EDUCATIONAL DATA MINING USING R PROGRAMMING AND R STUDIO

Sadiq Hussain*
Dibrugarh University,Dibrugarh-786004
*For correspondence. (sadiq@dibru.ac.in)

Abstract: Data Mining is the extraction of knowledge from the large databases. Data Mining had affected all the fields from combating terror attacks to the human genome databases. For different data analysis, R programming has a key role to play. R Studio, an effective GUI for R Programming is used extensively for generating reports based on several current trends models like random forest, support vector machine etc. It is otherwise hard to compare which model to choose for the data that needs to be mined. This paper analyses the performance of B.A. students of Dibrugarh University with respect to caste and gender.

Keywords: Educational Data Mining; R programming, R studio; one-way ANOVA; box-plot

1. Introduction:

Dibrugarh University, the easternmost University of India was set up in 1965 under the provisions of the Dibrugarh University Act, 1965 enacted by the Assam Legislative Assembly. It is a teaching-cum-affiliating University with limited residential facilities. The University is situated at Rajabheta at a distance of about five kilometers to the south of the premier town of Dibrugarh in the eastern part of Assam as well as India. Dibrugarh, a commercially and industrially advanced town in the entire northeastern region also enjoys a unique place in the fields of Art, Literature and Culture. The district of Dibrugarh is well known for its vast treasure of minerals (including oil and natural gas and coal), flora and fauna and largest concentration of tea plantations. The diverse tribes with their distinct dialects, customs, traditions and culture form a polychromatic ethnic mosaic, which becomes a paradise for the study of Anthropology and Sociology, besides art and culture. The Dibrugarh University Campus is well linked by roads, rails, air and waterways. The National Highway No.37 passes through the University Campus. The territorial jurisdiction of Dibrugarh University covers seven districts of Upper Assam, viz, Dibrugarh, Tinsukia, Sivasagar, Jorhat, Golaghat, Dhemaji and Lakhimpur. [1]

There are more than hundred numbers of Colleges/ Institutes offering TDC (Three Year Degree) Course affiliated/ permitted under the Dibrugarh University. Since the number of students in the Arts Stream is larger in comparison to the other stream (B.Sc., B.Com., B.Tech. etc), we considered the data for the B.A. (Bachelor of Arts) course for our present study of educational data mining. The required digitized data are collected from Dibrugarh University Examination Branch for the affiliated colleges of the University B.A. programme from 2011 to 2013. This paper evaluates performance gender wise as well as caste wise of the students. There are several data mining tools and statistical models available. This paper focuses on R programming and tried to fit the statistical models for such knowledge discovery.

2. Literature review:

2.1. Data mining:

Data Mining detects the relevant patterns from databases / data warehouses using different programs and algorithms to look into current and historical data which can be analyzed to predict future trends [2]. It is very difficult for any organization to extract hidden patterns from the huge data marts and data warehouses without the help of data mining tools and programs. It is like searching for the pearls in the sea of data. This knowledge set is extremely useful in developing a knowledge support system and making important decisions regarding the future trends predictions.

Statisticians have used different manual techniques for the benefit of the business, predicting trends and results based on data over the years. The business houses had developed huge databases or data warehouses to become

"data tombs". The data was never transformed into information. But with the help of data mining tools and algorithms now professionals from different areas may extract knowledge quickly and at ease.

2.2. Educational data mining:

Data mining, often called knowledge discovery in database (KDD), is known for its powerful role in uncovering hidden information from large volumes of data [3]. Its advantages have landed its application in numerous fields including e-commerce, bioinformatics and lately, within the educational research which commonly known as Educational Data Mining (EDM) [4]. EDM is defined by The Educational Data Mining community website, www.educationaldatamining.org as an emerging discipline, concerned with developing methods for exploring the unique types of data that come from the educational setting, and using those methods to better understand students, and the settings which they learn in. EDM often stresses with the improvement of student models which denote the student's current knowledge, motivation and attitudes [5].

2.3. R Programming:

R is a programming language for the purpose of statistical computations and data analysis. The R language is widely used by the data miners and statisticians on high dimensional pattern extraction. R's popularity has increased substantially in recent years which proved by the polls and surveys.

R is inspired by Scheme and an implementation of S programming language combined with lexical scoping semantics. S was designed by John Chambers while at Bell Labs. The creator of R was Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. R is named after the first names of the two R creators.
R is freely available under the GNU General Public License and the source code is written in FORTRAN, C and R. It is a GNU project. The pre-compiled binary versions are freely available for various flavors of operating system. R is basically command line interface (CLI) and various GUI interfaces are also available nowadays.

R provides numerous statistical techniques from modeling to analysis, clustering, classification and the list goes on. The packages developed by the R community plays an important role in this regards. The C,C++, Java, .NET or Python programmers may write their own code to manipulate the R objects. Advanced users may use algorithms of their choices for any computationally intensive tasks.

Graphical packages are also available in R. R produces dynamic, interactive and publication quality graphs for the data miners and statisticians.

2.4. R Studio:

R Studio is a free and open source integrated development environment (IDE) for R, the statistical computing language for the data miners. There are two editions of R Studio. One is R Studio Server, which may be accessed through web browser from a remote Linux Server. Another Edition is R Studio Desktop which available for Microsoft Windows, Mac OS X, and Linux. R Studio Desktop runs locally. R Studio uses the Qt framework for the GUI and is written in C++ language.

2.5. Use of analysis of variance (ANOVA) in R programming:

The analysis of variance is a commonly used method to determine differences between several samples. R provides a function to conduct ANOVA so: aov(model, data).The first stage is to arrange your data in a .CSV file. Use a column for each variable and give it a meaningful name. Don't forget that variable names in R can contain letters and numbers but the only punctuation allowed is a period. One needs to set out data file so that each column represents a factor in one's analysis. Usually the 1st column will be dependent variable and subsequent columns would be the independent factors. The second stage is to read your data file into memory and give it a sensible name. The next stage is to attach the data set so that the individual variables are read into memory. Finally it is required to define the model and run the analysis.

3. Experiments and evaluation

3.1. The data set:

We have included a small part of the Category and Gender based tables termed as Table 1 for the data analysis. The Examination Branch of Dibrugarh University provides various College Codes for different Colleges under its jurisdiction. The field 'ExamName' means the Final Examination B.A. Part-III. The 'Year' means the Year of Examination, 'Gender' means the Sex of the Candidate. The Category is terms as 'General','SC','ST' or 'OBC' category of candidates and 'CollegeCode' means the code of the College. The field 'PassPercentage' is the aggregate Percentage of the particular Candidate for all the three year examinations viz, B.A. Part-I, II and III Examinations. The Result is I for 'First Class',II for 'Second Class' and P for 'Simple Pass'. If the Pass Percentage is above 60%, it is First Class, above 45%, it is Second Class and below that it is 'Simple Pass'. The percentage for passing in the B.A. Examination changes over time between 30 and 40.

Table 1: College-wise Gender-wise Category-wise Pass Percentage of B.A. Third Year Candidates

Exam Name	Year	Gender	Category	CollegeCode	PassPercentage	Result
B.A. PART-III	2011	F	General	125	46.56	II
B.A. PART-III	2011	M	General	125	34.33	P

The Following is the SQL Procedure to extract the 1000 randomly selected First Class and Second Class students extracted from the Microsoft SQL Server Database.

```
--1000 random, First/Second class in Major
create OR alter procedure TDC_QuerryClass
@Year varchar(4),                --like 2013, 2014
@CoCode varchar(5),              --BA or BSc or BCom
@Class varchar(6),               --ID /I /II / or 'ID,I' meaning ID + I
@Gender varchar(5),      --(optional) M/F
@Cate varchar(10)                --(optional) Cate Code Like ST, SC, MOBC
as
set nocount on
if exists (select * from dbo.sysobjects where id = object_id(N'[tmpData]') and OBJECTPROPERTY(id, N'IsUserTable') = 1)
drop table [tmpData]
declare @ShId int, @EmId int
select @ShId = ShId, @EmId = EmId
from Schedules, Sessions,
(select EmId = Max(EmId) from Exams where EmCoCode = @CoCode) as Em
where ShEmId = EmId and SnNo = ShSnNo and SnYear = @Year
select RgId, EfId, ExamName = EmName, ExamYear = @Year, CandName = RgName,
        Gender = RgSex, Category = CdDesc, ColgCode = RgCgCode,
        Sub = StrCode, MajorSub = StrName, TotalMarks = EfAObtMarks,
        FirstYr = 0, SecondYr = 0, ThirdYr = 0, OverallPC = EfDivPc,
        Division = EfDiv
into tmpData
from ExamStr, ExamForms, RegnForms, Exams, CodeTable
where StrCoCode = @CoCode and StrCode <> 'GEN'
        and EfShId = @ShId and EfStrId = StrId
        and EfResult = 'P' and EfExcp = ''
        and RgId = EfRgId and EmId = @EmId
        and CdType = 'CS' and CdSeq <> 0 and CdCode = RgCaste

update tmpData set Sub = EsCode, MajorSub = EsName
from FormSub, ExamSub
where FsEfId = EfId and EsId = FsEsId
        and EsType = 'S' and EsName like 'MAJOR%'

update tmpData set FirstYr = ObtMarks
```

```
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 1

update tmpData set SecondYr = ObtMarks
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 2
update tmpData set ThirdYr = ObtMarks
from ResAggr, Exams
where DivSh = @ShId and Em = EmId and Rg = RgId
        and EmCoCode = @CoCode and EmSrNo = 3
select top 1000
        ExamName, ExamYear, CandName, Gender, Category, ColgCode,
        MajorSub, TotalMarks, FirstYr, SecondYr, ThirdYr, OverallPC, Division
from tmpData
where (',' + @Class + ',' like '%,' + Division + ',%' or @Class = '')
        and (Gender = @Gender or @Gender = '')
        and (Category = @Cate or @Cate = '')
order by EfId + RgId
return (0)
```

3.2. The R Code for the experiment:

The .csv file gender_cat is imported to the R studio for finding out the performance parameter in respect to gender and category i.e. General, SC,ST, OBC. F1 means that performance of the Female students of the year 2011, M1 means that the performance of the Male candidates of the year 2011 and so on. GEN1 means that the performance of the General candidates of the year 2011 and so on.

Anova Test w.r.t. Gender:

```
> gender_ba1 <- read.csv("C:/Documents and Settings/sadiq/Desktop/reseach/gender_ba1.csv")
>   View(gender_ba1)
> Percentage <- as.double (gender_ba1$Percentage)
> Gender <- as.factor (gender_ba1$Gender)
> results <- aov (Percentage~Gender)
> summary (results)
             Df Sum Sq Mean Sq F value   Pr(>F)
Gender        5   7645  1529.1   16.25 5.95e-16 ***
Residuals   5994 564100    94.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(results)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> TukeyHSD (results)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Percentage ~ Gender)

$Gender
         diff       lwr       upr     p adj
F2-F1  0.22932 -1.0074117  1.4660517 0.9950432
F3-F1  0.29798 -0.9387517  1.5347117 0.9834540
M1-F1 -2.56277 -3.7995017 -1.3260383 0.0000001
```

M2-F1 -1.86491 -3.1016417 -0.6281783 0.0002524
M3-F1 -1.55546 -2.7921917 -0.3187283 0.0045764
F3-F2 0.06866 -1.1680717 1.3053917 0.9999863
M1-F2 -2.79209 -4.0288217 -1.5553583 0.0000000
M2-F2 -2.09423 -3.3309617 -0.8574983 0.0000210
M3-F2 -1.78478 -3.0215117 -0.5480483 0.0005628
M1-F3 -2.86075 -4.0974817 -1.6240183 0.0000000
M2-F3 -2.16289 -3.3996217 -0.9261583 0.0000094
M3-F3 -1.85344 -3.0901717 -0.6167083 0.0002837
M2-M1 0.69786 -0.5388717 1.9345917 0.5928795
M3-M1 1.00731 -0.2294217 2.2440417 0.1853217
M3-M2 0.30945 -0.9272817 1.5461817 0.9804082

> pairwise.t.test (Percentage,Gender,p.adjust.method="holm")

Pairwise comparisons using t tests with pooled SD

data: Percentage and Gender

	F1	F2	F3	M1	M2
F2	1.00000	-	-	-	-
F3	1.00000	1.00000	-	-	-
M1	4.8e-08	1.9e-09	7.0e-10	-	-
M2	0.00017	1.6e-05	7.6e-06	0.53884	-
M3	0.00238	0.00032	0.00018	0.12166	1.00000

P value adjustment method: holm
>boxplot (Percentage~Gender)

In the figure below, the Y-axis shows the percentage of pass, and X-axis shows F1 means that performance of the Female students of the year 2011, M1 means that the performance of the Male candidates of the year 2011 and so on.

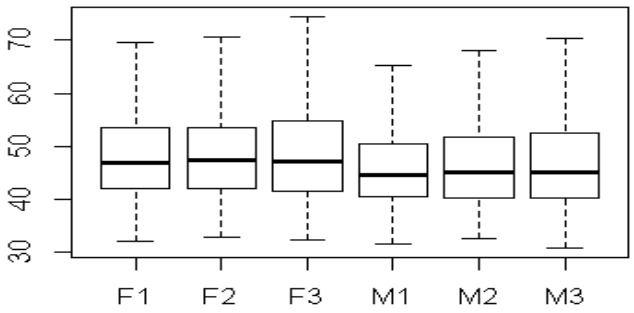

Figure 1: Box plot for year wise Gender wise B.A. Candidates

Anova Test w.r.t. Caste:

> ba_category2 <- read.csv("C:/Documents and Settings/sadiq/Desktop/reseach/Category/ba_category2.csv")
> View(ba_category2)
> Percentage <- as.double (ba_category2$Percentage)
> Category <- as.factor (ba_category2$Category)
> res1 <- aov (Percentage~Category)
> summary (res1)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Category	11	15796	1436.0	13.66	<2e-16 ***
Residuals	10142	1066327	105.1		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> plot(res1)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
> TukeyHSD (res1)
Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = Percentage ~ Category)

$Category

	diff	lwr	upr	p adj
GEN2-GEN1	0.000000000	-1.4989367	1.49893674	1.0000000
GEN3-GEN1	-0.376425912	-1.8761134	1.12326161	0.9996413
OBC1-GEN1	-1.802368899	-3.3016807	-0.30305710	0.0048797
OBC2-GEN1	-1.802368899	-3.3016807	-0.30305710	0.0048797
OBC3-GEN1	-1.797167395	-3.2968549	-0.29747988	0.0051220
SC1-GEN1	-1.599655567	-3.6043453	0.40503421	0.2748487
SC2-GEN1	-2.297165090	-4.3037200	-0.29061016	0.0099925
SC3-GEN1	-2.299376736	-4.3078047	-0.29094873	0.0099884
ST1-GEN1	-3.670100000	-5.1690367	-2.17116326	0.0000000
ST2-GEN1	-3.024240000	-4.5231767	-1.52530326	0.0000000
ST3-GEN1	-3.247433964	-4.7467458	-1.74812217	0.0000000
GEN3-GEN2	-0.376425912	-1.8761134	1.12326161	0.9996413
OBC1-GEN2	-1.802368899	-3.3016807	-0.30305710	0.0048797
OBC2-GEN2	-1.802368899	-3.3016807	-0.30305710	0.0048797
OBC3-GEN2	-1.797167395	-3.2968549	-0.29747988	0.0051220
SC1-GEN2	-1.599655567	-3.6043453	0.40503421	0.2748487
SC2-GEN2	-2.297165090	-4.3037200	-0.29061016	0.0099925
SC3-GEN2	-2.299376736	-4.3078047	-0.29094873	0.0099884
ST1-GEN2	-3.670100000	-5.1690367	-2.17116326	0.0000000
ST2-GEN2	-3.024240000	-4.5231767	-1.52530326	0.0000000
ST3-GEN2	-3.247433964	-4.7467458	-1.74812217	0.0000000
OBC1-GEN3	-1.425942987	-2.9260054	0.07411941	0.0806853
OBC2-GEN3	-1.425942987	-2.9260054	0.07411941	0.0806853
OBC3-GEN3	-1.420741483	-2.9211794	0.07969644	0.0835041
SC1-GEN3	-1.223229655	-3.2284809	0.78202155	0.6977560
SC2-GEN3	-1.920739179	-3.9278550	0.08637666	0.0760164
SC3-GEN3	-1.922950824	-3.9319392	0.08603757	0.0758672
ST1-GEN3	-3.293674088	-4.7933616	-1.79398657	0.0000000
ST2-GEN3	-2.647814088	-4.1475016	-1.14812657	0.0000005
ST3-GEN3	-2.871008052	-4.3710704	-1.37094566	0.0000000
OBC2-OBC1	0.000000000	-1.4996868	1.49968677	1.0000000
OBC3-OBC1	0.005201504	-1.4948609	1.50526390	1.0000000
SC1-OBC1	0.202713332	-1.8022569	2.20768356	1.0000000
SC2-OBC1	-0.494796192	-2.5016313	1.51203893	0.9996981
SC3-OBC1	-0.497007837	-2.5057158	1.51170010	0.9996876
ST1-OBC1	-1.867731101	-3.3670429	-0.36841930	0.0027492
ST2-OBC1	-1.221871101	-2.7211829	0.27744070	0.2446099
ST3-OBC1	-1.445065065	-2.9447518	0.05462170	0.0714104
OBC3-OBC2	0.005201504	-1.4948609	1.50526390	1.0000000
SC1-OBC2	0.202713332	-1.8022569	2.20768356	1.0000000
SC2-OBC2	-0.494796192	-2.5016313	1.51203893	0.9996981
SC3-OBC2	-0.497007837	-2.5057158	1.51170010	0.9996876
ST1-OBC2	-1.867731101	-3.3670429	-0.36841930	0.0027492
ST2-OBC2	-1.221871101	-2.7211829	0.27744070	0.2446099

```
ST3-OBC2  -1.445065065 -2.9447518  0.05462170 0.0714104
SC1-OBC3   0.197511828 -1.8077394  2.20276304 1.0000000
SC2-OBC3  -0.499997696 -2.5071135  1.50711815 0.9996665
SC3-OBC3  -0.502209341 -2.5111977  1.50677905 0.9996551
ST1-OBC3  -1.872932605 -3.3726201 -0.37324509 0.0026348
ST2-OBC3  -1.227072605 -2.7267601  0.27261491 0.2390538
ST3-OBC3  -1.450266569 -2.9503290  0.04979582 0.0692485
SC2-SC1   -0.697509523 -3.1054614  1.71044240 0.9986180
SC3-SC1   -0.699721169 -3.1092342  1.70979181 0.9985857
ST1-SC1   -2.070444433 -4.0751342 -0.06575466 0.0356523
ST2-SC1   -1.424584433 -3.4292742  0.58010534 0.4600983
ST3-SC1   -1.647778397 -3.6527486  0.35719184 0.2329746
SC3-SC2   -0.002211645 -2.4132766  2.40885335 1.0000000
ST1-SC2   -1.372934910 -3.3794898  0.63362002 0.5227432
ST2-SC2   -0.727074910 -2.7336298  1.27948002 0.9901911
ST3-SC2   -0.950268874 -2.9571040  1.05656625 0.9270876
ST1-SC3   -1.370723264 -3.3791513  0.63770474 0.5269143
ST2-SC3   -0.724863264 -2.7332913  1.28356474 0.9905087
ST3-SC3   -0.948057228 -2.9567652  1.06065071 0.9286505
ST2-ST1    0.645860000 -0.8530767  2.14479674 0.9621892
ST3-ST1    0.422666036 -1.0766458  1.92197783 0.9989246
ST3-ST2   -0.223193964 -1.7225058  1.27611783 0.9999982

> pairwise.t.test (Percentage,Category,p.adjust.method="holm")

        Pairwise comparisons using t tests with pooled SD

data:  Percentage and Category
GEN1    GEN2    GEN3    OBC1    OBC2    OBC3    SC1     SC2     SC3     ST1     ST2
GEN2 1.0000  -       -       -       -       -       -       -       -       -
GEN3 1.0000  1.0000  -       -       -       -       -       -       -       -
OBC1 0.0046  0.0046  0.0720  -       -       -       -       -       -       -
OBC2 0.0046  0.0046  0.0720  1.0000  -       -       -       -       -       -
OBC3 0.0046  0.0046  0.0720  1.0000  1.0000  -       -       -       -       -
SC1  0.2825  0.2825  1.0000  1.0000  1.0000  1.0000  -       -       -       -
SC2  0.0088  0.0088  0.0704  1.0000  1.0000  1.0000  1.0000  -       -       -
SC3  0.0088  0.0088  0.0704  1.0000  1.0000  1.0000  1.0000  1.0000  -       -
ST1  8.9e-14 8.9e-14 4.8e-11 0.0026  0.0026  0.0026  0.0325  0.7094  0.7094  -
ST2  2.7e-09 2.7e-09 4.7e-07 0.2553  0.2553  0.2548  0.5860  1.0000  1.0000  1.0000  -
ST3  9.7e-11 9.7e-11 2.4e-08 0.0688  0.0688  0.0680  0.2532  1.0000  1.0000  1.0000  1.0000

P value adjustment method: holm
> boxplot(Percentage~Category)
```

In the figure below, the Y-axis shows the percentage of pass, and X-axis GEN1 means that performance of the General students of the year 2011, SC1 means that the performance of the Schedule Caste candidates of the year 2011 and so on.

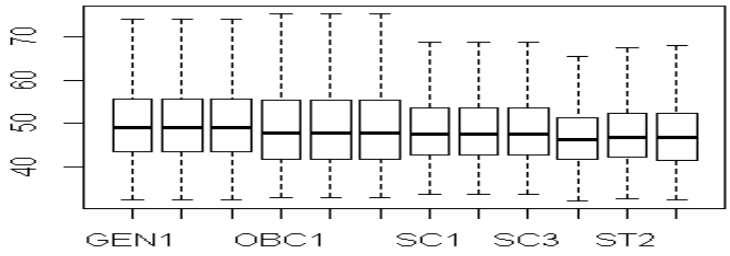

Figure 2: Box plot for year wise Caste wise B.A. Candidates

Anova Test w.r.t. the results of the candidates (Second Class):

```
> CategorySecondClass <- read.csv("C:/Documents and
Settings/sadiq/Desktop/reseach/Category/CategorySecondClass.csv")
>   View(CategorySecondClass)
> Percentage <- as.double (CategorySecondClass$Percentage)
> Category <- as.double (CategorySecondClass$Category)
> res1 <- aov (Percentage~Category)
> summary (res1)

             Df Sum Sq Mean Sq F value  Pr(>F)
Category      1    195  195.07   7.007 0.00815 **
Residuals   3769 104922   27.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

4. Conclusion and future study:

Depending on the p value from the above results, the author may conclude that there is statistically significant difference between the male and female candidates. From statistical analysis, it is also clear that the performance of the female candidates is better than the male candidates. Again on the basis of p value, the author observed that there is statistically significant difference among the performance of the candidates caste-wise. From statistical analysis, it is confirmed that the OBC students performance are better than the other category students. The results of the candidates further analyzed by grouping as the First Class Candidates and Second Class Candidates. It was found that the statistical difference between male and female candidates and among the caste of the candidates were significant because of the results of the Second Class Students. For future study, the author plans to evaluate the performance of other examinations of Dibrugarh University viz. B.Sc. and B.Com. with more parameters e.g. comparison of results of the H.S.L.C. and H.S. Examinations etc.

Acknowledgments:

The author expresses his gratefulness to Prof. Alak Kr. Buragohain, Vice-Chancellor, Dibrugarh University for his inspiring words and allowing us to use the Examination data of the University. We generously thank Mr. N.A. Naik, Senior Programmer, Mumbai based firm for helping us to extract the .csv files from the SQL Server database.

References:

[1] The Dibrugarh University website: www.dibru.ac.in
[2] John Silltow, (2006): Data Mining 101: Tools and Techniques, http://www.internalauditoronline.org/
[3] Witten, I.H. and Frank, E. (1999). Data Mining:Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kauffman, San Francisco, CA.
[4] Baker, R.S.J.d. (2010): Data Mining for Education. In: McGaw, B., Peterson, P., Baker, E. (eds.) To appear in International Encyclopedia of Education, 3rd edn. Elsevier, Oxford
[5] Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1(1), 3-17.
[6] Graham Williams(2009): Rattle: A Data Mining GUI for R, The R Journal Vol. 1/2, December 2009 ISSN 2073-4859.