Feature Selection of Student Performance Parameters using Statistical Techniques

Mr. Bhavesh Patel, Dr. Jyotindra Dharwa

Abstract-

Student information system holds a lot of information regarding demographic information and academic information of the students. It is very difficult to find out the highest affected parameters on student performance. This work aim is to perform statistical analysis and highlight the factors affected on the performance of students. We applied chi-square statistics technique to find out the most relevant academic and non-academic factors. This analysis is carried out by comparative analysis of various hypothesis testing. The outcome of this analysis would help the management to take timely and proactive decisions to improve the performance of students.

Keywords: Statistics, Chi-square, hypothesis testing. Student Performance

I. INTRODUCTION

Student database has very large datasets that involve many demographic and academic attributes. Student dataset involves many attributes those highly affected on student performance and some attributes are less affected on performance of students. In the Indian scenario the attributes of student dataset that affect performance is different from the students of foreign universities. Students in India do not come in different age groups nor there retention problems in courses. Online courses are very few in India. The attributes that influence a student's performance are entirely different in India. This research workis carried out by applying chi square technique on student dataset to identify the most probable factors contributing to the success of students in a course.

II. RESEARCH OBJECTIVE

The main goal of this research is to analyze the student data of UG Computer courses and identify the key performance indicators that affect the result of the student in the course.

III. ABOUT CHI-SQUARE STATISTICAL TECHNIQUE.

A. Why Chi-Square?

Chi-Square Analysis Pearson's chi-square test of independence is a statistical method used to identify degree of association between variables [1]. This technique is applied to analyze the dependency of all attributes (factors) on the outcome attribute.

B. When To Use Chi-Square Test?

The test procedure is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a <u>contingency table</u>, the expected frequency count for each cell of the table is at least 5

C. Steps And Procedure To Analysis of Chi Square Test:

This approach consists of four steps:(1) state the hypotheses, (2) formulate an analysis plan, (3) analyze sample data, and (4) interpret results.

State the Hypotheses

Suppose that Variable A has r levels, and Variable B has c levels. The <u>nullhypothesis</u> states that knowing the level of Variable A does not help you predict the level of Variable B. That is, the variables are independent.

- H₀: Variable A and Variable B are independent.
- Ha: Variable A and Variable B are not independent.

The <u>alternative hypothesis</u> is that knowing the level of Variable A *can* help you predict the level of Variable B.

Formulate an Analysis Plan

The analysis plan describes how to use sample data to accept or reject the null hypothesis. The plan should specify the following elements.

- Significance level. Often, researchers choose <u>significance levels</u> equal to 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.
- Test method. Use the <u>chi-square test for independence</u> to determine whether there is a significant relationship between two categorical variables.

Analyze Sample Data

Using sample data, find the degrees of freedom, expected frequencies, test statistic, and the P-value associated with the test statistic. The approach described in this section is illustrated in the sample problem at the end of this lesson.

- **Degrees of freedom.** The degrees of freedom (DF) is equal to:
 - DF = (r 1) * (c 1) Where r is the number of levels forone categorical variable, and c is the number of levels for the other categorical variable.
- Expected frequencies. The expected frequency counts are computed separately for each level of one categorical variable at each level of the other categorical variable. Compute r * c expected frequencies, according to the following formula.

$$E_{r,c} = (n_r * n_c) / n$$

where $E_{r,c}$ is the expected frequency count for level r of Variable A and level c of Variable B, n_r is the total number of sample observations at level r of Variable A, n_c is the total number of sample observations at level c of Variable B, and n is the total sample size.

■ **Test statistic.** The test statistic is a chi-square random variable (X^2) defined by the following equation.

$$X^2 = \Sigma [(O_{r,c} - E_{r,c})^2 / E_{r,c}]$$

where $O_{r,c}$ is the observed frequency count at level r of Variable A and level c of Variable B, and $E_{r,c}$ is the expected frequency count at level r of Variable A and level c of Variable B.

■ **P-value.** The P-value is the probability of observing a sample statistic as extreme as the test statistic. Since the test statistic is a chi-square, use the Chi-Square Distribution Calculator to assess the probability associated with the test statistic. Use the degrees of freedomcomputed above. [4]

IV. DATACOLLECTIONANDMETHODOLOGY

For this study, recent real world data were collected from various UGCollages of Computer Courses. A sample of 3558 students was taken from a group of 10000 student's records. Selection of students was at random.

The data was collected using a questionaries' which includes questions related to demographics, learning behavioral and related to the student's results. This questionnaire was developed in Google Spreadsheet shared among the students of various institute of computer courses to collect the students' data. This questionaries' was also uploaded on site and shared among the students for collection of student data. Data was collected based on following parameters. There are total 232 attributes. Attribute list is described as per the following.

Table 1: List of Parameters

Parameters	Data Type	Parameter_Val ues
Srno	Numeric None	
Gen	Numeric	{1- Male, 2-Female}
Percentagehsc	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Stream	Numeric	1-Commerce, 2- Science
F_annual_income	Numeric	1-Low, 2-Average, 3-Middle, 4-High, 5-Very High
attendance_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
attendance_2	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
attendance_3	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good,

		•
		5-Excellent
attendance_4	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
attendance_5	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
attendance_6	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Overall_attendance	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
W_L_H	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
W_Li_U	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
D_Re_H	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
E_W_L_U_H	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
sam1_sub_nm1	String	CLanguage
Assignement_Mark_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
INT_TH_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
INT_PR_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
EXT_TH_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
EXT_PR_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Result_1_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Sub_Attednace_1	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Sub_Faculty_Prformance	Numeric	1-Poor, 2- Average, 3-Good, 4-Very_Good, 5-Excellent
Same Procedure will be fo	llow for subjects	of all semesters

V. TOOLS ANDTECHNIQUES

We have used SPSS Statistical for analysis of the collected data. Here in the analysis we have used random sampling method for selection of the data. In the data collection, data are categorical so most preferable technique for analysis of the data is chi-square statistical technique.

To Select Chi-Square Technique from SPSS tool we have to go through following steps:

- A. Formulate a hypothesis about your variables
- B. Select crosstabs

Click on the "Analyze" tab at the top of the page Select "Descriptive Statistics" from the list Select "Crosstabs"

- C. Select your variables for the test
- D. Select the chi-square statistic from the "Statistics" button
- E. Select column percentages from the "Cells" button
- F. Examine the results

VI. RESULTS AND DISCUSSION

A total of 3558 records were taken randomly for the analysis. In this study, Chi-Square technique is used for analysis of randomly selected and categorical data. Chi-square test is one of the simplest and most widely used parametric as well as non-parametric we used Chi-square test to find the significance between the different attributes and grade obtained by student.

We have prepared following hypothesis to find out highly affected parameters on student performance.

A. Hypothesis1: Is that any Relationship exist between students learning behavior on result or not?

<u>INDEPENDENT VARIABLE:</u> WEEKLYLIBRARYHOURS, WEEKLYLAB USAGE, DAILYREADINGHOURS, EXTRA USAGE OF LIBRARYAND LAB

<u>**DEPENDENT VARIABLE:**</u> RESULT OF ALL SEMESTERS. **H0 (Null Hypothesis):** Students Behavioral parameters are Independent (Not Dependent) on Students Result of all the semester.

H1 (Alternative Hypothesis): Students Behavioral parameters are Dependent on Students Result of all the semester.

Table 2: Chi Square Analysis for Student behavior parameter vs. all semester results.

SEM	нт	DF	CV	SIG	co v	RE
FIRST SEM	W_L_H	16	81.394	0	-0.074	Rejected
SEC SEM	W_L_H	16	87.806	0	-0.023	Rejected
THIRD SEM	W_L_H	16	126.064	0	0.007	Rejected
FOURTH SEM	W_L_H	16	110.876	0	-0.04	Rejected
FIFTH SEM	W_L_H	16	64.018	0	-0.042	Rejected
SIXTH SEM	W_L_H	16	107.153	0	-0.07	Rejected
FIRST SEM	W_Li_U	16	118.437	0	-0.007	Rejected
SEC SEM	W_Li_U	16	26.714	0.45	0.008	Accepted
THIRD SEM	W_Li_U	16	96.861	0	-0.044	Rejected

FOURTH SEM	W_Li_U	16	74.218	0	0.056	Rejected
FIFTH SEM	W_Li_U	16	46.071	0	-0.029	Rejected
SIXTH SEM	W_Li_U	16	70.518	0	0.006	Rejected
FIRST SEM	D_R_H	12	58.967	0	0.053	Rejected
SEC SEM	D_R_H	12	13.092	0.362	-0.002	Accepted
THIRD SEM	D_R_H	12	59.29	0	0.014	Rejected
FOURTH SEM	D_R_H	12	54.122	0	0.049	Rejected
FIFTH SEM	D_R_H	12	74.752	0	-0.053	Rejected
SIXTH SEM	D_R_H	12	11.518	0.48	0.021	Accepted
FIRST SEM	E_W_L_U_ H	16	112.413	0	0.004	Rejected
SEC SEM	E_W_L_U_ H	16	95.601	0	0.16	Rejected
THIRD SEM	E_W_L_U_ H	16	97.82	0	0.86	Rejected
FOURTH SEM	E_W_L_U_ H	16	76.886	0	0.57	Rej ected
FIFTH SEM	E_W_L_U_ H	16	66.243	0	0.54	Rejected
SIXTH SEM	E_W_L_U_ H	16	84.804	0	0.96	Rejected

Table 3: Sum of Chi Square value for all behavioral parameters of students.

Hypothesis	Sum of ChiSquare Value
D_R_H	271.741
E_W_L_U_H	533.767
W_L_H	577.311
W_Li_U	432.819

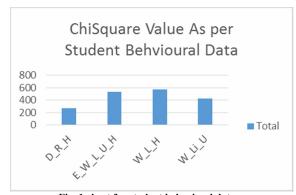


Fig. 1 chart for student behavioral data

B. Hypothesis2: Is this any relationship exist or not, between Internal Theory Exam, Internal Practical Exam, External Theory Exam, External Practical Exam, Assignment Regularity, Subject Attendance Ratio and Faculty performance on Result of All the semesters?

INDEPENDENT VARIABLE: INTERNAL THEORY MARKS, INTERNALPRACTICALMARKS, EXTERNAL THEORYMARKS, EXTERNAL PRACTICAL MARKS, ASSIGNMENT REGULARITY, SUBJECT ATTENDANCE RATIO, FACULTYPERFORMANCE

DEPENDENT VARIABLE: RESULT OF ALL SEMESTERS

H0 (Null Hypothesis): Parameters like Internal Theory and Practical Marks, External Theory and Practical Marks, Assignment Regularity, Subject Attendance Ratio and Faculty performance are independent (Not Dependent) on Result of all the semesters.

H1 (Alternative Hypothesis): Parameters like Internal Theory and Practical Marks, External Theory and Practical Marks, Assignment Regularity, Subject Attendance Ratio and Faculty performance are dependent on Result of all the semesters.

Analysis and Results to prove above hypothesis is as per the following:

First Semester Statistical Analysis:

Table 4: Chi Square analysis for all the subject of first semester

Table 4: Chi Square analysis for all the subject of first semester Chi Square Test for Subject 1 of First Semester						
SN	HT	DF	CV	SIG	COV	Result
	INT_TH	16	955.364	0	0.476	Rejected
	INT_PR	16	824.285	0	0.459	Rejected
	EXT_T H	16	908.8	0	0.479	Rejected
CLAN	EXT_P R	16	891.536	0	0.453	Rejected
	ASS_M	16	121.343	0	-0.75	Rejected
	ATT_M	16	73.907	0	-0.32	Rejected
	FAC_P	16	85.085	0	-0.00 4	Rejected
	Chi Square	Test f	or Subject 2	of First S	Semester	
SN	HT	DF	CV	SIG	COV	Result
	INT_TH	16	1042.52 1	0	0.499	Rejected
	INT_PR	16	762.101	0	0.432	Rejected
	EXT_T H	16	870.99	0	0.443	Rejected
OFFIC E	EXT_P R	16	1059.00 1	0	0.492	Rejected
	ASS_M	16	61.203	0	-0.02 1	Rejected
	ATT_M	16	14236	0	1	Rejected
	FAC_P	16	70.409	0	-0.00 6	Rejected
	Chi Square	Test f	or Subject 3	of First S	Semester	
SN	HT	DF	CV	SIG	COV	Result
	INT_TH	12	637.085	0	0.388	Rejected
	INT_PR	12	974.338	0	0.486	Rejected
	EXT_T H	12	886.474	0	0.466	Rejected
DE	EXT_P R	12	842.264	0	0.456	Rejected
	ASS_M	12	79.223	0	-0.00 5	Rejected
	ATT_M	12	28.317	0.00	0.03	Accepte d
	FAC_P	12	98.075	0	-0.01 5	Rejected

	Chi Square Test for Subject 4 of First Semester					
SN	HT	DF	CV	SIG	COV	Result
	INT_TH	16	2478.87 2	0	0.713	Rejected
	EXT_T H	16	2492.56 8	0	0.699	Rejected
CS_I	ASS_M	16	135.71	0	0.017	Rejected
	ATT_M	16	98.31	0	0.009	Rejected
	FAC_P	16	74.855	0	0.022	Rejected
	Chi Square	Test f	or Subject 5	of First S	Semester	
SN	HT	DF	CV	SIG	COV	Result
	INT_TH	16	2251.24	0	0.685	Rejected
	EXT_T H	16	2237.62 9	0	0.674	Rejected
IT	ASS_M	16	30.043	0.01 8	0.012	Accepte d
	ATT_M	16	144.429	0	-0.48	Rejected
	FAC_P	16	110.975	0	-0.00 7	Rejected

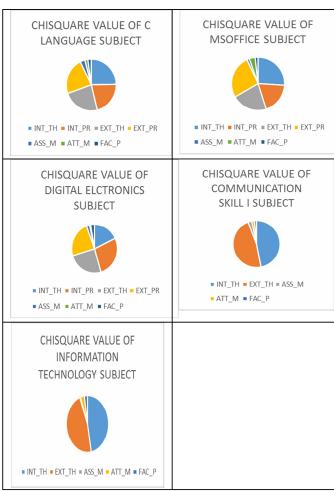


Fig. 2 pie chart for chi square value of all the subject of first semester.

Note 1: Same Mechanism is Repeat for all the Six Semester Note 2: List of Abbreviation described in appendix A.

C. Hypothesis3: Is this any relationship exist or not, between HSC Percentage, HSC Stream, Father Annual Income and Result of All the semesters?

INDEPENDENT VARIABLE: HSC PERCENTAGE, HSC STREAM, FATHER ANNUAL INCOME

DEPENDENT VARIABLE: RESULT OF ALL SEMESTERS

H0 (Null Hypothesis): Parameters like HSC Percentage, HSC Stream and Father Annual Income are independent (Not Dependent) on Result of all the semesters.

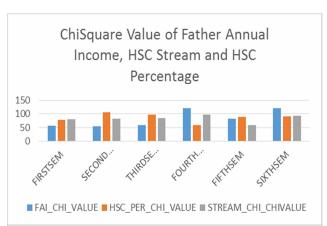
H1 (Alternative Hypothesis): Parameters like HSC Percentage, HSC Stream and Father Annual Income are independent (Not Dependent) on Result of all the semesters. are dependent on Result of all the semesters.

Analysis and Results to prove above hypothesis is as per the following:

Chi-square value analysis for Father Annual Income, HSC Stream, HSC Percentage:

Table 5: Chi-square value analysis for Father Annual Income, HSC Stream, HSC Percentage

	FAI	HSC_PER	STREAM
FIRSTSEM	55.97 7	78.535	80.199
SEC O NDSEM	54.71	105.246	82.858
THIRDSEM	58.26 2	96.697	85.477
FO URTHSEM	121.1 09	59.458	96.485
FIFTHSEM	81.37 6	89.726	58.668
SIXTHSEM	120.6 47	91.386	93.648



Fig, 3 Bar chart for chi square value of father annual Income, HSC stream and HSC Percentage.

Summary Finding of Chi Square Result Analysis of All the parameters affected on student performance

Table 6: Chi Square Result Analysis of All the parameters affected on student performance

PARAMETER	CO MBINE CHI_ VALUE	CO MBINE CO RELATIO N_ VALUE
INT_TH	1530.308	0.580

	INT_PR	855.123	0.424
	EXT_TH	1537.369	0.563
S	EXT_PR	891.288	0.464
	ASS_M	290.012	0.162
	ATT_M	368.808	0.202
Ī	FAC_P	82.232	0.090
	D_R_H	271.741	0.082
	E_W_L_U_H	233.767	0.094
	W_L_H	277.311	0.099
Ī	W_Li_U	132.819	-0.052
Ī	FAI	182.013	-0.057
	HSC_PER	82.841	-0.002
	HSC_STREAM	186.889	0.072

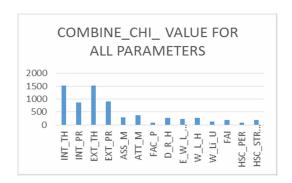


Fig. 4 Combine chi square value of all affected parameters on student performance.

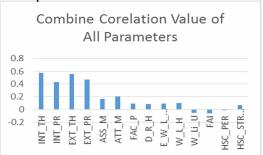


Fig. 5 Combine correlation value of all affected parameters on student performance.

VII. OUTCOME OF ANALYSIS:

By Analysis we have found that Students who have good score in Attendance and Assignment marks can get good score in the examination. Internal Theory and Practical have highest impact on result of examinations. We also found that Internal Marks and External Marks have strong correlation ship so those student who got good score in internal can have chance to get good score in external examination also. Based on statistics result we also prove, Father Income also dependent on student score but correlation shows it is negative co-relationship it means that students whose family income is low or average or middle may get good score but those students whose family income is Very high those students are not too much sincere in study and get low score.

This result also show that there not much impact of library usage on the result of examination.

VIII. CONCLUSION

In this study, we have used 232 attributes with 3558 randomly selected record for the analysis. This 232 attributes have all the demographic and academic records of the students. In this research study we have used SPSS tool and chi square technique for performing analysis on student's data. In this study student's data are categoricalso chisquare technique is most preferable technique for analysis of the data. We also have prepared some hypothesis for find out the highly affected parameters of student performance. And based on the outcome we conclude that students who are using maximum Computer Lab and Daily Reading Practice can get good score in the examination. Based on analysis we also concluded that students who are good in attendance and assignment can get good score in the examination. Based on the outcome of the analysis, highest affected parameters on student's results are their internal and external performance in theory and practical head. Based on statistics result we also prove, Father Income also dependent on student score but correlation shows it is negative co-relationship it means that students whose family income is low or average or middle may get good score but those students whose family income is very high those students are not too much sincere in study and get low score. This result also show that there not much impact of library usage on the result of examination. So if we set the parameter

IX. FUTURE SCOPE

This research study aim is only to find the highest affected parameters on the students' performance based on analysis of student's demographic information and academic records. But researcher can carry on further research by applying different data mining techniques and can do the prediction, classification, clustering and association on the student data. It willhelp the decision makers for getting improvement in the performance of students.

APPENDIX

APPENDIX A: 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		
SN	Subject Name		
HT	Hypothesis Testing		
DF	Degree of Freedom		
CV	Calculated value of Chi Squared		
SIG.	Significance value of Chi-Square		
COV	Corelation Value		
FAI	Father Annual Income		

ACKNO WLEDGMENT

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] Anne F. Maben, 2005, Chi-square test adapted from Statistics for the Social Sciences.
- [2] C.R Kothari, 2006, Research Methodology Techniques, New Age International (P) Limited.
- [3] J. Shana, T. Venkatachalam, 2011, "Identifying key performance indicators and predicting the result from student data", International Journal of Computer Application (0975 - 8587) Volume 25 – No. 9.
- [4] J. Garczynski, Chi-square Test in SPSS (PASW), Reference Department Albert S. Cook Library 410 -705-5168.
- [5] Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md Nasir B Sulaiman, Udiz, Feature Selection and Classification Techni Automatic Genre Classification of Traditional Mala Music ISMIR- Session 3A - Content Based Retrieval, Categorization and Similarity.
- [6] M.A Hall, and L.A Smith, 1998, Practical feature subset selection for machine learning, In Proceedings of th Aust ralian Computer Science Conference, pp181
- [7] S. Bhoite, S. Pathare, 2010, "Student Academic Performance Analysis Using Data Mining technique in Higher Education: A Case Study of Collage" proceeding of 4th National Conference: Indiacom – 2010, February 25-26, New Delhi
- [8] Rongrong LU1, Jun'e LIU 2, "The research of the knowledge management technology in the education", International Symposium on Knowledge Acquisition and Modeling, Beijing, China. 2008.
- [9] Thangamuthu, C., (2008) National Knowledge Commission Report on Higher Education, *University News*, Vol.45,48, Nov.-Dec., 7-15.
- [10] J. Pallant, 2013, SPSS Survival Manual: A Step by step guide to data analysis using IBM SPSS, 5th Edition
- [11] SPSS Support available at: www.spss.co.in/

AUTHO RS PRO FILE



Prof. Bhavesh Patel(First Author), 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



Dr. Jyotindra Dharwa is an associate professor in MCA programme of Acharya Motibhai Patel Instititue of Computer Studies, Ganpat University. He is Ph. D. in computer science and has 16 years of experience in research and academics. He has published more than 16 papers in National and International level Seminars & Conferences.