Identifying the Socio-Economic Factors affecting Students' Enrolment to Post Secondary Education Using Data Mining Techniques

R. Roseline Mary

Assistant Professor, Department of Computer Science

Christ University,

Bangalore, India

roseline.mary@christuniversity.in

***Abstract*---Education sector in India is a growing field that plays an essential role in improving the living status. The economic status or the rise of a country depends on the improved education system. According to statistics, post independent India gave more importance to primary education and expanded literacy rate to two thirds of its population. There are several efforts made by the government to improve the literacy rate in India. The education sector is growing gradually and still 25% of its population are illiterate and the number of students enrolled to higher education is still in decline. Data mining deals with the process in which we identify and extract all the hidden information from data bases affecting the different areas of life either directly or indirectly. Educational data mining plays a very important role in identifying, analyzing and visualizing the data to predict students' performance, their academic achievements, providing feedback for supporting instructors and so on. The main objective of this paper is to identify various socio-economic factors affecting students' enrolment for post-secondary education.**

***Keywords*---Data Mining, KDD, WEKA**

I. INTRODUCTION

Data mining is the growing field of research to identify the non-trivial patterns from huge databases by applying statistical and artificial intelligence techniques[1][2]. It is used to uncover the hidden information in a database [3]. Data mining techniques are used in various fields such as crime, medicine, engineering, education, marketing, and banking. It also provides many tasks that could help to study the students' performance [4]. Data mining is a creative and promising field of research in which it can be implemented and put into practice in education. "Educational Data Mining is a thriving discipline, concerned with developing methods for exploring the unique type of data that come from educational settings, and using those methods to better understand students, and the settings in which they learn."[5]. Day by day the growth of the data is very rapid and that data need to be transformed into useful information [6]. Educational data mining (EDM) tends to focus on new tools and techniques for discovering patterns in the data. It also gains popularity in the new research areas in higher education. Recent research findings in educational data mining help the students, institutions and government for improving the quality of education. Despite the rapid growth in the education sector, 25% of its population is still illiterate, 15% of the students reach high school, and only 7% graduate [7]. A statistical report says according to the year 2011, out of 74% of the literacy rate, only 47% have attained the diploma and post diploma courses [8]. Post-secondary education plays a vital role in country's development. But the statistical data prove that major population in India are still school dropouts. There are so many factors which affect the students' enrolment to post-secondary education such as family background, school infrastructure, facilities and their psychological behaviours and so on. The main aim of this paper is to identify the reasons for poor enrolment to post-secondary education and the result will help the students, management and policy makers to give a better solution. Data mining techniques particularly classification helps to analyze the input data and to develop a model describing important data classes or to predict future data trends.

II. LITERATURE SURVEY

In [11], the authors make use of classification as a process to improve the higher educational system and

its quality through the student data evaluation. They experiment this approach so that they can identify the major attributes that affect the performance of the students in their chosen courses. Ayesha et.al [12] handled a technique like clustering in data mining to investigate "student learning behaviour" that came as a motivation for teachers to identify why many students dropped out to a significant level and enhance the students' performance in higher education. Kan [13] has utilised classification, association, rules and clustering as data mining tools to design a course management system to improve the teaching learning process. In [15], the authors try to classify students, keeping their final grade that they obtained in their respective courses in mind. In this approach they implement four different classification methods for analysis. Paul [16], in his research used classification to evaluate previous year's student dropout data using Bayesian classification method.

III. STATEMENT OF THE PROBLEM

This paper aims to identify the socio-economic factors affecting the students' enrolment to post-secondary education using data mining techniques. The parameters comprises (1)personal information such as gender, occupation of the parents, age, family income, highest educational qualification of the parents, stay, and family size.(2)institution related information such as type of learning, usage of teaching aids, exposure to ICT, faculty qualification, etc. (3)psychological information such as social status, illness, disability, etc. are considered. These attributes were used to predict the students' enrolment to post-secondary education.

IV. METHODOLOGY

To build the classification, CRISP methodology is adopted. The proposed methodology is to build the classification model that tests the factors which affect the students' enrolment to post-secondary education.

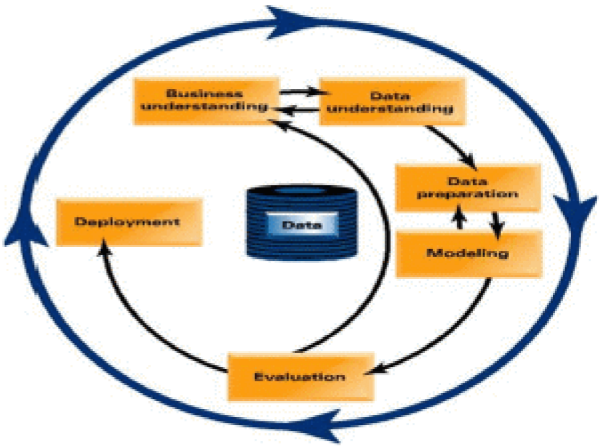


Fig. 1. CRISP Methodology

A. Data Mining Process

Data mining process is a step by step procedure to identify the problem, to collect the data sources, and to identify the relevant data sources from the data set. The next step is to prepare the actual data and build the model using the data mining techniques. The built model needs to be evaluated and the result obtained from evaluation has to be implemented for prediction.

1) Business Understanding

Any problem in terms of business understanding can be converted into a data mining problem. The objective of the problem can be defined, designed and accomplished further.

2) Data Understanding

Data set is used to identify the problem and discover useful information out of it. Data understanding also helps to analyze the quality of data such as "Is the data complete? or any missing values?"

3) Data Preparation

Data Preparation takes usually 90% of the time to collect, to assess, to clean and to select the data required to construct, integrate and format the data. To provide a solution to the identified business problem the data sources are very important and from all the data sources collected, the selected data sources should be identified and the actual data must be determined [17].

4) Building the Classification Model

To build the model relevant to the objective of our data mining problem, the classification technique is used. The author states that "Classification techniques are supervised learning techniques that classify data item into predefined class label" [18]. This technique in data mining is very useful from a data set to build the classification model that is used to predict future data trends. There are different classification algorithms to build the model like ID3,CART,J48 and

so on and with this we will able to identify a class for the actual data which we have determined in data preparation. The classification by using decision tree approach will produce a set of rules in a human understandable way for decision makers to interpret in their domain knowledge for validation and justification [18].

5) Evaluation

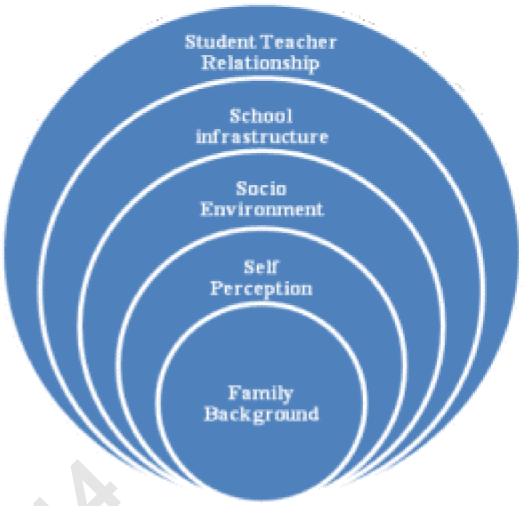
Evaluation is to check whether we correctly built the model and determine how to proceed with it to the next phase of implementing it. Evaluating the results assesses the degree to which the built model meets the objective of our problem, the additional challenges, and information for future directions. Selecting the proper data mining method is an important, critical and difficult task in Knowledge Discovery and Data Mining process. To implement this model, Waikato Environment for Knowledge Analysis Toolkit which has a collection of machine learning algorithms for solving data mining problems implemented in Java is used.

6) Deployment

The evaluated results need to be utilised in the deployment phase. This phase is the final phase to predict the future trends or patterns that have been gained out of this knowledge which will be applicable for the end user. The CRISP provides a uniform framework for experimenting, analyzing, evaluating and predicting the result.

V. ANALYSIS

To implement this model WEKA Toolkit which has a collection of machine learning algorithms for solving data mining problems implemented in Java is used. To analyze on the socio-economic factors that affect students' enrolment for post-secondary education using data mining techniques, a questionnaire which is a document listing a series of questions pertaining to the problem is used. The questionnaire is prepared based on five important indicators like Family Background, Self Perception, Socio-Environment, School Infrastructure and Student Teacher Relationship. Nine questions based on the family background, eight questions on self-perception, seven questions on socio-environment and eight questions on student teacher relationship are addressed in this questionnaire. There are four kinds of questions. Each is intricately related to each other. They are Factual Questions, Opinion-Attitude Questions, Information Questions, and Self Perception Questions. Factual questions are designed to elicit factual information from the respondents. The most common type of factual question is the background question. Background information is used to classify respondents into different categories. The family background information is concerned with their parent, sibling's details and also the family interest towards higher education. For the self-perception category of questions, respondent is asked to evaluate something about his/her own behaviour in relation to others. Such questions allow individual subjects to compare their ideas with those of other persons.



This research is intended to use decision tree which is the part of the induction class of data mining techniques. The decision tree is efficient for this dataset. The model makes use of the software Weka, which has the collection of machine learning algorithms for solving data mining problems implemented in Java. The data collected through questionnaire will help to analyse the factors and the tree analysis logistic regression can be used to estimate the probability of the non-enrolment feature based on the factors considered.

VI. CONCLUSION

This research is a preliminary attempt to help supporting the decision makers of the institution to improve their teaching methodology, teaching aids and all other infrastructure facilities that they lack. The result evaluated out of this project will motivate the parents of BPL (Below poverty line) towards the values of post-secondary education. This project will help the policy makers of our Indian government to help the children studying in government schools in a much better way towards their post-secondary education. The model proposed as an academician can be useful to build a software model to provide a solution by formulating the result.

REFERENCES

[1] Jawed Han, Michelin Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, 2006.

[2] Arun K. Pujari, "Data Mining Techniques", Universities Press (India) Private Limited, 2005.

[3] DunHam H.Margaret, Data Mining, Pearson Education, 2003.

[4] Samrat Singh, Vikesh Kumar, "Classification of Student's data Using Data Mining Techniques for Training & Placement Department in Technical Education," International Journal of Computer Science and Network (IJCSN) vol 1, Issue 4, August 2012.

[5] R. Baker, K. Yacef (2010). "The State of Educational Data Mining in 2009: A Review and Future Visions," Journal of Educational Data Mining, Volume 1, Issue 1 1: pp. 3– 17.

[6] Parack, S.; Zahid, Z.; Merchant, F., "Application of data mining in educational databases for predicting academic trends and patterns," Technology Enhanced Education (ICTEE), 2012 IEEE International Conference on, vol., no., pp.1-4, 3-5 Jan. 2012.

[7] "India still Asia's reluctant tiger", http://news.bbc.co.uk/2/hi/business/7267315.stm Zareer Masan of BBC Radio 4, 27 February 2008.

[8] Estimate for India, from India, the Hindu

[9] Ramasubramanian, Iyakutti and P. Thangavelu, "Enhanced data mining analysis in higher educational system using rough set theory," African Journal of Mathematics and Computer Science Research ,vol. 2, pp. 184-188, October, 2009

[10] Data Mining in Higher Education Roberto Lorene and Maria Morant Universidad Politécnica de Valencia, Spain.

[11] Qasem A. Al-Radaideh, Emad M. Al-Shawakfa, and Mustafa I. Al-Najjar," Mining Student Data Using Decision Trees ", The 2006 International Arab Conference on Information Technology ACIT'2006)".

[12] Ayesha, Sheela, T. Mustafa, A.R. Sattar, M.Inayat Khan, "Data Mining Model for Higher Education System," European Journal of Scientific Research, vol.43, pp.24-29, 2010.

[13] Liu Kan, Xiao Xingyuan, Liu Ping," DMCMS: A Data Mining Based Course Management System," 2010 Second International Workshop on Education Technology and Computer Science".

[14] Lin Zhang, Yan Chen, Yan Liang, Nan Li," Application of Data Mining Classification Algorithms in Customer Membership Card Classification Model".

[15] Varsha Namdeo Anju Singh Divakar Singh Dr. R.C Jain," Result Analysis Using Classification Techniques",2010 International Journal of Computer Applications (0975 - 8887),vol1 – No. 22

[16] Paul Saurabh," Mining Educational Data Using Classification to Decrease Dropout Rate of Students", International Journal of Multidisciplinary Sciences and Engineering, vol.3, No.5, May 2012.

[17] http://it.toolbox.com/blogs/enterprise-solutions/preparing-data-for-data-mining-17755.

[18] Qasem A. Al-Radaideh, Eman Al Nagi," Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2, 2012.