IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-10 E-ISSN: 2347-2693

Comparative analysis of classification algorithm in EDM for improving student performance

B.R. Patel

MCA , AMPICS, Ganpat University, Mehsana, India

Corresponding Author: bhavesr@gmail.com, Tel.: 8758422545

Available online at: www.ijcseonline.org

Received: 21/Sep/2017, Revised: 30/Sep/2017, Accepted: 15/Oct/2017, Published: 30/Oct/2017

***Abstract*—** Data mining techniques are useful to extract the useful information and to support in the decision making process. There are too many application in the educational domain where we can apply the data mining. Right now data mining in Educational domain is rapidly developing technique. In this research paper we analyse the student's result of every semester using data mining techniques. Data mining supports the too many techniques but here in the analysis we are using the various classification algorithm of data mining techniques. Here in this research analysis we worked on two model. Model A uses the dataset that contains all the student performance parameters and data mining classification techniques and generated the result based on Accuracy and Error Rate of the classifiers and Model B uses the dataset contains only statistically proved highly affected parameters on student performance and applied data mining techniques on this data sets and generate the results based on Accuracy and Error rate of the classifiers. This research work compare the result of both the model and check that which model is best. The comparison is done using the measurement of accuracy and measurements of Error Rate. This research work also shows that which algorithm is most suitable for predicting the performance of the students among the selected algorithms. The analysis work is done by considering various types of algorithm like decision tree algorithm, rule based algorithm, Bayesian algorithm and function based algorithms. This generic novel approach can be extended to other disciplines as well.

***Keywords* —** classification, error rate, data set, data mining, prediction.

I. Introduction

Right now university works in a competitive environment. The main issue of universities is to analyse their performance in depth, to determine their uniqueness, and to develop strategies for further development and future action. The objective of the proposed research work is to find out whether any pattern in the available data may help to predict student academic performance based on individual and pre-university characteristics. Data mining may be a promising and thriving frontier in data analysis.

So, in this research paper we have used data mining classification technique and its algorithms for predicting student's performance.

This research paper is organizing as per the following structure. Here, Section – I gives the introduction of research, Section – II specifies the objective of the research, Section – III gives the literature review, Section – IV specifies the data collection methodology, Section – V describe experiment result and discussion , Section – VI gives the conclusion and future work of research and Section – VII describe the appendix of the research work.

II. Objective of Research

In this paper different techniques of data mining suitable for classification have been compared such as Rules-based, Trees-based, Functions-based and Bayes-based algorithms. The first objective of our study is to find out the attributes which are dominant factors for prediction of student's performance. And other objective is to find out the best suited algorithm for predicting student performance.

III. Literature Survey

Bhise R.B, Thorat S.S and Supekar A.K. (2013) [1] used data mining process in a student's database using K-means clustering algorithm to predict students result.
Monika Goyal and Rajan Vohra (2012) [2] applied data mining techniques to improve the efficiency of higher education institution. If data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. This is an approach to examine

the effect of using data mining techniques in higher education.
K.Shanmuga Priya and A.V.Senthil Kumar (2013) [3] applied a Classification Technique in Data Mining to improve the student's performance and help to achieve the goal by extracting the discovery of knowledge from the end semester mark.

Varun Kumar and Anupama Chadha (2013) [4] used of one of the data mining technique called association rule mining in enhancing the quality of students' performances at Post Graduation level.
Tiwari & Yashpal Singh (2012)[5] evaluated performance of 4 clustering algorithms on different datasets in WEKA with 2 test modes. We presented their result as well as about tool and data set which are used in performing evaluation.
Sonali Agarwal, G. N. Pandey, and M. D. Tiwari (2012) [6] describes the educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. Data Mining is an emerging technique with the help of this one can efficiently learn with historical data and use that knowledge for predicting future behavior of concern areas. Growth of current education system is surely enhanced if data mining has been adopted as a futuristic strategic management tool. The Data Mining tool is able to facilitate better resource utilization in terms of student performance, course development and finally the development of nation's education related standards.
Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal (2012) [7] used decision tree classifiers are studied and the experiments are conducted to find the best classifier for retention data to predict the student's drop-out possibility.
Brijesh Kumar Baradwaj and Saurabh Pal (2011) [8] Used the classification task on student database to predict the students division on the basis of previous database.
Pallamreddy.venkatasubbareddy and Vuda Sreenivasarao (2010) [9] explained the Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal and use of decision trees is as a descriptive means for calculating conditional probabilities.
Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R (2011) [10] in their article "Comparison of different data mining techniques to predict hospital length of Stay" compared four data mining classification techniques MLP, Naïve Bayes, K-NN, J48 to predict length of stay for an inpatient in hospital on preprocessed dataset derived from electronic discharge summaries with 401 instances & 16 parameters. In result they found that MLP performs better than rest three classifiers with 87.8% correctly classified instances.

IV. DATA COLLECTION METHODOLOGY

For this study, we have collected total 1000 student's data from the various UG institutions of computer science. We have collected this data by considering the parameters like student's demographic, student's learning behavioural and student's academic information.

Table 1: List of Parameters

ATTRIBUTES	DATA TYPE	VALUES
Gen	Nominal	Male, female
Percentagehsc	Nominal	Poor, average, good, very_good, excellent
Stream	Nominal	Commerce, science
F_annual_income	Nominal	Low, average, middle, high, very high
F_qualification	Categorical	No formal education, primary, ssce, 1st degree, 2nd degree, phd
F_occupation	Categorical	Unemployed, government worker, private, self employed
M_qulification	Categorical	No formal education, primary, ssce, 1st degree, 2nd degree, phd
M_occupation	Categorical	Unemployed, government worker, private, self employed
No_of_sublings	Categorical	One, two, three, four
Overall_attendance	Nominal	Poor, average, good, very_good, excellent
W_l_h	Nominal	Poor, average, good, very_good, excellent
W_li_u	Nominal	Poor, average, good, very_good, excellent
D_re_h	Nominal	Poor, average, good, very_good, excellent
E_w_l_u_h	Nominal	Poor, average, good, very_good, excellent
Internal_marks	Nominal	Poor, average, good, very_good, excellent
Assignment_ marks	Nominal	Poor, average, good, very_good, excellent
Participation_extra_curriculum	Nominal	Poor, average, good, very_good, excellent
Practical_knowledge	Nominal	Poor, average, good, very_good, excellent
Theory_marks	Nominal	Poor, average, good, very_good, excellent
Internet_uses_learning	Nominal	Poor, average, good, very_good, excellent

Previous_sem_ marks	Nominal	Poor, average, good, very_good, excellent
Semester_wise_ result	Nominal	Poor, average, good, very_good, excellent

Ref: Abbreviation detail is described in Appendix A

V. EXPERIMENT RESULT AND DISCUSSION

To find out the first objective of research: Here we apply Statistical Analysis to find highly affected parameters on student performance.

Here we have applied Multiple Regression technique using SPSS tool is applied on collected data set and generate the following result.

Table 2: Anova Table

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3295.2	21	156.91	863.94	.000^b
	Residual	642.41	353	.182		
	Total	3937.6	355			

The F-ratio in the ANOVA table is nearer to 1 that shows that overall regression model is a good fit for the data.

Here p value is also 0.000 it is < 0.005 so it shows that model is a good fit of the data.

Table 3: Coefficient Table

Model	Unstandardized coefficients		Std. Coeffi.	T	Sig.
	B	Std. Error	Beta		
(constant)	.237	.075		3.149	.002
Gen	-.022	.015	-.010	-1.465	.143
Percentagehsc	-.009	.005	-.011	-1.648	.000
Stream	.018	.016	.008	1.129	.259
F_annual_income	-.019	.005	-.026	-3.717	.000
Fq	.001	.005	.001	.186	.853
Fp	.005	.006	.005	.798	.425
Mq	.012	.008	.015	1.589	.112
Mp	.025	.014	-.017	-1.736	.083
Nos	.095	.011	.057	8.301	.000
Overall_attendance	.207	.010	.183	21.05	.000
W_l_h	-.013	.005	-.018	-2.626	.004
W_li_u	-.007	.005	-.010	-1.412	.158
D_re_h	.001	.005	.002	.270	.787
E_w_l_u_h	.008	.005	.010	1.498	.134
Internal_marks	.265	.009	.249	29.749	.000
Assignment_marks	-.013	.007	-.014	-1.820	.000
Paticipation_extra_ Curriculam	.002	.008	.001	.196	.000
Practical_knoledge	.167	.013	.187	12.765	.000
Theory_marks	.021	.013	.022	1.592	.000
Internet_uses_ learning	-.248	.016	-.114	-15.145	.003
Previous_sem_marks	.390	.010	.422	37.925	.000

In the above coefficient table if significance level is < 0.05 then that variable is significant. If significance level is > 0.05 then the variable is not a good predicator and should be removed from the model.

Outcome of Statistical Analysis:

Based on the above analysis highly affected parameters on the students' performance are as per the following:

Table 4: Highly affected parameters on student's performance

Percent_HSC,	Internal_Marks
F_Annual_Income,	Assignment_Marks
W_L_H	Participation_Extra_Curriculam
Overall_Attendance	Practical_Knowledge
NOS	Theory_Marks
Internet_Uses_Learning	Previous_Sem_Marks

So, we have kept these research parameter for our further research work.

Classification and evaluation

In this research work we have used following classification algorithms on student's data set using the WEKA too.

Rule-based algorithms: oner

Trees-based algorithms: j48, decisionstump

function-based algorithms: logistic, multilayerperceptron and smo

bayes-based algorithms: bayesnet and naivebayes

In this work, there are two models, Model (A) and Model (B). Model (A) uses all the attributes to calculate the accuracy of classifiers, error rate of classifiers and class-wise accuracy of classifiers getting in each semester of course. Model (B) uses only statistically approved high affected attributes on the performance of students. So as per that Model (A) and Model (B) uses the following attributes where we applied different classification algorithms and calculate the various performance measures.

MODEL (A) - USES ALL THE ATTRIBUTES

(Gen, Percent_HSC, Stream, F_Annual_Income, FQ, FP, MQ, MP, NOS, Overall_Attendance, W_L_H, W_Li_U, D_Re_H, E_W_L_U_H, Internal_Marks, Assignment_Marks, Participation_Extra_Curriculam, Practical_Knowledge, Theory_Marks, Internet_Uses_Learning, Previous_Sem_Marks)

MODEL (B)

(Percent_HSC, F_Annual_Income, W_L_H, Overall_Attendance, NOS, Internal_Marks, Assignment_Marks, Participation_Extra_Curriculam,, Practical_Knowledge, Theory_Marks, Internal_Uses_Learning, Previous_Sem_Marks)

Analysis for Model (A):

Semester Wise Comparative analysis to find the Accuracy of Classifiers in Data Mining:

We have used following parameters to measure the accuracy.

(A) Timing to Build the Model
(B) Correctly Classified Instances
(C) Incorrectly Classified Instances.
(D) Root mean square error rate

Table 5: Time Taken to Build the Model

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	0.05	0.13	0.12	11.21	79.33	0.15	0.14	0.16	2.95
S2	0.08	0.14	0.110	10.29	78.05	0.11	0.14	0.160	2.96
S3	0.09	0.15	0.122	10.30	78.06	0.12	0.15	0.172	2.972
S4	0.08	0.14	0.119	10.29	78.05	0.11	0.14	0.169	2.96
S5	0.08	0.14	0.117	10.29	78.05	0.11	0.14	0.167	2.96
S6	0.06	0.12	0.09	10.27	78.03	0.09	0.12	0.14	2.94
MV	0.07	0.13	0.113	10.44	78.26	0.11	0.14	0.161	2.95

Table 6: Correctly classified instances

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	98.8	97.4	60.8	96.31	87.29	96.3	78.29	96.3	87.32
S2	98.1	96.4	62.8	95.31	90.29	96.3	80.29	94.3	92.32
S3	98.5	97.0	63.8	96.31	91.29	97.3	81.29	95.3	93.32
S4	98.6	97.1	64.8	96.41	92.29	97.5	82.29	95.5	93.52
S5	98.8	97.2	64.8	96.57	92.39	97.7	82.34	95.6	93.62
S6	98.9	98.4	59.8	98.31	89.29	98.3	79.291	98.3	89.32
MV	98.61	97.2	62.8	96.54	90.48	97.3	80.63	95.9	91.57

Table 7: In Correctly classified instances

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	1.2	3.57	39.12	3.69	13.7	3.6	22.7	3.15	13.67
S2	1.9	3.57	37.12	4.69	9.7	3.6	19.71	5.65	7.67
S3	1.5	2.99	36.12	3.69	8.7	2.6	18.71	4.65	6.67
S4	1.4	2.9	35.12	3.59	7.7	2.4	17.71	4.45	6.47
S5	1.2	2.79	35.12	3.43	7.6	2.23	17.66	4.32	6.37
S6	1.1	1.57	40.12	1.69	0.7	1.6	20.71	1.15	0.67
MV	1.3	2.9	37.12	3.46	8.02	2.67	19.53	3.9	6.92

Table 8: Root Mean Square Error Rate

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	0.0628	0.08	0.22	0.09	0.08	0.08	0.28	0.07	0.22
S2	0.0838	0.1	0.34	0.1	0.11	0.1	0.31	0.1	0.34
S3	0.0938	0.11	0.35	0.11	0.12	0.11	0.32	0.11	0.35
S4	0.1038	0.12	0.36	0.12	0.13	0.12	0.33	0.12	0.36
S5	0.1138	0.13	0.37	0.13	0.14	0.13	0.34	0.13	0.37
S6	0.0638	0.08	0.32	0.08	0.07	0.08	0.29	0.08	0.32
MV	0.087	0.1	0.32	0.1	0.11	0.1	0.31	0.1	0.32

Analysis for Model (B):

Same Parameters defined in model (A) are used here in Model (B)

Table 9: Time Taken to Build the Model

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	0.02	0.8	0.9	2.41	56.43	0.09	0.18	0.11	2.25
S2	0.0502	0.12	0.92	3.43	66.45	0.1	0.14	0.12	1.27
S3	0.0625	0.13	0.93	3.44	66.46	0.11	0.15	0.13	1.28
S4	0.0599	0.13	0.93	3.44	66.46	0.11	0.15	0.13	1.28
S5	0.0573	0.13	0.93	3.44	66.46	0.11	0.15	0.13	1.28
S6	0.03	0.1	0.9	3.41	66.43	0.08	0.12	0.1	1.25
MV	0.0467	0.23	0.92	3.26	64.78	0.1	0.15	0.12	1.43

Table 10: Correctly classified Instances

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	99.4342	97.42	60.88	97.53	87.16	97.48	78.29	97.39	87.1
S2	99.12	96.43	69.88	97.31	92.3	97.4	82.29	95.35	92.33
S3	99.129	97.44	71.89	97.82	93.9	98.9	84.29	96.45	93.53
S4	99.2695	97.54	71.99	98.02	93.91	98.92	84.59	97.44	94.03
S5	99.3257	97.66	72.03	98.42	94.71	98.93	84.75	97.46	94.13
S6	99.07	98.48	59.88	98.54	89.16	98.48	79.29	98.39	89.1
MV	99.2247	97.49	67.75	97.94	91.85	98.35	82.25	97.08	91.7

Table 11: In Correctly classified Instances

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	0.366	2.52	39.12	2.46	12.84	2.52	12.71	2.21	12.9
S2	0.88	3.57	30.12	2.69	7.7	2.6	17.71	4.65	7.67
S3	0.871	2.56	28.11	2.18	6.1	1.1	15.71	3.55	6.47
S4	0.731	2.46	28.01	1.98	6.09	1.08	15.41	2.56	5.97
S5	0.674	2.34	27.97	1.58	5.29	1.07	15.25	2.54	5.87
S6	0.92	1.52	40.12	1.46	0.84	1.52	20.71	1.21	0.9
MV	0.74	2.5	32.25	2.06	6.48	1.65	16.25	2.79	6.63

Table 12: Root Mean Square Error Rate

SEM	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
S1	0.054	0.07	0.22	0.06	0.06	0.07	0.19	0.17	0.22
S2	0.084	0.1	0.34	0.09	0.07	0.1	0.31	0.09	0.34
S3	0.094	0.11	0.35	0.1	0.08	0.11	0.32	0.1	0.35
S4	0.104	0.12	0.36	0.11	0.09	0.12	0.33	0.11	0.36
S5	0.114	0.13	0.37	0.12	0.1	0.13	0.34	0.12	0.37
S6	0.05	0.08	0.32	0.07	0.06	0.08	0.29	0.07	0.32
MV	0.083	0.1	0.32	0.1	0.08	0.1	0.29	0.11	0.32

Summary of Analysis for Model A:

	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
T	0.07	0.13	0.1	10.4	78.2	0.1	0.14	0.16	2.9
C	98.6	97.2	62.8	96.5	90.4	97.3	80.63	95.9	91.5
I	1.3	2.9	37.1	3.4	8.02	2.6	19.53	3.9	6.9
R	0.08	0.1	0.32	0.1	0.11	0.1	0.31	0.1	0.3

T- Time Taken to build model
C- Correctly classified model
I- Incorrectly classified model
R- Root Relative Squared error

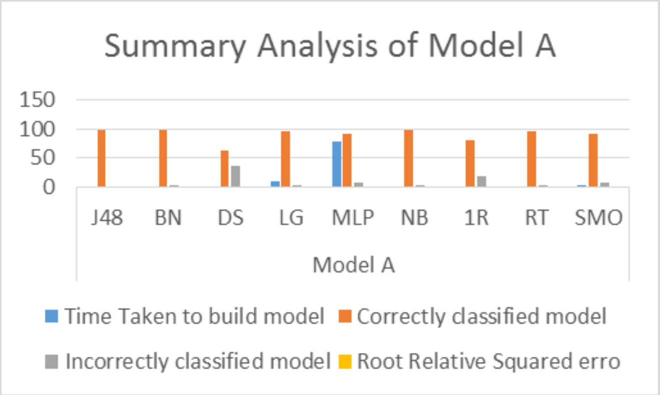

Figure 1: chart for summary analysis of model A

Summary of Analysis for Model B:

	J48	BN	DS	LG	MLP	NB	1R	RT	SMO
T	0	0.2	0.9	3.3	64.8	0.1	0.2	0.1	1.4
C	99.2	97.5	67.8	97.9	91.9	98.4	82.3	97.1	91.7
I	0.7	2.5	32.3	2.1	6.5	1.7	16.3	2.8	6.6
R	0.1	0.1	0.3	0.1	0.1	0.1	0.3	0.1	0.3

T- Time Taken to build model
C- Correctly classified model
I- Incorrectly classified model
R- Root Relative Squared error

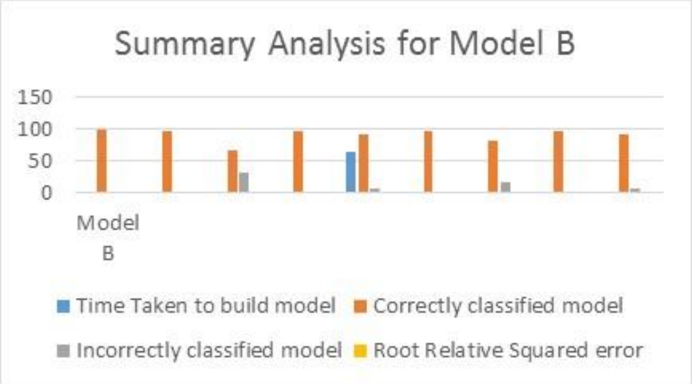

Figure 2: chart for summary analysis of model B

Result of Analysis

In this research paper, two models are used Model A and Model B. Model A Contains all the attributes of the dataset. Model B contains the only the statistically proved attributes those highly affected on the performance of students. In This research various data mining classification algorithms are used and these algorithms are applied on model A and model B and generate the semester wise result. These results are in the form of Accuracy of the classifiers and Error Rate of the classifiers. Theses generated results are compared and check that which algorithm is optimal for this types of dataset.After applying data mining algorithms on model A and model B. We have seen that in the both the model J48 algorithm gives the higher accuracy and Lower Error rate. Second Observation that if we select the only highly affected parameters on student performance in that case we got higher accuracy compared to by selecting all the attributes in the analysis.

VI. CONCLUSION AND FUTURE SCOPE

In this research paper, we have collected student data set with the student's learning behaviours, academic and demographic information. After that we have applied statistical technic to find the highly affected parameters on student performance. In this research paper, we have worked on two models model A and model B. In the model A we have used all the selected parameter while in model B we have taken only highly affected parameters on student performance. After that we have applied various classification algorithms on both of these model to predict the performances of students. In this research study we have used various classification algorithm to perform the analysis

The analysis result shows that tree based J48 algorithm is the best algorithm among all other algorithm for predicting the student's performance and second observation is that if we applied algorithm only on selected parameters in that case we got the higher accurate result rather than selection of all the algorithms. As a future scope we planned to develop a hybrid algorithm to get the higher accuracy as compared to J48.

VII. APPENDIX

LIST OF ABBREVIATION USED IN RESEARCH

ABBREVIATION	DESCRIPTION
W_L_H	WEEKLY LAB HOUR
W_LI_U	WEEKLY LIBRARY USAGE
D_RE_H	DAILY READING HOUR
E_W_L_U_H	EXTRA WEEKLY LAB USAGE HOUR
INT_TH	INTERNAL THEORY MARKS
INT_PR	INTERNAL PRACTICAL MARKS
EXT_TH	EXTERNAL THEORY MARKS
EXT_PR	EXTERNAL PRACTICAL MARKS
F_QUALIFICATIN	FATHER QUALIFICATION
M_QUALIFICATION	MOTHER QUALIFICATION
F_OCCUPATION	FATHER OCCUPATION
M_OCCUPATION	MOTHER OCCUPATION
F_ANNUAL INCOME	FATHER ANNUL INCOME
S1…..S6	SEMESTER 1 TO SEMESTER 6
DS	DECISION STUMP
LG, BN	LG – LOGISTIC, BN - BAYES NET
MLP	MULTI LAYER PERCEPTION
NB, 1R	NB - NAÏVE BAYES, 1R – ONE R
RT	RANDOM TREE

ACKNOWLEDGMENT

We are grateful to My Guide Dr. Jyotindra Dharwa for his guidance to improve my research work. We also thank to my friends and colleagues for their kindly support to complete my research paper timely.

REFERENCES

[1] Bhise R.B, Thorat S.S, Supekar A.K, "*Importance of Data Mining in Higher Education System"*, 2013.

[2] Monika Goyal ,Rajan Vohra2, "*Applications of Data Mining in Higher Education"*, 2012.

[3] K.Shanmuga Priya, A.V.Senthil Kumar, *"Improving the Student's Performance Using Educational Data Mining"*, 2013.

[4] Varun Kumar, Anupama Chadha, "*Mining Association Rules in Student's Assessment Data"*, 2012.

[5] Mahendra Tiwari, Yashpal Singh (2012), *"Performance Evaluation of Data Mining clustering algorithms in WEKA",* Global Journal of Enterprise Information System, vol 4, issue I

[6] Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, *"Data Mining in Education: Data Classification and Decision Tree Approach",* 2012.

[7] Surjeet Kumar Yadav, Brijesh Bharadwaj, and Saurabh Pal, "*Mining Education Data to Predict Student's Retention: A comparative Study"*, 2012.

[8] Brijesh Kumar Baradwaj, Saurabh Pal, "*Mining Educational Data to Analyze Students Performance"*, 2011.

[9] Pallamreddy.venkatasubbareddy, Vuda Sreenivasarao, *"The Result Oriented Process for Students Based On Distributed Data Mining",* 2010.

[10] *Tanuja S, Dr. U. Dinesh Acharya, and Shailesh K R (2011) , "Comparison of different data mining techniques to predict hospital length of Stay, Journal of Pharmaceutical and Biomedical Sciences (JPBMS)",* Vol. 07, Issue 07

Authors Profile

Prof. B. R. Patel, Ph.d (Pursuing), M.Phil (CS), M.C.A. Working as assistant professor in Ganpat University, M.C.A Department. Area of Interest is Data Mining, Software Engineering. Presented four National levels Paper. He has seven year teaching experience in the field of academic.