

# Enhanced Data Mining Techniques for Improving the Women Entrepreneurs Quality of Service

C.Grace Padma, Dr. Arunkumar Thangavelu

*Research Scholar, Manonmaniam Sundaranar University, Tirunelveli.  
Professor & Assistant Director (Centre for Ambient Intelligence), VIT University, Vellore.  
[gracepadma.phd@gmail.com](mailto:gracepadma.phd@gmail.com)*

## Abstract

Recent years, the dataflow in higher education field was generally straightforward and the application of innovation was restricted. Therefore, advance into a more integrated world where innovation has turned into a vital part of the business forms. Higher education data mining is a particular data mining field connected to data starting from educational environment, it depends on distinctive ways to find hidden knowledge from the available data. Data mining approach can be employed to solve various classification, regression, optimization and clustering problems. In this paper proposing a new data mining technique applied in quality function of higher education. This data mining knowledge is used to predict the understanding hidden pattern and uncover data in huge databases to suggest the way to improve the women entrepreneurs depends on higher education filed to improve the Women Entrepreneurs quality of service. Characterize the relationship between higher education development indices and different stages of women entrepreneurial activity through a data mining prediction analysis.

**Keywords:** Data Mining, Higher education data mining, Women Entrepreneurs, quality of service.

## 1. Introduction

The development of information technology in different fields has led to the huge volumes of data storage in different arrangements like records, documents, sound, image, scientific data, video and numerous new data format. The data gathered from distinctive applications require appropriate technique for extracting information from huge repositories for better decision making. Knowledge Discovery in Databases (KDD), often named as data mining, has the main goal for the discovery of valuable information from huge volume of data. The main capacities of data mining are applying different techniques and algorithm keeping in mind the end goal to find and extract pattern of available data [1]. Knowledge discovery and data mining applications have got much attention because of its importance in decision making and it becomes a vital component in different organization [9]. Data mining strategies have been introduced into new fields such as Databases, Artificial Intelligence, Statistics, Pattern Reorganization, Computation capabilities, Machine Learning and so on.

The data mining approaches are utilized in the increasing research interests of higher education. This new emerging

field, called Higher Educational Data Mining, concerns with developing routines that find knowledge from data capable of detecting unknown data. Higher education data mining uses numerous methods, for example, Neural Networks, Decision Trees, K- Nearest neighbor, Naïve Bayes and etc. Utilizing these methods numerous kinds of knowledge can be discovered, for example, classification association rules and clustering [10] [11]. The discovered knowledge can be utilized for prediction regarding how to bring noble laureates, how to become CEO, how to bring out Women Entrepreneurs, innovation, quality through accreditation assessments, massive open online courses etc.

The main objective of this paper is to utilizing data mining approach to predict the knowledge of higher education student success to become an entrepreneur especially for women. Data mining gives numerous tasks that could be utilized to study the student performance. The educational procedures and participants, by fulfilling their commitments through proper activities, make a tremendous amount of data which needs to be gathered, integrated and used in a useful way. By converting this information into knowledge, the delight of all members is attained: students, administrators, professors, social community and supporting administration. Since the data mining characterized the knowledge based data process from various perspectives, with the aim of extracting the interesting and implicit samples, information and trends from the data, it can significantly help to predict the knowledge of students are becoming highly educated and a good entrepreneur, in order to enhance the understanding of higher education.

## 2. Related work

Educational data mining is process of developing new methods that extract the needed data from the educational database. Since the education is significant for the development of the country, mining the education database and extract knowledge from the database support the educational system to over the issues in the knowledge discovery. The paper [2] suggests an educational data mining system that makes use of the K-means clustering algorithm. Using this mining process, the evaluation of students' academic reports is assessed and minimizes the drop out ratio. In future, the techniques used here will be enhanced and brings good quality outputs to enhance the students' performance.

The author [3] introduces a Student Advisory Framework which used to develop the intelligent system using clustering and classification. This classification and the clustering process is possible using the machine learning techniques. This system provides a way for a student to succeed in his/her academics by tracking the education details and minimize the academic failure. The paper used a dataset which has first year course results dataset and the pre-higher education dataset. The results of this work prove its better performance.

The data mining tools are widely used in the field of education and determine the performance of the students. The paper [4] reviewed the issues met by the educational institutions. It attempts to extract the knowledge and the performance of the students and also organize data utilizing weka data mining tool which supports to identify the data. It helps educator to act before drops a student for recourse allocation with certainty gained from knowing what number of students are likely failed or pass.

One of the biggest challenges in higher education is enhancing the student retention. Student retention characterized as two types of process such that enrolment management and academic performance. The paper [5] monitors the student performance using language processing and data mining technology. It help to identify students "at risk", the main objective of this paper is identifying the potential problems of students quickly with intervention options to improve the student retention.

Education Data Mining (EDM) is a developing research area that deals with enhancing of approach to explore data creating in an educational context. The paper [6] surveys the studies of EDM using Apriori algorithm. An improved support-matrix based Apriori algorithm was used in this research along with bottom up approach and standard deviation functional model for educational data pattern mining.

The author [7] presents an overview of EDM and EDM development which is used by different stakeholders. In EDM various methods and techniques for examining data formulate from different educational information approach can be progressed. It is popular application area in learning science because of the educational data growing availability. This research can be used to predict the how student can learn and also leaning education settings, it will make use of data driven decision making process and enhance the present educational process.

### 3. Proposed work

#### A. Women Entrepreneurship Overview

Entrepreneurs are significant for economies and national work market in that they denotes importantly to economic development additionally they give the majority of the employments and provide new business opportunities. According to the Global Entrepreneurship Monitor's "Report on Women and Entrepreneurship [8]" are interested in women's entrepreneurship which is an essential approach to exponentially increase the effect of new venture creation. Predicting approach to empower women's contribution and success in entrepreneurship are discriminating based on more economical and successful economic improvement.

The main target is subsequently to give a prediction analysis of the key parameters and perspective of female entrepreneurial activity and how that may contrast from of their male partners. The second aim of this proposed approach is to give an extensive analysis of the participated women's role in entrepreneurial activity in India based on higher education standard. Additionally, advance the understanding the requirements of the existing and aspiring women entrepreneurs, and will help to give some strategy knowledge helpful to create and improve a positive domain for women's entrepreneurship. The primary theories are that the male and female entrepreneurs contrast as for their individual and business profile: they initiate and run the businesses in various sectors, make different types of product and structure their organization and business in a different fashion.

#### B. Data Description

The information for the model was gathered through a survey conducted during the semester at the college, academic year 2013-2014, among the final year students. Incomplete data are eliminated from the survey list. The samples included 300 higher educated students information. The model of student's prosperity was made, where output measured with achievement in the course "Business Informatics" and "Economic".

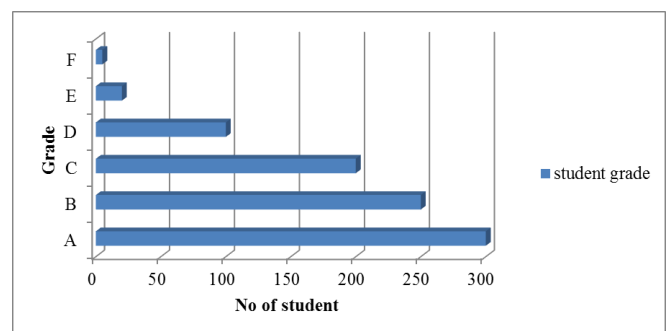


Figure 1 Distribution of grades in the course "Business Informatics" and "Economic"

Through the six classes labels are A, B, C, D, E and F and the same as students' final grades, as shown in figure 1. Grade A- Very important, Grade B- Important, Grade C- Somewhat important, Grade D- not important, and Grade E -Not important at all, Grade F- not considers. The proposed system is to predict the scope of the particular higher study course and recommend the student to take that course. This prediction is based on the utilizing the profile of the students who had taken that particular course.

#### C. C4.5 algorithm

Figure 1 shows the architecture of the proposed system in the factors of the enrolment framework. Initially, the proposed system utilizes the information of the historical database of the higher study students and the outcomes obtained by them, with the aim of acquiring C4.5 algorithm rules. In this module, the C4.5 algorithm utilized for filtering the training data based on generating rules.

The data cleaning is the preprocessing step in data mining process to eliminate the irrelevant data. The discovery of patterns will be helpful if the relevant data characterized in real academic performance and the decision and /or action taken by the student in the earlier stage.

Initially, the University provided a database of 2000 records corresponding to 300 students. After filtering process of the 1500 records were left. The data delivered is only from student at the school of Business Informatics” and “Economic”, enrolled over the year 2013-2014,

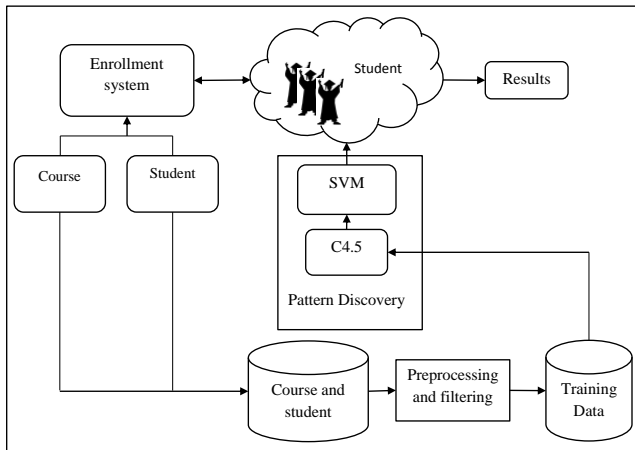


Figure 1 Propose System Frame work

The discovering pattern that utilized to improve the women entrepreneurs depends on higher education filed. C4.5 algorithm is based on the information gain ratio that is evaluated by entropy. The information gain ratio measure is used to select the test features at each node in the tree. Such a measure is referred to as a feature (attribute) selection measure. The attribute with the highest information gain ratio is chosen as the test feature for the current node.

Let  $D$  be a set consisting of  $(D_1, \dots, D_j)$  data instances. Suppose the class label attribute has  $m$  distinct values defining  $m$  distinct classes  $C_i$  (for  $i = 1, \dots, m$ ). Let  $D_j$  be the number of samples in  $D$ . The expected information needed to classify by using below equations.

$$Splitinfo_A(D) = - \sum (|D_j|/|D|) * \log(|D_j|/|D|) \quad (1)$$

$$Gain\ ratio(A) = Gain(A) / Splitinfo_A(D) \quad (2)$$

$$info(D) = - \sum P_i \log_2(P_i)$$

and

$$info_A(D) = - \sum (|D_j|/|D|) * info(D_j) \quad (3)$$

Where

$$Gain = info(D) - Info_A(D) \quad (4)$$

Where  $P_i$  denotes the probability of distinct class  $C_i$ ,  $D = data\ set$  and  $A$  denotes the sub attributes and  $(|D_j|/|D|)$  denotes the weight of  $j$ th partition. Same as  $Gain(A)$  is the expected reduction in entropy caused by knowing the value of feature  $A$ .

Entropy is an approach of measuring uncertainty or randomness in a given set of data. For calculating the entropy of the database file and questionnaire results, data set is a sequence of bytes in the file. In the C4.5 algorithm information gain is obtained from entropy. There are different sorts of entropies utilized in C4.5 classification algorithms which are as follows:

A discrete probability distribution  $P(i) = p_1, p_2, p_3, p_4, \dots, p_k$  with the entropy  $H(X)$  of a random variable  $X$  is define by

$$H(x) = - \sum_{i=1}^N P(i) \log_2 P(i) \quad (5)$$

'Quadratic entropy' is examined the complete probability collection with a finite discrete random variable  $X$

$$P_i \geq 0 (i = 1, \dots, n), \sum_{i=1}^n p_i = 1 \quad (6)$$

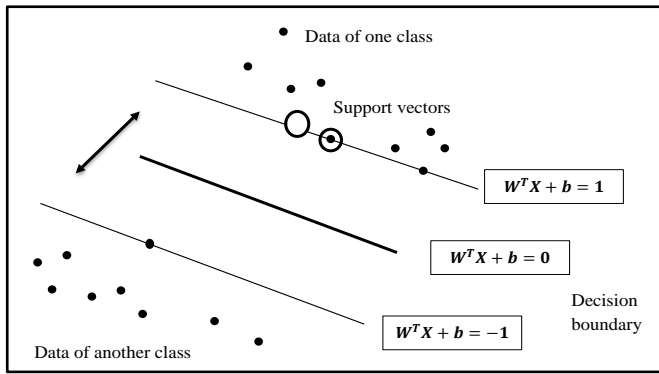
The C4.5 algorithm validation process is worth to utilizing enrolment data from year 2014 to test the classifier structure learnt with data between 2013 and 2014 replicates how the structure will be utilized in real situations: old data and questionnaire results is utilized to predict outcomes of new students thought about “becoming an entrepreneur”. Predicted results are fed into SVM classifier.

#### D. Support Vector Machine

SVM is utilized to take care of Binary classification problem. SVM direct to the linear algorithm into non-linear algorithm. This redirection purpose is uses the kernel function. Different kernel functions are there such as radial basis function, polynomial function. This kernel function can be utilized at the time of SVM training process to choose support vector. These support vectors are utilized by SVM to classify the information that outline the hyper plane in the vector space. The linear classifier structure as follows

$$F(x) = w^t + b \quad (7)$$

Where  $w$  is denotes weight vector and  $b$  denotes bias. The distance between the obtained information and hyper plane is presented in figure 2. The behind this statement is that SVM maps high dimensional feature space utilizing the inputs vectors nonlinearly and make the optimum separating hyperplane.



**Figure 2 Hyper Plane of Support Vector Machines**

Key concepts of Support Vector Machine (SVM) using key concept are define as follows

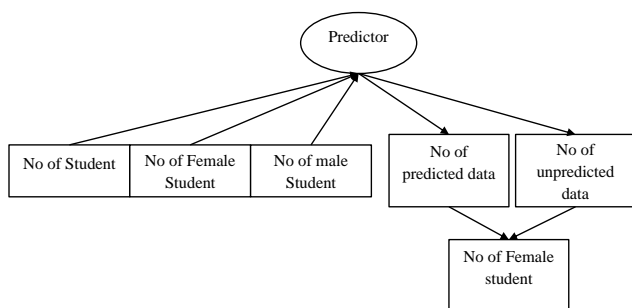
$y_i \in \{-1, 1\}$  are two classes and  $N$  labeled training samples  $(X_1, Y_1), \dots, (X_N, Y_N), X \in R^d$  where  $d$  is the dimensionality of vector.

$$w = \frac{1}{N_{interested}}$$

$$\sum_{i \in interested} X_i - \beta \frac{1}{N_{not interested}} \sum_{i \in not interested} X_i \quad (8)$$

Where  $N$  denotes the number of data and questionnaire results that are classified as interested and not interested female students to becoming an entrepreneur and  $\beta$  is used to training error minimization.

SVM approach using data mining technique which can be applied to efficiently assign higher education students with an accurately evaluates the male or female candidate will progress in a particular course by using the prediction techniques which is make use of “how many women candidates are like to became an entrepreneur” as shown in figure 3.



**Figure 3 Prediction of female students**

Women entrepreneur prediction is based on higher studies is a challenging task, is to extremely examine and effectively develop models for evaluating learning efforts with the combination of student academic data and questionnaire results. Student goal and female student outcomes clearly

relate to “promoting learning through effective higher graduates and academic course.”

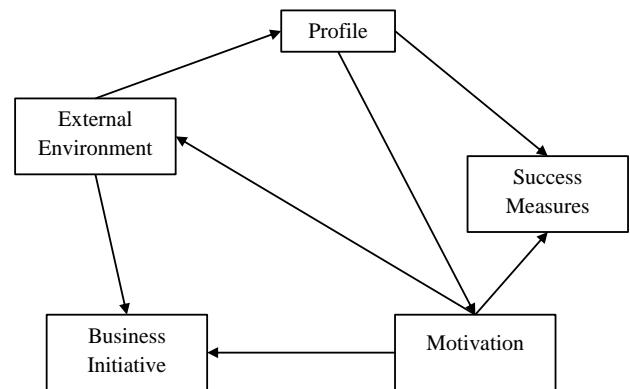
## Results and Discussion

### A. Research Methodology

This research study is fully based on primary data collected from higher education students (Sample size – 300) studying in different universities at Chennai (Refer Appendix 1 for trial Profile). The research can be identifying the impacts and influences among entrepreneurial intentions among higher study students. The hypothesis and regression analysis methods are used for calculates the higher study students to identify whether they take entrepreneurship as a career. An exploratory Varimax rotation with principle component analysis was performed by utilizing SPSS tool. Women Entrepreneurial intentions were measured for factor analysis, which eventually resulted in eight factors. Named these factors are shown in table 1 and analyzing method as shown in figure 4.

**Table 1. Attributes Factor**

No	Attributes
1	Name
2	Course
3	Mark
4	Psychological factor
5	About environment
6	Business thought
7	Motivation
8	Success factor



**Figure 4 Analyzing method**

### B. Metrics

Mean is define as the sum of the student detail observations divided by the number of student observations

$$M = \frac{\sum(X)}{N} \quad (1)$$

Where  $X$  is denoted as individual data points and  $N$  is denoted as number of datapoints.

The standard deviation is typically measures the variability of datasets is calculated equation is as follows

$$s^2 = \frac{\sum(X - M)^2}{n - 1} \quad (2)$$

Where  $X$  denotes sum of individual score,  $M$  denotes mean of all score and  $N$  is denoted as number of scores.

**C. Results**

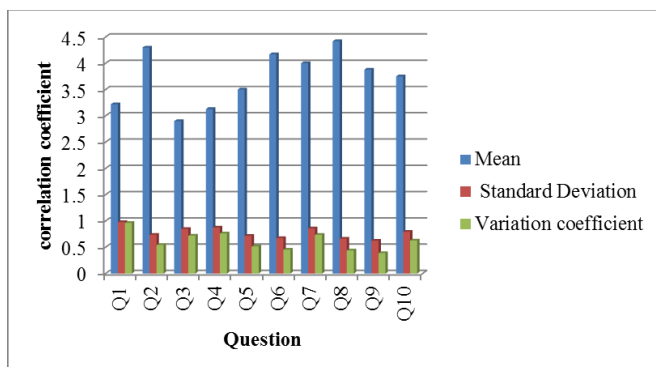
**Table 1 Questioner predicted values**

Group	variable	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Female	N	300	300	300	300	300	300	300	300	300	300
	Mean	3.22	4.30	2.90	3.13	3.50	4.17	4.00	4.42	3.88	3.75
	Std. Dev.	0.98	0.73	0.84	0.87	0.71	0.67	0.85	0.66	0.62	0.79
	Variance	0.96	0.54	0.71	0.76	0.51	0.45	0.73	0.43	0.38	0.62
		1	3	9	4	5	5	7	8	9	9

**Table 3 Correlation Coefficient between the Factors**

Variables	Family	Psychological	Financial	Environmental	Challenges
Family	-	0.83*	0.75	0.72	0.55
Psychological	-	-	0.65*	0.81*	0.42
Financial	-	-	-	0.61*	0.49
Environmental	-	-	-	-	0.52
Challenges	-	-	-	-	-

Table 2 shows the Questioner predicted values from the collected data and Table 3 shows the correlation coefficient between the measuring factors of motivate higher study women for becoming entrepreneurs.



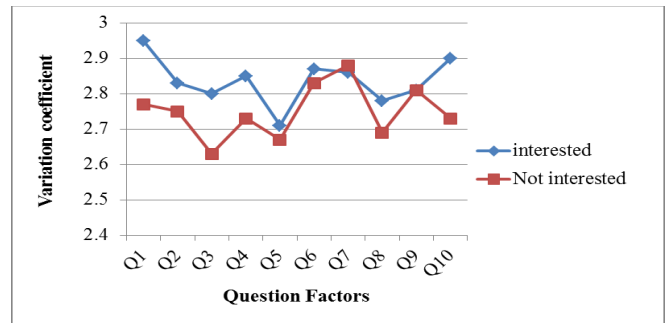
**Figure 5 Questioner predicted values**

**Table 4 interested women**

Group	variable	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
YES	N	300	300	300	300	300	300	300	300	300	300
	Mean	2.95	2.83	2.80	2.85	2.71	2.87	2.86	2.78	2.81	2.90
	Std. Dev.	0.57	0.88	0.85	0.84	0.75	0.88	0.82	0.79	0.83	0.72
	Variance	0.33	0.78	0.72	0.72	0.56	0.77	0.68	0.63	0.69	0.52
		1	1	8	0	8	9	5	3	4	8

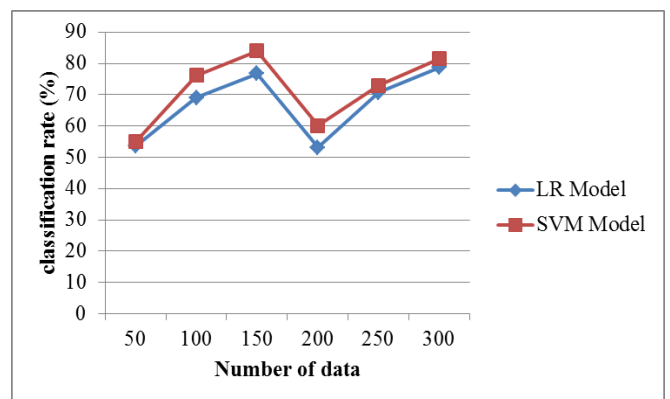
**Table 5 not interested women**

Group	variable	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
NO	N	300	300	300	300	300	300	300	300	300	300
	Mean	2.77	2.75	2.63	2.73	2.67	2.83	2.88	2.69	2.81	2.73
	Std. Dev.	0.69	0.77	0.76	0.76	0.80	0.78	0.78	0.80	0.79	0.61
	Variance	0.47	0.59	0.58	0.58	0.65	0.61	0.62	0.64	0.62	0.37
		8	5	0	5	2	0	2	5	4	2



**Figure 6 Predicted results**

Figure 6 indicate that higher education students with knowledge experience scored factor is better in all cases such the mean, standard deviation and variance. The maximum variances were recorded in case of Q1, Q3, Q4, and Q9. Finally these results are show that the women are very much interest in entrepreneur activation in present world. Higher study students have the opportunity to understand and take contact the life and individualities of real entrepreneurs.



**Figure 7 Classification rate**

Figure 7 shows the Logistic Regression (LR) and SVM model classification results. The classification rate is very high in SVM model to compare with parametric model such as LR. The proposed system shows the promising results to compare this LR model.

### Conclusion

Entrepreneurial education success is a major task to manage with the challenges that are considered from the burdened economies all over the world. Yet they are associated with the behavioral patterns and are having the same behaviors are included and the entrepreneurial competencies and the informal education are becoming more efficient to support the young people for the future while taking decisions and to become the victorious people in business. In order to make the student a better business professional, an additional special care has to be taken to find the opportunities for them. The students who are having experience in the business environment proves themselves in understanding the ethics and show better self-superior examination of acquired knowledge. This process of research have to be continued tile the low risk acceptance are recognized and the decision making pattern becomes slow that constitute common path for the young business people in order to establish their vision.

### References

- [1] Brijesh Kumar Baradwaj, Saurabh Pal, "Mining Educational Data to Analyze Students Performance", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.
- [2] Bhise R.B., Thorat S.S., Supekar A.K., "Importance of Data Mining in Higher Education System" Journal Of Humanities And Social Science (IOSR-JHSS), 2279-0837, Vol.6, Issue 6, PP 18-21, 2013.
- [3] Heba Mohammed Nagy, Walid Mohamed Aly, Osama Fathy Hegazy, "An Educational Data Mining System for Advising Higher Education Students", International Journal of Computer, Control, Quantum and Information Engineering Vol:7 No:10, 2013.
- [4] Khushbu N.Shah, Misika R. Patel, Neha V. Trivedi, Priyanka N. Gadariya, Richa H. Shah, Nehal Adhvaryu, "Study of Data Mining in Higher Education-A Review", International Journal of Computer Science and Information Technologies, Vol. 6 (1), 455-458, 2015.
- [5] Ying Zhang, Samia Oussena, Tony Clark, Hyeonsook Kim "Use Data Mining To Improve Student Retention in Higher Education – A Case Study", International Conference on Enterprise Information Systems (ICEIS), 8-12, 2010.
- [6] Jayshree Jha, Leena Ragma, "Educational Data Mining using Improved Apriori Algorithm", International Journal of Information and Computation Technology, 0974-2239, Vol.3, Number 5, pp. 411-418, 2013.
- [7] Suhas G. Kulkarni, Ganesh C. Rampure, Bhagwat Yadav, "Understanding Educational Data Mining (EDM)", International Journal of Electronics and Computer Science Engineering, 2277-1956, Vol.2, Number 2, 2013.
- [8] Mathias Rossi, Silna Borter, Marie Sansonnens, "Gender Differences in Entrepreneurship: Situation, Characteristics, Motivation and Entrepreneurial Behavior of Women Entrepreneurs in Switzerland", International Journal of Social, Education, Economics and Management Engineering Vol:7, No:8, 2013.
- [9] Poonam Yadav, "SR-K-Means Clustering Algorithm for Semantic Information Retrieval", International Journal of Inventions in Computer Science and Engineering, Vol.1 Issue 9, 2014.
- [10] S.Priyanka, T.K.P.Rajagopal, R.Thiyagarajan, "A Competitive Intelligence Measures In Business Forecast with Renewable Intensity Sources", International Journal of Inventions in Computer Science and Engineering, Vol.1, Issue 4, 2014.
- [11] Mahesh N, Jyothsna B, "Naive Bayes Training Phase Reduction Using Dataset Feature Filtering for Malware Detect", International Journal of Inventions in Computer Science and Engineering, Vol.1, Issue 3, 2014.