A STUDY MODEL ON THE IMPACT OF VARIOUS INDICATORS IN THE PERFORMANCE OF STUDENTS IN HIGHER EDUCATION

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

[1] *Research Scholar, Research & Development Centre, Bharathiar University, Tamilnadu, India*
[2] *Associate Professor, Department of Computer Science and Engineering, Roever Engineering College, Perambalur, Tamilnadu, India*

Abstract

In this technology revolutionized century knowledge has become a vital resource. Also, Education has been viewed as a crucial factor in contributing to the welfare of the country. Higher education does categorize the students by their academic performance. In higher education institutions a substantial amount of knowledge is hidden and need to be extracted using Knowledge Discovery process. Data mining helps to extract the knowledge from available dataset and should be created as knowledge intelligence for the benefit of the institution. Many factors influence the academic performance of the student. The study model is mainly focused on exploring various indicators that have an effect on the academic performance of the students. The study result shows the impact of various factors affecting the students of higher education system. The extracted information that describes student performance can be stored as intelligent knowledge for decision making to improve the quality of education in institutions.

Key Words: *Educational Data Mining, Academic Performance, Higher Education, Attribute Selection, Intelligent Knowledge*

---***---

1. INTRODUCTION

In today’s scenario, educational institutions are becoming more competitive because of the number of institutions growing rapidly. To stay afloat, these institutions are focusing more on improving various aspects and one important factor among them is quality learning. Today, learning has taken various dimensions such as online learning, virtual learning, socializing etc. For providing quality education and to face new challenges, the institutions need to know about their potentials which are explicitly seen and which are hidden. The truths behind today’s educational institutions are a substantial amount of knowledge is hidden. To be competitive, the institutions should identify their own potentials hidden and implement a technique to bring it out.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering and extracting hidden and potentially useful information from large amounts of data. Data mining is applied in various fields like medical, marketing, databases, machine learning, artificial intelligence, customer relations etc., Recently Data mining is widely used on educational dataset. Educational Data Mining (EDM) has become a very useful research area [1]. Educational Data Mining refers to techniques, tools, and research designed for automatically extracting meaning from large repositories of data generated by or related to people's learning activities in educational settings. Key uses of EDM [2] include learning and predicting student performance in order to recommend improvements to current educational practice. EDM can be considered as one of the learning sciences, as well as an area of data mining [3]. Romero and Ventura [13], did a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has specific requirements not presented in other domains. Some of the benefits of data mining in an education sector are identifying students’ needs and preferences towards course choices, and selection of specialisation , identifying students’ pattern trends, predicting students’ knowledge, grades, and final results, supporting automatic exploration of data, ‘constructing students’ profiles become easy, and helping management to understand business [10].

Sir Francis Bacon (1597) commented, “Knowledge is power” and in today’s context it may be rephrased as “Knowledge sharing is power”. The extracted information from the data can be transferred as knowledge and can be stored in decision making for the betterment of the institution. Institutions of Higher Learning (IHL) are similar to knowledge businesses, in that both are involved in knowledge creation, dissemination, and learning[11]. However, people in business world concerned with the profit they could gain by exploiting knowledge through the implementation of KMS whereas IHL consider that KMS could improve the quality of service deliveries and sustained competitive advantages in the academic world [12]. Different models have been used by these researchers to describe the factors found to influence student achievement, course completion rates, and withdrawal, along with the relationships between variable factors[14].

This paper makes a novel attempt to look into the higher educational domain of data mining to analyze the students' performance. Section 2 gives the overview of data mining techniques available to extract the hidden information and attribute selection methods. Section 3 provides the general account of the data under study and the pre-process stage of the data. Section 4 analyzes the impact of various indicators in the performance of students in higher education and applying various data mining techniques. Conclusion and a discussion on future work are in the final section.

1.1 Related Work

In [5] authors proved that data mining for small data sets has a real potential to become a serious part of higher education teachers' knowledge management systems. Also the study result show that student data, available to higher education teachers which falls into the category of a small data set carries enough student-specific characteristics in the sense of hidden knowledge which can be successfully associated with student success rates. In [9] authors used various different feature selection methods and have found out the influence of features affecting the student performance. The authors have used a selected number of attributes and have not taken attributes like attendance, theory, laboratory etc. The researchers in [6] conducted a study on a data set of size 50 MCA students for mining educational data to analyze students' performance. Decision tree method was used for classification and to predict the performance of the students. Different measures that are not taken into consideration were economic background, technology exposure etc. El-Halees.A [13] has done a work on mining students data to analyze learning behavior. The data size considered was 151. The details include personal and academic records of students. Classification based on Decision tree is done followed by clustering and outlier analysis. The knowledge extracted describe the student behaviour. Han and Kamber [4] depicted the data mining process and the methods to analyze data from different perspective and the steps to mine knowledge.

.

2. DATA MINING TECHNIQUES AND ATTRIBUTE SELECTION METHODS

Data mining also termed as Knowledge Discovery in Databases (KDD) refers to extracting or "mining" knowledge from large amount of data [4]. Knowledge Discovery process involve various steps in extracting knowledge from data as shown in Fig. 1.

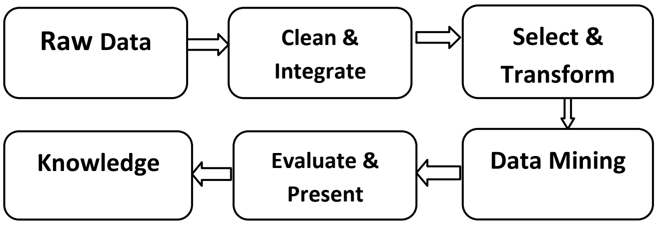

Fig. 1. Steps in the process of Knowledge Discovery

Data cleaning is the process, which is used to remove noise and inconsistent data. Data Cleaning routines do fill in missing values, smooth out noise and identify outliers and correct inconsistencies in the data [4]. Transformation is a technique which is used to make the data minable. To discover useful patterns within the data, we apply data mining methods. The hidden patterns, associations and anomalies in a dataset that are discovered by some Data mining techniques, can be used to improve the effectiveness, efficiency and the speed of the processes [6]. Different techniques and models are applied like neural networks, Bayesian networks, rule based systems, regression and correlation analysis to analyze educational data[3]. Evaluation is used to extract data with interest. Knowledge Discovery is involved in a multitude of tasks such as association, clustering, classification, prediction, etc. Classification and prediction are functions which are used to create models that are constructed by analyzing data and then used for assessing other data. Clustering is a way of identifying similar classes of objects. Association is mainly used to relate frequent item set among large data sets. Data mining for small data sets has a real potential to become a serious part of higher education teachers' Knowledge Management Systems [5]. This study is carried out using a small dataset with a number of attributes to analyze the performance of the students. Feature selection has been an active and fruitful field of research area in pattern recognition, machine learning, statistics and data mining communities [15, 16]. Various attribute selection methods do exists to identify the attributes that make great impact. Some of the notable methods are chi-square, information gain, correlation, gain ratio, and regression.

2.1 Chi-square

Chi-square test is a statistical method used to identify degree of association between variables [7]. The formula for calculating chi-square ($\chi 2$) is:

$$\chi 2 = \frac{\Sigma (o-e)^2}{\mathrm{e}}$$

That is, chi-square is the sum of the squared difference between observed (o) and the expected (e) data, divided by the expected data in all possible categories.

2.2 Information Gain

Information Gain is used to determine the best attribute among the attributes in a collection of samples S and if there are 'm' classes. The expected information needed to classify a given sample is

$$I(s_1 s_2, \dots . s_m) = -\sum_{i=1}^{m} \frac{s_i}{s} \log_2 \frac{s_i}{s}$$

An attribute 'A' with values {a1,a2,...,av} can be used to partition S into subsets where Sj contain those samples in S that have value aj of A. The expected information based on this partitioning by A is known as the entropy of A.

$$E(A) = \sum_{j=1}^{v} \frac{(s_{1j} + \cdots + s_{mj})}{s} I(s_{1j}, \dots . s_{mj})$$

The information gain, Gain(A) of an attribute A, in the sample set S, is given as

$$Gain\ (A) = I(s_1, s_2, \dots . s_m) - E(A)$$

2.3 Gain Ratio

Gain Ratio is also a measure to determine the best attribute. It can be calculated as

$$Gain\ Ratio(A) = \frac{Gain(A)}{Split\ Information\ (A)}$$

where 'A' is an attribute and Split Information(A) be calculated as

$$Split\ Information(A) = -\sum_{i=1}^{n} \frac{|s_i|}{|s|} \log_2 \frac{|s_i|}{|s|}$$

2.4 Linear Regression

Linear Regression involves finding the best line to fit two variables so that one variable can be used to predict other and to find a mathematical relationship between them.

$$Y = \alpha + \beta X$$

where Y is a response variable and X is a predictor variable and α, β are regression coefficients.

2.5 Correlation

Correlation is used to assess the degree of dependency between any two attributes. The correlation between the occurrence of A and B can be measured by computing

$$Corr\ (A, B) = \frac{P(AUB)}{P(A)P(B)}$$

If value is less than 1, it is negatively correlated and if greater than 1 it is positively correlated and if 1 then A and B are independent.

3. METHODOLOGY

The dataset used for this study for performance analysis was taken from PG Computer Application course offered by an Arts and Science College between 2007 and 2012. The data of 165 students were collected. Student personal and academic details along with their attendance were collected from the student information system. The collected information was integrated into a distinct table. Student dataset contains various attributes like Theory Scores, Laboratory scores, Medium of study, UG course, Family Income, Parental Education, First Generation Learner, Stay, Extracurricular activities etc. Among the 16 different attributes initially present, some of the relevant attributes which accounts to 13 was selected from the table for data mining process. The three irrelevant attributes are age, gender and community as the attribute values show only less variation. Feature selection can be useful in reducing the dimensionality of the data to be processed by the classifier, reducing execution time and improving predictive accuracy [8]. Listed below are the 12 attributes that are selected to act as predictors and the analysis will be carried using these different attributes. 'Result' is the attribute of the student dataset which act as the response variable. The Table 1 further shows the categorical values which define the possible set of values each attribute will take that can be used to analyze the given data.

Table 1 : Student Data Attribute Predictors

Attribute	Description	Categorical Values
FI	Family Income	{Good, Average, Poor}
PE	Parent Education	{No, One, Both}
PC	Previous Course	{C- Computer, NC- Non Computer}
FGL	First Gen. Learner	{Yes, No}
S	Stay	{H–Hosteller, D-Day Scholar}
LS	Living Setup	{R- Rural, U – Urban}
MS	Medium of Study	{T – Tamil, E – English}
ATD	Attendance	{Average, Good, Poor}
TY	Theory	{Average, Good, Poor, Excellent}
LAB	Laboratory	{Good, Excellent, Average, Poor}
ECA	Extra Curr. Act.	{Y, N}
UGP	UG Percentage	{Good, Excellent, Average, Poor}

The experiment is carried out with the support of SPSS statistical software. SPSS has all the capabilities for correlation, regression, classification, data reduction and clustering. Using the software, the data is cleaned by filling the missing values. Also, the selected data is transformed into categorical form if a numerical data exists. The categorical form is more suitable for applying various attribute selection techniques. For example, the family income attribute can be categorized as ‘Average’, ‘Good’ and ‘Poor’ instead of several numerical values. Thus the experimental data is pre-processed so that it is more suitable for feature selection with accuracy.

4. RESULTS AND DISCUSSION

Analysis is done to identify the dependency of predictor variables with that of the response variable. Techniques like correlation coefficient, chi-square test, information gain, gain ratio and regression are used. Feature selection is the process of removing features from the data set that are irrelevant with respect to the task that is to be performed [9]. Table 2 shows the list of factors which influence the performance of the student to a great extent in decreasing order based on various attribute selection methods.

Table 2 : Influence of Attribute using various selection methods

Chi square	Correlation	Info. Gain	Gain Ratio	Regression
TY	TY	TY	TY	S
MS	UGP	UGP	UGP	PE
PC	LAB	FI	FI	FGL
LS	ECA	ATD	ATD	FI
UGP	MS	MS	MS	UGP
S	PC	PC	PC	ATD
FGL	FGL	ECA	ECA	LS
ECA	S	S	S	PC
PE	PE	LS	LS	MS
FI	FI	FGL	FGL	ECA
LAB	ATD	PE	PE	TY
ATD	LS	LAB	LAB	LAB

Among the attributes listed in the table the first row depicts that it has high influencing value and it goes on decreasing as we move down the rows.

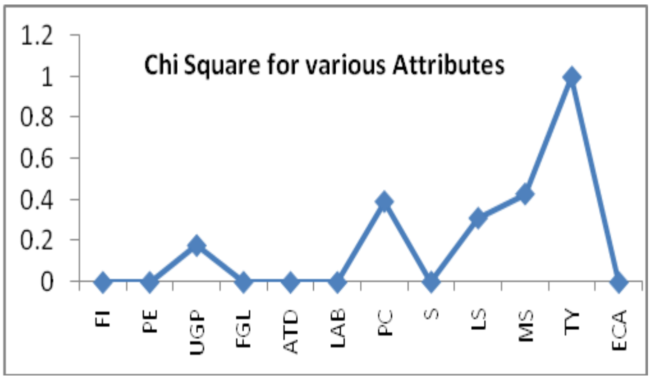

Fig - 2. Chi Square measure for various Attributes

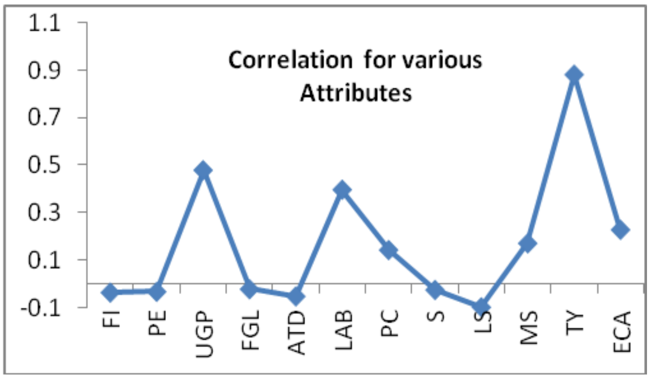

Fig - 3: Correlation measure for various Attributes

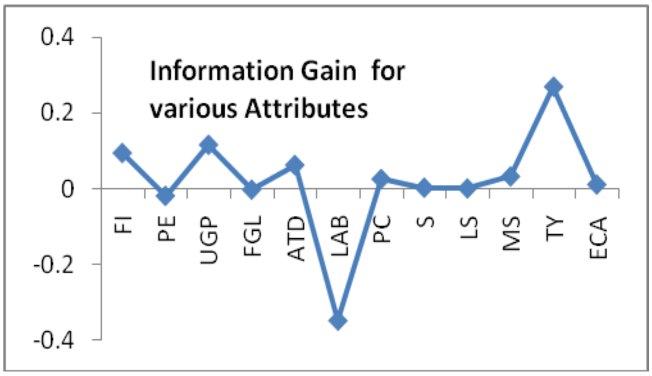

Fig - 4: Information Gain measure for various Attributes

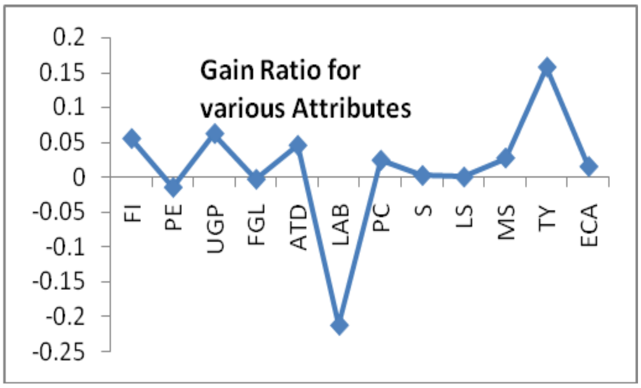

Fig - 5: Gain Ratio measure for various Attribute

The above figures show the significance of predictor attribute towards the response variable using various feature selection techniques. Fig. 2 shows the values calculated using chi square. Theory marks, Medium of Study and Previous Course Studied were the top indicators for performance prediction. Fig. 3 shows the result of using correlation analysis. Theory marks, UG percentage and Laboratory score were the top indicators. Fig. 4 shows the analysis using Information gain. The result identify that Theory marks, UG percentage and Family Income were the major influence factors. From Fig.5 it is observed that Theory marks, UG percentage and Family Income were the major influence factors using Gain Ratio.

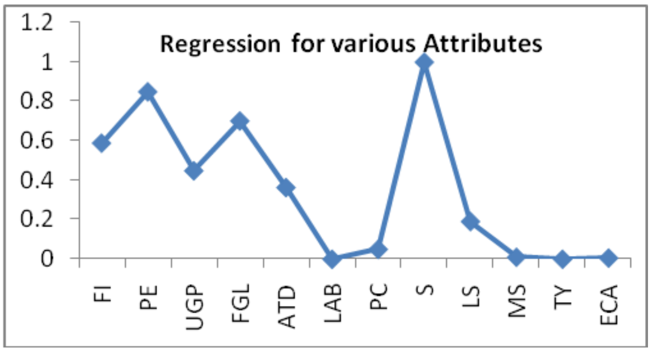

Fig - 6: Regression measure for various Attributes

Fig. 6 shows that Stay, Parent Education, First Generation Learners were the key influencers using regression. The influence factors are analyzed by categorizing the result obtained in Table 2 into 3 groups High, Medium and Low. ‘High’ is given if the attributes take 1 to 4 place and it is categorized as ‘Medium’ if it takes 5 to 8 place else it is termed as ‘Low’. Now, add weightage to High, Medium and Low category..

$$\text{Weightage} = [High * 5 + Medium * 3 + Low * 1]$$

From Table 3, we infer that the high impact attributes that contribute for the performance of the students are TY,MS, PC, UGP, S, ECA and FI (ie) Theory, Medium of Study, Previous Course studied, UG Percentage, Stay, Extra Curricular Activities and Family Income. So if we know the values of the above mentioned high influencing attributes we can predict student performance.

Table 3 : Ranking Of Attributes And Its Weightage

Attributes	High 1 - 4	Medium 5 - 8	Low 9 - 12	Weightage
TY	4	0	1	21
MS	1	3	1	15
PC	1	4	0	17
LS	1	1	3	9
UGP	3	2	0	21
S	1	4	0	17
FGL	1	2	2	13
ECA	1	3	1	15
PE	1	0	4	9
FI	3	0	2	17
LAB	1	0	4	9
ATD	1	2	2	13

5. CONCLUSION

This paper deals with the performance analysis of student. This study paper on performance analysis of student data help the institution to decide on the factors to concentrate for the better performance of the academic results of the students. The 7 attributes are selected from the 16 initial factors as more influencing for performance. Thus the hidden knowledge (performance influencing factors) was identified from a set of student data. The instructors can take steps to analyze and improve the student performance if they know the Medium of Study, UG Percentage, Theory marks obtained, Stay, Extra Curricular Activities and Family Income and whether the student was good in Previous Course studied. The study was carried out using only a small dataset and it can be extended to a large dataset and can use factors which are not dealt here. Predicting student performance by applying data mining techniques will be the future work.

REFERENCES

[1] Baker R.S.J.D., and Yacef K, 2009, The state of educational data mining in 2009: A review and future vision, Journal of Educational Data Mining, I, 3-17.

[2] Crist´obal Romero and Sebasti´an Ventura, “Educational Data Mining: A Review of the State of the Art,” IEEE Transactions on Systems, Man, and Cybernetics—Part c: Applications and Reviews, vol. 40, no. 6, 2010, pp. 601-618.

[3] Monika Goyal and Rajan Vohra, “Applications of Data Mining in Higher Education” IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012, pp.130-120.

[4] J. Han and M. Kamber “Data mining concepts and techniques”, Morgan Kaufmann, 2001.

[5] Srecko Natek, Moti Zwilling, 2013, “Data mining for small student data set – knowledge management system for higher education teachers”

[6] Brijesh Kumar Baradwaj, Saurabh Pal, IJACSA, Vol.2, No.6,2011,"Mining Educational Data to Analyze Students' Performance"

[7] Anne F. Maben, 2005, Chi-square test adapted from Statistics for the Social Sciences.

[8] Shyamala Doraisamy, Shahram Golzari, Noris Mohd. Norowi, Md Nasir B Sulaiman, Udiz, 2008, A study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Mala Music, ISMIR – Session 3 A – Content Based Retrieval, Categorization and Similarity.

[9] J. Shana, T. Venkatachalam, International Journal of Computer Applications (0975 – 8887) Vol.25-No.9 July 2011, "Identifying Key Performance Indicators and Predicting the Result from Student Data.

[10] Dr. Mohd Maqsood Ali, International Journal of Computer Science and Mobile Computing Vol.2 Issue. 4, April- 2013, pg. 374-383

[11] Rowley, J., "Is higher education ready for knowledge management?", International Journal of Educational Management, 2000, vol. 14(7), pp. 325–333.

[12] Lubega, J. T., Omona, W., & Weide, T. V. D., "Knowledge management technologies and higher education processes_: approach to integration for performance improvement", International Journal of Computing and ICT Research, 2011, vol. 5(Special Issue), pp. 55–68.

[13] El-Hales-A.(2008),"Mining Students Data to Analyze Learning Behavior: A Case Study", The 2008 International Arab Conference of Information Technology(ACIT2008)- Conference Proceedings, University of Sfax, Tunisia,Dec 15-18.

[14] "Mining Educational Data to Reduce Dropout Rates of Engineering Students", I.J. Information Engineering and Electronic Business, 2012, 2, 1-7, http://www.mecs-press.org/ DOI: 10.5815/ijieeb.2012.02.01

[15] P. Mitra, C. A. Murthy and S. K. Pal. "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312, 2002.

[16] Miller, "Subset Selection in Regression," Chapman & Hall / CRC (2nd Ed.), 2002.

BIOGRAPHIES

Jai Ruby is a Research Scholar in Research & Development Centre, Bharathiar University, Tamilnadu, India. She has 13 years experience in teaching field and research. Her current areas of research are Data Mining and Mobile Communications.

Dr. K. David is working as an Associate Professor, Department of Computer Science and Engineering, Roever Engineering College, Perambalur, Tamilnadu, India. He has over 15 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 is currently guiding 6 Ph.D scholars. His research interests include, UML, OOAD, Knowledge Management, Web Services and Software Engineering.