IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Significance of Influencing Factors' Relationship in Education Domain

Jai Ruby[1*], K. David[2]

[1]MCA Dept., Sarah Tucker College & Research Scholar, Research & Development Centre, Bharathiar University, Tamilnadu, India
[2]Dept. of Computer Science, H. H The Rajahs College, Pudukottai, Tamilnadu, India

Corresponding Author : ajairuby@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7si8.1015 | Available online at: www.ijcseonline.org

Abstract— In recent years, Educational Data Mining has developed into a research realm. All educational institutions are striving hard to prove themselves as the best to attract the student community. Academic performance of the students plays a vital role in determining the status of the institution. So, all the institutions strive hard to know their wards in advance and improve their performance to stand out among their competitors. Unfortunately most the information in the academic institutions are hidden and need to be extracted out. Data mining is a well known technique for bringing out the hidden potential of the institutions. For mining, the data need to be transformed and reduced for better performance. This model is mainly focused on finding out the significance of the relationship between the derived influencing factors.

Keywords *- Educational Data Mining, Academic Performance, Higher Education, Prediction, Contingency table, Chi-square.*

I. INTRODUCTION

In today's set-up, Educational Institutions are growing exponentially around the world. All the institutions are pulling hard to prove themselves to attract the students. Even though there exists various factors in determining the quality of the institutions, it's the academic quality and the output of the student plays a vital role. The students' academic performance can be improved if one can predict the student temperament and the academic stand of the student in advance. There exists various techniques to predict the student performance in advance. The hidden pattern can be revealed using mining techniques. Mining on the education data is said as Educational Data Mining. It is a multi-disciplinary research area that deals with artificial intelligence, statistics and data mining with the data generated from educational institutions.

The students have many factors to influence their academic performance. Various information related to socio-economic, psychological and environmental factors have great impact on the students' academic output. The significance of the impact of the factors need to be understood in order to predict the result accurately. It's also necessary to store the details in intelligent knowledge base for decision making and for creating self learning system for better prediction in advance. To improve the prediction accuracy is the key issue in the Educational Data Mining. In this paper, the researcher uses chi-square technique to find the significance of different attributes of the students in predicting the performance. The experiment strengthens the fact that various factors identified in the study model [1] are really influencing in predicting the results.

This paper makes an attempt to find the significance of influencing factors in predicting the academic performance of the students. Section 2 gives the terms and statistical techniques used. Section 3 provides the general account of the model, tools and the dataset under study. Section 4 deals with the experimental result and analysis. Last Session deals with conclusion and future work.

A. Related Work

Han and Kamber [2] says that various datamining tools can be used to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. To analyze, manage and make a decision on huge amount of data of various types we are in need of techniques in data mining. In Data Mining data sets will be explored to yield hidden and unknown predictions which can be used in future for efficient decision making. Data Mining involves various concepts mathematical and statistical approach to search data and help the analyst in recognizing significant trends, facts relationships and anomalies [3].

Data reduction has been used extensively in data mining for suitable analysis. Principal component analysis (PCA) and factor analysis (FA) methods are some of the popular techniques. The PCA and FA reduce the number of variables to avoid the curse of dimensionality [4]. The difficulty in handling of dimensionality which causes an increase in the

computing time exponentially is proportional to the number of variables. So, many methods have been published for dimension reduction [2,5]. Data reduction helps in reducing the dimensionality and hence improves the computational speed and memory. Categorical variables that are proved to be nominal and ordinal variables are described by tabulating their frequencies or probability. If two variables are associated, then it's always true that the probability of one will depend on the probability of the other. Among many techniques, chi square tests the association between two categorical variables and contingency table analysis allows us to quantify their association [6] in a better way. Even though there exists many correlation techniques to find the relationship between two variables, chi-test is used for finding the relationship when categorical attributes are involved. The correlation coefficient is only one of many possible relations between variables.

One of the most common statistical methods to measure associations between two categorical variables is the chi-square test [7]. The Chi-square test of independence is one of the most powerful statistics for testing hypotheses when the variables are nominal. Chi-square ($\chi 2$) can provide information not only on the significance of any observed differences, but also provides detailed information on exactly which categories account for it. The amount and detail of information this measure can provide shows that it is one of the most useful analysis tool. In addition, the $\chi 2$ is a significance test, and always be coupled with an appropriate test of strength[10]. Chi-square tests is said as one of the most utilized statistical analyses for the association or difference between categorical variables [11]. Chi-square test is a nonparametric test used for two purposes: (a) To test the hypothesis of no association between the groups. population or criteria (b) and to test how likely the observed distribution of data fits with the distribution that is expected which is said as ni-of-fit)[12].

II. TERMS AND TECHNIQUES

Knowledge Discovery in Databases (KDD) refers to extracting or "mining" information in the type of rules, patterns or models from large amount of data. It involves various process like Data pre-processing, data mining, pattern evaluation in extracting knowledge from data. Data cleaning, data integration, data transformation, data reduction, and data discretization are the various tasks involved in data pre-processing.

(i) Data Reduction

Data reduction means reducing or dropping the volume but producing the identical or similar analytical results. Data reduction can be done in various ways like (i) reducing the number of attributes (ii) reducing the number of attribute values (iii) reducing the number of tuples. The authors in [1] have chosen the third technique of reducing the number of attributes for data reduction. The study extracted the attributes that are high influential in predicting the academic results.

(ii) Kinds of Data

In computing, data is information that has been translated into a form that is efficient for processing. The educational data has attributes whose value falls into different categories and it needs to be processed so as to find the significance between the attributes. There are four types of data. They are nominal, ordinal, interval and ratio. Nominal data basically refers to categorically discrete data. Nominal scales are used for labelling attributes or variables, without any quantitative value. Ordinal refers to quantities that have a normal ordering. The ranking of students or the order of runners finishing a race is a good example for ordinal data. Items are ordered into some kind depending on the position on the scale. Interval data is like ordinal that it is measured along a scale in which each position is equidistant from one another. Interval data appears in the form of numerical or number values every time where the distance between the two points is standardized and equal. In a ratio scale, values are compared as multiples of one another. A ratio has all the property of an interval variable and also has a clear definition of 0.0. It is quantitative as interval data.

(iii) Contingency Table

A contingency table also known as a cross tabulation is a kind of table in a matrix form that exhibits the frequency distribution of the variables. A contingency table is a unique type of frequency distribution table, where two variables are shown simultaneously. They provide a basic representation of the interrelation between two variables and helps in finding the interactions between them. Each row refers to a specific sub-group in the population, the columns are sometimes referred to as banner points or cuts [8]. The table showing the distribution of one variable in rows and another in columns, used to study the correlation or association between the two variables. Usually a chi-square test could then be run on the table to determine if there is a relationship between the two variables.

Table 2.1 Contingency Table

	Column 1	Column 2	Grand Total
Category 1	x1	y1	x1+y1
Category 2	x2	y2	x2+y2
Grand Total	x1+x2	y1+y2	(x1+y1)+(x2+y2)

Table 1.1 shows the Contingency table where x1 and y1 are the count of the values of two categorical variables.

(iv) Chi-Square

Chi-square test is a statistical method that is used to build out the degree of association between variables [9].
'The formula for calculating chi-square ($\chi 2$) is:

$$\chi 2 = \frac{\sum(\text{ob} - \text{ex})^2}{\text{ex}}$$

Chi-square is the sum of the squared difference between the observed (ob) data and the expected (ex) data, divided by the expected data of all the attributes taken into consideration.

III. DATASET, TOOL AND METHODOLOGY

From the previous study in [1] it is understood that among various attributes of the educational dataset few are highly influential and they help for better prediction of academic activities. In order to find out the contribution of each attribute, it is necessary to find out the relationship between the two attributes. There are two different ways to find out the relationship between categorical or ordinal variables. The relationship can be found out using either distance metrics such as Euclidean distance or Manhattan distance or by using various statistical metrics such as chi-square test. Since distance measures are scale dependent, it is preferred to use chi square test. To represent the relationship between two variables contingency table is used where for each combination of attributes we have to specify the independent and the dependent variable. Independent variable is seen as the reason and the dependent variable is seen as the result.
Three different sets of data are considered for the study. The data is gathered from two different colleges using a questionnaire based on the high impact features identified in the previous study [1]. One of the dataset is the 165 records collected from an arts and science college through their ward register, assessment register and attendance. The dataset is related to socio economic, personal and academic details for the students. The partial manual data from the college is then digitized.

First educational dataset has 13 attributes which includes Theory Scores, Laboratory scores, Medium of study, Previous course score, Family Income, Parental Education, First Generation Learner, Stay, Extracurricular activities, Previous Course Studied, Living setup, Attendance and Result. Among these attributes family income, previous course studied, Previous course score, Extracurricular activities, Theory marks and Stay are found to be highly influential [1].

Two other datasets are collected from different colleges through questionnaire with questions mainly focussing on the influencing factors identified in the earlier study. Both the datasets that consists of 66 records each. Table 2.2 shows the college dataset prepared from various sources. Table 2.3 shows the sample questionnaire used for data collection.

Table 2.2 – Processed College dataset from Various sources

Family Income	Prev Course Studied	Prev Percentage	Stay	Medium of Study	Theory	Extra Curr.	Result
8000	CS	3	68	T	72.6	N	78.86
6000	CS	4	82	E	76.6	Y	81.86
8000	IT	3	70.8	E	72.8	Y	79.29
10000	CS	3	74	E	78.8	N	82.71
1000	PHY	3	75	T	80	N	81.71
20000	CS	2	53	E	68.2	N	74.14

Table 2.3 Sample Questionnaire used for Data Collection

Q. No.	Question	Answer Choice
1.	Family Income :	1.Upto 10000 2.10000-25000 3. Above 25000
2.	Area of Living :	a. Urban b. Rural
3.	Hostel / Stay :	0. No 1. Yes
4.	Medium of study :	0. Tamil 1. English
5.	Do you participate in Extra Curricular Activities ?	0. No 1. Yes
6.	Do you have a prior knowledge about your current course in school?	0.No 1. Yes
7.	Are you a first Generation Learner ?	0. No 1. Yes

Table 2.4 shows the users choice derived from the questionnaire of College Dataset-1 taking into account only the high influencing factors concerning the personal, socio economic and academic details.

Table 2.4 – User choice from questionnaire of College dataset-1

Family Income	Prev Course Studied	Prev %	Stay	Medium of Study	Theory	Extra Curr.	Result
2	1	3	1	1	2	0	1
2	1	4	1	0	3	1	2
1	1	3	0	0	2	1	1
1	1	3	1	0	3	0	2
3	0	3	1	1	3	0	0
3	1	2	1	0	2	0	1
3	1	3	1	1	3	1	2
2	1	3	1	0	3	1	2
3	1	3	0	0	2	0	2
1	1	3	0	1	3	0	2
2	1	3	1	0	2	1	2

The main objective is to find the relationship of the high influencing attributes in predicting the results. The regression analysis estimates the relationship between two or more variables. It indicates the significant relationship between dependent variable and independent variable. It also indicates the strength of the impact of multiple independent variables on a dependent variable.

The data gathered using the questionnaire falls under categorical data type. All the three datasets are subjected to chi-square test. Contingency table is prepared for the each independent attribute against the dependent variable. For eg. Contingency table for the attribute 'Stay' is prepared against the dependent variable 'Result'. Table 2.5 shows the contingency tables for the attributes family income, medium of study using dataset College Dataset-1.

Table 2.5 Contingency tables of attributes of College Dataset -1

Family Income				
Row Labels	**0**	**1**	**2**	**Total**
1 (Low)	3	36	28	67
2 (Avg.)	5	28	21	54
3 (Good)	6	24	14	44
Total	**14**	**88**	**63**	**165**

Medium of Study				
Row Labels	**0**	**1**	**2**	**Total**
0 (Tamil)	5	52	28	85
1(English)	9	36	35	80
Total	**14**	**88**	**63**	**165**

To find the association between the attributes of the dataset chi-square test is used. Chi-square is calculated using the observed and expected values of dataset for each attribute. The observed values are shown in the contingency tables of the above figure.

To calculate the expected value, the formula used is

$$\text{Exp. Attri} = \frac{total\ of\ row_i * total\ of\ col_i}{grand\ total}$$

Table 2.6 Observed values of attribute 'family income' of College Dataset-1

Family income – College Dataset-1				
	0	1	2	
1(low)	3	36	28	67
2(avg)	5	28	21	54
3(good)	6	24	14	44
	14	88	63	165

Table 2.6 shows the observed values of the 'family income' attribute of the 'College dataset-1' and Table 2.7 shows the calculated expected values for the same attribute.

Table 2.7 Expected values of attribute 'family income' of College Dataset-1

Family income – College Dataset-1				
	0	1	2	
1(low)	3	36	28	67
2(avg)	5	28	21	54
3(good)	6	24	14	44
	14	88	63	165

The formula for calculating chi-square ($\chi 2$) is:

$$\chi 2 = \frac{\sum (ob - ex)^2}{ex}$$

where 'ob' is observed and 'ex' stands for expected value.

For the experiment, the alpha level of significance is taken as 0.5 and the degrees of freedom is taken as (number of columns minus one) * (number of rows minus one) for the contingency table. For eg. For the attribute family income and result the degrees of freedom is (3-1) * (3-1) = 4. Table 2.8 shows the probability distribution for various degrees of freedom.

Table 2.8 Probability Distribution for Degrees of freedom

Df	0.5	0.10	0.05	0.02	0.01	0.001
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517

IV. RESULT AND DISCUSSION

Three different college datasets taken for the study is subjected to chi-square test. Contingency table is created for each of the attributes of dataset. Using the contingency table, the chi-square value for each attribute is calculated. Table 2.9 shows the chi-square values for different attributes of the three datasets.

Table 2.9 Chi-square values of attributes of three college datasets

Datasets	Family Income	Previous Course	Stay	Prev %	Medium of Study	Extra Curr	Theory
Data1	3.382	0.600	7.765	8.30	4.682	7.446	3.89
Data2	17.556	2.795	8.760	1.91	46.102	0.291	3.88
Data3	5.786	10.347	15.77	4.56	1.250	49.13	0.53

Using the chi square value and the degree of freedom, the corresponding probability is found out. If the chi-square value exceeds the critical value 0.5 then the null hypothesis is rejected and the two attributes are said to be dependent.

For the Attribute 'Stay':

H_o : Stay in hostel is independent of the Result

H_a : Stay in Hostel is associated with the Result

The 'df' for each attribute and the Chi-square values from Table 2.9 are used for the computation of significance of the relationship between the attributes with the help of the probability distribution values in Table 2.8

The significance calculated for the attribute 'Stay' for all the three datasets show that the values are greater than 0.5 and the null hypothesis Ho is rejected. It is concluded that 'Stay' is associated with the 'Result'. Also it is observed that increase in the attribute values shows an increase in 'result' values.

Table 2.10 Significance of freedom of attributes for 3 datasets

Datasets	Family Income	Prev. Course	Stay	Prev. %	Med of Study	Extra Curr	Theory
College Data-1	0.5 – 0.1	>.5	.05-.02	0.1 - .05	0.1 – .05	.05-.02	0.5-0.1
College Data-2	>.01	0.5-0.1	0.1 – .05	0.5-0.1	< .001	> 0.5	0.5-0.1
College Data-3	0.1 - .05	> 0.1	> 0.1	0.5-0.1	>.5	< .001	>.5

From the Table 2.10 it is observed that the result for the college dataset-1, except 'previous course' all the other attributes have a significance relationship with the 'result' attribute. For the College Dataset-2, other than 'extracurricular' others have a good dependency with the 'result' attribute. Considering College Dataset-3, 'medium of study' and 'theory' has less significance in prediction of the result. It is found that among the high influential attributes 'Family Income', 'Stay', and 'Prev. Percentage' are dependent on the result irrespective of the dataset.

V. CONCLUSION

The experiment proves the significance of influencing factors in educational domain for the academic performances is noteworthy. The identification of the performance influencing factors helps the institution to decide on the factors to concentrate for the better performance of the academic results of the students. The institutions can use this extracted knowledge and if they know the values of the above mentioned high influencing attributes they can predict student Performance in advance and enhance their quality of student output. Few parameters for a higher education institution is experimented and it can be extended to various levels using large datasets in primary and secondary education system

REFERENCES

[1] Jai Ruby & K. David, "A study model on the impact of various indicators in the performance of students in higher education", IJRET International Journal of Research in Engineering and Technology, Vol. 3, Issue 5, May-2014, pp.750-755.

[2] Han. J & Kamber. M, "Data mining concepts and techniques", San Francisco, USA, Morgan Kaufmann, 2001.
[3] A J. Chamatkar et al, "Importance of Data Mining with Different Types of Data Applications and Challenging Areas", Int. Journal of Engineering Research and Applications ISSN : 2248-9622, Vol. 4, Issue 5(Version 3), May 2014, pp.38-41
[4] Daiho Uhm, Sunghae Jun and Seung-Joo Lee,"A Classification Method Using Data Reduction", International Journal of Fuzzy Logic and Intelligent Systems, Vol. 12, no. 1, March 2012, pp. 1-5, pISSN 1598-2645
[5] J. H. Friedman, "On Bias, Variance, 0/1-loss, and the Curse of Dimensionality," Data Mining and Knowledge Discovery Vol. 1, pp. 55-77, 1997.
[6] Cynthia Fraser," Association between Two Categorical Variables: Contingency Analysis with Chi Square. In: Business Statistics for Competitive Advantage with Excel 2007", Springer, New York, NY
[7]] Jinwook Seo & Heather Gordish-Dressman , "Exploratory Data Analysis with Categorical Variables: An Improved Rank-by-Feature Framework and a Case Study",Journal International Journal of Human–Computer Interaction Volume 23, 2007 - Issue 3, Pages 287-314
[8] https://en.wikipedia.org/wiki/Contingency_table
[9] Anne F. Maben, 2005, Chi-square test adapted from Statistics for the Social Sciences.
[10] McHugh, Mary L. "The chi-square test of independence" Biochemia medica, Vol. 23(2), 2013, Pages 143-149.
[11] Franke, Todd & Ho, Timothy & A. Christie, Christina. "The Chi-Square Test Often Used and More Often Misinterpreted. American Journal of Evaluation. 33(3), 2012, Pages 448-458.
[12] Rana R, Singhal R. Chi-square test and its application in hypothesis testing. J Pract Cardiovasc Sci 2015;1 @ Pages 69-71.

Author Profile

Jai Ruby is a Research Scholar in Research & Development Centre, Bharathiar University, Tamilnadu, India. She is working as an Associate Professor in the Department of Computer Applications, Sarah Tucker College,Tirunelveli, Tamilnadu, India. She has 18 years experience in teaching field. Her research interest includes Data Mining,Mobile Communication and Data Analytics.

Dr. K. David is working as an Assistant Professor, Department of Computer Science, H.H. The Rajah's College, Pudukkottai, Tamilnadu, India. He has over 20 years of teaching experience and about 4.5 years of Industry experience. He has published scores of papers in peer reviewed journals of national and international repute and he had guided 2 Ph.D scholars and is currently guiding 5 Ph.D scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.

