"Using the Classification Method of Web Mining to Decrease the Dropout Rate in Educational Institute."

Rakesh Verma[#1], B. Dhanasekaran[*2]
Research Scholar, Bhagwant University, Ajmer, Rajasthan, India
rakeshverma@ipsacademy.org
Professor, A.K.T. Memorial College of Engg.&Tech.Kallakurichi,Tamilnadu,India
drbdhanasekaran@gmail.com

Abstract: **Improve the quality of educational process is most challenging part in dynamic era to reduce the dropout rate and monitor their overall growth in colleges and universities. Educational web mining is simply refers to digging out the required information from the large educational database which most of the institutions and courses opened in colleges are in self finance mode, so all time they focused to fill all the seats of the courses not on the quality of students. therefore a large number of students drop the course after first year is formed by the learning activities of students in institutions. The model use the C4.5(J48) classification technique and acquire a wide knowledge to find out the reasons of dropout from college. The accuracy of the work 82.5% by C4.5(J48). In this study in education system to reduce the dropout rate by applying web mining method decision tree to predicted by college through past performance of the students. A few derived variables (Grade of Higher School, Higher secondary school, Family annual income, Student branch, Student category. Living location of students) were selected. While some of the information for the variables was extracted from the database.**

Keywords: Educational Data Mining, Classification decision tree, C4.5(J48).Dropout management, web mining.

I Introduction

Web mining that can be applied on data related to the field of education this is emerging technique of web mining. This new emerging field, called educational web Mining, concerns with developing methods that discover knowledge from data originating from educational environments. Educational Web Mining uses many techniques such as Decision Trees, Neural Networks, Naïve Bayes, K- Nearest neighbor, and many others. Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for prediction.

One of the biggest challenge that higher education faces today is predicting the academic paths of student. Many higher education systems are unable detecting student population who are likely to drop out because of lack of intelligence methods. It remains a challenging task to accurately predict a currently enrolled student's likelihood of returning to colleges the next term develop learning and teaching initiatives to improve retention and progression in education process is an important academic concern, which mainly depends on monitoring student performance and exploiting student feedback. Student marks and achievement are the main sources to study student feedback and progress, yet university and educational centers can use to predict the student performance, student dropout, and study path. The main objective of this study is to identify those students who take dropout from college in first year. Early identification of these students is enough for the institution to accommodate its interventions and marketing strategies will greatly enhance the student persistence rate in specific majors. This paper focuses on predict reasons of dropout to use C4.5(J48) classification technique and acquire a wide knowledge to find out the reasons of dropout from colleges.

C4.5(J48)

This algorithm is a successor to ID3 developed by Quinlan Ross [2]. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle 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. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5(J48) uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification

II Proposed work

Information produced by web mining techniques can be represented in many different ways. In this paper. I have used the classification data mining technique to extract the important attribute like grade of Higher School, Higher secondary school, Family annual income, Student branch, Student category. Living location of students, Family annual income Status, father's qualification, mother's qualification, Father's occupation, Mother's occupation, Medium of Teaching that stored in a database to analyze reasons affecting the dropout of students in various college of higher education. The classifier algorithm J48 is used for predicting the main reasons of dropout from college.

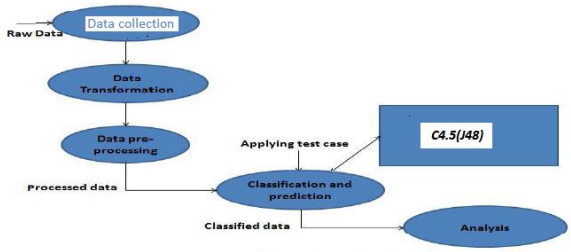

Figure 1:High level data flow diagram

III Collection of data

Data may be collected from many different and heterogeneous data sources. This stage comprises of collection all available information of students.. The data set used in this work was obtained from department School of computers IPS academy Indore of course integrated MCA (session 2015-2019) were taken for the experiment in MS Excel and converted into arff (Attribute-Relation File Format). This file was given as input to WEKA 3.9 tool to obtain results. Where WEKA stands for Waikato Environment for Knowledge Analysis is a popular suite of machine
learning software, developed at the University Of Waikato, New Zealand. The implementation of the dataset is done using a data mining tool WEKA. WEKA is open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications Initial size of the data was 40. Later on the data set was increased to 90. Most students were between the ages of 20 and 21, as this is the year when most of the students experience a new environment and infrastructure of study.

IV Selection of Attributes

Variable	Description	Possible Values
S.N	Serial Number	From 1 to 40 students
Branch	Students Branch	IMCA, MCA
Cat	Students category	{Unreserved, OBC, SC, ST}
PrevSemResult	Previous Semester Result	O:90-100% A:75-89% B:60-74% C:50-59% D:45-49% E:40-44% F: Less than 40%
SSC	Result in 10th	O :85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
HSC	Result in 12th	Distinction:85-100% First:60-84% Second:45-59% Third:35-44% Fail: Less than 35%
LLoc	Living Location of Student	{Village, Town, Tahseel, District}
FAIn	Family annual income status	(BPL, poor, medium, high}
FOcc	Father's Occupation	{Service, Business, Agriculture, Retired, NA}
Dropout	Dropout: Continue to enroll or not after one year	{Yes, No}

TABLE 1: List Of Attributes

V Result and Conclusion

This model is constructed by the C4.5(J48) classification method which provides the decision tree where each internal node is denoted by rectangles, and leaf nodes are denoted by ovals. All internal nodes have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. In this study in education system to reduce the dropout rate by applying web mining method decision tree to predicted by college through past performance of the students.

S. N	Father's Occupation	Ist Sem Mark/ Grade	SSC Mark/ Grade	HSC Mark/ Grade	Drop Out
1	Service	B	B	A	NO
2	Service	B	B	C	NO
3	Business	B	C	B	NO
4	Retired	B	B	E	YES
5	Retired	D	E	D	YES
6	Business	B	C	B	NO
7	Service	B	B	A	NO
8	Business	D	E	D	YES
9	Business	C	A	B	NO
10	Retired	D	D	E	YES
11	Retired	B	B	A	NO
12	Business	D	E	D	NO
13	Service	B	B	B	NO
14	Business	D	E	D	YES

Table:1 Dataset

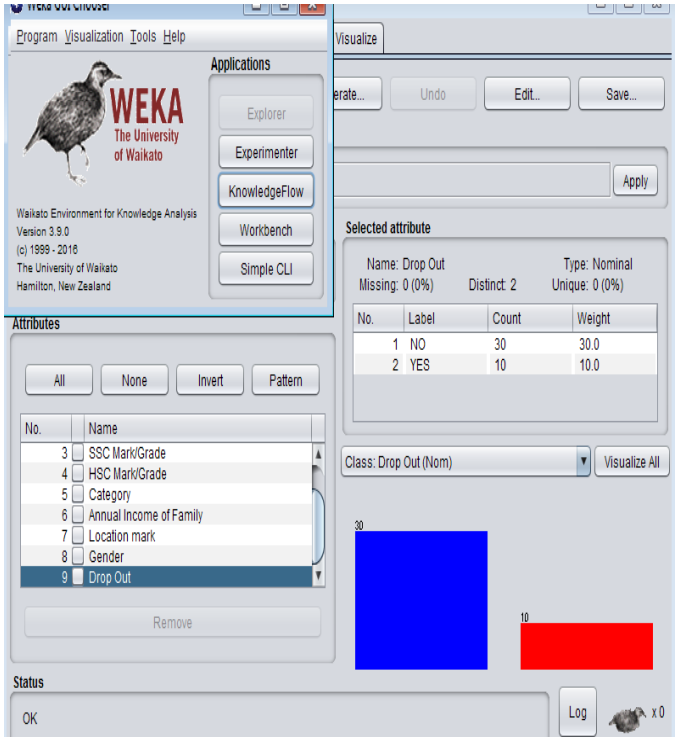

Figure :1

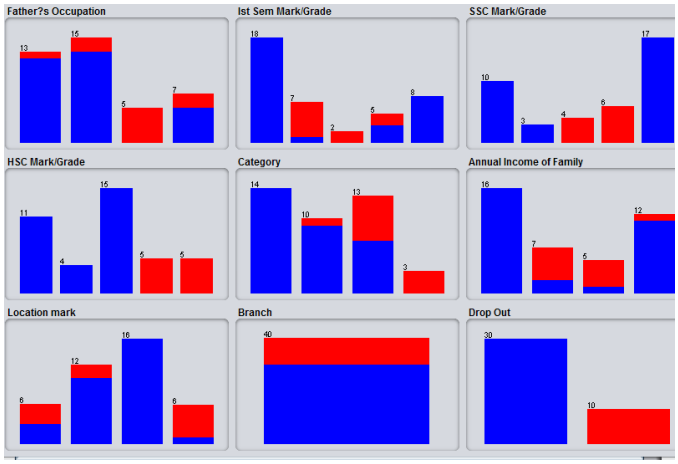

Figure2 – Histograms patterns of the Students Categorization

Figure:3

```
=== Confusion Matrix ===

 a  b   <-- classified as
30  0 |  a = NO
 0 10 |  b = YES
```

A few derived variables (Grade of Higher School, Higher secondary school, Family annual income, Student branch, Student category. Living location of students, Father's Occupation,) were selected. Here accuracy of result dropout rate with high school grade NO represent the 75% and YES represent the 25% applying by WEKA 3.9 in figure:1 and also given the result by confusion matrix in figure:3.

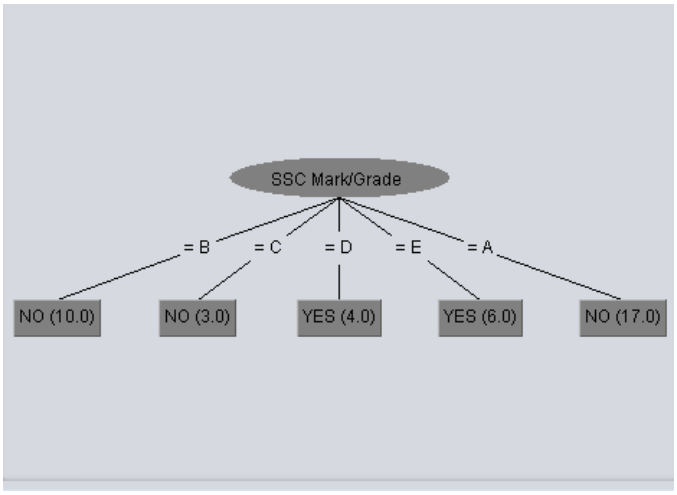

Figure: 4 A decision tree generated by J48 algorithm with SSC grade AB and C to represent NO and DE to represent YES to chance of dropout from the college.

The decision tree represent those student have grade D and E there is more chance to dropout the students from the college so through different internal assessment results and basis of these results teacher will guide to weak students and uplift them from downfall in advance. Same result obtains in HSS and first semester grade.

Accyracy Dropout

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          33               82.5    %
Incorrectly Classified Instances         7               17.5    %
Kappa statistic                          0.7544
Mean absolute error                      0.1208
Root mean squared error                  0.2583
Relative absolute error                 39.93   %
Root relative squared error             66.3848 %
Total Number of Instances               40
```

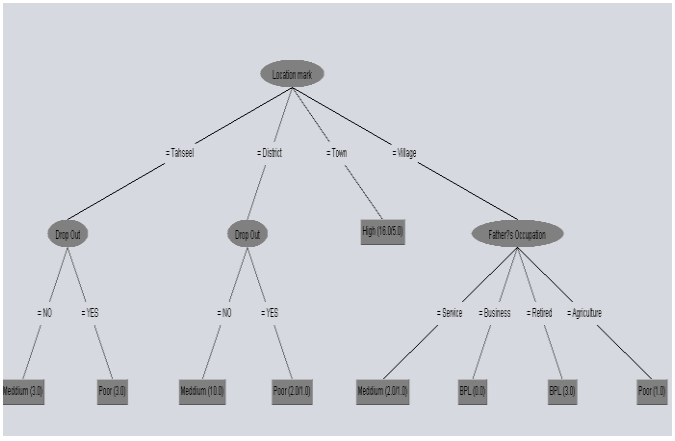

Figure: 5 A decision tree generated by J48 algorithm by family annual income in different category

An annual income of the students also play a important role in education system those student's family have minimum income and they are poor and not afford to pay fee in upcoming semester so we have to take special care to prevent drop out from the college

VI Conclusion

Those student have the grade D and E category the college have to pay more attention to all reach in ABC category .The increase the dropout rate of the students of any institution has become an alarming concern for the management. An early detection of students with risk would help the management to uplift them from downfall at the initial stage itself. Data mining techniques like classification be used to develop a support system to help authorities identify the dropout students in advance and take timely measures to curb the extremes. After the detailed study about the classification algorithms C4.5(J48) applied in this data set. Such specification could be useful in the educational system like Universities and Colleges for maintaining the overall quality of the education. By doing this we can know the academic status of the students in advance and can concentrate on reducing the dropout rate.

VII Future Scope

The classification model should also include the non-academic loss of student status, or a new model could be built to react to this specific situation. Further enhancements and various alternatives could be used to consider a larger number of parameters which will guide us to predict more in-depth information about the student under consideration. Predicting the academic outcome of a student needs a lot of parameters should be considered like Schedule and questionnaire format of personal interview to find out the reasons of dropout from college. In order to classify and predict dropout students, means Prediction models should include all personal, social, psychological and other environmental variables are necessitated for the effective prediction of the dropout of the students. Data from the admissions process are merged with the academic information that is collected from each academic period of a student; however, the reasons of low academic credential occur on a daily basis and waiting until the next academic session ends, could be crucial step. This forces us to think the new and possibly non-traditional ways, for collecting information close to real time are needed. This work focused on the loss of academic status due to low performance. The work could be extended in such a way where the co-curricular activity of the student could also be mined to find out the best we have in our institution.

References

[1] *Z. J. Kovačić, "Early Prediction of Student Success:Mining Students Enrollment Data," pp. 647-665, 2010.*

[2] *S. M. Patil and D. P. Kumar, "Data Mining Model for Effective Data Analysis of Higher Education Students Using MapReduce," International Journal of Emerging Research in Management &Technology, vol. 6, no. 4, pp. 177-183, 2017.*

[3] *K. B. Bhegade and S. V. Shinde, "Student Performance Prediction System with Educational Data Mining," International Journal of Computer Applications, vol. 146, no. 5, pp. 32-35, 2016.*

[4] *A. M. Shahiri, W. Husain and N. A. Rashid, "A Review on Predicting Student's Performance using Data Mining Techniques," Procedia Computer Science, vol. 72, pp. 414-422, 2015.*
G. S. Abu-Oda and A. M. El-Halees, "Data Mining in Higher Education : University Student Dropout Case Study," International Journal of Data Mining & Knowledge Management Process(IJDKP), vol. 5, no. 1, pp. 97-106, 2015.

[5] *S. Sultana, S. Khan and M. A. Abbas, "Predicting Performance of Electrical Engineering Students using Cognitive and Non-Cognitive Features for Identification of Potential Dropouts," International Journal of Electrical Engineering Education, vol. 54, no. 2, pp. 105-118, 2017*

[6] *L. Bonaldo and L. N. Pereira, "Dropout: Demographic profile of Brazilian university students," Procedia - Social and Behavioral Sciences, vol. 228, pp. 138-143, 2016.*

[7] *Kemenristekdikti, Statistik Pendidikan Tinggi Tahun 2017, Jakarta: Pusdatin Iptek Dikti, 2017.*

[8] *A. Utomo, A. Reimondos, I. Utomo, P. McDonald and T. H. Hull, "What happens after you drop out ? Transition to adulthood among early school-leavers in urban Indonesia," Demographic Research, vol. 30, pp. 1189-1218, 2014.*

[9] *T. Fahrudin, J. L. Buliali and C. Fatichah, "Predictive modeling of the first year evaluation based on demographics data: Case study students of Telkom University, Indonesia," International Conference on Data and Software Engineering (ICoDSE), pp. 1-6, 2016.*

[10] *A. K. Jain and C. K. Jha, "Dropout Classification through Discriminant Function Analysis: A Statistical Approach," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 2, no. 4, pp. 572-577, 2017.*

[11] *A. Katare and S. Dubey, "A Comparative Study of Classification Algorithms in EDM using 2 Level Classification for Predicting Student's Performance," International Journal of Computer Applications, vol. 165, no. 9, pp. 35-40, 2017.*

[12] *C. Márquez-Vera, A. Cano, C. Romero, A. Y. M. Noaman, H. M. Fardoun and S. Ventura, "Early Dropout Prediction using Data Mining: A Case Study with High School Students," Expert Systems Journal, vol. 33, no. 1, pp. 107-124, 2016.*

[13] *A. Cano, A. Zafra and S. Ventura, "An interpretable classification rule mining algorithm," Information Sciences, vol. 240, pp. 1-20, 2013.*

[14] *E. Osmanbegovic, M. Suljic and H. Agic, "Determining Dominant Factor for Students Performance Prediction by Using Data Mining Classification Algorithms," Tranzicija, vol. 34, no. 34, pp. 147-158, 2014.*