ANALYZING AND PREDICTING STUDENT PERFORMANCE USING DATA MINING

[1]Allando Raplang

[1]Assistant Professor
[1]Department of Computer Applications,
[1]Synod College, Shillong, India.

Abstract: One of the main issues in the educational institutions is finding out the cause of student's lack of performance in academics. In this study student's performance will be evaluated using association rule mining algorithm based on various attributes. The results will then be used to help students improve their performance.

Index Terms **- Association Rule Mining; Apriori algorithm, multiple regression**.

I. INTRODUCTION

Data Mining is the process of extracting useful information from large datasets. The useful information is then analyzed and summarized thereby converting into knowledge. Therefore, sometimes it also called data or knowledge discovery. The knowledge that is extracted can be useful for variety of purposes. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation).[1,2]

Predictive tasks are used to predict the value of a particular attribute based on the values of other attributes that are known. Predictive modeling refers to the task of building a model for the target variable as a function of the explanatory variables. There are two types of predictive modeling tasks: Classification, which is used for discrete target variables and Regression, which is used for continuous target variables.[3]

The objective of the paper is to analyze and predict the student's performance based on their previous and current academic performance in previous exams, unit test, assignments, attendance etc and family background by applying association rule and multiple regression analysis with two predictors.

II. DATA MINING TASK

Association Rule Mining:

Association Rule Mining is a popular and well researched method for discovering interesting relations between variables in large databases. The uncovered relationship can be represented in the form of association rules or set of frequent items.[4]

Table 1: an example of market transactions

TID	Items
1.	{Bread, Milk}
2.	{Bread, Diapers, Beer, Egg}
3.	{Milk, Diapers, Beer, Cola}
4.	{Bread, Milk, Diapers, Beer}
5.	{Bread, Milk, Diapers, Cola}

To illustrate the concepts, we use a small example from the dataset in Table 1. The set of items is ***I***= {milk, bread, beer, cola, diapers, egg**}**. An example rule can be extracted from table 1 could be {diapers} => beer} meaning that if diapers are bought, customers also buy beer.[4]

Regression:

Regression is a statistical perspective which can be used to evaluate the strength of a relationship between two variables. It is generally used to predict future values based on past values by fitting a set of a points to a curve.[7]

Linear regression assumes that a linear relationship exist between the input data and the output data.[7] In simple linear regression a criterion variable is predicted from one of the predictor variable. In multiple regression, the criterion variable is predicted by two or more predictors.

Base on the two tasks briefly discussed above, this study will try to answer the following questions:

1. What are the attributes that can be used to predict student's performance?
2. Which attributes affect the student's performance?

The common formula for a linear relationship is used in this model:

$$y = c_0 + c_1x_1 + c_2x_2 + \dots + c_nx_n$$

Here there are n input variables which are called predictors or regressors; one output variable, which is called as the response and n+1 constants. This is sometimes called multiple linear regression because there is more than one prediction.[7]

III. DATA SETS

In this study the data of the students who are pursuing their BCA 3rd semester will considered as training datasets. Association rules will be use for selecting attributes from the dataset and based on the accuracy of the relationship between the attributes rules are generated thereby strong attributes are identified. Attributes that will be used are as follows: SSLC %, Computer Background, HSSLC %, Maths HSSLC, 1st Semester %, 1st Semester Maths %, 1st Semester Attendance %, 1st Semester Assignment Marks, 1st Semester Practical Marks, and Unit Test Marks as shown in Table 2.

Table 2. Attributes and Possible Values

Attribute	Variable Name	Description	Values
SSLC %	SSLC	Percentage in Secondary Board Exam	1, 2, 3
Computer Background	Comp	Knowledge of computers before joining BCA	Yes/No
HSSLC %	HSSLC	Percentage in Higher Secondary Board Exam	1, 2, 3
Maths in HSSLC	Maths	Maths as one of the subjects in Higher Secondary	Yes / No
1st Semester %	1stSem	Percentage in the 1st Semester	P (pass), F (Fail)
1st Semester Maths %	1stMaths	Percentage of maths in the 1st Semester	P (pass), F (Fail)
1st Semester Attendance %	attend	Attendance during the 1st semester	A, B, C, D, E, F
1st Semester Assignment Marks	assign	Assignment score in the 1st semester	A, B, C, D, E, F
1st Semester Practical Marks	prac	1st semester internal practical score	A, B, C, D, E, F
Unit Test Marks	unit_test	Unit test score	A, B, C, D, E, F

IV. DATA PRE-PROCESSING

One of the important steps of data mining process is data pre-processing. Data pre-processing is used in identifying the missing values, noisy data and irrelevant and redundant information from dataset.

Table 3. Categorization of Attributes

Attributes	Range
SSLC & HSSLC	(SSLC% or HSSLC%) >= 60% → 1 45% <= (SSLC% or HSSLC%) < 60% → 2 (SSLC% or HSSLC%) < 45% → 3
1st Semester % & 1st Semester Maths %	Maths >=30% → P Maths < 30% → F Fail in one subjects → F else → P
Attendance, Assignment, Practical & Unit Test	(Attendance or Assignment or Practical or Unit Test) >= 90 → A 75<= (Attendance or Assignment or Practical or Unit Test) < 90 → B 60<= (Attendance or Assignment or Practical or Unit Test) < 75 → C 45<= (Attendance or Assignment or Practical or Unit Test) < 60 → D 30<= (Attendance or Assignment or Practical or Unit Test) < 45 → E (Attendance or Assignment or Practical or Unit Test) < 30 → F

V. METHODOLOGY

In this study, a free software tool, WEKA will be used. It is open source software that offers a collection of machine learning and data mining algorithms for data pre-processing, regression, association rules, clustering and classification.

Association Rule Mining [6]:

Association rule mining is a research method used for determining interesting relationships between the data items in a large itemsets and based on these relationships strong rules are generated using different measures of support and confidence.

The preliminaries necessary for performing data mining on any data are discussed below.

Let $\mathbf{I} = \{I_1, I_2, ..., I_n\}$ be a set of *items*. Let $\mathbf{D} = \{T_1, T_2, ..., T_n\}$ be a set of database transactions where each transaction $T \subseteq I$. Each transaction in **D** has a unique transaction ID and contains a subset of the items in **I**.

Association rule:

It is defined as an implication of the form **X**□**Y** where **X, Y** ⊆ **I** and **X** ∩ **Y** = Φ**.** The sets of items (for short *itemsets*) **X** and **Y** are called *antecedent* (left-hand-side or LHS) and *consequent* (right-hand-side or RHS) of the rule respectively.

Useful Concepts [2]

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

1. Support

The support **supp(*X*)** of an itemset *X* is defined as the proportion of transactions in the data set which contain the itemset.

supp(X)= no. of transactions which contain the itemset X / total no. of transactions

In the example database, the itemset {diapers,beer} has a support of 3/5 = 0.6 since it occurs in 60% of all transactions. To be even more explicit we can point out that 3 is the number of transactions from the database which contain the itemset {diapers,beer} while 5 represents the total number of transactions.

2. Confidence

The ***confidence*** of a rule is defined:

conf(X □ Y)= supp(X U Y)/supp(X)

For the rule {diapers}=>{beer} we have the following confidence:

supp({diapers,beer}) / supp({diapers}) = 0.6 / 0.8 = 0.75

This means that for 75% of the transactions containing diapers the rule is correct. Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Association rule generation is usually split up into two separate steps:

1. First, minimum support is applied to find all *frequent itemsets* in a database.
2. Second, these frequent itemsets and the minimum confidence constraint are used to form rules.

Apriori algorithm pseudo code:

```
procedure Apriori (T, minSupport){ //T is the database and minSupport is the minimum support
    L1= {frequent items};
    for (k= 2; Lk-1 !=∅ ; k++){
        Ck= candidates generated from Lk-1
        //that is cartesian product Lk-1 x Lk-1 and eliminating any k-1 size itemset that is not frequent
        for each transaction t in database do{
            #increment the count of all candidates in Ck that are contained in t
            Lk = candidates in Ck with minSupport
        }//end for each
    }//end for
    return ;
}
```

Multiple Regression (R) [4]:

Multiple Regression is a statistical tool that allows you to examine how **multiple independent variables** are related to a dependent variable. Once you have identified how these multiple variables relate to your dependent variable, you can take information about all of the independent variables and use it to make much more powerful and accurate predictions about why things are the way they are. This latter process is called "Multiple Regression".

The Formula for Multiple Regression

Y' = a + b1 X1 + b2 X2 [Y' – A predicted value of Y (which is a dependent variable or response)]

a – The Y intercept

b1 – The change in Y for each 1 increment change in X_1, b2 – The change in Y for each 1 increment change in X_2

X – an X score (which is a dependent variable or predictor)

How to calculate b_1 and b_2

$$b_1 = \left(\frac{r_{y,x1} - r_{y,x2}, r_{x1,x2}}{1-(r_{x1,x2})^2}\right)\left(\frac{SD_y}{SDx_1}\right),\ b_2 = \left(\frac{r_{y,x2} - r_{y,x1}, r_{x1,x2}}{1-(r_{x1,x2})^2}\right)\left(\frac{SD_y}{SDx_2}\right)$$

Where *r* is a correlation and SD is Standard deviation

Calculating "a"

a = $\overline{Y} - b_1\overline{X_1} - b_2\overline{X_2}$

$\overline{Y}$**= The mean of Y**

$b_1\overline{X}_1$ = Product of b_1 and mean of X_1, *$b_2\overline{X}_2$ = Product of b_2 and mean of X_2*

VI. RESEARCH AND DISCUSSION

A. Apriori Algorithm:

The dataset of 30 students of BCA 1st Semester batch 2015 & 2016, Synod College, Shillong was obtained. The datasets is then input into WEKA in which various association rules are generated between the attributes like maths background & unit test, assignment & unit test and further found out how these attributes affect the student's performance.

The analysis for generated association rules is as follows:

The rules generated for 90% confidence and 0.1 support are:

1. maths=No unit_test=C 11 ==> 1stSem=F 1stmaths=F 11 conf:(1)
2. maths=No assign=B unit_test=C 9 ==> 1stSem=F 1stmaths=F 9 conf:(1)
3. sslc=2 comp=Yes assign=B 13 ==> 1stSem=F 1stmaths=F 12 conf:(0.92)
4. comp=Yes hsslc=2 assign=B 12 ==> 1stSem=F 1stmaths=F 11 conf:(0.92)
5. sslc=2 comp=Yes maths=No assign=B 11 ==> 1stSem=F 1stmaths=F 10 conf:(0.91)
6. assign=B unit_test=C 10 ==> 1stSem=F 9 conf:(0.9)
7. assign=B unit_test=C 10 ==> 1stmaths=F 9 conf:(0.9)
8. comp=Yes hsslc=2 maths=No 10 ==> 1stSem=F 9 conf:(0.9)
9. comp=Yes hsslc=2 maths=No 10 ==> 1stmaths=F 9 conf:(0.9)

The rules generated for 80% confidence and 0.1 support are:

1. comp=Yes assign=B 17 ==> 1stSem=F 1stmaths=F 15 conf:(0.88)
2. hsslc=2 assign=B 15 ==> 1stSem=F 13 conf:(0.87)
3. hsslc=2 assign=B 15 ==> 1stmaths=F 13 conf:(0.87)
4. comp=Yes maths=No assign=B 15 ==> 1stSem=F 13 conf:(0.87)

From the above association rules of different confidence values the interpretation are as follows:

1. If student does not take maths in HSSLC and performed bad in unit test or doesn't do well in assignments then he/she is likely to fail in first semester maths and the first semester itself.

2. If student got second class in SSLC or HSSLC with computer background, does not do well in the assignments and /or no maths in HSSLC then he/she is likely to fail in first semester maths and the first semester itself.
3. If student does not do well in the assignments and performed poorly in the unit test then he/she is likely to fail in first semester maths and the first semester itself.
4. If student got second class in HSSLC with computer background and no maths in HSSLC then he/she is likely to fail in first semester maths and the first semester itself.

Based on the above interpretations, data (hsslc %, Computer background & maths in hsslc) of 16 students are being tested and the following results are found:

1. Eight Students with 2nd Class, with computer background or not and no maths in class 12 Fail in the 1st Semester
2. Two Students with 1st Class and no maths in class 12 Fail in the 1st Semester
3. One student with no maths but secured 1st Class in HSSLC cleared 1st Semester
4. Five students with 2nd Class above in HSSLC and with maths in HSSLC cleared the 1st Semester

B. Multiple Regression:

We have taken a student dataset consisting of 29 student's information (Table 4.) of a reputed institution considering the total marks obtained from the assignment, total unit test marks and their first semester percentage. The student's dataset as training dataset is then applied by the method of multiple regression (Table 5.) to predict the percentage of marks secured by the students in their final exams based on the total marks from the assignment and total unit test marks obtained in the first semester.

Table 4: Data collected from 29 students of first semester BCA (2015 & 2016)

Sl. No.	Assignment (X1)	Unit Test (X2)	1st Semester %
1.	26.5	26	56
2.	28	34	66
3.	27	25	39
4.	25.5	29.5	56
5.	27	19	46
6.	25	22.5	62
7.	26	28.5	65
8.	26	25	56
9.	27	29.5	43
10.	27	26	59
11.	27	30	47
12.	25	23.5	36
13.	26	22	42.5
14.	25	34	63
15.	26	21.5	50.5
16.	26	25	50
17.	26.5	24	47.5
18.	25	24	43
19.	26	28	51
20.	28	38	68
21.	25	33	61
22.	23.5	18	30.6
23.	24.4	35.5	63.5
24.	26.5	25.5	37
25.	24	26.5	38
26.	25	24	32
27.	24	29	56.3
28.	25	20	38.3
29.	26	38.5	55

Table 5: Dataset applied by the method of multiple regression

Sl no.	(X1)	X1^2	X2	X2^2	Y	Y^2	X1*X2	X1*Y	X2*Y	$(X1-M)^2$	$(X2-M)^2$	$(Y-M)^2$
1	26.5	702.25	26	676	56	3136	689	1484	1456	0.46	1.14	32.69
2	28	784	34	1156	66	4356	952	1848	2244	4.73	48.04	247.03
3	27	729	25	625	39	1521	675	1053	975	1.38	4.28	127.30
4	25.5	650.25	29.5	870.25	56	3136	752.25	1428	1652	0.11	5.91	32.69
5	27	729	19	361	46	2116	513	1242	874	1.38	65.11	18.34
6	25	625	22.5	506.25	62	3844	562.5	1550	1395	0.68	20.88	137.29
7	26	676	28.5	812.25	65	4225	741	1690	1852.5	0.03	2.05	216.60
8	26	676	25	625	56	3136	650	1456	1400	0.03	4.28	32.69
9	27	729	29.5	870.25	43	1849	796.5	1161	1268.5	1.38	5.91	53.04
10	27	729	26	676	59	3481	702	1593	1534	1.38	1.14	75.99
11	27	729	30	900	47	2209	810	1269	1410	1.38	8.59	10.78
12	25	625	23.5	552.25	36	1296	587.5	900	846	0.68	12.74	204.00
13	26	676	22	484	42.5	1806.25	572	1105	935	0.03	25.69	60.57
14	25	625	34	1156	63	3969	850	1575	2142	0.68	48.04	161.73
15	26	676	21.5	462.25	50.5	2550.25	559	1313	1085.75	0.03	31.01	0.05
16	26	676	25	625	50	2500	650	1300	1250	0.03	4.28	0.08
17	26.5	702.25	24	576	47.5	2256.25	636	1258.75	1140	0.46	9.42	7.74
18	25	625	24	576	43	1849	600	1075	1032	0.68	9.42	53.04
19	26	676	28	784	51	2601	728	1326	1428	0.03	0.87	0.51
20	28	784	38	1444	68	4624	1064	1904	2584	4.73	119.49	313.90
21	25	625	33	1089	61	3721	825	1525	2013	0.68	35.18	114.86
22	23.5	552.25	18	324	30.6	936.36	423	719.1	550.8	5.40	82.25	387.41
23	24.4	595.36	35.5	1260.25	63.5	4032.25	866.2	1549.4	2254.25	2.03	71.08	174.70
24	26.5	702.25	25.5	650.25	37	1369	675.75	980.5	943.5	0.46	2.46	176.43
26	24	576	26.5	702.25	38	1444	636	912	1007	3.33	0.32	150.87
27	25	625	24	576	32	1024	600	800	768	0.68	9.42	334.26
28	24	576	29	841	56.3	3169.69	696	1351.2	1632.7	3.33	3.73	36.21
29	25	625	20	400	38.3	1466.89	500	957.5	766	0.68	49.97	143.59
30	26	676	38.5	1482.25	55	3025	1001	1430	2117.5	0.03	130.67	22.25
Sum	748.9	19376.61	785	22062.5	1458.2	76648.94	20312.7	37755.45	40556.5	36.91	813.36	3,326.62
Mean	25.82	668.16	27.07	760.77	50.28	2643.07	700.44	1301.91	1398.5	1.27	28.05	114.71
SD										1.13	5.30	10.71

Correlation between Assignment (X1) and Unit Test Mark (X2), r(X1,X2) : 0.24

Correlation between Assignment (X1) and 1st Semester % (Y), r(X1,Y): 0.28

Correlation between Unit Test Mark (X2) and 1st Semester % (Y), r(X2,Y): 0.66

Standard Deviation of X1: 1.13

Standard Deviation of X2: 5.30

Standard Deviation of Y: 10.71

b1 = 0.01411, b2= 0.3104 a= 41.5151

Using the above findings, now we can predict final first semester results (Table 6.) of a student based on the attributes Assignment and Unit Test score. Below are the sample tested data of three students of the 1st Semester BCA 2017 batch:

Table6. Predicting 1st Semester results.

Sl. No.	Assignment	Unit Test	1st Semester % Result Predicted	1st Semester % Result Secured
1.	15	27	50	52
2.	15.5	24.5	49	52
3.	22	28	50	49

VII. CONCLUSION

The analysis using data mining association rule shows that there is a high chances a student with or without a computer background will fail in the 1st Semester exam if he/she has no maths in class 12 and there is a 100% chances a student with or without a computer background will clear the 1st Semester exam if he/she has maths in class 12. Also with multiple regression given Assignment marks and Unit Test score of a student we can predict the first semester percentage result.

REFERENCES

[1] Data Mining: What is Data Mining? http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm.

[2] http://software.ucp.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf.

[3] A Study on the Prediction of Student's Performance by applying straight-line regression analysis using the method of least squares: G.Narasinga Rao, Srinivasan Nagaraj, Assistant Professor, Department of Computer Science & Engineering, GMRIT, Rajam. Assistant Professor, Department of Computer Science & Engineering, GMRIT, Rajam.

[4] https://www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf.

[5] bcg_comp_chapter4.pdf: Excerpted from The Radical Statistician by Jim Higgins.

[6] Predicting Students Academic Performance Using Education Data Mining: Suchita Borkar, K. Rajeswar.

[7] Data Mining: Introduction and Advanced Topics: Margarnet H. Dunham.