineering Colleges. These results will be used by HODs in decision-making.

## REFERENCES

1. Jiawei Han. And Micheline Kamber,Jian Pei "Data Mining: Concepts and Techniques", 3rd edition.
2. Sunita B Aher, Mr. LOBO L.M.R.J. Data Mining in Educational System using WEKA, IJCA,2011
3. P. Sanjeev ve J. M, Zytkow. "Discovering Enrollment Knowledge in University Databases," 1th Conference on KDD (Montreal.20-21 August 1995).
4. M.Vranić, D. Pintar, Z.Skoćır, "The Use of Data Mining in Education Environment," ConTEL 2007. (Zagrep 13-15 June 2007), 243
5. Shearer, "The CRISP-DM model: The new blueprint for data mining" Journal of Data Warehousing, (2000). 5: 13-22.
6. http://www.the-datamine.com/Software/SPSS
7. http://homepage.univie.ac.at/marcus.hudec/Lehre/WS%202006/Methoden%20DA/IntroClem.pdf
8. Pushplata Pujari, Jyoti Bala Gupta," Classification of Multi Class Data Set By Using Data mining Classifiers" International Journal of Creative Research Thoughts, 2013.
9. Harleen Kaur and Siri Krishan Wasan "Empirical Studies on applications of Data Mining" Techniques in health care., Journal of computer Science, 2(2), 194-200, 2006, ISSN 1549-3636.

**Ramakrishna Gandi,** received his B.Tech in Computer Science and Engineering from RNEC Ongole, Affl. To JNTUH, India, in June 2006, and M.Tech in Computer Science from ANU, Guntur, India in June 2010.He has 7.5 years of experience in teaching. Currently he is working as an Asst. Professor in Department of CSE at Vignan's Institute of Information Technology, Visakhapatnam (A.P),India. His prime research interest includes Data Ware Housing & Data Mining, Image processing, Data structure. He has 3 International Publications & 2 Professional Certifications excluding this work.

**Prathima Rani Palla,** received her B.Tech in Computer Science & Engineering from Raghu Engineering College, Dakamari, VZM in 2006..M.Tech in Computer Science from Andhra University, VSKP. She has 7 years of experience in Teaching. Currently, She is working as Asst. Professor in Department of CSE at Vignan's Institute of Information Technology, Visakhapatnam (A.P),India. Her prime research includes Data Mining, Network Security.

**Madhuri Thimmapuram** received her B.Tech in Computer Science & Engineering from SAFA College of Engineering & Technology, Kurnool. M.Tech in Computer Science from JNTUA, Ananthapur. She has total 3 Years as experience in Teaching .Currently, She is working as Asst. Professor in Department of CSE at Vignan's Institute of Information Technology, Visakhapatnam(A.P),India Her prime research includes Data Mining, Cloud Computing. She has 1 International Publication & 1 Professional Certificates excluding this work.

**Daniel Prasanth T** received his B.Tech in Information Technology from Gudlavalleru Engineering College, Gudivada. M.Tech in Computer Science from JNTUK in 2014.He has 1 Year as experience in Teaching .Currently, She is working as Asst. Professor in Department of CSE at Vignan's Institute of Information Technology, Visakhapatnam(A.P),India. His prime research includes Data Mining,. Network Security.



*Published By:*
*Blue Eyes Intelligence Engineering*
*& Sciences Publication Pvt. Ltd.*