# A NOVEL APPROACH TO PREDICT THE STUDENTS ACADEMIC CONCERT THROUGH DATA MINING TECHNIQUES

**R. Dhanalakshmi[1], Dr. B. Umadevi[2]**
[1]Research Scholar, P.G. & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Tamil Nadu, India
[2]Assistant Professor & Head, P.G. & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Tamil Nadu, India

## Abstract

Today the educational qualities and standards of the colleges are needs to improve much in order to meet competitive challenges. The growth and success of the organizations are equally contributed by both students and faculties. The real role in academic performance of each student is more important for every teacher. In modern technical world, many statistical tools are available to caliber the students' concert. But the tools may not produce analytical or evolutionary reports. In this research the student's performance is analyzed using the data mining techniques naïve bayes algorithm and Support Vector Machine (SVM).The investigation was conducted with different types of metrics.

***Keywords**:* EDM, Support Vector Machine, Naïve Bayes, LMS.

## 1. Introduction

The Educational Data Mining (EDM) refers to techniques, tools, and research designed for automatically to extracts data generated by or related to people's learning activities in educational settings [1]. In general it is more precise and extensive. In several learning management system every student is accessed by the learning object by their access rates and their usage aspects. Due to the advancement in different fields produces large amount of data and also stored in different format. The stored data may be any one of the format like files, images, sound and videos. Now a day it is tough to manage and analyse huge data set. Data mining techniques helps in extracting knowledge from the large repositories. Data mining is a powerful analytical tool that gives critical information and knowledge, which can help to improve decision making processes.

Data mining due to its significance in decision making, it is successfully applied in diversity domains including education. Education Data Mining (EDM) is an emerging trend in research. The main focus is to explore the usefulness of educational learning system. In order to maximize the throughput student's performance in association with the curriculum, it requires various attributes to be considered. The attributes may be the good catalyst for Prediction and analysis. In many research work the attributes such as family factor, psychological profile, previous schooling, prior academic performance, and student interaction with their classmates and teachers plays the major role in determining their performance. The EDM algorithms are different from traditional data mining algorithms. The recent classification of EDM deals with predicting the output value based on input data. The major category of prediction classifications is (1) classification, (2) regression and (3) density estimation. Popular classification algorithm includes support vector machine, neural network, naïve bayes, Decision Tree and the predication is either a binary or categorical variable. In this research, we have used two different data mining classification algorithms (Naïve Bayes and Support vector Machine).

The paper is structured as follows. The section one describes about introduction. Section 2 deals about background study and its related works. The methodology of the research work is explained in section three. In section four portraits the Experiment results. Finally the paper is concluded in last section.

## 2. Background and related works

### 2.1 Literature Review

The diversified approach using various data mining technique are applied to analyse the academic performance of students at various levels. The academic progression used for in various models are mentioned below.
A comprehensive literature review of various researchers' works are discussed below.

Currently, The author *Amjad Abu Saa* has collected various sources of information for the predicting the students performance. The dataset emphasize the importance of the students attribute like gender first language, high school percentage, status, living location, etc. He applied the data mining techniques such as classification, clustering, Association Rule learning and Artificial intelligence etc.

Presently, The decision tree induction is supervised classification technique that builds a top-down tree model. The CART is another decision tree algorithm which uses minimal cost complexity pruning. In his Chi-Squared automatic interaction detection (CHAID), the splitting criterions used in other decision tree algorithm. At the outset the multiple decision tree techniques and algorithms. The CART has the best results CHAID has the accuracy 34.7 % rather than CART.

The author *Ahmed Mueen* in his research he analyzed students academic performance using the data mining classification techniques such as Naïve Bayes, neural networks and decision tree. The prediction performance of three classifies are measured and compared. He concluded that the naïve Bayes classifier outperforms other two classifiers by achieving the prediction accuracy of 86% than that of other approaches.

In Malaysia, the student progress and performance is not being addressed. The reasons are:
(i) The existing method is insufficient to predict the performance of the students. (ii) Due to the lack of investigations on the factors affecting student's achievements in particular courses within Malaysian context. The data mining technique Decision tree, neural network, k-nearest and also the SVM are used in his research. In his research it outcome focuses that Neural Network has the highest prediction accuracy by (98%) followed by Decision Tree by (91%). Next, Support Vector Machine and K-Nearest Neighbor gave the same accuracy, which is (83%). Lastly, the method that has lower prediction accuracy is Naive Bayes by (76%).

### 2.2 Predicting Students' Performance

Predicting the students' academic attainment is a significant part in higher learning institution. Understanding the factors that affect student performance is a difficult research task due to many different aspects like cultural, social, previous academic performance, interaction with teachers, etc. Several researchers have been working on these factors and they had produced promising results. Many researchers investigated the impact of socio-economic status. Some others studied the connection between student academic performance and their parent behaviors while others looked into the efficiency of teacher to improve student academic performance. It is also noticed that due to Learning Management System (LMS) such as Blackboard, Moodle, WebCT etc. most of the recent research conducted on EDM has been applied to web-based education. These system provide information about student assessments, activities in forums, and how many times students access teaching resources, which is very important information in predicting student performance and help teacher to detect course weaknesses.

## 3. Methodology

In educational data mining [2] method, predictive modeling is usually used in predicting student performance. In order to build the predictive modeling, there are several tasks used, which are classification, regression and categorization. The most popular task to predict student's performance is classification. Among the algorithms used are Decision tree, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor and Support Vector Machine for predictive modeling.

This section deals with the proposed methodology. The Support Vector Machine algorithm is used as the back bone of the method. It is adapted into the process of the educational data mining. The goal of this study is to improve the effectiveness of the proposed methodology to predict the students' academic performance better than Naïve bayes approach. The existing and proposed algorithm in predicting student performance will be described in the next section.

### 3.1 Naive Bayes Classifiers

The Naive Bayes (NB) algorithm is based on Bayes theorem with independence assumptions between predictors. A Naive Bayes model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Despite [3][4] its simplicity, the Naive Bayes classifier often does surprisingly well and is widely used because it often outperforms more sophisticated classification methods[3].

Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence. The posterior property is calculated by the following equation.

$$\boldsymbol{P(c/x) = \frac{P(x/c) * P(c)}{P(x)}}$$

In the above equation,
1. P(c|x) is the posterior probability of class (target) given predictor (attribute).
2. P(c) is the prior probability of class.
3. P(x|c) is the likelihood which is the probability of predictor given class.
4. P(x) is the prior probability of predictor.

## Pseudo code:

```
Step 1: Convert the data set into a frequency table.
Step 2: Create Likelihood table by finding the probabilities.
Step 3: Use Naïve Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.
```

### 3.2. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm [4] which can be used for both classification and regression [5][6] challenges. However, it is mostly used in classification problems. Even though it's considered that Neural Networks are easier to use than this, however, sometimes unsatisfactory results are obtained. A step in SVM classification [7] involves identification as which are intimately connected to the known classes. This is called feature selection or feature extraction [8]. Feature selection and SVM classification together have a use even when prediction [9] of unknown samples is not necessary. They can be used to identify key sets which are involved in whatever processes [10] distinguish the classes. The proposed algorithm for the efficient way of predicting students' academic performance is presented as in the pseudo code format.

**Pseudo code:**

```
Step 1 : CandidateSV = { closest pair from opposite classes }
Step 2 : while there are violating points do
Step 3: Find a violator
Step 4 : candidateSV = candidateSV U violator
Step 5 : if any $\alpha p < 0$ due to addition of c to S then
Step 6: candidateSV = candidateSV \ p
Step 7 : repeat till all such points are pruned [11]
Step 8 : end if
Step 9: end while
```

The method suggested in this paper is to improve the prediction of students' academic performance is belong to the process of data mining [12] which is given in Figure1. There are four main stages in this method. The stages are Data collection, preprocessing, classification and result interpretation. Data collection is gathering all information available on students considering factors affect student performance. This information can be collected for MCA students' semester examination marks. During pre-processing stage data cleaning, attributes selection, dimensionality reduction, and data partitioning are applied to get better prediction [13]. Whereas, in classification stage Data Mining algorithms are used for the classification of data. Normally, at this stage different Data Mining algorithms are executed [14] with different variables and compared to select algorithm which produce best results. Finally, in interpretation stage models obtained from previous stage are analyzed to predict student performance.

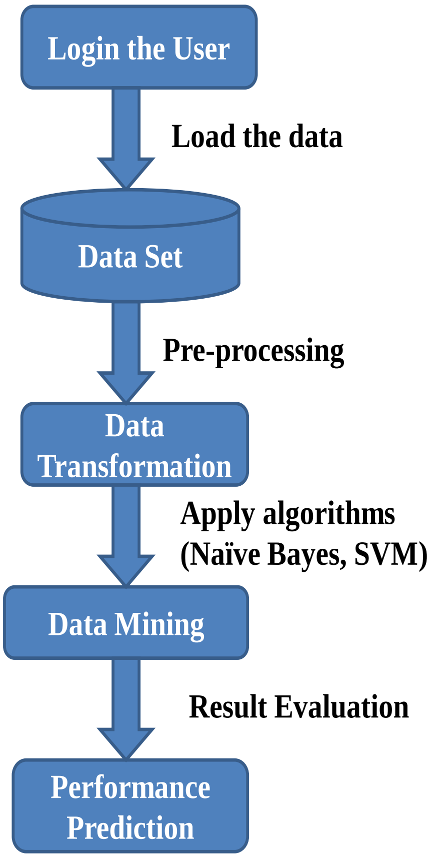
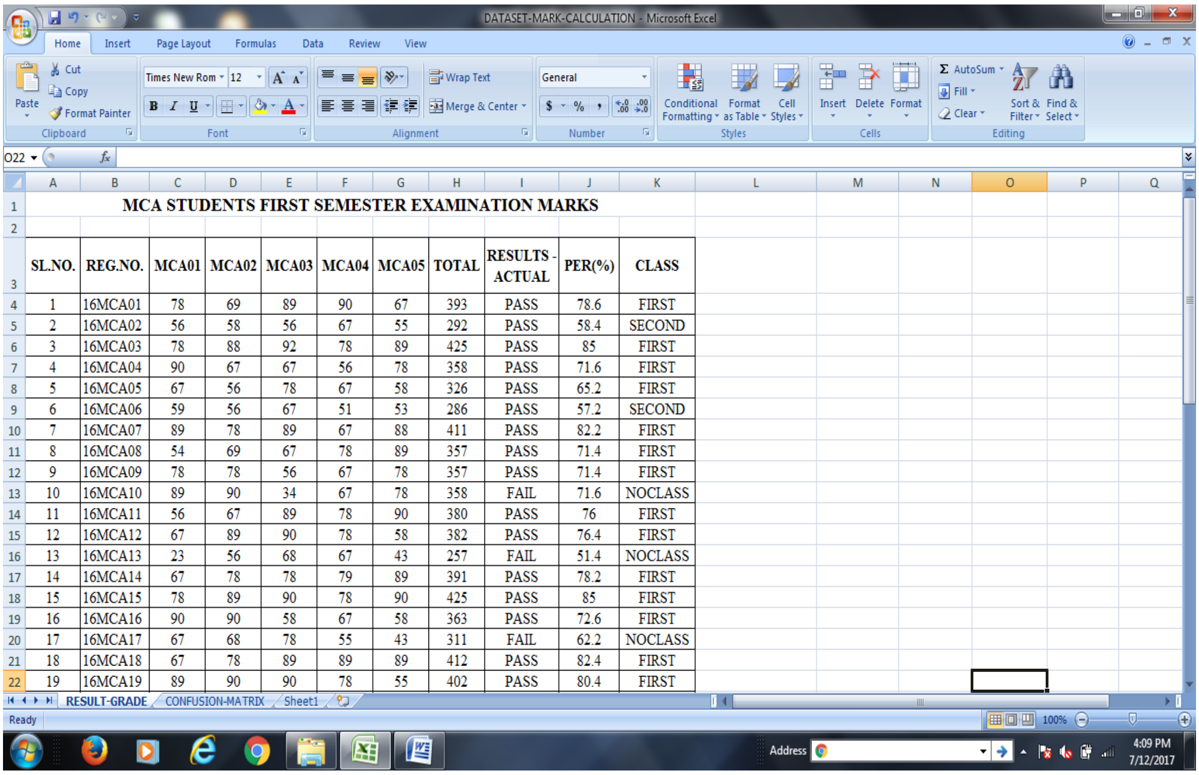


**Figure 1**: Method proposed for improving the prediction of Students' performance

## 4. Experiment results

The educational data mining is applied on prediction, analysis, visualization etc. The system predicts student's performance using Naïve Bayes and Support Vector Machine (SVM). The prime focus is towards improvement of educational process.

The dataset has been collected for MCA student's first semester Examination marks from the nearest leading educational institution. The initial approach is to calculate the total marks for the students based on their individual marks obtained in each subject. The second step is to evaluate their results based on their subject's marks. The next step is percentage computation. Final step is classifying the students into three different types such as of first, second and noclass. The sample dataset and the outcome of the actual data is mentioned in Figure 2.



MCA STUDENTS FIRST SEMESTER EXAMINATION MARKS

| SL.NO. | REG.NO. | MCA01 | MCA02 | MCA03 | MCA04 | MCA05 | TOTAL | RESULTS - ACTUAL | PER(%) | CLASS |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 16MCA01 | 78 | 69 | 89 | 90 | 67 | 393 | PASS | 78.6 | FIRST |
| 2 | 16MCA02 | 56 | 58 | 56 | 67 | 55 | 292 | PASS | 58.4 | SECOND |
| 3 | 16MCA03 | 78 | 88 | 92 | 78 | 89 | 425 | PASS | 85 | FIRST |
| 4 | 16MCA04 | 90 | 67 | 67 | 56 | 78 | 358 | PASS | 71.6 | FIRST |
| 5 | 16MCA05 | 67 | 56 | 78 | 67 | 58 | 326 | PASS | 65.2 | FIRST |
| 6 | 16MCA06 | 59 | 56 | 67 | 51 | 53 | 286 | PASS | 57.2 | SECOND |
| 7 | 16MCA07 | 89 | 78 | 89 | 67 | 88 | 411 | PASS | 82.2 | FIRST |
| 8 | 16MCA08 | 54 | 69 | 67 | 78 | 89 | 357 | PASS | 71.4 | FIRST |
| 9 | 16MCA09 | 78 | 78 | 56 | 67 | 78 | 357 | PASS | 71.4 | FIRST |
| 10 | 16MCA10 | 89 | 90 | 34 | 67 | 78 | 358 | FAIL | 71.6 | NOCLASS |
| 11 | 16MCA11 | 56 | 67 | 89 | 78 | 90 | 380 | PASS | 76 | FIRST |
| 12 | 16MCA12 | 67 | 89 | 90 | 78 | 58 | 382 | PASS | 76.4 | FIRST |
| 13 | 16MCA13 | 23 | 56 | 68 | 67 | 43 | 257 | FAIL | 51.4 | NOCLASS |
| 14 | 16MCA14 | 67 | 78 | 78 | 79 | 89 | 391 | PASS | 78.2 | FIRST |
| 15 | 16MCA15 | 78 | 89 | 90 | 78 | 90 | 425 | PASS | 85 | FIRST |
| 16 | 16MCA16 | 90 | 90 | 58 | 67 | 58 | 363 | PASS | 72.6 | FIRST |
| 17 | 16MCA17 | 67 | 68 | 78 | 55 | 43 | 311 | FAIL | 62.2 | NOCLASS |
| 18 | 16MCA18 | 67 | 78 | 89 | 89 | 89 | 412 | PASS | 82.4 | FIRST |
| 19 | 16MCA19 | 89 | 90 | 90 | 78 | 55 | 402 | PASS | 80.4 | FIRST |

**Figure 2** : Sample Dataset

The same dataset is applied through the Naïve Bayes and Support Vector Machine (SVM) classification algorithms. The confusion matrix will be generated by applying the N cross validation process. The results of this performance will applied for the both algorithms such as Naïve Bayes and Support Vector Machine (SVM). The outcome is listed in Table 1 and Table 2. The statistical results are evaluated on the basis of accuracy, sensitivity, fmeasure, specificity and error-rate from the confusion matrix. The outcomes are given in the above Table 3 and Figure 3. The result shows that SVM gives much better results rather than the naïve base.

**Table 1** : Confusion matrix for Naïve Bayes

| No.Of Samples = 50 | NAÏVE-BAYES-Predicted: FAIL | NAÏVE-BAYES-Predicted: PASS | TOTAL |
|---|---|---|---|
| **Actual : FAIL** | TN = 3 | FP = 4 | **7** |
| **Actual : PASS** | FN = 8 | TP = 35 | **43** |
| **TOTAL** | **11** | **39** | **50** |

**Table 2** : Confusion matrix for SVM

| No.Of Samples = 50 | SVM-Predicted: FAIL | SVM-Predicted: PASS | TOTAL |
|---|---|---|---|
| **Actual : FAIL** | TN = 6 | FP = 1 | **7** |
| **Actual : PASS** | FN = 2 | TP = 41 | **43** |
| **TOTAL** | **8** | **42** | **50** |

**Table 3**: Students' Performance Prediction based on SVM and Naïve Bayes

| Algorithm | Accuracy | Sensitivity | F-Measure | Specificity | Error rate |
|---|---|---|---|---|---|
| **Naïve Bayes** | 0.8600 | 0.8139 | 0.8529 | 0.4286 | 0.1400 |
| **SVM** | 0.9400 | 0.9530 | 0.9380 | 0.8571 | 0.0600 |

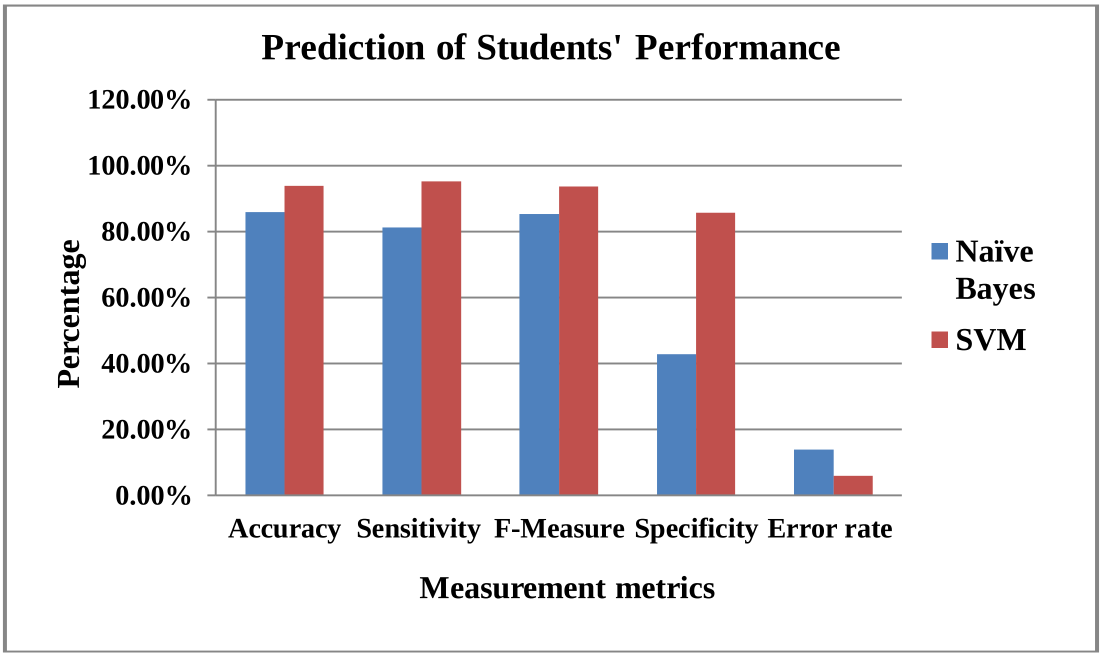


**Figure 3**: Students' Performance Prediction based on SVM and Naïve Bayes

## 5. Conclusion

The data mining techniques such as Naïve bayes and Support Vector Machine is used to predict the students' performance. The data set volume is used in this research is 50 students. The factors which mostly affect student's performance are taken as parameters. The students are classified in to three different types according to their percentage. The students are categorized according to their results as FIRST, SECOND and NOCLASS. The algorithms both Naïve bayes as well as SVM are compared with each other. The error rates at 14% and 6% respectively. Similarly the accuracy of the algorithms is 86% and 94% respectively. These statistical analyses will make the teachers to proact against the ill factors of the students. In addition to that it will also more useful for the students' community to understand their levels. The research helps also for the teachers, management and students to make a clear vision about their future plan.

# References

[1] B.Namratha, Niteesha Sharma, 2016, "Educational Data Mining –Applications and Techniques", *International Journal of Latest Trends in Engineering And Technology (IJLTET)*, Volume 7, issue 2.

[2] B. Umadevi, D.Sundar, Dr.P.Alli, 2011, " A Study on Stock Market Analysis for Stock Selection – Naïve Investors' Perspective using Data Mining Technique", *International Journal of Computer Applications* ,Volume 34– No.3.

[3] Springer, Informatics in Control, 2015, Automation and Robotics 12th International conference, *ICINCO* 2015 Colmar, France.

[4] Bhagyashri Wagh, , J. V. Shinde, , N. R. Wankhade, 2016, "Sentimental Analysis on Twitter Data using Naive Bayes", *International Journal of Advanced Research in Computer and Communication Engineering*, Volume 5, Issue 12.

[5] K.Gomathi, Dr. Shanmugapriyaa, 2016,"Heart Disease Prediction Using Data Mining Classification", *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, Volume 4, Issue II.

[6] http://www.analyticsvidhya.com/

[7] http://www.coursehero.com/

[8] Jyoti Devi, Nancy Seghal, 2017,"A Review of Improving Software Quality using Machine Learning Algorithms", *International Journal of Computer Science and Mobile Computing, IJCSMC*, Volume 6, Issue 3, pp.148 – 153.

[9] Harsh Singhal,Neelendra Badal,Amit Kumar Gupta,Devesh Singh Sisodia,Gautam Kumar Singh, Hemant Kumar Singh, 2015, "A Novel Approach for Load Balancing in Distributed System using FIFO-Support Vector Machine (FIFOSVM)", *International journal of science and research"*, Volume 4, Issue 12.

[10] Dr. Neeraj Bhargava, Abhishek Kumar, Devesh Kumar, Meenakshi, 2015, "A modified concept of PCA to reduce the classification error using kernel SVM classifier ", *International Journal of Scientific & Engineering Research*, Volume 6, Issue 6.

[11] B. Umadevi, D.Sundar, Dr.P.Alli, 2013, "An Optimized Approach to Predict the Stock Market Behavior and Investment Decision Making using Benchmark Algorithms for Naive Investors", *Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference* on *( IEEE Xplore Digital Library*), pp.1 -5.

[12] J.J. Little, Walter E. Gillett, 1990, "Direct evidence for occlusion in stereo and motion", *Research Gate*, November.

[13] Prof. G.P. Mohole, , Dhanashree Tambe, , Amey jadhav, ,Manoj khairnar, , Abhishek Sonar, 2015, "Application Opening Based on Emotion Using HOG Features", *International Journal of Advanced Research in Computer and Communication Engineering,* Volume 4, Issue 10.

[14] Dr.B. Umadevi, R. Dhanalakshmi, 2017, "A Comprehensive Survey of Students Performance Using Various Data Mining Techniques", *International Journal Of Science and Research (IJSR)*, Volume 6, Issue 4.

[15] B. Umadevi,,D.Sundar, Dr.P.Alli, 2014, "Novel Framework For The Portfolio Determination Using PSO Adopted Clustering Technique", *Journal of Theoretical and Applied Information Technology*", Volume 64 No.1.

[16] Ahmed Mueen, Bassam Zafar, Umar Manzoo, 2016, "Modeling and Predicting Students' Academic Performance Using Data Mining Techniques", *I.J.Modern Education and Computer Science*,Volume 11, pp.36-42, Published Online in MECS.

## Biography

**R.DHANALAKSHMI** is an M.Phil Research Scholar in PG & Research Department Of Computer Science, Raja Doraisingam Government Arts College, Sivaganga, Tamilnadu, India. Her research interest includes in Data mining and its applications.

**Dr. B.UMADEVI** has received her Doctoral degree in Computer Science from Manonmaniam Sundaranar University, Tirunelveli, India. Currently working as Assistant Professor & Head- P.G and Research Department of Computer Science, Raja Doraisingam Government Arts College, Sivagangai-Tamilnadu, India. She has over 22 years of Teaching Experience and published her research papers in various International, National Journals and Conferences. Her research interests include Data Mining, Soft Computing and Evolutionary Computing. She got the Best Paper Award for her publication in the IEEE International Conference on Computational Intelligence and Computing Research held on 27th Dec 2013 at VICKRAM College of Engineering and Technology.