IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 64-68
www.iosrjournals.org

Mining High Economic Status of Judokas through Clustering, KNN and Decision Trees

Reena Hooda[1]
[1](Dept. of Computer Sc. & Applications, Indira Gandhi University,Meerpur, Rewari, Haryana, India)

ABSTRACT: *Data mining turned into astimulating, operative, and widely acceptable methodology for analyzing central repository from different perspective even on the same data andattaining knowledge while elaborating various insights through the applicability of obtainable mining tools. The advantage of using these refined and efficient tools are that besides the required information, they aid in highlighting the same data sets in different modes to provide additionalinvestigation. The present paper underlining the implication of one of the finest tools of data mining for examining the data of 164 Judokas and to demonstrate various relationships between different entities, their grouping and pictographic representation of the classification through the k-means clustering, k-nearest neighbor method and decision tree classification.*
Keywords**:** *Clustering, Mining, Decision tree.*

I. INTRODUCTION

Data mining aims at providing the knowledge about the data in terms of their classification, summary, pictorial representation of data sets, hidden facts, prediction about some unknown values, pattern recognition, data distribution etc. The present paper attempts to enhance the received results of 164 Judokas [6], flow of data, interrelationships, and hierarchical distribution of data and partitions. These 164 students belong to a set of 300 students who gathered at Punjab University for Judo Championship. The work had been highlighting the relationship of achievement motivation and social status factors withthe position winning& participations of 360 Judokas [6]. The study had been conducted through considering the six factors such as social change, liberalism, social distance, nationalism, social revolution and untouchability to find the overall significance of these factors over the position winning, attitude and motivation[6]. The current paper signifying a more expressive examination by taking this database as an example data set while considering High Economic status containing 164 data sets in it and will be treated as inputs to various operations like modeling and classification, the tools used for this analysis is the RapidMiner 5.3.013[2].

II. EVALUATION AND IMPLEMENTATION

The very first step of the analysis is to create a local repository by importing the Excel sheet and transforms the attributes types, setting Id and Label field so as to retrieve the example repository into required format. Fig. 1 shows the database retrieval by importing Excel sheets and Fig 2 shows the retrieved example dataset. The repository is stored under the head Local Repository. Next step is to use the decision tree operator to generate a pictorial classification of the example dataset. This has been done through selecting this operator from the left plane under the head Modeling, subhead Classification & Regression [4] then drag & drop it on the process area. The operator shows a problem here as it is mandatory to assign a retrieve object to it that will be assigned to it later. Retrieve operator [2][4]contains the example dataset. Filtration is required to handle the missing values as well as duplicate values in example database to get the final tree classification. The decision tree is a descriptive and unsupervised methodology and even can work for non-numerical values have the biggest advantage to select it for detailed classification and ease of use. Fig.3 shows the decision tree classification.

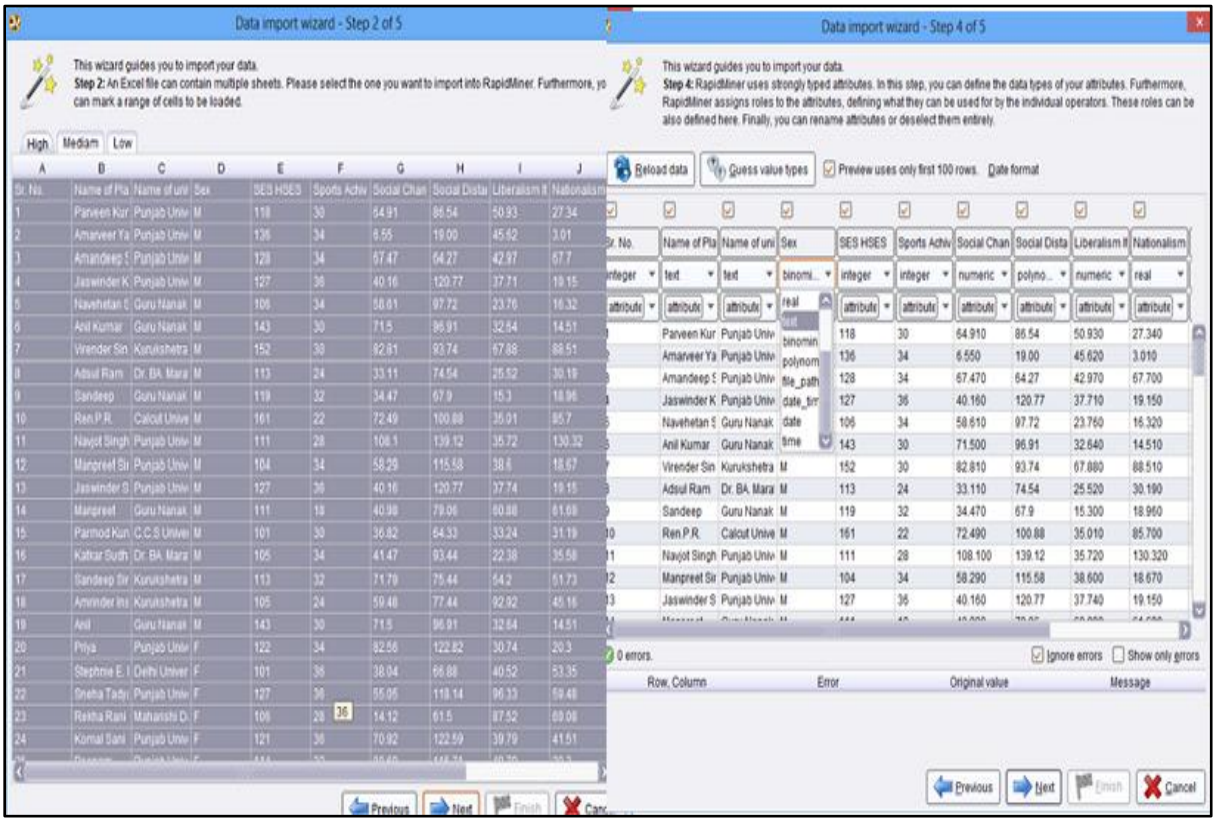

Fig.1. Shows the Database retrieval by importing Excel sheets and setting attributes types.

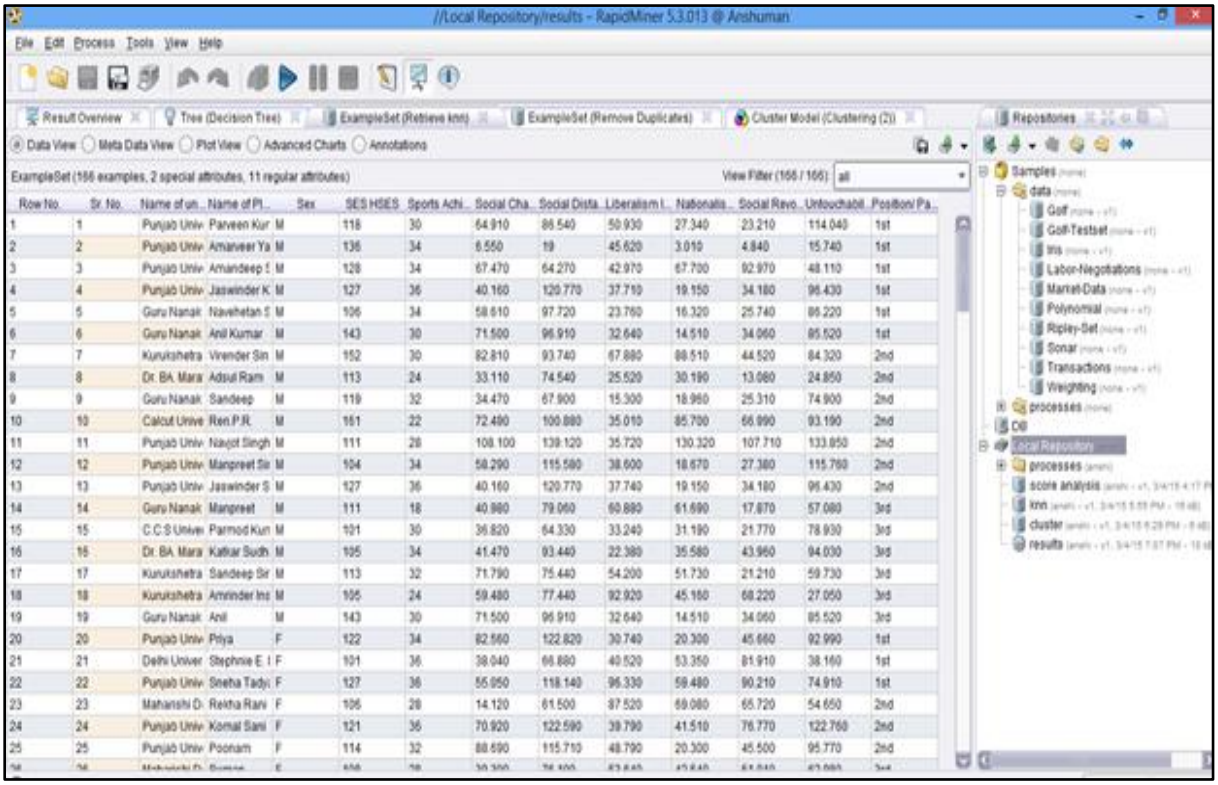

Fig.2. Shows the Retrieved Example dataset.

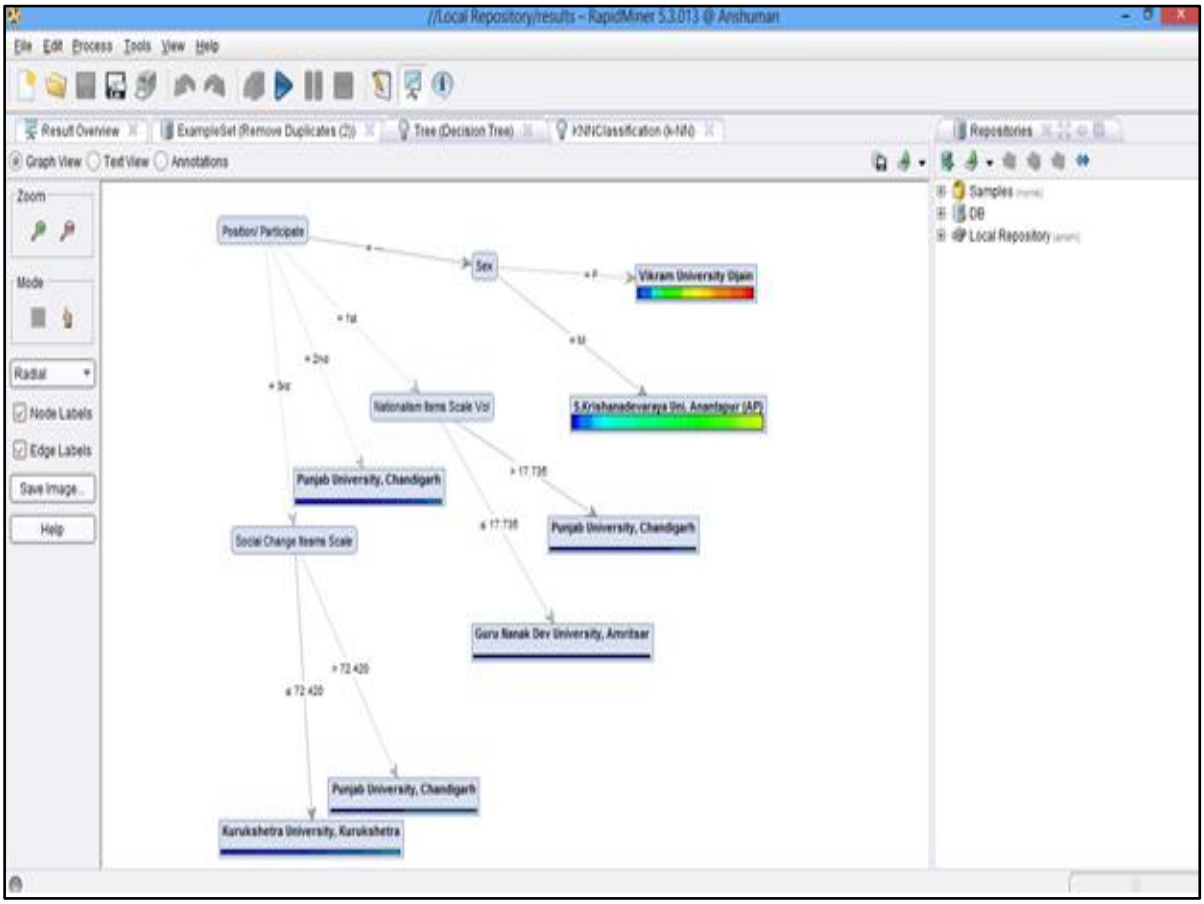

Fig. 3. Shows Decision Tree Classification.

To show the classification in a tabular form under the control of the user, clustering operator has been used that applies the k-means method for partitioning. The limitation of this operator is that it works only on non-nominal values that mean if the input database has the polynomial values etc. other than numerical, this operator shows an error. One option is to input the database excluding polynomials or convert polynomials to real or integer, another option is to use a select operator to filter only the selected attributes that has numeric types. This is the fundamental difference between decision tress classifications and clustering that decision tree considers all kind of attributes including nominal values. However, the advantage of clustering over decision tree is that it is supervised methodology whereas decision tree is unsupervised technique. The values of k in k-means and kNN is user defined i.e. set at 5 in k-means ranges from 0 to 4. The maximum runs value is set at 10 that is a default value and also be set by the user. This way, out of 164 Judokas, 5 clusters have been created containing 45, 38, 10, 23 and 48 items respectively. But if an option Good Start Value has also been considered, then the items will be 37, 10, 16, 48 and 53 in different clusters. User can also tick Add Cluster Attribute as additional attribute in result set to show the different entities belongs to which particular cluster. This attribute can also be added as a Label[1][3][5] in place of cluster in the result dataset. Further advantage of this clustering k-means operator is that the output can be further targeted to a separate file other than the results set through the use of Write Cluster operator. This Write operator can only work with the cluster model output so it is not suitable for kNN as kNN is a predication operator. The output of kNN is direcly linked to the result node. Fig.4, Fig.5 and Fig.6shows the different efficaciesk-means operator offers, such as graph view and text views, centroid table, centroid plot view and even clustered result set. It seems that kNN prediction operator doesn't fit for the classification as compared to k-means and decision trees as this operator is used for predicting a target and takes training set as well as example set to get the output and do not show the clear picture of classification.

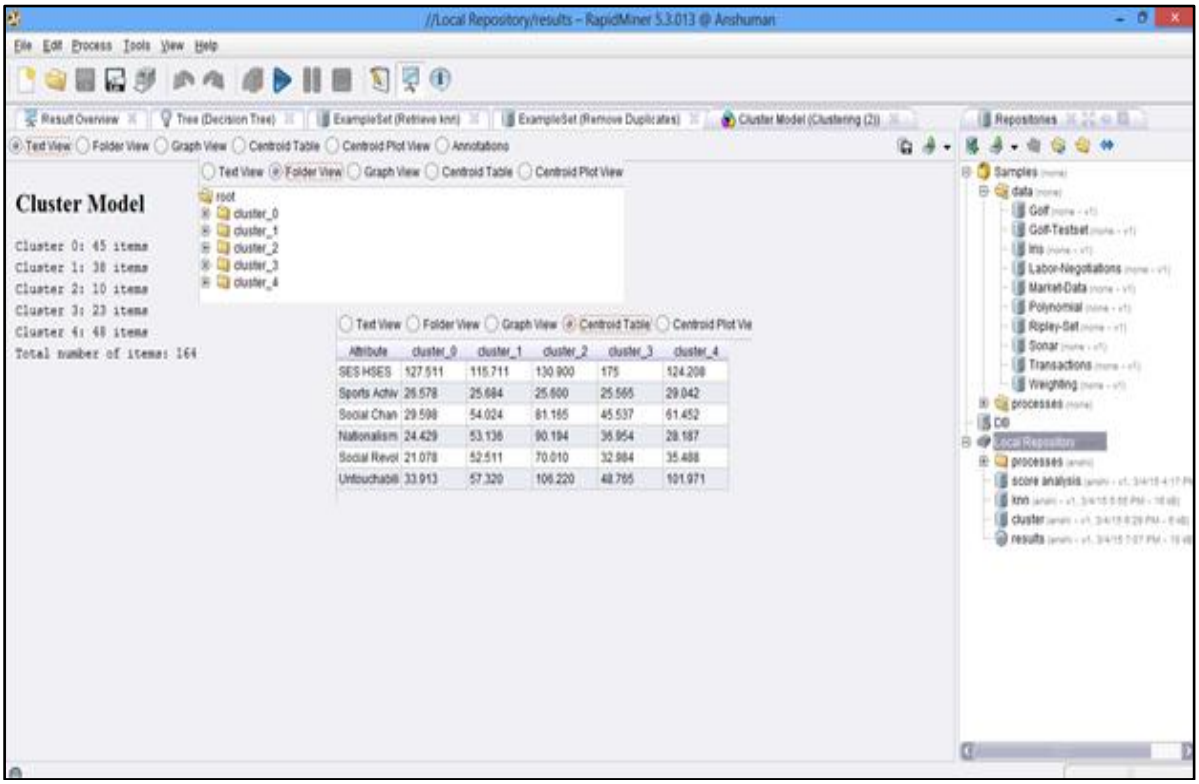

Fig.4. Shows Text View and Centroid View.

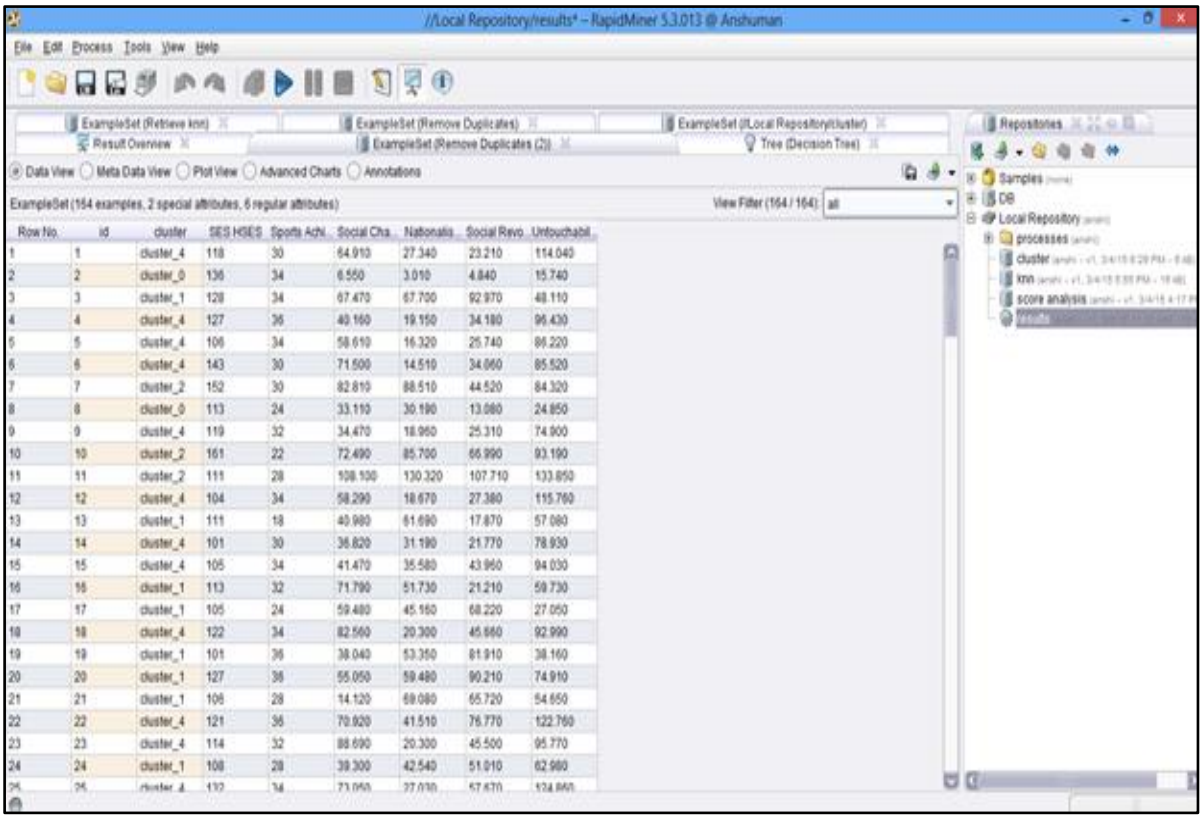

Fig.5. Shows Result file containing an additional Cluster attribute.

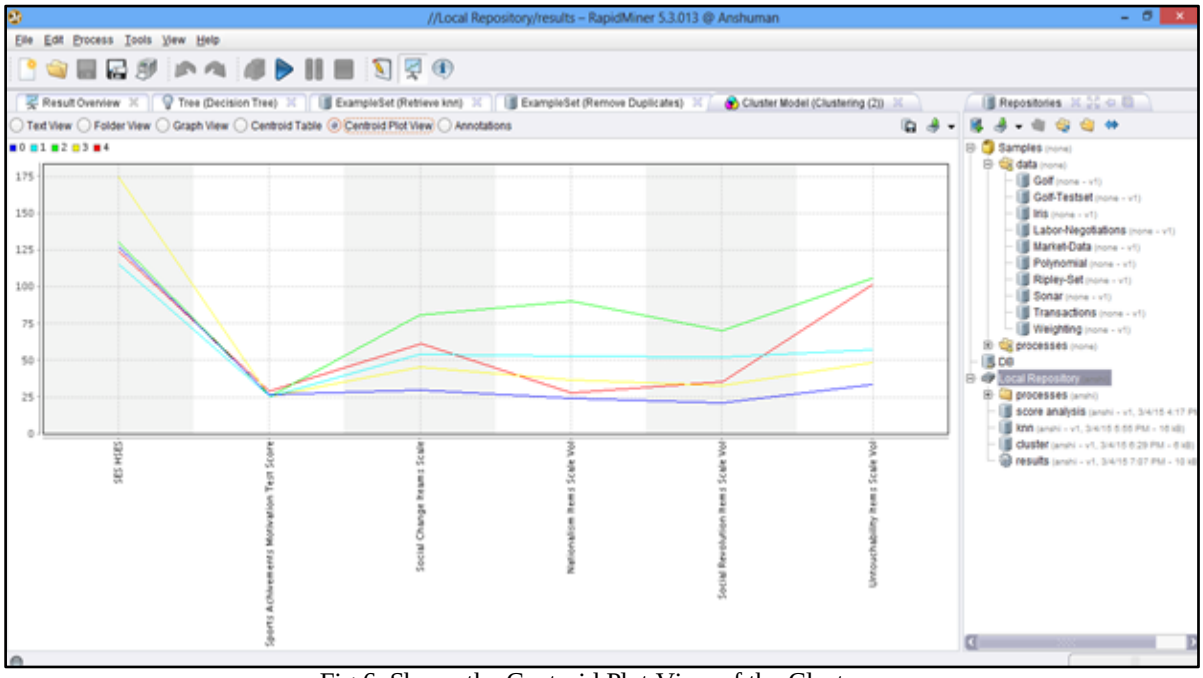

Fig.6. Shows the Centroid Plot View of the Clusters.

III. CONCLUSIONS AND FURTHER SCOPE OF WORK

The present paper demonstrated the significance of mining techniques for representing an acute and meticulous investigation of data items in tabular as well as pictorial forms through inputting the primary data of 164 Judokas to show detailed associations among the six given attributes and effectiveness of achievement motivation and High Economic Status over the position winning and partitioning of Judokas. The further scope of the work may include the enrichment of the mining methodologies and even to work with other operators and tools.

REFERENCES

[1] https://rapidminer.com/wp-content/uploads/2013/10/RapidMiner_RapidMinerInAcademicUse_en.pdf.
[2] https://rapidminer.com/
[3] http://sourceforge.net/projects/rapidminer/
[4] https://rapidminer.com/wp-content/uploads/2014/10/RapidMiner-v6-user-manual.pdf
[5] http://www.slideshare.net/dataminingtools/rapidminer-introduction-to-datamining
[6] D. Dhaka, *Relationship of Motivation and Attitude with Socio Economic Status and Performance of Judokas, Department of Physical Education,*MaharshiDayanand University, Rohtak. 2011.