International Journal of Information Science and Education.
ISSN 2231-1262 Volume 2, Number 1 (2012) pp. 21-41

http://www. ripublication.com

Predicting Student Academic Performance Using Temporal Association Mining

***[1]S.K. Althaf Hussain Basha, [2]Y.R. Ramesh Kumar, [3]A. Govardhan and [4]Mohd. Zaheer Ahmed**

[1]Asst. Professor,
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India
E-mail: althafbashacse@gmail.com
[2]Asst. Professor & Head, University Arts & Science College,
Kakatiya University, Warangal, India
E-mail: yrrkumar@yahoo.co.in
[3]Professor in CSE & Director of Evaluation, JNTUH, Hyderabad, India
E-mail: govardhan_cse@yahoo.co.in
[4]Kakatiya University, Warangal, India
E-mail: heera.2007@yahoo.co.in

Abstract

Predicting the performance of a student is of great concern to the higher education managements. In the real world, predicting the performance of the students is a challenging task. Many of the well known technical colleges are successful as they have meritorious students and faculty with them and a foot proof system working for them to grow continuously. The primary goal of Data mining in practice tends to be Prediction and Description. In this paper, we proposed a method to predict the intra-year Academic Performance of the student using the historic data. The idea is to identify existing patterns in the historic data, and maintain a database for it. Comparing the current performance of a student with the existing patterns, we predict the possible performance of the student in the future. In the process, we identify any new patterns that we come across. For the purpose of identifying the interestingness of the pattern we look into the percentage of increase or decrease we make, appropriate methodologies are used.

Keywords: Academic Student Performance, Data Mining, Higher Education, Temporal Association Mining, Prediction

Introduction

In recent years due to the rapid development of technology the amount of data has been growing tremendously in all areas. The need of discovering novel and most useful information from these large amounts of data has also grown. With the advent of data mining, different mining techniques have been applied in different application domains, such as, Education, banking, retail sales, bioinformatics, and Telecommunications. To extract useful information to fulfill the needs of the industry. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, though not necessary, to develop a powerful means for analysis as well as interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. It is intended to obtain meaningful and valuable information that is not previously known from these data by applying data mining techniques [1].

One of the significant facts in higher learning institutions is the explosive growth of educational data. These data are increasing rapidly without any benefit to the management. The main objective of any higher educational institution is to improve the quality of managerial decisions and to impart quality education. Good prediction of student’s success in higher learning institution is one way to reach the highest level of quality in the higher education system. Many prediction models available with a difference in approach to student performance were reported by the researcher, but there is no certainty that there are any predictors who can accurately determine whether a student will be an academic genius, a drop out, or an average performer.

In the ever changing global environment, the demand for the educated work force to meet the requirements is very high. Now-a- days, the important challenge is to strengthen the Universities and Educational Institutions to have more efficient, effective and accurate educational processes. With the vast available data, data mining is considered as the most suited technology appropriate to give additional insight into the teacher, student, alumni, manager, and other educational staff behavior and acting as an active automated assistant in helping them for making better decisions on their educational activities.

The higher education institutions use automated computer programs/tools developed with different technologies to predict the trades in the college. With the potential techniques in Data Mining and with the growth of technologies to handle huge databases, the predictive technologies have started growing tremendously. The academic research in Data Mining also contributed a lot to predictive technologies. The use of Data Mining is well founded on the theory that the historic data holds essential hidden and previously unknown knowledge that can be used for predicting the future direction and assist in decision making. The prediction of academic performance is regarded as a challenging task of temporal data prediction. Data analysis is one way of predicting increase or decrease of future academic performance.

The main objective of this paper is to use temporal association mining for identifying patterns in the student data and to predict the intra year Academic Performance of Student using the historic data (Predicting future value). As the

Academic environment hardly changes the prediction is based mainly on the historic data.

In this paper we are using temporal association mining to bring out the prediction hidden in the data. Our algorithm is the most preferred one for this purpose. Here we have used seven years of Under Graduate data of Kakatiya University, Warangal, from 2002 to 2007. The data has been preprocessed to suit the needs of our mining activity.

The rest of the paper is organized as follow: Section 2 discusses related work on Higher Education in data mining. Section 3 discusses on Temporal Association Rule Mining. Section 4 discusses Problem Statement. Section 5 discusses the Proposed Approach. Section 6 discusses Experiments results and analysis. Finally this paper ends with conclusions and directions for future work.

Related Work

The application of Data mining widely spreaded in Higher Education system. This have been in Education domain there have been many the researchers and authors have been explored and discussed various applications of data mining in higher education. The authors had gone through the survey of the literature to understand the importance of data mining applications in higher education, the use of data mining to investigate scientific questions within educational research for the quality improvements in this area.

V. Ramesh et al investigated the accuracy of Naïve Bayes Simple, Multilayer Perception, SMO, J48, REP Tree techniques for predicting student performance. From the results obtained they proved that Multilayer Perception algorithm is most appropriate for predicting student performance. MLP gives 87% prediction which is relatively higher than other algorithms. This study is an attempt to use classification algorithms for predicting the student performance and comparing the performance of NaiveBayesSimple, Multilayer Perception, SMO, J48, and REPTree [2].

Cortez and Silva [3] attempted to predict failure in the two core courses, namely Mathematics and Portuguese of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms, such as, Decision Tree (DT), Random Forest (RF), Neural Network (NN) and Support Vector Machine (SVM) were applied on a data set of 788 students, who sat for 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was also reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

Erdogan and Timor 2005 et al used educational data mining to identify and enhance educational process that can improve their decision making process. Finally Henrik, 2001 et al observed that clustering was effective in finding hidden relationships and associations between different categories of students[4].

Kotsiantis, et al [5] applied five classification algorithms namely, Decision Trees, Perceptron-based Learning, Bayesian Nets, Instance-Based Learning and Rule-learning, to predict the performance of computer science students from distance learning stream of Hellenic Open University, Greece. A total of 365 student records comprising several demographic variables like sex, age and marital status were used. In addition, the performance attribute, namely the marks in a given assignment was used as input to a binary (pass/fail) classifier. Filter based variable selection technique was used to select highly influencing variables and all the above five classification models were constructed. It was noticed that the Naïve-Bayes yielded high predictive accuracy (74%) for two-class (pass/fail) dataset.

Khan [6] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India, with the objective of establishing the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in the science stream. The selection was based on cluster sampling technique in which the entire population with interest was divided into groups, or clusters, and a random sample of these clusters was selected for further analyses. It was found that the girls with high socio-economic status had relatively higher academic achievement in the science stream and boys with low socio-economic status had relatively higher academic achievement in general.

Cristóbal Romero[7], et al compared different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses, and developed a specific mining tool for making the configuration and execution of data mining techniques easier for instructors. They also used real data from seven Moodle courses with Cordoba University students, also applied discretization and rebalance preprocessing techniques on the original numerical data in order to verify if better classifier models could be obtained. A classifier model appropriate for educational use has to be both accurate and comprehensible for instructors in order to be of use for decision making.

M.N. Quadri et al [8] have predicted student’s academic performance using the CGPA grade system where the data set comprised the students gender, his parents educational details, his financial background and so on. In [9] the author explored the various variables to predict the students who are at the risk of failing in the exam. The solution strongly suggests that the previous academic result strongly plays a major role in predicting their current outcome.

Sajadin Sembiring [10] et al applied the kernel method as data mining techniques to analyze the relationships between students behavior and their success, and to develop the model of student performance predictors .This is done by using Smooth Support Vector Machine (SSVM) classification and kernel k-means clustering techniques. The results of this study reported a model of student academic performance predictors by employing psychometric factors as variable predictors.

Temporal Association Rule Mining

The non-trival extraction of implicitly unknown and potentially useful information from data is dealt with here. The ultimate goal of temporal data mining is to discover hidden relations between sequences and subsequences of events.

Recent advances in data collection and storage technology have made it possible to collect vast amounts of data everyday in many areas of business and science. Examples are recordings of sales of products, stock exchanges, web logs, climate measures, and so on. One major area of data mining from these data is association pattern analysis. Association rules discover interrelationships among various data items in transactional data. Following the work of Agarwal and Srikant [11], the discovery of association rules has been extensively studied in [12], [13], [14], and [15]. In particular, in [16], [17], [18], they have paid attention to temporal information, which is implicitly related to transaction data, for example, the time that a transaction is executed, and discovered association patterns that vary over time. However, most works in temporal association mining have focused on special temporal regulation patterns of associated item sets such as cyclic patterns [16] and calendar-based patterns [11]. For example, it may be found that beer and chips are sold together primarily in the evening time on week days.

Other different DM techniques have been applied to provide feedback, such as, domain specific interactive data mining to find the relationships between log data and student behavior in an educational hypermedia system [19]; temporal data mining to describe, interpret and predict student behavior, and to evaluate progress in relation to learning outcomes in ITSs [20].

The Educational data is continuous time stamped data and a time series data. For the purpose of prediction we need to use a huge data set. Temporal data mining is of latest origin concerned with data mining of large sequential data. It is useful in discovering qualitative and quantitative temporal patterns in a temporal database or in a discrete valued time series dataset. Although there is no notion of time as such, the ordering among the records is very important and is central to the data description or modeling. Temporal data mining, however, is somewhat different with constraints and objectives rather than the traditional time series data. One main difference lies in the size of data sets and the way it is collected with little or no control over gathering process. Often the methods must be capable of analyzing large data sets. The second major difference lies in the kind of information that we want to estimate or unearth from the data, like trends and patterns in the data which are easily interpretable.

Problem Statement

Given the students admission data, predict their Performance in the current year and further identify possible result category wise, and grade wise on the previous year's performance.

In this process we are assuming that the result processing environment does not change and further the college continues to have the same academic environment.

Here, we wanted to study the students' academic performance in different social group categories in rural and urban areas in Government and private sector colleges, and different courses using Temporal association rule mining. The Data has been collected from 298 affiliated undergraduate colleges affiliated to Kakatiya University , over a period of six year from 2002to 2007.

Proposed Approach
Six years of data has been collected from the examinations branch of Kakatiya University for the purpose of this study. The collected data was associated only with examinations and hence several other data was also needed to be collected relating to the student's social status, the type and location of the college, and so on. The overall activities are broadly categorized into the following steps:
- Data collection and Data set preparation.
- Data preprocessing.
- Data processing.
- Results & Analysis.

Data Collection and Data set Preparation
We have collected student data set from 294 affiliated colleges of Kakatiya University from 2002 to 2007. The data set contains the result and marks for B.Sc.(M), B.Sc.(B),B.Com, B.A. Courses from these colleges. There are approximately 5,00,000 records in this data set. Further the personal data of the students containing their social status has been collected from the colleges. The data relating to the type of the college and the location (Rural/Urban), is added to this data. After combining all these data sets the resultant database record contains fourteen attributes, such as, different social groups and their categories, rural and urban areas, and Government and Private sector colleges in different courses of each student. As the data collected is from different sources, there needs to be a proper cleaning of data, such as, filling in missing values; smoothing noisy data, identifying or removing outliers, and resolving inconsistencies. Then, the cleaned data are transformed into a form of table that is suitable for data mining model.

Data Preprocessing
The data collected and brought together is very huge and contains a lot of unwanted details. The basic data has the following information.

Attribute list

Table 1: Data Structure of the Basic Data

SNO	ATTRIBUTE NAME	TYPE	DESCRIPTION
1	YEAR	Number	Year
2	DIST	Character	District Name
3	TOWN	Character	Town Name
4	CODE	Numeric	College code
5	CAT1	Character	'R' for Rural area , 'U' for Urban area
6	CAT2	Character	'G' for Government sector College, 'P' for Private sector college
7	COURSE	Character	'BSCB ' for BSc(Bio.sc.) Course, BSCM ' for BSc(Maths) Course

8	COLLEGE	Character	College Name
9	Regd No	Number	Student Hall Ticket Number
10	NAME	Character	Student Name
11	Social Status	Character	OC,BC,SC,ST Social Status Category
12	Gender	Character	‘M’ for Male , ‘F’ for Female
13	SMARKS	Number	Student Total Marks
14	PASS DIVISION	Character	First Class, Second Class, Third Class, Fail

This file contains the data related to an individual student details and hence cannot be used directly for further processing. Hence this file is processed further to aggregate to the basic data in order to produce the information regarding social status, gender, course and the college data. This data is represented in the following form:

Attribute list

Table 2: Data Structure for the aggregated data

SNO	ATTRIBUTE NAME	TYPE	DESCRIPTION
1	Social Status	Character	OC,BC,SC,ST Social Status Category
2	GROUP/Course	Character	‘BSCB ‘ for BSc(Bio.sc.) Course, BSCM ‘ for BSc(Maths) Course
3	YEAR	Number	Year
4	REGD	Number	Number of Students Registered
5	PASS	Number	Number of Students Passed

Further, the data is processed to find the information regarding the grade acquired by students social status wise.

Table 3: Data Structure for aggregated Grade wise data

SNO	ATTRIBUTE NAME	TYPE	DESCRIPTION
1	Social Status	Character	OC,BC,SC,ST Social Status Category
2	GROUP/Course	Character	‘BSCB ‘ for BSc(Bio.sc.) Course, BSCM ‘ for BSc(Maths) Course
3	YEAR	Number	Year
4	REGD	Number	Number of Students Registered
5	PASS	Number	Number of Students Passed
6	Grade A	Number	Number of Students Passed in between range 51-60%

7	Grade B	Number	Number of Students Passed in between range 61-70%
8	Grade C	Number	Number of Students Passed in between range 71-80%
9	Grade D	Number	Number of Students Passed in between range 81-90%
10	Grade E	Number	Number of Students Passed in between range 91-100%

The data contained in this form is chosen for further processing.

Data Processing

The collected basic data is processed to create tables, Table 2 and Table 3. The data in table 2 is processed to create support values, year wise and category wise for an overall pass aggregating the data college wise and year wise. The support value is calculated as:

Support value of overall pass =Total number of students passed in the specific course all social status wise for the year / Total number of students registered in that all social status for the year

The resultant data is represented in the following table 4

Table 4: Overall Pass Support Table

Social Status	Year	Regd	Pass	Support
Overall	2002	1572	900	0.5725
	2003	1643	985	0.5995
	2004	1744	1034	0.5929
	2005	2267	1378	0.6079
	2006	2544	1641	0.645
	2007	2569	1669	0.6497

Let us now we calculate the support table. The data in table 6.2 is processed to create support values, year wise for Social status wise pass aggregating the data College wise and year wise. The support value is calculated as

Support value of overall pass for Social status wise=Total number of students passed in the specific course Social status wise for the year / Total number of students registered Social status wise for the year.

The resultant data is represented in the following table 5

Table 5: Overall course wise Social status support table

Social Status	Year	Regd	Pass	Support	Social Status	Year	Regd	Pass	Support
OC	2002	600	386	0.6433	BC	2002	796	449	0.56407
	2003	607	385	0.6343		2003	857	532	0.62077
	2004	555	361	0.6505		2004	967	590	0.610134
	2005	669	479	0.716		2005	1241	746	0.601128
	2006	795	568	0.7145		2006	1329	876	0.6 59142
	2007	757	529	0.6988		2007	1347	916	0.68003
Social Status	Year	Regd	Pass	Support	Social Status	Year	Regd	Pass	Support
SC	2002	144	55	0.3819	ST	2002	32	10	0.3125
	2003	147	57	0.3878		2003	32	11	0.34375
	2004	170	63	0.3706		2004	52	20	0.384615
	2005	285	123	0.4316		2005	72	30	0.416667
	2006	322	152	0.4721		2006	98	45	0.459184
	2007	321	158	0.4922		2007	144	66	0.458333

Similarly, now we calculate Support table for College wise as below shown

Table 6: College wise support table

Social Status	Year	Regd	Pass	Support					
over all	2002	132	79	0.598485					
	2003	149	88	0.590604					
	2004	177	105	0.59322					
	2005	194	121	0.623711					
	2006	220	144	0.654545					
	2007	242	171	0.706612					
Social Status	Year	Regd	Pass	Support	Social Status	Year	Regd	Pass	Support
OC	2002	31	19	0.612903	BC	2002	69	45	0.652174
	2003	22	14	0.636364		2003	86	54	0.627907
	2004	31	20	0.645161		2004	95	58	0.610526
	2005	34	23	0.676471		2005	102	66	0.647059
	2006	38	27	0.710526		2006	115	80	0.695652
	2007	46	34	0.73913		2007	120	92	0.766667
Social Status	Year	Regd	Pass	Support	Social Status	Year	Regd	Pass	Support
SC	2002	20	9	0.45	ST	2002	12	6	0.5
	2003	26	12	0.461538		2003	15	8	0.533333
	2004	35	18	0.514286		2004	16	9	0.5625
	2005	38	20	0.526316		2005	20	12	0.6
	2006	43	23	0.534884		2006	24	14	0.583333
	2007	48	28	0.583333		2007	28	17	0.607143

As the work involves predicting the grade wise category wise pass, the basic data needs to be processed accordingly. The data thus processed is available in a table with the structure shown as in table 6.3. This data generated course wise , category wise, year wise and grade wise is as follow:

Support value for Overall course wise Grade wise Pass= Total number of overall students passed in the specific grade/Total number of overall students passed in that the year:

Table 7: Support table for Overall course wise grade wise

Social Status	Year	Regd	Pass	Grade A	Grade B	Grade C	Grade D	Grade E
Overall	2002	1572	900	414	336	130	20	0
	2003	1643	985	390	376	184	35	0
	2004	1744	1034	443	395	160	36	0
	2005	2267	1378	525	503	282	67	1
	2006	2544	1641	504	599	372	157	9
	2007	2569	1669	532	678	351	96	12
Social Status	Year	Regd	Pass	S A	S B	S C	S D	S E
Overall	2002	1572	900	0.46	0.373333	0.144444	0.022222	0
	2003	1643	985	0.395939	0.381726	0.186802	0.035533	0
	2004	1744	1034	0.428433	0.382012	0.154739	0.034816	0
	2005	2267	1378	0.380987	0.365022	0.204644	0.048621	0.00073
	2006	2544	1641	0.30713	0.365021	0.226691	0.095673	0.00548
	2007	2569	1669	0.318754	0.406231	0.210306	0.057519	0.00719

Now we calculate support table for Support value for Grade wise Pass

Support value for Grade wise Pass= Total number of students passed in the specific grade/Total number of students passed in that Category for the year

Table 8: Course wise, Category wise, Year wise and Grade wise Support Table

Social Status	Year	Regd	Pass	Grade A	Grade B	Grade C	Grade D	Grade E
OC	2002	600	386	160	146	69	11	0
	2003	607	385	136	145	85	19	0
	2004	555	361	110	158	76	17	0
	2005	669	479	145	189	117	28	0
	2006	795	568	153	188	140	79	8
	2007	757	529	132	215	142	38	2
Social Status	Year	Regd	Pass	S A	S B	S C	S D	S E
OC	2002	600	386	0.414508	0.37824	0.178756	0.0285	0
	2003	607	385	0.353247	0.37662	0.220779	0.04935	0

	2004	555	361	0.304709	0.43767	0.210526	0.04709	0
	2005	669	479	0.302714	0.39457	0.244259	0.05846	0
	2006	795	568	0.269366	0.33099	0.246479	0.13908	0.014085
	2007	757	529	0.249527	0.40643	0.268431	0.07183	0.003781

Similarly, now we calculate Support table for College wise as below shown

Table 9: College wise, Over all Year wise and Grade wise Support Table

Social Status	YEAR	REGD	PASS	Grade A	Grade B	Grade C	Grade D	Grade E
Overall	2002	132	79	31	38	10	0	0
	2003	149	88	34	36	14	4	0
	2004	177	105	38	39	22	6	0
	2005	194	121	36	44	31	10	0
	2006	220	144	33	46	44	21	0
	2007	242	171	36	70	43	22	0
Social Status	YEAR	REGD	PASS	SA	SB	SC	SD	SE
Overall	2002	132	79	0.3924	0.481	0.1266	0	0
	2003	149	88	0.3864	0.4091	0.1591	0.0455	0
	2004	177	105	0.3619	0.3714	0.2095	0.0571	0
	2005	194	121	0.2975	0.3636	0.2562	0.0826	0
	2006	220	144	0.2292	0.3194	0.3056	0.1458	0
	2007	242	171	0.2105	0.4094	0.2515	0.1287	0

Similarly, now we calculate Support table for College wise and social status wise as below shown

Table 10: Course wise, Category wise, Year wise and Grade wise Support Table

Social Status	YEAR	REGD	PASS	Grade A	Grade B	Grade C	Grade D	Grade E
OC	2002	31	19	9	8	2	0	0
	2003	22	14	4	7	2	1	0
	2004	31	20	7	8	4	1	0
	2005	34	23	4	12	6	1	0
	2006	38	27	6	11	7	3	0
	2007	46	34	10	12	8	4	0
Social Status	YEAR	REGD	PASS	SA	SB	SC	SD	SE
OC	2002	31	19	0.473684	0.421053	0.105263	0	0
	2003	22	14	0.285714	0.5	0.142857	0.071429	0
	2004	31	20	0.35	0.4	0.2	0.05	0

	2005	34	23	0.173913	0.521739	0.26087	0.043478	0
	2006	38	27	0.222222	0.407407	0.259259	0.111111	0
	2007	46	34	0.294118	0.352941	0.235294	0.117647	0

Once the required tables are ready we desire the required equations for predictions of given years.

Derivation of Expression

It is observed that the number of students who pass a course in a college obviously depends on the number of students who join the course in a college. The social mix of students being the same in all colleges due to reservation policy, the passing of a course from a college category wise and grade wise follows a specific pattern. It is also observed that every year there is a slight increase in the pass, as the students and the teachers use new methods of learning. In order to identify this we have calculated support value tables as specified in the tables.

Taking the help of the support values for the previous year and the registered candidates in a college, we can roughly estimate the overall pass in a college as follow:

OP_e=C_R*$SVOP_{Pr}$ ------------------------------(1)

Where OP_e =Overall Pass Estimate

$C_{R=}$ Current Registered Candidates

$SVOP_{Pr=}$ Support value for over all pass of Previous Year

As this is only an estimated value and the current year needs appropriate correction as an incremental factor, which is observed to be about 0.05.

$OP_{correction}$= C_R* $SVOP_{Pr}$ *0.05 ------------------------------- (2)

Where $OP_{correction}$=Overall Pass Correction

Hence the OP_P= OP_e + $OP_{correction}$

Using equation (1) and (2)

$OP_{P=}$ C_{R*} $SVOP_{Pr}$ + C_{R*} $SVOP_{Pr}$ *0.05

= C_{R*} $SVOP_{Pr}$ (1+0.05)

= C_{R*} $SVOP_{Pr}$ *1.05 --- (3)

Similarly we can drive the expression for Social status category wise pass and gradewise and categorywise pass as

SCP_P=OP_P*$SVCP_{cat}$*1.05-------------------- (4)

GPP=CP_P*$SVCP_{grade}$*1.05 -------------- (5)

Algorithm for Overall pass prediction

Inputs:

1. Current Registered Candidates CR

2. Overall pass support value table
3. Support value table for category wise pass
4. Support value table for class wise category

Variables: OP_P = Overall pass predicated value

SCP_P=Social status Category wise Predicated Pass Value
GP_P= Grade Wise Category wise Pass Predicated Value
$SVOP_{Pr}$=Support value for over all pass of Previous Year
$SVCP_i$ =Support Value for Category wise Pass i=1 to 4 representing (OC,BC,SC,ST)
$SVGP_{ij}$=Support value for Grade wise Category wise pass i=1 to 4 and j=1 to 5 (Grade A ,Grade B, Grade C, Grade D, Grade E)
C_R =Current year registered candidates

1. Begin
2. Input Registered Candidates C_R
3. Get Support value of pass for the previous year : $SVOP_{Pr}$
4. $OP_P=C_R*SVOP_{Pr}*1.05$
5. For all Categories do (i=1 to 4)
    a. Get support value of a Social status category SVSCPi
    b. $SCP_p= OP_P * SVSCP_i *1.05$
    c. For all grades-do(j=1 to 5)
        i. Get Support value of Grade wise Category wise Pass value $SVGP_{ij}$
        ii. $GP_P =SCP_p* SVGP_{ij} *1.05$
        iii. End –do (of j)
        iv. End-do (of i)
6. End.

Result & Analysis

Using the specified algorithm we have calculated predicated pass for the years for which we already have the actual data using the values of the previous year, as shown below.

Table 11: Prediction table for overall pass course wise

Social Status	Year	REGD	PASS	Predicted Value	Round off Predicted Value	Difference	Error %
Overall	2002	1572	900	-	-	-	-
	2003	1643	985	987.6813	988	3	0.304569
	2004	1744	1034	1097.828	1098	64	6.189555
	2005	2267	1378	1411.285	1411	33	2.394775
	2006	2544	1641	1623.694	1624	17	1.035954
	2007	2569	1669	1739.982	1740	71	4.254044

As the table shows the error percentage is less than 8% and hence the predictions made are significant.

Table 12: Prediction Table Course wise, Social Status wise

Social Status	Year	REGD	PASS	Predicted Value	Round off Predicted Value	Difference	Error %
OC	2002	600	386	-	-	-	-
	2003	607	385	410.0285	410	25	6.493506
	2004	555	361	369.619	370	9	2.493075
	2005	669	479	456.9089	457	22	4.592902
	2006	795	568	597.676	598	30	5.28169
	2007	757	529	567.8928	568	39	7.372401
Social Status	Year	REGD	PASS	Prediction	Round off Predicted Value	Difference	Error %
BC	2002	796	449	-	-	-	-
	2003	857	532	507.5787	508	24	4.511278
	2004	967	590	630.2989	630	40	6.779661
	2005	1241	746	795.0357	795	22	2.949062
	2006	1329	876	838.8442	839	37	4.223744
	2007	1347	916	932.2578	932	16	1.746725
Social Status	Year	REGD	PASS	Prediction	Round off Predicted Value	Difference	Error %
SC	2002	144	55	-	-	-	-
	2003	147	57	58.95313	59	2	3.508772
	2004	170	63	69.21429	69	6	9.52381
	2005	285	123	110.8985	111	12	9.756098
	2006	322	152	145.9168	146	10	6.578947
	2007	321	158	159.1043	159	5	3.164557
Social Status	Year	REGD	PASS	Prediction	Round off Predicted Value	Difference	Error %
ST	2002	32	10	-	-	-	-
	2003	32	11	10.5	11	0	0
	2004	52	20	18.76875	19	1	5
	2005	72	30	29.07692	29	1	3.333333
	2006	98	45	42.875	43	2	4.444444
	2007	144	66	69.42857	69	3	4.545455

The Predictions made social status wise is also significant as the error percentage is below 8%. The results are shown in table 12.

Table 13: Prediction Table of Overall course Grade Wise Pass

Social Status : Overall Wise grade Wise					
Year	2002	Regd	1572	Pass	900
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	414	336	130	20	0
Round off Predicted Value	-	-	-	-	-
Difference	-	-	-	-	-
Error %	-	-	-	-	-
Year	2003	Regd	1643	Pass	985
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	390	376	184	35	0
Round off Predicted Value	391	377	185	35	0
Difference	1	1	1	0	0
Error %	0.2564	0.266	0.5435	0	0
Year	2004	Regd	1744	Pass	1034
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	443	395	160	36	0
Round off Predicted Value	470	419	170	38	0
Difference	27	24	10	2	0
Error %	6.0948	6.0759	6.25	5.5556	0
Year	2005	Regd	2267	Pass	1378
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	525	503	282	67	1
Round off Predicted Value	538	515	289	69	1
Difference	13	12	7	2	0
Error %	2.4762	2.3857	2.4823	2.9851	0
Year	2006	Regd	2544	Pass	1641
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	504	599	372	157	9
Round off Predicted Value	499	593	368	155	9
Difference	5	6	4	2	0
Error %	0.9921	1.0017	1.0753	1.2739	0
Year	2007	Regd	2569	Pass	1669
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	532	678	351	96	12
Round off Predicted Value	555	707	366	100	13
Difference	23	29	15	4	1
Error %	4.3233	4.2773	4.2735	4.1667	8.3333

Table 14: Prediction Table of Social Status wise Grade Wise Pass

Social Status : OC Grade Wise					
Year	2002	Regd	600	Pass	386
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	160	146	69	11	0
Round off Predicted Value	-	-	-	-	-
Difference	-	-	-	-	-
Error %	-	-	-	-	-
Year	2003	Regd	607	Pass	385
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	136	145	85	19	0
Round off Predicted Value	145	154	91	20	0
Difference	9	9	6	1	0
Error %	6.617647	6.206897	7.058824	5.263158	0
Year	2004	Regd	555	Pass	361
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	110	158	76	17	0
Round off Predicted Value	113	162	78	17	0
Difference	3	4	2	0	0
Error %	2.727273	2.531646	2.631579		0
Year	2005	Regd	669	Pass	479
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	145	189	117	28	0
Round off Predicted Value	138	180	112	27	0
Difference	7	9	5	1	0
Error %	4.827586	4.761905	4.273504	3.571429	0
Year	2006	Regd	795	Pass	568
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	153	188	140	79	8
Round off Predicted Value	161	198	147	83	8
Difference	8	10	7	4	0
Error %	5.228758	5.319149	5	5.063291	0
Year	2007	Regd	757	Pass	529
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	132	215	142	38	2
Round off Predicted Value	142	231	152	41	2
Difference	10	16	10	3	0
Error %	7.575758	7.44186	7.042254	7.894737	0

The data is further processed to check the effectiveness of mechanism for students of a specific social status result.

Prediction for College Wise

Table 15: Prediction table for overall Pass College wise (Overall Prediction)

Social Status	Year	Regd	Pass	Round off Predicted Value	Difference	Error %
over all	2002	132	79	-	-	-
	2003	149	88	94	6	6.818182
	2004	177	105	110	5	4.761905
	2005	194	121	121	0	0
	2006	220	144	144	0	0
	2007	242	171	166	5	2.923977

Table 16: Prediction table for Social status wise pass college wise

Social Status	Year	Regd	Pass	Round off Predicted Value	Difference	Error %
OC	2002	31	19	-	-	-
	2003	22	14	14	0	0
	2004	31	20	21	1	5
	2005	34	23	23	0	0
	2006	38	27	27	0	0
	2007	46	34	34	0	0
Social Status	Year	Regd	Pass	Round off Predicted Value	Difference	Error %
BC	2002	69	45	-	-	-
	2003	86	54	59	5	9.259259
	2004	95	58	63	5	8.62069
	2005	102	66	65	1	1.515152
	2006	115	80	78	2	2.5
	2007	120	92	88	4	4.347826
Social Status	Year	Regd	Pass	Round off Predicted Value	Difference	Error %
SC	2002	20	9	-	-	-
	2003	26	12	12	0	0
	2004	35	18	17	1	5.555556
	2005	38	20	21	1	5
	2006	43	23	24	1	4.347826
	2007	48	28	27	1	3.571429
Social Status	Year	Regd	Pass	Round off Predicted Value	Difference	Error %
ST	2002	12	6	-	-	-
	2003	15	8	8	0	0
	2004	16	9	9	0	0
	2005	20	12	12	0	0
	2006	24	14	15	1	7.142857
	2007	28	17	17	0	0

Table 17: Prediction table for overall grade wise pass college wise

Social Status: Overall Wise grade wise college					
Year	2002	Regd	132	Pass	79
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	31	38	10	0	0
Round off Predicted Value	-	-	-	-	-
Difference	-	-	-	-	-
Error %	-	-	-	-	-
Year	2003	Regd	149	Pass	88
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	34	36	14	4	0
Round off Predicted Value	33	34	11	1	0
Difference	1	2	3	3	0
Error %	1.136364	2.272727	3.409091	3.409091	0
Year	2004	Regd	177	Pass	105
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	38	39	22	6	0
Round off Predicted Value	36	35	18	3	0
Difference	2	4	4	3	0
Error %	1.904762	3.809524	3.809524	2.857143	0
Year	2005	Regd	194	Pass	121
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	36	44	31	10	0
Round off Predicted Value	32	38	25	6	0
Difference	4	6	6	4	0
Error %	3.305785	4.958678	4.958678	3.305785	0
Year	2006	Regd	220	Pass	144
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	33	46	44	21	0
Round off Predicted Value	28	39	37	16	0
Difference	5	7	7	5	0
Error %	3.472222	4.861111	4.861111	3.472222	0
Year	2007	Regd	242	Pass	171
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	36	70	43	22	0
Round off Predicted Value	29	60	34	16	0
Difference	7	10	9	6	0
Error %	4.093567	5.847953	5.263158	3.508772	0

Table 18: Prediction table for Social Status grade wise pass college wise

Social Status : OC Category Grade Wise					
Year	2002	Regd	31	Pass	19
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	9	8	2	0	0
Round off Predicted Value	-	-	-	-	-
Difference	.-	.-	.-	.-	.-
Error %	-	-	-	-	-
Year	2003	Regd	22	Pass	14
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	4	7	2	1	0
Round off Predicted Value	3	6	1	1	0
Difference	1	1	1	0	0
Error %	7.142857	7.142857	7.142857	0	0
Year	2004	Regd	31	Pass	20
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	7	8	4	1	0
Round off Predicted Value	6	6	3	0	0
Difference	1	2	1	1	0
Error %	5	10	5	5	0
Year	2005	Regd	34	Pass	23
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	4	12	6	1	0
Round off Predicted Value	3	10	5	0	0
Difference	1	2	1	1	0
Error %	4.347826	8.695652	4.347826	4.347826	0
Year	2006	Regd	38	Pass	27
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	6	11	7	3	0
Round off Predicted Value	4	9	6	2	0
Difference	2	2	1	1	0
Error %	7.407407	7.407407	3.703704	3.703704	0
Year	2007	Regd	46	Pass	34
Grade	Grade A	Grade B	Grade C	Grade D	Grade E
Grade Pass	10	12	8	4	0
Round off Predicted Value	7	9	7	3	0
Difference	3	3	1	1	0
Error %	8.823529	8.823529	2.941176	2.941176	0

The Processing made so far is the overall data collected from all the colleges. In order to check correctness and applicability of the prediction mechanism we have used it on a specific college. The results are shown in the table 17. The error percentage is less than 7%. Similarly, we can calculate the grade wise predicted values for each college. The results are quite encouraging as the error percentage is significantly low.

Conclusions

The success of a college is mainly dependent on the results it produces in terms of student success rate. We have successfully derived a prediction mechanism for the success of students course wise, social status and Grade wise. The method has been proved to be effective from the error percentage calculated is less than 7. Further processing can be made taking into consideration of college environment. However, the method helps the college managements to improve their infrastructure and academic activities midway through the course in order to improve their performance.

References

[1] Han, J., Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.,2006, 2nd Edition

[2] V.Ramesh, P.Parkavi, P.Yasodha, Performance Analysis of Data Mining Techniques for Placement Chance Prediction ,International Journal of Scientific & Engineering Research Volume 2, Issue 8, August-2011

[3] P. Cortez, A. Silva, Using Data Mining To Predict Secondary School Student Performance||, In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[4] Ş. Z. ERDOĞAN, M. TİMOR . "A data mining application in a student database". Journal of aeronautics and space technologies, volume 2,number 2 , pp.53-57, 2005.

[5] Kotsiantis, S., Pierrakeas, C., Pintelas, P. (2003). Preventing student dropout in distance learning systems using machine learning techniques, In International Conference on Knowledge-Based Intelligent Information & Engineering Systems, Oxford, 3-5

[6] Z. N. Khan, ―Scholastic Achievement of Higher Secondary Students in Science Stream||, Journal of Social Sciences, Vol. 1, No. 2, 2005, pp. 84-87.

[7] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás ,Data Mining Algorithms to Classify Students,EDM2008 Procedings

[8] M. N. Quadri1 and Dr. N.V. Kalyanka- Drop Out Feature of Student Data for Academic Performance Using Decision Tree, Global Journal of Computer Science and Technology Vol. 10 Issue 2 (Ver 1.0), April 2010.

[9] Zlatko J. Kovacic, John Steven Green, Predictive working tool for early identification of 'at risk' students , Newzealand

[10] Sajadin Sembiring, M. Zarlis, Dedy Hartama,Ramliana S, Elvi Wani. Prediction of Student Academic Performance by An Application of Data Mining Techniques, International Conference on Management and Artificial Intelligence, Bali, Indonesia,IPEDR vol.6 ,pp.110-114,2011.

[11] R. Agarwal and R. Srikant, “Fast Algorithms for Mining Association Rules,” Proc. Int’l Conf. Very Large Databases(VLDB), 1994.

[12] J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD, 2000.

[13] J. Han and Y. Fu, “Discovery of Multi-Level Association Rules from Large Databases,” Proc. Int’l Conf. Very LargeDatabases (VLDB), 1995.

[14] J. Park, M. Chen, and P. Yu, “An Effective Hashing-Based Algorithm for Mining Association Rules,” Proc. ACM SIGMOD,1995.

[15] R. Srikant and R. Agrawal, “Mining Generalized Association Rules,” Proc. Int’l Conf. Very Large Databases (VLDB), 1995.

[16] B. Ozden, S. Ramaswamy, and A. Silberschatz, “Cyclic Association Rules,” Proc. IEEE Int’l Conf. Data Eng. (ICDE), 1998.

[17] Y. Li, P. Ning, X.S. Wang, and S. Jajodia, “Discovering Calendar- Based Temporal Association Rules,” J. Data andKnowledge Eng., vol. 15, no. 2, 2003.

[18] S. Ramaswamy, S. Mahajan, and A. Silberschatz, “On the Discovery of Interesting Patterns in Association Rules,” Proc.Int’l Conf. Very Large Databases (VLDB), 1998

[19] Beal, C. R. and Cohen, P. R. (2008). Temporal Data Mining for Educational Applications. In Proceedings of the 10th Pacific Rim international Conference on Artificial intelligence: Trends in Artificial intelligence, Hanoi, Vietnam,pp. 66-77.

[20] Hübscher, R., Puntambekar, S., Nye, A. (2007). Domain Specific Interactive Data Mining. In Workshop on Data Mining for User Modeling, at the 11th International Conference on User Modeling, Corfu, Greece,pp. 81-90.