# A Study On School Dropouts In Theni District: A Data Mining Analysis

S. Anitha
*II Pg Student,*
*Dept. of Computer Science,*
*Jayaraj Annapackiam College*
*for Women (Autonomous),*
*Periyakulam – 625 601,Theni*
*Dt, Tamil Nadu, India.*

B. Jasmine
*Asst. Professor,*
*Dept. of Computer Science,*
*Jayaraj Annapackiam College*
*for Women (Autonomous),*
*Periyakulam – 625 601,Theni*
*Dt, Tamil Nadu, India.*

S. Deepalakshmi
*II Pg Student,*
*Dept. of Computer Science,*
*Jayaraj Annapackiam College*
*for Women (Autonomous),*
*Periyakulam – 625 601,Theni*
*Dt, Tamil Nadu, India.*

## Abstract

*Education means expansion of cultural horizons and employment opportunities to an individual. There is a significant stride in India, in enhancing initial access to schooling and even enrolment of all children in primary schools. It has seen that unprecedented expansion of schooling infrastructure across the country, even ignoring the traditional framework of population size and distance norms. The 2009 Right of Children to free and Compulsory Education Act ensures the norms, standards and conditions essential to accessible, quality elementary education, at least 48 of every 100 students in India pursuing secondary education never go beyond that level. There is a growing concern from the World Bank pointing out that the country was doing worse than Vietnam and Bangladesh in enrolling students in secondary education. This data mining study evaluates the number of factors that may be responsible for raising the dropout ratio of girl children in Theni district, one of the South western districts of Tamilnadu State of India. Even though poverty is exposed as major reason for dropping out the girl children out of schools, the data mining analysis made in Theni district indicates, parents' attitude, quality of teachers, social environment also have played vital role in education female child. The data of the study have been collected through a survey of individual people' opinion, and treated by a data mining tool called XL miner. The result shows that parent's attitude towards girl child's education is positively related to the raising of school drop outs.*

**Keywords:** *Education; School Dropouts; Girls; Theni District; Data Mining; XL Miner*

## 1. Introduction

Education is the foundation of all development and a vital catalyst for growth. It is proven to be the key to ensuring sustained and equitable economic growth, improved health and social development in any country. Across the world, 171 million people could be lifted out of poverty if all children left school with basic reading skills. [1] UN Millennium Summit in 2000 asserted by stating that 'Education is development. It creates choices and opportunities for people, reduces the twin burdens of poverty and diseases, and gives a stronger voice in society'. [2] India as one of developing country, takes very much keen to provide basic education to all the girl children up to the age of 12. Unequal social, economic and power equations deeply influence children's access to education and their participation in the learning process. This is evident in the disparities in education access and attainment between different social and economic groups in rural India. The main objective of this case study is to find out the reason for dropping out of girl children from their school education in rural villages of Theni district. We have collected the inputs from several women in the villages of the district. The classification process is done by XL Miner tool for data mining. The findings paint a stark picture of the education status of the girl children in Theni district.

## 2. Problem Definition and Description

Theni is bounded on the north by Dindigul district, on the east by Madurai district, on the south by portions of Virudhunagar district and Idukki district of Kerala State and on the west by Idukki (Kerala). The total geographical area of the district is 3076.30 Sq. Km. According to the 2011 census, it has a population

of 1,243,684. The district has a population density of 433 inhabitants per square kilometre (1,120 /sq mi). Its population growth rate over the decade 2001-2011 was 13.69 %. [3].

Table 1. Total Population, Population in the Age-Group 0-6, Literacy Rate by Sensex-2011

|  | Population | | Population in age group 0-6 | | Literacy Rate (%) | |
|---|---|---|---|---|---|---|
|  | Tamil Nadu | Theni Dt. | Tamil Nadu | Theni Dt. | Tamil Nadu | Theni Dt. |
| Total | 72138958 | 1243684 | 6894821 | 110919 | 80.33 | 77.62 |
| Male | 36158871 | 624922 | 3542351 | 57258 | 86.81 | 85.48 |
| Fema | 35980087 | 618762 | 3352470 | 53661 | 73.86 | 69.72 |

## 3. Data Mining Process

Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various data mining techniques such as machine learning, artificial intelligence (AI) and statistical. Therefore the needs for a standard data mining process increased dramatically.

Any data that was collected by any means must be pre-processed to bring it to a form suitable for pattern recognition. Starting with the raw data in the form of images or meshes, we successively process these data into more refined forms, enabling further processing of the data and the extraction of relevant information. Detailed transaction information in the OLTP (On Line Transaction Processing) and legacy system is usually not suitable for data mining. For data mining to be effective, much careful work is needed in defining the aims of network of data mining then in selection, cleaning, transformation and separate storage of data that is suitable for data mining. A typical data mining process includes requirement analysis, data selection and collection, cleaning mining exploration and validation, implementing, evaluating and validation, monitoring and result visualization. [4] We implement classification algorithm to evaluate the predicted data of our case study. We generated the reports for result visualization and exploration.

### 3.1 Requirement Analysis

There are some goals that the data mining process is expected to achieve. The samples of the case study must be clearly defined. In requirement analysis, the problem is clearly defined as the objective of the case study. Objective of this case study is, to find out

the reason for leaving schools without completing their studies in Theni district.

### 3.2 Data Selection and Collection

This step includes finding the best source of databases for the analysis. If the data has implemented a data warehouse, then most of the data could be available there. If the data is not available in the warehouse, the source OLTP (On Line Transaction Processing) systems need to be identified and the required information extracted and stored in some temporary system. In some cases, only a sample of the data available may be required. In our case study, data were collected using questionnaire from the study area. Questionnaires were used to survey the reason for drop out in schools.

### 3.3 Cleaning and Preparing Data

This may not be an onerous task if a data warehouse contains the required data, since most of this must have already been done when data was loaded into the warehouse. Otherwise this task can be very resource intensive and sometimes more than 50% of effort in a data mining project is spent on this step. Essentially, a data store that integrates data from a number of databases may need to be created. When integrating data, one often encounters problems like identifying data, dealing with missing data, data conflicts and ambiguity. An ETL (Extraction, Transformation and loading) tool may be used to overcome these problems.

### 3.4 Data Mining Exploration and Validation

Once appropriate data has been collected and cleaned, it is possible to start data mining exploration. Assuming that, the user has access to one or more data mining tools, a data mining model may be constructed based on the sample details. It may be possible to take a sample of data and apply a number of relevant techniques. For each technique, the results should be evaluated and their significance is interpreted. This is likely to be an iterative process which should lead the selection of one or more techniques that are suitable for further exploration, testing and validation.

### 3.5 Implementing, Evaluating and Monitoring

Once a model selected and validated, the model can be implemented by the decision makers. This may involves software development for generating

reports or for results visualization and exploration which is used by the managers. It may be more than one technique that is available for the given data mining task. It is important to evaluate the results and choose the best technique. Furthermore, there is a need for regular monitoring the performance of the techniques that have been implemented.

In our case study, we have implemented Classification Algorithm to evaluate the predicted data. We generated the reports for result visualization and exploration. It may be more than one technique is available for given data mining task.

## 3.6 Result Visualization

Here, we classify the school drop outs according to the category. Although the problem is dealing with number of dimensions must visualized using two-dimensional computer screen or printout. And also see that the decision tree making of that problem. We derived the decision tree out of our processing and we have given it in Figure 1.
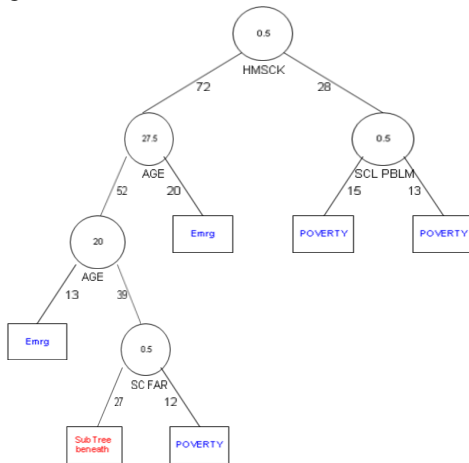


Figure 1. Classification Full Tree

## 4. Algorithm Implementation

In this case study, we have classified the school drop outs to the category we derived using data mining techniques. This study has adopted the survey method and descriptive research design. The population of the study consists of the entire set of school drop outs of Theni district. A total of 500 respondents of the study area have been covered as the sample for our case study. The survey was carried out for three days a week in continuation followed by similar exercise for four weeks alternatively. The information gathered by

means of questionnaire was also verified by personal interviews. The result is then compiled for both quantitative as well as qualitative data.

**Step 1:** Input variables are identified. Here we declare var1 as name, var2 as age, var 3 as place, var4 as size of the family, var5 as distance of the school, var 6 as high fees, var7 as no parent, var 8 as homesick, var 9 as poverty, var 10 as early marriage, var 11 as working in mill, var 12 as parent not allowed. Then the output we have taken is the tuple yes as 1, No as 0.



Figure 2. Input data and variables

**Step 2:** Here, we find out the probability value for each attribute by using information gained. The information gain is calculated from the Following Formula,

$$\text{Info (D)} = -\{_{ni=1}Pi \ \log2(Pi)$$

Where Pi is the Probability that an arbitrary tuple belongs to the class. Here all the classes are in the probability value of 0.2.

### Prior class probabilities

| Class | Prob. |
|---|---|
| Emrg | 0.333333333 |
| PNtAlwd | 0.333333333 |
| POVERTY | 0.333333333 |

Figure 3. Prior class probabilities

**Step 3:** Training log is used to find out the miss-classify of the given class. It also used for growing the full tree using data.

**Training Log (Growing the full tree using training data)**

| # Decision Nodes | % Error |
|---|---|
| 0 | 52 |
| 1 | 52 |
| 2 | 47 |
| 3 | 47 |
| 4 | 44 |
| 5 | 44 |
| 6 | 42 |

Figure 4. Training Log

**Step 4:** The full tree rules are used for find out the decision and terminal nodes, where the decision nodes are 5 and terminal nodes are 6.

**Full Tree Rules (Using Training Data)**

| #Decision Nodes | 6 |
|---|---|

| #Terminal Nodes | 7 |
|---|---|

| Level | NodeID | ParentID | SplitVar | SplitValue | Cases | LeftChild | Right Child | Class | Node Type |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | N/A | HMSCK | 0.5 | 100 | 1 | 2 | POVERTY | Decision |
| 1 | 1 | 0 | AGE | 27.5 | 72 | 3 | 4 | POVERTY | Decision |
| 1 | 2 | 0 | SCL PBLM | 0.5 | 28 | 5 | 6 | POVERTY | Decision |
| 2 | 3 | 1 | AGE | 20 | 52 | 7 | 8 | POVERTY | Decision |
| 2 | 4 | 1 | N/A | N/A | 20 | N/A | N/A | Emrg | Terminal |
| 2 | 5 | 2 | N/A | N/A | 15 | N/A | N/A | POVERTY | Terminal |
| 2 | 6 | 2 | N/A | N/A | 13 | N/A | N/A | POVERTY | Terminal |
| 3 | 7 | 3 | N/A | N/A | 13 | N/A | N/A | Emrg | Terminal |
| 3 | 8 | 3 | SC FAR | 0.5 | 39 | 9 | 10 | POVERTY | Decision |
| 4 | 9 | 8 | ILLNESS | 0.5 | 27 | 11 | 12 | POVERTY | Decision |
| 4 | 10 | 8 | N/A | N/A | 12 | N/A | N/A | POVERTY | Terminal |
| 5 | 11 | 9 | N/A | N/A | 17 | N/A | N/A | PNtAlwd | Terminal |
| 5 | 12 | 9 | N/A | N/A | 10 | N/A | N/A | POVERTY | Terminal |

Figure 5 (a). Tree rules using training data

**Training Data scoring - Summary Report (Using Full Tree)**

| Classification Confusion Matrix | | | |
|---|---|---|---|
| | **Predicted Class** | | |
| **Actual Class** | Emrg | PNtAlwd | POVERTY |
| Emrg | 18 | 3 | 12 |
| PNtAlwd | 5 | 8 | 6 |
| POVERTY | 10 | 6 | 32 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Emrg | 33 | 15 | 45.45 |
| PNtAlwd | 19 | 11 | 57.89 |
| POVERTY | 48 | 16 | 33.33 |
| Overall | 100 | 42 | 42.00 |

Figure 5 (b). Error report

**Training Data scoring - Summary Report (Using Full Tree)**

| Cut off Prob.Val. for Success (Updatable) | 0.5 | ( Updating the value here will NOT update value in detailed report ) |
|---|---|---|

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | 1 | 0 |
| 1 | 66 | 3 |
| 0 | 1 | 30 |

Figure 5 (c). Training data scoring – summary report

**Step 5:** The tree was constructed with corresponding attributes. The tree has taken the root node as the attribute Economic status. The root node is selected by the probability value. Finally in this step, to classify the actual class to predicted classes in given data such as Poverty, Early marriage, Parent not allow will be get from this step. The overall elapsed time to run the XL Miner for this case study is 3 sec.
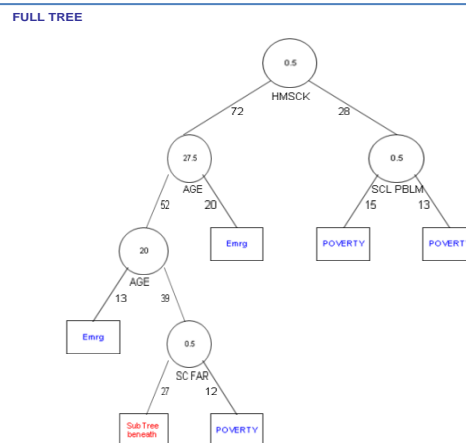


Figure 6. Classification full tree using training data

## 5. Data Mining Findings

The initial studies unveiled a number of relationships between variables as well as threshold values that justify further analysis. Improper attitude toward education, family background, employment / occupational structures, financial problems, in quality of life, unawareness about the importance of the education, improper understanding about the present situation of their environment [5], lack of realization and unhealthy emotions of the people and their generations, lack of proper management, improper building facilities, unhygienic condition, improper infrastructure, improper classroom facilities, improper incentives to the students, insecurity of girls children, lack of parental cooperation and no more positive co-relation between education and employment are some of the reason we found out of our case study.

**a. Family Size:** We found that if the family size is big, the economic condition is poor, parents sacrifice the education of their female child every easily. From their childhood, girls are expected to help their mothers look after animals, work in the fields, and help in the homes. This puts an additional burden on girls that many boys do not face when trying to find the time and energy to study for school. Many parents have less awareness about female education and give less importance for literacy, which is a necessary skill for girls to succeed in their life. The lack of parental or family support to provide academic help to the female child is a very big problem we face during the study time.

**b. Less Interest from Parents:** It has been found that parents neither have ever participated in school PTA meetings nor ever have visited the schools to keep monitoring of their children academic performance. Most of the selected cases of dropouts, parents were either least concern or gave less value of their children's education. [6]

**c. Single Parents:** It has been found that students from single-parent families are more likely to dropping out of school than students from two-parent families. In the absence of the mother in the family (who was dead), the elder female child does cooking and other household chores, so she does not get time to come school regularly. The child finds it increasingly difficult to remain awake long enough to study, after rising early and working late in their home. She has a poor academic record in school, mostly irregular, and lack of getting family's emotional and academic support have led her to take disinterest in studies and finally pushing her to leave the school as soon as possible before jumping to the next grade. Family size also plays a vital role in increasing the school dropout ratio in Theni district.

**d. High fees Structure:** Although none are rich, families make significant sacrifices to keep their girls in school. For virtually everyone, paying for education is a constant difficulty, and parents frequently share fields and finances with their siblings to help support the extended family's collective children. Even though the government of Tamil Nadu is supplying free books, free uniform, free mid-day meals for school students, most parents feel fortunate when they can afford the basic needs of their children. It is noteworthy to record that the free bicycle scheme for school student of high schools has allowed scores of young girls to attend schools who otherwise would have been stymied by the distance. [7]

**e. Less cooperation from the school teachers:** It also made the parents not to send their daughter to schools or drop her from her schoolings. In remote villages of Theni district, we could not find schools with high standards other than government schools. Even though government started schools in such remote villages, the teacher will not come regularly or not teaching properly. This also made the children to stay at home.

**f. Negative school experience:** The empirical study on dropouts has consistently shown that a host of negative school related experiences serve as powerful precursors to the decision to formally leave the school. Studies have shown that students who drop out of school are more likely than other students to have poor performance, disruptive behaviour, and poor attendance, negative attitudes towards school, and early school failure, particularly repeating grades. In many of the remote villages of Theni district, studying in a co-education schools are not acceptable by the parents. [8]

**g. Less financial support from Government:** The inadequate use of funds by school systems has led to a higher dropout rate among the students who need the most help, those students who are of low socio-economic status. The dropout rate is higher in school districts with a large number of students classified as having a low socio-economic status. Research attributes this high dropout rate among students of low socio-economic status to a lack of education among parents and the need to leave school early for familial support Therefore, socio-economic status was included to determine whether low socio-economic status significantly impacted dropout rates. [9],[10]

**h. Quality of education:** There is an urgent need for effective efforts for improvement in quality of education in rural villages. There is a need for the development of ICT policy with adequate budgetary provisions as well development of private education policy. Private education has become business in real terms with zero service elements in many parts of Theni district. [11]

## 6. CONCLUSION

The goal of data mining classification is to build a set of models that can correctly predict the class of the different objects. The input to this method is set of objects (training data); the classes in which these objects belong to (dependent variables), and a set of variables describing different characteristics of the objects (independent variables). We found out that not only financial problems, family and societal problems are also directly connected to the school drop outs of female child in Theni district of Tamil nadu. Appointment of quality teachers and improving the facilities in government schools particularly in rural areas of Theni district should be given top priority in development. So Capacity building of rural school teacher is one of the priority areas for consideration. Even though many obstacles found in providing education to younger India, if the parents are positive towards providing basic education to their child, the enrolments in schools will have considerably increased. The government's continuing focus will also be required to make the platform effective.

## 7. References

[1]. UNESCO, "*Education for All Global Monitoring Report*", UNESCO, 2010. www.efareport.unesco.org.

[2]. Shaharazad Abdul-Ealeh, Sam Barratt, Gohn Coventry, "*Back to School*", Global Campaign for Education in 2010, South Africa, 2008.

[3]. 'District Census 2011'. Census2011.co.in. 2011, www.census2011.co.in/district.php

[4]. Gupta G. K.,2011, "Introduction to Data Mining with Case Studies", 2ndEdision, PHI Learning publication, Page 7

[5]. Mujibul Hasan Siddiqui,"The Problems of School Drop Outs among Minorities with Special Reference to Muslims in India", International Journal of Management and Social Sciences Research (IJMSSR) ISSN: 2319-4421, Volume 2, No. 1, January 2013

[6]. Pankaj Das, "Process of Girls Dropout in School Education: Analysis of Selected Cases in India", In Engendering Empowerment: Education and Equality e-Conference. 12 April – 14 May. United Nations Girls' Education Initiative: New York. Available at: http://www.e4conference.org/e4e

[7]. "National Programme of Mid-Day Meal in Schools - Annual Work Plan & Budget 2012-13", Letter No19869/ NMP.1/2012-4Dated : 18.04.2012

[8]. Camilla A. Lehr, David R. Johnson, Christine D. Bremer, Anna Cosio, Megan Thompson, "Essential Tools Increasing Rates of School Completion: Moving From Policy and Research to Practice - A Manual for Policy makers, Administrators, and Educators", ICI Publications Office, May 2004

[9]. Jacqueline Stevenson, Mel Lang, "Social class and higher education: a synthesis of research", EvidenceNet is a Higher Education Academy resource, April 2010, www.heacademy.ac.uk/evidencenet.

[10]. David Bishop Collins, "Variables that impact high-school dropout rates in a large metropolitan area", December, 2009.

[11].Lydie Ehouman, Sandra Fried, Theresa Mann, Haroon Ullah, "Tamil Nadu: The Path to Becoming India's Leading State A Strategic Analysis of Health, Education Biotechnology and Tourism", A study conducted for the Government of Tamil Nadu Center for International Development, May 2002.

**S. Anitha** is a post graduate student in Jayaraj Annapackiam College for Women (Autonomous) a first grade college in Mother Theresa University, Kodaikanal. She has keen knowledge in programming and analysis of case studies. She has interest in data mining Techniques. She is available in Anitaa.skj@gmail.com

**B. Jasmine** is an Assistant professor at the Computer Science department of the Jayaraj Annapackiam College for Women (Autonomous). Have more than 3 years of teaching experience in the subject computer science, her areas of specialisation spread over Data Mining and Computer Networks. She also has a B.Sc and MCA degrees from the Madurai Kamaraj University, Madurai. Her contact can be made in

**S. Deepalakshmi** also involved in this case study. She is also a post graduate student of Department of Computer Science, Jayaraj Annapackiam College for Women (Autonomous), Periyakulam, affiliated to Mother Teresa University, Kodaikanal. She has rich skill set in programming and using them for analysing data. Her mail id is