Volume 3, Issue 10, October 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Admission Management through Data Mining using WEKA

Rakesh Kumar Arora
Dept. of Computer Science,
Krishna Engineering College,
Ghaziabad, UP, India

Dr. Dharmendra Badal
Dept. of Mathematical Science & Computer Applications,
Bundelkhand University,
Jhansi, UP, India

Abstract – Most of the institutions opened during last two decades are in self financing mode. In order to meet the expenditure, lot of effort is being taken by the institutions to have good strength of students. This paper proposes the use of data mining techniques in educational domain to improve the quality of admissions in any higher educational institute. The focus of this paper is to identify those admissions inquires which most likely to turn into actual admissions. The result of analysis will assist the academic planners to focus their efforts on the set of students that are likely to take admission in the institution after initial enquiry.

Keywords- Data Mining, Business Intelligence, WEKA, Data Visualization, Clustering, K–Means.

I. INTRODUCTION

With the rapid opening of educational institutions, admissions in any public universities and educational institutions are likely to face imminent admission crisis now a days. To have good quality of students, the institutes have some standard admission procedures like entrance test, interviews or group discussions but some of the institutes allow direct admission also. This paper uses Educational Data Mining Techniques (EDM) to improve the admission of any institute based on direct admission and try to find out to those admission inquiries which most likely turn into actual admissions. Data mining, the extraction of hidden predictive information from large databases is a powerful technology with great potential to help educational institute focus on the most important information in the data they have collected when admission inquires forms are filled. It discovers information within the data that queries and reports can't effectively reveal. After gathering data from the admission forms filled by students seeking admission collected over years, data mining technique need to be applied to determine set of patterns of students seeking admission in college.

With the help of data mining techniques, such as clustering, decision tree or association analysis it is possible to discover the key characteristics from the admission details of students and possibly use those characteristics for future prediction. This paper presents k-means clustering algorithm as a simple and efficient tool to analyze the admission taken by the students in previous years [1].

II. METHODOLOGY

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative [2]. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori.

The main idea is to define k centroids, one for each cluster. Associate each point belonging to a given data set and to the nearest centroid. After all the points in the data set are over, the first step is completed and an early grouping is done. Re-calculate k new centroids as barycenters of the clusters resulting from the previous step. After k new centroids has been calculated, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop the k centroids change their location step by step until no more changes are done. This algorithm aims at minimizing an objective function, in this case a squared error function. The objective function

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the n data points from their respective cluster centre's.[3]

Algorithmic steps for k-means clustering (4)
Let X = {$x_1,x_2,x_3,\ldots\ldots,x_n$} be the set of data points and V ={$v_1,v_2,\ldots\ldots,v_c$} be the set of centers.
1. Randomly select 'c' cluster centers.
2. Calculate the distance between each data point and cluster centers.
3. Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4. Recalculate the new cluster center using:

$$v_i = \left(\frac{1}{c_i}\right)\sum_{j=1}^{c_i} x_i$$

Where c_i represents the number of data points in i^{th} cluster.

5. Recalculate the distance between each data point and new obtained cluster centers.
6. If no data point was reassigned then stop, otherwise repeat from step 3.

The analysis using k-means clustering is being done with the help of WEKA tool. WEKA, formally called Waikato Environment for Knowledge Learning supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, visualization and feature selection. WEKA is an open source application that is freely available under the GNU general public license agreement. Originally written in C the WEKA application has been completely rewritten in Java and is compatible with almost every computing platform. It is user friendly with a graphical interface that allows for quick set up and operation. WEKA operates on the predication that the user data is available as a flat file or relation, this means that each data object is described by a fixed number of attributes that usually are of a specific type, normal alpha-numeric or numeric values. The WEKA application allows novice users a tool to identify hidden information from database and file systems with simple to use options and visual interfaces. [5]

III. RESULTS

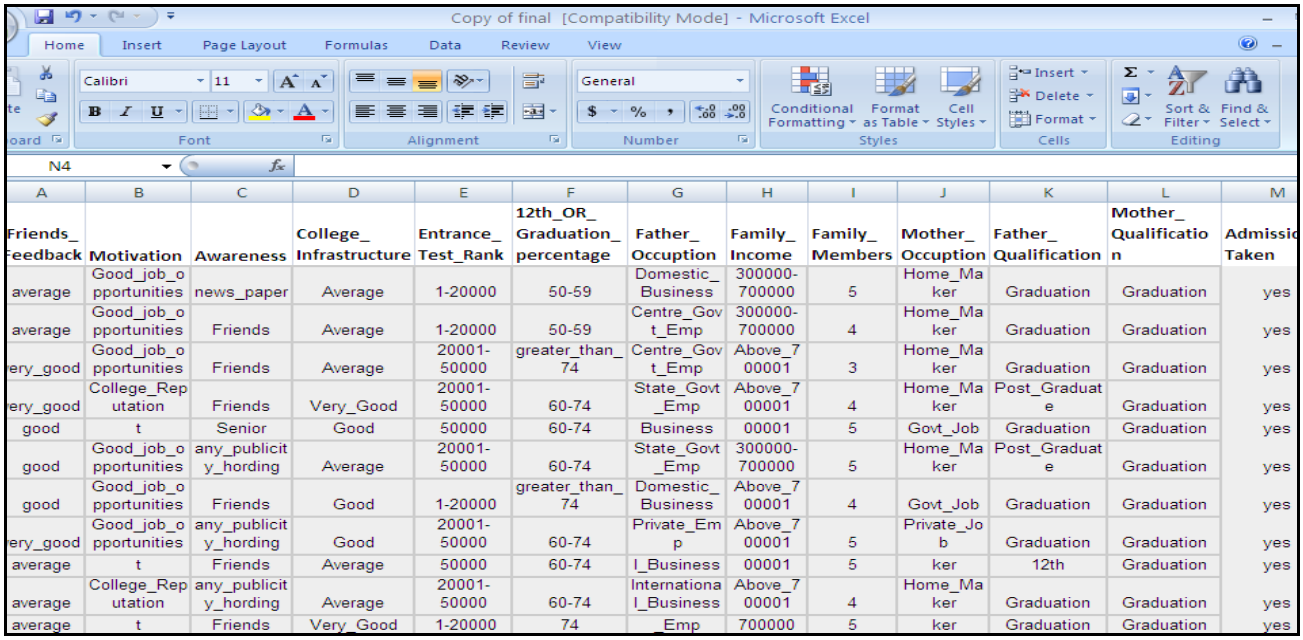
The model was applied on students that have filled up the forms to take admission through management quota in reputed Engineering College of Ghaziabad. The number of students involved in analysis are 129 and parameters analyzed in the paper includes rank of qualifying exam, reason for selecting the desired course , father's occupation, mother's occupation, father's income, father's qualification, mother's qualification and placement of institute. The possible values of parameters considered in this paper are shown in Table I.

TABLE I
PARAMETERS USED FOR ANALYSIS AND THEIR POSSIBLE VALUES

Parameters	Description	Possible values
Friends_Feedback	Feedback from friends regarding Institute	Poor, Average, Good, Very Good
Motivation	Reason for joining the Institute	Good_Job_Opportunities, College_Reputation, Your_Interest, Reference
Awareness	Awareness about Institute	Newspaper, Friends, Senior, Any_publicity_hording, Internet
College_Infrastructure	College Infrastructure	Poor, Average, Good, Very_Good
Entrance_Test_Rank	Rank obtained in entrance test conducted by University	1-20000, 20001-50000, 50001-100000, 100000+,
12th_OR_Graduation_ percentage	Percentage in qualifying exam	50-59, 60-74, greater_than_74,
Family_Income	Family income	300000-700000,Above_700000
Family_Member	No. of members in family	3,4,5
Father_Occuption	Fathers Occupation	Domestic_Business, International_Business Centre_Govt_Emp, State_Govt_Emp, Private_Emp
Mother_Occuption	Mothers Occupation	Home_Maker, Govt_Job, Private_Job
Father_Qualification	Fathers Qualification	12th, Graduation,Post_Graduate, Post_Graduate_Professional
Mother_Qualification	Mothers Qualification	12th,Graduation,12th,Post_Graduate_Professional ,Post_Graduate
Admission_Taken	Admission Taken in institute	Yes, No

The data file normally used by WEKA is in ARFF (Attribute-Relation File Format) file format, which consist of special tags to indicate different things in the data file. Figure 1 shows the sample view of dataset and Figure 2 shows the ARFF format of desired dataset. To convert an Excel format into .arff format an Excel to .arff convertor is being used.

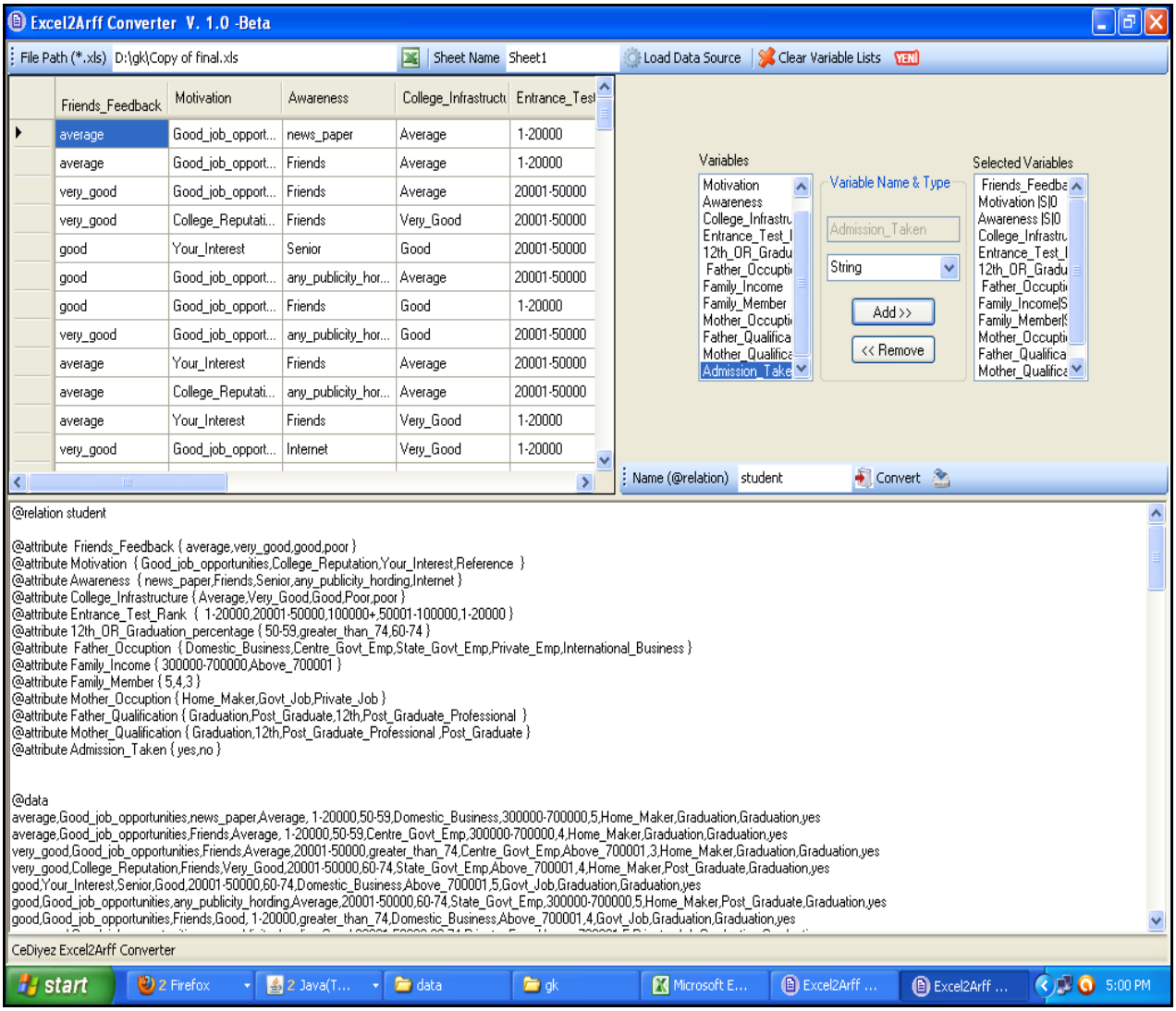
FIGURE 1: SAMPLE DATASET



Copy of final [Compatibility Mode] - Microsoft Excel

Friends_ Feedback	Motivation	Awareness	College_ Infrastructure	Entrance_ Test_Rank	12th_OR_ Graduation_ percentage	Father_ Occupation	Family_ Income	Family_ Members	Mother_ Occupation	Father_ Qualification	Mother_ Qualification	Admissio Taken
average	Good_job_o pportunities	news_paper	Average	1-20000	50-59	Domestic_ Business	300000-700000	5	Home_Ma ker	Graduation	Graduation	yes
average	Good_job_o pportunities	Friends	Average	1-20000	50-59	Centre_Gov t_Emp	300000-700000	4	Home_Ma ker	Graduation	Graduation	yes
ery_good	Good_job_o pportunities	Friends	Average	20001-50000	greater_than_ 74	Centre_Gov t_Emp	Above_700001	3	Home_Ma ker	Graduation	Graduation	yes
ery_good	College_Rep utation	Friends	Very_Good	20001-50000	60-74	State_Govt _Emp	Above_700001	4	Home_Ma ker	Post_Graduat e	Graduation	yes
good	t	Senior	Good	50000	60-74	Business	00001	5	Govt_Job	Graduation	Graduation	yes
good	Good_job_o pportunities	any_publicit y_hording	Average	20001-50000	60-74	State_Govt _Emp	300000-700000	5	Home_Ma ker	Post_Graduat e	Graduation	yes
good	Good_job_o pportunities	Friends	Good	1-20000	greater_than_ 74	Domestic_ Business	Above_700001	4	Govt_Job	Graduation	Graduation	yes
ery_good	Good_job_o pportunities	any_publicit y_hording	Good	20001-50000	60-74	Private_Em p	Above_700001	5	Private_Jo b	Graduation	Graduation	yes
average	t	Friends	Average	50000	60-74	l_Business	00001	5	ker	12th	Graduation	yes
average	College_Rep utation	any_publicit y_hording	Average	20001-50000	60-74	Internationa l_Business	Above_700001	4	Home_Ma ker	Graduation	Graduation	yes
average	t	Friends	Very_Good	1-20000	74	_Emp	700000	5	ker	Graduation	Graduation	yes

FIGURE 2: .ARFF FORMAT OF DATASET



Excel2Arff Converter V. 1.0 -Beta

File Path (*.xls) D:\gk\Copy of final.xls | Sheet Name Sheet1 | Load Data Source | Clear Variable Lists

Friends_Feedback	Motivation	Awareness	College_Infrastructu	Entrance_Test
average	Good_job_opport...	news_paper	Average	1-20000
average	Good_job_opport...	Friends	Average	1-20000
very_good	Good_job_opport...	Friends	Average	20001-50000
very_good	College_Reputati...	Friends	Very_Good	20001-50000
good	Your_Interest	Senior	Good	20001-50000
good	Good_job_opport...	any_publicity_hor...	Average	20001-50000
good	Good_job_opport...	Friends	Good	1-20000
very_good	Good_job_opport...	any_publicity_hor...	Good	20001-50000
average	Your_Interest	Friends	Average	20001-50000
average	College_Reputati...	any_publicity_hor...	Average	20001-50000
average	Your_Interest	Friends	Very_Good	1-20000
very_good	Good_job_opport...	Internet	Very_Good	1-20000

Variables: Motivation, Awareness, College_Infrastru, Entrance_Test_I, 12th_OR_Gradu, Father_Occupti, Family_Income, Family_Member, Mother_Occupti, Father_Qualifica, Mother_Qualifica, Admission_Take

Variable Name & Type: Admission_Taken | String | Add >> | << Remove

Selected Variables: Friends_Feedba, Motivation |S|0, Awareness |S|0, College_Infrastru, Entrance_Test_I, 12th_OR_Gradu, Father_Occupti, Family_Income|S, Family_Member|S, Mother_Occupti, Father_Qualifica, Mother_Qualifica

Name (@relation) student | Convert

```
@relation student

@attribute Friends_Feedback { average,very_good,good,poor }
@attribute Motivation { Good_job_opportunities,College_Reputation,Your_Interest,Reference }
@attribute Awareness { news_paper,Friends,Senior,any_publicity_hording,Internet }
@attribute College_Infrastructure { Average,Very_Good,Good,Poor,poor }
@attribute Entrance_Test_Rank { 1-20000,20001-50000,100000+,50001-100000,1-20000 }
@attribute 12th_OR_Graduation_percentage { 50-59,greater_than_74,60-74 }
@attribute Father_Occupation { Domestic_Business,Centre_Govt_Emp,State_Govt_Emp,Private_Emp,International_Business }
@attribute Family_Income { 300000-700000,Above_700001 }
@attribute Family_Member { 5,4,3 }
@attribute Mother_Occupation { Home_Maker,Govt_Job,Private_Job }
@attribute Father_Qualification { Graduation,Post_Graduate,12th,Post_Graduate_Professional }
@attribute Mother_Qualification { Graduation,12th,Post_Graduate_Professional ,Post_Graduate }
@attribute Admission_Taken { yes,no }


@data
average,Good_job_opportunities,news_paper,Average, 1-20000,50-59,Domestic_Business,300000-700000,5,Home_Maker,Graduation,Graduation,yes
average,Good_job_opportunities,Friends,Average, 1-20000,50-59,Centre_Govt_Emp,300000-700000,4,Home_Maker,Graduation,Graduation,yes
very_good,Good_job_opportunities,Friends,Average,20001-50000,greater_than_74,Centre_Govt_Emp,Above_700001,3,Home_Maker,Graduation,Graduation,yes
very_good,College_Reputation,Friends,Very_Good,20001-50000,60-74,State_Govt_Emp,Above_700001,4,Home_Maker,Post_Graduate,Graduation,yes
good,Your_Interest,Senior,Good,20001-50000,60-74,Domestic_Business,Above_700001,5,Govt_Job,Graduation,Graduation,yes
good,Good_job_opportunities,any_publicity_hording,Average,20001-50000,60-74,State_Govt_Emp,300000-700000,5,Home_Maker,Post_Graduate,Graduation,yes
good,Good_job_opportunities,Friends,Good, 1-20000,greater_than_74,Domestic_Business,Above_700001,4,Govt_Job,Graduation,Graduation,yes
```

CeDiyez Excel2Arff Converter

start | 2 Firefox | 2 Java(T... | data | gk | Microsoft E... | Excel2Arff ... | Excel2Arff ... | 5:00 PM

FIGURE 3: ANALYSIS OUTPUT

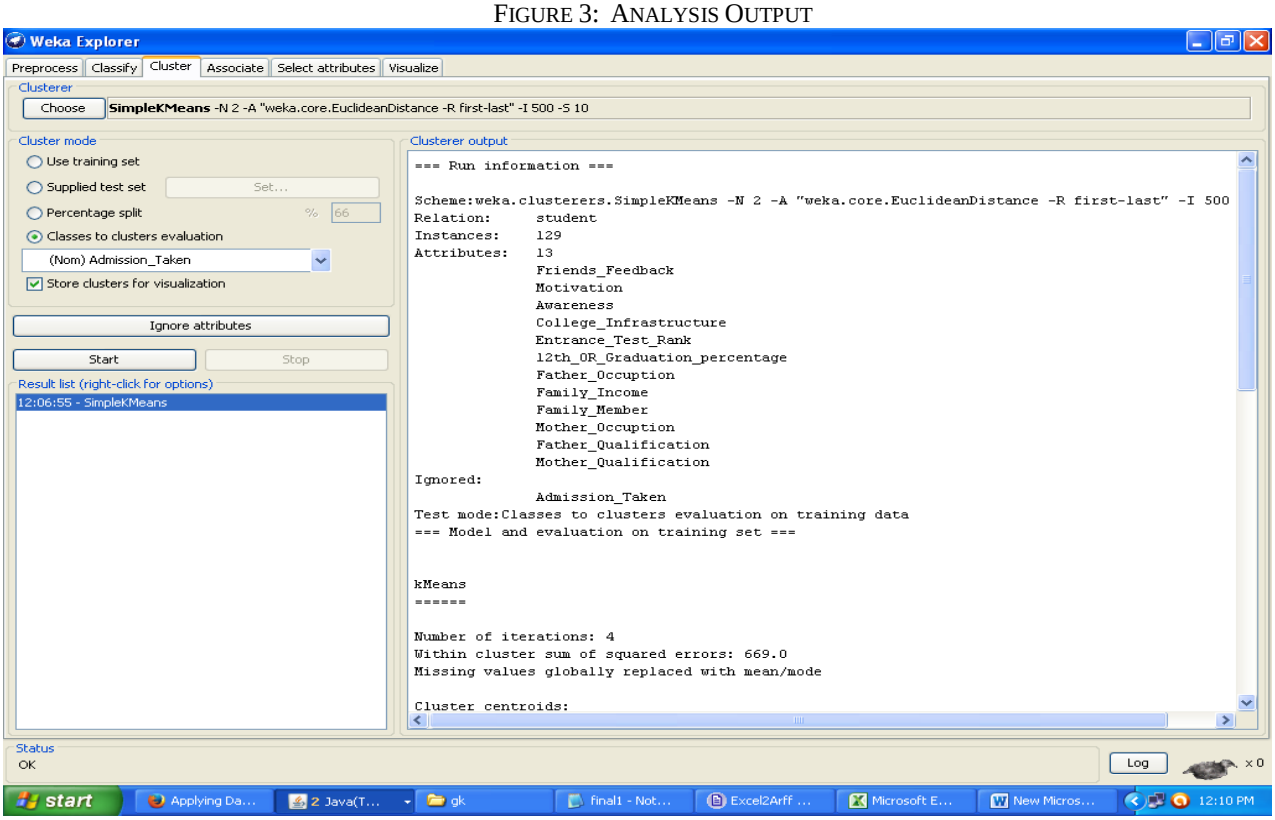


FIGURE 4: DETAILED DESCRIPTION OF FIGURE 3

Number of iterations: 4
Within cluster sum of squared errors: 669.0
Missing values globally replaced with mean/mode
Cluster centroids:

	Cluster#		
Attribute	Full Data (129)	Cluster 0 (68)	Cluster 1 (61)
Friends_Feedback	Good	Average	Good
Motivation	Good_job_ opportunities	College_Reputation	Good_job_ opportunities
Awareness	Friends	Friends	Friends
College_Infrastructure	Good	Good	Good
Entrance_Test_Rank	50001-100000	50001-100000	20001-50000
12th_OR_Graduation_percentage	50-59	50-59	Greater_than_74
Father_Occuption	Private_Emp	Domestic_Business	Private_Emp
Mother_Occuption	Home_Maker	Home_Maker	Home_Maker
Father_Qualification	Graduation	Graduation	Graduation
Mother_Qualification	Graduation	Graduation	Graduation
Family_Member	5	4	5
Family_Income	Above_700000	Above_700000	Above_700000

Time taken to build model (full training data) : 0.02 seconds
=== Model and evaluation on training set ===
Clustered Instances
0 68 (53%)
1 61 (47%)
Class attribute: Admission_Taken
Classes to Clusters:
0 1 <-- assigned to cluster
57 51 | yes
11 10 | no
Cluster 0 <-- yes
Cluster 1 <-- no
Incorrectly clustered instances : 62.0 48.062 %

The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. K-Means requires three parameters: number of clusters (k), cluster initialization and distance metric. The most critical choice is number of clusters (k) .The result after applying K Means algorithms on the data set using WEKA shows in Figure 3 and Figure 4. Figure 4 is a detailed description of Figure 3. Figure 4 shows the existence of two clusters (Cluster 0 and Cluster1) having 68 and 61 admission inquires respectively. It has been clearly indicated that at cluster 0, 57 out of 61 inquires are being converted into admission while at cluster 1, 51 out of 61 inquires are converted to admission. Although the success ratio of both clusters (83.82 and 83.60) have a minor difference but there is need to focus more on cluster 1 to further boost admission chances in institute.

IV. CONCLUSION

In this paper, a simple methodology based on k-means clustering algorithm is being used to analyze the data obtained from the admission form filled by admission seekers. This methodology will assist the academic planners to monitor admission details of students seeking admission in institute over the years. Hence this model will play important role in determining the reasons for decline in quality of admissions taken in the institute over the year and steps that need to be taken to improve performance from next academic session. This model will help in identifying the set of students that need to be focused to actually convert the inquiry into admission.

REFERENCES

[1] Arora K. Rakesh, Badal Dharmendra," Evaluating Student's Performance Using k-Means Clustering", IJCST Vol. 4, Issue 2, April - June 2013

[2][Online]Available:http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K_Means_ Clustering_Overview.htm

[3][Online]Available:http://www.home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[4][Online] Available: https://sites.google.com/site/dataclustering algorithms /k-means-clustering-algorithm

[5] [Online] Available: http://www.gtbit.org/downloads/dwdmsem6/dwdmsem6lman.pdf

[6] Romero, C., Ventura, S., Espejo, P.G., Hervas, C. (2008) Data Mining Algorithms to Classify Students. Proceedings of the First International Conference on Educational Data Mining, 8-17

[7] N.V.Anand Kumar Research Scholar, Department of Computer Science and engineering Anna university, Chennai G.V.Uma Assistant professor, Department of Computer Science and Engineering Anna university, Chennai "Improving Academic Performance of Student By Appling Data Mining Techniques

[8] Arora K Rakesh, Badal Dharmendra, "Location wise student admission analysis", International Journal of Computer Science, Information Technology and Security, Dec 2012.

[9] Arora K. Rakesh, Gupta K. Manoj ,"Data Mining: Scope Out Valuable Resources From Mountains Of Information", IITM Buisness Review Journal, July 10

\

