www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 1 January 2015, Page No. 10068-10071

Mining Educational Data for Predicting Higher Secondary School Student's Grade Using ID3 Algorithm

Nirmala Devi.R[1] *, Deepa.R*[2]*, Kalaiarasi. P* [3]

[1]Assistant Professor, Department of Computer Science
Nandha Arts and Science College
Erode,Tamil Nadu,India
E-mail: nibhunirmala@gmail.com

[2] Assistant Professor, Department of Computer Science
Nandha Arts and Science College
Erode,Tamil Nadu,India
E-mail: mdsupreet@gmail.com

[3] M.Phil Research Scholar, Department of Computer Science
Nandha Arts and Science College
Erdoe , Tamil Nadu, India
E-mail: kalai2p@gmail.com

Abstract: Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. The performance in higher secondary school education in India is a turning point in the academic lives of all students. It is essential to develop predictive data mining model for student's performance so as to identify the slow learners and make necessary steps for the improvement of the students.

In this paper, a new system that will predict students' higher secondary grades based on academic and personal details of the students. ID3 decision tree algorithm was used to train the data of the school students sets. The knowledge represented by decision trees were extracted and presented in the form of IF-THEN rules. A set if prediction rules were extracted from id3 decision tree algorithm and the efficiency of the generated model was found.

Keywords: Data mining, decision trees, id3 algorithm, prediction rules, if-then rules.

I. INTRODUCTION

Students, academic performance depends on diverse factors like personal, socio-economic, psychological and other environmental variables. The prediction of student performance with high accuracy is beneficial to identify the students with low academic achievements initially. This will help the educators to improve the performance of the students in future.

As data mining models have relatively higher degree of accuracy, we use such tools to develop predictive data mining model for student's performance in Indian educational system.

School education in Indian is divided into primary, secondary and higher secondary. The two – year education, which is known as Higher Secondary Education is important because it is deciding factor for opting desired subjects of study in higher education. In this connection, the objectives of the present investigation were fraMED so as to assist the low achievers in higher secondary level.

The main objective of this paper is to use data mining methodologies to study academic and personal details of the students that act as the dominant factors on their academic performance for the better performance in higher secondary level.

Data mining provides many tasks that could be used to predict the student's performance. In this research, the id3 decision tree algorithm is used.

II. RELATED WORKS

Although, using data mining in school education is a research field, there are many works in this area. That is because of its potentials to educational institutes.

Khan[1] conducted a performance study on 400 students comprising 200 boys and 200 girls selected from the senior secondary school of Aligarh Muslim University, Aligarh, India with a main objective to establish the prognostic value of different measures of cognition, personality and demographic variables for success at higher secondary level in science stream. The selection was based on cluster sampling technique in which the entire population of interest was divided into groups, or clusters and a random sample of these clusters was selected for further analyses. It was found that girls with high socio-economic status had relatively higher academic achievement in science stream and boys with low socioeconomic status had relatively higher academic achievement in general.

Walters and Soyibo [2] conducted a study to determine Jamaican high school students (population n = 305) level of performance on five integrated science process skills with performance linked to gender, grade level, school location, school type, and socio-economic background(SEB). The results revealed that there was a positive significant relationship between academic performance of the student and the nature of the school.

Cortez and Silva [3] attempted to predict failure in the two core classes (Mathematics and Portuguese) of two secondary school students from the Alentejo region of Portugal by utilizing 29 predictive variables. Four data mining algorithms were applied on a data set of 788 students, who appeared in 2006 examination. It was reported that DT and NN algorithms had the predictive accuracy of 93% and 91% for two-class dataset (pass/fail) respectively. It was reported that both DT and NN algorithms had the predictive accuracy of 72% for a four-class dataset.

Moriana et al. [4] studied the possible influence of extra- curricular activities like study-related (tutoring or private classes, computers) and /or sports-related (indoor and outdoor games) on the academic performance of the secondary school students in Spain. A total number of 222 students from 12 different schools were the samples and they were categorized into two groups as a function of student activities (both sports and academic) outside the school day. Analysis of variance (ANOVA) was used to verify the effect of extracurricular actives on the academic performance and it was observed that group involved in activities outside the school yielded better academic performance.

Hijazi and Naqvi [5] conducted as study on the student performance by selecting a sample of 300 students (225 males, 75 females) from a group of college affiliated to Punjab university of Pakistan. The hypothesis that was stated as "Student's attitude towards attendance in class , hours spent in study on daily basis after college, student's family income , students' mother's age and mother's education are significantly related with student performance" was fraMED. By means of simple linear regression analysis, it was found that the factors like mother's education and student's family income were highly correlated with the student academic performance.

From these specific studies, we observe that the student performance could depend on diversified factors such as demographic, academic, psychological, socio-economic and other environmental factors. Based on these observations, we constructed an ID3 prediction model with 7-class response variables obtained through feature selection techniques so as to evaluate the academic achievement of students at higher secondary level in India.

III. PROPOSED WORK

Data mining provides many tasks that could be used to predict the student's performance. In the present investigation, the secondary and primary data has to be collected from the schools. An ID3 decision tree algorithm is applied on data that are collected from various schools to predict the student's performance in the higher secondary examination. It helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising / counseling.

A. The Students Data Set and Preprocessing

The data set used in this paper contains students information collected from various schools. During preprocessing of data set, the irrelevant attributes or variables should be removed to get better input data for data mining

Table I represents the variables after preprocessing the dataset that were collected from source database.

Table I

ATTRIBUTE	DESCRIPTION	POSSIBLE VALUES
SEX	STUDENT'S SEX	{MALE ,FEMALE}
FAM-Size	STUDENT'S FAMILY SIZE	{ONE , TWO, THREE, FOUR, FIVE, MORE THAN FIVE}
LOC-AREA	STUDENT'S LIVING AREA	{CORPORATION, MUNICIPAL, RURAL}
STUME	TYPE OF SECONDARY SYLLABUS	{STATE BOARD , MATRIC}
XMARK-GRADE	GRADE OBTAINED AT SECONDARY LEVEL	{O – 90% TO 100% , A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49% , F - < 40%}
MED	MEDIUM OF INSTRUCTION	{TAMIL, ENGLISH}
LOC-SCH	LOCATION OF SCHOOL	{CORPORATION, MUNICIPAL, RURAL}
MEDU	MOTHER'S EDUCATION	{NO EDUCATION, PRIMARY, SECONDARY, GRADUATE, POST-GRADUATE}
FEDU	FATHER'S EDUCATION	{NO EDUCATION, PRIMARY, SECONDARY, GRADUATE, POST-GRADUATE}
FSAL	FATHER'S MONTHLY INCOME	{NO EARNINGS OR LESS THAN 5000,5000-10000,

		10,000-25,000, 25,000- 50,000, > 50,000}
MSAL	MOTHER'S MONTHLY INCOME	{NO EARNINGS OR LESS THAN 5000,5000-10000, 10,000-25,000, 25,000- 50,000, > 50,000}
(RESPONSE VARIABLE) HSCGRADE	GRADE IN HSC LEVEL	{O – 90% TO 100% , A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49% , F - < 40%}

Domain values for some of the high potential variables were defined for the present investigation as follows

STUME- STATE BOARD / MATRIC BOARD: Two types of secondary education is offered in India through state board and matric systems Students who has matric pattern syllabus up to their secondary may perform well in the higher secondary examination. Therefore this factor may influence the student performance at higher secondary level. Possible values are state and matric.

MED: Possible values are Tamil or English.

LOC-SCH: Exposure of the students varies according to the location of the school. Possible values are rural, municipal and corporation.

LOC-AREA: Students living area also may influence the student performance at higher secondary level. Possible values are corporation, municipal and rural.

MED: Mothers education also influences the student performance. Possible values are no-education, elementary, secondary, graduate, not – applicable.

XMARK-GRADE: Marks obtained at secondary level. Possible values are : O – 90% to 100% , A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49% , F - < 40%

FEDU: Fathers education. Possible values are no-education, elementary, secondary, graduate, not – applicable.

FSAL : Fathers salary. Possible values are no earnings or less than 5000, 5000-10000, 10,000-25,000, 25,000- 50,000, > 50,000.

HSCGRADE : Grade obtained at higher secondary level and it is declared as response variable. It is split into seven class values: O – 90% to 100%, A – 80% - 89%, B – 70% - 79%, C – 60% - 69%, D – 50% - 59%, E – 40% - 49%, F - < 40%

IV. TOOLS AND TECHNIQUE

A. *Decision Tree*

A decision tree is a tree-shaped structure that represents sets of decisions. These decisions generate rules for the classification of a dataset. A decision tree 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.

In this paper, the decision tree approaches are used to predict the grade of the higher secondary student and there are 7 grades (O, A, B, C, D, E and F) .The four widely used decision tree learning algorithms are: CART, CHAID, ID3 and C4.5.

B. *ID3 Algorithm*

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan. 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. In order to select the attribute that is most useful for classifying a given sets, we introduce a metric – information gain.

The present investigation used data mining as a tool with ID3 algorithm as a technique to design the student performance prediction model.

To find an optimal way to classify a learning set, what we need to do is to minimize the questions asked (i.e. minimizing the depth of the tree). Thus, we need some function which can measure which questions provide the most balanced splitting. The information gain metric is such a function used in this research.

V. RESULT AND DISCUSSIONS

A. *Rules set generated by Decision Tree*

The knowledge represented by decision tree can be extracted and represented in the form of IF-THEN rules.

Rules for HSCGrade = 'A'

If MED = 'English' and XMARK-GRADE = 'A' 0r 'O' and LOC-SCH = 'rural' and XMARK-GRADE = 'A' THEN HSCGrade = 'A'

If MED = 'English' and XMARK-GRADE = 'A' 0r 'O' and LOC-SCH = 'rural' and XMARK-GRADE = 'O' THEN HSCGrade = 'A'

If MED = 'English' and XMARK-GRADE = 'A' 0r 'O' and LOC-SCH = 'municipal' and XMARK-GRADE = 'O' THEN HSCGrade = 'A'

Rules for HSCGrade = 'B'

If MED = 'Tamil' and LOC-AREA = 'municipal' or LOC-AREA = 'rural' then HSCGrade = 'B'

If MED = 'Tamil' and LOC-AREA = 'corporation' and XMARK-GRADE = 'A' or 'O' and MSAL = 'not alive' or MSAL = 'above 5000 and less than 10000' THEN HSCGrade = 'B'

If MED = 'English' and XMARK-GRADE = 'B' or 'C' and STUME = 'matric' THEN HSCGrade = 'B'

If MED = ‘English’ and XMARK-GRADE = ‘A’ or ‘O’ and LOC-SCH = ‘municipal’ and XMARK-GRADE = ‘A’ THEN HSCGrade = ‘B’

Rules for HSCGrade = ‘C’

If MED = ‘Tamil’ and LOC-AREA = ‘corporation’ and XMARK-GRADE = ‘B’ or ‘THEN HSCGrade = ‘C’.

If MED = ‘Tamil’ and LOC-AREA = ‘corporation’ and XMARK-GRADE = ‘C’ or ‘E’ THEN HSCGrade = ‘C’.

If MED = ‘English’ and XMARK-GRADE = ‘B’ or ‘C’ and STUME = ‘state’ THEN HSCGrade = ‘C’.

Subsequently, the 10-fold cross method for the validation of the model was applied during ID3 prediction model construction process.

From the rule set for HSCGRADE = ‘A’, it was found that medium of instruction and previous academic achievement at secondary level had influence on academic achievement in higher secondary level. The English medium students maintained their academic performance both at secondary and higher secondary level despite of location of school. The information from the rule set for HSCGRADE = ‘C’ showed that the performance of the students with C grade in secondary level did not show any improvement in the higher secondary level. The rule set given above focuses only on the three classes namely A,B and C, since the number of objects in the rest of the classes was less, they were generated in the rule set.

The classification matrix has been presented in Table II, which compared the actual and predicted classifications.

Table II

HSCGRADE		% OF PREDICTION
OBSERVED	O	22.33
	A	56.60
	B	66.7
	C	64
	D	37.5
	E	0.00
	F	0.00

It was found that the overall model prediction accuracy of ID3 model was 64.67%. The accuracy of the present model was compared with other models and it was found to be higher than the accuracy of earlier model.

VI. CONCLUSION

Data mining is gaining its popularity in almost all applications of real world. One of the data mining technique i.e. classification is an interesting topic to the researcher as it is accurately and efficiently classifies the data for knowledge discovery. Decision trees are so popular because they produce classification rules that are easy to interpret than other classification methods. In decision trees, the ID3 prediction model was useful to analyze the interrelation between variables that are used to predict the outcome on the performance at higher secondary school education. This study is also helpful for those students who need special attention and will also lower failure ratio by taking proper action for the higher secondary education.

REFERENCES

[1] Z.N.Khan, *“Scholastic Achievement of Higher Secondary Students in Science Stream”,* Vol.1, No.2, 2005, Journal of Social Sciences, pp.84-87.

[2] Y.B.Walters and K.Soyibo, *“An Analysis of High School Student’s Performance on Five Integrated Science Process Skills”,* Vol.19, no.2, 2001, Research in Science and Technical Education, pp.133-145.

[3] 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.

[4] J.A. Moriana, F.Alos, R.Alcala, M.J.Pino, J.Herruzo, and R.Ruiz , *“Extra Curricular Activities and Academic Performance in Secondary Students”*, Vol.4, No.1, 2006, Electronic Journal of Research in Educational Psychology, pp.35-46.

[5] S.T.Hijazi and R.S.M.M.Naqvi, *“Factors Affecting Student’s Performance: A Case of Private Colleges”*, Vol.3, No.1, 2006, Bangladesh e-Journal of Sociology, pp.90-100.