# Mining Educational Data to Predicting Higher Secondary Students Performance

**A. Dinesh Kumar**
*Sri Krishna Arts and Science College*
*Coimbatore, India.*
*mail2thinesh@yahoo.com*

**V. Radhika**
*Sri Krishna Arts and Science College*
*Coimbatore, India.*

***Abstract*- Education means imparting knowledge to the students and developing their innate quality. Recent terms data mining techniques have been applied in educational field in order to find out the hidden knowledge from educational data. The great deal of research has been done identifying the factors that affect the student's performance those factors can be named as psychological factor and environmental factor. Student performance affected by different factors such as learning environment, economic condition and peer group family, this study focus on environmental factors and educational institute factors. Predicting student performance is very essential for higher secondary teachers to identify their students according to their performance by the name of excellent performers, average performers and below average performers. In this study student environmental and educational factors are compared. The C4.5 and ID3 decision tree algorithm applied on predicting the student performance with feature selection technique ranker.**

Keywords- Educational Data Mining, Prediction, Classification, Ranker.

## I. INTRODUCTION

Applying data mining education is an interdisciplinary research field also known as educational data mining (EDM). EDM analyze data generated by any type of information system supporting learning or education in schools, colleges, universities and other academic or professional learning institutions providing traditional model of teaching, as well as informal learning. Predicting students' results and student modeling have been the fundamental goals of educational data mining. These two issues are deeply connected with educational environment [1].

EDM has emerged as an independent research are in recent years for researchers all over the world from different and related research areas such as offline education try to transmit knowledge and skills based on face-face-contact and also study psychologically how humans learn. Psychometrics and statistical techniques have been applied to data like student performance, behavior, curriculum, that data was gathered from classroom environments. EDM is most used in e-learning and learning management system. E-learning provides online instructions [2].

The organization of thesis work includes related work in section 2. Section 3 describes the data mining process. In section 4 research work of this study is discussed. Section 5 includes result and discussion of thesis work. Section 6 discusses the conclusion of thesis work.

## II. RELATED WORK

Dr. N. Tajuniza et al [3] analyzed the factors affecting academic achievement that contribute the prediction of student academic performance. They have applied classification techniques and mapreduce concept for predicting the student performance and improve their results.

Alaa el-Halees represented data mining can be used in educational field to develop and understanding the learning process of student to focus on mining and assessing variables related to student academic performance [4]. Mining data in educational environment is called educational data mining.

Ryan S.J.D. Baker et al [5] analyzed the history and trends in the field of educational data mining (EDM) 2009. They focused about increased importance on prediction, the development of work using existing models to make scientific discoveries Pandey and Pal [6] investigated the student performance based by selecting 600 students from different colleges of Dr. R. M. L. Awadh University, Faizabad, India. By means of Bayes

Classification on category, language and background qualification, it was found that whether new comer students will performer or not.

Pandey and Pal [7] conducted study on the student performance based by selecting 60 students from a degree college of Dr. R. M. L. Awadh University, Faizabad, India. By means of association rule they find the interestingness of student in selecting class teaching language.

Yadav and Pal [8] obtained data from VBS University student's like Branch, Category, Student grade in high school, and Admission type, medium and family size from the previous student database to predict the students who are likely to fail with the help of ID3, C4.5 and CART algorithm. It was observed that C4.5 is the best algorithm for predict student result.AS Galathiya et al [9] conducted a research on classification with an improved decision tree algorithm using feature selection technique. They have used genetic search algorithm for improve the classification accuracy. By means of the classification accuracy is improved by implementing the diversities of algorithm using RGUI with weka packages.

Bharadwaj and Pal [10] had conducted a research on university student's data like attendance, class test, seminar and assignment marks from the student's prior database, to predict the performance at the end of the semester.

## III. DATA MINING PROCESS

In present day educational system, a student's performance is influenced by psychological and environmental factors. Students should be properly motivated to learn. Motivation leads to interest, interest leads to success. Proper assessment of abilities helps the students to perform better. Students requires proper study atmosphere both at school and home. Poor economic condition also affects the performance of the students as most of them are unable to get proper education. Uneducated family background also affects the students' performance. In this study consider environmental factors and educational institute factors. This helps the tutor to identify the factors that are related with the three types of learners and take appropriate action to improve their performance.

### *A. Data Preparations*

The data set used in this study was obtained from different colleges on the questionnaire method of Computer Science department of course B.Sc (IT), B.Sc, (CS) and B.E of session of 2013 to 2015. Initially size of the data is 300.In this step data stored in different tables was joined in a single table after joining process errors were removed.

### *B. Data Selection and Transformation*

In this step only those field were selected which were required for data mining. A few derived variables were selected. While some of the information for the variables extracted from the data base. All the predictor and response variables which were derived from the database are given in table I.

The parameter values for some of the variables have detailed below to give brief explanation about each attributes for the current investigation as follows:

- **FI** –to predict student level, Family Income (FI) plays vital role among all the students', by the help of given property values (i.e., Low, Medium and High).
- **ME**- If mothers are educated they can contribute to improve the performance of the students. In this study, ME considered to predict student's results with the help of selected property values by the students (i.e., Low, Medium and High).
- **MW**- how mother education is doing vital role to educate their children, likewise their working status has considered with the name of MW attribute. Because, in a situation a particular student mother doesn't work, then their mother can spend more time with them. Those data have been organized by the help of specified property values (i.e., Yes or No).
- **SH**- Study hours, it represents how many hours a student spends on study after attending the class in school. Again it shows how much serious the student takes studies. The possible values are High, Less, Never.

- **RE**- to predict student performance, relation or behaviors of the teacher with the student, which have collected by the name of handling basis (RE: Relation), and given to students to select according to their need. ( i.e., casual, strictly and friendly).
- **LS**- Learning style, students are following different learning styles. It's commonly believed that most of the students follow some particular method of interacting with, taking in and processing information. This collected by the help of specified property values (i.e., AL, VL, and TL)
- **RESULT**- it's our main constant which collects and keeps the entire students final results in separate place to predict student's performance with the help of allocated property values (i.e., Below Average, Average, Excellent).

Table I. Student Related Variables

| Variable | Description | Possible Values |
|---|---|---|
| Sex | Student Sex | {Male, Female} |
| Cat | Students Category | {General, OBC,SC,ST} |
| FI | Family Income | {Low, Medium, High} |
| ME | Mothers Education | {Low, Medium, High} |
| MW | Mothers Working Status | {Yes, No} |
| FE | Fathers Education | {Low, Medium, High} |
| FW | Fathers working Status | {Yes, No} |
| SH | Study Hours per day after school | {Less, High, Never} |
| RE | Student and Teachers Relation | {Casual, Strictly, Friendly} |
| LS | Students Learning Style | {Auditory, Visual, Tactile} |
| MED | Medium of Studying | {Tamil, English} |
| LOC | Location of the school | {Rural, Urban} |
| MO | Student Using Mobile Phone | {Own, Parents, No} |
| SN | Students Having Social Network Id | {Yes, No} |
| Result | Categories of the Student | {Excellent >=75, Average >=50 and <75, Below Average<50 } |

*C. Decision Trees*

Decision tree induction is the learning of decision trees from class- labeled training tuples. A decision tree is a flowchart- like tree structure, where each internal node (non-leaf node) denotes a test on attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node [11].

DT is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree are commonly used for gaining information for the purpose of decision- making. Decision tree starts with a root node on which it is for users to take actions. From this node, users will split each node recursively according to decision tree learning algorithm. The final result is a decision tree in which each branch represents a possible scenario of decision and its outcome [8].The most widely used decision tree learning algorithms are C4.5 and ID3

*D. The ID3 Decision Tree*

ID3 is a simple decision tree algorithm introduced by Ross Quinlan in 1986 [11]. It is based on Hunts algorithm. The basic idea of ID3 algorithm is to construct the decision tree by employing a top- down, greedy search through the given sets to test each attribute at every tree node. The tree is constructed in two phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It accepts only categorical attributes in building a tree model. It does not five accurate result when there is noise. To remove the noise pre- processing technique has to be used.

*E. C4.5*

C4.5 algorithm is developed by Quinlan Ross that generates the decision trees which can be used for classification problems [11]. It is the successor of ID3 algorithm by dealing with both categorical and continuous attributes to build a decision tree. It is also based on Hunt's algorithm. To handle the continuous attributes, C4.5

splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It also handles missing attribute values. It uses Gain Ratio as an attribute selection measure to build a decision tree. C4.5 removes the biasness of information gain when there are many outcome values of an attribute.

## IV. RESEARCH WORK

Several data mining solutions have been presented for educational data mining. Decision tree classification received significant attention in the area of predicting the student performance. In this section, a schematic overview is given of feature selection, Use full training set which is used for proposed algorithm. It is having only focus with the relevant attributes through feature selection method using Ranker Search.

### *A. Feature Selection*

Many unrelated attributes may be present in the data to be mined in educational data mining. Irrelevant attributes to be removed. Many data mining algorithm don't perform well with large amount of attributes. Therefore feature selection techniques needs to be applied before any kind of mining algorithm is applied. The main objectives of feature selection are to avoid over fitting and improve the performance of classification model [12].

### *B. Ranker*

Ranker Search method rank individual attributes according to their evaluation. Its sorts attribute by their individual evaluation and also perform attribute selection by removing the lower ranker ones [13]. Ranker is more suitable for attribute evaluation methods. Ranker search algorithm gives the best attribute with the highest information gain. The highest information gain attributes can be selected for classification and remaining attributes can be removed before applying classification for better classification.

### *C. Proposed Algorithm*

Here in the proposed system, feature selection is made with ranker search. It is having only focus with the relevant attributes through feature selection- Ranker search method.

**Input:** an attribute- valued dataset *D*
**If** *D* is "pure" OR other stopping criteria met **then**
    terminate
**end if**
Apply Feature Selection **Ranker** Search
**for all**
    attribute $a \in D$ **do**
    Compute information- theoretic criteria if we split on a
**end for**
$a_{best}$= Best attribute according to above computed criteria
Tree= Create a decision node that tests $a_{best}$ in the root
Dv= induced sub- data sets from D based on $a_{best}$
**for all** $D_v$ **do**
    $Tree_v$= C4.5 ($D_v$)
    Attach $Tree_v$ to the corresponding branch of Tree
**end for**
**return** Tree

The process of feature selection reduces the dimensionality of the data and enables learning algorithms to operate faster and more effectively. Using feature selection, classification accuracy was improved.

## V. RESULTS AND DISCUSSION

In this study, those variables whose probability values were greater than 0.60 were considered and highly influencing variables with high probability values have been shown in table II. These features were used for prediction mode construction. For both variable selection and prediction model construction were implemented in a week.

Table II. High potential Variable

| Variable | Description | Probability |
|---|---|---|
| ME | Mothers Education | 0.1889 |
| SH | Students Study Hour | 0.1752 |
| FI | Students Family Income | 0.1238 |
| RE | Teachers Relationship with Student | 0.0996 |
| MED | Medium of Instruction | 0.083 |
| FE | Fathers Education | 0.0711 |
| LS | Students Learning Style | 0.0697 |
| MW | Mothers Working Status | 0.0618 |

From above table, ME and SH have taken high priority over other factors. After both factors, FI has taken high priority then respectively RE, MED, FE, LS and MW have taken next priority.

Decision trees are considered easily understood models because a reasoning process can be given for each conclusion. The knowledge represented by decision tree can be extracted and represented in the form of IF-Then rules in table III.

Table III. Rule Set Generated by Decision Tree

| |
|---|
| **IF** ME= "High" **THEN** RESULT="Excellent". |
| **IF** ME = "Medium"**AND**FI="Medium" **AND** FE="Medium" **THEN** RESULT="Average". |
| **IF** ME = "Low"**AND**FI="Low" **AND** FE="Low" **AND THEN**RESULT="Below Average". |
| **IF** ME = "Medium"**AND**FE="Medium" **AND** FE="High" **THEN** RESULT="Excellent". |
| **IF** SH = "High" **THEN** RESULT="Excellent". |
| **IF** SH = "Less" **AND** RE="Strictly" **THEN** RESULT="Average". |
| **IF** SH = "Never" **AND** RE="Friendly" **AND** LS="TL" **THEN**RESULT="Below-Average". |

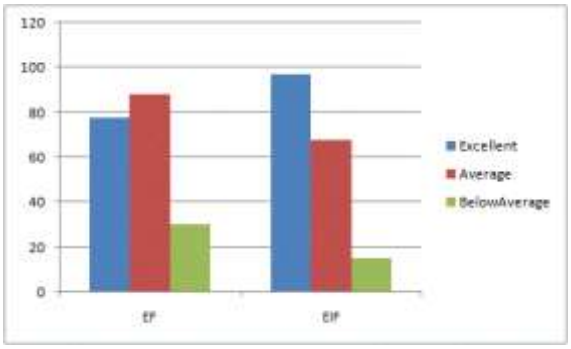


Figure 1. Comparison of EF and EIF factors

From figure 1, the Environmental factor heavily affects students' performance and which makes them to get low performance in their examinations but while considering Educational impacts from institute slightly helps students to achieve Excellent results in their studies.

Finally this study helped tutors to guide the students and parents to give high preference to improve their environmental factors like, ME, FI. Otherwise the students and the parents are supposed to give more preference to educational institute factors like, SH, and RE etc.

If they couldn't give high preference to environmental factors, then the students and the parents are advised to give full preference to Educational institute factors. But if they could give high preference to both factors most of the students will achieve excellent marks in their studies. From the above inferences it is clear that this study has helped the tutors to improve the performance of the students.

## VI. CONCLUSION

The need of prediction over student performance is to help teachers and parents to concentrating their students and children to improvise their performance as well as researcher to select among the decision tree classifier algorithm to find the best classifier for predicting the student performance.The results show thatME

(Mothers Education), SH (Students Study Hour), FI (Family income), FE (Fathers Education), FI (Family Income), MW (Mother Working Status) and RE (Teachers relationship) more affect the student performance. This Thesis work will also help to identify those students are low performers they needed special attention.As conclusion, we have met our objective which has evaluated the performance of students by the two decision tree classification algorithms based on Weka are implemented. This study successfully met accuracy of 83.66% with respected time 0.0 seconds. C4.5 is discovered as the best algorithm for predicting student performance.

## REFERENCES


[1] Romero and Ventura, Data Mining in Education, WIREs Data Mining Knowledge Discovery, 2013.

[2] Romero and Ventura, Educational Data Mining: A Review of the State of the Art, IEEE Transaction on Systems, Man, and Cybernetics, Vol.40, No.6, 2010.

[3] Dr.N.Tajunisha, M.Anjali, Predicting Student Performance Using MapReduce, International Journal of Engineering and Computer Science (IJECS),Vol.4, No.1, 2015.

[4] Birijesh Kumar, Suresh Pal, Mining Educational Data to Analyze Student's Performance, International Journal of Advanced Computer Science and Applications, Vol.2, N0.6, 2011.

[5] BakerRSJd, Yacef K, The state of educational data mining in 2009: A review and future visions, J EduData Min, 2009.

[6] Jiawei Han , Micheline Kamber , Data mining concepts and techniques, 2006.

[7] Chady EI Moucary, Data mining for Engineering Schools, International Journal of Advanced Computer Science and Applications, Vol.2, 2011.

[8] Surjeet Kumar, Data Mining: A Prediction for performance Improvement of Engineering Students Using Classification, World Computer Science and Information Technology Journal (WCSIT), Vol.2, 2012.

[9] As.Galathiya, AP.Ganatra, Classification with an improved Decision Tree Algorithm, International Journal of Computer Application (IJECS), Vol.46, 2012.

[10] B.K. Bharadwaj, S. Pal, Data Mining: A prediction for performance improvement using classification, International Journal of Computer Science and Information Security (IJCSIS), 2011.

[11] Ogor,E.N., Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques, Electronics Robotics and Automotive Mechanics Conference (CERMA), pp.25-28, 2007.

[12] SunitaBeniwal, JitenderArora, Classification and Feature Selection Techniques in Data Mining, International Journal of Engineering Research and Technology (IJERT), Vol.1, No.6, 2012.

[13] LanH.Witten, Eibe Frank, Mark A.Hall, Data Mining Practical Machine Learning Tools and Techniques, 2011.