Sachin Ahuja

Identification of Factors Influencing GER of Female Candidates in Higher Technical Education using Decision Trees

Dr. Sachin Ahuja
Ph.D(CSE), Associate Director, Chitkara University, Punjab
Sachin.ahuja@chitkara.edu.in

Abstract: This paper explores the various factors (social, economic, demographic) that may influence the GER of females in Higher Technical Education in Punjab (India). We studied to what extent these factors can be helpful in identifying the enrolled and not enrolled female candidates in Higher Education and to focus upon the most influencing factors to increase the participation of women in terms of enrolment in higher technical education.

Keywords: Data Mining (DM), Educational Data Mining (EDM), Knowledge Discovery in Database (KDD), Gross Enrolment Ratio (GER)

1. INTRODUCTION

Data Mining can be defined as the process involved in extracting interesting, interpretable, useful and novel information from data. It has been used by various organizations, businesses, scientists and Researchers to put through a sieve from large volume of data to extract some meaningful information to better understand the hidden patter or concept. This is also known as Knowledge discovery from Database [1].

EDM is an emerging interdisciplinary research area concerned with developing methods to explore the unique types of data in educational environment and using these methods to better understand students and the environment in which they study. EDM uses data repositories to better understand learners and learning and to develop computational approaches that combine data and to transform practice to benefit learners [2].

The EDM converts raw data coming from educational system into useful information that could potentially have a great impact on educational research and practice hence this process does not differ from other application areas of data mining like business, medicine, research, genetics etc. because it follows the same steps as general data mining process [5].

Female education is a catch-all term for a complex set of issues and debates surrounding education (primary education, secondary education, tertiary education, and health education in particular) for females. It includes areas of gender equality and access to education, and its connection to the alleviation of poverty. Also involved are the issues of single-sex education and religious education in that the division of education along gender lines as well as religious teachings on education have been traditionally dominant and are still highly relevant in contemporary discussions of educating females as a global consideration [3].

In Indian perspective, Women constitute 48% of the total population of India. The principle of gender equity is enshrined in Indian Constitution in its preamble, fundamental rights, fundamental duties and directive principles and also reducing the gender gap in higher education is a focus area. The enrolment of women students i.e. 56.49 Lakhs constituting 41.40% of the total enrolment is very low (Of the total women enrolment, only 14.72% women have been enrolled in professional courses making the situation worse.)[3][4].

2. RESEARCH OBJECTIVES

The main objective of this study is to explore factors that may impact the GER outcome in the Higher Technical education in India, especially in the state of Punjab. More specifically the data collected from the survey forms was used to achieve following objectives:

- Build a model for prediction of Women candidates' enrolment in Higher education.
- Present a result which can be easily understood by the users(Researchers, Government Organizations involved in empowerment of women in India and abroad)

3. LITERATURE REVIEW

Bresfelean worked on the data collected through the surveys from senior undergraduate students at the faculty of economics & Business administration in Cluj-Napoca [7]. Decision tree algorithms in the WEKA tool, ID3 and J48 were applied to predict which students are likely to continue their education with the postgraduate degree. The model was applied on two different specializations students' data and an accuracy of 88.68 % and 71.74 % was achieved with C4.5.

P. Cortez and A. Silva [8] worked on secondary students' data to predict their grade in contact education system. Past Performance as well as socio-economic information was collected and results were obtained using different classification techniques. It was found that the tree based algorithms outperformed the methods like Neural Networks and SVM.

Z. J. Kovacic presented a case study on educational data mining in [9] to identify up to what extent the enrolment data can be used to predict student's success. The algorithms CHAID and CART were applied on student enrolment data of information system students of open polytechnic of New Zealand to get two decision trees classifying successful and unsuccessful students. The accuracy obtained with CHAID and CART was 59.4 and 60.5 respectively.

M. Ramaswami and R. Bhaskaran [10] used the CHAID prediction model to analyze the interrelation between variables that are used to predict the outcome on the performance at higher secondary school education in India. The features like medium of instruction, marks obtained in secondary education, location of school, living area and type of secondary education were the strongest indicators for the student performance in higher secondary education. This CHAID prediction model of student performance was constructed with seven class predictor variables with accuracy 44.69%.

Thai-Nghe, Drumond, Krohn-Grimberghe, Schmidt-Thieme [11] have used recommender system technique in educational data mining to predict student performance.

In India, after higher secondary education students have to take crucial decision which branch to choose so that there will be good chances of placement.

Elayidom, Idikkula, J. Alexander, A. Ojha [12] created the decision tree which helps admission seekers to choose a branch with high industrial placement. The data was supplied by National Technical Manpower Information System (NTMIS) via Nodal center. Data was compiled by them from feedback by graduates, post graduates, diploma holders in engineering from various engineering colleges and polytechnics located within the state during the year 2000-2003. The standard database is processed to get a table, in which corresponding to each input combination, the percentage placement is computed.

Nghe, Janecek, and Haddawy [13] compared the accuracy of decision tree and Bayesian network algorithms for predicting the academic performance of undergraduate and postgraduate students at two very different academic institutes: Can Tho University (CTU), a large national university in Viet Nam, and the Asian Institute of Technology (AIT), international university in Thailand. It was found that decision trees are 3-12% more accurate than Bayesian Networks.

V. P. Bresfelean, M. Bresfelean and N. Ghisoiu [14] found that students success depends on students choice in continuing their education with post university studies or other specialization attribute, students admittance grade and the fulfillment of their prior expectation regarding their present specialization.

A. Merceron and K. Yacef [15] presented how pedagogically relevant knowledge can be discovered from web-based educational system. The authors built the decision trees from the student data of Logic- ITA web based tutoring tool used at Sydney university to generate if then rules which predict student marks he is likely to achieve.

Baradwaj and Pal [16] obtained the university students data like attendance, class test, seminar and assignment marks from the students' previous database, to predict the performance at the end of the semester.

4. DATA SELECTION AND PREPROCESSING

Data of 1000 females of the Punjab region was collected who appeared for the XII exam in the year 20011-12, 2012-13. The data was collected through the survey form filled by the student at the time of visit to their respective institutes. The survey form was specially designed for the purpose of data collection and study. The survey was conducted on 1000 female candidates who were willing to take admission in higher technical education. They have to enter their family details, demographic data (category, gender etc), previous academic performance (good, average), Location of residence, Family educational background, and sibling's education and family income, address and contact details. From these the attributes that possibly influence their enrolment in higher technical education are selected as shown in Table 1.

Table 1: Factors and their Categories

Sr. No.	Factors	Possible Category
1	Cultural Factor	Hindu
		Muslim
		Sikh
		Christians
2	Category	OPEN
		OBC
		SC,ST
		OTHERS
3	Family Type	Joint
		Nuclear
4	Age of Father at time of Child Birth	<30
		Between 30-35
		Above 35
5	Father's Education	Not Graduate
		Graduate
		Post Graduate
6	Father's Income	>50000 pm
		<50000 pm
7	Mother's Education	Not Graduate
		Graduate
		Post Graduate
8	Mother's Occupation	House Wife
		Self Employed
		Government Job
		Private Job
9	Number of Sibling's	0
		1
		=2 or >2
10	Sibling's Education	Not Graduate
		Graduate
		Post Graduate
11	Sibling's Occupation	Unemployed
		Employed
12	Location of Residence	Rural
		Urban
13	Driving(Two Wheeler/Car)	Yes
		No

5. RESULTS OBTAINED

The decision tree generated from the data is shown in Figure 1. The accuracy of the model is 69.9 %. That is out of 1000 instances 699 instances are correctly classified. The confusion matrix shows that out of 592 Not Enrolled female candidates 491 are correctly classified as NO (Not Enrolled) but 101 are classified as YES (Enrolled). And Out of 408 Enrolled students 208 are correctly classified as YES (Enrolled) but 200 are classified as NO (Not Enrolled).

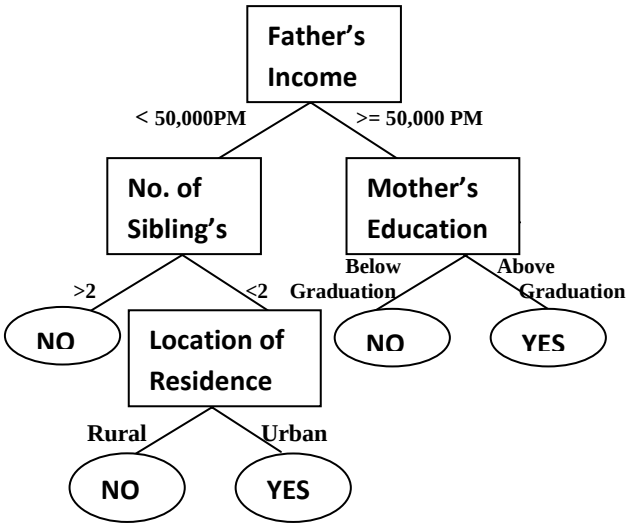

Figure 1: Decision Tree for Two Class Prediction

The Rules generated from this tree are

- If **Father's Income<50,000 PM** and **No. of Sibling's>2** then **Enrolment in Higher Education=NO**
- If **Father's Income<50,000 PM** and **No. of Sibling's<2** and **Location of Residence=Rural** then **Enrolment in Higher Education=NO**
- If **Father's Income<50,000 PM** and **No. of Sibling's<2** and **Location of Residence=Urban** then **Enrolment in Higher Education=YES**
- If **Father's Income>=50,000PM** and **Mother's Education= Below Graduation** then **Enrolment in Higher Education=NO**
- If **Father's Income>=50,000PM** and **Mother's Education= Graduate or above** then **Enrolment in Higher Education=YES**

It is clear from confusion matrix [6] in table 2 for two class prediction i.e. Enrolled or Not Enrolled, that out of 592 not enrolled students 491 are classified as not enrolled. So the true positive rate is 0.83.

		Predicted	
		Not Enrolled	Enrolled
Actual	Not Enrolled	491	101
	Enrolled	200	208

Table 2: Confusion matrix for two class prediction (Enrolled & Not Enrolled)

The most important attribute in predicting enrolments is found to be Father's Income. The other social attributes like category, Father's occupation, Nature of Family, Category and other attributes like medium of Education and driving skills are not appearing in the decision tree indicating less relevance of the prediction with such attributes.

From Figure 1 and Figure 2 it can be observed that the attributes Father's Income, Mother's Education, Number of sibling's play a major role in predicting the enrolment of female in higher education.

6. CONCLUSION

This study shows that Father's Income can be used to create the model using decision tree algorithm that can be used for prediction of enrolment of Female candidates in Higher Technical Education. From the confusion matrix it is clear that the sensitivity of the model is 0.83% and specificity of 0.50%, that means model is successfully identifying the students who are likely to drop out their idea of enrollment in Higher Education due to above reasons.

7. REFERENCES

[1]. C. Romero, S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert system with applications 33(2007), 135-146.

[2]. C. Romero, S. 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, November 2010.

[3]. J. Han, M. Kamber, Data Mining Concepts and Techniques, Second edition, Morgan Kaufmann, SanFrancisco, ISBN: 978- 81-312.

[4]. J. R. Quinlan,"Induction of decision trees", Machine Learning, Volume 1, Morgan Kaufmann, 1986, 81-106.

[5]. R. Kohavi, R. Quinlan, "Decision Tree Disovery", In Handbook of Data Mining and Knowledge Discovery, University Press,1999.

[6]. K. P. Soman, S. Diwakar, V. Ajay, Insight into Data Mining- Theory and Practice, Prentice Hall of India, New Delhi, ISBN: 81-203- 2897-3.

[7]. V. P. Bresfelean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment", Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28, 2007.

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

[9]. Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data", Proceedings of Informing Science & IT Education Conference (InSITE) 2010.

[10]. M. Ramaswami and R. Bhaskaran, "A CHAID based performance prediction model in educational data mining", IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010.

[11]. N. Thai-Nghe, L. Drumond, A. Krohn- Grimberghe, L. Schmidt- Thieme, "Recommender System for Predicting Student Performance", Elsvier B.V., 2010.

[12]. S. Elayidom, Dr. S. M. Idikkula, J. Alexander, A. Ojha, "Applying Data mining techniques for Placement chance prediction", International Conference on Advances in Computing, Control, and Telecommunication Technologies, 2009.

[13]. N. Thai Nghe, P. Janecek, and P. Haddawy, "A Comparative Analysis of Techniques for Predicting Academic Performance", 37th ASEE/IEEE Frontiers in Education Conference, October 2007.

[14]. P. Bresfelean, M. Bresfelean, N. Ghisoiu, "Determining Students' Academic Failure Profile Founded on Data Mining Methods", Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces, June 23-26, 2008.

[15]. A. Merceron and K. Yacef "Educational data mining: A case study", In Proceedings AIED, 2005, pp.467-474.

[16]. B. K. Baradwaj, S. Pal, "Mining Educational Data to Analyze Students' Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.